Python: California Housing Price Analysis Report

Goals

This analysis uses a multiple linear regression using California Housing dataset to acquire insights into which components have meaningful impacts on decision of California housing price. I am going to use the dataset included in scikit-learn library built in Python. This report summarizes findings from EDA, data preprocessing, multiple linear regression to understand crucial factors affecting housing prices in California. I compared the original multiple linear regression using OLS with stepwise linear regression after the log transformation to several independent and dependent variables.

Dataset Review

The California Housing dataset contains information derived from the 1990 U.S. Census, with data collected at the block group level. A block group typically includes a population of 600 to 3,000 people and represents the smallest geographical unit for which the U.S Census Bureau publishes sample data. The dataset contains a total of 20,640 instances. There are 8 numeric features (independent variables) and 1 target variable which is a dependent variable used in a multiple linear regression. The target variable is the median value of the California Housing, expressed in hundreds of thousands of dollars ($100,000). The 8 independent variables consist of 'MedInc' (Median income of households in the block group), 'HouseAge' (Median house age in the block group), 'AveRooms' (Average number of rooms per household), 'AveBedrms' (Average number of bedrooms per household), 'Population' (Total population in the block group), 'AveOccup' (Average number of household members), 'Latitude' (Geographical latitude of the block group) and 'Longitude' (Geographical longitude of the block group). There are non-missing values in the dataset. This dataset represents housing conditions across California districts and provides predictive independent variables used for inputs to determine the median house value. Furthermore, for both average room and bedroom variables, the values can be large due to a high number of vacant houses, such as in vacation resort areas, which should be considered when managing outliers.

Explanation of EDA and Analyses

(1) Distribution of Variables

Several variables are right-skewed including MedInc (median income), average number of rooms, bedroom, occupation, and population. This skewness can affect the accuracy of machine learning algorithms. There are also some variables including Median Income, Median Housing Age and Median Housing Value are capped. For example, the median housing age is capped at 52 years, making it hard for a machine learning algorithm to predict values accurately beyond the capped range. This issue is also shown on the scatter plot between the median income values and the median housing values. The correlation matrix between numerical variables, including the dependent variable shows a moderately strong correlation between the median income variable and the median housing value. According to the scatter plot between them, there are some horizontal lines shown on the plot, indicating that there are data-capped limits.

Regarding outliers in individual independent variables, skewed variables appear to have numerous outliers. However, I only removed outliers from a 'AveRooms' variable. First, a normal income distribution is typically skewed. I thought the distribution of the median income level follows the typical income distribution. Second, the population of the block groups within California ranges from 600 to 3,000 people with a wider range that could affect the right-skewed distribution of the population. Third, due to vacation resorts within California districts, the average of occupation can vary with some larger values seemed to be outliers which can be also applied to the average number of rooms and bedrooms. On the other hand, there is a strong correlation between average number of rooms and bedrooms in contrast of the average number of occupation. Therefore, I removed outliers from the average number of rooms not from both of them to avoid the data losses. In addition, after removing outliers from the average

number of rooms, there are still 20,174 instances, meaning that the proportion of outliers is only 2.26% so I removed outliers without huge data losses while reducing the correlation between the two variables.

A visual analysis of latitude and longitude variables reveals that housing prices along the southern coastlines are significantly higher than those along the northern coastlines, indicating different pricing trends. Populations tend to be concentrated in central locations but housing prices of these areas are lower compared to the southern coastlines with higher housing prices and condensed population. The disparity in housing prices may indicate the influence of other factors such as proximity to economic hubs or desirability of living in certain areas but these factors are now shown as variables in this dataset.

Develop a Multiple Linear Regression for Predicting California Housing Price

I divided dataset into 70% training data and 30% test dataset, of which the proportions between the two is widely accepted standard. I allocated 70% of the total dataset to sufficiently train the model and then secure 30% of the total dataset to test the model using some metrics for evaluation. After removing outliers from the average number of rooms, 20,174 instances remain which seems to be sufficient for allocating data for both the training and testing dataset in the chose proportion.

To remedy the skewed distribution of several independent variables and increase the accuracy of the multiple linear regression model, I used a log-transformation. The R-squared of this model is 61%, indicating the model 61% of the total variation of the dependent variable. The average squared difference between the actual values and the predicted values is very low, indicating the model performs well in predicting the target variable. The F-statistic and p-value indicate that the model is statistically significant. Based on p-values of beta coefficients for individual features, all coefficients are statistically significant. The magnitude of each coefficient indicates the most importance feature for predicting the target variable. The median value of income and the number of bedrooms is the most importance features associated with higher housing prices because they have all positive relationships with the target variable.

Multiple Linear Regression Assumption

I identified whether the original linear regression based on assumptions of linear regression such as linearity, independence, homoscedasticity and normality of residuals.

(1) Linearity

I identified the general trend of the linearity between the predicted values and the observed values of the dependent variable is visible but there are some deviations such as scattered points and a funnel-like pattern suggesting potential issues like heteroscedasticity. This indicates the model might not fully capture the relationship and transformations are necessary to address this issue. To address this issue, I applied the log transformation to the target variable and then the linearity between the predicted values and the observed values are improved.

(2) Normality of Residuals

The Q-Q plot indicates that there are deviations on most points of the residuals with the red diagonal line. There are deviations at both ends of tails due to the fact that there are several variables with outliers and also still deviations from normality in the middle parts due to non-linear effects. Overall, the normality assumption appears to be unsatisfied and further adjustments may be necessary to improve the model. After finalizing log transformation of y (dependent) variable, there are slight deviations at both ends of tails but most point follow the red diagonal line in the middle parts following normality.

(3) Independence of Residuals

The scatter plot of residuals against the order of data shows no specific patterns which indicates the autocorrelation between residuals violating the linear regression assumption. The residuals are randomly distributed, supporting the independence assumption between residuals as there doesn’t appear to be any systematic relationships or autocorrelations between the residuals.

(4) Homoscedasticity

In the first scatter plot of the homoscedasticity, the variance of residuals exhibits specific patterns such as funnel-shaped pattern observed. I applied a log transformation to the target variable to reduce residuals variance and improve its overall distribution. As shown in the second updated plot for homoscedasticity, the residuals exhibit a more uniform spread around zero, with a significant reduction in the funnel-shaped patterns shown in the first plot. This indicates that the log transformation effectively stabilized the variance of residuals and improved the linear regression model's assumptions, making it more reliable for prediction and ready to further improve the model.

In conclusion, the final original linear regression model includes the log transformation of the dependent variable and several independent variables with skewed distribution.

Model Improvement and Re-evaluation of the Updated Model

I included interaction terms while using stepwise regression for the final model selection to improve the overall model. I forcibly added individual variables if they are the components of the selected interaction terms. With updated multiple linear regression model with interaction terms, I re-evaluated the model for the comparison with the original model using metrics such as MSE, RMSE, R-squared and Adjusted R-squared. The adjusted R-squared is improved from 67.79% to 69.88% and RMSE reduces from 0.32 to 0.31.

Results and Practical Implications

(1) Median Income as a Key Determinant based on the beta coefficient

Median income is one of the most significant positive impact on housing prices, emphasizing the direct relationship between higher income levels and increased housing values. This finding aligns with the expectation that wealthier neighborhoods attract higher housing prices, making median income a crucial predictor for real estate valuation.

(2) Geographic and Demographic Influences

Housing prices are notably higher along the southern coastlines, reflecting the desirability of living in areas with favorable climates, proximity to economic hubs, and better amenities. The concentration of population in central regions with relatively lower housing prices suggests a trade-off between affordability and location desirability.

(3) Impact of Housing Characteristics

Among the housing features, the average number of rooms and bedrooms play significant roles in determining prices. The average number of rooms and bedrooms increase the median value of housing, aligned with typical real estate valuation. However, managing outliers in these variables is essential, as the presence of vacation resorts or vacant houses can distort their distributions. After addressing these outliers, the model improved in capturing the true impact of these features on housing prices.

(4) Improved Model with Log Transformations and Interaction Terms

Log transformations addressed skewness in the data, stabilized residual variance, and improved the model's assumptions of homoscedasticity as well as the linearity between the observed values and the predicted values particularity the log transformation with the target variable. Incorporating interaction terms further enhanced the model, resulting in better explanatory power (Adjusted R-squared increased from 67.79% to 69.88%) and reduced error metrics (RMSE

decreased from 0.32 to 0.31). These improvements highlight the importance of accounting for the log-transformation of the variables and variable interactions in complex datasets.