Housing Price Prediction Project

Name: Neel Adalja

SME: Ram Gupta

Date: Jan 02, 2025

**Model Report: Housing Price Prediction Project**

**Objective**

The objective of this project is to develop a machine learning model to predict housing prices based on various features such as Neighborhood, Lot Area, and amenities. The goal is to achieve a high-performing model to assist stakeholders in pricing decisions and market analysis.

**Dataset Overview**

* **Source:** [https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset]
* **Size:** 82 columns, 2930 rows
* **Target Variable:** SalePrice
* **Features:** Includes numerical, categorical, and binary data. Examples:
  + LotArea (Numerical)
  + Neighborhood (Categorical)
  + YearBuilt (Numerical)
  + BldgType (Categorical)
  + FullBath(Numerical)
  + BedoomAbvGr(Numerical)

**Data Preprocessing**

1. **Missing Values:**
   * Imputed missing numerical values with median.
   * Imputed categorical features with the mode.

A graph of a distribution of sales

Description automatically generatedA graph of a distribution of sales

Description automatically generated

* Before and after removing outlier in the dataset.

A graph of a box plot

Description automatically generatedA blue rectangular box diagram with numbers and lines

Description automatically generated with medium confidence

* Removed rows with sales prices above 300500 to normalize the data.

1. **Feature Engineering:**
   * Transformed and normalized the Bldg Type column to reduced unwanted categories.
   * Log transformation applied to skewed features like LotArea.
2. **Encoding:**
   * Categorical variables encoded using one-hot encoding (e.g., Neighborhood, BldgType).
3. **Train-Test Split:**
   * Split dataset into training (80%) and testing (20%) sets.

**Best Model**

* **Algorithm:** [Random Forrest Regressor]
* **Hyperparameters:** [n\_estimators = [100]]
* **Performance Metrics:**
  + Random Forest MAE: 18527.636972697186
  + Random Forest MSE: 24723.018802251023
  + Random Forest RMSE: 157.2355519666307
  + Random Forest R2 Score: 0.7657359852581718

A diagram of a graph

Description automatically generated with medium confidence

***Fig: Random Forrest Predictions***

* **Algorithm:** [Decision Tree Regressor]
* **Performance Metrics:**
  + Decision Tree MAE: 25513.96396396396
  + Decision Tree MSE: 35046.470180269585
  + Decision Tree RMSE: 187.2070249223292
  + Decision Tree R2 Score: 0.5292491100124799
* **Algorithm:** [Linear Regressor]
* **Performance Metrics:**
  + Linear Regression MAE: 20264.679126069095
  + Linear Regression MSE: 26230.836138961095
  + Linear Regression RMSE: 161.95936570313276
  + Linear Regression R2 Score: 0.7362898459515308

A blue and red dotted graph

Description automatically generated

***Fig: Linear Regression Predictions***

**Conclusion**

The best model selected for this project is Random Forest Regressor. Based on the Rsq score, this is better performing than linear regression model. However, future work on hyper tuing the model can achieve more accuracy and help better prediction.

The housing price prediction model provides a reliable tool for stakeholders to estimate housing prices based on available features. Future work will include incorporating additional data sources, improving feature engineering techniques, and fine-tuning model parameters further to enhance performance.

**Recommendations**

1. Continuously update the model with new data to ensure relevance.
2. Integrate external factors such as economic indicators for better predictions.
3. Regularly monitor model performance metrics to address potential data drift.

References:

Dataset: <https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset>