Celebal Assignment Week-5

Objective:

The objective of this project is to develop a **machine learning regression model** that predicts house sale prices using the **Random Forest Regressor**. The solution uses structured data and applies preprocessing techniques to ensure data quality and model accuracy.

Dataset Overview:

- **train.csv**: Contains 80+ features related to houses (e.g., number of rooms, area, location) and the target column SalePrice.
- **test.csv**: Contains the same features (excluding SalePrice) for which predictions are to be made.
- **house_price_predictions.csv**: Final output file containing predictions (Id, SalePrice).

Data Preprocessing Steps:

Remove Identifiers:

- Id column is dropped as it's not relevant to model learning.

Split Target and Features:

- Training data is split into X_train (features) and y_train (target: SalePrice).

Combine Train and Test Data:

- Feature sets from train and test are concatenated to ensure uniform preprocessing.

Missing Value Handling:

- **Numerical Columns**: Missing values are filled using the **median**.
- **Categorical Columns**: Missing values are filled with a constant **'Missing'**.

Feature Encoding:

- Categorical columns are **One-Hot Encoded** using ColumnTransformer.
- Numerical columns are **passed through** unchanged.

Model Building:

Model Used: RandomForestRegressor from sklearn.ensemble

Parameters:

- n_estimators=10 (number of decision trees)
- random_state=0 (for reproducibility)
- **Training**: Model is trained on the transformed training data.
- **Prediction**: Applied to test data to generate predicted sale prices.

Output Generation:

Predictions are paired with test Ids and saved as a CSV file named house_price_predictions.csv.

Visualization:

- A **line plot** is generated to visualize Id vs Predicted SalePrice.
- The plot uses:
  - color='yellow'
  - linewidth=2
  - Proper axis labels and a grid for clarity.

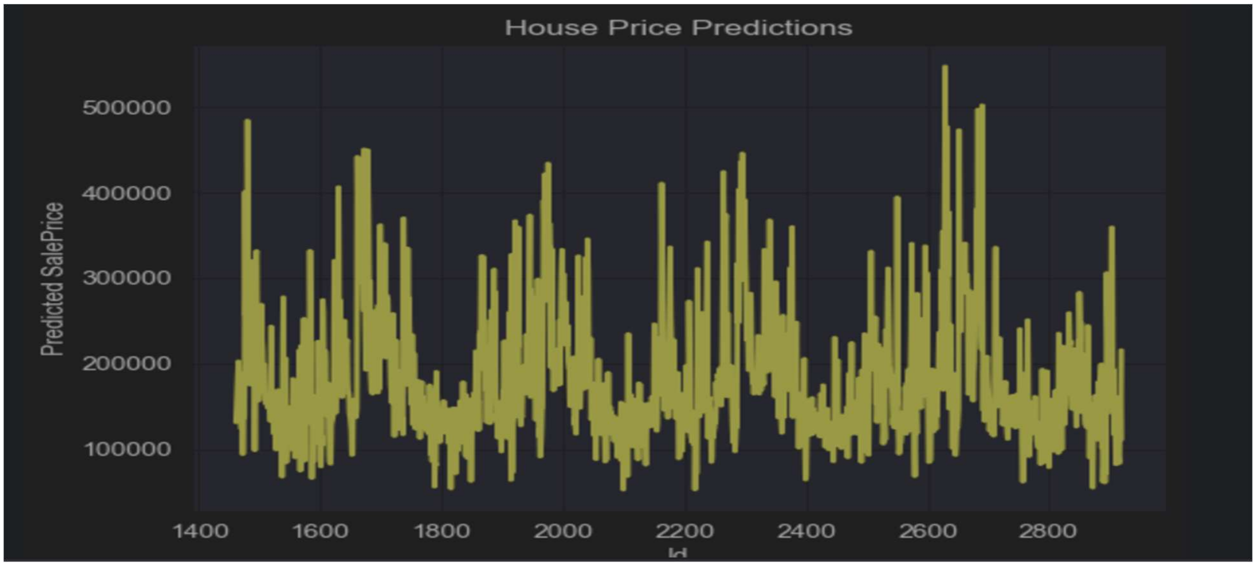
Conclusion:

This project successfully builds a pipeline for predicting house prices using:

- **Data cleaning**
- **Feature encoding**
- **Random Forest regression modeling**

The output is a clean CSV ready for submission or further evaluation.

Prediction values Graph


House Price Predictions
Predicted SalePrice
Id