Credit Report

Matthew Oehler

January 29, 2018

1 Introduction

Credit card companies, like any other company, want to be profitable. The biggest source of income for credit card companies is the interest collected from card holders that have outstanding balances. As a result, credit card companies are interested in maximizing their profits through optimizing the profits based on this type of interest. Card holders with a low monthly balance are neither a large risk, nor a large profit for credit card companies. Card holders with a high monthly balance yield a high risk to credit card companies, because there isn't a good way for the company to enforce payment of customers who default on their loans. So ideally, credit card companies want to maximize the number of card holders with a moderate monthly balance. This provides the happy medium for earning profits from interest, without a large risk of not getting paid. Since each potential card holder provides an opportunity to earn or lose money, credit card companies are interested in being able to predict the monthly balance of potential card holders before issuing them a card. In this study we will be assessing the relationships of certain demographics of card holders and their monthly balances. Using a statistical model, we performed an analysis to a) determine which factors are significantly related with one's credit card balance and b) determine how accurately we can predict an individuals credit card balance based on some of their demographic information.

2 Data

2.1 Data Description

For this study we will be using observations of 11 different variables for 294 different credit card holders. Each of the variables measured and their corresponding descriptions are shown below in Table 1.

	Variable	Description
1	Income	Card holders annual income in thousands.
2	Limit	Card holders credit limit.
3	Rating	Credit rating - similar to a FICO credit score but used internally by the company.
4	Cards	Number of open credit cards (including the current card) of the card holder.
5	Age	Age of the card holder.
6	Education	Years of education.
7	Gender	Gender of the card holder.
8	Student	Card holder is a full-time student.
9	Married	Card holder is married.
10	Ethnicity	Card holders ethnicity (Caucasian, Asian, or African-American.)
11	Balance	Current credit card debt.

Table 1: Description of Data

2.2 Data Exploration

First we examined the quantitative variables in the data set using a scatter-plot matrix shown in Figure 1. This figure displays each a scatter-plot of each possible pair of the quantitative variables

in the data set, including: Income, Limit, Rating, Cards, Age, Education, and Balance. Also included in the plot is the correlation between all possible pairs of quantitative variables, and an approximated density curve of values for each individual variable in the figure. Right away, it can be seen that the variables Limit and Rating are very highly correlated.

Next we created box plots to explore the characteristics of the categorical variables. The plots for each of the categorical variables including: Gender, Student, Married, and Ethnicity are shown in Figure 2. At first glance, we noticed that student status might have a significant effect on monthly balance, but the other categorical variables didn't seem overly significant.

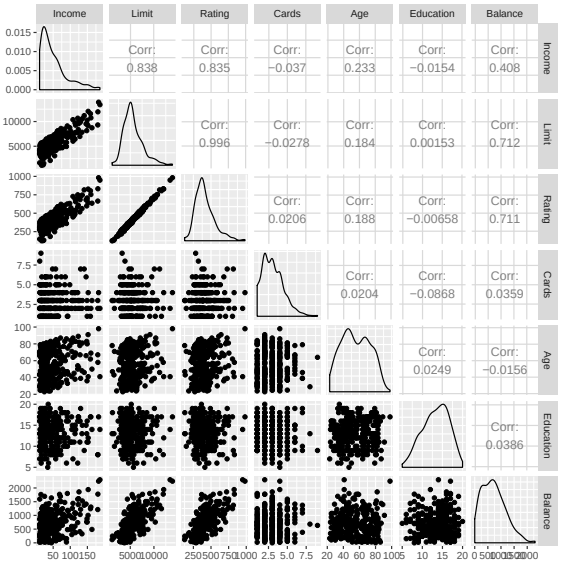

Figure 1: Scatter-plot matrix of the quantitative variables

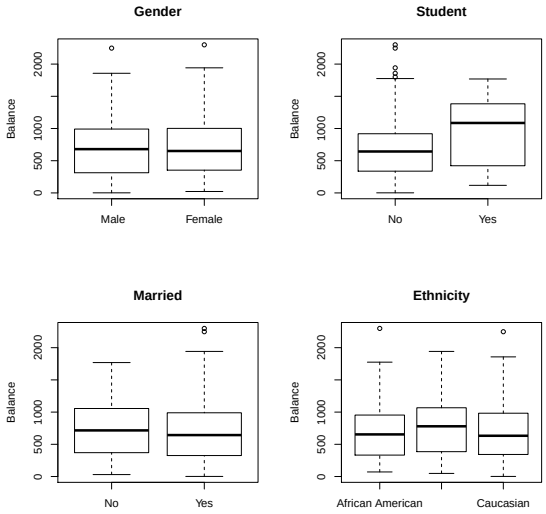

Figure 2: Box plots of categorical variables

3 Methodology

3.1 Model Description

As stated in the introduction, we wanted to focus on using the demographics of a card holder to predict their monthly balance. Since we have multiple covariates being used to predict a single quantitative response variables, we determined the multiple linear regression would be an appropriate modeling technique to address our questions of interest. Using a multiple linear regression model, we will be able to see which variables have a significant relationship with monthly credit card balance, and we will be able to use our model to predict the monthly balance for a potential card holder with certain demographics. Multiple linear regression requires that we make assumptions of linearity, normality, equal variance, and independence. These assumptions will be addressed in detail in Section 3.3.

The multiple linear regression model can be expressed with matrices and vectors as shown in Equation 1.

$$Y = \mathbf{X}\beta + \epsilon, \qquad \epsilon \sim \mathbf{N}\big(0, \sigma^2\big) \tag{1}$$

In Equation 1, Y refers to the 294×1 vector of credit scores for each observed person in the data set. All other observations for each individual are in corresponding rows of the $\mathbf{X}$ which is a 294×9 matrix. Each column of $\mathbf{X}$ corresponds with one of the explanatory variables (Income, Limit, ..., Ethnicity). Finally, ϵ is a 294×1 vector of residuals that represents the difference of the observed credit balance for each person and the mean credit balance of a person with identical demographic measurements.

3.2 Variable Selection

As was noted previously, Limit and Rating are very highly correlated. Since collinearity causes problems when doing linear regression we chose not to include Limit, because Limit is essentially a function of one's credit rating. To avoid the possibility of over-fitting the data, we decided to select a "best" subset of the variables to use in the model. Since we were primarily focused on prediction, we decided to uses the Akaike Information Criterion (AIC) to determine which variables to include in the model. The AIC has a weaker penalty than other information criterion, and so we knew that it would likely give us a model with more covariates, which often helps to improve the model's predictive accuracy. Using AIC, we ended up with four of the original covariates: Income, Student, Age, and Rating. Additionally, we were interested in seeing if there was a significant interaction between student status and income, and so we included a term for that in the model. The resulting model is shown in Equation 2, and has been written in standard algebraic notation for the sake of easier interpretation.

$$\hat{y} = \beta_0 + \beta_1(Student) + \beta_2(Income) + \beta_3(Rating) + \beta_4(Age) + \beta_5(Student \times Income) \quad (2)$$

In Equation 2, $\hat{y}$ represents the expected credit card balance of a person based on their demographics. The intercept term, β_0 is interpreted as the expected credit card balance of an individual who is 0 years old, isn't a student, and has no income or credit rating. By itself, β_0 doesn't provide any realistic insight, but it is necessary for improving the fit of the model. β_1 represents the average difference in credit card balance for a student as opposed to a non-student (holding all else constant). The coefficients $\beta_2 - \beta_4$ represent the average change in monthly credit card balance for each unit increase in the corresponding demographic. For example, holding all else constant, for each additional year of age, we an individuals credit card balance changes on average by β_4 dollars. Lastly, we have the interaction term, which expresses a multiplicative relationship between Student and Income, and is slightly more complicated to interpret. The average change in monthly balance per change in income for non-students is expressed entirely by β_2 (holding all else constant). However, for students we say that for each \$1000 increase in annual income, a student's credit card balance is expected to change by $\beta_2 + \beta_5$ dollars (again, holding all else constant). In the next section we will fit the model to calculate each of the β values, and verified that all of the required assumptions are maintained.

3.3 Model Diagnostics

In Section 3.1, we briefly mentioned that multiple linear regression requires four assumptions to hold. In this section, we show and/or justify that all of the assumptions held during this analysis.

We addressed the first assumption, of linearity using added variables plots which are displayed in Figure 3. These plots show the relationship of each variable with the response variable, after the effects of all other variables have been accounted for. We can see that the linearity assumption holds because the relationship for Age, Income, and Rating (all of the quantitative variables in the model) all appear to be linearly related with Balance.

Next we assessed the normality, and equal variance assumptions by looking at plots of the residuals. In Figure 4, we can see that the histogram of residuals approximately follows a normal distribution, which verified the necessary assumption of normality. The plot of fitted values vs. residuals shown in Figure 5, shows us that the variance throughout the dataset is approximately uniform, and so we concluded that the assumption of equal variance holds as well.

Lastly, examined the assumption of independence through critical thinking, rather than with a plot or other diagnostic test. It follows sound logic that, in general, the credit card balance of one individual, doesn't have an effect of the credit card balance of another individual. This implies that our response variable meets the independence assumption. We also used Cook's Distance to test for points that might be outliers. Some points were flagged based on Cook's distance rule of thumb, but after looking at those points in relation to other points on the fitted values vs. residuals plot, we determined that they aren't truly outliers.

Since we verified all the required assumptions we proceeded to inspect the results of the model, the details of which are contained in the following section.

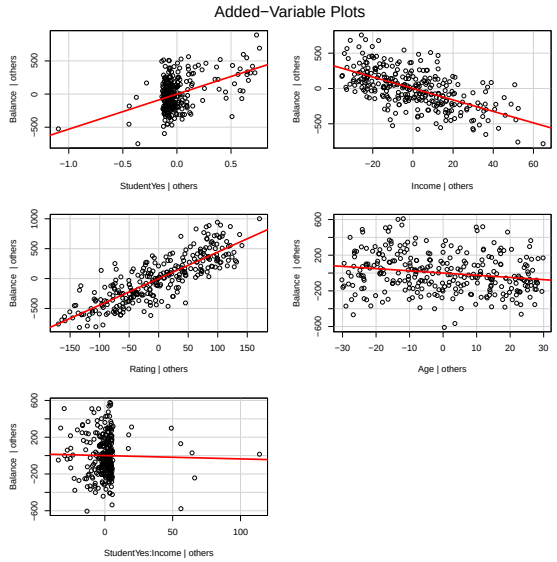

Figure 3: Added Variable Plots used to address the linearity assumption

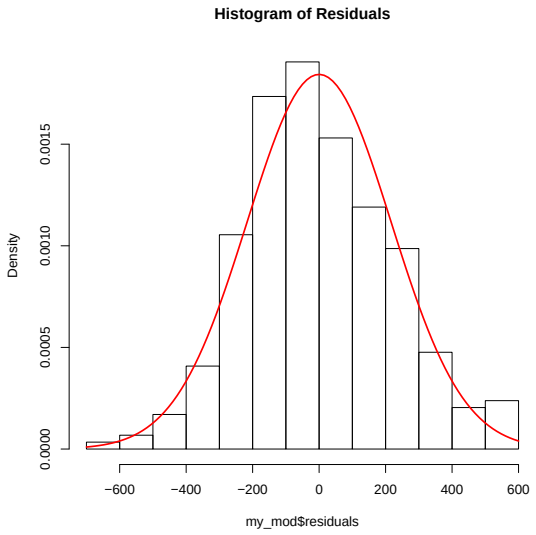

Figure 4: Histogram of the residuals

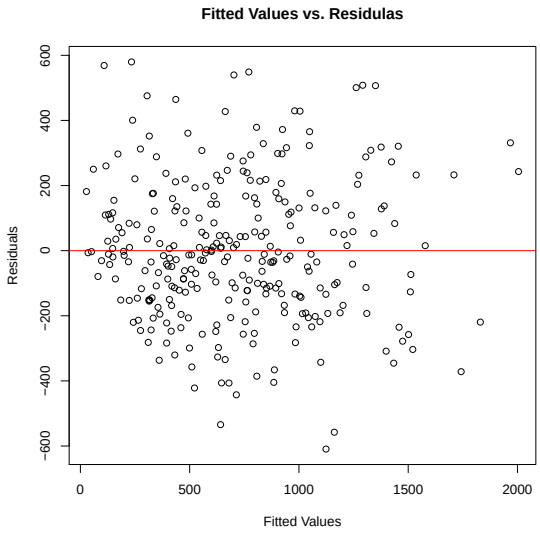

Figure 5: Fitted values vs. residuals

4 Results

4.1 Parameter Estimates

After verifying that all of the necessary assumptions hold we looked at the calculated coefficient estimates (see Table 2) to see which variables are significant. It turns out that all of the variables in our model except for the interaction between Student and Income are significant. The interaction term was determined to not be statistically significant with an associated p-value of 0.7149, which is greater than $\alpha = 0.05$. The resulting R^2 value was .7648, meaning that 76.48% of the variation in Balance is explained by the variables in our model.

We also calculated 95% confidence intervals for each of the model coefficients, which are displayed in Table 3. An example interpretation using these intervals would be as follows: We are 95% confident that holding all else constant, an individual's credit card balance would increase by an amount between 4.10 and 4.78 on average for each increase of 1 of their credit rating.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-629.5725	64.1021	-9.82	0.0000
StudentYes	529.5890	62.3208	8.50	0.0000
Income	-8.1121	0.6309	-12.86	0.0000
Rating	4.4391	0.1718	25.84	0.0000
Age	-2.4728	0.7548	-3.28	0.0012
StudentYes:Income	-0.3608	0.9865	-0.37	0.7149

Table 2: Model Summary

Coefficient	2.5 %	97.5 %
(Intercept)	-755.74	-503.40
StudentYes	406.93	652.25
Income	-9.35	-6.87
Rating	4.10	4.78
Age	-3.96	-0.99
StudentYes:Income	-2.30	1.58

Table 3: 95% confidence intervals for the estimates of the model coefficients

4.2 Performance Evaluation

The main goal of this analysis is to be able to predict the credit card balance of a potential card holder, before issuing them a credit card. Consequently, it is important to assess the prediction capability of the model. We assessed the model's prediction performance using a cross validation. For 10,000 iterations, we fit our model on a randomly selected 70% of the data set, then we used the remaining 30% of the data to test the model's prediction accuracy. We calculated the coverage, bias, and root prediction mean squared error (RPMSE), and interval width over all of the iterations. The average results are shown in Table 4. The results are good since we got the desired coverage of 95% for prediction intervals generated each iteration, and because the bias is close to zero. As a result we are satisfied with the prediction capability of the model.

Statistic	Result
Coverage	0.95
Bias	0.01
RPMSE	222.31
Interval Width	873.35

Table 4: Prediction Assessment

5 Conclusions

From this analysis we were able to draw several conclusions. Through the use of multiple linear regression we determined the relationships of several demographic variables of an individual and their credit card balance. We also determined that we can use this model to predict the credit card balance of an individual, given their demographic information. This analysis could be improved further by finding ways to reduce the width of the prediction intervals. Additionally, other demographic factors could be explored, and perhaps there are other variables that are better indicators of one's credit card balance. Overall, using the data that was provided, we have shown that, using statistical methods, it is possible to make an educated decision in regards to issuing credit cards to an individuals based on their demographic information.