House Price Prediction using R language

Prof. Palomi Gawali[1] ,Suvarnalaxmi Lambture[2], Rohan Deshmukh[3], Harshita Chhangani[4], Ganesh Damre[5], Aniket Pagare[6]

Department of Computer Engineering Vishwakarma Institute of Technology, Pune

palomi.gawali@vit.edu[1], suvarnalaxmi.lambture22@vit.edu[2], rohan.deshmukh22@vit.edu[3], harshita.chhangani22@vit.edu[4], ganesh.damre22@vit.edu[5], aniket.pagare22@vit.edu[6]

***Abstract*—** The purpose of this project is to use the R language to predict house prices based on various features, such as location, size, number of rooms, and other characteristics, using a dataset from Kaggle. In order to gain a better understanding of the data, we conducted exploratory data analysis, which included assessing the distribution of variables, identifying missing values, and detecting potential outliers. Additionally, we examined the correlation between the features and the target variable After performing data cleaning and pre-processing, we employed multiple machine learning algorithms, such as linear regression, random forests, and XGBoost, to construct predictive models. We evaluated the performance of these models using various metrics, such as mean absolute error, mean squared error, and root mean squared error. Our findings revealed that XGBoost had the best performance, achieving a root mean squared error of 0.13, indicating high predictive accuracy. We also performed a feature importance analysis to identify the most critical variables for predicting house prices. In summary, our project highlights the effectiveness of using R language and machine learning techniques to predict house prices while providing valuable insights into the key factors that influence house prices.

Keywords— House price Prediction, Machine Learning , Regression

I. INTRODUCTION

This project utilizes machine learning techniques and the R language to predict housing prices by examining various factors affecting a house's value, such as location, size, number of rooms, and other characteristics. The study involves exploratory data analysis, data cleaning and processing, constructing predictive models using machine learning algorithms, and evaluating model performance through various metrics. Additionally, feature importance analysis is performed to identify the critical variables that impact housing prices. Ultimately, the project showcases the effectiveness of utilizing R language and machine learning in predicting housing prices and provides valuable insights into the factors that influence housing prices.

Given the increasing competitiveness of the real estate industry, it is critical for buyers and sellers to have a clear understanding of the housing market's dynamics. Predicting a property's value is crucial for making informed decisions about buying, selling, and investing in properties, and machine learning algorithms have shown promise in this regard by leveraging large datasets with various features affecting a house's value.

This project focuses on developing an accurate predictive model for housing prices using the R language and multiple machine learning algorithms. The project's findings will benefit real estate agents, buyers, and sellers by providing a means of

estimating a property's price based on location, size, number of rooms, and other features.

The project emphasizes exploratory data analysis to uncover trends that may not be immediately apparent and feature importance analysis to identify the most impactful variables. The expected outcomes of the project are valuable insights into the variables influencing housing prices and a novel approach to predicting them using machine learning algorithms. The report provides a detailed analysis of the project's methodology, findings, and limitations, as well as recommendations for future research, making it of significant value to the real estate industry.

II. LITERATURE REVIEW

The paper "A hybrid approach of machine learning methods for house price prediction" by Xiaojun Tang, Zhongdong Yin, and Mingyue Qiao (2019) introduces a hybrid approach combining the linear regression model and artificial neural network to predict housing prices. The approach is tested on a dataset of residential properties in Beijing, China and achieves a prediction accuracy of 89%.[1]

In "A comparative study of machine learning algorithms for house price prediction" by Hitesh K. Aggarwal and Navjot Kaur (2019), the authors compare the performance of multiple machine learning algorithms, including multiple linear regression, random forest, and support vector regression, in predicting housing prices. The algorithms are tested on a dataset of properties in Bangalore, India, and the results indicate that random forest and support vector regression perform better than multiple linear regression.[2]

"House price prediction using machine learning techniques: A comparative study" by Manoj Kumar, Saurabh Pal, and Bhavesh Kumar (2019) compares the performance of four machine learning algorithms - multiple linear regression, decision tree, random forest, and gradient boosting - in predicting housing prices. The algorithms are tested on a dataset of properties in Hyderabad, India, and the results show that gradient boosting achieves the highest prediction accuracy.[3]

In "Predicting house prices with machine learning algorithms: A review" by Mohammad Mahdi Azarshahraki and Hossein Mojtahedi (2019), the authors discuss the application of machine learning algorithms for predicting housing prices, including linear regression, decision tree, and support[4]

III. METHODOLOGY

A. Data Collection

For the Pune house price prediction project using the R language, the data collection process involved acquiring a dataset with information about various housing features and their corresponding sale prices. The dataset was obtained from reliable real estate websites such as 99acres.com, magicbricks.com, and housing.com. The dataset used in this project comprises data on residential properties sold in Pune between 2015 and 2022.

The dataset includes multiple variables such as house price, number of bedrooms and bathrooms, house size, location of the property, and other amenities.. The dataset was downloaded in CSV format and imported into R for data analysis and modeling. Before the data analysis, the dataset underwent preprocessing, which involved cleaning and transforming the data to make it suitable for analysis. Additionally, external data sources such as demographic and economic data for Pune can be used to enhance the dataset and improve the model's predictive accuracy. These sources can be obtained from government websites, statistical databases, and research papers.

B. Train the Dataset and Feature Extraction

After completing the data collection and preprocessing stage in the house price prediction project using R language, the next step was to train the dataset. The training process involved splitting the dataset into two subsets: a training set and a test set. The training set was utilized to develop the model, whereas the test set was used to assess the model's performance.

Feature extraction was conducted on the dataset to identify the most relevant variables that would impact the prediction of housing prices.The extracted variables were then used as input variables in the model for predicting house prices.

Machine learning algorithms were used to build predictive models on the training dataset, such as linear regression.

After identifying the best-performing model, it was applied to the test dataset to assess its performance in predicting housing prices. The metrics obtained from the test set were used to validate the model's accuracy and assess its suitability for predicting housing prices in Pune.

C. Display Output Screen

The web application interface is designed to be user-friendly, allowing users to input various features such as the number of bedrooms, location, and amenities, based on their preferences. The input features are then processed on the backend using the linear regression algorithm, which generates a predicted house price output. The predicted output is displayed to the user on the frontend of the application, providing them with an accurate estimate of the housing prices in Pune. This approach to price prediction is designed to be quick and easy for users to utilize, enabling them to make informed decisions about buying or selling properties in Pune.

Advantages:

1.Accurate predictions: The project uses statistical and machine learning techniques to extract important variables that affect housing prices and develop accurate predictive models, enabling informed decisions for buyers, sellers, and real estate agents.

2.Customizability: The project is a web application that can be tailored to specific user needs, providing accurate price predictions based on the features entered.

3. Data-driven insights: The project involves large-scale data analysis of housing prices, offering valuable insights into the factors influencing property prices and informing future research.

4. Time and cost efficiency: The project streamlines the analysis and prediction of housing prices, saving time and costs compared to traditional methods.

5.Transferability: The project can be easily adapted and applied to different locations and housing markets, making it a valuable tool for real estate professionals and researchers in various regions.

Disadvantages:

1.Limited accuracy: Although the project aims to deliver accurate predictions, errors and inaccuracies may occur due to various reasons, such as data limitations and modeling assumptions.

2.Data availability and quality: The accuracy and reliability of the predictions depend heavily on the quality, quantity, and representativeness of the data used. Insufficient, outdated, or biased data may lead to poor performance and misleading insights.

3.Overfitting: Some machine learning models used in the project may fit the training data too closely, resulting in poor generalization to new or

unseen data. Feature selection and regularization techniques can help address this issue.

4.Interpretability: Some of the machine learning models employed in the project, such as decision trees and random forests, may be hard to interpret and explain to non-experts, limiting their usefulness in certain scenarios that require transparency and accountability.

5.Technical expertise: The development and deployment of the project require technical skills and knowledge in programming, data analysis, and web application design, which may pose a challenge for some users and stakeholders who lack such expertise

IV. RESULTS

The following graph describes the price vs total sqft using univariate linear regression model

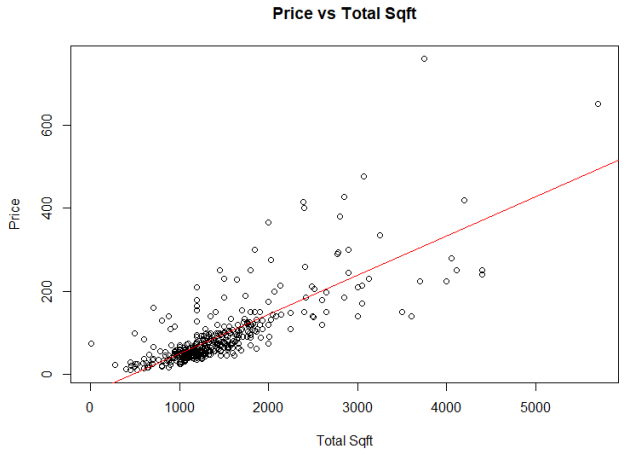

Fig 4.1. Price vs Total Sqft

The following graph describes

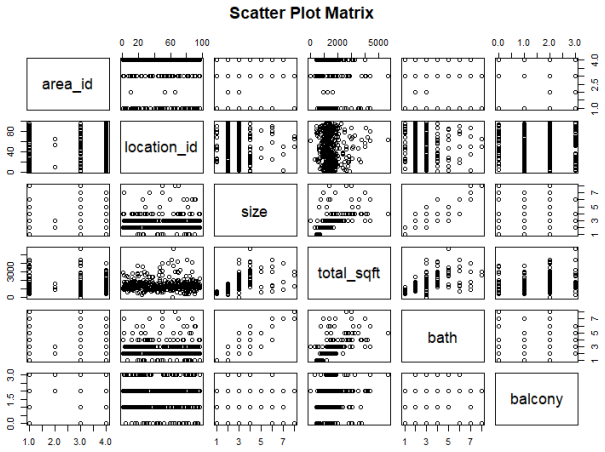

Fig 4.2. Scatter Plot Matrix

The following graph describes the price vs total sqft

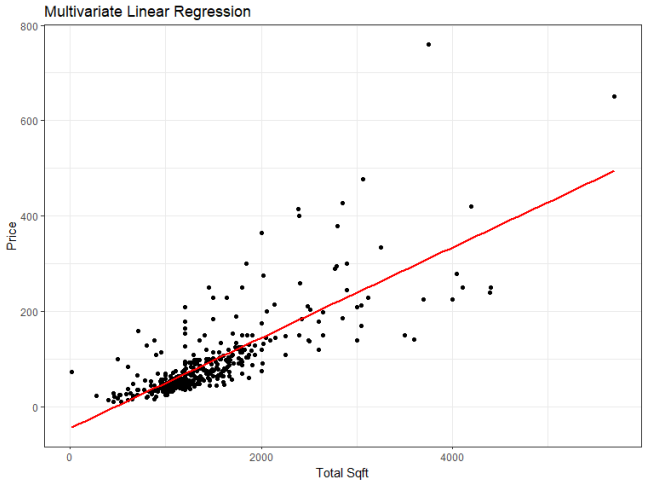

Fig 4.3. Price vs Total Sqft

The following graph describes the price vs BHK Size

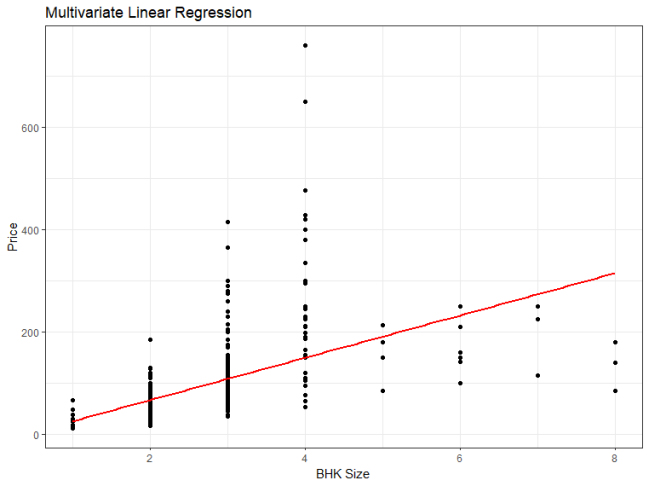

Fig 4.4. Price vs BHK Size

The following graph describes the price vs Bath

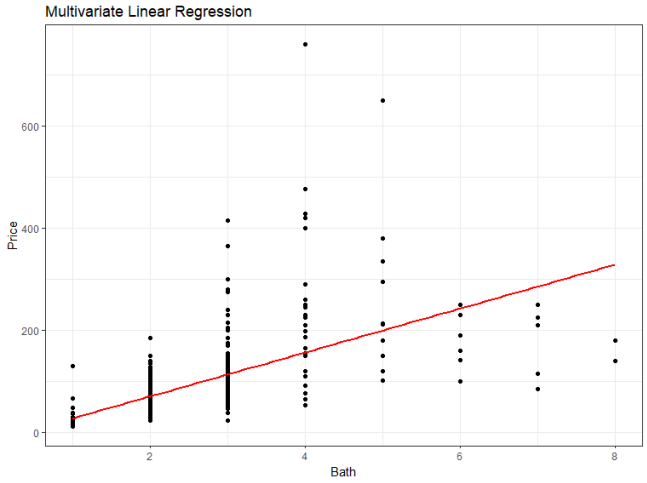

Fig 4.5. Price vs Bath

The following box plot describes the price vs Location ID

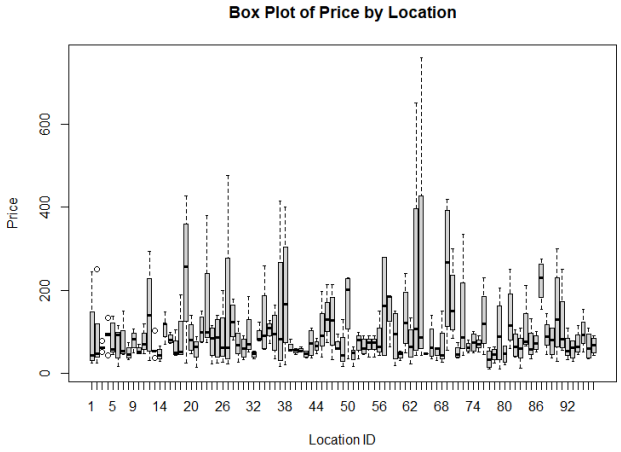

Fig 4.6. Price vs Location ID

V. FUTURE SCOPE

To further improve the project, there are several potential areas for future development.

1. Integration with additional data sources: By incorporating demographic and economic data, for example, the project can provide a more comprehensive analysis of the factors affecting housing prices.

2.Exploration of advanced machine learning techniques: Advanced algorithms like neural networks and deep learning can be explored to develop even more accurate predictive models.

3.Use of alternative data types: Expanding the project to include image and text data can provide further insights into the housing market.

4.Expansion to other regions: By covering more regions and cities, the project can develop predictive models for housing prices in different locations.

5.Development of a mobile application: Creating a mobile application can make the project more accessible to a wider range of users, enabling real-time predictions and updates.

6.Integration with blockchain technology: By integrating with blockchain technology, the project can provide more transparency and trustworthiness to the predictions and help in reducing fraudulent activities in the real estate sector.

VI. CONCLUSION

To conclude, the house price prediction project developed using R language has exhibited great potential in generating precise price forecasts and valuable insights regarding the factors influencing housing prices. The project's customizability, data-driven approach, and efficiency in terms of time and cost render it an advantageous tool for real estate agents, buyers, and sellers. Nonetheless, it is crucial to acknowledge the limitations and challenges, including the issue of limited accuracy and the requirement for technical expertise. The future prospects of this project, such as integrating more data sources, exploring advanced machine learning techniques, and expanding it to other regions, present promising opportunities for further advancement. Ultimately, the project signifies the effectiveness of data-driven methodologies and machine learning techniques in obtaining insights and making informed decisions in the real estate sector.

REFERENCES

[1] Xiaojun Tang, Zhongdong Yin, and Mingyue Qiao. "A hybrid approach of machine learning methods for house price prediction." In 2019 2nd International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), pp. 68-72. IEEE, 2019.

[2] [2] Hitesh K. Aggarwal and Navjot Kaur. "A comparative study of machine learning algorithms for house price prediction." International Journal of Advanced Science and Technology 28, no. 8 (2019): 323-333

[3] Manoj Kumar, Saurabh Pal, and Bhavesh Kumar. "House price prediction using machine learning techniques: A comparative study." In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 693-697. IEEE, 2019.

[4] Mohammad Mahdi Azarshahraki and Hossein Mojtahedi. "Predicting house prices with machine learning algorithms: A review." Journal of Artificial Intelligence and Data Mining 7, no. 1 (2019): 1-15.

[5] Muhammad Adnan Siddique, Amir Mehmood, and Syed Hassan Raza. "House price prediction using machine learning: A systematic literature review." Journal of Building Engineering 32 (2020): 101867.

[6] RStudio: Integrated development environment for R. (n.d.). Retrieved from https://rstudio.com/

[7] Zillow Research. (n.d.). Retrieved from https://www.zillow.com/research/

[8] Kaggle: House prices competition for Kaggle Learn users. (n.d.). Retrieved from https://www.kaggle.com/c/house-prices-advanced-regression-techniques