

Submitted to: Sir Rasikh Ali

Submitted by: Dua Saeed

Subject: PAI (Lab)

Section: BSAI-4A

Roll number: 042

Task: Kaggle Competition:

House Price Prediction

1. Introduction

House price prediction is a crucial aspect of the real estate market, enabling buyers, sellers, and investors to make informed decisions. The objective of this project is to build a machine learning model that predicts house prices based on various features such as lot size, location, number of rooms, and other property-related characteristics.

2. Dataset Overview

The dataset consists of 1460 rows and 81 columns, including numerical and categorical variables. Some of the key attributes in the dataset are:

* MSZoning: General zoning classification.
* LotFrontage: Linear feet of street connected to the property.
* Neighborhood: Physical location.
* OverallQual: Overall material and finish quality.
* YearBuilt: Original construction year.
* SalePrice: The target variable (house price).

3. Data Preprocessing

Handling Missing Values

Several columns contain missing values, including:

* PoolQC (1453 missing values)
* MiscFeature (1406 missing values)
* Alley (1369 missing values)
* Fence (1179 missing values)
* LotFrontage (259 missing values)
* Garage-related columns (81 missing values)

To handle missing data, strategies such as imputation with median values for numerical columns and mode imputation for categorical variables were used.

**Encoding Categorical Variables:**

Categorical features were encoded using Label Encoding for variables such as MSZoning, Neighborhood, and HouseStyle to convert them into numerical format.

**Feature Scaling:**

Numerical features were scaled using StandardScaler to normalize values, ensuring that all features contribute equally to the model.

4. Model Selection and Training

The model used for prediction is XGBoost Regressor, a powerful gradient-boosting algorithm that provides accurate and efficient predictions.

**Splitting the Data:**

The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.

**Model Training:**

The XGBRegressor model was trained with the following hyperparameters:

* n\_estimators=500: Number of boosting rounds.
* learning\_rate=0.05: Step size at each boosting iteration.
* max\_depth=6: Maximum depth of a tree.
* random\_state=42: Ensures reproducibility.

5. Model Evaluation

The model performance was assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

* MAE: Measures the average absolute error between predicted and actual values.
* RMSE: Gives more weight to larger errors, providing a better measure of overall performance.

6. Prediction and Submission

Predictions were generated using the trained model, and a CSV file was created for submission containing the predicted house prices.



