House Price Prediction Project Report

June 28, 2025

Introduction

This project aims to predict house sale prices using supervised machine learning techniques. The dataset includes various features such as lot area, year built, number of bedrooms, garage area, living area, number of bathrooms, overall quality, and neighborhood, with the target variable being the sale price.

Data Overview

- **Dataset Shape**: 500 rows and 10 columns.
- **Features**: LotArea, YearBuilt, BedroomAbVGr, GarageArea, GrLivArea, FullBath, HalfBath, OverallQual, Neighborhood, SalePrice.
- **Sample Data**:
  - LotArea: 6619-13512 sq ft
  - YearBuilt: 1958-2016
  - BedroomAbVGr: 1-4
  - GarageArea: 249-772 sq ft
  - GrLivArea: 921-2372 sq ft
  - FullBath: 1-3
  - HalfBath: 0-1
  - OverallQual: Varies
  - Neighborhood: Varies
  - SalePrice: 100000-500000

Methodology

Step 1: Import Libraries

Utilized Python libraries including pandas, numpy, matplotlib, seaborn, sklearn for data manipulation, visualization, and modeling.

Step 2: Load and Explore Dataset

The dataset was loaded from an Excel file and initial exploration revealed the structure and sample data points.

Step 3: Data Visualization

- **Correlation Heatmap**: A heatmap was generated to visualize the correlation between features. Strong positive correlations were observed between SalePrice and GrLivArea (0.05), and moderate correlations with GarageArea (-0.02).
- **Scatter Plots**:
  - **Garage Area vs Sale Price**: Showed a scattered distribution with no strong linear trend.
  - **Living Area vs Sale Price**: Indicated a positive trend with higher living areas correlating with higher sale prices.

Step 4: Model Development

- **Linear Regression and Random Forest Regressor** were implemented to predict SalePrice.
- **Train-Test Split**: The dataset was split for training and testing the models.
- **Evaluation Metrics**:
  - Mean Absolute Error (MAE): 954495.93
  - Root Mean Squared Error (RMSE): 111782.83

Step 5: Model Evaluation

A scatter plot compared actual vs. predicted sale prices using the Random Forest model, showing the model's performance.

Results

The correlation heatmap highlighted that GrLivArea and GarageArea have the most significant influence on SalePrice. The scatter plots confirmed a positive relationship between living area and sale price, while garage area showed a weaker correlation. The Random Forest model provided reasonable predictions, though the high MAE and RMSE suggest room for improvement in model accuracy.

Conclusion

The project successfully implemented a machine learning approach to predict house prices, with visualizations aiding in understanding feature importance. Future improvements could involve feature engineering, handling outliers, and testing additional models to reduce error metrics.

Recommendations

- Enhance the dataset with more diverse features (e.g., location specifics, additional quality metrics).
- Apply data preprocessing techniques like normalization and outlier removal.
- Explore advanced models like Gradient Boosting or Neural Networks for better accuracy.