A Machine Learning Enabled Data Quality Assessment Framework for Connected Vehicles Data

M.N. Wondie

Doctor of Data Science

2023

University of East London

A Machine Learning Enabled Data Quality Assessment Framework for Connected Vehicles Data

M.N. Wondie

A thesis submitted in partial fulfilment of the requirements of the University of East London for the degree of Doctor of Data Science

2023

University of East London

“More data beats clever algorithms, but better data beats more data.”— Peter Norvig

Abstract

Connected vehicles leverage innovations in sensors, IoT, cloud computing, AI, and 4G/5G to produce real-time vehicle data, enhancing applications in navigation, fleet management, diagnostics, and maintenance; improving cost-efficiency, revenue, customer satisfaction, and safety.

However, maintaining data quality in connected vehicles is challenging. Classical data quality assessment frameworks are inadequate for the complexity of connected vehicles, necessitating improved methods for assessing data quality in this domain.

This research integrates machine learning and statistical methods with classical frameworks to enhance data quality assessment. A literature review identifies data quality challenges, existing frameworks, their strengths and limitations. Implementing a classical framework with real-world connected vehicle data uncovers issues like missing, delayed, and invalid data but fails to answer some data quality requirements, which are identified as gaps leading to the development of three scenarios to leverage machine learning. Scenario I uses logistic regression to detect non-communicating vehicles addressing delayed and missing data issues. Scenario II forecasts missing mileage using a time series method. Scenario III assesses data accuracy using Light Gradient-Boosting Machine and Random Forest.

The implementation of these scenarios provided promising results. Scenario I detects non-communicating vehicles with F1-score of 0.85. Scenario II forecasts missing mileage with lower RMSE compared to state-of-the-art methods. Scenario III detects inaccurate fuel consumption with 97% accuracy and F1-score of 0.78, outperforming Isolation Forest.

In conclusion, implementing a classical data quality assessment framework with real-life vehicle data highlights various data quality issues and reveals certain limitations. Machine learning and statistical methods help to address these limitations. Therefore, a new framework that integrates classical data quality assessment with machine learning for connected vehicles data is proposed.

Keywords: *Data Quality, Assessment, Connected Vehicles, Machine Learning, Quality Control Chart*

Declaration

I hereby declare that the thesis entitled "A Machine Learning Enabled Data Quality Assessment Framework for Connected Vehicles Data" submitted by me, for the award of the degree of Doctor of Data Science to the University of East London is a work carried out by me under the supervision of Dr. Yang Li and Dr. Julie Wall, School of Architecture, Computing and Engineering, University of East London, London, UK.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this university or any other university or institute.

Name: Mulluken Nega Wondie

Date: September 22, 2023

Signature:

Mulluken Wondie

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACF - Autocorrelation function

AI - Artificial Intelligence

ANNs - Artificial neural networks

AR - Autoregressive

ARCH - Autoregressive Conditional Heteroskedasticity

ARFIMA - Autoregressive Fractionally Integrated Moving Average

ARIMA - Autoregressive Integrated Moving Average

ARIMA - Auto-Regressive Integrated Moving Average

ARMA - Autoregressive Moving Average

AWS - Amazon Web Services

CAN - Controller Area Network

CI/CD - Continuous Integration and Continuous Delivery

CL - Central line for the average

CV - Connected Vehicles

DBSCAN - Density-based spatial clustering of applications with noise

DQ - Data Quality

DQA - Data Quality Assessment

DSR - Design Science Research

DSRC - Dedicated Short-Range Communications

ETS - Error, Trend, Seasonality

GDPR - General Data Protection Regulation

GPS - Global Positioning System

GSM - Global System for Mobile Telecommunications

IaaS - Infrastructure as a service

IoT - Internet of Things

IQ - Information Quality

LCL - Lower line for the lower control limit

LightGBM - Light Gradient-Boosting Machine

MA - Moving Average

ML - Machine Learning

MLE - Maximum likelihood estimation

PaaS - Platform as a service

PACF - Partial autocorrelation function

RF - Random Forest

SaaS - Software as a service

SQL – Structured Query Language

SVM - Support vector machine

SWOC - Strength, Weakness, Challenge, Opportunity

UCL - Upper line for the upper control limit

V2I - Vehicle to Infrastructure

V2V - Vehicle to Vehicle

V2X - Vehicle to Everything

Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Dr. Yang Li and Dr. Julie Wall, for their unwavering guidance, invaluable insights, and support throughout my doctoral journey. Their expertise and mentorship have been instrumental in shaping the trajectory of my research, and I am profoundly grateful for their commitment. I extend a special appreciation to Dr. Julie Wall for her meticulous attention to detail in carefully reading my works and providing me with insightful and constructive criticism. Her dedication to the refinement of my research has been instrumental in shaping the quality of this thesis.

I also wish to express my gratitude to Professor Allan Brimicombe, my former supervisor, for his invaluable assistance in preparing the proposal for this research. His expertise and mentorship have been invaluable in laying the foundation for this project.

I am indebted to DAF Trucks N.V. for the generous financial assistance and for giving me the opportunity to work on this transformative project. The support not only made this research possible but also enriched my academic experience by providing real-world context and resources.

Last but not least, I extend my deepest appreciation to my family for their support, love, and encouragement. I am profoundly grateful for their sacrifices and encouragement.

Mulluken Nega Wondie
September 22, 2023
Eindhoven, The Netherlands

Dedication

I dedicate this work to the helpless mothers, children, and all those who suffer due to the toxic impact of ethnic politics in my home country, Ethiopia.

Chapter 1 : Introduction

1.1 Motivations and Challenges

The adoption of vehicle connectivity and fleet telematics is on the rise. According to some estimates, about 2 billion vehicles are expected to be connected by the end of 2025 (Mostefaoui *et al.*, 2022). McKinsey & Company also estimates that 95% of the global vehicles on the market in 2030 will be connected (Abdelkader, Elgazzar and Khamis, 2021). This increased connectivity presents new opportunities for organizations to leverage the vast amount of data generated to improve safety, reduce cost, and increase profitability. Analyzing this data allows organizations to identify meaningful patterns related to vehicle performance, vehicle health, and driver behaviour. By leveraging these insights, organizations can improve the quality, reliability, and safety of their vehicles. This, in turn, provides a competitive advantage and paves the way for new business opportunities, such as location-based services, value added services, and cost savings (Brimicombe and Li, 2009). Furthermore, the analysis of the data from connected services contributes to better traffic management and CO_2 reduction efforts (He *et al.*, 2019; Megler, Tufte and Maier, 2016). By understanding traffic patterns and optimizing routes, organizations can improve traffic flow and reduce emissions, making transportation more efficient and environmentally friendly. Currently, there are various applications of connected vehicles (CV), which can be categorized into four main groups, as described by Siegel, Erb, and Sarma (2017):

- Information Services: These applications focus on providing information and communication services to vehicle users. Examples include remote vehicle dashboards, diagnostics, fault prediction, data collection for decision-making, digital mapping, and communication. The goal is to enhance user comfort, enable remote monitoring, and optimize driving behaviour.
- Safety Services: These applications utilize connectivity and Advanced Driver Assistance Systems (ADAS) to improve safety. They encompass collision avoidance, hazard reporting, and driver monitoring. By sharing real-time data, CV can reduce accidents, enhance fleet-wide safety, and monitor driver impairment.
- Individual Motion Control: They utilize connectivity to either provide warnings to technicians or directly manipulate the actuators of an individual vehicle. Some instances encompass collision prevention, aided lane changing, and route optimization. The primary objective is to improve the level of individual vehicle control and safety.

- Group Motion Control: This category uses vehicle sensors and outside information to impact or regulate the actions of vehicles and drivers. Applications such as platooning entail the coordinated movement of vehicles in proximity, with the aim of optimizing fuel consumption, alleviating traffic congestion, and enhancing the efficiency of traffic flow. Intersection control utilizes connected data to optimize the flow and effectiveness of vehicles, leading to a decrease in delays, trip duration, CO_2 releases, and fuel usage.

Overall, the applications of connectivity of vehicles aim to improve user experience, enhance safety, optimize driving behaviour, and promote more efficient and eco-friendly transportation practices (Abdelkader, Elgazzar and Khamis, 2021; Jadaan, Zeater and Abukhalil, 2017; McQueen, 2017).

As a result, the automotive industry is experiencing a significant transformation as global automakers redirect substantial capital and resources toward the advancement of connectivity of vehicles (Yang *et al.*, 2018). This shift goes beyond mere changes in individual vehicle components; it encompasses the evolution of entire business models, transitioning from traditional ownership to service-based models (Abdelkader, Elgazzar and Khamis, 2021). To embrace this shift and explore different viewpoints, some automakers and suppliers have taken proactive measures, establishing new capital structures and organizations to adapt to the changing landscape (Jadaan, Zeater and Abukhalil, 2017). The aim is to stay competitive and innovative in an era of rapid technological advancements and shifting consumer preferences.

However, the potential of CV data is hindered by data quality (DQ) challenges (He *et al.*, 2019; Ricardo Perez-Castillo *et al.*, 2018). Telematics data, which includes multiple data sources like Controller Area Network (CAN) Bus, Global Positioning System (GPS), and embedded units, is susceptible to errors due to spatio-temporal variation and communication means like global system for mobile telecommunications (GSM). Back-end storage and processing also contribute to potential DQ issues. Poor DQ in CV can lead to financial losses, negative societal impacts, and increased operational costs (Strong, Lee and Wang, 1997). These effects include liabilities, risks, costs of regaining dissatisfied customers, lost shareholder value, and the need for information rework or cleansing (Batini *et al.*, 2009).

Poor DQ in CV can be categorized into real DQ issues such as missing data, implausible data, inaccurate data and soon, and uncertainty arising from a lack of knowledge and transparency regarding the state of the data. Addressing these challenges is crucial for the wider adoption of

connectivity for critical business purposes, and it requires further research on DQ issues in CV (Hamad, 2015).

However, research on DQ in the context of CV is limited compared to other areas, partly due to it being an emerging field (Juddoo *et al.*, 2018; Ricardo Perez-Castillo *et al.*, 2018)). Meanwhile, ensuring DQ in CV is of utmost importance, given its usage in critical applications. As CV plays a significant role in safety, efficiency, and decision-making processes, maintaining high DQ becomes crucial to ensure accurate and reliable operations, and to avoid potential risks associated with poor DQ. Further research and focus on DQ in CV is necessary to harness its full potential while ensuring safety and efficiency in this rapidly evolving technology landscape.

The literature shows that there are several general purpose DQ assessment frameworks. However, they cannot fully assess difficult DQ metrics such as accuracy. It is stated that the CV ecosystem is complex (Siegel, Erb and Sarma, 2017). There are also some methodologies leveraging advanced methods such as machine learning (ML) which are developed to handle more complex DQ issues in modern systems. However, they lack generalizability, and they only handle specific DQ topics, most of which involve outlier detection.

This shows that there is a need to develop a framework that can address the DQ assessment challenges in CV to increase confidence of users to develop data services. This was one of the main reasons that provided motivation for this research work.

The second motivation for this work is the researcher's background. The researcher works as a Data Scientist for DAF Trucks N.V., which is one of the biggest truck manufacturers in the world, in the Global Connected Services (GCS) department. GCS develops and manages connected solutions, and the researcher works with the connected data to develop services. In his day-to-day activity, the researcher faces different DQ issues.

The focus of this study is, therefore, the intersection of **Data Quality Assessment Frameworks, Connected Vehicles and Machine Learning** as shown in Figure 1-1.

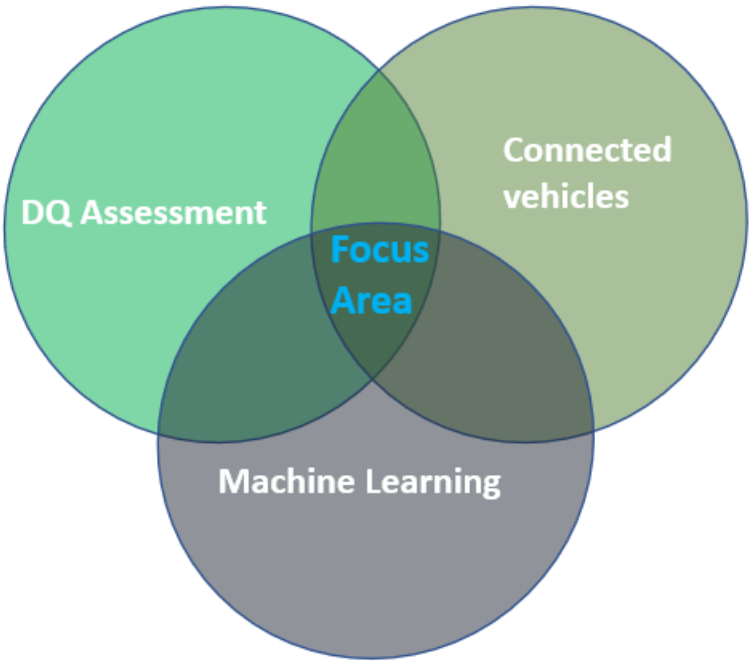


Figure 1-1 Research focus area: intersection of DQ Assessment, Connected Vehicles and Machine Learning

1.2 Research Objective and Questions

This study investigates methods and techniques that help assess CV DQ to improve reliability and confidence. To this end, this research has formulated the following main objective.

To develop a Machine Learning enabled Data Quality Assessment Framework for a better assessment of Connected Vehicles data.

The main research objective is further broken down into the following sub-objectives.

- Investigating Data Quality issues and challenges in Connected vehicles.
- Investigating existing Data Quality assessment methodologies.
- Developing Machine Learning Enabled Data Quality Framework to assess Connected Vehicles data

Based on the objectives set, the following research questions are designed.

- **Q1: What are the existing Data Quality assessment best practices, methodologies, and frameworks applicable to Connected Vehicles?**
- **Q2: What are the limitations of existing Data Quality assessment methods?**
- **Q3: To what extent does incorporating Machine Learning improve Data Quality assessment on Connected Vehicles data?**

1.3 Contributions

This study aims to contribute to the field of DQ assessment in CV by incorporating advanced techniques like ML and statistical methods. By doing so, this research presents several key contributions:

1. A comprehensive review of DQ assessment frameworks in the domain of CV systems: - A thorough review of DQ assessment in the domain of CV is provided, summarizing existing approaches and frameworks.
2. Highlighting limitations of general purpose classical DQ assessment frameworks when applied to CV data: - The study highlights the limitations of existing DQ assessment frameworks, emphasizing the need for more advanced techniques to address the complex nature of DQ in CV.
3. Application of ML to enhance CV DQ assessment: - The research demonstrates the application of advanced methods, specifically ML, for DQ assessment in CV, highlighting the potential of these techniques in improving DQ.
4. Proposed an ML enabled DQ assessment framework for CV: - An ML powered DQ assessment framework specifically designed for CV is proposed, providing a methodical and comprehensive approach to evaluate and enhance DQ in the domain.

The overall key contribution of the research is that ML methods can be used to enhance general purpose or classical DQ assessment frameworks for a better DQ assessment of CV data.

The combination of DQ assessment frameworks and ML in the context of CV is a unique aspect of this research. To the researcher's knowledge, no previous studies have specifically examined this precise combination.

Furthermore, the study highlights future research directions, identifying areas for further exploration and advancement in DQ assessment for CV.

1.4 Scope and Limitations

1.4.1 Scope of the Study

The objective of this study is to investigate if ML can improve the DQ assessment for CV data. Therefore, comparing performances of different ML algorithms is not in the scope of this study. In addition, this study considers only objective measures based on objective DQ dimensions. In other words, subjective DQ dimensions such as believability, accessibility and so on are not in this study's scope. Moreover, the study does not investigate other data sources.

1.4.2 Limitations of the Study

This research focuses on improving DQ assessment of CV data by incorporating ML and statistical methods. In conducting this study; there were various limitations encountered. Primarily, the availability of data was a challenge. This appears to be a widespread problem as described in (Zhou and Bridgelall, 2020). For experimental purposes, extensive data is available in the organization where this research's case studies were conducted. However, privacy and security issues limited its use for publications. To mitigate that, a search was performed for public datasets. This was partially successful as fuel consumption data was found, which was used for Scenario III in Chapter 5. However, comprehensive representative trajectory data was not found. Finally, a mix of anonymized data and public data was employed. Another limitation of the study is that it is limited to data generated from CV systems. However, the quality of connected data is subject to various aspects such as telecommunication and combining such data sources may reveal new insight. Further, this research did not investigate a wide range of algorithms. The quality of the research output could be improved by implementing more algorithms, making detailed comparisons, and selecting the algorithm that provides the best result.Finally, the research was limited to a selection of objective DQ metrics. However, DQ is part of a broader subject of data management which includes data ownership, governance, privacy, and security (Siegel, Erb and Sarma, 2017).

1.5 Thesis Layout

The dissertation is structured into seven chapters to comprehensively address DQ assessment in the context of CV systems. This chapter presents an introductory overview, outlining the research aims and underscoring the importance of assessing DQ in the CV ecosystem. Chapter 2 briefly provides fundamental concepts pivotal to the study, such as CV, DQ and its impact, and assessment methodologies. Chapter 2 also presents a literature review, offering insights into existing DQ assessment approaches in the domain of CV systems. In Chapter 3, the research methodology is discussed, explaining the steps taken to address research questions and achieve objectives. The adoption of a DQ assessment framework and its implementation in dashboard form for CV data is discussed in Chapter 4. Chapter 5 implements three scenarios by exploring the application of ML to enhance classical DQ assessment frameworks for CV. The proposed ML powered DQ assessment framework, specifically tailored to CV, is presented in Chapter 6. Finally, Chapter 7 concludes the dissertation by discussing research outcomes, addressing research questions, and suggesting future research directions.

Chapter 2 : Connected Vehicles, Data Quality and Data Quality Assessment Literature Review

This chapter presents a literature review on three fundamental concepts crucial to the research conducted in this study: 1. CV 2. DQ and its impact 3. DQ Assessment frameworks and methodologies. It establishes the theoretical foundation of the research by examining the current state of the art and identifying existing gaps. The chapter begins with an overview of CV, including its architecture and enabling technologies and DQ and impact of DQ on CV data. It then offers a systematic literature review of DQ assessment methodologies applicable to connected systems and examines selected articles relevant for the study presenting a critical review. Finally, the chapter highlights research gaps and presents concluding remarks.

2.1 Connected Vehicles

Global connectivity is seeing growth, and this trend extends to the integration of connectivity in vehicles (Siegel, Erb and Sarma, 2017). One crucial element of the CV ecosystem involves the effective implementation of wireless communication protocols. This enables vehicles to establish connections with other vehicles known as Vehicle-to-Vehicle (V2V), with infrastructure (V2I), or with a combination of vehicles and various entities (V2X). The V2X category encompasses both V2V and V2I interactions, as well as Vehicle-to-Person (V2P) communication (Mahmood, 2020). The term "V2I" pertains to traffic signs or stationery objects. According to Jadaan, Zeater and Abukhalil (2017), CV can be described as the ability of the various components within a vehicle to establish connections with external devices, networks, and services. These connections can include other vehicles, home, office, or infrastructure, and are facilitated through a connectivity module. Various terms are employed to denote V2V communication, including Vehicle-to-X (where X encompasses all entities), "Internet of Vehicles", “Connected Vehicles”, and "Talking Vehicles" (Mahmood, 2020). Nevertheless, the utilization of “Connected Vehicles” (CV) is prevalent among both practitioners and researchers.

CV can provide data-rich environments, which are widely recognized as crucial facilitators for numerous applications and services aimed at enhancing road safety, reducing traffic congestion, and promoting environmental sustainability (Siegel, Erb and Sarma, 2017). The provision of connectivity plays a crucial role in enabling various features and systems, such as dynamic routing, real-time navigation, and both conventional and near real-time infrastructure (Abdelkader, Elgazzar and Khamis, 2021). Additional examples include environmental

monitoring, self-driving vehicles, on-demand transportation solutions, transport as a service and so on (Abdelkader, Elgazzar and Khamis, 2021; Jadaan, Zeater, & Abukhalil, 2017). According to Mahmood (2020), within the context of the CV paradigm, the intelligence of vehicles is enhanced by the exchange of information with neighboring vehicles, interconnected infrastructure, and the surrounding environment.

With advancements in connectivity, sensors, communication technologies, processing capabilities, cloud computing and so on, which are enabling technologies, CV is exerting a considerable influence on society and demonstrating promising prospects for further development in the future.

2.1.1 Enabling Technologies

The rise of CV is facilitated by a range of technologies that support this paradigm, including "Internet of Things (IoT), AI, Dedicated Short-Range Communication (DSRC) protocols, Mobile technologies (both 4G and 5G), and cloud-based technologies" (Siegel, Erb and Sarma, 2017). Modern vehicles are also equipped with enabling technologies, including a sophisticated network of sensors that form a wide area network. This network facilitates the collection of numerous signals within the vehicle, as well as sensing of the surrounding environment (Siegel, Erb and Sarma, 2017). The subsequent section explains the technologies and methodologies that serve as the foundation for the development of CV.

2.1.1.1 Internet of Things

"Internet of things (IoT) refers to an interconnection of 'things' such as vehicles, devices, home appliances containing electronics, programs, sensors, and actuators with a technology that enables these objects to send and receive data" (Gubbi *et al.*, 2013). IoT is a network of interlinked computing devices, sensors, and actuators, physical as well as mechanical objects, animals or people with identification and the ability to send information (Gubbi *et al.*, 2013; Miraz *et al.*, 2015). IoT lets the things linked to the internet to be identified and monitored from a distance by using the connected control mechanism in the system. This facilitates seamless connection between the physical and virtual worlds. The result manifests as smart cities, industrial IoT, intelligent grids, intelligent transportation systems, smart healthcare, and similar advanced applications (Mahmood, 2020).

The application of IoT is increasing every day (Miraz *et al.*, 2015). IoT has taken the provision of services and the way of doing business to a new level of revolution (Leonardi *et al.*, 2016).

IoT is applied and being applied in various domains including health care, sport, academic systems, automating homes and offices, transportation and many more (Sethi and Sarangi, 2017). Today, due to the wide application of IoT, an extension of existing disciplines or business areas or applications have added prefixes such as “connected”, “smart”, “intelligent” and so on. Therefore, the existence of smart homes (Leonardi *et al.*, 2016), intelligent transport, CV, smart agriculture, and many more similar applications are all possible thanks to IoT. Its significance is recognized by businesses, individuals, non-governmental organizations, and governmental organizations. For example, the US intelligence council has included the “IoT as one of the disruptive technologies” (Atzori, Iera and Morabito, 2010). In summary, IoT serves as the fundamental framework that facilitates the advancement of ideas such as CV, intelligent systems, and the 4th industrial revolution, commonly known as Industry 4.0.

The industry has adopted two distinct architectural frameworks for IoT based on layers. The first is a three-layered architecture, comprising of the “perception layer, network layer, and application layer” (Sethi and Sarangi, 2017). The second is a five-layered architecture, comprising of the “perception layer, transport layer, processing layer, application layer, and business layer” (Sethi and Sarangi, 2017). For the sake of simplicity, the three-layered architecture will be elaborated in this section.

As described earlier, the three-layered architecture of IoT consists of three separate layers: “the Perception layer, Network Layer, and Application layer” (Sethi and Sarangi, 2017). These components can be associated with sensors, microcontrollers, internet connectivity, and service platforms, respectively.

I. The perception layer, also known as acquisition layer, employs sensors and electronic measuring devices to collect and transfer information from the physical world (Atzori, Iera and Morabito, 2010). Sensors convert physical characteristics to signals to capture complex inputs. Some of the common sensors in CV are used to measure safety and motion such as wheel-based sensors, speed sensors, steering and driver activity detecting sensors, power train signals like present choice of gear and engine-speed sensors.
II. The network layer, also called the processing layer, assumes the role of sharing and analyzing the information captured from the physical objects via sensors for subsequent tasks (Sethi and Sarangi, 2017). The most significant component of this layer is its networking capability that may assume both wireless and cabled forms (Atzori, Iera and Morabito, 2010).

III. The application layer or the utilization layer represents service platforms that are deployed to perform different actions on the physical platform such as adjustment, modification, maintaining and monitoring (Sethi and Sarangi, 2017). It enables specific intelligent services including smart transport, smart city, smart home and so on.

2.1.1.2 Cloud Computing

Cloud computing is a system of distributed computing where Information Technology services are provided by large and cheaper computing units linked by Internet Protocol (IP) networks (Qian *et al.*, 2009). It represents the provision of computational resources which include "servers, storage, databases, networking, software, analytics, and other intelligent services" (Qian *et al.*, 2009) through the internet which facilitates innovation, flexibility, and scaling. Cloud computing is characterized by five main things 1. Availability of massive computing resources 2. High scalability and elasticity 3. Easily resource sharing 4. Flexible scheduling, and 5. Multi-purpose capability (Lawson and Ramaswamy, 2015). The services offered by cloud computing can be classified into three main distinct categories: "Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS)" (Qian *et al.*, 2009). Cloud computing plays a crucial role in the realm of CV since it offers essential computational, storage, and communication facilities to handle the vast amounts of data created (Mahmood, 2020).

2.1.1.3 Artificial Intelligence and Machine Learning

AI and ML are disciplines widely used to create intelligent systems. While AI and ML are interrelated and sometimes used interchangeably, AI is more generic. Therefore, AI can be broadly described as the capability of machines to act intelligently (Jordan and Mitchell, 2015). And ML is a subfield of AI that facilitates the autonomous acquisition of knowledge and enhancement of systems via experience, without the need for human involvement (Mjolsness and DeCoste, 2001). It uses historical data and learns patterns without being explicitly programed using algorithms. ML algorithms include several types, such as "Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning algorithms" (Jordan and Mitchell, 2015).

Within the domain of CV, AI and ML techniques are employed for many purposes, including but not limited to "voice recognition, driver monitoring, virtual driving assistance, camera-based vision systems, and radar-based detection for other vehicles and roadside devices" (Wagstaff, 2012).

2.1.1.4 5G and DSRC Technologies

DSRC is a wireless communication system that facilitates “direct communication among vehicles, as well as between vehicles and other road users or surrounding infrastructure” (Abdelkader, Elgazzar and Khamis, 2021). This technology offers high-speed and secure communication capabilities, without relying on cellular networks or other existing infrastructure (Abdelkader, Elgazzar and Khamis, 2021). The term "5G" is commonly employed in the context of CV to denote the fifth-generation cellular chip utilized in Cellular Vehicle-to-Everything (C-V2X) communications (Mahmood, 2020). Although 4G technology can be utilized for vehicle communications, it is worth noting that 5G technology surpasses its predecessor in terms of both speed and dependability (Mahmood, 2020). The discourse frequently incorporates C-V2X technology, wherein the inclusion of 5G is commonly observed in conjunction with deliberations on autonomous vehicles due to its high velocity and communication capabilities (Yang *et al.*, 2018). The popularity of 5G technology has garnered the attention of vehicle manufacturers as they seek to incorporate it into their new inventions (Mahmood, 2020).

Comparing DSRC to C-V2V, it is noted that C-V2V has many main advantages over DSRC such as wider area coverage, more reliability, and better performance (Mahmood, 2020). While 5G is still developing, it has enormous potential.

CV became a reality by using all these technologies and other innovations. The diagram in Figure 2-1 below shows a typical architecture consisting of the components supporting CV.

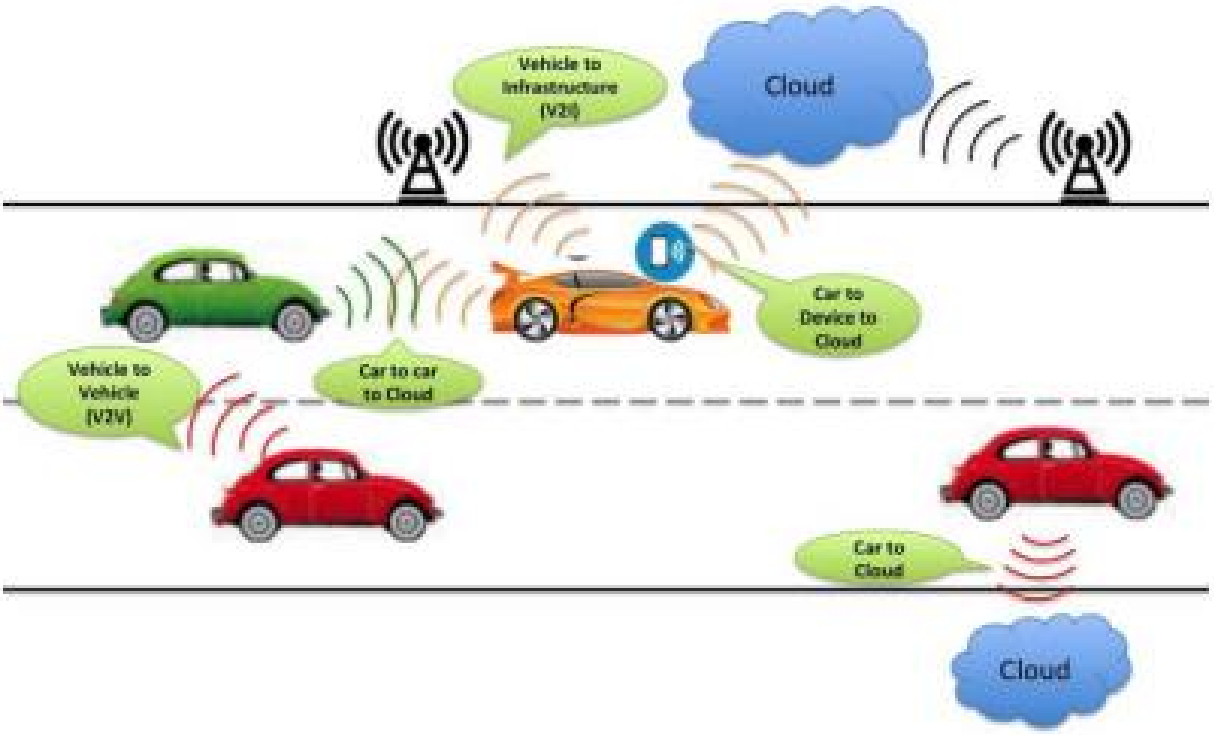


Figure 2-1 vehicle-to-everything communication supporting connected vehicles taken from (Siegel, Erb and Sarma, 2017)

2.2 Data Quality

The field of DQ holds a prominent position within the world of academia, as many researchers have put an effort to generate a substantial number of research literature addressing issues and solutions around it (Karkouch *et al.*, 2018). Good DQ, also known as information quality (IQ), is widely recognized as a critical requirement for the performance and growth of organizations (Lee *et al.*, 2002). High-quality data is an essential prerequisite for each decision or action made using data and information (Mazón *et al.*, 2012). To provide an illustration, the effectiveness of an ML solution in uncovering novel patterns and relationships within a given dataset is dependent upon the availability of a trusted dataset. In the presence of flawed data containing an excessive number of outliers, the visibility of patterns may be compromised, making them unrecognizable (Batini, Scannapieco and others, 2016).

The literature does not provide a universally accepted common definition for DQ; instead, various definitions are employed including “fitness for use” (Cai and Zhu, 2015; Juran and Godfrey, 1999), “conformance to requirements” (Fox, Levitin and Redman, 1994), “degree of data fitness for a given purpose” (Gudivada, Apon and Ding, 2017), “how much data satisfies user expectations” (Sebastian-Coleman, 2012) and many other related definitions. From the literature, it is evident that there is an agreement on the concept of “fitness for purpose”, i.e.,

DQ depends on the consumer, the situation, and the time (Cai and Zhu, 2015). Fitness for use, according to (Brimicombe, 2010), denotes the evaluated value of the outcomes of analysis applied for decision making. Fitness for use can be inferred from the reliability of the data if one topic is considered for decision making. However, if multiple themes and data sets are used, the analysis's outcome can be evaluated from the blend of data reliability of each topic. Therefore, Brimicombe (2010) emphasizes that fitness for use is not always derived from analytical outcomes, rather depends on contexts. In line with this, Cichy and Rass (2019) underscore the significance of recognizing the distinct characteristics of data across different domains. It is crucial to consider the classification of data structure and type into more precise categories, such as time-continuous data and event-based data (Micic *et al.*, 2017). Hence, it is important to prioritize the reliability of data through the implementation of a clearly defined methodology and the incorporation of uncertainty propagation modelling (Batini *et al.*, 2007). Additionally, when necessary, efforts should be made to minimize the extent of uncertainty. Fitness for use can be decomposed into three elements (Veiga *et al.*, 2017).

- Use: What is the goal the data is required to have the desired quality?
- Data: - What type of data must be available and exhibit the desired quality in consideration of the intended usage?
- Fitness: - What are the components and level of fitness needed for the data identified and the situation of the use defined?

Furthermore, in terms of overall quality, as defined by ISO 9000:2021, quality is defined as the extent to which a certain collection of inherent qualities satisfies a given set of requirements (International Organization for Standardization, 2021). Considering this, it is reasonable to assert that many researchers concur with the concept that may be summarized as follows: **DQ pertains to the extent or degree to which specific data aligns with the prerequisites of a certain use case** (Batini *et al.*, 2009; Sebastian-Coleman, 2010; Wang and Strong, 1996).

While the term DQ is commonly referred to as "fitness for use", since the phrase was initially coined by Juran and Godfrey (1999); due to the fact that DQ has multifaceted nature, there exists no exact or agreed-upon definition for DQ. Therefore, the most practical and easiest way to describe DQ is by using DQ dimensions. This is evident as many studies use DQ dimensions to formulate DQ assessment methodologies (Karkouch *et al.*, 2016). For example, Gudivada, Apon and Ding (2017) argue that the definition of DQ as "fit for business purpose" is broad and subjective. Therefore, they propose a more tangible and practical approach by incorporating

specific characteristics of DQ, such as “accuracy, currency, and consistency”, to establish a more concrete definition.

The different DQ dimensions capture one or more characteristics or aspects of DQ including “accuracy, completeness, timeliness” and so on which contribute towards the overall DQ measurement. The following section highlights the common DQ dimensions widely applied in the literature.

2.2.1 Data Quality Dimensions

The determination of DQ dimensions also lacks agreement among scholars and researchers. According to Cichy and Rass (2019) there is a notable degree of variability in the dimensions of DQ that are included within different frameworks. The extent of this variability is dependent upon the specific usage and subject matter being addressed. There are various definitions and prioritizations for the different dimensions of DQ, as outlined in several research works (Aziz, Saman and Jusoh, 2012; Loshin, 2010; Liaw *et al.*, 2011 and Strong, Lee and Wang, 1997). These studies identify conceptual frameworks for DQ that primarily vary in terms of the dimensions included and their classification. Several researchers have noted that there is a commonality across these approaches since they tend to employ descriptive and subjective definitions for the dimensions (Loshin, 2010; Pipino, Lee and Wang, 2002). These definitions often rely on phrases that have overlapping or ambiguous semantics. Additionally, Batini *et al.* (2015) present an extensive examination and comprehensive analysis of the various aspects and variations of DQ in their study. Nevertheless, the DQ framework put forth by Wang and Strong (Strong, Lee and Wang, 1997) is widely recognized as the most prominent conceptual framework within the research community. The concept of information has been characterized as a product by Wang (1998), and it is acknowledged that, like any other product, information is subject to quality requirements. A series of surveys and research were undertaken to identify, categorize, and prioritize dimensions of DQ that encompass 179 specific dimensions, based on their significance to customers. The outcome of their efforts is a hierarchical structure that categorizes the identified aspects of DQ into four distinct groups: intrinsic, contextual, representational, and accessibility. This framework has a total of fifteen primary dimensions of DQ, as illustrated in Figure 2-2. Within this framework of categorization, the group of intrinsic DQ includes several aspects or dimensions, such as “believability, accuracy, objectivity, and reputation”. The contextual group includes various attributes, namely “Relevancy, Value Added, Timeliness, Completeness, and Amount of data”.

The groups mentioned earlier, which include the primary aspects of DQ, are widely recognized. However, it is crucial to acknowledge the significance of the remaining two groups: "Accessibility, which includes accessibility itself and security", and "Representational, which includes interpretability, understandability, conciseness, and consistency". These additional groups play a vital role in extracting value from data

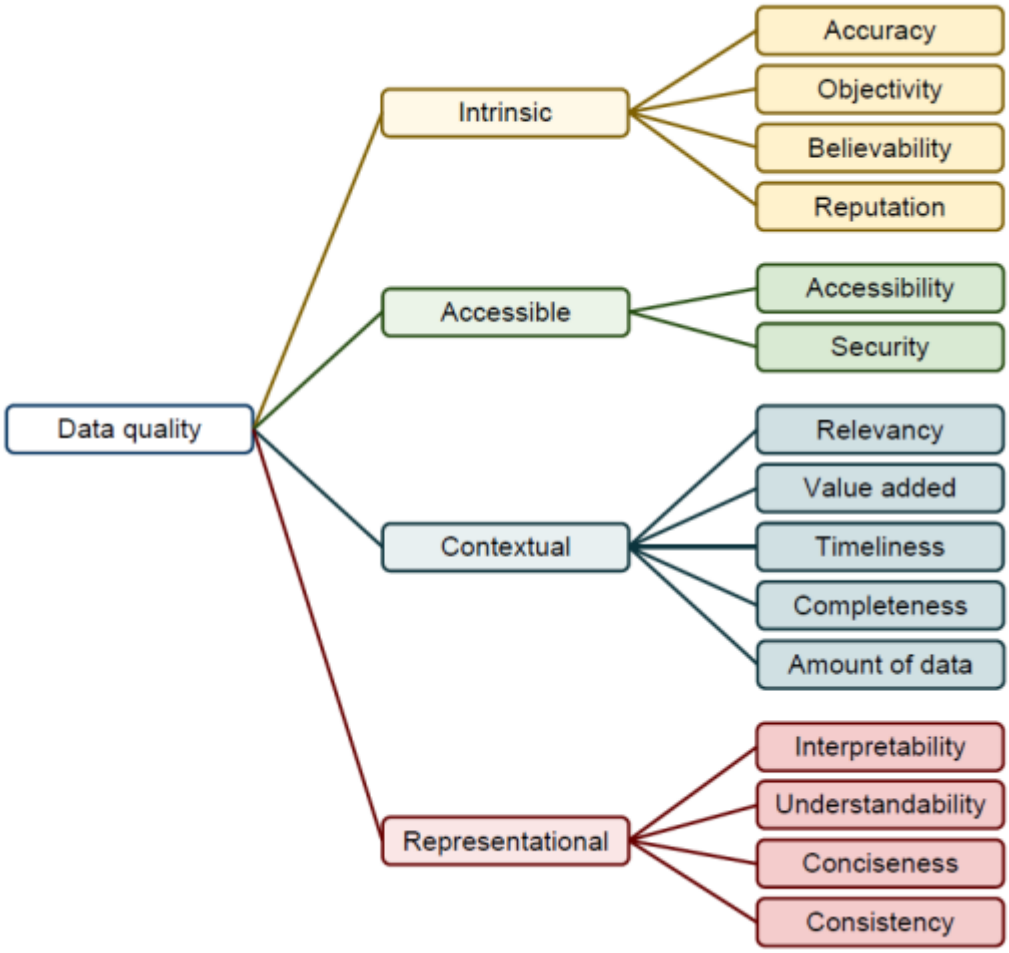


Figure 2-2 A DQ framework consisting of 15 dimensions identified by Wang and Strong (Strong, Lee and Wang, 1997)

The dimensions of DQ can be associated with the extension of data, consisting of both the actual data values and the metadata, which represents the intended meaning of the data (Eppler, 2006). This study concentrates on the characteristics of DQ that pertain to the extension and value of data, as identified by Eppler (2006) as being particularly pertinent to practical applications. Table 2-1 presents a concise compilation of the most employed measures of DQ, together with their respective definitions.

Table 2-1 DQ Dimensions representing extension of data or metadata

DQ Dimension	Definition
Accuracy	It is commonly used to describe the extent to which data precisely reflects the true value. Alternatively, it signifies the extent to which data is "accurate, reliable, and certified" (Cichy and Rass, 2019).
Completeness	The level of availability of values for all expected data elements by an entity, or the amount to which data meets the expected "breadth, depth, and scope" within the context under consideration (Wang and Strong, 1996).
Consistency	The degree to which a set of data is free from contradiction or the absence of difference when compared to a definition of the data being represented by two or more other representations of the same data (Group and others, 2013).
Timeliness	It is commonly used in reference to data, denoting the level of "currency or accuracy in reflecting the present state of reality at a particular moment" (Group and others, 2013).
Validity	For data to be deemed valid, it should adhere to the prescribed "syntax, including the format, type, and range", as stipulated in the data definition (Group and others, 2013).

DQ is well studied and researched, particularly focusing on DQ dimensions, contributing most of the research works in the domain. Accuracy, completeness, consistency, and timeliness are frequently discussed DQ dimensions (Batini *et al.*, 2009).

According to Batini *et al.* (2009), accuracy can be expressed as "the extent to which a given value, denoted as x, matches another value, referred to as x' which is regarded as the accurate representation of the real-world phenomenon under consideration". According to Strong, Lee and Wang (1997), accuracy is defined as "how correct, reliable, and certified a value is." According to the definition of Batini *et al.* (2009), data is seen as being accurate when the value it stores is consistent with the actual value in the real world.

As stated by Batini *et al.* (2009), completeness pertains to "the extent to which a collection of data fulfils the criteria of a specific objective in terms of its breadth, depth, and scope".

Validity, on the other hand, refers to the extent to which a given dataset conforms to a specific business rule, established norm, or area (Group and others, 2013). In contradiction to accuracy and completeness, validity does not incorporate any comparative analysis with respect to real-world phenomena (Sebastian-Coleman, 2010).

Consistency refers to the degree of conformity exhibited by data in relation to another dataset from the same situation or developed or generated using a similar method throughout a similar period of time (Group and others, 2013). The consistency of the data is essential to ensuring that it is accurate (Sebastian-Coleman, 2010). Controlling consistency can be accomplished through the utilization of rules and standards, other data pieces contained within the same source or data obtained from other sources, and distinct instances handled in the same manner (Batini *et al.*, 2009; Sebastian-Coleman, 2010).

The timeliness aspect of data is directly proportional to the duration elapsed since the data was last updated (Group and others, 2013, Sebastian-Coleman, 2010). According to Batini *et al.* (2009), data should be available and in the correct state when it is being used; otherwise, it is believed to be out of date.

In accordance with the data requirements and the nature of the domain, one can assign a higher or lower importance rating to a DQ dimension (Loshin, 2012). This means the emphasis on a specific or a set of DQ dimensions depends on the context. For instance, one application might require that the underlying data mostly be accurate, but it might not be complete. Another application could focus on completeness rather than consistency in the hope of achieving its goals (Group and others, 2013). In addition, it is feasible for the distinct aspects of DQ to sometimes contradict one another. For instance, the completeness dimension and the consistency dimension may be in conflict with one another due to the fact that an attempt to make a particular data piece complete may result in a reduction of the matching reference values in another circumstance or location, which, in turn, will lead to an increase in the amount of inconsistency (Fox, Levitin and Redman, 1994). According to research done by Wang and Strong (1996), rating the completeness and consistency of various aspects of decision making is not an easy task. In most cases, a decline in consistency follows an increase in completeness, which can be understood as the possession of more facts. Similarly, it is indicated in the same study that an increase in the currency dimension, specifically the acquisition of more up-to-date information, may lead to a decrease in accuracy.

When it comes to smart connected systems, the emphasis given for some DQ dimensions is higher than others in line with the findings of research described earlier, and some even introducing a new DQ dimension such as provenance. While many of the researchers agree on which DQ dimensions are more relevant for smart systems, some slightly differ. According to Juddoo and George (2018), for instance, important problems include maintaining consistency, correctness, completeness, and timeliness. Completeness, correctness, and timeliness are

highlighted as the most critical issues in (Cai and Zhu, 2015), while Olufowobi *et al.* (2016) adds provenance as an additional DQ concern. Provenance is defined as the record of the chronology (timeline) of data ownership, as well as any data transformations or alterations that were performed to a data set; and in practice, the DQ dimension of believability is the one that the method of data provenance as an assessment of DQ is typically utilized to measure (Olufowobi *et al.*, 2016).

(Hazen *et al.* (2014) has provided an in-depth analysis of the DQ issues that are present in CV, particularly with regard to spatio-temporal elements. (Juddoo and George, 2018; Olufowobi *et al.,* 2016 and Ricardo Perez-Castillo *et al.,* 2018) agree on the fact that smart connected systems are made up of three parts: "the physical, the smart, and the connectivity" components as described in section 2.1.1. These parts come together to form a complex network that has three layers: "acquisition layer through sensors, processing layer, and utilization layers". This results in an increase in complexity regarding things like the scale of deployment, the lack of resources, the network complexity, the variation in sensors, the situational context, the destruction of components, fail-dirty occurrences, security vulnerabilities, data stream processing complexities, and so on (Ricardo Perez-Castillo *et al.,* 2018). In addition to producing substantial amounts of data, the various components are also prone to malfunctioning. Big data is known for having a sloppy structure and having values that are missing (Gudivada, Apon, and Ding, 2017; Juddoo and George, 2018). Machine generated data does not have the appropriate information. Errors such as offset, continuously varied or drifting, crashed or jammed, trimming error, outlier, noise, and so on are prevalent (Megler, Tufte and Maier 2016; Ricardo Perez-Castillo *et al.,* 2018). These kinds of errors are common in IoT based connected systems. As a result, words such as "trustworthiness," "confidence," "credibility," and so on are frequently employed in research works about DQ in smart connected systems (Juddoo *et al.*, 2018).

In conclusion, it can be noted that the DQ aspects in IoT applications, one of which is CV, stay the same. However, certain aspects of the data, such as its timeliness, accuracy, and completeness, have become significantly more important than others (Farooqi, Khattak, and Imran, 2018).

2.2.2 Data Quality Impact

It is important to place an appropriate emphasis on DQ, and one way to do so is to investigate the influence that it has on businesses. A problem with the quality of the data causes problems

for businesses in a variety of diverse ways. According to Cichy and Rass (2019), high-quality data is essential to the success of businesses since sound decisions are predicated on easy access to relevant information. An insufficient degree of DQ will have far-reaching impacts, including poor decision-making, increased costs, poor operational performance, and non-compliance with rules (Redman, 1998).

Strong, Lee and Wang (1997) describe DQ issues as the inability to use data for the needed purpose if the data becomes unfit due to difficulties in one or more of the DQ dimensions. This definition applies when the data becomes unfit because of issues in one or more of the DQ dimensions. According to Floridi (2013), failing to have a clear concept of DQ can lead to misunderstandings, mistakes that are costly, or missed opportunities.

In general, the impact of DQ is explored by many researchers. For instance, (Spruit, Linden and others, 2019) conducted a comprehensive study on businesses and uncovered 11 implications of DQ issues. However, the classification given by Loshin (2011) provides a reasonable summary by grouping the implications of poor DQ into four distinct categories: (1) financial cost; (2) Lack of confidence and satisfaction; (3) Loss of productivity; and (4) In-adherence to risk and compliance. These categories are briefly described as follows:

1. Financial Cost

The financial cost of problems with DQ is the one that receives the most attention, given that all the other impacts will affect financial performance. Several studies have been undertaken to demonstrate the financial consequences of inadequate DQ. For instance, according to an estimation by the Data Warehousing Institute (TDWI), issues related to DQ result in an annual cost of 700 billion dollars for organizations in the United States (Gudivada, Apon and Ding, 2017). According to the findings of yet another study conducted by Gartner (Spruit, Linden and others, 2019), poor DQ results in an average loss of 15 million dollars for enterprises. Based on another study conducted by IBM in 2016, it was predicted that poor DQ costs organizations in the United States more than three trillion dollars annually (Spruit, Linden and others, 2019). It is also stated in (Redman, 2017) that a study that was carried out in 2017 found that the cost of poor DQ approximately ranges from 15% to 25% of the income of most businesses. These findings clearly show the magnitude of the financial burden poor DQ places on businesses.

2. Lack of Confidence and Satisfaction

In addition to the financial costs, poor levels of DQ might have a negative effect on the decision-making process that occurs within an organization. According to a study done by KPMG in

2017, titled "Global CEO Outlook," it was shown that the majority of CEOs included the study, accounting for 56%, express concern regarding the validity and integrity of the data that underpins their decision-making processes (Cichy and Rass, 2019). Many businesses, because of problems with DQ, make assumptions about the accuracy and reliability of the data they collect (Sarfi *et al.*, 2012). This leads to all the negative consequences stated, including low customer satisfaction, which damages a company's reputation and negatively affects profitability and efficiency.

3. Loss of Productivity

A deterioration in the quality of the data also brings about a decrease in productivity since it makes it more difficult to perform straight-through processing using automated services (Spruit, Linden and others, 2019). As soon as it enters the system, it calls for a significant amount of work to be done in order to mitigate all of the adverse effects, assuming that this is even possible (Redman, 2017). This means that more data problems will arise, which will result in more valuable time being spent by personnel trying to fix them (Shardt, Yang and Ding, 2016). It has an impact on staff members at all levels of the firm, including managers, those working in customer service, data experts, and others. Inaccurate decision making and a failure to capitalize on business possibilities are yet other negative consequences of poor DQ (Redman, 2017).

4. In-adherence to Risk and Compliance

Inadequate DQ can also give rise to a compliance risk, wherein the level of DQ fails to align with the anticipated standards set forth by regulatory bodies (Loshin, 2011). In other words, poor DQ might make it difficult to comply with regulations. Various nations and organizations are beginning to implement new policies as a direct result of the collection and exploitation of enormous volumes of data as well as sensitive data. For instance, one of the obligations that are set by General Data Protection Regulation (GDPR) is that businesses are required to correct erroneous or incomplete privacy sensitive data (Hoofnagle, van der Sloot and Borgesius, 2019).

2.2.3 Data Quality Impact in Connected Systems

The nature and volume of data created by connected systems is complex, which makes the problem of poor DQ in IoT-based applications, such as CV, much more severe. Data from telematics and CV are referred to as opportunity data in a study by Keller *et al.* (2017) since it enables to develop new services there by introducing new revenue stream. Massive amounts of

data are generated, but the logic that underlies this process is not obvious, and neither the nature of errors nor the statistical characteristics are clear (Keller *et al.*, 2017). While there is currently no empirical research that specifically measures the extent to which inadequate DQ affects connected systems, it is evident that the impact will be amplified due to the complex nature of such systems, which involve numerous interlinked components.

The data that is produced by IoT or smart systems is of various nature (Kim *et al.*, 2019; Perez-Castillo *et al.*, 2018). This is because these systems incorporate many different components. The generation of data in the IoT differs from that of other systems in that its data sources are both many in number and varied in type. These differences can be attributed to the following factors (Karkouch *et al.*, 2015).

- Multiple elements, such as sensors, are engaged.
- Both humans and machines are involved in the process.
- Both space and time play a major role in the equation.

The data that is produced by the IoT can be arranged in a variety of ways and Perez-Castillo *et al.* (2018) groups it into the following four.

- The information that is gathered by sensors, such as readings of temperatures.
- Instrument data that is sensor data that has been augmented with metadata (for example, the time of the reading and when the instrument was released).
- Generic data that is data which is pertaining to a business area (for example, client information).
- IoT data which is the union of instrument data and generic data.

Due to all the reasons listed above, DQ has been identified as the primary challenge for IoT-based smart systems (Gubbi *et al.*, 2013; Perez-Castillo *et al.*, 2018; Prathiba, Sankar and Sumalatha, 2016). It is impossible to implement the potential benefits of the IoT if the data is unreliable, as is clearly described in (Davenport and Redman, 2015) with a practical example in which an incorrect clock caused the postponement of an important meeting agenda. According to the same study, DQ problems are more prevalent in smart systems that are based on the IoT because devices bring extra errors on top of the common human faults that are familiar in conventional systems. In their research, Fekade *et al.* (2018) identified a number of factors that can contribute to poor DQ, such as issues with connection, interference from the surrounding environment, or sensor malfunctions. This could result in data that is either missing or inconsistent, which in turn could result in monetary loss, a decrease in customer satisfaction, a decrease in productivity, and a failure to comply with regulatory standards (Loshin, 2011).

Time is yet another aspect of IoT and connected systems that is quite significant. According to Cai and Zhu, (2015), the "timeliness" of data is brief and the rate at which data change is very rapid in modern connected systems, both of which imply higher requirements for processing technology that delivers timely data. In this regard, a preventive warning system that is deployed based on a streaming CV will not be able to serve the intended purpose if data is delayed or missing, which depending on the application area may lead to costs up to and including loss of life (Jia *et al.*, 2014).

2.2.4 Data Quality Assessment

The primary aim of a DQ Assessment process is to identify inaccurate and erroneous data elements and assess their potential effects on diverse data-centric business processes, hence facilitating the development of a strategy to mitigate these impacts. Therefore, before using data for any application, it should be assessed for fitness-of-use. This is because, as stated in (Brimicombe, 2010), it is unlikely to achieve 100% data accuracy as errors and uncertainty are inevitable, and further deterioration can even be caused when two or more data sets are combined, which will have an impact on the quality of the output from the data. Therefore, it is important to know the level of DQ and the means of mitigating DQ issues. For this, DQ assessment strategy should be developed from the outset when embarking on a project of a data centric application or service. Many researchers and practitioners apply methodologies and frameworks to systematically assess DQ.

For a long time, DQ has been considered as a multidimensional concept and as a result its assessment is viewed as a complex process with multifaceted challenges (Cichy and Rass, 2019). Besides, DQ is understood as a multidisciplinary problem spanning subjects such as computing, quality control, human factors, and statistics (Cichy and Rass, 2019). Therefore, many researchers approached DQ assessment from this perspective of multidisciplinary nature. However, addressing all aspects of DQ at one time for general purposes is challenging. For example, according to Eppler (2006), it has been observed that most DQ assessment frameworks are intended to be tailored to certain domains, with only a limited number of frameworks possessing the versatility to be applied across several domains.

The DQ assessment process often relies on measures defined upon dimensions of DQ. DQ dimensions are tools to describe the concept of DQ and researchers suggest starting with DQ dimensions in DQ assessment endeavor (Juddoo *et al.*, 2018). DQ dimensions are features of data which may provide the overall fitness level if measured properly and it is recommended to

begin DQ assessment effort by clearly listing relevant DQ dimensions for the situation in context (Cichy and Rass, 2019). To this end, there has been several research in DQ and it is evolving to respond to changes such as various data types and innovations (Karkouch *et al.*, 2016) such as shift from monolithic to networked systems (Batini *et al*., 2009), expansion of IoT and connectivity as these changes introduce complexities. Consequently, many frameworks and methodologies have also been developed using advanced techniques, including ML and statistical methods.

To be able to understand the current state of the art and identify gaps regarding CV DQ assessment frameworks and methodologies, a systematic review of the literature was done, and the findings are reported in the subsequent section.

2.3 Data Quality Assessment on Connected Systems - Systematic Literature Review

This section provides an overview of the findings of analysis of systematic literature review related to the DQ assessment in connected systems. Since CV belongs to the family of smart connected systems, the systematic literature review was performed for the wider group of smart connected systems after initial preliminary search evidence returned only a few results for CV DQ assessment. The systematic literature review was conducted for the following reasons:

1. To summarize the existing body of knowledge concerning the use of methodologies applied to DQ assessment with a specific focus on connected systems.
2. To identify shortcomings in existing research around DQ assessment for connected systems.
3. To provide a foundation to build on for subsequent chapters of this research.

The systematic literature review included a range of scholarly articles published between 2011 and 2022. The justification for selecting this specific time interval is based on the emergence of connected systems during this period, as well as the increasing importance placed on accurately assessing DQ within the given field. The time of the review was also restricted to the year 2022 due to the period in which this research was performed.

2.3.1 Research Questions for Systematic Literature Review

This systematic literature review is formulated to answer the following questions.

- ***Q_1:*** **What is the intensity of research on Data Quality issues and solutions in smart connected systems?**
- ***Q_2:*** **What kind of Data Quality issues in smart connected systems are being addressed by researchers?**
- ***Q_3:*** **What approaches and techniques are researchers investigating to enhance Data Quality assessment in smart connected systems?**

2.3.2 Search Process and Strategy

To get acquainted with and to have a general understanding, a generic search was conducted. What was evident in this preliminary practice is that, unlike most well-studied areas for which the journals and conferences where publications are available are known, the sources pertaining to the subject of DQ assessment in connected systems appear to be scattered throughout several publications. The topic selected for this study is multi-disciplinary spanning computer science, Information science, Sensor technologies, the IoT and so on. Therefore, it is expected that the literature will be spread across different journals. Therefore, the strategy devised here is to search from well-known digital libraries which are famous in publishing studies of related areas of the study including DQ, IoT, connected systems and sensors. The digital libraries selected for this study are:

- IEEE Xplore
- Science Direct
- JSTOR
- Scopus
- Ebsco
- ACM

In addition to the digital libraries mentioned above, open-access journals are also explored.

To construct a search query, it is advised to split the study question to separate items based on context, approach, result, subject of study and so on (Khan *et al.*, 2003). Next, terms with similar meanings, acronyms, and other forms of written terms must be sorted out. In addition, journals and databases must be consulted to collect potential keywords from titles, abstracts, index terms or other meta data items. After selecting keywords, advanced search queries can be formulated by applying logical operators such as AND, OR and NOT.

For this systematic review, the strategy explained above is followed. First, an initial search is conducted to find available systematic reviews in DQ issues and solutions in CV systems and to have a global view of the magnitude of the literature around the topic considered. During this process, it is understood that there are few studies on the subject. Therefore, related systematic reviews, for example, the work of Bashir and Gill (2017), have been identified and studied. Using this as a starting point, an experimental search was conducted using the various keywords obtained from the research questions. The search included the digital libraries listed above and other sources, including the Internet. The keywords used at this stage are data quality, assessment, issues, connected systems, Internet of Things, smart systems, and enhancements. While data quality and the Internet of Things return a big search hit when applied separately, the combination does not result in many hits. The main lesson obtained at this stage is that since the study is multi-disciplinary, it is important to search for related fields. According to Budgen and Brereton (2006), publication bias is one of the issues identified in conducting a systematic literature review since most published materials focus on positive results. Consulting an expert, employing statistical analysis, investigating grey literature, and conference proceedings helps to tackle this specific issue.

With the lessons learned above, the search query is formulated by combining logical operators. There are three main components in the search: connected systems, data quality and alternating assessment/enhancement and related words. From the preliminary investigation, it is understood that there are various forms of synonyms or abbreviations of the terms used. For example, connected systems are used interchangeably with smart systems and the Internet of Things is usually contracted as IoT. Many researchers and practitioners also use connected devices, smart systems, or intelligent systems. Besides, it is understood that many research works on the subject are found in sensors journals. Enhancement is also similar to improvement and solution. To include all forms of spelling, wild card (*) is used where a variation is expected. It is also believed that some of the words should appear together, for example, Internet of Things and data quality. In such cases, the words are enclosed with “”. The search query formed with this process is given below.

(IoT OR "Internet of Things" OR sensor OR "Intellig*" OR "smart*" OR "connect*") and ("Data quality") AND ("enhanc*" OR "improv*" OR "solution")

Later, it became apparent that some articles use only data quality; and IoT or Connected Systems related terms but leave assessment, enhancement or improvement or any related term even if the research is targeted at investigating solutions for connected systems DQ issues.

Therefore, the last part is omitted to increase the recall. Instead, it is included in the inclusion and exclusion criteria at the later stage of the process. As a result, the search string becomes;

(“data quality") AND (IoT OR "Internet of Things" OR connect* OR sensor OR intellig* OR smart*)

The search string's syntax is adapted to conform to certain journal requirements during the search process. During the search process, there were several filters applied. According to Suresh *et al.* (2014), it has been since 2011 that an enormous appetite for smart systems and the Internet of Things has developed. Further, it is explained that this interest is initiated by the introduction of IPv6, and many organizations undertake different experiments regarding IoT. Therefore, a filter with the publication year 2011 and after is applied. In addition, due to time and resource constraints, only English literatures are considered. Books, book chapters, thesis, and dissertations are also excluded. To avoid publication bias, grey literature is included. The snowball approach is used to identify works cited by the retrieved articles and considered relevant. Applying this search strategy explained, 732 articles were retrieved. The number of articles obtained from each database that was queried is provided in Table 2-2.

Table 2-2 Articles retrieved per journal

Database	**Number of articles**
IEEE Xplore	256
EBSCO	135
Scopus	221
Science Direct	57
ACM	63
Total	**732**

The citations of the 732 articles are extracted and imported to Mendeley. After removing duplicates, 540 articles remained. As a subsequent process, screening is made based on the title, abstract, conclusion and full-text scan. The process is explained as follows.

Inclusion/Exclusion criteria:

a) Remove redundant and unrelated publications that lack terms 'Connected systems', 'IoT', 'Sensor' or any of the following words in various forms prefixed with smart or intelligent: city', 'transport', 'education', 'governance', 'energy', 'safety', 'environment', or 'healthcare' in their title, abstract, or meta-data.

b) Exclude publications that do not specifically discuss challenges related to "data quality" within the context of smart connected systems.
c) Exclude articles that exclusively focus on discussing DQ concerns without exploring potential solutions, assessments, or strategies for improvement or enhancement.
d) Exclude systematic literature reviews from the analysis and include only primary studies.

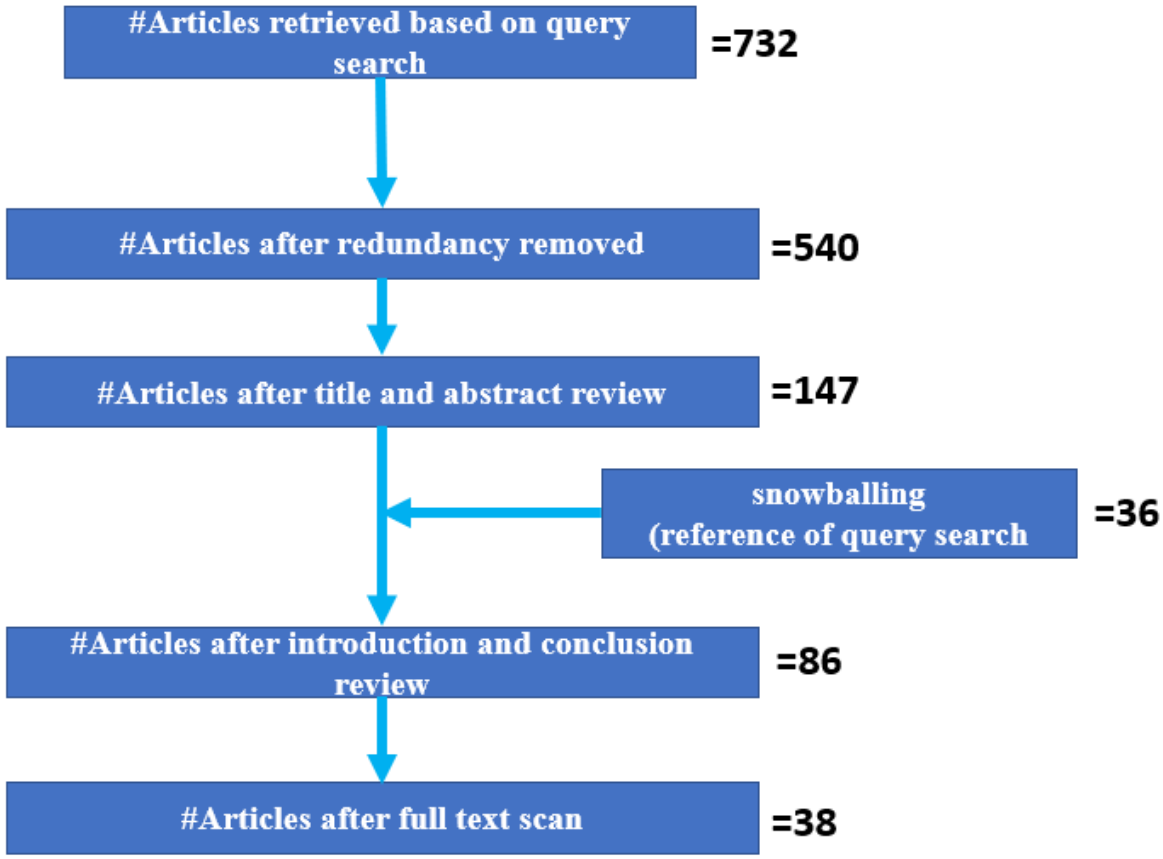


Figure 2-3 Step by-step article filtering process for systematic literature review

2.3.3 Quality Assessment

The process of quality assessment is intended to validate the quality of the original studies included in the study. In other words, the strength of the secondary study being conducted depends on how solid the building blocks (which are the primary research) are. The selection of peer-reviewed articles is one of the efforts to this end. But conducting a quality assessment gives thorough criteria. In the systematic review, a quality assessment matrix is developed to guide this process. The content of the quality assessment matrix varies depending on the type and context of the study. For this systematic review, primarily relevance and clarity as a general guideline is used. The definition of the detailed quality assessment criteria is provided in Table 2-3. To provide an objective assessment, points are provided for each assessment criterion.

Based on the defined assessment criteria, there is a maximum attainable 24 points. The maximum possible point for each criterion is given in parentheses.

Table 2-3 assessment criteria for literature selected for final review

Quality criteria	Description
Number of DQ issues addressed (5)	A score was given based on the number of DQ dimensions (listed in section 2.2.1) handled. 1 DQ dimension = 1 point, 2 DQ dimensions = 2 points, …, 4 or more DQ dimensions = 5 points.
Currency (3)	This is used to determine if current smart connected systems DQ issues are handled by the research. In connected systems, timeliness, completeness, consistency, accuracy, and validity are the main problems according to the literature. If the research handles three of these, a score of 3 is given, if it handles any one of the five the DQ dimensions listed here, a score of 2 is given or else a score of 1 is given.
Complexity level (3)	Complexity is determined based on the number of data sources, the volume of data, number, and variety of sensors. If all three are manifested, then a score of 3 is given, if any two of them are described then 2, if only one is described then a score of 1 is given, or else 0 points will be given.
Data (3)	This is determined by the number of data categories specified in Section 2.2: sensor data, device data and general data. If the dataset used includes all three, then a score of 3 is given, if it touches only two of them, then a score of two is given; if only one of them is used, then a score of 1 is given, or else a score of 0 is given.
Research method (1)	If it uses a clear and explicit research methodology, then a score of 1 is given; or else a score of 0 is given.
Effectiveness of the method (2)	If the method's effectiveness is shown compared to other methods, then a score of 2 is given, if no comparative result is provided but simply explained, a score of 1 or 0 is given.
Tested on real-life data/system (2)	If the method is tested in a real-life working system, then a score of 2 is given, if it is tested using sample data taken from a real-life system or simulated data then a score of 1 is given, or else a score of 0 is given.
Quantified Result (2)	If a measuring approach is used and a quantified result is given, then a score of 2 is given, if a measuring approach is used but no quantified result is given, then a score of 1 is given, or else a score of 0 is given.
Generalizability of the method (2)	If generalizability is demonstrated by applying to a different domain or area, then a score of 2 is given, if generalizability is only assumed, then a score of 1 is given, else a score of 0 is given.
Limitation or Validity (1)	If the limitation of the study is discussed, then a score of 1 is given or else a score of 0 is given.

To assess the quality of the selected papers for the final study, a matrix is developed based on the DQ criteria defined in Table 2-3. The developed matrix is presented in Appendix A. The final score is given by adding the individual scores of each criterion.

2.3.4 Analysis and Findings

Q1: What is the intensity of research on Data Quality issues and solutions in smart connected systems?

This question can be answered by observing the distribution of the studies for different DQ dimensions. In this case, the publication year, and the journals where the articles were extracted may give insight. The distribution of the journals per year is provided in Figure 2-4 below. As can be seen from the graph, the number of publications increased in recent years.

Another indicator for the first question is the distribution of the selected articles per journal. As shown in Table 2-2, most articles are published in IEEE (256) followed by Scopus and ACM.

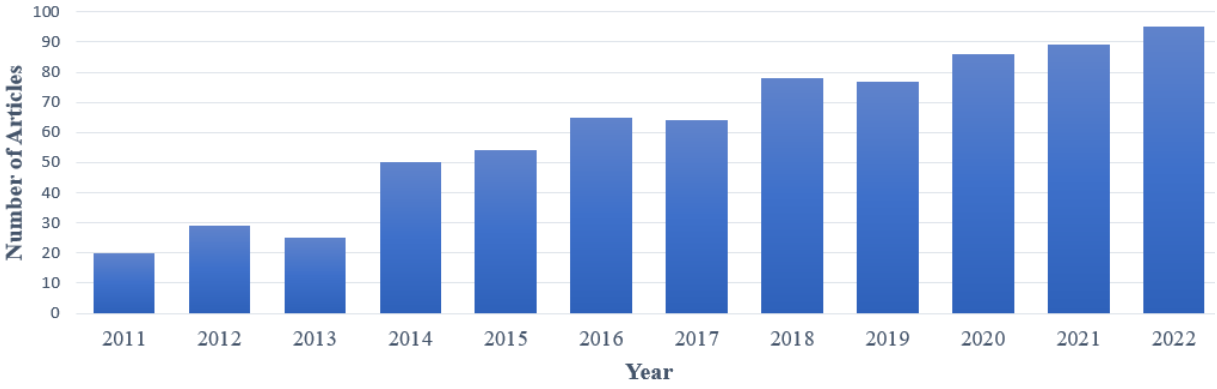


Figure 2-4 Number of retrieved articles per year (distribution per year)

Q2: What kind of Data Quality issues in smart connected systems are being addressed by researchers?

This question can be answered by looking into the distribution of the selected articles per the DQ dimensions handled given in Figure 2-5. As shown in the figure, most of the research focused on validity followed by completeness.

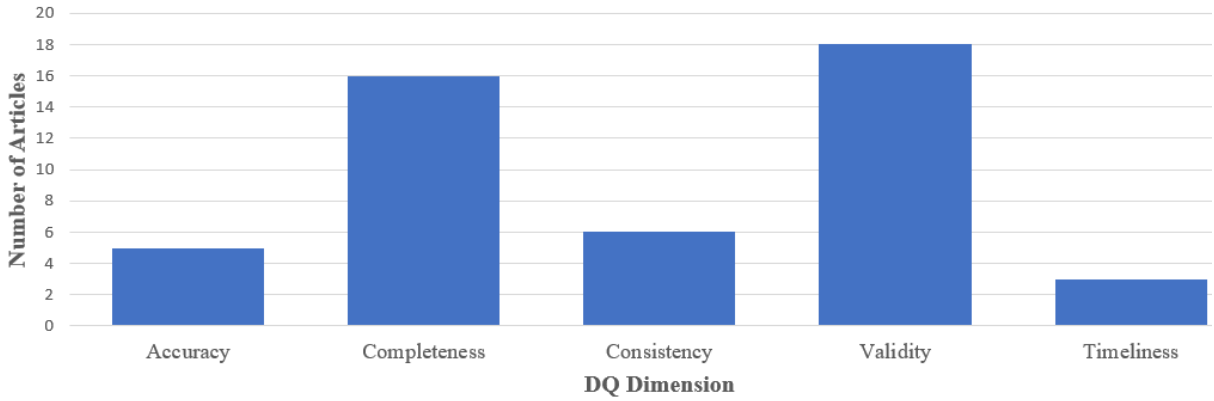


Figure 2-5 Number of selected articles per DQ dimension (distribution per DQ per year)

Q3: What approaches and techniques are researchers investigating to enhance Data Quality assessment in smart connected systems?

To answer this question, the selected articles are categorized based on the approach they follow. If the research employs some sort of ML or advanced statistical analysis, then it is grouped under the ML approach. If it uses simple statistics, human expertise, standards, business rules, classical frameworks or reports and dashboards, then it is grouped under the classical approach. According to this categorization, 22 of the articles fall under the ML group and the remaining 16 fall under the traditional approach.

To assess the strength of the studies, the scores allocated in the assessment matrix were added and the articles were grouped based on the total score obtained which is given in Figure 2-6. According to the results obtained, scores range from 8 to 19 from a possible 24-point score.

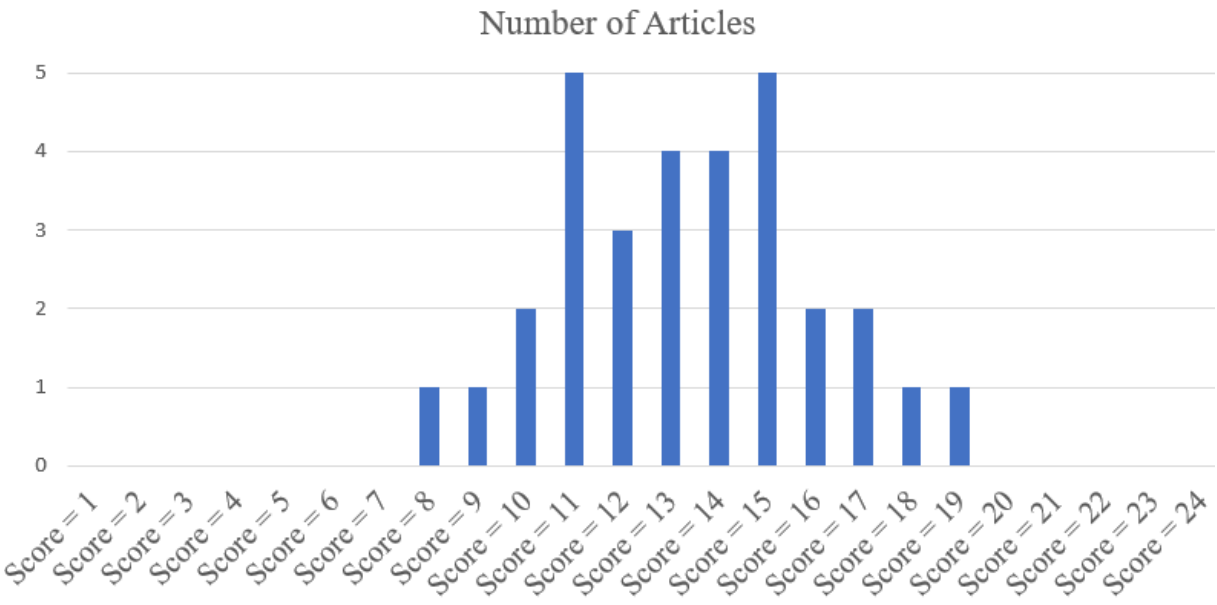


Figure 2-6 Number of selected articles according to the total score allocated (Distribution per score)

2.3.5 Discussion

This section has provided an overview of the different DQ assessment methods regarding smart connected systems using a systematic literature review. Statistics are presented according to the research questions.

It appears that even though the DQ assessment research has been there relatively for longer period, it is only recently that attention is given to smart connected systems DQ assessment. As shown in Figure 2-4, the number of articles published has increased in recent years. In addition, due to the interdisciplinary nature of the domain, the studies are sparse across multiple journals.

The findings also show that different DQ issues including accuracy, completeness, validity, timeliness, and consistency are subjects of investigation in the domain of smart connected systems and attracted researchers. However, studies are concentrated on the validity and completeness DQ dimensions.

The study also shows that different approaches have been investigated to tackle the issues including ML and statistical and classical frameworks. However, the majority have employed ML and advanced statistical techniques. This literature review also shows that most of the articles have employed simulated data which may not reflect the characteristics of the real-world situation. In this regard, there may be room for improvement.

Another important observation is the gap between the maximum attainable points and the points allocated to the selected articles is high. This is an indication that the articles focused on a specific issue which is only part of the problem or limited to the experimental stage. This is another gap which necessitates a comprehensive approach to DQ assessment.

2.4 Review of Data Quality Assessment Frameworks and Methods

Section 2.3 demonstrated that there exist many DQ assessment frameworks and methodologies in the literature, which may differ based on an organization's specific needs and data characteristics. These frameworks utilize various techniques, including rule-based approaches, adherence to standards, statistical methods, and advanced ML techniques (Mohammed *et al.*, 2020). For simplicity, they are categorized as classical and ML based DQ assessment frameworks and methodologies as described in section 2.3.4. In this section, a methodologically organized detailed review of DQ assessment frameworks and methods from each category is presented followed by a comparative analysis.

2.4.1 Classical Data Quality Assessment Frameworks and Methodologies

Many of the classical DQ assessment frameworks and methodologies draw inspiration from the work of Wand and Wang (1996), who conducted an extensive survey to define DQ dimensions. These methodologies, in some form or another, incorporate the DQ dimensions identified in their research. As such, the development of DQ assessment frameworks can be traced back to 1998 when Wang introduced "Total Data Quality Management (TDQM)", which has gained widespread acceptance in various domains (Pipino, Lee and Wang, 2002). TDQM has also found applications in the context of big data and connected systems (Batini *et al.*, 2015; Rasta, Nguyen and Prinz, 2013). One of the main advantages of TDQM is its iterative approach to DQ management, emphasizing a systematic and integrated approach. TDQM operates by having data users specify their requirements, which are then translated into measurable DQ dimensions by data engineers or information product engineers. The validation of these requirements against the engineers' output is conducted using expert knowledge. TDQM proposes 15 DQ dimensions that may be effectively utilized in several domains and have garnered substantial acceptance within the field (Wang and Strong, 1996). Wang and Strong (1996) recommend that when implementing TDQM, an organization should adhere to the following steps: (i) gain a comprehensive understanding of the information production process; (ii) form a DQ team comprising a senior manager serving as the TDQM champion, a DQ engineer well-versed in the TDQM approach, and participants representing information providers, producers, users, and

data managers; (iii) provide training on IQ evaluation and management to all stakeholders involved in the information product; and (iv) establish a culture of continuous improvement in information production. The main strength of TDQM is that it tries to address DQ throughout the entire data life cycle, from creation to usage. However, it can be complex and time consuming to have a complete understanding of the entire data life cycle posing a challenge to implement it effectively. Building upon TDQM, Lee *et al.* (2002) proposed the "A methodology for information quality assessment (AIMQ)", which classified DQ dimensions into four distinct groups: intrinsic, contextual, representational, and accessibility. AIMQ employs surveys to collect data about the organization's information quality level to benchmark in four quadrants by analyzing the survey's response. AIMQ identifies two broad roles in the DQ assessment namely IT professionals and users. AIMQ's benchmarking approach helps organizations to compare their performance against best practices. The development of questionnaires and surveys to measure IQ is also considered as a strength of the framework as it provides a standardized tool for DQ assessment. However, AIMQ is challenging for implementation as it is less adaptable, and it relies on manual processes significantly. AIMQ's application to large organizations may also limit its applicability to smaller organizations with fewer resources.

Also, many other methodologies have been developed rooted in the principles of TQDM. TDQM itself evolved into "Total Information Quality Management (TIQM)" (Cichy and Rass, 2019), emphasizing the involvement of the entire organization in the continuous improvement process of DQ (Francisco *et al.*, 2017). Both TDQM and TIQM view information as a product, highlighting the importance of managing DQ throughout its lifecycle.

Another framework known as the "Hybrid Information Quality Management (HIQM)" framework is designed to address real-time DQ issues (Cappiello, Ficiaro and Pernici, 2006). The HIQM methodology involves several key steps. Firstly, it defines the objectives and scope of the information quality assessment, including identifying the information domains, sources, and stakeholders involved. Next, quality criteria and metrics are developed to evaluate the information, considering factors such as accuracy, timeliness, completeness, and consistency. The methodology includes collecting and analyzing the relevant information, using techniques like data profiling, sampling, and auditing to identify DQ issues and their root causes. Based on the result, recommendations for improvement are provided, which may involve data cleansing, standardization, or enrichment processes. HIQM emphasizes the importance of data governance and stewardship, and it establishes monitoring mechanisms such as DQ dashboards and

automated checks to ensure ongoing DQ assessment and improvement. Continuous measurement and feedback are also emphasized, with the establishment of performance indicators and regular reviews of the impact of DQ initiatives. While HIQM brings major improvements to classical DQ frameworks including real-time monitoring, continuous improvement, and recovery; its focus on large organizations, its complexity, and the manual processes it involves pose a limitation to its adoption.

The "Comprehensive methodology for data quality management (CDQ)", developed by Batini *et al.* (2008), is yet another framework which aims to provide a complete framework by integrating and enhancing existing approaches. CDQ emphasizes the importance of DQ education and training, promoting a culture of DQ awareness and accountability within the organization. It suggests different DQ dimensions based on data structure, including accuracy, completeness, and currency for structured and semi-structured data, and dimensions like condition and originality for unstructured data. CDQ encourages selecting dimensions based on observed DQ issues in the organization. The methodology encompasses phases such as understanding the data, defining quality requirements, measuring, and assessing quality of data, improving DQ, and controlling DQ level. Continuous improvement and measurement are emphasized, with the use of measurable performance indicators and a feedback loop.

The "Heterogeneous Data Quality Management (HDQM)" is a framework specifically designed for managing DQ challenges and issues in heterogeneous data sources (Carlo *et al.*, 2011). It extends the CDQ methodology to address diverse data types, formats, and structures. HDQM focuses on data integration and harmonization, mapping and transforming data from various sources into a unified format or data model. Quality metrics and criteria are defined based on data types and structures, assessing data accuracy, completeness, consistency, and other relevant aspects. Data profiling is executed to identify areas for improvement, and techniques such as data cleansing, transformation, enrichment, or data source-specific techniques are suggested. Monitoring mechanisms, data governance policies, and continuous improvement are emphasized, along with measurable performance indicators and a feedback loop. The main strength of HDQM is its meta-model approach which provides a structured approach to integrating and understanding various data types. It also focuses on stakeholder involvement ensuring that the DQ management process is comprehensive and inclusive. However, it also has the common limitations available in other classical DQ assessment frameworks including manual processes, lack of flexibility and complexity.

The “Data Quality Assessment Framework” (DQAF) developed by Sebastian-Coleman (2010) provides an integrated approach to comprehensive DQ management. It includes five objective dimensions of DQ including completeness, consistency, accuracy, timeliness, and integrity. DQAF involves understanding the data, defining quality requirements, measuring, assessing, improving, and controlling DQ level. It is widely used in the industry and emphasizes practicality, effectiveness, and simplicity. Focusing on objective DQ dimensions and having a clear focus on well-defined metrics makes it easy for implementation. However, it fails to handle other DQ dimensions and lacks good data governance.

The “Observe-Orient-Decide-Act methodology for DQ assessment (OODA DQ)” is another framework which is inspired by the OODA loop and proposes a cyclical and adaptive approach to DQ management (Sundararaman and Venkatesan, 2017) in four stages which are observe, orient, decide, and act. Data profiling, monitoring tools, feedback, and metrics are used to observe the current state of DQ status. Analysis and interpretation are performed to gain insights and understand patterns and trends (orient stage). Informed decisions are made to address DQ issues (which is the “decide” stage), and appropriate actions are taken (which is the act stage). The process is repeated iteratively, allowing for adaptability and refinement of the DQ management strategy. The OODA loop adds flexibility hence making the framework adaptable and it provides an opportunity for situational awareness. It also gives flexibility and simplicity. However, it may lead to information overload during the “observe” stage and bias during the “orient” stage making it challenging for implementation.

Another framework known as the “Task-Based Method to Data Quality Assessment and Improvement (TBDQ)” provides a structured and task-oriented approach to managing DQ (Vaziri, Mohsenzadeh and Habibi, 2016). It involves steps such as task identification, requirement gathering, data profiling, assessment, improvement, monitoring, and feedback. Based on the outcome, prioritization is made to align with tasks.

There are also frameworks developed based on statistical methods. To this end, the Six Sigma DQ framework can be mentioned (Pyzdek and Keller, 2014). This methodology is a data-driven framework that emphasizes continuous improvement and defect reduction. It focuses on reducing process and product variability to achieve high-quality outcomes. The Six Sigma methodology employs a systematic approach to the assessment of DQ, covering several stages such as problem identification, measuring the current state, analyzing data, enhancement of processes, and process control. It uses statistical tools and methodologies to gain insights into

root causes of DQ problems. The six sigma DQ framework focuses on defect reduction and process improvement which may lead to a better DQ but fails to address all aspects of DQ.

Frameworks are also developed based on standards. For example, ISO 8000 is a set of international standards for DQ management (Standardization, 2022). ISO 8000-61 specifically guides assessing and measuring DQ attributes using appropriate techniques and tools. It emphasizes the importance of defining DQ requirements, establishing quality metrics, and conducting periodic assessments for ongoing improvement. Another ISO standard called ISO/IEC 25012 (Gualo *et al.*, 2021), which emphasizes that data lives longer than the software life cycle, specifies a DQ model to be able to define and assess DQ requirement in data generation, capturing and usage but also to define DQ acceptance criteria and compliance with regulations. According to this model, identification of any DQ issue should lead to an improvement of underlying components including data, software, hardware, human and processes. The main benefit of ISO based DQ frameworks is ensuring consistency across industries by providing standards and comprehensive guidelines. But standards and guidelines can be rigid making it challenging to adapt to specific organizational needs.

2.4.2 Machine Learning based Data Quality Assessment Methods

In today's complex multiple inter-connected systems such as CV, the generic frameworks may not capture all the required metrics of DQ (Cichy and Rass, 2019). Some DQ dimensions such as accuracy are also inherently difficult to assess. As stated by (Sebastian-Coleman, 2012), the best way to prove that a piece of information is correct is to compare it with some trusted source, that is a source which is correct all the time. Such sources may not exist or at least may not be readily available (Maydanchik, 2007). To illustrate this with an example, assume that an organization needs to verify the ages of its 4,000 staff members. This could be done by checking necessary documents showing birthdates. But if such a document does not exist, it may be necessary to call and ask each individual staff member which is difficult and costly for a large scale. This issue is compounded in the era of connected systems where a big volume of data is generated, multiple layers are involved and the data is of diverse types which includes timestamps, device data, location data and others (Kim *et al.*, 2019). Therefore, classical DQ assessment methodologies alone may not work as a proper DQ assessment method for smart connected systems. For example, Ricardo Perez-Castillo *et al.* (2018) stress that DQ in IoT is subjective and recommends applying context-specific DQ assessment methods. Some work has been done to adapt the generic frameworks to modern IoT applications. For example, in the

works of Alrae, Nasir and Abu Talib (2020), a new framework called the "House of Information Quality framework for IoT systems", by correlating DQ dimensions with the technical aspects of IoT, is proposed. This method depends on expert opinions for validation which is difficult to apply on high volume, continuous, and disparate sourced systems (Cichy and Rass, 2019).

On the other hand, developments in big data and ML have presented new opportunities for an automatic DQ assessment and improvement such as detecting outliers, inaccurate and incorrect values, and imputing of missing values (Gudivada, Apon and Ding, 2017). The application of ML for assessing and improving DQ is increasingly prevalent, particularly in IoT-based smart systems. For instance, clustering algorithms can identify similar records and highlight potential duplicates, while classification models can detect patterns and relationships within the data to identify inconsistencies or predict DQ issues. Realizing this fact, some researchers have employed ML methodologies for DQ assessment. Some of the methodologies that applied ML are reviewed as follows.

The use of ML in the assessment of DQ has been subject to investigation since as early as 2003, as indicated in the research conducted by NASA (Isaac and Lynnes, 2003) in developing an automated quality assessment architecture powered by ML for earth science data. However, research in this context has grown with the development of big data and IoT systems (Karkouch *et al*., 2018). Many ML methods for DQ assessment have primarily focused on sensors and anomaly detection, which aligns with the validity dimension of DQ (Barnes and Hu, 2013; Diop *et al.*, 2017; Vasta *et al.*, 2017). For example, Rahman, Smith and Timms (2013) applied ML for marine sensor DQ Assessment by flagging the data as good, probably good, probably bad and bad. They emphasized that on the one hand DQ assessment is a critical aspect of any sensor based IoT systems, on the other hand getting a good DQ level in such systems is difficult. They also emphasized that unsupervised methods such as anomaly detection fail to meet expectations. Therefore, they proposed an innovative approach to flag the data. In the flagging scheme, they proposed a domain expert to normally inspect sensor values and give a class label. Then a supervised classification model would be developed using the labelled data. However, labelling substantial amounts of data by human experts turned out to be difficult. Therefore, they adopted a clustering method to identify the different class labels. To improve performance of the supervised classification model, they have proposed training multiple classification methods and applying a majority vote hence forming ensemble methods. They used six features, most of which are derived features including duration since last calibration and gradient filter to calculate the sudden changes in the readings between consecutive samples. They have used

Bayesian Network and Decision tree with 5-fold cross validation on the training dataset for evaluation. According to the report, the method achieved better performance compared to the state-of-the art of other bagging methods. The main contribution of the approach is that it demonstrated that combining different methods can be effective in assessing sensor data. In addition to using different classification models, they have applied clustering methods to perform labelling of the data. In addition, it showed that constructing relevant features specific to the domain is particularly important for the success of such methods. They also emphasized that the method can be applied to other sensors but the feature set to be used may be different according to the domain in consideration. The main limitation of the method is that the quality flags used, i.e., “good, probably good, probably bad and bad” are too generic to apply on other critical systems.

Laptev, Amizadeh and Flint (2015) developed a framework called "Generic and Scalable Framework for Automated Time-Series Anomaly Detection (EGADS)" to better detect anomalies in large-scale time series data. They argue that existing anomaly detection methods are inadequate to handle large-scale time-series data, are not scalable, are limited by specific use-case constraints and tend to produce a high rate of false positives which makes them unfit for reliable DQ assessment of large and dynamic datasets. The proposed framework integrates a collection of anomaly detection algorithms and forecasting models, enhanced by an anomaly filtering layer to reduce false positives. First, the anomaly detection method applies multiple models, such as statistical methods, ML algorithms, and deep learning techniques to identify potential anomalies. Then the forecasting models are used to predict future data points to compare against actual data, aiding in the detection of anomalies that deviate from expected patterns. Finally, the anomaly filtering layer is used to filter out false positives by cross-referencing detected anomalies with historical data and contextual information. They claim that they rigorously tested the framework on both real-world and synthetic datasets with varying characteristics. The experiments demonstrated a significant improvement in precision and recall by 50-60% compared to existing anomaly detection systems. This enhanced performance highlights the framework's ability to accurately detect genuine anomalies while minimizing false positives. The framework is designed to be scalable for large datasets and improves accuracy by using the extra filtering layer to remove false positives. It is also versatile in that it can be applied for time-series data of various domains. The framework is open sourced which promotes transparency, enabling researchers and practitioners to benchmark and improve upon the system. However, the framework is complex which poses technical challenges to

implement, and it is resource intensive. Besides, While the framework effectively detects anomalies, understanding the underlying reasons for the detected anomalies can be challenging, particularly with complex models like deep learning which undermines interpretability.

Another work which is developed to improve Intelligent Transport System (ITS) data employs a combination of unsupervised ML techniques and supervised methods including Random Forest and support vector machine (Megler, Tufte and Maier, 2016). They stated that the data collected from various sensors is often noisy and incomplete which affects sound decision making. Therefore, they proposed a novel approach to identify suspect data. Their proposed framework consists of two main ML components. First, using labeled data, they trained supervised ML methods to clean suspect data. The second component employs k-means clustering for outlier detection according to traffic patterns resulting in clusters such as congested and light traffic which are known as 'regimes'. Using this method, each observation is assigned to a certain cluster, and distance is calculated to the cluster. This framework helps to detect two possible anomalies, i.e., an anomalous record within each cluster and an anomaly from the whole cluster. The identified anomalies are then marked as "bad data". They incorporated the developed data cleaning method into travel time prediction, and they reported a significant improvement on the prediction result. The framework is scalable and efficient which can be used without human intervention and for large volume of data near real-time. However, like most of the ML based frameworks, it is resource intensive and complex. In addition, it is highly dependent on pre-defined metrics, for example completeness threshold of 95% as good, which requires frequent updates to reflect changes.

Wang *et al.* (2017) conducted research on utilizing deep learning techniques to detect wireless sensor drift, which aids in automatic sensor calibration and thereby enhances the quality of the data. These approaches aim to identify anomalies or deviations from expected patterns in sensor data, enabling the detection and mitigation of DQ issues related to validity DQ dimension. The main goal of the framework is to develop a calibration method without ground truth since it is difficult to rely on ground truth for large scale sensors over extended period which is difficult to find. This would help ensure DQ in various applications such as environmental monitoring and smart cities. The proposed method employs a novel deep learning method called the Projection-Recovery Network (PRNet) to perform blind calibration of sensor measurements online. It involves: (i) projecting the drifted data into a feature space, and (ii) using a deep convolutional neural network (CNN) to recover the estimated drift-free measurements. The authors deployed a 24-sensor testbed to evaluate the performance of their method and reported

an accuracy of 80% recovery rate compared to previous methods. However, the method is difficult to interpret and complex to implement.

Random Forest is one of the most employed ML methods for IoT DQ assessment as an outlier detection method (Liu *et al.*, 2017). One example is the works of Farooqi, Khattak and Imran (2018) where they used Random Forest regression to assess the accuracy of IoT data. In this work, historical weather data is used to train the Random Forest regression model. Using the results obtained from this model, rules are established. The proposed DQ assessment approach for operationalization is that after predicting the new value, if it does not comply with the rules established, then the data is considered inaccurate, and the recommendation is to remove this specific data. One of the strengths of this work is that the authors aim to develop a model that ensures DQ standards provided by ISO 8000. While it is reported that the method has improved DQ assessment in IoT compared to traditional rule-based systems, there was no empirical evidence or comparison result comparing it to other ML methods.

Advanced methods such as deep learning are also investigated in the works of Dai, Yoshigoe and Parsley (2018). They have combined deep learning and statistical control chart to improve DQ via outlier detection. They argue that traditional DQ assessment methods rely on user experience or predefined business rules. Therefore, they are often labor-intensive, inefficient and struggle with the complexity and volume of modern datasets. To solve this problem, they proposed a novel DQ assessment framework that combines deep learning and statistical control chart. Initially, a deep learning model is developed, and subsequently, the model is applied to the data to generate a predicted value. Then statistical quality control is employed to assess the disparity between the predicted value and the actual value. This method facilitates the visual identification of outliers. They tested the framework using an open salary dataset. According to the paper, the method detected anomalies with good accuracy, improving DQ assessment. The framework showed that integrating ML with statistical methods could significantly enhance DQ assessment. The other strength of the framework is automation which avoids human error. In addition, it is scalable to big data. Combining deep learning with control chart not only increases the level of accuracy but also facilitates validation as it gives the opportunity to ascertain that the detected anomalies are genuinely indicative of DQ issues. Besides, the control chart provides visual inspection capability. Although the actual application of this method to CV data was not undertaken, it represents a significant effort towards demonstrating the use of ML and advanced statistical techniques for big data DQ assessment. The proposed framework

also has some weaknesses. First, the use of deep learning makes it challenging to interpret the reasons behind some detected anomalies. It is also complex and resource intensive.

One of the comprehensive works that covers multiple DQ issues combining classical methods with advanced ML methods is found in the works of Shrivastava *et al.* (2019). The proposed method tries to address the challenge of ensuring DQ in the era of Big Data and IoT. The authors stated that with the exponential growth of data, traditional DQ management techniques struggle to keep up. Therefore, the authors aim to develop a scalable, automated, and interactive DQ advisor (DQA) that can efficiently handle large volumes of data and provide timely, accurate assessments by enhancing exisitng tradional DQ assessment methods. In this work, a DQ advisor is developed which helps in the assessment and improvement of DQ. The DQ advisor is supported by a visual inspection functionality so that experts can have a judgement on the proposed DQ improvement by the DQ advisor. Different modules are included in the framework. The general validator module ranging from simple null, uniqueness and duplicate checks to statistical functions such as correlation and summary enables users to assess relatively easier DQ issues. The AI and the Time series modules help to identify relatively complex DQ issues such as anomaly detection. One of the main strengths of the proposed method is it automatically generates dynamic executable graphs for performing data validations fine-tuned for a given dataset. However, the framework combines multiple modules which makes it complex and requires advanced level knowledge.

Tsai *et al.* (2019) applied Bayesian Principal Component Analysis on environmental sensors with a five-component architecture. The first component is a sensor correlation agent to identify the correlation among sensors and produce the Bayesian models for each sensor based on correlated sensors from which its value is predicted. The second component, which is the group analyzer, is used to group problematic sensors together and compute error estimation. Another component, which is called the System Recorder, is used to log the real sensor values and the error estimation values from the previous component and store that in a repository for further analysis. On the stored data, an algorithm is developed to detect erroneous values. The algorithm works according to a threshold set derived from the group in such a way that if the predicted value is beyond the set threshold, then it will be marked as an error value otherwise it will be marked as a "good value". The real sensor reading stored in the repository is used for validation. They measured the method using accuracy metric on 12 different sensors and reported from 49% to 98%. The method may result in high false negative and false positive depending on the threshold, which is its main limitation. To produce proper performance

evaluation, it is good to use another performance metric such as F1-score, which is not the case in this work.

Okafor, Alghorani and Delaney (2020) applied ML methods to assess DQ of low-cost environmental monitoring sensors. Specifically, they employed linear regression and artificial neural networks (ANN) for calibration of sensors. They emphasized the importance of feature selection to build an effective model, and they have investigated three different approaches including forward feature selection, backward elimination, and exhaustive feature selection. They reported that both learning algorithms trained on exhaustive feature selection performed well and ANN resulted in slightly a better outcome based on performance metrics MAE, RMSE and R^2. They applied the method for calibration of sensors O_3.

Azimi and Pahl (2021) Developed an ML augmented multi-layered framework to assess DQ to enable reliable Data-as-a-Service (Daas) offering to build continuous DQ management. They used traffic count data with weather data. In their framework, they have formulated three functions. The first is an estimator which they used to estimate the expected volume. For this function, they have employed supervised ML methods. To validate their method, they used historical data, i.e., data from previous years. Second, they built a predictor function to predict the expected volume of data for a required future date. For the predictor function, three features are employed namely temperature, count of vehicles, and weekday. The third function proposed is an adapter taking two features namely number of cars and speed. This is used to suggest some actions. They emphasized that while DQ assessment methods exist, there are situations where some DQ issues are not possible to detect or assess using the existing methods. Therefore, they proposed the ML based framework to fill this gap. In their framework, they tried to tackle DQ issues including completeness, timeliness, consistency, accuracy, and validity including format validity. In this regard, it is one of the most comprehensive ML DQ assessment frameworks. The method also emphasizes conducting root cause analysis enabling users to trace back to the origins of the error. However, the method is complex, and it also did not provide numerical evidence of performance measures.

Lesouple *et al.* (2021) developed an advanced version of the Isolation Forest algorithm called Generalized Isolation Forest for Anomaly Detection (GIF) to enhance the effectiveness and robustness of anomaly detection across various datasets. They argue that traditional anomaly detection methods, including the Isolation Forest algorithm, face challenges when dealing with complex data structures and are often sensitive to specific data artifacts. These limitations reduce their efficiency and accuracy, necessitating a more robust and adaptable approach to

anomaly detection. The GIF algorithm builds upon the traditional Isolation Forest by incorporating generalization techniques that improve robustness and accuracy. To enhance the robustness of the model, the authors introduce generalization methods that address common shortcomings of the Isolation Forest algorithm, such as sensitivity to data artifacts and inefficiencies in handling complex data structures. The authors conducted extensive experimental evaluations using both real-world and synthetic datasets with varying characteristics. The results demonstrated significant improvements in the performance of the GIF algorithm compared to the traditional Isolation Forest as reported in the article. Enhanced performance, robustness, scalability, and versatility are the main strengths of the framework. However, this method has its weaknesses including complexity, computational intensity, and less interpretability.

Han, Wu and Yang (2022) applied multiple supervised ML algorithms for anomaly detection including Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, and Artificial Neural Networks to identify DQ issues. They tested on two different datasets, one of which is IoT data. Measuring its performance based on F1-score, they reported that except Naïve Bayes, most of the ML algorithms performed well. Initially, they split their dataset on an 80%/20% basis for training and test, respectively. An important experiment they performed later is to apply different proportion of train/test split including 90%/10%, 70%/30%, 50%/50% and 30%/70%. One interesting observation is that even though Artificial Neural Networks and Support Vector Machine resulted in slightly better F1-score on average, Logistic Regression performed consistently well on the different splits. They also trained by changing the size of the dataset and F1-score of Logistic Regression was again consistent. They have made extensive validation, and they have also released the core code for practitioners. However, they were limited to supervised methods which may not cover the full spectrum of anomaly detection.

From all articles reviewed, it is demonstrated that no single method was enough to produce an effective assessment, rather a combination of methods is applied. In addition, different ML methods are used by different researchers including Logistic Regression, Artificial Neural Networks, time series and so on. One common observation is that most of the articles emphasized effective feature selection. They all focus on wisely selected limited but powerful feature sets. The most used method is anomaly detection by leveraging different ML methods such as Isolation Forest. From supervised methods, Logistic Regression, Random Forest, Artificial Neural Networks and Support Vector Machine are used and many of them combined

different ML algorithms. In terms of DQ metrics, many of the articles focused on limited DQ dimensions, specifically validity. Another crucial point is that there is always a trade off in complexity, performance, and accuracy. Some of the methods also did not use appropriate performance metrics. From the literature, it is also evident that various ML methods are explored for the different DQ issues. However, it is still not enough and most of the existing works do not have a comprehensive approach. The following section provides a comparative analysis of classical and ML based DQ assessment frameworks.

2.4.3 Comparative Analysis of Classical Data Quality Assessment Frameworks and Machine Learning Methodologies

The classical DQ assessment frameworks primarily measure metrics defined based on DQ dimensions tailored to requirements in context and aim to support multiple domains. For validation, they depend on expert opinion and hence incorporate subjective assessments. Notable differences to this are the DQAF and the six sigma DQ assessment frameworks which focus on objective assessments. While the Six Sigma framework uses statistical methods, DQAF employs metrics defined based on only objective DQ dimensions. This makes implementation much simplified. The classical frameworks also provide a more structured approach in general for measuring as well as monitoring DQ using scorecards and dashboards (Gitzel, Turring and Maczey, 2015) and they are easier to implement. However, they lack the means to handle complex topics. This is reflected in their reliance on expert opinion and those that use objective assessment, such as DQAF omit some DQ dimensions altogether. As an example, DQAF did not define a metric to measure the accuracy DQ dimension, instead it suggests using validity DQ dimension metrics as a proxy measure. In terms of data, only a few of the classical frameworks tackle unstructured and heterogeneous data. In addition, there is not enough depth of coverage given to real-time DQ assessment. Classical frameworks also place emphasis on the importance of defined roles when approaching the DQ assessment process. Even though, there exist differences in the established roles from framework to framework, most of the frameworks suggest the role of a senior manager who advocates for the DQ project (Cappiello, Ficiaro and Pernici, 2006; Pipino, Lee and Wang, 2002), a data steward responsible for defining DQ requirements, goals, and measurement approach (Cappiello, Ficiaro and Pernici, 2006; Sebastian-Coleman, 2012) and a DQ analyst tasked with conducting and executing DQ assessment activities (Cappiello, Ficiaro and Pernici, 2006; Pipino, Lee and Wang, 2002).

The ML based methodologies, on the other hand, tackle more difficult topics and do not primarily depend on expert validation. ML based methods act autonomously, adapt to changes easily and can scale to handle large volume of data. They can easily detect complex patterns and relationships in the data. Not only can they be used as an autonomous DQ assessment method, but they can also be used to generate rules to be used as an input for classical DQ assessment frameworks. ML based DQ assessment frameworks are particularly useful for IoT based smart connected systems since these systems generate huge volumes of data from various sensors which are susceptible to various DQ issues. However, ML based DQ assessment frameworks lack generalizability. Most of them focus on specific DQ dimensions and the majority deal with outlier detections. In addition, such methodologies are difficult to implement, resource intensive and complex to interpret. Moreover, they do not take a comprehensive approach. While most of the frameworks, both classical and ML based, agree on the multidimensional nature of DQ, there are variations in the number and specifications of dimensions. Therefore, standardization of DQ dimensions remains a gap in effective implementation (Fox, Levitin and Redman, 1994). Table 2-4 presents summary of the classical DQ assessment frameworks and ML based DQ assessment methods.

Table 2-4 Comparison of classical and ML based DQ assessment frameworks

Classical DQ Assessment Frameworks	**ML Based DQ Assessment Frameworks**
They are easy to implement and understand.	They can handle complex patterns and relationships.
They require lower cost as they do not use advanced computational resources and knowledge.	They are costly since advanced knowledge and resources are required, but they can also be cost-effective in the long run by automating repetitive tasks.
Results obtained are often straightforward and easy to interpret.	They can provide insights regarding difficult DQ issues using advanced methods
They are well established and widely used in various domains.	They are usually applicable to specific domains.
They may struggle with large datasets and complex data structures.	They require significant computational resources and expertise.
They are less flexible for different data types and changing DQ issues.	They can adapt to new data types and DQ issues.
They often require manual intervention and periodic updates.	They can automate DQ assessment process, reducing the need for manual intervention
They may not identify DQ issues not defined in the metrics list.	They can identify more DQ issues by detecting relationships.
They may not capture complex DQ issues.	ML models can be difficult to interpret and require advanced knowledge to implement.

2.5 Identification of Gaps on Existing Research

Research into connected systems DQ assessment is still in its early stage and recent research has begun to recognize the significance of emphasizing the concept of DQ. Most existing discussions do not take a comprehensive approach either. There are very few papers which investigate DQ assessment methods specifically focusing on CV to achieve an improved level of DQ, which is the main gap in the literature.

Looking at smart connected systems, it can be identified that there is a gap between smart connected systems and the quality level of data required by consumer systems. There is also a gap in the existing classical DQ assessment frameworks and DQ requirements by smart connected systems which bring new challenges such as high volume of data, real time requirement and autonomy without human intervention on which classical DQ assessment frameworks struggle. Although there have been several research works applying advanced methods such as ML, which are reviewed in section 2.4.2, on general connected systems, they might not work for CV and new techniques need to be developed as CV brings extra challenges and new requirements with both efficiency and effectiveness. The ones available either focus on a specific DQ dimension or they lack rigor. Especially, the extra dimension attributed to CV such as variation on space and time is not well investigated on the papers reviewed.

CV use cases such as predictive and preventive maintenance, location-based services and so on require real-time (currency/timeliness) and accurate (accuracy) data. However, not much research is available to tackle these DQ issues with enough depth even within the wider connected systems scope. Therefore, a study is required to address timeliness as well as accuracy DQ issues with a specific focus on connected systems. There is also not enough study done on the completeness DQ dimension. In the context of CV, it is possible for data to be missing at various stages due to the involvement of multiple components. Data loss can occur because of many event failures, including network disruptions, intrusion attacks, connection errors, errors in data transformation, and issues with data storage, among others. Therefore, a technique to tackle this DQ dimension is required.

In general, there are only few studies specifically touching on DQ assessment in CV. While many studies about CV are available, they only elaborate on the potential benefits of CV. But to unleash the potential, the DQ issues surrounding these systems should be tackled.

2.6 Conclusion

In this chapter, the most relevant literature was reviewed and evaluated to understand the research gaps step by step. First, background on important topics including CV and enabling technologies as well as DQ and its impact is presented. Then, with a particular focus on DQ on smart connected systems, a systematic literature review is presented. This part shows the gap in the present DQ assessment research work. Subsequently, a review of classical DQ assessment frameworks is presented and then frameworks that leveraged ML are reviewed. In summary, DQ issues in smart connected systems are identified and the gap that currently exists in the research is pointed out. Particularly, the absence of enough studies focusing specifically on CV was demonstrated. In addition, it was shown that even though advanced methods such as ML and statistical methods have been explored by some researchers, those studies only attempted to tackle only a few DQ dimensions. For example, the timeliness DQ dimension, which is an important DQ dimension for CV is not explored with enough depth. These research gaps motivate this thesis to seek a systematic solution for DQ assessment in CV, with the ultimate objective of achieving a more intelligent and integrated ecosystem that includes the DQ assessment and improvement for CV.

Chapter 3 : Methodology

3.1 Introduction

This chapter gives an overview of the research methodology employed in this study. It comprehensively describes the research method used in designing and evaluating the proposed **Machine Learning enabled Data Quality Assessment Framework**. It also describes the study's design according to the adopted methodology and the rationale behind the selection of specific research methods, techniques, and tools utilized in this research endeavor. The chapter is organized as follows: Section 3.2 introduces Design Science Research and provides the reasoning behind the selected methodology. Section 3.3 describes the strategy followed to accomplish the objective of this research. Sections 3.4 and 3.6 highlight data collection, and environment and tools used respectively. Section 3.5 highlights the ethical considerations. Finally, Section 3.7 presents the evaluation method used for the experiments and prototypes implemented and the proposed framework.

3.2 Design Science Research

Design Science Research (DSR) was chosen as the guiding research methodology since it is deemed to be suitable for the execution of this research project. The DSR methodology and approach is predominantly employed within the area of information systems and computer science (Hevner, 2007). The origin of DSR can be traced back to engineering the science of the artificial (Kotzé, van der Merwe and Gerber, 2015). The purpose of research is to address issues, enhance current solutions, or advance the body of knowledge. To this end, this study's main objective is to enhance the assessment of DQ by using a framework based on ML and advanced statistical techniques. The DSR approach was developed to help guide the study, since it facilitates the research procedures for several reasons.

First, the primary focus of DSR is the provision of a solution. According to Kotzé, van der Merwe and Gerber (2015), the main objective of DSR is to produce and enhance artefacts that can effectively address current problems or generate innovative solutions. An artefact refers to several entities inside a given context, such as an instantiation, method, model, or construct. As stated by Von Alan *et al.* (2004), the concept of DSR is distinguished by its emphasis on well-defined design, the process of inventing rather than discovering, a focus on purposeful objectives, the generation of value, and a pragmatic approach. In other words, the objective of DSR is to generate novel and valuable artefact (utility). To do this, it is important to implement ongoing evaluation. The significance of novelty is especially evident in the context of DSR, as it relates to the effective resolution of unresolved problems or known solutions in a better and

efficient way. According to Von Alan *et al.* (2004), it is essential to provide a comprehensive and structured depiction of the artefact, employing a formal representation that is both coherent and consistent. The DSR methodology prioritizes the construction of a problem-solution space through implementing a well-defined and efficient approach to attain the optimal solution. In addition, it is important to establish efficient communication means with relevant stakeholders, including implementers and customers.

Second, the DSR methodology adheres to seven well-defined guiding principles, as proposed by Von Alan *et al.* (2004). These principles include artefact generation, relevance, evaluation, contribution to knowledge, rigor, optimal search for solution, and communication. According to Von Alan *et al.* (2004), it is emphasized that the research process should involve a constant and iterative interaction between processes and artefacts in order to effectively address complex challenges and provide valuable outcomes. Subsequently, it is important to conduct an evaluation that provides feedback to facilitate a deeper comprehension of the issue at hand, hence leading to potential improvements and refinements in the product. The iterative process persists until an enhanced and validated artefact is achieved. Therefore, evaluation plays a crucial role when DSR is applied as a research methodology.

Finally, the DSR aims to attain the following potential contributions as outlined by Von Alan *et al.* (2004).

- A well-defined and clearly stated problem.
- A clear evidence that there is no optimal solution available for the problem.
- Formulation, construction, and presentation of an artefact aimed at resolving the identified problem or improving a target artefact.
- A thorough evaluation of the proposed or developed artefact.
- A detailed explanation of both the theoretical and practical value and use the artefact offers.
- An effective communication on the impact of the artefact to stakeholders.

Other methodologies were also explored including action research and applied research. DSR was selected as a good fit for this research project as it provides the required structure to enable the artefact created to be evaluated in iterative cycles.

In conclusion, the utilization of DSR is deemed suitable for this study due to its problem-solving approach, which seeks to generate and evaluate novel solutions to practical challenges. DSR is an approach that integrates rigorous scientific inquiry with the development of tangible

outcomes, such as models, methods, processes, or systems, to tackle real-world problems and enhance current practices.

3.3 Strategy

Before delving into the strategies defined for this research, it is important to reiterate the objectives set for this research project, which are:

- Investigating Data Quality issues and challenges in CV.
- Investigating existing Data Quality assessment methodologies.
- Developing **Machine Learning Enabled Data Quality Framework to assess connected vehicles data.**

Therefore, to accomplish the study objectives, a strategic plan was developed in accordance with the principles of DSR. The study was structured into three iterations.

- **Iteration 1**. **Classical Data Quality Assessment Framework Adoption and Prototype Dashboard Development:** The first iteration involved adopting a classical DQ assessment framework informed by the literature review findings. Subsequently, a prototype of the adopted DQ assessment framework was developed and implemented as a DQ dashboard. The prototype of the initial framework from this iteration was deployed followed by a thorough evaluation. One of the major outcomes of this iteration was identifying gaps or limitations of the initial prototype framework with respect to metrics defined for CV based on DQ requirements.
- **Iteration 2. Enhancement of the Framework using ML and Advanced statistical methods**: This iteration employed case studies, known as scenarios. At this stage, some critical data elements were selected informed by the findings from Iteration 1, and ML methods were applied to fill gaps identified in the second iteration.
- **Iteration 3. Proposed Framework for Connected Vehicles Data Quality Assessment**: In this iteration, the findings from the literature review, the initial adopted framework from iteration 1 and the enhancement of the initial framework with ML from the second iteration are combined and a new framework which is fit for a better DQ assessment of CV data is proposed.

As stated earlier, the third iteration aims to enhance classical DQ assessment frameworks by incorporating ML methods and advanced statistical techniques. To demonstrate this, the following three scenarios were developed and evaluated.

A. Scenario I: Detecting missing data or delayed data - Completeness and Timeliness Data Quality dimensions

In the context of a CV system, the timely flow of information may be hindered by a range of factors, including communication problems. Consequently, the data remains held within the embedded device until the communication problem is resolved. Nevertheless, the storage capacity of the embedded unit is limited. For example, in the organization where this research was conducted, the embedded unit of the connected system implemented can only store data for a maximum of 2 weeks. If the problem's duration exceeds the storage capacity of the embedded unit, older information will likely be erased. This leads to two main problems:

1. Unavailability of information for timely utilization if information is delivered delayed, which affects timeliness DQ dimension, or
2. Total information loss, which is the complete absence of information, which affects completeness DQ dimension.

Early detection of the problem enables the implementation of preventive actions. Nevertheless, detecting this phenomenon proves to be challenging for human beings or basic rules. Hence, this research aims to examine ML methodologies, with special attention on classification learning methods, specifically logistic regression, to analyze historical data to develop a mechanism that detects the issue as early as possible.

B. Scenario II: Forecasting Mileage - Completeness Data Quality dimension

This is a continuation of the first scenario where the focus shifts to forecast the missing data detected in Scenario I. Due to several reasons throughout the data flow process, data can be lost before it gets to the target systems. This will have negative consequences in the decision-making process. To identify instances of missing data, it is possible to formulate rules. Nevertheless, it is challenging to identify all rules especially in a CV ecosystem where spatio-temporal variation is big, and the process of handling missing data presents a multifaceted challenge. Hence, the exploration of ML approaches is undertaken to forecast missing mileage. Specifically, time series is employed to forecast missing mileage.

C. Scenario III: Detecting inaccurate values – Accuracy and validity Data Quality dimensions

Data may deviate from actual values due to several reasons, including computation errors, device failures, signal interference, and fraudulent activities. The complex characteristics of the

CV ecosystem pose various challenges in identifying these anomalies. Hence, this work aims to investigate the application of ML techniques to assess the credibility of the incoming data or reported data values. For this scenario, fuel consumption data element was selected and regression models including Light Gradient Boosting Machine (LightGBM) and Random Forest were employed to predict the likely value. Further statistical quality control was applied to detect inaccurate values by comparing reported values against predicted values.

Throughout each iteration, a range of methods were used to achieve the defined objectives. Table 3-1 below shows the target objectives, and the methods used per iteration.

Table 3-1 Objectives and methods per iteration of the thesis

Iteration	**Objectives**	**Method**
Iteration 1: Initial framework development	Investigating DQ issues and challenges in CV	(Systematic) Literature Review Observation Document Review
	Investigating existing DQ assessment frameworks and methodologies	(Systematic) Literature Review
	Adopting a framework and developing prototype dashboard	Implementing adopted framework from candidate frameworks identified in the literature with a dashboard
Iteration 2: Enhancement of the framework using ML and statistical methods	Evaluating ML methodologies to improve DQ assessment	ML (DBSCAN, Logistic Regression, LightGBM, Random Forest) Time series Statistical Control Chart
Iteration 3: Developing ML enabled DQ assessment framework for CV data	Developing ML enabled DQ framework to assess CV data	Combining classical DQ assessment frameworks and ML methodologies

Each iteration follows the three cycles outlined by Hevner (2007), which include relevance, rigor, and design. Additionally, an evaluation is conducted, which may need the revision of the requirements analysis or the entire restart of the process if the artefact fails to match the

specified requirements and expectations. Within each iteration, the five step DSR approach proposed by Vaishnavi, Kuechler and Petter (2004) is applied as described below.

- Step 1 - Awareness: This step involves raising awareness by recognizing the existence of a problem; and acknowledging the necessity to create an artefact and formulate a theory to devise a solution.
- Step 2 - Suggestion: Drawing upon existing knowledge or theoretical frameworks, a potential solution is put forth.
- Step 3 - Development: An artefact is created based on the suggestion proposed.
- Step 4 - Evaluation: The artefact is subjected to evaluation.
- Step 5 - Conclusion: The artefact has been thoroughly documented and effectively communicated.

Finally, drawing up on the outcomes of these iterations, a **Machine Learning enabled Data Quality Assessment Framework** is proposed which combines classical DQ assessment frameworks and ML techniques.

3.4 Data Collection and Analysis

To address the proposed research inquiries, it is important to gather and analyze relevant data. Hence, acquiring data holds significant importance in research. In this study, two primary data sources and a business requirements document have been utilized.

1. Real-Life Connected Vehicles Data

The primary data source employed in this study is real-life CV data. The data is obtained from connected trucks through the DAF Connected system solution. The data is stored within Snowflake, a cloud-based data warehouse that is hosted on the Amazon Web Services (AWS) platform. The connected system is used to gather data in the form of real-time messages. There exist three distinct types of messages.

i. Monitoring messages: these types of messages are created when the connected truck sends messages every time a predefined event occurs. The event can be a technical problem of the vehicle, for example, when AdBlue which is a diesel exhaust fluid is too low, when a driver changes working state such as from work to available or from work to rest and so on. Each message includes data elements about the vehicle when the event is captured, such as speed, external conditions including ambient temperature, and GPS information.

ii. Index messages: these types of messages are created when the vehicle sends data from key-on until key-off. The vehicle continuously sends information based on predefined time intervals, currently set to a time ranging from 1 to 5 minutes. The message contains detailed and predefined information such as distance covered, fuel used, total weight carried, speed of the vehicle, location of the vehicle and so on.
iii. Status messages: in addition to monitoring and index messages, aggregate information is sent at the end of a trip, i.e., when the journey of the complete trip ends and hence the engine is turned off. In terms of content, this is equivalent to a summary or aggregate of index messages of the corresponding journey.

For this research, only selected trip related data from index messages and status messages are used. Table 3-2 below provides description of the data elements from the DAF Connect system used in this research.

Table 3-2 Description of selected data used from real-life connected system

Data Element	Description	Example Value
VID	Vehicle ID	1
TRIPID	The trip/trajectory identifier	1669398b-5c8a-48ef-ba18-abd937a01808
DATETIME_BEGIN	Start time of the trip	2019-11-02 10:00:05
DATETIME_END	End time of the trip	2019-11-02 10:20:55
TOTALDISTANCE_BEGIN	Cumulative mileage of the vehicle at the start of the trip in meters	1,7404,210
TOTALDISTANCE_END	Cumulative mileage of the vehicle at the end of the trip in meters	1,7407,480
DISTANCEDONE	The distance covered by the trip in KM	3.27
AVG_SPEED	The average speed of the trip in KM per HR	6.89
BRAKE_DURATION	The total brake duration during the trip in seconds	92
TRIPDURATION	The total time duration of the trip in seconds	1,708
BRAKE_DURATION_RATIO	The ratio of the brake duration to the total trip duration	0.05
HARSHBRAKE_DURATION	The harsh brake duration of the trip in seconds	8
HARSHBRAKE_DURATION_RATIO	The ratio of the harsh brake duration to the total trip duration	0.01
IDLING_DURATION	The idling duration of the trip in seconds	1,124
IDLING_DURATION_RATIO	The ratio of the idling duration to the total trip duration	0.66
GPS_ELEVATIONLOSS	The total elevation loss in meters during the trip	11
GPS_ELEVATIONGAIN	The total elevation gains in meters during the trip	30
PTO_COUNT	The number of power take-off events in the trip	0
PTO_DISTANCE	The total distance covered while power take off being set active	0
PTO_DISTANCE_RATIO	The ratio of the distance covered while PTO is active to the total distance covered	0
PTO_DURATION	The total duration while PTO is active in seconds	0
PTO_DURATION_RATIO	The ratio of the PTO duration to the total duration	0

Table 3-2 Continued

Data Element	Description	Example Value
TOTALFUELCONSUMPTION_BEGIN	The cumulative fuel consumed at the start of the trip in ML	2,994,673
TOTALFUELCONSUMPTION_END	The cumulative fuel consumed at the end of the trip in ML	2,996,273
FUELCONSUMED	The total fuel consumed in the trip in L	1.6
IDLING_FUELCONSUMPTION	The total idling fuel consumed in ML in the trip	709
ACCELERATION_DURATION	The total time duration of acceleration in the trip in seconds	284
ACCELERATION_DURATION_RATIO	The ratio of the acceleration duration to the total trip duration	0.17
MAXTHROTTLEPADDLE_DURATION	How long does a throttle/accelerator pedal position sensor last in seconds	1
MAXTHROTTLEPADDLE_DURATION_RATIO	The ratio of throttle/accelerator duration to the total trip	0.00
DPABRAKINGSCORE_SUM	The sum of the driver performance braking score point	0
DPAANTICIPATIONEVENT_COUNT	The sum driver performance assistant anticipation points	0
CRUISECONTROL_DISTANCE	The total distance travelled while cruise control on	0
CRUISECONTROL_DISTANCE_RATIO	The ratio of cruise control distance to the total distance	0
CRUISECONTROL_FUELCONSUMPTION	The total fuel consumed while cruise control on	0
FUEL_INDEX	The fuel consumption in liter/100km (L per 100 KM)	48.93
GROSSCOMBINATIONWEIGHT	The total carried weight in the trip in tons	7.7

2. Public dataset

To ensure the reproducibility of the results, a publicly available dataset is utilized in scenario III. The dataset used in this study consists of fuel consumption data obtained from European buses. The data was acquired using sensors and was extracted from a publicly available GitHub repository (Rosameo, 2021). Table 3-3 provides a detailed description of the dataset, including the specific data elements collected.

Table 3-3 Data elements of bus public dataset for fuel consumption (Rosameo, 2021)

Field name	Example value	Description
Date-time	2019-01-14	Trip datetime
VehicleID	0	identifier of the vehicle
avg_slope	0.01	average slope of the path
Mass	19.61	mass in ton of the vehicle including passengers
aircond_ptime	0	percentage of travel time with air conditioning on
stop_ptime	0.13	percentage of the travel time with the vehicle stopped and with the engine on

3. Business Requirement Document

In addition to the datasets described earlier, various documents are consulted. Specifically, the Business Requirement Document (BRD), which describes the high-level functional and non-functional requirements of the connected system, is reviewed. This document was helpful to extract the DQ requirements of the CV solution implemented in DAF. While this document consists of multiple sections regarding the functionality of the connected system, the section that describes about quality is given more attention for this research. Specifically, the following requirements were extracted.

- All messages generated should be available. There should not be missing data.
- 100% of the trips should have no delay, trips with a delay of equal or less than 15 minutes are not considered as delays.
- All data should be available as generated from the vehicle (without any manipulation). Information received should be accurate.

Measures defined based on these high-level requirements served as the basis to develop the initial prototype dashboard for Iteration 1.

Furthermore, in addition to historical data from the connected system, DAF Vehicle Information Electronics (DAVIE) is used. The DAVIE tool is used to read actual values from the vehicle on demand and this value is compared to the predicted values. This is used for validating Scenario III as ground truth by reading fuel consumption value, which is the target variable for Iteration 2.

3.5 Ethical Considerations

Since the primary data source for this research is real-life CV data, many measures are taken to avoid privacy concern in accordance with GDPR.

- Data elements that can potentially be used to identify an individual are excluded.
- Only aggregate information is displayed in dashboards and reports.
- When the information shown in the dashboard is considered as sensitive, the actual information is hidden
- Only relevant data elements are extracted.
- For Scenario III, only test vehicles are used, i.e., vehicles owned by customers are excluded.

- When the data element is necessary for reproducibility and when it is identified as privacy sensitive, the actual value is replaced with anonymized data.

3.6 Techniques (Environment, Tools, and Libraries)

Various tools and libraries are used to extract, load, analyze, model, and visualize the collected data. Table 3-4 illustrates the environment, tools, and libraries employed to fulfill various objectives in order to accomplish the overall study objective.

Table 3-4 Environment, tools and libraries used

Purpose	**Tool/Library**	**Environment**
Data extraction from snowflake database	Python Structured query language (SQL) snowflake.connector snowflake.sqlalchemy	AWS
Data Exploration, Visualization, Dashboarding	Python Pandas, Plotly, Tableau, Power BI	Local instances/AWS
Feature Construction	Python, Geopy, shapely	AWS (SageMaker - ml.t2.2xlarge)
Clustering	DBSCAN from Sklearn.cluster	
Classification	GridSearchCV, train_test_split , RepeatedStratifiedKFold from sklearn.model_selection Pipeline from sklearn.pipeline LogisticRegression from sklearn.linear_model f1_score, make_scorer from sklearn.metrics RobustScaler from sklearn.preprocessing	
Time series	Pmdarima ARIMA, acf, adfuller, kpss from statsmodels.tsa.stattools seasonal_decompose from statsmodels.tsa.seasonal plot_acf, plot_pacf from statsmodels.graphics.tsaplots lowess from statsmodels.nonparametric.smoothers_lowess lag_plot from pandas.plotting	
Regression	Pandas LightGBM, RandomForest from Pycaret IsolationForest Matplotlib	

The datasets described in Table 3-2 and Table 3-3 were extracted, and then put into the selected environments, analyzed, and ultimately presented in the form of a dashboard or graphs. To

address the research inquiries, suitable models are constructed, and the findings are examined using the tools and techniques described in Table 3-4.

3.7 Evaluation

To verify the validity of the research outcomes, various methods are employed depending on the type of activities undertaken. In this research, only objective evaluation is employed. The various evaluation methods utilized in this research are described as follows.

3.7.1 Iteration 1 - Initial Data Quality Dashboard based on DQAF

The initial dashboard was developed by adopting a classical DQ assessment framework from the literature and defining metrics extracted from available requirement documents in accordance with the adopted framework. Therefore, the validation method is developed based on the inputs from literature and requirement documents. The following steps are taken into consideration to develop the evaluation method.

1. Informed argument: - A claim for the artefact's utility is established by applying support or evidence from the knowledge base (the literature).
2. Requirement versus artefact's utility: - Comparing requirements against the developed artefact, in this case, the prototype dashboard.
3. Scenarios: - Extended scenarios to prove the artifact's utility is developed.

Specifically, the initial prototype dashboard is validated based on answers for the following objective questions with respect to fulfilling the requirements.

- Were all metrics defined based on the selected DQ dimensions included? Or does the developed dashboard answer all the questions for the defined metrics?
- Were the results conclusive? In other words, is there any doubt on the result displayed on the dashboard?

3.7.2 Iteration 2 - Enhancement with Machine Learning and Statistical Methods

The enhancement of the initial framework employs various ML and statistical methods. Therefore, respective performance metrics for each method are applied. Planning appropriate performance indicators is crucial prior to the implementation of any ML model. The performance metrics may vary based on the specific model type, such as classification or

regression, and the contextual factors, such as the presence of imbalanced data sets. The performance metrics applied for the three scenarios are described as follows.

3.7.2.1 Scenario I: Classification

The first scenario employed classification algorithm, specifically logistic regression. Therefore, classification performance metrics are used. In classification performance evaluation, it is important to start with the following four significant concepts (Sokolova, Japkowicz and Szpakowicz, 2006).

- True Positive (TP): refer to instances where both the prediction and the actual outcome align as Yes.
- True Negative (TN): refer to instances where both the prediction and the actual outcome align as No.
- False Positive (FP): refer to instances that were predicted as Yes and those that were actually observed as No.
- False Negative (FN): refer to instances where a prediction is made as No, but the actual outcome is Yes.

Having this context, one or more of the following performance measures can be applied depending on situations and preferences.

Accuracy

One of the comprehensible metrics in classification is accuracy, which may be defined as the proportion of accurate predictions relative to the total number of predictions, represented as a percentage and mathematically given as:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \qquad 3.1$$

Recall

Recall is another essential metric expressed as the proportion of instances from a category correctly predicted using the classification method and mathematically represented as.

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP} + \text{FN})} \qquad 3.2$$

Precision

Precision is yet another important metric that is determined by dividing the count of accurately predicted positive cases by the overall count of positive instances and can be calculated using the following formula.

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})} \qquad 3.3$$

F1- score

The F1-score, referred to as the precision and recall modulation index, is a widely accepted metric that integrates accuracy and recall. It is represented mathematically as follows.

$$\text{F1} - \text{score} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \qquad 3.4$$

Confusion matrix

Classification measures (TP, TN, FP, FN) are frequently depicted in a confusion matrix, which serves as a tabular representation of the comprehensive performance of the model, and from which the performance metrics can be calculated.

Receiver operating characteristic curve (ROC) and Area under Curve (AUC-ROC)

The process of generating the ROC curve involves the plotting of the recall, i.e., the true positive rate against the false positive rate, at various threshold levels. The region below the ROC curve is often represented as the area under the curve (AUC). A higher rate of AUC indicates better performance of the classification model.

In general, according to Sokolova, Japkowicz and Szpakowicz (2006), the selection of performance metrics is dependent upon the specific objectives and characteristics of the given situation. The measure of accuracy is appropriate for datasets that are balanced, but it may result in misleading results when dealing with imbalanced classes. The F1-score is a metric that effectively combines precision and recall, making it particularly valuable in scenarios involving imbalanced classes and the optimization of threshold values. The ROC-AUC metric gives valuable insights into the discriminatory capacity of a model across various thresholds. It is particularly well-suited for imbalanced datasets. However, it may not comprehensively capture

the performance of individual classes. It is commonly wise to employ various metrics to have a comprehensive understanding of the strengths and limitations of a classification method.

In this research, it was apparent that the dataset used for training exhibited a significant class imbalance. Hence, despite the examination of other performance measures such as accuracy, the F1-Score was employed as the primary performance metric for evaluation of the developed classification algorithm by applying on a validation and test dataset that was kept from the historical data.

3.7.2.2 Scenario II: Time series

The second scenario was based on time series forecasting**.** In time series forecasting, to facilitate model comparison, one can utilize statistical measurements such as the Akaike Information Criterion (AIC) and/or the Bayesian Information Criterion (BIC) (Cerqueira, Torgo and Mozetič, 2020). Lower scores for these criteria would imply a more optimal model fit.

AIC:

The AIC was formulated by Hirotsugu Akaike (Cerqueira, Torgo and Mozetič, 2020). The relative quality of a statistical model for a specific dataset is quantified by this measure. The AIC considers both the model's goodness of fitness and complexity of the model as determined by the number of parameters used. The fundamental principle that underlies the AIC is to strike a balance between the model's goodness of fit, which quantifies its ability to account for the actual data, and the model's complexity, which is influenced by the number of parameters it includes. The mathematical expression for determining AIC is given as follows.

$$\mathrm{AIC} = -2\ln(\mathrm{L}) + 2\mathrm{k} \qquad 3.5$$

In which:

- L is the log-likelihood that the model could result in.
- k is the number of feature sets in the model.

Models with lower AIC values are considered better. The AIC tends to penalize models with a higher number of parameters, indicating a preference for simpler models but at the same time that they can adequately represent the data.

BIC:

The BIC, also referred to as the Schwarz Information Criterion, was initially introduced by Gideon E. Schwarz (van der Aalst, Bichler and Heinzl, 2017). The BIC is a model selection criterion that, like the AIC, considers both the goodness of fit and the complexity of the model. However, BIC applies a more stringent penalty on models with more parameters than the AIC. Mathematically, the BIC can be represented in the following manner.

$$\mathrm{BIC} = -2\ln(\mathrm{L}) + \mathrm{k}\ln(\mathrm{n}) \qquad 3.6$$

L and k are similar as in AIC, whereas n is the total number of data points.

Like the AIC, models with lower BIC values are indicative of better fit. However, BIC exhibits a stronger preference for simpler models compared to AIC due to its additional penalty term. In practical applications, both the AIC and the BIC offer a quantitative approach for comparing several models and determining the optimal choice that achieves a harmonious trade-off between model fit and complexity. The selection of the suitable criterion is dependent upon the specific problem being addressed and the fundamental assumptions made regarding the data and model.

In this research, the application of Akaike Information Criteria (AIC) was employed to select the most suitable model among a range of potential candidate models, particularly in the context of applying auto-arima to construct individual models for multiple vehicles.

Furthermore, different model diagnostic techniques are used. It is necessary to fit potential models to the data and assess their overall effectiveness. Diagnostic plots can be used to evaluate the randomness of the residuals and their conformance to the assumptions of zero mean and constant variance. In addition, the accuracy of the model can be assessed by employing evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE), which are described in the next section.

3.7.2.3 Scenario III: Regression

The third scenario employed regression models. Again, the selection of metrics to evaluate the performance of a regression model depends on the specific context and objectives of the work. The following performance metrics are explored in this research, a detailed description of which can be found in (Chicco, Warrens and Jurman, 2021).

Mean Absolute Error

MAE shows the average absolute mismatch between the predicted outcomes and the actual observations. It gives equal weight to all errors, and is mathematically represented as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{Actual}_i - \mathrm{Predicted}_i| \qquad 3.7$$

Mean Squared Error

The MSE calculates the mean of the squared discrepancies between the predicted and observed outcomes. It gives higher weight to larger errors. It is given as the following equation.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{Actual}_i - \mathrm{Predicted}_i)^2 \qquad 3.8$$

Root Mean Squared Error

RMSE is used to gauge the magnitude of error in regression models and is in the same unit as the target variable. It can be calculated as the square root of MSE as follows.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \qquad 3.9$$

R-squared (Coefficient of Determination):

R-squared quantifies the ratio of the variation in the target variable that is predictable from the feature set in a regression model. It spans from 0 to 1, where higher values indicate a better fit, and mathematically given as.

$$\mathrm{R-squared} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} \qquad 3.10$$

Where SSR is Sum of Squares of Residuals, which is the total of the square of discrepancy between predicted and actual outcomes) and SST is the Total Sum of Squares, which is the total of the square of differences between actual outcomes and their average.

Adjusted R-squared

The adjusted R-squared metric considers the number of predictors (p) present in the model, hence making appropriate adjustments to the R-squared value. It penalizes adding unnecessary predictors. Adjusted R-squared is calculated as follows.

$$\text{Adjusted R} - \text{squared} = 1 - \frac{(1 - \text{R} - \text{squared})(\text{n} - 1)}{\text{n} - \text{p} - 1} \qquad 3.11$$

Mean Absolute Percentage Error

MAPE quantifies the average percentage deviation between predicted and true values, with normalization based on the true values. This approach is frequently employed when there is a need to quantify inaccuracies by representing them as a proportion relative to the true values and is calculated as follows.

$$\text{MAPE} = \frac{1}{\text{n}} \sum_{\text{i}=1}^{\text{n}} \left(\frac{|\text{Actual}_\text{i} - \text{Predicted}_\text{i}|}{\text{Actual}_\text{i}} \right) \text{x}100 \qquad 3.12$$

As stated earlier, each metric has its strengths and limitations, and the choice rests on the specific context, goal, and characteristics of the dataset. R-squared quantifies the proportion of variance explained by the model, providing a measure of goodness of fit; MSE and RMSE emphasize larger errors and are sensitive to outliers, with RMSE having units similar to the target variable; MAE is robust to outliers and measures average absolute error in the target variable's units; MAPE expresses errors as proportion of the true values and is useful for business contexts, but it's undefined for zero values and can be inflated by small values. In this research, evaluation metrics including R-Squared, RMSE and MAPE are employed.

To validate the method on the public dataset, noise was introduced systematically. Accordingly, the developed model is evaluated based on the percentage of noises it detects.

3.7.3 Iteration 3 - Overall Proposed Data Quality Assessment Framework Evaluation

The overall framework evaluation went back to the evaluation criteria set for the initial prototype dashboard, since both strive to answer the same questions raised by identified requirements. Specifically, it is evaluated by checking **whether the questions not answered by the adopted classical DQ assessment framework are answered by the new proposed**

framework or not. In other words, the same questions asked to evaluate the prototype dashboard are also raised with a small modification as follows.

- Were all metrics defined based on the selected DQ dimensions included? Or does the developed framework answer all the questions for the defined metrics?
- Were the results conclusive? In other words, is there any doubt on the results obtained by applying the framework?

In addition, to achieve a comprehensive understanding of the proposed framework, a holistic evaluation is conducted using the Strengths, Weaknesses, Opportunities, and Challenges (SWOC) analysis. SWOC, which is similar to SWOT where "T" is replaced with "C" to create a more constructive approach provides a straightforward yet effective approach to evaluate the current situation, allowing to pinpoint comparative advantages and discover potential strategies for performance enhancement with a visual representation of four simple quadrants (Noreen *et al.*, 2020). Although traditionally used for strategic planning, it has recently been applied to scientific research, as illustrated by Noreen *et al.* (2020). The summary of the evaluation methods adopted for this research are summarized in Table 3-5 below.

Table 3-5 Summary of evaluation methods for each iteration

Iteration	**Method used**	**Evaluation**
Iteration 1: Initial framework development	DQ Assessment Dashboard based on DQAF	Comparing Dashboard with requirements Answering questions formulated in 3.7.1
Iteration 2: Enhancement with ML and Statistical Methods	**Scenario I**	
	DBSCAN	1. Silhouette Coefficient 2. Accuracy by using existing dealer dataset
	Logistic regression	F1-score and Accuracy using historical data
	Scenario II	
	Time series	AIC, RMSE
	Scenario III	
	LightGBM Random Forest Control Chart	R^2, RMSE, MAPE, Accuracy, F1-score
Iteration 3: Overall Proposed Data Quality Assessment Framework Evaluation	ML enabled DQ Assessment framework	Comparing with requirements and Dashboard prototype of Iteration I Answering questions formulated in 3.7.1 SWOC

3.8 Summary

This chapter outlines the methods, techniques, and methodologies employed to conduct the overall research project. It begins by explaining the rationale for selecting DSR as the guiding framework for this project. Following that, a detailed discussion on the research techniques, tools, data analysis and evaluation approaches used is presented. The chapter also presents a three-iteration design plan developed according to the various phases of the DSR methodology. Figure 3-1 provides an overall view of the research methodology.

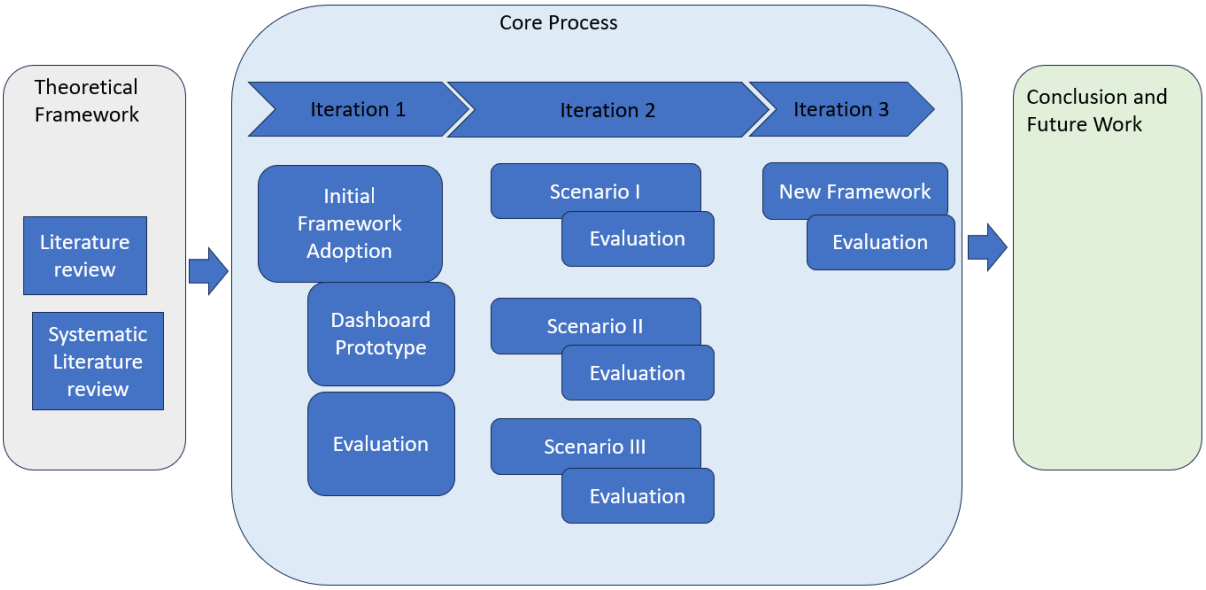


Figure 3-1 Structure of Research Methodology

Chapter 4 : Data Quality Assessment Framework Adoption and Prototype Development – Iteration 1

4.1 Introduction

It is essential to reiterate that DSR as a research methodology was used in this research, as was covered in Chapter 3, and that the three iterations listed below were established.

- Iteration 1: Adopting an existing framework, developing an initial prototype dashboard and evaluation of the initial framework.
- Iteration 2: The enhancement of the initial framework through the application of ML and statistical methods.
- Iteration 3: The proposal of an ML enabled DQ Assessment framework.

This chapter deals with the first iteration and is structured in the following manner. Firstly, the adoption of the classical DQ assessment framework is discussed. Next, development of the prototype for the initial DQ assessment dashboard is explained. Finally, the developed DQ assessment dashboard is evaluated to identify the shortcomings of classical DQ assessment frameworks specifically regarding DQ assessment for CV.

4.2 Classical Framework Adoption

The framework adoption involved reviewing existing DQ assessment frameworks. To this end, the study adopted the DQAF framework for ongoing improvement, which was developed by Laura Sebastian-Coleman (Sebastian-Coleman, 2012) and identified through the literature review. Using this framework, i.e., DQAF, a dashboard prototype is created for initial assessment purposes. Chapter 2 provided a detailed discussion and comparison of classical DQ assessment frameworks. This section gives a detailed discussion of the adopted framework i.e., DQAF, and the rationale why this framework is adopted.

4.2.1 Why DQAF is Adopted?

According to Sebastian-Coleman (2012), the DQAF is a conceptual framework or a set of definitions that offers a guideline for measuring and assessing DQ.

The assessment of DQ is a challenging process because defining DQ and developing assessment and improvement mechanisms is large in scope and complex in nature that requires a lot of effort. Therefore, according to the author, DQAF has developed a more simplified and targeted strategy for evaluating DQ to address this complexity. This is achieved through a set of

strategies: First, the DQAF has established objective measures based on objective DQ dimensions, which are task-independent features of data that may be evaluated independent of the context in which the data will be used. In contrast, subjective characteristics such as believability and relevancy require input from data consumers, either through a survey or some other instrument, who have specific applications in mind, "reflecting their needs and experiences" (Wang *et al.* 2002). This input is necessary to evaluate data quality when subjective DQ dimensions are considered. Second, the DQAF has defined assessments that individuals with only a technical background from the IT department can use for overall data management. These controls include basic controls to confirm the receipt of data, efficacy measures of technical processes in the data chain, and content measures to understand the "completeness, consistency, and validity" of data. In most organizations, the IT department oversees data processing, and as a result, it is also able to measure the integrity of the data linkages. This narrows down the stakeholders involved. Third, because there is a necessity for objective measurement types, the options for DQ dimensions have been cut down to exclude measures that could be easily established. In the end, the goal of DQAF is to establish in-line measures. These are measurements obtained during the processing of data within an application or data store, such as during an Extract, Transform, and Load (ETL) process. These three strategies or decision choices would make it possible to measure the quality of vast volumes of complicated data in an automated manner for the purpose of ensuring and monitoring the ongoing quality of the data continuously.

It is worth noting that one of the most significant advantages of the DQAF is that it makes use of a streamlined form that consists of the following six objective DQ dimensions (Sebastian-Coleman, 2012).

- Completeness: the amount to which all necessary data is present.
- Accuracy: the degree to which the data accurately reflect reality.
- Consistency: whether the data is consistent with a predetermined set of business rules or logical relationships.
- Validity: whether the data falls within a predetermined range of acceptable values.
- Timeliness: whether the data is available when needed.
- Integrity: whether a relational constraint is applied between two different data sets.

The main reason behind the selection of these DQ aspects by the DQAF was the need to simplify and facilitate the process of implementing the DQ evaluation. To accomplish this, DQAF uses fundamental building blocks discussed above and summarized as follows:

1. The DQ dimensions are objective, which means that they can be evaluated based on the data itself.
2. These dimensions serve as fundamental components to establish the responsibility of the IT department for the data, and they help in establishing the practical expectations for data management.

As a result, the DQAF deconstructs the dimensions of "completeness, timeliness, validity, consistency, and integrity" to identify repeatable patterns by means of which certain measures can be taken in a manner that is consistent (Sebastian-Coleman, 2012). The DQAF measurement types can be thought of as basic business needs for taking repeatable kinds of measurements, in the same way that a "thermometer can be thought of as a generic method for measuring temperature" (Sebastian-Coleman, 2010).

The other fundamental principle of DQAF is that it places a significant emphasis on data measurement as an integral component of the data management process. It asserts that data may be measured in the same way that any manufactured item can. It employs some DQ dimensions discussed earlier to develop the measurement criteria. The DQAF prioritizes the customer's expectations and recommends maintaining continuous communication to understand the accessible data, its purpose, the risks connected with it, and the mitigation strategies that correspond to those risks. This will help identify critical data to adjust prioritization when the DQ assessment does not meet expectations. In addition to this, there should be a follow-up to ensure that the requirements are being met. Furthermore, it emphasizes automating the DQ assessment process to ensure it is reliable and consistent.

The DQAF framework offers a method structured and organized for assessing DQ in relation to the selected DQ dimensions. To this end, the framework includes the following six processes: 1. identifying data quality needs, 2. profiling data, 3. analyzing quality level of the data, 4. prioritizing quality issues found in the data, 5. developing an improvement strategy for DQ, and 6. monitoring the quality level of the data. These processes are logically grouped and are laid out in Figure 4-1 below.

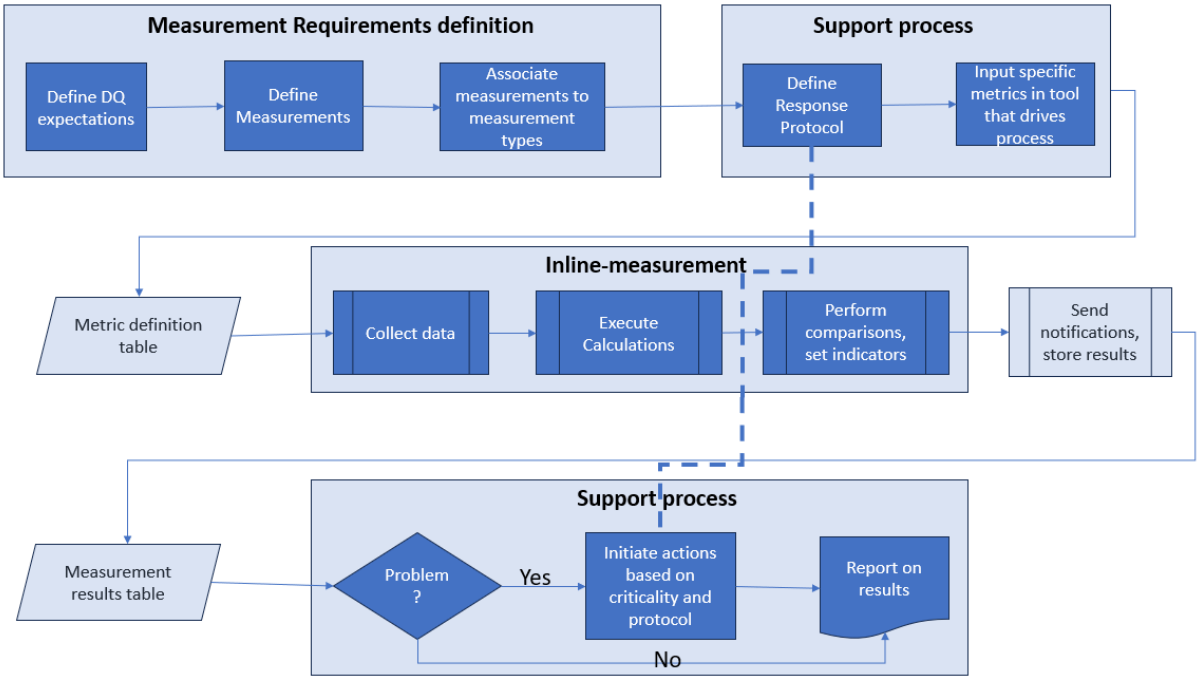


Figure 4-1 DQAF in-line DQ measurement Process diagram [taken from (Sebastian-Coleman, 2012)*]*

Defining DQ requirements involves understanding the context of the data and defining DQ objectives. Profiling data entails analyzing the data to understand its characteristics and locate potential issues. Assessing DQ entails evaluating the DQ across the six dimensions of "completeness, accuracy, consistency, timeliness, validity, and uniqueness". Prioritizing DQ issues involves identifying the most critical DQ issues and determining the order in which they should be addressed. Developing a DQ improvement plan involves developing a plan to address the identified issues, while monitoring DQ involves continuously monitoring DQ to ensure that it remains at the required level.

However, DQAF does not specify implementation details, and the tools and techniques required to implement it. Therefore, the implementation depends on the availability, architecture, and processes in a certain context. It also does not determine what data element to measure. This depends on the specific business process by identifying what data element is critical and should be given a priority. It is also not "all or nothing", rather businesses should adopt a subset of measures based on the criticality of the data element to the business or the situation in context. Some measures may be more useful than the others depending on the situation. It is also worth noting that DQAF is not a "magic bullet", and its success depends on the availability of professionals, tools, and a mechanism to review the findings and devise an effective way to act accordingly.

Based on the DQ dimensions identified, the DQAF framework has defined 48 measures, of which some selected measures are presented in Table 4-1. In defining the measures, DQAF takes a simple approach by focusing on measurements that can be used to detect unexpected conditions or changes in data content that indicate a problem with the quality of the data for each of the six DQ dimensions. Each measurement type is defined discreetly to understand the individual functions to take it. In some cases, there is more than one way of looking at a type. For example, some completeness measures, and even complex validity measures, can also be understood as integrity measures and vice versa. In that regard, there can be an overlap of measures across DQ dimensions.

It also emphasizes that to define measurements one needs to have metadata knowledge such as the concept the data represents, data creation process, the system used to keep, update, or remove data, the data model of the target system and data processing rules of the target system. In addition, it stresses the objects of measurement as processes, models, and content.

Finally, it specifies three simple steps to take on measurements:

1. Collect which represents what features to measure such as “file sizes, process duration, record counts, and grouping”.
2. Calculate which represents “compute averages, percentages” and so on and
3. Compare (against a signifier of the quality for example threshold, or other standards such as results of past measurements in a stable process).

A selected set of measures and corresponding DQ dimension mappings, relevant to this study, is given in Table 4-1 below.

Table 4-1 DQAF Measures mapping to DQ dimensions and descriptions (Sebastian-Coleman, 2012)

DQ Dimension	Measurement Type	Measurement Description	Measurement Purpose
Completeness	Historical trend completeness	Conducting a rationality evaluation by comparing the magnitude of the current input with that of previous inputs.	Assuring reception of data
Completeness	Row length completeness	verifying that the length of rows aligns with a predetermined expectation.	The State or quality of data upon its reception.
Completeness	Mandatory data elements completeness	verify that all mandatory data elements properly populated	The State or quality of data upon its reception.
Completeness	Data element content completeness	Verify that all data elements are available	The State or quality of data upon its reception
Validity	Plausibility verification of single data element	comparing values in the data being received with acceptable values within a specified domain, which may be represented by a reference table, a range, or a predefined rule.	record counts based its content
Validity	Plausibility verification of aggregate	Aggregate outcomes of detailed plausibility evaluation, compare aggregate totals and proportions of plausible/implausible outcomes to historical trends and benchmarks	Aggregate overview
Timeliness	Currency of data when required	verify the time data is received to the time greed and expected by users	Conforming to agreed process
Integrity/Consistency	Consistency of data over time across different entities	verify the reasonability of cross-entity data by checking timestamps against a set of business rules that govern the order of events	Consistency over time
Consistency	Uniformity in applying default for a data element	Verify data type properties and default values for each data element and each field whether uniform values are used or not	Data structure

It is worth noting that in the measurement and DQ dimension mapping, DQAF does not specify any accuracy DQ related measure out of the 48 measures defined.

The DQAF framework has several advantages over other DQ assessment frameworks Cichy and Rass (2019). Firstly, it provides a comprehensive approach to evaluating DQ across six dimensions, ensuring that all aspects of DQ are considered. Secondly, the framework provides a structured approach to DQ assessment, making it easier to identify and prioritize DQ issues. Finally, the framework is flexible and can be customized to meet the specific needs of different organizations.

Several examples of the implementation of the DQAF framework have been reported in the literature. For example, the framework has been used to evaluate DQ in healthcare organizations, financial institutions, and government agencies (Sebastian-Coleman, 2010). It is indicated in the same study that the framework was effective in identifying DQ issues in a healthcare organization where the framework was implemented, leading to the development of a DQ improvement plan.

In conclusion, the DQAF framework is a comprehensive and structured approach to evaluating DQ across six dimensions. The framework provides a useful tool for organizations to evaluate the level of DQ and develop DQ improvement plans accordingly. By using the framework, organizations can ensure that their data is of high quality, leading to better decision-making and improved business operations.

4.2.2 Initial Prototype Dashboard

4.2.2.1 Defining measurement metrics

As described earlier, DQAF was chosen for this research to create an initial DQ assessment dashboard. To facilitate the implementation, the following assumptions have been taken into consideration, consistent with the DQAF framework:

1. It is solely concerned with objective DQ Assessment:
   DQ is a broad subject that requires several processes and the participation of many stakeholders, both of which are beyond the scope of this study. This was one of the main reasons that DQAF was adopted for this study. Therefore, this research focuses only on the objective assessment of DQ by referencing an existing DQ requirement and translating these requirements to DQAF set of measures.
2. It focuses on a small number of the available data elements:

A significant amount of information is gathered for a variety of signals and data elements by the CV system, which is too big for this study. Because of this, only a few data elements, such as the mileage and the fuel usage are taken into consideration.

3. It only considers a limited number of measures:
   Only a few selected measures are given, all of which are based on specified objective DQ dimensions and make use of the selected data elements. The is done in such a way that the usefulness and limitations of the classical method of assessing the DQ of CV data can be demonstrated.

As a result, based on the DQAF, the dashboard presented in Figure 4-2, which contains the metrics that were selected, is constructed as a prototype. During the building of the initial DQ assessment dashboard, the following steps are considered.

3. Defining or Adopting a DQ assessment framework:
   For this step, a review of different DQ frameworks is conducted and DQAF is adopted because of the advantages described in section 4.2.1 including practicality and ease of implementation.
4. Defining DQ measures and mapping to corresponding DQ dimensions according to DQAF:
   Measures are identified for the selected data elements and mapped to the identified DQ dimensions. As explained in Chapter 2, the CV architecture follows a layered architecture. Data from the vehicle and the environment is collected and transmitted to the processing engine. These data elements include mileage, fuel consumption, fault codes, driver actions and so on. So, measures are defined to assess the selected data elements.

On the other hand, the DQ dimensions completeness, Accuracy, Timeliness, Consistency, Validity, and Integrity are considered, which are described as attributes to define objective measure according to DQAF. Furthermore, the DQ measures from DQAF are taken as input to define the measures for this prototype dashboard.

Besides, consulting the BRD document as explained in Chapter 3, the following high-level requirements are extracted.

A. All messages generated should be available. There shall not be a message or a data element missing.

B. 100% of the trips should have no delay, trips with a delay of equal to or less than 15 minutes are not considered as delays.

C. All data should be available as generated from the vehicle (without any manipulation). Information received should be accurate.

The BRD document did not present a detailed breakdown of the quality requirements. Therefore, what is presented here is the exact copies of the high-level requirements. This necessitated to decompose the high-level requirements into measurable components. Hence, combining the concepts explained earlier which are the DQAF framework building blocks, business requirements, and selected connect data elements; the following measurement matrix is developed. The measures are defined at message level (which is the lowest level granularity) and aggregated to vehicle level.

Table 4-2 Defined measures for initial DQ Assessment according to DQAF Measures

Measure	**DQ dimension**	**Definition**	**DQAF Measure type**
From all vehicles connected, a stable proportion of vehicles should generate messages	Completeness / Timeliness	[Number of vehicles making trips]/ [Total vehicles connected]	Dataset completeness
*All data including mileage from vehicles making trips should be available.	Completeness / Consistency	A. [Number of records missing the data element]/ [All records] B. [Number of vehicles with missing the data element] / [Total vehicles making trips]	Record or Row completeness
All messages generated from all vehicles should be available timely	Timeliness	A. [Number of messages with delay [received 15 mins or more after generation]/ [Total messages generated] B. [All vehicles with delayed messages]/[All vehicles with messages]	Process/ Adherence to schedule
**All data generated from each vehicle should be valid	Validity	A. [Number of invalid occurrences]/ [Total messages] B. [Number of vehicles with invalid data]/[All vehicles with messages]	Content
***All data generated should reflect the accurate representation of the data.	Accuracy	[Number of inaccurate values]/ [Number of all values]	N/A

Data element mileage is used

***Data elements mileage and fuel consumption are used*

****There is no defined measure type in DQAF*

4.2.2.2 Development of the initial Data Quality Assessment Dashboard

Using real-life CV data, measures are defined and calculated values are computed as presented in Table 4-2. The implementation of calculating these measures is done with python functions. These values are visualized, and part of the dashboard is provided in Figure 4-2 below. The data is obscured within the dashboard so as not to compromise privacy issues due to GDPR.

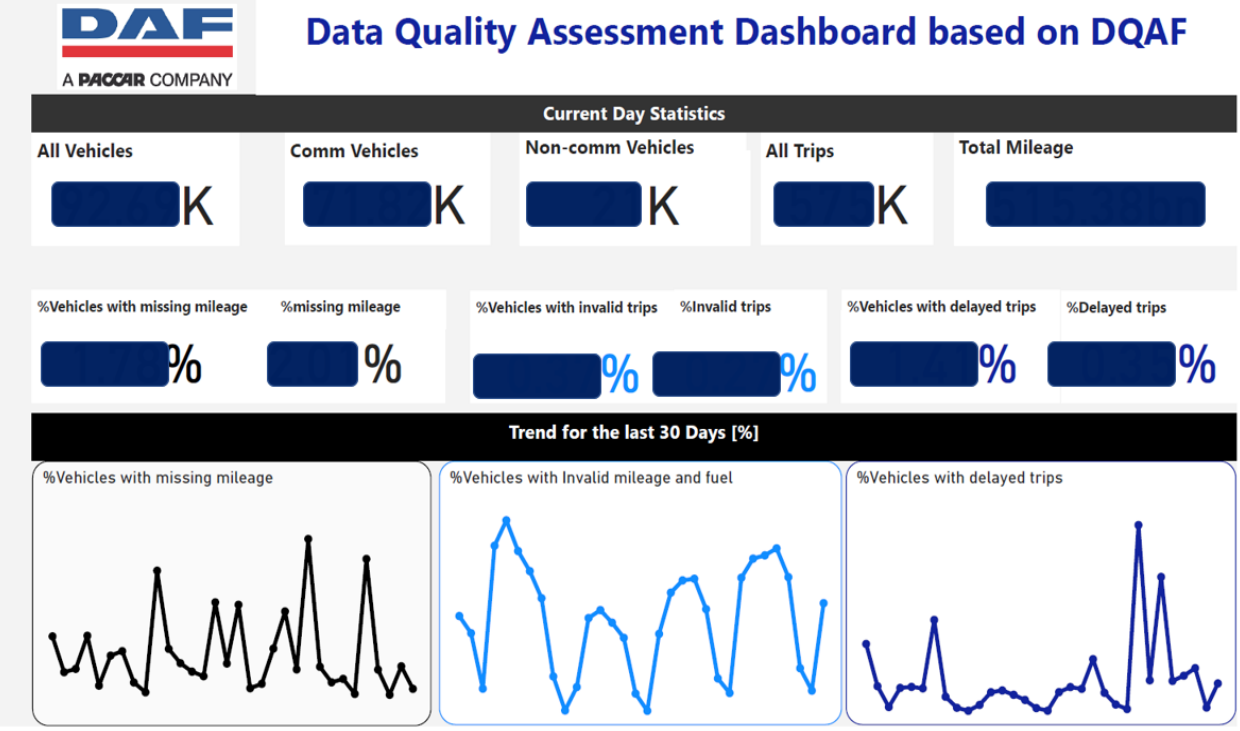


Figure 4-2 Initial DQ Assessment Dashboard according to DQAF showing the overview page

The prototype dashboard also provides drill-down possibility to learn further details. An example of such a possibility is given in Figure 4-3 which shows the last location of vehicles marked as non-communicating in the overview dashboard of Figure 4-2.

Figure 4-3 DQ Dashboard drill down example to geo-location level

The implementation of DQAF has provided particularly useful insights. Considering that the metrics chosen for this activity are only objective, DQAF is more appropriate compared to the other classical DQ assessment methodologies reviewed in Chapter 2. To understand the value of the adopted methodology and the implemented dashboard, a detailed evaluation is presented in section 4.3 below.

4.3 Evaluation

The developed prototype dashboard offers helpful insights regarding the DQ dimensions chosen and the metrics defined. But at the same time, there are questions that cannot be answered with this dashboard, which implies that there are still some uncertainties regarding the quality of the data even after implementing the prototype dashboard. In this section, the answers that were obtained from the prototype dashboard, as well as the uncertainties and limitations that are associated with using the dashboard are discussed.

4.3.1 Insights from the Prototype Dashboard

The dashboard provided helpful information regarding data missing from multiple points of view. It is essential to compare to the high-level requirements set earlier to ensure that the evaluation is accurate, which is described as follows.

A. All messages generated should be available. There shall not be a message or a data element missing.

Regarding this requirement, the dashboard displays two crucial metrics for CV DQ assessment.

First, it shows the percentage of the total number of CV that are not transmitting data or which are not communicating. This percentage is based on the expectation that all CV will send data if they are not experiencing technical problems. Communication is extremely vital, as it serves as the foundation for any company operation that is determined by the data collected from CV. Next, it displays the amount of missing mileage as a percentage as well as the number of vehicles that have missing mileage. The first metric identifies the vehicles that are not transmitting data, while the second metric determines whether there is any mileage that is missing from the vehicles that are communicating. In this regard, the metric for missing mileage shows a DQ issue that is certainly an issue; to put it another way, this metric reveals clearly both the total amount of mileage that is missing and the total number of vehicles that are impacted. The second metric, on the other hand, indicates that there might be a problem; specifically, the vehicle may not be sending data because it is currently turned off and parked, which is not a concern. Because of this, it is impossible to tell, based on the overall number of non-communicating vehicles in the dashboard, which ones are having a problem and unable to send data, and which are just parked and turned off.

B. 100% of the trips should have no delay, trips with a delay of equal to or less than 15 minutes are not considered as delays.

The dashboard also displays other information giving insight for other requirements defined, such as the number of messages that were delayed and the total number of affected vehicles. However, the dashboard does not indicate or tell if there is a delay of non-communicating vehicles. The metrics displayed are only about messages received and vehicles sending data. The timely availability of data is essential to the success of certain business functions, including uptime monitoring and vehicle health monitoring. This measure contributes to the confidence that may be achieved when providing services that require timely data. As a result, this metric gives an important insight to take an informed decision and to produce a mitigation plan for potential issues.

C. All data should be available as generated from the vehicle (without any manipulation). Information received should be accurate.

Another important metric displayed on the dashboard is the total amount of invalid data and the number of vehicles affected over time. The validity DQ dimension is used for fulfilling this criterion, as recommended by DQAF as a proxy measure for accuracy. As a result, the metrics that are defined on this basis reflect the quantity of invalid data, which may be characterized as

data that falls outside of the range of values that have been defined or data that breaks the rules that have been set. However, the numbers displayed as valid on the dashboard do not reflect whether they are accurate or not. For the data to be useful, it must first be reliable, and ensuring its reliability is a crucial step to prevent making wrong decisions.

4.3.2 Limitations of the Prototype Dashboard

Even though the dashboard displayed essential measures and provided valuable insights that helped understand the data's level of quality, it did not show all measures defined. To be more specific, the dashboard does not provide access to the following measurements, both of which can be attributed to the space and time, i.e., spatio-temporal characteristics of CV which cannot be captured easily to define DQ measurements.

1. **Whether the vehicles not transmitting data are not communicating due to issues or parked properly.**

The first limitation is related to the completeness and timeliness DQ dimensions defined to measure the first DQ requirement. The dashboard displays, as discussed earlier, the number of vehicles that are communicating and sending data together with the number of vehicles that are not transmitting data, or not communicating. However, it does not show if the vehicles that are not sending data are properly parked and power is switched off, or whether they are experiencing undesirable issues that prevent them from communicating, in which case the data that is being created is either missing or delayed. In the second scenario, it will result in the vehicle missing data or experiencing a delay when it communicates back by the time the issue is fixed. However, in the first scenario, where power is intentionally switched off and the vehicle is parked, it will not experience any problems when it communicates back and there will not be delay or missing data.

2. **Whether the data received is accurate or not.**

Again, even though the dashboard shows the proportion of valid values versus invalid values, it does not show whether the values in the valid category denote the correct real-world representation or not. From the literature, it is understood that accuracy is a difficult DQ dimension to measure as it requires a known and verified reference value to compare (Sebastian-Coleman, 2010). The DQ assessment frameworks reviewed including the selected framework for this implementation, i.e., DQAF did not provide a simple and practical measure for accuracy DQ dimension. As further emphasized in DQAF, measuring accuracy is difficult

as there should be the real-world entity or a surrogate value which is validated and verified to make a valid comparison (Sebastian-Coleman, 2012). As stated in (Galarus and Angryk, 2016), one of the primary obstacles encountered in evaluating the quality of spatio-temporal data is the limited availability of ground-truth data. Galarus and Angryk (2016) also stated that error, as defined in academic literature, refers to the discrepancy between observed data and the true or accurate value. Therefore, in situations where conclusive proof is lacking, the evaluation of accuracy is dependent upon simple judgement based on other related information.

Therefore, this research aims to augment the classical DQ assessment frameworks by applying ML by filling the gaps identified in this research according to the findings from the literature review and the implementation of this prototype dashboard. The following sections describe the use of ML to improve the DQ assessment process by using three scenarios; specifically, it attempts to detect non-communicating vehicles and whether data being received is accurate or not.

At this point, equipped with the information obtained from the literature and the implemented prototype dashboard, it is possible to answer the generic questions raised in section 3.7.1 as evaluation of the adopted classical DQ assessment framework. This evaluation is given in Table 4-3.

Table 4-3 Evaluation of the adopted classical framework based on the implemented prototype.

Question	Answer
Were all metrics defined based on the selected DQ dimensions included?	**No:** Accuracy DQ dimension and its associated metrics are missing.
Were the results conclusive?	**No:** 1. The non-communicating metrics fail to provide information regarding whether the vehicle is actively creating data but not transmitting it, or if the vehicle is parked and switched off. 2. Missing mileage metrics does not include the missing amount from non-communicating vehicles

4.4 Conclusion

This chapter started by adopting a classical DQ assessment framework from the candidate frameworks reviewed in Chapter 2. Accordingly, DQAF was adopted. There are several reasons DQAF was selected. DQAF has developed a more simplified and targeted strategy for evaluating DQ to address the complex nature of DQ assessment. First, DQAF established 48 objective measures based on objective DQ dimensions, which are task-independent features of data that may be evaluated independent of the context in which the data will be used. Second, the DQAF has defined assessments that individuals with only a technical background from the IT department can use for overall data management. This narrows down the stakeholders involved. Third, because there is a necessity for objective measurement types, the options for DQ dimensions have been cut down to exclude measures that could be easily established. In the end, the goal of DQAF is to establish in-line measures. Moreover, it asserts that data may be measured in the same way that any manufactured item can.

Based on the adopted DQAF framework, a prototype DQ dashboard is developed for CV data. First, measures were defined from the requirements collected and mapped to DQAF measure types and DQ dimensions. Next, measure calculation results are stored in a repository and finally the results are visualized.

Then a detailed evaluation of the prototype dashboard is presented. The dashboard provided valuable insights that helped to understand the data's level of quality. However, it failed to show all measures that were defined, especially spatio-temporal characteristics of CV which cannot be captured easily with classical DQ measurements. One of the possible outcomes of research according to DSR is: "*A clear evidence that there is no optimal solution available for the problem*". In this regard, one of the main outcomes of this iteration is proof that classical DQ assessment framework fails to give a complete assessment of CV data. The gaps identified in Table 4-3 are used to define the scenarios for the next iteration.

Therefore, the next chapter explores advanced methods such as ML and statistical methods to augment classical frameworks for a comprehensive DQ assessment by defining three scenarios that are identified as gaps of the prototype dashboard developed.

Chapter 5 : Data Quality Assessment Framework Enhancement with Machine Learning for Connected Vehicles data – Iteration 2

5.1 Introduction

The purpose of this chapter is to investigate the potential enhancement of classical DQ assessment methods through the utilization of advanced techniques, specifically ML and statistical methods. This research aims to determine whether these methods can effectively improve the evaluation of DQ in CV. This is the second iteration as explained in section 3.3. To address the limitations identified in iteration I of the initial dashboard used for assessing DQ in CV, three scenarios were created to explore the potential of advanced techniques like ML and statistical methods. These scenarios aimed to overcome the shortcomings identified in the classical DQ assessment framework that was initially applied.

The first scenario (Scenario I), developed to tackle the second limitation identified from the initial prototype dashboard in Table 4-3, focuses on addressing the issue of non-communicating vehicles and determining whether they are genuinely experiencing problems or simply parked with their power supply turned off. This DQ assessment metric was not possible to include in the prototype dashboard discussed in Chapter 5, highlighting a gap in classical DQ assessment frameworks. By utilizing real-life data and employing a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, a new feature was generated which served as an important feature for the trained ML model to detect this issue in a timely manner. A logistic regression model was then trained to differentiate genuine issues from false alarms, achieving a performance score of 0.85 using the F1-score metric. This scenario contributes to improving dimensions of DQ including timeliness and completeness. The second scenario (Scenario II), builds upon the first one, aiming to forecast the mileage of non-communicating vehicles, further enhancing DQ assessment in CV. The third scenario (Scenario III), developed in response to the first limitation of the prototype dashboard highlighted in Table 4-3, involves detecting erroneous fuel consumption values, addressing the accuracy dimension of DQ assessment. This is done by training an ML model to predict fuel consumption using real-life data and a publicly available dataset. The discrepancy between observed and predicted values is then computed and analyzed using a control chart to identify any inaccuracies. The approach was evaluated using real-life reference data, achieving a successful detection rate of inaccurate values in the test datasets with accuracy of 97% and F1-score of 0.78.

The organization of this chapter is as follows: Firstly, the application of supervised and unsupervised ML approaches to identify missing data and delayed data for scenario I is explained. Next, the process of forecasting mileage for non-communicating vehicles for Scenario II is discussed. Lastly, the utilization of ML methods in combination with a statistical control chart to detect inaccurate data received from CV for Scenario III is described.

5.2 Scenario I: Detecting missing data or delayed data (Completeness and Timeliness Data Quality dimensions)

CV rely on the integration of spatial and temporal elements as essential components (Cerqueira *et al.*, 2018). However, the connectivity of CV can be affected by several factors, including disruptions in telecommunication, network-related problems, issues with cabling infrastructure, and unforeseen incidents (Ashton and others, 2009), as discussed in Chapter 2. When faced with these circumstances, vehicles may encounter failures in transmitting data, which can result in delays or data loss. The purpose of this section is, therefore, to differentiate vehicles that are experiencing issues as part of DQ assessment for CV data to prevent or reduce missing data and delayed data.

5.2.1 Problem description of non-communicating vehicles

To fully reap the benefits of CV, reliable data flow is crucial. However, there are instances where data may be delayed or missing due to several reasons. As described earlier, a vehicle may not be sending data for two reasons: 1. If it is properly parked with the main switch off, which is a normal behaviour, or 2. due to technical issues preventing the vehicle from communicating or sending data. If the latter occurs, there are two possible undesired outcomes: delayed data will be received once the issue is resolved, or the data will be lost entirely. Both situations negatively impact business operations by affecting DQ, specifically the timeliness and completeness DQ dimensions, among others (Juddoo *et al.*, 2018). For this research, all the vehicles are classified as follows depicted in Figure 5-1 below. Depending on the age of the last message received from the vehicle, it can be classified as either communicating or non-communicating. The non-communicating vehicles can be sub-divided into new non-communicating, i.e., last message is not older than 24 hours (one day) or continuously non-communicating, i.e., last message received is older than one day. The time limit can be adjusted according to the requirements. When the vehicle communicates back, it can be categorised in one of the following.

- It may communicate back with no delay (data is buffered in the embedded unit) and hence no missing data is reported.
- It may communicate back with no missing data but with a delayed data or buffered data is reported.
- It may communicate back with no delayed data but with missing data reported, or
- It may communicate back with both missing data and delayed data reported.

The last three situations are undesired situations and should be avoided. To avoid or minimize the impact, several actions can be taken including rebooting the embedded unit remotely by sending messages, manual reboot or replacing the embedded unit all together. However, investigating all vehicles that are not sending data is costly and difficult since vehicles are driving in various locations and times which is spatio-temporal dependency. So, distinguishing the vehicles that are likely to communicate back with delay or missing data from those that are likely parked properly is particularly important to take an effective preventive action.

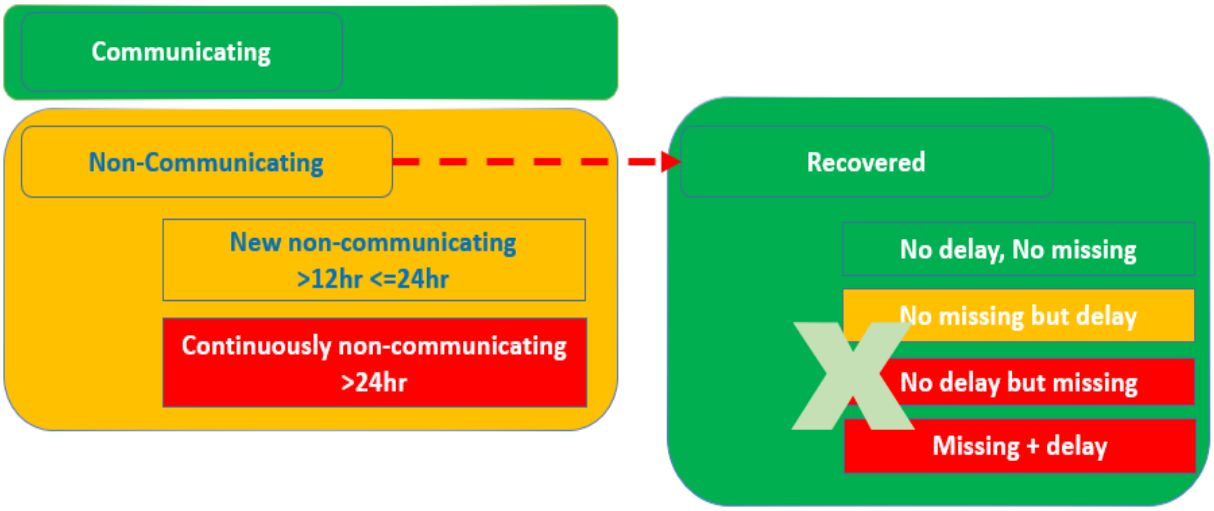


Figure 5-1 Daily classification of vehicles according to their communication state

Addressing this DQ problem requires identifying the problematic vehicles from those that are properly parked. While a DQ assessment dashboard can be developed to monitor the problem (Gitzel, Turring and Maczey, 2015), classical DQ assessment methods may not suffice in the context of CV ecosystem since the spatio-temporal characteristic or behaviour cannot be easily captured, as described in section 4.3.2, and demonstrated with a prototype dashboard. In this research, ML methods are explored to differentiate normally parked vehicles from those not sending data due to technical issues or problems preventing the vehicle from communicating, which is described in the next section.

5.2.2 Proposed solution to detect missing data or delayed data

The main objective of this scenario is to detect missing data and delayed data hence to assess completeness and timeliness DQ dimensions. The proposed approach involves applying two ML components:

1. An unsupervised ML method to capture the spatio-temporal aspect by generating a new feature, And
2. A supervised binary classification method using the newly constructed feature together with other features in the data.

The data used in this study is collected from the Controller Area Network (CAN) bus of the vehicle through a connectivity module and transferred to a backend system as described in section 3.4 and briefly re-iterated here. A chunk of information transmitted at a time is called a message. Three types of messages are received, including time trigger messages, event trigger messages, and summary or trip level messages. Time trigger messages are generated based on a predefined time interval from the moment the engine is turned on until the engine is turned off consisting of a number of data elements or signals, while event trigger messages are generated when a certain event including a driver action such as braking, vehicle health related events such as oil temperature too high, is triggered. Summary messages, also known as trip messages, are generated at the end of a trip and contain a summary of the entire trip. The data elements extracted for this scenario are described in Table 5-1, and the metadata for the complete dataset is presented in section 3.4.

Table 5-1 Data elements collected from connected system and used for scenario I

Field name	Example value	Description
VEHICLEID	1	Vehicle identifier
EVENTID	4	The reason the data was generated (4 = trip start message, 1 = timer interval message, 6 = GPS fix messages, 5 = trip end message)
EVTDATETIME	2020-06-12 12:00:13	The datetime when the message was generated
GPSLATITUDE	47.75903702	The GPS latitude when the data was generated
GPSLONGITUDE	9.889554024	The GPS longitude when the data was generated
VDIST	148812965	The cumulative vehicle mileage when the data was collected
RECEIVEDDATETIME	2019-05-04 10:09:30	Datetime when the data was received in the back end
COUNTRY	Germany	Country where the vehicle was driving by the time data was collected

In addition to these primary features, a number of derived features are computed and presented in Table 5-2 below. These features are generated mostly by using historical values and comparing the previous or next states of the record in consideration. For example, the number of times the vehicle has communicated back with a missing distance is captured as NUMBER_MISSING, the total cumulative mileage of the vehicle just before it stopped communicating is captured as PREV_VDIST and the state of the vehicle when it communicates back is captured as NOCOMM. The rationale for this approach is that trajectory data or trip data is expected to be consistent. For example, a vehicle should begin its next trip from where it ended its previous trip under normal circumstances. While calculation of most of the features is relatively easier, which involves simple calculations or aggregations, the determination of feature DIST_CLOSEST_PARKING involved an application of unsupervised ML method, which is described in the next section. The complete list of derived columns is given in Table 5-2.

Table 5-2 Derived data elements or features for scenario I

Field name	**Calculation**
PREV_GPSLATITUDE	The GPS latitude immediately before the current event
PREV_GPSLONGITUDE	The GPS longitude immediately before the current event
PREV_EVTDATETIME	The timestamp immediately before the current event
PREV_EVENTID	The eventid immediately before the current event
PREV_VDIST	The mileage immediately before the current event
VDIST_DIFF	The gap in mileage of the previous trip and next trip
PREV_RECEIVEDDATETIME	The datetime when event immediately before the data was received in the back end
PREV_DAY	The communication day name of the immediate previous event
COMM_DAY	The communication day name of the event
DELAY_HRS	The delay in hour from the time the trip was made, and the information is received to the back end
EVT_TIME_DIFF_HR	The time difference between the event datetime of the previous immediate event and the current event
TIME_DIFF_HR	The time gap in between previous trip and next trip
NEXT_STARTING_POINT_KM	The haversine distance in kilometer between the gps points of the immediate previous event and the current event
SNP_PREV_VDIST	The total vehicle distance /mileage of the vehicle at the immediate previous timer event
SNP_PREV_EVENTID	The immediate previous timer eventid

Table 5-2 Continued

Field name	Calculation
PREV_SNP_RECEIVEDDATETIME	The received datetime of the immediate previous timer event
SNP_NEXT_VDIST	The total vehicle distance of the current timer event
SNP_NEXT_EVENTID	The eventid of the current event
NEXT_SNP_RECEIVEDDATETIME	The receive datetime of the current timer event
MISSING_SNP	The mileage difference between the current timer event and the current timer event
SNP_PREV_GPSLATITUDE	The GPS latitude of the immediate previous timer event
SNP_PREV_GPSLONGITUDE	The GPS longitude of the immediate previous timer event
SNP_NEXT_GPSLATITUDE	The GPS latitude of the current timer event
SNP_NEXT_GPSLONGITUDE	The GPS longitude of the current timer event
SNP_NEXT_DIR_EVTDATETIME	The generated datetime of the current timer event
SNP_NEXT_DIR_RECEIVEDDATETIME	The received datetime of the current timer event
SNP_NEXT_DELAY_HRS	The time difference between the receive datetime and the generated datetime of the current timer event
PREV_SNP_DAY	The day name of the immediate previous timer event
NEXT_SNP_DAY	The day name of the current timer event
SNP_MISSING_GPS	The distance gap between the GPS points of the current timer and the immediate previous timer event
*NOCOMM	An indicator of whether the vehicle is communicating or not (Yes = non-communicating, No = Communicating). If a trip has a delay of 13 hours or more or it has a missing trip, then it is marked as non-communicating
NUMBER_MISSING	How many times has the vehicle showed a missing trip in the past
NR_BUFFER	How many times has the vehicle been delayed in the past
**DIST_CLOSEST_PARKING	The closest parking location in meters from the current event GPS point

NOCOMM: is the target feature or dependent variable for the classification learning model which provides a value of "Yes**" which represents a vehicle likely to face an issue and hence labelled as non-communicating, "**No**" which represents the vehicle is not facing an issue and expected to communicate back normally without an issue.*

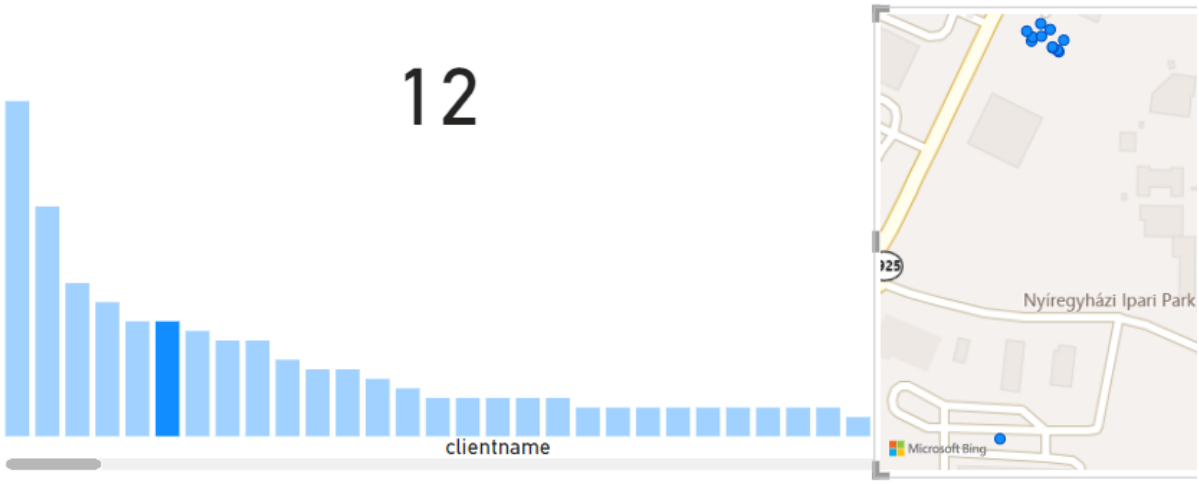
***DIST_CLOSEST_PARKING: is a feature constructed with the help of the unsupervised learning step described in section 5.2.2 below.*

The data elements of Table 5-1 and Table 5-2 are merged as non-communicating dataset for the classification algorithm.

As previously described, the proposed approach consists of two main components. Firstly, an unsupervised ML technique is employed to capture the spatio-temporal characteristics of the CV. This is achieved by generating a novel or a new feature which is not available in the original data set. Secondly, a supervised ML method, specifically classification learning model, is utilized to determine whether a vehicle that is not transmitting data is experiencing an issue and is likely to encounter missing data or delays. This section describes the two components.

5.2.2.1 Unsupervised learning method to generate a new feature with DBSCAN

During the process of exploring the data, it became apparent that the location where vehicles are parked can have an impact on the vehicles facing an issue of delayed delivery of data or missing data or both missing data and delayed data. To investigate this further, dealer locations were considered since the dealer location dataset was readily available. It was discovered that many vehicles that stopped at dealer locations were typically parked normally and returned without any issues. This has provided a good insight. However, dealer locations are only a small fraction of the places where vehicles can be normally parked, with many other parking locations being unknown such as customer locations known as home base or other parking areas. To further investigate, the data is visualized on aggregate of missing or delay occurrences per customer using bar chart and corresponding GPS locations on a map as shown in Figure 5-2. The figure shows an overview containing the number of vehicles not sending data per client/customer. On the right side of the same plot, the map shows the last known locations for the vehicles (for the number of vehicles given on the left) in the bar graph. And below in the table is shown, the status of the vehicles such as how many times they had an issue of missing distance, how many delays or buffered data occurrences, whether they were driving or stopped when they last communicated, and how many vehicles were parked on the same location. As depicted in the same figure, zooming in to a single client or customer revealed that the selected customer (dark blue) has 12 vehicles which did not communicate, i.e., no data is received. Out of these 12 vehicles which did not send data, 11 of them are shown in the right side of the same plot, parked on the same location and one of them is also parked a little further as shown on the map. All of them did not have any missing data or delayed data as shown in the table of Figure 5-2 with columns **nr_missing** and **nr_buffer** showing 0 values indicating that there was neither delayed data nor missing data experienced by these vehicles.



mileage_last_two_weeks	nr_close_locs	prev_non_comm_locs	nr_missing	nr_buffer	nr_recovery	last_trip_status
10.71	3.00		0.00	0.00	3.00	stopped
25.08	37.00	1.00	0.00	0.00	3.00	stopped
6,276.06	37.00	3.00	0.00	0.00	4.00	stopped
7,086.26	37.00		0.00	0.00	1.00	stopped
9,110.89	37.00		0.00	0.00	1.00	stopped
9,514.39	37.00		0.00	0.00	1.00	stopped
10,623.38	37.00		0.00	0.00	0.00	stopped

Figure 5-2 vehicles not sending data of the same customer at known location came back with no issue

Since this feature is essential to the investigation for this research, a clustering algorithm was utilized to extract potential parking locations from the dataset in order to identify all possible locations where a vehicle may be parked for longer period (excluding short parking events such as parking for the purpose of loading and unloading of goods or items). The process involved using GPS coordinates from the dataset to determine the locations where vehicles were parked by comparing the GPS point where they ended their previous trip or journey and where they began their next trip. This is the duration that the vehicle has been in a continuous parking state graphically represented as the duration between time **t2** and **t3** as illustrated in Figure 5-3 below.

For understanding, Figure 5-3 shows a trip or trajectory model where from t1 to t2 is one trip and from t3 to t4 is a second trip.

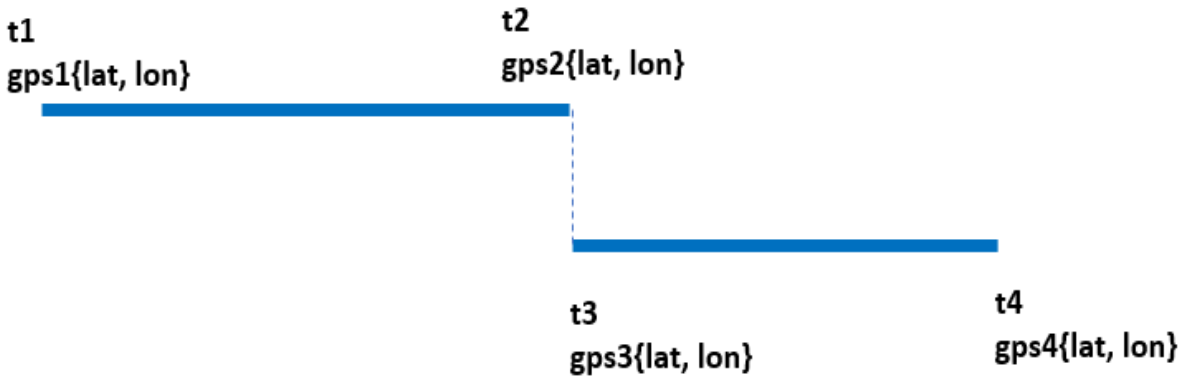


Figure 5-3 Trip sequence and GPS coordinates of consecutive trips

If the GPS points of t2 and t3 are the same and the time duration is greater than or equal to 9 hours, the location is assumed to be a parking location. The 9 hours threshold is selected since the minimum rest period for drivers is 9 hours according to European Union law (European Union, 2006). This will exclude working sites such as locations where the vehicle stops to load and unload goods. This is because any issue experienced at working locations should also be captured and hence, they should not be excluded as parking locations identification.

To identify the parking locations using ML (specifically DBSCAN), the GPS points corresponding to t2 and t3, i.e., gps2 and gps3 (which represent the same location) in Figure 5-3, from the data set are collected as described in section 3.4 and DBSCAN is trained. The data set includes 1,304,516 locations. According to Ester *et al.* (1996), DBSCAN is one of the most extensively used clustering algorithms. This technique is a density-based non-parametric algorithm that produces sub-groups of closely situated points together and marks the less dense points as outliers. The following justifications underpin the selection of DBSCAN (Khan *et al.*, 2014).

1. It works well with data that has an irregular shape.
2. There is no requirement to predefine the total number of clusters in advance.
3. It performs well with data coming from a variety of distributions (it does not assume normal distribution).

According to Ester *et al.* (1996), the DBSCAN algorithm makes use of two parameters:

1. **minPts**: which refers to the minimum number of points required for a group to be considered dense enough to form a cluster.
2. **eps (ε)**: which indicates how close things need to be to one another for them to be deemed to be a part of the same group or cluster.

This method is used to determine the parking spots based on where the cluster centroids are located. The resulting collection is saved so that it can be used later to develop a new feature for the supervised learning module, which is used in the classification process to identify vehicles that are facing communication issues as explained in this section earlier.

To complete this activity, the following steps are implemented to ascertain that potential parking spots are captured using the DBSCAN algorithm.

1. GPS locations where vehicles stopped (9 hours or more in the same location) in the past were collected as shown in Figure 5-3.
2. Then, DBSCAN is trained on the dataset described in number 1 by setting parameters eps to 0.2 and minPoints to 10. As described earlier, DBSCAN relies on two key parameters: eps (epsilon) and minPts (minimum points). Determining optimal values for these parameters is crucial for achieving meaningful clustering results. The parameter minPts determines the minimum number of points required to form a dense region. A common heuristic is to set minPts to at least D + 1, where D is the number of dimensions in the dataset (Ester *et al.*, 1996). For instance, in a two-dimensional dataset, Ester *et al.* (1996) suggest minPts should be at least 3 and recommend a default value of 4, and in a three-dimensional dataset, it should be at least 4. On the other hand, Sander *et al.* (1998) suggest choosing MinPts of 2*dim for data set of more than 2 dimensions, where dim represents the dimensions of the data set. Domain specific knowledge can also guide the selection of minPts. For example, if prior knowledge suggests that clusters should contain a minimum number of points, this information can be used to set minPts appropriately. Finally, experimentation with different values of minPts can help identify an optimal value by observing changes in the number and quality of clusters formed. Analysis of the data shows that the average number of vehicles per customer is 10. Using this and domain knowledge, minPoints is set to 10 which implies that either 10 vehicles are parked at a time, or 1 vehicle of the same customer is parked at the same location at least 10 times, 2 vehicles at least 5 times and so on. Similarly, finding the right value of eps is crucial for a reliable performance of DBSCAN. Determining the optimal value for eps involves evaluating the density of points within a cluster. A commonly used method is the k-distance graph, where k is

typically set to minPts - 1. This method involves several key steps: first, calculating the k-distances for each point in the dataset by determining the distance to its k^{th} nearest neighbor is performed. Next, sorting these distances in ascending order should be done. After sorting, creation of a plot of the k-distances, with the x-axis representing the index of the points and the y-axis representing the k-distance values is performed. Finally, identifying the elbow point on the plot, which represents the point of maximum curvature and serves as a strong candidate for the eps parameter, as noted by Sander *et al.* (1998) is determined. Domain knowledge can also inform the selection of eps, particularly if there is an understanding of the scale of clusters or the density of points in the dataset. To determine eps for this study, the known parking locations, i.e., dealer locations, described earlier were analyzed. The k-distance plot was also investigated and presented in Figure 5-4. Both sources suggested that a 0.2 km (200 meters) radius is an optimal value for eps.

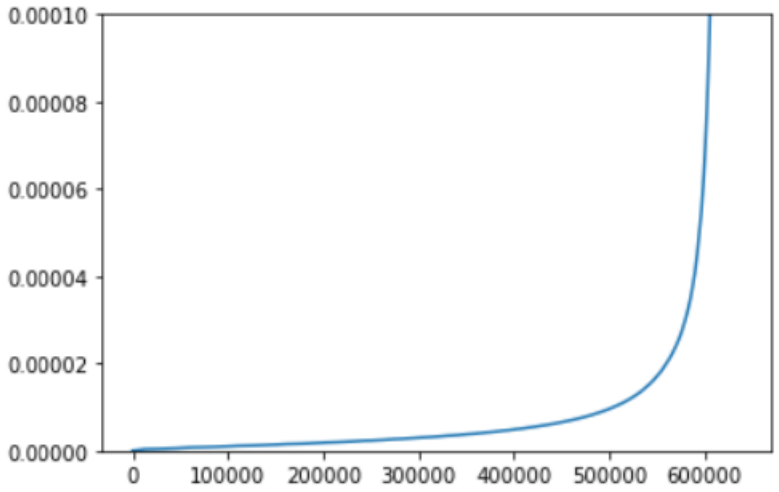


Figure 5-4 Plot of k^{th} nearest distance to determine optimal eps

In this plot, the x-axis represents the indices of the points ordered, and the y-axis represents the k^{th} nearest neighbour distances in kilometers.

3. After training of the DBSCAN, centroids of each cluster are retrieved. One cluster is considered as one parking location. Therefore, the centroid of each cluster is taken as a representation of the identified parking locations.
4. The centroid points as parking locations are stored in a database as Parking Locations dataset to enrich feature sets for later use in the supervised ML component.

Figure 5-5 shows the result of the DBSCAN, and 13,219 parking locations are identified and stored with this approach.

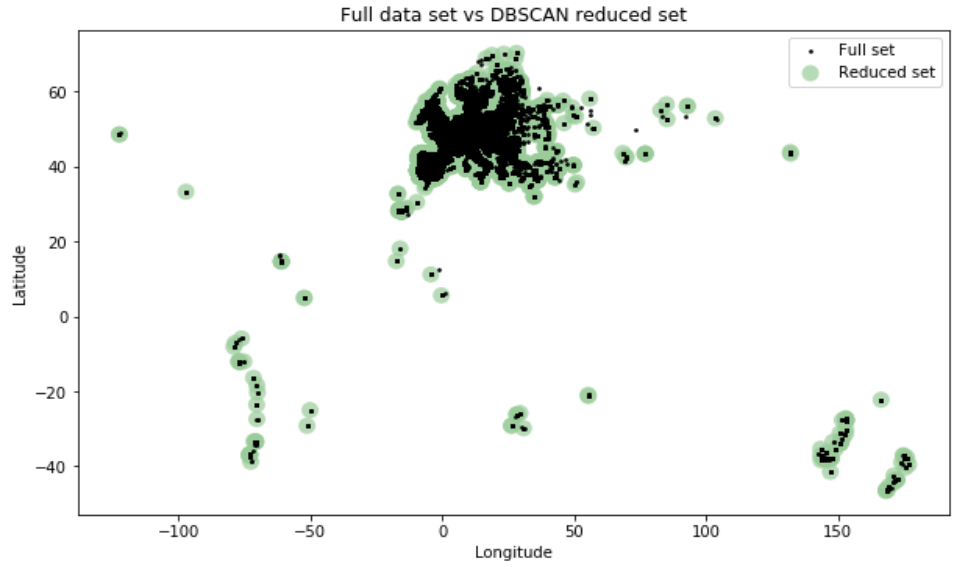


Figure 5-5 DBSCAN result plot for parking location points (clusters and corresponding centroids)

For better visibility, the above plot is filtered for the Dutch regions of Limburg and North Brabant and presented in Figure 5-6 below

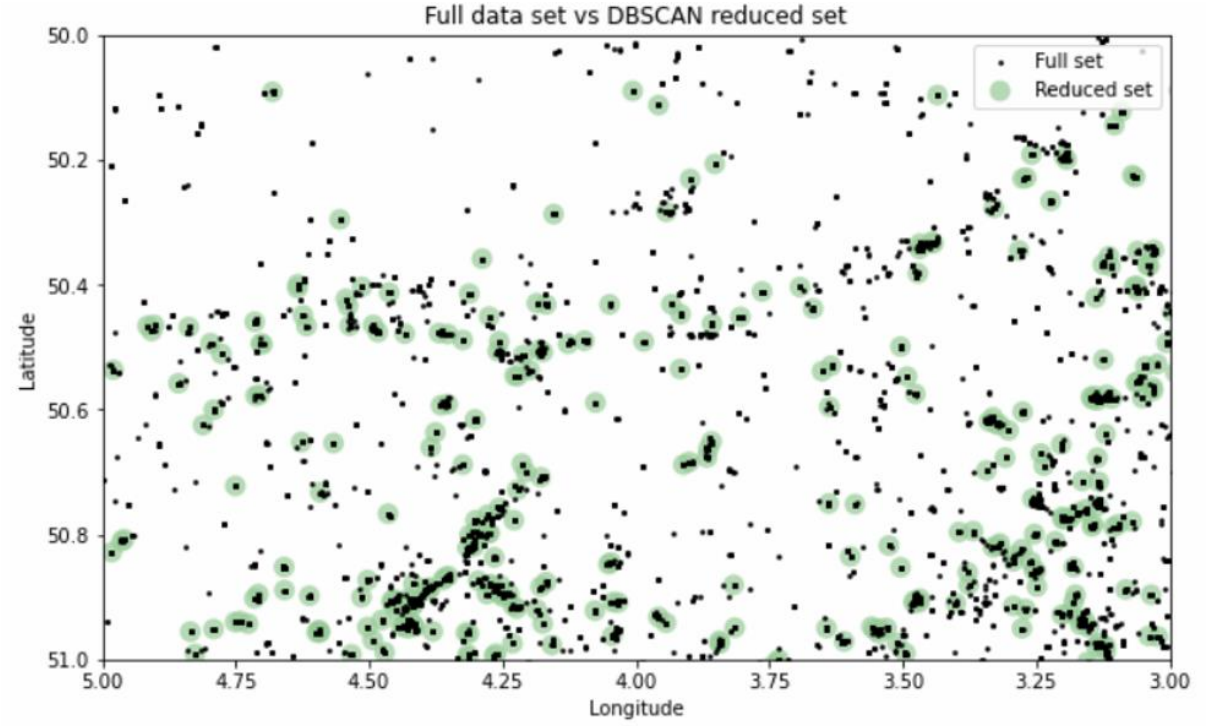


Figure 5-6 subset of DBSCAN result filtered for Limburg and North Brabant regions, The Netherlands

To have a better understanding, visualization of part of the reduced points or clusters which are representations of parking locations on the map, is given in Figure 5-7 below.

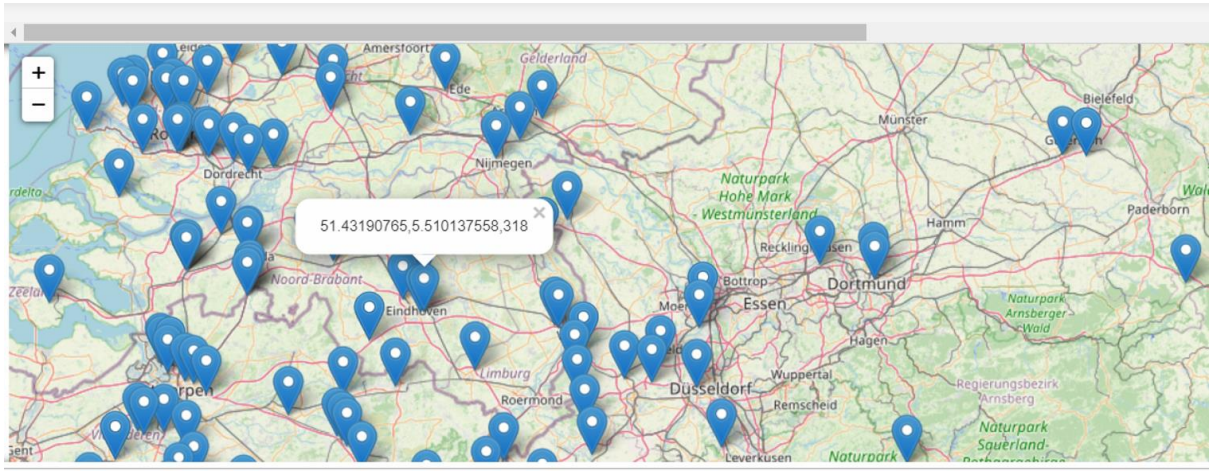


Figure 5-7 Part of DBSCAN result on the map

Further, zooming in to individual points helps verify if the identified location is a parking location. For example, Figure 5-8 given below shows one example of an identified parking location from Figure 5-5. As shown in the figure, there are multiple parking spots. This algorithm identifies the centroid (blue point) and stores the corresponding GPS coordinates.

Figure 5-8 Example of identified parking locations (blue point or centroid used as the GPS point for the parking location)

The complete result is stored for later stage, i.e., supervised ML step to derive a new feature by calculating how far the vehicle/vehicles in question is from the nearest parking location. Figure 5-9 below shows the sample location centroids of the identified parking locations.

GPSLATITUDE	GPSLONGITUDE	NUMBER Of PARKING EVENTS
[redacted]	[redacted]41	4328
[redacted]	[redacted]33	2936
[redacted]	[redacted]79	2867
[redacted]	[redacted]05	2657
[redacted]	[redacted]34	2193
[redacted]	[redacted]92	2184
[redacted]	[redacted]97	1970
[redacted]	[redacted]48	1941
[redacted]	[redacted]39	1939
[redacted]	[redacted]89	1868
[redacted]	[redacted]77	1734
[redacted]	[redacted]33	1687
[redacted]	[redacted]27	1571
[redacted]	[redacted]27	1566
[redacted]	[redacted]92	1507
[redacted]	[redacted]11	1471
[redacted]	[redacted]97	1462
[redacted]	[redacted]38	1460
[redacted]	[redacted]45	1443
[redacted]	[redacted]25	1434
[redacted]	[redacted]42	1419
[redacted]	[redacted]08	1410
[redacted]	[redacted]24	1406
[redacted]	[redacted]38	1402

Figure 5-9 Sample coordinates of the identified parking locations

5.2.2.2 *Supervised learning to identify non-communicating vehicles*

To determine if a vehicle is communicating or not, a supervised ML algorithm is trained following the architecture given in Figure 5-10 below. The main goal of this exercise is that: given a set of vehicles which are not sending messages, the metric displayed as the number corresponding to "Non-comm Vehicles" in the prototype dashboard of Figure 4-2 , it tries to distinguish the ones likely experiencing an issue and therefore not able to communicate versus the ones that are properly parked, main power is switched off and therefore not experiencing any issue and will communicate back without any missing data or delayed data when the issue is fixed. As such, the problem is formulated as a two-class classification, i.e., non-

communicating due to an issue as **Yes** or **No,** classification problem and therefore a supervised ML algorithm is investigated.

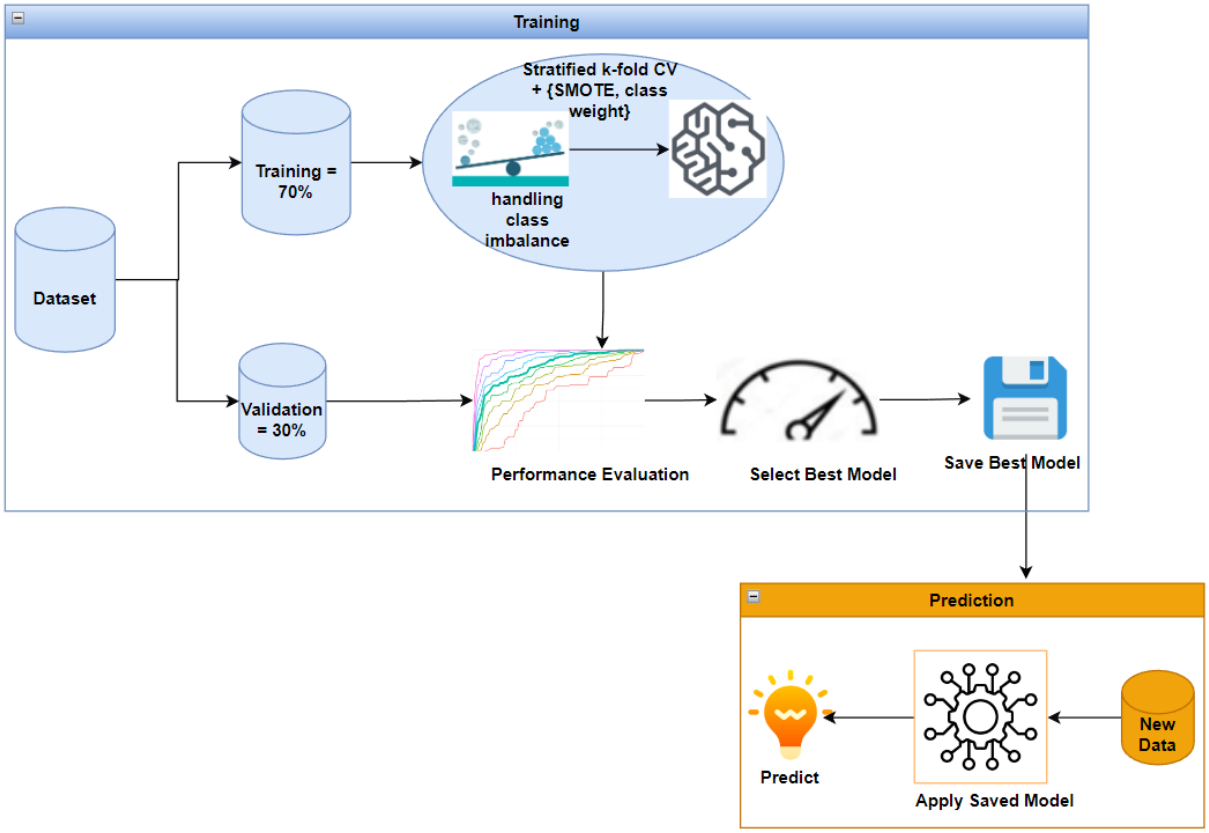


Figure 5-10 Supervised machine learning architecture adopted to identify non-communicating vehicles

Since this is a continuation of Iteration 1, much of the data understanding and exploration work is already performed in Chapter 4, and this chapter describes only the remaining subsequent steps in the process. The proposed architecture starts by splitting the dataset in 70% to 30% ratio for training and validation, respectively. Subsequently, a logistic regression model is trained, wherein the issue of imbalanced data is addressed using different techniques including adjusting class weights, oversampling minority class and under sampling majority class in combination with k-fold cross validation. Next, the best performing model is selected and saved for future prediction on new data. The detailed description of the steps depicted in Figure 5-10 is described as follows.

A) Feature Engineering

First the dataset (described in Table 5-1 and Table 5-2) is enriched with distance to the nearest parking location when it was last communicating with the help of the parking locations data generated using the DBSCAN algorithm discussed in this section earlier. In other words, for

each row in the source data, the minimum haversine distance to the GPS locations of the parking dataset, i.e., minimum of distance between (PREV_GPSLATITUDE, PREV_GPSLONGITUDE) from source data and (GPSLATITUDE, GPSLONGITUDE) from the parking data is calculated. The algorithm used to find the distance to the nearest parking location is presented below.

Algorithm 5.1: Calculation of the nearest parking location

***Inputs**: Non-communicating dataset and ParkingLocations dataset*

***Output**: Non-communicating dataset enriched with distance to nearest parking*

Initialization:

Distance_to_each_parking[] ←*empty // distance of a GPS point in the input dataset to each parking location, initially empty*

DIST_CLOSEST_PARKING[]←*empty //the minimum of Distance to each parking location, initially empty*

For each record (i) in Non-communicating dataset:

For each record (k) in ParkingLocations dataset:

Distance_to_each_parking[k]← *Haversinedistance(TrainingDataset[i](latitude, longitude), ParkingLocation[i](latitude, longitude))*

DIST_CLOSEST_PARKING[i] ←*minimum (Distance_to_each_parking[])*

Merge DIST_CLOSEST_PARKING[] to Non-communicating dataset

End

The haversine distance, sometimes referred to as the great circle distance (Maria *et al.*, 2020), is applied to compute the distance to the closest parking location. The haversine distance is a mathematical formula used for determining the distance between two sets of geographic coordinates. This calculation requires the longitude and latitude values as input parameters (Maria *et al.*, 2020). The formula of haversine distance is presented in equation 5.1

$$= 2\mathrm{r}\arcsin\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)} \quad 5.1$$

in which

- r is the radius of the Earth, and it is equal to 6371 km.
- φ_1, φ_2 are the latitude of the first GPS point and latitude of the second GPS point.

- λ_1, λ_2 are the longitude of the first GPS point and longitude of the second GPS point.

Without incorporating the new feature, i.e. DIST_CLOSEST_PARKING, the distribution of the target variable, i.e. NOCOMM as described in Table 5-2, is 76.29% (**No**) to 23.71% (**Yes**), which means out of the vehicles which are identified as non-communicating or not transmitting data using the initial assessment prototype dashboard in section 4.2.2, 76.29% of them communicated back without any missing data or delayed data and 23.71% communicated back with missing data or delayed data or both as shown in Figure 5-11.

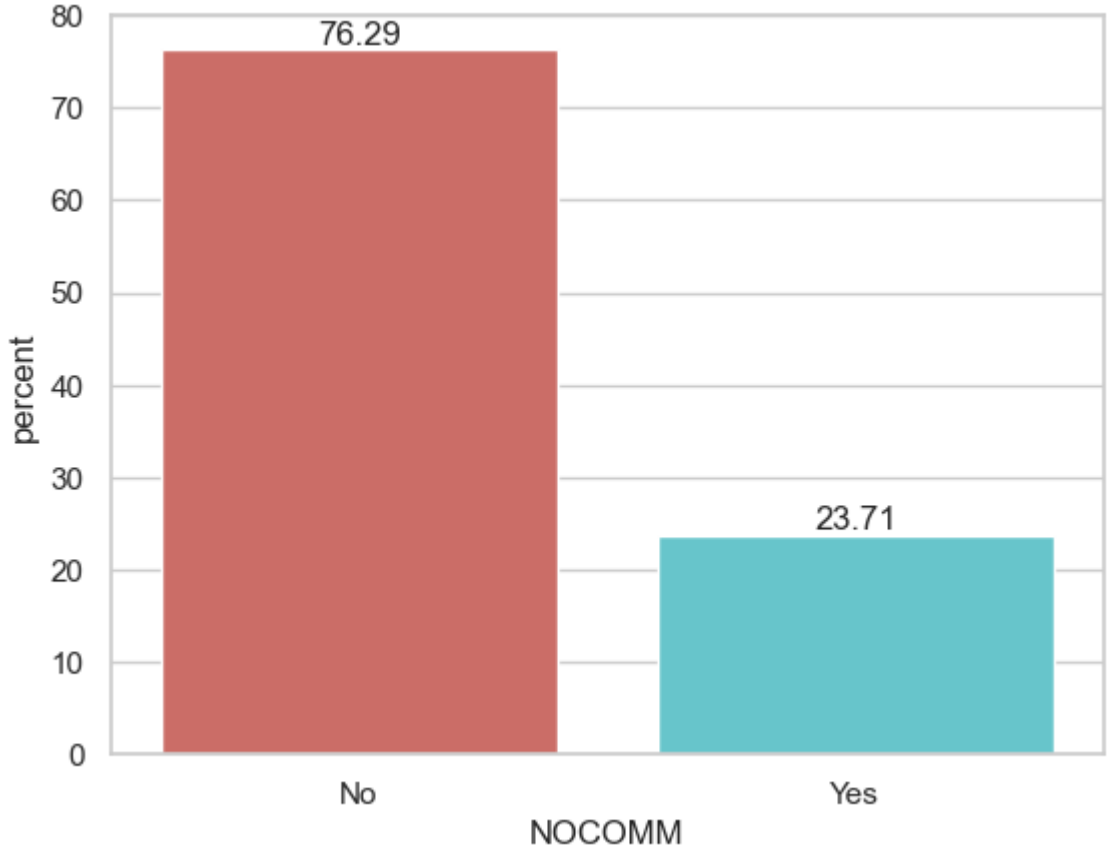


Figure 5-11 Distribution of NOCOMM [Yes versus No] in the data collected for non-communicating classification

By adding this feature and filtering for closest parking distance to less than or equal to 400m (<=400m), which is equivalent to filtering the non-communicating dataset where the last known communication point was near one of the identified parking locations within a 400 meters distance radius from the centroid of the identified parking locations, the distribution becomes 92.63% (**No**) to 7.37% (**Yes**), which means only 7.37% instead of 23.71% of vehicles whose last position was within 400m radius of the identified parking location are actually non-communicating or facing real communication issue. Extending the filter to nearest parking

locations to less than or equal to 200m (<=200m) strengthens this statement further making the distribution 98.00% (**No**) to 2.00% (**Yes**) which means only 2.00% of vehicles whose last position was within 200 meters of one of the identified parking locations communicated back with either missing data or delayed. The remaining 98.00% communicated back without delayed data or missing data. After enriching the dataset with this new feature, the shape of the dataset looks as follows.

```
(298476, 45) Index(['RONAME', 'EVENTID', 'EVTDATETIME', 'ROWNUM', 'GPSLATITUDE', 'GPSLONGITUDE', 'PREV_GPSLATITUDE', 'PREV_GPSLONGITUDE', 'PREV_ROWNUM', 'PREV_EVTDATETIME', 'PREV_EVENTID', 'PREV_VDIST', 'VDIST', 'VDIST_DIFF', 'RECEIVEDDATETIME', 'PREV_RECEIVEDDATETIME', 'PREV_DAY', 'COMM_DAY', 'DELAY_HRS', 'EVT_TIME_DIFF_HR', 'TIME_DIFF_HR', 'NEXT_STARTING_POINT_KM', 'SNP_PREV_VDIST', 'SNP_PREV_EVENTID', 'PREV_SNP_RECEIVEDDATETIME', 'SNP_NEXT_VDIST', 'SNP_NEXT_EVENTID', 'NEXT_SNP_RECEIVEDDATETIME', 'MISSING_SNP', 'SNP_PREV_GPSLATITUDE', 'SNP_PREV_GPSLONGITUDE', 'SNP_NEXT_GPSLATITUDE', 'SNP_NEXT_GPSLONGITUDE', 'SNP_NEXT_DIR_EVTDATETIME', 'SNP_NEXT_DIR_RECEIVEDDATETIME', 'SNP_NEXT_DELAY_HRS', 'PREV_SNP_DAY', 'NEXT_SNP_DAY', 'SNP_MISSING_GPS', 'NOCOMM', 'COUNTRY', 'CUSTOMER_LINKED', 'NUMBER_MISSING', 'NR_BUFFER', 'Dist_Closest_parking'], dtype='object')
```

Note: the extra feature ***'Dist_Closest_parking'****, which is generated using the result of the DBSCAN result by applying haversine distance to each data point.*

Another important data element that shows an influence on the distribution of the target variable is the event which was observed just before the vehicle stopped communicating or sending data, i.e., last known event type. The meaning of the events is given in Table 5-1. As shown in Figure 5-12 below, about 92.00% of the vehicles which were not sending data communicated back without missing data or delayed data when the last event is 5 which represents trip end event. However, when examining the vehicles that did not send data and last event was trip start or timer (while on trip), it was found that approximately 57% and 42% of them reported issues such as missing data and delayed data, respectively. These issues occurred specifically during event 1 (timer event or time interval event) and event 4 (trip start event) respectively. This finding is logical, as it is more probable for a vehicle to encounter problems if it stops sending data during a trip (event 1 – timer event) rather than after completing the trip (event 5 - trip end event).

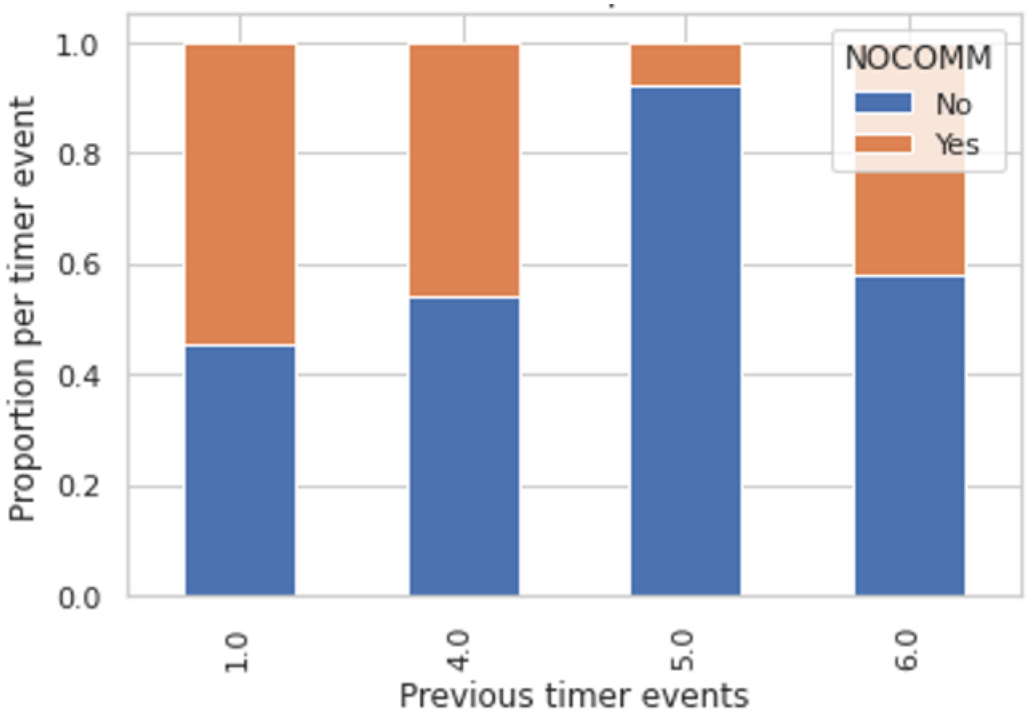


Figure 5-12 Influence of last known event on non-communication

Other features in the dataset are also investigated but no clear correlation was observed. For example, the day of the week is reported as an important feature in (Azimi and Pahl, 2021). However, no clear distinction was observed for communication state in this dataset as depicted in Figure 5-13 except a slight difference in the weekends.

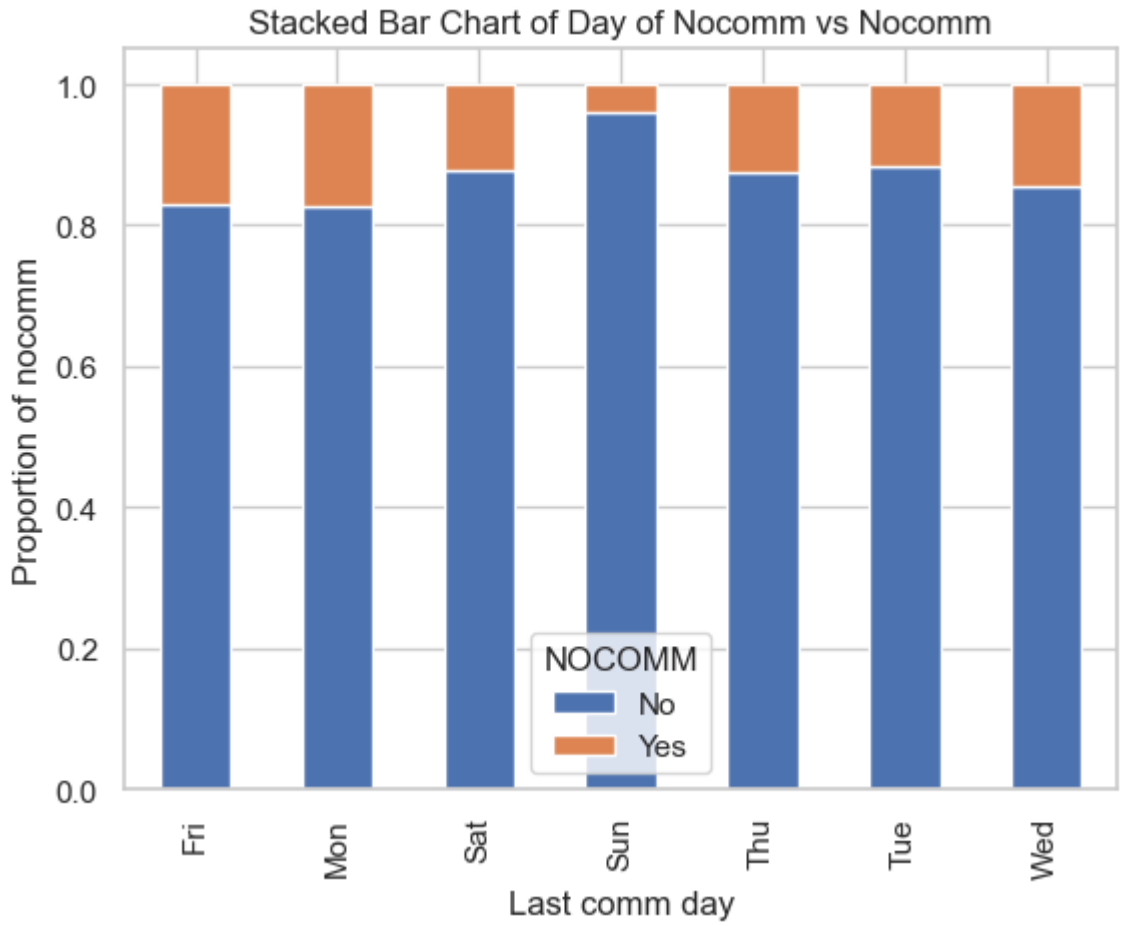


Figure 5-13 Influence of days of the week on non-communication

B) Feature importance

To evaluate the importance of each feature, a Random Forest Feature selection technique is employed. The outcome is presented in Figure 5-14 given below. The diagram illustrates the significance of the features, considering their weight and the direction of their influence, whether it is positive or negative. During the data pre-processing stage, categorical features were transformed using one-hot encoding. Several ML algorithms are unable to directly process label data (Kursa and Rudnicki, 2010). Therefore, it is important for all input variables and output variables to have numerical values. As a result, the above-mentioned figure exhibits additional features that are not included in the lists provided in Table 5-1 and Table 5-2 earlier. For example, the first one-hot encoded category is labelled as NUMBER_MISSING, which denotes the frequency of instances in which the vehicle has encountered a discrepancy in mileage or trip records in previous occurrences (as outlined in Table 5-2). So, one-hot encoding method is used to convert the buckets created as categorical values for the variables to numeric

values. The description of the most crucial features encoded using the one-hot encoding technique is presented below.

- NUMBER_MISSING =

{

MN_0-1: True or 1 if it has no missing event historically, 0 otherwise.

MN_1-5: True or 1 if it has 1 or 2 or 3 or 4 missing events historically, 0 otherwise.

MN_5-10: True or 1 if it has 5 or more but less than 10 missing events historically, 0 otherwise.

MN_10+: True if it has 10 or more times missing events historically, 0 otherwise.

}

- NR_BUFFER =

{

BN_0-1: True or 1 if it has no delay/buffer event historically, 0 otherwise.

BN_1-5: True or 1 if it has 1 or 2 or 3 or 4 delayed events historically, 0 otherwise.

BN_5-10: True or 1 if it has 5 or more but less than 10 delayed events historically, 0 otherwise.

BN_10+: True if it has 10 or more delayed events historically, 0 otherwise.

}

- DIST_CLOSEST_PARKING =

{

0-200: True or 1 if last position is within 200m from one of identified parking locations.

200-400: True or 1 if the last position is more than 200m but less than 400m distance from one of the parking locations identified.

400-800: True or 1 if the last position is more than 400m but less than 800m distance from one of the parking locations identified.

800+: True or 1 if the last position is more than 800m to any of the parking locations identified.

}

- EVENTID=

{

1 = True or 1 if the last message was a timer message.

4 = True or 1 if the last message was a trip start message.

5 = True or 1 if the last message was a trip end message.

}

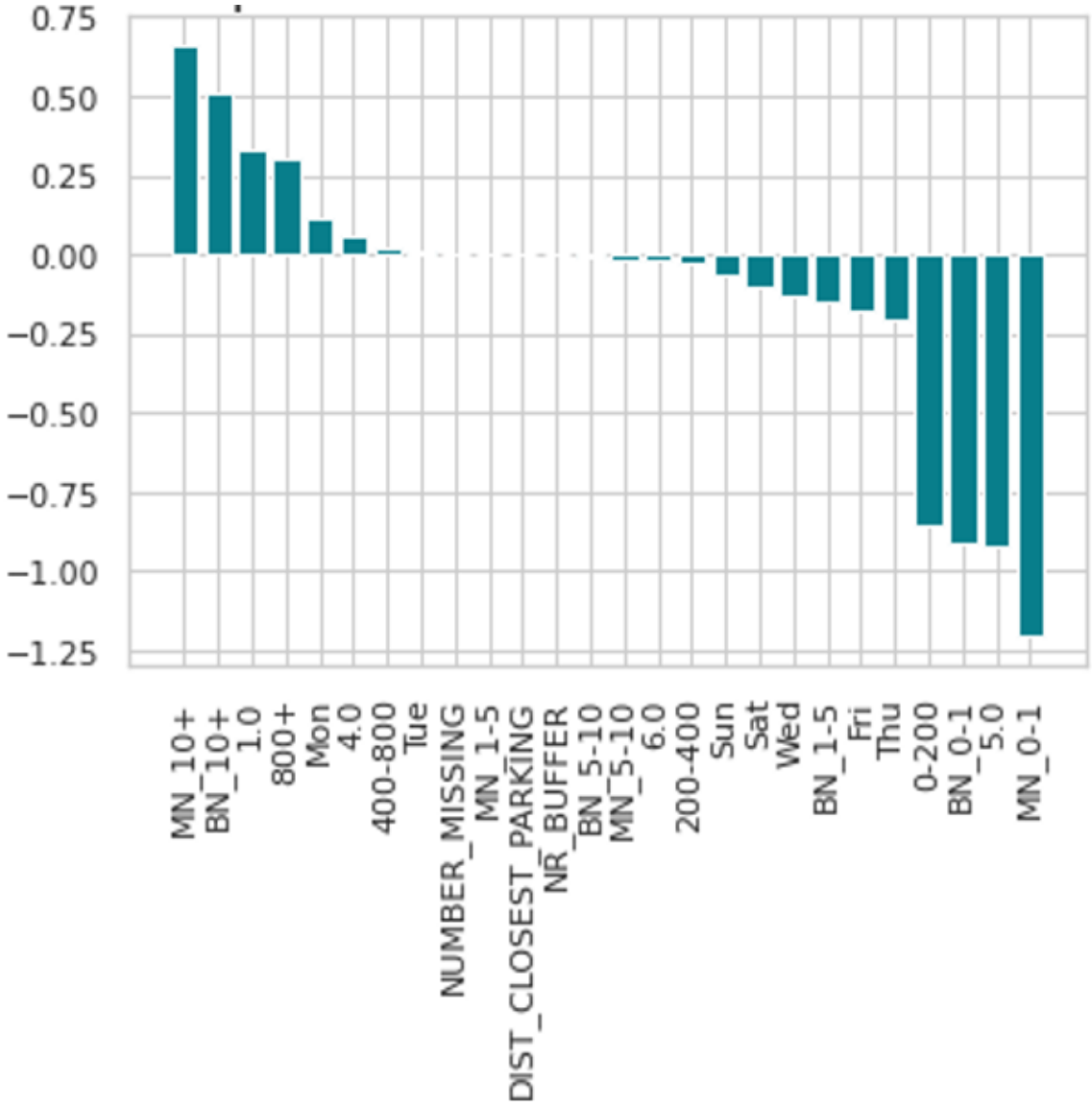


Figure 5-14 Feature importance of variables used for classification of nocomm [Yes/No]

From Figure 5-14 above, it is clearly shown that some features increase the chance of a vehicle of being in a non-communicating state (positive side of the Y-axis), and others decrease the likelihood of the vehicle being in a non-communicating state, i.e., the chance of missing data or communicating back with a delayed data. The top four features which have high influence for a vehicle to be in non-communicating state with real issue are {MN_10+, BN_10+,1.0, 800+} and the top four features which have high influence for the vehicle to come back with no missing or no delayed data are {MN_0-1, 5.0, BN_0-1,1.0, 0-200}. This proves that NUMBER_MISSING, NR_BUFFER, DIST_CLOSEST_PARKING, EVENTID are key features. In other words, a vehicle which has experienced a non-communication state before

more frequently is highly likely to be in non-communication state again. Next, the last received message type has an influence in such a way that if it is 1 = time interval message or 4 = trip start message, then it is highly likely that the vehicle is facing some kind of issue which prevents the vehicle from sending data. On the other hand, if it is 5 = trip end message, then it is highly likely that the vehicle is not facing an issue. This expectation is reasonable as, in a typical scenario, a moving vehicle is expected to consistently transmit data unless it faces an undesired situation that prevents it from doing so. In addition, the DIST_CLOSEST_PARKING is an important feature. If a vehicle is marked within 200m from a known parking location, then the chance that it will have missing data or delayed data is small. However, if the last known location is far from any parking location by more than 800m, then it is highly likely that the vehicle is facing a problem and hence will have missing data or delay.

C) Model Building

To train a model which generalizes effectively on new data, it is essential to develop a good strategy to split data into training, validation, and test sets to avoid evaluating the developed model on the same dataset employed for training.

Selecting an optimal train-test split depends on a range of factors including dataset size, problem complexity, and analysis objectives. While there is no one-size-fits-all ratio approach, certain guidelines and best practices are often recommended based on different influencing factors. One factor to consider is the size of the training data. For example, (Kohavi, 1995) recommends 80%/20% split ratio for a moderate size dataset and this is the most applied train/test ratio. For small datasets, 90%/10% can also be considered. For a larger dataset, (James *et al.*, 2013) recommends 70%/30% for robust evaluation. There are also some researchers using 75%/25% split ratio to provide a reasonable compromise (Hastie *et al.*, 2009). To mitigate the dependency on a single train/test split and to make better use of the data, cross-validation is often employed. Cross-validation is a technique that allows for the estimation of a model's performance with reduced variance compared to a single train/test split (Delen, Walker and Kadam, 2005). In cross-validation, the dataset is divided into k equally sized folds and the model is trained k times, each time using k-1 folds for training and the remaining fold for testing. Then, the results are averaged to provide a more robust estimate of model performance (Kohavi, 1995). In case of class imbalance, an enhanced variant of k-fold cross-validation called stratified k-fold cross validation is recommended to get a more balanced training and evaluation sets by ensuring that each fold has a similar class distribution to the entire dataset (Kohavi,

1995). Preserving the class distribution also results in a reliable and consistent performance estimate and reducing variance in model evaluation (James *et al.*, 2013).

Another factor to consider is the complexity of the chosen learning algorithm. If the learning algorithm is relatively simpler, a smaller training dataset may be enough while a bigger training dataset is recommended for complex ML algorithms in order to capture the complexity of the dataset (Hastie *et al.*, 2009).

Another factor to consider is the purpose of the analysis. If the main objective of the study is to assess the effectiveness of the trained model, a bigger test set such as 70%/30% split ratio is recommended which increases the chance of getting a more reasonable model that can generalize well. However, if the objective of the research is to develop the best model, bigger training data with split ratio of 80%/20% is recommended (James *et al.*, 2013).

Another crucial factor to consider is the target variable's proportion. If the data exhibits class imbalance, the normal k-fold cross validation may not be optimal. Instead, stratified k-fold cross validation may be used to produce a fair evaluation ((James *et al.*, 2013).

Considering all the factors mentioned above, for this research, to produce a fair evaluation the training dataset is split on a basis of 70%/30% ratio for training and validation since: 1. the dataset is large 2. the dataset exhibits a class imbalance and 3. The objective of this study is to build a well generalizing model.

Traditionally, it was commonly believed that evaluating a model's performance based on the validation set was enough to accept that the developed model is effective. However, recent studies have challenged this understanding. For example, (Westerhuis *et al.*, 2008) discovered that performance metrics derived from cross-validation are often overly optimistic. Similarly, (de Boves Harrington, 2006) demonstrated that relying on a single split of training and test sets can lead to inaccurate estimates of model performance. These findings highlight the necessity of an additional blind test set, which is not involved in the model selection and validation processes, to achieve a more accurate assessment of the model's generalization performance.

Therefore, this study utilizes a test set alongside the validation set. The train-validation-test split design of this study is illustrated in Figure 5-15 below.

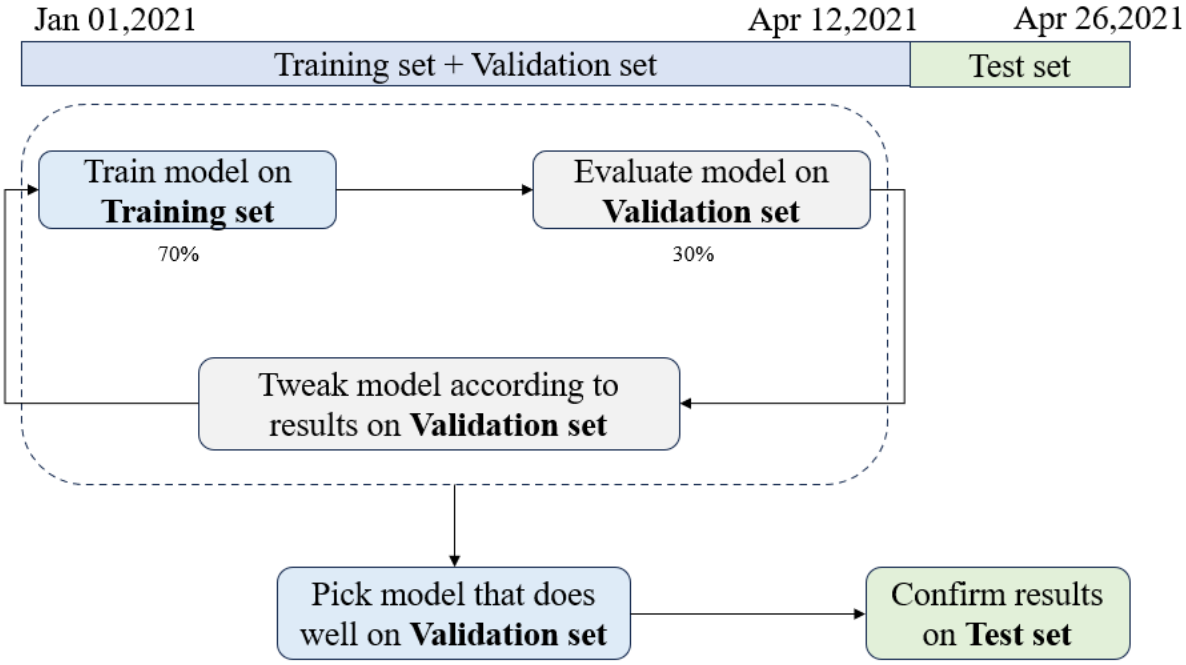


Figure 5-15 train/validation/test split strategy of the dataset

The learning algorithm chosen for this study is the logistic regression supervised learning method. The use of logistic regression as a classification algorithm for this activity is motivated by the need for interpretability. The main goal of scenario I is to avoid or minimize missing data and delayed data hence addressing completeness and timeliness DQ dimensions by detecting non-communicating vehicles and taking necessary action to resolve the issue. This requires sorting based on the likelihood that the vehicle is in non-communicating state since acting on all non-communicating vehicles is costly. Logistic regression works based on underlying probabilistic model and returns the probabilities to each data row in the data based on which sorting can be done, and action can be taken. Logistic regression is considered as one of the most interpretable algorithms (Li, 2013). As stated in (Han, Wu and Yang, 2022), "logistic regression is the most popular algorithm because of its simplicity, effectiveness, and strong interpretability". Besides, logistic regression is less affected by various training datasets. Therefore, following the adopted approach, the logistic regression model is trained using stratified k-fold cross-validation, where k is set to 10. During the data exploration phase, it became evident that the dataset exhibits a significant class imbalance. Therefore, different methods have been investigated to handle the observed class imbalance. Initially, the Synthetic Minority Oversampling Technique (SMOTE) (Kotsiantis *et al.*, 2006) has shown an

improvement. Further, following numerous iterations employing various techniques including the use of class balancing through class weights, provided the best outcome. Table 5-3 presents the different configurations employed in the experiment, together with their respective outcomes. In the context of class imbalance, the F1-Score metric is commonly used for performance evaluation due to its superiority in assessing unbalanced data sets, as opposed to other metrics like accuracy (He and Garcia, 2009). Therefore, F1-score is used as a primary performance metric together with other classification metrics as outlined in section 3.7.2.1.

Table 5-3 Selected results of training experiment

Configuration	F1-score
Logistic regression	0.51
Logistic regression + SMOTE	0.68
Logistic regression + SMOTE + Grid Search CV	0.73
Logistic regression + Class weight	0.76
Logistic regression + Balance with class weight + Grid Search CV	0.81

The basic model configuration resulted in an F1-score of 0.51. Subsequently, the SMOTE was employed, resulting in an increase in the F1-score to 0.68. The application of SMOTE in combination with grid search has resulted in a further improvement of the F1-score performance metric to 0.73. However, the optimal outcome was achieved by employing class weights in combination with grid search to determine the most suitable parameter for class weights. Logistic regression has several parameters that can be tuned to improve the model's performance. For Scenario I, the Grid Search for best model has set the parameters given below.

Best: 0.806369 using {'C': 0.01, 'class_weight': {0: 1, 1: 10}, 'penalty': 'l2', 'solver': 'liblinear'}

Where:

- *C* represents the inverse of regularization and is used to prevent overfitting. A lower value indicates a strong regularization, which helps to prevent overfitting. In this experiment, C is set to a lower value of 0.01.
- *Class_weight* represents the weights associated with the classes.
- *Penalty* is used to specify the norm used in the penalization and determine the type of regularization to be applied. The possible values are l1, l2, elasticnet, and none. In this experiment, the grid search has set penalty to l2 (Ridge) which tries to distribute the error among all terms.
- *Solver* is used to specify the algorithm to find the optimal parameters.

Each parameter plays a vital role to tune the model for performance and adapt to a given dataset (James *et al.*, 2013).

The achieved F1-score with this configuration is 0.81 for non-communicating class, 0.88 macro average and 0.94 weighted average as presented in Table 5-4. The confusion matrix of the selected configuration which resulted in the best outcome is presented in Figure 5-16 below.

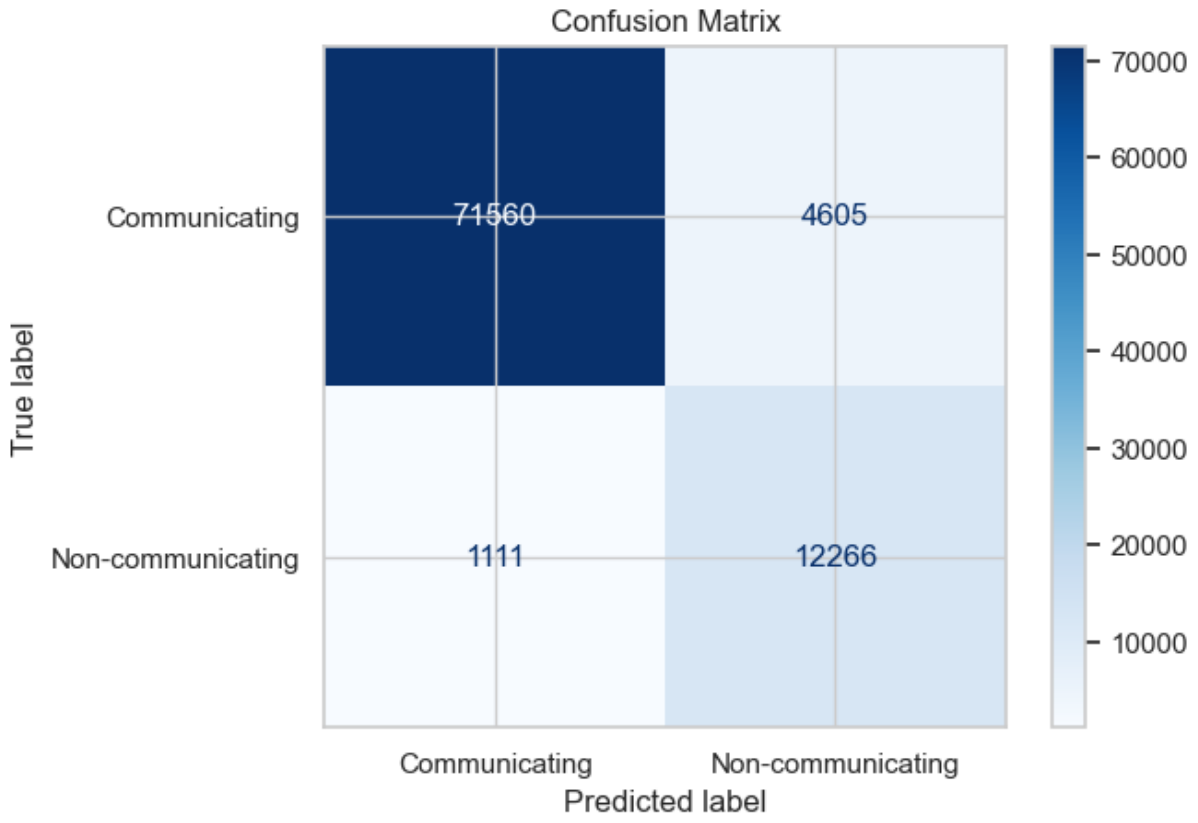


Figure 5-16 Confusion matrix of the train validation set

The corresponding classification report of the selected configuration which resulted in the best outcome is presented in Table 5-4 below.

Table 5-4 Classification report of the best model on the training validation set

	precision	**recall**	**f1-score**	**support**
Communicating	0.99	0.94	0.96	76165
Non-communicating	0.72	0.92	0.81	13377
accuracy			0.94	89542
macro avg	0.85	0.93	0.88	89542
weighted avg	0.95	0.94	0.94	89542

In the end, the best performing model is stored and applied to predict new non-communicating vehicles on new blind test dataset following the strategy depicted in Figure 5-15.

5.2.3 Evaluation of Scenario I

As explained in section 3.3, this scenario employs two ML methods. First, DBSCAN is used to identify potential parking locations of vehicles which is used to generate a new feature for the next step of classification. Then, a logistic regression is trained to classify vehicles as communicating or non-communicating. In this section evaluation of the two methods is presented.

5.2.3.1 Unsupervised learning - DBSCAN

For the DBSCAN algorithm, two evaluation methods are employed. First, the model's performance is assessed using the silhouette coefficient and then its effectiveness is evaluated using the dealer locations data described in section 5.2.2 as a test dataset.

1. Using silhouette Coefficient

The Silhouette Coefficient is a method used to evaluate clustering algorithms (Rousseeuw, 1987). The Silhouette Coefficient ranges between -1 and 1, where 1 indicates the best clustering, and -1 indicates the worst clustering outcome. Higher scores suggest well-defined, dense clusters, while values near 0 signify overlapping clusters. Negative values usually indicate wrongly assigned data points. Typically, a Silhouette score above 0.5 is considered acceptable for DBSCAN. However, the main drawback of the Silhouette Coefficient is its high runtime for large datasets (n), requiring O(n) calls. In this experiment, for instance, it took 6 hours on ml.t2.2xlarge Notebook on AWS to process 1,304,516 records, resulting in 13,219 clusters, equivalent to the number of parking locations identified in section 5.2.2. However, this is required once and is not expected to have a significant impact on operationalization. The DBSCAN model in this experiment achieved a Silhouette Coefficient of 0.632, which is higher than the acceptable threshold of 0.5.

2. Comparison with available parking locations

As stated earlier in section 5.2.2, the dealer locations are already available. The dealer locations were found to be strongly correlated with the communication state of the vehicle. However, since vehicles park in other locations in addition to dealer locations, this activity was initiated to identify all potential parking locations. On the other hand, this method is supposed to capture

all the parking locations including dealer locations. In other words, the dealer locations can serve as a test dataset for the DBSCAN model. Therefore, a comparison is performed to see how many of the dealer locations are captured in this method. With the help of a haversine distance formula, a look up is performed to find the nearest identified parking location for each dealer in the database. Table 5-5 below presents examples of identified parking locations.

Table 5-5 Partial view of closest identified parking location to dealer locations

Dealer	LATITUDE	LONGITUDE	Gps (lat, long)	Closest_parking
Dealer 1	52.1347	-0.4371	(52.1347, -0.4371)	5
Dealer 2	38.2275	-3.631	(38.2275, -3.631)	5
Dealer 3	53.8889	10.6912	(53.8889, 10.6912)	8
Dealer 3	48.149	13.9819	(48.149, 13.9819)	8
Dealer 4	44.9066	8.8923	(44.9066, 8.8923)	8
Dealer 5	51.3957	0.516	(51.3957, 0.516)	10
…	…	…		

There are in total 835 dealer locations in the database. In line with section 5.2.2 for the impact of parking locations on non-communicating; filtering for distance to the closest parking location less than or equal to 400m resulted in 672 records, which is 80.47%. In other words, this method has identified or captured 80.47% of the dealer locations correctly as parking locations.

5.2.3.2 Supervised Classification with logistic regression

The second component of Scenario I is a logistic regression classification model. To evaluate this model, historical data is used. Following the strategy to split the data into training set, validation set and hold out a test set shown in Figure 5-15, the data from April 12 to April 30, 2021, is used to assess the model on a blind new test dataset.

The logic to prepare the historical data is illustrated in Figure 5-1. And the same metadata as described in Table 5-1 and Table 5-2 is applied. Using this logic, a vehicle is labeled as nocomm (**Yes/No**) depending on the availability of missing data or delayed data. A total of 34,637 data points from the recovered class of Figure 5-1 were collected over the specified period. Subsequently, a prediction is executed by employing the model stored earlier as the best model.

The output of the prediction results in the *vehicleid* together with the probability that the vehicle will come with a problem (with missing data or delayed data) when it communicates back as presented in Table 5-6 below. The output is used to filter vehicles which are highly likely facing

an issue so that action can be taken to avoid further delay or missing data. The default cut-off point, i.e. 0.5, is used as a threshold to filter NOCOMM "**Yes**" or "**No**".

Table 5-6 provides sample prediction output with likelihood of non-communicating

vehicleid	predicted
9748	0.31
49749	0.96
58830	0.98
59205	0.97
65956	0.99
67443	0.28
69754	0.27

The confusion matrix and the classification report on the prediction result of this dataset by applying the best model is presented in Figure 5-17 and Table 5-7 below respectively.

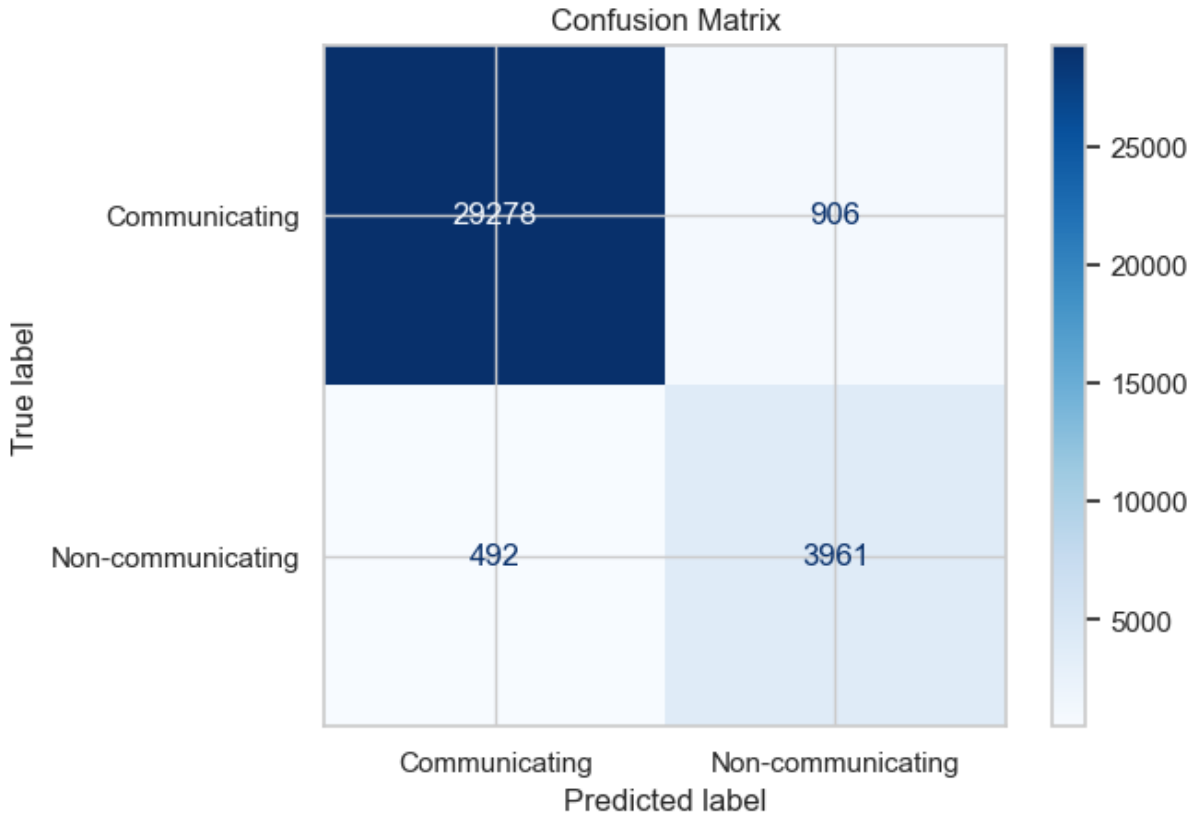


Figure 5-17 Confusion matrix of the independent test set

Table 5-7 Classification report of the best model on the independent test dataset

	precision	recall	f1-score	support
Communicating	0.98	0.97	0.98	30184
Non-communicating	0.82	0.89	0.85	4453
accuracy			0.96	34637
macro avg	0.9	0.93	0.91	34637
weighted avg	0.96	0.96	0.96	34637

The F1-score on this test dataset was 0.85 on non-communicating class, macro average 0.91 and weighted average 0.96. The F1-scores obtained for the test dataset are close to the F1-scores obtained on the training validation set which are 0.81, 0.88 and 0.94 for non-communicating class, macro average and weighted average respectively as presented in Table 5-4.

5.3 Scenario II: Predicting Mileage

5.3.1 Problem description of missing mileage

This scenario builds upon Scenario I by shifting the focus from assessment to improvement. Its main goal is to predict the likely mileage of a given set of vehicles that have been identified as non-communicating or experiencing some form of issue through the method developed in scenario I, which detects missing data or delay. Mileage, which represents the total distance driven by the vehicle, is selected for this scenario as it is one of the most important and commonly used data elements, for example in warranty contracts. Reporting incorrect mileage will, therefore, have a financial consequence directly or indirectly.

5.3.2 Proposed solution to forecast mileage

Taking the latest mileage status for all vehicles and including that in a report to be distributed to customers may lead to taking a wrong decision as the vehicle may not be communicating while it is still driving. Therefore, it is important to know the status of the vehicles before distributing the latest information. Whether a vehicle is sending data or not can be easily identified as explained in section 4.2.2, and whether a vehicle identified as not sending data is due to facing an issue or normally parked can be predicted with the method explained in section 5.2. Given the status of vehicles as communicating and non-communicating (facing an issue), it is also useful to forecast the current mileage for vehicles that are identified as non-communicating while they may be driving. The purpose of this section is, therefore, to forecast the missing mileage of vehicles that are predicted as non-communicating due to an issue. For

this purpose, time series forecasting using historical data is employed. This section describes the application of time series analysis to forecast missing mileages of vehicles not communicating and identified as facing issues.

Dataset

For this experiment the dataset of 166 test vehicles from January 1, 2019, till December 31, 2022, is used. The objective here is to forecast mileage (distance driven) for a specific period using historical data. Since monthly reporting of mileage is required, the dataset is aggregated monthly. Therefore, to develop a time series model, the average daily mileage per month of all vehicles is calculated as given in Table 5-8 below.

Table 5-8 Dataset for timeseries forecasting

Month	Average Daily Mileage (KM)
1/31/2019	120.80
2/28/2019	161.21
3/31/2019	136.87
4/30/2019	138.29
5/31/2019	140.18
…	…
12/31/2022	269.08

Approaches for implementation

For this implementation, the following two different approaches are investigated.

A. Generic time series model
B. Individual time series models

A. Generic Model

Initially, one generic time series model is developed using the historical data given in Table 5-8 above. In other words, one time series model for all vehicles was fitted. The steps followed to develop the forecast model are described as follows.

5.3.2.1 Data exploration

The data is checked for completeness and any anomalous behaviour and it is found to be clean and readily good for usage. The aggregate data is then plotted as shown in Figure 5-18 below.

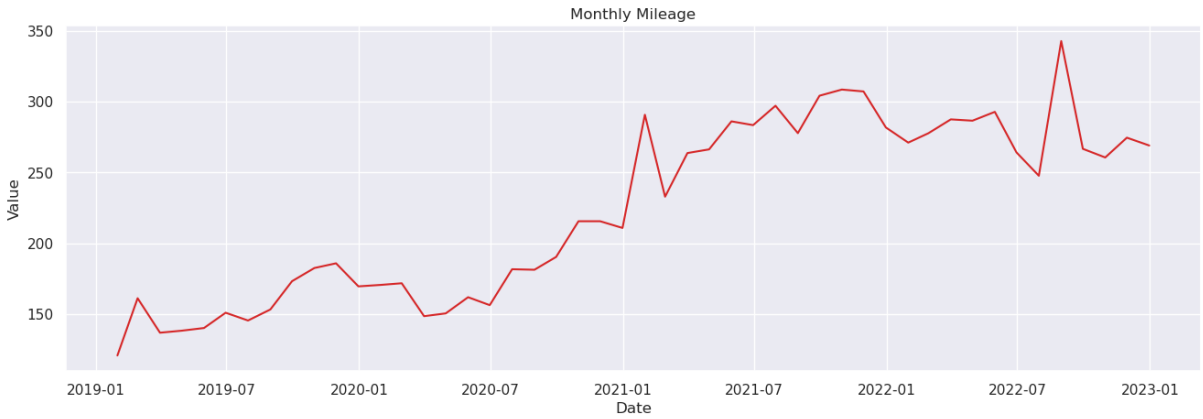


Figure 5-18 Plot of the original time series data

5.3.2.2 Pre-processing

To achieve effectiveness in time series modelling, it is important that the data exhibits stationarity, which refers to the removal of trends, seasonality, and other relevant factors. The breakdown presented in Figure 5-19 demonstrates the presence of both trend and seasonality aspects in the data.

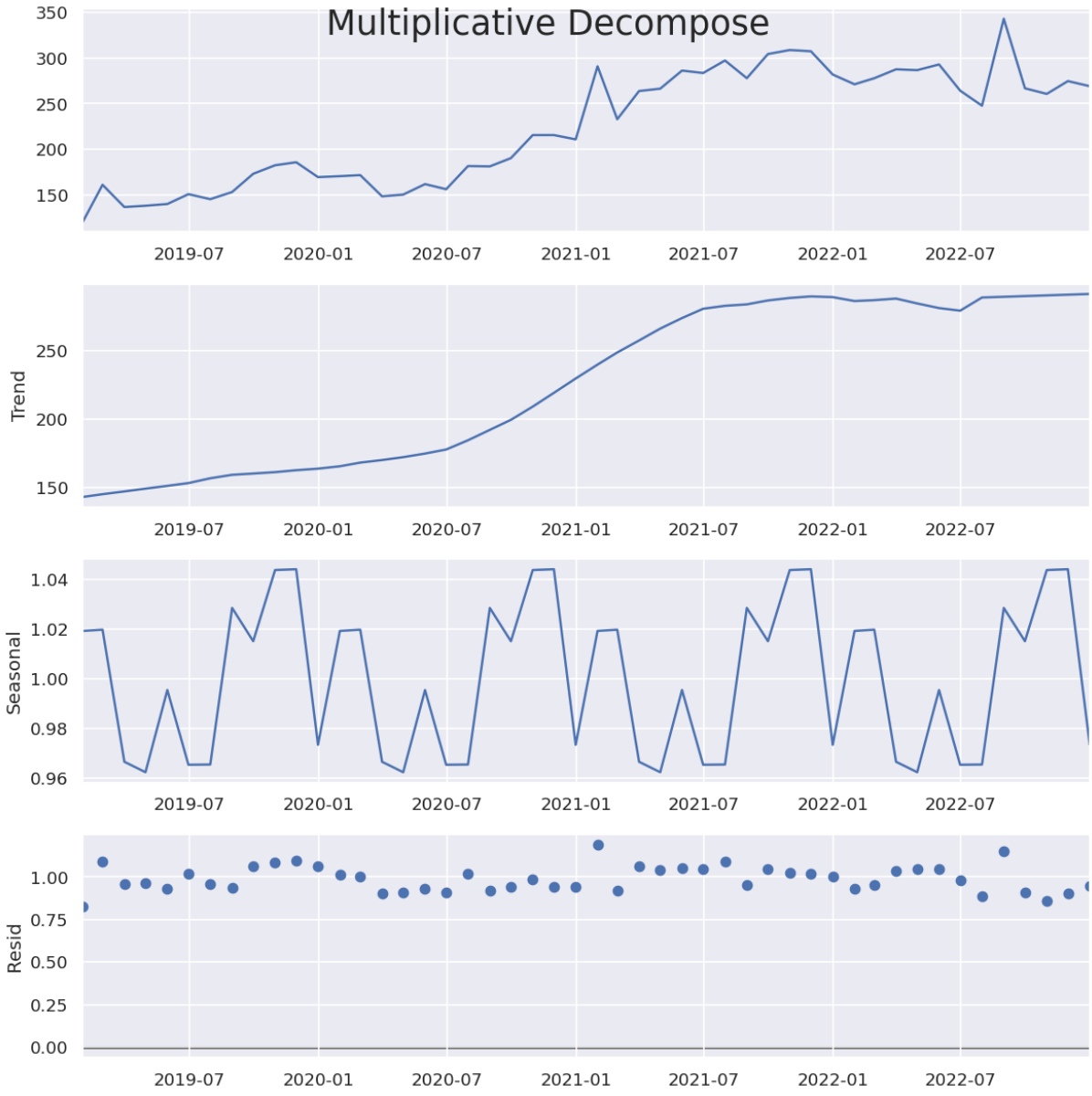


Figure 5-19 the time series data decomposition plot

Furthermore, the results of statistical stationary tests indicate that the data exhibits non-stationarity. The Augmented Dickey Fuller (ADF) test is widely employed in the literature as a stationary test (Cerqueira, Torgo and Mozetič, 2020). Its null hypothesis states that the time series exhibits a unit root and is therefore non-stationary. If the p-value obtained from the ADF test is smaller than the predetermined significance level of 0.05, it is recommended to reject the null hypothesis. Another testing method, known as the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, is employed to inspect the presence of trend stationarity (Adhikari and Agrawal, 2013). The null hypothesis and the interpretation of the p-value is opposite to the approach taken by the ADF test. The outcomes of the Augmented ADF test and the KPSS test on the time series dataset are presented below in Table 5-9.

Table 5-9 ADF and KPSS test results

Test	Statistic	p-value	Critical Values	Critical Values	Critical Values	Critical Values
ADF	-1.196390196236740 3	0.67514662 74345228	1%, -3.584828853 223594	5%, -2.928299149519890 7	10%, -2.60234382 71604937	
KPSS	0.489271	0.044083	10%, 0.347	5%, 0.463	2.5%, 0.574	1%, 0.739

As shown in Table 5-9, the p-value obtained from the ADF test is 0.68, which is above the commonly accepted significance level of 0.05. Consequently, the null hypothesis should be rejected, indicating that the time series data is non-stationary. In addition, it can be observed that the p-value for the KPSS test is 0.04, which is below the significance level of 0.05. Consequently, the null hypothesis should not be rejected, indicating that the series does not exhibit trend stationarity.

5.3.2.3 De-trending

According to the previous ADF and KPSS tests conducted, the results indicate that it is important to apply de-trending techniques to achieve stationarity of the data. Differencing is considered to be one of the simplest techniques for detrending (Box *et al.*, 2015). The diagram presented in Figure 5-20 illustrates the outcome of the differencing technique applied to the dataset.

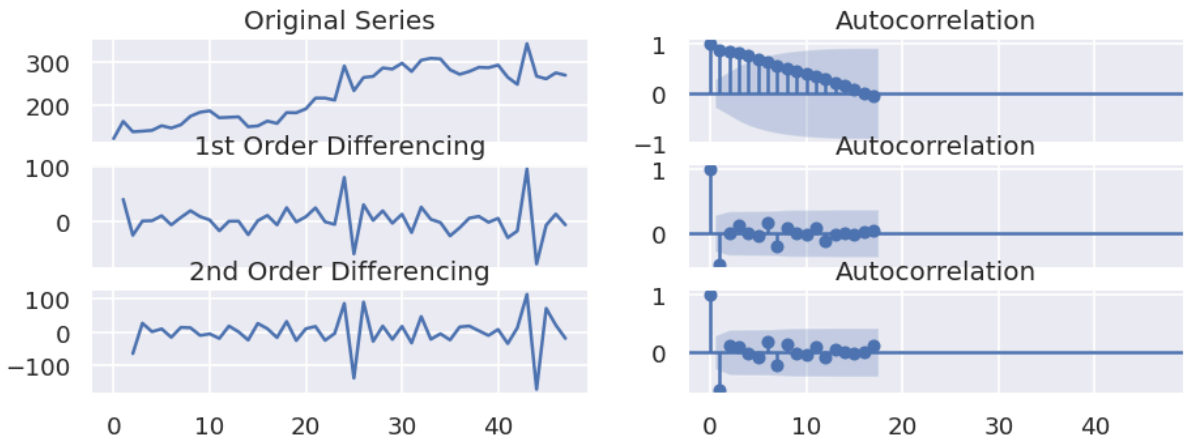


Figure 5-20 Detrending of the time series data

The result indicates that after applying one order of differencing, the series exhibits a noticeable stationarity pattern. Therefore, while it may not be ideal, one order of differencing can be applied.

Further investigation of the PACF plot shows that PACF lag 1 is below the significance level, denoted by the blue line as shown in Figure 5-21. Therefore, it is reasonable to set p to 1.

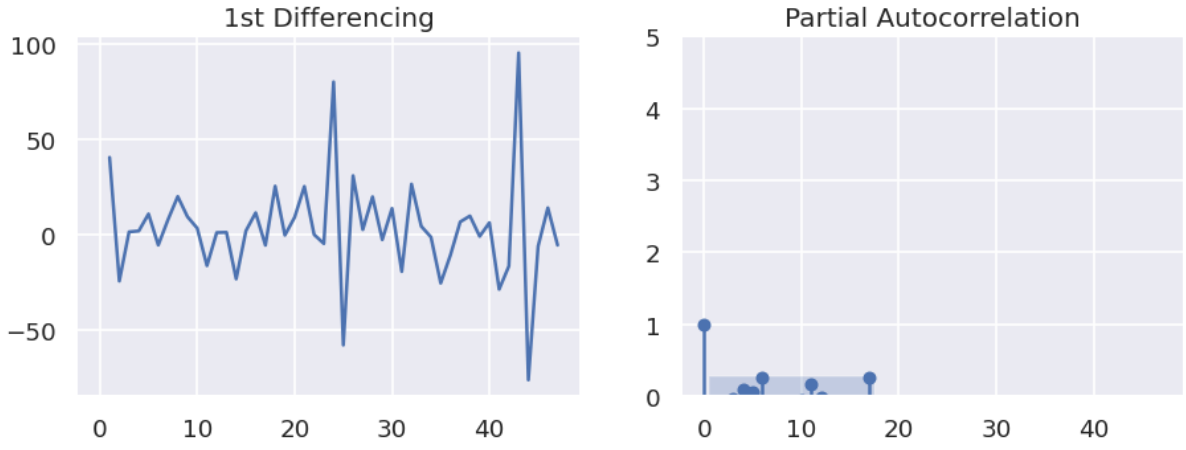


Figure 5-21 Partial Autocorrelation (PACF) plot of the time series dataset

Similarly, the ACF plot, depicted in Figure 5-22, indicates that all lags are positioned below the significance level. Consequently, it is reasonable to set the value of q to 1.

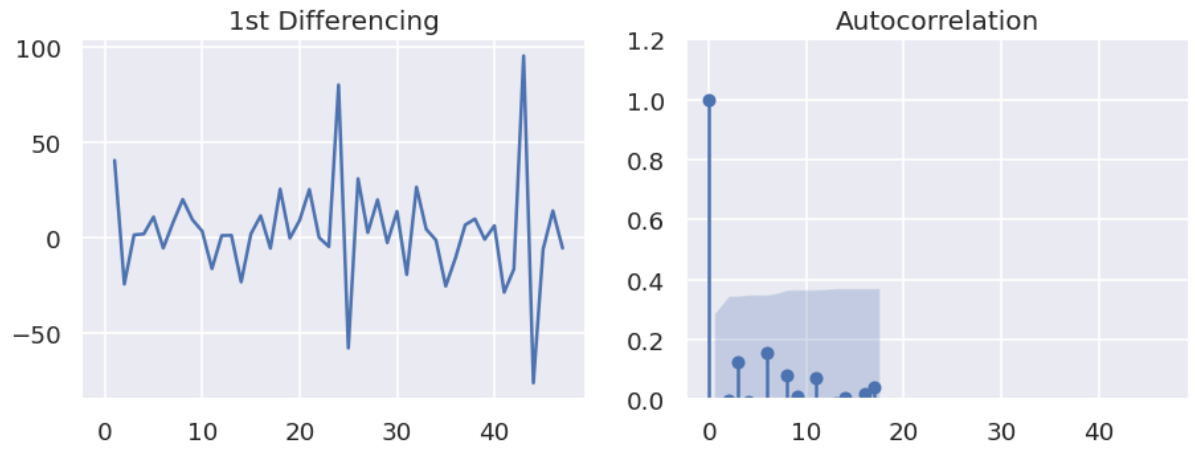


Figure 5-22 Autocorrelation (ACF) plot of the time series dataset

In addition, the lag plots depicted in Figure 5-23 demonstrate a decrease in correlation starting from lag 2. Hence, this can be taken as a further confirmation that the values of p, d, and q can be set to 1.

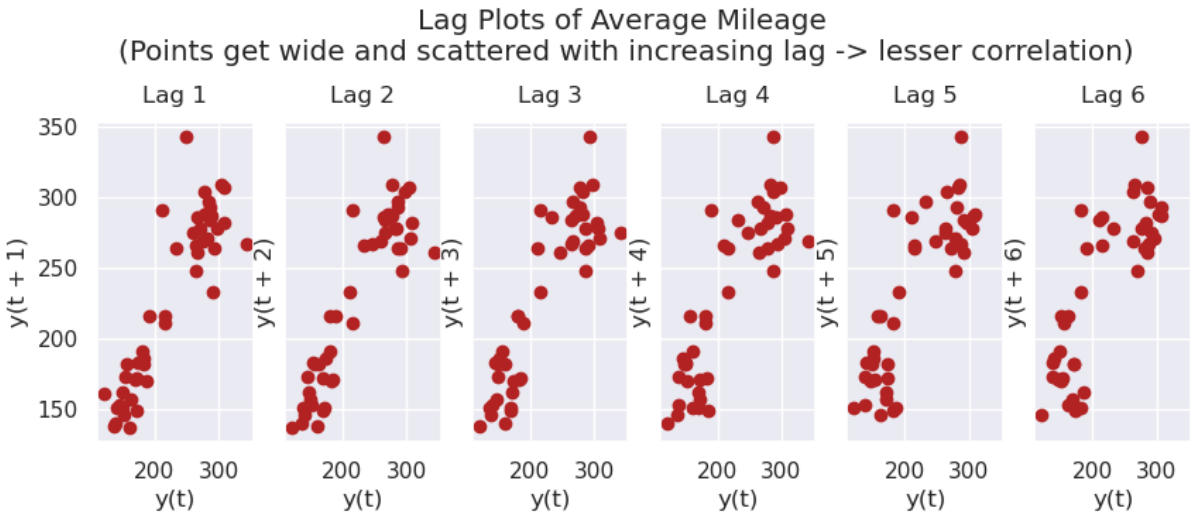


Figure 5-23 Lag plots of the time series dataset

5.3.2.4 Model Building

With the parameters set above, i.e., p =1, d=1 and q =1, ARIMA (p, d, q) model, is executed. Nevertheless, the outcome is unsatisfactory because the p-values associated with the predictors above the threshold of 0.05, as indicated in Table 5-10 below.

.

Table 5-10 Initial ARIMA model outcome

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.9581	1.662	1.779	0.075	-0.3	6.216
ar.L1.D.AVG_DIST	-0.2374	0.209	-1.137	0.256	-0.647	0.172
ma.L1.D.AVG_DIST	-0.3893	0.177	-2.204	0.028	-0.735	-0.043

Furthermore, it can be clearly noted that the residual plot, as depicted in Figure 5-24 below, exhibits a deviation from normal distribution, specifically, the right-side tail of the plot is skewed.

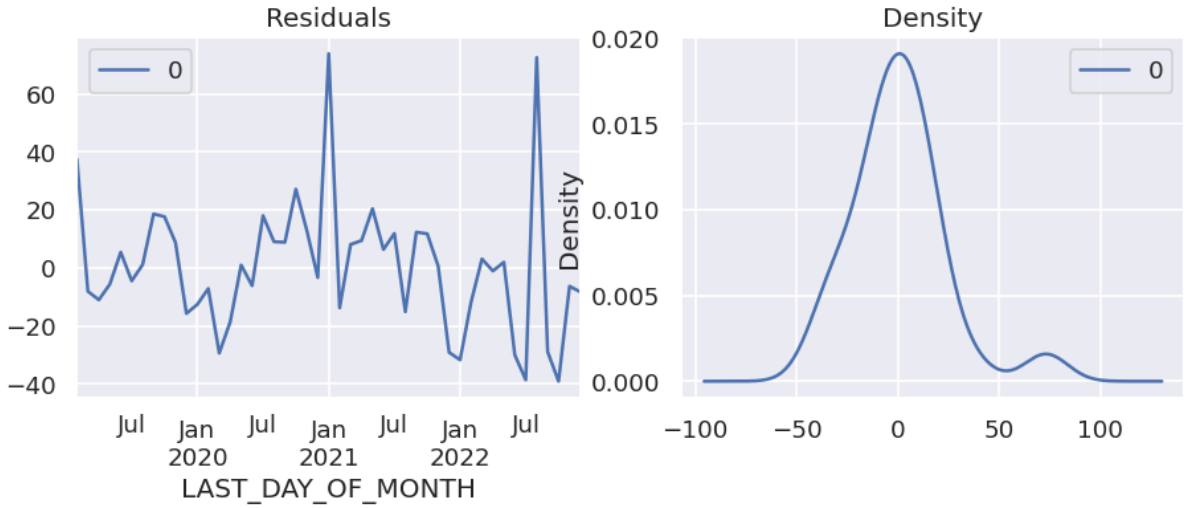


Figure 5-24 Residual plot of the initial ARIMA Model

To improve the effectiveness of the forecasting model, the addition of a seasonal component is implemented, and the auto-arima method is employed to identify the optimal values of parameters. Hence, the ARIMA (p, d, q) model is effectively modified to the SARIMA (p, d, q) (P, D, Q) model, where:

- P refers to the seasonal Auto Regressive order.
- D refers to the seasonal difference order and.
- Q refers to the seasonal Moving Average order.

Based on the results obtained, it is evident that the ARIMA (2,2,0) (0,1,1) model results in the most optimal outcome, as measured by its lowest AIC value which is described in section 3.7.2.2. The forecasted result plot is presented in Figure 5-25.

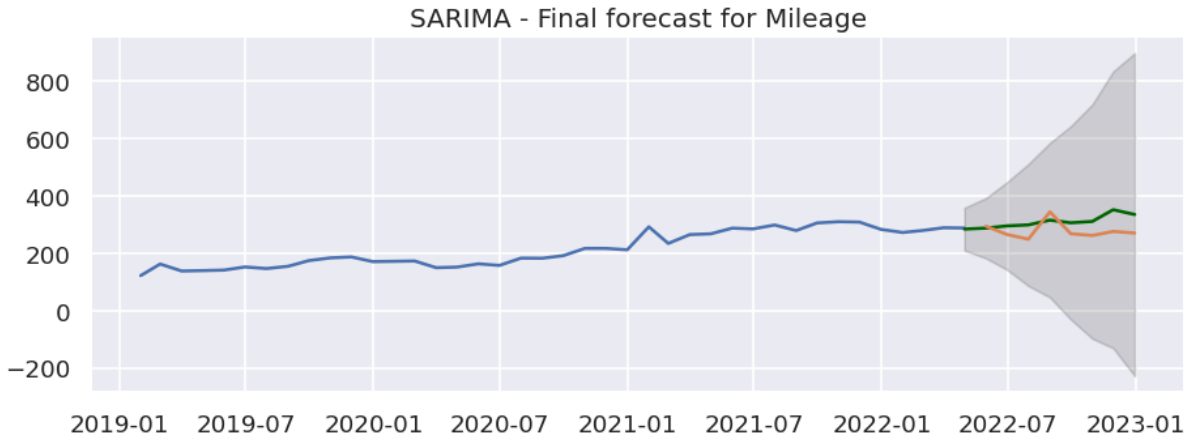


Figure 5-25 Forecast using the SARIMA model

While the SARIMA model created results in a good forecast for the full population, its application to individual vehicles shows that there is significant discrepancy for some vehicles. The observed discrepancy can be attributed to differences in vehicle types and their respective usage patterns.

One potential approach for addressing this issue involves the addition of additional variables, referred to as exogenous variables, into the time series model. This modification transforms the time series model from a SARIMA model to a SARIMAX model in which X represents the presence of exogenous variable, as discussed by Ensafi *et al.* (2022). In this research, however, incorporating the different potential exogenous variables was not possible due to the following limitations.

1. Most of the variables which might have an impact on the mileage driven by vehicles, such as the sector in which the vehicle is used, were not available or accessible for this study.
2. Some of the available variables, for example client information, location information and so on are privacy sensitive.

An alternative approach involves the development of an individual time series model that does not rely on the inaccessible or unavailable features, the implementation of which is described as follows.

B. Individual models

The rationale behind this approach is that individual models may result in better performance as they more accurately represent the behaviour that is unique to a given vehicle. The methodology employed in this study draws inspiration from the successful outcomes observed in individual demand forecasting, as demonstrated by the research conducted by Widiarta,

Viswanathan and Piplani (2008). However, constructing an individual time series model for many vehicles is a significant challenge. The implementation of automation can serve as a practical option. Hence, the auto-arima methodology, in combination with looping algorithm, is used in this study to construct a time series model for each different non-communicating vehicle.

By iterating over all the vehicles in scope, auto-arima chooses the best model automatically and saves the result as a pickle file for further prediction. The proposed algorithm is presented as follows.

Algorithm 5.2: Individual time series model training using auto-arima

***Inputs**: time series training dataset*

***Output**: arima model for all vehicles in the training dataset*

Initialization:

unique_vehicle_ids[] ← *unique identifiers of all vehicles in the training dataset*

Candidate_models[] ← *empty //to find all auto-arima models*

Best_arima_model ← *empty //to store the best model*

For each record (i) in unique_vehicle_ids[]:

Training_data_subset ← *taining _dataset[i] //filter for single vehicle*

Sort(Training_data_subset) //sort the training data by month

Candidate_models[] ←*Auto_arima() //call auto-arima algorithm and store results*

Best_arima_model ← *Candidate_models[] with minimum AIC*

Save (Best_arima_model) //save best model to pickle file

End

The algorithm presented is designed to autonomously train, optimize, and determine the most suitable time series model for each individual vehicle within the given dataset. The last step involves storing the optimal model for each vehicle as a pickle file, enabling its application in future predictions on new datasets. Next, to make predictions for an unknown mileage of vehicles, the previously saved pickle file is loaded using the algorithm presented below, with the required vehicle identifier or set of vehicle identifiers passed as a parameter.

Algorithm 5.3: Forecasting mileage by applying individual auto-arima model

Inputs: *Vehicle identifier or set of vehicle identifiers to forecast*

Output: *Forecasted mileage for the vehicle or vehicles in the input data*

Initialization:

unique_vehicle_ids ← *all unique identifiers of vehicles to forecast mileage*

n_periods ← *integer value of the number of periods to forecast*

forecast_output[] ← *forecasted values, initially empty*

For each record (i) in unique_vehicle_ids:

Load (Best_arima_model) //Load best model saved for the vehicle

Individual_forecast ← *Best_arima_model.predict(n_periods)*

Forecast_output.append(Individual_forecast)

End

save(Forecast_output)

5.3.3 Evaluation of Scenario II

Historical data is used to evaluate the developed auto-arima forecasting model. Using the historical data of 166 vehicles mentioned earlier in section 5.3.2, auto-arima models stored are used to forecast for 6 periods and part of the forecasting result is given in Table 5-11 below.

Table 5-11 Prediction result using individual time series model

Vehicleid	Period	Individual model prediction	Actual value
1	1	4.26	4.53
1	2	4.58	4.15
1	3	8.14	7.76
1	4	7.08	8.93
1	5	2.82	1.93
1	6	6.29	7.80
2	1	262.82	245.37
2	2	297.07	310.21
2	3	197.92	200.64
2	4	317.18	298.81
2	5	283.66	303.44
2	6	299.90	276.68
…	…	…	…

The evaluation employed the RMSE metric as it is widely applied to compare the forecasted value with the actual value by many research literatures including (Cerqueira, Torgo and Mozetič, 2020). The different evaluation techniques for time series forecasting including RMSE and their description is provided in section 3.7.2.2. To have a reasonable evaluation, the moving average is used as a reference base line. The findings, as presented in Table 5-12 partially for individual vehicles together with average of all vehicles, indicate that the individual models show a better performance compared to the moving average in terms of the selected performance metric, i.e., RMSE.

Table 5-12 Comparison of Individual time series models with moving average as a baseline

Vehicleid	RMSE Moving average	RMSE Individual model
1	1.07214	1.07214
2	21.64738	17.09012
3	91.35362	45.35472
4	37.45362	31.10004
5	32.84995	26.35728
	…	…
Average	92.57412	41.54637

5.4 Scenario III: Detecting Inaccurate Values (Accuracy Data Quality dimension)

Chapter 2 has presented an explanation of the different components of the CV ecosystem. Also, in the same chapter, it is indicated that the introduction of data inaccuracies might occur due to a failure or malfunction in any of these components, including sensor drift and processing errors, among others. This section describes the application of ML and statistical quality control chart for enhancing the evaluation of accuracy metrics in the DQ dimension for CV data. The approach starts by selecting a specific data element for analysis and is described in this section.

5.4.1 Problem description of inaccurate data

From previous chapters in this thesis, it is understood that: CV data is collected using multiple sensors and passes through different points. Inaccurate values may be reported to end users due to sensor issues or other technical difficulties. It also became clear that: assessing CV data is complex. Invalid data, which may be evaluated using metrics established in the validity DQ dimension, can be easily examined by comparing the provided values with the accepted range of values or defined rules. However, metrics associated with the accuracy dimension of DQ

pose challenges when assessing using classical DQ assessment frameworks. Establishing standards for assessing data accuracy is a possible option, however, the precise determination of an accurate value poses considerable challenges. The only way of ascertaining the accuracy of data is by its comparison with a certified source that is entirely devoid of any degree of uncertainty. For instance, when two different records or places contain varying values for the same real-world fact or event, it is logical to conclude that at least one of the values is inaccurate. However, without an established factual reference, it is difficult to determine which one is accurate or what the accurate value is.

Also from the literature, it is understood that: some researchers employ the metrics of validity DQ dimension to assess the accuracy DQ dimension. Nevertheless, it is important to note that this approach has certain limitations as valid data, which adheres to predefined rules or range of values, does not automatically guarantee its accuracy. As an illustration, a temperature measurement of 30 degree Celsius obtained from a temperature sensor by a driving CV can be considered a valid or plausible value. However, its accuracy cannot be assured unless it is cross-validated or compared against a confirmed reference. For this particular case, in the absence of a certified reference source, a complete assessment requires the consideration of various aspects, including the region, season, and time of day at which the measurement was conducted, to arrive at a reasonable conclusion. An alternative method that has gained increased attention for evaluating the accuracy of DQ is the use of outlier identification techniques. Nevertheless, this methodology identifies anomalies that are relatively easier to identify, which is inadequate for assessing accuracy.

Therefore, this section explores the use of ML and statistical techniques to assess the accuracy of a given data element. The study is conducted with real-life CV data from real-world scenarios. Furthermore, to ensure the ability to reproduce the results, a publicly available dataset is also employed. The fuel consumption data element is chosen and employed as the dependent variable or target variable in both datasets.

5.4.2 Proposed solution to detect inaccurate data

As demonstrated in section 4.3.2, it was found to be difficult to assess accuracy DQ dimension using the classical DQ assessment method employed. Therefore, it is necessary to devise another mechanism for proper accuracy DQ evaluation. As discussed earlier, to assess accuracy DQ, most experts propose utilizing alternative measures intended for validity DQ dimension

which is not sufficient because validity does not always imply accuracy, as there is no measurement defined for accuracy DQ dimension in classical DQ assessment methodologies in the absence of reference data. On the other hand, some recent research employs the same approach used for outlier detection, but they concentrate on relatively easier outliers. This study tries to assess accuracy with better precision.

In this research advanced methods, specifically ML and statistical methods are proposed and described as follows. This proposal takes inspiration from the different methods discussed in Chapter 2. Specifically, the works of Dai, Yoshigoe and Parsley (2018), where they combined deep learning and statistical methods to detect outliers and the method developed by yahoo known as Extensible Generic Anomaly Detection System (EGADS) where two modules are combined namely a time series module generating an expected value at a certain point in time and a second module checking the actual value with the expected value and generates the errors (Laptev, Amizadeh and Flint, 2015).

For this experiment, fuel consumption data element is selected. The proposed solution consists of two main modules: 1. ML module and 2. Statistical quality control module. The architecture of the proposed solution is given in Figure 5-26 below.

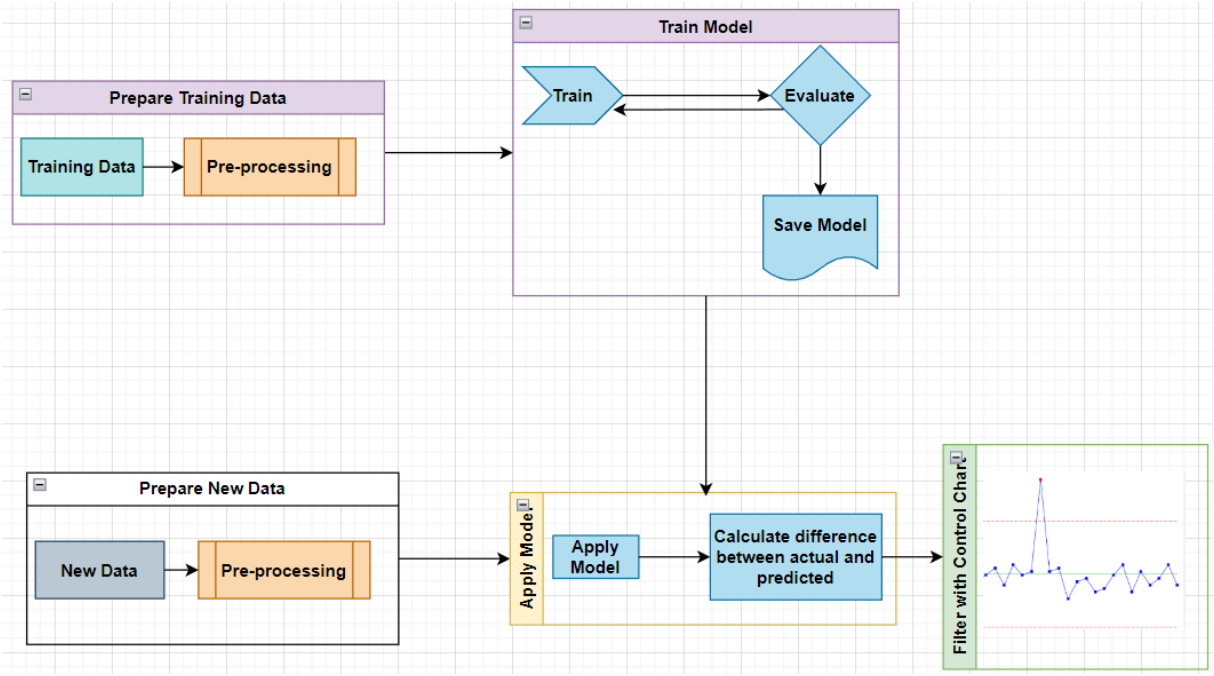


Figure 5-26 Adopted architecture to assess accuracy data quality

1. ML module:

The purpose of this module is to use historical data and apply an ML algorithm to obtain a predicted value for the selected data element, i.e., fuel consumption in this experiment.

2. Statistical quality control module:

This module is used to apply statistical quality control on top of the difference between the actual values and the predicted values obtained from the ML module. This is motivated by the statistical control process, the basis of which is the central limit theorem (Benneyan, 1998). This method is used to determine if the difference between predicted value and actual value is within normal variation or not.

1. Application on real-life Connected vehicle data

A. Dataset

As described earlier, real-life CV trajectory or trip data is used in this research. And for this scenario, fuel consumption data element is selected. The complete dataset description is provided in section 3.4; however, it is important to re-iterate that the data set contains attributes about the vehicle, the driver behaviour and the trip characteristics including the weight the vehicle carried, the trip distance covered and so on. It includes 23 numerical features and 2 categorical features.

B. Pre-processing

To apply the data for modelling, several pre-processing is performed. Specifically, the following steps are carried out.

- Outlier treatment: -

  From the original data, it was observed that several outliers exist across the different data elements. To understand the statistical behaviours of each data element; different plots such as Density plot, Scatterplot versus the target variable, i.e., fuel consumption and Quantile-Quantile plots are generated for each numeric variable. An example of those plots before and after outlier treatment for gross combination weight data element is presented in Figure 5-27 and Figure 5-28 respectively.

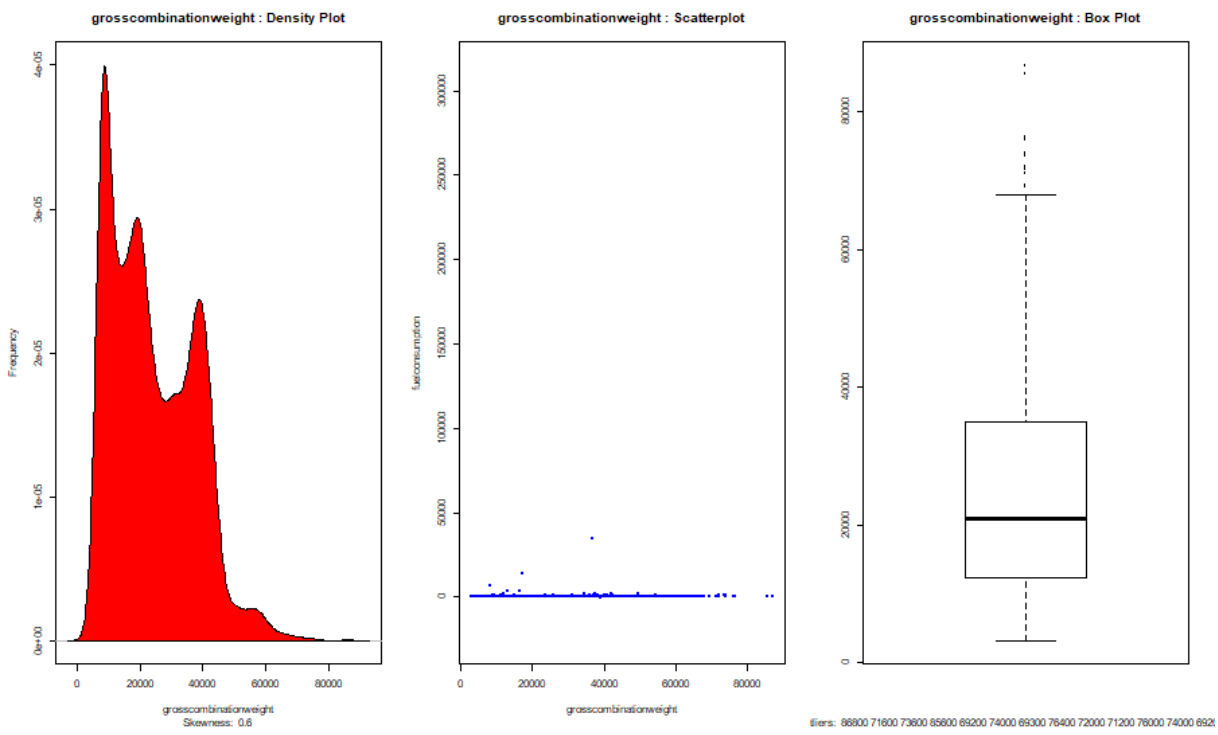


Figure 5-27 Gross combination weight before outlier treatment

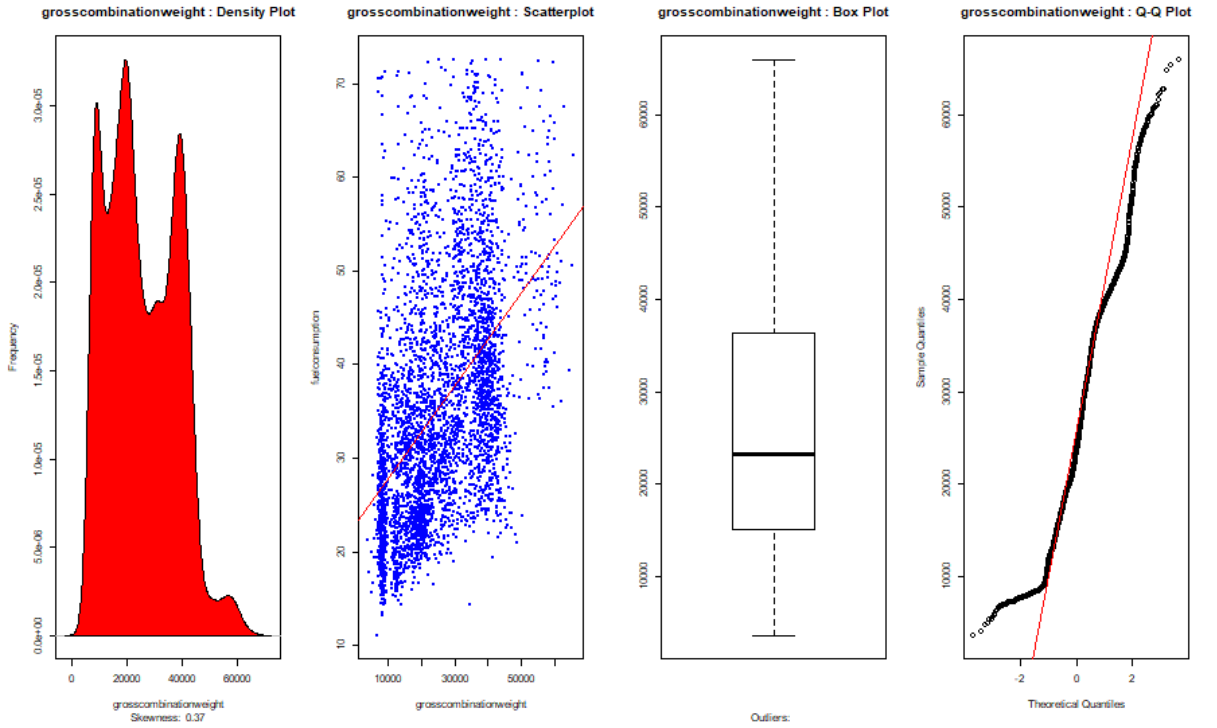


Figure 5-28 Gross combination after outlier treatment

In addition, studying the metadata description of the dataset and statistical characteristics of each data element, a number of filters is applied and presented in Table 5-13 as follows.

Table 5-13 Filters applied on data elements before training

Data Element	Min	Max	Remark
AVG_SPEED	0	125 km/hr	Based on traffic regulation rules.
DISTANCEDONE	0	500	According to European union regulation, 4 hours is the maximum a driver is allowed to drive continuously. With maximum allowed speed of 90 km/hr, the maximum distance of a trip will be 360 km.
BRAKE_DURATION	0	1000	
TRIPDURATION	0	20000	
HARSHBRAKE_DURATION	0	1000	
IDLING_DURATION	0	50	
GPS_ELEVATIONGAIN	0	2000	
GPS_ELEVATIONLOSS	0	2000	
PTO_COUNT	0	10	
ACCELERATION_DURATION	0	4000	
MAXTHROTTLEPADDLE_DURATION	0	4000	
DPABRAKINGSCORE_SUM	0	3000	
DPAANTICIPATIONEVENT_COUNT	0	80	
CRUISECONTROL_DISTANCE	0	400000	
GROSSCOMBINATIONWEIGHT	0	60	
FUEL_INDEX	12	55	

- Mean and mode imputation: -

  Missing values were inputted with mean values, particularly the gross combination weight data element exhibited missing values, and this method was applied to replace the missing values. For categorical variables, the mode value is applied to replace the missing data.

- Normalization with z-score: -

  The following formula was used to normalize the data to minimize the impact of scaling variation.

$$x' = {(x - \mu)}/{\sigma} \qquad 5.2$$

where x' is the new value of x, μ is the mean and σ is the standard deviation.

C. Feature Selection

The number of potential features available in the dataset and presented in Table 3-2 is many. Therefore, feature selection is necessary to choose the most relevant features while maintaining the model performance. In this section, SHapley Additive exPlanations (SHAP) is used by combining feature selection with parameter tuning (Lombardi *et al.*, 2022). The integration of the tuning process with the selection of optimal features is a potential requirement for any ranking-based selection system. The process of ranking selection involves iteratively eliminating features that are deemed less significant, while simultaneously retraining the model, until a state of convergence is achieved.

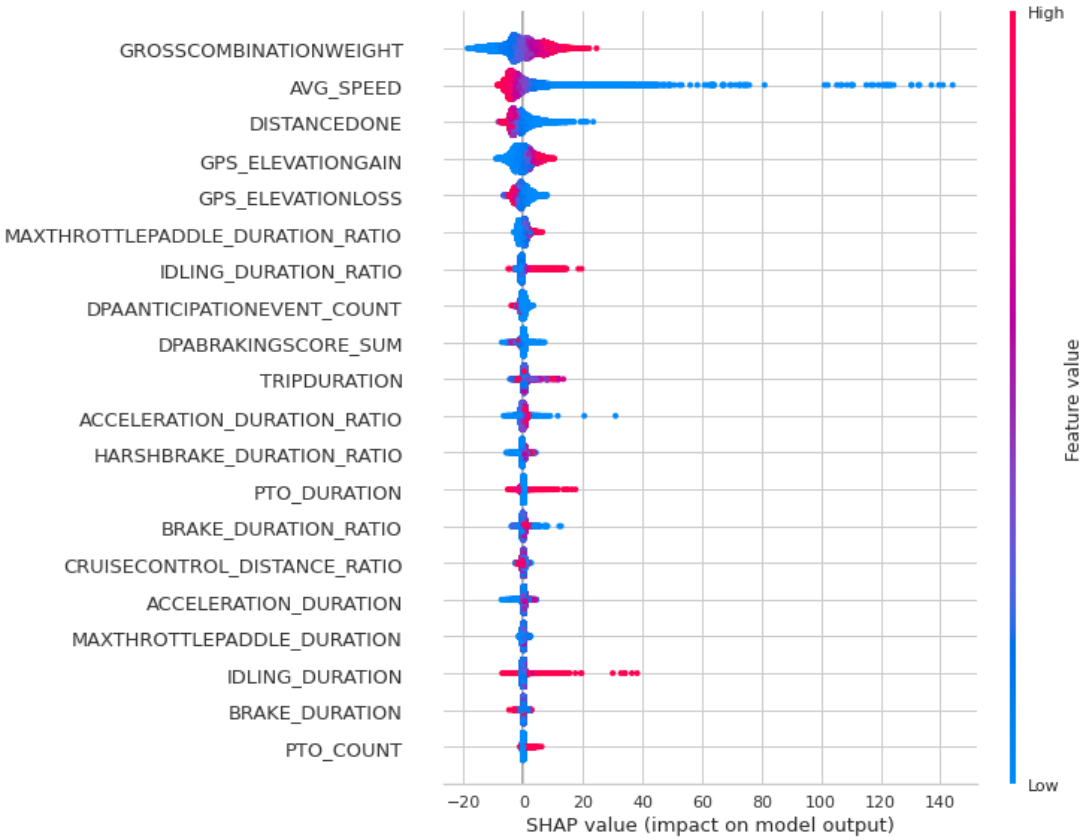


Figure 5-29 Feature rankings using SHAP

The diagram given above in Figure 5-29 shows the list of features ranked on their importance. The gross combination weight, the average speed and distance travelled appeared to be more important features.

D. Multicollinearity

Multicollinearity occurs when two or more features are strongly correlated, which may cause regression models to underperform. When such a situation occurs where two features are highly correlated, then the better feature of the two should be kept by removing the other. Variance inflation factor (VIF) is a common method to identify multicollinearity. In other words, VIF is used to measure the degree of severity of multicollinearity in regression analysis. In general, a VIF value of 4 and above indicates that collinearity might exist. In this exercise, all features with VIF equal to 4 or above are removed and the resulting feature set for the final prediction contains data elements: {GROSSCOMBINATIONWEIGHT, AVG_SPEED, DISTANCEDONE, CRUISECONTROL_DISTANCE_RATIO, MAXTHROTTLEPADDLE_DURATION, TRIPDURATION, IDLING_DURATION, DPAANTICIPATIONEVENT_COUNT, BRAKE_DURATION_RATIO, GPS_ELEVATIONGAIN}.

Besides, when there are related features and one of them is a ratio with respect to another feature, the ratio is selected. For example, in between BRAKE_DURATION and BRAKE_DURATION_RATIO, the latter is selected.

E. Model Building

To have a baseline, linear regression model was fitted by using the features selected, resulting in R^2 of 0.67. Since linear regression is easier to interpret, this activity was helpful to understand the influence of each variable by inspecting the coefficients and the different diagnostic plots. Having this as a baseline, multiple regression algorithms including the following are trained.

- Random Forest Regressor
- Gradient Boosting Regressor
- Elastic Net
- Lasso Regression and so on

The implementation of these algorithms is available as a package in a single python library called pycaret as outlined in Table 3-4. The output of the execution of the different algorithms is given in Table 5-14 below.

Table 5-14 Training output of different algorithms for fuel consumption prediction

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
CatBoost Regressor	2.67	14.14	3.76	0.86	0.11	0.09	14.28
Extreme Gradient Boosting	2.77	15.11	3.89	0.85	0.11	0.09	0.999
Light Gradient Boosting Machine	2.83	15.51	3.94	0.84	0.11	0.09	0.995
Random Forest Regressor	2.94	16.90	4.11	0.83	0.12	0.09	78.95
Extra Trees Regressor	2.96	17.10	4.14	0.82	0.12	0.10	31.77
Gradient Boosting Regressor	3.14	18.77	4.33	0.81	0.12	0.10	27.13
K Neighbors Regressor	3.54	24.45	4.94	0.75	0.14	0.11	8.771
Ridge Regression	4.16	31.48	5.61	0.68	0.17	0.14	0.136
Linear Regression	4.16	31.48	5.61	0.68	0.17	0.14	0.979
Bayesian Ridge	4.16	31.48	5.61	0.68	0.17	0.14	0.182
Least Angle Regression	4.16	31.48	5.61	0.68	0.17	0.14	0.135
Huber Regressor	4.09	32.19	5.67	0.67	0.16	0.13	0.655
Decision Tree Regressor	4.29	35.61	5.97	0.63	0.17	0.14	1.641
Lasso Regression	4.58	39.21	6.26	0.60	0.18	0.15	0.144
Lasso Least Angle Regression	4.58	39.21	6.26	0.60	0.18	0.15	0.133
Elastic Net	4.98	46.15	6.79	0.53	0.20	0.17	0.147
Passive Aggressive Regressor	5.47	52.11	7.20	0.47	0.22	0.19	0.244

Based on the output given in Table 5-14, boosting based ensemble algorithms such as Light Gradient Boosting Machine (LightGBM), CatBoost Regressor (catboost) and Extreme Gradient Boosting (xgboost) resulted in the best result with the highest R^2, low RMSE and MAE and lowest RMSLE of 0.11 and MAPE of 0.09. Bagging based ensemble methods such as Random Forest also performed well. Considering other factors especially efficiency, LightGBM is selected. Further tuning is performed on LightGBM and the result of the different fold executions and summary is given in Table 5-15.

Table 5-15 Tuning result of LightGBM for fuel consumption prediction

	MAE	MSE	RMSE	R2	RMSLE	MAPE
Fold						
0	2.84	15.64	3.96	0.84	0.11	0.09
1	2.81	15.35	3.92	0.84	0.11	0.09
2	2.82	15.62	3.95	0.84	0.11	0.09
3	2.84	15.52	3.94	0.84	0.11	0.09
4	2.81	15.31	3.91	0.84	0.11	0.09
5	2.83	15.59	3.95	0.84	0.11	0.09
6	2.83	15.54	3.94	0.84	0.11	0.09
7	2.81	15.27	3.91	0.84	0.11	0.09
8	2.85	15.56	3.95	0.84	0.11	0.09
9	2.83	15.70	3.96	0.84	0.11	0.09
Mean	2.83	15.51	3.94	0.84	0.11	0.09
Std	0.01	0.14	0.02	0.00	0.00	0.00

The final model after tuning sets the following parameter values.

```
LGBMRegressor(bagging_fraction=0.6, bagging_freq=2,
              feature_fraction=0.4, min_child_samples=41,
              min_split_gain=0.9, n_estimators=260, n_jobs=-1,
              num_leaves=70, random_state=123, reg_alpha=2,
              reg_lambda=3)
```

F. **Model diagnosis**

Model diagnosis is a crucial step to understand that the model trained is appropriate for the data in consideration. There are several techniques used for model diagnosis. In this experiment, some of them are investigated. To begin with, the prediction error is given in Figure 5-30 below.

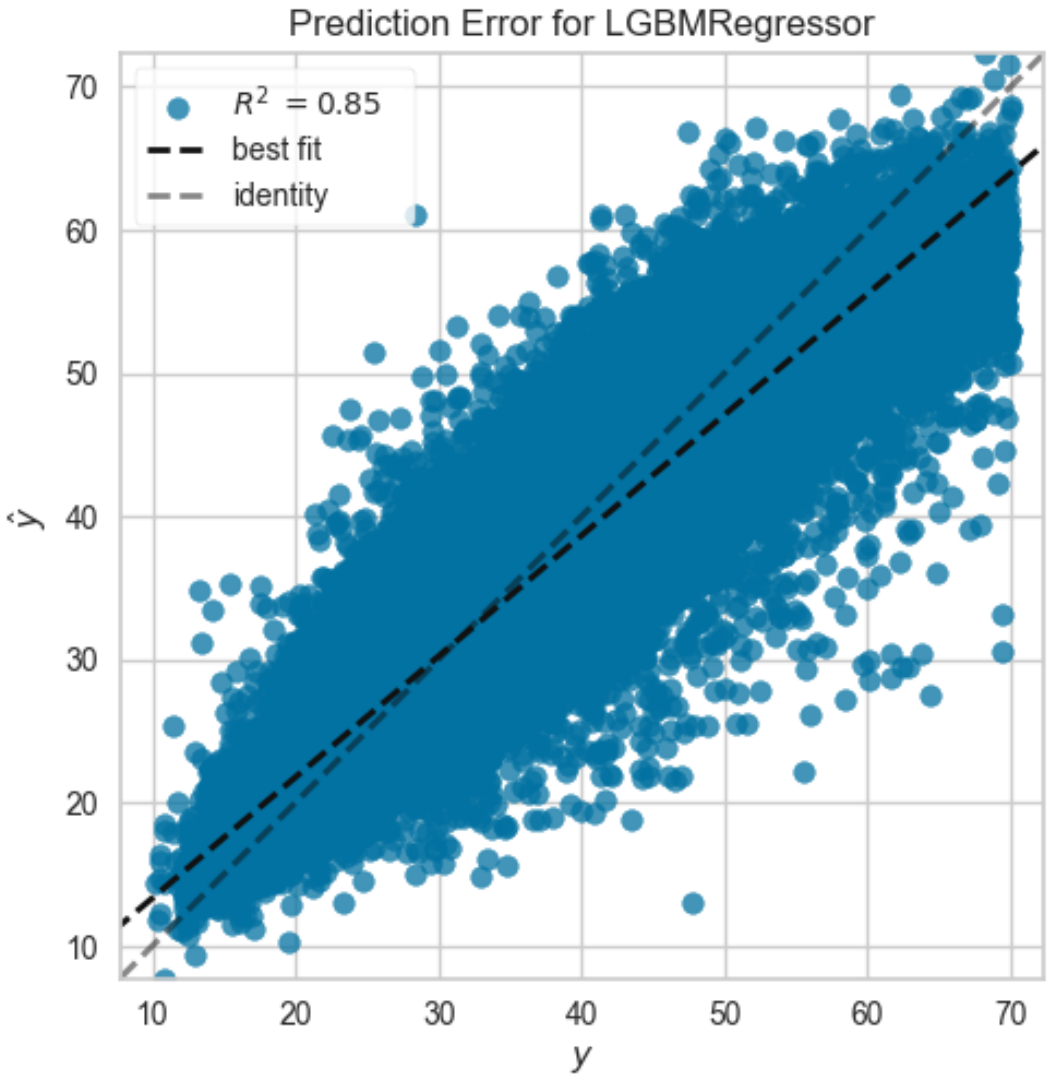


Figure 5-30 Prediction error of LGBM (best fit vs identity)

In Figure 530, the identity line represents the ideal scenario where all predicted results are the same as the corresponding actual values, and the best fit line shows the trend of the predicted results. As the two lines are closer to each other, the model has performed well. Next, the residual plot is investigated as presented in Figure 5-31.

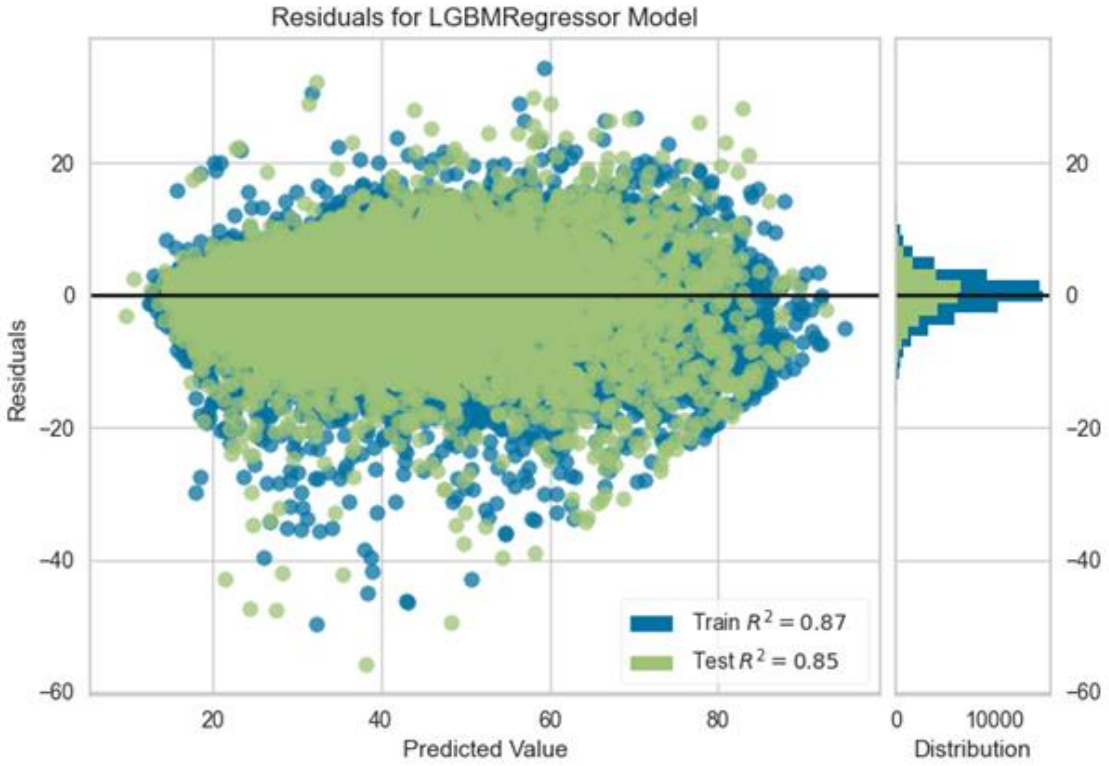


Figure 5-31 Residual plot of LGBM

The residual represents the difference between the actual and the predicted values. The residual plot therefore shows the deviation of the observed values from the best fit line given in Figure 5-30. Since the points are distributed randomly around the residual line, i.e., residual = 0 with no clear pattern, the model can be considered as a good fit of the data.

Finally, the tuned model is saved as a pickle file for future prediction on a new dataset.

G. Statistical Control chart

Using the trained model saved earlier as reference, a prediction is made on a new dataset. Then, a control chart is built using the difference between the actual values and the corresponding predicted values. For easy, interactive, and user-friendly presentation; a tableau visualization is created and given in Figure 5-32 below. A control chart helps to identify whether a process is in control or out of control (Benneyan, 1998). It consists of the following main components.

- Data points: these are points of measurements in the process to be monitored.
- Central Line (CL): this represents the average value of the points in the process.
- Upper Control Limit (UCL): this represents the threshold above which a data point is considered a potential issue.

- Lower Control Limit (LCL): this represents the threshold below which a data point is considered a potential issue.

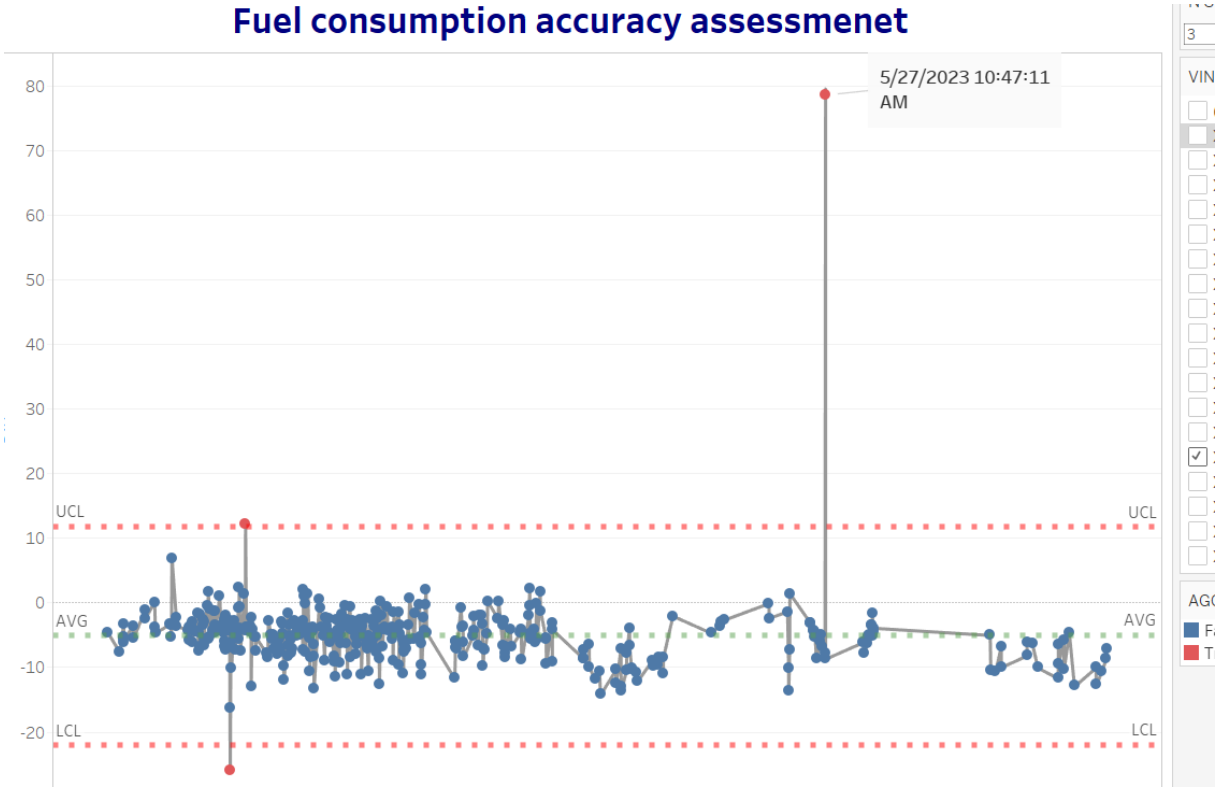


Figure 5-32 Control chart of fuel consumption actual value versus predicted value - Tableau visualization

As shown in Figure 5-32 above, the inaccurate values crossing the thresholds of UCL and LCL are marked as red and lies outside the control limits.

5.4.3 Evaluation of Scenario III

To evaluate this approach, data from 17 test vehicles is used. First, a new dataset for the 17 test vehicles is extracted from the connect system from January 1, 2023, till Aug 1, 2023. 4317 records/trips were found. Then the DAVIE system, which is described in section 3.4, is used to extract corresponding values. Only 552 values are found in DAVIE.

Then FUEL_INDEX values, which represent the fuel consumption in litre per 100 kilo meter as presented in Table 3-2, are compared and an indicator is added as 1 if the values match or 0 if a difference is observed. 43 values which do not match with DAVIE data are detected.

Then by referencing the model developed and saved earlier, the fuel consumption value is predicted, and the result is saved.

After getting the predicted values, the difference between the predicted and the actual values is calculated. On top of the result, quality control is applied and data points that lie beyond the control limits are marked as inaccurate as shown in Figure 5-33 below. Of the 43 inaccurate values extracted from Davie, 33 of them lied beyond the control limit, which amounts to 76%. More examples of the control chart showing detected values of the proposed method are presented in Figure 5-34, Figure 5-35, and Figure 5-36.

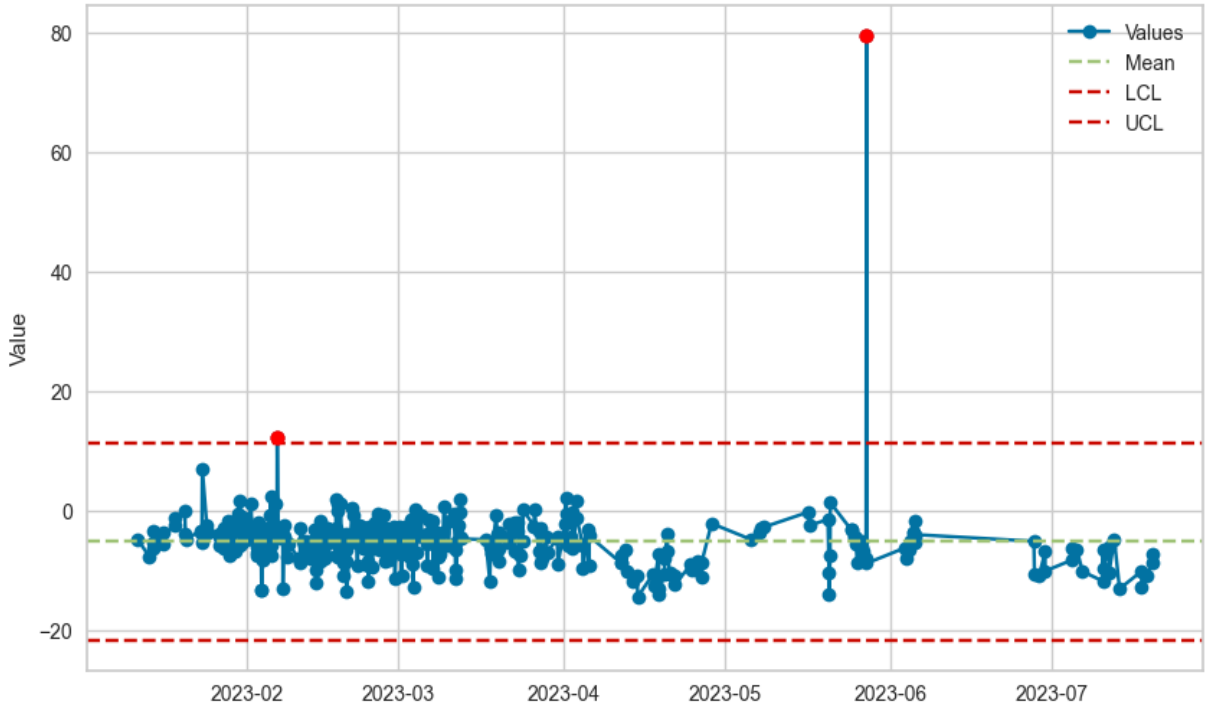


Figure 5-33 Control chart - Example 1

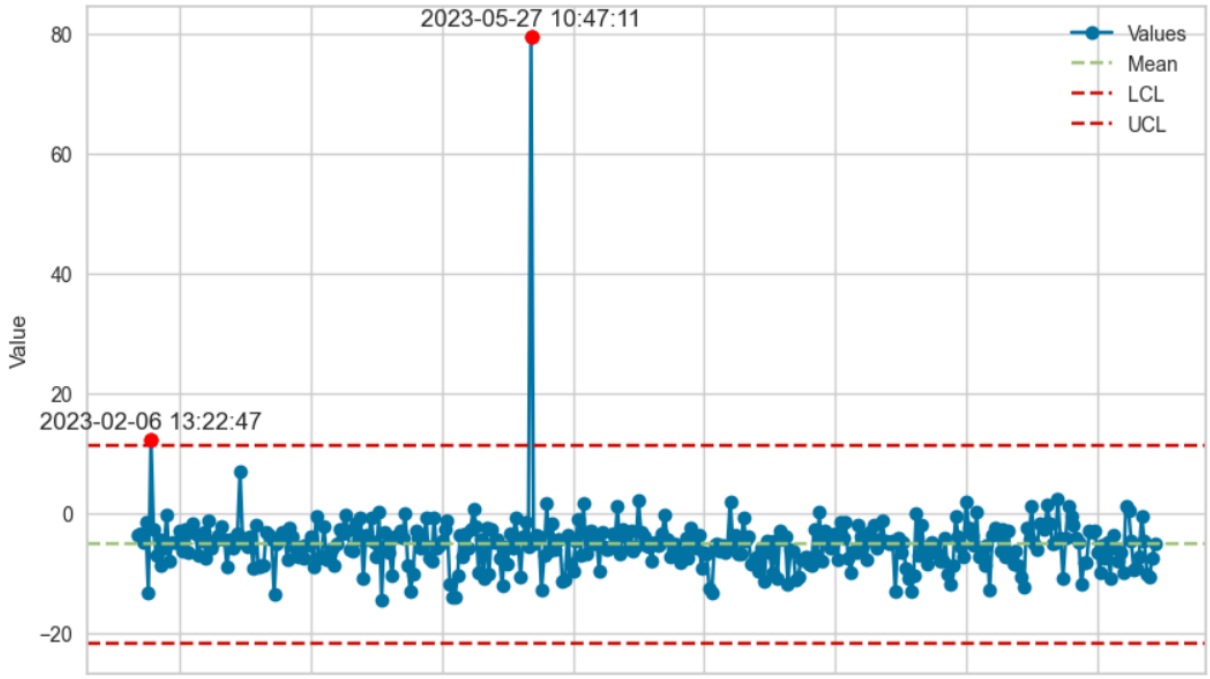


Figure 5-34 Control chart - Example 2

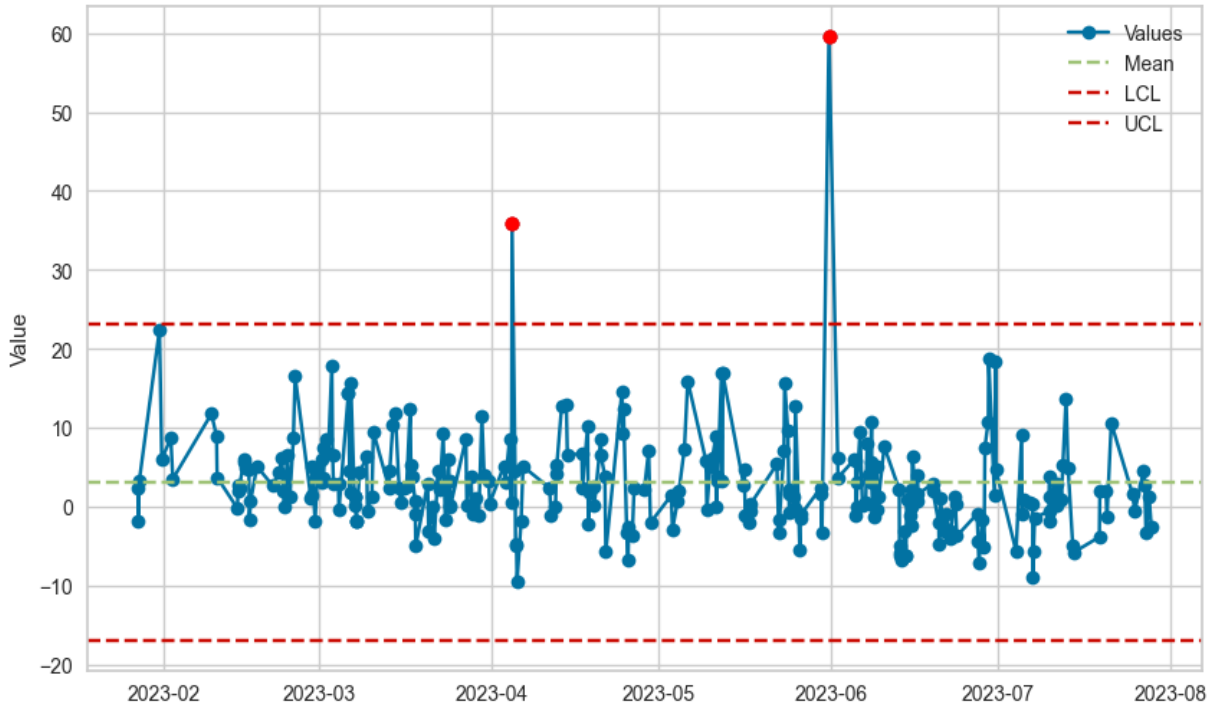


Figure 5-35 Control chart - Example 3

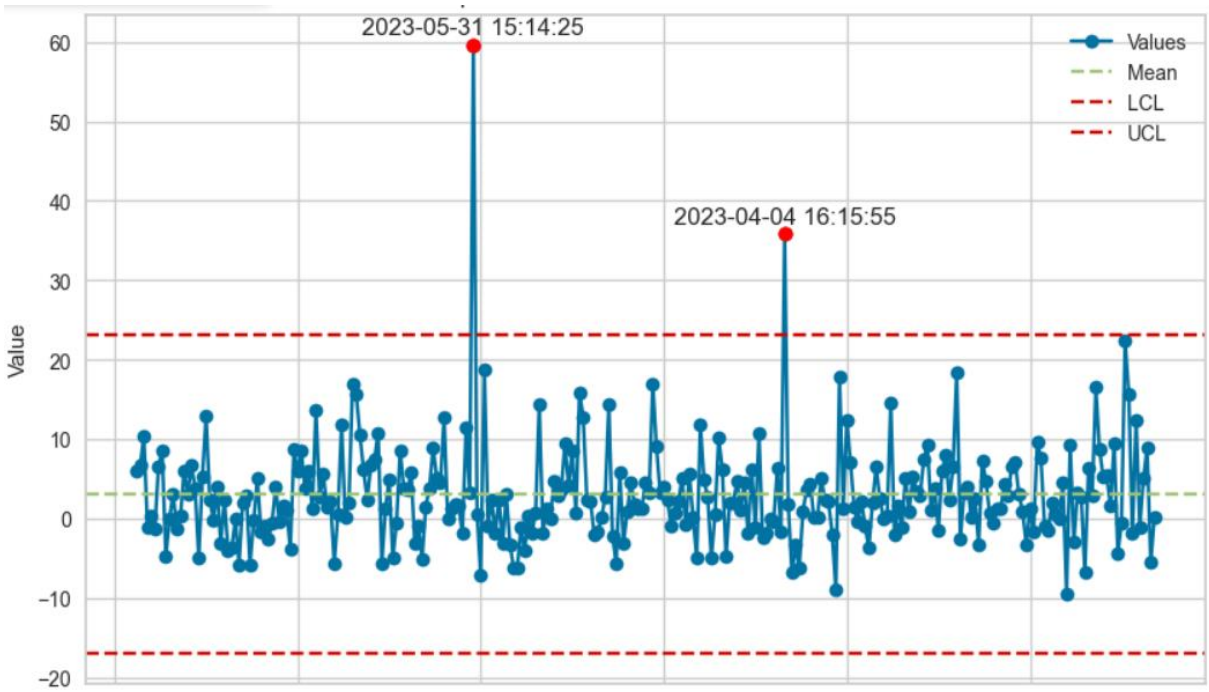


Figure 5-36 Control chart - Example 4

To compare to the state of the art, an outlier detection approach called Isolation Forest (Liu, Ting and Zhou, 2008) is trained using the same set of features. This method detected 30 of the inaccurate values correctly which is about 69.76%. Example plots on the result of the Isolation Forest on the same test setares given in Figure 5-37 and Figure 5-38.

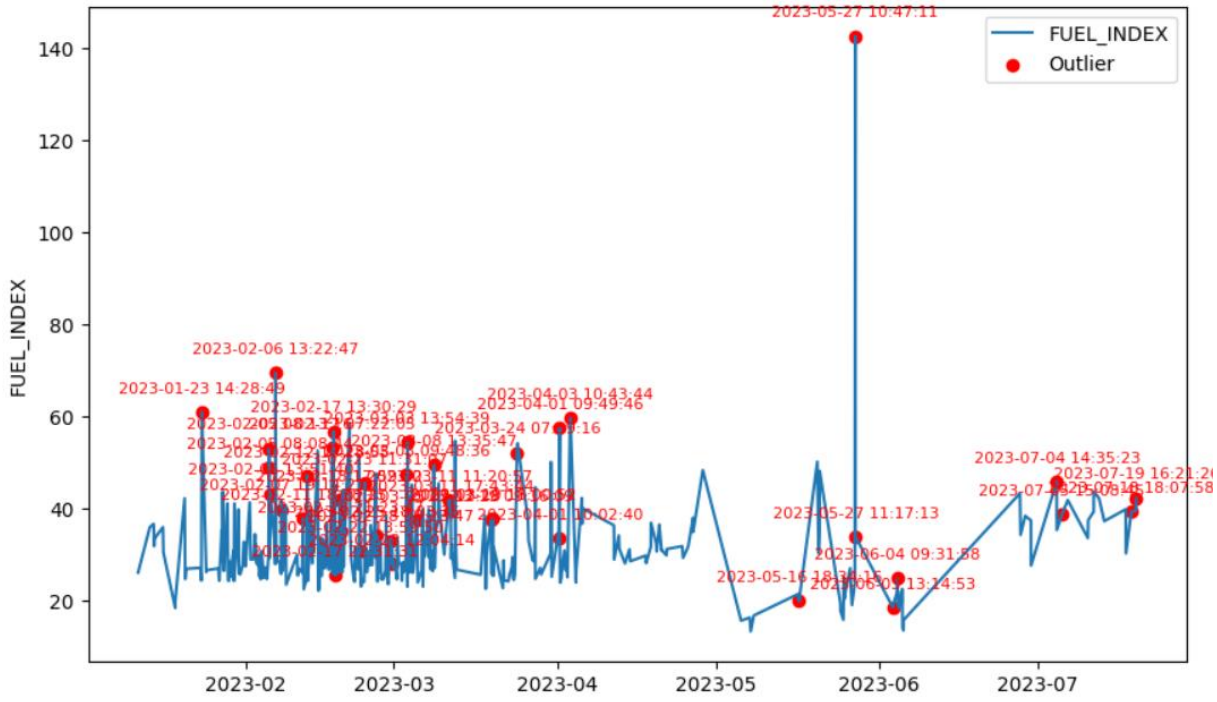


Figure 5-37 Isolation Forest - Example 1

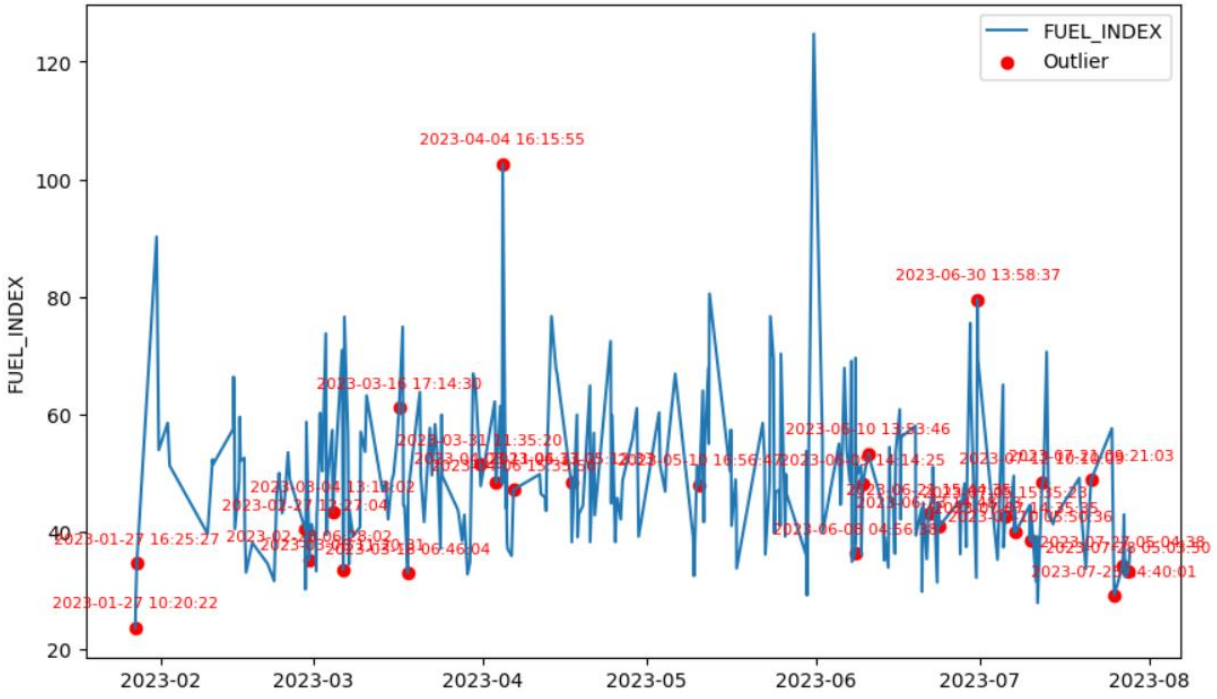


Figure 5-38 Isolation Forest - Example 2

Even though this method detected 69.76% of the inaccurate values, it has wrongly identified multiple accurate values as inaccurate, which is a frequent problem in many outlier detection methods based on the literature review summarized in section 2.4.

Performance metrics comparison

To get a good understanding of the proposed method and the selected state of the art, i.e., Isolation Forest in this case, the confusion metrics of both approaches are presented in Figure 5-39 and Figure 5-40 respectively.

1. Confusion matrix of the proposed approach

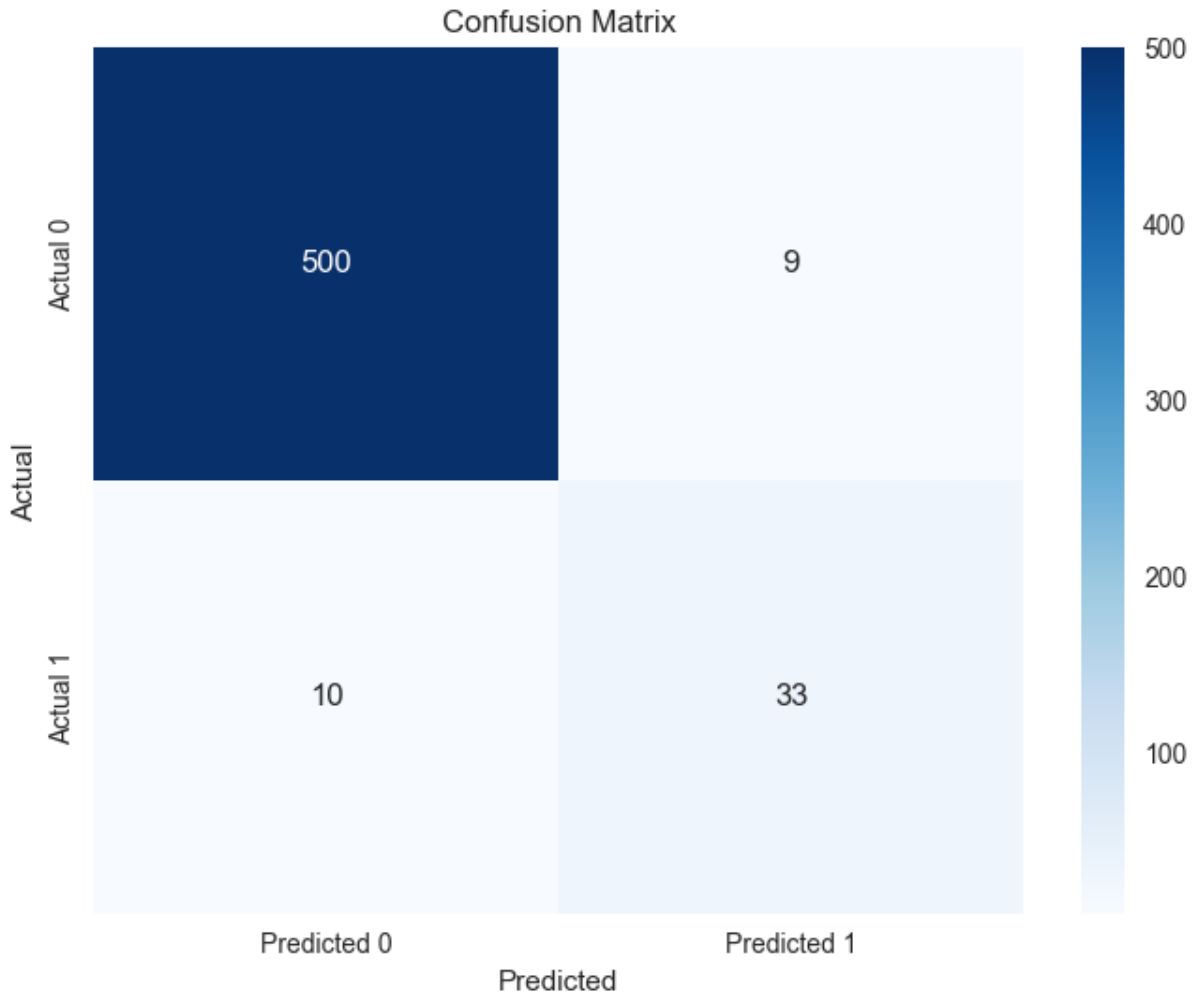


Figure 5-39 Confusion matrix of the adopted approach

2. Confusion matrix of Isolation Forest

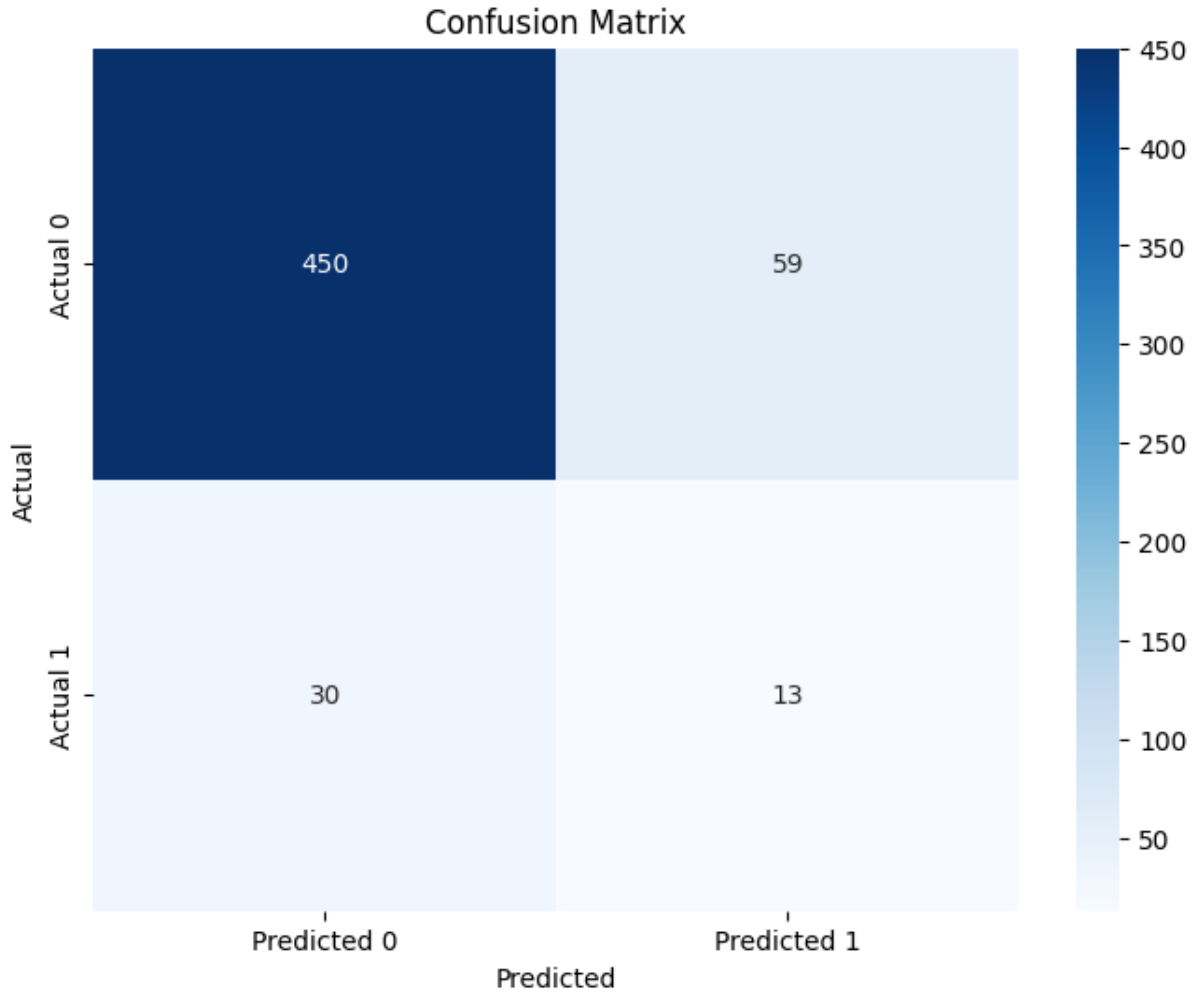


Figure 5-40 Confusion matrix of Isolation Forest

The summary of different performance metrics of both the proposed method and Isolation Forest is presented in Table 5-16 below.

Table 5-16 Performance metrics of proposed method vs Isolation Forest

Performance Metrics	Proposed Method	Isolation Forest
Accuracy	0.97	0.84
Precision	0.79	0.18
Recall	0.78	0.30
F1-score	0.78	0.23

As shown in Table 5-16, the proposed method provided a better result in all metrics.

2. Application on Public dataset

For reproducibility reasons, the above process is repeated using a public dataset. The public data set used for this scenario is fuel consumption data collected from European buses with sensors retrieved from a public github repository (Rosameo, 2021) described in Table 5-17 below, containing the following data elements.

Table 5-17 Data elements of bus public dataset for fuel consumption

Field name	Example value	Description
Date-time	43480.25762	Trip datetime
VehicleID	0	identifier of the vehicle
avg_slope	0.009036145	average slope of the path
mass	19.614	mass in ton of the vehicle including passengers
aircond_ptime	0	percentage of travel time with air conditioning on
stop_ptime	0.12244898	percentage of the travel time with the vehicle stopped and with the engine on
brake_usage	0.367346939	percentage of the travel time with the brake and with the engine on
accel	0.617674419	percentage of the travel time with the accelerator pedal pressed
fuel_per_km	0.75	fuel used in the trip

Examination of this dataset shows that it does not have any missing value. It also does not have an outlier. In addition, it consists of relatively few features. Therefore, the only pre-processing performed on this dataset was normalization with z-score.

Feature selection

The same method implemented earlier is used to select and rank features according to their importance and is presented in Figure 5-41 below. The feature importance ranking result is similar to a previous study conducted on the same dataset (Rexeis, Röck and Hausberger, 2018).

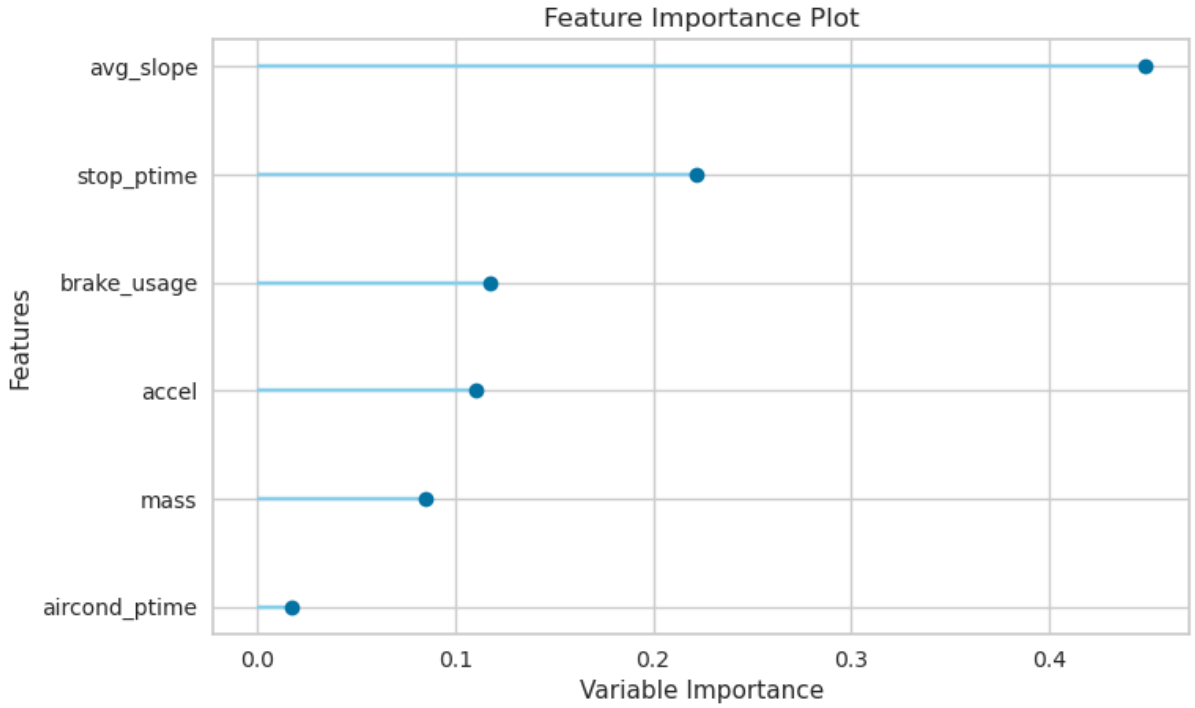


Figure 5-41 Feature importance of the public dataset

Model Building

Using the features presented in Figure 5-41 and fuel consumption as the target feature, different machine learning models including the following are trained: -

- Random Forest Regressor
- Gradient Boosting Regressor
- Elastic Net
- Lasso Regressor and so on

Table 5-18 Results of different algorithms for fuel consumption prediction on public dataset

Model	MAE	MSE	RMSE	R2	RMSLE	M APE	TT (Sec)
RandomForest Regressor	0.0602	0.0060	0.0778	0.6785	0.0521	0.1338	3.0290
GradientBoosting Regressor	0.0674	0.0072	0.0849	0.6171	0.0570	0.1524	1.2740
Eastic Net	0.1112	0.0188	0.1372	-0.0002	0.0936	0.2838	0.0210
Lasso Regressor	0.1112	0.0188	0.1372	-0.0002	0.0936	0.2838	0.0210

In this case, Random Forest is selected as it resulted in the best output as given in Table 5-18 above. Further tuning is performed on Random Forest and the result is given in Table 5-19 below.

Table 5-19 Random Forest tuning result for fuel consumption on public dataset

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Random Forest Regressor	0.0336	0.0024	0.0492	0.8719	0.033	0.075

Following the same steps performed on the connect dataset, the tuned model is saved for future prediction. Then the saved model is referenced on new data to predict fuel consumption. Sample prediction result is given in Table 5-20 below.

Table 5-20 Prediction result on public dataset for fuel prediction

Date-time	VehicleID	avg_slope	mass	aircond_ptime	stop_ptime	brake_usage	accel	fuel_per_km	Label
1/15/2019 6:10	0	0.009036	19.61	0	0.122449	0.367347	0.61767	0.72	0.70149
2/1/2019 12:47	11	0.035714	20.41	0	0.090909	0.090909	0.32356	0.74	0.72365
2/1/2019 13:37	0	0.001786	22.68	0	0.487395	0.201681	0.49623	0.73	0.72115
2/1/2019 16:50	8	0	20.67	0	0.153846	0.261538	0.33939	0.69	0.68906
2/1/2019 18:30	0	0.001799	24.3	0	0.23913	0.130435	0.28367	0.68	0.68816

Control chart and evaluation on the public dataset:

As explained earlier, the chosen public dataset is clean and there is no knowledge of erroneous value in the target variable, i.e., fuel consumption. On the other hand, the objective of this scenario is detecting inaccurate values of the chosen data element in the data. Therefore, error values (noise) are used to evaluate the results on this dataset. As described in (Kalapanidas *et al.*, 2003), introducing noise is a common method. There are various techniques of introducing noise including adding a randomly distributed error value multiplied by the standard deviation to each value as applied in (Kalapanidas *et al.*, 2003). In this research, the same proportion of errors in the connect dataset is used to introduce noise in the public dataset, which is explained as follows. 10% of the data set, which amounts to 6660, is kept for testing purposes. Then for 10% of this set (which is 666), the actual value is replaced with systematically calculated inaccurate values (noise). As mentioned earlier, the experiment is performed using real-life connected data. The outcome shows that the inaccurate values deviate by a range of values from

10% to 80% from the actual value. Therefore, new inaccurate values are introduced with the same proportion for the 10% of the data kept. Accordingly, eight different error rates are used with equal proportion. Therefore, for the first 12.5% of 666, i.e., 83, the actual value is replaced with a value of 10% higher, for the next 12.5%, the actual value is replaced with 20% higher and so on.

Then the trained model is applied to find the predicted values. Next, the difference to the corresponding actual values (i.e., the systematically introduced error values for the added noise) are calculated. Last, UCL and LCL values are calculated for each record using the calculated difference, and values beyond UCL are identified as inaccurate. To visualize this, a control chart is developed with an option to filter for a specific vehicle and required standard deviation.

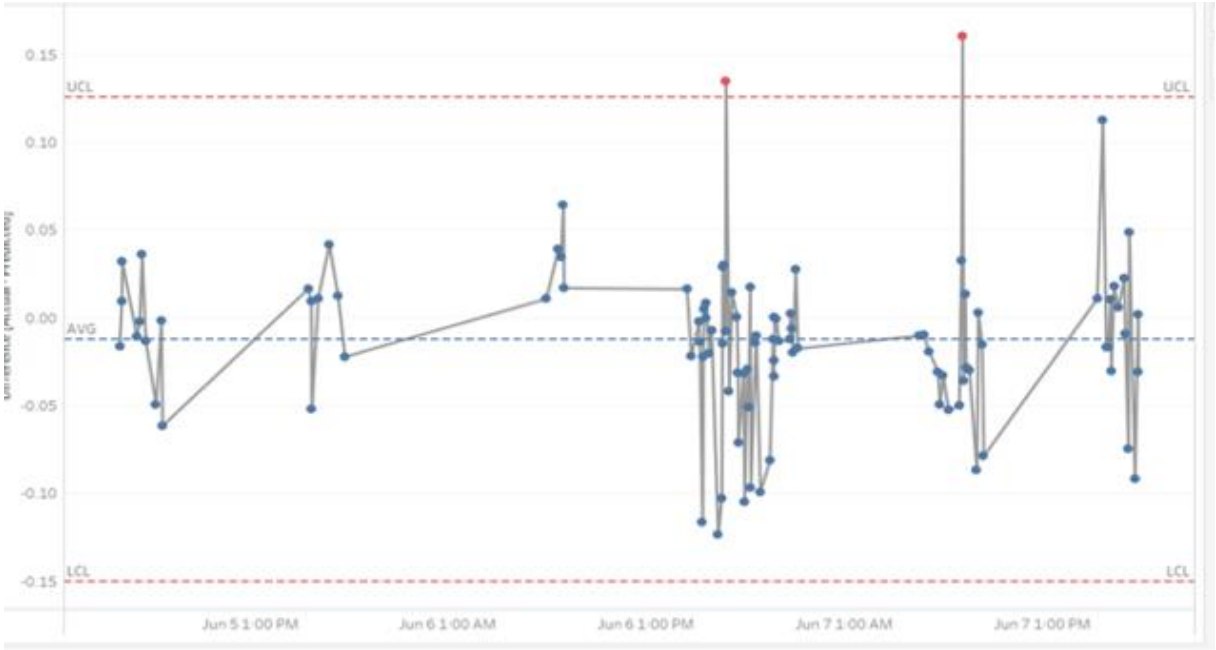


Figure 5-42 Control chart showing inaccurate values on public dataset

Table 5-21 below shows the introduced inaccurate values for a single vehicle and Figure 5-42 gives the corresponding control chart detecting two of the incorrect values accurately when 3 standard deviation is selected.

Table 5-21 Example of noise, predicted value and actual values for validation

Date-time	fuel_per_km_error_value	predicted_fuel_per_km	actual_value
6/7/2019 8:09	0.667651403	0.506997609	0.517651
6/7/2019 16:39	0.748088235	0.635558882	0.658088235
6/6/2019 17:49	0.603460722	0.468844247	0.473460722

The outcome shows that 566 of the error values or the introduced noise out of the 666, which is 85%, were correctly detected. This represents 85% of the inaccurate values correctly identified as inaccurate.

Chapter 6 : The Proposed Framework for Connected Vehicles Data Quality Assessment – Iteration 3

6.1 Introduction

This proposal introduces an ML enabled framework for assessing the quality of CV data. It integrates classical and general-purpose methods with ML techniques to take advantage of the strength of both approaches. According to (Azimi and Pahl, 2021), DQ is often not directly observable or measurable, necessitating innovative approaches to infer quality. This methodology, therefore, intends to improve the precision and effectiveness of CV DQ assessment.

The proposed framework in this chapter was developed using the knowledge gained from previous chapters by organizing the different elements into two main views namely **Process View** and **Implementation View**.

The structure of this chapter is organized as follows. First, the understandings gained from earlier chapters are reiterated. Next, the process view of the framework is discussed followed by the implementation view of the proposed framework. Finally, evaluation of the proposed DQ assessment framework against requirements and overall, Strength, Weakness, Opportunity, and Challenge (SWOC) analysis is presented.

6.2 Foundational Work from Previous Chapters

Chapter 1 stresses the motivation and challenges; and sets research objectives: CV can generate vast amounts of data that have multiple applications, including traffic management, predictive maintenance, and autonomous driving. However, the quality of the data can be compromised by several factors, including sensor errors, network issues, and data corruption. Poor DQ can result in inaccurate insights, poor decision-making, and increased safety risks. Therefore, establishing a reliable DQ assessment framework for CV is crucial. Therefore, this chapter defines the study objectives.

Chapter 2 provides a review of the literature: As was covered earlier in Chapter 2, the research literature has discussed several different frameworks that have been established for evaluating the quality of data. Some are better suited for one industry or business than they are for others, while others are better suited for a different industry or business. This is because diverse types of companies encounter varying degrees of difficulty with regards to the quality of their data. Some organizations have issues with the data's completeness and consistency,

while others have issues with the data's lineage and the content of the gaps they find in the data. As a result, not all problems with the quality of the data may be assessed using the same set of procedures and methodologies. This is where a DQ framework is used – one designed for a specific business case. This is even more important in the CV domain, where the ecosystem is already more complicated due to the additional spatial and temporal dimensions introduced by the CV system. This chapter also underlines the present gap in CV DQ assessment.

Chapter 4 adopts and implements a selected classical DQ assessment framework in the form of a prototype dashboard: By adopting a candidate framework from Chapter 2, it provides evidence regarding the importance and utility of classical general purpose DQ assessment frameworks. It also shows the limitations of these frameworks by demonstrating the inability of the implemented dashboard to capture all defined metrics especially those related to CV characteristics such as space and time.

Chapter 5 implements three scenarios using ML and statistical methods to assess DQ requirements that were not captured in Chapter 4: By implementing ML and statistical quality control methods for selected data elements, it demonstrates that classical DQ assessment frameworks can be enhanced with advanced methods such as ML for CV DQ assessment.

6.3 Process View of the Proposed Framework

Figure 6-1 below provides the process view of the proposed framework. The framework considers the common DQ assessment approaches and extra complexities introduced in the CV ecosystem. The Framework is constructed on the following set of fundamental components derived from the insights acquired in the preceding chapters.

A. Phased Approach
B. Shared Responsibility and Collaboration
C. Use of Framework
D. Advanced Methods (ML and Statistical Control Charts)
E. Continuous Assessment

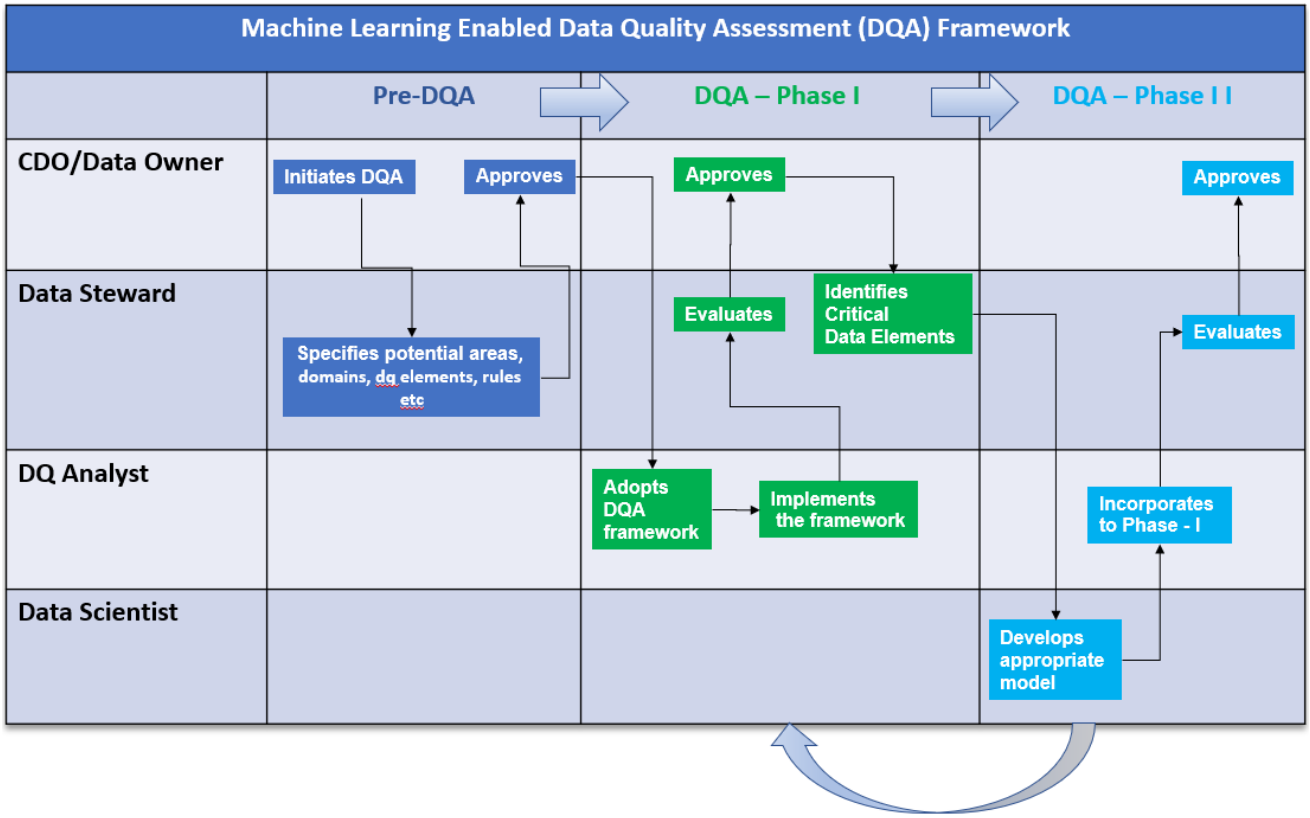


Figure 6-1 Process view of the proposed Data Quality Assessment (DQA) Framework for connected vehicles data

A. Phased Approach

The implementation of DQ assessment should be done in phases. Firstly, it is important to establish a consensus regarding the importance of the DQ assessment endeavor. This facilitates garnering the support of all relevant parties involved. This conceptual framework proposes three distinct phases namely pre-DQA, DQA-Phase I and DQA-Phase II. The pre-DQ assessment (pre-DQA) phase includes the initial stages of the project, involving the start of the project, the establishment of alignment among stakeholders, and the completion of essential preparations such as resource allocation. During DQ assessment phase I (DQA-Phase I) of the DQ assessment process, a framework for evaluating DQ is adopted. This phase includes the establishment of measures or metrics to assess DQ, as well as the implementation of the actual DQ assessment. The implementation is then evaluated, the outcome of this evaluation is assessed, and limitations or gaps are recognized. Specifically, this phase identifies critical data elements that cannot be assessed using the framework adopted. The second phase of DQ assessment (DQA-Phase II) involves the use of advanced techniques, such as ML and statistical methods, to analyze critical data elements that were not fully assessed in the previous phase of DQ assessment (DQA-phase I).

B. Shared Responsibility

The process of DQ assessment requires a coordinated effort. The roles within organizations regarding DQ or data management in general might differ, with some organizations combining one or more roles together. According to the findings of Dama International (2017), it is evident that organizations allocate one individual to multiple roles or combine distinct roles into a singular function. Nevertheless, Dama International (2017) proposes to form a DQ team which includes the following members: Data owner, Data Steward, Data Users, Data Producers, Data Analyst, and Data Custodian. Wende (2007) also recommended a similar set of roles in the data governance model for DQ management she proposed. The relevant aspect to consider here is that the assessment of DQ requires a collaborative endeavor involving a well-defined allocation of roles and responsibilities. For this study, the roles defined by Wende (2007) are adopted and described as follows.

1. Chief Data Officer (CDO)/Executive Data Steward

This role's main responsibility includes overseeing the overall level of DQ. Therefore, this function is responsible for initiating a DQ assessment. Furthermore, this role must effectively persuade senior management to allocate appropriate resources and bears the responsibility of evaluating and approving the outcomes of the DQ assessment at each stage.

2. Data Steward

This role is responsible for establishing the business objectives related to DQ improvement, identifying the individuals or groups who have ownership or a stake in the data, determining the business processes that are affected by or affecting DQ, and defining the rules that govern the management and usage of the data. Additionally, it evaluates the execution of the DQ assessment activities, finds areas of improvement, and highlights areas of uncertainty.

3. Data Quality Analyst

This role performs an evaluation of the available data against the rules outlined in the definition phase by implementing DQ assessment procedures. This role assesses data based on various aspects specifically DQ dimensions, including the accuracy of data elements, the completeness of all necessary data elements, the consistency of data elements across numerous data sets, the timeliness of data, and other relevant factors. Depending on the scale and complexity of the DQ project in context, it may be necessary to conduct multiple assessments using various tools and

techniques. The decision to employ these tools will depend on factors such as the amount and variety of data involved.

4. Data Scientist

Developing an ML model to evaluate the critical data elements that have not been accurately assessed in DQA-Phase I is necessary. Therefore, the proposed framework introduces a new role that complements the existing roles in the classical DQ assessment frameworks to design and implement appropriate ML models.

C. Use of Framework

Considering the variations in challenges and risks associated with DQ across different organizations, it is crucial to implement a methodical approach. Utilizing a structured framework enhances the chance of success in DQ assessment endeavor. It also helps to have a standard that can be uniformly followed across different organizations and domains.

D. Advanced Methods

When the existing classical DQ assessment frameworks are applied in complex systems, like CV, frequently exhibit limitations in terms of covering all the crucial metrics required for a thorough assessment. Therefore, the use of advanced techniques such as ML and statistical components is essential to enhance the thorough evaluation of DQ. The development of the model may exhibit variability depending on the specific data element identified and the contextual aspects inherent to the given situation. Therefore, the skills and knowledge of a data scientist is required for the proposal, design, and development of an appropriate model.

E. Continuous Assessment

Continuous assessment of DQ holds significant importance to make sure that data remains fit for purpose (Azimi and Pahl, 2021). Following the initial assessment, a comprehensive evaluation should be performed. Ideally, the DQ should achieve the expected level and be verified through ongoing assessments. However, the outcome of the DQ assessment may result in unsatisfactory results. Certain data elements that resulted in unacceptable outcomes may be considered as critical data elements. In such instances, it is possible to undertake improvement projects, the outcomes of which ought to be evaluated. Furthermore, it is important to note that the quality of data cannot be assumed to be consistently high through time, even if it may be satisfactory at present. Consequently, it is important that the assessment of DQ is conducted as an ongoing process.

6.4Implementation View of the Proposed Framework

The implementation view of the proposed framework is depicted in Figure 6-2 below. It comprises of four distinct yet interrelated components. First, the data that is to be evaluated is represented as a source. The data source can be stored either in a cloud-based storage system or within an on-premises infrastructure. The subsequent and principal component of the framework is the DQ assessor, which comprises three sub-components. The data reader component is utilized to extract the necessary data from various sources. The DQ Metrics repository is responsible for storing the metrics established based on the requirements of DQ assessment and mapped to the various dimensions of DQ, as shown in Chapter 4. The DQ Assessor repository is responsible for storing the implementations of rules in the form of functions and ML models that determine the state of each record extracted from the data sources using the data reader according to the DQ metrics. There are two essential methods via which this repository ought to be continuously updated: 1. By adding new rules and functions or making changes to existing rules using knowledge from subject matter experts or historical trends from the data. 2. updating existing ML models by retraining or introducing new models via training. The backend of the DQ Assessor utilizes function calls and ML model references to apply DQ Metrics on the extracted dataset. The status component is a repository of data that stores the outcome or result of the DQ assessor. It includes several properties such as the name and description of the DQ metrics, the status of the record, the score (if applicable), and any recommendations (if applicable). The presentation component serves the purpose of displaying outcomes through the utilization of dashboards and reports. The pipeline and orchestration component are used to manage workflows and scheduling tasks.

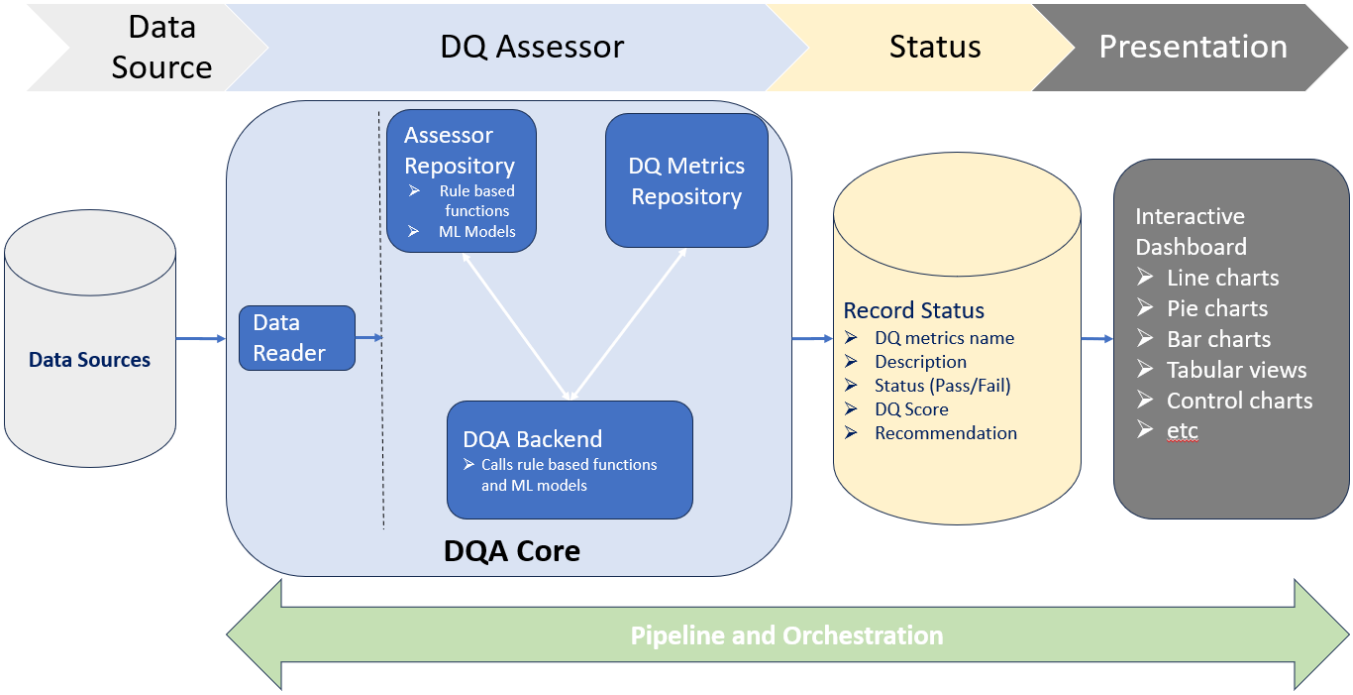


Figure 6-2 Implementation view of the proposed Data Quality Assessment (DQA) framework for connected vehicles data

This framework can be implemented in any appropriate tool and programing language of convenience. In this research the framework is implemented based on the designed methodology in Chapter 3 and explained as follows. The data reader is implemented with Structured Query Language (SQL) embedded in Python. The data used in this implementation is stored in snowflake cloud data warehouse system. The DQ assessor is implemented primarily in Python. First, a series of python functions are written for defined metrics in Chapter 4 during adoption of classical DQ assessment framework. Next, ML models were developed and stored as a set of pickle files for reference. AWS SageMaker notebook was used as a development environment for the ML models. Three models were developed and stored as pickle files: 1. Logistic regression as described in section 5.2, 2. SARIMA Time Series model as described in section 5.3, and 3. RandomForest and Light GBM as explained in section 5.4. The DQ metrics repository is stored in the snowflake database. The DQ assessment backend is implemented in Python to call the assessor repository functions and ML models according to the DQ repository and to write the result to the status repository which is implemented in a snowflake database. The presentation component is implemented with Tableau and Power BI visualization tools. Finally, the CI/CD pipeline and orchestration is implemented using Github actions.

6.5 Evaluation and SWOC Analysis

6.5.1 Evaluation against Requirements

The queries raised during the evaluation of the implemented classical DQ assessment framework in Chapter 4 are equally relevant for the overall DQ assessment framework. Given that the proposed framework builds upon the existing classical DQ assessment framework, this evaluation specifically tries to validate if the proposed framework addresses the gaps identified in Table 4-3, with the context of the evaluation criteria set in section 3.7 and re-iterated as follows:

- Were all metrics defined based on the selected DQ dimensions included?
- Were the results of the DQ assessment conclusive?

Table 6-1 Evaluation of the proposed DQ assessment framework with respect to DQ requirements

Requirement	Gap Identified in classical framework	Enhanced framework
All messages generated should be available. There shall not be a message or a data element missing.	➢ Non-communicating vehicles are not known whether they are facing issue or not. ➢ Missing data from non-communicating vehicles is not detected or not known.	**Improvement:** ➢ Non-communicating vehicles due to issues were detected as demonstrated in section 5.2, which indicates a potential issue of missing data or delay. ➢ Missing mileage was forecasted as demonstrated in section 5.3. **Limitation:** ➢ Accuracy rate depends on the quality of the ML model.
100% of the trips should have no delay, trips with a delay of equal to or less than 15 minutes are not considered as delayed.	➢ Delay from non-communicating vehicles is not known.	**Improvement:** ➢ Delay was detected as demonstrated in section 5.2. **Limitation:** ➢ Accuracy is dependent on the quality of ML model. ➢ Does not quantify amount of delay.
All data should be available as generated from the vehicle (without any manipulation). Information received should be accurate	➢ Accuracy of received data is not known	**Improvement:** As demonstrated in section 5.4, inaccurate values could be detected with reasonable accuracy (80% to 85%). **Limitation:** Accuracy is dependent on the quality of ML model.

6.5.2 Strength, Weakness, Opportunity, and Challenge (SWOC) Analysis

In this section, a summary of the critical analysis of the proposed DQ assessment framework employing strength, weakness, opportunity, and challenges (SWOC) analysis is presented.

- **Strength**

Firstly, the proposed framework enables the inclusion of additional features that are not readily available in the data, hence providing valuable insights for assessing DQ. In this research, proximity to a parking location is added to the feature set. This variable has been identified as a strong predictor for determining the presence of missing data or delayed data in CV dataset, as outlined in section 5.2.

Additionally, the framework makes it possible to integrate DQ dimensions and associated metrics, such as accuracy, that cannot be evaluated using classical DQ assessment frameworks unless there is reference data available. This is because unavailability of reference data is common in CV systems as data is generated while the vehicle is driving with varying space and time attributes (spatio-temporal aspects). In this research, the accuracy of the data element pertaining to fuel consumption is evaluated by the utilization of ML and statistical control chart, as demonstrated in section 5.4.

- **Weakness**

Nevertheless, the proposed framework has its own limitations. To begin with, it is important to note that the outcome is not entirely definite (not 100% certain). Since the enhancement of the proposed framework is based on predicted values, it depends on the quality of the ML models developed. Moreover, it should be noted that even with a very proficient ML model, achieving absolute accuracy is not entirely guaranteed. In the assessment of accuracy DQ dimension in section 5.3, the framework only indicates whether the given value is trustworthy or not trustworthy. Values deemed untrustworthy merely display the predicted value, expected to be close to the true value. The ML algorithms developed also depend on using constructed features. For example, Scenario I uses engineered features including previous status such as number of times it has missing data or delayed data in the past and proximity to parking locations. If the vehicle is new and if there is a new parking location, this feature will not be accurate.

Additionally, the proposed framework exclusively considers metrics that are capable of being objectively measured on the basis of objective DQ dimensions listed in Table 4-2. The framework does not incorporate subjective assessment.

- **Opportunity**

The proposed framework is expected to improve the comprehensive assessment of DQ for CV systems, as demonstrated in the implemented scenarios outlined in sections 5.2, 5.3 and 5.4. The use of the framework will enhance users' trust in utilizing the data for decision-making by eliminating uncertainty and providing a clear indication of the level of DQ. Additionally, it helps to avoid the use of inaccurate data, as the assessment result intends to clearly highlight the problematic part of the data. This will encourage organizations to use CV data to develop applications to generate new revenue, reduce costs and improve processes.

- **Challenges**

The framework has included an additional element, specifically ML, into the existing classical DQ assessment frameworks. This necessitates additional resources in terms of skilled manpower as well as tooling compared to the resources normally used by classical DQ assessment frameworks. This presents itself as one of the challenges. Another significant challenge encountered during the development of this framework relates to concerns around data privacy and security. The implementation of the GDPR has posed challenges in using data for ML purposes. Also, availability of public data within the field of CV is limited.

One further challenge that arises is the absence of standardization. The development of ML models depends on several factors, including the domain, data characteristics, and special needs and objectives. Furthermore, the effectiveness of the models depends on the developer's expertise and the setting's contextual factors.

Chapter 7 : Conclusions and Future Work

7.1 Introduction

This chapter provides a comprehensive summary of the findings and conclusions derived from the dissertation. The chapter commences with restating the objective set in Chapter 1 and findings from the remaining chapters by highlighting the interconnectedness of the different chapters. The subsequent section emphasizes the research contributions by addressing the research questions and concluding remarks. Finally, this chapter closes by drawing attention to specific aspects that may necessitate further investigation in the future.

7.2 Summary of the Thesis

The primary goal of this dissertation was to develop a ***Machine Learning Enabled Data Quality Assessment Framework*** specifically tailored to CV data. While DQ research is well-established, there is limited work related to DQ assessment for CV data identified as a gap in the field as described in the literature review. This research aimed to address this specific gap by studying existing solutions and by proposing a new enhanced solution. Chapter 1 sets the research objectives. Chapter 2 lays the foundation by introducing key concepts related to CV, DQ, and DQ assessment frameworks and methodologies. The subsequent chapters, including Chapter 2, were designed to answer the research questions specifically Chapter 2 through a literature review, Chapter 4 through the adoption and implementation of a selected classical DQ assessment framework informed by the literature review, and Chapter 5 by implementing ML methods for selected scenarios to fill the identified gaps informed from Chapter 2 and Chapter 4.

The research findings indicate that DQ assessment in CV is challenging and cannot be adequately addressed by general purpose classical DQ assessment frameworks due to the complex nature of the CV ecosystem. On the one hand, existing general purpose classical DQ assessment frameworks lack the means to assess difficult DQ issues but provide valuable insights and are easier to implement. On the other hand, methodologies using ML manage to tackle difficult DQ issues but lack generalizability. In addition, they are difficult to implement. Therefore, the proposed solution lies in applying advanced methods like ML and statistical techniques to complement classical DQ assessment frameworks which takes advantage of the strengths of both classical DQ assessment frameworks and ML methodologies. The results suggest that incorporating ML and statistical methods can improve DQ assessment in CV.

To ensure the objectives have been achieved, the research questions, formulated in the beginning, are revisited, and discussed in relation to the research's contributions as follows.

7.3 Conclusions Related to Research Questions

RQ1: What are the existing Data Quality assessment best practices, methodologies, and frameworks applicable to connected vehicles?

To answer this research question, ***Chapter 2: Connected Vehicles, Data Quality and Data Quality Assessment Literature Review*** is employed. This chapter includes systematic literature review.

The literature review has identified several methodologies and frameworks for DQ assessment. However, it became evident that research specifically focusing on DQ assessment in the context of CV is limited. On the other hand, general DQ research is thriving and actively evolving. In recent years, the growth of IoT, connectivity, and big data has captured researchers' attention, leading to an increase in research outputs in this domain. Researchers are actively exploring ways to address the unique challenges presented by these complex systems.

For the sake of simplicity and convenience, the DQ assessment frameworks reviewed were categorized into two main groups:

- General purpose classical DQ assessment frameworks: These frameworks are widely used and considered applicable across various domains. These frameworks are also relatively easier to implement. Most of these DQ frameworks measure DQ based on metrics developed based on DQ dimensions. The finding also suggests that there are variations on DQ dimensions specified in each framework and there is lack of standardization. Some of the identified DQ assessment frameworks in this category include Total Data Quality Management (TDQM), A methodology for information quality assessment (AIMQ), Hybrid Information Quality Management (HIQM), Comprehensive methodology for data quality management (CDQ), The Data Quality Assessment Framework (DQAF) and so on.
- Frameworks that apply ML methodologies: These methodologies leverage ML techniques to address complex DQ challenges and are gaining popularity in DQ assessment. The popularity can be attributed to the growth of IoT, big data and connected systems. The systematic literature review shows that the methodologies in this category discuss only specific domains and most of them focus on outlier detection. However, they can handle difficult DQ topics such as accuracy compared to general purpose classical DQ dimensions. The findings also show that they are difficult to implement.

In conclusion, the research emphasizes the necessity for further investigation focused on evaluating the quality of data in CV. It acknowledges the significance of existing frameworks for assessing DQ and emphasizes on the potential benefits of integrating advanced techniques such as ML to improve the assessment of DQ in this field.

RQ2: What are the limitations of existing Data Quality assessment methods?

Both ***Chapter 2: Connected Vehicles, Data Quality and Data Quality Assessment Literature Review*** and ***Chapter 4: Data Quality Assessment Framework Adoption and Prototype Development – Iteration 1*** are employed to address this research question.

The literature suggests that various DQ assessment methodologies encounter different limitations. General purpose classical DQ assessment frameworks, although widely used, seem to struggle in capturing complex dimensions of data such as accuracy, which is crucial in the context of CV. On the other hand, ML methodologies show promise in handling these complex aspects; however, they lack generalizability and often require a lot of customization when applied to different domains. It is also understood that CV architecture is inherently complex, incorporating spatio-temporal aspects, which further complicates DQ assessment. The dynamic nature of CV ecosystem, involving multiple components and data sources, poses challenges for traditional classical DQ assessment methods.

Moreover, the use of the chosen classical DQ framework, as discussed in Chapter 4, has revealed that not all DQ concerns can be effectively addressed with only classical DQ assessment frameworks. For example, the implemented dashboard failed to show if a vehicle not sending data was driving and experiencing issue or parked and switched off. This means it was not possible to assess the completeness and timeliness of DQ dimensions. In addition, it was not possible to determine if received data is accurate or not, i.e., the accuracy DQ dimension could not be handled.

Overall, the findings from both the literature and the adopted classical framework implementation highlight the need for better DQ assessment approaches that can effectively address the complexities of CV. Balancing the benefits of advanced techniques like ML with the need for generalizability, adaptability, and simplicity of general-purpose classical frameworks remain an important consideration.

RQ3: To what extent does incorporating Machine Learning improve Data Quality assessment on connected vehicles data?

Chapter 5: Data Quality Assessment Framework Enhancement with Machine Learning for Connected Vehicles data – Iteration 2 is employed to answer this research question.

The first scenario demonstrated that ML could help to capture unknown features. In this regard, unsupervised ML, specifically DBSCAN clustering, was used to identify parking locations. This helped to generate a new feature namely distance to the nearest parking location, which captures the spatio-temporal aspect of CV. This feature is combined with other features and used to predict whether vehicles not sending data is because it is facing a real issue and hence information is missing, or it is simply parking, and power is switched off. This helped to detect missing data and delayed data. This scenario has therefore demonstrated that the completeness and timeliness DQ dimensions can be assessed better by applying ML.

In the second scenario, time series forecasting was employed to demonstrate that missing data can be forecasted by using the mileage data element, which involved forecasting values for vehicles that are missing data. This proves that ML can be leveraged to improve DQ.

The third scenario demonstrated that ML and statistical quality control can be used to determine if a certain data value received is accurate or not. Specifically, the fuel consumption data element was used to implement this scenario. Using historical data, a predictive model was developed. Subsequently, the discrepancy between the predicted value and the actual value was computed. On the difference, statistical quality control is applied. Using the statistical quality control, values crossing the control lines of UCL are identified as inaccurate. This helps to increase the confidence of users by highlighting how trustworthy a given data is.

7.4 Contributions

The findings suggest that using advanced techniques like ML and statistical methods can significantly enhance DQ assessment in CV by addressing the limitations of classical frameworks. In this regard, the research has made several valuable contributions:

1. A comprehensive review of DQ assessment frameworks in the domain of CV systems:- The study provided an exhaustive and thorough review of DQ assessment approaches, frameworks and methods in relation to CV.
2. Highlighting limitations of general purpose classical DQ assessment frameworks when applied to CV data:- The study highlighted the limitations of existing DQ assessment

frameworks, emphasizing the need for more advanced techniques to address the complex nature of DQ in CV.

3. Application of ML to enhance CV DQ assessment:- The study has also demonstrated the application of advanced methods, specifically ML and advanced statistical methods, for DQ assessment in CV, highlighting the potential of these techniques in improving DQ.
4. Development of a new framework:- The main contribution of this research was the development of **a Machine Learning Enabled Data Quality Assessment Framework for Connected Vehicles** providing a methodical and comprehensive approach to assess and enhance DQ in the domain.

The evaluation of the new framework showed that it effectively addressed the gaps identified in classical frameworks using ML and statistical methods while also acknowledging that classical frameworks still provide valuable insights.

7.5 Future Work

The research has generated some additional research recommendations that need future exploration in the domain of CV DQ assessment. These recommendations are described in this section.

The first one is the generation of synthetic data. To validate any ML method, the availability of data is important, preferably real representative data. However, there is no comprehensive representative trajectory or trip data publicly available to use. Moreover, it is difficult to use company owned data due to restrictions imposed by GDPR. One solution is to generate synthetic trajectory data which is representative of the real-world trajectory without risking any violation of the restrictions of GDPR. Therefore, future research may focus on generating synthetic trajectory data to facilitate research in CV DQ assessment.

Another potential work is to incorporate new data sources. This research employed only data generated from CV. CV data is affected by other factors such as telecommunication, traffic situation, weather condition and so on. Combining such data sources could give more insight.

Yet another potential improvement to this work is to incorporate more ML algorithms to build a more effective ML model. This study focused only on a limited set of algorithms, and potential future research endeavors could include examining recently developed algorithms that may offer improved performance.

Finally, comprehensive CV DQ management is a potential topic of investigation to improve DQ in CV systems as DQ is one aspect of overall data management and any wrongdoing in any of the components of data management affects DQ. This research only investigated DQ from the perspective of objective measurement of accuracy and reliability. However, DQ is dependent on the overall data management strategy including security and privacy, data integration, data governance and ownership. In addition, this research was limited to data collected from the vehicle to the back end and did not explore V2V, V2I and other related data which can be a potential future research topic.

References

van der Aalst, W. M. P., Bichler, M. and Heinzl, A. (2017) 'Responsible data science'. Springer.

Abdelkader, G., Elgazzar, K. and Khamis, A. (2021) 'Connected vehicles: technology review, state of the art, challenges and opportunities', *Sensors*. Multidisciplinary Digital Publishing Institute, 21(22), p. 7712.

Adhikari, R. and Agrawal, R. K. (2013) 'An introductory study on time series modeling and forecasting', *arXiv preprint arXiv:1302.6613*.

Von Alan, R. H. *et al.* (2004) 'Design science in information systems research', *MIS quarterly*. Springer, 28(1), pp. 75–105.

Alrae, R., Nasir, Q. and Abu Talib, M. (2020) 'Developing house of information quality framework for IoT systems', *International Journal of System Assurance Engineering and Management*. Springer, 11, pp. 1294–1313.

Ashton, K. and others (2009) 'That "internet of things" thing', *RFID journal*. Jun, 22(7), pp. 97–114.

Atzori, L., Iera, A. and Morabito, G. (2010) 'The internet of things: A survey', *Computer networks*. Elsevier, 54(15), pp. 2787–2805.

Azimi, S. and Pahl, C. (2021) 'Continuous Data Quality Management for Machine Learning based Data-as-a-Service Architectures.', in *CLOSER*, pp. 328–335.

Aziz, A. A., Saman, M. Y. M. and Jusoh, J. A. (2012) 'Data investigation: Issues of data quality and implementing base analysis technique to evaluate quality of data in heterogeneous databases', *Journal of Theoretical and Applied Information Technology*, 45(1), pp. 360–372. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84874528600&partnerID=40&md5=c0a93b1f761e14519748e0717d2f0282.

Barnes, B. B. and Hu, C. (2013) 'A Hybrid Cloud Detection Algorithm to Improve MODIS Sea Surface Temperature Data Quality and Coverage Over the Eastern Gulf of Mexico', *IEEE Transactions on Geoscience and Remote Sensing*, 51(6), pp. 3273–3285. doi: 10.1109/TGRS.2012.2223217.

Bashir, M. R. and Gill, A. Q. (2017) 'IoT enabled smart buildings: A systematic review', in *Intelligent Systems Conference (IntelliSys), 2017*, pp. 151–159.

Batini, C. *et al.* (2007) 'A Framework And A Methodology For Data Quality Assessment And Monitoring.', in *ICIQ*, pp. 333–346.

Batini, C. *et al.* (2008) 'A comprehensive data quality methodology for web and structured

data', *International Journal of Innovative Computing and Applications*. Inderscience Publishers, 1(3), pp. 205–218.
Batini, C. *et al.* (2009) 'Methodologies for data quality assessment and improvement', *ACM computing surveys (CSUR)*. ACM, 41(3), p. 16.
Batini, C. *et al.* (2015) 'From data quality to big data quality', *Journal of Database Management*, 26(1), pp. 60–82. doi: 10.4018/JDM.2015010103.
Batini, C., Scannapieco, M. and others (2016) 'Data and information quality', *Cham, Switzerland: Springer International Publishing*. Springer.
Benneyan, J. C. (1998) 'Use and interpretation of statistical quality control charts', *International Journal for Quality in Health Care*. JSTOR, 10(1), pp. 69–73.
de Boves Harrington, P. (2006) 'Statistical validation of classification and calibration models using bootstrapped Latin partitions', *TrAC Trends in Analytical Chemistry*. Elsevier, 25(11), pp. 1112–1124.
Box, G. E. P. *et al.* (2015) *Time series analysis: forecasting and control*. John Wiley \& Sons.
Brimicombe, A. (2010) 'GIS, environmental modeling and engineering'. CRC Press/Taylor & Francis Group.
Brimicombe, A. and Li, C. (2009) *Location-based services and geo-information engineering*. John Wiley & Sons.
Budgen, D. and Brereton, P. (2006) 'Performing systematic literature reviews in software engineering', in *Proceedings of the 28th international conference on Software engineering*, pp. 1051–1052.
Cai, L. and Zhu, Y. (2015) 'The challenges of data quality and data quality assessment in the big data era', *Data Science Journal*. Ubiquity Press, 14.
Cappiello, C., Ficiaro, P. and Pernici, B. (2006) 'HIQM: A methodology for information quality monitoring, measurement, and improvement', in *Advances in Conceptual Modeling-Theory and Practice: ER 2006 Workshops BP-UML, CoMoGIS, COSS, ECDM, OIS, QoIS, SemWAT, Tucson, AZ, USA, November 6-9, 2006. Proceedings 25*, pp. 339–351.
Carlo, B. *et al.* (2011) 'A data quality methodology for heterogeneous data', *International Journal of Database Management Systems*, 3(1), pp. 60–79.
Cerqueira, V. *et al.* (2018) 'On Evaluating Floating Car Data Quality for Knowledge Discovery', *IEEE Transactions on Intelligent Transportation Systems*, 19(11), pp. 3749–3760. doi: 10.1109/TITS.2018.2867834.
Cerqueira, V., Torgo, L. and Mozetič, I. (2020) 'Evaluating time series forecasting models: An empirical study on performance estimation methods', *Machine Learning*. Springer, 109,

pp. 1997–2028.
Cheng, H. *et al.* (2018) ‘Data quality analysis and cleaning strategy for wireless sensor networks’, *Eurasip Journal on Wireless Communications and Networking*, 2018(1). doi: 10.1186/s13638-018-1069-6.
Chicco, D., Warrens, M. J. and Jurman, G. (2021) ‘The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation’, *PeerJ Computer Science*. PeerJ Inc., 7, p. e623.
Cichy, C. and Rass, S. (2019) ‘An overview of data quality frameworks’, *IEEE Access*. IEEE, 7, pp. 24634–24648.
Dai, W., Yoshigoe, K. and Parsley, W. (2018) ‘Improving Data Quality Through Deep Learning and Statistical Models’, in *Information Technology-New Generations*. Springer, pp. 515–522.
Davenport, T. and Redman, T. (2015) ‘Build data quality into the internet of things’, *Wall Str. J.*
Delen, D., Walker, G. and Kadam, A. (2005) ‘Predicting breast cancer survivability: a comparison of three data mining methods’, *Artificial intelligence in medicine*. Elsevier, 34(2), pp. 113–127.
Diop, M. *et al.* (2017) ‘A methodology for prior management of temporal data quality in a data mining process’, in *2017 Intelligent Systems and Computer Vision (ISCV)*, pp. 1–8. doi: 10.1109/ISACV.2017.8054906.
Ensafi, Y. *et al.* (2022) ‘Time-series forecasting of seasonal items sales using machine learning--A comparative analysis’, *International Journal of Information Management Data Insights*. Elsevier, 2(1), p. 100058.
Eppler, M. J. (2006) *Managing information quality: Increasing the value of information in knowledge-intensive products and processes*. Springer Science \& Business Media.
Ester, M. *et al.* (1996) ‘A density-based algorithm for discovering clusters in large spatial databases with noise.’, in *kdd*, pp. 226–231.
European Union (2006) *EU rules for working in road transport*, *EU rules for working in road transport*. Available at: https://europa.eu/youreurope/citizens/work/work-abroad/rules-working-road-transport/index_en.htm (Accessed: 15 December 2022).
Farooqi, M. M., Khattak, H. A. and Imran, M. (2018) ‘Data quality techniques in the internet of things: Random forest regression’, in *2018 14th International Conference on Emerging Technologies (ICET)*, pp. 1–4.
Fekade, B. *et al.* (2018) ‘Probabilistic recovery of incomplete sensed data in IoT’, *IEEE*

Internet of Things Journal. IEEE, 5(4), pp. 2282–2292.
Floridi, L. (2013) 'Information quality', *Philosophy & Technology*. Springer, 26(1), pp. 1–6.
for Standardization, I. O. (2021) *Quality Management Systems--Fundamentals and Vocabulary*. International Organization for Standardization.
Fox, C., Levitin, A. and Redman, T. (1994) 'The notion of data and its quality dimensions', *Information processing & management*. Elsevier, 30(1), pp. 9–19.
Francisco, M. M. C. *et al.* (2017) 'Total data quality management and total information quality management applied to costumer relationship management', in *Proceedings of the 9th international conference on information management and engineering*, pp. 40–45.
Galarus, D. E. and Angryk, R. A. (2016) 'A smart approach to quality assessment of site-based spatio-temporal data', in *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–4.
Gitzel, R., Turring, S. and Maczey, S. (2015) 'A data quality dashboard for reliability data', in *2015 IEEE 17th Conference on Business Informatics*, pp. 90–97.
Group, D. Q. D. W. and others (2013) 'The Six Dimensions of EHDI Data Quality Assessment'. DAMA UK.
Gualo, F. *et al.* (2021) 'Data quality certification using ISO/IEC 25012: Industrial experiences', *Journal of Systems and Software*. Elsevier, 176, p. 110938.
Gubbi, J. *et al.* (2013) 'Internet of Things (IoT): A vision, architectural elements, and future directions', *Future generation computer systems*. Elsevier, 29(7), pp. 1645–1660.
Gudivada, V., Apon, A. and Ding, J. (2017) 'Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations', *International Journal on Advances in Software*, 10(1), pp. 1–20.
Hamad, K. (2015) 'QUALITY CONTROL OF ARCHIVED INTELLIGENT TRANSPORTATION SYSTEMS DATA.', *International Journal for Traffic & Transport Engineering*, 5(3).
Han, S., Wu, Q. and Yang, Y. (2022) 'Machine learning for Internet of things anomaly detection under low-quality data', *International Journal of Distributed Sensor Networks*. SAGE Publications Sage UK: London, England, 18(10), p. 15501329221133764.
Hastie, T. *et al.* (2009) 'Model assessment and selection', *The elements of statistical learning: data mining, inference, and prediction*. Springer, pp. 219–259.
Hazen, B. T. *et al.* (2014) 'Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications', *International Journal of Production Economics*. Elsevier, 154, pp. 72–80.

He, H. and Garcia, E. A. (2009) ‘Learning from imbalanced data’, *IEEE Transactions on knowledge and data engineering*. Ieee, 21(9), pp. 1263–1284.
He, Z. *et al.* (2019) ‘Vehicle sensor data-based transportation research: Modeling, analysis, and management’. Taylor & Francis.
Hevner, A. R. (2007) ‘A three cycle view of design science research’, *Scandinavian journal of information systems*, 19(2), p. 4.
Hoofnagle, C. J., van der Sloot, B. and Borgesius, F. Z. (2019) ‘The European Union general data protection regulation: what it is and what it means’, *Information \& Communications Technology Law*. Taylor \& Francis, 28(1), pp. 65–98.
International, D. (2017) *DAMA-DMBOK: data management body of knowledge*. Technics Publications, LLC.
Isaac, D. and Lynnes, C. (2003) ‘Automated data quality assessment in the intelligent archive’, *White Paper prepared for the Intelligent Data Understanding program*. Citeseer, 17.
Jadaan, K., Zeater, S. and Abukhalil, Y. (2017) ‘Connected vehicles: an innovative transport technology’, *Procedia Engineering*. Elsevier, 187, pp. 641–648.
James, G. *et al.* (2013) *An Introduction to Statistical Learning*.
Jia, L. *et al.* (2014) ‘Impact of Data Quality on Real-Time Locational Marginal Price’, *IEEE Transactions on Power Systems*, 29(2), pp. 627–636. doi: 10.1109/TPWRS.2013.2286992.
Jordan, M. I. and Mitchell, T. M. (2015) ‘Machine learning: Trends, perspectives, and prospects’, *Science*. American Association for the Advancement of Science, 349(6245), pp. 255–260.
Juddoo, S. *et al.* (2018) ‘Data governance in the health industry: investigating data quality dimensions within a big data context’, *Applied System Innovation*. Multidisciplinary Digital Publishing Institute, 1(4), p. 43.
Juddoo, S. and George, C. (2018) ‘Discovering most important data quality dimensions using latent semantic analysis’, in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1–6.
Juran, J. and Godfrey, A. B. (1999) ‘Quality handbook’, *Republished McGraw-Hill*, 173(8), pp. 34–51.
Kalapanidas, E. *et al.* (2003) ‘Machine learning algorithms: a study on noise sensitivity’, in *Proc. 1st Balcan Conference in Informatics*, pp. 356–365.
Karkouch, A. *et al.* (2015) ‘Data quality enhancement in Internet of Things environment’, in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pp. 1–8. doi: 10.1109/AICCSA.2015.7507117.

Karkouch, A. *et al.* (2016) ‘Data quality in internet of things: A state-of-the-art survey’, *Journal of Network and Computer Applications*. Elsevier, 73, pp. 57–81.
Karkouch, A. *et al.* (2018) ‘A model-driven framework for data quality management in the Internet of Things’, *Journal of Ambient Intelligence and Humanized Computing*, 9(4), pp. 977–998. doi: 10.1007/s12652-017-0498-0.
Keller, S. *et al.* (2017) ‘The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches’, *Annual Review of Statistics and Its Application*. Annual Reviews, 4, pp. 85–108.
Khan, K. *et al.* (2014) ‘DBSCAN: Past, present and future’, in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232–238.
Khan, K. S. *et al.* (2003) ‘Five steps to conducting a systematic review’, *Journal of the royal society of medicine*. SAGE Publications Sage UK: London, England, 96(3), pp. 118–121.
Kim, S. *et al.* (2019) ‘Extending data quality management for smart connected product operations’, *IEEE Access*. IEEE, 7, pp. 144663–144678.
Kohavi, R. (1995) ‘A study of cross-validation and bootstrap for accuracy estimation and model selection’, *Morgan Kaufman Publishing*.
Kotsiantis, S. *et al.* (2006) ‘Handling imbalanced datasets: A review’, *GESTS international transactions on computer science and engineering*, 30(1), pp. 25–36.
Kotzé, P., van der Merwe, A. and Gerber, A. (2015) ‘Design science research as research approach in doctoral studies’, *AMCIS 2015 Proceedings*.
Kursa, M. and Rudnicki, W. (2010) ‘Feature Selection with the Boruta Package’, *Journal of Statistical Software, Articles*, 36(11), pp. 1–13. doi: 10.18637/jss.v036.i11.
Laptev, N., Amizadeh, S. and Flint, I. (2015) ‘Generic and scalable framework for automated time-series anomaly detection’, in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1939–1947.
Lawson, V. and Ramaswamy, L. (2015) ‘Data Quality and Energy Management Tradeoffs in Sensor Service Clouds’, in *2015 IEEE International Congress on Big Data*, pp. 749–752. doi: 10.1109/BigDataCongress.2015.124.
Lee, Y. W. *et al.* (2002) ‘AIMQ: a methodology for information quality assessment’, *Information & management*. Elsevier, 40(2), pp. 133–146.
Leonardi, A. *et al.* (2016) ‘Dealing with data quality in smart home environments - Lessons learned from a smart grid pilot’, *Journal of Sensor and Actuator Networks*, 5(1). doi: 10.3390/jsan5010005.

Lesouple, J. *et al.* (2021) 'Generalized isolation forest for anomaly detection', *Pattern Recognition Letters*. Elsevier, 149, pp. 109–119.
Li, J. (2013) 'Logistic regression', *Course Notes. URL http://sites. stat. psu. edu/\~{} jiali/course/stat597e/notes2/logit. pdf*.
Liaw, S.-T. *et al.* (2011) 'Data quality and fitness for purpose of routinely collected data--a general practice case study from an electronic Practice-Based Research Network (ePBRN)', in *AMIA Annual Symposium Proceedings*, p. 785.
Liu, F. T., Ting, K. M. and Zhou, Z.-H. (2008) 'Isolation forest', in *2008 eighth ieee international conference on data mining*, pp. 413–422.
Liu, S. *et al.* (2017) 'Context-aware data quality estimation in mobile crowdsensing', in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9. doi: 10.1109/INFOCOM.2017.8057033.
Lombardi, A. *et al.* (2022) 'A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease', *Brain informatics*. SpringerOpen, 9(1), pp. 1–17.
Loshin, D. (2010) *The practitioner's guide to data quality improvement*. Elsevier.
Loshin, D. (2011) 'Evaluating the business impacts of poor data quality', *Information Quality Journal*.
Loshin, D. (2012) *Enterprise data quality: The six dimensions of data quality*. Morgan Kaufmann.
Mahmood, Z. (2020) 'Connected vehicles in the IoV: Concepts, technologies and architectures', in *Connected vehicles in the internet of things*. Springer, pp. 3–18.
Maria, E. *et al.* (2020) 'Measure distance locating nearest public facilities using Haversine and Euclidean Methods', in *Journal of Physics: Conference Series*, p. 12080.
Maydanchik, A. (2007) *Data quality assessment*. Technics publications.
Mazón, J.-N. *et al.* (2012) 'Open Business Intelligence: On the Importance of Data Quality Awareness in User-friendly Data Mining', in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. New York, NY, USA: ACM (EDBT-ICDT '12), pp. 144–147. doi: 10.1145/2320765.2320812.
McQueen, B. (2017) *Big Data Analytics for Connected Vehicles and Smart Cities*. Artech House.
Megler, V. M., Tufte, K. and Maier, D. (2016) 'Improving data quality in intelligent transportation systems', *arXiv preprint arXiv:1602.03100*.
Micic, N. *et al.* (2017) 'Towards a Data Quality Framework for Heterogeneous Data', in *2017*

IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 155–162. doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.28.

Min, H., Scheuermann, P. and Heo, J. (2013) 'A Hybrid Approach for Improving the Data Quality of Mobile Phone Sensing.', *International Journal of Distributed Sensor Networks*, pp. 1–10. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=95291152&site=ehost-live.

Miraz, M. H. *et al.* (2015) 'A review on Internet of Things (IoT), Internet of everything (IoE) and Internet of nano things (IoNT)', in *2015 Internet Technologies and Applications (ITA)*, pp. 219–224.

Mjolsness, E. and DeCoste, D. (2001) 'Machine learning for science: state of the art and future prospects', *science*. American Association for the Advancement of Science, 293(5537), pp. 2051–2055.

Mohammed, F. *et al.* (2020) 'A Framework for Measuring IoT Data Quality Based on Freshness Metrics', in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 1242–1249.

Mostefaoui, A. *et al.* (2022) 'Big data architecture for connected vehicles: Feedback and application examples from an automotive group', *Future Generation Computer Systems*. Elsevier, 134, pp. 374–387.

Noreen, K. *et al.* (2020) 'SWOC analysis of e-learning educational services at Rawalpindi Medical University in the midst of COVID-19', *Journal of Rawalpindi Medical College*, 24(Supp-1), pp. 37–43.

Okafor, N. U., Alghorani, Y. and Delaney, D. T. (2020) 'Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach', *ICT Express*. Elsevier, 6(3), pp. 220–228.

Olufowobi, H. *et al.* (2016) 'Data provenance model for Internet of Things (IoT) systems', in *International Conference on Service-Oriented Computing*, pp. 85–91.

Perez-Castillo, R *et al.* (2018) 'DAQUA-MASS: An ISO 8000-61 based data quality management methodology for sensor data', *Sensors (Switzerland)*, 18(9). doi: 10.3390/s18093105.

Perez-Castillo, Ricardo *et al.* (2018) 'Data Quality Best Practices in IoT Environments', in *2018 11th International Conference on the Quality of Information and Communications*

Technology (QUATIC), pp. 272–275.
Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002) 'Data quality assessment', *Communications of the ACM*. ACM, 45(4), pp. 211–218.
Prathiba, B., Sankar, K. J. and Sumalatha, V. (2016) 'Enhancing the data quality in wireless sensor networks — A review', in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pp. 448–454. doi: 10.1109/ICACDOT.2016.7877626.
Pyzdek, T. and Keller, P. (2014) *Six sigma handbook*. McGraw-Hill Education.
Qian, L. *et al.* (2009) 'Cloud computing: An overview', in *IEEE international conference on cloud computing*, pp. 626–631.
Rahman, A., Smith, D. V and Timms, G. (2013) 'A novel machine learning approach toward quality assessment of sensor data', *IEEE Sensors Journal*. IEEE, 14(4), pp. 1035–1047.
Rasta, K., Nguyen, T. H. and Prinz, A. (2013) 'A framework for data quality handling in enterprise service bus', in *Third International Conference on Innovative Computing Technology (INTECH 2013)*, pp. 491–497. doi: 10.1109/INTECH.2013.6653640.
Redman, T. C. (1998) 'The impact of poor data quality on the typical enterprise', *Communications of the ACM*. ACM, 41(2), pp. 79–82.
Redman, T. C. (2017) 'Seizing opportunity in data quality', *MIT Sloan Management Review. https://sloanreview. mit. edu/article/seizing-opportunity-in-data-quality. November*, 29.
Rexeis, M., Röck, M. and Hausberger, S. (2018) *Comparison of fuel consumption and emissions for representative heavy-duty vehicles in Europe*.
Robinson, S. *et al.* (2014) 'Methods for pre-processing smartcard data to improve data quality', *Transportation Research Part C: Emerging Technologies*, 49, pp. 43–58. doi: 10.1016/j.trc.2014.10.006.
Rodríguez, C. C. G. and Servigne, S. (2013) 'Managing sensor data uncertainty: A data quality approach', *International Journal of Agricultural and Environmental Information Systems*, 4(1), pp. 35–54. doi: 10.4018/jaeis.2013010103.
Rosameo (2021) *Sensors-Data-about-Fuel-Consumption-in-Buses*. Available at: https://github.com/rosameo/Sensors-Data-about-Fuel-Consumption-in-Buses.
Rousseeuw, P. J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics*. Elsevier, 20, pp. 53–65.
Sarfi, R. J. *et al.* (2012) 'Data quality as it relates to asset management', in *PES T D 2012*, pp. 1–5. doi: 10.1109/TDC.2012.6281418.
Sebastian-Coleman, L. (2010) 'Data Quality Assessment Framework', in *The 4th MIT Info.*

Qual. Ind. Symp.
Sebastian-Coleman, L. (2012) *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes.
Sethi, P. and Sarangi, S. R. (2017) 'Internet of things: architectures, protocols, and applications', *Journal of Electrical and Computer Engineering*. Hindawi, 2017.
Shardt, Y. A. W., Yang, X. and Ding, S. X. (2016) 'Quantisation and data quality: Implications for system identification', *Journal of Process Control*, 40, pp. 13–23. doi: https://doi.org/10.1016/j.jprocont.2016.01.007.
Shi, W. *et al.* (2015) 'Improving Power Grid Monitoring Data Quality: An Efficient Machine Learning Framework for Missing Data Prediction', in *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, pp. 417–422. doi: 10.1109/HPCC-CSS-ICESS.2015.16.
Shrivastava, S. *et al.* (2019) 'DQA: Scalable, Automated and Interactive Data Quality Advisor', in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 2913–2922.
Siegel, J. E., Erb, D. C. and Sarma, S. E. (2017) 'A survey of the connected vehicle landscape—Architectures, enabling technologies, applications, and development areas', *IEEE Transactions on Intelligent Transportation Systems*. IEEE, 19(8), pp. 2391–2406.
Smith, D. *et al.* (2012) 'A Bayesian Framework for the Automated Online Assessment of Sensor Data Quality.', *Sensors (14248220)*, 12(7), pp. 9476–9501. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=79279157&site=ehost-live.
Sokolova, M., Japkowicz, N. and Szpakowicz, S. (2006) 'Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation', in *Australasian joint conference on artificial intelligence*, pp. 1015–1021.
Solomakhina, N. *et al.* (2014) 'Extending statistical data quality improvement with explicit domain models', in *2014 12th IEEE International Conference on Industrial Informatics (INDIN)*, pp. 720–725. doi: 10.1109/INDIN.2014.6945602.
Spruit, M., Linden, V. van der and others (2019) 'BIDQI: The business impacts of data quality interdependencies model', *Technical Report Series*. UU BETA ICS Departement Informatica, (UU-CS-2019-001).
Standardization, I. O. for (2022) *ISO 8000-1:2018 Data quality – Part 1: Overview and general principles*. Available at: https://www.iso.org/standard/81745.html.

Strong, D. M., Lee, Y. W. and Wang, R. Y. (1997) ‘Data quality in context’, *Communications of the ACM*. ACM, 40(5), pp. 103–110.
Sundararaman, A. and Venkatesan, S. K. (2017) ‘Data Quality Improvement Through OODA Methodology.’, in *ICIQ*.
Suresh, P. *et al.* (2014) ‘A state of the art review on the Internet of Things (IoT) history, technology and fields of deployment’, in *Science Engineering and Management Research (ICSEMR), 2014 International Conference on*, pp. 1–8.
Tsai, F.-K. *et al.* (2019) ‘Sensor abnormal detection and recovery using machine learning for IoT sensing systems’, in *2019 IEEE 6th international conference on industrial engineering and applications (ICIEA)*, pp. 501–505.
Vaishnavi, V., Kuechler, W. and Petter, S. (2004) ‘Design science research in information systems’, *January*, 20, p. 2004.
Vasta, R. *et al.* (2017) ‘Outlier Detection for Sensor Systems (ODSS): A MATLAB Macro for Evaluating Microphone Sensor Data Quality.’, *Sensors (14248220)*, 17(10), pp. 1–14. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=125917379&site=ehost-live.
Vaziri, R., Mohsenzadeh, M. and Habibi, J. (2016) ‘TBDQ: A pragmatic task-based method to data quality assessment and improvement’, *PLoS One*. Public Library of Science San Francisco, CA USA, 11(5), p. e0154508.
Veiga, A. K. *et al.* (2017) ‘A conceptual framework for quality assessment and management of biodiversity data’, *PloS one*. Public Library of Science San Francisco, CA USA, 12(6), p. e0178731.
Wagstaff, K. (2012) ‘Machine learning that matters’, *arXiv preprint arXiv:1206.4656*.
Wand, Y. and Wang, R. Y. (1996) ‘Anchoring data quality dimensions in ontological foundations’, *Communications of the ACM*. ACM, 39(11), pp. 86–95.
Wang, R. Y. (1998) ‘A product perspective on total data quality management’, *Communications of the ACM*. Association for Computing Machinery, Inc., 41(2), pp. 58–66.
Wang, R. Y. and Strong, D. M. (1996) ‘Beyond accuracy: What data quality means to data consumers’, *Journal of management information systems*. Taylor & Francis, 12(4), pp. 5–33.
Wang, Yuzhi *et al.* (2017) ‘A deep learning approach for blind drift calibration of sensor networks’, *IEEE Sensors Journal*. IEEE, 17(13), pp. 4158–4171.
Wende, K. (2007) ‘A model for data governance--Organising accountabilities for data quality management’.

Westerhuis, J. A. *et al.* (2008) 'Assessment of PLSDA cross validation', *Metabolomics*. Springer, 4, pp. 81–89.
Widiarta, H., Viswanathan, S. and Piplani, R. (2008) 'Forecasting item-level demands: an analytical evaluation of top--down versus bottom--up forecasting in a production-planning framework', *IMA Journal of Management Mathematics*. OUP, 19(2), pp. 207–218.
Yang, D. *et al.* (2018) 'Intelligent and connected vehicles: Current status and future perspectives', *Science China Technological Sciences*. Springer, 61(10), pp. 1446–1471.
Yu, R. *et al.* (2014) 'Improving data quality with an accumulated reputation model in participatory sensing systems', *Sensors (Switzerland)*, 14(3), pp. 5573–5594. doi: 10.3390/s140305573.
Zhang, C. *et al.* (2019) 'A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data', in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1409–1416.
Zhang, P. *et al.* (2017) 'Data quality in big data processing: Issues, solutions and open problems', in *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 1–7. doi: 10.1109/UIC-ATC.2017.8397554.
Zhou, Y. and Bridgelall, R. (2020) 'Review of usage of real-world connected vehicle data', *Transportation Research Record*. SAGE Publications Sage CA: Los Angeles, CA, 2674(10), pp. 939–950.

Appendix A: Selected articles assessment matrix of the systematic literature review

Article	**Quality assessment criteria**									
	1-DQ issues	**2-Currency**	**3-Complexity**	**4=Data**	**5=Method**	**6=Effectiveness**	**7=real data?**	**8=quantified?**	**9=Generalizeable?**	**10=Limitation?**
(Zhang *et al.*, 2019) Score=11	Consistency, Validity **2**	**2**	Two data sources are used **2**	Sensor data **1**	Case study **1**	Reported that inconsistency is minimized and the anomaly is detected **1**	Synthetic and real data were used. **2**	Better performance compared to similar methods reported. **2**	'Can be applied to any time series data' stated. **1**	Limitations not discussed. **0**
(Dai, Yoshigoe and Parsley, 2018) Score=10	Validity 1	2	One data source 1	Open data 1	Case Study 1	Evidence presented 2	It only used open data for research. **1**	NA 0	It can be generalized	Inability to test/apply it to real-world data 1
Casado-vara, R., Prieto-castrillo, F. and Corchado, J. M. (2018) Score=13	Validity 1	1	Indoor surface sensors and assumes a static system 2	Temperature sensor data and meta data 2	Case study 1	Mentioned energy efficiency improvement 1	Experimental data 1	Stated 'noise was reduced from 15% to 5%' 2	'Can be applied to any sensor data' 1	Many assumptions used 1
(Cerqueira *et al.*, 2018) Score=15	Accuracy, Consistency, Completeness 3	2	Transport sector High volume 2	Two taxi data sets from the US cities of	Case study 1	A comparison to show effectiveness is provided. 2	Simulated noise added 1	No quantified measure. 0	'Can be applied to more domains' stated. 1	Inability to test/apply it to real-world data 1

				Nanjing and San Francisco Taxi Transport) 2						
(Cheng *et al.*, 2018) Score=16	Accuracy, Completeness and timeliness 3	2	High volume 54 Mica2Dot sensor nodes 3	Sensor data, meta data 2	Case study 1	Reliability increased. No comparison. 1	Simulated data used 1	Completeness raised to 90.21% and correctness to 84.79% 2	'Can be applied to any sensor data' stated. 1	No limitation was discussed. 0
(Karkouch *et al.*, 2018) Score=14	Completeness, Validity, Consistency 3	3	Not applied to a specific domain. 1	UCI Repository data 0	Combines case study and survey 1	Compared case study result to survey result. 2	No real-life data. Data from UCI repository. 1	'Mean Absolute Error' was used as a measure but no quantified result was given. 1	'Can be applied for social media and phone data' stated. 1	'Limited comparisons are done' is mentioned as a limitation. 1
(Fekade *et al.*, 2018) Score=13	Completeness 1	2	Sensors (from 54 Mica dot sensors) 1	Sensor data and device data 2	Case study 1	'Performs better than KNN and NN' 2	It only used simulated data 1	An accuracy increase of 10% is reported compared to the next best method	Generalizability is not mentioned or implied. 0	'Only proximity as a main parameter for clustering' 1

								explored. 2		
(Cai and Zhu, 2015) Score=8	Consistency, Validity 2	2	Sensors ,Autonomous vehicles 2	Survey result 0	Survey 1	NA 0	NA 0	NA 0	Generalizability is not mentioned or implied. 0	'study not a rigorous' mentioned as limitation 1
(Megler, Tufte and Maier, 2016) Score=11	Validity 1	2	Intelligent Transport 3	Traffic Sensor data 1	1	Comparative result provided 2	Tested on real-life data 2	A quantified measure is given. 2	Do not indicate generalizability 0	no limitation stated 0
(Olufowobi *et al.*, 2016) Score=12	Completeness, Accuracy 2	2	Sensor 1	Temperature sensor data, metadata 2	Case study 1	'Data quality improved' 1	Simulated data is used. 1	No explicit measure is given 0	'To any sensor. Platform independent' 1	'Lack of flexibility of the approach' 1
(Min, Scheuermann and Heo, 2013) Score=11	Completeness 1	1	Mobile Phone sensing 1	Temperature sensor data, metadata 2	Case study 1	'Data quality improved by estimating missing data' 1	Simulated sensor data from Berkeley Intel lab 1	10% improvement in accuracy 2	'It is specific to mobile sensor' 0	'Limited data set' 1
(R Perez-Castillo *et al.*, 2018) Score=10	Validity, Completeness, Timeliness 3	3	Smart connected platform 2	NA (no experiment is done) 0	NA 0	NA 0	NA 0	NA 0	'Framework can be applied in any kind of data' 1	'No empirical evidence' 1

(Wang *et al.*, 2017) Score=16	Validity, Completeness 2	2	Environmental sensor network 2	Environmental sensor data, meta data 2	Case study 1	'Missing and erroneous readings replaced by more accurate values' 1	Only simulated noise was introduced 1	'Root mean error was used and the method results in the lowest' 3	'It can be applied to any WSN' 1	'Limited dataset' 1
(Robinson *et al.*, 2014) Score =15	Completeness, Validity 2	2	The transport sector, high volume 2	Smart card reading from transport (sensor data, device data, and business data) 3	Case study 1	'Missing data and inaccurate data are improved' 1	Real-life data from Singapore transport is used. 2	The missing data was brought down from 7% to 0.7%. 3	'Applicable only for transport system' stated 0	No limitation described 0
(Rodríguez and Servigne, 2013) Score=17	Completeness, validity, Accuracy 3	2	An environmental monitoring system, Sensor 3	Volcano activity monitoring sensor data, meta data 3	Case study 1	An interface is provided for the user with dashboard to decide. 1	Real-life volcano activity monitoring data. 1	No quantified measure was given. 0	Can be applied to any sensor data (model-based). 1	Did not mention limitations. 0

(Shi *et al.*, 2015) Score=17	Completeness 1	2	Power grid monitoring, Sensor 3	Sensor data, device data, business data 3	Case study 1	'Missing data was efficiently predicted' 1	No real-life data is used (only sample) 1	Mean square error was used and resulted in the lowest 0.0021. 3	'Can be applied to any data' 1	'Limited data set (only few samples were used)' stated 1
(Smith *et al.*, 2012) Score=15	Completeness, Validity 2	2	Marine Analysis Network 2	Marine sensor data, meta data 2	Case study 1	'Missing data and outliers were removed' 1	Tasmanian marine network system (yes) 2	Accuracy (34% improvement) 3	'Limited to continuous (in terms of time) data' 0	Does not state limitations. 0
(Solomakhina *et al.*, 2014) Score=19	Validity, Completeness, Consistency 3	2	Industrial environment, high volume 3	Industrial sensor data (compressor pressure), meta data, business data 3	Case study 1	'Outlier was smoothed and missing data were replaced' 1	Tested on actual turbine measurement data 2	Accuracy measure is used and 99.8% is reported. 2	'Can be applied to more systems' 1	'Inability to work with streaming data and long run time of the method' 1
(Shrivastava *et al.*, 2019) Score=12	Validity 1	2	IoT sensors 2	drilling sensor data, meta data	Case study 1	'Anomalous data was detected 1	Real data. 2	No quantified measure was given. 0	'The framework can be applied to any domain' 1	No limitation is described 0

				2						
(Yu *et al.*, 2014) Score=11	Validity 1	2	Generic sensor 2	Simulated sensor data 1	Case study 1	'Outlier was detected and replaced' 1	Only simulated data 1	No quantified measure was given. 0	'Model-based framework applicable to any system' 1	'Not tested on real data' stated as a limitation 1
(Zhang *et al.*, 2017) Score=15	Validity, Completeness, Consistency 3	2	Social media, IoT 3	Public dataset 0	Case study 1	'Anomaly was detected and replaced' 1	No real data was used 1	Accuracy was used as a measure and 82% accuracy was reported. 2	'Can be applied to any big data domain' 1	'The approach is not tested rigorously' 1