Report : House Price Prediction Using Regression Models

Introduction

The real estate market has always been a complex domain influenced by various factors such as location, property size, age, and amenities. Accurately predicting house prices is essential for buyers, sellers, and real estate agents. This report outlines a project aimed at predicting house prices using multiple regression techniques. The project utilizes a dataset containing various features of houses and their corresponding sale prices.

Dataset Overview

The dataset used for this analysis is derived from the **House Prices: Advanced Regression Techniques** competition on Kaggle. It contains numerous features that contribute to house pricing, including but not limited to:

- **GrLivArea**: Above ground living area in square feet.
- **BedroomAbvGr**: Number of bedrooms above ground.
- **Neighborhood**: The location of the property.
- **OverallQual**: Overall material and finish quality.
- **YearBuilt**: Year the house was built.
- **SalePrice**: The target variable representing the sale price of the house.

The dataset consists of both numeric and categorical variables, necessitating a thorough exploration and preprocessing phase to ensure optimal model performance.

1.Exploratory Data Analysis (EDA)

Data Loading

The first step involved loading the dataset and checking for missing values and data types.

python

Data Visualization

To understand the relationship between features and the target variable, a correlation heatmap was generated. The numeric features were filtered to create a heatmap visualizing correlations.

Feature Selection

Based on the correlation heatmap, relevant features for the regression model were identified:

- GrLivArea
- BedroomAbvGr
- Neighborhood (to be encoded)
- OverallQual
- YearBuilt

The features were selected to capture the size, location, quality, and age of the houses, which are critical factors affecting house prices.

2.Data Preprocessing

The next step involved preprocessing the data, including handling missing values, encoding categorical variables, and splitting the data into training and testing sets.

Handling Missing Values

Missing values were handled through imputation or removal, depending on the feature's significance.

Encoding Categorical Variables

Categorical features such as Neighborhood were encoded using one-hot encoding to convert them into numerical values.

3.Data Splitting

The dataset was split into training and testing sets, with a typical split of 80% for training and 20% for testing.

4.Model Implementation

Various regression models were implemented to predict house prices, including:

1. **Linear Regression**
2. **Decision Tree Regression**
3. **Random Forest Regression**

5.Model Training and Evaluation

Each model was trained on the training data and evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to gauge performance.

Results Summary

The performance of the models was summarized as follows:

- **Linear Regression**: MAE = XX, RMSE = XX
- **Decision Tree Regression**: MAE = XX, RMSE = XX
- **Random Forest Regression**: MAE = XX, RMSE = XX

These results indicate the effectiveness of the chosen features and the suitability of different regression techniques for this task.

Conclusion and Future Work

This project demonstrated the importance of feature selection and data preprocessing in building predictive models for house prices. The models implemented provided valuable insights into the factors influencing prices in the real estate market.

Future work may involve:

- Exploring additional features such as garage size, lot area, and other amenities.
- Implementing advanced algorithms like Gradient Boosting or XGBoost.
- Fine-tuning the models through hyperparameter optimization to enhance prediction accuracy.