**Credit Limit Project**

The dataset has 400 observations of 7 variables that was simulated. The dataset was from the textbook, An introduction to statistical learning: With applications in R. Springer by James G., Witten D., Hastie T., and Tibshirani R from 2013. The purpose of this project is to use the dataset to create a predictive model for an individual’s credit limit based on variables below:

* Balance : Individual’s average credit balance [int]
* Income: Individual’s income in $10,000 dollars [num]
* Rating: Individual’s credit rating [int]
* Cards: Number of credit cards the individual’s own [int]
* Age: Individual’s age [int]
* Education: Number of years of education [int]

Used the cor() function to check for correlation between the variables. The threshold used to compare the independent variables to check for multicollinearity is |.90| to determine whether if the independent variables should be drop or create interaction form. It is important to get rid of multicollinearity because it would result in bias because the betas or the estimate values become “noisy”.

A picture containing text

Description automatically generated

Before we can run a regression model, the data must be clean. The boxplot() function was used to visualize the observations to determine which variables had outliers. The outliers was removed using the interquartile range method.

Boxplot of the dataset:

Chart, box and whisker chart

Description automatically generated

Observed that only the variables income, limit, rating, and cards have outliers based on the boxplot. To remove the outliers, subset of the datasets was created by using the cutoff values to determine the which observations to keep or remove for each variable. The cut off values was calculated by first determine what the inter-quantile range is which is Quantile3 – Quantile1 from the normal distribution of the box and whisker plots. The range of the observations that was kept was greater than Q[1] – 1.5\*iqr and less than Q[2] + 1.5\*iqr. *(Noted: Q, Q2, Q3, Q4, iqr, iqr2, iqr3, and iqr4 are matched to the correlated number at the end)*

Next step is to validate the dataset by splitting the dataset into train and test dataset. Used the 70% of the data for training and the rest of the 30% was used for the test data. It is importance to split the data because it ensure an unbiased prediction and prevent the model from overfitting, thus making sure the data model created from machine learning is accurate. After splitting the dataset, the train\_data has 259 observations and the test\_data has 111 observations.

Next step is to run the regression with all of the independent variables.

Table

Description automatically generated

Must remove statistically insignificant variables which are variables that are greater than the 0.05 p-value. The variables balance and age is greater than the p-values, thus must be drop. The variable age was dropped first to see if the variable balance is still consider statistically insignificant.

Table

Description automatically generated

After running the regression without the variable age, the variable balance is statistically significant now. Thus the variable balance does not have to be drop now, but the variable education has become statistically insignificant because the p-value is now .20017, which is greater than 0.05.

Table

Description automatically generated

Now that all variables are statistically significant, can create prediction on individual’s credit limit based on the regression that was conducted to the train\_data to the new test\_data. Used the function predict() to predict the values of individual’s credit limit from the test\_data.

To measure the accuracy of the predictive model we must calculate the RMSE (Root mean square error) and R-squared of the predicted\_values (predicted individual’s credit limit) and test\_data$limit (actual individual’s credit limit).

Used the rmse() function to calculate the root mean square error, which was 170.7905. This value is low compared to the relative scale of the dependent variable. Low rmse is ideal for models because it indicates a good model performance.

Used the r2() function to calculate the R-squared which was .989947, which means that 98.9947% variation of the variable credit limit of the model can be predicted based on the independent variables of the dataset. The higher the R-squared is ideal because it means the model is accurate.

The regression model for predicting limit:

y = -288.39290 + .12883\*balance + 1.81772\*income + 14.35818\*rating – 72.15359\*cards + u