Team member

Vanisa Achakulvisut	UIN667903568	vachak2@uic.edu
Matt Carey	UIN653922201	mcarey21@uic.edu
Abhijit Zarekar	UIN676178928	azarek3@uic.edu

Questions:

1. Describe the business model for Lending Club. Consider the stakeholders and their roles, and what advantages Lending Club offers. How does the platform make money? (Not more than 1.5 pages, single spaced, 11 pt font. Please cite your sources).

Lending Club has a simplistic business model when you look at how the revenue is generated. Lending Club breaks the company into different sectors that directly operate with each other through borrowing and investing. The borrowing portion is broken down into business and personal loans, auto refinancing, and medical and dental financing. The investing can be into loans, retirement accounts, or different types of institutions such as banks and financial advisors. With these two segments, Lending Club makes their income from the fees associated with these transactions. These are the fees that are associated with borrowers, investors, or the accounts from 3rd party institutions. The entire business cycle is summed up that the investors provide the capital to the borrowers and in return, the investors are receiving the interest on the loan along with the original investment. The whole concept behind this business was through social media connecting people with the ability to help each other. After huge growth and venture capitalists jumping on the idea Lending Club went public as their own platform.

With our examples below we are breaking down the system in which Lending Club sets up how they perform evaluating loans. Once a loan is submitted and approved through Lending Club the borrower's profile is put into a note, each of these notes are viewed by the investors. Each note has a particular risk factor depending on the borrowed investment health. Just like we see in the example below in problem 2, there is a grading system on how these loans are distributed. Lending Clubs “grades range from ‘A1’ indicating the lowest risk to grade ‘G5’ as the highest, in total there are 35 grades. The lower the rating the higher the return is for the investor.”[1]

The stakeholders role in a company run by investments would hold a different weight in the type of atmosphere the business runs in. The stakeholders refers to everybody that has a vested interest in Lending Club as a corporation. From employees on all hierarchy of the company down to all the consumers invested. This company does not deal with a product that is what we would refer to as simply supply and demand of a product. Lending Club is a continuous cycle of borrowing and lending of loans… and then back around again in a continuous cycle. There are real world problems with making all your revenue as a company directly from the fees incurred in all the contracts of these loans. Say something happens with either side of the spectrum not being able to hold up to their end of the bargain financially. This could have crucial outcomes for the stakeholders involved with the Lending Club model. One thing that all stakeholders should be aware that this business model deals with processing unsecured loans. “Since the loans are not secured, LendingClub cannot sell borrowers' assets to pay back the investors. Without any collateral, LendingClub must take collection action against the borrower

in case of a default."[2] Little nuances written into the details of these contracts that either the borrower or lender don't see puts strain on the business as a whole as a company.

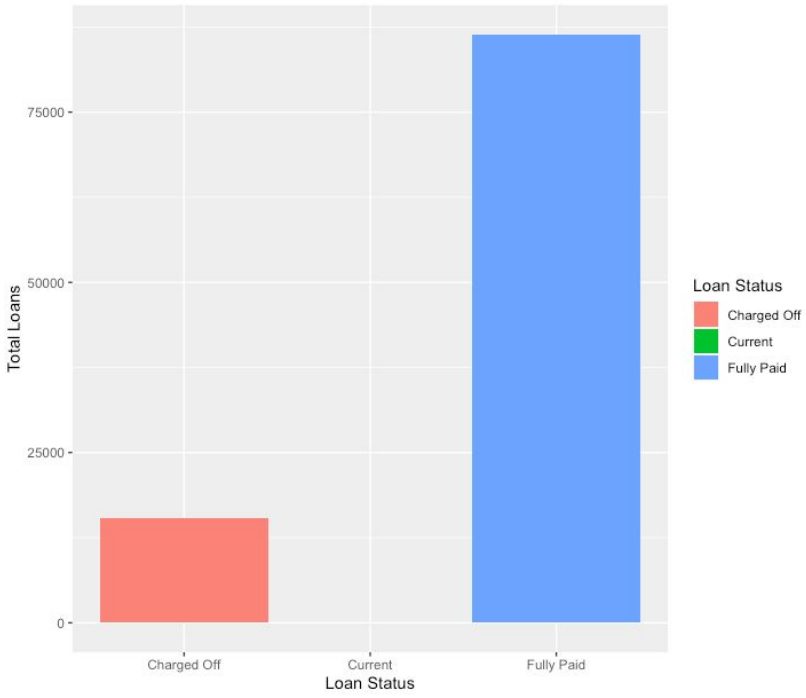
With the essence of how Lending Club has set their business protocols, they are in the market to appease customers that may not be accepted with other loan providers. Lending Club sets up standards where even though they have a lower grading for a loan, certain investors take advantage of the higher payoff for these riskier endeavors. Certain advantages that Lending Club offers that hit this target market are long term loans, no hard credit inquiry, along with accepting low credit scores. With long term loans you can actually "stretch the loan to repayment terms of three years and five years.[3] Gives customers the ability to prolong payments for longer periods of time when they are unable to meet the terms of their note. Without having to worry about your credit getting hit looking into different loan options, Lending Club does not check credit rates, which "allow you to conveniently shop around without hurting your credit score."[3] Even though you may not be in the best credit standings, Lending Club reaches customers with credit scores as low as 600.

1. "Lending Club." *Businessmodelzoo*, 2016, www.businessmodelzoo.com/exemplars/lending-club.
2. P, Kim. "Lending Club Review." *CreditDonkey*, CreditDonkey, 27 June 2018, www.creditdonkey.com/lendingclub-review.html.
3. Writer, AuthorBill FayStaff. "Lending Club Review: How It Works, Requirements and Alternatives." *Debt.org*, 28 Aug. 2019, www.debt.org/credit/loans/personal/lending-club-review/.

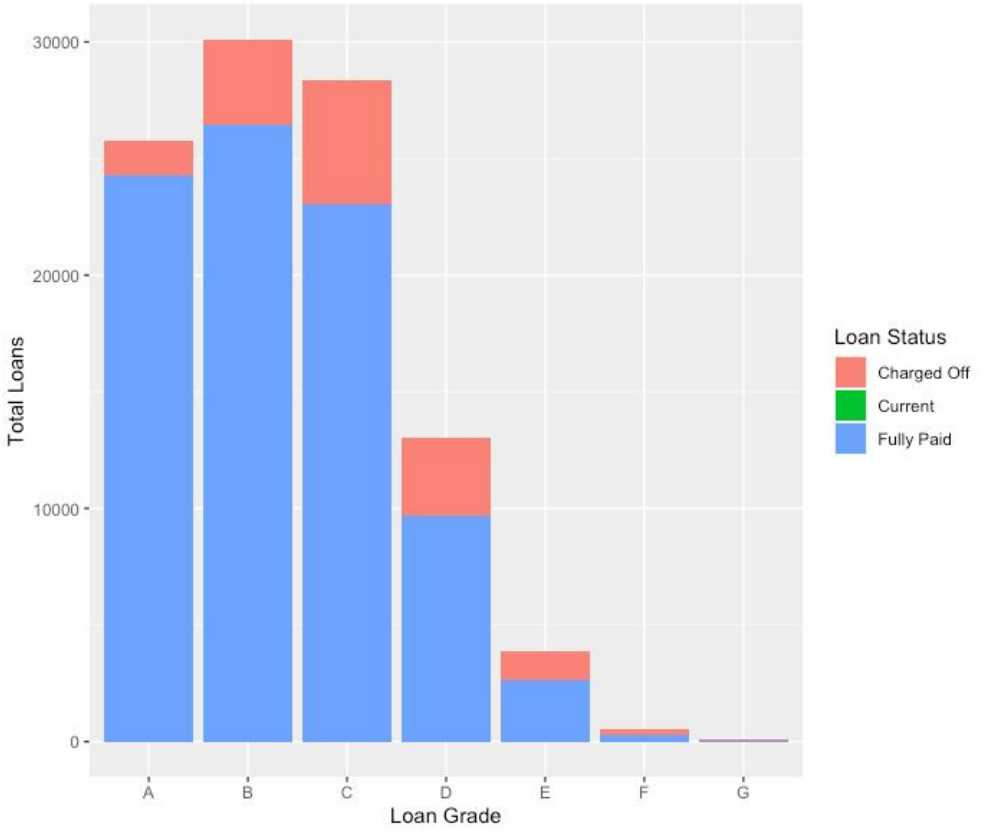
2. Data exploration

(a) some questions to consider,

(i) 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?



loan_status	nLoans	loan_prop
Charged Off	15,341	0.151
Fully Paid	86,385	0.849



By Count (No. of Customers)	A	B	C	D	E	F	G
Charged Off	1,486	3,657	5,344	3,375	1,238	214	27
Fully Paid	24,265	26,453	23,025	9,665	2,628	296	53
Charged Off+Fully Paid	25,751	30,110	28,369	13,040	3,866	510	80
Proportion of defaults(Charged Off)	5.77%	12.15%	18.84%	25.88%	32.02%	41.96%	33.75%

How does default rate vary with loan grade?

grade	nLoans	prop_defaults
A	25,751	5.77
B	30,110	12.15
C	28,369	18.84
D	13,040	25.88
E	3,866	32.02
F	510	41.96
G	80	33.75

- As the proportion of default goes up the grade gets worse with one exception for G grade.
- The ranking of grades from High to Low are A,B,C,D,E,F respectively except G which does not follow the trend.
- The proportion of default goes up as the grade gets worse.

Does it vary with sub-grade?

sub_grade	nLoans	prop_defaults
A1	3,202	2.3
A2	4,247	3.8
A3	4,037	5.4
A4	6,207	6.1
A5	8,058	8.1
B1	5,229	9.6
B2	5,911	10.4
B3	5,995	11.5
B4	6,075	13.1
B5	6,900	15.3
C1	6,653	15.9
C2	6,417	18.4
C3	5,777	18.7
C4	5,044	21.2
C5	4,478	21.4
D1	3,695	23.5
D2	2,903	25.5
D3	2,611	26.0
D4	2,024	27.7
D5	1,807	29.1
E1	1,318	31.4
E2	966	29.0
E3	781	32.9
E4	468	34.8
E5	333	37.2
F1	196	40.3
F2	122	41.8
F3	81	37.0
F4	66	48.5
F5	45	48.9
G1	28	28.6
G2	26	38.5
G3	17	29.4
G4	7	28.6
G5	2	100.0

- Yes, it does vary with the subgrade. Suffix 1 to 5 after each Grade denotes the sub_grades, for example, A1 is the best grade which gives the minimum proportion of default. 2,3,4,5 denote the sub-ranked with higher proportion of default.

And is this what you would expect, and why?

- Yes. The proportion of default should increase as the grades are getting worse.

(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? And is this what you expect, and why?

How many loans are there in each grade?

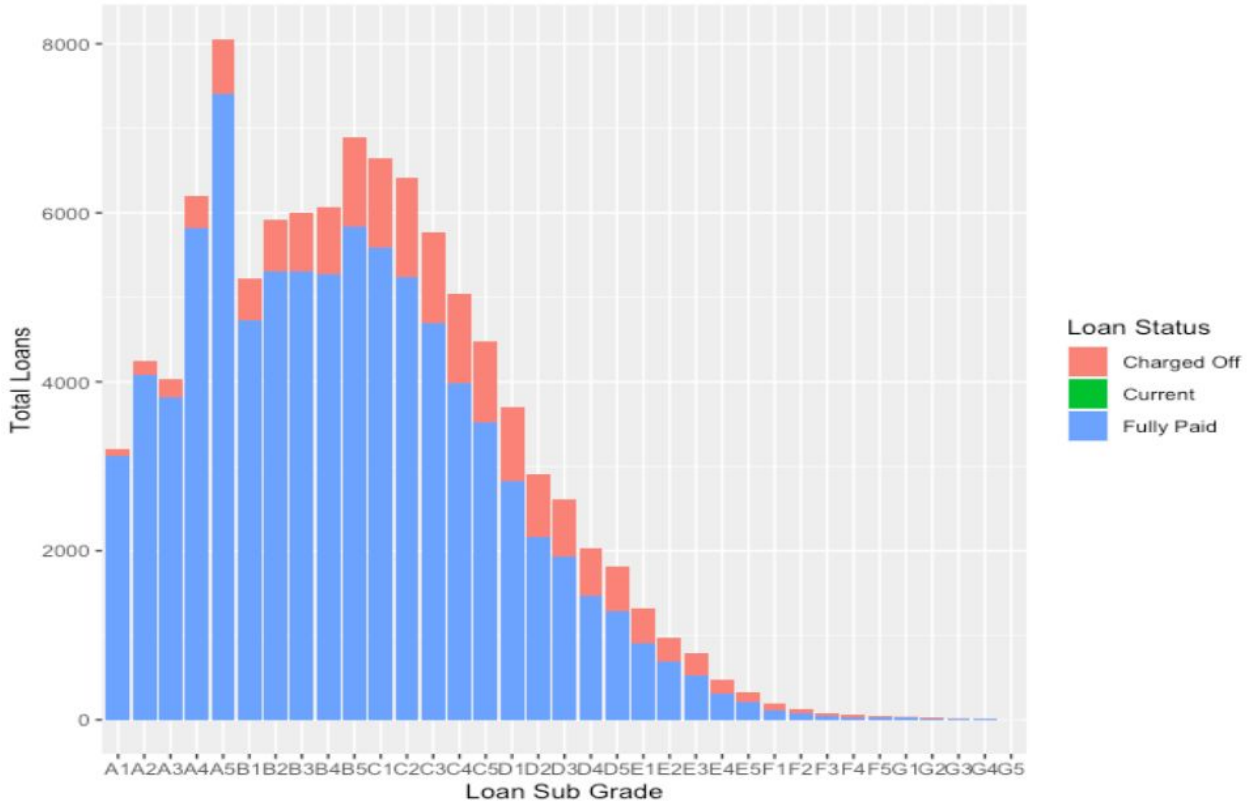
grade	nLoans	sum_loan_amt	avg_loan
A	25,751	369,886,150	14,364
B	30,110	379,700,225	12,610
C	28,369	333,016,600	11,739
D	13,040	162,293,650	12,446
E	3,866	52,422,150	13,560
F	510	5,665,900	11,110
G	80	790,000	9,875

And do loan amounts vary by grade?

- Yes, loan amounts do vary by grade. As we traverse from grade A to grade G the loan amounts generally show a decreasing trend. Grade B shows the highest loan amount.

Does interest rate for loans vary with grade, subgrade?

grade	mean(int_rate)
A	7.06
B	10.20
C	13.34
D	16.62
E	19.31
F	23.90
G	25.90



sub_grade	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
int_rate	5.9	6.4	6.8	7.1	8.0	8.3	9.3	10.1	11.1	11.7	12.3	12.8	13.4	14.1	14.8
sub_grade	D1	D2	D3	D4	D5	E1	E2	E3	E4	E5	F1	F2	F3	F4	F5
int_rate	15.6	16.4	16.8	17.4	17.9	18.3	18.9	19.6	20.6	21.7	22.9	23.9	24.5	25.0	25.6
sub_grade	G1	G2	G3	G4	G5										
int_rate	25.9	25.8	25.9	26.0	26.1										

- Yes, the interest rate goes up as we traverse from Grade A to Grade G.
- With respect to Sub-grades, the interest rate varies. AAs we traverse from Sub-grade A1 to Sub-grade G5 the interest rate decreases.

And is this what you expect, and why?

- Yes, the results correspond to what we expected because the customers who get the high grade have lower default rate. Therefore, they should be accessible to higher amount and lower loan interest rate

(iii)What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose?

Purpose	Count	Mean(loan_amnt)
car	950	8,278
credit_card	24,051	13,839
debt_consolidation	60,669	13,161
educational	1	2,200
home_improvement	5,502	12,406
house	385	12,429
major_purchase	1,705	10,302
medical	991	7,507
moving	705	6,879
other	5,119	8,677
renewable_energy	62	8,259
small_business	885	14,271
vacation	699	5,907
wedding	2	3,600

- People borrow money for several reasons : car, credit_card, debt_consolidation, educational, home_improvement, house, major_purchase, medical, moving, renewable_energy, Small_business, Vacation, Wedding, other.
- The average loan amount varies by the purpose of usage. Small business and credit cards hold the highest credit amount while car and renewable energy are the lowest credit amount along these purposes of lending .
- The most popular purpose for lending money is for debt consolidation and credit cards respectively.

And within grade?

grade	purpose	nLoans	defaults	nondefau	percent_defaults_bygrad
A	vacation	46	4	42	8.70%
A	debt_consolidation	13,440	794	12,646	5.91%
A	credit_card	9,532	551	8,981	5.78%
A	car	260	15	245	5.77%
A	major_purchase	432	22	410	5.09%
A	other	434	22	412	5.07%
A	home_improvement	1,481	75	1,406	5.06%
A	small_business	21	1	20	4.76%
A	medical	79	2	77	2.53%
A	house	12	-	12	0.00%
A	moving	13	-	13	0.00%
A	renewable_energy	1	-	1	0.00%
B	vacation	155	24	131	15.48%
B	credit_card	7,930	984	6,946	12.41%
B	debt_consolidation	18,149	2,250	15,899	12.40%
B	major_purchase	455	55	400	12.09%
B	moving	77	9	68	11.69%
B	medical	244	26	218	10.66%
B	house	38	4	34	10.53%
B	home_improvement	1,561	161	1,400	10.31%
B	car	281	28	253	9.96%
B	other	1,132	109	1,023	9.63%
B	small_business	80	7	73	8.75%
B	educational	1	-	1	0.00%
B	renewable_energy	6	-	6	0.00%
B	wedding	1	-	1	0.00%
C	medical	386	87	299	22.54%
C	credit_card	4,910	992	3,918	20.20%
C	small_business	253	51	202	20.16%
C	debt_consolidation	17,999	3,438	14,561	19.10%
C	car	267	47	220	17.60%
C	renewable_energy	23	4	19	17.39%
C	home_improvement	1,490	248	1,242	16.64%
C	other	1,892	308	1,584	16.28%
C	major_purchase	465	70	395	15.05%
C	vacation	290	43	247	14.83%
C	house	102	15	87	14.71%
C	moving	291	41	250	14.09%
C	wedding	1	-	1	0.00%
D	credit_card	1,355	386	969	28.49%
D	renewable_energy	18	5	13	27.78%
D	moving	231	63	168	27.27%
D	major_purchase	261	71	190	27.20%
D	home_improvement	717	190	527	26.50%
D	debt_consolidation	8,421	2,206	6,215	26.20%
D	car	102	26	76	25.49%
D	small_business	296	74	222	25.00%
D	medical	212	47	165	22.17%
D	other	1,155	256	899	22.16%
D	vacation	157	32	125	20.38%
D	house	115	19	96	16.52%
E	car	36	13	23	36.11%
E	moving	77	26	51	33.77%
E	debt_consolidation	2,374	781	1,593	32.90%
E	other	413	132	281	31.96%
E	major_purchase	79	25	54	31.65%
E	credit_card	288	89	199	30.90%
E	renewable_energy	13	4	9	30.77%
E	house	79	24	55	30.38%
E	home_improvement	228	67	161	29.39%
E	vacation	41	12	29	29.27%
E	small_business	176	49	127	27.84%
E	medical	62	16	46	25.81%
F	renewable_energy	1	1	-	100.00%
F	major_purchase	11	8	3	72.73%
F	other	74	35	39	47.30%
F	debt_consolidation	266	115	151	43.23%
F	home_improvement	25	10	15	40.00%
F	house	28	11	17	39.29%
F	credit_card	31	12	19	38.71%
F	small_business	44	15	29	34.09%
F	car	3	1	2	33.33%
F	medical	6	2	4	33.33%
F	vacation	9	2	7	22.22%
F	moving	12	2	10	16.67%
G	small_business	15	8	7	53.33%
G	major_purchase	2	1	1	50.00%
G	medical	2	1	1	50.00%
G	other	19	8	11	42.11%
G	credit_card	5	2	3	40.00%
G	debt_consolidation	20	5	15	25.00%
G	house	11	2	9	18.18%
G	car	1	-	1	0.00%
G	moving	4	-	4	0.00%
G	vacation	1	-	1	0.00%

- Customers' grade do vary by purpose.
- For grade A,B (good customers) tend to spend their loan on vacation, debt consolidation and credit card
- For grade C , the main purpose of borrowing is to spend on credit card, debt consolidation and medical
- Inside every grade we can see that the ranges of percent defaults are rising accordingly. In group A our default ranges from 2.53-8.70%, group B: 8.75-15.48%, group C: 14.09-22.54%, group D: 16.52-28.49, group E 25.81-36.11%, group F: 16.67-72.73%, group G: 18.18-53.33%. So at the grade grouping levels go from A to G the percentage of defaults by grades rise. The only grade class we see lower in ranges is grade G. The reason that we can see ranges fall for grade G is from the total number of observations lowering from 509 to 80. With grade F having 6.4 times the numbers of observations in lower grades we can expect that.

- We can also see that within every group letter the purposes for defaulting on a loan vary completely within that class. Grade classes are not affected by the results of the purposes from other grade classes.

Do defaults vary by purpose?

purpose	nLoans	defaults	percent_defaults
educational	1	-	0.00
wedding	2	-	0.00
credit_card	24,051	3,016	12.54
home_improvement	5,502	751	13.65
car	950	130	13.68
major_purchase	1,705	252	14.78
debt_consolidation	60,669	9,589	15.81
vacation	699	117	16.74
other	5,119	870	17.00
medical	991	181	18.26
house	385	75	19.48
moving	705	141	20.00
renewable_energy	62	14	22.58
small_business	885	205	23.16

- Yes, Defaults vary by purpose. From the table below we observe that percent defaults are less for educational or wedding purposes, while it is high for purposes such as renewable energy and small businesses.

iv) Calculate the annual return. Show how you calculate the percentage annual return. Compare the average return values with the average interest_rate on loans – do you notice any differences, and how do you explain this?

- Annual Return = (Total Payments - Funded amount)*(12/36)
- Percentage of Annual Return = (Total Payments - Funded amount)/(Funded amount) *(12/36)*100
- The actual term represents the duration of each account from the issued date to the last payment date, generally speaking it's the age of each loan account.
  Therefore, the calculation of actual return will interpolate by its duration of the loan.
  This reveals the actual profit that Lending Club had made.

nLoans	defaults	defaultRate	avgInterst	avgLoanAmt	avgRet	avgActualRet	avgActualTerm	minActualRet	maxActualRet
101,726.0	15,341.0	15.08	11.53	12,816.53	2.29	4.69	2.26	(33.33)	41.19

- When we compare the interest rate (mean(int_rate)= 11.5%) collected from customers with average actual return (4.7%) , there is a big difference between average interest rate and Annual return. This is the indicator showing that Lending Club gets a low return in business.

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?

grade	nLoans	defaults	defaultRate	avgInterst	avgLoanAmt	avgRet	avgActualRet	avgActualTerm	minActualRet	maxActualRet
A	25,751	1,486	5.77	7.06	14,363.95	2.27	3.67	2.27	(33.33)	16.37
B	30,110	3,657	12.15	10.20	12,610.44	2.50	4.61	2.25	(33.33)	23.20
C	28,369	5,344	18.84	13.34	11,738.75	2.33	5.24	2.24	(32.27)	39.73
D	13,040	3,375	25.88	16.62	12,445.83	2.05	5.50	2.28	(33.33)	36.05
E	3,866	1,238	32.02	19.31	13,559.79	1.39	5.18	2.32	(33.33)	41.19
F	510	214	41.96	23.90	11,109.61	0.62	4.83	2.42	(32.06)	36.56
G	80	27	33.75	25.90	9,875.00	2.36	7.86	2.27	(24.71)	30.88

- Yes, returns vary by grade. As we observe from above table Grade G has the highest return, whereas Grades A through E has higher return (high risk, high return). The F Grade is an exception, it has very low percent returns.
- Grade G has the highest return, however the number of loans is less than other grades which mean the number of customers who will get this grade is low.
- If we have to target by volume. Grade A, B, C are the highest proportion of this data. We will get the highest payback money.

sub_grade	nLoans	defaults	defaultRate	avgInterst	avgLoanAmt	avgRet	avgActualRet	avgActualTerm	minActualRet	maxActualRet
A1	3,202.00	73.0	2.28	5.87	14,508.77	2.30	3.44	2.27	(25.31)	10.35
A2	4,247.00	162.0	3.81	6.40	14,182.67	2.26	3.53	2.26	(31.31)	11.35
A3	4,037.00	216.0	5.35	6.79	14,312.51	2.23	3.57	2.27	(33.33)	12.24
A4	6,207.00	381.0	6.14	7.11	14,609.97	2.22	3.64	2.26	(32.31)	12.71
A5	8,058.00	654.0	8.12	7.97	14,238.22	2.32	3.92	2.27	(33.33)	16.37
B1	5,229.00	503.0	9.62	8.32	13,263.81	2.22	3.90	2.26	(33.33)	14.53
B2	5,911.00	613.0	10.37	9.26	13,197.17	2.43	4.31	2.25	(32.30)	20.21
B3	5,995.00	688.0	11.48	10.13	12,931.65	2.59	4.65	2.26	(32.27)	21.78
B4	6,075.00	795.0	13.09	11.11	11,948.70	2.72	5.08	2.24	(32.53)	20.92
B5	6,900.00	1,058.0	15.33	11.67	11,916.19	2.50	4.96	2.25	(33.33)	23.20
C1	6,653.00	1,059.0	15.92	12.32	11,641.48	2.50	5.20	2.23	(32.27)	25.09
C2	6,417.00	1,181.0	18.40	12.78	11,566.56	2.19	4.98	2.23	(32.26)	24.18
C3	5,777.00	1,079.0	18.68	13.42	11,862.68	2.47	5.35	2.25	(32.22)	30.40
C4	5,044.00	1,068.0	21.17	14.08	12,115.61	2.20	5.20	2.25	(32.26)	24.46
C5	4,478.00	957.0	21.37	14.75	11,545.65	2.23	5.55	2.23	(32.25)	39.73
D1	3,695.00	869.0	23.52	15.60	11,797.49	2.32	5.64	2.26	(32.20)	31.33
D2	2,903.00	741.0	25.53	16.36	11,970.94	2.01	5.49	2.26	(33.33)	29.91
D3	2,611.00	679.0	26.01	16.84	12,949.75	2.14	5.63	2.28	(33.33)	31.06
D4	2,024.00	561.0	27.72	17.43	12,607.70	1.63	5.17	2.29	(32.17)	36.05
D5	1,807.00	525.0	29.05	17.86	13,625.08	1.91	5.39	2.33	(32.21)	34.06
E1	1,318.00	414.0	31.41	18.33	13,980.41	1.45	5.07	2.32	(33.33)	34.20
E2	966.00	280.0	28.99	18.93	13,580.80	1.74	5.59	2.29	(30.99)	30.94
E3	781.00	257.0	32.91	19.62	13,869.08	1.18	4.79	2.36	(33.33)	33.33
E4	468.00	163.0	34.83	20.62	12,439.53	1.05	5.42	2.28	(30.88)	36.74
E5	333.00	124.0	37.24	21.69	12,683.11	1.11	4.98	2.41	(30.84)	41.19
F1	196.00	79.0	40.31	22.90	11,125.13	0.92	5.69	2.33	(28.76)	36.56
F2	122.00	51.0	41.80	23.91	11,906.56	0.18	3.57	2.53	(28.19)	35.45
F3	81.00	30.0	37.04	24.46	9,720.37	2.40	6.74	2.42	(32.06)	32.81
F4	66.00	32.0	48.48	24.99	9,792.42	(0.62)	3.08	2.50	(30.78)	31.17
F5	45.00	22.0	48.89	25.58	13,313.89	(0.88)	3.65	2.40	(29.42)	34.95
G1	28.00	8.0	28.57	25.94	9,550.89	3.60	8.55	2.26	(24.71)	26.58
G2	26.00	10.0	38.46	25.83	9,663.46	1.65	6.92	2.30	(22.25)	27.07
G3	17.00	5.0	29.41	25.89	9,255.88	2.94	7.98	2.39	(24.69)	30.88
G4	7.00	2.0	28.57	25.99	14,671.43	3.61	14.84	1.68	(10.86)	27.33
G5	2.00	2.0	100.00	26.06	5,637.50	(14.98)	(14.98)	3.00	(22.07)	(7.89)

- Yes, returns vary by sub-grade. For sub-grades A1 through D5 they show returns around 3.4 to 5.6 percent.
- Returns for sub-grades E1 to E5 are lower than sub-grade A1 to D5.
- Returns for subgrades G1- G4 are exceptionally high but G5 is negative.
- So, there is some amount of consistency in returns for sub-grades A1 through C5.

If you wanted to invest in loans based on this data exploration, which loans would you invest in?

- We would invest in loans graded from A to B because the actual return from these grade are high with return of 3.4 to 5.1% with default rate below the overall default rate of (15.1%)

(v) Generate some new derived attributes which you think may be useful for predicting default., and explain what these are.

We have generated following new attributes:

1. ratio_annualinc_loanamt = annual_inc / loan_amnt

loan_status	mean(ratio_annualinc_loanamt)
Charged Off	6.83
Fully Paid	7.62

- This attribute is a ratio of annual income to the loan amount.
- The “charged off” borrowers have a relatively low ratio of annual income to the loan amount compared to “fully paid” borrowers.
- The higher income of borrowers, the better characteristic they are (potential to be Fully paid customers)

2. ratio_install_loanamt = installment/loan_amnt

grade	loan_status	mean(ratio_install_loanamt)
A	Charged Off	0.0310
A	Fully Paid	0.0309
B	Charged Off	0.0325
B	Fully Paid	0.0323
C	Charged Off	0.0339
C	Fully Paid	0.0339
D	Charged Off	0.0355
D	Fully Paid	0.0354
E	Charged Off	0.0369
E	Fully Paid	0.0368
F	Charged Off	0.0392
F	Fully Paid	0.0392
G	Charged Off	0.0402
G	Fully Paid	0.0402

- This attribute is a ratio of installment to the loan amount.
- The “charged off” borrowers have slightly higher ratio of installment to the loan amount compared to “fully paid”

(b) 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?

- Yes, There are missing values
- The proportion of missing values in different variables is as follows:

no	Variable	Proportion of NA
1	id	1.00
2	member_id	1.00
3	loan_amnt	0.00
4	funded_amnt	0.00
5	funded_amnt_inv	0.00
6	term	0.00
7	int_rate	0.00
8	installment	0.00
9	grade	0.00
10	sub_grade	0.00
11	emp_title	0.06
12	emp_length	0.00
13	home_ownership	0.00
14	annual_inc	0.00
15	verification_status	0.00
16	issue_d	0.00
17	loan_status	0.00
18	pymnt_plan	0.00
19	url	1.00
20	desc	1.00
21	purpose	0.00
22	title	0.00
23	zip_code	0.00
24	addr_state	0.00
25	dti	0.00
26	delinq_2yrs	0.00
27	earliest_cr_line	0.00
28	inq_last_6mths	0.00
29	mths_since_last_delinq	0.48
30	mths_since_last_record	0.81

30	mths_since_last_record	0.81
31	open_acc	0.00
32	pub_rec	0.00
33	revol_bal	0.00
34	revol_util	0.00
35	total_acc	0.00
36	initial_list_status	0.00
37	out_prncp	0.00
38	out_prncp_inv	0.00
39	total_pymnt	0.00
40	total_pymnt_inv	0.00
41	total_rec_prncp	0.00
42	total_rec_int	0.00
43	total_rec_late_fee	0.00
44	recoveries	0.00
45	collection_recovery_fee	0.00
46	last_pymnt_d	0.00
47	last_pymnt_amnt	0.00
48	next_pymnt_d	1.00
49	last_credit_pull_d	0.00
50	collections_12_mths_ex_med	0.00
51	mths_since_last_major_derog	0.70
52	policy_code	0.00
53	application_type	0.00
54	annual_inc_joint	1.00
55	dti_joint	1.00
56	verification_status_joint	1.00
57	acc_now_delinq	0.00
58	tot_coll_amt	0.00
59	tot_cur_bal	0.00
60	open_acc_6m	1.00

61	open_act_il	1.00
62	open_il_12m	1.00
63	open_il_24m	1.00
64	mths_since_rcnt_il	1.00
65	total_bal_il	1.00
66	il_util	1.00
67	open_rv_12m	1.00
68	open_rv_24m	1.00
69	max_bal_bc	1.00
70	all_util	1.00
71	total_rev_hi_lim	0.00
72	inq_fi	1.00
73	total_cu_tl	1.00
74	inq_last_12m	1.00
75	acc_open_past_24mths	0.00
76	avg_cur_bal	0.00
77	bc_open_to_buy	0.01
78	bc_util	0.01
79	chargeoff_within_12_mths	0.00
80	delinq_amnt	0.00
81	mo_sin_old_il_acct	0.04
82	mo_sin_old_rev_tl_op	0.00
83	mo_sin_rcnt_rev_tl_op	0.00
84	mo_sin_rcnt_tl	0.00
85	mort_acc	0.00
86	mths_since_recent_bc	0.01
87	mths_since_recent_bc_dlq	0.73
88	mths_since_recent_inq	0.11
89	mths_since_recent_revol_delinq	0.63
90	num_accts_ever_120_pd	0.00

91	num_actv_bc_tl	0.00
92	num_actv_rev_tl	0.00
93	num_bc_sats	0.00
94	num_bc_tl	0.00
95	num_il_tl	0.00
96	num_op_rev_tl	0.00
97	num_rev_accts	0.00
98	num_rev_tl_bal_gt_0	0.00
99	num_sats	0.00
100	num_tl_120dpd_2m	0.04
101	num_tl_30dpd	0.00
102	num_tl_90g_dpd_24m	0.00
103	num_tl_op_past_12m	0.00
104	pct_tl_nvr_dlq	0.00
105	percent_bc_gt_75	0.01
106	pub_rec_bankruptcies	0.00
107	tax_liens	0.00
108	tot_hi_cred_lim	0.00
109	total_bal_ex_mort	0.00
110	total_bc_limit	0.00
111	total_il_high_credit_limit	0.00
112	revol_bal_joint	1.00
113	sec_app_earliest_cr_line	1.00
114	sec_app_inq_last_6mths	1.00
115	sec_app_mort_acc	1.00
116	sec_app_open_acc	1.00
117	sec_app_revol_util	1.00
118	sec_app_open_act_il	1.00
119	sec_app_num_rev_accts	1.00
120	sec_app_chargeoff_within_12_mths	1.00
121	sec_app_collections_12_mths_ex_med	1.00
122	sec_app_mths_since_last_major_derog	1.00
123	hardship_flag	0.00
124	hardship_type	1.00
125	hardship_reason	1.00
126	hardship_status	1.00
127	deferral_term	1.00
128	hardship_amount	1.00
129	hardship_start_date	1.00
130	hardship_end_date	1.00
131	payment_plan_start_date	1.00
132	hardship_length	1.00
133	hardship_dpd	1.00
134	hardship_loan_status	1.00
135	orig_projected_additional_accrued_interest	1.00

Explain how you will handle missing values for different variables:

General rules

1. Understanding the meaning of variable
2. Understanding the type of variable (Categorical or numerical)
3. Depending on the role of the variable on the predicted value various methods could be employed to replace the NULL values. For example, Impute the data with statistical value i.e. mean/ median , “other” or “blank” for unknown categorical value

- Emp_title
  - Employment title
  - categorical variable
  - We could replace NA by “Others”

- mths_since_last_delinq
  - Months since delinquency
  - Numerical variable

    - We can use use max function to impute data because we treat the customers as non-delinquent
- Bc_open_to_buy
    - Total open to buy on revolving bank cards.
    - Numerical variable
    - We can use use mean function to impute data
    -
- Bc_util
    - Ratio of total current balance to high credit/credit limit for all bankcard accounts.
    - Numerical variable
    - We can use use mean function to impute data

- Mo_sin_old_il_acct
    - Months since oldest installment account opened
    - Numerical variable
    - We can use zero to impute data because we treat the customers do not have installment account opened

- Mths_since_recent_bc
    - Months since the most recent bankcard account opened.
    - Numerical variable
    - We can use max to impute data because we treat the customers' recent bankcard account opened.

- Mths_since_recent_inq
    - Months since the most recent inquiry.
    - Numerical variable
    - We can use max function to impute data because we do not have information about the recent inquiry.

- Num_tl_120dpd_2m
    - Number of accounts currently 120 days past due (updated in past 2 months)
    - Numerical variable
    - We can use the mean function to impute data because we consider conservative approaches.
- Percent_bc_gt_75
    - Percentage of all bankcard accounts > 75% of limit.
    - Numerical variable
    - We can use the mean function to impute data because we assume that the account with N/A has the same proportion of other accounts

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). For example, it has been noted that the FICO scores on loan applicants are updated periodically, and the data can carry thus FICO scores from after the loan issue_date. So, even though FICO score can be useful, the values in the data may not be usable. Identify and explain which variables will you exclude from the model.

Potential Data Leakage Variables

- Emp_title - change over the time
- Emp_length - change over the time
- Delinq_2yrs - change over the time
- Inq_last_6mths - happen after lending
- Open_acc - change over the time
- Pub_rec - change over the time
- Revol_bal - change over the time
- Title - unnecessary for calculation
- zip_code - unnecessary for calculation
- Addr_state - unnecessary for calculation
- Out_prncp - change over the time
- Out_prncp - change over the time
- Out_prncp_inv - change over the time
- Recoveries - change over the time
- Collection_recovery_fee - change over the time
- Total_acc - change over the time
- Last_pymnt_d - notice after lending
- Last_pymnt_amnt - notice after lending
- Last_credit_pull_d - notice after lending
- Collections_12_mths_ex_med - notice after lending
- Acc_now_delinq - notice after lending
- Tot_coll_amt - inconsistent timing
- Tot_cur_bal - inconsistent timing
- Total_rev_hi_lim - inconsistent timing
- Acc_open_past_24mths - change over the time
- Avg_cur_bal - inconsistent timing
- Bc_open_to_buy - change over the time
- Bc_util - change over the time
- Chargeoff_within_12_mths - inconsistent timing
- Delinq_amnt - inconsistent timing
- Mort_acc - inconsistent timing
- Mths_since_recent_bc - change over the time
- Mo_sin_old_il_acc t- change over the time
- Num_tl_120dpd_2m - change over the time
- Percent_bc_gt_75 - inconsistent timing

- ❖ Mths_since_recent_inq - inconsistent timing
- ❖ Hardship_flag - inconsistent timing

- We **drop** column which consisted of **NA >= 60%** of each entire column
- If the column with NA< 40% will be filled out with appropriate value.
- **Data leakage** is inconsistent with the time of data which is not available at the time we make predictions.
- For this problem, the objective is to predict the loan status from data of Jan 2015 - May 2015 (Range of loan issued date) and this dataset was retrieved in Sep 2018 (Max of Last payment date). Therefore, the data in the table which does not belong to 2015 could be eliminated. For example, delinquency within 2 years. The values which keep updating consistently are considered to be the datakage.
- We not only eliminate the data leakage, but we also drop out the **single value data**, namely term, pymnt_plan, policy_code, application_type. These variables contain all similar values in the column. Thus we can delete them in our data preparation process.

4. Develop decision tree models to predict default.

(a) Split the data into training and validation sets. What proportions do you consider, why?

- We split the data into 70% training sets and 30% into Test set. We need to ensure that our model is trained on maximum data so that it gets maximum accuracy. 70% training data will help the model to learn much better.
- Total number of rows is 101,726. 71,208 Training dataset and 30,518 for Testing dataset

(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 parameters that we use for this experiment
(after deleted the NA, data-leakage, single value):
"Loan_amnt", "funded_amnt", "funded_amnt_inv", "term","int_rate","installment" ,
"grade", "sub_grade", "home_ownership", "annual_inc", "verification_status", "issue_d",
"loan_status", "purpose", "dti", "earliest_cr_line", "mths_since_last_delinq", "revol_util",
"initial_list_status", "total_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int",
"total_rec_late_fee", "mo_sin_old_rev_tl_op", "mo_sin_rcnt_rev_tl_op", "mo_sin_rcnt_tl",
"num_accts_ever_120_pd", "num_actv_bc_tl", "num_actv_rev_tl", "num_bc_sats",
"num_bc_tl", "num_il_tl", "num_op_rev_tl", "num_rev_accts" , "num_rev_tl_bal_gt_0",
"num_sats", "num_tl_30dpd" ,"num_tl_90g_dpd_24m" ,"num_tl_op_past_12m",
"pct_tl_nvr_dlq", "pub_rec_bankruptcies", "tax_liens", "tot_hi_cred_lim", "total_bal_ex_mort",
"total_bc_limit", "total_il_high_credit_limit", "disbursement_method", "debt_settlement_flag"

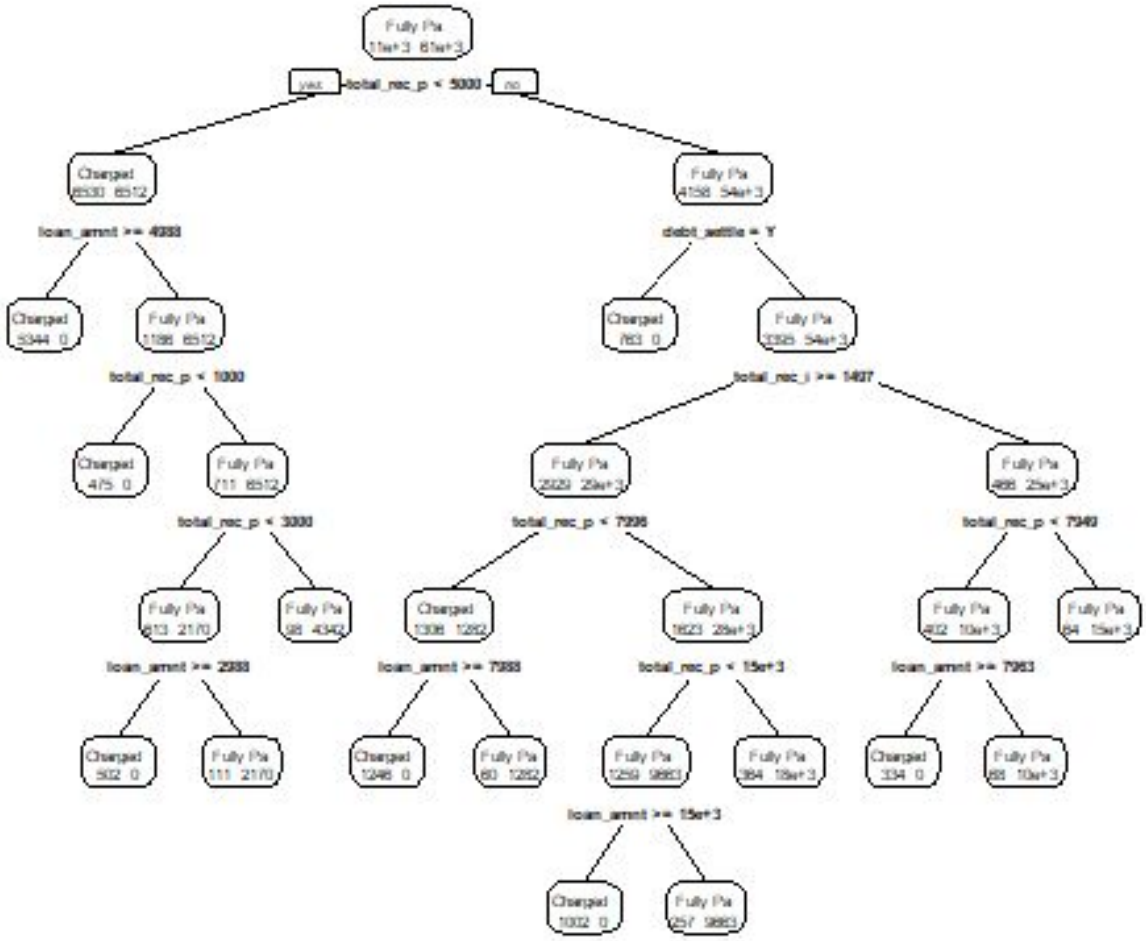
Decision Tree 1 with minimum split 30:

```
lcDT1 <- rpart(loan_status ~., data=lcdfTrn, method="class",
+               parms = list(split = "information"), control = rpart.control(minsplit = 30))
> print(lcDT1)
n= 71208

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 71208 10688 Fully Paid (0.150095495 0.849904505)
    2) total_rec_prncp< 4999.605 13042  6512 Charged Off (0.500690078 0.499309922)
      4) loan_amnt>=4987.5 5344     0 Charged Off (1.000000000 0.000000000) *
      5) loan_amnt< 4987.5 7698  1186 Fully Paid (0.154065991 0.845934009)
       10) total_rec_prncp< 999.885 475     0 Charged Off (1.000000000 0.000000000) *
       11) total_rec_prncp>=999.885 7223   711 Fully Paid (0.098435553 0.901564447)
         22) total_rec_prncp< 2999.645 2783   613 Fully Paid (0.220265900 0.779734100)
           44) loan_amnt>=2987.5 502     0 Charged Off (1.000000000 0.000000000) *
           45) loan_amnt< 2987.5 2281   111 Fully Paid (0.048662867 0.951337133) *
         23) total_rec_prncp>=2999.645 4440    98 Fully Paid (0.022072072 0.977927928) *
    3) total_rec_prncp>=4999.605 58166  4158 Fully Paid (0.071485060 0.928514940)
      6) debt_settlement_flag=Y 763     0 Charged Off (1.000000000 0.000000000) *
      7) debt_settlement_flag=N 57403  3395 Fully Paid (0.059143250 0.940856750)
       14) total_rec_int>=1496.725 32002  2929 Fully Paid (0.091525530 0.908474470)
         28) total_rec_prncp< 7996.08 2588  1282 Charged Off (0.504636785 0.495363215)
           56) loan_amnt>=7987.5 1246     0 Charged Off (1.000000000 0.000000000) *
           57) loan_amnt< 7987.5 1342    60 Fully Paid (0.044709389 0.955290611) *
```

29) total_rec_prncp>=7996.08 29414 1623 Fully Paid (0.055177806 0.944822194)
  58) total_rec_prncp< 14995.26 10922 1259 Fully Paid (0.115271928 0.884728072)
    116) loan_amnt>=14987.5 1002   0 Charged Off (1.000000000 0.000000000) *
    117) loan_amnt< 14987.5 9920   257 Fully Paid (0.025907258 0.974092742) *
  59) total_rec_prncp>=14995.26 18492   364 Fully Paid (0.019684188 0.980315812) *
15) total_rec_int< 1496.725 25401   466 Fully Paid (0.018345734 0.981654266)
  30) total_rec_prncp< 7948.55 10482   402 Fully Paid (0.038351460 0.961648540)
    60) loan_amnt>=7962.5 334   0 Charged Off (1.000000000 0.000000000) *
    61) loan_amnt< 7962.5 10148   68 Fully Paid (0.006700828 0.993299172) *
  31) total_rec_prncp>=7948.55 14919   64 Fully Paid (0.004289832 0.995710168) *



Decision Tree 1 with minimum split 50:

```
> lcDT1 <- rpart(loan_status ~., data=lcdfTrn, method="class",
+               parms = list(split = "information"), control = rpart.control(minsplit = 50))
> print(lcDT1)
n= 71208

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 71208 10688 Fully Paid (0.150095495 0.849904505)
    2) total_rec_prncp< 4999.605 13042  6512 Charged Off (0.500690078 0.499309922)
      4) loan_amnt>=4987.5 5344     0 Charged Off (1.000000000 0.000000000) *
      5) loan_amnt< 4987.5 7698  1186 Fully Paid (0.154065991 0.845934009)
       10) total_rec_prncp< 999.885 475     0 Charged Off (1.000000000 0.000000000) *
       11) total_rec_prncp>=999.885 7223   711 Fully Paid (0.098435553 0.901564447)
         22) total_rec_prncp< 2999.645 2783   613 Fully Paid (0.220265900 0.779734100)
           44) loan_amnt>=2987.5 502     0 Charged Off (1.000000000 0.000000000) *
           45) loan_amnt< 2987.5 2281   111 Fully Paid (0.048662867 0.951337133) *
         23) total_rec_prncp>=2999.645 4440    98 Fully Paid (0.022072072 0.977927928) *
    3) total_rec_prncp>=4999.605 58166  4158 Fully Paid (0.071485060 0.928514940)
      6) debt_settlement_flag=Y 763     0 Charged Off (1.000000000 0.000000000) *
      7) debt_settlement_flag=N 57403  3395 Fully Paid (0.059143250 0.940856750)
       14) total_rec_int>=1496.725 32002  2929 Fully Paid (0.091525530 0.908474470)
         28) total_rec_prncp< 7996.08 2588  1282 Charged Off (0.504636785 0.495363215)
           56) loan_amnt>=7987.5 1246     0 Charged Off (1.000000000 0.000000000) *
           57) loan_amnt< 7987.5 1342    60 Fully Paid (0.044709389 0.955290611) *
         29) total_rec_prncp>=7996.08 29414  1623 Fully Paid (0.055177806 0.944822194)
           58) total_rec_prncp< 14995.26 10922  1259 Fully Paid (0.115271928 0.884728072)
            116) loan_amnt>=14987.5 1002     0 Charged Off (1.000000000 0.000000000) *
            117) loan_amnt< 14987.5 9920   257 Fully Paid (0.025907258 0.974092742) *
           59) total_rec_prncp>=14995.26 18492   364 Fully Paid (0.019684188 0.980315812) *
       15) total_rec_int< 1496.725 25401   466 Fully Paid (0.018345734 0.981654266)
         30) total_rec_prncp< 7948.55 10482   402 Fully Paid (0.038351460 0.961648540)
           60) loan_amnt>=7962.5 334     0 Charged Off (1.000000000 0.000000000) *
           61) loan_amnt< 7962.5 10148    68 Fully Paid (0.006700828 0.993299172) *
         31) total_rec_prncp>=7948.55 14919    64 Fully Paid (0.004289832 0.995710168) *
```

As per the result form the output console, we get the description of our decision tree.
2),3) are the splits of Total_rec_prncp
4)*,5) are the splits under 2)
10)*,11) are the splits under 5)
22),23) are splits under 11)
44)*,45)* are splits under 22)

which “*” represents a leaf node.
We ended up with 14 leaf node with balance number of splits in each state.

Confusion Matrix with Train dataset:

		TRUE	
		Charge Off	Fully Paid
Predict	Charge Off	9666	0
	Fully Paid	1022	60520

> mean(predTrn == lcdfTrn$loan_status): **0.9856477**
Accuracy: .0.98564768 **≈ 98.6%**
TP Rate: .9044 **≈ 90.4%,**
FP Rate: **.00 ≈ 0%**
Precision: **1.00**
F1 score: 0.9498

Confusion Matrix with Test dataset:

		TRUE	
		Charge Off	Fully Paid
Predict	Charge Off	4161	0
	Fully Paid	492	25865

> mean(predict(lcDT1,lcdfTst, type='class') ==lcdfTst$loan_status): **0.9838784**
Accuracy: .0.98388 **≈ 98.4%**
TP Rate: .89246 **≈ 89%,**
FP Rate: **.00 ≈ 0%**
Precision: **1.00**
F1 Score = 0.9442

#Performance Evaluation with different classification threshold
CTHRESH=0.35

- Table with Train dataset

		TRUE	
		Charge Off	Fully Paid
Predict	Charge Off	9666	0
	Fully Paid	1022	60520

- Table with Test dataset

		TRUE	
		Charge Off	Fully Paid
Predict	Charge Off	1022	60520
	Fully Paid	9666	0

We tried thresholds 0.5, 0.99. The result for different thresholds is the same.

Performance evaluation based on confusion matrix with CARET package:

1. **Confusion matrix of Train Dataset:**

```
Confusion Matrix and Statistics

             Reference
Prediction    Charged Off Fully Paid
  Charged Off        9666          0
  Fully Paid         1022      60520

               Accuracy : 0.9856
                 95% CI : (0.9847, 0.9865)
    No Information Rate : 0.8499
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9414

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9044
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9834
             Prevalence : 0.1501
         Detection Rate : 0.1357
   Detection Prevalence : 0.1357
      Balanced Accuracy : 0.9522

       'Positive' Class : Charged Off
```

a.

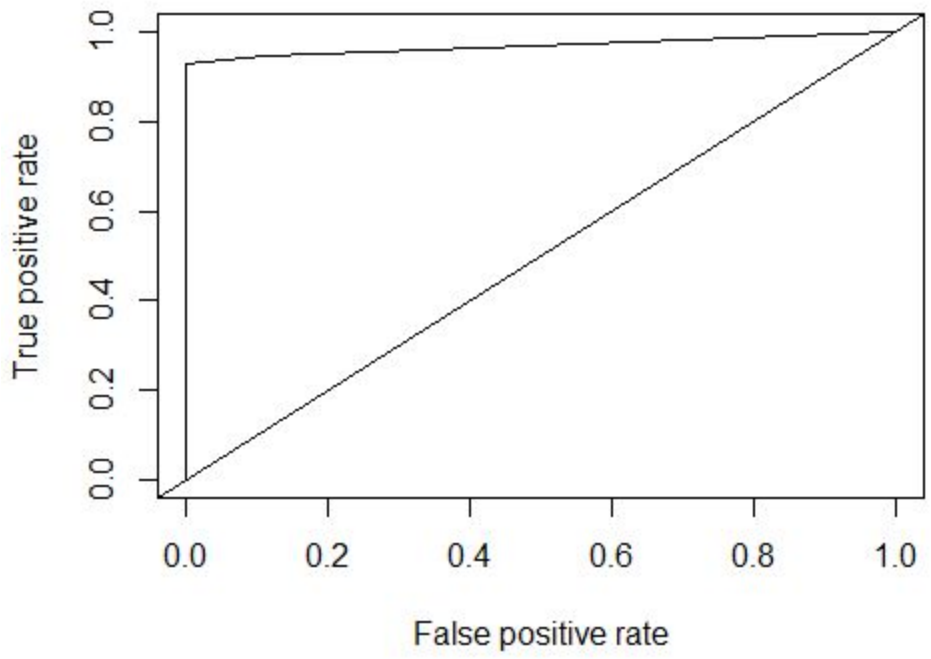
(c) Identify the best tree model. Why do you consider it best? Describe this model – in terms of complexity (size). Examine variable importance. Briefly describe how variable importance is obtained in your best model.

- For our experiment, we try to make a Decision Tree 1 with minimum split 30 and 50. According to the result we found the accuracy of the model approximately the same. Therefore, Decision Tree 1 with minimum split 30 is our best model.

- Variable importance is the goodness of split measures for each split for which it was the primary variable.

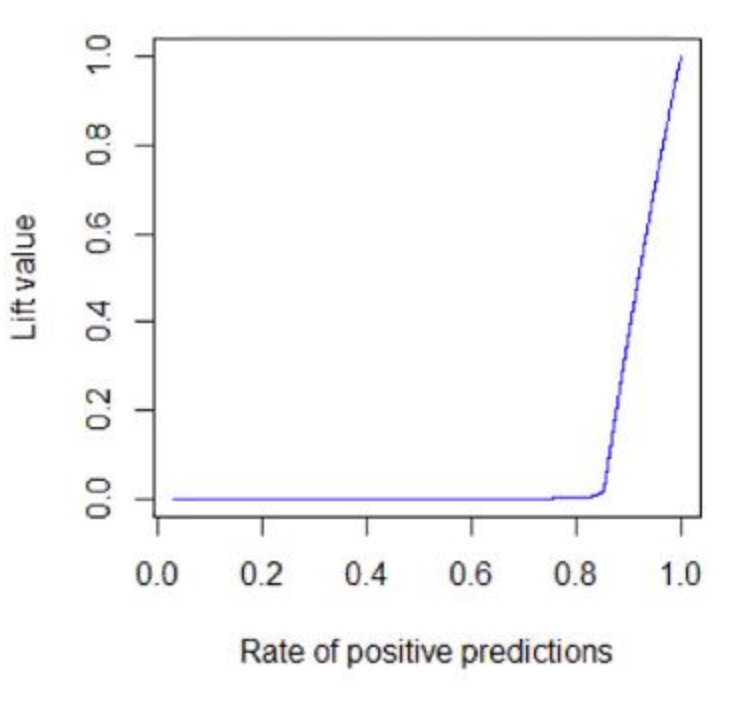
```
> lcDT1$variable.importance
       funded_amnt          loan_amnt      funded_amnt_inv         installment           total_pymnt
      18050.961020        18050.961020        17707.852877        11143.047060
9995.731563
    total_rec_prncp      total_pymnt_inv       total_rec_int debt_settlement_flag            int_rate
       9571.942529         7554.566149         4572.821119         2104.296214           98.809614
             grade     earliest_cr_line mo_sin_old_rev_tl_op
        96.399623           36.396499            4.149163
```

ROC Curve:



According to the ROC curve, the cut-offs the true positive rate is higher positive at the beginning of the graph (low false positive rate) Plus, the area under the curve is large nearly to 1 which refers to high efficiency tests.

Lift Curve:



5. Develop a random forest model. 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 random forest and best decision tree model from Q 4 above. Do you find the importance of variables to be different? Which model would you prefer, and why.

For evaluation of models, you should include confusion matrix related measures, as well as ROC analyses and lifts. Explain which performance measures you consider, and why.

What parameters do you experiment with, and does this affect performance?
The parameters that we use for this experiment
(after deleted the NA, data-leakage, single value)
Plus, we use exactly the same variable that we use for decision trees because we want to compare their performance, so we start with the same training and testing data set.

The parameters as follows:
"Loan_amnt", "funded_amnt", "funded_amnt_inv", "term","int_rate","installment" , "grade", "sub_grade", "home_ownership", "annual_inc", "verification_status", "issue_d", "loan_status", "purpose", "dti", "earliest_cr_line", "mths_since_last_delinq", "revol_util", "initial_list_status", "total_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "total_rec_late_fee", "mo_sin_old_rev_tl_op", "mo_sin_rcnt_rev_tl_op", "mo_sin_rcnt_tl", "num_accts_ever_120_pd", "num_actv_bc_tl", "num_actv_rev_tl", "num_bc_sats", "num_bc_tl", "num_il_tl", "num_op_rev_tl", "num_rev_accts" , "num_rev_tl_bal_gt_0", "num_sats", "num_tl_30dpd" ,"num_tl_90g_dpd_24m" ,"num_tl_op_past_12m",

"pct_tl_nvr_dlq", "pub_rec_bankruptcies", "tax_liens", "tot_hi_cred_lim", "total_bal_ex_mort", "total_bc_limit", "total_il_high_credit_limit", "disbursement_method", "debt_settlement_flag"

1. Building Random Forest = 300 trees

```
> rflcdf
Ranger result
Call:
 ranger(loan_status ~ ., data = lcdfTrn, num.trees = 300, importance = "permutation",
probability = TRUE)

Type:                         Probability estimation
Number of trees:                   300
Sample size:                     71208
Number of independent variables:  48
Mtry:                          6
Target node size:                  10
Variable importance mode:        permutation
Splitrule:                     gini
OOB prediction error (Brier s.):  0.007992539
```

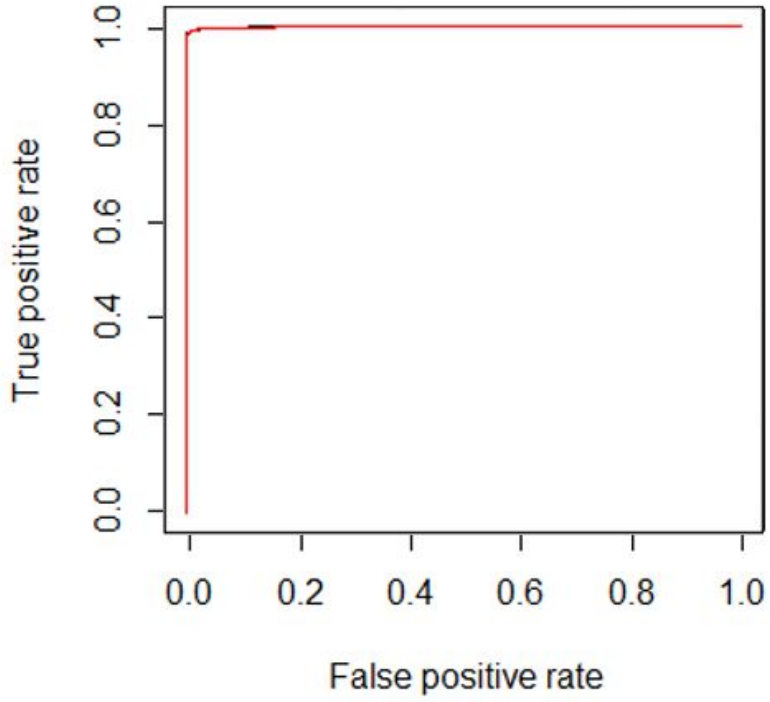
2. Building Random Forest = 700 trees

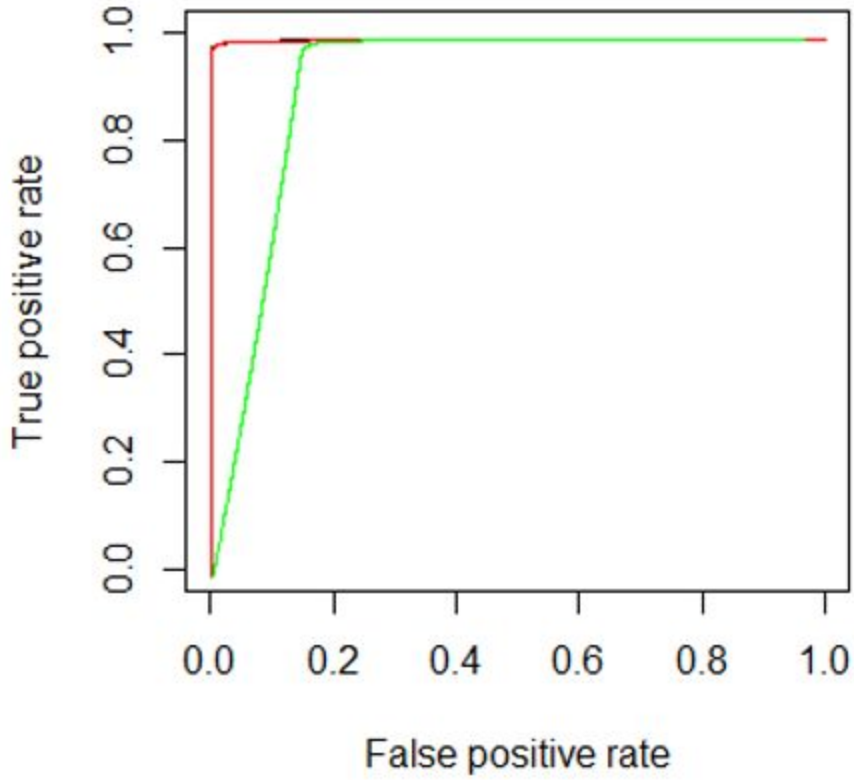
```
Ranger result

Call:
 ranger(loan_status ~ ., data = lcdfTrn, num.trees = 700, importance = "permutation",
probability = TRUE)

Type:                         Probability estimation
Number of trees:                   700
Sample size:                     71208
Number of independent variables:  48
Mtry:                          6
Target node size:                  10
Variable importance mode:        permutation
Splitrule:                     gini
OOB prediction error (Brier s.):  0.008143247
```

- The higher number of tree, the lower the error rate.


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Variable importance: error running

Confusion Maxtrix

- 300 Trees

		TRUE	
		Charge Off	Fully Paid
Predict	Charge Off	231	25958
	Fully Paid	4329	0

- 700 Trees

		TRUE	
		Charge Off	Fully Paid
Predict	Charge Off	231	25958
	Fully Paid	4329	0

Compare the random forest and best decision tree model from Q 4 above.
Do you find the importance of variables to be different?

- Random forest tends to give a lower error rate in each model.

6. The purpose of the model is to help make investment decisions on loans. How will you evaluate the models on this business objective? Consider a simplified scenario - for example, that you have $100 to invest in each loan, based on the model's prediction. So, you will invest in all loans that are predicted to be 'Fully Paid'. Key questions here are: how much, on average, can you expect to earn after 3 years from a loan that is paid off, and what is your potential loss from a loan that has to be charged off ? One can consider the average interest rate on loans for expected profit – is this a good estimate of your profit from a loan? For example, the average int_rate in the data is 11.2%; so after 3 years, the $100 will be worth (100 + 3*11.2) = 133.6, i.e a profit of $33.6. Now, is 11.2% a reasonable value to expect – what is the return you calculate from the data? Explain what *value of profit* you use. For a loan that is charged off, will the loss be the entire invested amount of $100? The data shows that such loans have do show some partial returned amount. Looking at the returned amount for charged off loans, what proportion of invested amount can you expect to recover? Is this overly optimistic? Explain which *value of loss* you use. You can also consider the alternate option of investing in, say in bank CDs (certificate of deposit); let's assume that this provides an interest rate of 2%. Then, if you invest $100, you will receive $106 after 3 years (not considering reinvestments, etc), for a profit of $6. Considering a confusion matrix, we can then have profit/loss amounts with each cell, as follows:

		Predicted	
		FullyPaid	ChargedOff
Actual	FullyPaid	*profitValue*	$6
	ChargedOff	*lossValue*	$6

For the perfect model, we assume that all customers are fully paid. The actual profit we will be 3*4.78 = 14.34 calculated from overall annual return rate of fully paid.

Overall interest rate is not a reasonable value for calculating the value of profit because this rate is the interest that Lending Clubs collect from their customers which does not represent the real profit of the company.

The potential of charge of calculated from

(a) Compare the performance of your models from Qs 4 and 5 above based on this. Note that the confusion matrix depends on the classification threshold/cutoff you use. Evaluate different thresholds and analyze performance. Which model do you think will be best, and why.

From Q4 we developed the decision tree and developed the Random Forest model in Q5. Random forest is better due to less error.

(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 analysis 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).