

*Lending Club Micro-Loans:*  
**Predicting Loan Defaults  
via Machine Learning.**

**Michael Halpert**  
**5.28.17**  
**Springboard Capstone Project**

## *Agenda*



- 1. Introduction to Lending Club**
- 2. Problem Explanation**
- 3. Data Sets & Summary Statistics**
- 4. Predictive Models**
- 5. Results**
- 6. Additional Considerations**

<b>Build a Portfolio</b> Per Loan: <input type="text" value="\$25"/>								
<b>Filter Loans</b> <a href="#">Save</a>   <a href="#">Open</a> Inquiries in the last 6 months ▼ 10 <input type="range" value="10"/> 0 10 Loan ID ▼ <input type="text"/> Location State ▶ Open Credit Lines ▼ 0 >30 <input type="range" value="30"/> 0 >30 Interest Rate ▶ Loan Purpose ▶ Public Records ▶								
<input type="checkbox"/>	Investment	Rate	Term	FICO®	Amount	Purpose	% Funded	Amount / Time Left
<input type="checkbox"/>	\$0	G 5 30.99%	60	660-664	\$20,000	Loan Refinancing & Consolidation	<div><div></div></div> 87%	\$2,450 18 days
<input type="checkbox"/>	\$0	G 5 30.99%	60	670-674	\$18,000	Loan Refinancing & Consolidation	<div><div></div></div> 92%	\$1,350 22 days
<input type="checkbox"/>	\$0	C 1 12.62%	36	690-694	\$16,000	Other	<div><div></div></div> 97%	\$375 26 days
<input type="checkbox"/>	\$0	G 2 30.84%	60	660-664	\$15,000	Home Improvement	<div><div></div></div> 81%	\$2,800 19 days
<input type="checkbox"/>	\$0	C 3 14.08%	60	665-669	\$14,400	Medical Expenses	<div><div></div></div> 76%	\$3,325 20 days
<input type="checkbox"/>	\$0	F 3 30.17%	60	705-709	\$26,000	Loan Refinancing & Consolidation	<div><div></div></div> 93%	\$1,750 25 days
<input type="checkbox"/>	\$0	E 5 26.30%	60	695-699	\$15,575	Loan Refinancing & Consolidation	<div><div></div></div> 77%	\$3,550 24 days
<input type="checkbox"/>	\$0	G 5 30.99%	60	660-664	\$11,400	Moving - Relocation	<div><div></div></div> 67%	\$3,750 24 days

LendingClub is a crowd lending platform that enables pools of individual investors to act in the capacity of a bank and make unsecured loans to individual borrowers.

## Medical expenses for 48002972

[Sell Notes](#) [Glossary](#)

Loan ID: 109746317 | [Lending Club Prospectus](#)

[« Previous](#) | [Next »](#)

[Add to Order](#)

Amount Requested **\$14,400**  
Loan Purpose **Medical expenses**  
Loan Grade **C3**  
Interest Rate **14.08%**  
Loan Length **5 years (60 payments)**  
Monthly Payment **\$335.67 / month**

Review Status **Approved** ✓  
Funding Received **\$11,075 (76.91% funded)**  
Investors **267 people funded this loan**  
Listing Expires in **20d 13h (6/19/17 6:00 AM)**  
Note Status **In Funding**  
Loan Submitted on **5/19/17 7:29 PM**

### ■ **Member\_117989031's Profile** (all information not verified unless noted with an "\*\*")

Home Ownership	<b>RENT</b>	Gross Income	<b>\$3,833 / month</b>
Job Title	<b>Supervisor</b>	Debt-to-Income (DTI)	<b>6.08%</b>
Length of Employment	<b>9 years</b>	Location	<b>986xx</b>

### ■ **Member\_117989031's Credit History** (as reported by credit bureau on 5/19/17)

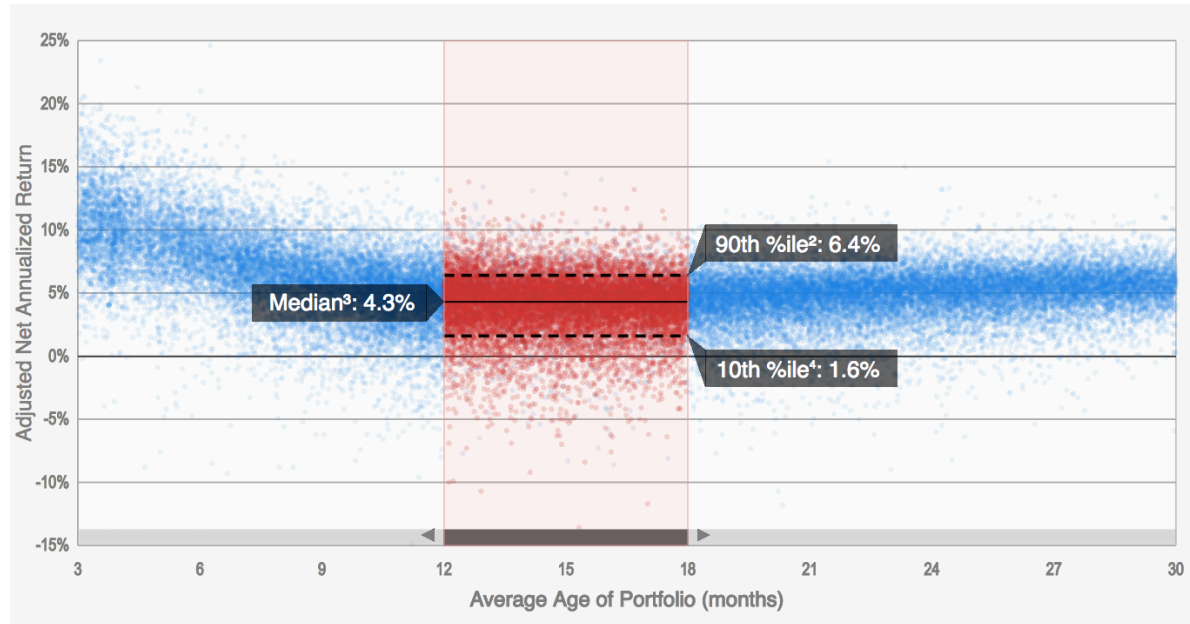
Credit Score Range:	<b>665-669</b>	Delinquent Amount	<b>\$0.00</b>
Earliest Credit Line	<b>10/2003</b>	Delinquencies (Last 2 yrs)	<b>1</b>
Open Credit Lines	<b>3</b>	Months Since Last Delinquency	<b>16</b>
Total Credit Lines	<b>15</b>	Public Records On File	<b>0</b>
Revolving Credit Balance	<b>\$4,466.00</b>	Months Since Last Record	<b>n/a</b>
Revolving Line Utilization	<b>81.20%</b>	Months Since Last Major Derogatory	<b>16</b>
Inquiries in the Last 6 Months	<b>0</b>	Collections Excluding Medical	<b>0</b>
Accounts Now Delinquent	<b>0</b>		

### ■ **Loan Description**

LendingClub provides credit history information on each borrower to potential lenders. This enables lenders to evaluate the risk level of each individual loan.

# While loan yield rates appear attractive, loan defaults substantially lower investor returns.

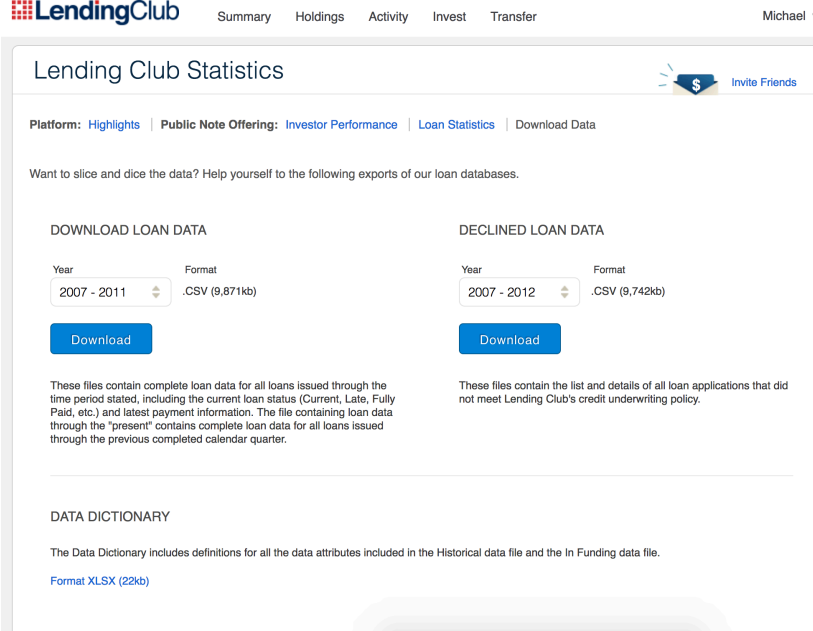
Adjusted Net Annual Return



Even as some loans offer annual yields as high as 30%, their risk is quite substantial. The best investors with diversified portfolios only yield around 6.4%

**Can machine learning help  
us predict loan defaults?**

# Lending Club maintains very thorough data sets on all borrowers and it's full loan history.



The screenshot shows the 'Lending Club Statistics' page. At the top, there's a navigation bar with the Lending Club logo and links for Summary, Holdings, Activity, Invest, and Transfer. A user profile 'Michael' is visible in the top right. Below the navigation bar, the page title 'Lending Club Statistics' is displayed. A sub-header 'Platform: Highlights | Public Note Offering: Investor Performance | Loan Statistics | Download Data' is present. A message states: 'Want to slice and dice the data? Help yourself to the following exports of our loan databases.' There are two main sections: 'DOWNLOAD LOAN DATA' and 'DECLINED LOAN DATA'. Each section has a 'Year' dropdown menu (set to '2007 - 2011' and '2007 - 2012' respectively) and a 'Format' dropdown menu (set to '.CSV'). Below each dropdown is a 'Download' button. A note at the bottom states: 'These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.'

<https://www.lendingclub.com/info/download-data.action>

Many loans in the entire lending club portfolio have reached their full maturity. The data sets go back to 2007, and have extensive credit information and outcomes for all borrowers. The data dictionary is also very well documented.

# Can we use data to make predictions?

The Loan Data Files includes application and outcome information including:

- Borrower demographics, employment and income status
- Credit Histories & Credit Scores.
- Information on existing credit accounts, limits, loan status', and delinquencies
- Loan amounts and purposes
- Lending Club repayment histories and late fees

By utilizing the outcomes of loans with earlier vintages, we should be able to train reasonably high quality predictive models to assist with our investigation.

**We will use the models to predict default – or a loss of the investors principal in a loan.**



# Data Ingestions

# Cleaning the Data:

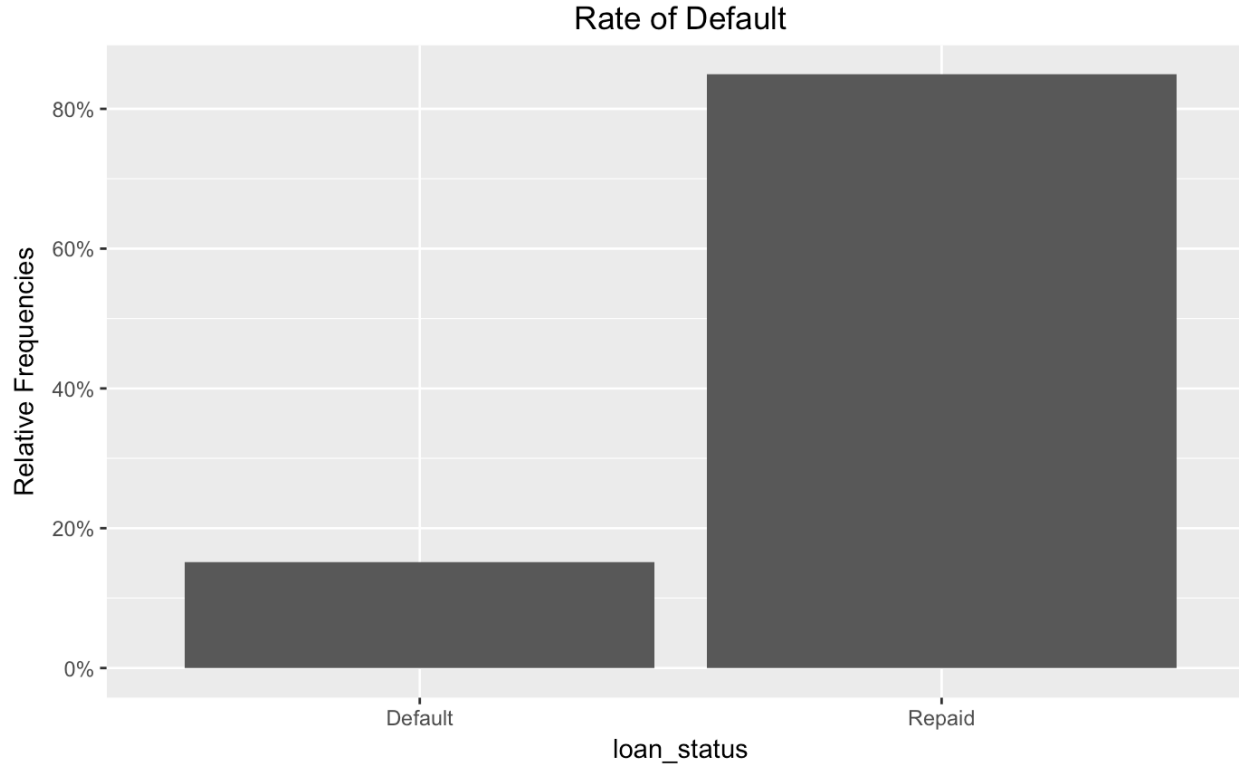
While the Lending Club data is detailed, it is slightly messy and requires quite a bit of cleaning:

- There are a number of loans that have not matured, and we are only interested in early vintage loans that have either been fully repaid or defaulted.
- Many of the fields feature text which R interprets as a string instead of a numeric or factor.
- The data set features a number of empty columns, NA's and other columns where a null result is a 0. These need to be normalized.
- There are a number of features where the factors have too many levels. In many of the outputs, our need is to find a binary yes/no output.
- The data set also features extraneous loan outcome fields that are heavily correlated to the loan outcome. We are only interested in working with data features that we would have when a new loan is published.

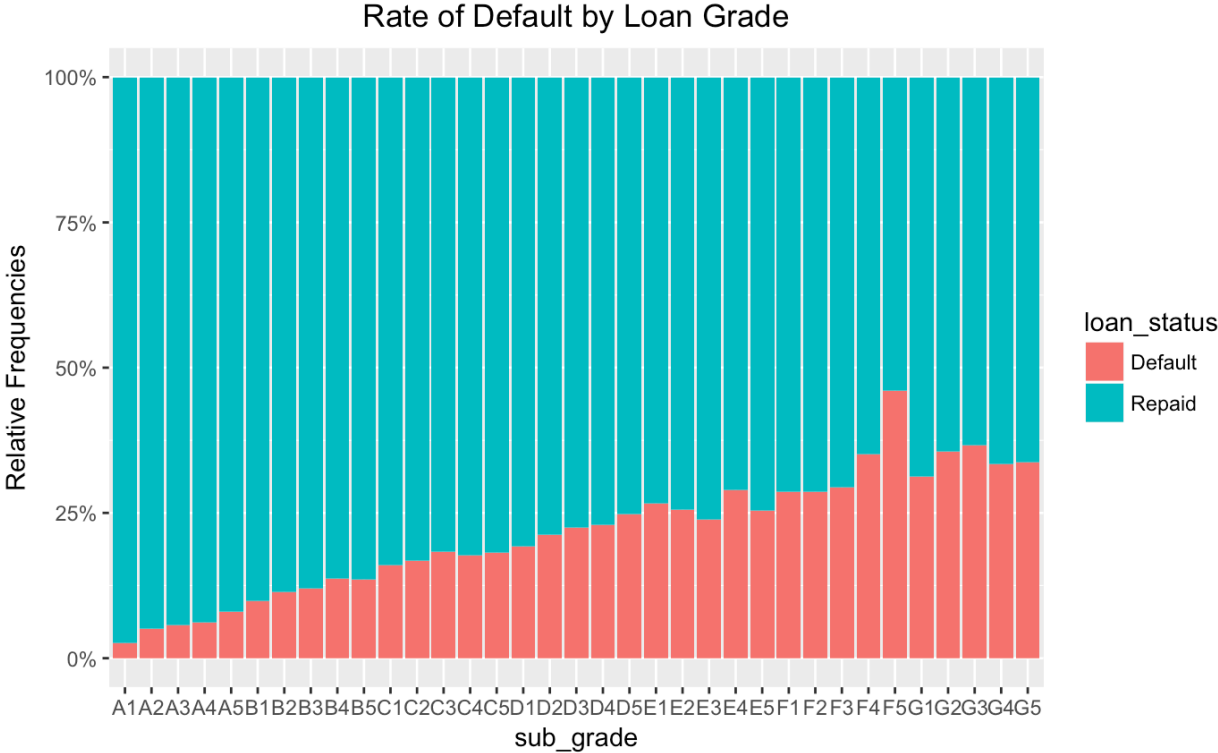
# Most Important Variables:

Variable	Definition
Term	Length of the loan. Either 36 or 60 Months
Int_rate	Rate of interest charged on loan
Sub_grade	Grade of the loan that is assigned by lending club. This is tied to the interest rate on the loan.
Emp_length	How long has the borrower been employed with their current employer.
Home_ownership	The borrowers housing is either fully owned, mortgaged, rented or they have another arrangement.
Verification_status	Have the details of the loan, employment and income been verified by lending club.
purpose	What is the loan being used for?
DTI	Debt to Income. What is the borrowers total debt burden as reported by the credit bureau.
Delinq_2years	Debt to Income. What is the borrowers total debt burden as reported by the credit bureau.
fico_range_high	The high end of the reported FICO credit score of the borrower.
inq_last_6mths	How many credit inquiries has the borrower had on their credit file over the last 6 months.
mths_since_last_delinq	How many months has it been since the last delinquency on the borrowers credit file.
mths_since_last_record	How many months has it been since the last record on the borrowers credit file.
open_acc	How many open credit accounts a borrower has on their credit file.
pub_rec	The number of public records a borrower has on their credit file.
revol_bal	How much debt does the borrower currently posses.
revol_util	What is the proportion of credit utilized by the borrower.
total_acc	Number of credit accounts that a borrower has open on their credit file.
pub_rec_bankruptcies	Number of times a borrower has declared bankruptcy.
dti_lc	Debt burden of monthly lending club installment payments relative to a borrowers income.
loan_status	Was the loan repaid, or did the borrower default.

# Only 15.1% of all loans default.

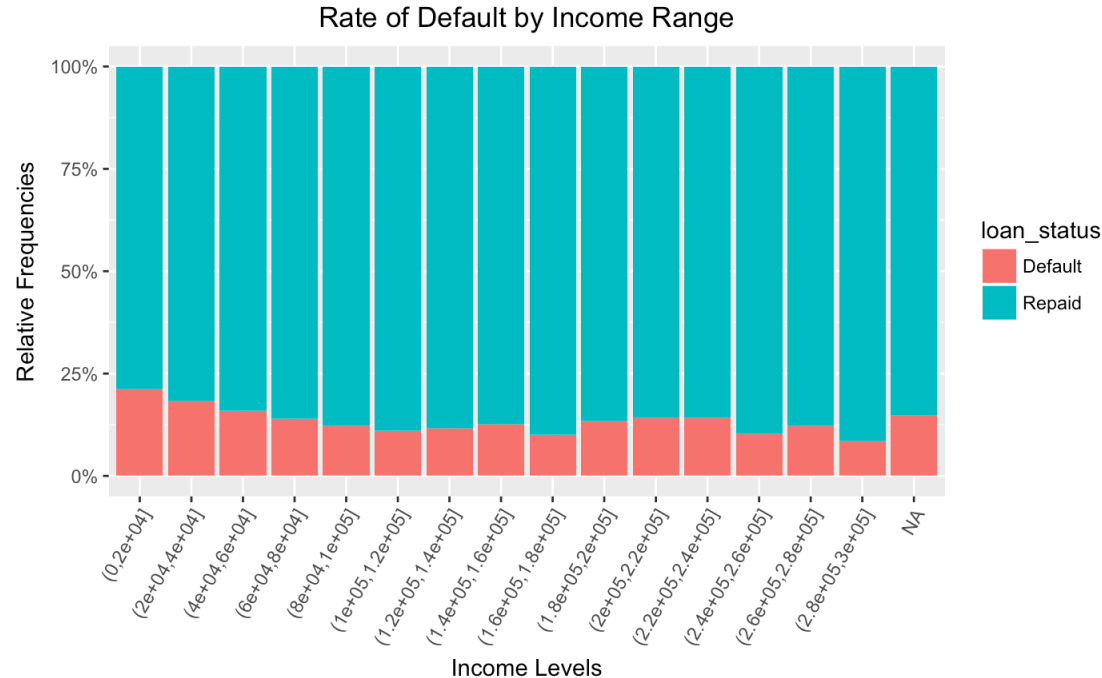


# The rate of default increases by loan grade.



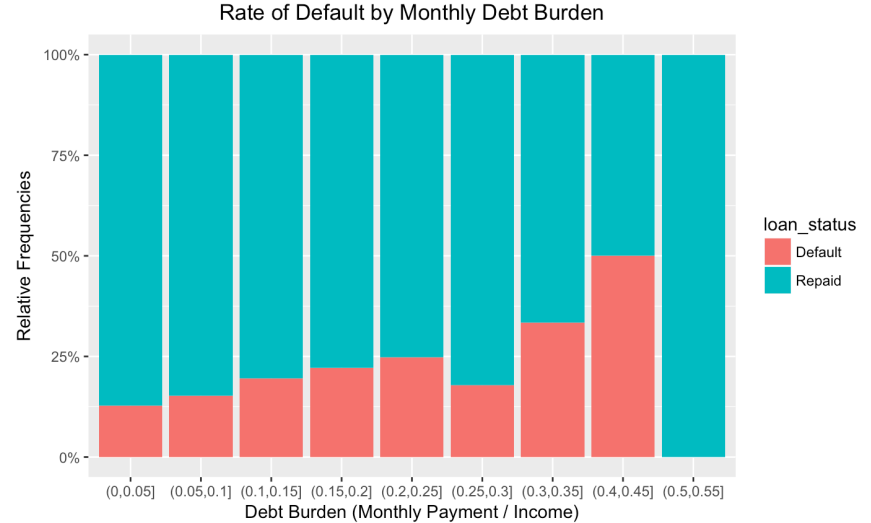
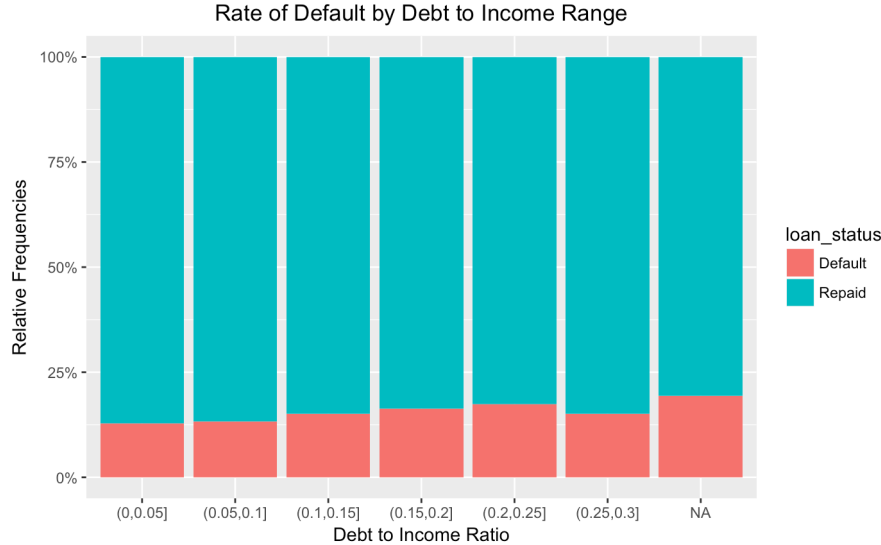
More interestingly we can see that the rate of default goes up substantially with the grade of the loan. The A1 loans rarely default, but the G rated loans default almost 1/3 of the time.

# Income levels play a smaller part in loan default.



Income does not seem to have a substantial impact on the rate of default. The rate decreases as incomes approach \$40K per year. Then however it seems to increase again.

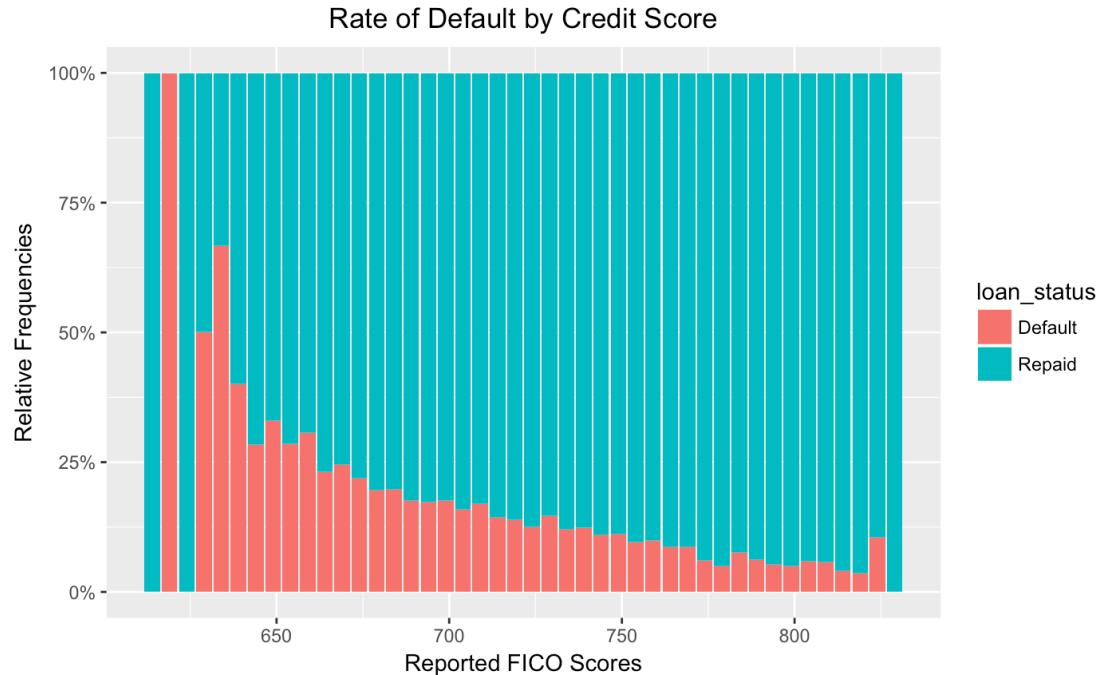
# Post funding debt burden has a larger impact than a debt to income ratio.



The rate of default does not seem to be highly impacted by Debt-to-Income as reported by the credit bureaus. The variance seems quite small, although steadily increasing as people carry more debt. This does not seem entirely logical. It is likely more important to assess their debt levels after a loan is received.

Debt burden seems to have a substantial impact on repayment. People whose post-funding debt burden is over 40% seem to default on loans nearly 50% of the time. If their debt burden is 50% for a lending club loan, it is unlikely that they will be able to keep up with other basic expenses such as housing, car and food payments.

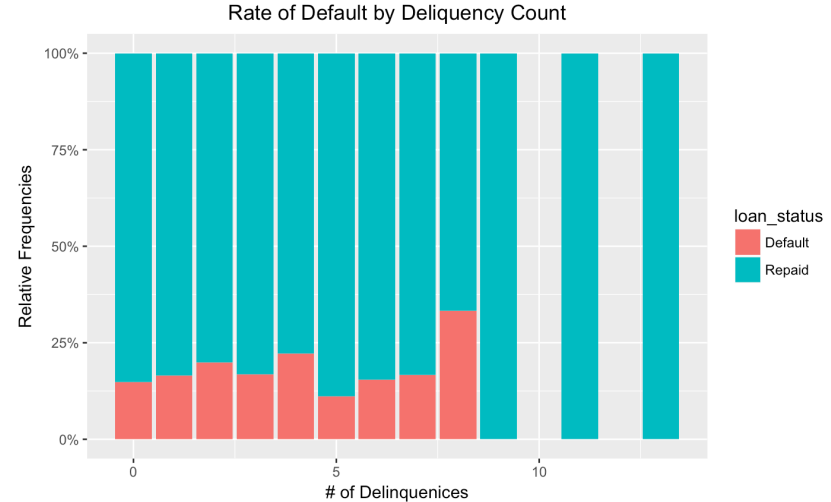
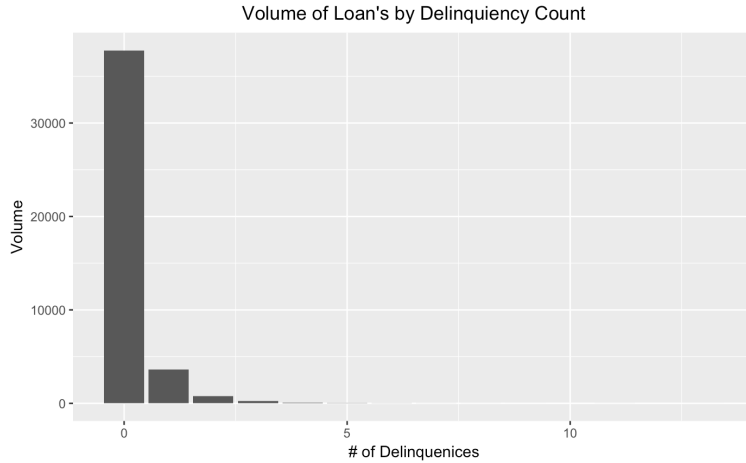
# Lower credit score applicants are clearly more risky.



It appears that defaults are highly influenced by FICO credit scores. Lower scores seems to have a much higher likelihood of default. It appears that investing in loans below a score of 640 is highly risky.



# While few borrowers have multiple delinquencies, they show risk increase.



It appears that having a delinquency in the last 2 years does have an impact on the rate of default. If borrowers have 0 defaults, they are .2% more likely to repay their loan. Having between 1 and 4 defaults in the previous 2 years seems to increase the rate of default at a linear rate. However, having over 5 defaults gets into the outlier territory and does not seem to have enough data to display any trends.

# **Machine Learning Models to Predict Loan Defaults**

# **We will do some final cleanup of the data to prepare for predictive modeling.**

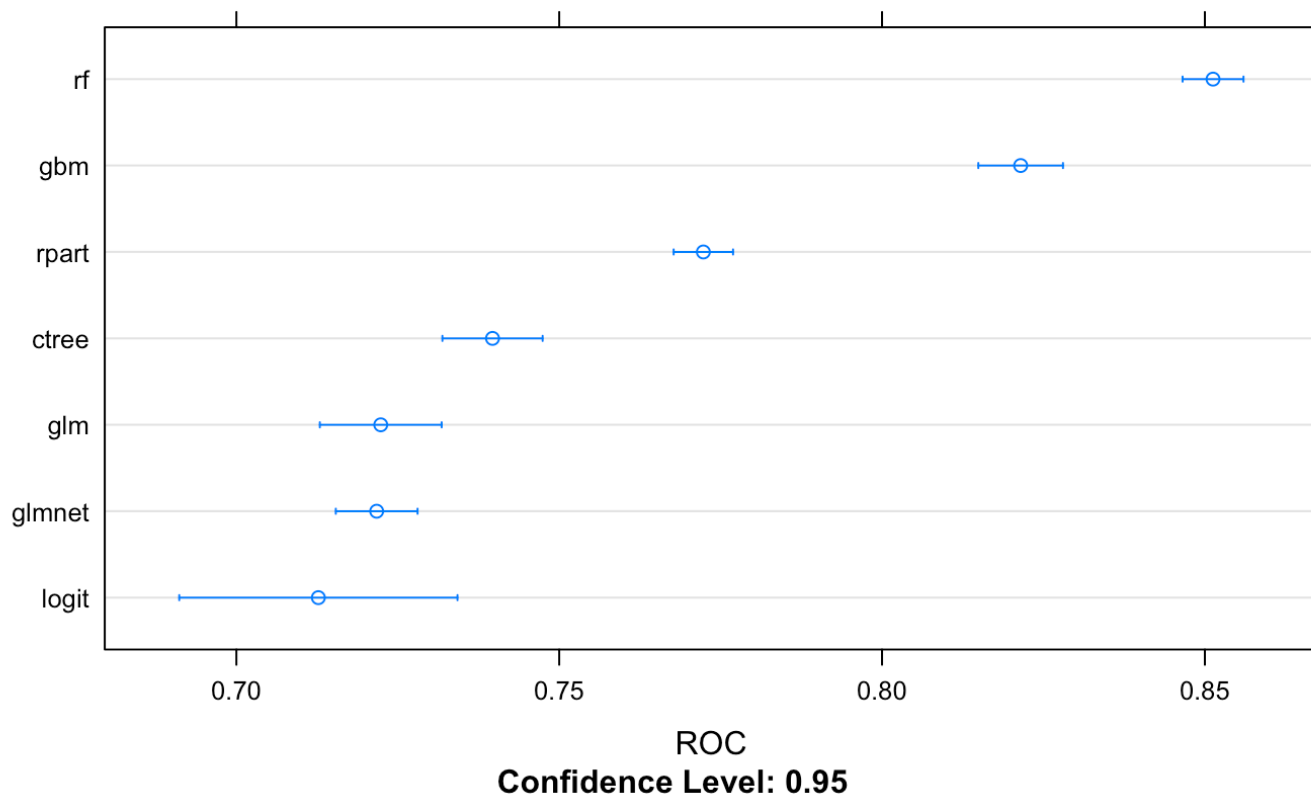
- We will use the SMOTE package to rebalance our data set in order to make up for the substantial class imbalance. Although this reduces the size of our data set, it creates a 50/50 balance between defaulted and repaid loans.
- We divide our data set into a 70% training set and a 30% test set.
- We set universal parameters for the Caret function to make sure that we can do apples to apples comparison between algorithms. We will use ROC as our comparison metric between algorithms.
- We will utilize 10 fold cross validation in our models.

# We will test various algorithms:

- **glm** - Generalized Linear Model
- **glmnet** - Lasso and Elastic-Net Regularized Generalized Linear Models
- **LogitBoost** - Boosted Logistic Regression
- **ctree** - Conditional Inference Tree
- **rpart** - Categorization and Regression Trees
- **rf** - Random Forest
- **gbm** - Stochastic Gradient Boosting

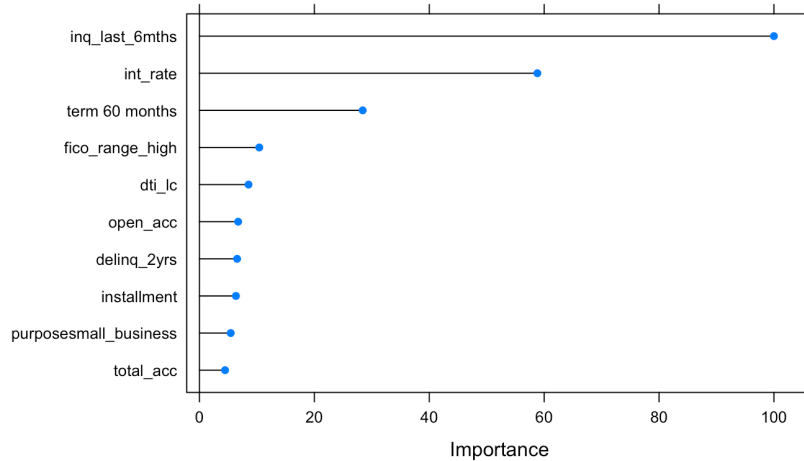
**The results of the models  
are highly performant.**

# Random Forest was the best performing model based on ROC.

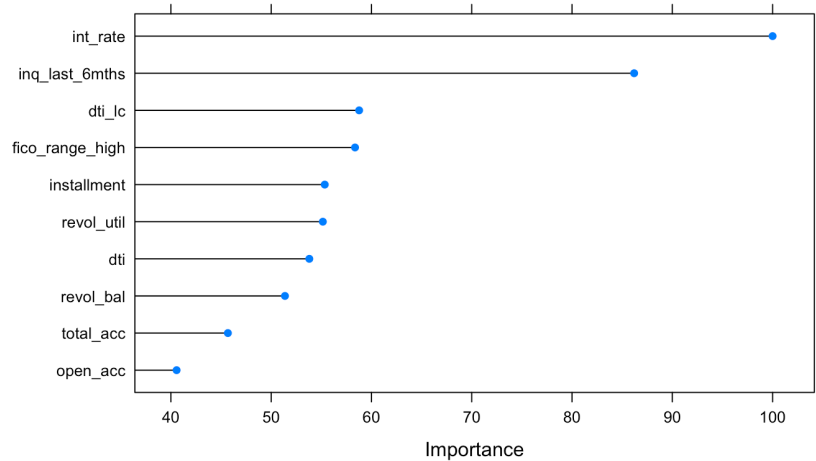


# Both the RF and GBM models had similar variables of importance.

GBM - Variable Importance



RF - Variable Importance



Both models seemed to find quite a bit of importance in the borrower's credit activity over the recent past. The most important variable it found were the number of credit inquiries. This signifies that a borrower may have been applying for more traditional loans elsewhere unsuccessfully before utilizing the LendingClub platform as a last resort. The interest rate and term of the loan also seem to be very important predictors in the likelihood to default.

# Predictions - GBM

Reference		
Prediction	Default	Repaid
Default	2714	869
Repaid	1139	2986

Accuracy :	0.7395
95% CI :	(0.7295, 0.7493)
No Information Rate :	0.5001
P-Value [Acc > NIR] :	< 2.2e-16

Kappa :	0.479
McNemar's Test P-Value :	1.937e-09

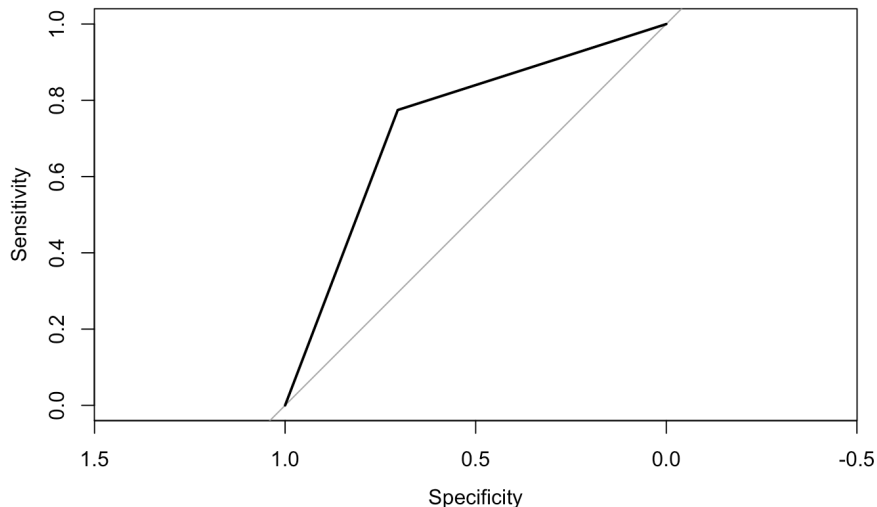
Sensitivity :	0.7044
Specificity :	0.7746
Pos Pred Value :	0.7575
Neg Pred Value :	0.7239
Prevalence :	0.4999
Detection Rate :	0.3521
Detection Prevalence :	0.4648
Balanced Accuracy :	0.7395

'Positive' Class :	Default
--------------------	---------

Area under the curve :	0.7395
------------------------	--------



The GBM model performed well, with an accuracy of 73%. The sensitivity was .67 and the specificity was .79. The AUC for predicted values was .7281. What is troubling is that the model still predicted that 1139 loans would be repaid, but in fact defaulted. This is a very high false negative rate of 14.7% which would result in a fairly substantial loss of principal. By adjusting the cutoff value, it would be likely that we could improve this false negative rate, as false positives would incur lower losses for the investor.



# Predictions – Random Forest

```
Reference
Prediction Default Repaid
Default      2860   795
Repaid       993  3060

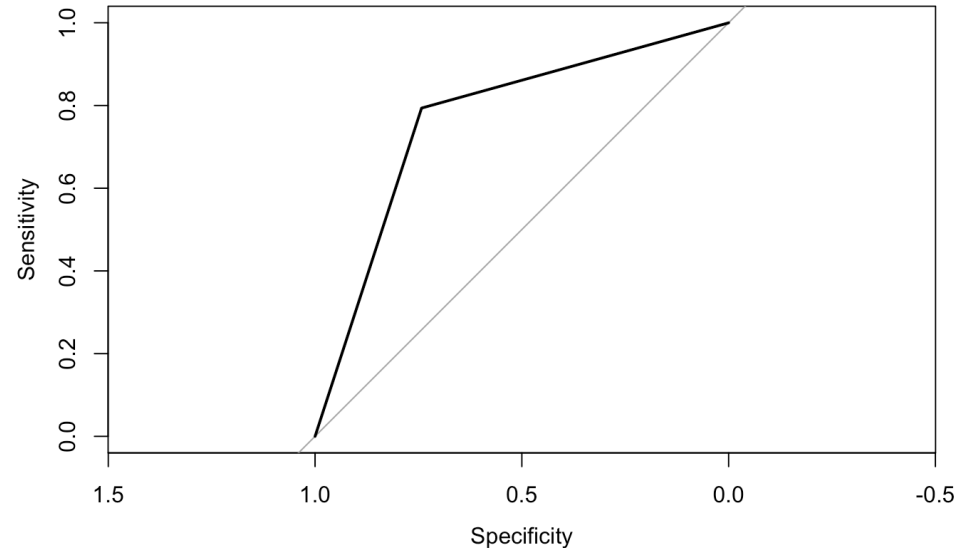
Accuracy : 0.768
95% CI : (0.7584, 0.7774)
No Information Rate : 0.5001
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5361
McNemar's Test P-Value : 3.179e-06

Sensitivity : 0.7423
Specificity : 0.7938
Pos Pred Value : 0.7825
Neg Pred Value : 0.7550
Prevalence : 0.4999
Detection Rate : 0.3710
Detection Prevalence : 0.4742
Balanced Accuracy : 0.7680

'Positive' Class : Default

Area under the curve: 0.768
```



The Random Forest model performed well, with an accuracy of 75%. The sensitivity was .742 and the specificity was .79. The AUC was .7557. While the accuracy was only slightly higher than the GBM model predictions, the false negative rate was much lower with 12.8% with 993 predicted repaid loans defaulting. While better than GBM, this is also still a very high false negative rate which would result in a fairly substantial loss of principal. By adjusting the cutoff value, it would be likely that we could improve this false negative rate, as false positives would incur lower losses for the investor.

**Random Forest was the clear winner amongst all models, with GBM in a very close 2<sup>nd</sup> place.**

**The results of the models  
are highly performant.**

# **The analysis of the Lending Club data set proves that there are clear predictive patterns in the provided data sets:**

- Individuals with high debt burdens are more likely to default.
- Borrowers with substantial credit history activity in the past 6 months seem to be much more likely to default.
- Lending Club appears to be the loan source of last resort for lower quality borrowers.
- Low quality borrowers prefer longer 60 month term loans.

# **Using machine learning - investors can predict the likelihood of default with a very high accuracy of nearly 75%.**

This is much better than an investor could do by merely guessing or flipping a coin. It would suit an investor to use these models before making any investments.

In future iterations of this project, we can put together an API that automates the investment process to auto-invest in less risky loans. This will likely create a balanced portfolio with lower levels of risk than the average investor.