An Analysis of the Give Me Some Credit Challenge

Jonah Calvert

Date: March 30, 2024
Department of Mathematics
University of Cincinnati

https://github.com/Jonah-Calvert/Capstone.git

Abstract

This paper delves into the quantitative assessment of borrower default risk, an important aspect of financial risk management. Utilizing the dataset from the 2011 Kaggle "Give Me Some Credit" competition, construction of predictive models that solely focus on the Probability of Default (PD) component of the Expected Loss (EL) equation is done. The analysis employs both multiple linear and logistic regression techniques in R[7] to attempt to find the most accurate prediction. Using data preparation the data is cleaned and adjusted to help build the model.
The models' performances are evaluated using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores, providing insights into their predictive accuracy. Additionally, the paper tries to solve the problem of multicollinearity among predictors, implementing variance inflation factor (VIF) analysis to improve model reliability, using a proxy variable in place of highly correlated variables to achieve a AUC score of .8359 using logistic regression with the proxy variable, placing the model in the low 700's for the competition.

1 Introduction

Credit Scoring and Rating is a method used by financial institutions around the world to analyze the reliability of individuals, and determine the amount of money that these persons can be loaned, and the particular rates money can be loaned at. While scoring and rating are very widely used in today's world, many institutions have their own proprietary methods of calculation, making it incredibly difficult for the average person to find exactly how their scores will be calculated and evaluated.

Credit rating analysis can typically be thought of by using four elements[1].

$$EL = PD \times LGD \times EAD$$

where EL(Expected Loss) is the dependent variable the financial institutions are trying to find. Typically from the eyes of the consumer, it is assumed that credit scoring is used to find the reliability of the loan Borrowers so that the financial institution can evaluate the risk assigned to each Borrowers. However, that is only one part of equation PD(Probability of Default), where as EL still comes with two other factors, LGD(Loss given Default) and EAD(Exposure at Default). This is to say rather than institutions trying to find just the reliability of the Borrowers, they are trying to find a loan configuration that maximizes the profits of their loans while taking on the least risk they can.

Using a data set from a competition ran by Kaggle in 2011 called Give Me Some Credit[4], I will be using Multiple Linear and Logistic Regression to formulate a model for evaluating risk of individual Borrowers. Because this dataset only includes Borrowers information, I will only be focusing on the PD portion of the EL equation, as LGD and EAD would depend on the specific market information of the financial institution, which is not something I can access within the competition.

2 Data

The data for this paper comes from the Give Me Some Credit competition ran by Kaggle. It is a collection of 250,000 loan applications, split into 2 parts, the training and testing set. Each application has 10 independent variables, and the testing set has an additional independent testing variable.

Focusing on the 10 independent variables the model will be built from,

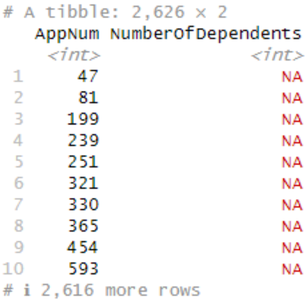
- Age.
- Number of dependents.

- Revolving utilization of unsecured lines: Amount of credit debt divided by current credit limits [Percentage].
- Number of times the Borrowers has been past due on a loan between 30-59 days within the past 2 years.
- Number of times the Borrowers has been past due on a loan between 60-89 days within the past 2 years.
- Debt ratio: Current monthly debt payments added with cost of living divided by monthly income [Percentage].
- Monthly income.
- Number of open loans.
- Number of real estate loans.

We also have an additional variable, Delinquency in 2 years, that we will then test the model against to find its efficacy.

3 Data Preparation

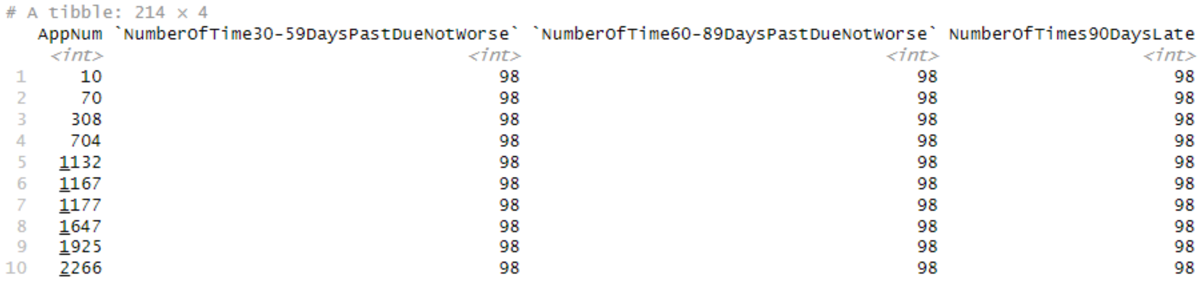
Because the data was collected by a third-party, an attempt has to be made to try to limit the effect of mistakes in the collection, as well as remove outliers in the data that would overly effect the modelling. Taking a look at each variable independently, there are some some simple mistakes. When taking a look at number of dependents and monthly income, several thousand null values appear (Figure 1). For this paper, it is assumed that those values should be 0, and were intentionally left blank by the Borrowers during the collection process.



```
# A tibble: 2,626 × 2
   AppNum NumberOfDependents
    <int>              <int>
 1     47                 NA
 2     81                 NA
 3    199                 NA
 4    239                 NA
 5    251                 NA
 6    321                 NA
 7    330                 NA
 8    365                 NA
 9    454                 NA
10    593                 NA
# ℹ 2,616 more rows
```

Figure 1: Null Values Within Dependents

We also find several instances of Borrowers with Number of times the Borrowers has been past due on a loan between 30-59 days, 60-89 days, and 90 days within the past 2 years much higher than the rest of the Borrowers (Figure 2). There is an odd pattern of them having 96 or 98 instances of a loan past due which would not be possible within 2 years. This data will be excluded from the modelling due to errors.



```
# A tibble: 214 × 4
   AppNum `NumberOfTime30-59DaysPastDueNotWorse` `NumberOfTime60-89DaysPastDueNotWorse` NumberOfTimes90DaysLate
    <int>                                  <int>                                  <int>                   <int>
 1     10                                     98                                     98                      98
 2     70                                     98                                     98                      98
 3    308                                     98                                     98                      98
 4    704                                     98                                     98                      98
 5   1132                                     98                                     98                      98
 6   1167                                     98                                     98                      98
 7   1177                                     98                                     98                      98
 8   1647                                     98                                     98                      98
 9   1925                                     98                                     98                      98
10   2266                                     98                                     98                      98
```

Figure 2: Errors within Data

There is also some troubling behavior in the Revolving Utilization of Credit as seen in Figure 3.

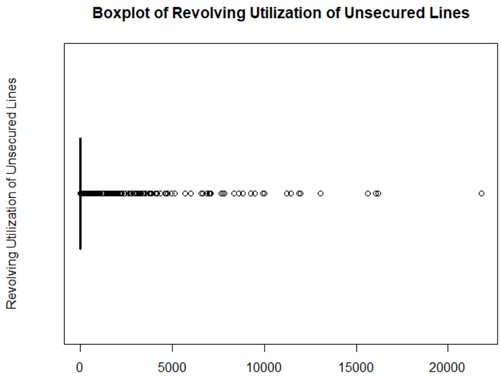


Figure 3: Errors within Data

Taking a closer look, around 2,000 people have a revolving utilization of credit higher than 1. This means that their current usage of credit cards is higher than their current credit card limits, which should not be possible. Because these outliers are so large, I will be limiting this variable to 1, and removing the data with higher than 1.

4 Linear Regression

Simple linear regression was an idea first thought upon by Sir Francis Dalton, and later his biographer and fellow mathematician Karl Pearson[8]. Dalton first came across the idea when studying the offspring of pea plants. Although not fully conceptualized by Dalton, an eventual form of the regression would be

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where Y_i is the dependent variable, or the value you are trying to predict with a model, β_0 is the constant/intercept of the model, or the value of the dependent variable when all predictor variables are 0. β_1 is the predictor coefficient, while X is called the independent variable, that when multiplied by β_1 gives you the contribution of the independent variable X to the predicted value of Y_i. Lastly, ϵ stands in for the error/residual of the model, or the discrepancy between your observed and predicted values.

In Dalton's pea plant model, Y_I (the predicted value) would have been the size of the pea plant trying to be predicted. X_i would stand for the predictor variable, the size of the seed, and when multiplied by the coefficient β_1 would give the influence of the seed size on the pea plant size.

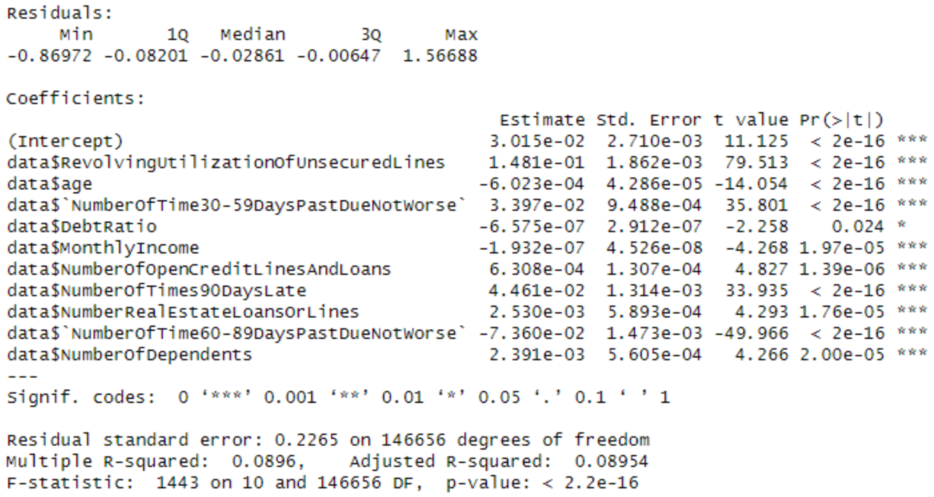
Because the model uses several predictors (Our variables from before) simple linear regression cannot be used, and instead need to use multiple linear regression, an extension of simple linear regression. In multiple linear regression there are familiar components, Y being the predicted value, β_1 being the coefficient for the dependent variable X_1. However because the model has several predictor variables, each is given their own corresponding coefficient $/beta_i$. This gives the model,

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \epsilon_i$$

where $\beta_i X_i$ is the ith predictor coefficient and variable in the model. Each $\beta_i X_i$ combination will correspond to one of the predictor variables mentioned in section 2. This leaves us with the complete model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{10} X_{10} + \epsilon_i$$

However, although this model includes every predictor variable, some of them may not be relevant to the final predicted value. This is where predictor selection comes in to help reduce the complexity of the model, and possibly make it a better fit for the data. Using R, I input the model using all 10 predictors to get a summary of the model.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.86972 -0.08201 -0.02861 -0.00647  1.56688 

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                 3.015e-02  2.710e-03  11.125  < 2e-16 ***
data$RevolvingUtilizationOfUnsecuredLines   1.481e-01  1.862e-03  79.513  < 2e-16 ***
data$age                                   -6.023e-04  4.286e-05 -14.054  < 2e-16 ***
data$`NumberOfTime30-59DaysPastDueNotWorse`  3.397e-02  9.488e-04  35.801  < 2e-16 ***
data$DebtRatio                             -6.575e-07  2.912e-07  -2.258    0.024 *  
data$MonthlyIncome                         -1.932e-07  4.526e-08  -4.268 1.97e-05 ***
data$NumberOfOpenCreditLinesAndLoans        6.308e-04  1.307e-04   4.827 1.39e-06 ***
data$NumberOfTimes90DaysLate                4.461e-02  1.314e-03  33.935  < 2e-16 ***
data$NumberRealEstateLoansOrLines           2.530e-03  5.893e-04   4.293 1.76e-05 ***
data$`NumberOfTime60-89DaysPastDueNotWorse` -7.360e-02  1.473e-03 -49.966  < 2e-16 ***
data$NumberOfDependents                     2.391e-03  5.605e-04   4.266 2.00e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2265 on 146656 degrees of freedom
Multiple R-squared:  0.0896,	Adjusted R-squared:  0.08954 
F-statistic:  1443 on 10 and 146656 DF,  p-value: < 2.2e-16
```

Figure 4: Global Regression model

As seen in Figure 4, all of the predictor variables are significant using an $\alpha = .05$, signified by at least 1 * next to each predictor. However, the R-squared value, or the proportion of variance explained by the predictors, is fairly low. R-squared, said to be the worth of the model [5], is a range of [0,1] where a higher number typically results in a better model. The R-squared of the complete model is .0896, meaning around 9% of the variance of the model is explained by the predictors. While normally this would indicate a problem in a model, it is expected the model would have a low r-squared value, as credit scoring is an incredibly complex topic with a large amount of predictors, including human behavior which cannot be properly explained in a regression model. The data includes 10 of the possible 100's of predictors that could go into the model, meaning it won't correlate perfectly with the data.

In figure 4 the estimates for the predictor variables are very small. This is because the predicted variable Serious Delinquency in 2 years is a binary variable, either 0 or 1, so theoretically once we multiply all predictor coefficients by their variables and add them together, we should get a value somewhere between 0 and 1, and we can make a determination later on how to analyze those values.

4.1 Predictor Analysis

When building a model, simplicity is usually a benefit to models for a variety of reasons. It makes it easier to interpret, better for future predictions, and the model becomes more robust. If there are too many predictors for a too small sample size, we have over-fitting[5] which can cause the model to make inaccurate predictions on new data. I did some initial testing using AICc curves and RCME to try to find a better fit model for predicting new data. However AICc scores suggested I use models based on one or two predictors, while the RCME scores were much closer, however due to the nature of them, a higher number of predictors always resulted in a higher RCME.

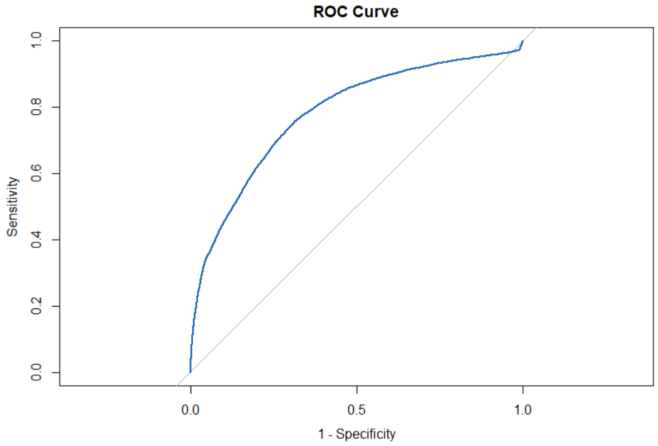
4.2 ROC and AUC

When trying to find the efficacy of a model, we can use an ROC(receiver operating characteristic) curve. The origins of ROC come from radio operators during WWII detecting enemies in combat. This is a curve of True Positive Rate(TPR) and False Positive Rate(FPR) given by [6]:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

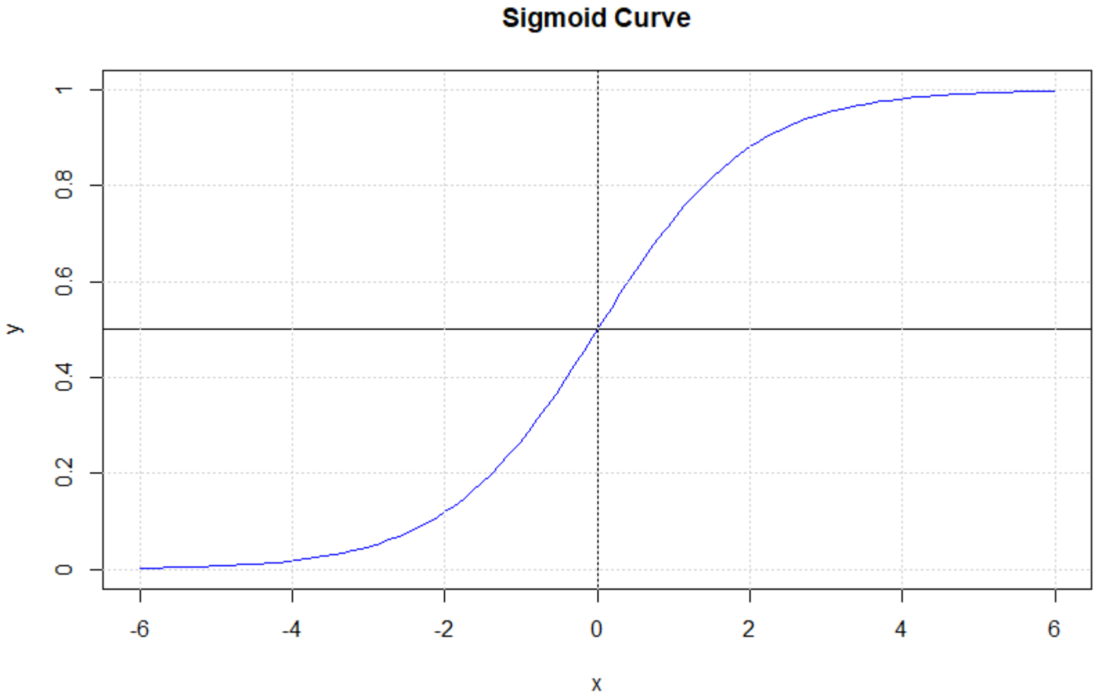
Where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives and TN is the number of true negatives. Then, we plot these together:



Then, we do an Area under the curve(AUC) analysis on the ROC, finding the area under the ROC. AUC is given by a value between 0 and 1, where .5 is said to be equivalent to guessing the outcomes, anything below .5 is worse than guessing and would be better by simply doing the opposite of what was done, and the higher than .5 the better the model at predicting the training data. The complete model in figure 4 gives an AUC score of .7789, meaning the model got 78% of the predictions correct. This places the model in the high 700's place of the Kaggle competition.

5 Logistic Regression

Our prediction variable is a binary response variable. However, linear regression gives a continuous outcome. We then have to find some threshold for the ROC curve to correctly predict whether the Borrowers defaulted on their loan. Logistic Regression is a modelling technique that can solve this problem by creating a Sigmoid curve instead of a regression line:



$$y = \frac{e^x}{e^x + 1}$$

which then allows us a better fitting regression.

Logistic regression has been a technique that has been around since the 1800's where several mathematicians used it for various problems. First by mathematician Pierre Verhulst, then popularized separately by Raymond Pearl and Lowell Reed in the 1920's[2]. It can be used when the outcome variable is binary, calculating the probability of an event occurring, rather than a continuous outcome. The generalized model comes from using a linear function:

$$Y = \beta_0 + \beta_1 X$$

and substituting it into the logistic function:

$$y = \frac{e^x}{e^x + 1}$$

to get:

$$p = \frac{e^{\beta_0 + \beta_1 X}}{e^{\beta_0 + \beta_1 X} + 1}$$

rearranging this function gives us:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X + \epsilon$$

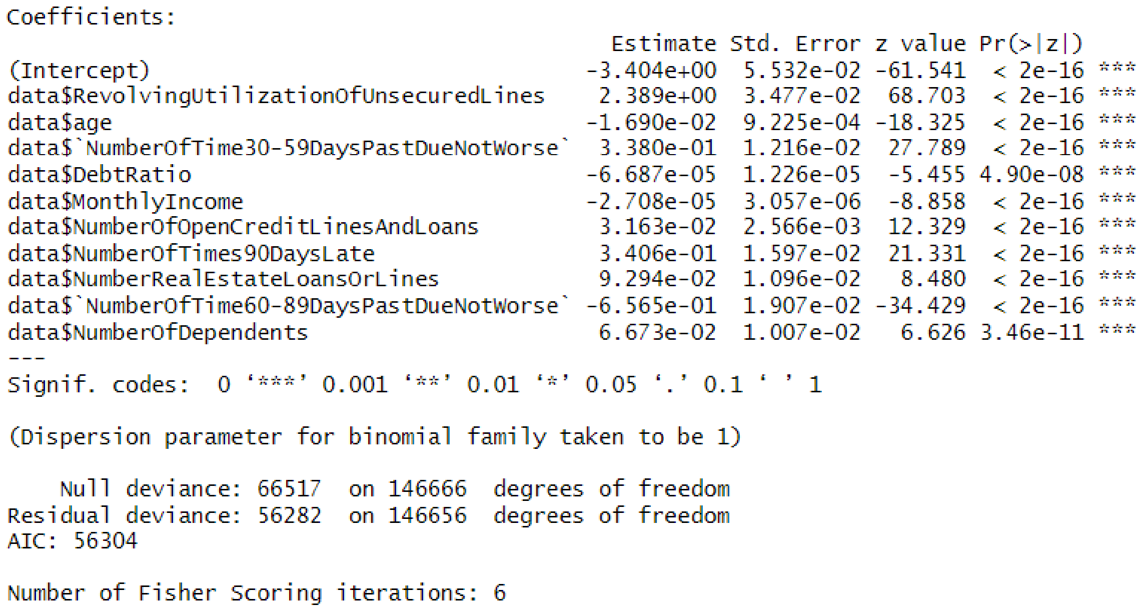
This equation is called the logit equation(coined by Joseph Berkson in 1944 [2]) where p is probability of the outcome being 0 or 1, with the remaining variables being the same as simple linear regression. This can easily be revised to fit the multiple predictor variables by inputting the equation,

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{10} X_{10} + \epsilon_i$$

into the logit equation,

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{10} X_{10} + \epsilon_i$$

giving us a usable logistic model for predicting defaults in loans. Similar to linear regression, this equation is inputted into R to find a complete model to analyze.

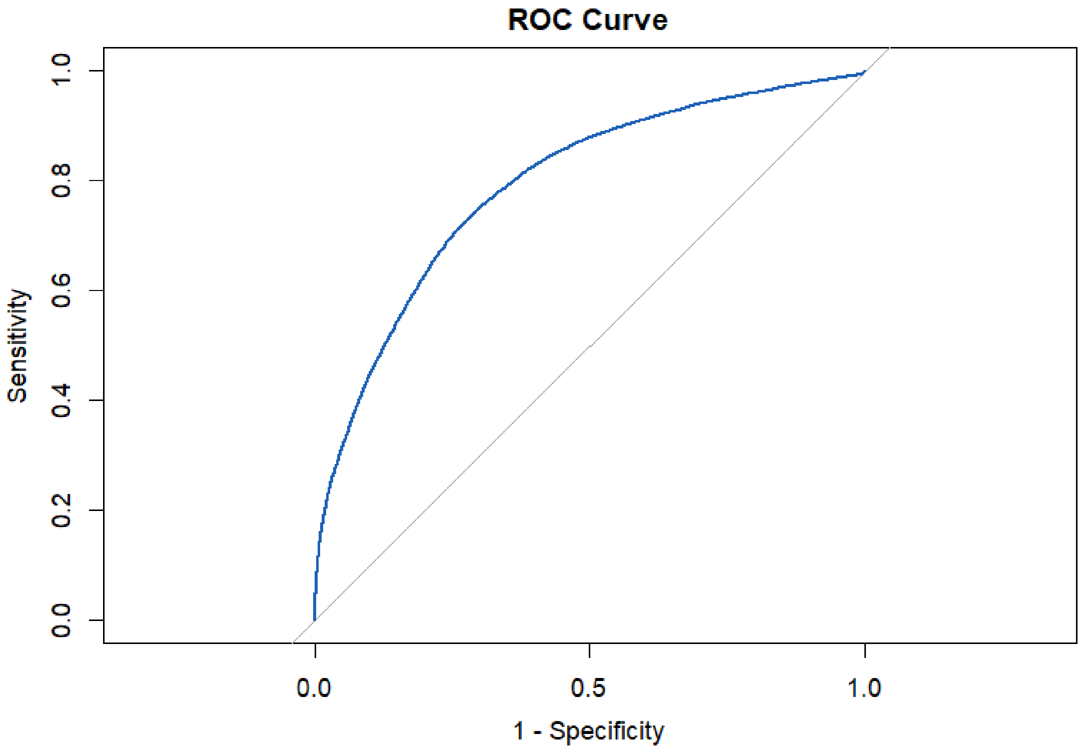


```
Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                 -3.404e+00  5.532e-02 -61.541  < 2e-16 ***
data$RevolvingUtilizationOfUnsecuredLines    2.389e+00  3.477e-02  68.703  < 2e-16 ***
data$age                                    -1.690e-02  9.225e-04 -18.325  < 2e-16 ***
data$`NumberOfTime30-59DaysPastDueNotWorse`  3.380e-01  1.216e-02  27.789  < 2e-16 ***
data$DebtRatio                              -6.687e-05  1.226e-05  -5.455 4.90e-08 ***
data$MonthlyIncome                          -2.708e-05  3.057e-06  -8.858  < 2e-16 ***
data$NumberOfOpenCreditLinesAndLoans         3.163e-02  2.566e-03  12.329  < 2e-16 ***
data$NumberOfTimes90DaysLate                 3.406e-01  1.597e-02  21.331  < 2e-16 ***
data$NumberRealEstateLoansOrLines            9.294e-02  1.096e-02   8.480  < 2e-16 ***
data$`NumberOfTime60-89DaysPastDueNotWorse` -6.565e-01  1.907e-02 -34.429  < 2e-16 ***
data$NumberOfDependents                      6.673e-02  1.007e-02   6.626 3.46e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 66517  on 146666  degrees of freedom
Residual deviance: 56282  on 146656  degrees of freedom
AIC: 56304

Number of Fisher Scoring iterations: 6
```

Each predictor has a very high significance which means they all contribute to the response variable. However, looking towards the bottom the AIC is very high. Unfortunately, it cannot be looked at in a vacuum and must be compared with several other logistic models with less predictors to give us an accurate representation of the model. To understand the estimates of each predictor coefficient, a positive value means the larger the predictor variable gets, the more likely the borrower is to default on their loan. If the value is negative, the higher the predictor variable the less likely a borrow is to default. For instance, according to our model, The older you are the less likely you are to default. While the more real-estate loans you take on, you are more likely to default. It is important we pay attention to the magnitudes of the estimates, as something like Age has a much higher magnitude than Debt-Ratio, meaning it has a larger impact on our response variable.

5.1 ROC and AUC

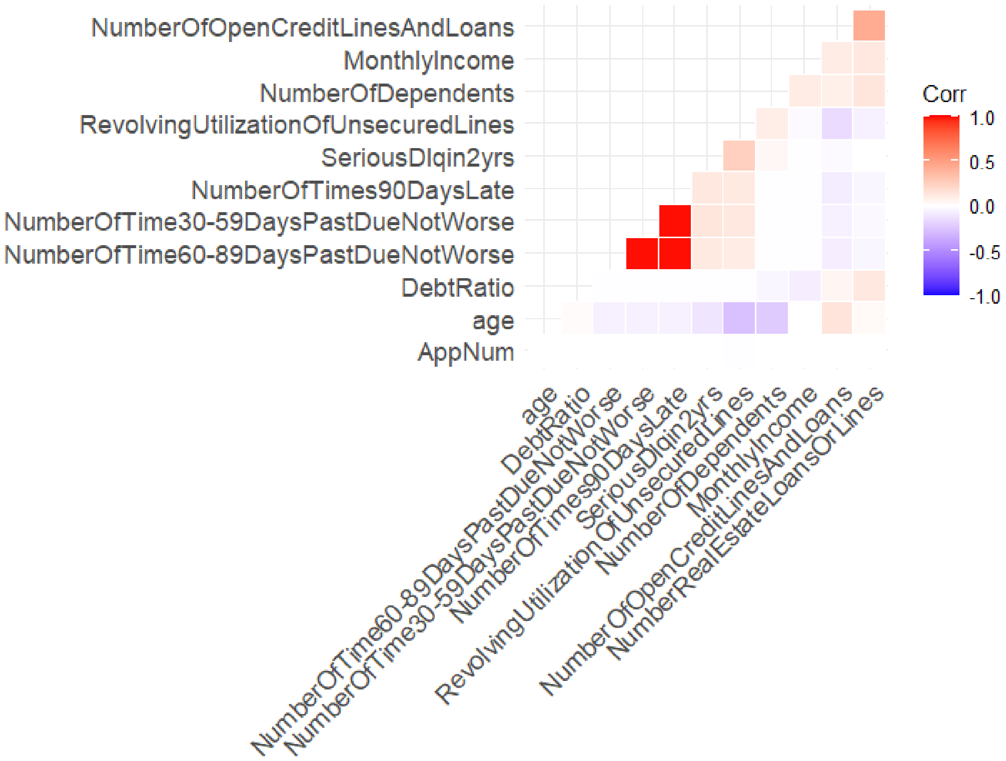
Once again we will get an ROC curve and AUC score from our complete model.



We are given an AUC score of 0.7893 meaning the model got 79% of the predictions correct. This is not a significant difference from the linear regression, however it is higher, meaning it is considered better in the Kaggle competition.

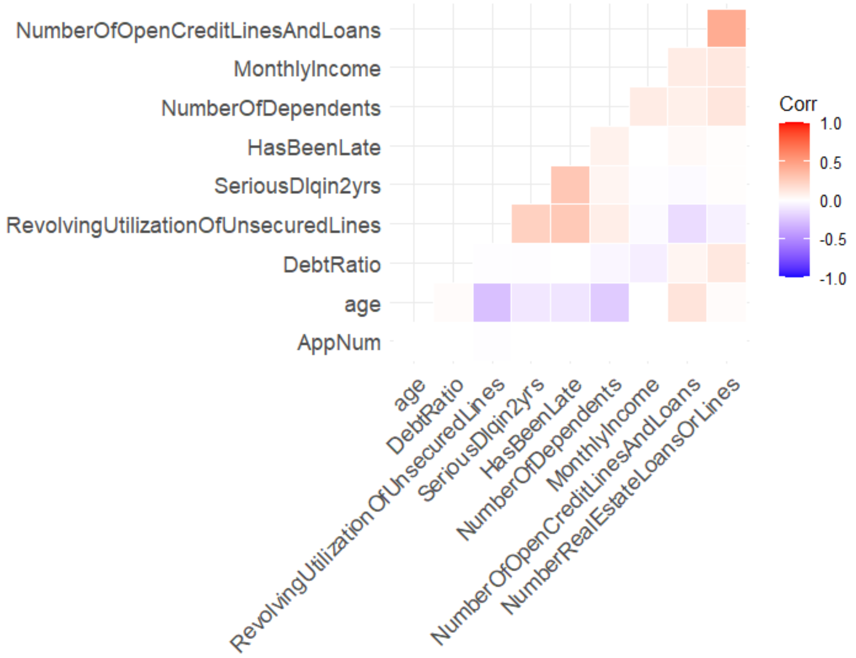
6 Multicollinearity

Multicollinearity in a set of predictors happens when two or more predictors in a regression model have a high correlation to each other. Using the car[3] library in R we are able to find a VIF(Variance Inflation Factor) score. A higher score means a larger correlation between variables. I can then turn the VIF into a correlation matrix.



the correlation matrix shows 3 problem predictors, including all of the number of times late within certain number of days. Looking at their VIF scores which paint a picture of how correlated, they are 87, 213, 150 respectively. A VIF of less than 5 is usually okay meaning that we have a very high level of multicollinearity.

To remedy this all three predictors are dropped from the models and replaced them with one boolean variable of whether or not they had been late on a payment. This changes the correlation matrix to



which has much better VIF scores, fixing our multicollinearity problem.

7 Conclusion

Using the Kaggle competition rules, the higher AUC score a model has, the better it places. This leads to models that very specifically look for high AUC scores, but often leave out crucial parts of analysis. When looking at the logistic growth model, the AIC scores were very low, but fixing them using model selection reduces the AUC score. It is a limitation of the competition to allow for ease of scoring. Using this AUC score, we find the logistic modeling with a proxy variable filling in for number of times payment has been late wins over the other models with an AUC score of .8359 placing around 700th place in the competition. Although this is a fairly low placing, it is competing with machine learning models and likely cannot get much higher without significant changes.

References

[1] Christine Bolton. *Logistic regression and its application in credit scoring.* University of Pretoria (South Africa), 2009.

[2] Jan Salomon Cramer. “The origins of logistic regression”. In: (2002).

[3] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage, 2019. URL: `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

[4] *Give Me Some Credit*. `https://www.kaggle.com/competitions/GiveMeSomeCredit/overview`. Accessed: 2010-09-30.

[5] Frank Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Jan. 2015. ISBN: 978-3-319-19424-0. DOI: `10.1007/978-3-319-19425-7`.

[6] Charles E Metz. “Basic principles of ROC analysis”. In: *Seminars in nuclear medicine*. Vol. 8. 4. Elsevier. 1978, pp. 283–298.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: `https://www.R-project.org/`.

[8] Jeffrey M. Stanton. “Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors”. In: *Journal of Statistics Education* 9.3 (2001). DOI: `10.1080/10691898.2001.11910537`.