Step by Step House Price Prediction ML Regression Project

# Project Overview:

This project aims to predict house prices using machine learning regression techniques. It involves data preprocessing, exploratory data analysis, feature selection, model training, and evaluation. This repository contains code and resources necessary to replicate the project and understand the methodologies used.

This repository contains a comprehensive project demonstrating the process of building a machine learning model to predict house prices. The project covers various stages, from data loading and preprocessing to model training and evaluation.

The goal is to create a model capable of accurately predicting house prices based on various features, such as:

* Lot size
* Dwelling type
* Overall condition rating
* Year of construction
* Basement area

The project implements and compares different regression models, including:

* Support Vector Regression (SVR)
* Random Forest Regression (RFR)
* Linear Regression (LR)

The models are evaluated using the Mean Absolute Percentage Error (MAPE) metric.

Project Structure:

├── HousePricePrediction.xlsx

└── HousePricePrediction.ipynb

* HousePricePrediction.xlsx: The dataset containing house price information.
* HousePricePrediction.ipynb: The Jupyter Notebook containing the code for this project, including data exploration, preprocessing, model training, and evaluation.

Project Steps:

1. Dataset Loading and Exploration:
   * Load the dataset (HousePricePrediction.xlsx).
   * Explore the data dimensions, column names, and data types.
2. Data Preprocessing:
   * Identify categorical, integer, and float features.
   * Drop irrelevant columns (e.g., 'Id').
   * Handle missing values (using mean for 'SalePrice').
   * One-hot encode categorical features.
3. Exploratory Data Analysis (EDA):
   * Create a correlation heatmap to analyze relationships between numerical features.
   * Visualize the distribution of unique values in categorical features.
4. Model Training:
   * Split the dataset into training and validation sets.
   * Train three regression models:
     + Support Vector Regression (SVR)
     + Random Forest Regression (RFR)
     + Linear Regression (LR)
   * Evaluate the performance of each model using MAPE.

### Dataset:

The dataset used for this project is sourced from Kaggle, containing various features related to houses and their sale prices. Key features include:

* MSSubClass
* YearBuilt
* BldgType
* Exterior1st
* SalePrice (target variable)

#### Technologies Used:

* Python
* Pandas
* NumPy
* Scikit-Learn
* CatBoost
* Matplotlib / Seaborn for visualization

##### Model Evaluation:

* SVM Model: Achieved a Mean Absolute Error of approximately 0.18.
* CatBoost Regressor: Achieved an R² Score of approximately 0.8936.

###### Results:

The evaluation results indicate that the SVM model provided the best accuracy among the models tested. CatBoost also showed strong performance, suggesting that both models can be utilized for predicting house prices effectively.

###### Conclusion:

The project demonstrates the application of various regression techniques for house price prediction. While the SVM model yielded the best results in this implementation, exploring ensemble learning methods like Bagging and Boosting could further improve performance.

###### 

###### Questions:

* Feature Engineering: How could you further improve the model by exploring feature engineering techniques, such as creating new features based on existing ones?
* Hyperparameter Tuning: How can you optimize the hyperparameters of each model to improve performance?
* Model Evaluation: How can you use cross-validation for a more robust evaluation of model performance?
* Model Visualization: How can you visualize the predictions of the model against the actual values to gain insights into the model's performance?

This project is licensed under the MIT License. See the LICENSE file for more details.