National College of Ireland

Project Submission Sheet – 2022/2023

Student Name: 1. Akimuddin Aslam Shaikh 2. Sanjay Girish Dialani 3. Saif Ali Khan

Student ID: 1. X22123245 2. X22102442 3. X22123296

Programme: H9DAPA Year: 2023

Module: Domain Application of Predictive Analysis

Lecturer: Qurrat Ul Ain

Submission Due Date: 11 August 2023

Project Title: **Data-Driven House Price Prediction: Predictive Analytics for House Price Estimation**

Word Count: 10

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.
Signature: 1. Akimuddin Aslam Shaikh 2. Sanjay Girish Dialani 3. Saif Ali Khan

Date: 11 August, 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only
Signature:
Date:
Penalty Applied (if applicable):

Data-Driven House Price Prediction: Predictive Analytics for House Price Estimation

Akimuddin Aslam Shaikh
School of Computing
National College of Ireland
x22123245@student.ncirl.ie

Sanjay Girish Dialini
School of Computing
National College of Ireland
x22102442@student.ncirl.ie

Saif Ali Khan
School of Computing
National College of Ireland
x22123296@student.ncirl.ie

***Abstract*—This is the second and final part of two-part project endeavor that focuses on domain application, specifically house price value in Indian context, implementing advanced machine learning technique. Building on the groundwork laid in the first phase, and to explore the intricacies of the real estate market, this research uses the Kaggle "House Price Prediction Challenge" dataset. The dataset provide the research with the thorough details on real-estate property characteristics, offering an opportunity to investigate the variables affecting value of a property and build a solid predictive model empowering stakeholders to make informed decisions in the ever-changing real estate landscape. Beyond simple technique comparison, the project places higher priority in domain-specific application. For the development of the project, machine learning techniques like Linear regression, Random forest, and decision tree were implemented to build a powerful predictive model. The team gains invaluable experience in data analysis and machine learning, as project progresses, applying the acquired knowledge to address real-world challenges. The project seeks to make a meaningful impact, through responsible, fair, and transparent practices, in the dynamic world of real estate analytics and decision-making.**

I. Introduction

In the ever-changing real estate market environment, understanding the critical factors that affects the house price has become crucial. For analyzing house prices, this study sets out on a thorough journey of domain applications with cutting-edge machine learning algorithms to unearth valuable insights. After laying the foundation in first phase, the project present the cumination of the research, including a potent predictive model capable of predicting house prices. The fundamental objective of the project is to understand the complex interplay of factors that affects property values, with a focus on domain applications that provide practical solutions for stakeholders in the real estate sector. By deploying machine learning models such as linear regression [1], decision trees [2], and random forests [3], the project had build a powerful predictive model that aids in precise house price estimation. For many stakeholders in the real-estate industry, the domain application of the project initiative is of immense significance. In order to make informed decision, homeowners, prospective buyers, real estate professionals, and investors can obtain useful information into market trends, pricing strategies, and investment prospects. Better risk management benefits financial instiution by enhancing their lending procedures based on reliable house price forecasts. As the results have practical applications, the project's business value goes beyond transcends academic research.

Incorporating the outcomes from the coding part, the implementation phase entailed employing a different machine learning approaches such as linear regression [1], decision tree [2], random forests [3] to construct a reliable predictive model for house price prediction. The project showed promising performance. The Decision Tree [2] Regressor achieved the highest R-squared value of 0.701, indicating a strong correlation between the features and target variable price. However, the Random Forest Regressor [3] demonstrated a reasonable R-squared value of 0.606 and comparatively lower RMSE of 407.93 and MSE of 166409.27, making it a favorable choice for precise predictions.

II. Research and investigation into applicable techniques

For the housing price prediction project, a thorough analysis of machine learning algorithms is part of the research and investigation into the applicable methodologies.To identify the most appropriate techniques for our specific domain application, this procedure comprises examining the advantanges, disadvantage and characteristics of several approaches.

1. Linear Regression: It is is a fundamental statistical technique used in supervised learning for predicting continuous numerical values. For housing price prediction analysis, It was a natural choice because it models the relationship between the dependant variable and an independant variables. In this project, the dependant variables were BHK or RK, stakeholder, under construction, location, square feet and independant variables was Target price (in Lacs), allowing the research to estimate the impact of each feature on the target variable. To minimize the error between the predicted and actual values, the model is represented by a linear equation, and its coefficients are determined through mathematical optimization.

Advantages of using Linear regression:

a. Simplicity: Linear regression [1] prove as a good starting point for modeling, as it is straightforward, easy to understand, simple to implement and easier to interpret the output coefficient.

b. Interpretability: The model's coefficients shed light on the magnitude and direction of each characteristic's impact on housing prices.

c. Speed: Due to its processing efficiency, It is appropriate for large datasets.

Disadvantages of using Linear regression:

a. Linearity assumption: The assumption of a linear relationship between the features and the target variables in it may not always be true in real-world scenario.

b. Limited flexibility: Sometimes it may prove incapable of capturing complex non-linear relationships between features and target variables.

c. Sensitivity to outliers: It can potentially have an impact on performance of model as it is sensitive to outliers.

2. Decision Trees: Decision trees [2] are hierarchical and adaptable supervised learning models as it is capable of handling both classification and regression tasks. It split the data into branches recursively, and partition the data into subsets based on the values of features, leading to a tree-like structure. In decision trees [2], each branch represents a decision rule, each leaf node represents a predicted value, and each internal node represents a characteristic. It can be helpful in predicting house prices since they have the ability to identify the important factors that affects a house prices. This information can prove helpful as it used to make more precise predictions about the price of a house, which can benefits both buyers and sellers.

Advantages of using Decision trees:

a. Simple to use and implement: It has a hierarchical nature and generate a tree like graphical representation that is easy to understand and can be explainable to technical teams and stakeholders.

b. Non-linearity: It is capable of capturing complex non-linear relationships in the data, where multiple factors can influence property values, making it suitable for housing price prediction.

c. Robustness: They are less affected by missing values and outliers values as compared to linear regression [1].

d. Interpretable: Decision trees [2] provides with clear insights into the decision-making process, and are easy to interpret and visualize .

Disadvantages of using Decision trees:

a. Overfitting: Decision trees [2] can impair generalization to new data, as they are suspectible to overfitting, especially with deep trees, which

b. Instability: The model is less stable as it generate different decision tree [2] result from small changes in the data.

c. Computational and Time complexity: It can be computationally expensive to train a decision tree [2], especially when dealing with large datasets or complex tasks.

3. Random Forest: Random forest [3] is a machine learning technique that make use of an ensemble of decision trees [2] to make predictions and reduce overfitting. On several randomly selected subsets of the data, multiple decision trees [2] are created, and averaging the predictions of individual trees (for regression tasks), to make a final prediction. They are an efficient algorithm that can be useful in house price prediction as it can handle both regression and classification tasks, and can provide visuals into important features.

Advantages of using Random forest:

a. Robustness: It is less suspectible to overfitting as compared to individual decision trees [2].

b. Feature importance: Random Forest [3] allow us to pinpoint the most critical variables that affects the house prices.

c. Reduced risk of overfitting and improved accuracy: By combining many decision trees [2] into a single ensemble model, Random forests [3] can reduce the risk of overfitting and produces forecast that are more reliable.

Disadvantages of using Random forest:

a. Complexity: The disadvantage in random forest [3] are similar to decision trees [2] as it can also prove to be computationally expensive and challenging to interpret compared to individual decision trees [2].

b. Black-box nature: It provides accurate predictions, However, it is called as 'black-box' model as it is difficult to understand how the model is making predictions.

We can make informed decisions about which models to implement for housing price prediction, by carefully examining these relevant methodologies, considering their strengths and weaknesses in domain application contexts.

III. Implementation of selected techniques

In this section, the reseaarch delve into the practical implementation of the selected machine learning techniques on the housing price prediction dataset that was downloaded from kaggle website. The implementation involves a series of steps, including data preprocessing, data transformation, Feature engineering and selection, model creation, training, and evaluation.

A. Load the dataset

In this initial step, the dataset was first loaded and carried out some statistical calculation to understand the nature of dataset.

B. Data Preprocessing

Data preprocessing is a crucial part of the machine learning as in this stage, we can analyze the data and gain some valuable insights. To ensure the dataset's suitability for analysis, data preprocessing was conducted thoroughly, before applying the machine learning models. The dataset consisted of 68720 rows and 12 column, with a variety of features capturing important aspects of residential properties.

a. Property Types Distribution: With 68662 instances of BHK (bedroom, hall and kitchen) properties and 58 instances

of RK (room and kitchen properties, the dataset showcased a predominant distribution of property types.

b. Location Insights: By analyzing the property locations it was found that certain cities held prominence in property listings. Among all the cities in India, Bangalore, Lalitpur, Pune, Mumbai, and Kolkata emerged as the most listed cities, indicating significant market activity.

c. Stakeholder's Information: The distribution of the Seller highlighted the participation of various entities in real estate transactions. The highest count were of Dealers that contributes to 42437 properties, followed by Owners that stands at 24920, and the least were of Builders that had listed only 1363 properties.

d. Construction Status: Further research into the construction status of properties revealed that 56587 properties were not under construction while 12133 properties were under construction.

C. Data Visualization

This section of the research offers explanations for some of the important visualizations obtained from the dataset.

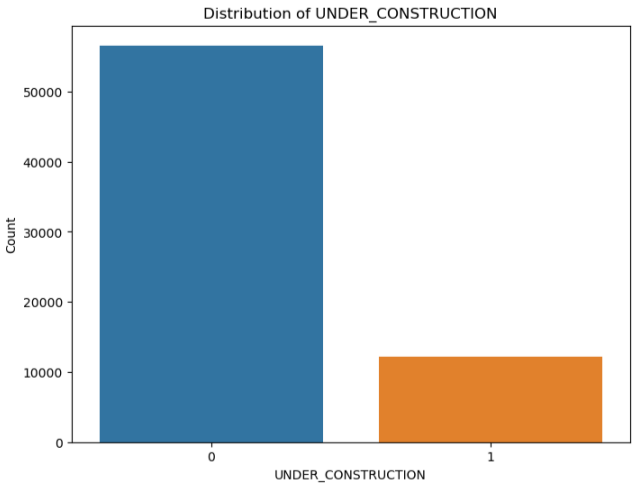

Fig. 1. Histogram of construction status

Figure 1 depicts the distribution of properties construction according to their status. It can be analyzed that more than 50,000 properties are alreadt built, wheras on the other hand, nearly around 10,000 properties are under construction.

Figure 2 depicts the distribution of stackholders according to their construction status.

Figure 3 displays the number of properties posted by sellers for sale. From the graph it was found that highest number of properties posted were by Dealer, followed by Owner and the least by builder.

Figure 4 depicts the average price of properties listed by stackholders. It was analyzed that the highest average price were listed by builder that stands around 17.5 lakhs (INR), followed by dealer that contributes around nearly 15 lakhs, and at last by owner of properties at around more than 12.5 lakhs.

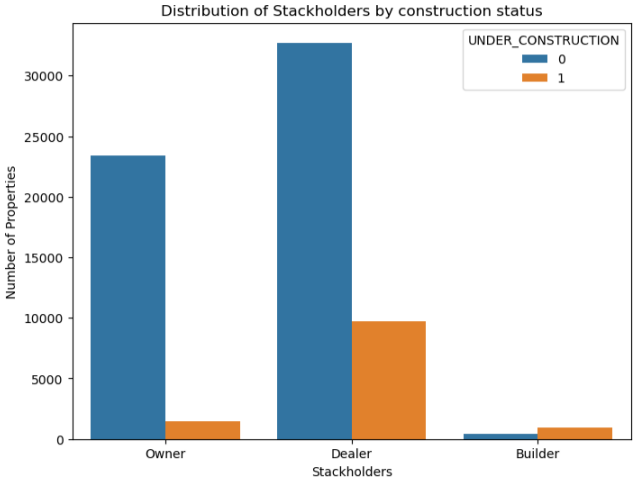

Fig. 2. Properties of stackholders by construction status

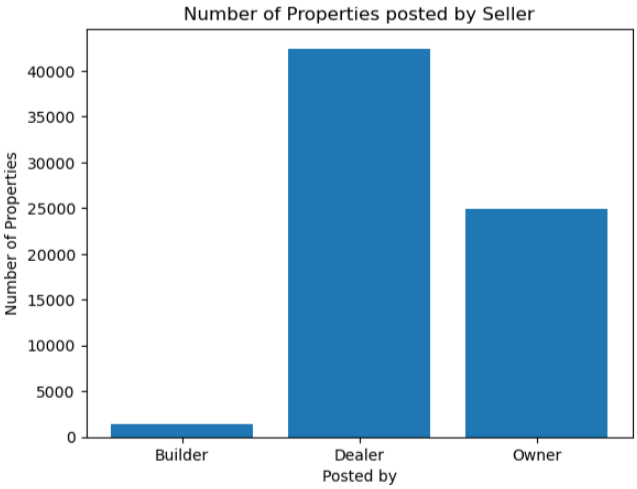

Fig. 3. Histogram of BHK or RK properties

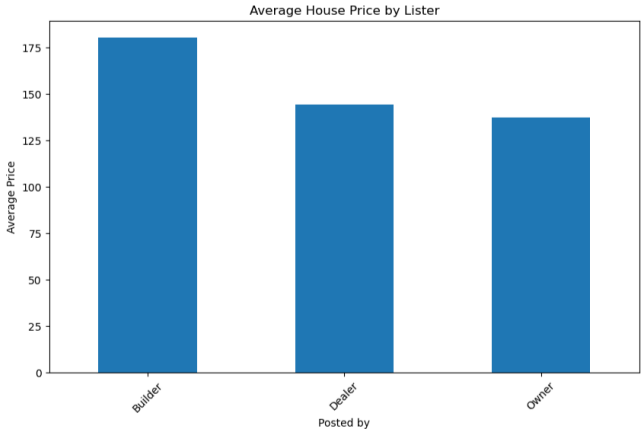

Fig. 4. Histogram of BHK or RK properties

All the graphs provide valuable insights that are beneficial

to stackholder and customers to make informed decision about the properties price.

D. Data Transformation

In this step, some column such as 'POSTED BY', 'BHK OR RK', and 'ADDRESS', were dropped, to prepare the dataset for model training, as they were deemed to be less influential for predicting housing prices. The remaining features from dataset were selected against the target variable 'PRICE'.

E. Feature Scaling and Encoding

To standardize the feature scales and encode categorical variables, the implementation of pipeline method allowed the construction of data transformation pipeline. In this pipeline process two key transformations were employed, where MinMaxScaler was used for scaling numerical features to bring them within a consistent range and OneHotEncoder [4] was used to encode Categorical using OneHotEncoder [4] to convert them into a format suitable for machine learning algorithms.

F. Model Training and Evaluation

In this step, Three distinct machine learning algorithms were implemented and evaluated:

a. Linear Regression: Linear regression [1], a fundamental statistical method for forecasting numerical values, was applied to the preprocessed dataset. An evaluation of model's performance resulted an R-squared value of 8.86e-05, signifying that it had insufficient predictive capability for this complex domain application.

b. Decision Tree Regressor: The transformed data was trained on decision tree [2] regressor as it is capable of capturing non-linear relationships. Evaluation of the decision tree model resulted in an R-squared score of 0.70, indicating that it performed better as compared to linear regression [1].

c. Random Forest Regressor: for more robust predictions, The random forest regressor [3] was employed, as it is an ensemble method that combine multiple decision trees [2]. Evaluation yielded an R-squared score of 0.61, indicating a reasonable predictive performance.

G. Interpretation and Conclusion

The use of the chosen approaches revealed the complexity of the dataset and the difficulties in making reliable predictions of property prices. The three models were compared, and the results showed how crucial it is to use more advanced algorithms in order to capture complex correlations in the data.

The following sections of this study will go further into the findings, interpret them, explore the consequences for the domain application, and offer practical advice for real estate stakeholders.

IV. Quantitative Results and Business Value Interpretation

The precise prediction of house prices has a significant impact on the real estate business, as it directly influences buyers' purchase decisions, sellers' listing methods, and investors' profits. The goal of this research was to create a predictive model for estimating house prices based on pertinent features. We used a variety of regression approaches to accomplish this, including linear regression [1], decision tree regression [2], and random forest regression [3]. The goal was not only to evaluate the models' quantitative outputs, but also to analyze their business value and practical consequences in the real estate arena.

A. Quantitative Results

Following the data preparation step, three regression approaches were used: linear regression [1], decision tree regression [2], and random forest regression [3]. The dataset had 68,720 samples with 12 characteristics, one of which was the target variable 'PRICE'.

1. R-squared: The coefficient of determination, often known as the R-squared value, quantifies the amount of variance in house prices explained by the predictor factors. The R-squared value for the linear regression [1] model was determined to be 8.86e-05, indicating that the model explains a very little part of the variance. The decision tree regression model [2], on the other hand, achieved a substantially higher R-squared value of 0.7018, indicating greater explanatory power. Similarly, the random forest regression model [3] had an R-squared value of 0.6063, indicating that it performed well in explaining variance.

The R-squared values show the models' goodness-of-fit. A higher R-squared value shows that the model captures a greater fraction of the variance in house prices, indicating better predictive potential. When analyzing R-squared values, it is critical to keep the problem context and domain-specific aspects in mind. Numerous complex elements influence property prices in real estate, making it difficult to get high R-squared values with restricted characteristics.

2. Root Mean Squared Error (RMSE): The RMSE is a metric that estimates the average magnitude of prediction mistakes. The RMSE for the linear regression [1] model was calculated to be 650.11, suggesting that forecasts differed from real house values by roughly 650.11 units on average. With an RMSE of 355.05, the decision tree regression model [2] surpassed linear regression [1], implying higher prediction accuracy. The RMSE of the random forest regression model [3] was 407.93, indicating its capacity to provide more accurate predictions.

The RMSE values represent the models' average prediction error. Lower RMSE values indicate better forecast accuracy. The RMSE provides a knowledge of the average variance between projected and actual values in the context of housing price prediction. Lower RMSE values signify more accurate estimations, which is critical in the real estate market for both buyers and sellers.

3. Mean Squared Error (MSE):

Larger errors are penalized more strongly by the MSE, which squares the RMSE. The linear regression [1] model produced an MSE of 422,647.53, indicating that the model's forecasts are far off from average house values. In contrast,

the decision tree regression [2] and random forest regression models [3] performed better, with MSE values of 126,061.72 and 166,409.27, respectively.

MSE values are a quantifiable measure of prediction accuracy that may be used to compare models. Lower MSE values, like the RMSE, indicate superior forecasting accuracy. Lower MSE values indicate models that produce forecasts that are closer to the actual prices, implying greater confidence in the model's projections.

B. Business Value Qualitative Interpretation

While quantitative results provide useful insights into model performance, they must be interpreted in the context of the real estate domain in order to comprehend their practical consequences and business value.

1. Limited Predictive Power: The linear regression [1] model's low R-squared value shows that it might not be the best match for the dataset. As a result, the model's ability to anticipate property prices effectively based on the selected features may be jeopardized. The decision tree regression [2] and random forest regression models [3], on the other hand, beat the linear regression [1] model, demonstrating more predictive power and the possibility for more accurate estimations.

Numerous complicated and dynamic elements influence property prices in the real estate market, making accurate projections difficult. As a result, it is critical to recognize that no single model can represent the total intricacy involved in determining housing values.

2. Moderate Predictive Accuracy: All three models' RMSE data show that the average prediction error spans from 355 to 650.11 units of the target variable. While this degree of precision may be adequate for preliminary price estimations and broad trend research, it may not be exact enough for key business choices like setting listing prices or determining property valuations.

In the real estate industry, even little forecasting errors can have huge financial consequences for buyers, sellers, and investors. As a result, while the models might provide useful insights, they should be employed as supplementary tools rather than primary decision-makers.

3. Business Decision Support: Despite their shortcomings, prediction models can be useful tools for real estate brokers and sellers. Predictions from the model can provide significant insights into prospective price ranges, assisting in determining initial listing pricing and understanding overall market trends. However, when depending entirely on the model's estimations, it is critical to exercise caution, as human expertise and domain knowledge are still required for sophisticated decision-making in the real estate business.

Real estate stakeholders can use the models to get preliminary estimates of property prices, which will help them better understand market trends and make data-driven pricing decisions. To arrive at more accurate and context-specific pricing methods, the model's predictions must be combined with expert judgment.

4. Features Importance: Through feature importances, the decision tree [2] and random forest models [3] have the advantage of discovering key elements that strongly influence house values. This information can help real estate developers and investors focus on key qualities and make informed investment decisions. Understanding the relevance of property features also aids in selecting property traits that have a direct impact on pricing methods.

Identification of significant features enables real estate stakeholders to focus on aspects that influence property pricing. Stakeholders can target renovations or alterations to increase property prices and appeal to potential buyers by evaluating the relative relevance of certain aspects.

5. Conclusion Finally, the quantitative results show the merits and limits of the employed regression models for predicting housing prices. Although the models have a moderate prediction accuracy, their business value comes from providing real estate stakeholders with significant insights regarding feature relevance, initial price estimations, and overall market trends. Despite their limitations, these models might be useful as supplemental tools for real estate decision-making

However, their dependability must be carefully assessed, and human knowledge is still required in evaluating and applying model projections for essential business choices. As the real estate market evolves, using modern methodologies and improving models will be critical for maximizing corporate value and contributing to the real estate industry's growth and success. Continuous refinement, domain expertise, and a holistic approach are required to fully realize the potential of machine learning models in house price prediction and to serve stakeholders in the ever-changing real estate sector.

REFERENCES

[1] Ningyan Chen et al. House price prediction model of zhaoqing city based on correlation analysis and multiple linear regression analysis. *Wireless Communications and Mobile Computing*, 2022, 2022.

[2] Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. Determinants of house price: A decision tree approach. *Urban Studies*, 43(12):2301–2315, 2006.

[3] Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, and Gbenle Oluwadara. House price prediction using random forest machine learning technique. *Procedia Computer Science*, 199:806–813, 2022.

[4] Weinan Weng. Research on the house price forecast based on machine learning algorithm.

PROJECT LOG

PHASE	DATE	AKIMUDDIN SHAIKH	SANJAY DIALINI	SAIF ALI KHAN
CA1	27/06/2023	Dataset collection, explored and selected machine learning techniques	Conducted research on ethical concerns	Focused on target audience
	28/06/2023 TO 29/06/2023	projected the goals of the project	Research on business value of the project	worked on preliminary visualization, Listed the applicable technique
CA2	7/8/2023	Data exploration, data preprocessing, Identified Key Insights, data visualization	Research on Linear regression, and its advantage, disadvantage	collaborated on the selection of machine learning techniques and began quantitative analysis.
	8/8/2023	Data Transformation, Feature Scaling and Encoding,	Research on Decision tree, and its advantage, disadvantage	Analyzed dataset statistics, interpreted the business value of quantitative results,Focused on evaluating model performance
	9/8/2023	model training and evaluation, wor	Research on Random forest, and its advantage, disadvantage	Concluded quantitative analysis, provided qualitative interpretations, and highlighted implications for the real estate domain.
	10/8/2023	work on report	Compiled Research Findings, Evaluated applicability	work on report
	11/8/2023	Video Presentation and cross check the report work		