IDS 572 Assignment 1

Britney Scott, Abdullah Saka, Shourya Narayan, Chaitra Srirama

2/8/2020

Background

LendingClub is an American peer-to-peer lending company that offers an online platform for matching borrowers seeking loans and lenders looking to make an investment. It provides an online platform which enables borrowers and investors to pair with each other. Both individuals and institutions can participate as investors if they satisfy financial stability standards put forth by LendingClub ("Lending Club" 5-6).

LendingClub is appealing to investors because they can choose how much to fund each borrow at $25 increments ("Alternative Investments"). Investors who hold diverse portfolios with LendingClub historically have a positive return ("Your Return"). Investors have control over the amount of risk they choose to take on, and have access to risk grades from LendingClub. LendingClub grades all loans from A to G, with each grade being further divided into five subgrades based on factors such as the borrower's FICO score and loan amount ("LendingClub" 8-9). Because the notes have the status of unsecured creditors, there is a risk that investors may lose all or part of the money if LendingClub becomes insolvent, even if the ultimate borrower continues to payback money ("LendingClub" 12).

Interest rates vary 6.03% to 26.06% between different types of loans and depend on a large number of factors regarding the borrower ("LendingClub" 3). A background check performed by LendingClub takes into consideration the borrower's credit score, credit history, income, and other attributes which help to determine the loan grade. The minimum credit criteria for borrowers to obtain a loan is:

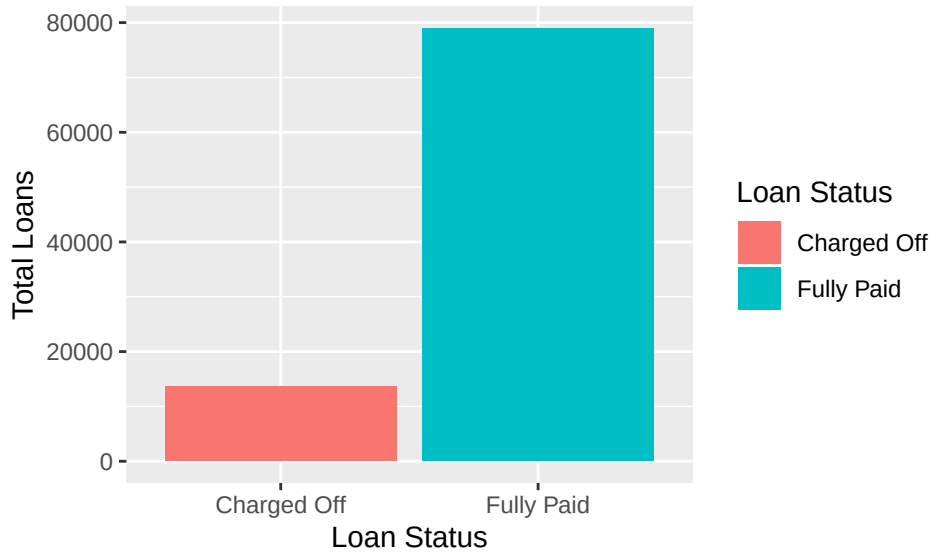
- A minimum FICO score of 660
- Below 35% debt-to-income ratio excluding mortgages
- Good debt-to-income ratio including mortgages
- At least 36 months of credit history
- At least two open accounts
- No more than 6 recent (last 6 months) inquiries ("LendingClub" 6)

LendingClub makes money by charging fees to both the borrowers and the lenders. Borrowers pay an origination fee when the loan is given, and investors pay a service fee of 1% ("LendingClub" 10). LendingClub also charges investors collection fees when payments are missed by the borrower, if applicable ("Interest Rates and Fees").

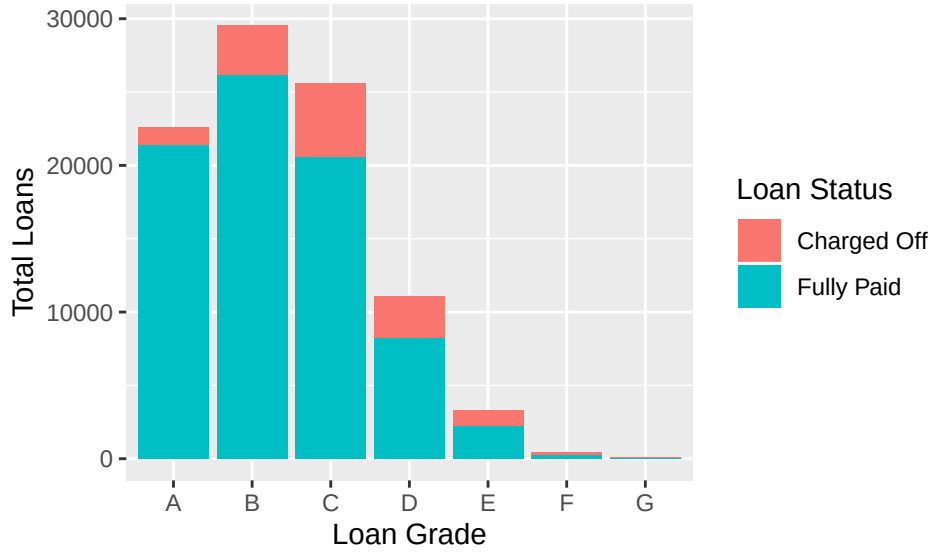
Data Exploration

The analysis will begin with some exploration of the provided data. The output vatiable indicates whether a loan defaulted or not.

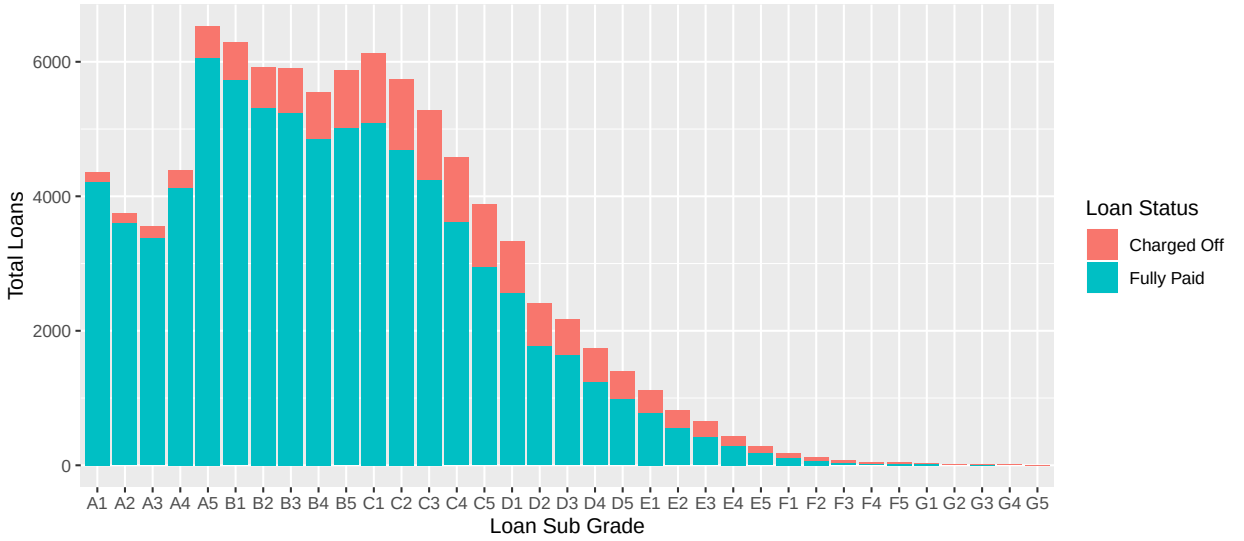
There are 13,652 defaulted loans in the dataset and 78,972 loans which were fully paid. About 14.74 per cent of the data represents defaulted loans.



Loan grade seems to correlate with loan defaulting, as evident in the following graph. This is to be expected, because loans with better grades such as 'A' and 'B' are less risky. Only 5.17 per cent of the A grade loans defaulted as opposed to 45.07 per cent of G grade loans, the lowest grade.

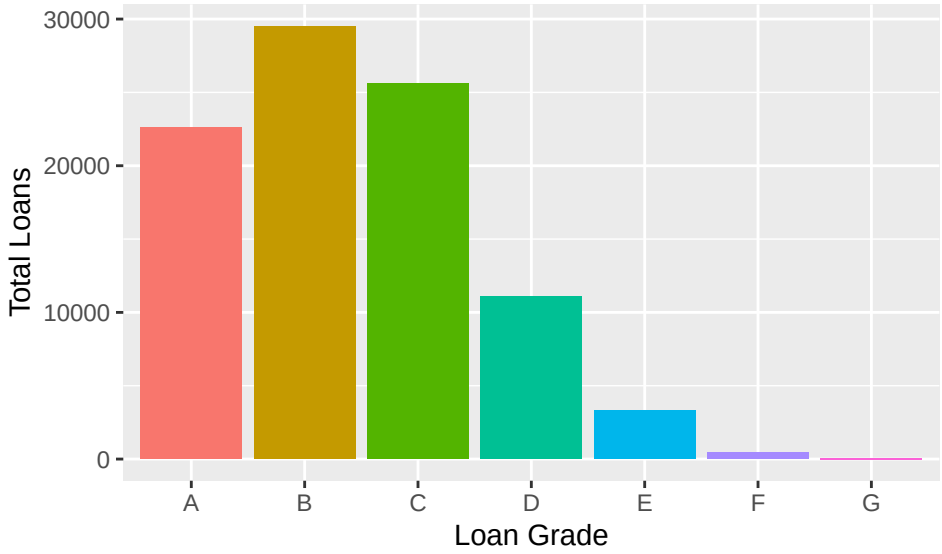


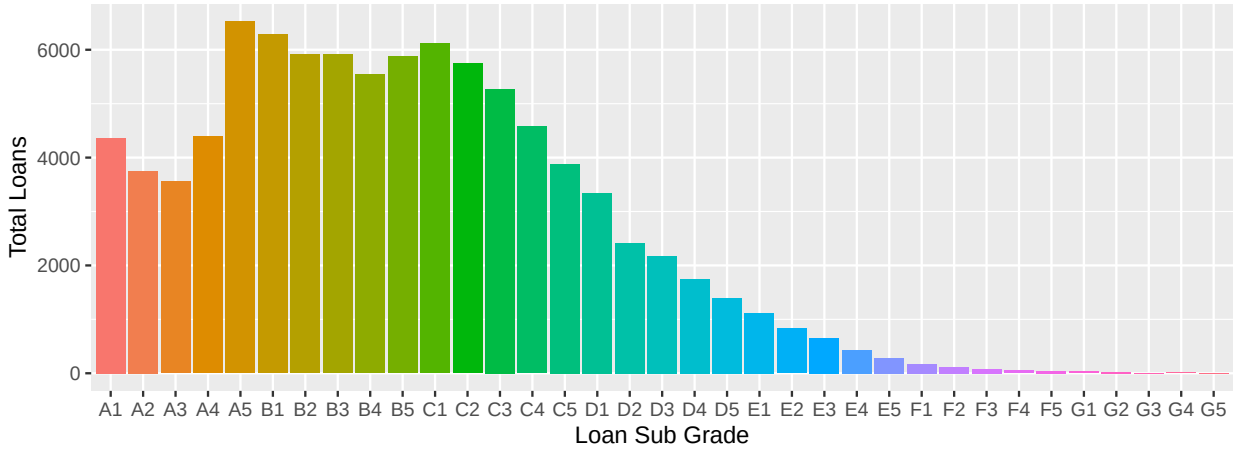
Taking a closer look at the subgrades, we see even more variation. Within the 'B' rating, for example, 8.96 per cent of B1 rated loans, 11.39 of B3 rated loans, and 14.46 per cent of B5 rated loans defaulted. This, again, is to be expected as the ratings of the loans get progressively lower.



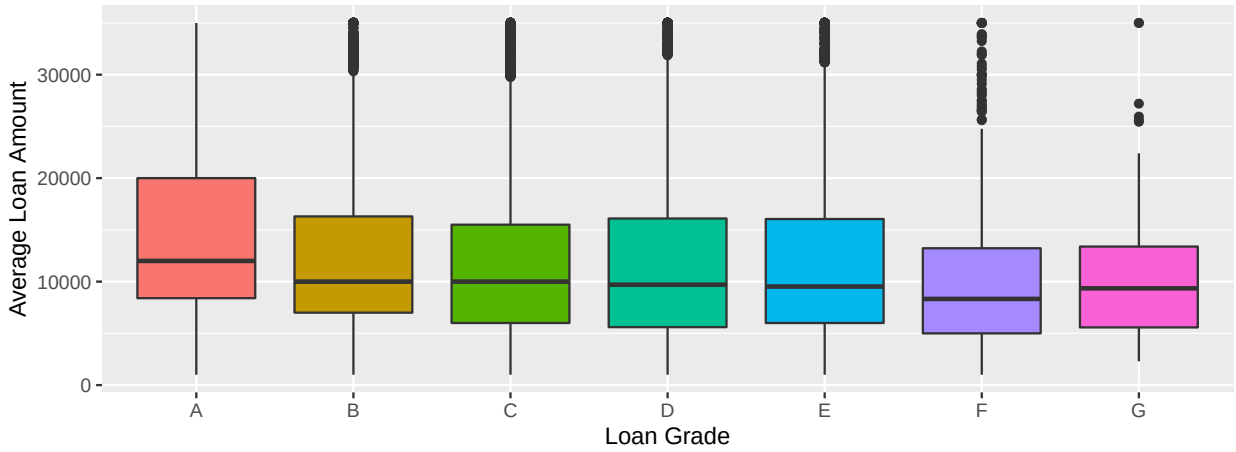
The number of loans within each grade category vary quite a bit, with B loans having the highest count of 29,523 loans. E, F and G categories contain only 3,309, 463, and 71 loans respectively.

The following plots show the number of loans in each grade, as well as in each subgrade. The large variation is evident.





We also wanted to examine the average loan amount for each grade in the data. As observed from the plot , the average loan amount decreases as the grade worsens. This is to be expected as investors would invest in lower amount of loans as the the grade worsens.

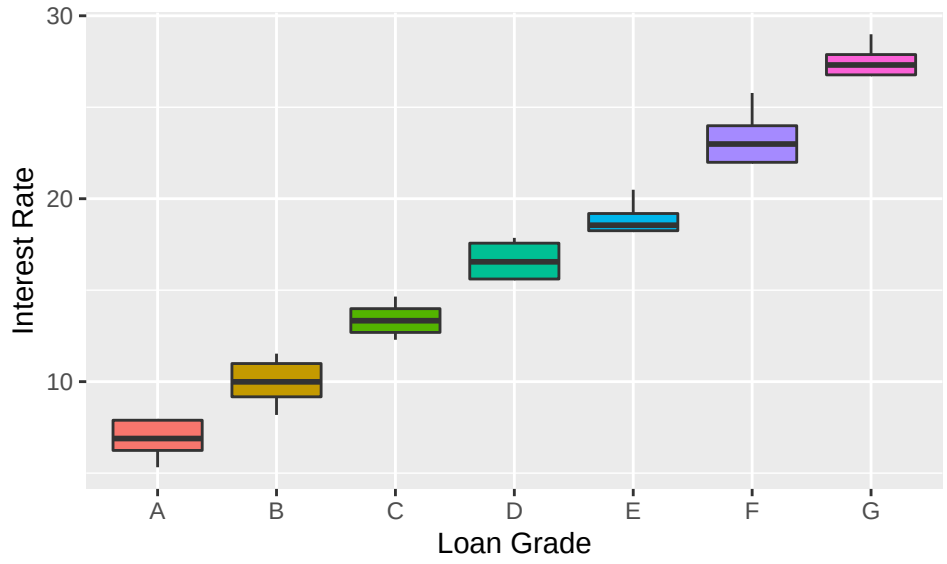


Interest rate varies drastically by the grade of the loan, as shown in the table below The same applies when subgrades are examined, and a steady increase in the interest rate can be seen with each step lower in subgrade. This is to be expected, since a lower grade indicates higher risk and therefore requires a higher rate of return.

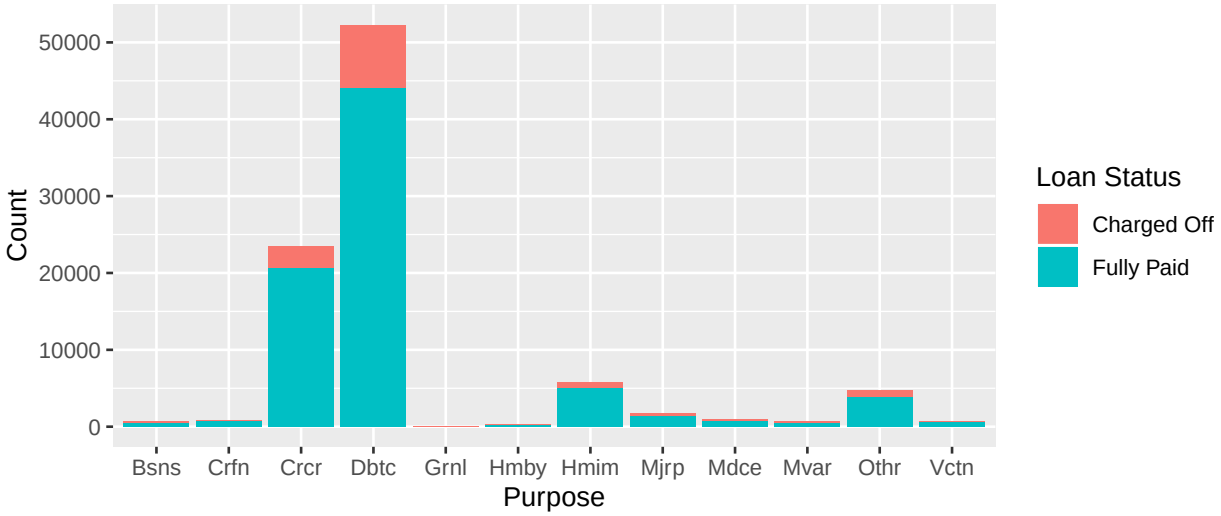
Loan Grade	Average Interest Rate
A	6.839653
B	9.932979
C	13.252912
D	16.674971
E	18.969649
F	23.247862
G	27.374789

Loan Sub Grade	Average Interest Rate
A1	5.321119
A2	6.240192
A3	6.889587
A4	7.259016
A5	7.890000
B1	8.179653
B2	9.170000
B3	9.988650
B4	10.985499
B5	11.528117
C1	12.287944
C2	12.684176
C3	13.328611
C4	13.988255
C5	14.645541
D1	15.610000
D2	16.542175
D3	16.990000
D4	17.563366
D5	17.851516
E1	18.238994
E2	18.550700
E3	19.192535
E4	19.991155
E5	20.994806
F1	21.990000
F2	22.990000
F3	23.990000
F4	24.990000
F5	25.780000
G1	26.770000
G2	27.310000
G3	27.880000
G4	28.490000
G5	28.990000

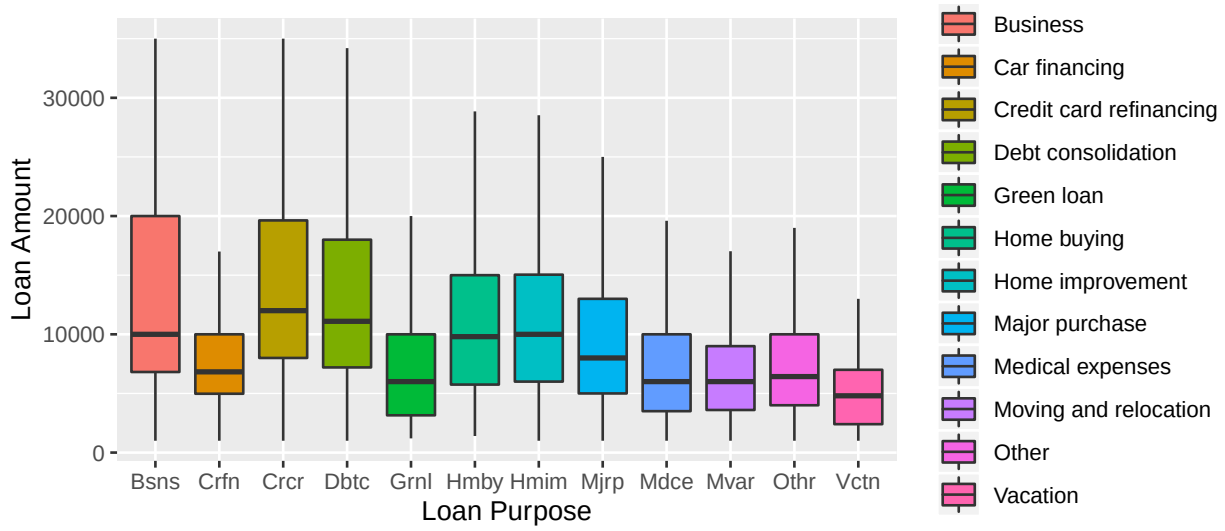
The following boxplot helps to illustrate the increase in interest rate as the grade of the loan worsens.



It's also important to look at what people are borrowing their money for. The vast majority of the loans in the dataset are for debt consolidation, with credit card refinancing in second place. The above graph shows the count of each purpose, as well as the proportion of that type of loan that defaulted. Credit card refinancing has the lowest default rate in the dataset. The highest default rate is for green loans, but there are only 59 total in the dataset of this category.



The amount of money given varies depending on the purpose of the loan. The following boxplot illustrates these differences well. Vacation loans have the smallest average amount, while credit card refinancing loans are typically quite large.



We also checked to see whether there was any change in loan purpose across grades. Debt consolidation is consistently the most frequent purpose across different grades, and green loans are always the rarest.

	A	B	C	D	E	F	G
Business	25	77	219	232	137	38	11
Car financing	247	313	233	86	38	8	0
Credit card refinancing	8615	8625	4678	1228	259	16	3
Debt consolidation	11082	16494	15618	6861	1996	233	29
Green loan	3	8	17	22	6	3	0
Home buying	12	43	82	92	70	28	10
Home improvement	1564	1886	1515	645	188	22	0
Major purchase	461	549	494	192	71	9	2
Medical expenses	107	245	372	188	53	8	3
Moving and relocation	23	98	299	230	82	18	2
Other	396	1014	1750	1107	374	74	10
Vacation	56	171	319	188	35	6	1

Finally, we exawmined annual return for various loans. We can calculate annual return for each loan using the following equation:
$((TotalPayment - FundedAmount)/FundedAmount) * (12/36) * 100$

Comparing the average return to the average interest rate, the two are negatively correlated. Across the different loan grades, as the interest rate increases, the annual return is decreasing. The average annual return of some of the lowest graded loans is even negative. This makes sense since we know the loans with lower grades are more likely to default. For the most part, the annual return increases as subgrade worsens too. The difference between interest rate and annual return is the smallest for the loans with better grades.

Loan Grade	Average Interest Rate	Average Annual Return	Difference
A	6.839653	2.2726688	4.566984
B	9.932979	2.5188551	7.414124
C	13.252912	2.2559848	10.996927
D	16.674971	1.9790544	14.695916
E	18.969649	1.2404470	17.729202
F	23.247862	-0.8086581	24.056520

Loan Grade	Average Interest Rate	Average Annual Return	Difference
G	27.374789	-0.1643031	27.539092

Loan Sub Grade	Average Interest Rate	Average Annual Return	Difference
A1	5.321119	2.0152225	3.305896
A2	6.240192	2.1950239	4.045168
A3	6.889587	2.3341449	4.555442
A4	7.259016	2.3552057	4.903810
A5	7.890000	2.4003205	5.489680
B1	8.179653	2.2749860	5.904667
B2	9.170000	2.4667824	6.703218
B3	9.988650	2.5716188	7.417031
B4	10.985499	2.7489118	8.236587
B5	11.528117	2.5620572	8.966060
C1	12.287944	2.5117403	9.776203
C2	12.684176	2.2944368	10.389739
C3	13.328611	2.3188210	11.009790
C4	13.988255	2.1402260	11.848029
C5	14.645541	1.8469623	12.798579
D1	15.610000	2.1199230	13.490077
D2	16.542175	1.8607832	14.681392
D3	16.990000	2.4726608	14.517339
D4	17.563366	1.5534816	16.009884
D5	17.851516	1.6114281	16.240088
E1	18.238994	1.1567569	17.082237
E2	18.550700	1.2429400	17.307760
E3	19.192535	1.2089593	17.983575
E4	19.991155	1.3314083	18.659746
E5	20.994806	1.4955443	19.499261
F1	21.990000	0.0656086	21.924391
F2	22.990000	-1.2789691	24.268969
F3	23.990000	-0.8775832	24.867583
F4	24.990000	0.6068922	24.383108
F5	25.780000	-4.6783841	30.458384
G1	26.770000	-1.4032071	28.173207
G2	27.310000	-1.7841390	29.094139
G3	27.880000	1.8874117	25.992588
G4	28.490000	4.9065252	23.583475
G5	28.990000	3.2243207	25.765679

As an investor, the type of loan you want to invest in depends on the level of risk you are willing to take on. While the lower grade loans are riskier, the potential return is clearly higher since the interest rate negatively correlated with the grade. We would choose the higher grade loans because we are not as interested in the risk associated with the lower loan grades.

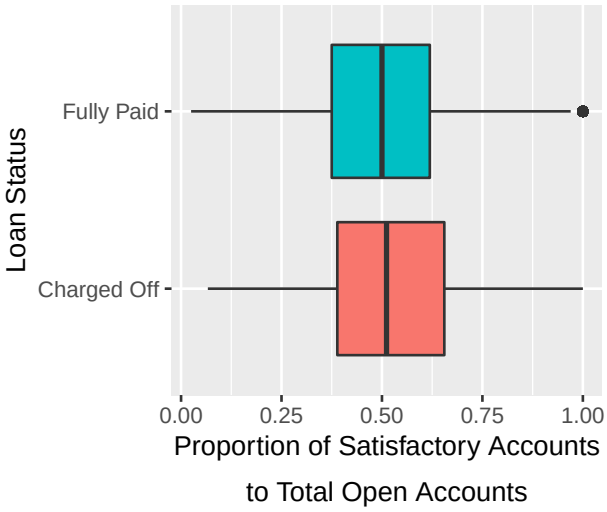
Variable Exclusion and Manipulation

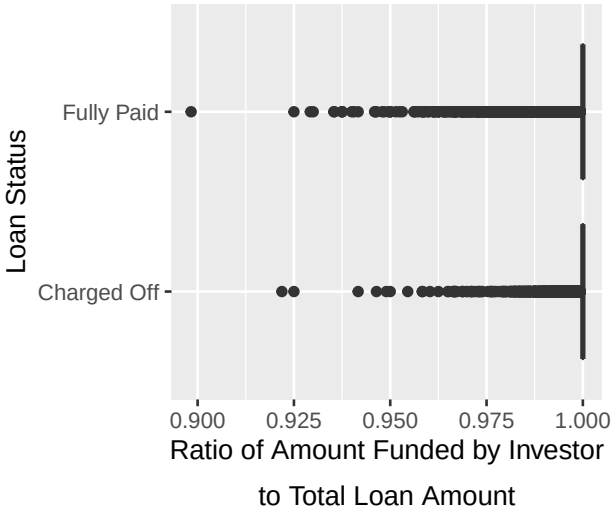
We chose to add in a few additional derived attributes.

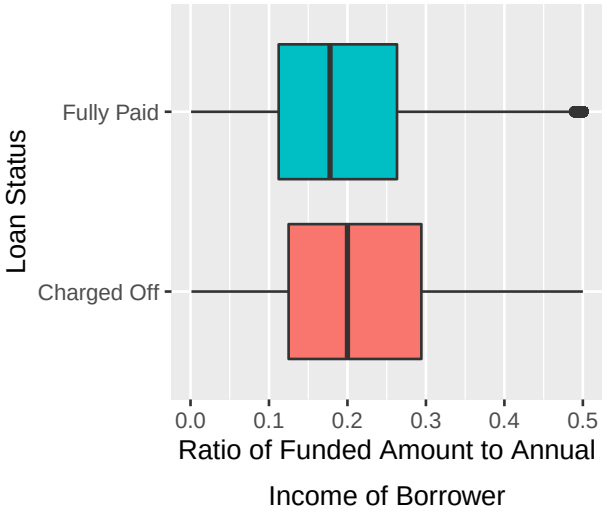
- Proportion of satisfactory bankcard accounts

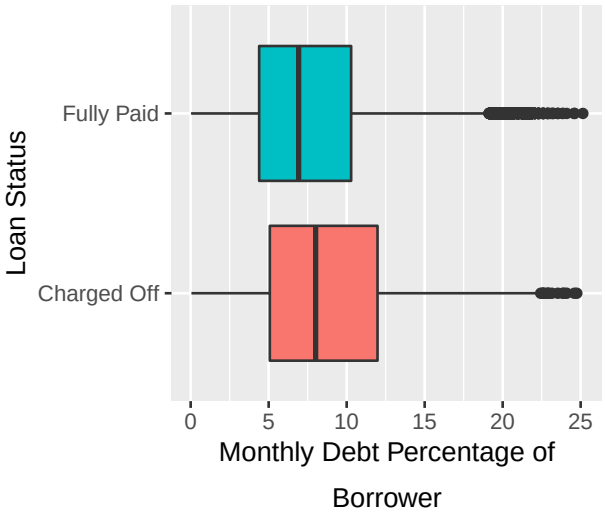
- Proportion of open accounts that are satisfactory
- Ratio of amount funded by investor to total loan amount
- Ratio of funded amount to annual income of borrower
- Monthly debt percentage of borrower
- Ratio of open acocunts to total accounts

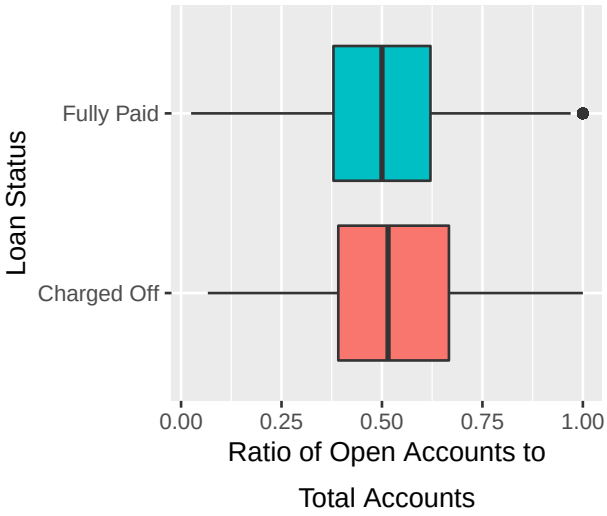
Boxplots for all of these attributes show how they vary between loans that were paid off and ones that defaulted.











We decided to remove all of the attributes with more than 60% missing values. This decreases the number of independent variables from 150 to 92.

Next, some variables which may cause leakage need to be removed. These are variables which have been updated after the loan was given. For example, FICO score is updated every time an individual goes through a credit check, so all variables including FICO score have been removed. Other variables which are updated include total payment and interest payments received to date. After removing these unnecessary columns, the total number of independent variables decreases further to 60 columns.

Next, missing values must be addressed. For some columns, the absence of a value is meaningful. For example, a missing value for months since recent inquiry indicates that there has not been an inquiry. We cannot fill these fields with a zero, as that would indicate a very recent inquiry. For such columns, we filled the missing values with a number much higher than the maximum value for the column. Other columns with which we used this approach include months since oldest installment account opened, months since most rencent bankcard account opened, and months since last delinquency.

In other cases, the NA truly indicates a missing value. For these columns, we replaced the missing values with the median for that column. We used this approach for revolving line utilization rate, total open to buy on revolving bankcards, ratio of current balance to credit limit for all bankcard accounts, and percentage of bankcards over 75% percent of their limit.

Decision Tree Models

The first step of building decision tree models is splitting the data between training and testing sets. We chose to split the data at a ratio of 70:30.

Information Model

For the first decision tree, we used the information method with a minimum split of 30 and a complexity parameter of 0.0001. This performed at 88 per cent accuracy on the training set.

Metric	Result
Training Accuracy	0.88

As a different scenario, we have used 0.35 rather than 0.5 as a threshold. It indicates that when we determine the lower threshold for labeling, training accuracy is almost same with previous model, yet testing accuracy is likely to be lower (it is possible that distribution of different classes in terminal nodes are alike). Training and test accuracy of the model whose threshold is 0.35:

Metric	Result
Train Accuracy	0.88
Test Accuracy	0.81

The first model's accuracy seemed rather high and leads to concerns about overfitting. After generating the model, we pruned it using a complexity parameter of 0.0003 in order to keep it at a manageable size and avoid small nodes which can lead to overfitting and lower accuracy on the validation data.

This pruned model performed well on the training data, and has 86 per cent accuracy. On the testing data, the model performance decreases to 85 per cent accuracy.

The confusion matrix and accuracy of the first model after pruning for the training data:

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       54962        8698
  Charged Off        294         883
```

Metric	Result
Training Accuracy	0.86

The confusion matrix and performance metrics of the first model after pruning for the testing data:

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       23399        3884
  Charged Off        317         187
```

Metric	Result
Test Accuracy	0.85
Precision Score	0.86
Recall Score	0.99

Gini Model

Next, we created a second decision tree model using the same training and testing sets. All parameters were kept the same except for the method, which was changed from information to gini. Before pruning, this tree performed at 88 per cent accuracy on the training data.

Metric	Result
Training Accuracy	0.88

Once again, we chose to prune the tree to avoid overfitting. This model performs similarly on the training and testing data to the information model, with 86 per cent training and 85 per cent testing accuracy.

The confusion matrix and accuracy of the second model after pruning for the training data:

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       55004        8819
  Charged Off        252         762
```

Metric	Result
Training Accuracy	0.86

The confusion matrix and performance metrics of the second model after pruning for the testing data:

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       23476        3913
  Charged Off        240         158
```

Metric	Result
Test Accuracy	0.85
Precision Score	0.86
Recall Score	0.99

C5.0 Model

Next, we chose to run a model using C5.0 to see how it compared to the rpart models. We selected confidence factor as 0.45 and the number of trials as 3. Overall, the C5.0 decision tree model performs slightly worse than other models on the valdation data with 83 per cent accuracy.

The confusion matrix and accuracy of the C5.0 model for the training data:

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       54752        5974
  Charged Off        504        3607
```

Metric	Result
Training Accuracy	0.9

The confusion matrix and performance metrics of the C5.0 model for the testing data:

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       22662        3680
  Charged Off       1054         391
```

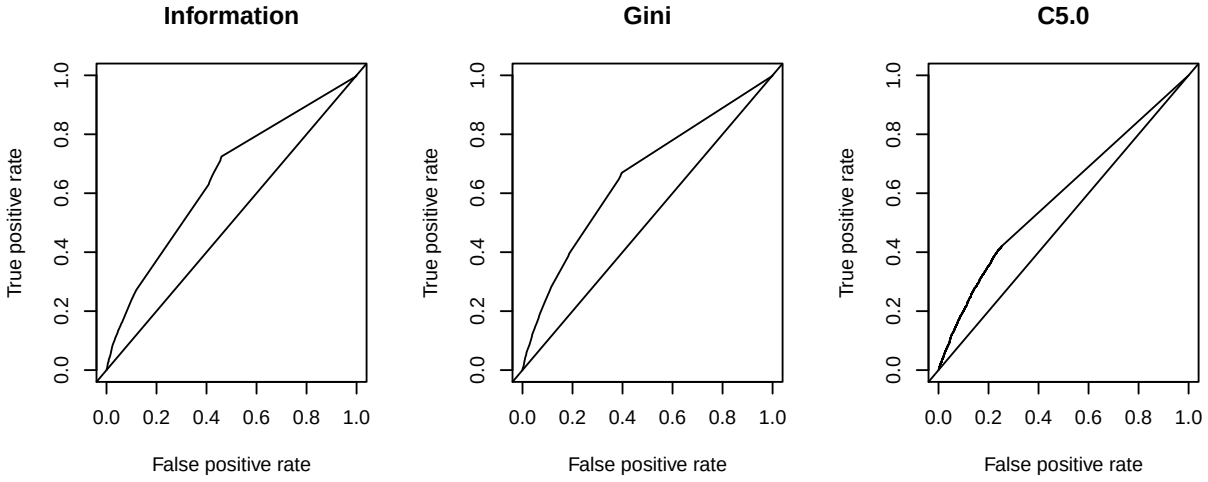
Metric	Result
Test Accuracy	0.83
Precision Score	0.86
Recall Score	0.96

Comparing the Models

We created an additional scenario for each of the three models by modifying the parameters to see how they would affect the metrics. For the rpart models, we modified the complexity parameter. Decreasing it lead to a slight increase in recall score but a slight decrease in precision score for both the information and the gini models. For C5.0, we modified both the trials and the confidence factor and were able to slightly improve the test accuracy and the recall score of our previous C5.0 model. Overall, the rpart models are slightly better performing than C5.0.

	Test Accuracy	Precision Score	Recall Score
Information, minsplit=30, cp=0.0001	0.85	0.86	0.99
Information, minsplit=30, cp=0.001	0.85	0.85	1.00
Gini, minsplit=30, cp=0.0001	0.85	0.86	0.99
Gini, minsplit=30, cp=0.001	0.85	0.85	1.00
C5.0, trials=3, cf=0.45	0.83	0.86	0.96
C5.0, trials=8, cf=0.55	0.84	0.86	0.97

ROC curves for each of the models are displayed below.



Finally, we checked which variables are more important for decision tree. The model looks at the improvement measure to each variable in its split. The values of these improvements are summed up, and are then scaled relative to the best variable.

These are the top ten attributes which carry more weight than other attributes (more statistically significant).

	Overall
grade	100.00
sub_grade	100.00
int_rate2	100.00
acc_open_past_24mths	90.45
emp_length	67.08
mnthDebt	64.94
revol_util	58.25
purpose	57.87
avg_cur_bal	57.21
pct_tl_nvr_dlq	56.25
mths_since_recent_inq	52.92
home_ownership	52.74
num_actv_rev_tl	52.14
total_il_high_credit_limit	50.44
pub_rec_bankruptcies	50.07

Random Forest Model

In order to further improve the accuracy of our predictions, we built some random forest models. These have an advantage to a single decision trees because they generate multiple trees in order to create a more robust model. In order to maximize the performance of the tree, we decided to build several models with increasing number of trees to see which performs best. First, we build a model with 40 trees.

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       23626        3998
  Charged Off         80          68
```

Second, we built a smiliar model with 70 trees.

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       23658        4001
  Charged Off         48          65
```

Finally, we built a model with 200 trees.

```
             Reference
Prediction    Fully Paid Charged Off
  Fully Paid       23686        4037
  Charged Off         20          29
```

Looking at the three confusion matrices, it can be seen that increasing the number of trees improves the ability of the model to predict fully paid loans correctly. However, the amount of correctly predicted charged off loans decreases as the number of trees is increased. Predicting charged off loans incorrectly as fully paid is much more costly than predicting fully paid loans as charged off. Therefore, the smaller model is actually better in this scenario.

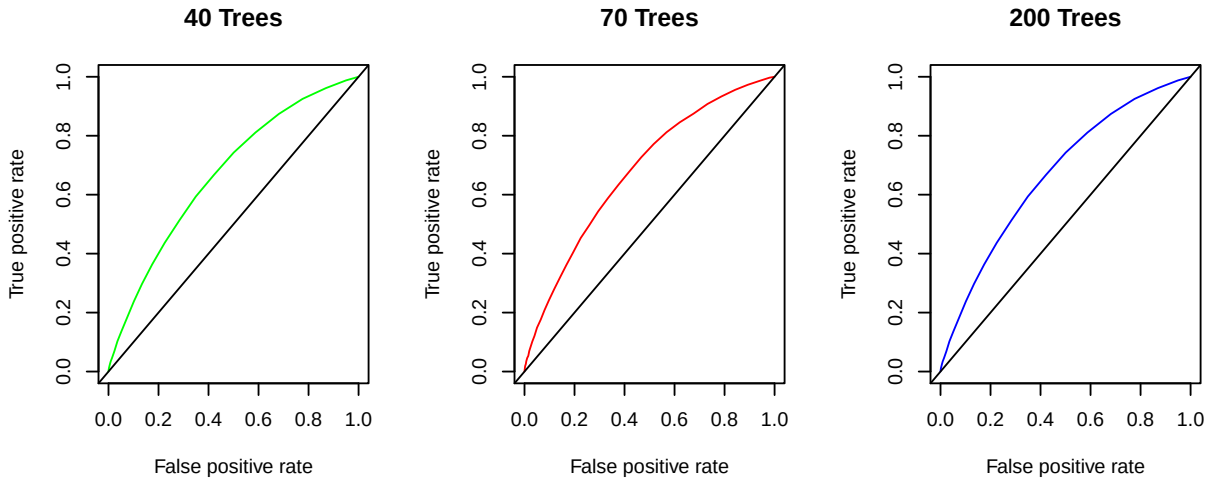
We saw something similar when comparing the performance metrics. 200 trees performs slightly worse than 40 and 70. 40 and 70 trees result in the same performance metrics, so it makes sense to use the less costly

model of 40 trees.Besides, when tree size increases in random forest, model's ability to predict 'charged off' is lower so that model affects investment decision negatively.

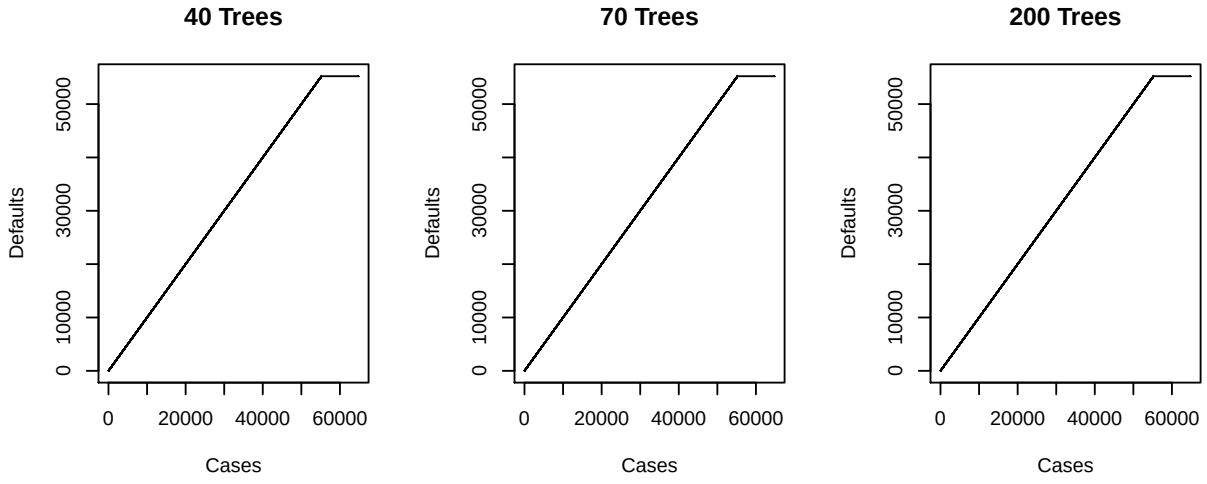
Overall, though, the random forest models do not performs much differently than the single trees.

Metric	40 Trees	70 Trees	200 Trees
Test Accuracy	0.85	0.85	0.85
Precision Score	0.86	0.86	0.85
Recall Score	1	1	1

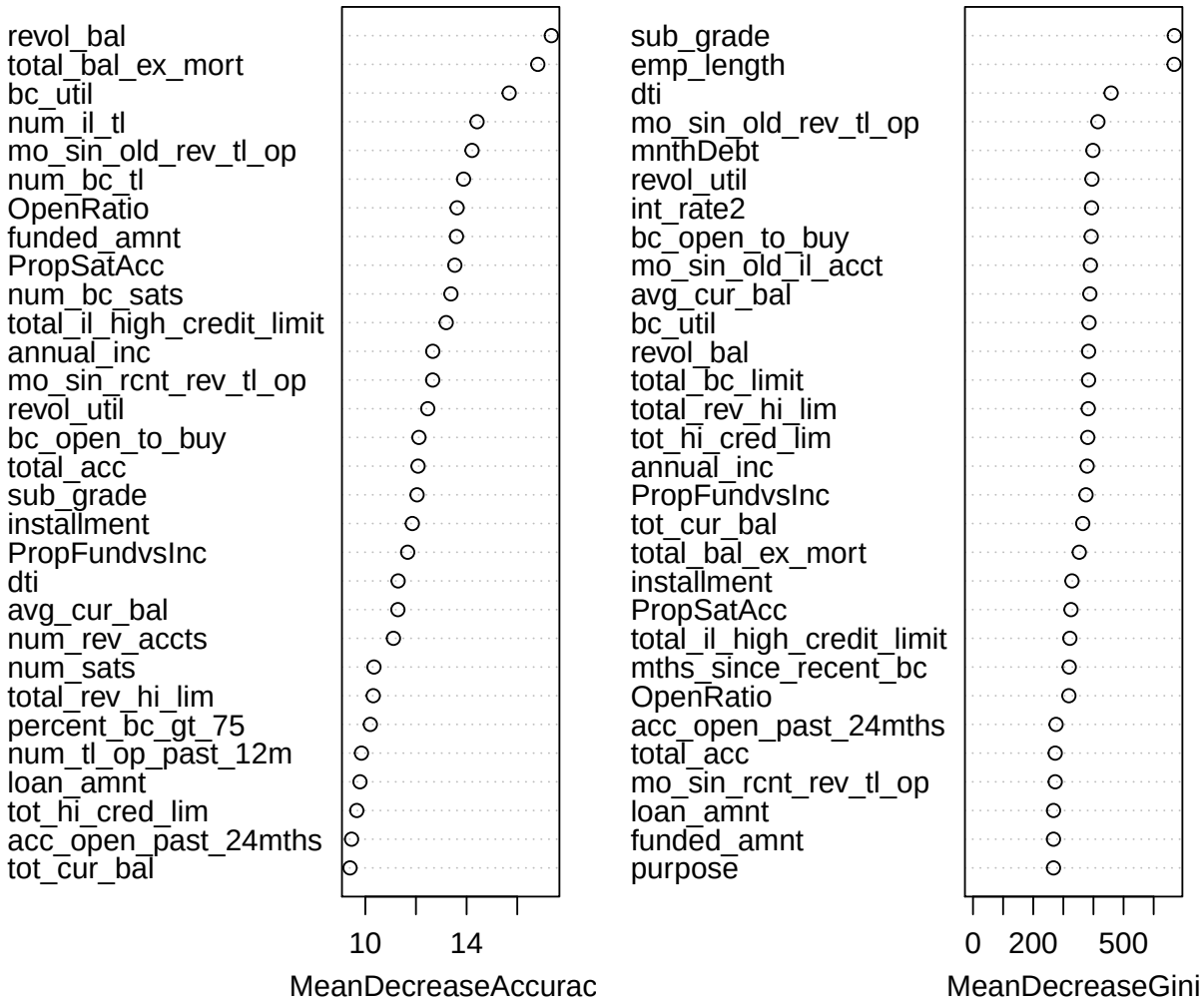
We plotted ROC curves for all random forest models and it can be clearly seen that there is no significant difference among models. However, the random forest model created with 40 trees is slightly better than other models. ROC curves for all three random forest models:



Lift curves for all three random forest models :



For the best random forest model, which had 40 trees, the most important variables are:



Variables' importance are calculated based on two different selection methods such as MeanDecreaseAccuracy and MeanDecreaseGini. When the MeanDecreaseAccuracy is determined during the out of bag error calculation phase, the MeanDecreaseGini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the random forest. Therefore, the significant variables vary from method to method. Besides, these significant variables in which random forest found are different than variables in which C5.0 determined.

Cost-Based Performance

To begin the cost analysis, we are calculating the return on the loans. To begin, we simply calculate the return based on the annual return of the loans, which was calculated during the data exploration.

Grade	Avg. Interest	SD of Interest	Avg. Amount	Avg. Payment	Avg. Return	SD of Return
A	6.839653	0.9264953	14517.58	15529.612	2.2726688	3.877575
B	9.932979	1.2220691	12689.20	13666.653	2.5188551	6.079256
C	13.252912	0.8422211	11986.27	12807.051	2.2559848	8.396707
D	16.674971	0.8401295	12284.90	12964.566	1.9790544	10.275733
E	18.969649	0.8756259	12694.85	13021.988	1.2404470	11.829564
F	23.247862	1.2633375	10518.52	9918.955	-0.8086581	13.809136
G	27.374789	0.6813744	10892.61	9987.720	-0.1643031	16.095761

This isn't the best approach, though, because sometimes the loans do not receive their full interest rate if the loan is paid back early. We can use the period between the loan issue and the last payment date to determine how long it took to pay each loan off, and then use this to calculate a more accurate return rate based on the actual term of the loan.

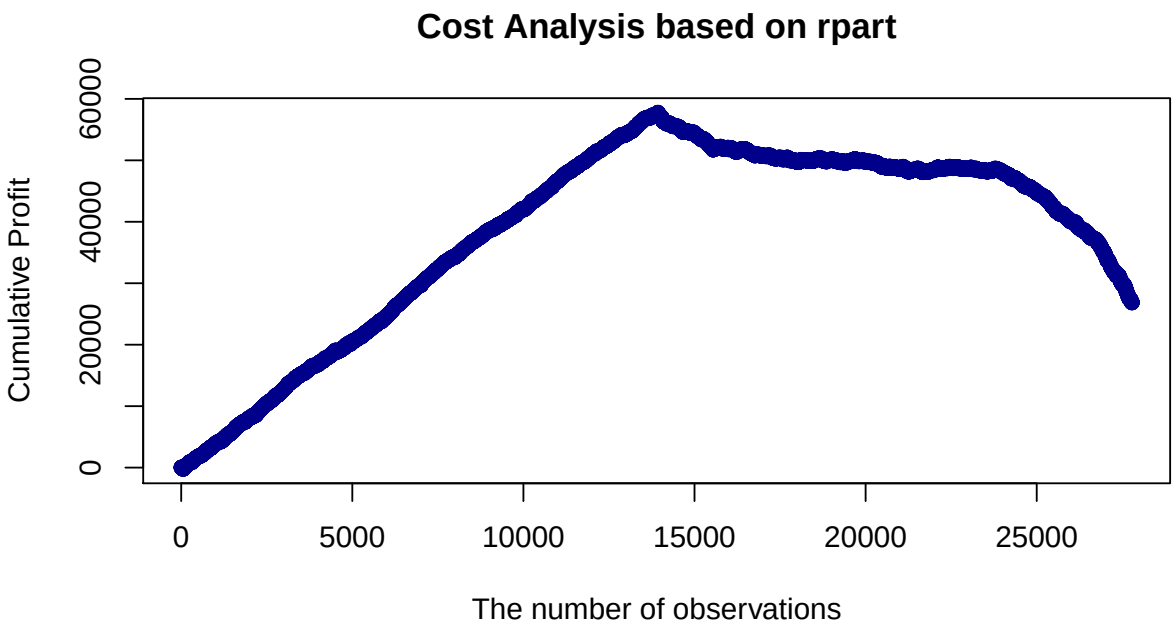
Grade	Default Rate	Avg. Interest	Avg. Return	Avg. Actual Term	Avg. Actual Return
A	0.0517020	6.839653	2.2726688	2.265799	3.618867
B	0.1140467	9.932979	2.5188551	2.255265	4.547946
C	0.1947961	13.252912	2.2559848	2.269992	4.966492
D	0.2558938	16.674971	1.9790544	2.266261	5.494573
E	0.3215473	18.969649	1.2404470	2.321960	4.945214
F	0.4362851	23.247862	-0.8086581	2.330850	3.811167
G	0.4507042	27.374789	-0.1643031	2.428902	4.428403

Finally, in order to determine the costs for the cost analysis, we can use this more accurate return on the loans based on their outcome status. The following chart shows about a 12.3 per cent loss on loans that are charged off, and a 7.5 per cent profit on paid off loans.

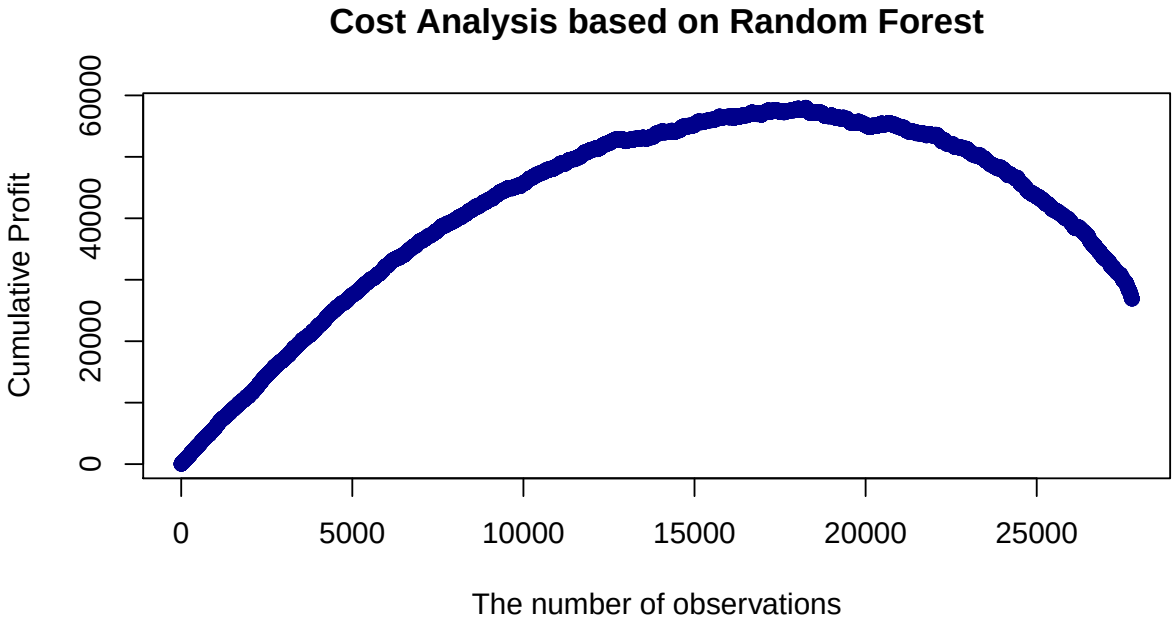
Loan Status	Interest Rate	Total Return	Total Actual Return
Charged Off	13.34667	-0.3690510	-12.301699
Fully Paid	10.95154	0.1432236	7.475575

We performed cost analysis regarding cost table above. Profit value is 8 and loss value (penalty) is 40. We selected bigger loss value since we wanted to decrease the risk. Also, threshold of labeling to points in random forest is 0.5. Lastly, we checked current interest rate of certificate of deposit. The nominal interest rate is 2%. This means you can get 5.6 profit if you put $100 into deposit account. Then, we plotted cumulative profit in this dataset. This plot demonstrates that we take into consideration 15,000 observations as an investment. After that point, risk of the investment increased gradually and you can face considerable loss.

Here is the cost table based on the rpart decision tree.



In contrast, here is the cost table based on the optimal random forest with 40 trees.



This plot points out that we should consider around 18,000 observations as an investment. After that point, risk of the investment increased gradually and it is more likely to face loans at risk of default.

According to cost tables, we can say that random forest is slightly better than the rpart pruned decision tree because it detects more labels correctly, allowing us to make better profits.

Conclusions

Detecting whether or not loans will default is important for all stakeholders in the LendingClub. Investors stand to lose about 12.3 per cent of their money when investing in loans that default, so accurately predicting this occurance is crucial. While most of the models in the report performed similarly, the best one we built is the random forest model with 40 trees. This model performs at 85 per cent accuracy on validation data, and while single trees meet similar performance standards, the random forest performs much better in the cost analysis.

References:

"Alternative Investments: How It Works." LendingClub, LendingClub Corporation, 2020, www.lendingclub.com/investing/peer-to-peer.

"Interest Rates and Fees." LendingClub, LendingClub Corporation, 6 Aug. 2019, www.lendingclub.com/investing/investor-education/interest-rates-and-fees.

"LendingClub." 424B3, U.S. Securities and Exchange Commission, 30 Apr. 2014, www.sec.gov/Archives/edgar/data/1409970/000119312514173269/d719822d424b3.htm.

"Your Return: Three Key Factors." LendingClub, LendingClub Corporation, 2020, www.lendingclub.com/investing/investment-performance.