CAP 4773 – Introduction to Data Science and Analytics

Department of Computer and Electrical Engineering and Computer Science

Final Paper Report

Student: Artur Nascimento

Professor: Raquel Assis

The dataset used during this experiment is called “Credit” with the source being a package called “ISLR,” which encloses data for a book known as *Introduction to Statistical Learning with Applications in R*. This dataset has a large focus on credit card balance and the ability to predict which customers will default on their credit card debt, the database of values contains 400 observations relating to 12 distinct variables. Out of the 12 variables, 4 are considered categorial values which assign the data into specific groups such as Gender, Student, Married and Ethnicity. The remaining 8 features are split between integer and numerical datatypes, the integer variables contain only whole numbers, those being ID, Limit, Rating, Cards, Age, Education and Balance, with numerical data fluctuating between decimal and whole numbers alike, this datatype is found on the feature known as Income.

Diving deeper into the breakdown of the categorical data, the following table shows the split of data points based on grouping of the 400 observations present on the dataset:

Category	Yes	No	
Student	40	360	
Married	245	155	
Category	**Male**	**Female**	
Gender	193	207	
Category	**Asian**	**African American**	**Caucasian**
Ethnicity	102	99	199

This written report will address which variable shows the higher prediction to the Balance response from the set of given features. The best approach on solving this problem relates to feature selection, by using such technique the model will be fit to a response based on the determination of which specific features are associated with the wanted response. With the results gathered, the problem will be broken down into a multiple linear regression model, where the test static and p value will be taken into account. Lastly, a simple linear regression model will be created where the most strongly related feature to the Balance response will be shown.

Using best subset, forward, and backward selection on all 8 features in order to construct models for predicting the response of average credit card balance. The appropriate number of features predicted to create the best model based on the estimated mean squared error (MSE), using regression model accuracy metrics like adjusted R^2, Mallows' C_p and BIC, can be seen on the table below:

Metrics	Best Subset	Forward Selection	Backward Selection
Adjusted R^2	5 Features	5 Features	5 Features
Mallows' C_p	5 Features	5 Features	5 Features
BIC	2 Features	2 Features	2 Features

Since the experiment requires all 2^p possible models to be compared, the ideal algorithm to predict the number of features in this scenario would be the best subset selection. The metrics were used because it shows an indirect estimate that adjusts training error for model complexity. Although two metrics in this problem happen to be similar, being the adjusted R^2 and Mallows' C_p, they do not lead to the same conclusions. Adjusted R^2 describes how well the model explains the observed data, while Mallows' C_p which is a variant of AIC measures how well the model will fit new data. Additionally, BIC which was the only metric that showed a distinct difference, does so by having a stronger penalty for including additional variables to the model. Therefore, a conclusion can be drawn that the overall best fit model will have an estimated number of features around 5 and 2. The names of features mentioned on the table below will then be applied to a multiple linear regression model in a later section, thus allowing the assumption that the given response can be estimated by a set of linear combination of features, shown by a comparison between their p value and f statistic:

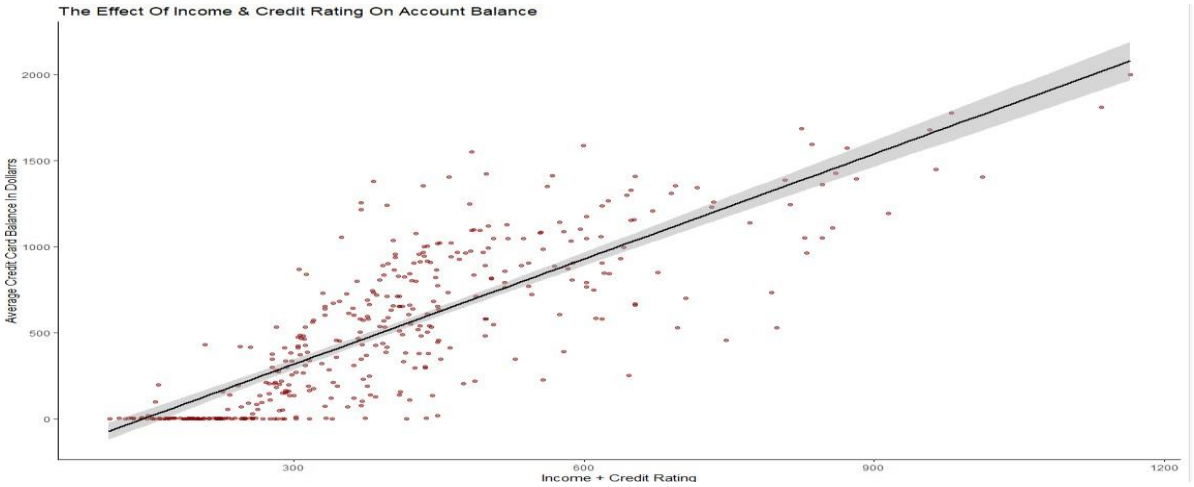
Experiment	Names
Model 1 -> 2 Features	Income, Rating
Model 2 -> 5 Features	Income, Limit, Rating, Cards, Age

Knowing what features must be tested for both models in this experiment, a mathematical notation for the null hypothesis for both multiple linear regression models can be made:

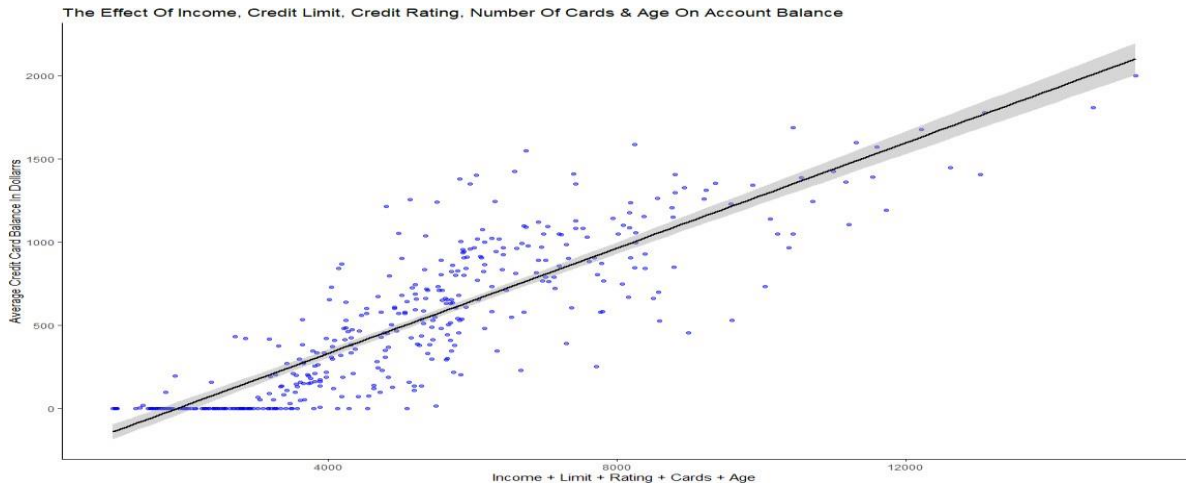
$$Model\ 1 \rightarrow H_0: \beta_1 = \beta_2 = 0$$
$$Model\ 1 \rightarrow H_A: At\ least\ one\ \beta_i \neq 0$$

$$Model\ 2 \rightarrow H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$Model\ 2 \rightarrow H_A: At\ least\ one\ \beta_i \neq 0$$

Model 1 Scatterplot:



Model 2 Scatterplot:



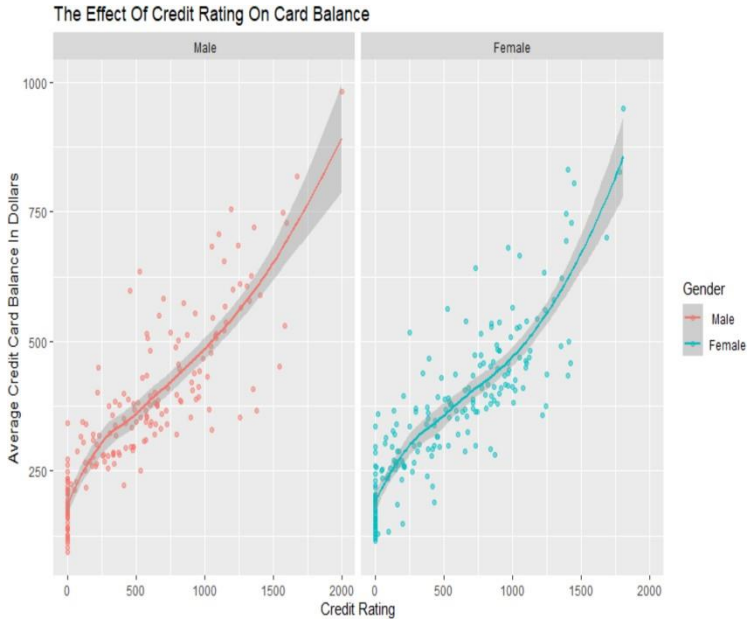
Looking at both graphs there is a clear linear relationship between the feature of both models and the Balance response. However, even though both plots look similar their summary outcomes are surprisingly different.

Metrics	Model 1	Model 2
Residual Standard Error	162.9	161.6
Adjusted R-Squared	0.8751	0.8781
F – Statistic	1391	567.4
P – Value	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$

The table above shows that even though model 1 only has 2 features, Income and Credit Rating, it shows the same level of predictability at a much lower complexity level than the 5 features of model 2. Meaning, since the test statistic F = 1391 is so large that it is located far into the right tail of the distribution under H_0, given that the P value found is equal to $2.2 * 10^{-16}$ satisfying the comparison of $P < 0.001$, an assumption can be made that the null hypothesis in this multiple regression problem will be rejected. Therefore, the probability of the data appearing under the null hypothesis is close to zero, since at least one of the population intercept and slope will be equal to 0, which in this case both features are. Nevertheless, model 2 falls under the same category and the null hypothesis will also be rejected, thus both models show a definite and consequential relationship between at least one of their used features or all of them when it comes to the response known as Balance. Furthermore, the percentage value of the adjusted R^2 for both models have shown a smaller difference between the observed data and the fitted values. In easier terms, about 87% of the variation in Balance can be attributed to changes within the features found on both models. Therefore, a conclusion can be made that there may exist a linear relationship between each unique feature used to the response in question, although, some features may not be as strongly related to the response as the others.

Arriving at the end of the report we must go back to the main experiment goal. Which out of the 8 features shows the greatest relationship to the response? During the feature selection, the one variable that was present on every model happened to be the one called Rating, which is to be expected since it measures the creditworthiness of a borrower, thus the likelihood that the customer will repay their debts. With that said, using a simple linear regression we are able to find the variations of the Balance response that is attributed to the Rating feature. The number happens to be alarmingly higher than others, having an attributed prediction response around 75% according to the adjusted R^2 metric. Which if taken as a direct comparison with models 1 and 2 seen previously, which also contained the feature on their calculations. A conclusion can be made that the extremely high confidence level on prediction was directly

related to the sheer presence of the Rating variable, leading to the similarity in the linear regression shown by their scatterplots. Additionally, Since the T statistic measures the number of standard errors the coefficient is from zero, we can conclude that the greater the T value, the greater the confidence that the feature in question can be considered a predictor. Thus, as the T value for Rating for both model 1 and 2 was the highest at 45.81 and 2.553 respectively, it shows the most relation to the Balance response since it provides the greater evidence against the null hypothesis. Below is the last graph showing the linear relationship between the Rating feature to the Balance response with different color points showing the group relationship between Male and Female borrowers:



As shown by the scatter plot, male and female have the same correlation strength when it comes to average credit card balance, which means that the distribution is very similar along the curve, without a color class difference between them we would be unable to witness a major difference.

As this experiment comes to an end, the major limitation found with interpreting this dataset derives from the short number of observations. This problem becomes more apparent once trying to use the K-nearest neighbors to predict whether the variables fit a specific model grouping, like the 4 categorical variables mentioned at the very start. Only the feature known as Student shown an acceptable predication error of 8%, most likely because it provided the greatest majority, while every other categorical example fell around the range of 50% prediction error as their distributions were much similar throughout as shown by the scatterplot above.