

MAT 355 001 SPRING 2021: Generalized Linear Models and Predictive Modeling

Final Report

Credit Risk Analysis

Research By: Gavin Bougadis

3/29/2021

Instructor: Maochao Xu

**Abstract**

For this project, the assignment was to find a data source in which you could identify a statistical modeling problem and solve it using modern regression techniques. This required us to come up with various potential questions, create or test various models in R that were covered in our lectures, and report on our findings and report them in an easy-to-understand manner for a general audience.

For this project, I utilized a popular website called Kaggle to retrieve my dataset. In this case, I decided to look at data uploaded by a user that contains credit bureau data that contains borrower information such as person income, loan intent, loan status (whether a person has defaulted or not), and credit history length. For the most part the data was fairly clean with an overall usability rating of 7.1 on Kaggle’s rating system. My goal when I set out on this project was to identify different variables that may impact whether a borrower may default on a loan as well as attempt to predict whether one may default in the future.

**Introduction**

* Background information
  + According to Corporate Finance Institute, “Credit risk arises when corporate or individual borrower fails to meet their debt obligations. It is the probability that the lender will not receive the principal and interest payments of a debt required to service the debt extended to a borrower…”. More background information needed to understand this research project are the various terms used to identify a borrower. You need to know this in order to understand the factors play an important role in predicting whether a borrower will default on a loan or not. The following variables are person\_age, which is defined as the age of customer. Person\_income is the annual income the borrower makes, person\_home\_ownership is the type of home ownership customer has. This would include ownership such as if they are owning, renting, or having a mortgage. Person\_emp\_length is the employment length of customer in years. Loan\_intent is the reason/intent the customer has taken out the loan to use. Loan\_grade is the classification system that assigns quality of a borrower’s credit based on their past history among other things (A being best, G being worst). Loan\_amnt is the amount the credit bureau has granted the customer. Loan\_int\_rate is the interest rate the customer is paying back on. And lastly, loan\_status represents that status of the current loan, with 0 representing non-default and 1 representing default.
* Research Motivation
  + I chose this project idea because MAT 355 is my only, and final class before I graduate in May and as I had begun to apply to more jobs, I noticed many of the analytical positions were looking for proficiency in R. Not only R, but also, more specifically financial modeling. When I researched financial modeling, I came across credit risk modeling and thought that it would be an interesting project to take on. Additionally, at the time of choosing my topic I was in the middle of a couple of interview processes where the company model dealt specifically with loan services.
* Research Objectives
  + Identify any significant variables that may help in understanding and predicting a borrower’s ability to repay a loan.
  + Create generalized linear models that are able to predict whether a customer will default on a loan.
  + Generate visual analysis that will help the reader understand the biggest impacts and best customers to give loans out to.
* Research Questions and Approach

**Data Preparation**

When you look at the summary of credit, which is the title of the original, uncleaned data, you can see that there are 895 NA values for person employment length and 3116 NA values for loan interest rate. There was also a total of 32,579 observations which we cleaned into 31,094 observations with 14 variables. Additionally, by using the str() method in R we were able to see what types of variables we would be dealing with. One change I wanted to make for this was the variable titled cb\_person\_default\_on\_file was distinguished by a “Y” or “N”. To make this easier to work with I changed this to a binary factor variable where 1 represents yes, and 0 represents no. To fix the 3,116 entries without an interest rate entered, I decided I would create dummy categorical variables (coarse classification) for variables customer\_emp\_length and loan\_int\_rate. For interest rate, I decided to split the entries into 5 groups, 0-6, 6-12, 12-18, 18+, and missing. I did this to discretize the data. Additionally, to deal with NA values, I replaced the NA values with them with the median interest rate. I used the same concept for employment length to discretize the data by splitting the data into 8 groups, split up by a range of five for each group. To deal with NA values, one of the bins was titled “Missing”. After doing this we were able to proceed with our renamed as credit\_new in R.

It is important to note that we did not use any of the new categorical variables (employment\_cat, cat\_int\_rate) we created outside of exploratory analysis.

**Exploratory Data Analysis**

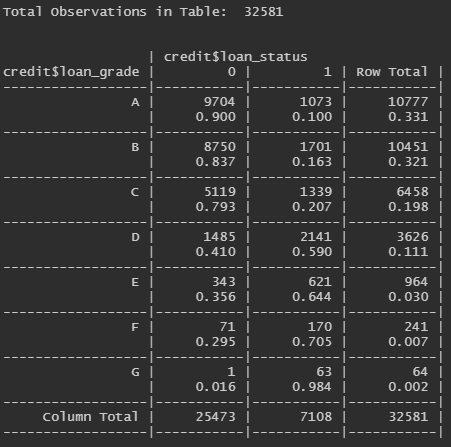
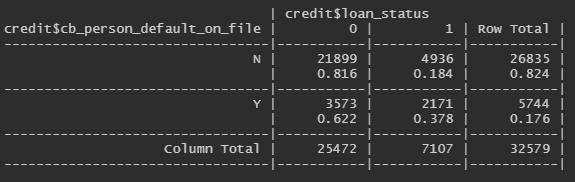
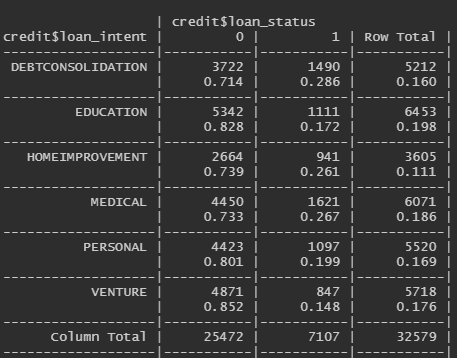
This histogram (Figure 1) gives an idea of the annual income our borrowers have when they go to receive a loan. By looking at annual income, initial thoughts might indicate that someone with a higher income may be safer to give a loan to depending on the proportional size of their loan. Here, it appears the highest frequency is roughly 60,000. This would match similarly with our summary of a person’s income which states the mean is $58,704.

Next, we looked at the histogram (Figure 2) that examines the frequency of loan amounts for customers in the data set. We see the frequency is heavily concentrated around the ranges of $500 to $15,000. The highest frequency is a little less than $10,000 which also adds up with the summary of the loan amount data, which states the mean is $9,288.

Following that, we used a cross table to look deeper into the loan grade of a borrowers in this data set I used a cross table which generates table proportions of each loan grade with respect to their loan status. This is great for categorical variables as it automatically analyzes the proportions of each categorical variable with respect to the total number of observations or another variable (in this case, binary). In this case, the main categorical variables I wanted a better look at and understanding of were loan grade, loan intent, and default on file.

For loan grade, loan intent, and previous default history I decided to compare these with loan status for a quick analysis on hypotheses that might be significant contributing factors when it comes to defaulting on a loan. Upon examination, one can see that when looking at loan grade in proportion to loan status, the proportion of defaults increases as the loan grade moves from grade A (best possible) to G (worst possible).

Additionally, because the focus of my study is on the probability and analysis of a borrower defaulting on a loan, I also compared the loan intent with loan status. Here, it is easy to see that those who take out loans for debt consolidation, home improvement and medical all have the highest default rate with a range of 0.261-0.286. The lowest loan default rate is with those who take out loans for venture purposes, with a rate of 0.148.

The last cross table was the proportion of those who had a default on file from the past compared with to their current loan status. Those who did not have a default on file prior to their current loan had defaulted on the current loan at a rate of 18.4%, while the proportion of those who had a default on file prior to this loan was 37.8%. Here, it is quite clear that this is something I would want to keep in mind throughout my analysis, as it appears on the surface that there is a significant correlation between whether a person has defaulted in the past and their current loan status.

One more type of visual analysis I wanted to run to get a better understanding of the data was a box plot (Figure 3). Here, I ran a box plot on the loan grade, comparing it with the loan amount. What I found was interesting. Based on the box plot, one can see that the median is roughly $7,500 for group A, whereas the median loan amount for group G is a little less than $20,000 and it appears that from group A to group G, the median tends to trend up for each group. However, there an incredibly large number of outliers in the groups A and C as well as plethora of outliers in B, whereas there little to no outliers in groups E, F, G and the box plot extends throughout the entire range of the loan amount values. This would lead us to conclude that the loan amount given to group A tends to be much more similar for each borrower as opposed to someone in group G where the tends to be a wide variety of the loan amount.

Now for basic linear models, for this data set we used generalized linear models for every model. Here (Figure 4), we used the factor variable of loan grade, with 1 representing A and 7 representing G, we can say that as the loan grade increases by 1 (worse loan grade) the odds of defaulting on a loan is multiplied by 0.76, We can also see that impact of loan grade is extremely significant with respect to defaulting.

One other basic linear model I wanted to look at was with loan status as the response variable and the categorical interest rate as the independent variable in the model (Figure 5). Here, we see that each bin is significant in relation to the loan default. As for the coefficients, when compared with a reference of 0-6, we can say that the probability of default increases.

**Existing Methods**

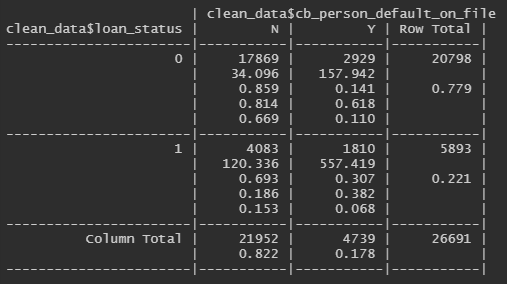
Now, before I started conducting my own, I wanted to review existing methods other people may have used to analyze the data. Because Kaggle is a public data platform anyone is free to download data and upload their findings, I will be referencing the users have uploaded to the site using the same data as I have for this research project.

The first person’s (manuelcuberli) work I reviewed performed their analysis in R. They explored the data using ggplot, which is a package in R to create a few scatter plots which looked at the distribution of some of the variables, such as age. Another good tactic they used was to run the most complex linear model, which as discussed in class, is also known as the additive model which contains the maximum number of parameters but excludes the interaction effects. After identifying the significant variables from the additive model, they created a new model that only included the significant variables identified in the summary. One thing I think Manuel could have approved upon was predictive analysis as he stopped after just creating a model that only contained significant variables.

The other remaining people who conducted analysis on this data set used Python to analyze it. The user named Mun Hong Lai, does a great job of cleaning the data and making clear to the reader why they are doing it. They also look at the histograms of different variables in the data set during the exploratory data analysis phase, similar to what I did, to identify the skewness among other details. One thing I think they could have improved upon was discritizing some of the variables such as employment length and interest rate. One approach I had in common with this user was the use of a decision tree(s) to help with decision making on borrowers. Overall, they were able to conduct more advanced machine learning using nearest neighbor analysis and were able to get a high prediction accuracy in their study of 0.939.

**Statistical Methods/Analyses/Model**

* **Which variable has the most significant impact on whether a person defaults on a loan?**
  + This question more so falls in line with exploratory data analysis. However, in order to determine which variable may have the most significant impact, we looked at the correlation matrix as well as the covariance matrix. In simple terms, the correlation matrix is able to compare the correlation between each variable in the data set and the covariance matrix will show that, if there is a correlation, what direction the variables move together in (positive or negative). I was unable to figure this out in R. However, I have experience doing this in Python, so I was able to get the results using the Pandas library with the cleaned data from R. This was the only question I used Python for.
* **Using binary analysis, what information are we able to identify by constructing a generalized linear model that only takes into account the continuous variables in our data with loan status as our response variable and what model might fit best ?**
  + For this model we will be using a binary response variable. The binary analysis is most useful when at least one of the explanatory variables is continuous (not categorical). In our data set the only continuous variables we have are loan interest rate and percent of income interest rate so this is what the model we based on. Some examples of this would be when a patient is labeled dead or alive, or in our case whether a person has defaulted on a loan or not. Mathematically, if the probability that a person has defaulted on a loan is ‘p’ then the probability of obtain y (0 or 1, i.e., default or non-default) it is given by where y has a mean of p and variance of p(1-p). We will compare two models; one will just have the main effects and the other one will be estimated using the interaction terms.
* **Are there differences between those who have a default on file and those who do not?**
  + The model we ended up using for this data is a log-linear model for categorical data. We will address this question using analysis of deviance to compare the fit of two competing models. One in which the default on file is independent of loan status and a second model that has the interaction between default on file and loan status. To do this, first we created a cross table using loan status and default on file. We could see the following information:

From this, we created another data frame and gave it a 2x2 factorial design to conduct analysis with a new column ‘y’ that includes the counts of the 4 possible combinations of loan\_status and default on file. We also converted both loan status and default on file to factors to conduct our log-linear analysis in R. Because we are using y as an observation of the counts, we are trying to model the expected number of counts as:

For this problem, ϒ1 is the proportion of did not default on the loan, ϒ2 is the proportion of those who did default on the loan. α1 is the proportion of those who do not have a default on file, α2 is the proportion of those who do have a default on file. The expected value of y in any cell in the table is a function of the (total # of observations) \* (the relative proportion of loan status) \* (the relative proportion of those who have a default on file). Before talking about the analysis, it is important to note that since we are looking at the independence of counts in this table, the distribution to use is Poisson.

* **Can we predict whether a borrower will default?**
  + As mentioned previously, the loan status is binary variable which takes two values (0: non-default, 1: default). That said, the best model to use is binomial logistic regression to classify borrowers into defaulted and non-defaulted. We will estimate our parameters based on a binomial model and find the best model with significant variables. In the end, the final model might be used by lenders to predict the status of borrowers which are not in the data set.

Before we start analysis, I will talk a little about the mathematical side of logistic regression.

Logistic regression models are based on the following equation:



Where ϴ(x) = E(Y|X = x) = P(Y = 1| X = x)

And here, x is the vector of predictor values and X is the matrix containing all the predictor values for all observations.

**Results**

* **Which variable has the largest correlation on whether a person defaults on a loan?**

This question is fairly easy to address. When we look at the correlation table, using Python. You can see this in Figure 17, the variable that has the highest correlation with loan\_status is loan\_percent\_income, this suggest that this has the largest impact on whether a person defaults or not. Additionally, if we look at the covariance matrix in Figure 18, we can get an idea of which direction the two variables move together. It can be seen that the covariance is stated as 0.0169 so we can expect an increase (closer to 1, default) as loan\_percent\_income increases.

* **Using binary analysis, what information are we able to identify by constructing a generalized linear model that only takes into account the continuous variables in our data with loan status as our response variable and what model might fit best ?**

As mentioned previously, for this question we would be creating a binomial model that used loan\_status as the response variable and we would the compare 2 models. One with only main effects, the other would include the interaction effect and main effects. To examine our simpler model (containing only main effects), we looked at the summary in which showed the residual deviance is 21,665 with a AIC of 21,675. For the interaction effect model, we also saw that in the summary, however, the AIC and residual deviance were slightly lower (AIC: 21354, Residual Deviance: 21,346). This would suggest that the interaction effect model might be slightly better but to confirm this we would have to run an analysis of variance to see if there is any statistical significance between the two models. When we ran this, we saw that the p-value was less than 0.05 for our interaction model, confirming our hypothesis that it would be a better model. Now using this model, we can see from the coefficient estimates that loan interest rate and proportion of their loan in comparison with their income are both positively related, while the interaction effect between interest rate and loan\_percent\_income is negatively related. These estimates, however, are the estimates in logits. The logit is the log of the probability of success so we cannot directly interpret the coefficients directly, but we are able to see the positive effect of loan\_int\_rate and percent\_income\_rate, this means the higher the interest rate and the higher the percentage of their annual income rate the loan is, the more likely the borrower is to default on a loan.

Next, to make sure that this model was a good fit I used glm diagnostic function in R to see how well the data fit. This looked at ordered deviance residuals vs. the quantiles of the standard normal, it is said that if the fitting is good, observations should be plotted around the lined in Figure 19, and as we can see for the most part it is plotted around the line, thus we can further conclude the model is a good fit. To examine this closer, I plotted the fitted model for each variable separately (Figure 6). This will return logistic curves for the increase or decrease in the incidence as a function of loan\_percent\_income and loan\_percent\_income increasing using the predicted values for loan status using models that contain only one independent continuous variable. Here we can see a logistic increase for both loan interest rate and loan percent income. As interest rate increases, the predicted value of loan status increases as well. This is a similar case for the predicted value of loan status when compared with loan\_percent\_income on the x-axis (Figure 6). We can see logistic growth once again, and it levels off right around 1, meaning it is almost always expected a customer will default on a loan with this large of loan compared to their income when the proportion of their loan is 80% or higher.

From this, we can say that in order to prevent future borrowers from defaulting, it might be beneficial to give out large loans with respect to their salary. Furthermore, it would be better to give borrowers’ a lower interest rate. Although a bank or whoever is collecting on a loan will make more money with this, it will result in a higher percentage of customers defaulting, thus being more costly than offering a lower interest rate.

* **Are there differences between those who have a default on file and those who do not?**

When conducting analysis to answer this question we first constructed our model for our null hypothesis which consisted of only main effects. We found through the summary of the model that the AIC was 834.5 with a residual deviance of 787.52 (Figure 7). Because the residual deviance for this model is higher the degrees of freedom, it may imply that the interaction effect model may be better. When we look at the fitted values in R, to compare what our count values will be based on the model, we see that it gets more inaccurate as our count variable increases. After comparing our alternative hypothesis model in Figure 8, we can see that this model is an extremely good fit, in fact it can’t get any better. When you plot the model on residuals vs. fitted (Figure 10) there is a perfect fit of the data This can further be confirmed when we look at the summary of the model in (Figure 8) which shows an AIC of 48.94, significantly better than our original model and a residual deviance of zero. Our model check of this was to look at the predicted values, and as we already figured, the model predicted and matched up exactly with our count data column.

Although, it had already been confirmed to this point that our alternative model would be more appropriate, we used an ANOVA test to accurately deny the null hypothesis. In Figure 9, we can see, as expected, the p-value is less than 0.05, suggesting that the interaction effect model is significantly better than the main effects model, and we reject the null hypothesis. Thus, resulting in a conclusion that there is an association between loan status and having a default on file.

* **Can we predict whether a borrower will default?**

Starting with practical estimation, we first used all the variables in the data set to generate the model. This is also known as the most complex model. When looking at the summary in Figure 11, using a significance value of 5%, we can see the following variables are not significant:

cb\_person\_default\_on\_file, cb\_person\_cred\_hist\_length because the p-value is greater than 0.05. Because our study has a prediction purpose, we will exclude these variables.

In our next table, we create another model that excludes the two variables that are not significant. The model summary can be seen in Figure 12. Before interpreting the deviance, it should be noted that a low null deviance means the predictors did not add any performance to the model and we can just use the null model (with only intercept). On the other hand, a low residual deviance means that the model containing the predictors is appropriate. In Figure 12, the null deviance is big (greater than degrees of freedom) and the residual deviance is small compared to the degrees of freedom. That means that our model is a good fit and it performs well with the data.

Next, in our ANOVA test in Figure 13, we compare the two models. It clearly shows that the second model we ran (where we removed the two insignificant variables) is better than our original complex model, thus we will stick with this for predictions.

We then looked at a 95% confidence interval for all of our variables from the model. In Figure 14, it can be seen that all the intervals do not contain zero, except for some classes of the categorical variables (i.e. the category homeownership OTHER is not statistically significant from the reference category for the variable person\_home\_ownership). The category B is also not statistically different from the reference category, A, for the variable loan\_grade.

Now for our coefficients, the beta estimates in Figure 12 are in log-odds form. That said, for instance, an increase of person income by one unit would lead to a change in the log-odds with -4.851e-05. Another important thing to note is the sign of each estimate. This sign will give us information about the effect (positive or negative) of the predictor variable on the dependent variable. For categorical variables, we can interpret the estimates as the change in the log-odds while using a reference category. For instance, borrowers with a home ownership listed as “OTHER” will increase the chance of them defaulting by 0.445 greater than the reference category, home ownership type of “MORTGAGE”.

An easier indicator to interpret is the Odds ratio. The odds ratio is a transformation of estimates by exponentiating the model parameters. In Figure 15, the table gives us the odds ratio for each model parameter. Taking the person age as an example, when age increases by one unit, the odds of default decrease by 1.3%. On the other hand, a supplementary unit of loan interest rate will lead to a 6.7% chance that the observation will be a default. The same interpretation is correct for the rest of variables in the model.

It should be noted that the loan grades have a pattern in terms of odds of each category compared to the reference (Loan Grade A). Indeed, once we move from grade A to grade G, the odds of default become higher. As a result, we can see that these categories will help us conduct a better prediction.

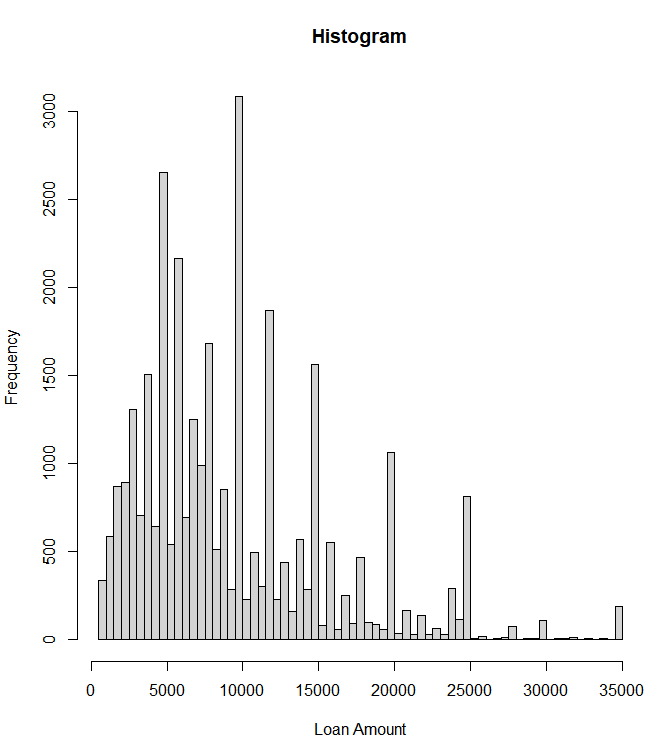
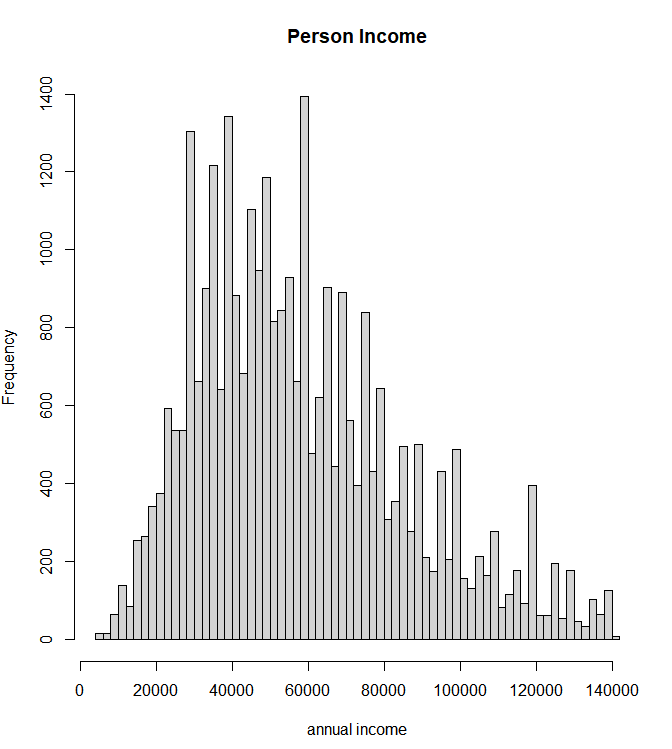
When we created this model, our goal was to predict the probability of a borrower defaulting on a loan. To do this, we needed to compare the predicted dependent variable versus its observed values using the confusion matrix. The confusion matrix describes the classification performance for each model (it could be on test data if data is split into train and test). For our case, in Figure 16, we can see the confusion matrix. The row values (0/1) indicate whether individuals have defaulted or not. The column values (FALSE, TRUE) represent whether the same individual is predicted to default or not. We can see that the accuracy is 0.748 +0.108 = 0.856, resulting in 85.6% accuracy.

To wrap up our remarks on this model and research question, we can see that this model is not the best. However, although not discussed in lecture, one way we might be able to improve on the model is to build a decision tree which may be better for classification and in predicting the borrowers’ default.

**Conclusion**

For our research study our purpose was to predict the probability of a borrower defaulting on a loan and find any underlying relationships between variables using linear and predictive modeling in R. We created a set of questions to determine what we needed to focus on for this project. Some things we wanted to focus on were correlations between variables and identifying if there were different variables, in our case we looked at differences between those who have a default on file and those who do not. One of the main limitations we came across was that we could not run some of the models we wanted to such as longitudinal models because our data did not fit the requirements to run that type of model. The managerial implications of this study would be that this could be very useful to the hypothetical loan company because with this knowledge, the company would be able to make better decisions on what types of loans to give and to who while decreasing loan default rates by carefully selecting who to give loans to.

**Appendix:**

Figure 1 Figure 2

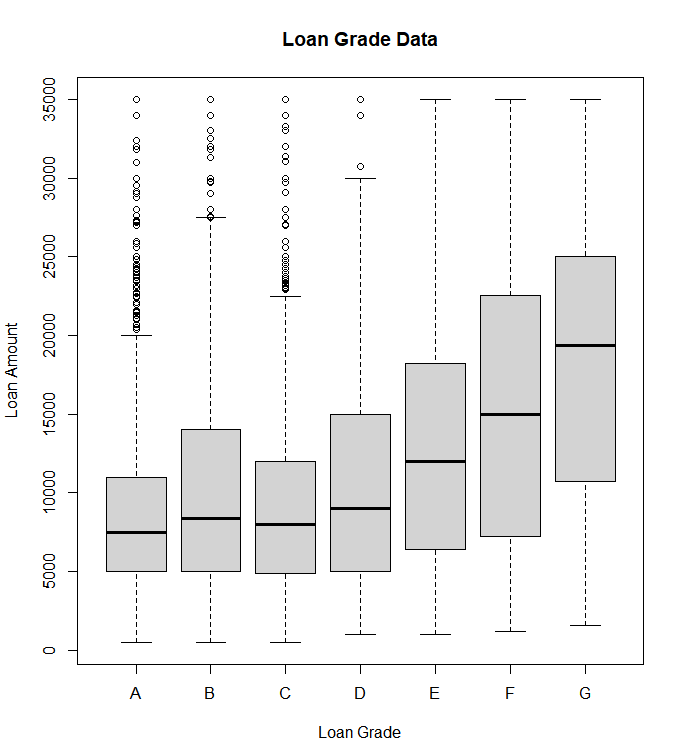
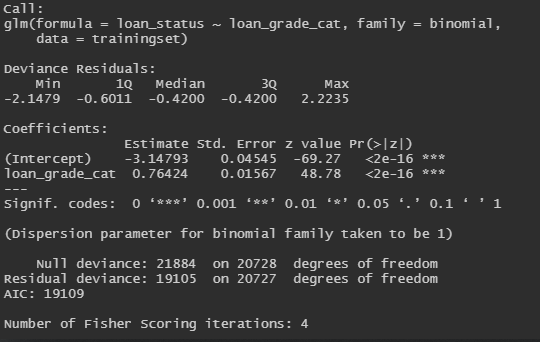
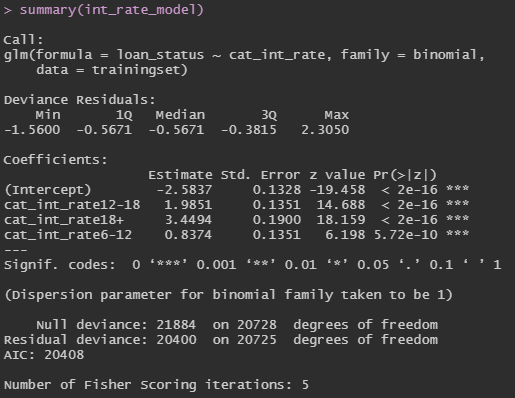
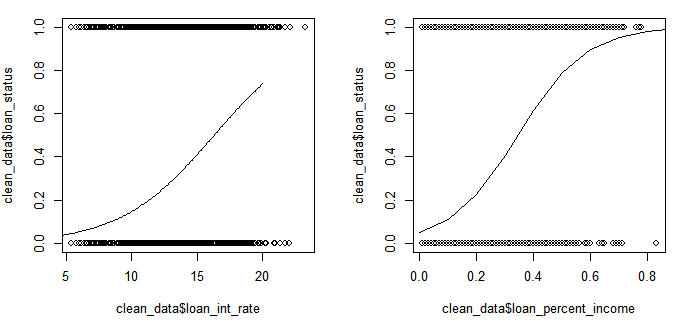
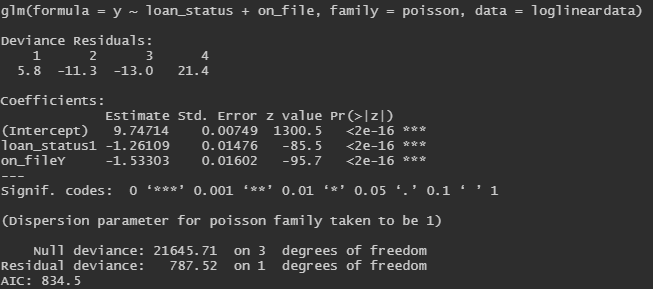
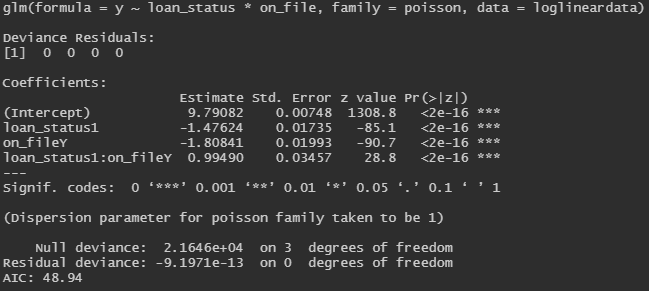
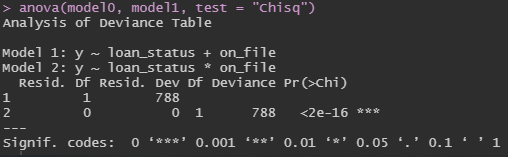
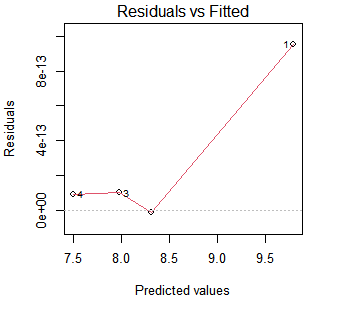
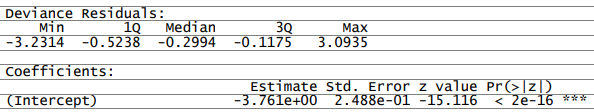
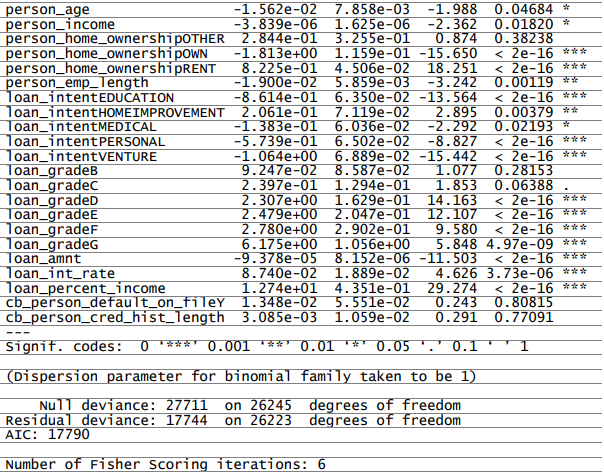
Figure 3 Figure 4

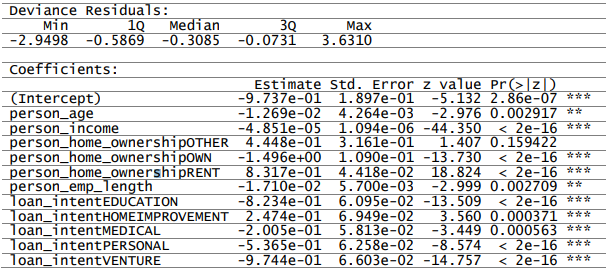
Figure 5 Figure 6

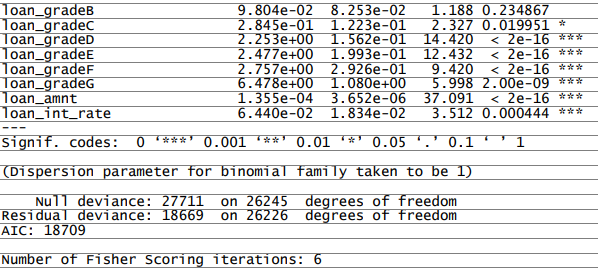


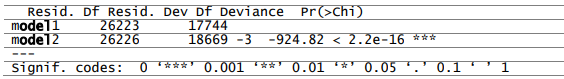
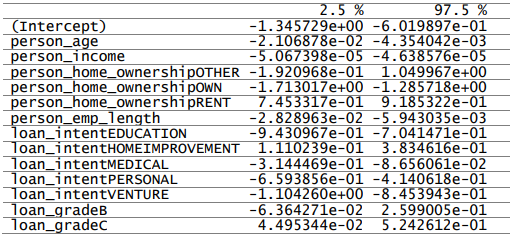
Figure 7 Figure 8

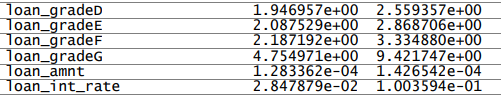
Figure 9

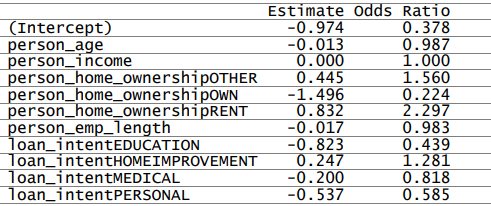
Figure 10 Figure 11

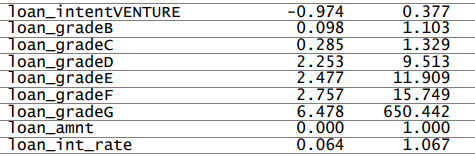
Figure 12



Figure 13 Figure 14



Figure 15 Figure 16



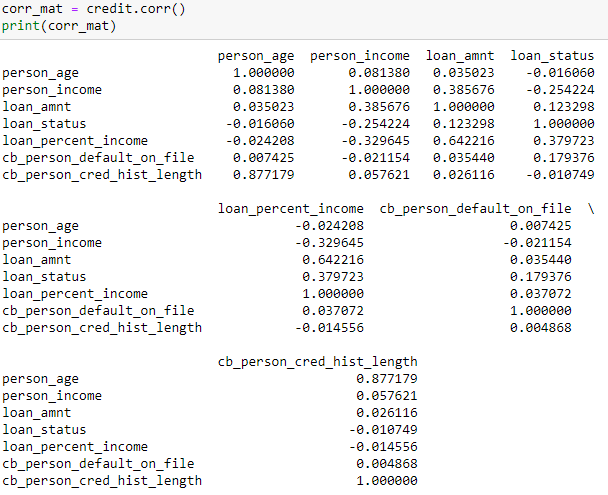
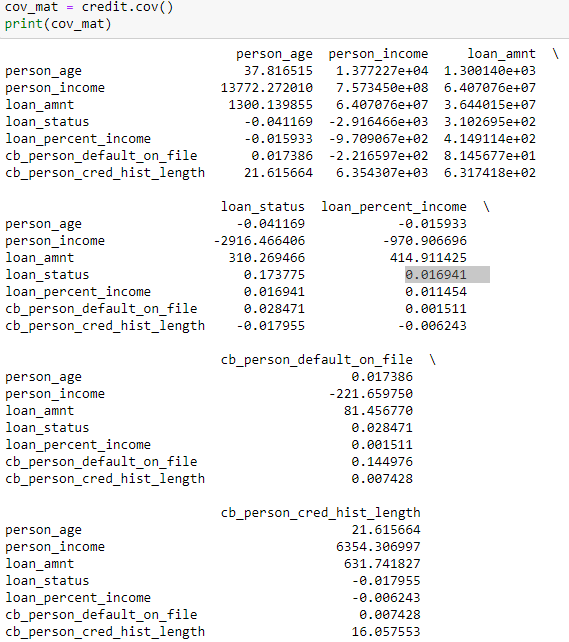
Figure 17 Figure 18

Figure 19

Work Cited

“Credit Risk Analysis Models - Overview, Credit Risk Types, Factors.” *Corporate Finance Institute*, 20 Feb. 2020, corporatefinanceinstitute.com/resources/knowledge/credit/credit-risk-analysis-models/.