Project Report
Mumbai Rent Guru

Project Title: Mumbai Rent Guru
Your Name: Mihir Panchal
Date: July 2024

Abstract

This project addresses the challenge of predicting housing prices in Mumbai, a key concern for potential buyers and investors. Utilizing a linear regression model, the project develops a systematic approach to analyze various features affecting housing prices. Key findings reveal the effectiveness of the model in accurately predicting prices based on specified features, demonstrating the potential of data-driven methodologies in real estate.

Introduction

The real estate market in Mumbai is characterized by its complexity and volatility, making accurate price prediction essential for buyers and investors. This project aims to develop a predictive model that leverages historical housing data to estimate prices based on various influential features.

Objectives

- To preprocess and clean the Indian housing dataset for effective analysis.
- To build a linear regression model for predicting housing prices.
- To evaluate the model's performance using relevant metrics.

Importance and Relevance

Accurate housing price predictions can facilitate informed decision-making in the real estate market, enabling stakeholders to make better investments and policy decisions. This project highlights the significance of applying machine learning techniques to real-world problems.

Methodology

Description of Methods and Technologies Used

The project employed Python along with libraries such as Pandas, NumPy, scikit-learn, and Pickle for data manipulation, modeling, and saving the trained model. The methodology included data loading, preprocessing, model development, and evaluation.

Steps Taken

1. Data Loading and Preprocessing:

- Data Import: The dataset was loaded from a CSV file *('Indian_housing_Mumbai_data.csv')*.

- Exploratory Data Analysis (EDA): Basic EDA was conducted using *`describe()`, `info()`, and `shape`* to understand the dataset.

- Feature Removal: Unnecessary columns (latitude, longitude, isNegotiable, priceSqFt, description, currency, city) were dropped to focus on relevant features.

- Handling Missing Values: Missing values in *`numBathrooms`* were replaced with the mean, while *`numBalconies`* missing values were filled with 0.

- Data Cleaning: Various columns were cleaned for consistency, including *'house_type', 'house_size', 'verificationDate',* and *'SecurityDeposit'*.

- Data Type Conversion: Certain columns were converted from object to integer type for proper analysis.

- Categorical Encoding: The *'house_type'* column was simplified for better readability.

2. Modeling:

- Train-Test Split: The dataset was split into *training (80%)* and *testing (20%)* sets using *`train_test_split`*.

- One-Hot Encoding: Categorical features were encoded using *`OneHotEncoder`*.

- Column Transformer: A *`make_column_transformer`* was employed to apply transformations only to categorical columns while passing numerical features unchanged.

- Linear Regression: A linear regression model was established to predict housing prices.

- Pipeline: A *`make_pipeline`* was used to streamline the process of transformation and model training.

- Model Training and Evaluation: The model was trained on the training data, and predictions were made on the test set, evaluated using metrics such as R^2 *Score, Mean Absolute Error (MAE)*, and *Mean Squared Error (MSE)*.

- Hyperparameter Tuning: A loop was implemented to find the best *`random_state`* for the train-test split, optimizing model performance.

- Model Saving: The trained pipeline was saved using *`pickle`* for future use.

Implementation

The implementation process began with data import and exploratory analysis, followed by comprehensive data preprocessing. The model was developed through a structured workflow, employing various techniques to ensure effective predictions.

Challenges Faced

- Handling Missing Values: Addressing missing values required careful analysis to ensure minimal impact on model accuracy.

- Data Cleaning: Ensuring consistent data formats and types presented challenges, particularly with varied inputs in categorical columns.

Results

The linear regression model achieved a R^2 Score of ***76.24%*** on the test set, indicating a strong predictive capability.

Data Visualization

- Scatter Plots: Scatter plots of predicted versus actual prices illustrate the model's performance.

- Error Analysis: A histogram of prediction errors helped identify any biases or patterns in the model's predictions.

Analysis of Results

The results indicate that the model effectively captures the relationship between housing features and prices. The evaluation metrics confirm the model's reliability and potential application in real estate decision-making.

Conclusion

This project successfully developed a linear regression model to predict housing prices in Mumbai. The approach demonstrates the power of machine learning in addressing real-world issues and provides a foundation for further research and application in real estate analytics.

Recommendations for Future Work

Future efforts could explore advanced modeling techniques, such as regularization methods or ensemble learning, to further enhance prediction accuracy. Additionally, integrating real-time data could improve the model's applicability in dynamic market conditions.