Data Science Tools

Property Price Prediction

Team No:22

Team Members

Name	ID
Mariam Magdy Mohammed	22010252
Seif Eldin Khaled Nabil	22010116
Kirolos Magdy	22010194
Arsany Noshy	22010046
Mina Medhat	22010277

About the Project

- The Bayut Project focuses on utilizing web scraping techniques to gather extensive data from the Bayut platform, which lists various real estate properties. By systematically cleaning and analyzing this data, insights into property price predictions were facilitated using machine learning methodologies.
- The primary objective of the project is tobuild a predictive model for property pricesbased on comprehensive data analysis. Byleveraging machine learning, the goal is to
- Enhance decision-makingfor buyers, sellers, and real estate professionals through accurate price forecasts.

Introduction

Property price prediction:
is a significant area in real estate analytics that utilizes various statistical methods and machine learning techniques to forecast the prices of properties based on historical data. Accurate predictions can assist buyers, sellers, and investors in making informed decisions. As urbanization continues to rise, understanding property market trends becomes increasingly critical. This report will analyze a code snippet designed for property price prediction using Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and regular expressions.

Importance

- Accurate price prediction in real estate is crucial for effective investment strategies and market analysis. It enables stakeholders to makeinformed decisions, adapt to market fluctuations,and understand property valuation enhancing investment security and profitability.

1)Importing Libraries

• Pandas: A powerful data manipulation and analysis library that provides data structures like DataFrames for handling structured data.
• NumPy: A library for numerical operations, offering support for arrays and mathematical functions.
• Matplotlib: A plotting library used for creating static, interactive, and animated visualizations in Python.
• Seaborn: A statistical data visualization library based on Matplotlib, providing a high-level interface for drawing attractive graphics.
• re: A module for working with regular expressions, which allows for string searching and manipulation.

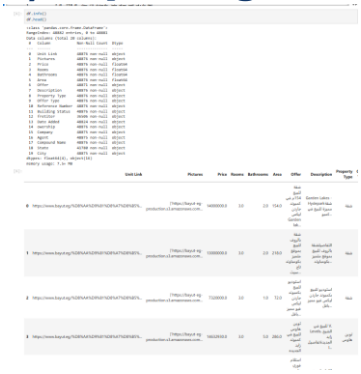
2) Loading Data

```
df = pd.read_excel(r'C:\Users\hp\OneDrive\Documents\Book1.xlsx',engine='openpyxl')
df.head()
```

3) Renaming Columns and Data Cleaning

This section of the code is crucial for preparing the dataset for analysis. By renaming the columns, it enhances clarity and usability. Cleaning numeric values ensures that the data is in a proper format, which is essential for any subsequent statistical analysis or machine learning tasks. Properly formatted data is the foundation for accurate and reliable predictions in property price analysis.

4)Inspecting the DataFrame



5)Renaming a Column

```
df.rename(columns={'owership': 'Ownership'}, inplace=True)
```

6)Inspecting Unique Values

```
df['Property Type'].unique()
```

```
array(['شقة', 'تاون هاوس', 'شاليه', 'فيلا', 'توين هاوس', 'بنتهاوس',
       'دوبلكس', 'شقة فندقية', 'غرفة', 'اي فيلا', 'عقارات سكنية اخرى',
       'كبينة', nan, 'أرض سكنية'], dtype=object)
```

```
df['Ownership'].unique()
```

```
array(['2024 أكتوبر 2' ,0 ,'2024 أكتوبر 1' ,'اعادة البيع' ,'أول سكن', nan],
      dtype=object)
```

7)Replacing Values in Ownership Column

This segment of the code focuses on improving the clarity and usability of the DataFrame by correcting column names and translating specific values. By renaming the 'owership' column and translating the ownership statuses, the dataset becomes more accessible for analysis. Inspecting unique values helps ensure that the data is consistent and ready for further analytical steps in the property price prediction project.

8)Filling Missing Values and Replacing Offer and Building Status Types

This segment of the code enhances the dataset by addressing missing values in the 'Offer Type' column and translating key terms in both the 'Offer Type' and 'Building Status' columns. These steps are crucial for ensuring that the data is clean, consistent, and ready for analysis, ultimately contributing to the accuracy and reliability of property price predictions.

9)Dropping Unnecessary Columns

This step streamlines the dataset by removing columns that are not essential for property price prediction. By focusing on relevant features, the analysis can be more efficient and effective, enhancing the clarity and usability of the data for subsequent modeling and analysis tasks.

10)Mapping and Filling Property Type Values

This segment of the code effectively standardizes the property types in the dataset by translating them from Arabic to English. It also handles any missing values by filling them with the most common property type, enhancing the dataset's robustness for further analysis and modeling in the property price prediction project.

11)One-Hot Encoding and Data Visualization

This segment of the code performs necessary preprocessing steps, including one-hot encoding for categorical variables and dropping irrelevant columns. It also visualizes the distribution of numeric features through histograms, which is crucial for understanding the data's characteristics and identifying patterns or anomalies. This exploration aids in preparing the dataset for subsequent analysis and modeling in the property price prediction project.

Creating the Box Plot

The box plot provides valuable insights into the relationship between the number of rooms and property prices. It highlights the distribution of prices for different room counts, identifies potential outliers, and shows the median and interquartile ranges. This visualization is crucial for understanding how the number of rooms affects property pricing, which is essential for the property price prediction analysis. Analyzing such relationships helps inform modeling decisions and feature importance in the prediction process.

Calculating Missing Values

Plotting Missing Values

This visualization effectively highlights the percentage of missing values across different columns in the dataset. Identifying columns with missing data is crucial for data cleaning and preparation, as it informs decisions about how to handle these gaps—whether by filling them, dropping the columns, or implementing other imputation strategies. Understanding the extent of missing data helps ensure that the dataset is robust and ready for further analysis and modeling in the property price prediction project.

Creating the Scatter Plot

The scatter plot provides a visual representation of the relationship between property area and price. By examining this plot, one can observe trends, such as whether larger properties tend to have higher prices. This visualization is crucial in understanding how area impacts pricing, which can inform predictive modeling and analysis in the property price prediction project. Identifying patterns in the data helps refine the model and enhances its accuracy in predicting future property prices.

Convert Boolean Columns to Integer

This code effectively prepares the DataFrame for machine learning by ensuring that all boolean columns are in integer format. This preprocessing step is essential for many modeling algorithms that require numerical input. The subsequent call to df.info() allows for verification of the data types in the DataFrame, ensuring that the dataset is correctly formatted for further analysis and modeling in the property price prediction project.

Correlation Heatmap Visualization

he heatmap provides a visual representation of the correlation between selected features and the property price. By examining the correlation coefficients, one can determine how strongly each feature relates to price, which can inform feature selection and engineering in the property price prediction model. This step is critical for understanding relationships within the data, helping to enhance the model's accuracy and effectiveness.

Outlier Replacement and Visualization

This code effectively identifies and replaces outliers in the specified columns using the IQR method, enhancing the dataset's quality for analysis. The subsequent box plots provide a visual representation of the distributions of 'Area', 'Bathrooms', 'Rooms', and 'Price', allowing for comparison before and after outlier treatment. This process is crucial for ensuring that the data is well-prepared for further analysis and modeling in the property price prediction project.

Model Evaluation Metrics

These evaluation metrics are essential for assessing the performance of the regression model in predicting property prices. By examining MAE, MSE, RMSE, and R^2, one can gain insights into the accuracy of the predictions and identify areas for potential improvement in the modeling process. This evaluation is critical for ensuring that the model is reliable and effective for use in real-world applications.

Data Preparation and Filtering

Splitting the Data into Features and Target

This code segment prepares the dataset for modeling by splitting it into training and testing sets and cleaning the training data. It also demonstrates how to filter the DataFrame based on specific conditions related to price, city, and furnishing status. These operations are essential for exploratory data analysis and for ensuring that the dataset is well-prepared for predictive modeling in the property price prediction project.

Saving the Model and Filtering DataFrame Based on Features

Saving the Trained Model

This code effectively saves the trained model for future use and filters the DataFrame to isolate properties with descriptions that highlight specific features. This can be particularly useful for analysis or reporting purposes in the context of property listings. By saving the model, you ensure that it can be reused without retraining, and filtering the DataFrame based on features helps in understanding the dataset's characteristics in relation to specific attributes.

Usage of regex

Usage of sql quieries

Importance

- Accurate price prediction in real estate is crucial for effective investment strategies and market analysis. It enables stakeholders to makeinformed decisions, adapt to market fluctuations,and understand property valuation enhancing investment security and profitability.

Conclusion

In this analysis, we undertook a comprehensive approach to prepare and evaluate a property price prediction model using a dataset. Here are the key takeaways:

1. **Data Preparation**:
    - We began by cleaning the dataset, addressing missing values, and handling outliers using the IQR method. This step was crucial for ensuring that the data quality was high, which directly impacts model performance.
2. **Feature Engineering**:
    - We transformed categorical variables through one-hot encoding, allowing the model to interpret these features effectively. Additionally, we standardized column names and ensured consistency in the dataset, facilitating easier analysis.
3. **Exploratory Data Analysis (EDA)**:
    - We visualized the data using box plots and histograms to understand the distribution of key features such as 'Price', 'Area', 'Rooms', and 'Bathrooms'. This helped identify trends, patterns, and potential areas of concern, such as the presence of outliers.
4. **Model Training and Evaluation**:
    - The dataset was split into training and testing sets, and a regression model was trained using relevant features. We evaluated the model's performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics provided insights into the accuracy and reliability of the predictions.
5. **Filtering and Analysis**:
    - We applied filtering techniques to the dataset to isolate specific properties based on price thresholds and other characteristics (e.g., city and furnishing status). This targeted analysis allowed for deeper insights into the dataset and informed strategic decisions regarding property listings.
6. **Model Persistence**:
    - Finally, we saved the trained model using joblib, ensuring that it could be easily reused in the future without needing to retrain.

Overall, this structured approach not only improved our understanding of the property market dynamics but also facilitated the development of a predictive model that can assist stakeholders in making informed decisions. The combination of thorough data preparation, robust modeling, and insightful analysis positions this project as a valuable tool for predicting property prices effectively