Loan Analysis

Matt Rock

4/25/2020

Section 1: Executive Summary

This report was commissioned to adjust how the bank approaches issuing loans to consumers to reduce default rates and increase the financial stability of the bank's loan portfolio.

When reviewing the provided data, it became clear that loans only become profitable after the debtor has made the most of the payments. But, a debtor may default on a loan early on. The loss from one bad loan can eat the profits from five successful ones. Therefore, taking on more loans and hoping they will be successful is the wrong approach. The attached report shows that by being more selective when approving loans, the amount of loans that default will be cut in half. Compared to the old methodology, the new block of business will show a profitability increase of 242%.

This report investigates what elements of the old method were not indicative of loan quality. For example, job title and length of employment were not useful indicators and are not included in the new model. The reason for a loan is now the yes or no question "Is this to consolidate credit card debt?" Many of the questions around current credit balances and limits had information that was redundant, and many were removed.

Two methods were used to test the model. First, a random portion of the data was held back to validate the testing, and the results for both portions matched closely. Second, there was another model to test against - the model the bank initially used to build the block of business. While there was an ability to make a more accurate model than the original, "more accurate" is misleading when one bad loan can undo the work of five successful ones.

There are several additional downstream benefits to the new method. Much of the information needed about an applicant can be sourced from a credit check. This means targeted marketing efforts can help draw in customers that would pay back a loan before they even know they want to apply. Loan officers will certainly appreciate needing less information to complete a loan application, especially when they are confident the requested data is relevant.

Finally, current account holders that have one or two issues preventing them from being acceptable risks can be contacted by a bank representative and offered financial counseling on how to improve their financial status. The reduction in loan approvals could be changed by building a better pipeline to quality loan applications.

Section 2: Introduction

In the aftermath of the 2008 recession, the question everyone asked the banking industry was "How did you miss this financial collapse? You had all the information, but ignored it." The pressure to create more mortgage-backed tranches meant housing lenders were actually competing to lend to individuals with riskier credit, leading to a recession that still continues to have repercussions.

Now, with the specter of financial ruin still lingering, banks need to protect themselves from bad risks of all kinds. The purpose of this project is to assist banks in being able to protect their investment portfolio with safe investments, while still creating a satisfying application experience for the loan officer and loan seeker.

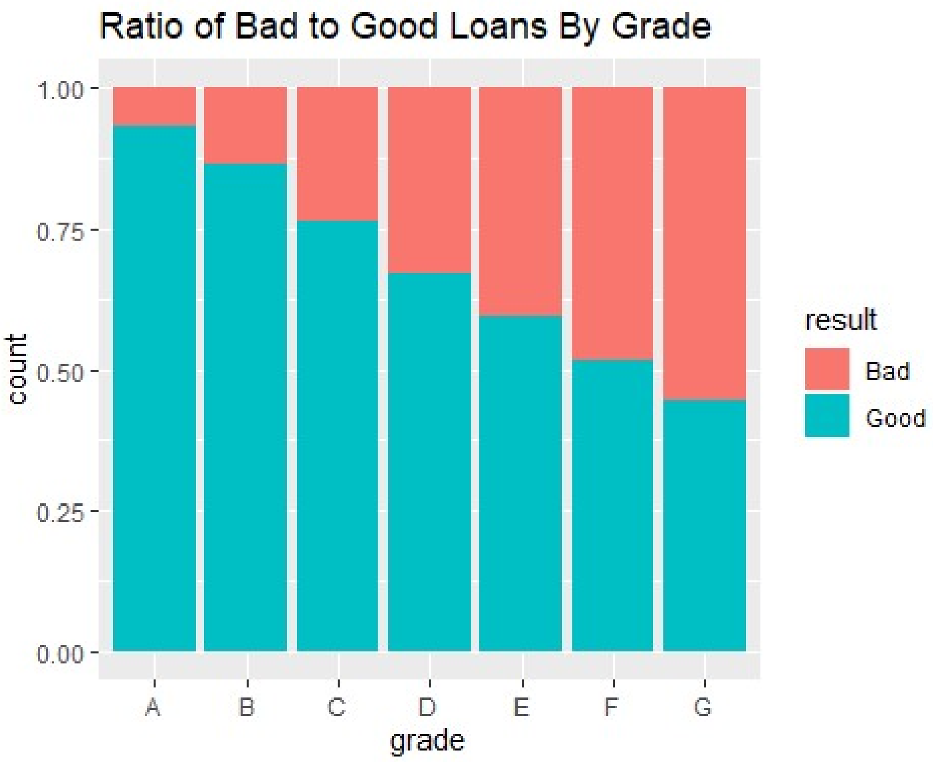
Section 3: Preparing and Cleaning the Data

The initial data set had 50,000 observations and 32 variables. Selecting only on the relevant loan status left 34,655 data points. The initial variable trimming removed loanID (randomized number), employment (covered by job length and salary), and state (with 50 states, random chance would lead to false positives). The interest rate of the loan, and therefore the loan payment, comes from the loan officer, so neither would be useful to help the bank project if a loan will fail.

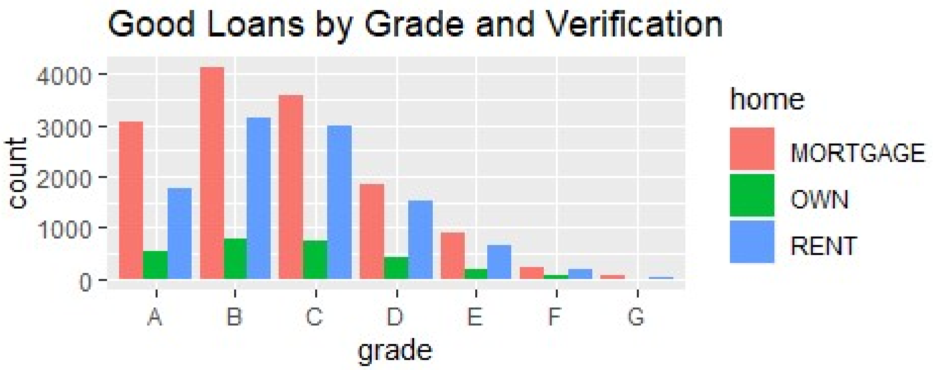
Finally, many of the variables were very similar. Total accounts is removed, as open accounts would work just as well. Likewise, credit card limits and installment limits fell under total limits. Total balance shows financial problems more than average balance. Finally, the ratio of used credit to total credit and total open credit are heavily correlated with the total credit used and total credit limits.

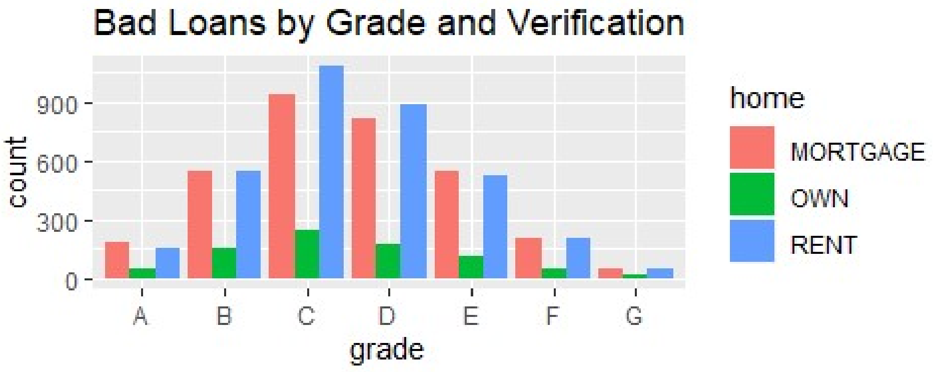
Section 4: Exploring the Data

The grade of the loan is the biggest indicator of its success. However, since the loan grade is the result of the loan process, rather than information around the application, it doesn't make sense to use it as part of a model predicting if a loan is successful. Where the grade will be useful is benchmarking what variables indicate a good or bad loan.

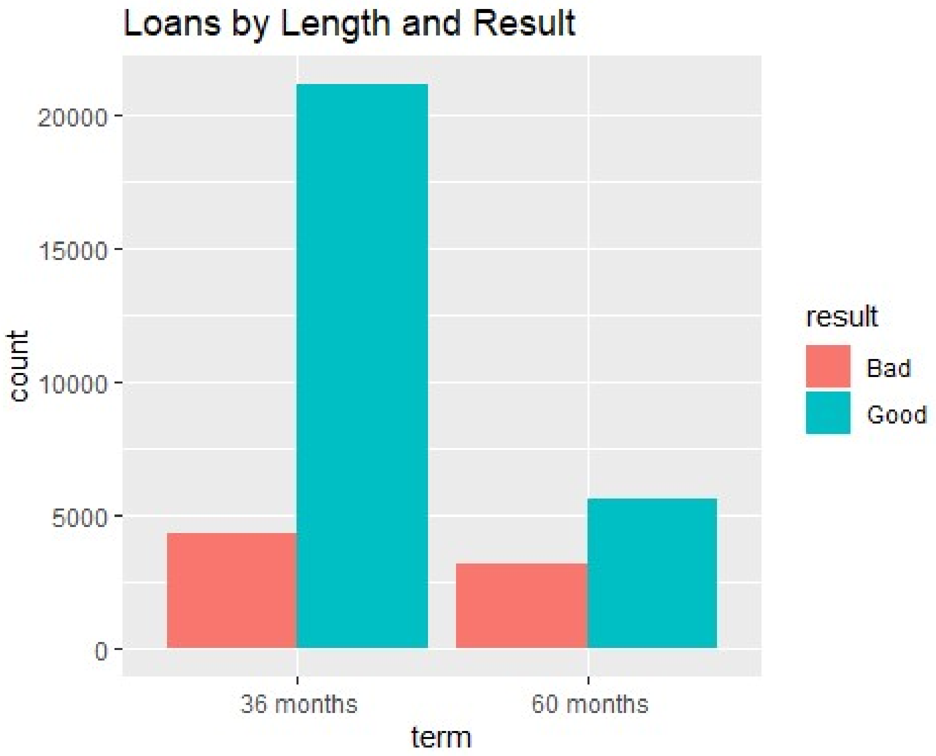


Loans to people with a mortgage were more successful than those to renters. About one out of every five loans given to someone with a mortgage failed, compared to renters failing with one of four loans. This was heavily influenced by the grade of the loan

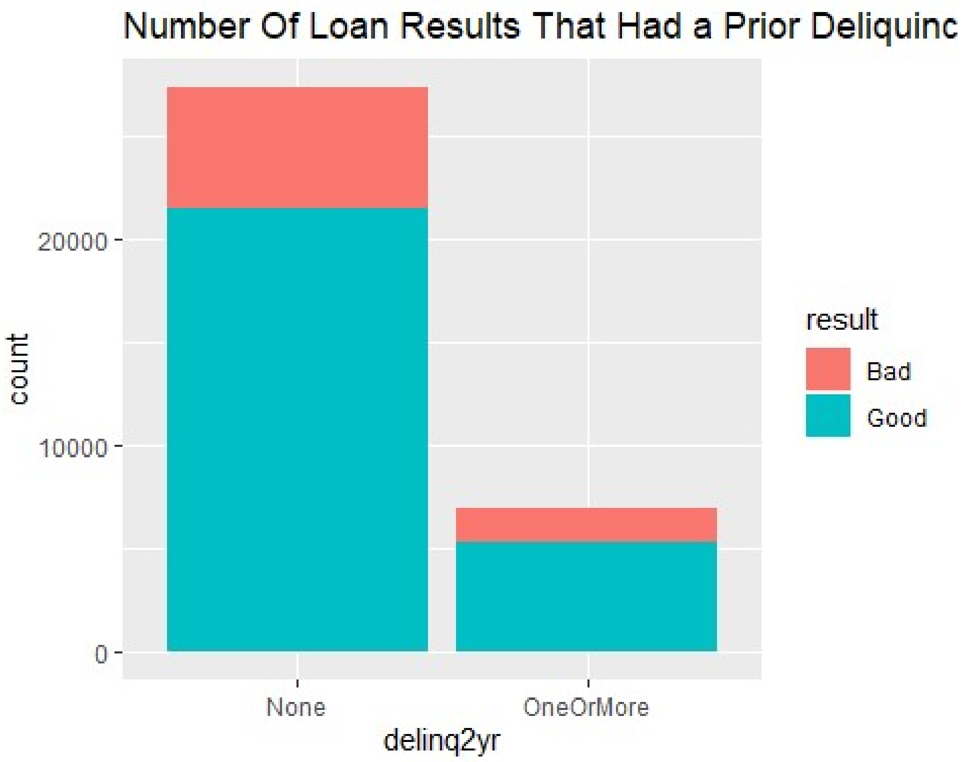




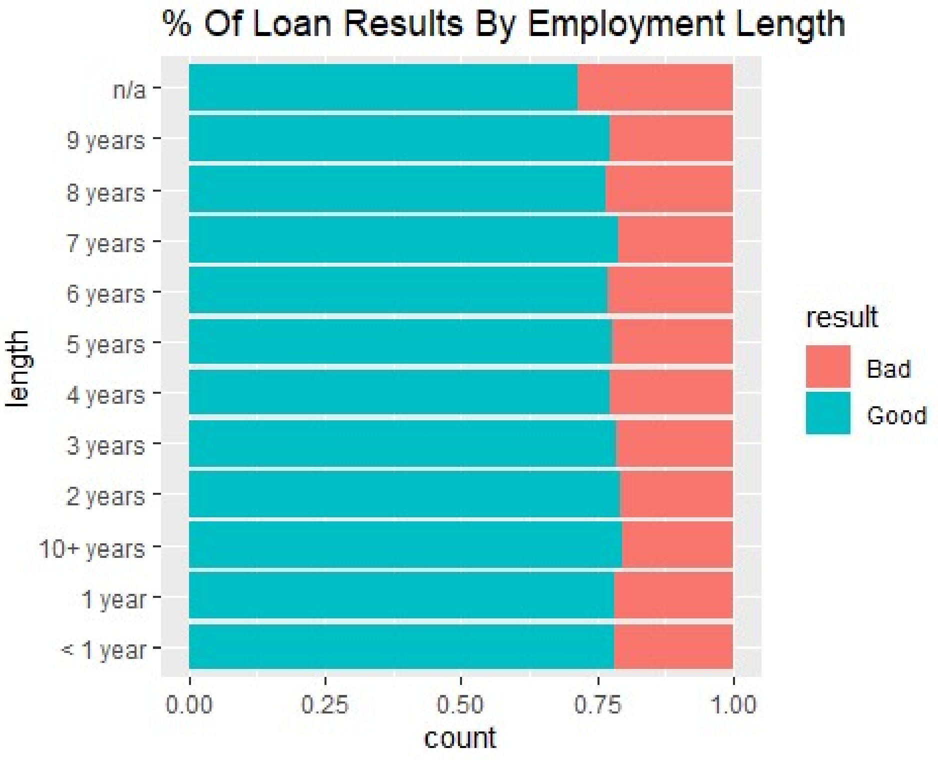
Similarly, the length of the loan was a very strong indicator of success or failure. Around one of six loans with a 3 year term failed, but over a third of 5 year loans failed.



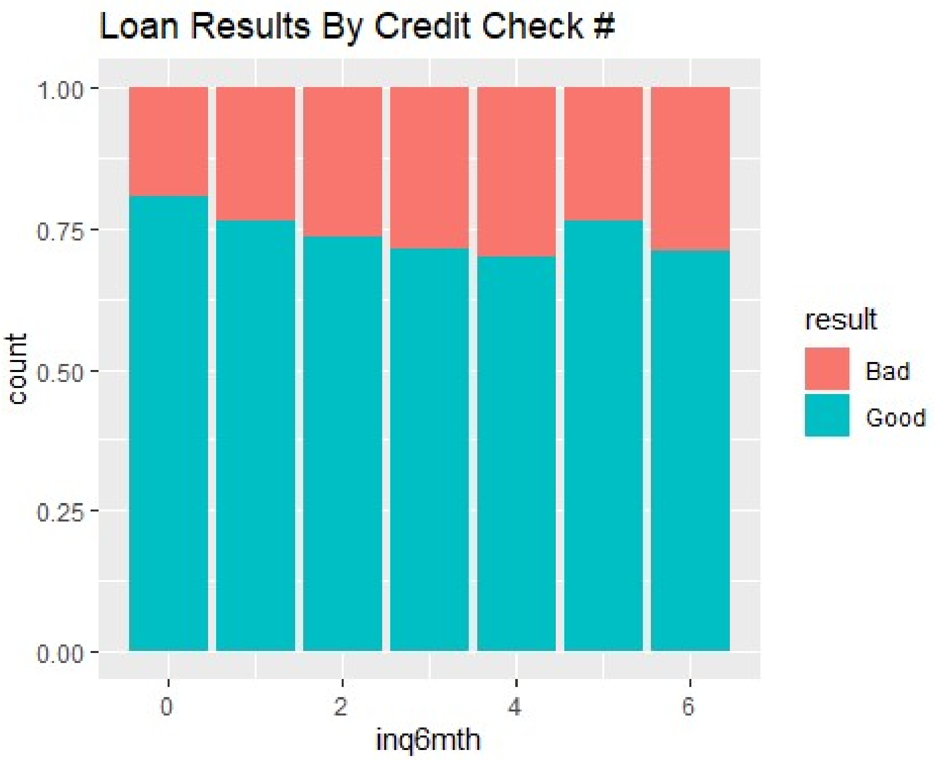
Some variables involved loan approvals more than loan results and were removed. The majority of loans were given to people with no late payments in the last 2 years, but a loan was more likely to default if it was issued to someone without a late payment than with. Length of employment, public record hits, and 6 month inquiries also had no real clear bearing on loan quality.

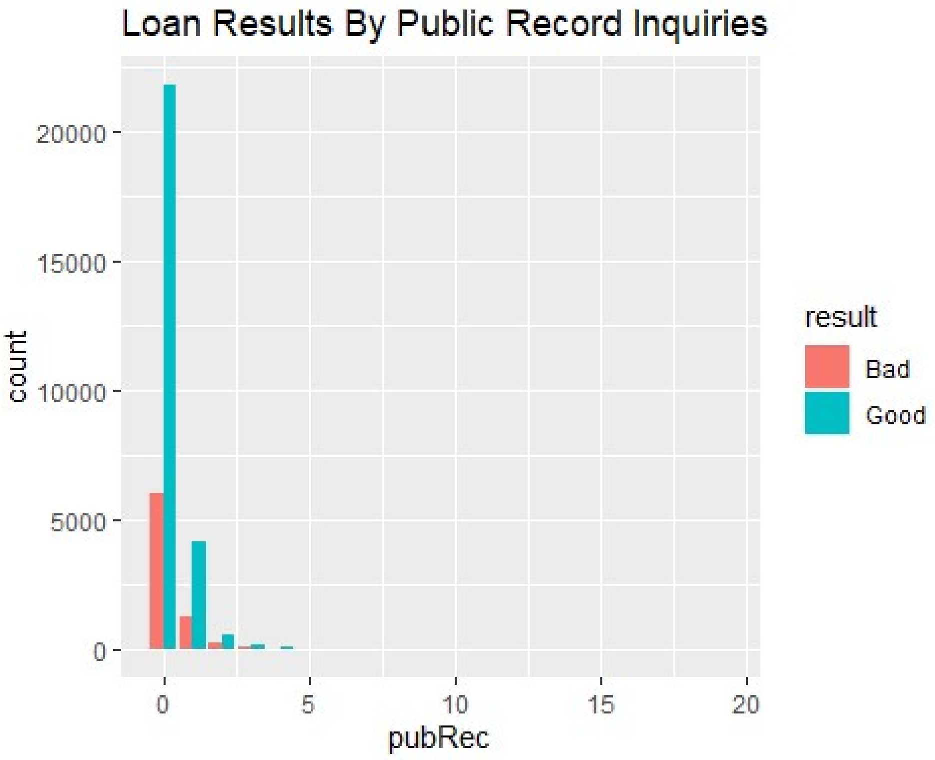


Number Of Loan Results That Had a Prior Deliquinc

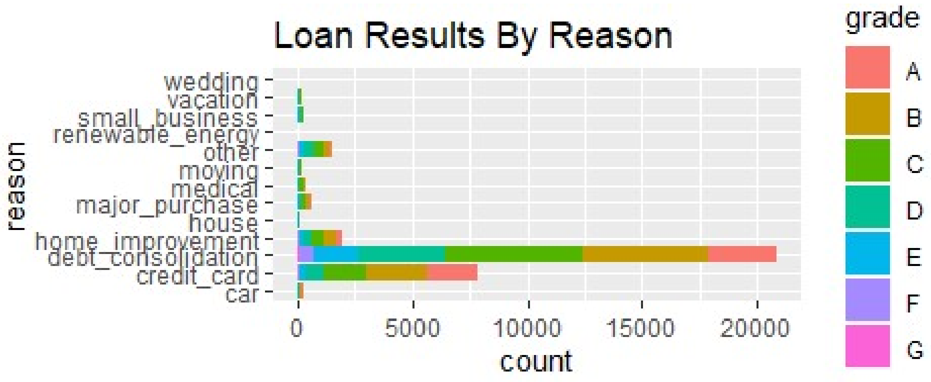


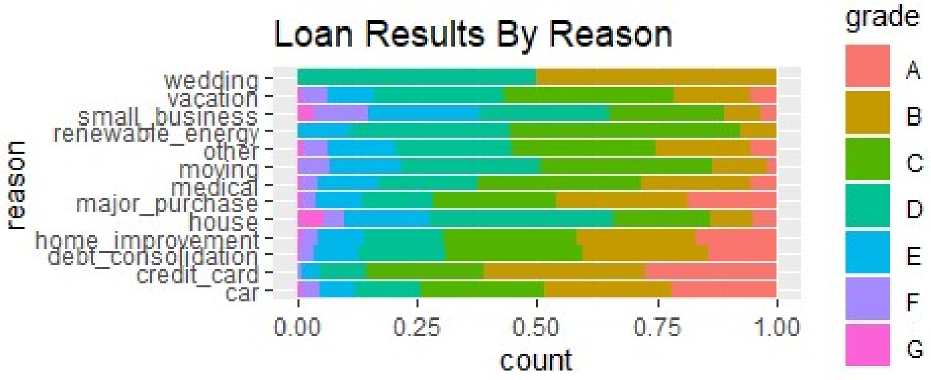
% Of Loan Results By Employment Length





Looking at the reasons for loans, credit card not only had enough data points to be useful, it also was strongly associated with higher grades. I turned the reason into a binary indicator variable - 1 to indicate it was a credit card-associated loan, 0 otherwise.

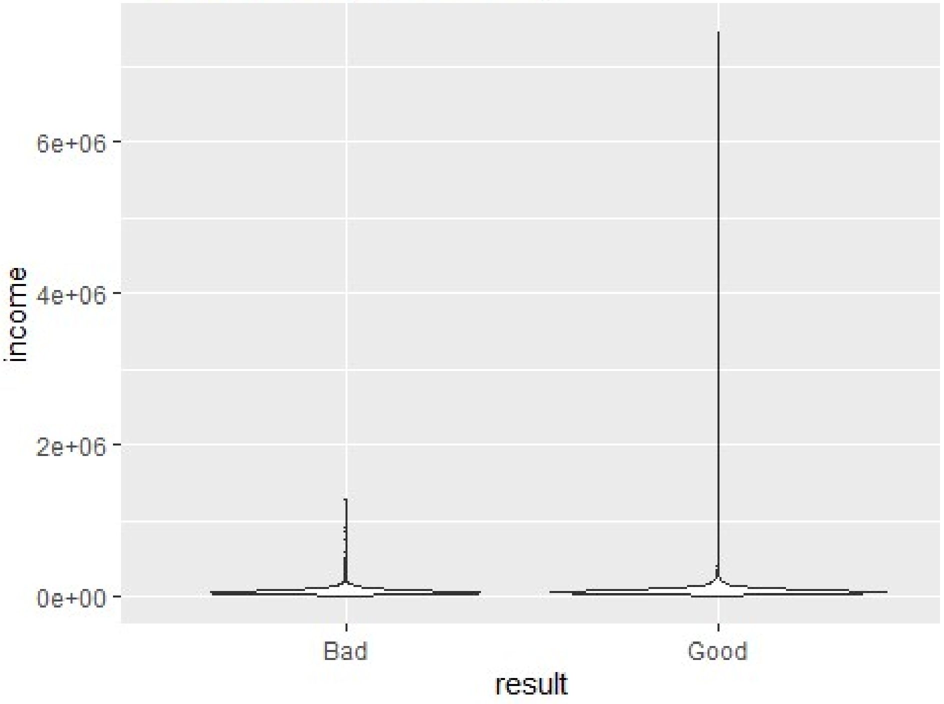




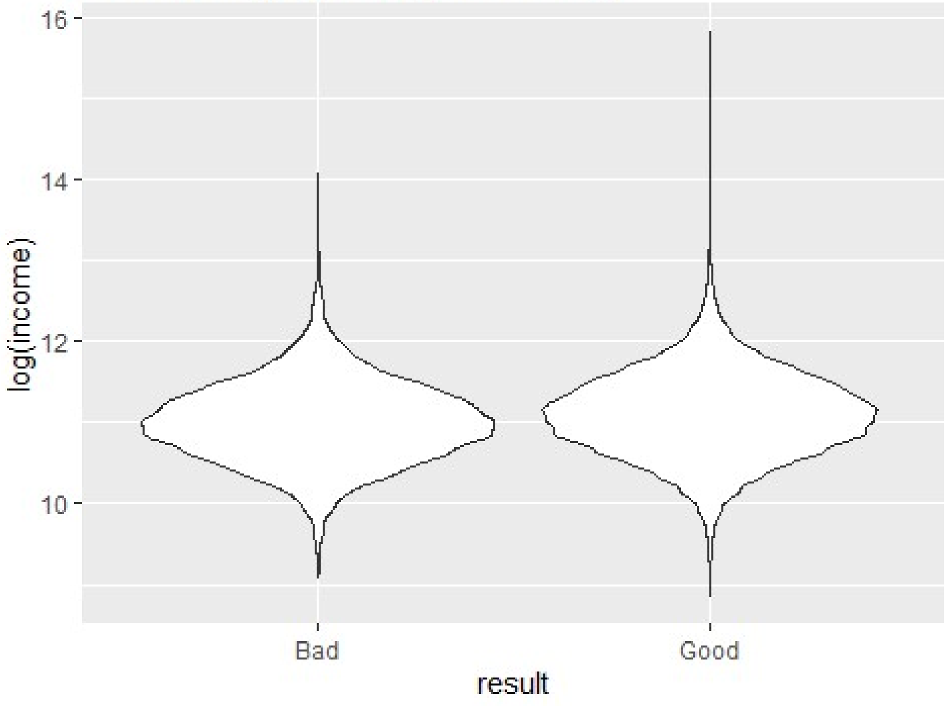
The variables that involved total income, debt or credit levels were heavily skewed. Logarithmically transforming those variables helped reduce skewness. There were still a few outliers, but the transformation helped normalize the data significantly.

Loan amounts were not altered, as they did not span nearly the same range as the transformed variables.

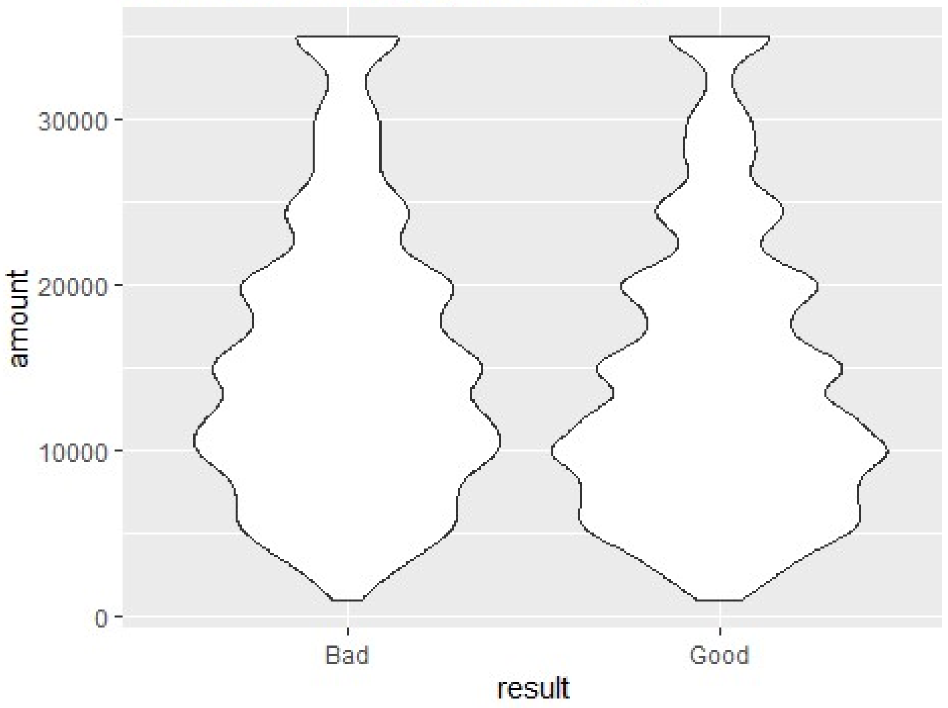
Income Displayed Linearlly



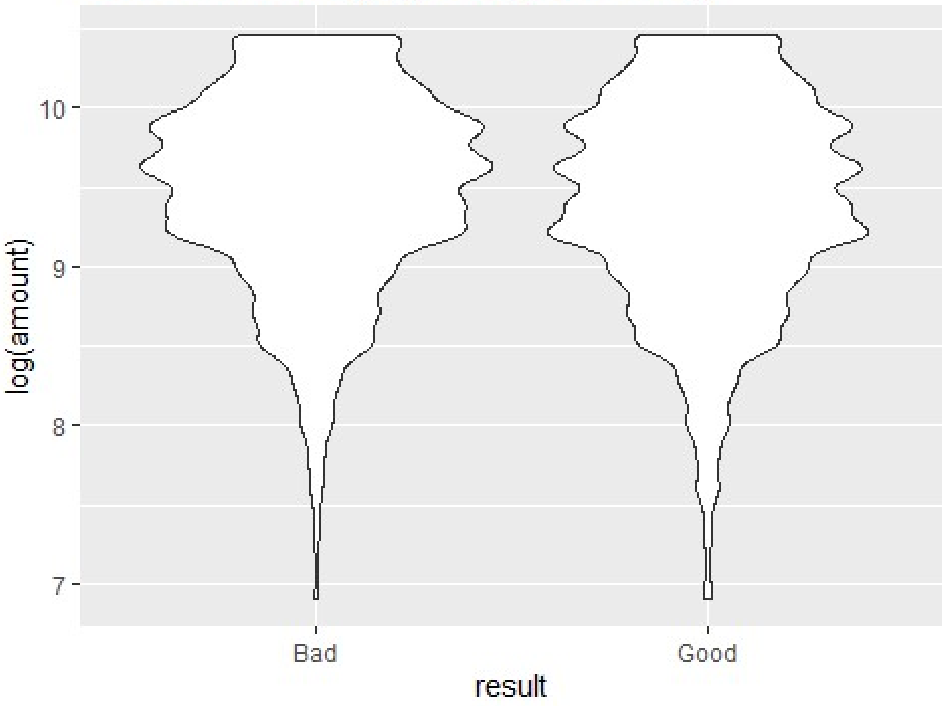
Income Displayed Logarithmically



Loan Amount Displayed Linearlly



Loan Amount Displayed Logarithmically



Section 5: Building The Model

The data was split into two sets. 80% of the loans were used to build the models, and the grade variable was dropped from them. The remaining 20% would be held to verify the model's accuracy.

The initial pass of using only first-order variables to build the model looks like the current selection was very good. Only open accounts had a p-value above .1 and was removed.

A model using higher-order variables was considered, but it ended up with a slightly worse success rate as the first-order model. More complicated and worse made it easy to discard.

```
## [1] "First-order model"

##                     trueResponses
## estimatedResponses FALSE  TRUE
##              FALSE   574   424
##              TRUE   5429 20989

## [1] 0.7865115

## [1] "Higher order model"

##                     trueResponses
## estimatedResponses FALSE  TRUE
##              FALSE   633   491
##              TRUE   5370 20922

## [1] 0.7862197
```

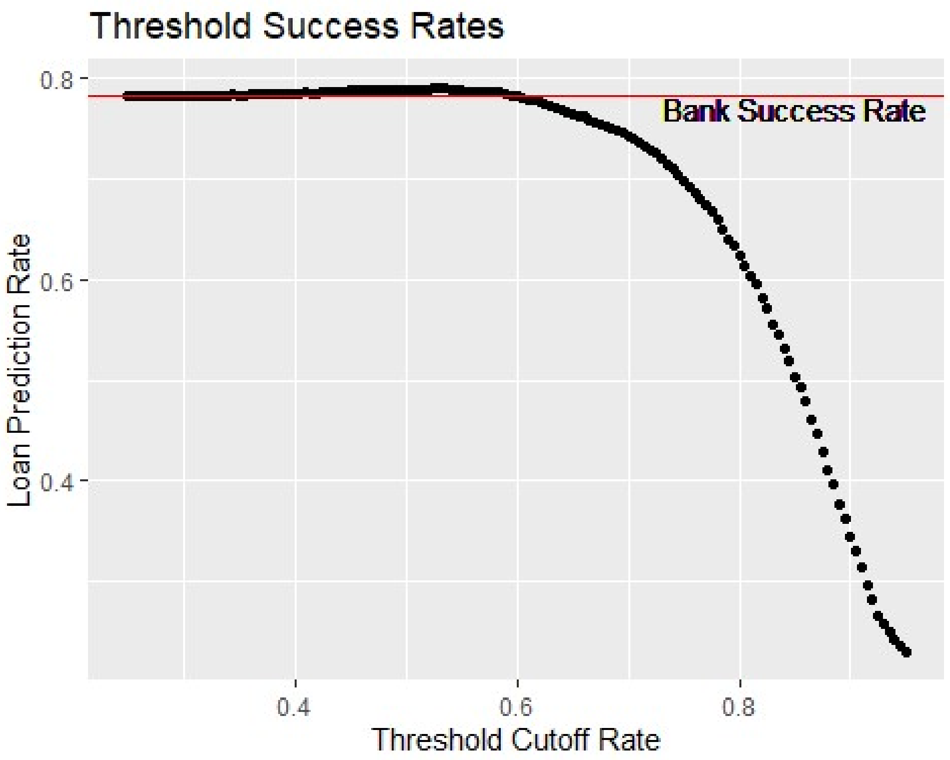
Now to use the testing dataset to see how we did.

```
##                     trueResponses
## estimatedResponses FALSE TRUE
##              FALSE   153  106
##              TRUE   1348 5247
```

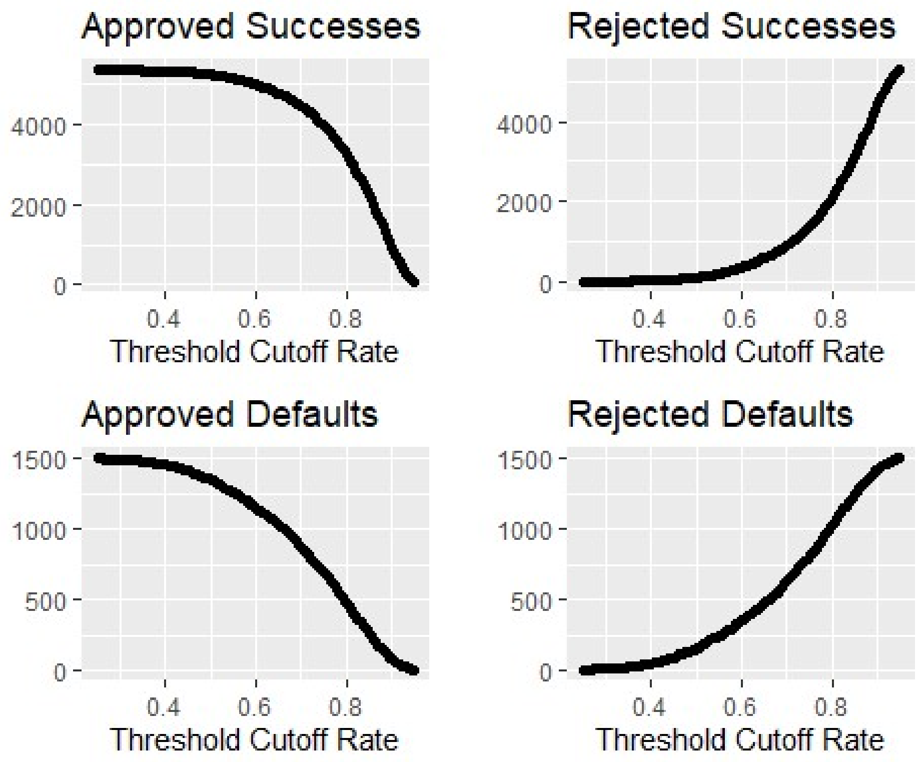
We have a success rate in this data set very close to the training set. Let's take it one step further. How can we see if our roughly 78.8% success rate at identifying if a loan will succeed or not is actually worthwhile? Compare the model to the bank's success rate of 78.1%, we're slightly ahead of them. Let's see if we can improve.

Section 6: Optimizing the Threshold for Accuracy

What's the optimal threshold? Let's try other thresholds to see if we can get more accurate results.



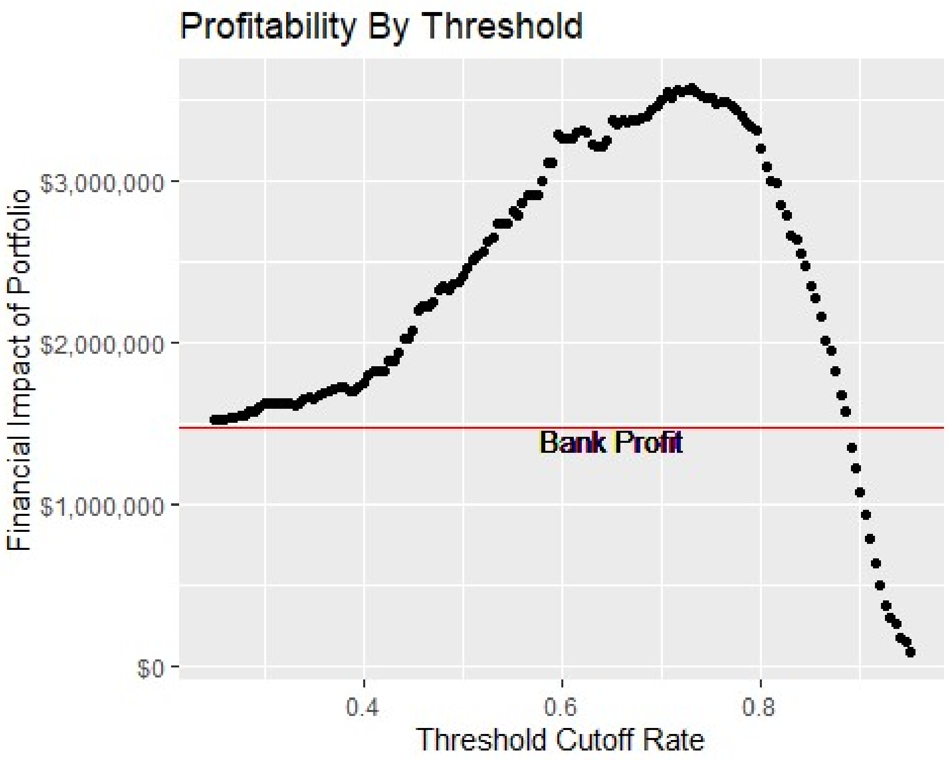
Instead of using .5, the maximum accuracy of 79% results from using a threshold of 0.535. There wasn't much difference in prediction quality until around the .65 threshold, when it began to decline slowly, then rapidly. Looking at each loan's prediction and result helps see that trend happening.



While "being right" is a great reason to build a model, "making money" is another one. How does changing the threshold alter how much money the bank makes?

Section 7: Optimizing the Threshold for Profit

Using the same general idea as maximizing accuracy, but instead of pass/fail I'll focus on how profitable the loan ends up being. We know that the bank profit of loans that are either fully paid or written off is $1,474,366. If our model was perfect, the maximum possible profit in our sample data is $12,523,305. What's the best the model can do?



Here, the graph peak is at the threshold of 0.73, with a profit of $3,571,514. Interestingly, peak profitability comes with the model's threshold set at a point where it loses accuracy. We can see why if we break the loans into chunks based on the value the model assigned to each one.

Effect of Threshold Changes on Profits

From	To	LoanCount	SuccessfulLoanCount	DefaultLoanCount	Profits
0.000	0.535	387	164	223	$-1,260,889.90
0.535	0.730	1515	995	520	$-836,258.50
0.730	1.000	4952	4194	758	$3,571,514.06

The poorly ranked loans didn't just fail, they exploded. The loans rated between .535, our most accurate setting, and .730 included 1515 loans that cost the bank $-836,258. While there are some successful loans the model is rejecting, by sticking to the highest-ranked loans we can significantly increase profitability. Most importantly, by reducing the number of loans that default from 1501 to 758 profitability increases 242%.

Section 8: Results Summary

I had three goals for creating a method to improve the performance of personal loans: speed, accuracy, and profitability. How did each turn out?

Speed

The amount of information needed to complete a loan application shrunk dramatically. Gone are questions about occupation, or public record queries. Loan officers can do more and better work, and loan seekers feel less stress around a stressful situation.

Accuracy

While it’s possible to use the model to predict loan success at a higher rate than a bank, it’s not prudent. The recommended threshold of $3,571,514 performs under the bank’s success rate of 78.1%, but with good reason.

Profit

That reason is being accurate is a $-836,258 drain on the balance sheet. For any two fully paid loans that returns a slight profit, those proceeds are eaten up by one loan that went so sour it’s a drain on the financial books.