It Is All About Who: A Data Analysis on Loan Quality and Default

STAT 471 (Modern Data Mining) Mini Project

David Fan

I. Abstract

Lending Club's legacy

An innovator in the long held institutional based lending process, a global leader in the space of peer-to-peer monetary platform, and an enabler to low-cost loans and alternative investment options, Lending Club is undoubtedly a leader in its field today. Over its now 10 years of existence, Lending Club has helped millions of individuals across various purposes (from business development to personal finances) to achieve their dreams and goals. Its commitment to excellence and customer experience has anchored its current position and is helping them set sail for future success.

The Challenge with Growth

The heightened demand comes with its unique challenges. Lending Club, being a fast growing and scaling company faces the prominent problem of balancing its reach (continually serving a massive span of individual) while maintaining the quality of its product. It is paramount to Lending Club to ensure the quality of loans in its portfolio without comprising and denying too much services. **In this paper I will want to take the lense of as a Lending Club employee or executives trying to optimize for its loan selection.** This paper seeks to examine this particular area with four main components:

1. **To Understand The Business**
   The first section of this analysis-driven paper will focus on scoping out the main driver of Lending Club's success and unprecedented growth, while also outlining the potential issue that the business currently faces. I seek to use this section to illustrate the importance and main value proposition of this study. In particular, under this section I will be looking into the three main areas that constituted the success of lending club (the quantity of loans, the retention of borrowers and the average amount per loan) and also an area of potential improvement (stagnant loan quality).

2. **To Explore Potential Risk Factors**
   The second part of this paper will introduce the scope of the analysis, the features involved and engineered, posing hypotheses and explore related variables that may contribute to the default rates (such as interest rate, geographic locations, words in the description). This section will also examine the appropriate risk ratio, or in other words the appropriate weight given to different kind of errors (False Positive and False Negative), to be employed in the evaluation of our final model.

3. **To Build a Prediction**
   The third part of this paper will employ standard modelling techniques ranging from logistic regression to elastic net in order to select the best model that outputs the most accurate prediction. A base benchmark model will be employed and hyperparameter tuning will be done. Analysis will also be made and drawn from the model output to verify previously held hypothesis from Part 2.

4. **To Conclude and Recommend**
   The last part of this paper will focus on the assessment of the potential impact if the model were to be employed. It will also offer possible recommendations to Lending Club in how it may use the analyses derived to prevent loan default. Furthermore, I will also address the limitations of the paper and room for various improvements

The data employed in this analysis is the openly available lending club data on Kaggle. It contains relevant fields of information (to be further explained later on) on all approved Lending Club loans from **2007 -2011.**

II. The Story So Far

The Drivers of Success

I will begin this paper with an analysis as to why lending club has been so successful in the past decade. Specifically, I believe that it can be broken down to its success in 3 main areas: Loan Volume, Borrower Retention, and Average Loan Value.

Loan Volume

Loan Volume, or better stated, total loan applications, is quite hard to track. However I think it is reasonable to use a substitute, that is the number of approved loans, as a measure of total loan volume. This measure of loan volume is essentially an indirect way of measuring the demand for Lending Club's service. As a lending agency, the amount of loans directly correlate to the amount of revenue that Lending Club can generate as an intermediary, but more so serves as a clear indicator of the market position of Lending Club in terms of shares and popularity.

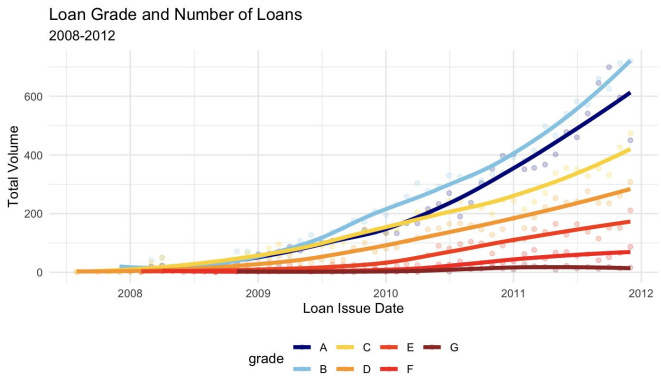


figure3: Lending Club Analysis

As seen from the graphics above Lending Club has evidently had success with growing amounts of loans through its platform. In fact, from 2007 to 2011, approved loans under Lending Club increased by about 5000 yearly, and in 2011, the total number of approved loans have grown by about 10000. Beyond just that, the graph also indicates an overall increase of loans across different grades with a stronger swing towards high quality loans (grades A and B).This is a strong indicator of a growing market demand for Lending Club's product, especially from higher quality borrowers.

I believe this is not only a result of the natural product market fit of Lending Club as a platform but more so a testament to the successful marketing the Lending Club team has launched around that period.

Borrower Retention

Another crucial piece that defines the success of an intermediary such as Lending Club is its customer loyalty. While loans are not something a person or an agency would get on a day-to-day basis, there are quite a number of organizations that requires infiltration of capital and debt to function.

Lending Club has been excellent in this regard as well. In particular, it has established strong ties and relationships with huge organizations in the United States with histories of success and reliability. Some of these notable partners include (within parentheses is the number of transactions they've had with Lending Club from 2007-2011):

- **US Army** (133)
- **Bank of America** (108)
- **IBM** (67)
- Kaiser, UPS, DoD, JP Morgan and etc.

As seen from the graphics below, these transactions have also been extremely strong and rapidly growing between 2008 and 2011 - clearly signaling the growing business-client relationship that Lending Club has cultivated. This form of "dependencies" creates a sustainable source of income that is absolutely beneficial for Lending Club

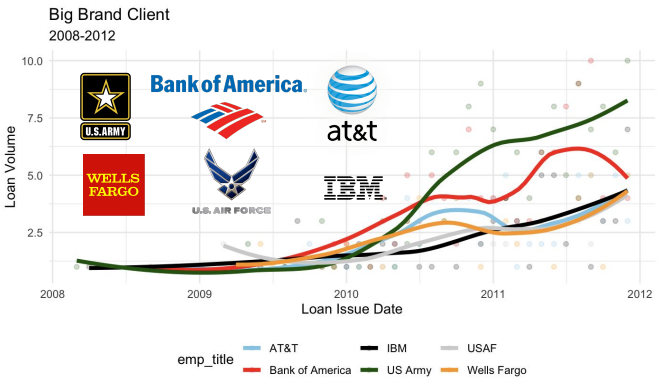


figure5: Lending Club Analysis

II. The Story Continued

Average Loan Value

Apart from the quantity of transactions that the past two factors have been more concerned with, an equally important aspect to examine is the per loan value that Lending Club currently houses. Similar to the first factor, the higher the loan value, the greater the principal payment/profit that Lending Club can obtain.

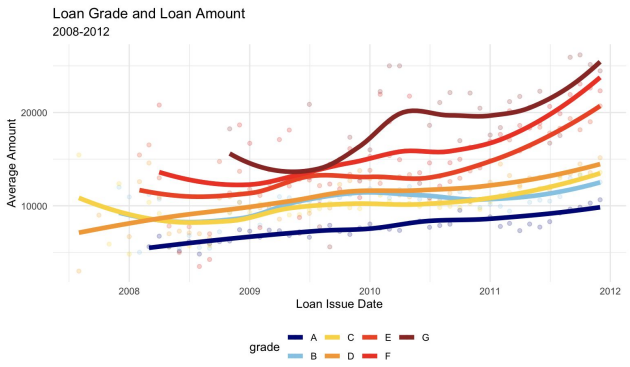


figure1: Lending Club Analysis

The graphics shows a steady trend of increasing loan values across time for all loan grades. This sends a strong signal - as long as Lending Club can control the default rate, it should be well positioned for more profit.

The Missing Piece

The past three factors indicated a strong and healthy market condition well suited for Lending Club to push its product. It is easily foreseeable that Lending Club will attract even more lenders and borrowers in the market and certainly more applications for loans. However, an increase in those applications not only provides an opportunity but also poses logistical and strategic problems of which loans to approve and reject.

In the period between 2007-2011, the default rate of LC is about 13%, equivalent to about 1 in 7.5 loans defaulting. From the historical data, if a loan is defaulted/charged off, we could usually only salvage about **60%** of the value, while the rest is unrecoverable. A good loan on the other hand could potentially generate us an income that is about 6-30% of the original loan amount.

This problem is of even greater importance now, since as loan value increases, payoff increases, as well as potential amount lost. This means with the right classification that helps us target the good loans (unlikely to default) and avoid the bad ones (likely to default) we could increase our income quite significantly.

As opposed to times when there were only small quantities of loans and thorough due diligence could be done on every application, the increase in loan volume means that Lending Club has a to seek a more automated approach that can maintain quality loans and reject those that are likely going to bring losses to the company.

Looking at the graph below, we can see that the default rate has been quite constant between 2007-2011. However, there seems to be a small spike towards the end of 2011, especially among lower grade loans.

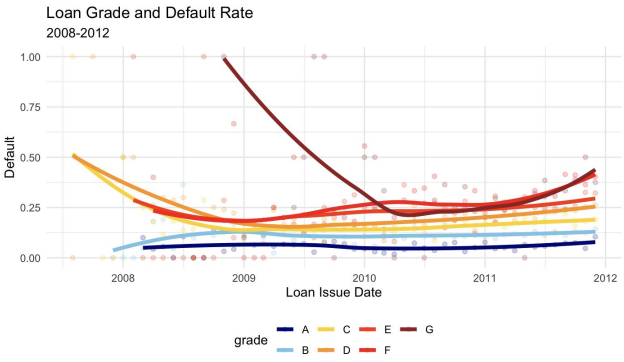


figure4: Lending Club Analysis

This trend should not be taken lightly, but instead needs to be enacted upon as soon as possible. The rest of the paper will primarily focus on understanding various factors related to default rate and devising the right decision criteria to maximize the potential payoff for Lending Club.

III. EDA

Default and The Factors

I will be diving into the aforementioned dataset further to summarize the relevant fields that may affect our target variable, the default rate.

The Data

The dataset has generally 4 buckets of feature:

- **Pre-funded loan data:** this bucket contains details about the loan, including the amount, the grade and other relevant information.
- **Borrower information:** this bucket includes information pertaining to the loan applicant such as job title, home ownership, and income.
- **Borrower credit data:** this bucket includes the past credit history of the borrowers. It includes his/her credit history, records of delinquency and the like.
- **Post loan data:** this includes information after the approval of the loan, such as our target variable (loan status) and also other information such as total payment to date and etc.

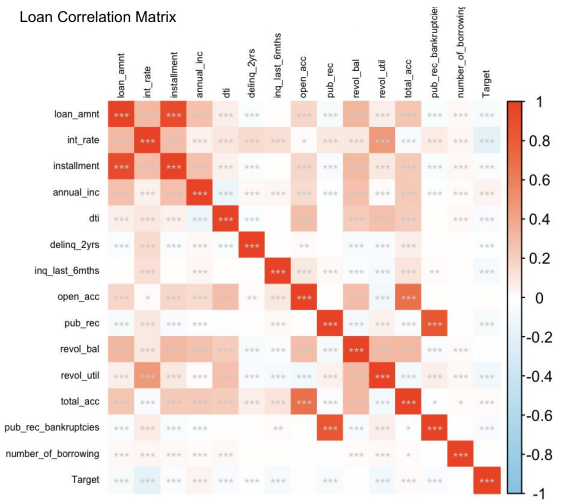
The exploratory data analysis will focus on the pre-loan data and how it may relate to default rates. The post loan data will only be used slightly towards the end of the section to calculate the appropriate risk ratio. This is due to the fact that the if we need to predict the status of the loan in real time, we would not have access to these set of post loan data. Using this information could make an artificially great model that has no real world relevance.

On top of the given data in the dataset I also feature engineered four additional columns:

- **Number_of_borrowing:** this looks at the number of times we have had a transaction with a customer till the “current” transaction. I believe that customers that have had a long history with Lending Club are less likely to default due to the long history of collaboration.
- **Smog_score/undefined_smog_score/num_words:** smog score is a readability index developed to evaluate how overly complex a piece of text is. It is analogous to the Fog index which is frequently used in finance to evaluate how transparent and truthful a company is with their public statements. I employed the same logic, and applied the smog score on the description of the loan in hopes to get some relevant insights on how loan description may affect the default rate. As an extension of this, I also created some subordinate columns such as undefined_smog_score and num_wrods to give labels for those text that either have too little content or just utterly makes no sense.

Correlation Analysis

First, I will be tackling the numerical data in the original dataset just to gauge some basic pairwise correlation that may exist.



Looking at the correlation matrix above, we can clearly observe multiple variables significantly and positively correlating with our Target variable (defined as any loan status that is not Fully Paid). These variables include loan amount, interest rate, installment, dti ratio, delinquencies, inquiries, # of derogatory public records, revolving line utilization rate, total accounts and public record of bankruptcies. The only variable that is significantly yet negatively correlated with our Target variable is the annual income. Personally, I think the results are rather sensible, and led me to form 4 main hypothesis:

III. EDA

Hypothesis 1: *individuals who have higher loan obligations to be paid off are more likely to default*: the variables of interest here are loan amounts, interest rate and installment. These variables characterize the obligation required for a person to pay back their loan on a monthly basis (higher interest rate or installment means more payment).

Hypothesis 2: *individuals who have a worse off credit history are more likely to default*: the variable of interest here are the dti ratio, delinquencies, inquiries, # of derogatory public records and public record of bankruptcies. A higher value in these variables likely indicates a track record of worse off credit.

Hypothesis 3: *individuals who have greater credit demand are more likely to default*: the variables of interest here are the revolving line utilization rate and total accounts. Higher value in these fields seems to indicate a greater demand for loans and potentially a less stable source to payoff loans.

Hypothesis 4: *individuals who earn more are less likely to default*: the variable of interest here is annual income and points to whether a person can pay off their loan.

One caveat to note is that despite the strong positive relationship, multiple of these variables are also pairwise correlated themselves, which means that not all of them will likely return significant individually. However, the hypothesis may still hold.

Discrete Factor Analysis

Next I will look more in depth into the relationship between discrete variables and the default rate.

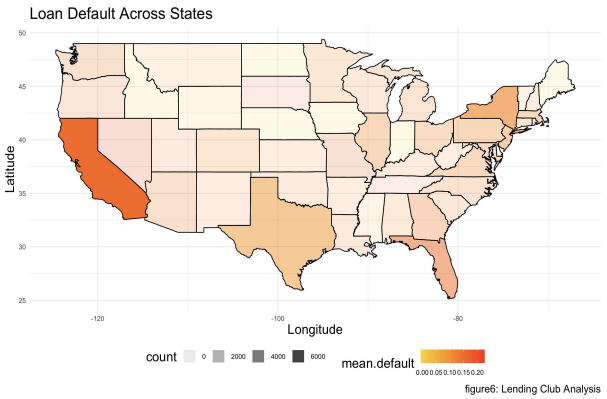


figure6: Lending Club Analysis

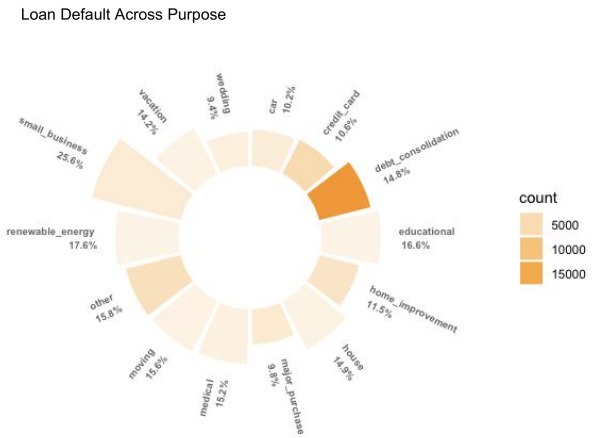
Hypothesis 5: *Loan default rate varies from region to region:* From the heatmap to the left, it seems like the amount of loans and default rate does not remain uniform across the United States. In particular, California seems to have the highest default rate and also the most amount of loans given.

Home Ownership

figure7: Lending Club Analysis

Hypothesis 6: *Loan default rate is higher for ambiguous home owners:* The graphics above shows a higher loan default rate for individuals who state Others as the home ownership status. However, this is a rather weak hypothesis since the sample size for individuals who state Others as their home ownership is rather small.

Purpose



Hypothesis 7: *Loan default rate varies across the purpose of the loan:* The circular bar chart above shows apparent disparity of loan default rates across different purposes. Specifically loans meant for small businesses seems the most likely to default (possibly due to the shaky start up system). On the other hand, loans used for weddings seem to be the safest of all others

III. EDA

Employment Length

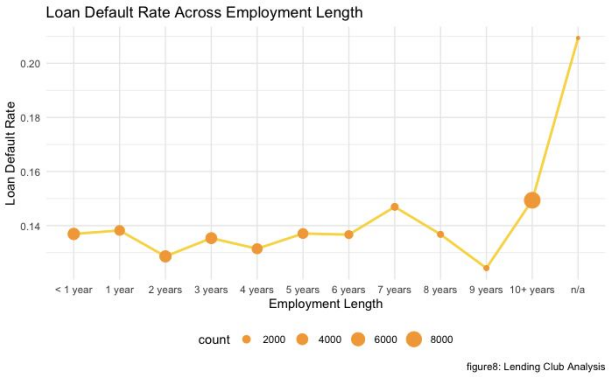


figure8: Lending Club Analysis

Hypothesis 8: Loan default does not vary across employment length but varies if employment is not specified: Looking at the graph above, the default rate is rather stable across employment period but peaks if the the length is NA.

Engineered Factor Analysis

I also specifically wanted to examine how the factors I engineered links with the default rate

Number_of_borrowing

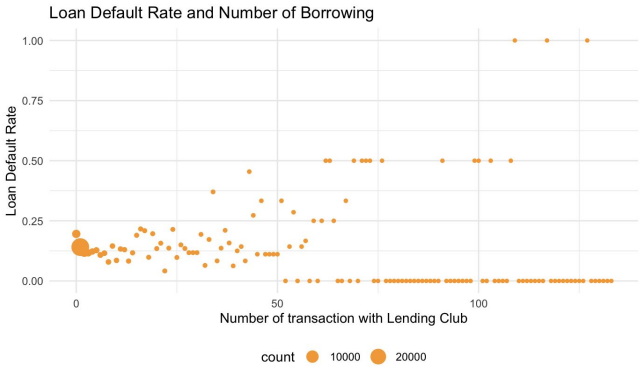


figure9: Lending Club Analysis

As opposed to my initial conjecture, the number of borrowing, if anything seems to be trending slightly upward in terms of contributing to default rate.

Smog Score

The smog score on the other hand appears quite interesting. Most of the earlier smog score results share similar loan default rates, but past the 20 threshold, the default value have increased quite strongly.

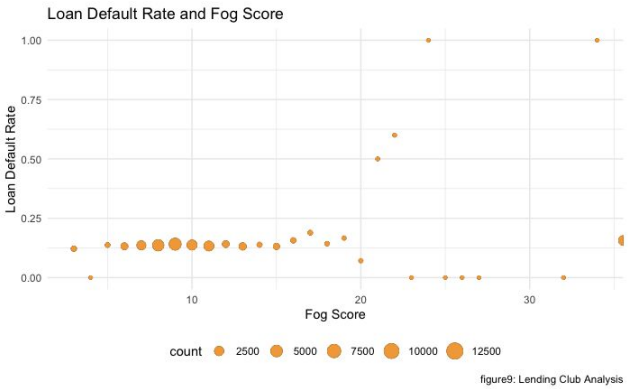


figure9: Lending Club Analysis

Since the relationship between the engineered factors and the default rate does not appear too clearly, I will refrain from making any tangible hypothesis for this part

Risk Ratio

The last part of the EDA will be spent on explaining the objective function to be optimized. The key element to realize here is that the expected payoff of investing in a good loan and the expected loss of investing in a bad loan is not actually commensurate, but weighted. To find the specific weighting that is appropriate, I have to estimate the expected value of both aforementioned conditions. To do so, I used the following calculation:

E[Investing in a good loan] = I[loan did not default] * (installment * term - loan amount)

E[Investing in a bad loan] = I[loan did default] * (total payment - loan amount)

Applying that the the whole dataset and taking the median and mean of expected value yields the following table:

Type	Median E[x]	Mean E[x]
Expected Value Good Loan	1618.72	2630.84
Expected Value Bad Loan	-3581.15	-5139.64

To avoid the presence of high leverage variables, I will take median as the expected value, resulting a risk ratio of 1:2.2 for False Positive to False Negative (with positive as default).

IV. Modelling

The Models

Carrying the the hypothesis and the variables listed and engineered, I will be moving on the the modelling phase of this paper. I reserved 10% of the data points in the o use as the final test data. The Bayes Classifier was used to determine the classification threshold across all models except for decision tree

Base Model

The model to start with is the Null model. This model predict all loans to not default. This effectively represents the status quo that needs to be beaten.

Full Logistic

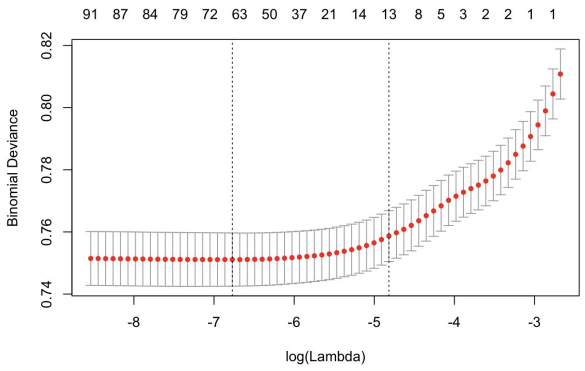
The second model attempted is a brute force logistic model, whereby no variable selection is performed and all the variables were employed (significant or not).

Pure Lasso

The third model attempted is a lasso model. I constructed the lasso model by doing a cross validation to find the the most appropriate lambda that ensures a simple model while guaranteeing a good cv test result. Two models were made from this technique, one was using the lambda that generates the minimal binomial deviance, while the other uses the lambda that generates the simplest model with 1 standard deviation of the deviance

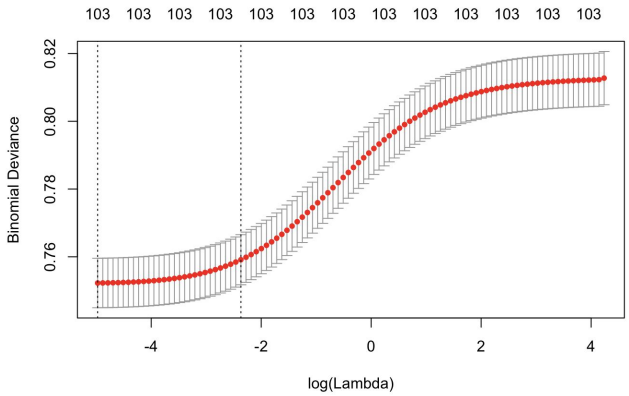
Lasso with backward selection

Building on top of the pure lasso model. Backward selection were then done to both of these models to ensure only significant variables are left.



Pure Ridge

The fourth model tried was the ridge model. Similar to lasso I used the cross validation to find the best lambda value to maximize the binomial deviance. The model with both lambda min and 1se were used once again.



Pure Elastic Net

The fifth model used was the elastic net model, a hybrid of both lasso and ridge. I fitted a number of elastic net model using various alpha level including 0.3, 0.5 and 0.7. For each of these I tested the model on the test data using both lambda parameters as above

Decision tree

The sixth model tried on this dataset, was the decision tree using the rpart package. I hyperparameter tuned the complexity parameter of the decision tree using 10 fold cross validation and employed the weighted MCE as the selection criteria.

Upsampling with Decision Tree

I also attempted to upsample the data to deal with the weight imbalance. Specifically I bootstrapped more entries from the the data which is labeled as default and created a more balance dataset, in terms of # of loans being paid off and defaulted. Then I applied a similar process using the decision tree method earlier, including hyperparameter tuning and cross validation

VI. Conclusion

Model Summary

While models such as lasso,ridge, elastic net and logistic were able to outperform the Null Model. Decision Tree, even after upsampling was not able to do so (in fact upsampled made it even worse). Although most models are quite similar in terms of performance (ROC and Test Error), the final model chosen was the lambda 1se model by elastic net (alpha - 0.3), since it has the least test error (which in this case given the unique nature of weight takes precedence in my opinion).
see model performance summary in the appendix

Final Model: Raw Elastic Net with Alpha = 0.3 and Lambda = lambda min

Variables <fctr>	coefficients <dbl>
(Intercept)	-3.213578e+00
term60_months	4.056440e-01
int_rate	9.181208e+00
gradeD	1.616725e-02
emp_lengthn/a	3.452109e-01
annual_inc	-2.488570e-06
purposecredit_card	-6.924330e-02
purposeother	5.442322e-02
purposesmall_business	5.067672e-01
addr_stateCA	4.336741e-02
addr_stateFL	8.084556e-02
addr_stateNV	1.717185e-01
dti	4.099372e-04
inq_last_6mths	7.749700e-02
pub_rec	9.683460e-02
revol_util	2.642959e-01
pub_rec_bankruptcies	8.871947e-02

Looking at the final model above, we can notice that most of the initial hypothesis were in fact verified. For instance interest rate is highly positively related with default, and annual income negatively related. Interestingly, D grade loans is singled out as a loan with high default value. Credit card exchanges seems to have rather low default rate. California, as expected, along with Nevada and Florida, is of the states with a rather high default probability.

Unfortunately, the features that were engineered did not appear significant in the final model. Homeownership was also not listed as a significant factor as opposed to what I initially outlined.

Assessment

The confusion matrix of the our final model looks as follows:

	Actually Paid Off	Actually Defauled
Predicted Pay Off	**3310**	**488**
Predicted Defaulted	**56**	**43**

If this formula were to be used in making the decision it would have salvaged 43 loans that ended up defaulting while missing 56 good loans. As per our expected calculation earlier this trade off is way better compared to the risk attached to both side. It could in fact potentially save Lending Club 43*(3581) + 56*(-1618) = **60k dollars.**

Recommendations

Overall, it is nonetheless quite evident that the model did not significantly outperform the null cases, indicating Lending Club likely does have a strong algorithm in place. However I would recommend two main strategic initiative to LC. **First, Lending Club should reconsider some of its interest rate**, while it is in accordance to FICO standard, the interest rate right now may be slightly too high - causing more defaults to happen. By lowering the interest rate, while the income may slightly decrease, it will likely be offsetted by the decreasing default rate as well. **Second, Lending Club should also reconsider their marketing strategy**, despite being the area with the highest default rate, most of Lending Club's loan re still in California, it may be of the company's best interest to diversify its loan location to reduce the risk of its portfolio.

Improvements & Next Steps

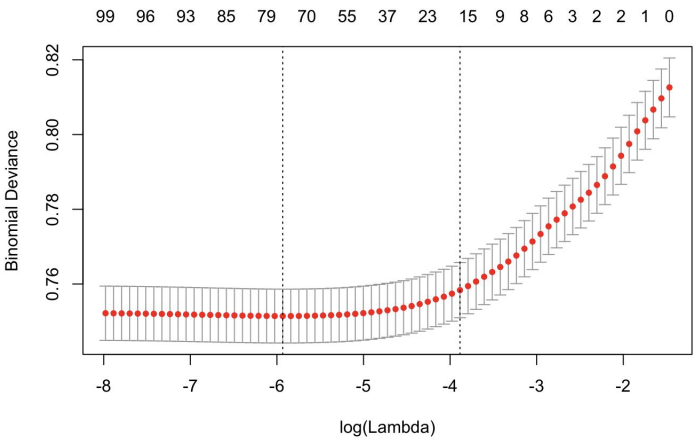
I believe this paper can definitely improved with the addition of more advanced set of modelling tools, and better hyperparameter tuning. Furthermore, it may also help significantly if macro-trends (recession) could be taken into account in making the prediction.

I. Appendix

Model Summary

Model	AUC ROC	Test Error
Null Model	**0.5**	**0.2998**
Full Logistic	**0.6969**	**0.2978**
Pure Lasso (lambda 1se)	**0.6983**	**0.2921**
Pure Lasso (lambda min)	**0.6898**	**0.2930**
Lasso BSelect (1se)	**0.7002**	**0.296**
Lasso BSelect (min)	**0.698**	**0.2948**
Pure Ridge (lambda 1se)	**0.6926**	**0.2919**
Pure Ridge (lambda min)	**0.6968**	**0.2928**
Pure Elastic (alpha 0.3 lambda 1se)	**0.6921**	**0.2898**
Pure Elastic (alpha 0.3 lambda min)	**0.6981**	**0.2932**
Pure Elastic (alpha 0.5 lambda 1se)	**0.6912**	**0.2912**
Pure Elastic (alpha 0.5 lambda min)	**0.6981**	**0.2929**
Pure Elastic (alpha 0.7 lambda 1se)	**0.6927**	**0.2925**
Pure Elastic (alpha 0.7 lambda min)	**0.6982**	**0.2918**
Decision Tree	**0.5**	**0.2998**
Decision Tree (upsampled)	**0.5535**	**0.3900**

Elastic Net Visuals (alpha = 0.3)



All other relevant codes will be in the RMD file.