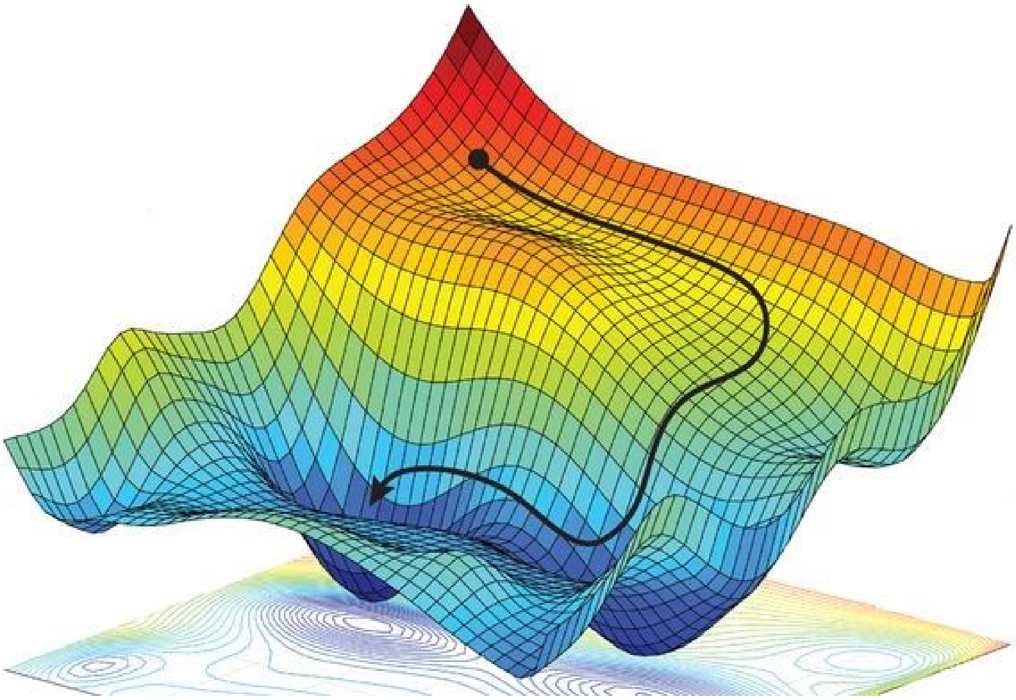
Predictive Analytics in the Personal Loan Market

Linear Regression and Classification Approaches to Estimating Income and Default Likelihood



Daniel Furlong |Gannon Condon | Kyle Gutstadt | Yasmine Kalhor

11.08.2020
GSBA-576

Executive Summary

This report explores different methods that can be used to predict an individual's income and whether or not individuals will default on their loans. Specific features connected to holders of personal loans, such as age, loan amount, and loan interest rate, will be key factors in this analysis. Regression and classification techniques played a key role in determining the variables of interest: income (continuous) and loan status (default: yes or no). This report provides a comprehensive analysis on simulated credit bureau data, including, but not limited to, data gathering, cleaning, preprocessing and descriptive statistics, as well as the building and implementation of multiple supervised machine learning modeling methods. In conclusion, final results and metrics regarding model selection will be shown and discussed.

Readers will gain a deeper understanding of how dependent and independent variables can be used to build effective, efficacious, and understandable models with varying levels of predictive power, while taking in-sample and out-sample accuracy/error into consideration. We also believe our findings will give insight into other ways of approaching similar objectives, including feature engineering, model selection, and hypothesis testing.

Data Collection

The data used for this project was collected from a dataset available on Kaggle, an online community of data scientists and machine learning practitioners. Privacy concerns surrounding actual personal loan data made it exceedingly difficult to find datasets with well-defined variables and used measurements. For these reasons, we decided to use a simulated dataframe to avoid any confusion.

The credit risk dataset used can be found here.

The original, uncleaned dataset contained 32,581 rows (observations) and 12 columns (variables), including both categorical and numeric variables. There will be two variables of interest throughout, one for ordinary least squares linear regression and another for the default risk classification task. They will be referred to as “person_income” and “loan_status” throughout.

From a high level, the variables can be divided between loan holder demographics and quality/quantity of the loan itself.

VARIABLE	DESCRIPTION
Person_age (numeric)	Age of the person at the time of the loan
Person_income (numeric)	Annual income of the person at the time of the loan
Person_home_ownership (categorical)	Type of ownership of the home: Mortgage, Own, Rent, Other
Person_emp_length (numeric)	Amount of time in years that person is employed
Loan_intent (categorical)	Is the aim of the loan: Grade A-G
Loan_grade (categorical)	Dimension of the loan taken (U.S. Currency)
Loan_amnt (numeric)	Dimension of the loan taken (U.S. Currency)
Loan_int_rate (numeric)	Interest rate paid for the loan
Loan_status (categorical; binary)	Dummy variable where 1 is default, 0 is not default
Loan_percent_income (numeric)	Ratio between the loan taken and the annual income
Cb_person_default_on_file (categorical; binary)	Whether the person has defaulted before (Yes = 2; No =1)
Cb_person_cred_hist_length (numeric)	Number of years of personal history since the first loan taken from that person

Cleaning

In order to avoid data loss, we used the *deletion when necessary* approach in order to keep and supplement NA values as needed.

While the simulated nature of the dataset mitigated inherent rawness/messiness, greatly easing the cleaning process, cleaning was still needed. Initial exploratory analysis quickly led to the discovery of erroneous outliers in both the *person_age* and *person-emp_length*. Figure 1.1 shows values related to age and employment greater than 120 years. Since someone cannot be employed longer than the number of years they are old, the value was changed to “NA”. The same approach was taken for any value in the person_age variable that was greater than 120 years. This accounted for a total of 7 altered values in the dataset.

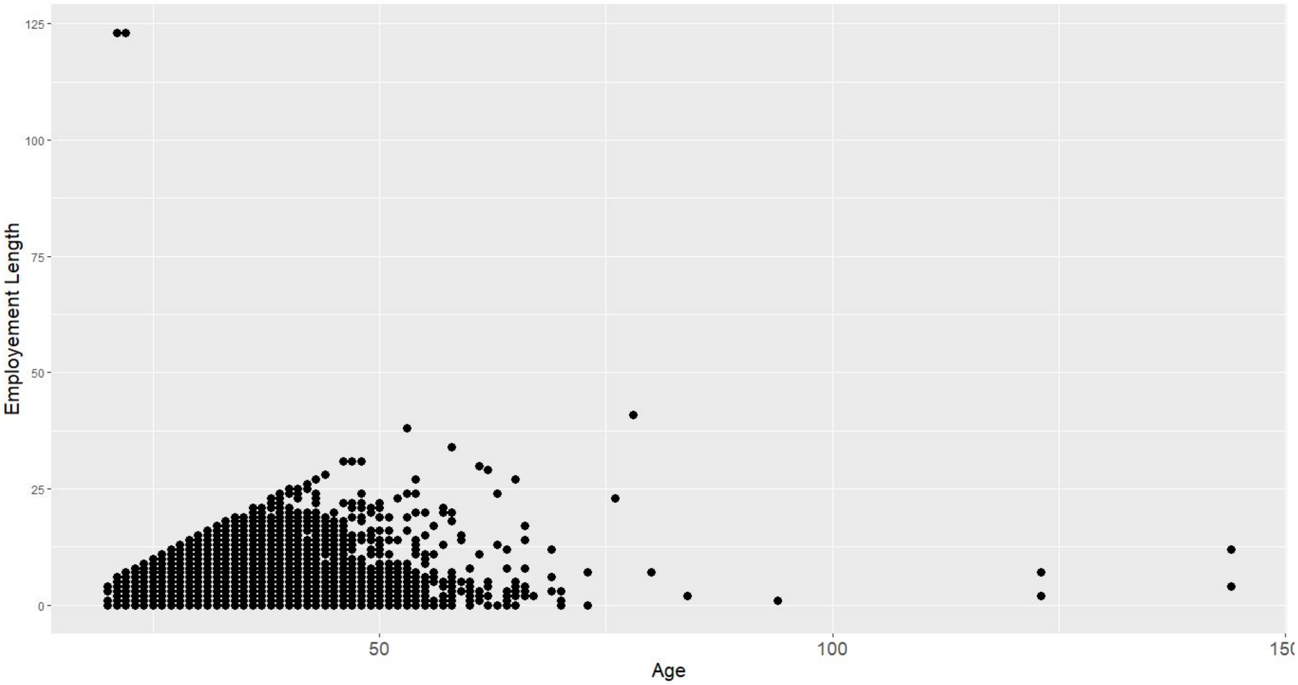


Figure 1.1

Omitting all NA values, using R Studio’s *na.omit* function, led to 12.12% data loss from the original, leaving a cleaned dataframe with a dimension of 28,632 x 12.

For further exploratory analysis, a heatmap matrix was created using the *reshape2* R library in order to visually show the correlation between all variables. As expected, there was a high correlation between: age and credit history length (.88) and the grade and interest rate of the loan (.93). There was moderate correlation between: loan amount and loan amount/income ratio (.58), history of default and loan grade (.54), and history of default and interest rate (.5). There was weak correlation between: home ownership category and income (-.24), loan amount/ income ratio and income (-.3) and home ownership categories and employment length. Further, loan status was weakly correlated (.34 to .38) with loan amount/ income ratio, loan grade, and interest rate.

From the heatmap in Figure 1.2, it is clear that certain variables have the potential to add bias and multicollinearity during the modeling process. This includes loan amount/income ratio (a function of loan amount divided by person income), age and credit history length, and loan grade and interest rate.

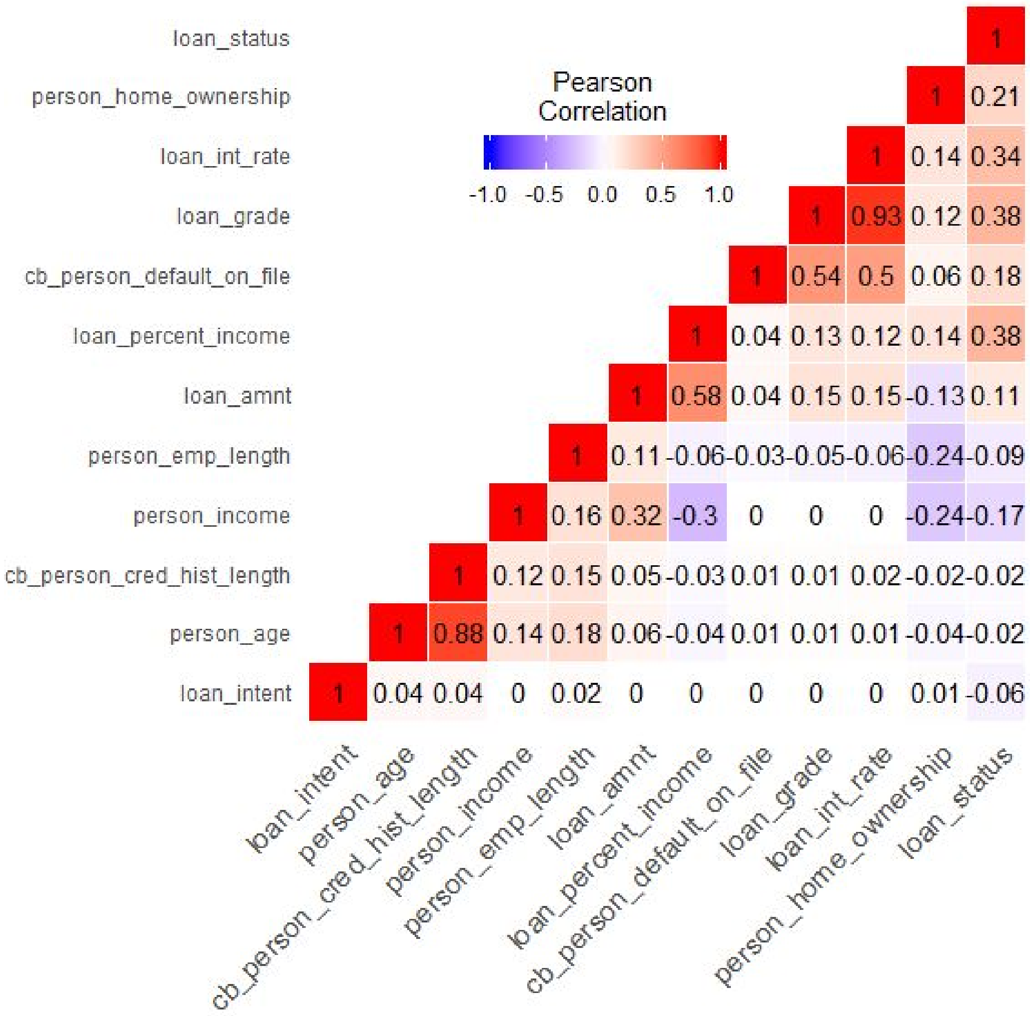


Figure 1.2

In order to visualize relationships, the *ggplot2* library was utilized to create scatter plots. Figure 1.3 shows the positive correlation between income and loan amount, as well as default occurring at all loan amounts, but with a potential bias towards lower incomes.

Note: all numeric variables were converted to factors using the *as.factor* function in order to show levels.

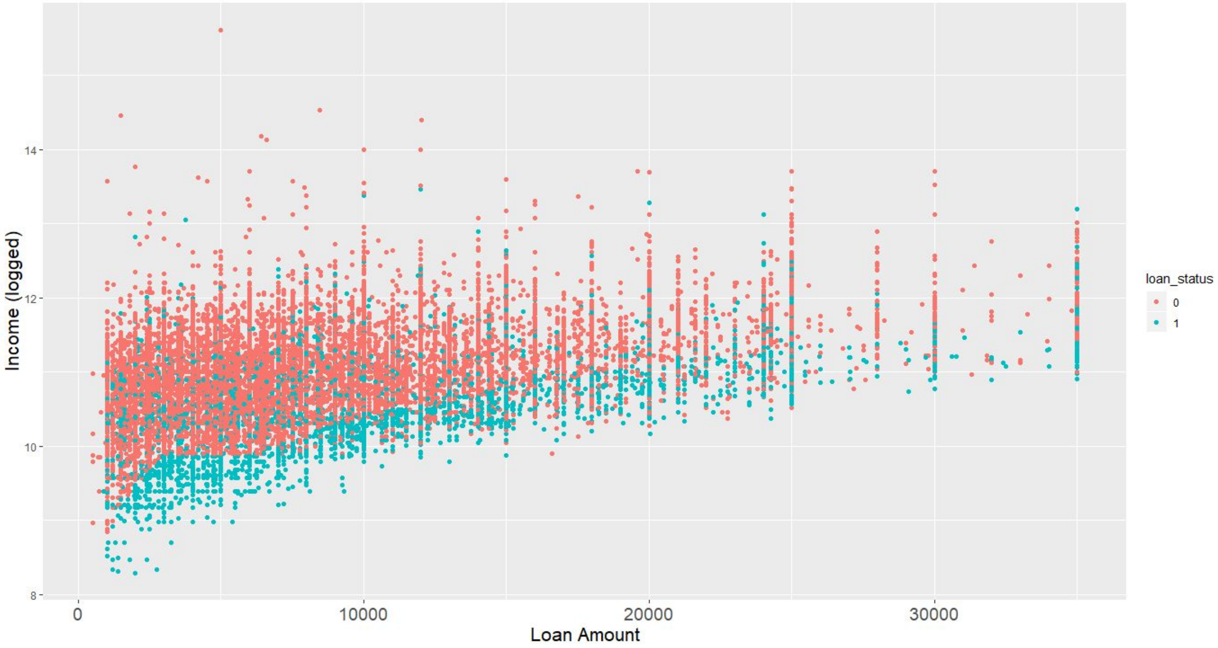


Figure 1.3

Figure 1.4 is the most visually instructive plot generated showing the relationship between income, interest rate and loan grade. There are clear interest rate ranges that encapsulate each grade category, with a tapering effect toward lower rated loans and lower income.

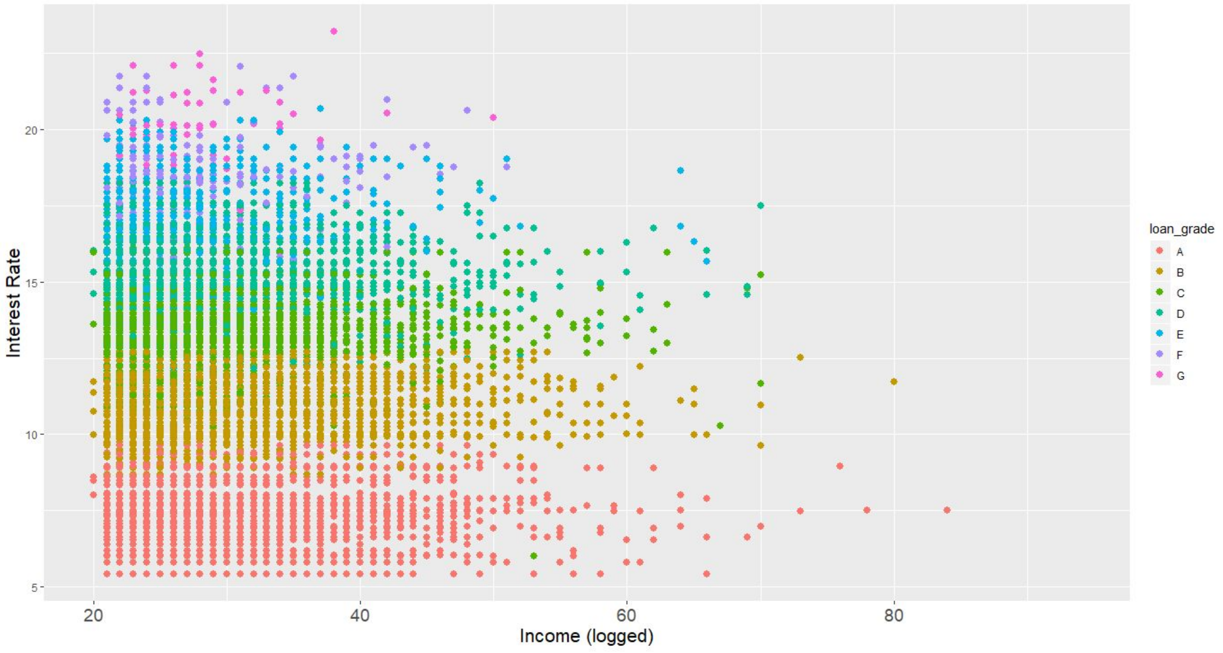


Figure 1.4

Based on Figure 1.4, we tested the assumption that higher interest rates were affiliated with higher default rates. Figure 1.5 supports this assumption showing a significant proportion of defaults occur in loan grades D through G.

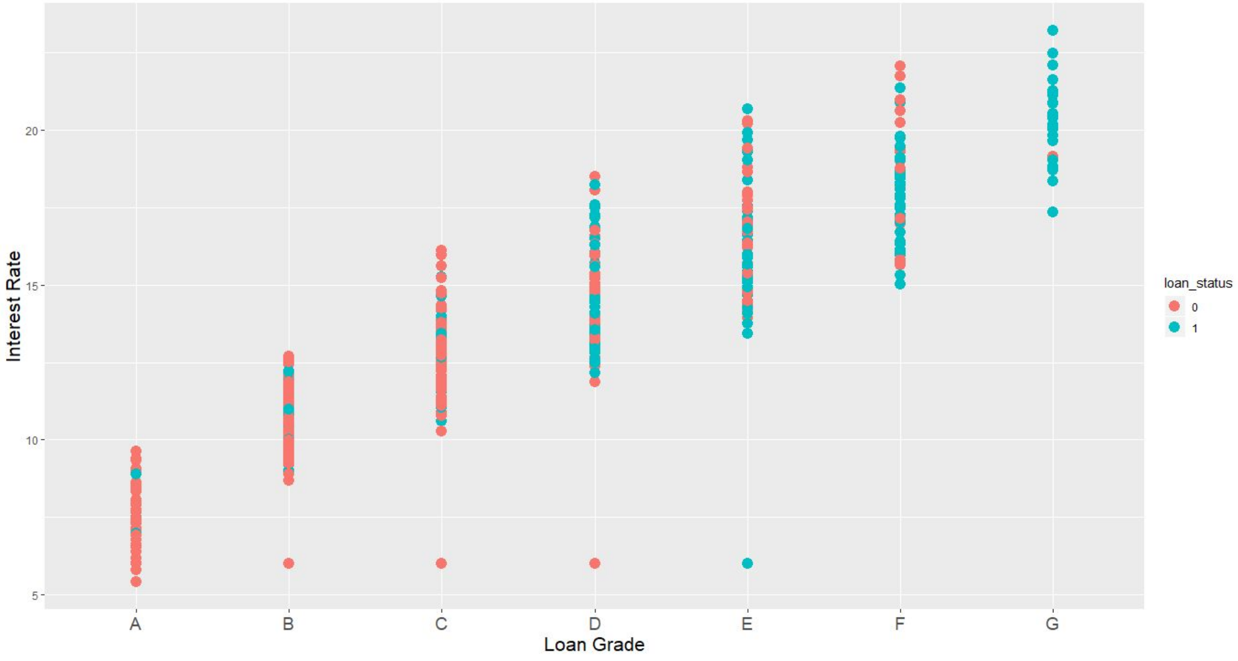


Figure 1.5

Figure 1.6 demonstrates that those who rent their living accommodations have a significantly higher change of defaulting on their loan. This can be seen in the right side of the chart with data points labeled “1”. This continues to suggest that those with lower incomes and thus less access to capital, generally have a lower credit rating and higher chance of defaulting.

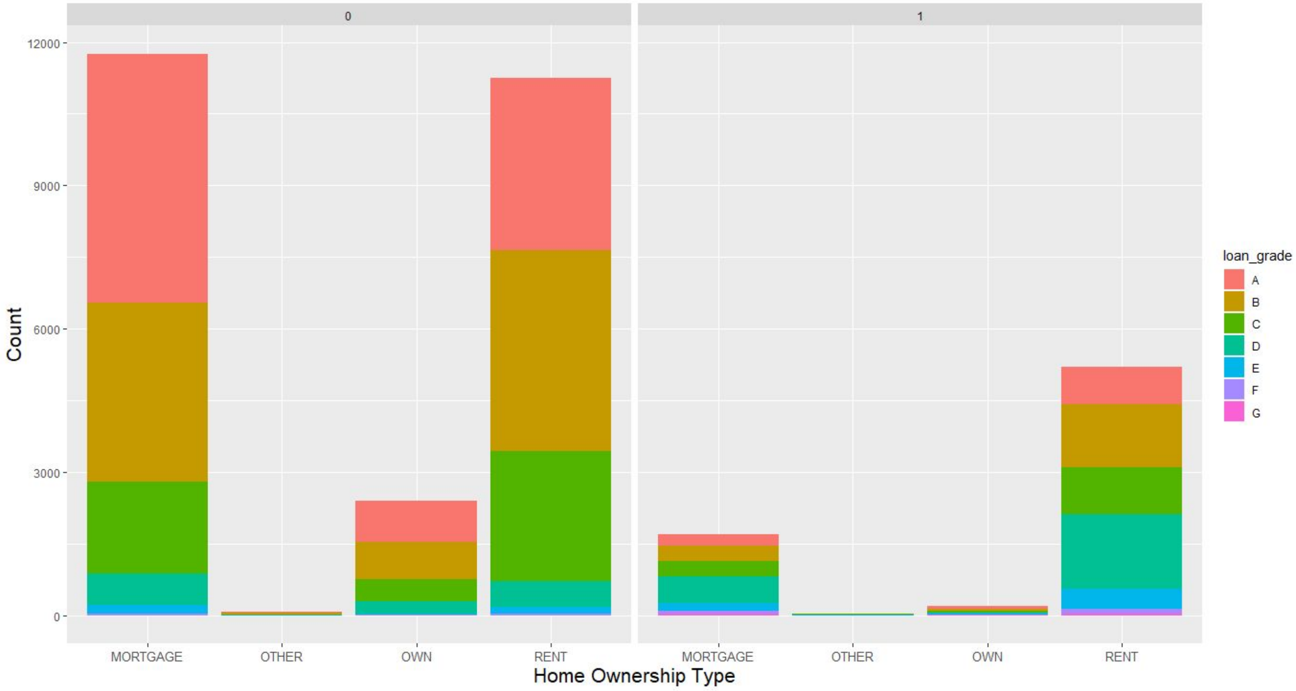


Figure 1.6

*For step-by-step cleaning procedure, please refer to the R code chunk provided with the report.

Partitioning

In order to test and verify results, the master data was partitioned into three separate sets: Train, Test, and Valid. Using the *createDataPartition* function in R Studio, we decided to run a 70/15/15 random stratified split to keep the loan_status variable of interest in equal proportions across each set.

Since the default rate in the master dataframe was imbalanced, as only 22% defaulted, balanced Training and Testing sets were created (split 70/30) to test for differences in accuracy during classification model selection. This method resulted in 56.37% data loss.

OLS Linear Regression Task

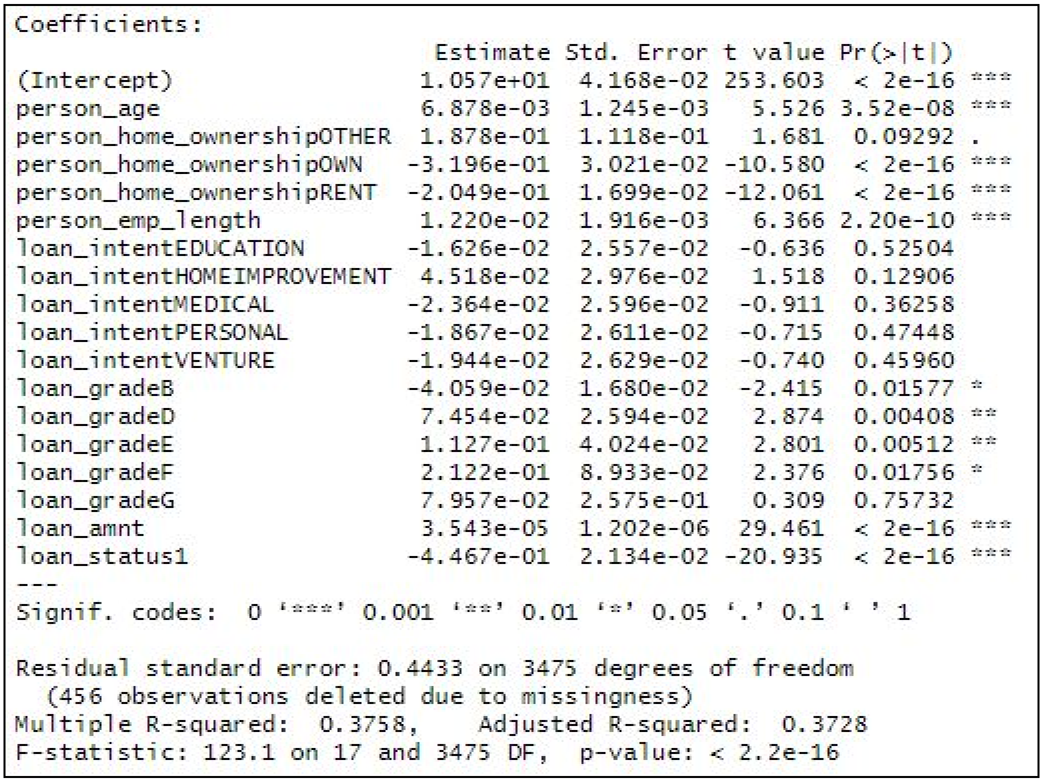
Our first linear regression model, Model1, was built around the person_income variable. Due to the range of incomes, $4,000-6,000,000, a log-level approach was used to avoid issues revolving around scaling -- this means that when interpreting Beta coefficients $\%\Delta\ Y = (100 \text{ x } Bi)\Delta\ x$. First, we kept all variables except: log_person_income and loan_percent_income because of their direct correlation with the variable of interest. Next, due to their low significance in the model output, we then removed cb_person_default_on_file, cb_person_cred_hit_legnth, and loan_int_rate (Model2), which resulted in an out of sample RSE of .4476. To further investigate we excluded the “OTHER” factor level from person_home_ownership and factor level “C’ from loan_grade as they were the two remaining insignificant variables in the model. To compare, we built three additional models, two in which we removed each variable individually and another removing both. Our results show that of the four models, TrainModel3 had the lowest out of residual standard error of.4458.

Model	In-Sample RSE	Out-Sample RSE
Model1	0.447	0.4476
Model2 (!= “OTHER” level)	0.447	0.448
Model3 (!= Loan Grade “C” level)	**0.4471**	**0.4458**
Model4 (!= “OTHER” or “C”	0.4471	0.4463

A residual analysis was run for each OLS model. This involved plotting the residuals and calculating the mean -- for each model the residuals mean was zero. A multicollinearity check was also run on each independent variable using *caret’s vif* function. Using a VIF greater than 10 as a threshold value, none of the models returned values causing concern.

The final step of the OLS regression task was to run the validation set through the chosen model, model3. The RSE of the validation set was .4433, which was lower than the in-sample training model and out-sample testing model, indicating the efficacy of model3 on unseen data. From the output table below, it can be seen that a one unit increase in the loan_status variable leads to a 44.67% decrease in income. This speaks to the relationship between income and the default rate. This relationship will be further examined in the classification section.

Model	Validation Set RSE
Model3	**0.4433**



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                1.057e+01  4.168e-02 253.603  < 2e-16 ***
person_age                 6.878e-03  1.245e-03   5.526 3.52e-08 ***
person_home_ownershipOTHER  1.878e-01  1.118e-01   1.681  0.09292 .  
person_home_ownershipOWN  -3.196e-01  3.021e-02 -10.580  < 2e-16 ***
person_home_ownershipRENT -2.049e-01  1.699e-02 -12.061  < 2e-16 ***
person_emp_length          1.220e-02  1.916e-03   6.366 2.20e-10 ***
loan_intentEDUCATION      -1.626e-02  2.557e-02  -0.636  0.52504    
loan_intentHOMEIMPROVEMENT  4.518e-02  2.976e-02   1.518  0.12906    
loan_intentMEDICAL        -2.364e-02  2.596e-02  -0.911  0.36258    
loan_intentPERSONAL       -1.867e-02  2.611e-02  -0.715  0.47448    
loan_intentVENTURE        -1.944e-02  2.629e-02  -0.740  0.45960    
loan_gradeB               -4.059e-02  1.680e-02  -2.415  0.01577 *  
loan_gradeD                7.454e-02  2.594e-02   2.874  0.00408 ** 
loan_gradeE                1.127e-01  4.024e-02   2.801  0.00512 ** 
loan_gradeF                2.122e-01  8.933e-02   2.376  0.01756 *  
loan_gradeG                7.957e-02  2.575e-01   0.309  0.75732    
loan_amnt                  3.543e-05  1.202e-06  29.461  < 2e-16 ***
loan_status1              -4.467e-01  2.134e-02 -20.935  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4433 on 3475 degrees of freedom
  (456 observations deleted due to missingness)
Multiple R-squared:  0.3758,	Adjusted R-squared:  0.3728 
F-statistic: 123.1 on 17 and 3475 DF,  p-value: < 2.2e-16
```

Validation set output

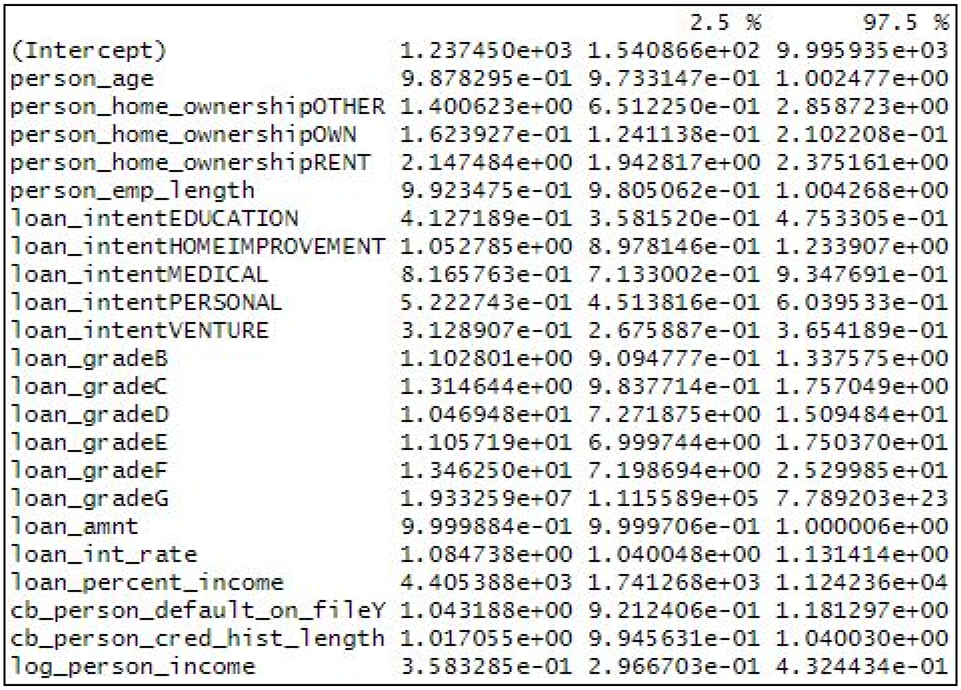
*Note: a low R-Squared of 0.3758 shows that approximately 62% of the variation of the output cannot be explained by the input variables.

Classification Task

A. Logistic Regression Model 1

To begin, we wanted to run the regression on loan_status with all the variables in the model as a baseline. Looking at the resulting summary, we decided to cut person_emp_length, loan_amnt, cb_person_default_on_file, and cb_person_crd_hist_length due to their insignificance to the model's predictive ability. This led to a model with worse in and out of sample error. Deciding to evaluate the original model's results in a different way, we noticed that person_age was only significant at the .01 level and also highly correlated with all of the previously removed variables (except loan_amnt). Further manipulation of the data, led us to discover that removing person_age had a .01% cumulative change across both in and out of sample testing accuracy. At this point the model appeared to be most effective including all variables, but we wanted to see the effect of increasing/decreasing the threshold rate. When set to 0.6, accuracy across samples produced the lowest accuracy witnessed from previous testing. Using a threshold of 0.4, led to almost as poor results. One notable observation was that lowering the threshold to 0.4 resulted in the highest Kappa values in both in and out of sample showing a high level of agreement between predicted and realized results while accounting for a randomness penalization. Realizing that both threshold alterations did not lead to more accurate models, we concluded that the original model regressing all variables was the most accurate predictor of loan_status.

The chart below shows the odds-ratio interpretation, along with the confidence interval, of the coefficients, using base e for conversion. It can be seen that loan grade G has the largest positive effect on the odds-ratio, increasing the likelihood of the event occurring. Unsurprising considering that 98.44% of everyone in the master dataset with a Grade G loan defaulted. The greatest negative effect on the odds-ratio comes from OWN factor level, decreasing the likelihood of defaulting on a loan. This also aligns with the fact that only 7.5% of those who own their home in the master dataset default on their loan.

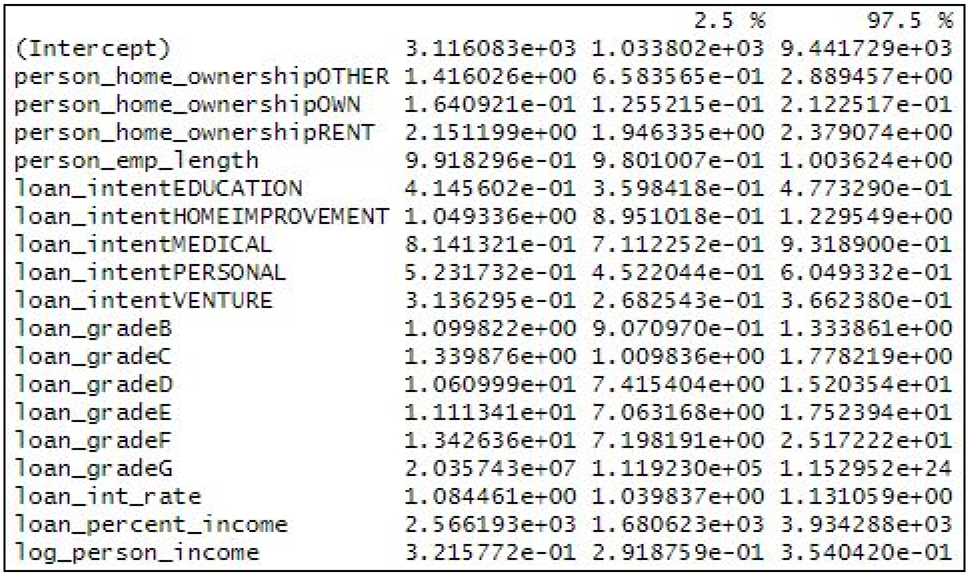


```
                                             2.5 %       97.5 %
(Intercept)                 1.237450e+03 1.540866e+02 9.995935e+03
person_age                  9.878295e-01 9.733147e-01 1.002477e+00
person_home_ownershipOTHER 1.400623e+00 6.512250e-01 2.858723e+00
person_home_ownershipOWN   1.623927e-01 1.241138e-01 2.102208e-01
person_home_ownershipRENT  2.147484e+00 1.942817e+00 2.375161e+00
person_emp_length           9.923475e-01 9.805062e-01 1.004268e+00
loan_intentEDUCATION        4.127189e-01 3.581520e-01 4.753305e-01
loan_intentHOMEIMPROVEMENT 1.052785e+00 8.978146e-01 1.233907e+00
loan_intentMEDICAL          8.165763e-01 7.133002e-01 9.347691e-01
loan_intentPERSONAL         5.222743e-01 4.513816e-01 6.039533e-01
loan_intentVENTURE          3.128907e-01 2.675887e-01 3.654189e-01
loan_gradeB                 1.102801e+00 9.094777e-01 1.337575e+00
loan_gradeC                 1.314644e+00 9.837714e-01 1.757049e+00
loan_gradeD                 1.046948e+01 7.271875e+00 1.509484e+01
loan_gradeE                 1.105719e+01 6.999744e+00 1.750370e+01
loan_gradeF                 1.346250e+01 7.198694e+00 2.529985e+01
loan_gradeG                 1.933259e+07 1.115589e+05 7.789203e+23
loan_amnt                   9.999884e-01 9.999706e-01 1.000006e+00
loan_int_rate               1.084738e+00 1.040048e+00 1.131414e+00
loan_percent_income         4.405388e+03 1.741268e+03 1.124236e+04
cb_person_default_on_fileY 1.043188e+00 9.212406e-01 1.181297e+00
cb_person_cred_hist_length 1.017055e+00 9.945631e-01 1.040030e+00
log_person_income           3.583285e-01 2.966703e-01 4.324434e-01
```

B. Logistic Regression Model 2

When first thinking about which variables to include for the logistic regression for the loan_status variable of interest, person_income was removed since a logged version of the variable had been created to mitigate scaling issues. Furthermore, loan_amnt, cb_person_cred_hist_length, cb_person_default_on_file, and person_age were removed while monitoring accuracy after every iteration of the model. To figure out the ideal threshold value, a while loop function identified 0.5 as the optimal value to maximize accuracy. The resulting model led to accuracy of 86.7% in-sample and out-sample accuracy of 87.49%. A low p-value signifies that this accuracy rate is significantly different from the no information rate.

The converted coefficient odds-ratio table below shows very similar results to the logistic regression model discussed above. Loan grade G still has the largest positive effect on the odds-ratio, increasing the likelihood of the event occurring, and home ownership has the greatest negative effect on the odds-ratio.



```
                                                2.5 %       97.5 %
(Intercept)                 3.116083e+03 1.033802e+03 9.441729e+03
person_home_ownershipOTHER 1.416026e+00 6.583565e-01 2.889457e+00
person_home_ownershipOWN   1.640921e-01 1.255215e-01 2.122517e-01
person_home_ownershipRENT  2.151199e+00 1.946335e+00 2.379074e+00
person_emp_length           9.918296e-01 9.801007e-01 1.003624e+00
loan_intentEDUCATION        4.145602e-01 3.598418e-01 4.773290e-01
loan_intentHOMEIMPROVEMENT 1.049336e+00 8.951018e-01 1.229549e+00
loan_intentMEDICAL          8.141321e-01 7.112252e-01 9.318900e-01
loan_intentPERSONAL         5.231732e-01 4.522044e-01 6.049332e-01
loan_intentVENTURE          3.136295e-01 2.682543e-01 3.662380e-01
loan_gradeB                 1.099822e+00 9.070970e-01 1.333861e+00
loan_gradeC                 1.339876e+00 1.009836e+00 1.778219e+00
loan_gradeD                 1.060999e+01 7.415404e+00 1.520354e+01
loan_gradeE                 1.111341e+01 7.063168e+00 1.752394e+01
loan_gradeF                 1.342636e+01 7.198191e+00 2.517222e+01
loan_gradeG                 2.035743e+07 1.119230e+05 1.152952e+24
loan_int_rate               1.084461e+00 1.039837e+00 1.131059e+00
loan_percent_income         2.566193e+03 1.680623e+03 3.934288e+03
log_person_income           3.215772e-01 2.918759e-01 3.540420e-01
```

C. Classification and Regression Tree

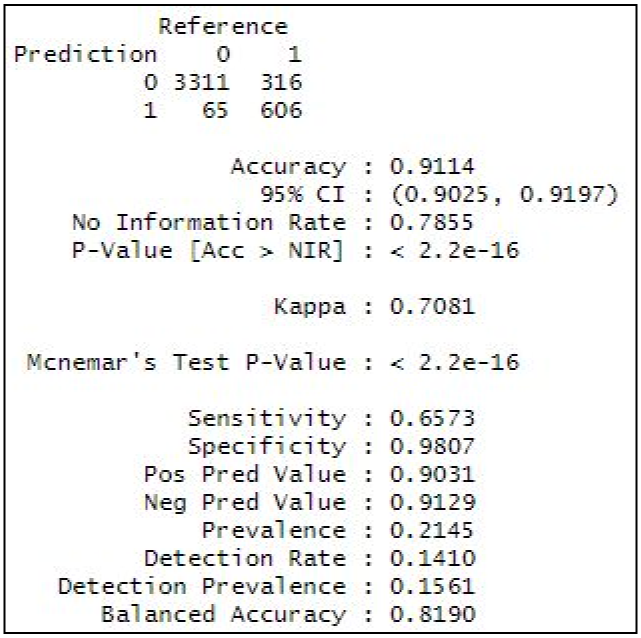
The third classification modeling method used was the CART decision tree methodology, which was accessed through the *rpart* function in R. This package implements a greedy algorithm for feature selection, implementing an iterative process until enough features have been found or the tree cannot split the data any further.

Our CART decision tree settled on eight internal nodes and five layers, leading to 92.24% accuracy in-sample and 92.35% out-sample, the highest percent among our classification models. This nonparametric technique created an intuitive, easy to visualize decision tree (below) and ultimately ended up being the model we selected to test the validation set. However, it is important to note that due the nonparametric nature of CART, no generalizations or inferences can be made regarding the underlying probability distribution.

The CART Dendrogram can be found here.

D. Support Vector Machine

When building the SVM model in R, we input our classification variable ‘loan_status’ and the other 11 independent x variables. For the model to run we converted the loan_status variable into a factor. Additionally, we had to implement the deletion method with respect to cleaning the na’s from both the testing and training data. This reduced the observations from 22,808 to 20,023 for the training data, and from 4,887 to 4,298 for the testing data. The model (SVM0) was then built using the training data and then we used the built model to make predictions using the testing data. Below is the confusion matrix for the out of sample forecast using the testing data.

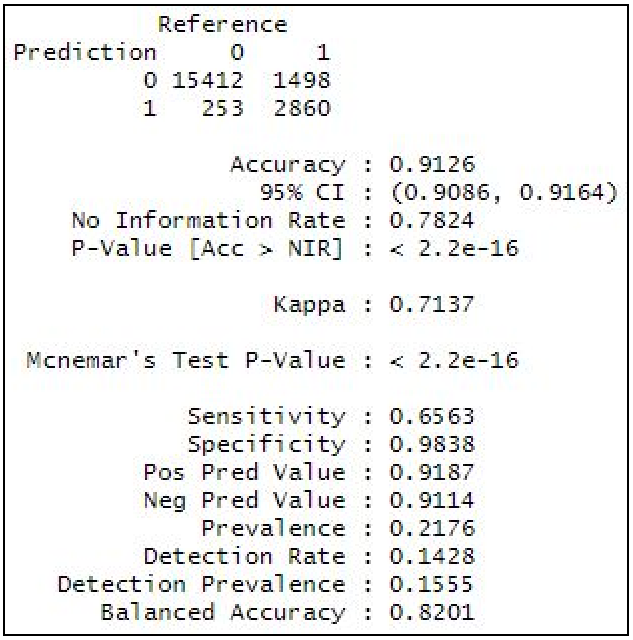


```
          Reference
Prediction    0    1
         0 3311  316
         1   65  606

               Accuracy : 0.9114
                 95% CI : (0.9025, 0.9197)
    No Information Rate : 0.7855
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7081

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6573
            Specificity : 0.9807
         Pos Pred Value : 0.9031
         Neg Pred Value : 0.9129
             Prevalence : 0.2145
         Detection Rate : 0.1410
   Detection Prevalence : 0.1561
      Balanced Accuracy : 0.8190
```

Based on the output of the confusion matrix for the testing data we can see that we only have an error rate of 8.86% due to the high accuracy. We can also see that the accuracy of the model is 12.59% greater than if we simply assume that no clients would default (No Information Rate). Additionally, we can see from the P-value that due to it being smaller than 5% we have sufficient evidence to reject the null hypothesis that accuracy is greater than the no information rate. Therefore, we can say that the model is statistically better than the no information rate. These statistics are informing us that we have quite an effective model that can be used to predict loan defaults.

Moving onto the other statistics, we can see from the Kappa value that there is quite a large amount of agreement between the reference class and the predicted class, despite being penalized for the potential of random agreement. The miniscule Mcnemar’s Test p-value indicates that there is in fact a difference between the clients that would default from the clients that would not default, based on the independent variables that were used to create this model. Regarding the sensitivity value, we can see that if in reality some clients would be in the loan default class, we would correctly predict that 65.73% of them would default. The specificity class is the inverse of the above statistic, i.e given clients that are in the non-default class, we would correctly identify the non-default clients 98.07% of the time. The positive predictive value indicates that given that we predict that a client would default, the probability that the prediction was correct is 90.31%. The negative predictive value is the opposite of this, i.e given that we predict a client would default, what is the probability the prediction is correct? The prevalence indicates the proportion of the positive class (default) within the total sample data, ergo 21.45% of the clients in the sample data do default. The detection rate informs us that we correctly predicted 14.10% of the clients in the positive class, from the total sample of 4,298 clients. The detection prevalence illustrates the total number of positive class predictions as a percentage of the total sample size. Finally, the balanced accuracy is the average of the sensitivity and the specificity statistics.

Below is the confusion matrix for the in sample forecast, and you can see the figures overall are quite similar to the out of sample forecast that we have described above.



```
          Reference
Prediction     0     1
         0 15412  1498
         1   253  2860

               Accuracy : 0.9126
                 95% CI : (0.9086, 0.9164)
    No Information Rate : 0.7824
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7137

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6563
            Specificity : 0.9838
         Pos Pred Value : 0.9187
         Neg Pred Value : 0.9114
             Prevalence : 0.2176
         Detection Rate : 0.1428
   Detection Prevalence : 0.1555
      Balanced Accuracy : 0.8201
```

We attempted to tweek the cost and gamma values in the svm module using the *tune()* function in order to improve the predictive power of the SVM model we built. However, this workload was very intensive on our computers and we decided to stop after 30 minutes. We felt that if tuning the model took that amount of time, it would not be feasible and the opportunity cost too high to implement such tweeks.

Model	In-Sample Accuracy	Out-Sample Accuracy
Logistic Regression 1 *No Information Rate : 0.7824	0.8676	0.8751
Logistic Regression 2 *No Information Rate : 0.7824	0.867	0.8749
CART	**0.9224**	**0.9235**
SVM *No Information Rate : 0.7855	0.9126	0.9114

Running the chosen CART technique on the validation set results in a 92.88% accurate predictive model.

Model	Validation Set Accuracy
CART	**0.9288**

*Note: each selected classification model was also used to test the balanced datasets resulting in a 10-15% drop in accuracy. However, in each case the p-value was below the .05 threshold signalling a significant difference between the accuracy and no information rate.

CONCLUSION

Using supervised machine learning techniques we were able to generate models with higher accuracies and significant p-values, supporting the hypothesis they are effective methods to calculate credit risk and loan default. However, calculating a continuous variable such as income proves to be more inaccurate due to the limited amount of contributing independent variables available in the dataset.

For the classification task, after comparing all four model's accuracy rates we determined that the CART model was the most successful at predicting whether an individual would default. Unfortunately, when we pursued using a random forest the computation was too intensive for our computers and we were not able to find out if it might have given us an improved result. One other aspect that we thought about pursuing but decided against was evaluating the models based on sensitivity and specificity values. We believe that a more risk averse financial institution would prioritize maximizing the sensitivity value for a given model, due to the greater financial loss that would be incurred from incorrectly predicting defaulting clients from the defaulting population. Analyzing the models based on sensitivity and specificity would allow us to find which model would cater to different needs a client might have or an institution’s risk appetite, possibly to minimize their risk by prioritizing loans being declined correctly or to grow their business through an increase in successful loans.