

The Superior University

|  |  |  |
| --- | --- | --- |
| Name: Muhammad Bilal Sajid | Roll No: 018 | Course: PAI Lab |
| Semester: 4th | Section: BSAI 4A | Department: |
| Submitted To: | Total Marks: | Date: |

**Lab 1**

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

**House Price Prediction Using Machine Learning**

**Introduction**

This project involves predicting house prices using machine learning models, specifically Random Forest and XGBoost regressors. The dataset includes various features related to house characteristics, and the goal is to predict the SalePrice of homes using 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.

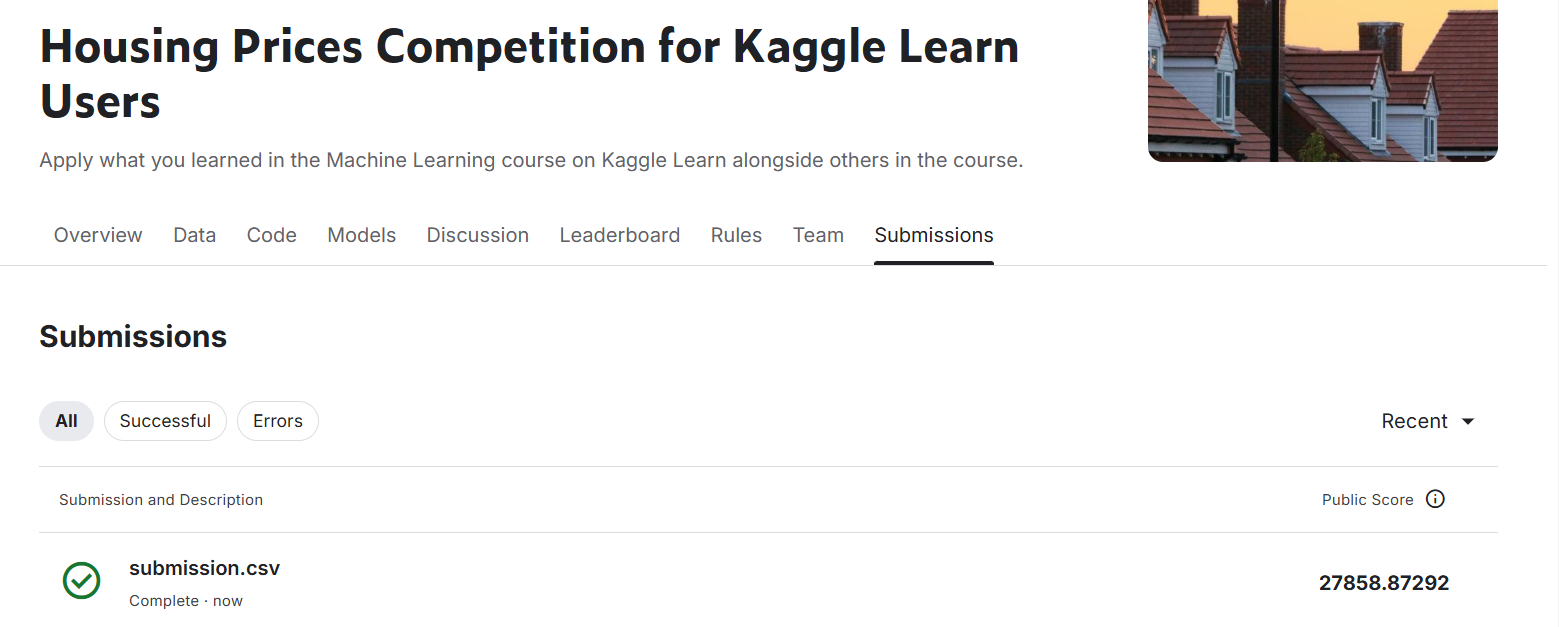
**5. Final Predictions and Submission**

* The test dataset is preprocessed similarly to the training dataset.
* Missing columns in the test dataset are aligned with the training dataset.
* The trained XGBoost model is used to predict house prices.
* The predictions are exponentiated (since the log transformation was applied earlier) and 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**

****