House Price Prediction Model

K Nikhila
21CE31007

Duration: May '24 – Jun '24

1. Introduction

This project focuses on developing a data-driven model to predict house prices in Bengaluru, India. In the real estate industry, accurate price prediction is a critical component for both buyers and sellers, enabling them to make informed decisions. This model leverages a dataset of real estate listings from Kaggle, employing a range of data science and machine learning techniques to account for key factors influencing property value. The final output is a robust and tested prediction model that can be used to estimate a home's price based on its characteristics, with an impressive accuracy rate of 84.77%.

2. Objectives

The primary objectives of this data science project were to:

- Clean and preprocess a raw dataset to handle missing values, inconsistencies, and outliers.
- Perform feature engineering to create new, relevant variables that can enhance the model's predictive power.
- Implement dimensionality reduction to effectively manage categorical features with many unique values.
- Develop and evaluate a regression model capable of accurately predicting house prices.
- Validate the model's accuracy using cross-validation techniques.
- Export the final model for potential integration into a real-world application.

3. Scope of Works

The scope of this project encompassed the full lifecycle of a data science regression task:

- Data Acquisition and Exploration: A dataset containing Bengaluru house prices was downloaded from Kaggle. Initial exploration was performed to understand the dataset's structure, including its 13,320 rows and 9 columns covering attributes like area_type, location, total_sqft, and price.
- Data Cleaning: Features deemed irrelevant for the model, such as area_type, society, balcony, and availability, were removed. The remaining dataset was cleaned by handling all missing (NA) values, specifically in the bath and size columns.
- Feature Engineering and Outlier Removal: New features were created, and a multi-step process was used to remove outliers, including applying business logic and statistical methods to ensure data integrity and model reliability.
- Model Building and Training: The preprocessed data was used to train a machine learning model, with 80% of the data designated for training and 20% for testing.

- Validation and Evaluation: The model's accuracy was assessed on the test set and through K-fold cross-validation.
- Model Export: The final model and the associated column data were exported to files for potential use in a production environment.

4. Methods Used

The project utilized a structured methodology to ensure the development of an accurate and reliable model. The entire process was documented in a Jupyter Notebook using Python, with key libraries including Pandas for data manipulation and scikit-learn for machine learning.

Data Preprocessing and Feature Engineering:

- The initial dataset was loaded, and four columns (area_type, society, balcony, availability) were dropped to simplify the model.
- Missing values were handled by dropping all rows with NA entries.
- The total_sqft column contained values in various formats, including ranges like "2100 - 2850". A function was created to convert these ranges to their average value (e.g., "2475.0") and drop other irregular string values.
- A new feature, bhk, was created by extracting the numerical bedroom count from the size column (e.g., "2 BHK" becomes "2").
- A new feature, price_per_sqft, was calculated to assist with outlier detection and removal.

Outlier Removal:

- Business Logic: Based on business knowledge that a typical bedroom is at least 300 sqft, properties with less than 300 sqft per bedroom were identified as outliers and removed.
- Statistical Methods: To address extreme price variations, a function was developed to remove outliers based on price_per_sqft for each location. It filtered out any property with a price_per_sqft value outside of one standard deviation from the mean for its specific location.
- Bedroom-Bathroom Logic: An additional outlier removal step was applied based on the principle that the number of bathrooms in a home is typically not more than two greater than the number of bedrooms (bath > bhk + 2). Listings that violated this rule were removed.

5. Results

The model was successfully trained and evaluated, and the following results were obtained:

- Model Accuracy: The Linear Regression model achieved an accuracy score of 86.29% on the test data. This is a strong result, indicating a high level of predictive capability.
- Cross-Validation: To ensure the model's stability and prevent overfitting, K-fold cross-validation (ShuffleSplit with 5 splits) was performed. The scores from the five iterations were consistent, reinforcing the model's reliability. The average accuracy rate across the folds was 84.77%.
- Model Prediction: The predictive function was successfully tested with example properties in different locations. For a 1000 sqft, 2-bedroom home in 1st Phase JP Nagar, the predicted

price was ₹83.86 lakhs. For a similar home in Indira Nagar, the predicted price was ₹193.31 lakhs, demonstrating the model's ability to capture the significant impact of location on price.

6. Discussion

The project successfully navigated a common challenge in data science: dealing with real-world, messy data. The iterative process of data cleaning and outlier removal based on both statistical analysis and business rules was crucial to building a reliable model. The use of One Hot Encoding for the location feature was effective in converting a highly cardinal categorical variable into a format that the linear regression model could understand. The high accuracy score and consistent cross-validation results validate the model's performance and demonstrate its potential for practical use. The final model, saved as a pickle file, along with the column information, provides a deployable solution for a prediction application.

7. Conclusion

This project demonstrates the complete process of building a machine learning model for house price prediction in Bengaluru. By focusing on meticulous data cleaning, feature engineering, and robust model validation, a Linear Regression model was developed that achieved a high accuracy rate of approximately 84.77%. The model effectively predicts house prices based on key features, including location, square footage, number of bathrooms, and bedrooms. The success of this project highlights the practical application of data science principles in the real estate sector.

8. References

Building Predictive Models with Machine Learning

Link:https://www.researchgate.net/publication/379097335_Building_Predictive_Models_with_Machine_Learning

Declaration by the Student

I, K Nikhila, hereby declare that the work presented in this report is an authentic record of my own self-initiated project. This work has been completed in accordance with the guidelines provided for a self-study project. The material has not been submitted to any other university or institution for any academic purpose.

Date: K Nikhila

Place: IIT Kharagpur 21CE31007