Real Estate Price Prediction using Machine Learning

Vinitha Chowdary A
CSE-AI
Amrita School of Engineering
Bengaluru, India-560035
bl.en.u4aie22066@bl.students.amrita.edu

Gamidi Rohan
CSE-AI
Amrita School of Engineering
Bengaluru, India-560035
bl.en.u4aie22019@bl.students.amrita.edu

Sneha Saragadam
CSE-AI
Amrita School of Engineering
Bengaluru, India-560035
bl.en.u4aie22057@bl.students.amrita.edu

***Abstract*— The Real Estate Price Prediction project aims to leverage machine learning techniques to develop an accurate model for predicting property prices. The dataset undergoes extensive preprocessing, including handling missing values, outlier removal, and feature engineering. The model employs a Linear Regression algorithm, achieving a reasonable accuracy score during evaluation. Cross-validation is utilized to ensure robust performance. The project also features data visualisation, showcasing insights into property distributions, price per square foot, and the impact of different features on pricing. The predictive model allows users to estimate property prices based on essential characteristics, providing valuable information for individuals navigating the dynamic real estate market in Bengaluru.**

I. INTRODUCTION

The real estate sector is a dynamic landscape, influenced by various factors that contribute to the fluctuating values of properties. In this project, we delve into the intricate world of real estate price prediction, employing advanced machine-learning techniques to unravel patterns and insights that drive market dynamics. The primary motivation behind this endeavour is to empower stakeholders – from potential buyers and sellers to real estate investors – with a predictive tool that aids in making informed decisions.

Understanding the intricacies of the real estate market requires a nuanced approach. We start by sourcing our data from Kaggle, focusing on comprehensive features such as location, size, number of bedrooms, bathrooms, and various other attributes that play pivotal roles in determining property values. The dataset undergoes a meticulous preprocessing phase, including data cleaning to rectify inconsistencies, and feature engineering to distil meaningful information from the raw data. This rigorous preparation sets the stage for our machine learning model to learn and extrapolate patterns from historical real estate transactions.

The project is not just about prediction, it's about creating a comprehensive tool. The model goes beyond numerical outputs by providing visual representations of trends and insights. Users can explore and interpret the data, gaining a deeper understanding of the factors influencing property values in specific locations or property types.

The real estate price prediction project is an amalgamation of data science and machine learning. It aims to be a valuable resource for anyone navigating the complex world of real estate, offering predictions that are accurate and reflective of the dynamic nature of the market. By harnessing the power of data, our project aspires to contribute to more informed decision-making in the realm of real estate transactions.

II. LITERATURE SURVEY

In recent years, a surge of research has been witnessed in the domain of real estate price prediction using machine learning techniques. One notable endeavour is the development of an estate price prediction system that leverages temporal and spatial features along with a lightweight deep learning model [1]. This project introduces two frameworks: the Basic Framework for Estate Price Prediction (BEPP) and the Lightweight Framework for Estate Price Prediction (LEPP). BEPP incorporates a Gaussian mixture model (GMM) to categorize estates based on features and a novel spatiotemporal data structure. Meanwhile, LEPP employs an extraction model to rank influential features, feeding them into a lightweight model for accurate predictions. The lightweight model not only significantly reduces computational costs compared to traditional CNN-LSTM models but also enhances prediction accuracy [1].

Another study delves into house price prediction using a variety of machine learning algorithms, catering to the complexities of the housing market [2]. This research considers

factors such as area, location, population, size, and the number of bedrooms and bathrooms. Various algorithms, including linear regression, least absolute shrinkage and selection operator (LASSO), and decision tree regression, are employed to predict house prices with high accuracy. The study compares the performance of these algorithms and discusses the advantages and disadvantages of the proposed model [2].

Similarly, a project focusing on predicting house prices employs machine learning algorithms such as Decision Tree regressor and trains the dataset using Light GBM and Random Forest [3]. Achieving a commendable 90% accuracy in predicting house prices, this system offers valuable insights for buyers and property investors. The research acknowledges the importance of accurate predictions based on customer needs and financial income, though it highlights the dependence on the availability and accuracy of the dataset [3].

In the context of incorporating spatio-temporal dependencies into machine learning algorithms for housing price prediction, a study conducted in Metropolitan Adelaide, Australia, stands out [4]. Utilizing a 32-year housing price dataset with 428,000 sale transaction records and 38 explanatory variables, the research employs non-linear tree-based models and ensemble machine learning techniques. The study underlines the implications for real estate investors, builders, policymakers, and property valuation methods, while also recognizing limitations related to macroeconomic forces affecting the Australian property market [4].

Furthermore, a study focusing on housing price prediction in Taiwan from 2013 to 2018 explores the efficacy of deep learning methods, particularly CNN and BPNN [5]. By constructing a dataset comprising housing attributes and macroeconomic data, the study demonstrates the effectiveness of deep learning methods in predicting housing prices and suggests avenues for further research in refining housing price prediction models [5].

A comprehensive study comparing traditional and advanced machine learning approaches for housing price prediction has been conducted using the "Housing Price in Beijing" dataset [6]. Employing Random Forest, XGBoost, and LightGBM, the research showcases the advantages of the Stacked Generalization method. While providing promising results, the study opens avenues for future research and suggests the need for considering the combination of different machine learning models [6].

The significance of machine learning in predicting housing prices is further emphasized in a study that explores linear regression, artificial neural networks, support vector regression, and gradient boosting [7]. This research underscores the importance of location and structural characteristics in determining property prices. It outlines the advantages and disadvantages of different machine learning models, highlighting the complexity of accurately predicting housing prices [7].

A study on house price prediction introduces a model based on machine learning, comparing methods such as XG Boost, Random Forest, LightGBM, hybrid regression, and stacked generalization regression techniques [8]. While showcasing the effectiveness of machine learning algorithms in predicting housing prices, the research acknowledges the potential challenges, such as overfitting in Random Forest. The study suggests future directions for research, including further validation of different datasets and exploration of additional machine-learning techniques [8].

The paper[9] aims to delve into the intricate realm of interaction effects within small samples, scrutinizing their role in amplifying rates of false-positive and false-negative findings. To achieve this, the study employs regularized linear regression as a proposed technique to counter overfitting, complemented by a thorough utilization of bias-variance analysis to discern the learning algorithms' performance. Utilizing a dataset centred around the flow of water from a dam and the corresponding water levels, the methodology showcases the efficacy of the chosen techniques in handling the complexities of the interaction effects. The paper[9] concludes by emphasizing the insightful contributions of bias-variance analysis in unravelling the intricacies of learning algorithm behaviour, shedding light on the detrimental impact of overfitting, and underscoring the imperative need for regularization to mitigate its effects. Results demonstrate the enhanced predictive capabilities of the proposed regularized linear regression method, especially when accompanied by feature normalization and polynomial regression cost functions. The bias-variance analysis elucidates the delicate trade-off between bias and variance within the models, providing a comprehensive understanding of their dynamics.

The aim of the study in paper[10] is to compare existing supervised machine learning approaches for predicting heart disease diagnosis and to improve the accuracy of KNN by changing K values. The study involves a comparative analysis of performance measures, including accuracy, precision, recall, and F1 Score, using different Train-Test Split Ratios into Data Sets. Seven ML methods and neural networks were developed and tested to assess their accuracy in predicting heart disease. The study found that the Random Forest Classifier provides high values for all performance measures, including Recall, Precision, Accuracy, and F1-Score. Additionally, the study observed that when the training ratio decreases or the test ratio increases, the scores of all performance measures decrease. The study concludes that ML algorithms, such as logistic regression, decision tree, random forest, and SVM, can be used

to forecast the diagnosis of heart disease. The outcomes of the study can be utilized for early detection and management of cardiac disease. Further research is recommended to analyze and evaluate additional machine learning algorithms to enhance precision and performance.

III. METHODOLOGY

A. Data Exploration and Cleaning:

In this phase, the dataset ("Bengaluru_House_Data.csv") was loaded into a Pandas DataFrame (df1). The initial exploration involved checking the shape and basic statistics of the dataset. Categorical features, such as 'area_type', 'society', 'balcony', and 'availability', were identified and subsequently dropped to focus on relevant information. The presence of missing values was addressed by dropping rows containing NaN values. The 'size' column was transformed to extract the number of bedrooms ('bhk'), and the 'total_sqft' column was cleaned and converted to numerical values.

B. Outlier Removal and Feature Engineering:

Outliers in the dataset were addressed through two main functions. The first function (remove_pps_outliers) focused on outliers in the price per square foot, using statistical measures for each location. The second function (remove_bhk_outliers) targeted outliers related to the number of bedrooms. These outlier removal processes ensured a more robust dataset. Feature engineering included creating a new feature, 'price_per_sqft,' and cleaning up the 'location' column by handling less frequent data points.

C. Exploratory Data Analysis (EDA):

This phase involved exploring the distribution of key features and understanding patterns within the dataset. Scatter plots were used to visualize the relationship between total square feet area and price for different bedroom configurations in specific locations. Statistical measures, histograms, and visualizations were employed to gain insights into the distribution of price per square foot and the number of bathrooms.

D. Model Training and Evaluation:

The project utilized Linear Regression as the predictive model. Data preparation included one-hot encoding for the 'location' column and splitting the dataset into features (X) and the target variable (y). The dataset was further divided into training and testing sets. The Linear Regression model was trained on the training set and evaluated on the test set. Cross-validation was performed to assess the model's performance under different splits, utilizing ShuffleSplit as a cross-validator.

E. Prediction Function and Results:

A prediction function (predict_price) was developed to estimate house prices based on input parameters, including location, square footage, bathrooms, and bedrooms. The trained Linear Regression model was employed for predictions. The project concluded with the presentation of results, showcasing the model's accuracy and the insights gained from the analysis.

IV. RESULTS

The machine learning project focused on predicting house prices in Bengaluru using a Linear Regression model. The accuracy of the model is 0.8452277697874357. Notable steps included outlier handling, feature engineering (such as creating the 'bhk' feature), and cleanup of location data for enhanced model generalization. Exploratory Data Analysis (EDA) visualizations provided insights into variable distributions. The model's success relies on its predictive accuracy for new data, and a user-friendly prediction function was implemented to estimate house prices based on location, square footage, bathrooms, and bedrooms.

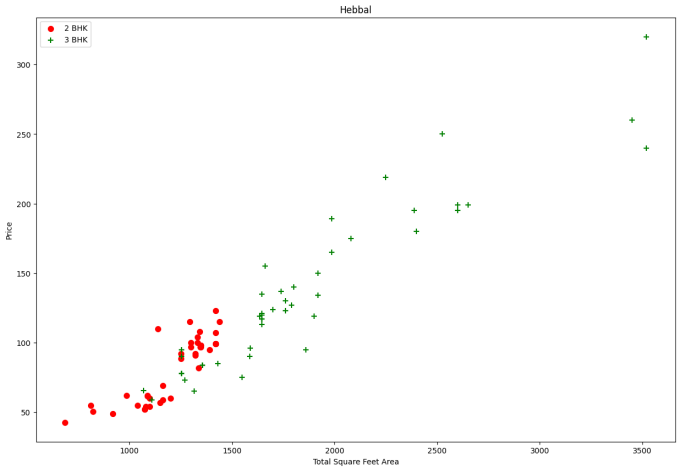

figure 1. Scatter chart before removing outliers

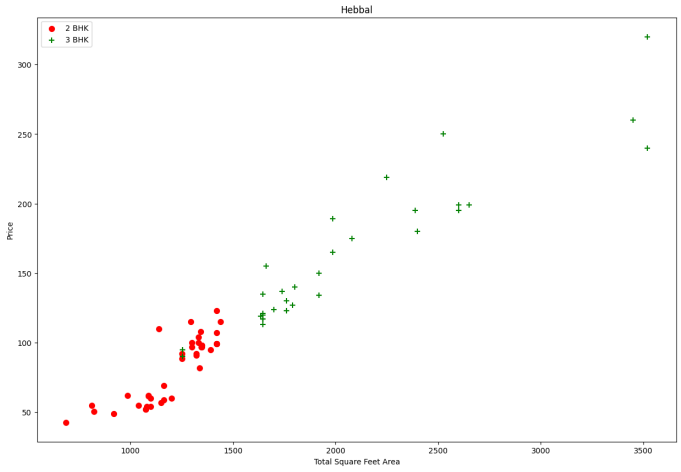

figure 2. Scatter chart after removing outliers

```
predict_price('1st Phase JP Nagar',1000, 2, 2)
✓ 0.0s

C:\Users\sneha\AppData\Local\Packages\PythonSoftwareFou
  warnings.warn(

83.49904677198772
```

V. CONCLUSION

In the course of the project, a comprehensive analysis of Bengaluru's housing market was undertaken, leveraging a dataset featuring crucial attributes such as total square feet area, bedroom count, and location. The exploratory data analysis uncovered pivotal insights, leading to the need for meticulous data cleaning and outlier removal. Innovative techniques, including the categorization of less frequent locations and the application of statistical measures to eliminate outliers, were employed to refine the dataset, bolstering the robustness of the predictive model.

The predictive model, anchored in Linear Regression, demonstrated promising accuracy throughout training and cross-validation, underscoring its efficacy in forecasting house prices in Bengaluru. The model not only functions as a valuable tool for predicting property values based on specific attributes but also enhances our understanding of the intricate dynamics shaping housing prices. As we conclude this project, the developed model emerges as a reliable resource for homebuyers and real estate professionals navigating the complexities of Bengaluru's real estate landscape. It lays the foundation for informed decision-making in this dynamic market, contributing to a more insightful and data-driven approach to property transactions.

VI. FUTURE SCOPE

Advanced Predictive Models: Implement more sophisticated machine learning models, such as ensemble methods (Random Forest, Gradient Boosting) or deep learning approaches, to potentially improve predictive accuracy.

Feature Engineering: Explore additional relevant features that could influence housing prices, such as proximity to amenities, transportation hubs, or the overall economic development of an area.

Dynamic Price Index: Develop a dynamic price index that considers changing economic conditions, inflation rates, and other external factors to provide more accurate and real-time predictions.

User Interface and Accessibility: Build a user-friendly interface or a mobile application that allows users to input property details and receive instant price predictions, making the model more accessible to a broader audience.

Integration with Real Estate Platforms: Collaborate with real estate platforms to integrate the predictive model, providing users with valuable insights while browsing property listings.

Regional Expansion: Extend the model to cover housing markets in other cities or regions, tailoring the analysis to the unique dynamics of each location.

Incorporate Time Series Analysis: Integrate time series analysis to capture trends and seasonality in housing prices, enabling more accurate long-term predictions.

In-depth Market Insights: Conduct in-depth market trend analysis, identifying patterns and factors that influence housing prices over time.

- Feedback Mechanism: Implement a feedback loop where users can provide information about actual property transactions, enabling continuous improvement and validation of the model.

- Collaboration with Stakeholders: Collaborate with real estate experts, government bodies, and urban planners to incorporate domain knowledge and ensure the model aligns with industry standards and regulations.

VIII. REFERENCES

[1] Chiu, SM., Chen, YC. & Lee, C. Estate price prediction system based on temporal and spatial features and lightweight deep learning model. Appl Intell 52, 808–834 (2022).

[2] Bhagat, Ayushi and Gosavi, Mayuri and Shahasane, Aditi and Mishra, Nandini and Nerurkar, Amit, House Price Prediction using Machine Learning (April 9, 2023).

[3] 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, doi: 10.1109/ICAC3N53548.2021.9725552.

[4] A. Soltani, et al., Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms, Cities 131 (2022), 103941

[5] Truong, Q., Nguyen, M., Dang, H., et al., 2020. Housing price prediction via improved machine learning techniques. Proc. Comput. Sci. 174 (Jun.)

[6] A. Chaurasia and I. U. Haq, "Housing Price Prediction Model Using Machine Learning," 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), Ghaziabad, India, 2023, pp. 497-500, doi: 10.1109/ICSEIET58677.2023.10303359.

[7] C. Zhan, Z. Wu, Y. Liu, Z. Xie and W. Chen, "Housing prices prediction with deep learning: an application for the real estate market in Taiwan," 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), Warwick, United Kingdom, 2020, pp. 719-724, doi: 10.1109/INDIN45582.2020.9442244.

[8] H. Prakash, K. Kanaujia and S. Juneja, "Using Machine Learning to Predict Housing Prices," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 1353-1357, doi: 10.1109/AISC56616.2023.10085264.

[9] M. Rajasekhar Reddy, B. Nithish Kumar, Madhusudana Rao Nalluri, and B. Karthikeyan, “A New Approach for Bias–Variance Analysis Using Regularized Linear Regression”, Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals. Springer Singapore, Singapore, pp. 35-46, 2020.

[10] R. M Sampreeth, Sravani, N., and Dr Tripty Singh, “Real Estate Price Prediction”, in International Conference on Recent Trends in Electronics, Information & Communication Technology, Sri Venkateshwara College of Engineering, Bengaluru, India, 2019.