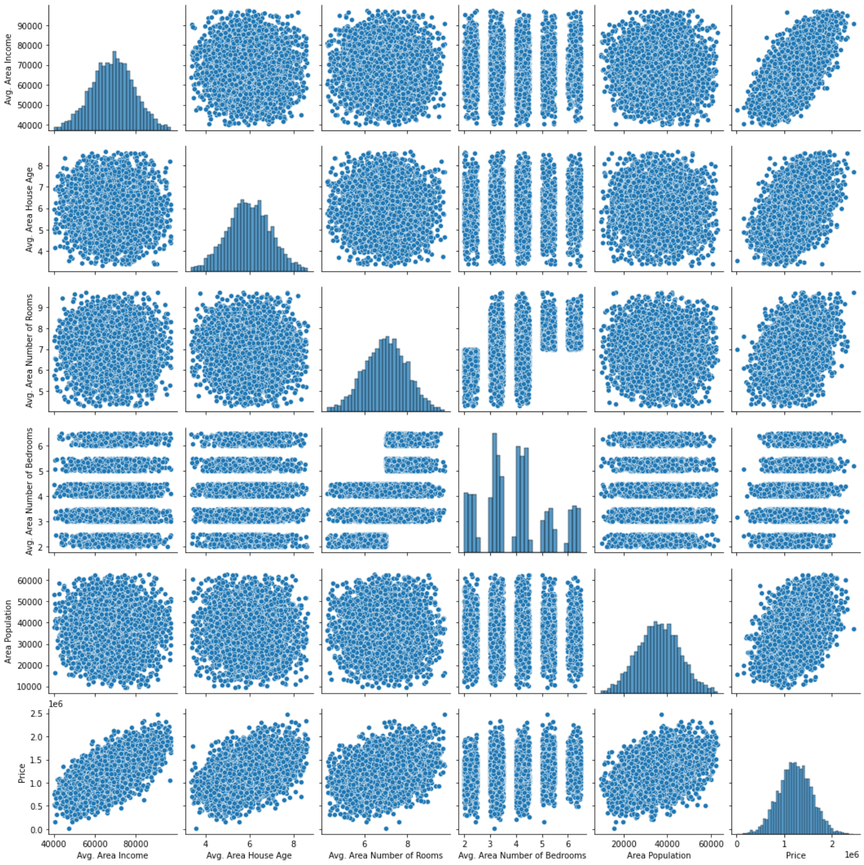
US HOUSING PRICE PREDICTION

Business Understanding

Within the housing market, a Comparative Market Analysis, or CMA, is utilized by the broker to present the seller with a proposed sale price and a comprehensive justification for this price. In this assignment, we are presented with the US housing market data and our analysis aims to develop a regression model for price prediction that could assist a broker/ institution in confidently evaluating the realistic selling price of properties in any of the targeted locations. The analysis aims at pricing homes closer to their "real value" (based on a concrete model) will result in lower resource use on the part of the broker/agency, and thus, a quicker (and more lucrative) sale.

Data Understanding

The housing dataset that we are given contains 7 variables recorded for 5000 properties scattered randomly in different parts of the country.



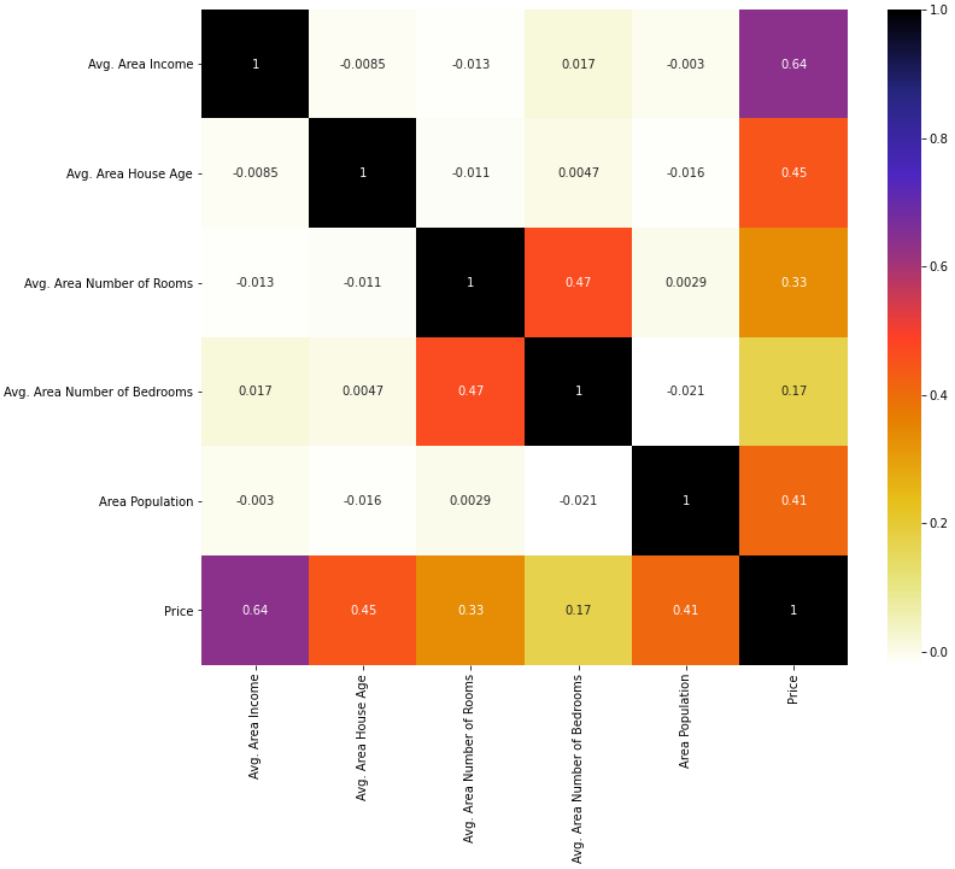
This data now serves as an input for our price prediction model that could help us identify the relationship between independent and dependent variables. Since, the model aims at predicting the price of the listed property, which is a continuous variable, we need to identify and eliminate all categorical variables from our consideration. Based on the observations, we identified that only one categorical variable i.e., Address is present in our data rest all are continuous variables. Moreover, before training the machine learning model, we need to ensure that the data doesn’t contain any missing / NA values, which in our case is not present.

Data Exploration and Transformation

In order to understand the impact of the given variable with respect to Price, we tried to implement the sklearn Feature selection model to identify the most impactful variables. Based on the univariate selection method, we observed that ‘Area Population’ is the most impactful variable while ‘Avg. Area Number of Rooms’ has the lowest impact in the determination of the Price of a house.

Results from Correlation Matrix were also in sync with the findings from the previous method which identified ‘Avg. Area Income’ as strongly correlated to the Pricing of a house while ‘Avg. Area Number of Rooms’ still has the least correlation among all.

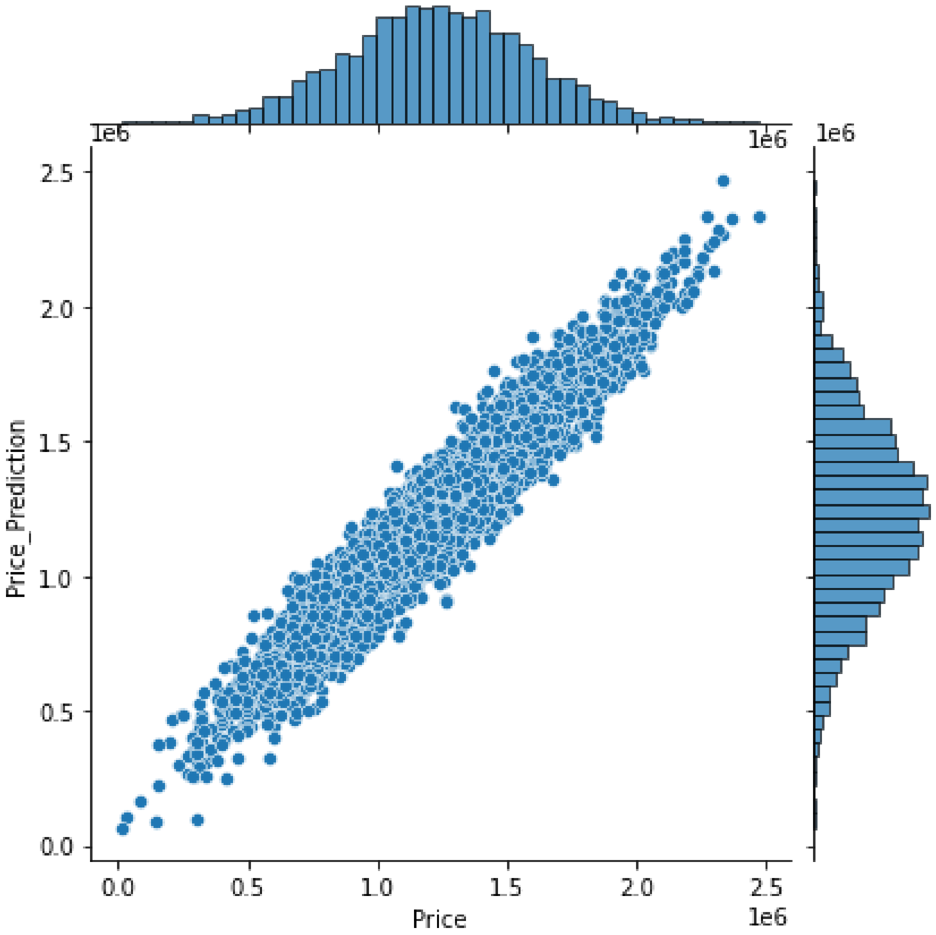
Moreover, in order to make sure that the data fed into the model is normally distributed and doesn’t contain any extreme input values, the data has been treated by identifying and removing outliers present in the given data. A total of 111 values were identified as outliers and were removed during data preprocessing.



	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1	-0.0085	-0.013	0.017	-0.003	0.64
Avg. Area House Age	-0.0085	1	-0.011	0.0047	-0.016	0.45
Avg. Area Number of Rooms	-0.013	-0.011	1	0.47	0.0029	0.33
Avg. Area Number of Bedrooms	0.017	0.0047	0.47	1	-0.021	0.17
Area Population	-0.003	-0.016	0.0029	-0.021	1	0.41
Price	0.64	0.45	0.33	0.17	0.41	1

Modeling

In order to build an optimal pricing model for the US housing market, we used a Multiple Linear Regression model. The model was trained on 80% of data and all of the explanatory variables including Avg. Area Income, Avg. Area House Age, Avg. Area Number of Rooms, Avg. Area Number of Bedrooms and Area Population were used to predict the response of the dependent variable – Price. The model consistently had an accuracy above 90.43% which was validated using a K-Fold cross-validation technique. Apart from the Linear Regression Model, the model's accuracy was also compared against other models such as Lasso Regression and Decision-Tree model. Linear Regression had a similar accuracy as that of the Lasso Regression model, which is 90.9%, however, Decision-Tree just turned out to be about 73.3% accurate in this given situation. Lastly, Hyperparameter Optimization was conducted using 'GridSearchCv' for the chosen model in order to fine-tune the performance of the model.



Evaluation

Three different supervised learning models namely the Linear Regression model, Lasso Regression model, and Decision Tree model have been used to make price predictions about the given housing data. A number of performance metrics including MAE, MSE, RMSE, R-Squared, Model Scoring, and AIC have been used to compare and evaluate the performance of each model. Based on results from various comparison markers both the Linear Regression Model and the Lasso Regression Model have similar results for the given dataset, with a mean model accuracy score of 90.89%. Furthermore, the R-squared values for both models are 0.9171, indicating that the model's predictions are extremely close to the best fit line. Since the evaluation factors in both of the models listed above are relatively close, the broker/real estate business may adopt any of the two models, Linear Regression or Lasso, to develop a Price prediction model.

Interpretation of Result

P-value: In hypothesis testing, a p-value is used to assist you to support or reject the null hypothesis. The method requires us to develop a Null Hypothesis and an Alternative Hypothesis and then test them against a given dataset to determine their significance. The null Hypothesis asserts that there is no association between the predictor and the response variable, and hence a change in one will have no influence on the other. While alternative hypotheses examine the data in relation to the idea that both variables are connected. The P-Value findings for all three models are all in sync, i.e., all models have a P-value of zero. We could now reject the null hypothesis with a 95 percent confidence level since the P-value of 0.00 is less than Alpha 0.05, indicating that there is an association between the predictor and the response variable for the provided dataset.

The coefficients for each explanatory variable can alternatively be interpreted as follows:

- 1 unit increase in Avg. Area Income is associated with an increase of $21.542
- 1 unit increase in Avg. Area House Age is associated with an increase of $164954.602
- 1 unit increase in Avg. Area Number of Rooms is associated with an increase of $121022.512
- 1 unit increase in Avg. Area Number of Bedrooms is associated with an increase of $1846.959
- 1 unit increase in Area Population is associated with an increase of $15.045

Conclusion

After analyzing the housing dataset, we were able to construct a supervised machine model that leverages a Linear Regression model to predict the price of a house. The model is also able to identify and verify the existing relationship that exists between the Price of a house and other variables such as Avg. Area House Age, Avg. Area Number of Rooms and more. Although the intercept is negative, this means that the model is overestimating the y values on average, necessitating a negative correction in the predicted values. However, once this issue is rectified, this model could assist the real estate broker/ institution in effectively predicting the housing prices in the US, allowing them to evaluate the realistic selling price of houses in these areas with confidence and lower their resource usage.