A Comparative Study of Sales Prediction Using Machine Learning Models: Integration of PySpark and Power BI

MhD Saeed Sharif
Intelligent Technologies Research Group, School of Architecture, Computing and Engineering, UEL
London, United Kingdom
s.sharif@uel.ac.uk

Madhav Theeng Tamang
Intelligent Technologies Research Group, School of Architecture, Computing and Engineering, UEL
London, United Kingdom
u1430774@uel.ac.uk

Anup Nepal
Intelligent Technologies Research Group, School of Architecture, Computing and Engineering, UEL
London, United Kingdom
u2533745@uel.ac.uk

Wael Elmedany
College of Information Technology, University of Bahrain
Kingdom of Bahrain
welmedany@uob.edu.bh

***Abstract*— Retail management requires accurate sales forecasting for strategic planning, inventory control, and revenue maximisation. In this research, we are predicting sales using PySpark-built ML models and the Big Mart dataset. We evaluate and demonstrate the prediction skills of numerous machine learning algorithms, concentrating on XGBoost. Big Mart offers item weight, visibility, type, and outlet data. We use these properties as prediction model features. PySpark, a strong distributed computing platform, manages massive datasets for analysis and model training. In massive trials, we test decision trees (DT), XGBoost, linear regressions (LR), and random forests (RF). Training and testing enhance model accuracy. RMSE and R-squared measure our model quality. Our metrics evaluate the model's data fit and prediction accuracy. XGBoost performed best with an RMSE of 1081 and an R-squared of 0.59. The XGBoost algorithm accurately predicts Big Mart sales. The model performs well because of its ensemble learning and understanding of intricate dataset links. We also used Power BI to present analytical insights, helping decision-makers design sales-estimated business plans. This study employed several machine learning algorithms, XGBoost gave the best performance. This research provides insights into how organisations might use these technologies to improve resource allocation and inventory management decision-making.**

Keywords*— *sales prediction, machine learning, pyspark, xgboost, power bi, predictive analytics, decision tree, linear regression, random forest

I. INTRODUCTION

Businesses in many sectors require reliable sales forecasting. Predicting sales helps companies allocate resources, make informed decisions, and find development prospects [1]. Due to data availability, companies use machine learning to enhance sales forecasting. Accurate sales forecasting improves inventory management. Businesses may prevent shortages and overstock by accurately estimating sales. Customer happiness and loyalty rise with lower storage costs and item availability to match demand. Effective sales forecasting helps companies allocate resources. Production, staffing, and marketing methods may change based on expected sales volume. Financial resources, supply chain procedures, and human capital are optimised, saving money and improving performance. Effective sales forecasting improves operations and finds growth possibilities. Analysing sales trends and patterns may help companies respond to changing market demand and customer behaviour. An organisation can increase b production and marketing to achieve a high market share if its prediction meets demand. Companies have nowadays started using machine learning algorithms to predict and determine their sales forecasting [2]. These ML algorithms can go through large amounts of data, and filter out trends, making more accurate results than previous traditional methods. Big data analytics for sales forecasting is beneficial [3]. Previous sales data help companies create, learn from previously done work and implement that information in decision-making. Businesses can identify underperforming stores or products and change their strategy. It is a must thing for Companies to predict sales well to be stable in today's competitive market.

Looking into the present years, predictive analytics has grown into the most advanced form in the case of data studying and advanced analytical tools [4]. The change in traditional sales forecasting methods like time series analysis and regression modelling using machine learning algorithms plays a game-changing role in this development. Using machine learning helps organisations capacity to manage complex datasets and make more accurate insights. The flexibility of machine learning algorithms provides development potential in comparison to traditional statistical processes [5]. The utilisation of a machine learning model helps to provide much better and more valid sales prediction results rather than traditional methodologies.

In this research, we mainly focus on the comparison of multiple machine learning (ML) algorithms for sales prediction. PySpark for model creation and Power BI for data visual representation are used for this study. We aim to compare different machine learning algorithms for analysing sales and evaluate the performance given. Here we are using the Big Mart datasets in our research study. The aim is to utilize Pyspark for ML algorithm comparisons and Power BI for visualization. To accomplish this project we use Google Colab and Pyspark for data preparation and machine learning models, and power bi for visualisation. The combination of these three tools will provide efficient big data predictive

analysis. This study illustrates an innovative integration of PySpark for scalable machine learning with Power BI for dynamic visualisations. The scalability of PySpark enables the real-time administration of large datasets, an essential aspect for organisations handling big data. This integration of technology allows businesses to easily make data-driven decisions, beyond traditional sales forecasting methods.

Regardless of advancements in predictive analytics and machine learning, companies are facing issues to effectively recognise sales patterns. Traditional approaches may not handle the large amounts of data which is required for meaningful insights. Additionally, the complexity of modern corporate scenarios, fluctuating market and customer behaviour, makes predictions more tough and challenging. For sure, creative and complex prediction techniques are necessary to properly handle this challenge.

Machine learning is an area of computer science and AI, aiming to represent the way of human learning behaviour [6]. Consistency in data processing and giving meaningful outcomes is the major focus of advanced ML. Machine learning model input and create output without predetermined instructions. Machine learning algorithms can manage big datasets and discover patterns helping it become more powerful in the area of prediction analysis [7]. Machine learning algorithms like linear regression (LR), ensemble techniques, and deep learning, are being used for complex sales identification. These models use powerful algorithms to assess sales data, identify patterns, and provide insights. Forecasting goes beyond company growth. Researchers have created prediction models for weather, stock prices, market trends, and patient health using machine learning and statistical analysis [8].

II. Related Work

A data scientist, Michael Crown used time series forecasting and non-seasonal Autoregressive Integrated Moving Average (ARIMA) models to make predictions [9]. ARIMA modelling provided weekly estimates for one year utilising 2.75 years of Sales dataset from store sales, weekly sales, department, date, and holiday data as attributes. Moreover, these models need stable data, typically unattainable in sales datasets due to seasonality, trends, and external variables. While seasonal differencing has been used to alleviate restrictions, the inflexibility of standard time series models may hamper accurate sales trend prediction in complex corporate organisations. To enhance sales forecasting accuracy in dynamic and multidimensional corporate contexts, innovative machine-learning approaches must be explored to manage subtle nuances in sales data.

Chouskey et al. created a weather forecasting system that provides notifications for uncertain weather conditions, assisting people and businesses in preparing for it [10]. MapReduce and Spark were used to create models and collect data from several weather sensors. All human activity depends on weather predictions. The authors use temperature, humidity, pressure, and wind speed to improve their forecasts.

Big Mart may be able to better understand its customers by using machine learning to analyse consumer data and activity. These include anticipating future buying patterns, identifying client groups based on demographic and purchase information, and making item recommendations based on past purchases. Big Mart may gain valuable insights from integrating machine learning into its sales processes, which would empower it to make more informed choices and improve operational efficiency [11].

By analysing the design and management techniques that make it possible to estimate sales based on user activity, Yuan et al. concentrate on maximising the sales performance of e-commerce [12]. They unveiled a methodology that tries to find the best possible relationship between customer preferences and product choices while also forecasting current sales.

One of the main participants in large-scale data processing is PySpark, the Python API for Apache Spark [13]. The program's scalability, efficiency, and user-friendliness have made it a popular choice. It is renowned for its capacity to handle massive datasets on distributed computer clusters [14]. PySpark, which runs concurrently and manages data ranging from terabytes to petabytes, is built on Apache Spark [15]. PySpark's seamless interaction with Apache Beam was emphasised by Akidau et al. [16], who also emphasised PySpark's use in distributed data processing tasks and Beam's flexibility. Studies comparing cloud-based platforms like Google Dataflow and Microsoft Azure Databricks have been conducted, highlighting the unique characteristics and trade-offs of each framework. Resilient Distributed Datasets (RDDs) are a set of coherent programming abstractions provided by the Spark programming paradigm [17]. High-level operations are used to perform computations on RDDs without having to worry about underlying processes like workload allocation and fault tolerance. Worker nodes disseminate immutable copies of RDDs, which are processed simultaneously by the processes. Driver software is written by application developers and connects to a cluster of employees. The driver manipulates one or more RDDs after defining them.

Data visualisation is very important for today’s data-driven world, it makes people and organisations understand complex information and retrieve the gist of information from the visual [18]. Microsoft Power BI is the most popular tool that creates interactive and informational visualisations. Power BI has a user-friendly interface and several visualisation options for data-driven decision-making in multiple areas [19]. The software gives users the freedom to create interactive dashboards, analyse data, and apply several filters, helping stakeholders identify crucial insights. Power BI’s visual data presentation helps in making perfect business decision-making [19]. The use of Power BI simplifies complex information, making its stakeholders quickly understand and work on key findings. Metrics have been created by Yitong Liu and Xi Chen to measure variables including sales accumulation, forecast accuracy, and forecast accumulation [20]. Create visual charts using Power BI to examine the difference between the forecasted and actual sales and analyze the accuracy rate trend of sales estimates. This will help businesses adopt budget management more methodically and efficiently.

III. Methodology

In the study, sales are predicted using machine learning. For any task, gathering and analysing data is necessary. Achieving objectives requires choosing the right data processing methods and instruments. Regression analysis

and classification are essential for sales forecasting. Machine learning models were used to predict sales factors that are included in the dataset. Machine learning regression algorithms, like LR, DT, RF and XGBoost, were used in this project to identify the best predictive model for sales analysis.

A. Data description

We obtained data from the Kaggle repository [21]. The datasets consist of 15000 sales-related data, including the factors that are directly related to the sales prediction. In this dataset, we have 7 categorical and 5 numerical columns. Information about item qualities such as fat content, type, Maximum price, visibility, weight and store types and locations is mentioned in the dataset. In general, the dataset has the proper sales information related to the food stores and supermarkets. Using machine learning methods like ensemble or regression organisation might increase their income and sales ratio.

B. Data processing

To stay competitive and make smart decisions about their sales and marketing organisation must have to do the best research. They need to get sales information from a lot of different resources and make it clean to get the best output. The dataset we used has information about past sales, items (like weight, fat content, type, and MRP), and stores (IDs, established years, sizes, locations, and types). A lot of work goes into preparing data to make sure it is correct and reliable. The preparation step is very important for keeping the data clean and ready to be analysed. As part of preprocessing, missing data is fixed, outliers are found, category variables are turned into numbers, and numerical features are standardised so that scales are consistent. Organisations can order and standardise datasets so they are easier to use and analyse by cleaning and preparing them in a lot of detail. Cleaning and preparing data thoroughly makes it possible to find insights and trends that could help with sales and marketing choices. Comparative analysis helps businesses figure out how well their sales are doing, spot trends, and make smart marketing decisions based on data. A careful look at sales data might help businesses find places to improve, make their marketing more effective, and increase their growth and profits.

Once the data cleaning and normalization were completed, the suitable machine learning algorithms were selected. The main dataset was divided into training and testing sets, the distribution of training and testing data may impact the model’s performance. The proportion for training is 80% and the remaining 20% is for testing.

Machine learning enhances computer performance by letting computers learn from data and predict without scripting. Several practical methods to improve computer abilities are in this area. Supervised learning, which trains computers to anticipate new data using labelled datasets, is essential to machine learning. In predictive analytics, regression and classification are popular methods for continuous and categorical variables data respectively [22]. In contrast, unsupervised learning finds patterns or structures in unlabelled data to expose its basic relationships. Semi-supervised learning uses tagged and unlabelled data. Reinforcement learning, an important machine learning topic, teaches agents to maximise rewards via environmental interaction. This strategy is useful when the algorithm learns from testing and alters its behaviour. Machine learning trains prediction models using LR, DT, RF and XGBoost. Machine learning requires algorithms to help computers make data-driven judgements and predictions.

The real-world situations should be reflected while evaluating the model's effectiveness. Evaluating prediction performance is the most do thing while evaluating machine learning models. Regression metrics like R-squared, MAE, MSE, and RMSE are used for comparison. These metrics extract information about the precision and predictive capability of the model. To overcome the overfitting in module holdout and cross-validation can be considered. Various performance analyses like feature significance, and model interpretation are also used by the researchers to get a deeper understanding of the model. These are done to know about the predictive power of models according to the data type used.

C. Integration with Power BI and Visualization

Power BI and machine learning models can make sales forecasts much more accurate, which helps businesses make better choices. Companies can more accurately predict sales trends and patterns when they use PySpark to train models and share them in PMML or CSV forms that work with Power BI. These models can be easily added to Power BI apps and reports, giving users useful information about what the future holds.

Decision-makers need to be able to interact with visible images of data. Key performance indicators (KPIs) are summed up in Power BI's dynamic graphs to give a full picture of sales success. For instance, scatter plots help people understand how different factors affect sales by showing how different elements are related. People with access to time series plots can see trends and predict sales over time.

Power BI's powerful tools and easy-to-use interface make it possible for businesses to get deep insights from their sales data. You can use these lessons to help your business grow, set smart goals, and be more successful overall. Using the machine learning models in Power BI can help you make better predictions and decisions. Power BI's analysis and visualisation tools can give companies a competitive edge in today's fast-paced business world by helping them look at data, find trends, and make accurate predictions about future sales. The Figure 1. illustrates the different stages of this study.

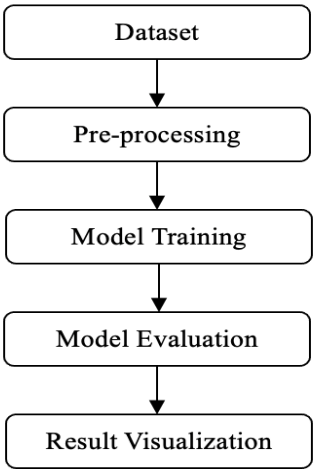

Fig. 1. Stages of this study.

IV. IMPLEMENTATION

Going through supermarket data with machine learning, helps businesses estimate their future sales and recognise the pattern and also opens space for data-driven decisions.

A. Exploratory Data Analysis

A deep understanding of the dataset plays an important role in accurate prediction. Extracting hidden patterns or trends in data is challenging that's why exploratory data analysis is necessary. Thorough analysis is essential for the structure of the dataset and a new understanding of the soundness study. For this study, a data analysis is required for features that are important for understanding and forecasting. In this process of our exploratory data analysis, we have created several visualisations to help with feature selection.

Exploratory data analysis is used to find missing values, which is considered as first step in making sure the information is reliable and correct. This helps analysts to learn valuable output about the quality information by finding missing values. If we don't account for lost data, it could cause studies to be biased, which could lead to wrong results or models. If we don't handle missing variables correctly in machine learning, it can slow down the model and make it harder to apply to new data. Finding missing values helps us make better decisions by making it clear what the limits of the dataset are and making sure that the data insights you get are correct and fair. Finding missing numbers is important for making sure that data is correct, for research, and for making smart business choices.

Mostly, missing values are handled by finding the mean, mode, or median for numerical and categorical variables to find the most frequent value. Another option is to remove columns with missing values. We used mean and mode for numerical variables. We adopted a different approach to obtain the mode value to address missing data. The mode is a statistical indicator that indicates the most common value in a collection. In data imputation, restoring missing values with the mode includes replacing them with the most common value in the column. The most typical approach for managing missing values is to compute the mode.

B. Converting Categorical Columns into Numerical Columns

Most machine-learning applications need categorical variables to be converted to numerical representations for better training and prediction. Categorical data encoding is essential since most machine learning methods need numerical input characteristics. In this research, PySpark is employed to convert category columns to numerical columns, it used String Indexer which assigns an integer index to each distinct categorical column category. It assigns numerical values to each category. Categorical column names and values are translated to index columns using the Pyspark string indexer. For instance, low fat and regular fat item content values are transformed to 1 and 0, whereas outlet type, item types, and other categorical data are converted to integers.

C. Correlation Matrix

Correlation matrices are statistical tools that show the link between variables in a dataset. The correlation between variables in the table helps identify those with a stronger link. Examining the correlations between the several factors in this dataset was vital. It helps to find the correlation employing Pearson's Correlation Coefficient, measuring linear associations between variables from -1 to +1. Greater correlation values signify a better linear relationship between variables.

D. Application of Machine Learning algorithms

1) Linear Regression Model

Using this algorithm, we are finding the connection between predictor variables (features) and the target variable, "ItemOutletSales." This algorithm assumes a linear connection between the target and predictor variables. Several characteristics, including weight, fat content, and visibility, affect the prognosis for item outlet sales. Linear regression models aim to reduce the gap between anticipated and actual values by finding the best-fitting line or hyperplane. The trained model may utilise gained coefficients to forecast new data by assessing observation properties. Root Mean Squared Error (RMSE) is used to compare the difference between expected and actual target variable values and RMSE also helps to evaluate the model performance. The RMSE result is presented below after data processing.

2) Decision Tree Model

Using a decision tree and the dataset, outlet sales are predicted. Weight, fat content, visibility, kind, MRP, outlet type, establishment year, size, location type, and outlet type are some of the elements taken into account in this process. The decision tree method enhances target variable similarity within each group by breaking the dataset into smaller groups according to characteristic values. The method, which often makes use of measures like Gini impurity or information gain, determines which attribute at each node is most useful for breaking up the data into smaller groups. By starting at the root node of the tree structure and following the feature values of each new observation until it reaches a leaf node, the trained decision tree may make predictions about future events.

3) Randon Forest Model

This method is a combination of several decision trees, each tree is trained on a randomly selected dataset and its characteristics. The final prediction is determined by

averaging all tree forecasts for regression tasks or by a majority vote for classification tasks. Sales of products at outlets using factors like weight, fat content, and visibility in the dataset are calculated using Random Forest.

4) XGBoost Model

XGBoost was used to forecast shop sales in this study. Variables in the dataset were weight, fat content, visibility, kind, MRP, outlet establishment year, size, and location type outlet kind. The XGBoost ensemble learning approach uses gradient boosting to generate a reliable prediction model via repeated operations. To increase model performance and avoid overfitting, XGBoost gradually incorporates decision trees and regularisation techniques.

V. Result and Discussion

Important conclusions are drawn by analysing Root Mean Square Error (RMSE) values from various machine learning models. The worst prediction accuracy among the models is shown by Linear Regression (LR), which has the highest RMSE score of 1198. LR assumes a linear connection between qualities and the target variable, which may lead to an incomplete representation of data complexity. With RMSE of 1106 and 1099, respectively, DT and RF algorithms outperformed the LR model. Even while RF and DT algorithms perform poorly in generalisation due to over-fitting, they are superior to logistic regression at capturing non-linear interactions. On the other hand, XGBoost (XGB), which had an RMSE score of 1081, fared better than other models. XGBoost's ensemble learning technique, which combines gradient boosting with decision trees to prevent overfitting and promote generalisation, is responsible for its improved prediction accuracy. For best results, these results highlight how crucial it is to choose and fine-tune machine learning models.

The model having the lowest RMSE score is best for predictive analysis representation. The RMSE score indicates the number of mistakes in our research. The accuracy of a model is determined by comparing predicted values to input values. Reducing RMSE values improves model performance overall. The results show that XGBoost, with its ensemble learning technique, more successfully finds detailed patterns in sales data compared to linear models. This highlights the model's practical utility in business, as its better performance may result in more exact inventory and personnel selections.

A. Power BI Dashboard for Visualization

The Power BI dashboard offers detailed information on Big Mart sales data and predicted numbers. Most EDA trials will be presented as narratives on this dashboard, enabling users to pick data based on their requirements. After saving the predicted weekly sales in 'cleaneddf.csv', Power BI data transformations split the id column into columns for item types, outlet sales, and date information. Figure 2. shows the sales dashboard for Big Mart data analysis.

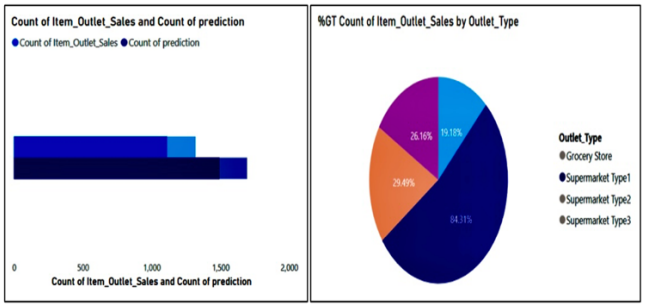

Fig. 2. Sales Dashboard for Big Mart Data Analysis.

Similarly, Figure 3. compares the average of outlet sales versus the predicted sales for different items.

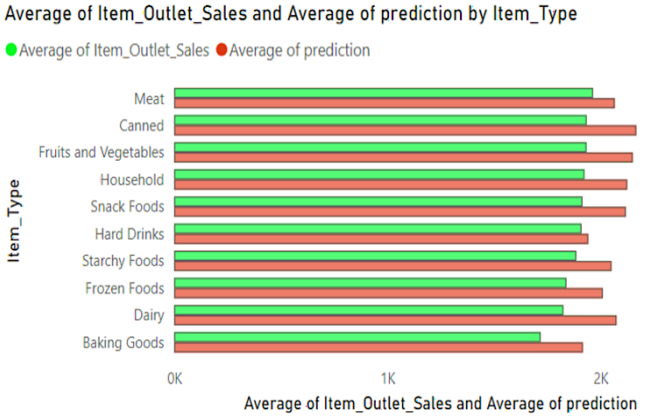

Fig. 3. Average Outlet sales vs the predicted outlet sales for different items

In this study, we used RMSE and R-squared to evaluate the performance of the algorithms as the project is done based on regression technologies. In addition, we have also worked for precision, recall and F1-Score for the Random forest and XGBoost while these are not suitable for Linear regression and Decision Tree. Table 1. below shows the performance obtained from several machine learning algorithms.

TABLE 1. PERFORMANCE OF DIFFERENT MACHINE LEARNING MODELS

Model	RMSE	R-Squared	Precision	Recall	F1-Score
Random Forest	1099	0.57	0.78	0.75	0.76
XGBoost	1081	0.59	0.81	0.79	0.80
Linear Regression	1198	0.50	-	-	-
Decision Tree	1106	0.55	-	-	-

VI. Conclusion

We compared machine learning models for retail sales prediction to derive actionable insights and support data-driven decision-making. Using a rich dataset with variables like fat content, outlet size, and location, we demonstrated that ensemble methods—particularly Random Forest and XGBoost—outperform single models in capturing complex relationships within retail data. Random Forest mitigates overfitting through collective learning, while XGBoost's

gradient boosting fine-tunes model parameters for precise predictions. The integration of PySpark for scalable model training and Power BI for dynamic data visualization transformed our forecasting approach, enabling efficient handling of large datasets and offering stakeholders intuitive, interactive insights.

This study opens multiple avenues for further investigation, including exploring additional predictive variables such as market trends and economic indicators to enhance model accuracy. Ongoing research should also examine the resilience of these models over time and in varying business environments. Finally, ethical considerations are crucial as companies adopt predictive analytics, and future research must ensure responsible data management and transparency in model interpretation. By leveraging ensemble learning and big data visualization, this study provides a robust framework for advancing sales prediction in the retail sector.

REFERENCES

[1] Bohanec, M., Robnik-Šikonja, M. and Borštnar, M.K., 2017. Organizational learning supported by machine learning models coupled with general explanation methods: A Case of B2B sales forecasting. *Organizacija*, *50*(3), pp.217-233.

[2] Cheriyan, S., Ibrahim, S., Mohanan, S. and Treesa, S., 2018, August. Intelligent sales prediction using machine learning techniques. In *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 53-58). IEEE.

[3] Hofmann, E. and Rutschmann, E., 2018. Big data analytics and demand forecasting in supply chains: a conceptual analysis. *The international journal of logistics management*, *29*(2), pp.739-766.

[4] Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. *Journal of business research*, *70*, pp.263-286.

[5] Raj Theeng Tamang, M., Sharif, M.S., Al-Bayatti, A.H., Alfakeeh, A.S. and Omar Alsayed, A., 2020. A machine-learning-based approach to predict the health impacts of commuting in large cities: Case study of London. *Symmetry*, *12*(5), p.866.

[6] Barbierato, E. and Gatti, A., 2024. The challenges of machine learning: A critical review. *Electronics*, *13*(2), p.416.

[7] Tsoumakas, G., 2019. A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, *52*(1), pp.441-447.

[8] Sharif, M.S., Tamang, M.R.T., Elmedany, W. and Fu, C.H., 2022, November. Evaluating the Stressful Commutes Using Physiological Signals and Machine Learning Techniques. In *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)* (pp. 175-180). IEEE.

[9] Michael Crown - ARIMA Models for Walmart Sales Forecasting (no date) Michael Crown. Available at: http://mxcrown.com/walmart-sales-forecasting/ (Accessed: 27 February 2024).

[10] P. Chouksey and A. S. Chauhan, "A Review of Weather Data Analytics using Big Data", International Journal of Advanced Research in Computer and Communication Engineering, 2017.

[11] Praveen, S.P., Chaitanya, P., et al. (2023) 'Big Mart Sales using Hybrid Learning Framework with Data Analysis', in 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS). 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 471–477. Available at: https://doi.org/10.1109/ICACRS58579.2023.10404941.

[12] Yuan, H., Xu, W. and Wang, M. (2014) 'Can online user behaviour improve the performance of sales prediction in E-commerce?', in 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2347–2352. Available at: https://doi.org/10.1109/SMC.2014.6974277.

[13] Shaikh, E., Mohiuddin, I., Alufaisan, Y. and Nahvi, I., 2019, November. Apache spark: A big data processing engine. In *2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)* (pp. 1-6). IEEE.

[14] Srivastava, P., 2021. A Case Study on Clustering of Datasets Using K-Means Algorithm in Spark. *Mathematical Statistician and Engineering Applications*, *70*(2), pp.1741-1750.

[15] Singh, P. (2022c) 'Manage Data with PySpark', in P. Singh (ed.) Machine Learning with PySpark: With Natural Language Processing and Recommender Systems. Berkeley, CA: Apress, 15–37. Available at: https://doi.org/10.1007/978-1-4842-7777-5-2.

[16] Akidau, T., Chernyak, S. and Lax, R., 2018. *Streaming systems: the what, where, when, and how of large-scale data processing*. " O'Reilly Media, Inc.".

[17] Zaharia, M. et al. (no date) 'Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing'.

[18] Simon, P., 2014. *The visual organization: Data visualization, big data, and the quest for better decisions*. John Wiley & Sons.

[19] Ferrari, A. and Russo, M. (2016) Introducing Microsoft Power BI. Microsoft Press.

[20] Liu, Y. and Chen, X. (2022) 'Application of Big Data Analysis Based on Power BI in Sales Forecasts', in Proceedings of the 5th International Conference on Computer Science and Software Engineering. New York, NY, USA: Association for Computing Machinery (CSSE '22), pp. 722–726. Available at: https://doi.org/10.1145/3569966.3571272.

[21] Brij (2018) *BigMart sales data*, *Kaggle*. Available at: https://www.kaggle.com/datasets/brijbhushannanda1979/bigmart-sales-data/data (Accessed: 12 February 2024).

[22] Kim, K. and Hong, J.S., 2017. A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. *Pattern Recognition Letters, 98*, pp.39-45.