Lending Club Project - SDS 323

Alt Nayani and Conor McKinley

5/11/2020

Abstract:

About Lending Club: Lending Club is a platform that matches those requesting loans to investors wishing to fund those loans. The purpose the loans range to fund automobile purchases, mortgage payments, or debt consolidations. After providing adequate amounts of data, Lending Club provides a 'rating' from A to G, which determines the applicable interest rate for the specified loan. These loans are then available for investors to fund. Each loan may be funded fully by one investor or by multiple investors that seek to gain exposure to the specific loan.

Lending Club Diversification: While funding individual loans are an option on Lending Club, the majority of investors fund a diversified basket of notes to diversify any idiosyncratic risks of an individual loan. The concept here is: if an investor were to fund one large loan and the loan defaults, then the investor loses all the capital. However, if an investor funds small amount of a large portfolio of loans, then an single loan cannot bankrupt the portfolio and reduces the overall variance of returns. Lending Club's diversified investing program has historically returned 3.77% to 6.01% annually (14.81% over a 3-year period), which we will use as the null performance metric we are attempting to beat.

Goal of Project: The goal of the project is to be able to identify a portfolio that outperforms the 'default' diversified portfolio of loans that Lending Club offers. Definitionally, we will define outperformance as the ability to earn an additional percentage return (profit over the funded amount). While in the traditional sense, investors will care about the riskiness and variability of returns of the different strategies we try, we can assume the riskiness of the underlying loans are equal in each strategy since all loans will be A-D quality and of 36 month term.

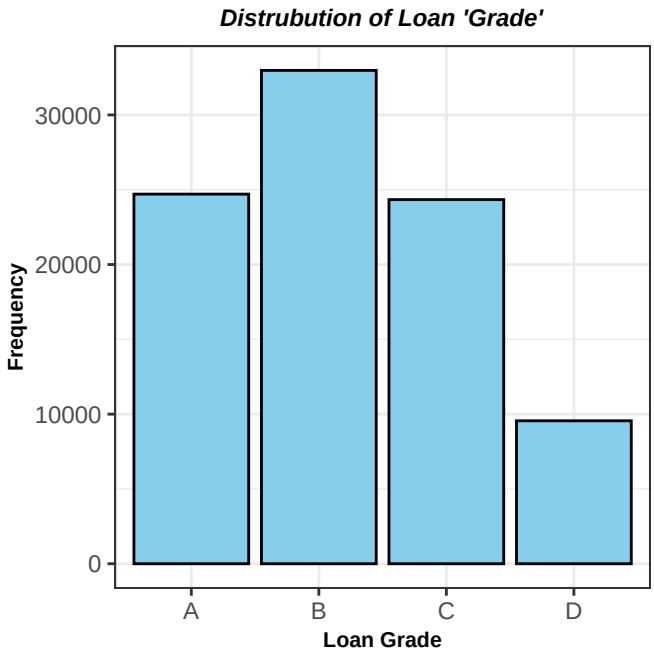
To be able to maintain credibility in the performance of the project, we will split the data set into a training and testing split then determine the appropriate outputs of the results.

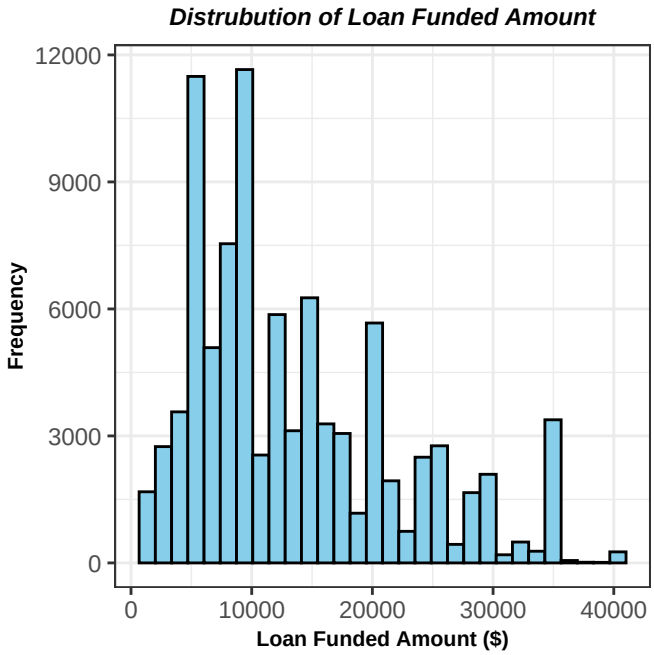
Data Cleaning and Variables of Interest:

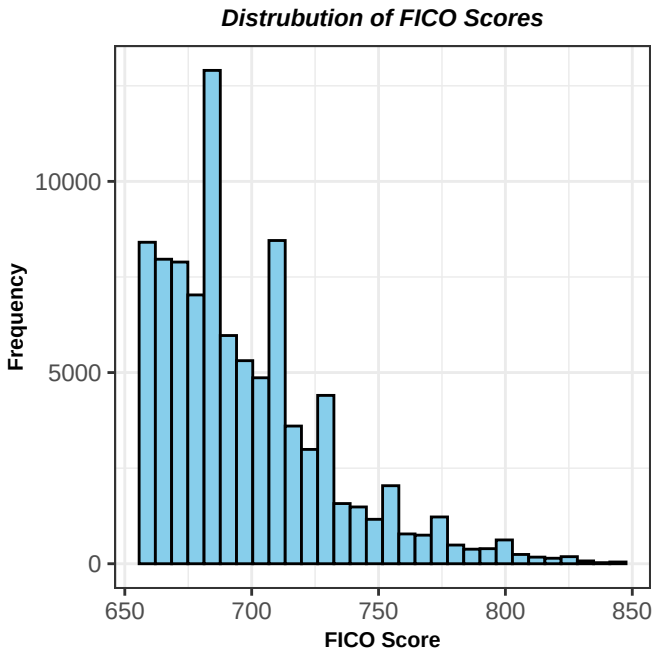
The data specifically looked at which loans originated in Q1 2016 with a term of 36 months. All 60 month loans were excluded due to the possible on-going nature of the loans. Furthermore, Lending Club's diversified portfolio contains only A through D 'grade' loans; therefore, to make an apple to apples comparison, we will exclude any loan that is not A through D.

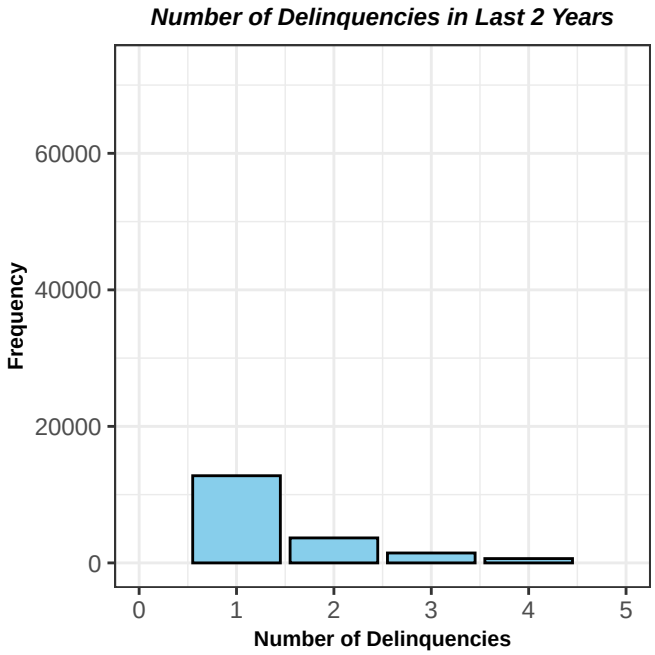
Prior to running any supervised learning techniques, we will view the initial variables of interests to better understand any patterns and the raw data.

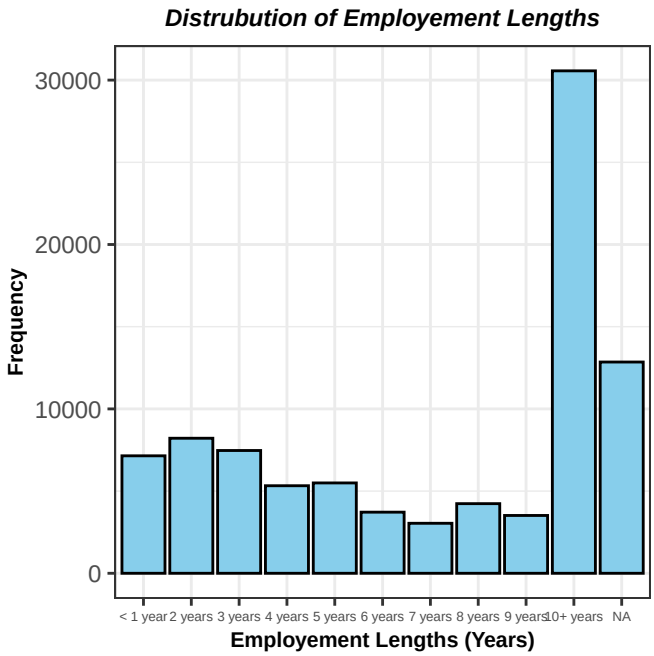
First, we will look at the distribution of key variables in the dataset such as: grades, funding amount, FICO score, delinquencies in the last 2 years, employment length and verification status.

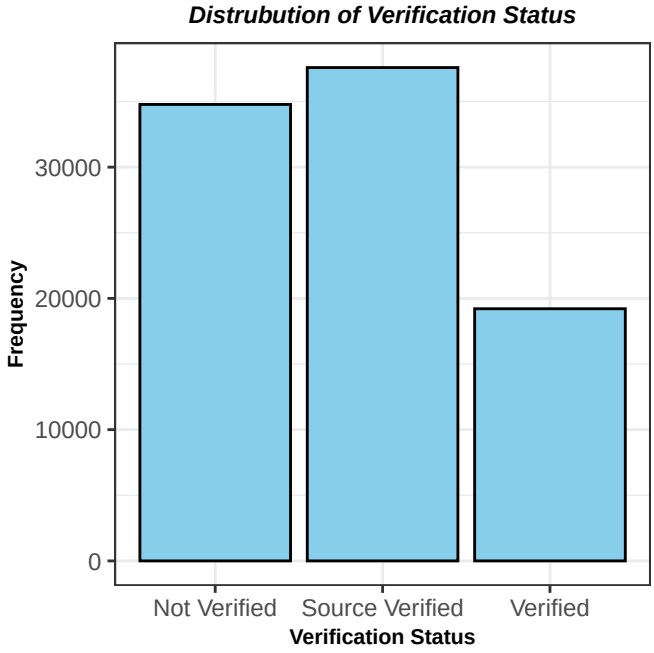

Distrubution of Loan 'Grade'
30000
20000
10000
0
Frequency
A
B
C
D
Loan Grade


Distrubution of Loan Funded Amount
12000
9000
6000
3000
0
Frequency
0
10000
20000
30000
40000
Loan Funded Amount ($)

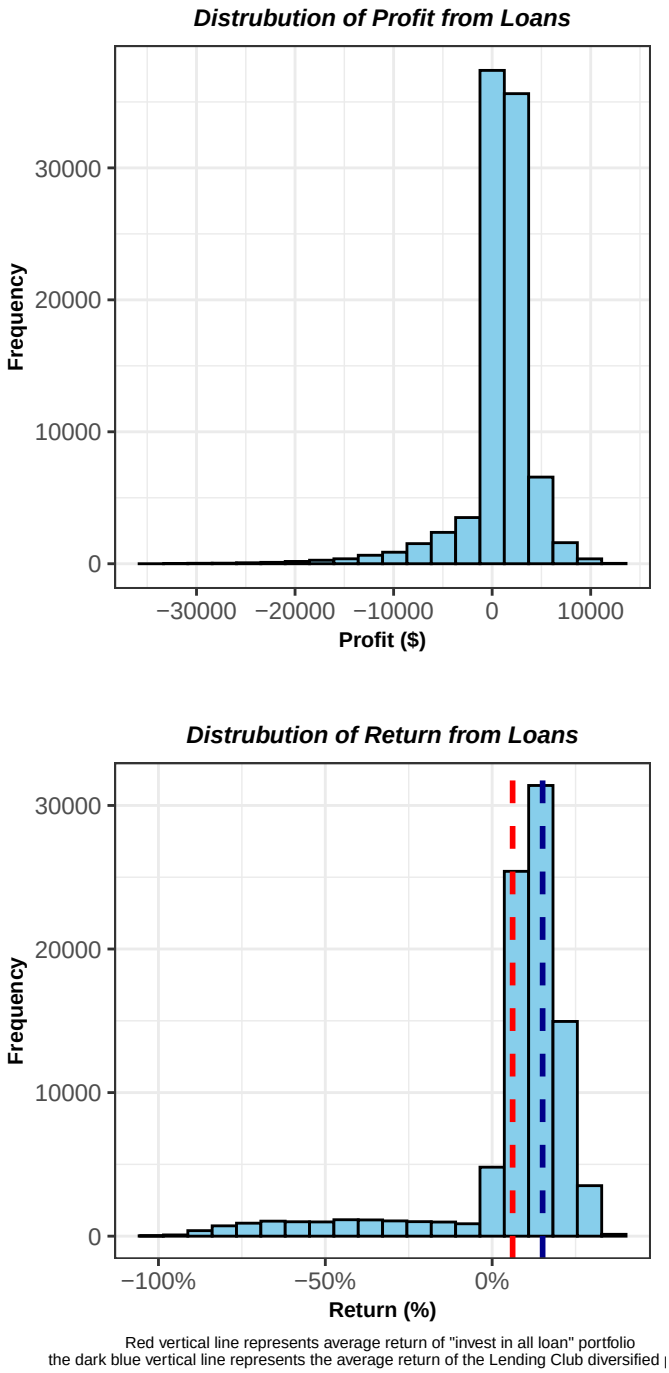

Distrubution of FICO Scores
Frequency
10000
5000
0
650
700
750
800
850
FICO Score


Number of Delinquencies in Last 2 Years
Frequency
60000
40000
20000
0
0
1
2
3
4
5
Number of Delinquencies





Now we will look at the variables that we are trying to predict: profit and % return (defined as profit over funded amount):



The return that is shown above is a simple equal investment in all loans available in Lending Club. We can see that the return is centered slightly positive and highly variable with returns as low as -100% and as high as +38%. Furthermore, the average return is 6% over a 3 year period, less than the annualized return of 4.81% quoted by Lending Club (15.1% over a three year period).

Linear Regression:

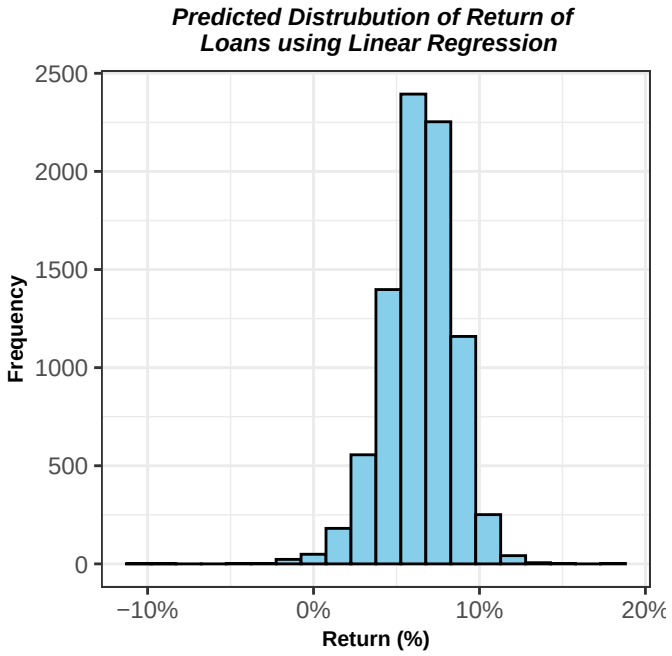
We will begin by using linear regression to predict returns of loans originated on the Lending Club platform.

First, we will begin by attempting a simple linear regression with a list of 30 variables that seem promising. All variables included were individually selected to avoid any data that would not be available at the time of the loan origination.

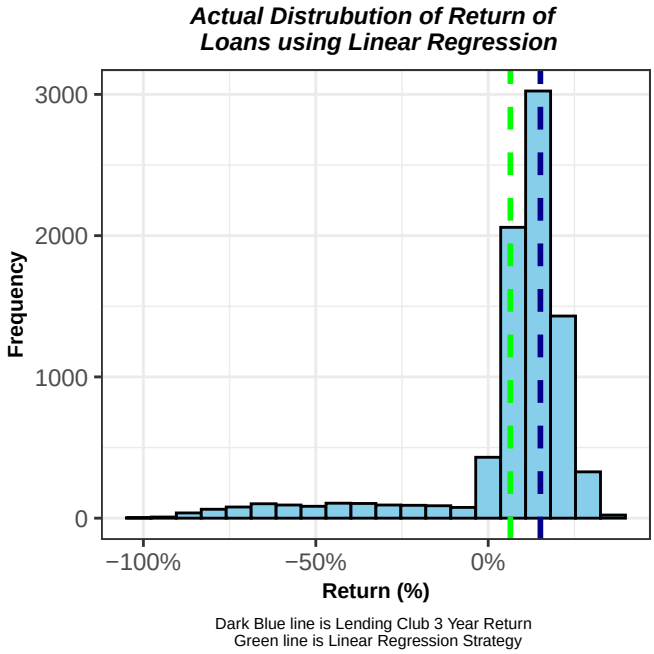
Furthermore, we split the dataset into a training and testing split over multiple iterations to get a low variance estimate of our results.

The overall model performed pretty terribly in-sample with an average R-squared of 0.9%.

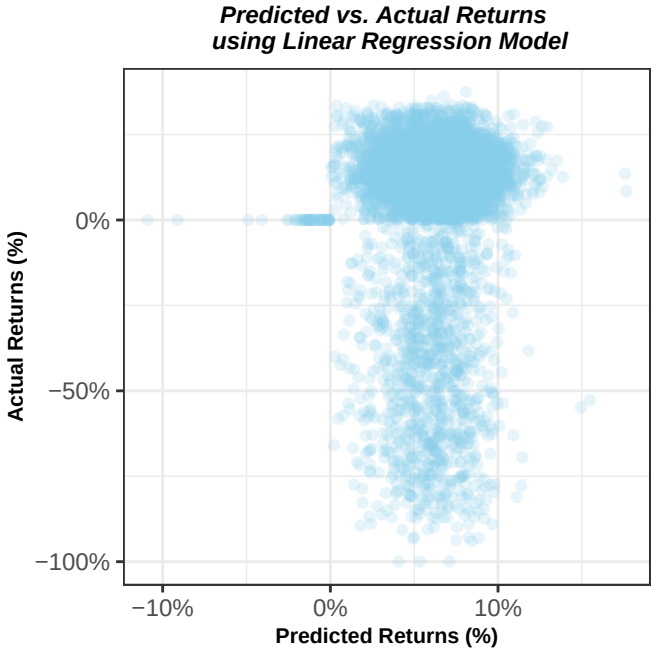
Below you can see the predicted returns based of the simple linear model:



All loans that we predict to be harmful to the portfolio (negative returns), we will not invest in them. Therefore, we will compare the loans that we projected with positive returns to the actual returns. The distribution those loans can be seen below:



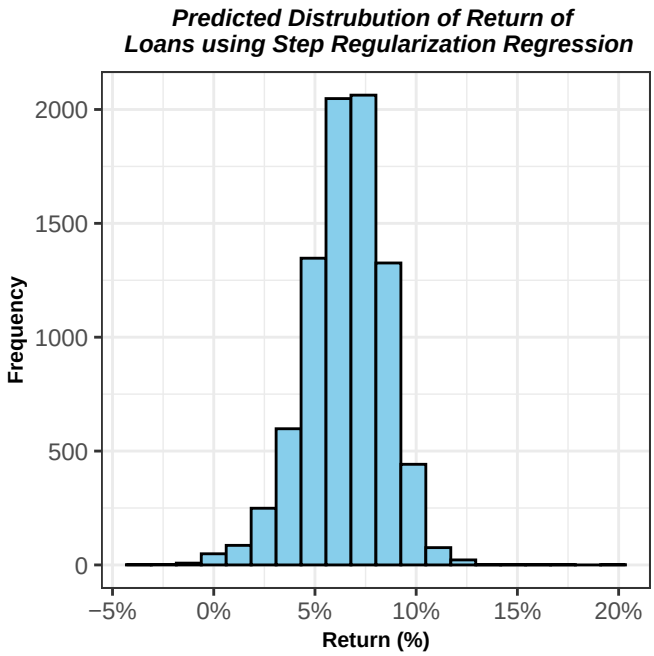
Finally, we will look at the predicted returns v. the actual returns on a scatterplot:



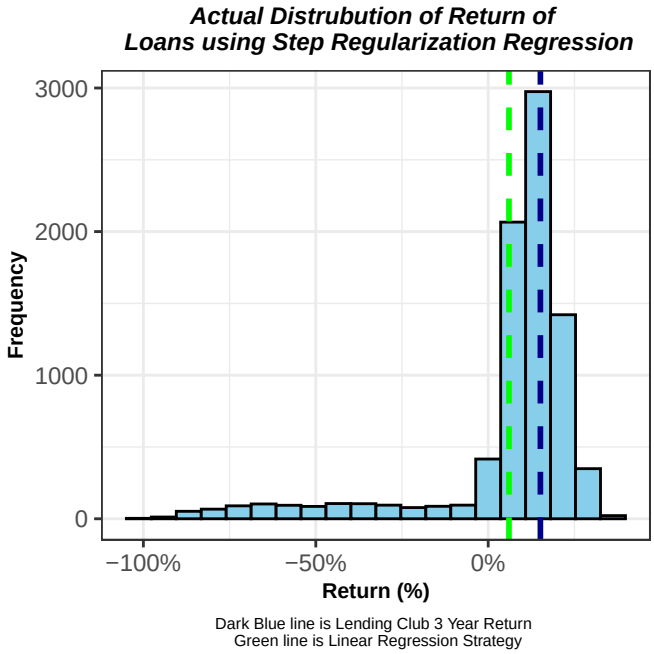
The next type of linear regression we will look at is a regularization model that fits the training data by trying to minimize the error through addition and subtraction of terms and interactions. Similar to the other linear regression model, we split the dataset into training and testing splits as well as removed any lookahead bias in the data.

The regularization model had an average R-squared of 0.8%, a tenth of a percentage point below the R-squared of the simple linear model.

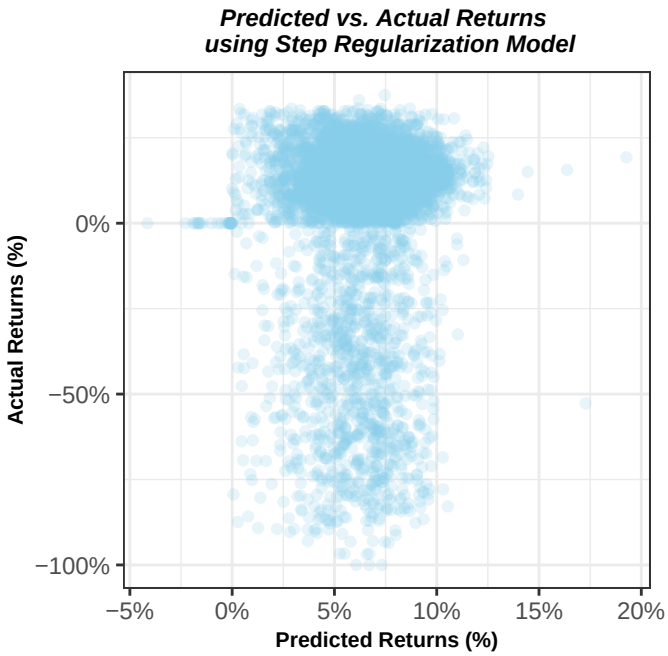
Below we can see the distribution of projected returns:



Similarly, we removed any loans with a projected negative return and compared the remaining loans to the actual returns. Below is the distribution of how we performed out of sample:



Finally, we included a scatterplot of the actual vs. predicted return of the strategy. As seen earlier with the simple linear model, there is no distinct pattern and the model has almost no ability to predict the return of the actual loan.

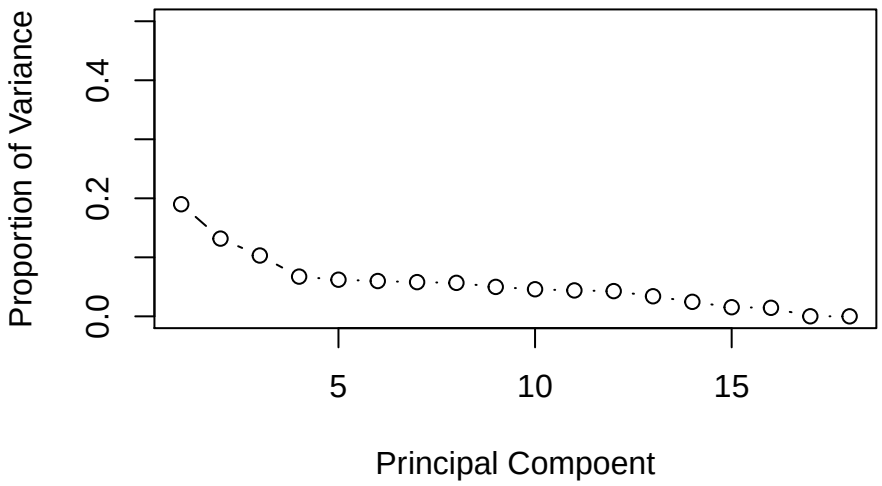


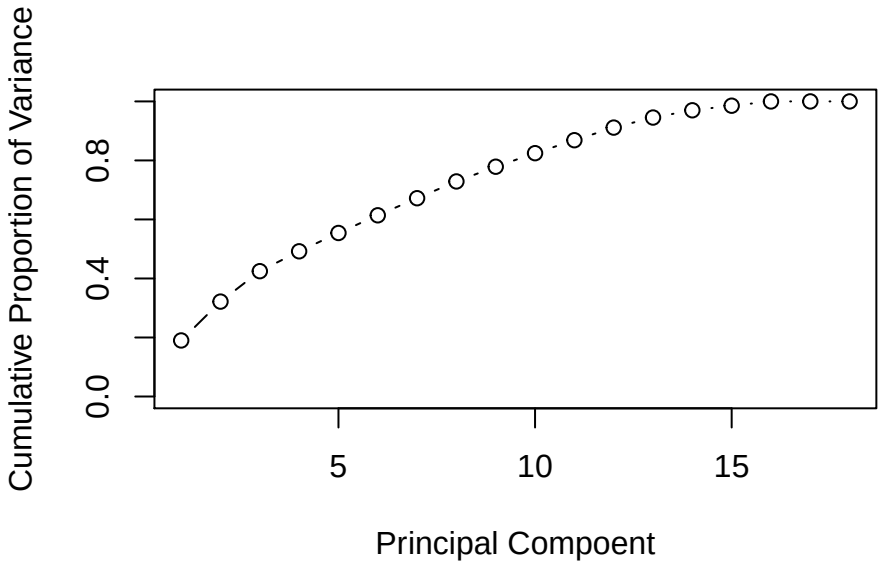
To conclude this section, we can see both the simple linear model and regularization model did a poor job at predicting the return of the loans. We underperformed the Lending Club diversified return of ~15% over a 3-year period, compared to our ~6% return over a 3-year period.

Principal Components Analysis (PCA):

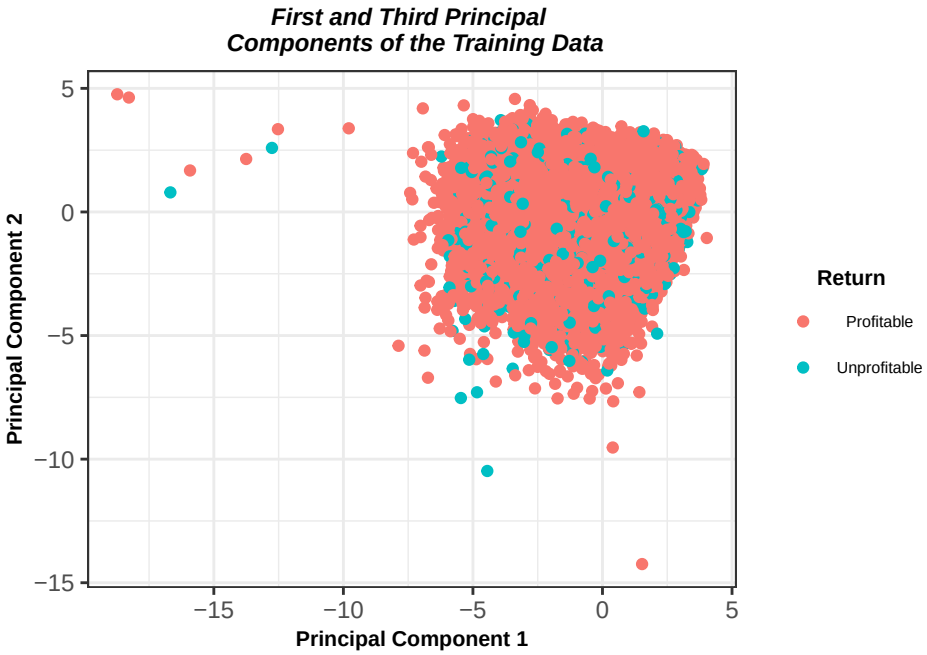
Next we will be using Principal Components Analysis to better predict returns of specific loans, then invest in loans that are predicted to be profitable and avoid predicted unprofitable loans. Since we have a large number of features in the dataset, a dimensionality reduction technique such as PCA could be useful to determine the best indicators of performance.

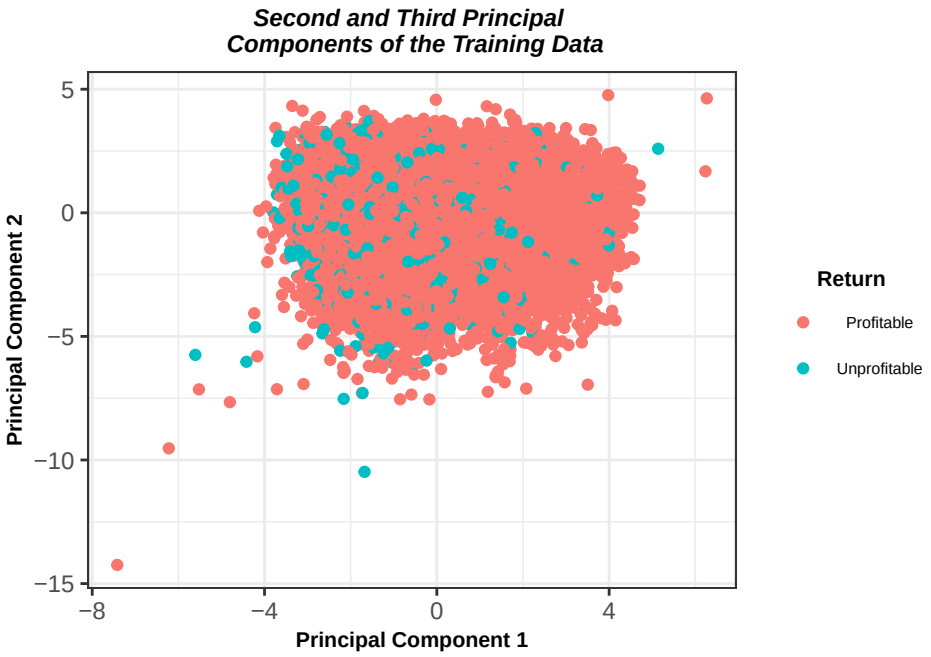
We will begin by splitting the dataset into a training and testing dataset then running PCA on the training data. Furthermore, we will remove any variables that would not have been known at the origination of the loan, removing any lookahead bias. Let us plot the summary of PCA and the first three components against each other.


Proportion of Variance
0.0
0.2
0.4
5
10
15
Principal Compoent


Cumulative Proportion of Variance
0.0
0.4
0.8
5
10
15
Principal Compoent

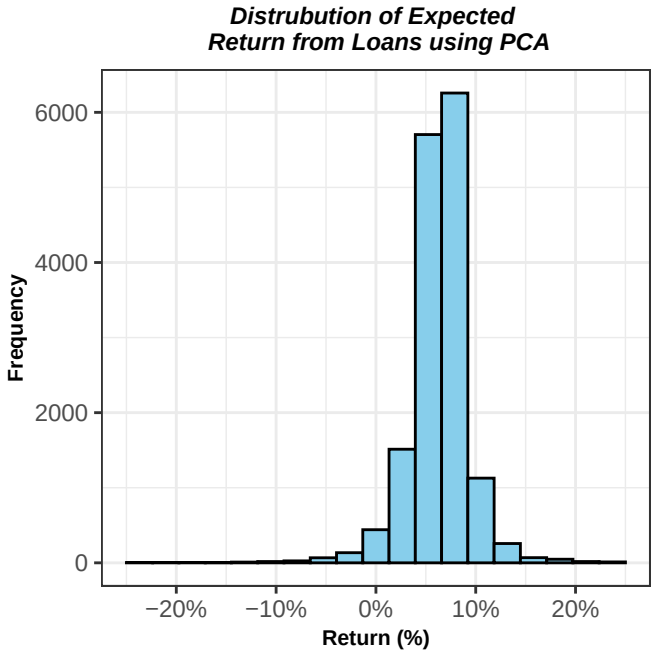

First and Second Principal
Components of the Training Data
Principal Component 2
4
0
−4
−8
−15
−10
−5
0
5
Principal Component 1
Return
Profitable
Unprofitable


First and Third Principal
Components of the Training Data
Principal Component 2
5
0
−5
−10
−15
−15
−10
−5
0
5
Principal Component 1
Return
Profitable
Unprofitable

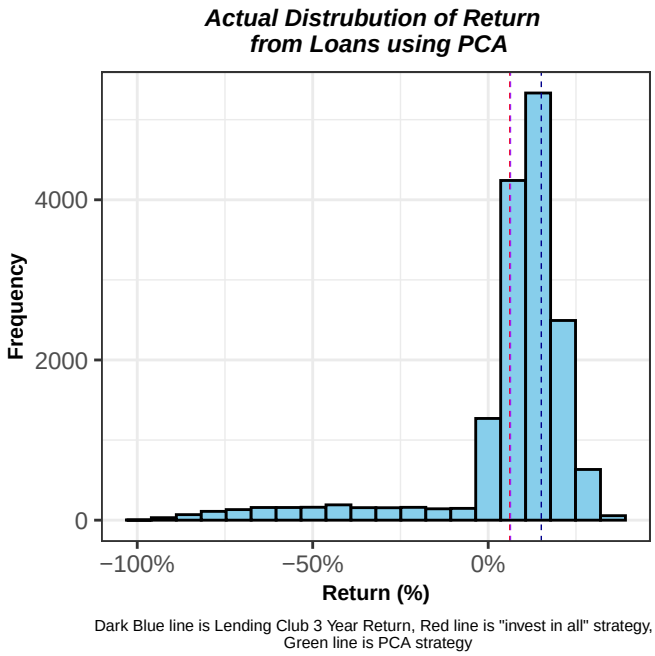


As we can see, the components do not do a great job of identify profitable vs. unprofitable loans with the variables present. Next, we can create a linear model that regresses the principal components (including interactions) on percent return of the loans. The first ten principals resulted in approximately 2% R-squared for return. We then applied the PCA model and regression to the testing data set.

Below we can see the distribution of expected returns based on the PCA analysis:



Since we are interested only in loans that we expect to be profitable, we will test our results only on the expected positive return loans. We can see below how we would have done if we invested in loans based on the PCA analysis.



The PCA strategy underperformed the 'invest in all' strategy and the Lending Club portfolio by over 7%. Furthermore, the PCA underperformed both Linear Regression strategies.

Conclusion:

The goal of the project was to be able to beat the diversified program offered by Lending Club based on the data available to us at the origination of the loan. Below is the summary of our results:

Table 1: Return per Strategy

	Lending Club Diversified	Invest All Strategy	Simple Linear Regression	Regularization Model	PCA
Avg. Return	0.1513	0.0614	0.0643	0.0602	0.0635

Overall, we can see that neither the linear models or the Principal Component Analysis was able to outperform the Lending Club diversified program. On average, the PCA was the worst-performing strategy followed by the invest in all strategy. Of our models, the regularization performed the best at a 6.57% return over a 3-year period.

We can conclude that we cannot beat the Lending Club strategy. We believe there are a couple of reasons for this. First, Lending Club has a team of professionally staffed employees that selectively chose loans to be included in their portfolio. While we are Finance students, we do not think we are able to accurately determine undervalued loans by looking individually at loans. Secondly, we were unable to match the percent allocation to each 'grade' of loan equivalently to Lending Club. This may provide a rationale for the lower return, since we may be over allocating to 'D' grade loans that may create a loss of capital or over allocating to 'A' grade loans and not getting enough risk thus return in our portfolio.

We can conclude that Lending Club has some expertise in investing in a loan portfolio as opposed to a retail investor looking to outperform the market.