**Predicting students' dropout and academic success rates**

By: Chukwudi Nkamnebe

DS 535 Capstone Project

October 2023

**BACKGROUND OF THE STUDY**

The US Census Bureau published census data which has 7 variables and 20,000 observations. A simple random sampling was performed to take a sample of 300 from the data.

The data was collected by districts or block groups and in other to have a correct representation of the data, the median of each district was used to input each data.

The X variables are:

1. Housing Median Age
2. Total Rooms
3. Total Bedrooms
4. Population
5. Households
6. Median Income

The Y variable is the Median House Value

**Problem Objective**

The project aims at building a model of house price to predict the median house values in California using the provided data set. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

In this chapter, an exploratory data analysis was performed in other to understand the dataset and find relationships between the variables. We also checked the data for outliers. All rows with missing values were deleted in other to clean the data.

**Histogram of the median House Value**

Chart, histogram

Description automatically generated

The graphs show the distribution of the values of houses in California. We can see that majority of the price of the house are between 150,000 and 200,000.

Chart, histogram

Description automatically generated

The graphs show the distribution of the median income of households in California. We can see that majority of the price of the house are between 3 and 4.

Box plot showing Income

Diagram

Description automatically generated

We can see from the box plot that the average income of people in California is about 30k per year. At the 25% and 75 percentage we can see a value of 20k and 45k per year respectfully.

**Box plot showing House price**

**Chart, box and whisker chart

Description automatically generated**

We can see from the box plot that the average house price of people in California is about 200k per year. At the 25% and 75 percentage we can see a value of 150k and 250k respectfully.

**Correlation metrics**

Table

Description automatically generated with medium confidence

The diagram above shows the relationship between all variables. We can see that the variable with the highest correlation with the Y variable is Total rooms at 0.22. this means that there is a little relationship between the total rooms in the house and the price of the house. The more the rooms the higher the price of the house.

Chart, scatter chart

Description automatically generated

The graph above shows the relationship between income of California residence and the price of the house. The higher the income of households and higher the price of the house. A correlation of 0.66 was recorded between these 2 variables

Chart, scatter chart

Description automatically generated

The graph above shows the relationship between the 2 variables Total bedrooms and Population. The higher the population in a district the more the bedrooms. The correlation heat map shows that there is a 0.86 correlation metrics between both variables.

**Regression Analysis**

The dataset has 300 rows and 7 columns. For the regression analysis, the dataset was partitioned in a 60% by 40% ratio with 60% of the data assigned to training dataset and 40% assigned to validation dataset.

A regression model was built on the training dataset. The image below shows that all the regressors are significant at 5% except the ‘households’ variable which as a p-value of 0.7. we can see a residual standard error of 72360 which is high. The smaller the residual standard error, the better a regression model fits a dataset. Also, R-squared measures the goodness of fit of a regression model. Hence, a higher R-squared indicates the model is a good fit while a lower R-squared indicates the model is not a good fit. In this case, an R squared and Adjusted R squared values is 60% which shows the low performance of the model.

Table

Description automatically generated

Sum of Squared Errors (SSE) tells how much of Y is left unexplained. It tells how much cannot be attributed to a linear relationship. The Mean Square Error (MSE) on the other hand tells how close a regression line is to a set of points. It takes the distance from these points to the regression line and squares them. A high SSE/MSE suggests that the data have a reasonable degree of differences between them and may not usable. Root Mean Squared Error (RMSE) is simply the square root of the MSE. Lower values of RMSE indicate a better fit. In this case, a high **RMSE** of **70925.59** for the training and 8587.26 for the test data shows the model is not reliable. A high Mean Absolute Error (**MAE**) of **54796** for the training and 8587 for the test validation data also further confirms that the model does not fit well fit.

**Text

Description automatically generated**

**Variable selection**

3 variable selection methods were used to perform variable selection. The names of these methods are exhaustive search, backward elimination, and forward selection**.**

The result from the exhaustive search shows that the model with 5 subsets would perform better than the model with 6 subsets. We can see that the model with 5 subsets have an adjusted R of 0.597 and a bic of -136.2123 while the model with 6 subsets has an adjusted r of 0.595 and a bic of -131.1663.

**Table

Description automatically generated**

Using the backward elimination and forward selection method, we can also see that the model with 5 subset is the best model mirroring the same result as the exhaustive search method.

Using all 3 methods, the best model will be the model with 5 subsets, followed by the model with 6 subsets and finally the model with 4 subsets. Below is the graphical representation of the backward elimination and the forward selection method of variable selections.

**Chart, waterfall chart

Description automatically generated**

From the above analysis, we can conclude that the model with 5 subsets is the better model. Even though it has a small, adjusted r square of 0.597, it still performed better than the rest of the model. The closest model to it is the model with 6 subsets with has an adjusted r square of 0.595.

The variable ‘Household’ with a p-value of 0.7 does not have a significant effect on the price of the house in California. Removing this variable from the model give a better performance.

**Chart, histogram

Description automatically generated**

Looking at the histogram of the residuals, we can see that we have a normal distribution.

**Summary and conclusion**

looking at the overall data, we were able to pick out the variable that determine the price of the house in California and predict the future house price. All the regressors are significant at 5% except the ‘households’ variable which as a p-value of 0.7. This means that the number of households do not play a significant role in determining the value of the house.

3 variable selection method was used, and they all gave the same results.

The best model was the model with 5 subsets which had high error rate and low adjusted R square. We can say that this model is not a good model.