Modeling FICO Score and Loan Amount

Ashleigh Romer

Supervisor: Dr. Jebessa Mijena

Department of Mathematics

Georgia College & State University 2019

Contents

1 Abstract

In this research, we use Lending Club data from Kaggle to analyze FICO scores and loan amounts funded using multiple predictors in R. Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California. This particular data set contains information on all of the loans issued through Lending Club from 2007 to 2015. First, we will clean our big data with 1,048,575 rows and 97 columns and then perform exploratory data analysis. We will also use feature engineering and subset selection methods to build a linear model to accurately predict FICO score and loan amount funded.

2 Introduction

By modeling loan amount and FICO score, we may obtain a better understanding of what goes into predicting these values. Lending Club is a US peer to peer lending group that "has become one of the more reputable destinations for online personal loans, usually an ideal method to borrow for a special need or credit card debt consolidation. It helped to originate peer-to-peer marketplace lending, which matches borrowers with investors who are willing to fund the loans." (debt.org) FICO score is the most well known credit scoring system. The FICO score is used to help investors, lending groups and banks assess your mortgage, credit line amounts and your interest rate, and, therefore, plays a large role in people's financial decisions. Furthermore, loans are used to make many of our daily purchases, and most of our large investments, including cars, homes, college, etc. First, we will use exploratory analysis to look at correlation and skewness of the data and then use linear regression modeling to predict amount funded. We will use forward, backward, and sequential replacement linear regression modeling in R to determine our predictors and their coefficients for each model. This allows us to describe how each predictor relates to our predicted variables. Next, we will creating a training data set comprised of 70% of the data and a testing data set with the remaining 30%. Finally, we will compare the accuracy of the models using MAE, MSE, RMSE, MAPE and minimum maximum accuracy, in order to find our most useful linear model for each: amount funded and FICO Score. Our amount funded model will be based of off the forward or sequential replacement regression model, as will we see that they produce the same model, which include the predictor variables number of payments on the loan (term), loan title provided by the borrower (title), debt to income ratio (dti), number of open trades in last 6 months (open_acc_6m), balance to credit limit on all trades (all_util), ratio of total current balance to high credit/credit limit for all bankcard accounts (bc_util), total bankcard high credit/credit limit (total_bc_limit), FICO score and home ownership. Our FICO score model will be based off of the backward regression model, which includes the predictor variables LC assigned loan grade (grade), the loan title provided by the borrower (title), initial listing status of the loan, payments received to date for portion of total amount funded by investors (total_pymnt_inv), post charge off gross recovery (recoveries), months since oldest bank installment account opened (mo_sin_old_il_acct), number of mortgage accounts (mort_acc) and number of currently active bankcard accounts (num_actv_bc_tl).

3 Cleaning the Data

With such a large data set, we need to ”clean” the data, so that we may begin building our models with the most complete and accurate information. By doing this, we will be able to process the data quicker, and come up with more accurate predictions from our models based on this complete information.

To begin the data cleaning process, we remove incomplete data or NA values. We also remove predictors that have no variation within the set as these will not contribute helpful information to our models. Further, we remove certain predictors containing many levels, such as predictors with dates and job titles; note that these are not removed because we find them unimportant, but because our system cannot process this amount of information. Finally, we take the average of the high and low range of FICO scores before taking out the loan, as the difference is always 4, and also do the same for the high and low range of FICO score after taking out the loan.

After cleaning the data, we end with a data set of 60,795 rows and 92 columns.

4 Exploratory Analysis

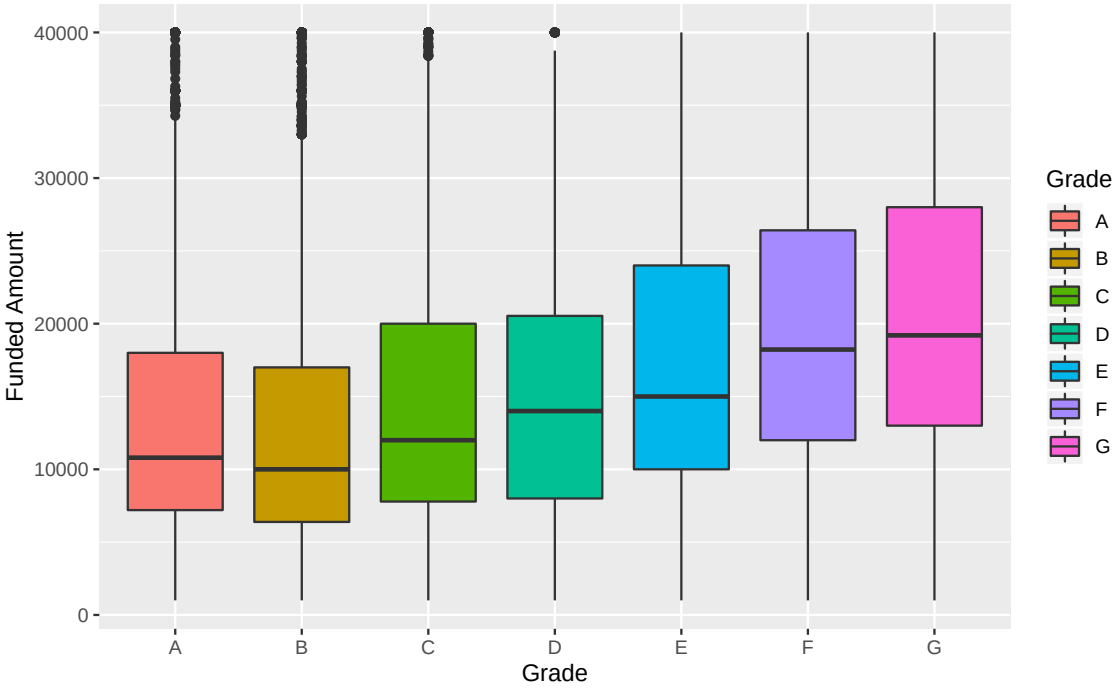

Figure 1: Boxplot of Amount Funded by Loan Grade

In Figure 1, we see that those with a lower loan grade generally receive a higher mean amount than those with a higher loan grade, with a grade of G receiving the highest mean loan amount and those with a grade of B receiving the lowest mean loan amount. It is also important to notice the outliers for those with better loan grades (A, B, and C) as they are receiving larger amounts of money.

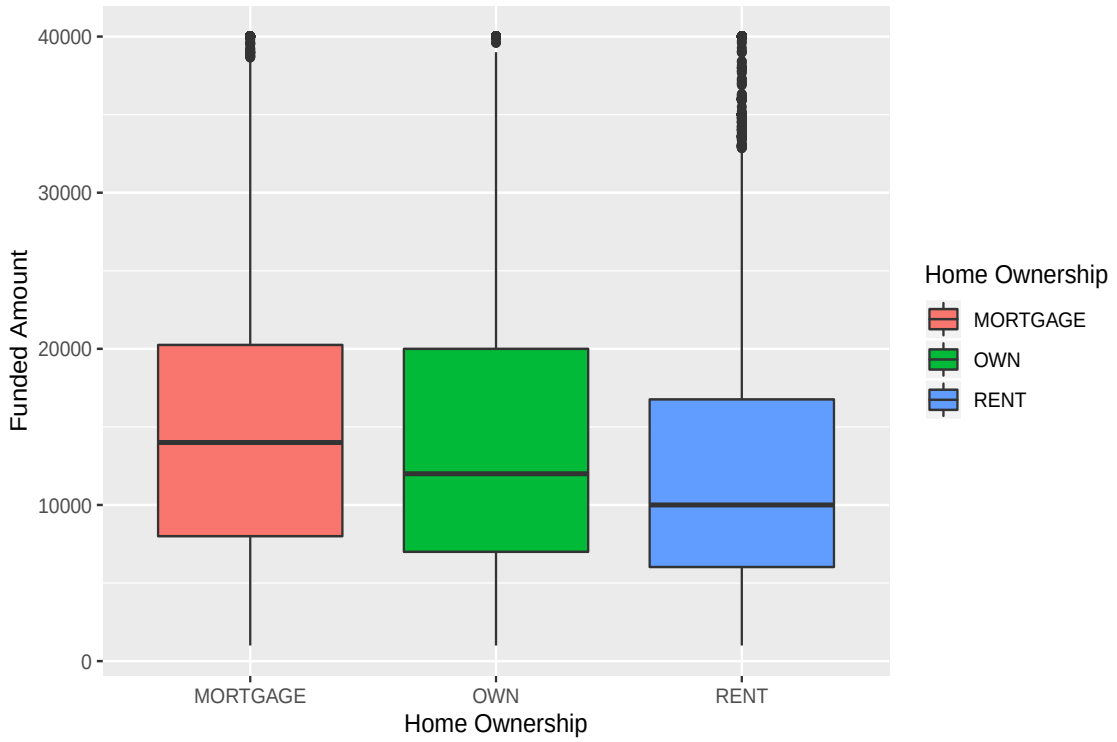

Figure 2: Boxplot of Amount Funded by Home Ownership

In Figure 2, we see that loan applicants receive a higher mean amount if they have a mortgage and receive the least mean amount if they are renting. Again, notice the outliers for those who are renting.

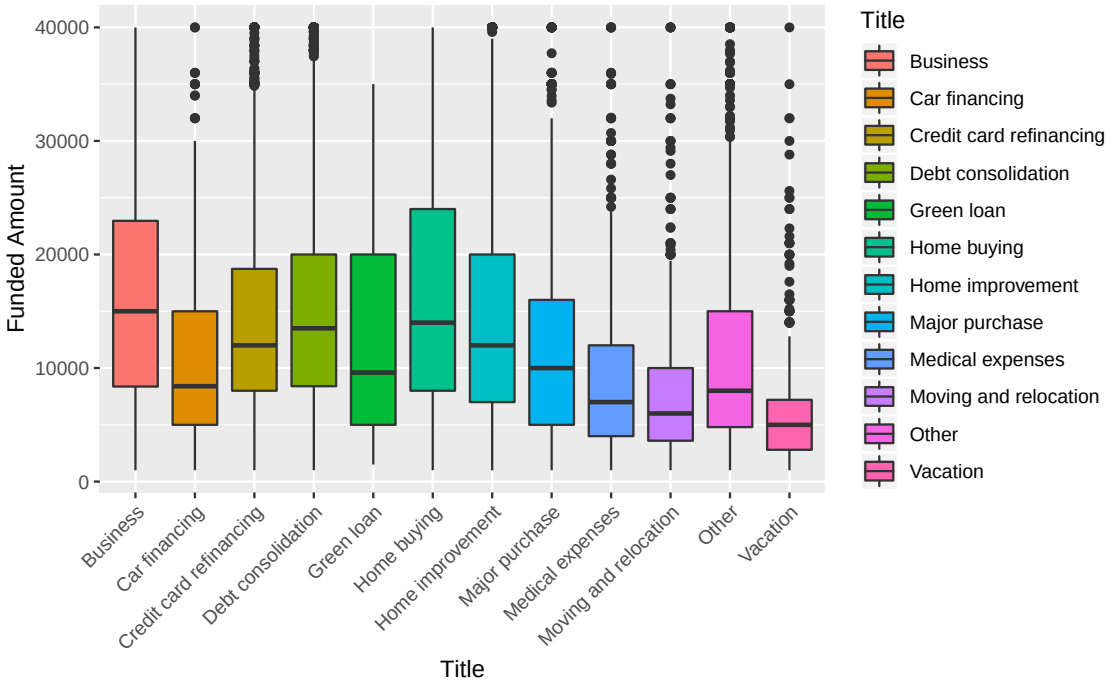

Figure 3: Boxplot of Amount Funded by Title

In Figure 3, we see how different loan purposes get funded differently. We notice that business loans and home buying loans are funded the highest amounts of money, while vacation and moving or relocation loans receive the least.

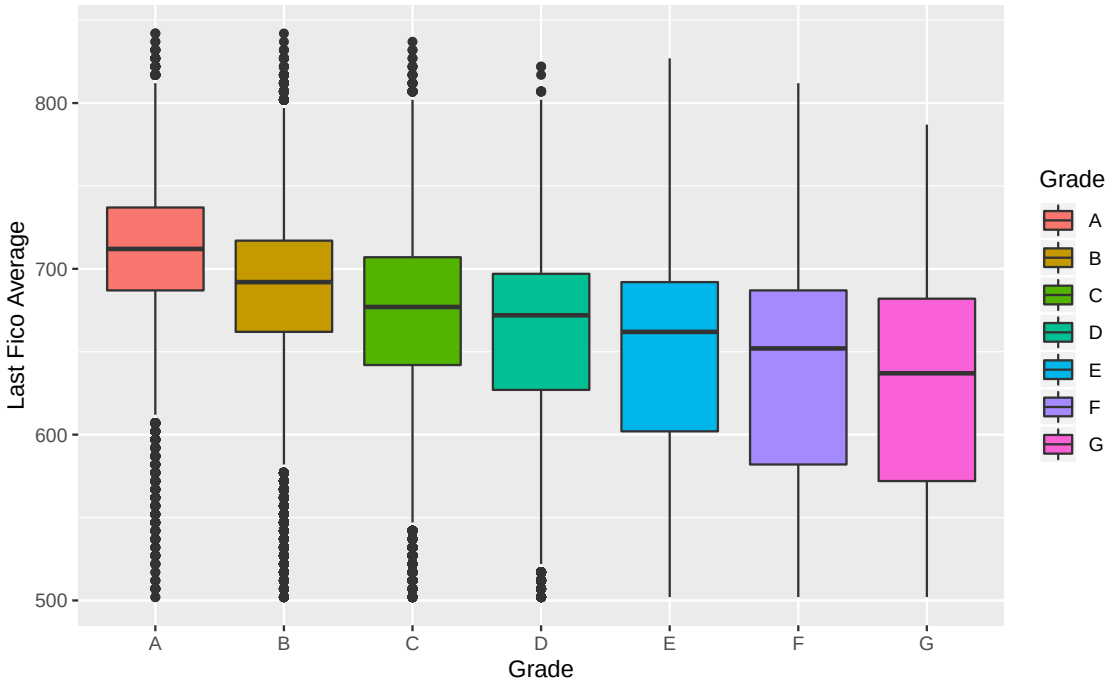

Figure 4: Boxplot of FICO Score by Loan Grade

In Figure 4, we can observe a trend in mean FICO score as it increases with a higher loan grade, but notice the outliers which suggest that FICO scores depend highly on other factors as well.

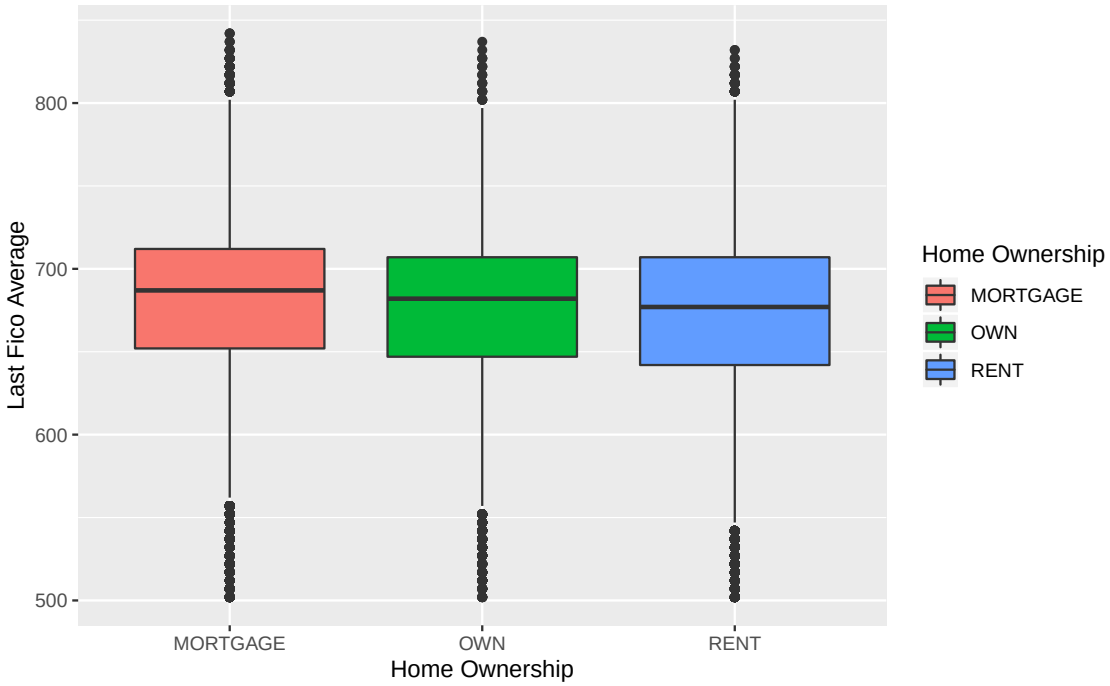

Figure 5: Boxplot of FICO Score by Home Ownership

In Figure 5, we see that the mean FICO score is only slightly higher for those with a mortgage, but again with many outliers.

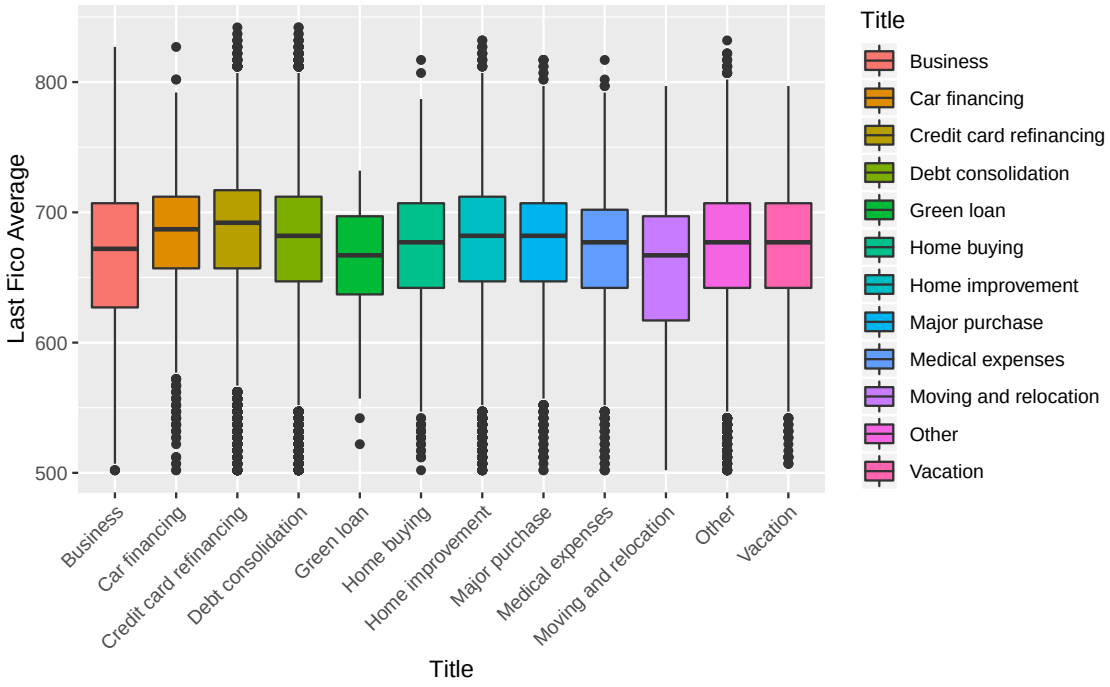

Figure 6: Boxplot of FICO Score by Title

In Figure 6, we can see that credit card refinancing and car financing loans have the highest mean FICO score association.

5 Methods

As stated earlier, we are now working with a data set containing 92 predictors, so, in order to quickly and efficiently pick the best predictors to use in our FICO score and amount funded models, we will use three subset selection methods: forward step-wise selection, backward step-wise selection, and sequential replacement, which are all different types of linear regression. Using these methods, we will also have an intercept and the coefficients for each predictor variable. Also from the exploratory analysis, we can see that the data is scewed, so we will be taking the logarithm of both FICO score and amount funded to make it a more normal distribution.

In order to understand how linear regression works, we must first understand the residual sum of squares (RSS), which is composed of the following equation:

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

where Y represents the predicted quantitative response to a change in the predictor variable X, and β_0 and β_1 are unknown constants. Now, we take

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2}$$

to be the "prediction for Y based on the i^{th} value X", to come up with

$$e_i = y_i - \hat{y}_i \tag{3}$$

which is called the i^{th} residual or "the difference between the i^{th} observed response value and the i^{th} response value that is predicted by our linear model." Based off of these equations, we now can say that

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \tag{4}$$

where n is the number of observations. Notice that you would like the difference between your predicted and observed values to be as small as possible, meaning that the smaller your RSS value, the more accurate

your model is, so we want to minimize the RSS.

Also, we must have a method to choose the number of predictor to include in our subset selection, so we now also look at the adjusted R-squared value (adjusted RSq), Mallow's Cp (Cp), and the Bayesian information criterion (BIC). Adjusted R-squared describes the amount of variation explained by the model, so we would like this value to be as close to 1 as possible. Mallows Cp is another look at the residual sum of squares that also penalizes for the amount of variables in the model. This is important because in the RSS value, you will continue to see a decrease in variation between predicted and observed values the more predictors you have, so Cp addresses this issue of over-fitting. If you were to add a predictor variable that increases the accuracy of the model more than it would by chance, the Cp value will decrease, but if we were to choose a predictor variable that does not increase the accuracy of the model more than it would by chance, our Cp value would increase; we would also like to see this value as small as possible. Finally, we will look at the BIC, which is similar to Cp in that it adds a penalty for the number of variables added to the model, thus we would also like for this value to be very small. Note that in our model, since there are so many variables, we will choose these values based on where the changes between the n^{th} and $(n-1)^{th}$ becomes relatively insignificant.

"Forward selection: We begin with the null model—a model that contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. We then add to that model the variable that results in the lowest RSS for the new two-variable model. This approach is continued until some stopping rule is satisfied.

Backward selection: We start with all variables in the model, and remove the variable with the largest p-value—that is, the variable that is the least statistically significant. The new (p-1) variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached. For instance, we may stop when all remaining variables have a p-value below some threshold.

Sequential Replacement: This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one. Note that the p-values for variables can become larger as new predictors are added to the model. Hence, if at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model. We continue to perform these forward and

backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model."

These will each give us a model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n \tag{5}$$

where Y is our predicted value, β_0 is the intercept value, β_n is the coefficient for each X_n respectively, and n is the number of variables used in the model.

After going through these subset selection methods, we now need a way to compare and choose the best of the three models we have come up with. We do this by first splitting the data up into test and training data, and then looking at five equations: mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (rMSE), mean squared error (MSE), and min. max. accuracy.

$$MAE = mean(|y_i - \hat{y}_i|) \tag{6}$$

$$MAPE = mean(\frac{|y_i - \hat{y}_i|}{y_i}) \tag{7}$$

$$rMSE = \sqrt{mean((y_i - \hat{y}_i)^2)} \tag{8}$$

$$MSE = mean((y_i - \hat{y}_i)^2) \tag{9}$$

$$MinMaxAccuracy = mean(\frac{min(y_i - \hat{y}_i)}{max(y_i - \hat{y}_i)}) = mean(1 - \frac{\epsilon_i}{y_i}) \tag{10}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value. Looking at all of these values, we will choose the model with the lowest error values (MAE, MAPE, rMSE, and MSE) and the high accuracy in predicting the test data values, which will give us our final model.

6 Results for Amount Funded Model

First, we will look at amount funded using backward step-wise selection.

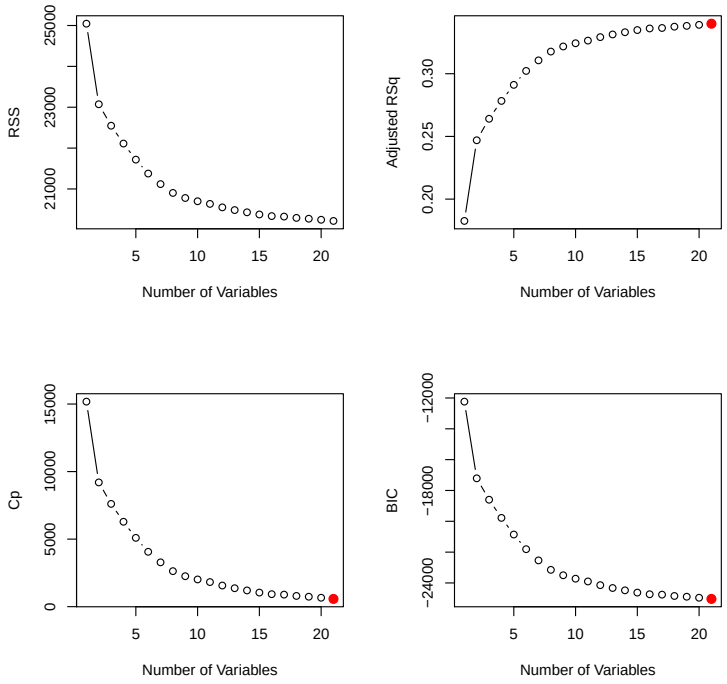

Figure 7: Backward Regression Variable Selection for Amount Funded

Looking at Figure 7, we see that the RSS, Cp, BIC, and adjusted R-squared values stop significantly changing around 15 variables, so we will perform backward step-wise selection with a 15 variable parameter, which gives us the following as our model:

log(funded amount)=term + title + debt to income ratio + number of open trades the past 6 months + balance to credit limit on all trades + ratio of total current balance to high credit/credit limit for all bankcard accounts + total bankcard high credit/credit limit + FICO score average where the coefficients are given below.

```
            (Intercept)                    term          titleCar financing titleCredit card refinancing
           6.600061e+00            2.574475e-02               -2.367260e-01                -3.330285e-02
titleDebt consolidation   titleMajor purchase   titleMedical expenses   titleMoving and relocation
           7.730515e-02           -1.935521e-01               -4.566483e-01                -5.108508e-01
             titleOther           titleVacation                         dti                  open_acc_6m
          -3.152463e-01           -7.911576e-01               -1.015012e-03                -9.053135e-03
               all_util                  bc_util           total_bc_limit                    ficoAverage
          -4.047598e-04            2.446203e-03                1.048405e-05                 2.057418e-03
```

Next, we will look at amount funded using forward step-wise selection.

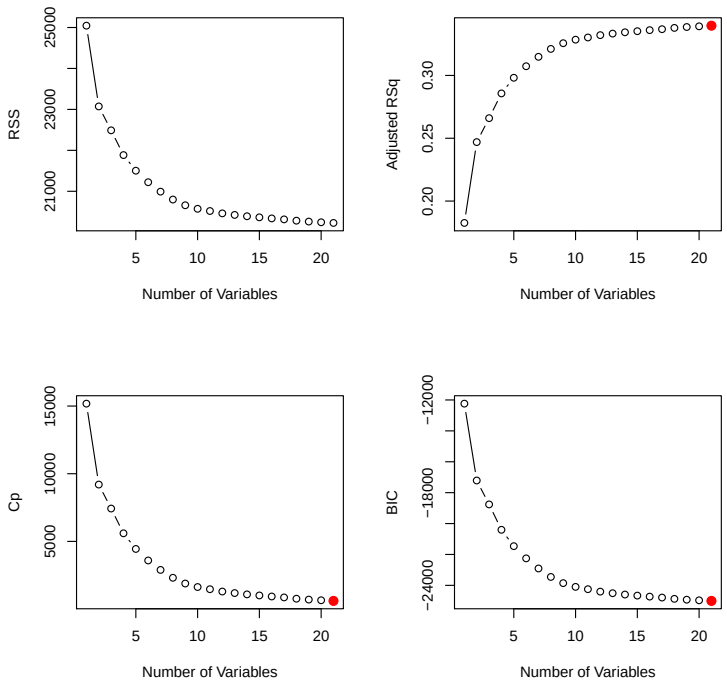

Figure 8: Forward Regression Variable Selection for Amount Funded

Looking at the Figure 8, we see that the RSS, Cp, BIC, and adjusted R-squared values stop significantly changing around 15 variables, so we will perform forward step-wise selection with a 15 variable parameter, which gives us the following as our model:

log(funded amount)=term +title +debt to income ratio +number of open trades past 6 months +balance to credit limit on all trades +ratio of total current balance to high credit/credit limit for all bankcard accounts +total bankcard high credit/credit limit +FICO score average +home ownership with the following

coefficients.

(Intercept)	term	titleCar financing	titleCredit card refinancing
6.892090e+00	2.559725e-02	-2.882126e-01	-8.586064e-02
titleGreen loan	titleMedical expenses	titleMoving and relocation	titleOther
-2.182807e-01	-5.086758e-01	-5.338856e-01	-3.637878e-01
titleVacation	dti	open_acc_6m	all_util
-8.391632e-01	-8.515882e-04	-1.036083e-02	-4.037973e-04
bc_util	total_bc_limit	ficoAverage	home_ownershipRENT
2.480267e-03	1.046518e-05	1.758956e-03	-8.279887e-02

Finally, we will look at amount funded using sequential replacement.

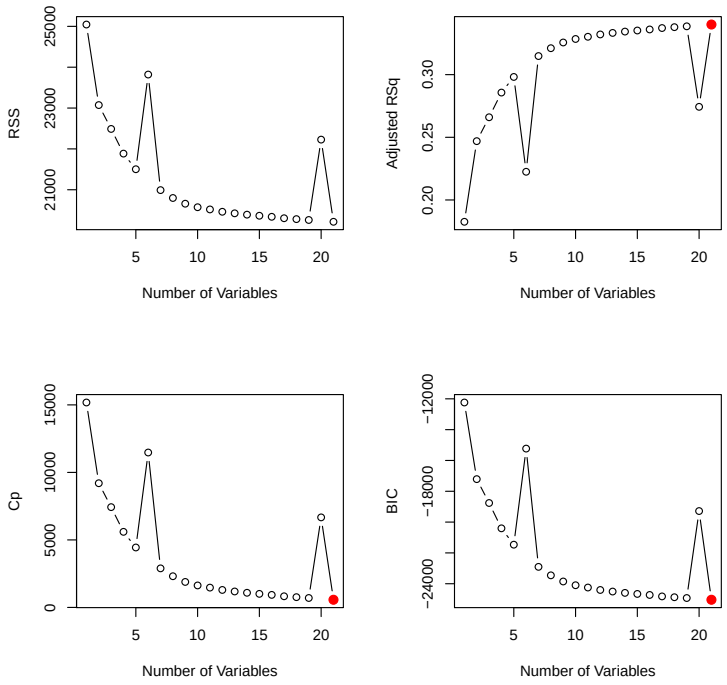

Figure 9: Sequential Replacement Variable Selection for Amount Funded

Looking at Figure 9, we see that the RSS, Cp, BIC, and adjusted R-squared values stop significantly changing around 15 variables, so we will perform sequential replacement with a 15 variable parameter, which gives us the following as our model:

log(funded amount)=term +title +debt to income ratio +number of open trades past 6 months +balance to credit limit on all trades +ratio of total current balance to high credit/credit limit for all bankcard accounts +total bankcard high credit/credit limit +FICO score average +home ownership with the following coefficients.

(Intercept)	term	titleCar financing	titleCredit card refinancing
6.892090e+00	2.559725e-02	-2.882126e-01	-8.586064e-02
titleGreen loan	titleMedical expenses	titleMoving and relocation	titleOther
-2.182807e-01	-5.086758e-01	-5.338856e-01	-3.637878e-01
titleVacation	dti	open_acc_6m	all_util
-8.391632e-01	-8.515882e-04	-1.036083e-02	-4.037973e-04
bc_util	total_bc_limit	ficoAverage	home_ownershipRENT
2.480267e-03	1.046518e-05	1.758956e-03	-8.279887e-02

Now we will choose the most accurate model of the three using the following values:

	MAE	MSE	RMSE	MAPE	Min Max Accuracy
Backward	0.46309164	0.35341458	0.59448682	0.05070347	0.9518224
Forward	0.46108202	0.35057323	0.59209225	0.05048464	0.9520226
Sequential	0.46108202	0.35057323	0.59209225	0.05048464	0.9520226

Notice that forward step-wise selection and sequential replacement both have the same formula for the model and therefore also produce the same resulting values. We find the lowest error values and the highest accuracy for the model produced by these selection methods, thus we have our funded amount model.

7 Results for FICO Score Model

First, we will look at FICO Score using backward step-wise selection.

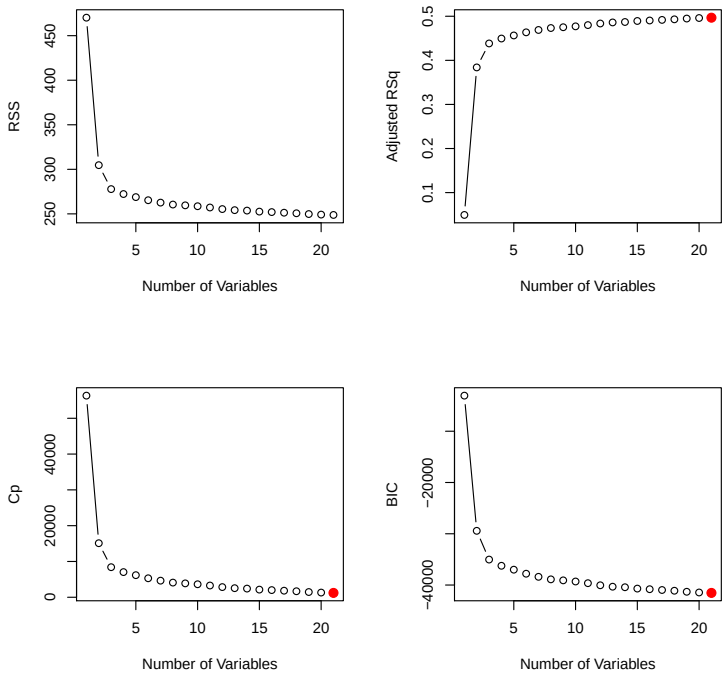

Figure 10: Backward Regression Variable Selection for FICO Score

Looking at Figure 10, we see that the RSS, Cp, BIC, and adjusted R-squared values stop significantly changing around 15 variables, so we will perform backward step-wise selection with a 15 variable parameter, which gives us the following as our model:

log(FICO Score)=grade +title +initial list status +total payment (rounded) +recoveries +months since oldest bank installment account opened +number of mortgage accounts +number of currently active bankcard accounts.

```
        (Intercept)               gradeB               gradeD               gradeE
       6.497831e+00         7.377873e-03        -2.799146e-02        -4.245504e-02
             gradeF titleDebt consolidation      titleHome buying titleHome improvement
      -5.339780e-02         -5.237449e-03        -7.089146e-03        -5.176233e-03
titleMajor purchase  titleMedical expenses initial_list_statusw       total_pymnt_inv
       2.338625e-03         -8.068342e-03         6.085741e-03         1.668501e-06
          recoveries     mo_sin_old_il_acct             mort_acc       num_actv_bc_tl
      -3.785018e-05         1.184357e-05         3.371237e-03        -1.316343e-03
```

Next, we will look at FICO Score using forward step-wise selection.

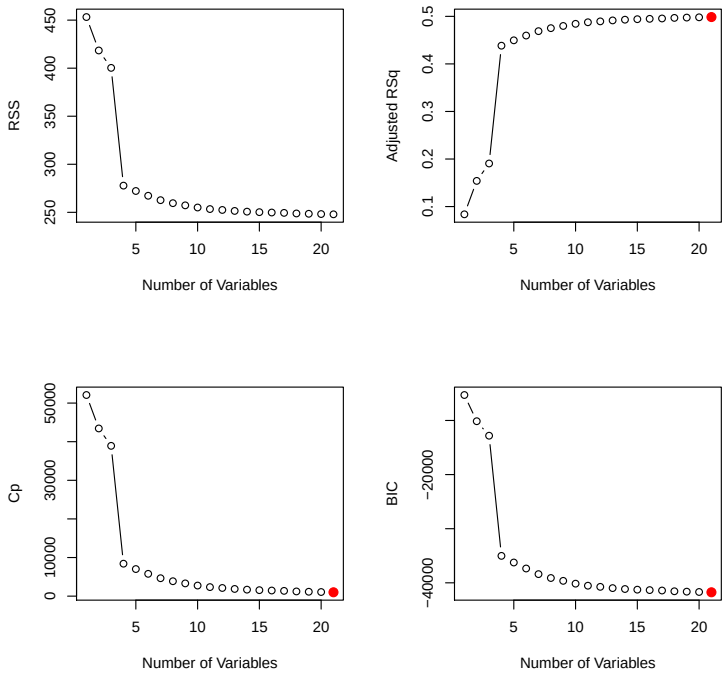

Figure 11: Forward Regression Variable Selection for FICO Score

Looking at Figure 11, we see that the RSS, Cp, BIC, and adjusted R-squared values stop significantly changing around 15 variables, so we will perform forward step-wise selection with a 15 variable parameter, which gives us the following as our model with the coefficients listed below:

log(FICO Score)=installments +title +total payments (rounded) +months since most recent 90-day or worse rating +number of installment accounts opened in past 12 months +months since most recent install-

ment accounts opened +number of mortgage accounts +months since most recent bankcard delinquency +number of currently active bankcard accounts +subgrade +FICO average.

(Intercept)	installment	titleDebt consolidation	titleHome buying
5.686287e+00	-5.014135e-05	-4.696033e-03	-1.309745e-02
titleHome improvement	titleMajor purchase	titleMedical expenses	total_pymnt_inv
-8.082996e-03	-2.623809e-03	-1.315489e-02	2.416191e-06
mths_since_last_major_derog	open_il_12m	mths_since_rcnt_il	mort_acc
1.468604e-04	-6.669001e-03	3.980872e-05	3.252191e-03
mths_since_recent_bc_dlq	num_actv_bc_tl	sub_gradeB5	ficoAverage
2.849298e-04	-1.372000e-03	5.040108e-03	1.182521e-03

Finally, we will look at FICO Score using sequential replacement.

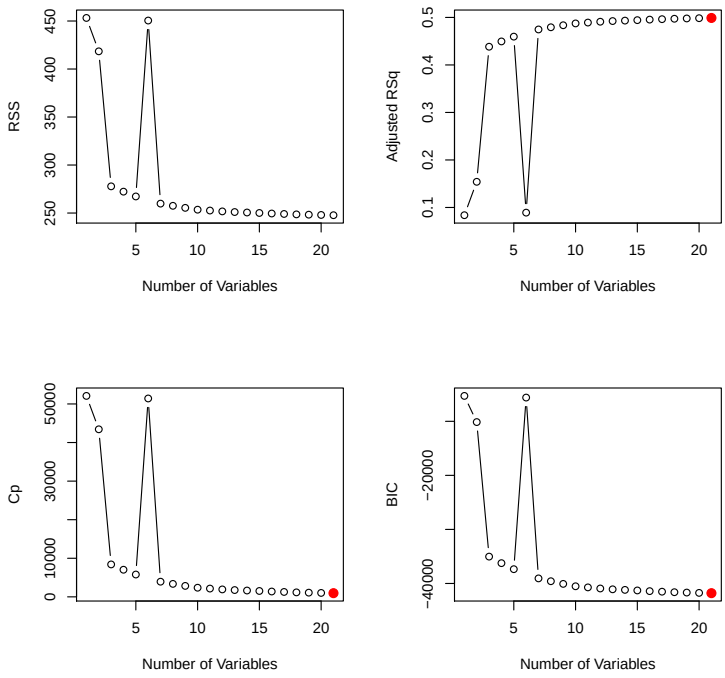

Figure 12: Sequential Replacement Variable Selection for FICO Score

Looking at Figure 12, we see that the RSS, Cp, BIC, and adjusted R-squared values stop significantly changing around 15 variables, so we will perform sequential replacement with a 10 variable parameter, which gives us the following as our model with the coefficients listed below:

log(FICO Score)=loan amount +installment +title +total payment (rounded) +number of currently active bankcard accounts +FICO average.

```
      (Intercept)                 loan_amnt                installment titleDebt consolidation
     5.687636e+00              1.128494e-06              -8.384314e-05              -5.280165e-03
titleHome buying titleHome improvement     titleMajor purchase   titleMedical expenses
    -1.484496e-02             -7.741219e-03              -4.242208e-03              -1.348795e-02
  total_pymnt_inv           num_actv_bc_tl                 ficoAverage
     2.432610e-06             -1.194629e-03               1.205691e-03
```

Now we will choose the most accurate model of the three using the following values:

	MAE	MSE	RMSE	MAPE	Min Max Accuracy
Backward	0.06106499	0.00661867	0.08135521	0.00943959	0.9906634
Forward	0.06218038	0.0067967	0.08244208	0.00961751	0.9904915
Sequential	0.06418237	0.00725348	0.08516739	0.00992685	0.9901904

Notice that the backward step-wise selection model produces the lowest error values and the highest accuracy, thus we have our FICO score model.

8 Further Research

In the future, if we were to expand upon these models, we could predict quantitative values pertaining to this data in a similar way. It would also be interesting to see, if we had a computer with higher capabilities, if using all of the variables, including those with many levels, would change our models. Also, since our analysis capabilities were limited during this research by the computer's capabilities, we were not able to look at the last type of regression which is exhaustive subset selection. In our case it would have tested 2^{92} possible models to come up with the best fit, unfortunately we could not complete that for now.

9 References

- https://www.kaggle.com/wendykan/lending-club-loan-data
- James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2017.
- Fay, Bill. “Credit Scoring: FICO, VantageScore & amp; Other Models.” Debt.org, www.debt.org/credit/report/scoring-models/.