Predicting Home Selling Price Using Machine Learning

Introduction:

In this document, we explore the process of predicting home selling prices using machine learning techniques. We utilize a dataset containing various attributes of homes in California, such as median income, housing median age, and geographical location.

Data Preprocessing:

Initially, we loaded the dataset and identified missing values in the 'total_bedrooms' column.
We handled missing data using the KNNImputer, which replaced missing values with estimates based on the values of neighboring instances.
Additional features were engineered, such as 'rooms_per_household' and 'bedrooms_per_room', to provide more insights into the dataset.
Categorical variables were encoded using one-hot encoding to convert them into numerical form.

Exploratory Data Analysis (EDA):

We visualized the relationships between various features and the target variable 'median_house_value' using scatter plots, histograms, box plots, and correlation heatmaps.
EDA helped us understand the distribution of data, identify outliers, and explore correlations between different features.
Model Building:

We split the data into training and testing sets with a 70:30 ratio.
Three different machine learning models were employed: Linear Regression, Random Forest Regression, and XGBoost's XGBRegressor.
Each model was trained on the training data and evaluated on the testing data using metrics such as R-squared and Root Mean Squared Error (RMSE).

Results:

Linear Regression: Test Accuracy - 55.39%, RMSE - 76518.24
Random Forest Regression: Test Accuracy - 78.12%, RMSE - 53592.39
XGBRegressor: Test Accuracy - 77.43%, RMSE - 54429.21

Conclusion:

The Random Forest Regression model performed the best among the three models, with the highest test accuracy and the lowest RMSE.
Our analysis demonstrates the effectiveness of machine learning techniques in predicting home selling prices based on various attributes.

Future Work:

Further optimization of model hyperparameters could potentially improve predictive performance.
Exploring additional features or different algorithms may yield better results.