Lending Club part A and B

Adrian Blamires
Michael Gannon
Jinrong Qiu

IDS 572

1. Introduction

Lending Club is a fintech company that started as a platform for peer-to-peer loans between investors and borrowers. Since their inception in 2005 they have grown to offer a variety of financial products and technology driven services. Borrowers were able to obtain personal loans for anywhere between $1,000 and $40,000, with a standard loan period of 3 years. Lenders were able to search loan listings based on the information provided by the borrower and use that information to determine if they were viable candidates for a personal loan. Most recently in 2020, they acquired Radius Bank thus ceasing all peer-to-peer lending[1].

Prior to this acquisition, Lending Club's business model was based on their ability to connect borrowers with individual investors. This data backed platform used analytic metrics, other than a traditional FICO score, to deem the creditworthiness of these borrowers[2]. Individuals once approved, receive loans divided up into $25 fragments with an associated risk profile determined by the company's analytics and algorithm[3]. These fragments are given a grade, A1-G5, which will dictate the risk and return on a lender's investment. The lower the grade, the higher risk and interest rate for the borrower.

In the fintech sector, peer-to-peer loans is a growing industry surpassing $67.93 billion in 2019 and the expectation that it will reach $558.91 billion in 2027. The main factors driving this growth are the lower operating costs and low market risk. With no physical infrastructure, and lower average workforce in comparison to traditional institutions, companies like Lending Club are able to maximize their profit while still offering much more competitive pricing. Peer-to-peer lending also reduces an investor's market risk. By creating a plan based on the data, market factors such as interest rate, unemployment rates, and property prices risk will not affect the repayment process[4].

Alternative lending companies, such as Lending Club, generate revenue by charging fees to both investors and borrowers. Borrowers seek loans for a number of reasons such as debt consolidation or paying down credit card balances. Investors are able to purchase loans after they have already been issued with the cash flow being serviced from the company in question. By acting as the intermediary between a borrower and investor, the customer relationship is maintained by Lending Club. For this service of connecting the two parties, Lending Club collects a service fee from the investor, and a loan origination fee from the borrower[5].

Peer-to-peer lending companies are essentially a marketplace for people to connect. They take the risk that a traditional bank would assume and transfer it to an individual investor. This transfer makes individual investors the main stakeholders in this transaction although the company hosting the platform does retain some inherent risk. They assume the role of market maker and are expected to classify and grade the borrowers profiles accurately so that investors know the extent of the risk that they are assuming. One company, Prosper, had up to half of their loans fail between 2006 and 2008 resulting in a class action lawsuit from investors[6]. By assuming the role of marketplace and market maker, companies

---

[1] https://en.wikipedia.org/wiki/LendingClub

[2] https://www.cnbc.com/2019/02/21/personal-loans-surge-to-a-record-138-billion-in-us-as-fintechs-lead-new-lending-charge.html

[3] https://www.businessmodelzoo.com/exemplars/lending-club/

[4] https://www.alliedmarketresearch.com/peer-to-peer-lending-market

[5] https://www.morganstanley.com/im/en-us/financial-advisor/insights/investment-insights/an-introduction-to-alternative-lending.html

[6] https://foundationcapital.com/wp-content/uploads/2020/04/FC_CharlesMoldow_TrillionDollarMarket.pdf

like Prosper and Lending Club rely on the strength of their analytics and algorithms to drive customer satisfaction by both the borrowers and the investors.

Ultimately, fintech companies that are built on alternative lending create an opportunity to disrupt traditional banking markets. Connecting individuals and providing them with the necessary analysis to make informed decisions allows for growing adoption across the industry. With Lending Club's acquisition of Radius they will remove the peer-to-peer aspect of their marketplace but they continue to rely on the analytics that is the foundation of their business.

2. Data exploration

(ai) What is the proportion of defaults ('charged off' vs 'fully paid' loans) in the data? How does default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why?

The proportion of "charged off" and "fully paid" can be seen in the chart below. The majority of loans are listed as fully paid. As loan grade decreases from A-G the default rate increases. This trend is consistent with sub-grades as well. Please refer to the appendix for the table summarizing the sub-grade results. This should be expected as lower graded loans will have higher risk, and therefore, increased probability of defaulting on the loan.

Table 1 - Number and percentage of loans in the LC data set by loan status

	Charged Off	**Fully Paid**
Number of Loans	**13785**	**86215**
Percent of Total	**13.8%**	**86.2%**

Table 2 - Default rate of loans by grade.

	A	B	C	D	E	F	G
Charged off	1187	3723	4738	2858	1010	239	30
Fully Paid	21401	30184	21907	9635	2569	469	50
Default rate	5.26%	11.0%	17.8%	22.9%	28.2%	33.8%	37.5%

(ii) How many loans are there in each grade? And do loan amounts vary by grade? Does interest rate for loans vary with grade, subgrade? Look at the average, standard-deviation, min and max of interest rate by grade and subgrade. Is this what you expect, and why?

Graph 1 - A histogram can be seen below showing loan amount vs count and is color coded for the different grades and sub-grades.

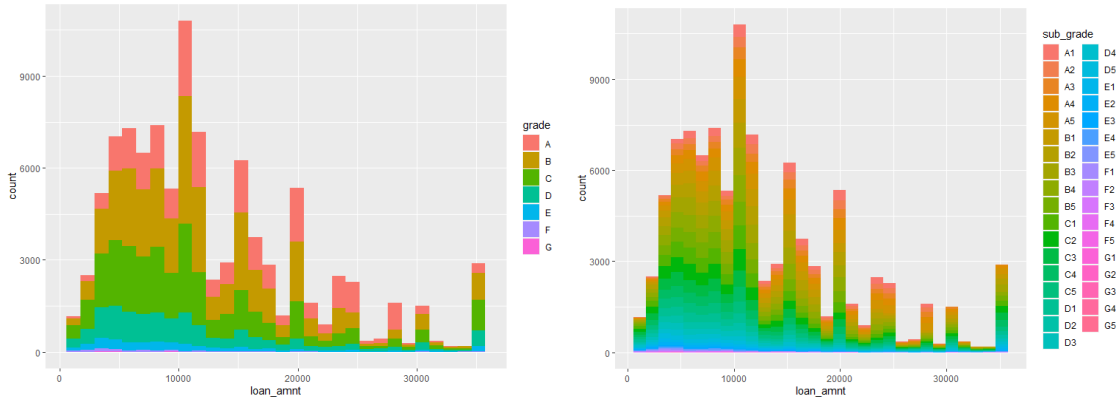


Table 3 - Shows the percentage of loans in each grade. The majority of the loans are in the first three grades (83.1%) while grade B has the highest percentage of loans in the data set.

Loan Grade	Number of Loans	Percent (%)
A	22588	22.6
B	33907	33.9
C	26645	26.6
D	12493	12.5
E	3579	3.58
F	708	0.708
G	80	0.08

The majority of the loans in dollars are in grades A through C. The average loan amount is highest in grade A and decreases as grade decreases. This would make sense because a lender is likely to provide more cash for a loan that is less risky. The standard deviation increases as the grade increases which demonstrates that high grade loans have a larger spread around the average. Please see the appendix for the analysis by sub-grade.

Table 4 - Summary statistics on loans by grade.

grade	TotalLoanAmt	AvgLoanAmt	stdevLoanAmt	MinLoanAmt	MaxLoanAmt
A	$327,649,125	$14,505	$7,441	$1,000	$35,000

B	$428,494,575	$12,637	$7,418	$1,000	$35,000
C	$319,762,050	$12,001	$8,153	$1,000	$35,000
D	$148,590,825	$11,894	$8,571	$1,000	$35,000
E	$41,583,800	$11,619	$8,914	$1,000	$35,000
F	$6,564,925	$9,272	$7,750	$1,000	$35,000
G	$946,075	$11,826	$9,417	$1,600	$35,000

Table 5 - Summary statistics for interest rates by loan grade. Numbers are represented in percentages.

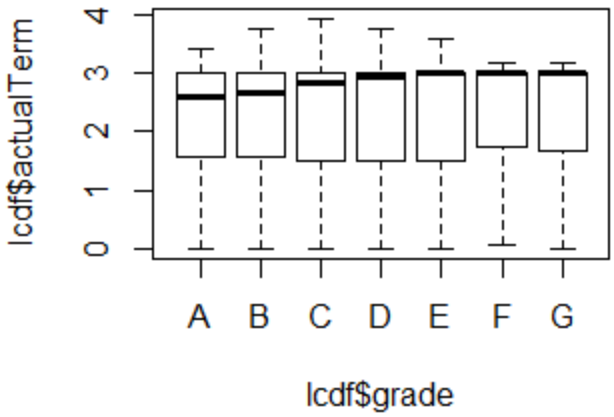
grade	Avginterestrate	stdevinterest	Mininterstrate	Maxinterestrate
A	7.17	0.967	5.32	9.25
B	10.8	1.44	6	14.1
C	13.8	1.19	6	17.3
D	17.2	1.22	6	20.3
E	19.9	1.38	6	23.4
F	24	0.916	22	26.1
G	26.4	0.849	25.8	29

Table 5 demonstrates that the interest rates increase as the loan grade increases. Additionally there appears to be a standard minimum interest rate for graded B thru E while lower rated loans have a higher minimum. There does appear to be some overlap in interest rates, suggesting that interest rates alone are not attributed to one specific grade. For example, Grade B has a max of 14.1 interest rate, which is consistent with the average for grade C. Please refer to the appendix for the table that summarizes interest rates by sub grade.

The data is consistent with what would be expected. Lower grade loans are riskier and therefore would require a high return for investors to choose those loans. The interest rate would therefore need to increase as the grade/sub-grade decreases.

(iii) For loans which are fully paid back, how does the time-to-full-payoff vary? For this, calculate the 'actual term' (issue-date to last-payment-date) for all loans. How does this actual-term vary by loan grade (a box-plot can help visualize this).

Graph 2 - Time to pay-off for fully paid loans.



The graphic above shows that there is variation in the distribution of payback time for fully paid back loans based on loan grade. You can see that the median converges closer to 3 years as the loan grade decreases. This suggests that a larger proportion of lenders pay back the loan before three years at higher loan grades. Higher loan grades have an increased probability of being paid back in a shorter period of time. The actual term of the loan was determined by subtracting the issue date from the last payment date and normalizing on a yearly basis.

(iv) Calculate the annual return. Show how you calculate the percentage annual return. Is there any return from loans which are 'charged off'? Explain. How does return from charged - off loans vary by loan grade? Compare the average return values with the average interest_rate on loans – do you notice any differences, and how do you explain this? How do returns vary by grade, and by sub-grade. If you wanted to invest in loans based on this data exploration, which loans would you invest in?

The table in the appendix provides details on the annual return rate by grade and sub-grade for fully paid and charged off loans. The percentage annual return is calculated by dividing the total collected payments divided by the original funded amount of the loan. It is then divided by the actual term of the loan in years and multiplied by 100 to determine the return rate.

There is a negative return for charged off loans. This suggests that lenders collect something but it is on average less than the original investment provided. The return on charged off loans is consistent across grades and sub-grades, with a few outlier exceptions. Most of the charged off loans have a return between -11% and -13%. A few poorly rated loans have lower returns.

There is a difference between average return rate and interest rate. For fully paid loans the interest rates are often higher than the actual return rate. This is due to the fact that some loans are paid off before the three year period and as a result the full interest expected is not collected from the loans. This lowers the return as the investor will collect less on the original investment. For fully paid loans, the actual returns

are higher as the loan grade decreases, which is what should be expected as lower graded loans are riskier and a higher return will be expected by the investor.

In this example, we would want to invest in loans that have the lowest risk of default but still generate a competitive return. It would probably be wise to diversify funds across different loan types at higher grades (A through C) to minimize risk of default while also generating competitive returns.

(v)What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose? Do loan amounts vary by purpose? Do defaults vary by purpose? Does loan-grade assigned by Lending Club vary by purpose?

Looking at the chart below, the main reasons for loans were debt consolidation, credit card repayment, and home improvement. The highest number of defaults was for debt consolidation and the highest rate of defaults was for small businesses. The average loan amounts range from around $5600 to $13660. The distribution of loan grade by purpose is similar with the majority of loans attributed to grade B and C.

Table 6 - Summary of loan characteristics by purpose.

	purpose <chr>	nLoans <int>	defaults <int>	defaultRate <dbl>	avgLoanAmt <dbl>
1	car	928	107	0.1153017	7955.038
2	credit_card	24989	2865	0.1146504	13660.144
3	debt_consolidation	57622	8319	0.1443719	13227.955
4	home_improvement	5654	682	0.1206226	11911.059
5	house	354	63	0.1779661	12756.568
6	major_purchase	1823	266	0.1459133	9948.286
7	medical	1119	172	0.1537087	7313.248
8	moving	691	144	0.2083936	6882.308
9	other	5091	838	0.1646042	8304.920
10	renewable_energy	58	11	0.1896552	8806.897
11	small_business	893	203	0.2273236	13603.415
12	vacation	678	101	0.1489676	5674.410
13	wedding	100	14	0.1400000	9123.750

Table 7 - Summary table of total loans by purpose and grade.

	A	B	C	D	E	F	G
car	253	306	238	92	27	8	4
credit_card	8349	9809	5008	1518	266	37	2
debt_consolidation	11573	19745	16497	7534	1954	292	27
home_improvement	1457	1777	1496	673	215	33	3
house	37	74	83	74	48	27	11
major_purchase	441	553	479	252	70	26	2
medical	84	270	382	251	97	34	1
moving	10	96	207	234	108	32	4
other	324	1036	1702	1321	551	139	18
renewable_energy	3	5	22	18	8	2	0
small_business	15	100	249	300	159	62	8
vacation	42	127	257	180	59	13	0
wedding	0	9	25	46	17	3	0

(vi) Consider some borrower characteristics like employment-length, annual-income, fico-scores (low, high). How do these relate to loan attributes like, for example, loan_amout, loan_status, grade, purpose, actual return, etc.

Looking at the employment-length, we discovered that the more working experience the borrowers have, the smaller the default rate gets. Borrowers with one year or less employment-lengths have 14.44% to 21.08% default rate. We also observed that as employment length increases, the annual income increases, and the number of loans decreases. Although the number of loans decreases, the average loan amount has increased. This represents the more experienced borrowers who usually borrow more money than people who have less employment length. Despite the number of loan amount increases, the average of actual return has increased while the actual term was shorter. This means borrowers were able to pay off the loan at a shorter time and the lender generates more revenues as the default rate is low.

Table 8 - Summary statistics for loan attributes by employer length.

	emp_length	nLoans	Annualinc	defaults	defaultRate	avgIntRate	avgLoanAmt	avgActRet	avgActTerm
1	n/a	6148	47518	1296	0.2108	12.6249	10152.3	3.95266	2.42677
2	< 1 year	8104	68622.3	1204	0.14857	12.1141	12170.9	5.00501	2.23395
3	1 year	6649	68580.5	960	0.14438	12.1566	12137.3	5.0983	2.2492
4	2 years	8987	70891	1206	0.13419	12.0589	12251.6	5.28841	2.22225
5	3 years	8046	71488.2	1088	0.13522	11.9791	12432.9	5.31042	2.25295
6	4 years	5892	73320.6	775	0.13153	11.9793	12556.2	5.37192	2.23635
7	5 years	6046	74068.3	841	0.1391	12.096	12658.4	5.2832	2.22886

8	6 years	4712	72868	632	0.13413	12.1716	12588.9	5.47917	2.2329
9	7 years	5124	71440.2	712	0.13895	12.1508	12563.1	5.38665	2.22716
10	8 years	4990	74639.5	698	0.13988	11.9313	13029.1	5.17112	2.24448
11	9 years	3908	74253.4	522	0.13357	11.982	13077.1	5.19045	2.2365
12	10+ years	31394	81773.7	3851	0.12267	11.7776	13740.7	5.55871	2.23816

Comparing borrower characteristics like annual income to attribute loan status, we can see that the average annual income is higher for people in the fully paid category with an average of $74744. On the other hand, the charged off loans borrowers have a lower annual income of $64678.

Table 9 - Avg. income of borrowers by loan status.

Loan_status	Annual Income
Charged Off	64678
Fully Paid	74744

(vii) Generate some (at least 3) new derived attributes which you think may be useful for predicting default., and explain what these are. For these, do an analysis as in the questions above (as reasonable based on the derived variables).

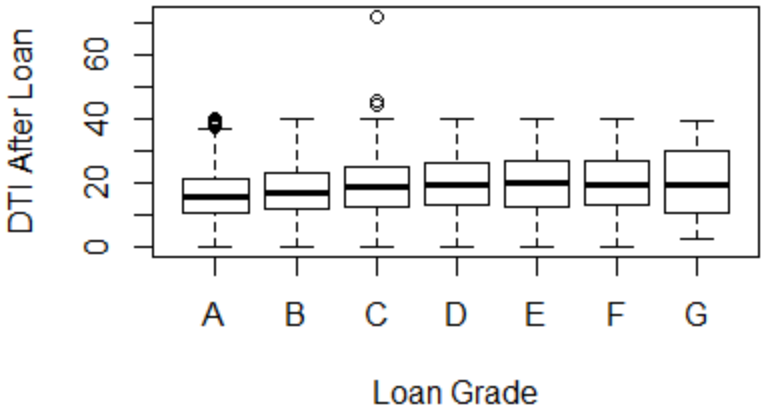
DTI_After Loan Origination - This will update the debt to income ratio to take into account the monthly payment expected from the new LC loan. The DTI currently in the database does not include the new loan, so this may be a more robust data point to evaluate potential liquidity of the borrower to make payments of the existing loan. Monthly income was determined by dividing the annual income by 12. The monthly income was then multiplied by the existing DTI to identify the monthly debt obligations before the loan Finally the monthly payment from the new loan was added to the monthly debt obligations before the loan to determine the new debt load and divided by monthly income.

Table 10 - Summary statistics for DTI_After Loan Origination.

grade	loan_statu	AvgDTI_AfterLoa	MedianDTI_AfterLoa	stdev	Min	Max

	s	n	n			
A	Charged Off	17.5	17.2	7.43	0.47	39.8
A	Fully Paid	15.9	15.4	7.36	0.01	39.9
B	Charged Off	18.4	18.1	8.05	0.42	40
B	Fully Paid	17.5	17	7.94	0.04	40.1
C	Charged Off	20.1	20	8.52	0.02	40
C	Fully Paid	18.8	18.3	8.48	0.05	71.8
D	Charged Off	20.9	20.7	8.74	0.14	40
D	Fully Paid	19.3	18.9	8.96	0.02	40.1
E	Charged Off	21.4	21.7	9.24	0.2	40
E	Fully Paid	19.5	19.2	9.39	0.05	40.1
F	Charged Off	20.8	20.9	9.19	0.59	39.9
F	Fully Paid	19	18.8	9.11	0.01	39.9
G	Charged Off	23.7	24.6	10.4	2.18	39
G	Fully Paid	18.1	16.5	10.6	2.9	39

Graph 3 - Boxlplot showing the quartiles for the DTI_After Loan Origination variable.



The DTI_After Loan variable does demonstrate higher values for loans identified as charged off compared to fully paid, and this trend also increases as loan grade decreases.

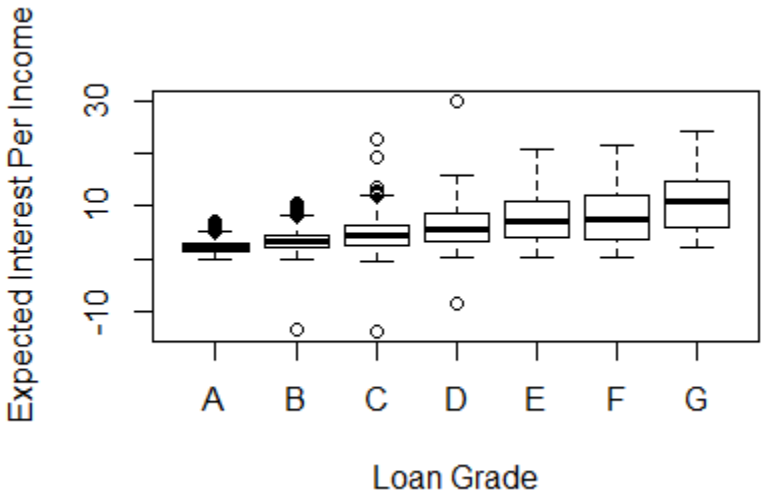
Expected Interest as a Percent of Annual Income - Loans are increasingly more difficult to pay off when interest keeps accruing. This could occur in situations where the borrower is not generating enough income to pay down the principal balance each month. This created variable is intended to quantify the total expected interest accumulated for the loan over the three year period and divide by the annual income. Expected interest was determined by multiplying the monthly payment variable by 36 months and then subtracting the loan amount. The expected interest was then divided by the annual income.

Table 11 - Summary statistics for the Expected Interest as a Percent of Annual Income Variable

Grade	loan_status	AVGexpint_perinco ~	Medianexpint_perinc ~	stdev	Min	Max
A	Charged Off	2.35	2.15	1.22	0.2	6.73
A	Fully Paid	2.14	1.94	1.15	0.01	7.16
B	Charged Off	3.66	3.32	1.93	0.09	10.4
B	Fully Paid	3.38	3.07	1.85	-13.3	10.4
C	Charged Off	4.92	4.6	2.58	-13.8	12.7
C	Fully Paid	4.54	4.15	2.57	-0.42	22.7
D	Charged Off	6.45	6.1	3.35	0.39	15.3

D	Fully Paid	5.9	5.4	3.42	-8.49	29.9
E	Charged Off	7.72	7.58	3.99	0.07	17.8
E	Fully Paid	7.17	6.73	4.1	0.19	20.9
F	Charged Off	8.86	8.67	4.78	0.65	21.6
F	Fully Paid	7.51	7.09	4.88	0.38	21.6
G	Charged Off	12.9	14.2	5.91	2.8	24.3
G	Fully Paid	9.96	9.79	5.26	1.99	22.3

Graph 4 - Boxplot summarizing the percentiles for the Expected Interest as a Percent of Income variable.



The expected interest Per Income variable shows that higher graded loans have a very tight distribution with this statistic. Additionally the average and median increase as loan grade decreases and varies slightly between charged off and fully paid when evaluated at the grade level (i.e, charged off is higher in both instances). Note that there are a few negative values for the min. When evaluating, there were four instances where the data returned a negative value. In each of these scenarios the cumulative amount collected over 36 months was much smaller than the total loan amount. This could signal an error in either the payment field or the loan amount field. Since it was four instances total, it is likely to have a negligible effect on the results.

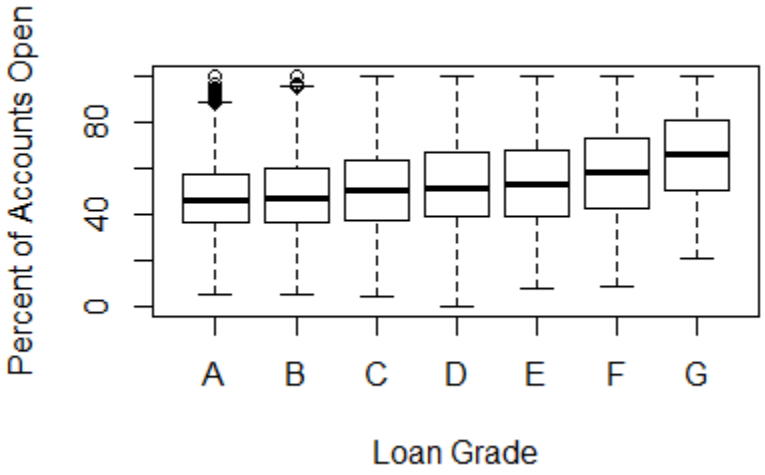
Percent of Accounts still Open - This variable is intended to quantify how many credit accounts the borrower has at the time the loan is originated. Borrowers that have successfully closed past accounts will be more likely to have success in paying off the loan compared to borrowers that start accumulating

open accounts. This variable was determined by dividing the number of open accounts by the total number of accounts.

Table 12 - Summary statistics for the Percent of Accounts Still Open variable.

grade	loan_status	AVGPercentOpenAcc	MedianPercentOpenAcc	stdev	Min	Max
A	Charged Off	48.8	46.7	16.4	12.5	100
A	Fully Paid	47.4	45.8	15.5	5.56	100
B	Charged Off	49.9	48.2	17.5	9.09	100
B	Fully Paid	48.9	47.1	17.2	5.26	100
C	Charged Off	52.8	50	18.8	6.76	100
C	Fully Paid	51.2	50	18.8	4.17	100
D	Charged Off	54.1	52.4	19	0	100
D	Fully Paid	53.2	50	19.5	2.94	100
E	Charged Off	54.3	53.3	19.5	10.7	100
E	Fully Paid	54.3	52.6	20.4	7.53	100
F	Charged Off	60.3	60	19	15	100
F	Fully Paid	57.4	56.7	21.1	9.09	100
G	Charged Off	71.9	73.7	20.3	30.4	100
G	Fully Paid	62.2	61	21.1	20.8	100

Graph 5 - Boxplot summarizing the percentiles for the percent of accounts still open variables.



The data shows that this variable has similar trends to other continuous variables in the data set. The average and median for the percent of accounts open is lower for the higher grade loans and it increases as the loan grade decreases. This is also true when segmenting each grade by fully paid and charged off, with charged off having higher means.

(b) Summarize your conclusions and main themes from your analyses

Looking at the data Lending Club does a good job assigning grades to applicants based on the default and interest rates associated. There are a number of other variables that allow for differentiation for both the grade of the loan and whether or not the loan is charged off or fully paid. This is seen with variations in averages, standard deviations, and min/maxes for each variable with relation to loan status. The variations between charged off and fully paid, and the trends across multiple different loan grades, should allow for a decision tree to predict the loan outcome.To predict the outcome of future applicants there needs to be enough differentiation of these variables to ensure a decision tree can be developed by assessing purity in the leaf nodes of a decision tree.

There are certain variables, such as employment length, which may function better as binary input when looking at the effect on default rate.

(c) Are there missing values? What is the proportion of missing values in different variables? Explain how you will handle missing values for different variables. You should consider what the variable is about, and what missing values may arise from – for example, a variable monthsSinceLastDeliquency may have no value for someone who has not yet had a delinquency; what is a sensible value to replace the missing values in this case? Are there some variables you will exclude from your model due to missing values?

The first step was to characterize the number of variables that have missing values. Before we drop any variables with all NA values, we use the code dim(lcdf) to calculate the total number of variables which is 145. Then, we use the code lcdf <- lcdf %>% select_if(function(x){!all(is.na(x))}) to drop the variable that contains only NA values. Then we use dim(lcdf) again to check the number of variables left, which is 108. This means 37 variables with all empty values have been dropped. The proportion of missing values varies across different variables. For example, we have a 97.31% for open_re_12m, open_il_12m, all_util, and etc… On the other hand, we have a very minimal percentage for last_credit_pull with 0.004% and last_pymnt_d with 0.064%. We disregard variables that contain more than 60% of missing values since it will be difficult to build a good model with such a high number of missing values.

Before removing vars wit all empty values
> dim(lcdf)
[1] 100000 145

After removing vars wit all empty values
> dim(lcdf)
[1] 100000 108

Proportion of missing values in different variable

emp_title	title	mths_since_last_delinq
0.06705	0.00012	0.49919
mths_since_last_record	revol_util	last_pymnt_d
0.82423	0.00041	0.00064
last_credit_pull_d	mths_since_last_major_derog	open_acc_6m
0.00004	0.71995	0.97313
open_act_il	open_il_12m	open_il_24m
0.97313	0.97313	0.97313
mths_since_rcnt_il	total_bal_il	il_util
0.97393	0.97313	0.97694
open_rv_12m	open_rv_24m	max_bal_bc
0.97313	0.97313	0.97313
all_util	inq_fi	total_cu_tl
0.97313	0.97313	0.97313
inq_last_12m	avg_cur_bal	bc_open_to_buy
0.97313	0.00002	0.00964
bc_util	mo_sin_old_il_acct	mths_since_recent_bc
0.01044	0.03620	0.00911
mths_since_recent_bc_dlq	mths_since_recent_inq	mths_since_recent_revol_delinq
0.74329	0.10612	0.64746
num_rev_accts	num_tl_120dpd_2m	pct_tl_nvr_dlq
0.00001	0.03824	0.00016
percent_bc_gt_75	hardship_dpd	settlement_term
0.01034	0.99955	0.99535

Variables that has less than 60% missing values

emp_title title mths_since_last_delinq revol_util

0.06705	0.00012	0.49919	0.00041
last_pymnt_d	last_credit_pull_d	avg_cur_bal	bc_open_to_buy
0.00064	0.00004	0.00002	0.00964
bc_util	mo_sin_old_il_acct	mths_since_recent_bc	mths_since_recent_inq
0.01044	0.03620	0.00911	0.10612
num_rev_accts	num_tl_120dpd_2m	pct_tl_nvr_dlq	percent_bc_gt_75
0.00001	0.03824	0.00016	0.01034

For mths_since_last_delinq, there are 49.919% missing values. The missing values means the customer does not have any delinquency. We don't want to put a zero in the missing value because it will mean they have 0 days since delinquency. So, we need to place a value higher than the existing max value such as 500 days to replace the missing values.

For revol_util variables we decided to use the median to replace missing values because revol_util stands for the amount of credit the borrower is using relative to all available revolving credit. We don't want to use the maximum number or any number that is higher or lower than the median because it will give us an inaccurate prediction on the revolving line utilization rate.

Variable that we want to exclude from the model due to missing values will be bc_util since it is a ratio. In addition, we also want to exclude percent_bc_gt_75 because it is a percentage.

3. Consider the potential for data leakage. You do not want to include variables in your model which may not be available when applying the model; that is, some data may not be available for new loans before they are funded. Leakage may also arise from variables in the data which may have been updated during the loan period (ie., after the loan is funded). Identify and explain which variables you will exclude from the model.

emp_title title mths_since_last_delinq revol_util

last_pymnt_d last_credit_pull_d avg_cur_bal bc_open_to_buy

mths_since_recent_inq

num_rev_accts num_tl_120dpd_2m

There are a number of variables which have no information such as id, member_ID, url, desc, and many others. These inputs with majority n/a output can be removed up front to clean up the data set. There are also variables that are dependent on a variety of other factors and are only relevant to the snapshot in time that this data was extracted. Variables like revolBal and revolUtil change over time so using it on training data may not carry over when applying the model to new data. Also, variables the loan being distributed such as funded_amnt may prove relevant on test data but won't be available when

looking at new applicants. Any variables that are redundant also need to be filtered out so that the model can work on relevant data.

Other examples of data that would not be available at the time the loan is generated included fields that summarize the principle paid to date and outstanding loan principal. Although this could be important to consider if evaluating already originated loans, it wouldn't help a user predict whether a borrow would be more or less likely to default. Total accounts would be another example, as the total accounts may change after the loan is originated. The last credit pulled/credit score obtained would be biased to any repayments or delinquencies on the loan and therefore may bias the prediction capability of identifying default risk when developing a model for evaluation at loan origination. The amount in collections would also be irrelevant, because this would only apply after the loan had been originated. Furthermore any data fields that summarize current information or information in the last 12 or 24 months would represent information that is updated after the loan had been originated and could potentially bias the prediction capability of the model.

4. Do a univariate analysis to determine which variables (from amongst those you decide to consider for the next stage prediction task) will be individually useful for predicting the dependent variable (loan_status). For this, you need a measure of relationship between the dependent variable and each of the potential predictor variables. Given loan-status as a binary dependent variable, which measure will you use? From your analyses using this measure, which variables do you think will be useful for predicting loan_status? (Note – if certain variables on their own are highly predictive of the outcome, it is good to ask if this variable has a leakage issue).

In order to determine the predictability of individual variables on loan_status the predictor variables must first be set to numeric as the auc function measures area under a curve. Once converted, the first test is for all variables with AUC greater than 0.5. This returns 55 variables in total which is too many variables so further reduction is done by looking at AUC greater than 0.54. This test returns 33 variables but the variables at the higher end of the range were identified earlier as problematic for their susceptibility to leakage. After further testing, the ideal range to determine strength of individual variables as predictors, while reducing exposure to leakage, is a range between 0.55 and 0.59. Those 12 variables can be seen below. They include expected strong indicators such as annual income, credit limit, and DTI.

Table 13 - summarizes the AUC for variables for the predictor variable loan status.

names	x
annual_inc	0.57678
tot_hi_cred_lim	0.57355
total_bc_limit	0.57301
DTI_AfterLoan	0.56853

dti	0.56827
total_rev_hi_lim	0.56557
mort_acc	0.55832
home_ownership	0.55531
mo_sin_old_rev_tl_op	0.55112

Part B: we will next develop predictive models for loan_status.

5. Develop decision tree models to predict default. (a) Split the data into training and validation sets. What proportions do you consider, why?

The data was split with 70% in training data and 30% in validation data. The purpose of the model is to differentiate loans that are at risk of default, which represents approximately 14% of the claims. Since there are almost 6X as many loans that are fully paid compared to charged off, we would want to use a training set that includes a larger sample of claims to ensure there are enough examples of the charged off loans to build the model. There is a risk that the model may not be as accurate If our training set did not have an adequate number of charged off loans in the data set.

(b) Train decision tree models (use both rpart, c50) [If something looks too good, it may be due to leakage – make sure you address this] What parameters do you experiment with, and what performance do you obtain (on training and validation sets)? Clearly tabulate your results and briefly describe your findings. How do you evaluate performance – which measure do you consider, and why?

The first tree was built using a minimum split quantity of 30 observations. These parameters did not build a tree because the default complexity parameter of 0.01 was too large for the function to build anything. This did provide a baseline as the confusion matrix for this default was 86.3%. This matrix assumes every loan will be issued and the data contains 86.% fully paid with the remainder charged off. The next iteration lowered the complexity parameter to 0.0 with a minsplit of 10 and minbucket of 5 in order to grow a full tree. This yielded a fully-grown tree but had clear indications of overfit when looking at the confusion matrix below.

	Actual	
pred	Fully Paid	Charged Off
Fully Paid	58517	2606

Charged Off	1914	6963

This matrix above is based off the training data with an accuracy rate of 93.5% which indicates that the model may be overfit. The next step to building the tree using rpart was to prune the tree from the x-error based on the complexity parameter. There appeared to be some stability in the model at a cp of 9.5795e-05 so that value was used in building the next tree. The confusion matrix for both the training and test data can be seen below.

Training

	Actual	
pred	Fully Paid	Charged Off
Fully Paid	59404	4803
Charged Off	1027	4766

Test

	true	
pred	Fully Paid	Charged Off
Fully Paid	23722	3701
Charged Off	2062	515

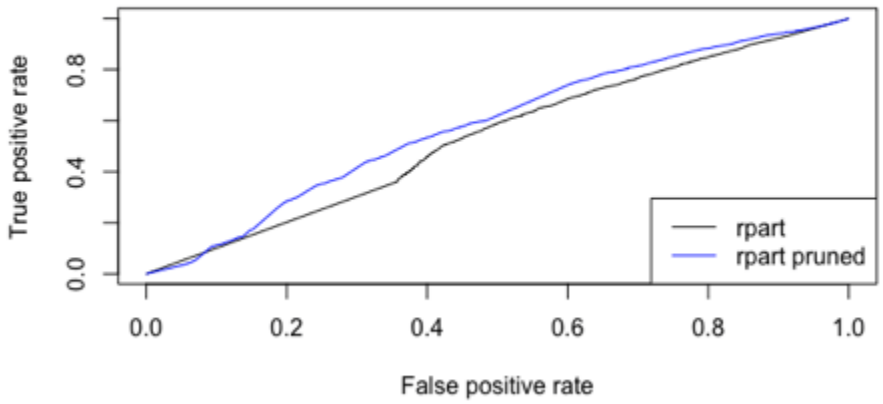
The errors for these two trees yielded 91.7 % and 80.8% accuracy respectively. The concern with the results on the test data is that the model is overpredicting the fully paid off which will have implications for the financial impact of these decisions. Looking at the complexity parameter, the chart shows stability around 1.2410e-04 so this value was used to prune the tree further. This yielded a lower training accuracy of 89.6% but slightly raised the accuracy on the test data with 82.9%. The confusion matrix for the test data can be seen below.

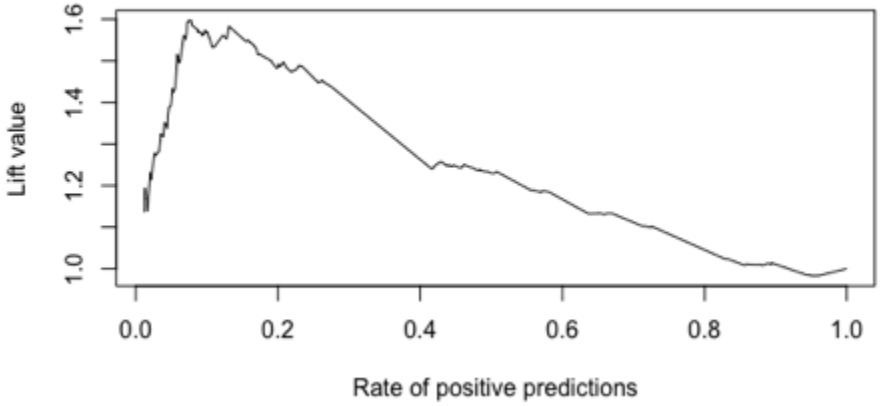
	Actual	
pred	Fully Paid	Charged Off
Fully Paid	24556	3910
Charged Off	1228	306

The final test was run to determine if there are any significant findings when using the gini parameter as opposed to information. No significant error was found. The C50 testing only resulted assuming all loans would be fully paid, due to the imbalance in the data, so that model was not used in analysis.

(c) Identify the best tree model. Why do you consider it best? Describe this model – in terms of complexity (size). Examine variable importance. How does this relate to your uni-variate analyses in Question 4 above?

In order to evaluate performance the below ROC curve was derived from the original rpart model and the pruned tree. This shows a slight increase from the pruned tree to the original model. At lower probabilities there is larger lift. As the probability increases the lift value is creasing. The likelihood of having a positive response is highest at approximately 0.15. If the goal is to have the highest accuracy, then the tree lcDT4 would be the best according to the below charts. Further analysis will have to be done depending on the financial impact in later sections.





Variable Importance is shown below. The most important is a created variable that determines the expected interest payments as a function of income. There is also DTI which was expected as well as a few variables regarding the credit limit and balances.

names	x
1 expint_perincome	1131.
2 total_rev_hi_lim	842.
3 total_bc_limit	710.
4 tot_hi_cred_lim	695.
5 dti	684.
6 total_bal_ex_mort	661.
7 bc_util	656.
8 total_il_high_credit_limit	597.
9 mo_sin_old_rev_tl_op	589.
10 DTI_AfterLoan	566.

6. Develop a random forest model. (Note the 'ranger' library can give faster computations) What parameters do you experiment with, and does this affect performance? Describe the best model in terms of number of trees, performance, variable importance. Compare the performance of random forest and best decision tree model from Q 5 above. Do you find the importance of variables to be different ? Which model would you prefer, and why ?

The parameters that have the biggest impact on a random forest model are the number of trees and the mtry variable, which quantifies how many variables can be included in each decision tree within the forest. A series of random forests were run to identify the impact of these input variables. The number of trees was originally set at 500 and the default mtry is 6, which the square root of the variables in the data set rounded down to the next closest integer. The base case had an accuracy of 86.1% and a specificity of 0.24%. The specificity is poor, and results in potential risk of investing in a loan that is predicted to be fully paid when in fact it is a loan that may be charged off. The number of trees were then decreased to 300 and mtry remained at 6. This had no impact on accuracy, but specificity decreased slightly. This effect also occurred when we decreased mtry from 6 to 3. The model with the best specificity occurred when the mtry variable was increased to 10, but overall the sensitivity does not increase significantly. Random forests also allow for a min number of variables in a leaf node and the maximum depth of a tree. Models that quantified these variables at 30 and 10 respectively were performed, but there was no significant impact on accuracy and improved specificity. Below are examples of the confusion matrix on training data, test data, and the ROC chart for the model.

Table 6B1 - Confusion matrix for Random Forest Model on Training Data.
Note: Accuracy = 86.25%, Recall = 99.9%, Specificity = 0.66%

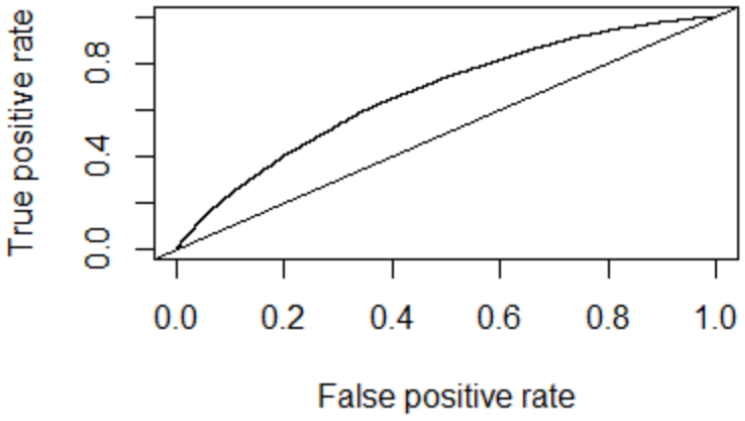
	predicted	
true	Charged Off	Fully Paid
Charged Off	64	9554
Fully Paid	68	60314

Table 6B2 - Confusion matrix for Random Forest Model on Test Data
Note: Accuracy = 99.9%, Recall = 99.9%, Specificity = 0.43%

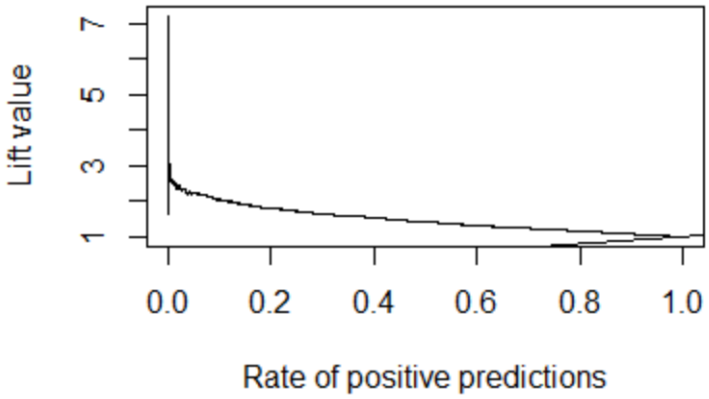
	Actual	
pred	Charged Off	Fully Paid
Charged Off	18	24
Fully Paid	4149	25809

>

Graph 6B - ROC plot



Graph 6B2 - Lift Plot for the best Random Forest Model



The major importance variables in the random forest are DTI_After loan, DTI, installment, funded amount to investors, funded amount, and expected interest per income. These variables are slightly different from the decision tree model using Rpart. This could be due to the fact that the split variables on a given node may be influenced by prior splits in one decision tree, and therefore variables that could be important may not be considered. In a random forest, multiple trees are built using a small subset of variables. The smaller subset of variables allows for more variables to split at different nodes, and then averaged out across all the decision trees built. Therefore there may be an impact on the order of importance for certain variables. . The random forest model would be preferred because we are using more decision trees in order to arrive at consensus prediction. Additionally the accuracy is improved, and the threshold can be adjusted in order to improve the specificity associated with the model.

(a) Compare the performance of your models from Questions 5, 6 above based on this. Note that the confusion matrix depends on the classification threshold/cutoff you use. Evaluate 6 different thresholds and analyze performance. Which model do you think will be best, and why.

Assuming that each loan is for $100 and each loan returns 8% on average. Charged off loans lose 12% on average. It is assumed. that all fully paid loans will be paid off in 2.12 based on the average of the data set. This means that the average profit on a $100 loan is .08*2.12*$100 = $16.96. Each 100 that does not get spent on a loan can be assumed to be invested in a CD that returns 2% annual compounding. This gives $6.12 profit per loan not

issued using the formula FV = PV * (1+r)^2. The profit lost by issuing a loan that gets charged off is 36% without taking into account any opportunity cost. Since it is known that every $100 spent could generate $6.12 in three years then this must be added to the profit lost giving a total of $42.12 lost per loan. These financial metrics were applied to the confusion matrix below at the given threshold. The best financial decision based on the rpart models would be using lcDT2 with a CTHRESH of 0.50 with a total of $275,494.20. The data heavily favors giving out loans with an 86% paid off rate. If loans were given to every individual in the test data set, the total profit would be $259,718.72.

lcDT2 (CTHRESH 0.50)

	Actuals	
pred	Fully Paid	Charged Off
Fully Paid	22506	3506
Charged Off	3278	710

lcDT2 (CTHRESH 0.86)

	Actuals	
predictions	Fully Paid	Charged Off
Fully Paid	20547	3101
Charged Off	5237	1115

lcDT4 (CTHRESH 0.50)

	Actuals	
pred	Fully Paid	Charged Off
Fully Paid	24556	3910
Charged Off	1228	306

lcDT4 (CTHRESH 0.86)

	Actuals	
predictions	Fully Paid	Charged Off
Fully Paid	20358	2783
Charged Off	5426	1433

For the random forest model the confusion matrix for different thresholds is exhibited in teh table below. The table demonstrates that the different threshold levels all have impacts on the

overall accuracy, recall and specificity. As the threshold increases, the specificity and recall decreases while accuracy increases.

Threshold	5%	Charged Off	Fully Paid			Accuracy	0.2573
Pred	Charged Off	4006	22120	26126		Recall	0.958441
	Fully Paid	161	3713	3874		Specificity	0.961363
		4167	25833	30000			
Threshold	10%	Charged Off	Fully Paid			Accuracy	0.430233
Pred	Charged Off	3518	16444	19962		Recall	0.935346
	Fully Paid	649	9389	10038		Specificity	0.844252
		4167	25833	30000			
Threshold	20%	Charged Off	Fully Paid			Accuracy	0.703333
Pred	Charged Off	2014	6747	8761		Recall	0.89863
	Fully Paid	2153	19086	21239		Specificity	0.483321
		4167	25833	30000			

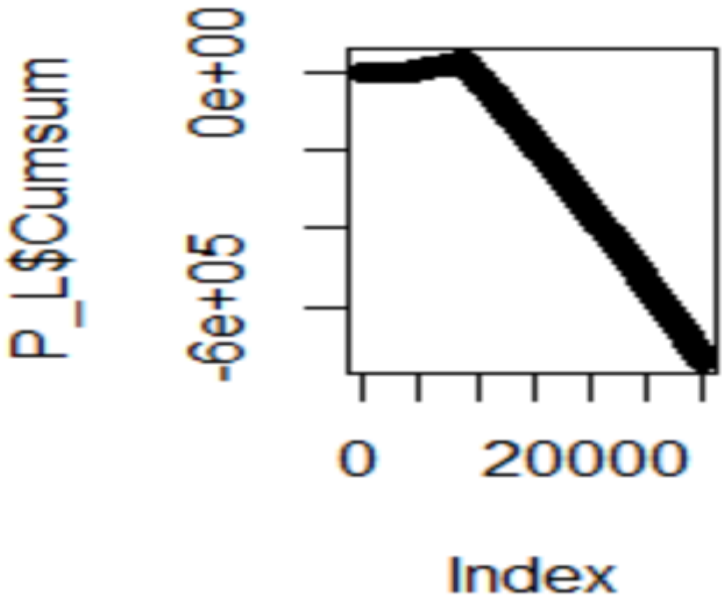
Threshold	30%	Charged Off	Fully Paid			Accuracy	0.8289
Pred	Charged Off	708	1674	2382		Recall	0.874756
	Fully Paid	3459	24159	27618		Specificity	0.169906
		4167	25833	30000			
Threshold	40%	Charged Off	Fully Paid			Accuracy	0.857567
Pred	Charged Off	127	233	360		Recall	0.863698
	Fully Paid	4040	25600	29640		Specificity	0.030478
		4167	25833	30000			
Threshold	50%	Charged Off	Fully Paid			Accuracy	0.860767
Pred	Charged Off	12	22	34		Recall	0.861343
	Fully Paid	4155	25811	29966		Specificity	0.00288
		4167	25833	30000			
Threshold	60%	Charged Off	Fully Paid			Accuracy	0.861133
Pred	Charged Off	1	0	1		Recall	0.861129

	Fully Paid	4166	25833	29999		Specificity	0.00024
		4167	25833	30000			

(b) Another approach is to directly consider how the model will be used – you can order the loans in descending order of prob(fully-paid). Then, you can consider starting with the loans which are most likely to be fully-paid and go down this list till the point where overall profits begin to decline (as discussed in class). Conduct an analyses to determine what threshold/cutoff value of prob(fully-paid) you will use and what is the total profit from different models. Also compare the total profits from using a model to that from investing in the safe CDs. Explain your analyses and calculations. Which model do you find to be best and why. And how does this compare with what you found to be best in part (a) above.

For this analysis a profit of $16.96 was used to represent the average return for a loan that is fully paid, while -42.12 was used to represent the lost income from a negative return plus an opportunity cost for forgoing a CD that would have generated $6.12 in profit over three years. Any loans that were not predicted as fully paid were given a profit of $6.12 to represent the CD investment. Probabilities were arranged in descending order and the formula's were applied to identify the cumulative sum. It was determined that the probability threshold of 0.2 occurring in row 8761 generated the highest profitability of $29,599.

Graph - Summarizing the cumulative profit based on average return over 3 years of $16.96, loss of $42,12 for charged off loans, and CD profit of $6.12 for those not predicted by the model. Based on this calculation, it appears that the rpart model mentioned above would result in a higher profitability.



Appendix:

Questions 2AI - Table summarizing Default Rate by Sub-grade

Grade	Sub-Grade	Default Rate (%)
A	A1	2.78
A	A2	3.38
A	A3	4.83
A	A4	6.21
A	A5	7.16
B	B1	7.88
B	B2	9.00
B	B3	11.47

B	B4	12.04
B	B5	14.35
C	C1	15.03
C	C2	16.25
C	C3	18.53
C	C4	19.91
C	C5	20.99
D	D1	21.58
D	D2	22.95
D	D3	22.72
D	D4	24.66
D	D5	23.60
E	E1	26.48
E	E2	27.58
E	E3	27.65
E	E4	30.26
E	E5	33.51
F	F1	25.00
F	F2	31.21
F	F3	36.20
F	F4	48.45
F	F5	47.27
G	G1	38.71
G	G2	42.86
G	G3	26.32

G	G4	40.00
G	G5	50.00

Question 2A(III) - Table summarizing interest rates by sub-grade

grade	sub_grade	Avginterestrate	stdevinterest	Mininterstrate	Maxinterestrate
A	A1	5.680069	0.347485	5.32	6.03
A	A2	6.415494	0.166259	6.24	6.97
A	A3	7.094107	0.324701	6.68	7.62
A	A4	7.475851	0.357395	6.92	8.6
A	A5	8.241788	0.424467	6	9.25
B	B1	8.87001	0.721752	6	10.16
B	B2	9.959382	0.815586	6	11.14
B	B3	10.84593	0.887329	6	12.12
B	B4	11.73146	0.839794	6	13.11
B	B5	12.22738	0.851215	6	14.09
C	C1	12.86153	0.786176	11.99	14.33
C	C2	13.3082	0.873285	6	15.31
C	C3	13.97528	0.865608	6	15.8
C	C4	14.56803	0.854714	6	16.29
C	C5	15.22136	0.883442	6	17.27
D	D1	16.09891	0.870687	6	17.77
D	D2	16.95641	0.886628	6	18.55
D	D3	17.44531	0.873474	6	19.2
D	D4	18.07453	0.831805	17.14	19.52

D	D5	18.48426	1.002095	6	20.31
E	E1	18.97299	0.98727	6	21
E	E2	19.57885	1.058906	18.49	21.7
E	E3	20.14332	1.032144	18.99	22.4
E	E4	20.99339	0.952378	19.99	23.1
E	E5	21.97003	0.762833	20.99	23.4
F	F1	23.12476	0.59623	21.99	23.7
F	F2	23.74262	0.47617	22.99	24.08
F	F3	24.38534	0.247137	23.63	24.5
F	F4	24.95299	0.214472	23.76	25.09
F	F5	25.59546	0.272905	23.83	26.06
G	G1	26.12	0.472927	25.8	26.99
G	G2	26.39381	0.736468	25.83	27.31
G	G3	26.73368	1.016706	25.89	27.99
G	G4	26.99	1.369306	25.99	28.49
G	G5	26.7925	1.465	26.06	28.99