Real Estate Price Prediction Project

Author: Imon Deb

Date: 15/9/2024

1. Project Overview

This project aims to predict real estate prices based on features like location, square footage, number of bathrooms, and number of bedrooms (BHK). Using a dataset containing property listings, I developed a machine learning model that accurately estimates the price of a house given relevant input features.
The project involved data preprocessing, feature engineering, handling outliers, and training several regression models to identify the best-performing model. The final model uses key predictors and applies scaling to ensure optimal performance.

2. Dataset Description

2.1. Source

The dataset consists of real estate listings of Bengaluru from kraggle and contains information about various properties, including location, area type, size, number of rooms, and the price of the property.

2.2. Features

- location: Categorical feature representing the area where the property is located. One-hot encoding was used to represent different locations.
- total_sqft: Numeric feature representing the size of the property in square feet.
- BHK: Numeric feature indicating the number of bedrooms in the property.
- bath: Numeric feature representing the number of bathrooms.
- price: Target variable representing the property price.

Other features such as availability, area_type, and balcony were included initially but were later excluded due to low correlation with the target variable, as explained further in Section 4.

3. Data Preprocessing

3.1. Handling Missing Values

Missing values in the dataset were identified and handled appropriately. For features like bath, missing values were imputed based on the median or mode of relevant categories. Any rows with critical missing data were removed. Duplicate rows in the dataset were identified and removed to ensure that each entry was unique and did not skew the analysis.

3.2. Outlier Detection and Treatment

Outliers were handled in multiple stages:
- Price per Square Foot Analysis: Outliers in price_per_sqft were detected based on domain-specific logic, and extreme anomalies (such as properties with prices disproportionate to their size) were flagged and removed.
- Custom Outlier Handling: Rules such as removing properties where total_sqft/BHK < 250 were applied to avoid unrealistic data points.

The final approach kept data points within a 1 standard deviation threshold from the mean price per square foot.

3.3. One-Hot Encoding

For the location feature, one-hot encoding was applied. Locations with fewer than 10 listings were grouped into a single "Other" category to prevent sparsity.

4. Feature Selection and Engineering

4.1. Feature Importance

Exploratory Data Analysis (EDA) revealed that features such as availability, area_type, and balcony had minimal correlation with the target variable. As a result, these features were excluded from the model training process to avoid adding noise. The "society" feature was dropped due to having approximately 50% missing values. This decision was made to ensure the quality and reliability of the dataset.

4.2. Log Transformation

The price feature was log-transformed to address heteroscedasticity in the data, leading to better model performance and more stable predictions.

4.3. Standardization

To ensure that features like total_sqft, BHK, and bath were on a comparable scale, standardization was applied. This helped improve the performance of machine learning models, especially in algorithms sensitive to feature scaling (e.g., linear regression).

5. Model Selection and Training

5.1. Model Comparison

Multiple regression models were evaluated using GridSearchCV for hyperparameter tuning:
- Linear Regression
- Lasso Regression
- Ridge Regression

- Random Forest Regression

These models were compared based on cross-validated performance scores. Lasso Regression was included to handle feature selection by penalizing less important features.

5.2. Final Model

The best-performing model was selected based on the highest accuracy and lowest error metrics, which was Linear Regression. The final model's accuracy improved after effectively dealing with outliers and log transforming target variable price.

6. Results and Inference

6.1. Key Predictors

The following features were found to be the most important predictors for determining real estate prices:

- Location: One of the strongest predictors, as prices vary significantly across different areas.
- Total Square Footage: Directly proportional to price.
- BHK and Bathrooms: Positively correlated with price, with more bedrooms and bathrooms generally leading to higher prices.

6.2. Model Performance

The final model achieved an accuracy of 0.8358676487455482 and exhibited low residuals when predicting prices on the test set. The predictions were generally well-aligned with actual prices, though some variability remained in the upper price range.

6.3. Insights

- Location Impact: Premium locations exhibited a much higher price per square foot than less developed or suburban areas.
- Size Correlation: Larger properties (total_sqft) consistently commanded higher prices, although this trend was less pronounced in very high-end locations where the price per square foot overshadowed size.
- Non-Predictive Features: Features like availability and balcony did not contribute meaningfully to price prediction and were excluded from the final model to avoid unnecessary complexity.

7. Future Work

Several potential improvements can be made in future iterations of the project:

- Additional Features: Including amenities like proximity to schools, transportation, and other infrastructure could further improve the accuracy of price predictions.
- Time Series Analysis: Incorporating time-based factors like market trends or the year of construction could provide more temporal insights into price movements.
- Advanced Modeling: Exploring advanced machine learning models such as XGBoost or

Random Forest might yield better results, especially for capturing non-linear relationships in the data.

8. Conclusion

This project successfully developed a machine learning model to predict real estate prices using essential features like location, size, and number of rooms. The final model offers a reliable tool for estimating property prices based on readily available data and can be expanded further with additional features and model enhancements.