Lending Club Analysis

Scott Monaco

1/22/2022

Contents

Executive Summary

Lending Club is an online peer-to-peer platform that connects individual borrowers with individual investors. The compnay is the largest such platform that provides an important source of liquidity in a market where institutional investors have historically neglected. As of September 30, 2019, Lending Club has originated over $53 billion in total loan issuance.

However, as more borrowers sign on to the platform, the supply of loans plays a role in the evolving underwriting standards, which can impact investor results based on how loans perform. As a result, it is increasingly important to understand what factors are most associated with creditworthiness.

Goal of the Analysis

The goal of this analysis is to identify a set of important features that will predict loan status, which we classify as "good" for a loan that is fully paid off and "bad" for a loan that is charged off. Our analysis uses data from the period between 2007-2011 for which we have 38,971 observations and 38 attributes. We have developed four categories of models to predict loan status, which are:

- Logistic Regression based model (Backwards Selection)
- LASSO model
- Elastic Net model
- Random Forest model

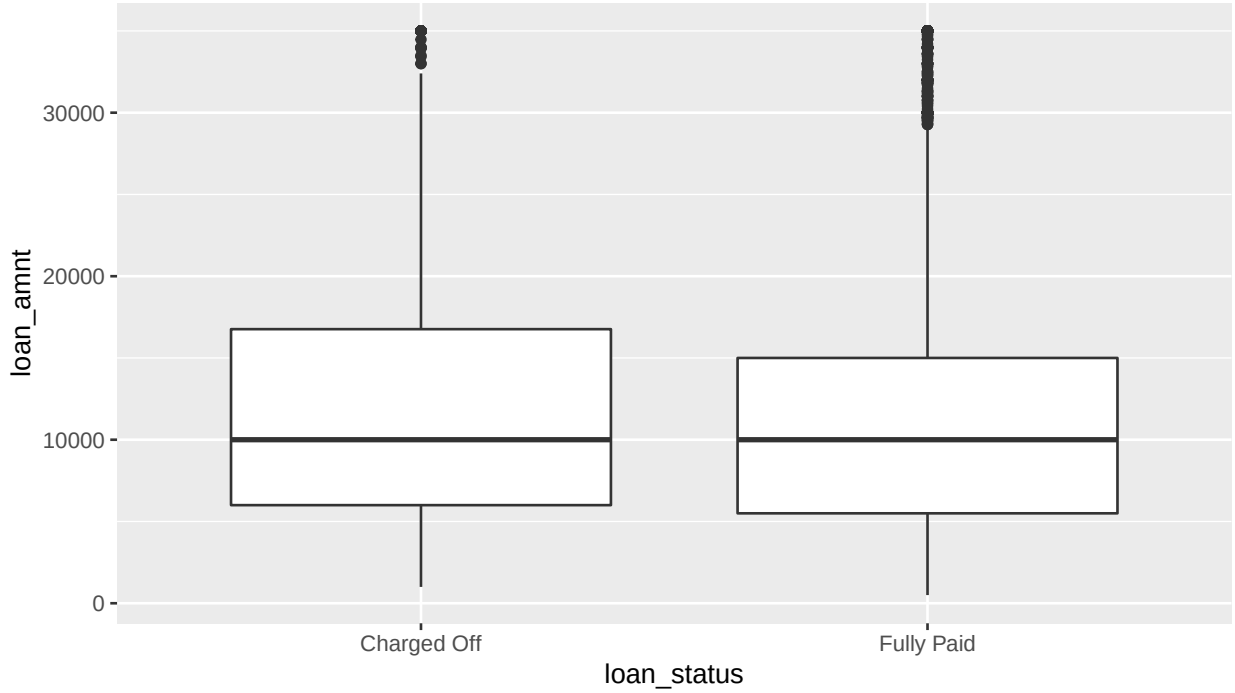
Main Findings

The final analysis reveals that significant predictive power exists among variables given in the data. In particular, term, interest rate, the ratio of monthly debt obligations to self reported income (dti), revolving line utilization rates, the number of credit inquiries in the past 6 months, the number of derogatory public records, and the number of bankruptcies have a negative association with the response variable of loan status = fully paid off. That is to say, the greater numbers or percentages of these variables correspond to a lower chance of a loan being fully paid off. Conversely, annual income is positively associated with the response variable. That is, the greater the annual income, the greater chance of fully paying off the loan.

Data Summary and Exploration

Data Overview

The dataset on the loans includes attributes such as loan amount, home ownership status, interest rate, loan status and grade of the loan among many others. Of the 38 predictors, 15 of them are categorical in nature (see **Exhibit 1**). The response variable for the analysis is loan status, which is a two-level response variable. Of the 38,971 observations in the data set, 33503 have loan status fully paid off (86%), while the remaining 14% of the data are charged off. Below is a graphical summary of loan status by loan amount, which shows significant overlap in the size of the loan and whether they were paid off or not.



Quantitative and Graphical Summaries

To get a better understanding of the predictor variables we are working with in this analysis, we looked at a correlation table of the numerical variables to understand where there may be some collinearity.

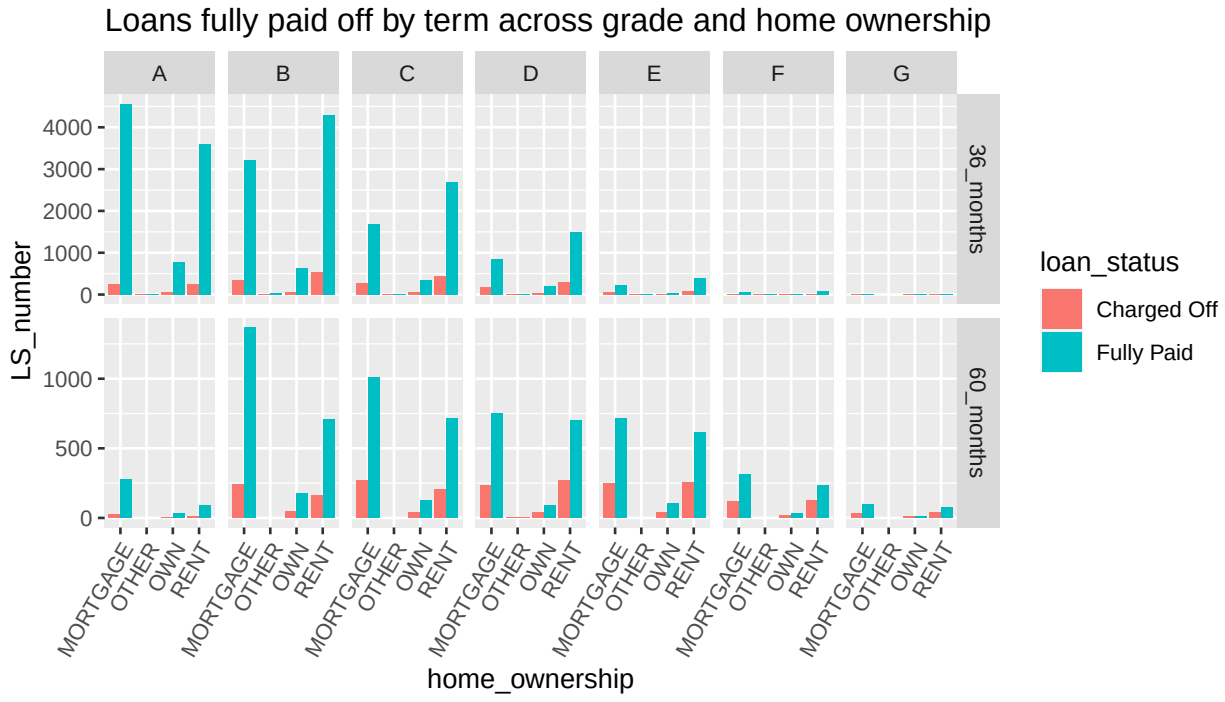
```
##                          loan_amnt   int_rate          dti revol_util installment
## loan_amnt              1.00000000 0.31161507  0.066145107 0.06644093  0.92967151
## int_rate               0.31161507 1.00000000  0.107009902 0.46698558  0.28497767
## dti                    0.06614511 0.10700990  1.000000000 0.27559822  0.05454552
## revol_util             0.06644093 0.46698558  0.275598221 1.00000000  0.09638883
## installment            0.92967151 0.28497767  0.054545522 0.09638883  1.00000000
## pub_rec               -0.05163309 0.09695236 -0.005232117 0.05872691 -0.04666102
## pub_rec_bankruptcies -0.03694740 0.08239579  0.005461790 0.06122984 -0.03386322
##                             pub_rec pub_rec_bankruptcies
## loan_amnt            -0.051633087          -0.03694740
## int_rate              0.096952356           0.08239579
## dti                  -0.005232117           0.00546179
## revol_util            0.058726907           0.06122984
```

```
## installment          -0.046661017          -0.03386322
## pub_rec               1.000000000           0.84575407
## pub_rec_bankruptcies  0.845754067           1.00000000
```

Here, we can see that installment and loan amount are highly correlated, which makes sense since one is a derivation of the other. Also the number of derogatory public records is highly correlated with the number of public record bankruptcies. This information will be useful in model building later.

With this correlation matrix helping us understand key summaries and relationships of numerical variables, we know look at similar graphical summaries of potentially important categorical variables.

```
## `summarise()` has grouped output by 'term', 'grade', 'home_ownership'. You can override using the `.g
```



The above graph provides a breakdown of loan status by term across grade and homeownership. First, it appears that the higher the grade (i.e. A, B, C), the greater number of loans paid off. Second, loans that have a term of 36 months show greater frequency of fully paid loans. Finally, mortgage and rent show associations with fully paid loans at the higher grade levels. This association weakens as the grade lowers, and more charged off loans appear among both mortgage owners and renters.

The last part of the data exploration involves a summary of the association between geographic features (i.e. state) and other predictors. We saved a heatmap for **Exhibit 2** that shows which states have the highest average loan amounts. Below is a breakdown of the worst performing states by % defaulted:

```
## # A tibble: 5 x 3
##   addr_state     n percentage
##   <fct>      <int>      <dbl>
## 1 NV           494      0.221
## 2 AK            80      0.2
## 3 TN            10      0.2
## 4 SD            62      0.194
## 5 FL          2816      0.173
```

Problems with the Data & Data Cleaning

First, we note there is no missing data. However, the data does need to be cleaned. The levels of verification status were duplicated so we merged the data into two levels. Second, states and earliest credit line had categories with too few observations in it, therefore we grouped observations among many categories into one bin. Third, we decided to remove highly correlated variables from our consideration, such as grade since grade and sub grade are collinear. Fourth, we had to remove the variables related to post loan data because they are not useful in predicting the quality of loans before they are issued. Finally, we split the dataset into three sets. The training data set (50%) that will be used to train our models. The testing data set (25%) that will be used for model selection. The validation data set (25%) that will be used to summarize the final model.

Variables Considered as Input

After cleaning the data in accordance with the previous section, we have the following variables under consideration as input for our model building

```
##  [1] "loan_amnt"             "term"                  "int_rate"
##  [4] "installment"           "sub_grade"             "emp_length"
##  [7] "home_ownership"        "annual_inc"            "loan_status"
## [10] "purpose"               "dti"                   "delinq_2yrs"
## [13] "inq_last_6mths"        "open_acc"              "pub_rec"
## [16] "revol_bal"             "revol_util"            "total_acc"
## [19] "pub_rec_bankruptcies"  "verification.status"   "state"
## [22] "earliest.cr.line.year"
```

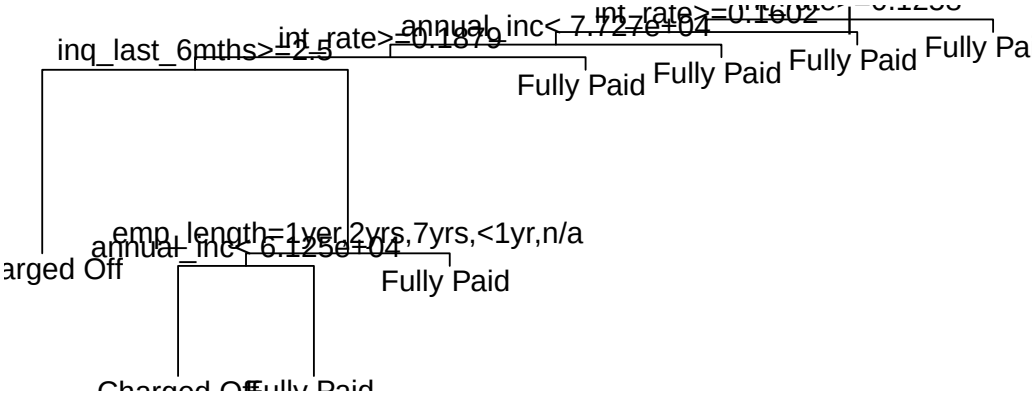
Identify Important Risk Factors

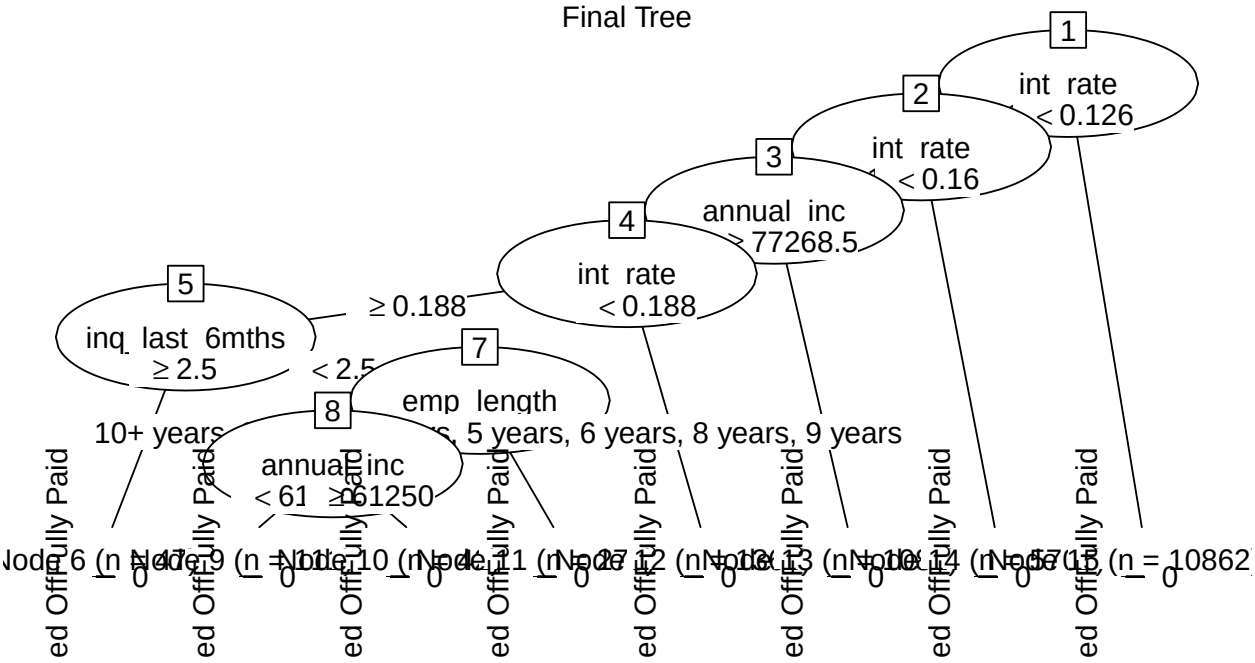
Before we build our models for our predictive analysis, we will examine a tree model based on minimum deviance (cp) to identify risk factors.

Tree-based model

To identify risk factors that a loan will be defaulted from a tree based model, we used an r partition with minimum cp = .00078 that contains all of the variables under consideration. This effectively creates a tree based on minimum deviance that will help us identify important variables for predicting loan status and possible interactions to consider.

```
## n= 19485
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 19485 2733 Fully Paid (0.1402617 0.8597383)
##     2) int_rate>=0.12575 8623 1731 Fully Paid (0.2007422 0.7992578)
##       4) int_rate>=0.16015 2921  764 Fully Paid (0.2615543 0.7384457)
##         8) annual_inc< 77268.5 1839  543 Fully Paid (0.2952692 0.7047308)
##          16) int_rate>=0.18785 475  183 Fully Paid (0.3852632 0.6147368)
##            32) inq_last_6mths>=2.5 47   21 Charged Off (0.5531915 0.4468085) *
##            33) inq_last_6mths< 2.5 428  157 Fully Paid (0.3668224 0.6331776)
##              66) emp_length=1 year,2 years,7 years,< 1 year,n/a 153   69 Fully Paid (0.4509804 0.549
##               132) annual_inc< 61250 111   54 Charged Off (0.5135135 0.4864865) *
##               133) annual_inc>=61250 42   12 Fully Paid (0.2857143 0.7142857) *
##              67) emp_length=10+ years,3 years,4 years,5 years,6 years,8 years,9 years 275   88 Fully
##          17) int_rate< 0.18785 1364  360 Fully Paid (0.2639296 0.7360704) *
##         9) annual_inc>=77268.5 1082  221 Fully Paid (0.2042514 0.7957486) *
##       5) int_rate< 0.16015 5702  967 Fully Paid (0.1695896 0.8304104) *
##     3) int_rate< 0.12575 10862 1002 Fully Paid (0.0922482 0.9077518) *
```





Based on the tree output, interest rate is determined to be the key first split. Interest rates are negatively associated with the probability of paying off a loan. That is, the lower the interest rate, the greater chance the loan will be paid off. Another factor that is negative associated with a fully paid loan status is term. The lower the term, the greater chance of paying off a loan. Annual income shows a positive association with the response. In particular the tree diagram shows that incomes over ~$77000 have a higher chance of paying off a loan.

Model Build and Evaluation

In building models for our analysis, we want to consider a range of options, including general linear models, lasso equations with different lambdas, elastic net models, and random forests with different mtry.

Model Building (10 models under consideration)

As a start, we explore logistic regression as a basic analytical approach to predicting loan status. Our methodology explores different ways to develop a range of models

- **Model 1**: Logistic regression model, using backwards selection to hone in on significant predictors based on p-value 0.05.

Our first model under consideration was arrived by eliminating, one-by-one, the highest p-value predictor based on the Anova output. **Exhibit 3** shows the final Anova output and summary of the model.

- **Model 2**: Relaxed LASSO model corresponding to lambda = lambda min

Our second model is the model obtained from a LASSO, but relaxed to fit a glm model. This corresponds to the same variables as a LASSO model, but due to the shrinkage of LASSO estimates, the coefficient estimators for model 2 should be slightly larger than those for traditional LASSO (see **Exhibit 4**).

- **Model 3**: Relaxed LASSO model corresponding to lambda = lambda first

Our third model is the model obtained from the LASSO with lambda first, but relaxed to fit a glm model. This corresponds to the same variables as traditional LASSO, but due to the shrinkage of LASSO estimates, the coefficient estimators for model 3 should be slightly larger than those a typical LASSO using lambda first (see **Exhibit 5**).

- **Model 4**: "Kitchen sink" model, using backwards selection and interaction terms

Our fourth model takes the first model as a base model and considers the impact of interactions. The first significant interaction included was the one employment length and annual income. This interaction shows that the impact of employment length on loan status depends on annual income. In particular, the higher the annual income, the lower probability of paying off a loan for those borrowers who are employed at the company for 2 or more years. The second significant interaction included was the one between interest rate and revolving line utilization rate. The impact of interest rate on loan status depends on revolving line utilization rate. As the utilization rate increases, then the slope of the interest rate variable increases, which means the probability that a loan gets fully paid increases (see **Exhibit 6**).

- **Model 5**: Best General Linear Model (Bestglm) with smallest AIC

Our fifth model was built using the bestglm methodology with the smallest AIC as the model selection criterion and nvmax = 15. Furthermore, this model was built using an exhaustive method. Once the variables were identified from this model, we then fit the variables into a glm fit, as shown in **Exhibit 7** (i.e. relaxed).

- **Model 6**: Parsimonious Model

Model 6 was chosen to be the parsimonious model in the group. It was selected based on the relaxed fit of the best glm arrived at by model 5, and then shrunken down to eliminate all variables, one-by-one via backwards selections, that have p-value greater than 0.05 according to the Anova output (see **Exhibit 8**).

- **Model 7**: Random Forest

Our seventh and final model was a random forest arrived at by using mtry = 4 (rounded down from square root of 18 predictors) and ntree = 250 (see **Exhibit 9**).

Model Evaluation

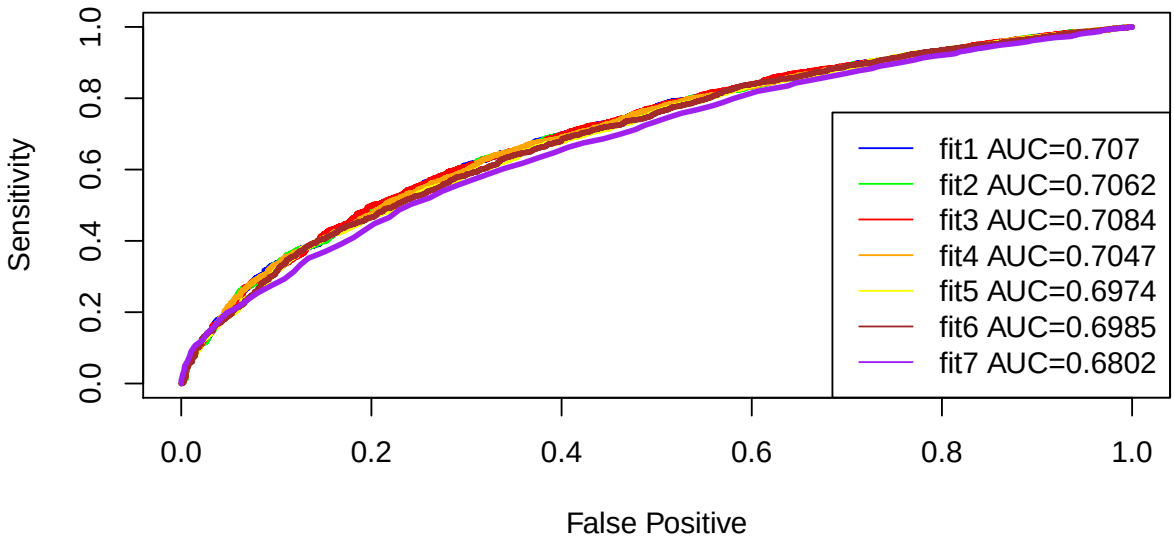
With seven models in hand, we will now apply these classifiers to the testing data and evaluate the classifiers using a method that maximizes the return per the estimate for a sensible loss ratio for picking up a bad loan. In this case, we will evaluate each of the classifiers using testing misclassification error, which will be chosen to maximize the return for a two-to-one loss ratio of picking up a bad loan to that of missing a good loan.

For illustrative purposes, we will also provide a comparison of the ROC curves for each of our classifiers built.

Since the loss ratio of picking up a bad loan to that of missing a good loan is 2:1, that means that false positives are twice as costly as false negatives. Given our notation for false positive, $a_{0,1} = L(Y = 0, \hat{Y} = 1)$, and false negative, $a_{1,0} = L(Y = 1, \hat{Y} = 0)$, then a risk ratio of $a_{0,1} = 2a_{1,0}$ implies that the optimal rule for classifying a good loan (i.e. denoted "1" or "fully paid" is

$$\hat{P}(Y = 1|x) > \frac{2}{(1+2)} = 0.67$$

Below is a summary of the ROC curves for each of the 10 model fits, along with a summary output for testing AUC. However, misclassification error remains our key criterion for selecting a model.



Finally, the data frame below shows the output for the testing MCE assuming a risk ratio of 1/2 for each of our 10 models.

```
##   Model 1 MCE Model 2 MCE Model 3 MCE Model 4 MCE Model 5 MCE Model 6 MCE
## 1   0.2779431   0.2777379   0.2755825   0.2764036   0.2778405   0.2777379
##   Model 7 MCE
## 1    0.283075
```

As we can see above, model 4 has the lowest testing overall weighted misclassification error assuming a risk ratio of 1/2. Therefore, based on testing MCE as the model selection criteria, we will choose model 4.

Model Selection and Conclusion

Model Selection

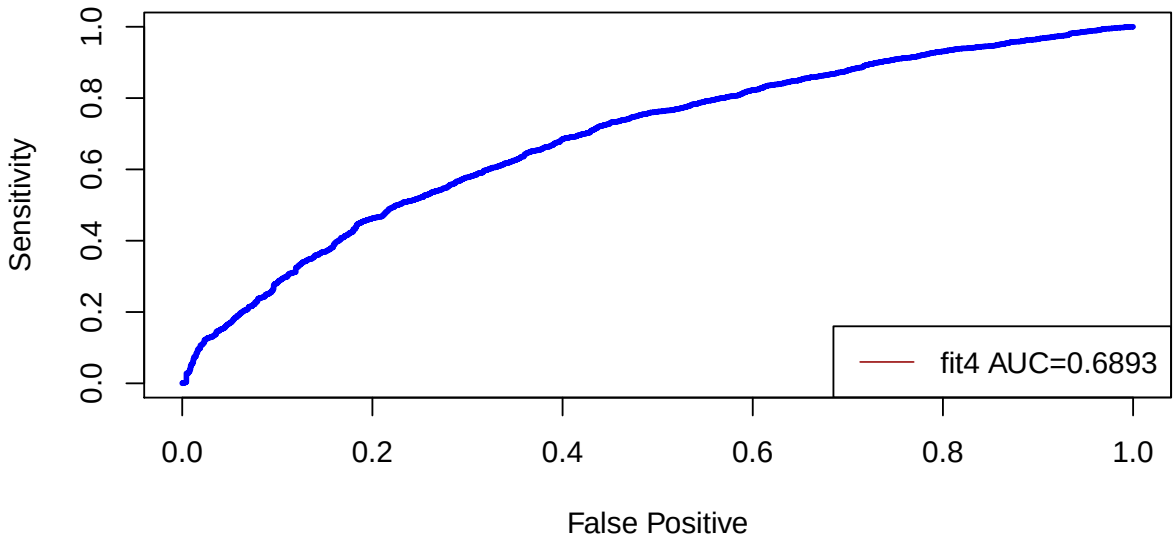
The last step of our analysis is to validate our model using the validation data set in loan.validation.Final.

```
## [1] 0.2827671
```

The final misclassification rate is 0.283, based on validation data.

```
## Setting levels: control = Charged Off, case = Fully Paid
```

```
## Setting direction: controls < cases
```



The final validation AUC is 6893.

Conclusion

Based on the final model chosen from the validation data set, we conclude that the model accounting for interactions should be the selected model for predicting loan status for the following reasons:

- It has significantly superior predictive power, as measured by MCE, compared to the LASSO, elastic net and random forest models.
- It still maintains some interpretability so that we can draw insights into the risk factors that affect the status of a loan.

These insights could provide enormous value to Lending Club and potential investors as they assess new loans that apply on the online platform. For example, there are specific states and sub grades that tend to be positively associated with good borrowers. Using these insights, we could target sub-gradeG1 loans, which appear to have the best association with loans fully paid off while sub-gradeD3 has the worst association with loans fully paid off. Illinois is the best state to lend to while Washington is the worst, based on the data. Employment length of 5 years has the strongest association with fully paid off loans. If the purpose of the loan is for paying of a credit card bill or for a wedding, this has positive associations with paying off the loan. However, if the purpose is education, then these types of loans have the worst associations with fully paid off loans. Finally, if the earlist credit line was 1996, then these types of loans have positive associations with the probability of fully paying off the loan.

Despite the wide range of data mining techniques used in this analysis, if we had more time or more processing power, we would wish to improve upon the study by gathering more information existing assets of the borrower or the FICO score of the borrower. This information could have predictive power in determining whether a loan will be fully paid off or not.

Appendix

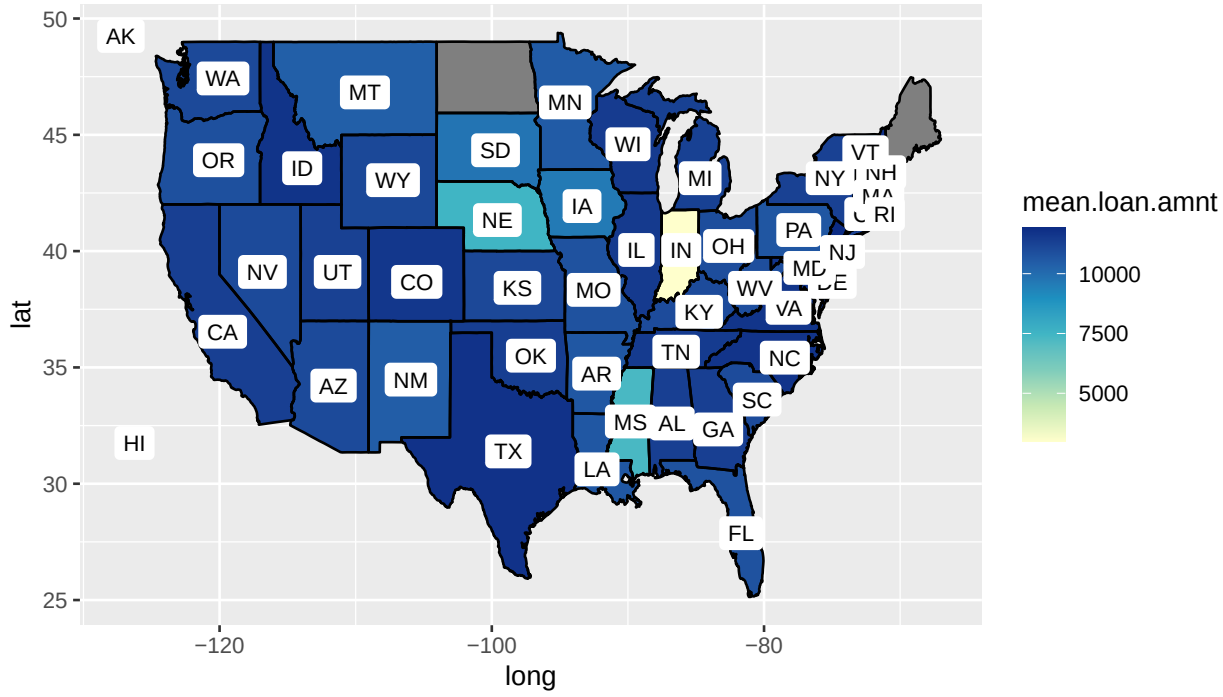
- **Exhibit 1**

```
## Classes 'data.table' and 'data.frame':   38971 obs. of  38 variables:
##  $ loan_amnt              : int  5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...
##  $ funded_amnt            : int  5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...
##  $ funded_amnt_inv        : num  4975 2500 2400 10000 3000 ...
##  $ term                   : Factor w/ 2 levels "36_months","60_months": 1 2 1 1 2 1 2 1 2 2 ...
##  $ int_rate               : num  0.106 0.153 0.16 0.135 0.127 ...
##  $ installment            : num  162.9 59.8 84.3 339.3 67.8 ...
##  $ grade                  : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...
##  $ sub_grade              : Factor w/ 35 levels "A1","A2","A3",..: 7 14 15 11 10 4 15 21 27 10 ...
##  $ emp_title              : Factor w/ 28303 levels "","  old palm inc",..: 1 18662 1 329 23262 23645
##  $ emp_length             : Factor w/ 12 levels "1 year","10+ years",..: 2 11 2 2 1 4 9 10 5 11 ...
##  $ home_ownership         : Factor w/ 4 levels "MORTGAGE","OTHER",..: 4 4 4 4 4 4 4 4 3 4 ...
##  $ annual_inc             : num  24000 30000 12252 49200 80000 ...
##  $ verification_status    : Factor w/ 3 levels "Not Verified",..: 3 2 1 2 2 2 1 2 2 3 ...
##  $ issue_d                : Factor w/ 52 levels "Apr 2008","Apr 2009",..: 14 14 14 14 14 14 14 14 14 14
##  $ loan_status            : Factor w/ 2 levels "Charged Off",..: 2 1 2 2 2 2 2 2 1 1 ...
##  $ purpose                : Factor w/ 14 levels "car","credit_card",..: 2 1 12 10 10 14 3 1 12 10 ..
##  $ zip_code               : Factor w/ 810 levels "007xx","010xx",..: 701 277 493 737 786 695 248 722
##  $ addr_state             : Factor w/ 49 levels "AK","AL","AR",..: 4 11 15 5 36 4 27 5 5 42 ...
##  $ dti                    : num  27.65 1 8.72 20 17.94 ...
##  $ delinq_2yrs            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line       : Factor w/ 526 levels "Apr 1964","Apr 1966",..: 192 34 426 161 203 429 254
```

```
##  $ inq_last_6mths         : int  1 5 2 1 0 3 1 2 2 0 ...
##  $ open_acc               : int  3 3 2 10 15 9 7 4 11 2 ...
##  $ pub_rec                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal              : int  13648 1687 2956 5598 27783 7963 17726 8221 5210 9279 ...
##  $ revol_util             : num  0.837 0.094 0.985 0.21 0.539 0.283 0.856 0.875 0.326 0.365 ...
##  $ total_acc              : int  9 4 10 37 38 12 11 4 13 3 ...
##  $ total_pymnt            : num  5863 1015 3006 12232 4067 ...
##  $ total_pymnt_inv        : num  5834 1015 3006 12232 4067 ...
##  $ total_rec_prncp        : num  5000 456 2400 10000 3000 ...
##  $ total_rec_int          : num  863 435 606 2215 1067 ...
##  $ total_rec_late_fee     : num  0 0 0 17 0 ...
##  $ recoveries             : num  0 123 0 0 0 ...
##  $ collection_recovery_fee: num  0 1.11 0 0 0 0 0 0 2.09 2.52 ...
##  $ last_pymnt_d           : Factor w/ 107 levels "Apr 2009","Apr 2010",..: 43 5 61 43 45 43 80 43 4 8
##  $ last_pymnt_amnt        : num  171.6 119.7 649.9 357.5 67.3 ...
##  $ last_credit_pull_d     : Factor w/ 108 levels "Apr 2009","Apr 2010",..: 9 99 9 8 46 37 108 26 99 1
##  $ pub_rec_bankruptcies   : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

- **Exhibit 2**

```
## Warning: Removed 1 rows containing missing values (geom_label).
```



- **Exhibit 3**

Logistic Regression:

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: loan_status
##                       LR Chisq Df Pr(>Chisq)
## term                    80.569  1  < 2.2e-16 ***
## int_rate                 5.746  1  0.0165250 *
## sub_grade               42.187 34  0.1581177
## emp_length              38.466 11  6.526e-05 ***
## annual_inc             117.498  1  < 2.2e-16 ***
## purpose                136.457 13  < 2.2e-16 ***
## dti                      1.937  1  0.1639816
## inq_last_6mths          57.759  1  2.963e-14 ***
## pub_rec                 14.586  1  0.0001339 ***
## revol_util              11.630  1  0.0006491 ***
## state                   74.722 38  0.0003464 ***
## earliest.cr.line.year   45.735 29  0.0249276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Exhibit 4**

Relaxed LASSO (lambda min):

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: loan_status
##                        LR Chisq Df Pr(>Chisq)
## term                     82.090  1  < 2.2e-16 ***
## int_rate                  5.915  1  0.0150161 *
## sub_grade                41.961 34  0.1639407
## emp_length               39.057 11  5.180e-05 ***
## home_ownership            4.184  3  0.2422802
## annual_inc              107.892  1  < 2.2e-16 ***
## purpose                 135.516 13  < 2.2e-16 ***
## dti                       1.952  1  0.1623702
## inq_last_6mths           58.501  1  2.032e-14 ***
## pub_rec                   3.195  1  0.0738823 .
## revol_util               11.194  1  0.0008205 ***
## pub_rec_bankruptcies      0.221  1  0.6385518
## state                    71.726 38  0.0007670 ***
## earliest.cr.line.year    46.479 29  0.0210178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Exhibit 5**

Relaxed LASSO (lambda first):

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: loan_status
##                       LR Chisq Df Pr(>Chisq)
## term                    86.077  1  < 2.2e-16 ***
## int_rate                 5.006  1  0.0252573 *
## sub_grade               41.623 34  0.1729164
## emp_length              44.482 11  5.983e-06 ***
## annual_inc             117.111  1  < 2.2e-16 ***
## purpose                134.296 13  < 2.2e-16 ***
## dti                      2.033  1  0.1539457
## inq_last_6mths          58.401  1  2.137e-14 ***
## pub_rec                  3.111  1  0.0777490 .
## revol_util              12.252  1  0.0004648 ***
## pub_rec_bankruptcies     0.393  1  0.5304780
## state                   75.984 38  0.0002459 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Exhibit 6**

Logistic Regression with Interactions

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: loan_status
##                         LR Chisq Df Pr(>Chisq)
## sub_grade                34.966 34  0.4219721
## emp_length               38.515 11  6.402e-05 ***
## annual_inc              118.420  1  < 2.2e-16 ***
## purpose                 133.221 13  < 2.2e-16 ***
## dti                       1.619  1  0.2032221
## term                     81.777  1  < 2.2e-16 ***
## inq_last_6mths           57.698  1  3.056e-14 ***
## pub_rec                  14.095  1  0.0001738 ***
## int_rate                  5.614  1  0.0178203 *
## revol_util               11.534  1  0.0006835 ***
## state                    75.306 38  0.0002958 ***
## earliest.cr.line.year    45.818 29  0.0244623 *
## emp_length:annual_inc    15.212 11  0.1729863
## int_rate:revol_util       8.932  1  0.0028016 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Exhibit 7**

Relaxed Best GLM

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

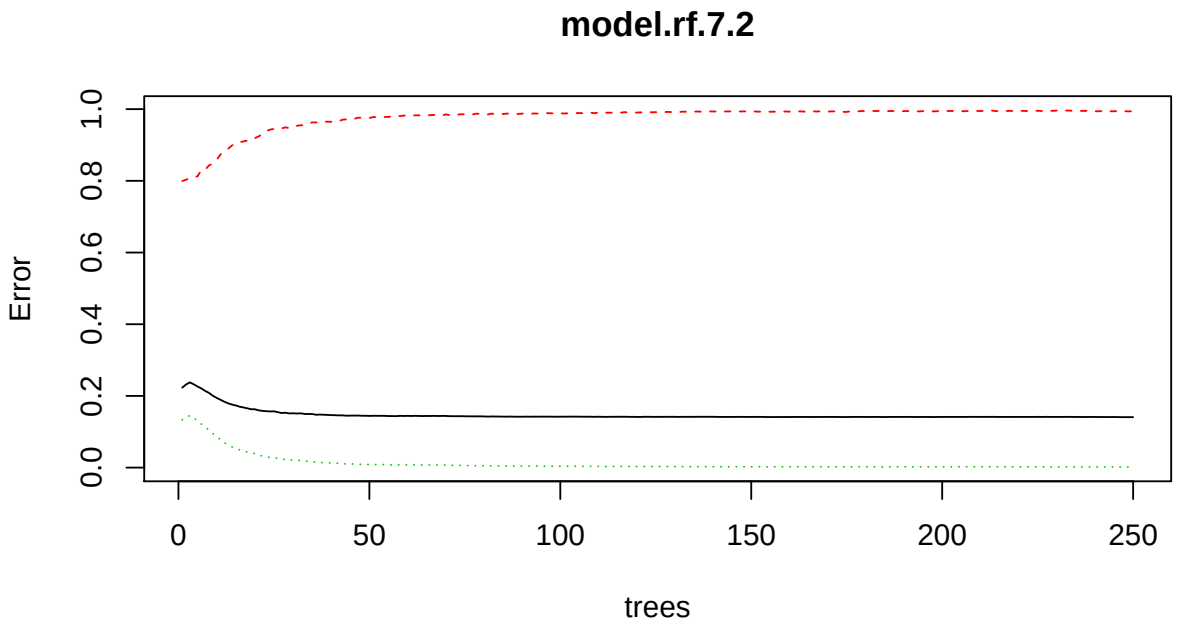
```
## Analysis of Deviance Table (Type II tests)
##
## Response: loan_status
##                LR Chisq Df Pr(>Chisq)
## term             80.916  1  < 2.2e-16 ***
## int_rate        233.380  1  < 2.2e-16 ***
## home_ownership    5.031  3   0.169525
## annual_inc      118.036  1  < 2.2e-16 ***
## dti               0.024  1   0.877914
## inq_last_6mths   60.413  1  7.689e-15 ***
## pub_rec          20.835  1  5.007e-06 ***
## revol_util        7.684  1   0.005571 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Exhibit 8**

Relaxed Best GLM. Significance Level 0.05 (Parsimonious Model)

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: loan_status
##                LR Chisq Df Pr(>Chisq)
## term             79.144  1  < 2.2e-16 ***
## int_rate        243.389  1  < 2.2e-16 ***
## annual_inc      141.386  1  < 2.2e-16 ***
## inq_last_6mths   58.889  1  1.668e-14 ***
## pub_rec          20.148  1  7.168e-06 ***
## revol_util        9.027  1    0.00266 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Exhibit 9**



Random Forest