

The Superior University

|  |  |  |
| --- | --- | --- |
| Name: Abdul Moeez | Roll No: 017 | Course: PAI Lab |
| Semester: 4th | Section: BSAI 4A | Department: Artificial Intelligence |
| Submitted To: Sir Rasikh | Total Marks: | Date: |

**Lab 1**

**Task: Kaggle Competition: House Price Prediction**

**House Price Prediction Using Machine Learning**

**Introduction**

This project focuses on forecasting house prices by leveraging machine learning techniques, particularly Random Forest and XGBoost regressors. The dataset encompasses a range of features related to property attributes, with the objective of predicting the SalePrice of homes based on historical data.

**Code Explanation**

**1. Data Loading and Preprocessing**

* The dataset is loaded using pandas.read\_csv() from a specified file path.
* The Id column is explicitly converted to integers to maintain consistency.
* Basic exploration is performed using .shape, .columns, .dtypes, and .describe().
* Missing values are identified and handled using SimpleImputer (median for numeric and most frequent for categorical features).

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

* A histogram is plotted using seaborn to visualize the distribution of SalePrice.
* A correlation heatmap is generated to examine relationships between numerical features.

**3. Feature Engineering and Scaling**

* The dataset is split into numeric and categorical columns.
* Missing values in numeric columns are filled with the median.
* Missing values in categorical columns are filled with the most frequent values.
* One-hot encoding is applied to categorical features using pd.get\_dummies().
* The features are standardized using StandardScaler to normalize the dataset.

**4. Model Training and Evaluation**

* The dataset is split into training and validation sets (train\_test\_split() with an 80-20 split).
* Two models are trained:
  + **Random Forest Regressor**
  + **XGBoost Regressor**
* Models are evaluated using Root Mean Squared Error (RMSE) to measure prediction accuracy.
* The results of both models are displayed.

**5. Final Predictions and Submission**

* The test dataset is preprocessed in the same way as the training dataset.
* Missing columns in the test dataset are aligned to match those in the training dataset.
* The trained XGBoost model is used to predict house prices.
* Predictions are exponentiated to reverse the log transformation applied earlier.
* The final predictions are saved to a CSV file for submission.

**Model Performance**

* **Random Forest RMSE:** Displayed in the console.
* **XGBoost RMSE:** Displayed in the console (typically better performance due to boosting).

**Output Screenshot**

