California Housing Prices Dataset

Creation of a machine learning regression model, Alessandro Sciorilli

The following report outlines the procedures and findings of a machine learning regression model constructed using the California Housing Prices Dataset from Kaggle. The dataset provides details on various factors, including the median house value, the house's geographical location (longitude, latitude, and ocean proximity), median housing age, total rooms, total bedrooms, as well as demographic information about California (population, households, and median income). The objective of this analysis is to develop a machine learning regression model that can accurately predict the price of a house in California.

After importing the dataset into a Pandas dataframe, I explored the data and checked for missing values. I found that 207 values were missing for the "total bedrooms" variable, so I proceeded to remove the rows with missing values from the dataset. After cleaning the dataset, I performed one-hot encoding on the "ocean proximity" variable. One-hot encoding is necessary because ocean proximity is a qualitative (categorical) variable that cannot be used in a linear regression model. I created a dummy variable (0-1) for each sub-category, namely "<1h ocean," "inland," "island," "near bay," and "near ocean." To visualize this information, I plotted the data in a bar chart. The chart revealed that the majority of the houses were located within a distance of less than 1 hour from the ocean or inland. Only a small percentage of houses were located near the bay or near the ocean, and there was a very small number of houses located on the island.

Observing the scatter plot (Plot 5), two main patterns become evident. Firstly, there is a direct relationship between the proximity of a house to the two main cities in California (LA and San Francisco) and its median house value. The closer a house is to these cities, the higher its median house value. Additionally, there is a positive correlation between a house's proximity to the ocean and its median house value. In fact, the correlation coefficient (-0.48) indicates a negative correlation between the "inland" position and the median house value. Furthermore, the correlation matrix reveals a positive correlation (+0.68) between the median income and the median house price. This implies that areas with higher median incomes tend to have higher house prices.

To predict the median house value as the target variable, I constructed a machine learning regression model by splitting the dataset into training and test sets with a 70/30 ratio. Following this approach, I utilized 70% of the dataset to estimate the model's parameters and determine the best fitting line that minimized the sum of squared errors. Subsequently, I evaluated the performance of the trained model using the remaining 30% of the "unseen" data from the dataset, calculating metrics such as R2 (coefficient of determination), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error).

The calculated coefficients (Betas) and intercept provide me with the following multiple linear regression equation:

House price (y) = -**2212845** **-2.64**(*longitude)* **-2.51***(latitude)* **+1.11***(housing_median_age)* **-5.01***(total*_rooms) **+9.35***(total_bedrooms)* **-3.60***(population)* **+4.45***(households)* **+3.92***(median_income)* **-2.31***(<1H OCEAN)* -**6.33***(INLAND)* **+1.30***(ISLAND)* **-2.6***(NEAR BAY)* **-1.77***(NEAR OCEAN)*

Analyzing the regression model, it seems to exhibit some level of unreliability. The estimated betas deviate from the observed correlation between variables in the initial phase of the analysis. Specifically, it is difficult for me to justify a beta of -5.01 for the variable "total_rooms" and a positive beta of +9.35 for the variable "total_bedrooms." Both variables should be positively correlated with the house price, as a larger number of rooms usually indicates a larger square footage and, therefore, a higher value. It is also true that larger houses are typically located in rural areas or away from major city centers, which could potentially decrease their value. The presence of a -3.60 beta, indicating a negative correlation between population and house value, is also suspicious because house prices should, on average, be higher in areas with a larger population, such as big cities. Additionally, the negative correlation between being "NEAR BAY" or "NEAR OCEAN" and house price is also questionable, as houses located close to the bay or seaside usually have higher values. The weakness of this machine learning model is further emphasized by a relatively low R2 value (0.64), suggesting that only approximately 64% of the variance in the dependent variable (house price) can be explained by the independent variables in the model. Therefore, the model does not ideally capture the relationship between Y and the X variables. Moreover, there is a considerable average absolute difference (MAE) between the predicted and actual values of the target variable. The disparity between the two distributions is illustrated in a line chart, highlighting the difference in median house values between the actual and predicted distributions. Suggestions for further improvement include:

- Since the model is a multiple regression model, adding more dependent variables would increase the value of R2. It would be better to consider the calculation of Adjusted R2 to address this issue.

- Checking for multicollinearity is crucial. The presence of multicollinearity among independent variables in a multiple regression model poses a problem and reduces the reliability of the estimation. In this dataset, I observed correlations between the variables "total_rooms" and "total_bedrooms," as well as between "population" and "households." Therefore, removing the variables "total_bedrooms" and "households" could potentially enhance the reliability of the model.

Dataset link : https://www.kaggle.com/datasets/camnugent/california-housing-prices