Lending Club Risk Analysis

James Young

## Summary

## Using statistical modeling methods, a credit scoring system was devised that limited the “BAD” loan rate to approximately 1 in 16 in the top score bucket compared to 1 in 4 based on random investment or 1 in 7 if investment was limited to A grade loans. “BAD” is defined further below.

## Definition of The Problem

Lending Club is a website that allows crowd-sourced lending as an investment vehicle for interested parties and as a low-fee alternative to traditional loans for the loan seekers. Like traditional lending, there is an associated risk with funding these loans. This project aims to provide a model that produces a lending score for prospective loans, which, in turn, aims to mitigate risk better than a randomized investing model.

## Description of The Data

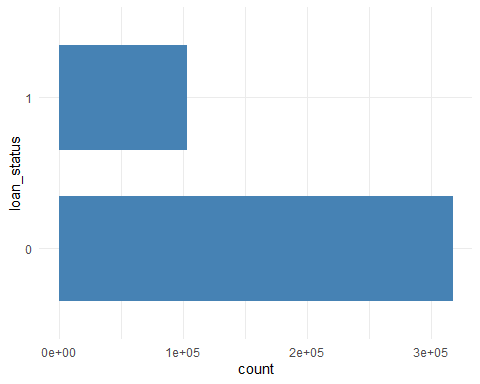
We have a dataset composed of roughly 1.3 million loan cases where each case is a row in the dataset. Each case has 110 variables in the original dataset, however many of these columns contain a lot of *NAs*, which will cause problems later for building our predictive models. We will start the data cleaning by removing any column that has more than 10% of rows filled with NA. In the future I may try making dummy variables for these columns.

Next, I removed any data that we would not have privelage to at the time of investment, such as total received interest and other payment information. This information could only be known after the loan was in progress. This removes another 11 columns and leaves us with 77 columns now, which are either numerical (54 columns) or factorial (22 columns), although that doesn’t tell the whole story. We have numerical data that describe *Dates* such as “earliest\_credit”, multiple columns describing money amounts, and credit related scores. We have factors describing varibales such as State, Zip Code, and even various descriptions of what the loan is for. Since this model is for a peer lending site, the model does not need to be limited in variables for the sake of simplicity. If this were a more regulated situation, I would use a step function to make the model optimized for simplicity and predictive power (determined by AIC).

## Define “Good” and “Bad”

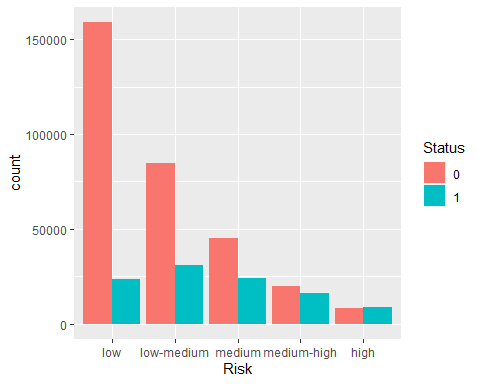
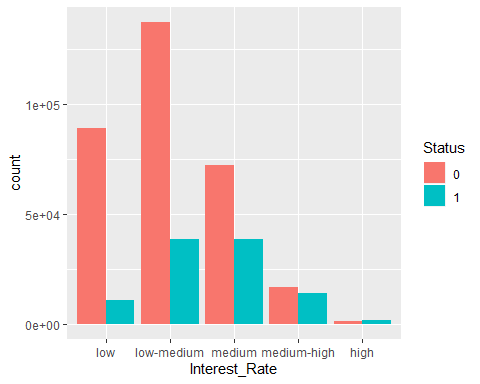
On the next page we can see the count of how many individuals we have data on in each of the loan\_status categories. We want to build a model that predicts whether an individual will pay back their loan or not. This means we only need to keep the *GOOD* “Fully Paid” group and the multiple groups I will consider *BAD* including “Default”, “Late (31-120 days)”, “Late (16-30 days)”, “Charged Off”, and “Does not meet the credit policy. Status: Charged Off”. The groups “Current”, “In Grace Period”, and “Issued” will be removed since they have not completed their term or earned the “*BAD”* label. Future work will factor profitability into “GOOD” or “BAD” labels

**The Model**

We will make a model that stringently tries to predict the *GOOD* customers over all the bad categories. We will start by merging the *BAD* categories into one group which will receive the number *1*. The fully paid *GOOD* group will receive the number *0*. We will use these numbers to represent *GOOD* and *BAD* to our predictive models. However, before we make models, we will do additional exploratory data analysis. Here we can see there are approximately 100,000 *BAD* accounts and 300,000 *GOOD* accounts. Randomly funding a loan in this case would have an approximate 1 in 4 theoretical chance of being *BAD*. Seen another way, you have an 75% chance of randomly funding a *GOOD* loan, which we will use as the benchmark to beat with our models.

**Exploratory Data Analysis**

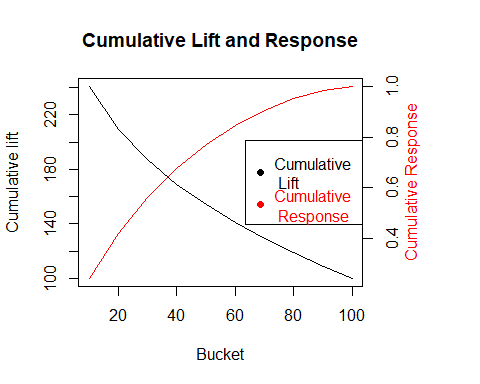
The most striking data feature that are easily visualized are the trend of loan rating vs “BAD” rate and interest rate vs “BAD” rate. If a lender only invested in “low risk” grade A loans, they would have a “BAD” rate of approximately 1 in 7, almost twice as good as random investing. (See next page).



**Statistical Modeling Results**

While trying various models, we achieved a range of predictive abilities (AUC as metric). All model’s showed respectable generalization between train and test sets (80/20 train/test split of approximately 440,000 observations). Below, to the right, we can see the cumulative lift and response of the ensemble model. Approximately 70% of the “GOOD” loans can be captured in the top 40% loans (counting from low probability to high probability for “BAD” loan). Further below, the gains table presents the exact percentages of positive response captured at each 10% interval (again, counting from low probability to high probability for “BAD” loan). The ensemble method had slightly fewer observations (~10,000) due to the model cutting out observations that previous models gave a numerical probability that was rounded to 0. This will be adjusted for in the future.

|  |  |  |
| --- | --- | --- |
| Model | Train AUC | Test AUC |
| Logistic | 0.724 | 0.729 |
| MARS | 0.728 | 0.732 |
| GAM | 0.704 | 0.711 |
| XGBoost | 0.75 | 0.746 |
| Ensemble | 0.754 | 0.748 |



Depth Cume Cume Pct Mean

of Cume Mean Mean of Total Lift Cume Model

File N N Resp Resp Resp Index Lift Score

-------------------------------------------------------------------------

10 9184 9184 0.60 0.60 25.3% 244 244 0.59

20 8542 17726 0.43 0.52 42.4% 177 212 0.44

30 8808 26534 0.36 0.46 56.9% 145 190 0.35

40 8814 35348 0.28 0.42 68.1% 112 170 0.28

50 8872 44220 0.23 0.38 77.5% 93 155 0.23

60 8855 53075 0.18 0.35 85.1% 76 142 0.18

70 8892 61967 0.14 0.32 91.0% 59 130 0.14

80 8729 70696 0.10 0.29 95.2% 42 119 0.11

90 8831 79527 0.08 0.27 98.3% 31 109 0.08

100 8835 88362 0.04 0.24 100.0% 17 100 0.04

**Credit Scoring Model**

The credit score is created by taking the model output (probability of “BAD” status) and subtracting it from 1 then multiply by 100, creating a score between 0 and 100. Below, to the left, we can see the top score bucket (90-100) has approximately a 1 in 16 chance of ending up “BAD”. We can also see this in the confusion matrix (below, right), which shows more details of the prediction breakdown when the decision threshold is set at 0.1 probability for “BAD”.

|  |  |  |
| --- | --- | --- |
| Prob. thresh < 0.1 | Actual Good | Actual Bad |
| Predicted Good | 17819 | 1217 |
| Predicted Bad | 48927 | 20399 |

