A Comparative Study for Predicting House Price Based on Machine Learning

Elias Eze
Computer Science and Digital Technologies Department, ACE, UEL, London, UK
eeze@uel.ac.uk

Sujith Sujith
Computer Science and Digital Technologies Department, ACE, UEL, London, UK
u2240271@uel.ac.uk

Mhd Saeed Sharif
Computer Science and Digital Technologies Department, ACE, UEL, London, UK
s.sharif@uel.ac.uk

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

***Abstract*—This study presents the development and comparison of four machine learning approaches, namely, Random Forest, Decision Tree, Linear Regression and k-Nearest Neighbours (k-NN), to perform house prices prediction based employing the Boston Housing Dataset. The approaches performance was evaluated using root mean square error as well as R^2, with the aim of identifying the model that best predicts housing prices. The dataset was thoroughly analyzed for features, correlations, multicollinearity, and overfitting. Results indicate that the RF model outperformed the other models in predicting house prices, as it has ability to address non-linearity and complex interactions among variables and reduce the impact of outliers. The DT model also performed well but may have been more prone to overfitting. LR, on the other hand, may have been limited by its assumptions of linearity and independence among variables.**

***Keywords*—Time-series, machine learning, k-nearest neighbours (k-NN), linear regression, decision tree, random forest**

I. INTRODUCTION

The market of housing is a critical measure of a nation's economic prosperity, as it reflects both the growing population and the shifting dynamics of urbanization. As people move from rural to urban areas, the population in cities increases, which in turn leads to greater demand for housing. As demand for housing increases, so does the price of houses [1]. Furthermore, infrastructure development in a specific region can result in sudden price hikes for housing. For example, once problems such as poor road conditions and unreliable power supplies in residential areas are addressed, homeowners tend to raise house prices in that area.

Housing prices are affected by various factors that have been identified by experts. These factors include physical condition, location, and concept. Physical condition refers to observable characteristics such as the size of the house, number of rooms in the house, availability of a yard (front and back garden), and age of the property. Other physical features such as the size of the structure, number of bedrooms and bathrooms, and interior design may also influence the price. Meanwhile, concepts pertain to marketing tactics used by developers to attract potential investors, such as proximity to major roads, educational institutions, markets, hospitals, and airports [2-3]. Finally, the location of the property also significantly impacts its price, as the prevalent cost (price) of land is largely influenced by the area it is situated in.

Hence, understanding the trends of housing prices and the factors influencing them is not only essential for tenants, but it is also crucial for homeowners, analysts, policymakers in the real estate industry, as well as urban and regional planning authorities [3-4]. These stakeholders can benefit from a computerized forecasting system that can aid in making informed decisions about the acquisition of properties and the optimal timing for such transactions.

ML [5-7] is a subfield of artificial intelligence (AI) that involves creating algorithms that can learn from and make predictions based on data. In recent years, ML has become a popular tool in the field of real estate to predict housing prices [8-17]. The Boston Housing Dataset is a widely used dataset for predicting housing prices. It consists of data collected by the U.S Census Service and includes fourteen attributes related to housing, such as the number of rooms per dwelling, per capita crime rate by town, and the percentage of lower status of the population.

The primary objective of this research paper is to develop a ML prediction model for the Boston Housing Dataset, using the DT, RF, k-NN and LR models. These models will be compared, and their performance evaluated to determine which model best predicts housing prices. The evaluation of the performance of these ML models were carried out using metrics such as R-squared (R^2), and RMSE.

II. LITERATURE REVIEW

Housing price prediction is a crucial task in the field of real estate, and accurate prediction models are of great importance for both buyers and sellers. In recent years, the use of ML techniques for housing price prediction has gained significant attention due to their ability to capture complex patterns and relationships between different variables [11-17]. This literature review survey aims to provide an overview of the existing research papers on housing price prediction models using ML techniques. The review will cover the different types of ML techniques used for housing price prediction, including regression, classification, and clustering, and highlight the strengths and limitations of each approach. The review will also examine the different datasets used in these studies, the evaluation metrics employed, and the performance of the models in predicting housing prices.

Bai [17] and Sanyal et al. [18] carried out a study on Boston house price prediction using regression model that focused on the development and evaluation of several regression models for predicting house prices using the Boston house dataset. The authors employ Simple LR, Polynomial Regression (PR), Lasso Regression (LaR), and Ridge Regression (RR) to create advanced automated ML models. This study also explores the impact of various attributes of the dataset on the prediction accuracy of the models. The authors begin by highlighting the importance of predicting house prices in the real estate industry and how ML models can help in this regard. They then provide an overview of the four regression models used in the study, along with the measuring metrics used to evaluate their performance. The authors then describe the methodology used in the study, which involves pre-processing the dataset, handling outliers, and splitting the data into training and testing sets. They then train the models on the training set and evaluate their performance using several measuring metrics such as RMSE, R^2, and Cross-Validation. The authors found that Lasso Regression outperformed all other models in terms of prediction accuracy, while Simple LR performed poorly. They attributed this result to the ability of Lasso Regression to handle complex data and reduce the impact of irrelevant features on the prediction accuracy. The authors also explored the correlation between various attributes of the Boston house dataset using a heat map and found that some attributes had a strong positive or negative correlation with the house prices.

The research work in [19], the authors investigated the use of various ML algorithms such as LR, DT, and RF for predicting housing prices. The authors used a dataset from 2015 to 2019. They explore the impact of location, area, and the number of rooms on housing prices and apply traditional and advanced ML approaches to predict individual housing prices. Their results demonstrate that RF outperforms LR and DT on both training and test data. The accuracy of the predictions made by LR and DT was lower than that of RF. Additionally, Adetunji et al. [20] studied house price prediction using RF ML Technique and explored the use of the RF ML algorithm for predicting housing prices. This approach involves treating price prediction as a classification issue. This suggests that the proposed model can be useful for predicting housing prices, especially when compared to other prediction models. Overall, the paper highlights the potential of the RF algorithm for housing price prediction and suggests that this approach could be useful for many real-world applications.

The study by Henriksson and Werlinder [21] aims to compare the predictive performance of XGBoost and RF regressor models in terms of housing price prediction. The study uses two datasets and considers various evaluation metrics, including R2, RMSE, and MAPE, as well as training and inference times. The authors conduct substantial data cleaning and hyperparameter tuning to find optimal parameters for both models. The results show that XGBoost outperforms the RF model for two types of data sets. The study highlights the potential of XGBoost as an effective and efficient model for predicting housing prices, especially when compared to the RF model.

A related study by Kumar [4] presented an analysis of different ML algorithms for predicting housing prices using a data from 2015 to 2019. The study compares the performance of traditional ML algorithms such as LR, DT, and RF, as well as a more advanced method, the CNN RF. The paper highlights the importance of considering various factors such as location, area, and number of rooms while predicting individual housing prices. It aims to provide an optimistic result for housing price prediction by exploring various impacts of features on prediction methods. The study finds that all three traditional ML algorithms accomplished the desired outcomes, but they have their pros and cons.

Though these reviewed related studies have recorded some good and impressive results by applying different ML algorithms and methods, but our study aims to carry out a thorough comparative investigation of the performance capability of four difference ML methods namely - LR, DT, RF, and k-NN models, as opposed to performing house price prediction using only one, or two ML technique. The study will use a dedicated openly available Boston housing dataset to train, test, evaluate, and compare overall models' performance to determine which model best predicts housing prices. Finally, the findings of this study can be utilized in real estate applications, policy-making.

III. Model Selection

The importance of model selection lies in the fact that different models have different strengths and weaknesses. Some models may perform well on certain types of datasets while underperforming on others. Furthermore, different models have different hyperparameters that can be tuned to improve their performance on a given dataset. Choosing the wrong model or failing to optimize the hyperparameters can lead to poor predictions, decreased model performance, and adversely affect the overall results prediction accuracy. Therefore, when adopting ML model for predicting the housing prices, selecting an appropriate model is crucial to achieve accurate predictions and maximize the overall performance of the model. In this process, it is important to evaluate multiple models and select the best one based on a set of evaluation metrics, which typically includes measures such as RMSE, and R^2 score.

A. Linear Regression (LR)

LR is a simple and widely used ML algorithm for predicting a continuous target variable based on one or more independent variables. The relationship between the dependent variable Y and the independent variables X can be model as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon \quad (1)$$

where β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_n$ are the coefficients or weights that quantify the impact of each independent variable on the target variable, and ε is the error term that captures the unexplained variation in Y.

The coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_n$ can be estimated using various techniques, such as gradient descent or ordinary least squares. Once the coefficients are estimated, we can use the linear equation to make predictions on new data by plugging in the values of the independent variables:

$$Y\_pred = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \quad (2)$$

where X_1, X_2, ..., X_n are the values of the independent variables for the new data point.

In the context of Boston housing price prediction, we can use LR to model the relationship between the features of a given house (such as the crime rate in a specific neighbourhood, the average number of rooms per dwelling, and the distance to work centres) and its market value. By fitting a LR model to the training data, we can estimate the coefficients that best capture this relationship and use them to predict the median value of owner-occupied homes for new houses based on their features.

B. Decision Tree (DT)

DT regression is a ML technique that uses a decision tree as a predictive model to map the input features of a new instance to a predicted output value. In the context of Boston housing price prediction, we can use DT regression to model the non-linear relationship between the features of a given house and its market value.

The DT regression, at each node of the tree, the algorithm selects the feature and the split value that result in the highest reduction in the mean square error (MSE) of the target variable (i.e., the median value of owner-occupied homes). The process continues until the tree reaches a stopping criterion, such as minimum number of instances per leaf or maximum depth.

To predict the median value of owner-occupied homes for a new house using the DT regression model, we would start at the root of the tree and follow the path that corresponds to the values of the input features of the new instance. The final prediction would be the mean value of the training instances that belong to the leaf node reached by the new instance.

The formula used for prediction in DT regression is as follows:

$$y\_pred = \Sigma\, y_i/n \quad (3)$$

where y_pred is the predicted value of the target variable (i.e., the median value of owner-occupied homes) for a new instance, yi is the actual value of the target variable for the i-th training instance that belongs to the leaf node reached by the new instance, and n is the total number of training instances that belong to the leaf node. The predicted value is the average of the actual values of the training instances that belong to the same leaf node as the new instance.

C. Random Forest (RF)

RF regression is a ML algorithm that combines multiple DTs to predict the target variable. In the context of Boston housing price prediction, we can use RF regression to model the relationship between the features of a house and its market value.

To build a RF regression model for Boston housing price prediction, we would first split the dataset into a training set and a testing set. We would then use the training set to fit a RF model with the features as independent variables and the median value of owner-occupied homes as the dependent variable. The RF model consists of an ensemble of DTs.

The formula for the RF regression model can be expressed as:

$$y = f(X) + \varepsilon \quad (4)$$

where y is the median value of owner-occupied homes, X is a vector of features, f(X) is the RF regression function, and ε is the random error term.

To predict the median value of owner-occupied homes for a new house, we would input its features into each tree in the RF, and then take the average of the predicted values across all the trees. Finally, we can evaluate the performance of the RF regression model on the testing set by calculating metrics such as the mean squared error or the coefficient of determination (R-squared). RF regression is a powerful and flexible algorithm that can capture non-linear and interaction effects between the features and the target variable. However, it may require more computational resources and hyperparameter tuning than LR or other simpler models.

D. k-Nearest Neighbours (k-NN)

In the k-NN method the value of k, which represents the number of neighbours to consider, is a hyperparameter that can be tuned to find the optimal value for a given problem.

k-NN methodology involves several steps. First, the dataset is split into training and testing sets. The training set is used to build the k-NN model, while the testing set is used to evaluate the model's performance. Next, the distance metric is chosen, which determines how the similarity between data points is calculated. Common distance metrics used in k-NN include Euclidean distance, Manhattan distance, and cosine similarity. Then, the value of k is selected, and the k-NN model is trained. Finally, the performance of the model is evaluated on the testing set using evaluation metrics such as RMSE, and R^2. Once trained, the model can be used to predict housing prices on the testing set. The performance of the model can be evaluated using metrics such as RMSE, and R-squared to assess its accuracy in predicting housing prices in Boston. k- NN can be a useful methodology for predicting housing prices, as it can capture local patterns in the data and provide interpretable results.

IV. Results and Discussions

The housing values in Boston's suburbs are detailed in the Boston Housing dataset. The dataset has 14 features as described in Table I and its data structure in Table II.

TABLE I. BOSTON HOUSING DATASET FEATURES AND DESCRIPTION

S/N	Feature	Description
1	CRIM	Per capita crime rate by town
2	ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
3	INDUS	Proportion of non-retail business acres per town.
4	CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5	NOX	Nitric oxides concentration (parts per 10 million)
6	RM	Average number of rooms per dwelling
7	AGE	Proportion of owner-occupied units built prior to 1940
8	DIS	Weighted distances to five Boston employment centres
9	RAD	Index of accessibility to radial highways
10	TAX	Full-value property-tax rate per $10,000
11	PTRATIO	Pupil-teacher ratio by town
12	B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13	LSTAT	% lower status of the population
14	MEDV	Median value of owner-occupied homes in $1000's

The other attributes describe various neighbourhood traits in the Boston area, including the proportion of lower-class residents, crime rates, and highway accessibility. After loading the dataset into Python, the initial step was to inspect the data by displaying a sample view of the first 10 rows. This provided an immediate overview of the dataset's structure, allowing for a quick assessment of the data's format, column names, and the values within each column.

TABLE II. BOSTON HOUSING DATASET STRUCTURE

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
5	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
6	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
8	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9

A detailed exploratory analysis on the dataset derived a useful summary statistic and provided a comprehensive overview of the data's central tendencies and dispersion. Various measures such as mean, median, mode, standard deviation, and range were calculated, shedding light on the dataset's distribution and variability (see Table III).

TABLE III. BOSTON HOUSING DATASET'S DISTRIBUTION AND VARIABILITY

	count	mean	std	min	25%	50%	75%	max
CRIM	506.0	3.613524	8.601545	0.00632	0.082045	0.25651	3.677083	88.9762
ZN	506.0	11.363636	23.322453	0.00000	0.000000	0.00000	12.500000	100.0000
INDUS	506.0	11.136779	6.860353	0.46000	5.190000	9.69000	18.100000	27.7400
CHAS	506.0	0.069170	0.253994	0.00000	0.000000	0.00000	0.000000	1.0000
NOX	506.0	0.554695	0.115878	0.38500	0.449000	0.53800	0.624000	0.8710
RM	506.0	6.284634	0.702617	3.56100	5.885500	6.20850	6.623500	8.7800
AGE	506.0	68.574901	28.148861	2.90000	45.025000	77.50000	94.075000	100.0000
DIS	506.0	3.795043	2.105710	1.12960	2.100175	3.20745	5.188425	12.1265
RAD	506.0	9.549407	8.707259	1.00000	4.000000	5.00000	24.000000	24.0000
TAX	506.0	408.237154	168.537116	187.00000	279.000000	330.00000	666.000000	711.0000
PTRATIO	506.0	18.455534	2.164946	12.60000	17.400000	19.05000	20.200000	22.0000
B	506.0	356.674032	91.294864	0.32000	375.377500	391.44000	396.225000	396.9000
LSTAT	506.0	12.653063	7.141062	1.73000	6.950000	11.36000	16.955000	37.9700
MEDV	506.0	22.532806	9.197104	5.00000	17.025000	21.20000	25.000000	50.0000

Next, the boxplots (Fig. 1) of all the independent features were visualized to provide a comprehensive overview of the data distribution and identify any potential outliers. Upon analysing the boxplots in Fig. 1, it was observed that the features CRIM, ZN, and B exhibit the presence of outliers.

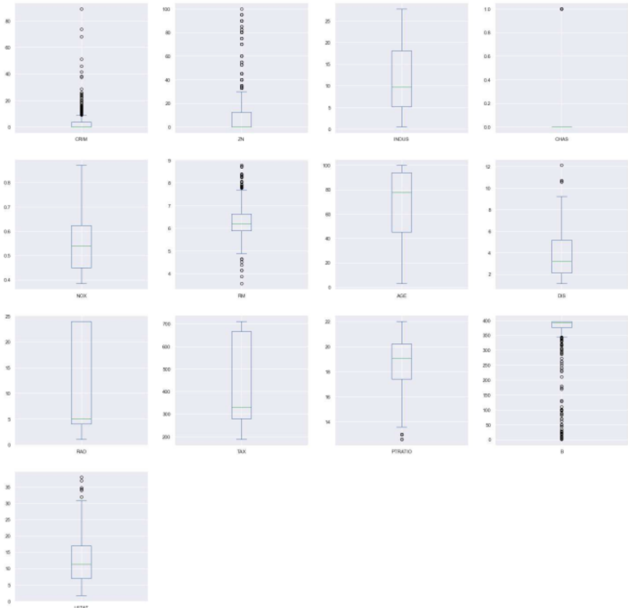

Fig. 1: Boxplots of all the independent features.

models. To address this issue, the Interquartile Range (IQR) method was employed as an effective approach for outlier detection and treatment. Any data point falls outside of this range (1.5 times the IQR) is considered an outlier. By identifying and removing these outliers, the adverse impact of outliers was mitigated on the analysis and ensured that our results are not biased by extreme values. Upon examining Fig. 2, Fig. 3, and Fig. 4 (the boxplots for the three variables), it was observed that there was a remarkable decrease in the number of outliers. This suggests that the use of the IQR method for identifying, and removing outliers was effective in this context.

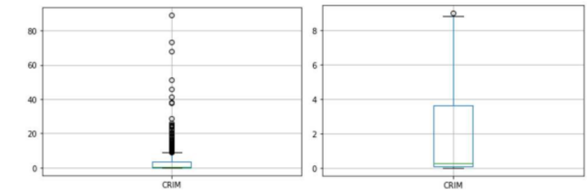

Fig. 2: Boxplots comparing the outliers in “CRIM” before and after IQR rectification

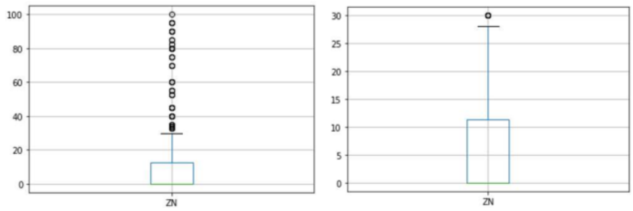

Fig. 3: Boxplots comparing the outliers in “ZN” before and after IQR rectification

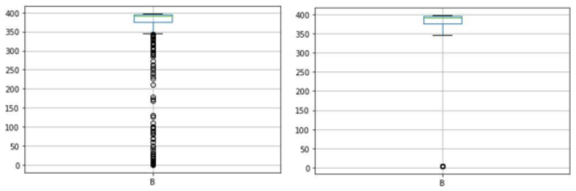

Fig. 4: Boxplots comparing the outliers in “B” before and after IQR rectification

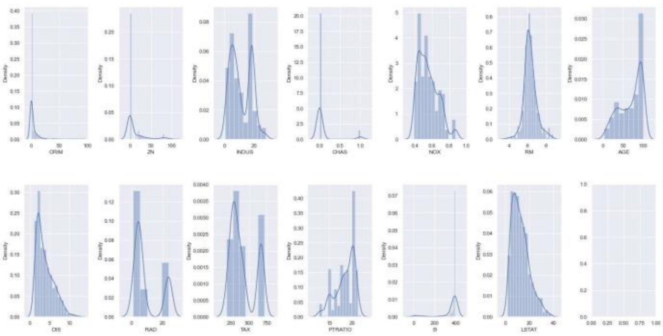

Fig. 5: Density graphs for independent variables.

From density analysis as shown in Fig. 5 and Fig. 6, it can be observed that the distribution of the "MEDV" variable, along with most of the other independent variables exhibit a relatively normal distribution. However, it is worth noting that the "ZN" and "CRIM" variables display positive skewness, as indicated by their longer tails on the right side of the distribution plots.

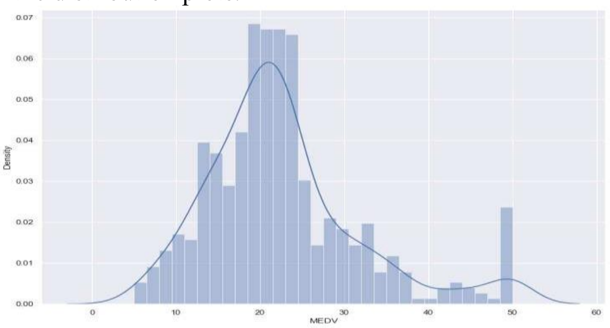

Fig. 6: Density graphs for independent variable “MEDV”.

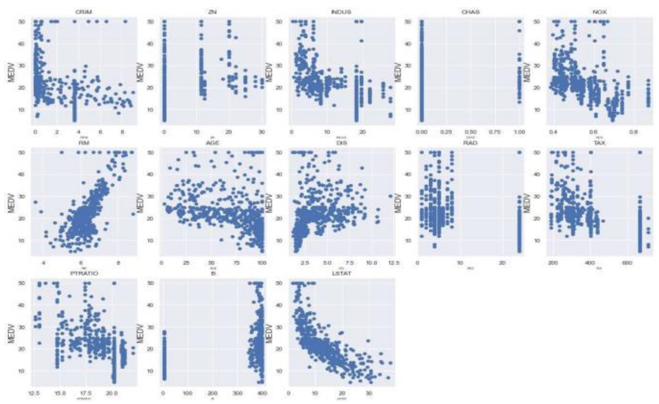

Fig. 7: Scatter plots.

The relationship between the features and the target variable MEDV (Median Price) was analysed as demonstrated in Fig. 7 using visualizations, which allowed for a clear understanding of the presence of linear relationships. By plotting the features against the target variable on scatter plots, it became evident which features showed a linear trend with MEDV. These visualizations provided valuable insights into the nature of the relationships between the features and the target variable, helping to identify which features may have a significant impact on the median price.

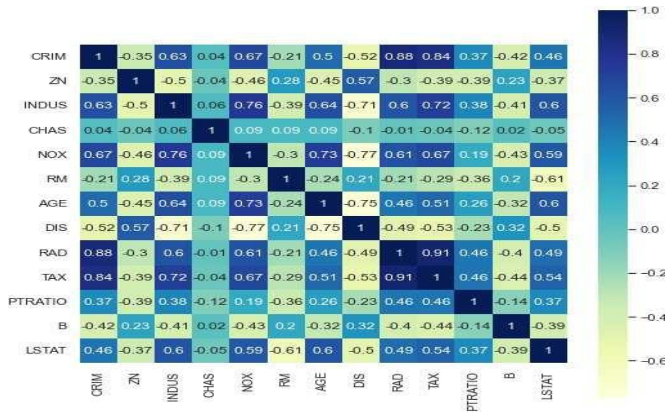

Fig. 7: Heat-map correlation between independent variables.

Finally, the heat-map as shown in Fig. 7 was used to analyse the intercorrelation among all the independent variables in the dataset. This was done to gain insights into the strength and direction of relationships between variables, which can help in identifying potential multicollinearity issues and understanding the overall structure of the data. Multicollinearity can impact the accuracy and stability of regression models, as it can result in inflated standard errors, reduced predictive power, and difficulties in interpreting the individual effects of variables.

The results shown in Table IV demonstrate a comparison between the actual and predicted values for the target variables for LR model, DT regression model, RF model and k-NN method.

TABLE IV. PREDICTION MODELS PERFORMANCE COMPARISON

Linear Regression Model		**Decision Tree Regression Model**		**Random Forest Model**		**k-NN Model**	
Actual	*Predicted*	*Actual*	*Predicted*	*Actual*	*Predicted*	*Actual*	*Predicted*
23.6	28.8	23.6	20.5	23.6	23.3	23.6	27.9
32.4	34.3	32.4	32.0	32.4	30.9	32.4	28.1
13.6	16.4	13.6	17.4	13.6	15.8	13.6	17.6
22.8	24.1	22.8	21.7	22.8	23.6	22.8	26.7
16.1	18.8	16.1	15.2	16.1	15.7	16.1	16.6

One commonly used performance evaluation metric for ML models is the RMSE. RMSE provides a measure of how well the model's predicted values align with the actual values. A lower RMSE value indicates a better fit of the model to the dataset, as it represents the average of the squared differences between predicted and actual values. By comparing RMSE values across different regression models or against a benchmark, we can make informed decisions about the model's performance and potential adjustments needed. RMSE is a valuable tool in assessing and comparing the accuracy of ML models, aiding in model selection and optimization for predictive modelling tasks. The RMSE values in Table V show that the RF model outperformed the other models with the lowest RMSE of 3.3261.

TABLE V. PREDICTION MODELS USED AND THEIR RMSE VALUES

Model	**RMSE**
Linear Regression Model	5.1437
Decision Tree Regression Model	5.3381
Random Forest Model	3.3261
k-NN Model	4.3927

Furthermore, the R-squared value, which is a measure of the proportion of variance in the dependent variable explained by the model, was calculated for the RF model. The obtained R-squared value of 0.8491 indicates that the model is able to explain approximately 84.91% of the variability in the median house prices. This high R-squared value indicates a better fit of the model to the data, suggesting that the RF model has a good level of explanatory power in predicting the target variable. Additionally, feature importance estimation [23-25] was calculated for the RF model, which provides insights into the contribution of each feature in predicting the median house prices. This was done by analysing the splits and node impurities across all the trees in the RF ensemble. The calculated feature importance values can help us understand which features are the most influential in determining the target variable. The RF model's feature importance results in Table VI show that RM, LSTAT, and DIS are the most important features for predicting median house prices, with importance values of 0.5141, 0.3284, and 0.0875, respectively, while CHAS has the least importance with a value of 0.0018.

Finally, the visualization of the regression fit of the RF model using a scatter plot as shown in Fig. 8 further provided valuable insights into the model's overall prediction performance. By comparing the actual and predicted values on a scatter plot, it is possible to visually assess how well the model's predictions align with the ground truth. This can allow for easy identification of any patterns, trends, or discrepancies in the model's predictions.

In generally, the results obtained from the results analysis clearly indicate that the RF model outperformed the other models in predicting house prices. This may be due to its ability to handle non-linearity, capture complex interactions among variables, and reduce the impact of outliers. The DT model also performed well, but it may have been more prone to overfitting, as DTs tend to be sensitive to noise in the data. LR, on the other hand, may have been limited by its assumptions of linearity and independence among variables, which may not have been fully met in the dataset. Additionally, the analysis of correlation among variables and the significance of variables helped identify important features that strongly influenced house prices. This information can be useful for decision-making in real estate investments or policy making related to housing markets.

TABLE VI. FEATURE IMPORTANCE OF RF MODEL

Feature	Importance
RM	0.5141
LSTAT	0.3284
DIS	0.0875
PTRATIO	0.0298
B	0.0202
INDUS	0.0181
CHAS	0.0018

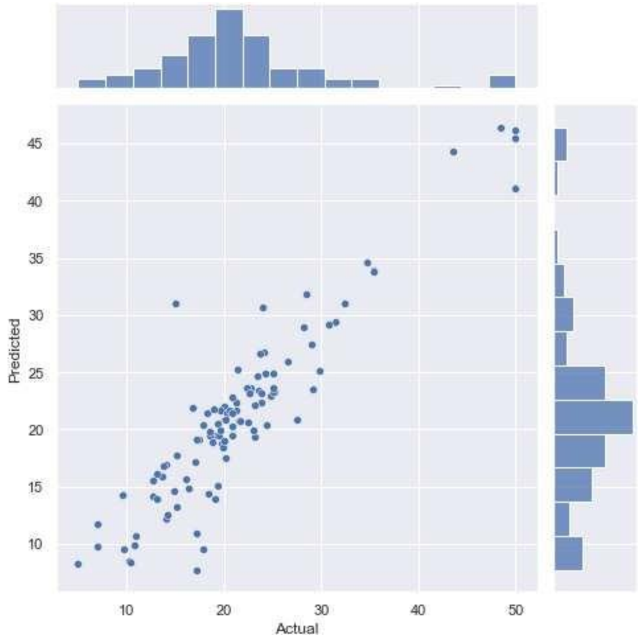

Fig. 8: Scatter plot of the regression fit of RF model.

CONCLUSION

In conclusion, the analysis of predicting house prices using ML techniques on the Boston dataset yielded valuable insights into the performance, strengths, and limitations of different models. The RF model was identified as the best performing model based on the RMSE evaluation criterion. The findings of this study can be utilized in real estate applications, policy-making, and further research in the field of ML for housing price prediction. Future research can focus on exploring other advanced modelling techniques, incorporating additional relevant variables, and validating the findings on different datasets to further enhance the accuracy and robustness of the predictions.

REFERENCES

[1] Glaeser, E.L., Gyourko, J., Morales, E. and Nathanson, C.G., 2014. Housing dynamics: An urban approach. Journal of Urban Economics, 81, pp.45-56.

[2] UN-Habitat (United Nations Human Settlements Programme). (2008). "The State of the World's Cities 2008/2009: Harmonious Cities. Available online https://unhabitat.org/books/the-state-of-the-worlds-cities-20082009-harmonious-cities. Accessed 24/06/2023

[3] Joint Center for Housing Studies of Harvard University. (2021). "The State of the Nation's Housing 2021. Available online https://unhabitat.org/books/the-state-of-the-worlds-cities-20082009-harmonious-cities. Accessed 25/06/2023

[4] Un-Habitat, 2012. State of the World's Cities 2008/9: Harmonious Cities. Routledge.

[5] Kreuzberger, D., Kühl, N. and Hirschl, S., 2023. Machine learning operations (mlops): Overview, definition, and architecture. IEEE Access.

[6] Zhang, M., Kujala, P., Musharraf, M., Zhang, J. and Hirdaris, S., 2023. A machine learning method for the prediction of ship motion trajectories in real operational conditions. Ocean Engineering, 283, p.114905.

[7] Masini, R.P., Medeiros, M.C. and Mendes, E.F., 2023. Machine learning advances for time series forecasting. Journal of economic surveys, 37(1), pp.76-111.

[8] Pham, T.Q.D., Le-Hong, T. and Tran, X.V., 2023. Efficient estimation and optimization of building costs using machine learning. International Journal of Construction Management, 23(5), pp.909- 921.

[9] Møller, S.V., Pedersen, T., Montes Schütte, E.C. and Timmermann, A., 2023. Search and predictability of prices in the housing market. Management Science.

[10] M. Ahtesham, N. Z. Bawany and K. Fatima, "House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan," 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, 2020, pp. 1-5

[11] A. P. Singh, K. Rastogi and S. Rajpoot, "House Price Prediction Using Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 203-206

[12] S. Sharma, D. Arora, G. Shankar, P. Sharma and V. Motwani, "House Price Prediction using Machine Learning Algorithm," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 982-986

[13] N. T. M. Sagala and L. H. Cendriawan, "House Price Prediction Using Linier Regression," 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia, 2022, pp. 1-5

[14] M. Cekic, K. N. Korkmaz, H. Müküs, A. A. Hameed, A. Jamil and F. Soleimani, "Artificial Intelligence Approach for Modeling House Price Prediction," 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), Istanbul, Turkey, 2022, pp. 1-5

[15] C. Chee Kin, Z. Arabee Bin Abdul Salam and K. Batcha Nowshath, "Machine Learning based House Price Prediction Model," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 1423-1426

[16] A. Gupta, S. K. Dargar and A. Dargar, "House Prices Prediction Using Machine Learning Regression Models," 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1-5

[17] S. Bai, "Boston house price prediction: machine learning," 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2022, pp. 1678-1684

[18] S. Sanyal, S. Kumar Biswas, D. Das, M. Chakraborty and B. Purkayastha, "Boston House Price Prediction Using Regression Models," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-6

[19] Begum, A., Kheya, N.J. and Rahman, Z., 2022. Housing Price Prediction with Machine Learning. International Journal of Innovative Technology and Exploring Engineering, 3075(3), pp.42-46.

[20] Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F. and Oluwadara, G., 2022. House price prediction using random forest machine learning technique. Procedia Computer Science, 199, pp.806-813.

[21] Henriksson, E. and Werlinder, K., 2021. Housing Price Prediction over Countrywide Data: A comparison of XGBoost and Random Forest regressor models.

[22] Kumar, D. (2022). House Price Prediction using Random Forest and CNN Algorithm. International Research Journal of Modernization in Engineering Technology and Science, Vol. 4, no. 8, pp.921-925.

[23] Molnar, C., König, G., Bischl, B. and Casalicchio, G., 2023. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Mining and Knowledge Discovery, pp.1-39.

[24] Zhu, M., Yang, Y., Feng, X., Du, Z. and Yang, J., 2023. Robust modeling method for thermal error of CNC machine tools based on random forest algorithm. Journal of Intelligent Manufacturing, 34(4), pp.2013-2026.

[25] Regier, P., Duggan, M., Myers-Pigg, A. and Ward, N., 2023. Effects of random forest modeling decisions on biogeochemical time series predictions. Limnology and Oceanography: Methods, 21(1), pp.40- 52.