Housing Price Analysis and Prediction Project

--

Project Overview

This project focuses on analyzing and predicting housing prices using Databricks, PySpark, and Spark SQL. The dataset includes various features related to housing, such as location, property type, area, and price category. The main goal is to preprocess the data, perform feature engineering, and build a predictive model to estimate housing prices.

Objectives

1. **Data Cleaning -** Remove duplicates and handle missing values to ensure data quality.
2. **Data Transformation -** Perform feature engineering and encode categorical variables for analysis.
3. **Exploratory Data Analysis (EDA) -** Understand the dataset through descriptive statistics and correlation analysis.
4. **Model Building -** Train machine learning models to predict housing prices.
5. **Prediction** - Use the trained model to make predictions on new data.
6. **Visualization** - Visualize the results and insights for better understanding and interpretation.

Methodology

1. **Data Loading**

   - The dataset is loaded from an Amazon S3 bucket into Databricks for processing using PySpark.

2. **Data Cleaning**

    - Duplicates are removed, and missing values are handled to ensure the dataset is clean.
    - Data types are corrected where necessary to facilitate accurate analysis.

3. **Data Transformation**

    - Feature engineering is performed to prepare the data for modeling.
    - Categorical variables are encoded using StringIndexer and OneHotEncoder to convert them into numerical form.

4. **Exploratory Data Analysis (EDA)**

    - Descriptive statistics are generated to summarize the dataset.
    - Correlation analysis is conducted to understand the relationships between different features.

5. **Model Building**

    - The features are assembled using VectorAssembler.
    - The dataset is split into training and testing sets.
    - A linear regression model is trained on the training data to predict housing prices.

6. **Prediction**

    - The trained model is used to make predictions on the test data.
    - The predictions are compared with actual prices to evaluate the model's performance.

7. **Visualization**

- The results and insights are visualized using Databricks visualization tools to aid interpretation and decision-making.

Challenges

1. **Handling Missing Values**

   The dataset had missing values that needed to be addressed to avoid biased results.

   - Solution: Used the fillna() method to handle missing values appropriately.

2. **Categorical Variables**

   Converting categorical variables into a numerical format without losing information was crucial.

   - Solution: Applied StringIndexer and One-Hot Encoding to transform categorical variables.

3. **Model Performance**

   Ensuring the model's accuracy and generalization to unseen data.

   - Solution: Split the data into training and testing sets and used cross-validation techniques.

4. **Handling Outliers**

   Outliers in the data can significantly skew the results and affect the model's performance.

- Solution: Used statistical methods to identify outliers and apply appropriate techniques like capping or transformation to mitigate their impact.

Recommendations and Future Steps

1. **Feature Enhancement -** Incorporate additional features that may influence housing prices, such as economic indicators or geographical data, to improve the model's accuracy.
2. **Model Experimentation -** Experiment with different machine learning models (e.g., decision trees, random forests) to compare their performance and identify the best model for prediction.
3. **Real-time Data -** Implement real-time data processing to update the model continuously with the latest data for more timely and relevant predictions.

Conclusion

The project successfully demonstrated the process of analyzing and predicting housing prices using Databricks, PySpark, and Spark SQL. The final model provided valuable insights and predictions for housing prices. Future steps include further tuning of the model, adding more features, and exploring different machine learning algorithms to improve accuracy.