Predictive Modeling Process

Abbey Chaver and Tina Huang

November 4, 2016

Abstract

In this report, we will be examining the variables in the Credit.csv dataset from *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. We will also be building different regression models in order to determine how to best predict Balance given ten predictors, using similar approaches as in Chapter 6.

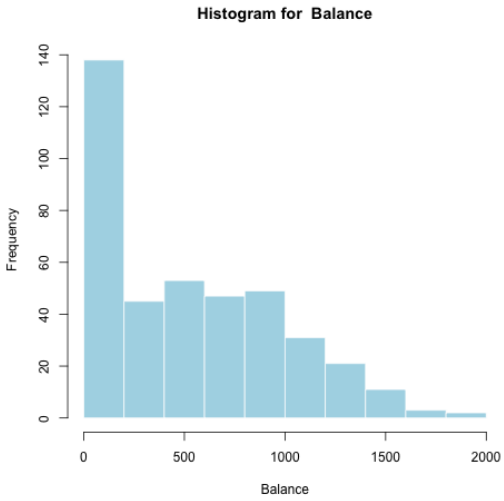
Introduction

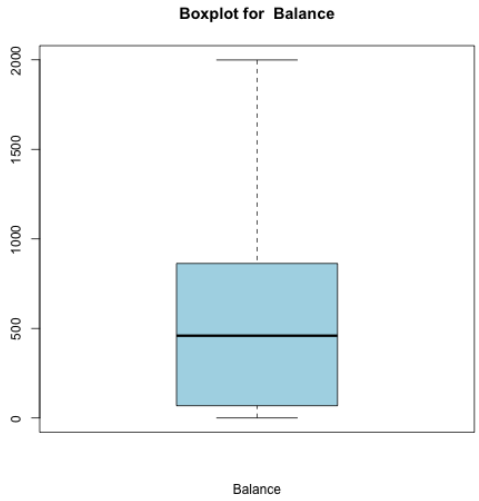
The purpose of this report is to determine the best model for predicting Balance given ten different predictors, including quantitative variables such as income and qualitative variables such as ethnicity. The distributions of these variables will be examined through summaries and plots, and five different regression models will be applied to the data; ordinary least squares, ridge, lasso, principal components, and partial least squares. These five models will be compared by looking at their respective coefficients and also by comparing their mean squared errors.

Data

The Credit.csv dataset has the balance (or the average credit card debt) for individuals with a number of quantitative descriptors including age, cards (the number of credit cards), education (the number of years of education), income (thousands of dollars), limit (their credit limit), and rating (their credit rating), as well as qualitative descriptors such as gender, student (whether or not they're a student), married (whether or not they're married), and ethnicity.

Our target variable is `Balance`. We get a sense of how it's distributed with a boxplot and histogram.



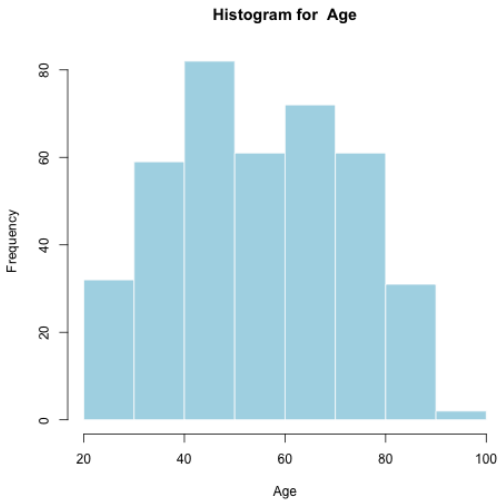


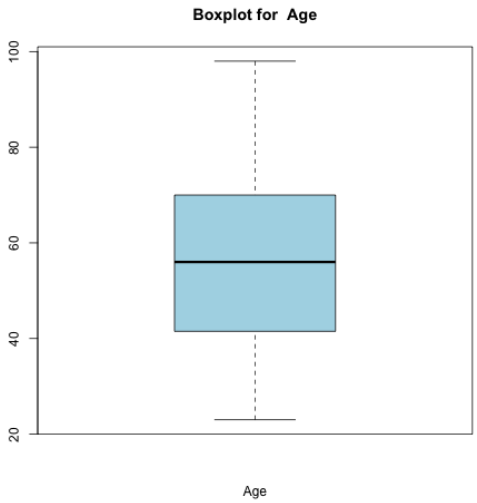
Statistic	Value
Minimum	0
Maximum	1999
Range	1999
Median	459.5
First Quartile	68.75
Third Quartile	863
Interquartile Range	794.2
Mean	520
Standard Deviation	459.8

The average balance falls at about $500,000, with a long tail.A fairly unusual distribution - a huge portion of accounts have balances below $200,000.

Now we look at our predictors, with an initial visual exploration of the quantitative variables.

Age

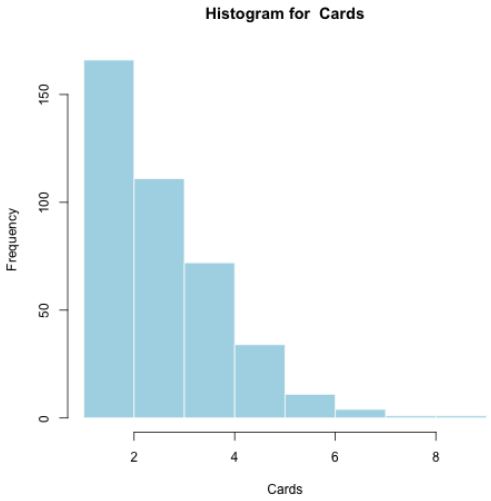


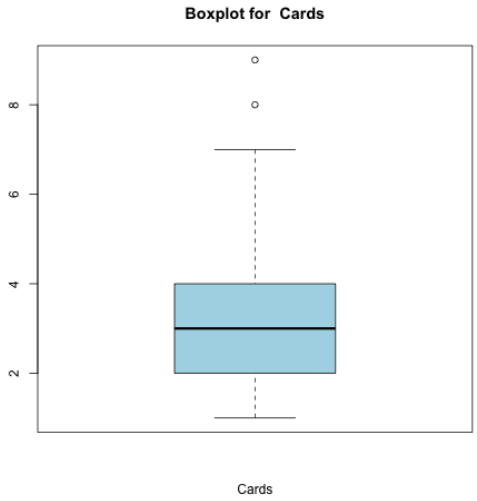


Statistic	Value
Minimum	23
Maximum	98
Range	75
Median	56
First Quartile	41.75
Third Quartile	70
Interquartile Range	28.25
Mean	55.67
Standard Deviation	17.25

Account holders are on average in their fiftes, with a generous spread. Ages are fairly normally distributed.

Credit Cards

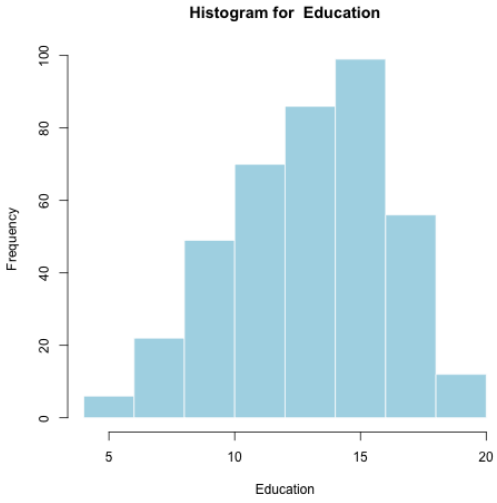


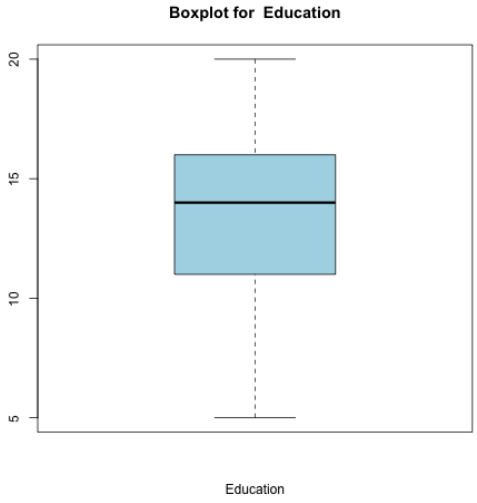


Statistic	Value
Minimum	1
Maximum	9
Range	8
Median	3
First Quartile	2
Third Quartile	4
Interquartile Range	2
Mean	2.958
Standard Deviation	1.371

The average account holder has around 3 cards. Cards look more exponentially distributed.

Years of Education

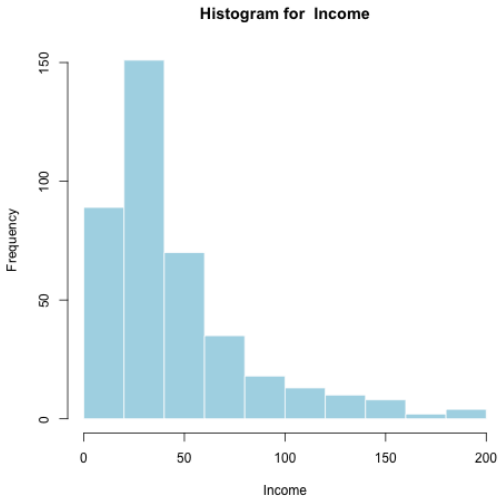


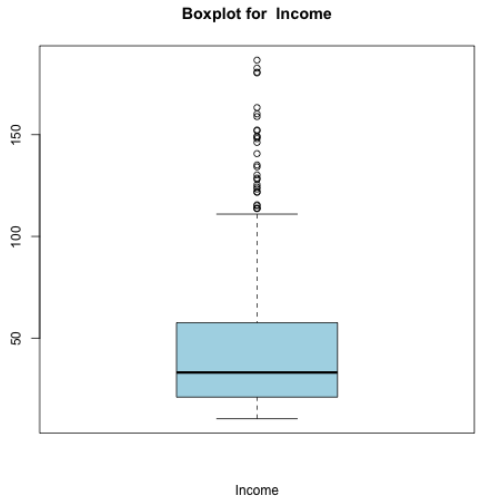


Statistic	Value
Minimum	5
Maximum	20
Range	15
Median	14
First Quartile	11
Third Quartile	16
Interquartile Range	5
Mean	13.45
Standard Deviation	3.125

The interquartile range of education falls between 11 and 16 years - highschool and college. Years of education is fairly normal with a bit of a left skew.

Income

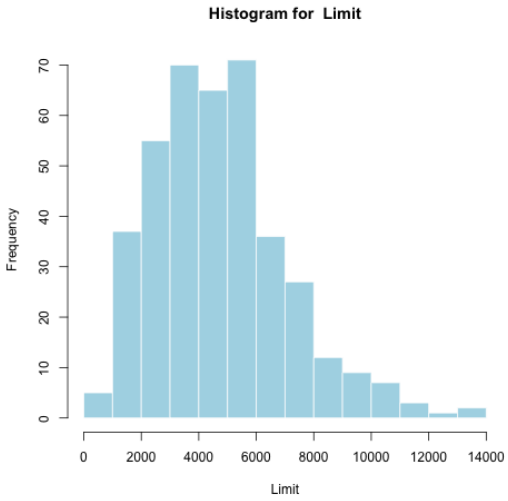


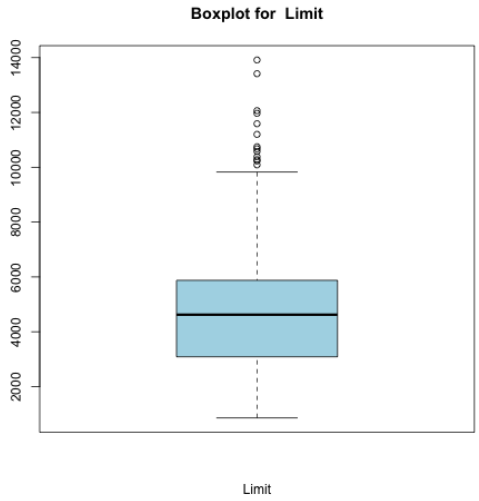


Statistic	Value
Minimum	10.35
Maximum	186.6
Range	176.3
Median	33.12
First Quartile	21.01
Third Quartile	57.47
Interquartile Range	36.46
Mean	45.22
Standard Deviation	35.24

The mean seems to be around $30,000 a year. Incomes skew right in this data set.

Credit Limit

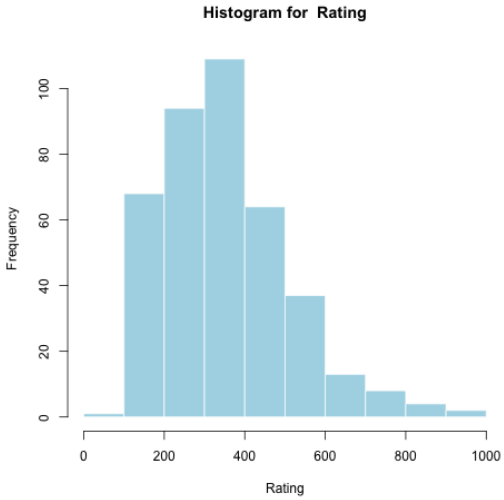


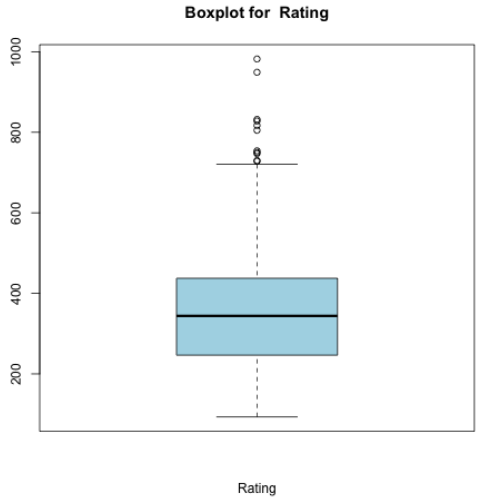


Statistic	Value
Minimum	855
Maximum	13913
Range	13058
Median	4622
First Quartile	3088
Third Quartile	5873
Interquartile Range	2785
Mean	4736
Standard Deviation	2308

The average credit limit is around $5000, with a long tail. Credit limits are distributed normal with a right skew.

Credit Rating





Statistic	Value
Minimum	93
Maximum	982
Range	889
Median	344
First Quartile	247.2
Third Quartile	437.2
Interquartile Range	190
Mean	354.9
Standard Deviation	154.7

Ratings are a bit less than 400, on average. Rating is distributed similarly to Credit Limit- probably because Credit ratings are used to determine credit limits.

Correlations

We look at the correlation matrix of the quantitative variables to get a visual sense of how they are related.

	Income	Limit	Rating	Cards	Age	Education	Balance
Income	1	0.7921	0.7914	-0.01827	0.1753	-0.02769	0.4637
Limit	0.7921	1	0.9969	0.01023	0.1009	-0.02355	0.8617
Rating	0.7914	0.9969	1	0.05324	0.1032	-0.03014	0.8636
Cards	-0.01827	0.01023	0.05324	1	0.04295	-0.05108	0.08646
Age	0.1753	0.1009	0.1032	0.04295	1	0.003619	0.001835
Education	-0.02769	-0.02355	-0.03014	-0.05108	0.003619	1	-0.008062
Balance	0.4637	0.8617	0.8636	0.08646	0.001835	-0.008062	1

Most variables look positively correlated with `Balance`. `Income`, `Limit`, and `Rating` all seem to have particularly high correlations and less noise. However, they look closely correlated with each other, which can be a problem for linear modeling. Fortunately, our dimension reduction techniques can ameliorate some

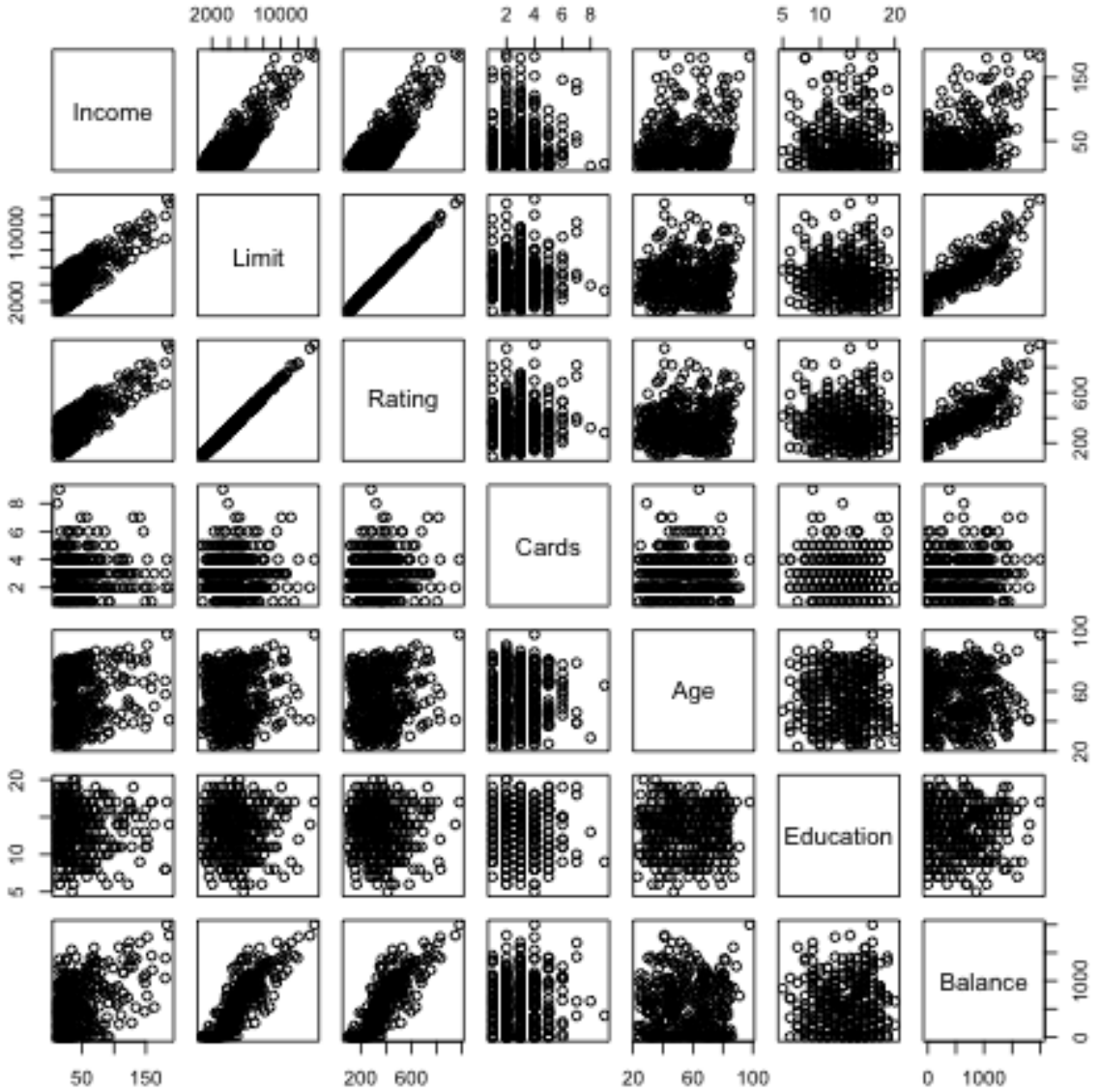
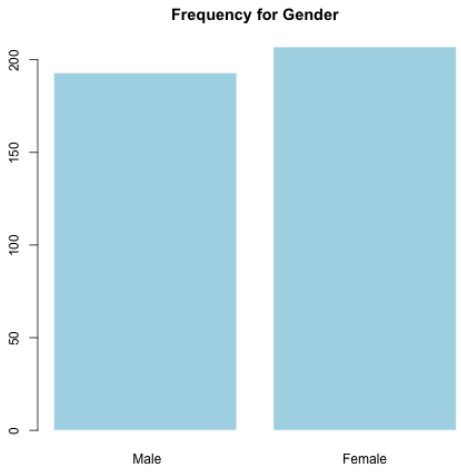

Income
Limit
Rating
Cards
Age
Education
Balance
2000 10000
2 4 6 8
5 10 20
50 150
200 600
20 60 100
0 1000

Figure 1:

of those issues- we may hypothesize that they will do particularly well on this data.

Now we consider qualitative variables, with histograms and conditional boxplots (we look at the values of `Balance` grouped by the variable in question).

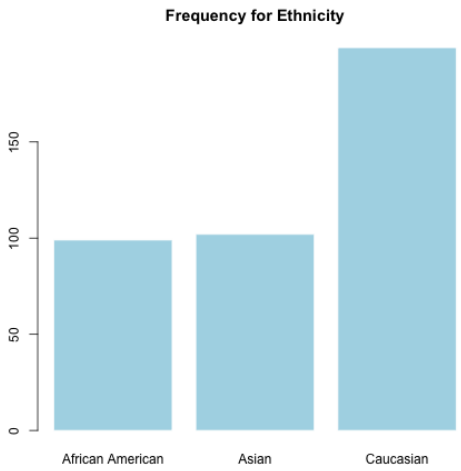
Gender

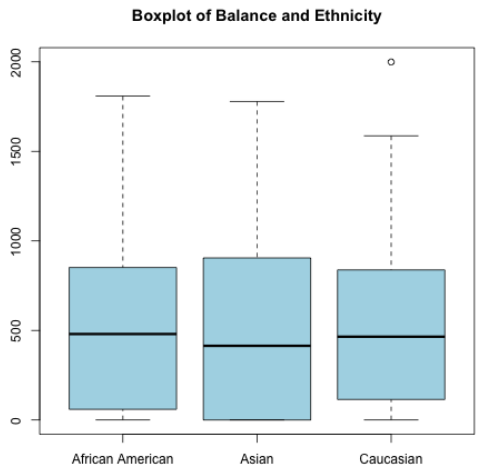


Gender	Frequency
Male	0.4825
Female	0.5175

Accounts are split fairly evenly, with a few more held by women. Average account balance is about the same, but the spreads are slightly different between genders.

Ethnicity

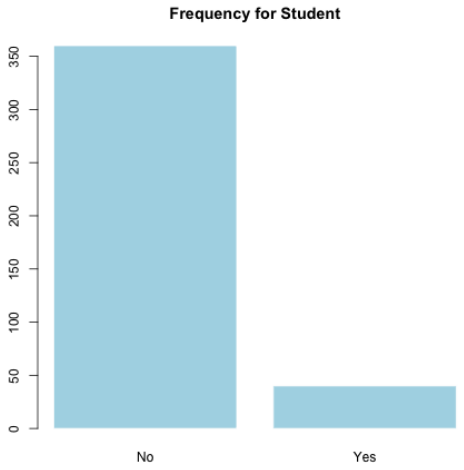


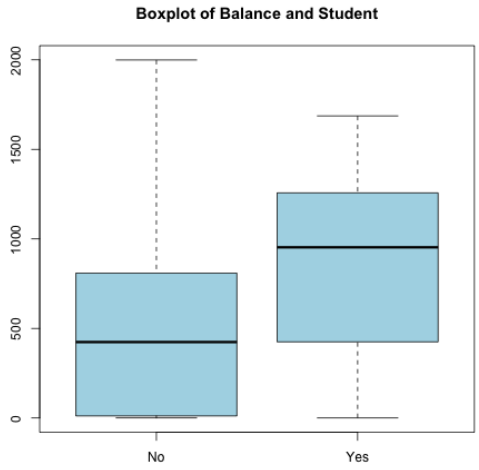


Ethnicity	Frequency
African American	0.2475
Asian	0.255
Caucasian	0.4975

Account holders are about a quarter African American, a quarter Asian, and half Caucasian. Average account balance is about the same across ethnicities, with Asian accounts being distributed a bit more widely.

Student

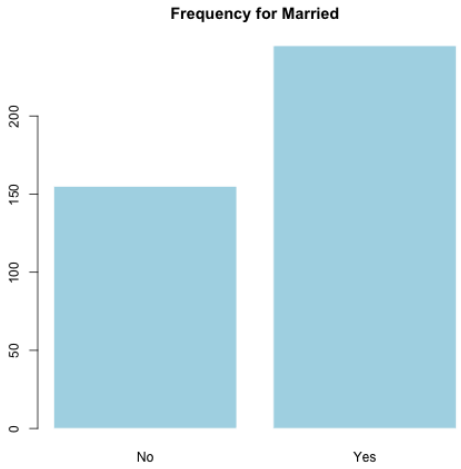


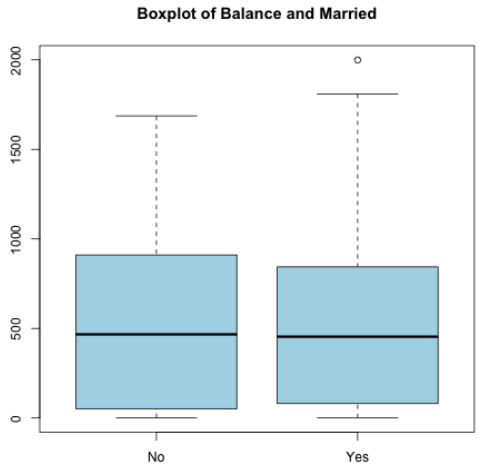


Student	Frequency
No	0.9
Yes	0.1

Most account holders are not students. Students generally seem to have much higher balances - surprising considering that they rarely have significant income.

Married





Married	Frequency
No	0.3875
Yes	0.6125

A majority of account holders are married, but there are plenty of nonmarried account holders. Balances seem to be similarly distributed for married and nonmarried accounts.

Anova on Categorical Variables

Table 13: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data$Gender	1	38892	38892	0.1949	0.6591
data$Student	1	5623889	5623889	28.19	1.841e-07
data$Married	1	16976	16976	0.08509	0.7707
data$Ethnicity	2	58038	29019	0.1455	0.8647
Residuals	394	78602117	199498	NA	NA

Only Student produces a significant F statistic - the other categorical variables do not have a strong relationship with Balance.

Overall, we get the sense that `Income`, `Rating`, `Limit`, and `Student` will be particularly useful predictors for `Balance`.

Data Processing

Since we are building regression models in this project, we need to convert all of our categorical variables to quantitative ones. We do this by creating new binary variables, or dummy variables. Additionally, we

need to mean-center and standardize all the variables so that they all have comparable scales, as dimension reduction methods can be greatly thrown off by predictors with wildly different scales.

Methods

In this project, we attempt to create an accurate predictive model for `Balance` given our predictors using regression methods. We will evaluate 5 methods: ordinary least squares on the data to serve as our benchmark regression model. In addition, we will also perform two shrinkage methods - ridge regression and lasso regression - and two dimension reduction methods - principal components regression and partial least squares regression. These last four methods are all variations on least squares regression which attempt to mitigate particular weaknesses of ordinary least squares.

Generally, predictive models wrestle with what is known as the *bias-variance tradeoff*. When we fit a model to data, we want the model to be as close as possible to the actual underlying distribution. For a linear model, we use the assmuption of linearly related variables and a normally distributed error term centered at zero. Under this model, we can show that least squares estimates of our variable coefficients are *unbiased*: that given infinite samples, we will find the true coefficient. But in reality, we rarely have unlimited samples, and a result of this is that ordinary least squares regression has high *variance*: estimates vary quite a bit based on the training data. The result is *overfitting*, when the model is too specific to the training data, and fails to provide accurate predictions when given new data.

Regularization

Regularization methods attempt to reduce variance by "shrinking" coefficients towards zero, which they do by adding a function of the coefficients to the sum or squared residuals typically minimized in least squares regression. The difference between lasso and ridge regression is simply the function - lasso uses the l1 norm (the sum of the absolute value of the coefficients), while ridge regression uses the l2 norm (the sum of the square of the coefficients). The result of this difference is that lasso often sets coefficients to zero and thus functions as a method of subset selection on the predictors, for reasons beyond the scope of this paper. From our original data exploration, we saw that many of the predictors had weak relationships with `Balance`, and are likely to not provide any valuable information for our prediction. Eliminating them from the model by setting their coefficients to zero may decrease our risk of overfitting, so we may hypothesize that lasso will perform particularly well.

Dimension Reduction

Dimesion reduction models take an alternative approach: rather than fitting a model with the original predictors, we create new predictors from linear combinations of the original predictors. Typically, we end up fitting a model with fewer variables than we had originally, which reduces the dimension of the predictor space. This is valuable because in high-dimensional spaces, "distance" as measured by residuals becomes larger and less useful. Another benefit is that when multiple predictors have a strong relationship with each other, we can combine them into a single variable, and reduce the problems created by multicollinearity. As we saw, many of our quantitative variables were highly correlated, so these methods are particularly valuable. The two methods we use are Principal Components and Partial Least Squares, which use similar methods to create these linear combinations of original predictors. Neither clearly dominates the other, so we will evaluate both.

Model Validation

Our ultimate goal is to build an accurate predictive model, so it is critical that we choose the model which will produce the lowest error on new data. To estimate our error statistic, mean squared error, on the different

models, we will reserve a portion of our data as a validation set. We will train our models on the majority of the data, and then calculate mse on the data which was not used to fit the data. This will help us select the model with the lowest bias (closest to the true distribution) that is not overfit on the training data.

Analysis

To look at the distributions of all the variables, we created functions in `code/functions/data_functions.R`, `descr_stats()` for quantitative variables and `qual_descr()` for quantitative variables to create summaries of the variables, as well as any relevant plots such as histograms, boxplots, barplots, and conditional boxplots. To perform the ANOVA tests, we used the `aov()` function.

To prepare the data through processing, we "dummified" the factors using the function `model.matrix()`. We also mean-centered and standardized all the variables using the `scale()` function.

For the ordinary least squares regression model, we used the `lm()` function and retrieved the coefficients by calling `$coefficients` on the output. We calculated the mse of the model by calling `$residuals` on the lm output, squaring the residuals, and finding the mean of the result.

For the shrinkage methods, we used the package `glmnet`. The function `cv.glmnet()` was used to perform cross-validation, and the best value of lambda was found by calling `$lambda.min` on the output. Using this value of lambda, we are able to use the test data set and find the mse by using the function `predict()`, and finding the mean of the squared differences between the predicted outputs and the actual test outputs. We also called the `glmnet()` function to fit the model on the full data set, and used `predict()` again to find the official model coefficients.

For the dimension reduction methods, we used the package `pls`. The functions `pcr()` and `plsr()` were used depending on which regression model we were fitting, with the argument `validation = "CV"` to perform 10-fold cross-validation. We call *validation*PRESS on the output of this function and then find the number of components that yields the lowest mse. To plot the cross-validation errors, we used the `validationPlot()` function. Similar to above, we find the mse by using the `predict()` function, and refit the model using the full data set with the functions `pcr()` or `plsr()` to find the official model coefficients.

To generate the final plots in the results, the `ggplot2` package was used. To format the summary tables, the `pander` package was used.

Results

Table 14: Coefficients for Different Models

	OLS	Ridge	Lasso	PCR	PLSR
Income	-0.5982	-0.56871	-0.55143	-0.5985	-0.59894
Limit	0.9584	0.71866	0.78157	0.9565	0.67806
Rating	0.3825	0.59306	0.51106	0.38449	0.6641
Cards	0.05286	0.04425	0.03884	0.0532	0.04059
Age	-0.02303	-0.02538	-0.01677	-0.02362	-0.02382
Education	-0.00747	-0.00588	0	-0.00747	-0.0065
GenderFemale	-0.01159	-0.01068	-2e-05	-0.01039	-0.0112
StudentYes	0.2782	0.27318	0.26608	0.27843	0.27603
MarriedYes	-0.00905	-0.01103	0	-0.00932	-0.0114
EthnicityAsian	0.01595	0.01638	0	0.016	0.01651
EthnicityCaucasian	0.01101	0.01101	0	0.01012	0.01021

As we can see in Table 1, which has the official coefficients from the Credit dataset for each of the models (Ordinary Least Squares, Ridge, Lasso, Principal Components Regression, and Partial Least Squares Regression), most of the coefficient estimates are similar across the different models but the lasso model does not have coefficient estimates for four of the predictors.

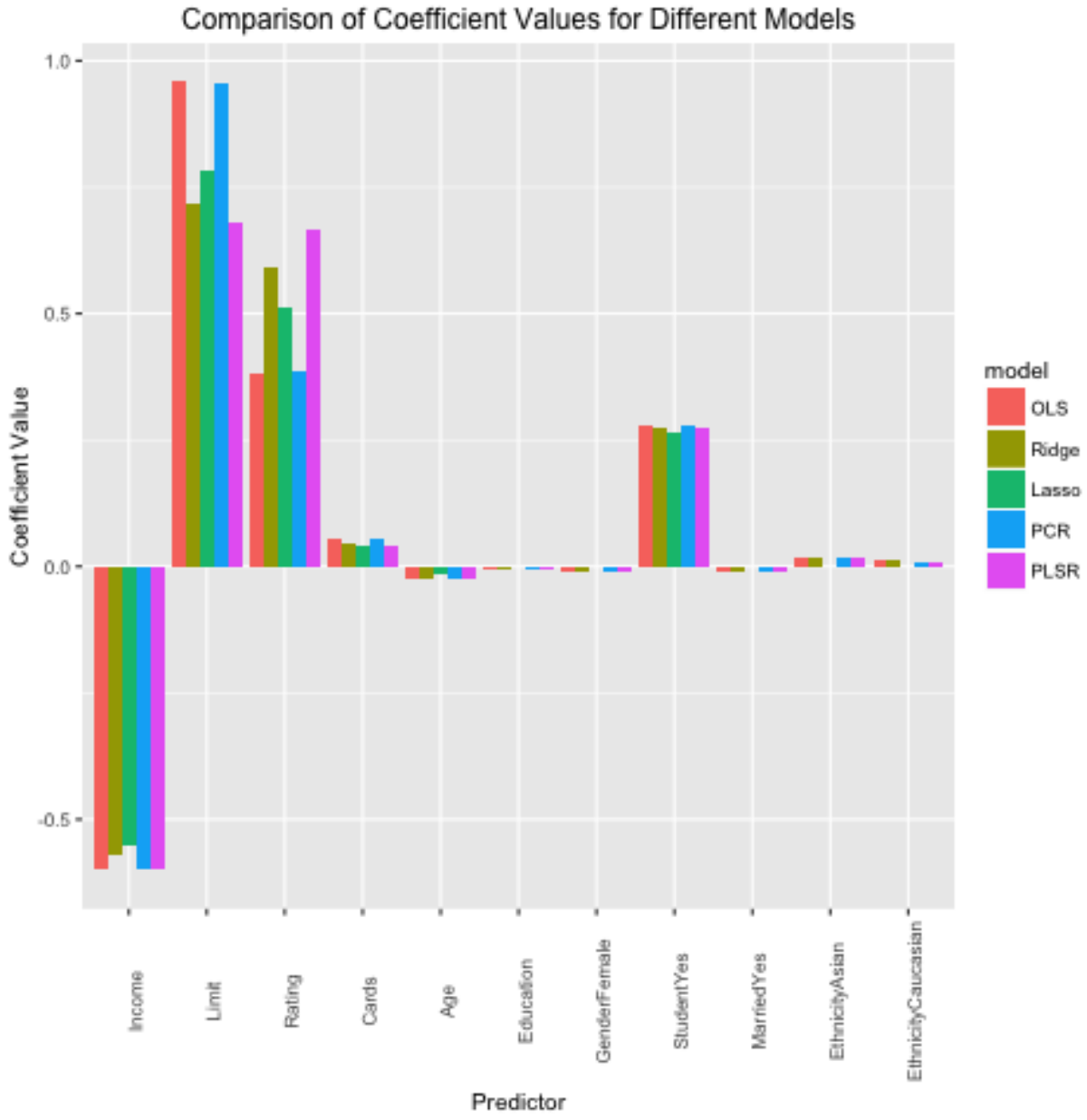


Figure 1: Coefficients Plot

Looking at a plot of these coefficients separated by model, we can more easily see for which predictors the estimated coefficients differ. The estimates for most of the predictors are very similar across the models, but for limit and rating, the estimated coefficients differ quite significantly.

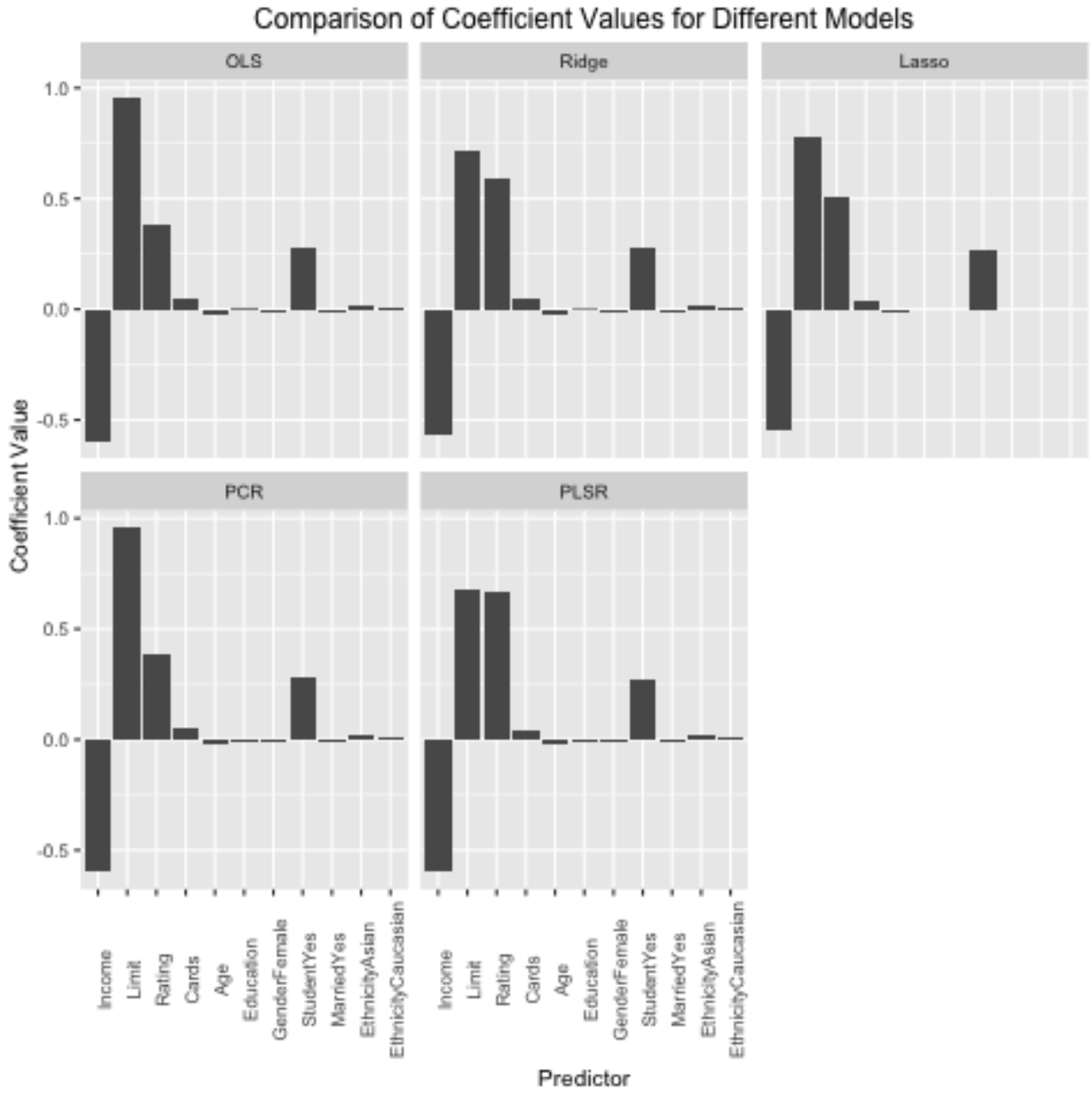


Figure 2: Coefficients Plot (separated by model)

Another plot of these coefficients is given in Figure 2, which has separate plots for each of the models.

Table 15: MSE for Different Models

	OLS	Ridge	Lasso	PCR	PLSR
MSE	0.04479	0.05103	0.04899	0.04926	0.04863

Now looking at Table 4, we can see the mean squared errors for our models, and that the lowest mean squared error for the regression alternatives (not the ordinary least squares model) is for the partial least squares regression model, with a value of 0.0486271. However in general, the ordinary least squares model still has the lowest mse of 0.0447862.

Conclusions

In this project, we have built five different regression models to predict Balance based on ten predictors from the dataset Credit.csv. Based on our results, we see that the ordinary least squares regression model yields the lowest mean squared error overall, and the partial least squares regression model yields the lowest mean squared error of the alternative regression models. Therefore we conclude that the ordinary least squares model is the most accurate predictor for Balance based on the ten predictors.