🏠 House Price Prediction Project

In this project, I implemented a house price prediction model using advanced regression techniques. The aim was to accurately predict the sale prices of houses based on various features. Here's a detailed overview of the methodology I followed:

1 Importing Libraries

I began by importing essential libraries such as pandas, numpy, matplotlib, seaborn, and xgboost. These libraries are fundamental for data manipulation, visualization, and building machine learning models. Additionally, I used sklearn for model evaluation and tuning.

2 Data Loading and Inspection

I loaded the training and test datasets from CSV files. Initial data inspection allowed me to understand the dataset's structure and features. I utilised pandas to view the first few records and the data types of each column.

3 Data Exploration, Cleaning, and Engineering

3.1 Exploratory Data Analysis (EDA) 🔍

I performed EDA to identify missing values and understand the distribution of the target variable, **SalePrice**. This included visualizing missing values with a bar plot and understanding the target distribution with a histogram. Additionally, I explored the correlation between numerical features using a heatmap, which helped identify significant predictors for the model.

3.2 Handling Missing Values ⚠️

To address missing values, I used the SimpleImputer from sklearn. The median value was employed for imputation of numerical features common to both training and test datasets, ensuring that the integrity of the data was maintained.

3.3 Feature Engineering 🔧

I applied one-hot encoding to categorical variables to convert them into a numerical format suitable for modeling. The training and test datasets were aligned to ensure they had the same feature set, filling in any missing columns with zeros.

4 Train-Test Split 📊

I separated the features from the target variable and performed a train-test split. 80% of the data was allocated for training, while 20% was reserved for validation, ensuring robust model evaluation.

5 Feature Scaling ⚖️

To standardize the features and improve model performance, I applied StandardScaler. Scaling ensures that all features contribute equally to the model, avoiding dominance by variables with larger magnitudes.

6 Hyperparameter Tuning using Grid Search 🔍

I implemented hyperparameter tuning through Grid Search to optimize the XGBoost model. The parameter grid included various options for n_estimators, learning_rate, max_depth, and more. This systematic approach helped identify the best-performing model configuration.

7 Model Evaluation 📈

Using the optimal model, I made predictions on the validation dataset. The model's performance was quantified using the R^2 score, which achieved a remarkable value of **0.912**, indicating a strong correlation between predicted and actual sale prices.

8 Visualization of Results 📊

To visually assess the model's performance, I plotted predicted values against actual sale prices. This scatter plot provided insight into prediction accuracy. Additionally, I created a pie chart to illustrate the proportion of errors in predictions, further aiding in understanding the model's effectiveness.

9 Final Predictions and Submission 📤

For the final step, I prepared the test dataset for predictions using the best model. The predicted sale prices were saved to a CSV file for submission, ensuring that I could present the results effectively.

🏆 Results and Interpretation

The project yielded significant insights, including:

- A high R^2 score of **0.912**, demonstrating the model's predictive accuracy.
- Visualizations helped in identifying areas of improvement in prediction and understanding model behaviour better.

💡 Skills Developed

Through this project, I honed various skills, including: Data Cleaning, Exploratory Data Analysis (EDA), Feature Engineering, Model Evaluation, Hyperparameter Tuning, Visualisation, XGBoost Implementation

Hashtags

#DataScience #MachineLearning #HousePricePrediction #XGBoost #FeatureEngineering #DataAnalysis #ExploratoryDataAnalysis #DataVisualization #ModelTuning #HyperparameterTuning #PredictiveAnalytics #BigData #Python #Pandas #Numpy #Seaborn #Matplotlib #DataCleaning #ModelEvaluation #AI