AIDI 1003-01 - Machine Learning Frameworks

Project Proposal: Real estate price prediction

Project Team:

Name	Student ID
Satish Kumar	200574904
Dhyan Trivedi	200565531
Anmol Toor	200578497
Harsh Patel	200575814

Dataset Link: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Project Scope

This project aims to predict the sale prices of homes based on various property features using the House Prices dataset from Kaggle. The dataset includes 81 attributes related to property characteristics, such as building type, neighborhood, lot size, and overall quality, providing a comprehensive set of factors that influence housing prices. The project will encompass data preparation, feature selection, model development, and evaluation on test dataset to achieve the most accurate predictions.

Objective

The primary objective is to develop a robust machine learning model capable of predicting housing prices with high accuracy. This will involve exploring advanced regression techniques and feature engineering methods to identify the most influential features and construct an optimal model. The results will benefit stakeholders such as real estate agencies and prospective home buyers by providing insights into the property value determinants.

Intended Approach

1. **Data collection & Preparation**: The dataset will undergo preprocessing steps to handle missing values, remove duplicates, and standardize/normalize numerical features. Categorical features will be encoded using appropriate methods (Label Encoding for binary and One-Hot Encoding for multi-class categories).

2. **Exploratory Data Analysis (EDA):** An EDA phase will involve visualizing the relationships between key features and the target variable, identifying outliers, and evaluating feature correlations. Box plots, histograms, scatter plots, and correlation heatmaps will help uncover patterns and refine feature selection.

3. **Feature Engineering:** Feature selection techniques, such as SelectKBest, will identify the top features contributing to the price prediction. This step ensures that the model focuses on the most relevant attributes, potentially improving both performance and interpretability.

4. **Model Training and Tuning**: Various regression models will be trained, including Linear Regression, Random Forest, Gradient Boosting, and XGBoost. GridSearchCV will be employed for hyperparameter tuning to optimize model performance

5. **Model Evaluation and Validation**: The models will be evaluated based on metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. Cross-validation will be used to assess model robustness, and the best-performing model will be selected based on these metrics.

6. **Model Deployment:** The most efficient model will be deployed using Flask python library and an API will be created in which, user can provide values of selected features. The API will return predicted sales price based on the values provided.

Expected Outcome

By applying feature selection and advanced regression techniques, this project aims to produce a model with high predictive accuracy for house prices. The results will demonstrate how certain property attributes drive price variations and will serve as a valuable reference for industry applications.