California Housing Price Prediction

Introduction/Problem statement:

Investment is a common business pursuit for many individuals, and there are various assets that can be utilized for this purpose, such as stocks, ETFs, cryptocurrencies, and real estate. Real estate investment, in particular, has seen a significant increase in both demand and sales. The value of a house is influenced by various factors, including its size, location, number of bedrooms, and the prices of comparable properties. Real estate investors aim to accurately assess the cost of a house before buying or selling, as purchasing at a higher price and selling at a lower price can result in financial loss. Banks may also request the current market value of a property when using it as collateral for a loan. Additionally, home buyers may seek mortgage loans from financial institutions and may consider the asking price of a house to determine if it is overpriced. Sellers, on the other hand, may consider the predicted value of their property to determine a fair market price. This project aims to analyze the factors that impact the housing prices in California through the use of machine learning models.

Dataset source and information:

The dataset is a csv file from Kaggle.
The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. The columns are as follows; their names are described below:

1) Median House Value: Median house value for households within a block (measured in US Dollars) [$]

2) Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]

3) Median Age: Median age of a house within a block; a lower number is a newer building [years]

4) Total Rooms: Total number of rooms within a block

5) Total Bedrooms: Total number of bedrooms within a block

6) Population: Total number of people residing within a block

7) Households: Total number of households, a group of people residing within a home unit, for a block

8) Latitude: A measure of how far north a house is; a higher value is farther north [°]

9) Longitude: A measure of how far west a house is; a higher value is farther west [°]

10) Distance to coast: Distance to the nearest coast point [m]

11) Distance to Los Angeles: Distance to the center of Los Angeles [m]

12) Distance to San Diego: Distance to the center of San Diego [m]

13) Distance to San Jose: Distance to the center of San Jose [m]

14) Distance to San Francisco: Distance to the center of San Francisco [m]

Data Wrangling and EDA

During the data wrangling process, I assessed the dataset for any missing values and performed necessary cleaning steps. After this initial inspection, I conducted exploratory data analysis to uncover trends and insights within the data. Through this analysis, I discovered a positive relationship between Median Income and Median House Value. I also utilized histograms to examine the distribution of several variables, including "Tot_Room," "Tot_Bedrooms," "Population," and "Households," and found that they were skewed. Additionally, I plotted the data using ggplot and observed that house prices tend to increase in areas closer to the coast, particularly in Southern California and the Bay Area. The analysis also revealed a positive relationship between population and total rooms, and I used a heat map to confirm that "Tot_Room," "Tot_Bedrooms," "Population," and "Household" were highly correlated, while median income and median house value had a medium level of correlation.

Preprocessing/Modeling

Next, the code separates the data into input features (X) and the target variable (y), with X representing all of the columns except for the first one and y representing the first column. The data is then split into training and testing sets, with 80% of the data being used for training and the remaining 20% being used for testing. The input data is also standardized using scikit-learn's StandardScaler.

The code then fits and trains four different machine learning models on the training data: a random forest regressor, a linear regressor, a Lasso regressor, and an XGBoost regressor. The models are then evaluated on the testing data and the R-squared value is calculated to assess their performance. The code also prints the root mean squared error (RMSE) for each model.

Finally, the code visualizes the feature importance for the random forest model using a bar chart and plots the predictions made by the random forest model against the observed values for the testing data.

Model	R^2
Random Forest	0.97
Linear Regression	0.66
Lasso	0.66
XGBRegressor	0.98

Limitation/Recommendations:

- One limitation of this analysis is that the dataset used is from 1990 census data, meaning that the resulting model may not accurately reflect current trends in housing prices. However, the concepts used in this analysis could potentially be applied to predict future housing prices if more up-to-date data becomes available.
- In addition, there are other factors that can impact the value of a home, such as the condition, age, and size of the property. Including these variables in the dataset could provide a more comprehensive understanding of housing prices in real life. It is recommended to consider incorporating these variables in future analyses.