Assignment 1

Contributors:
Saraschandrika Addanki (saddan5@uic.edu : 658694881)
Mukesh Choudhary (mchoud20@uic.edu : 671652774)
Akash Bunde (abunde2@uic.edu : 665383604)

- Describe the business model for online lending platforms like Lending Club. Consider the stakeholders and their roles, and what advantages Lending Club offers. What is the attraction for investors? How does the platform make money?

LENDING CLUB

Business Model

Online Money Lending Platform is essentially a ‘Marketplace’ that brings together borrowers looking for unsecured loans with the investors *(Lenders)* who want to get high returns on their investments. So, Peer-to-Peer(p2p) Lending is a system where a borrower who needs money can get a loan from an investor who can provide loans at a certain interest rate where both parties’ respective needs meet *(or on agreed terms & conditions)*.

The total lending process can be done online. Here the company collects the information of the borrower like his annual income, credit score, the purpose of the lending, his plans on how to spend and get returns, etc. It then sends this information to the potential investors. After reviewing the details of the borrower, the investors decide whether to invest fully, fund partially, or not lend at all. The funding responses were then put forward to the borrower and he can decide from whom he wants to borrow.

These Platforms are useful for people who do not have a credit score, collateral for a secured loan, and want to get funding very quickly.

Lending has a huge cost advantage over traditional banks. Lending Club's cost savings are passed on to borrowers with the finest credit histories, resulting in reduced interest rates. Ideally Lending Club passes the risk of default loans to the Lenders.

Advantages of Lending Club

Advantages to Customers/Borrowers:

- One of the major advantages is convenience. Borrowers can shop through various loans and interest rates by sitting at home
- Borrowers enjoy a lower rate of interest as compared to traditional lending systems/banks
- The process is simplified, secure and transparent
- It’s not an institution like a bank. All the operations are online, and savings are passed on to the stakeholders

Advantages to Creditors/Investors:

- Lenders or investors enjoy higher and faster yield
- It becomes a new asset class, which enables diversification, for the investors
- It is also transparent as all the information related to loans is available for investors

How does it make money?

Lending Club basically acts as a traditional bank between Investors and Borrowers. Because Lending Club only accepts applications from borrowers with excellent credit, investors may almost always expect a favorable return.

Main sources for Lending Club to make money are through Origination and service fees

- Depending on the loan grade and term, borrowers pay a one-time origination charge ranging from 1.11 percent to 5% of the total loan amount
- Investors, on the other hand, pay a 1% service fee on each payment received from a borrower

Lending Club also has cost advantage over traditional banks

- In the banking industry, efficiency is assessed by the operating ratio, which is defined as marketing costs divided by the total number of loans outstanding. This normally equates to roughly 5% to 7%. As a result, a $100 loan will cost the bank $5 to $7
- Lending Club, on the other hand, has an operating ratio of just 2%, which means that issuing a $100 loan costs the company only $2
- Since they operate fully online, they save on the number of Employees and do not incur any location cost

References:

1. https://foundationcapital.com/wpcontent/uploads/2020/04/FC_CharlesMoldow_TrillionDollarMarket.pdf

2. https://help.lendingclub.com/hc/en-us/articles/215437458-How-does-LendingClub-make-money-

3. https://vator.tv/news/2014-03-21-how-does-lending-club-make-money

2.a.

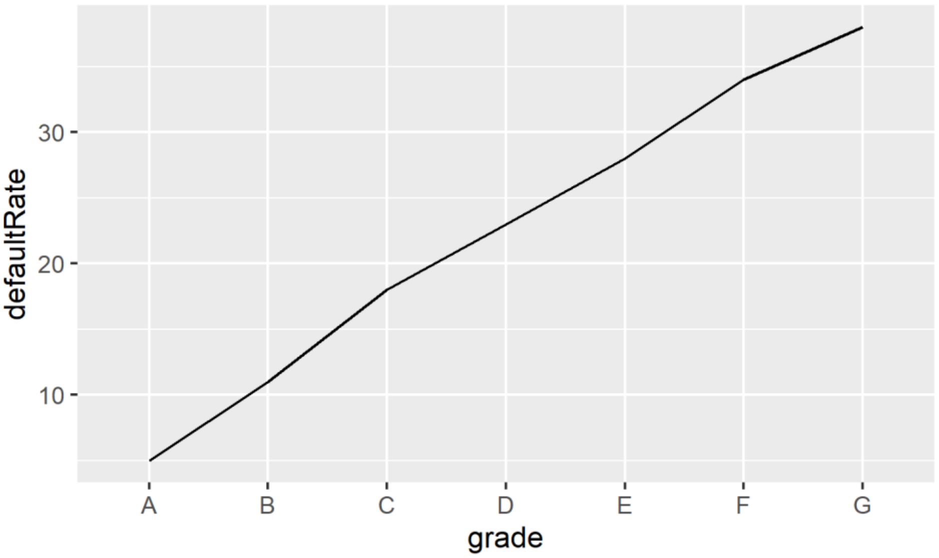
(i) What is the proportion of defaults ('charged off' vs 'fully paid' loans) in the data? How does default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why?

Ans:

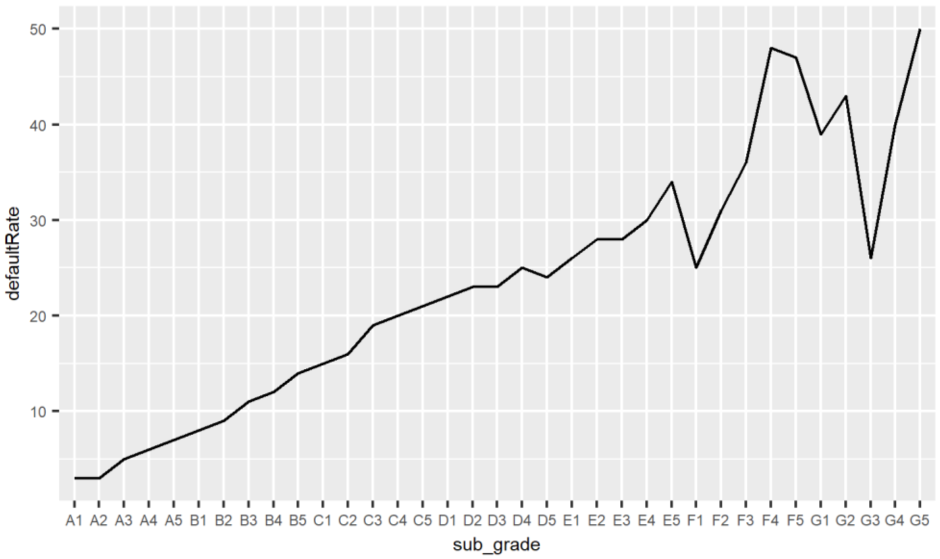
Defaults of charged off vs Fully paid: [13:86]

As the grade varies from A to G loan default rate increases. This is expected result as loan grades are assigned based on borrower's credit rating, income, employment factors, etc. Grade A are the safest ones and it is evident from the default rate that we have observed. Similar is the case with subgrades. Although there's variation in subgrades for grade F and G, overall trend seems to follow the initial reasoning.

Following chart shows variation in rate of defaults by loan grade.



Following chart shows variation in rate of defaults by loan sub-grade.



(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?

Ans:

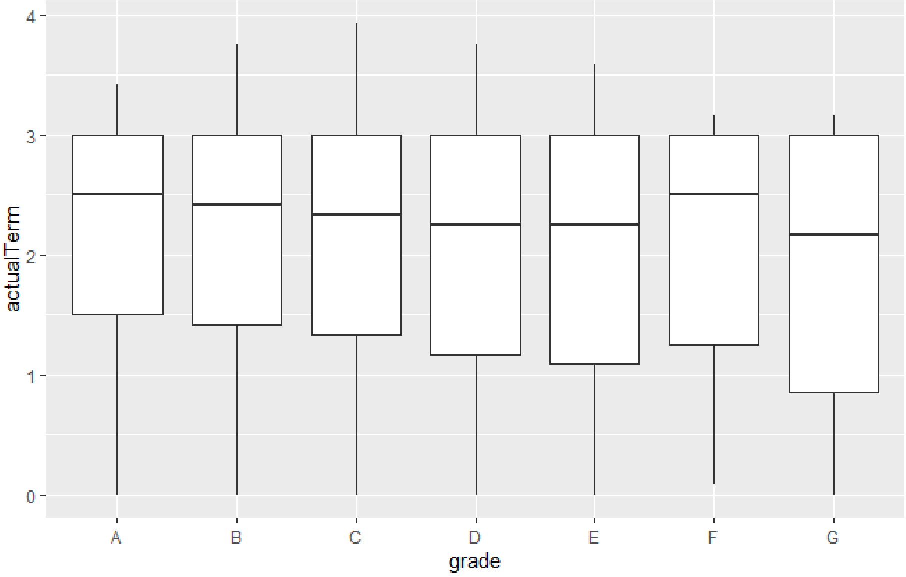
Loan amount varies by grade. We can see a decreasing trend in number of loans by grade. This can be explained by the risk appetite of the lenders. Majority of lenders will usually prefer loans that are safer and give regular returns.

Average loan amount also shows a decreasing trend with grade. Average loan amount for grade G is higher than F. From histogram plot, we can see that majority of loans for G are bucketed in the range of 1600-9000 loan amount but the average increases because of outliers.

Average interest rate follows an increasing trend for both Grades and Sub-Grades. This is expected as interest rate depends on credit ratings, employment factor, etc. These factors are showcased using Grade and Sub-Grade. Higher the risk, higher will be the interest rate.

(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 boxplot can help visualize this).

For all grades, people having 3 years actual period lie in the 75th percentile whereas the median lies between 2 and 2.5 years.



(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?

Percentage annual return = 100*((Payment – Funded Amount)/(Funded Amount))*(1/Actual Term)

Returns from 'charged off' loans are negative as expected. As loan grade changes, i.e., as it becomes riskier, the returns become more negative.

Average return values are less than the average interest rates. There are multiple reasons behind this. One of the reasons is that the loans are completed before their original term. Other reasons include adjustments in return rate by Lending Club based on future charged off rate and the service fee(1%) that the lender pays.

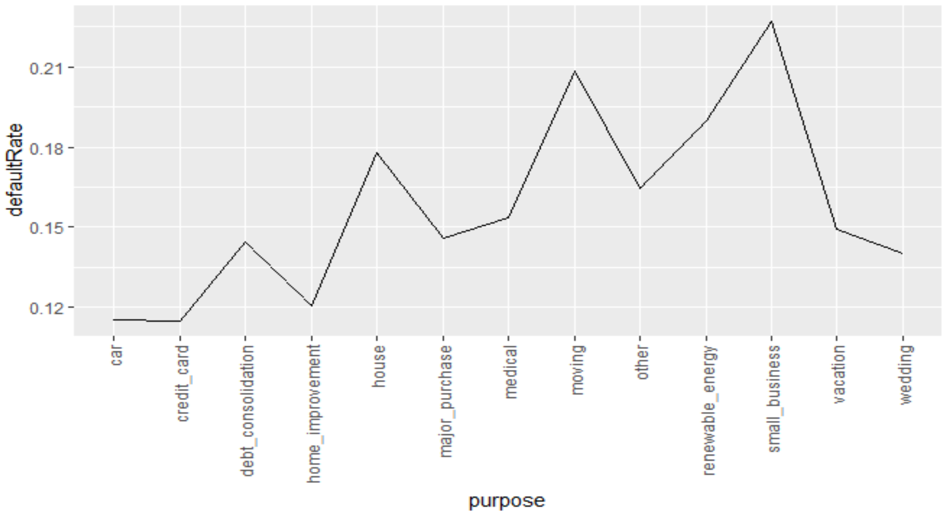
If I wanted to invest in loans, I would diversify my investments across loans from different grades based on average returns and default rate.

(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?

Ans:

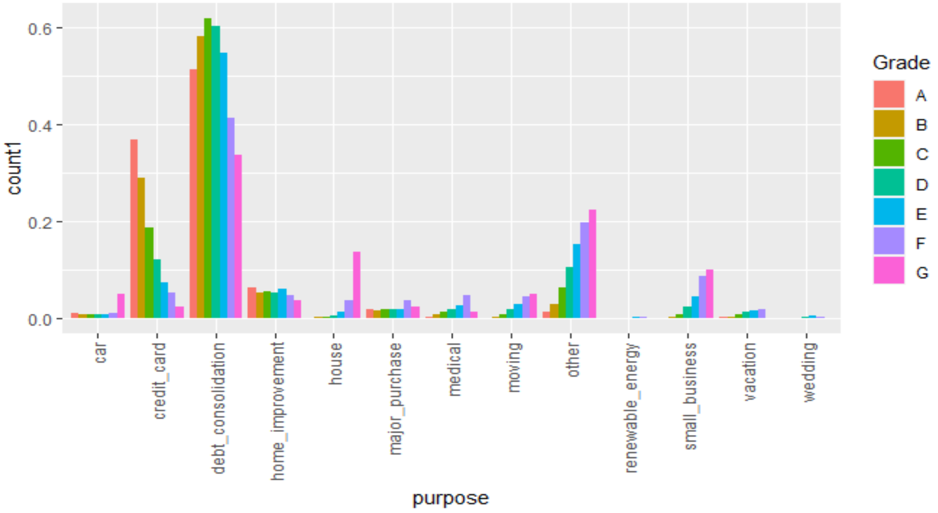
From the data, borrowers wanted money for 12 distinguished purposes. Loan amounts varied by purpose. Highest average loan was taken for ‘credit_card’ and it closely followed by ‘debt_consolidation’.

Following chart shows variation in default rate by purpose



We can observe that default rate is highest for ‘small_business’ loans.

The following chart shows variation in grade within purpose



Debt consolidation has the highest number for C grade loans. Most of the credit_card loans are tagged with grade A and most of the loans under ‘other’, ‘house’, ‘small business’, ‘car’ are tagged with grade G

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

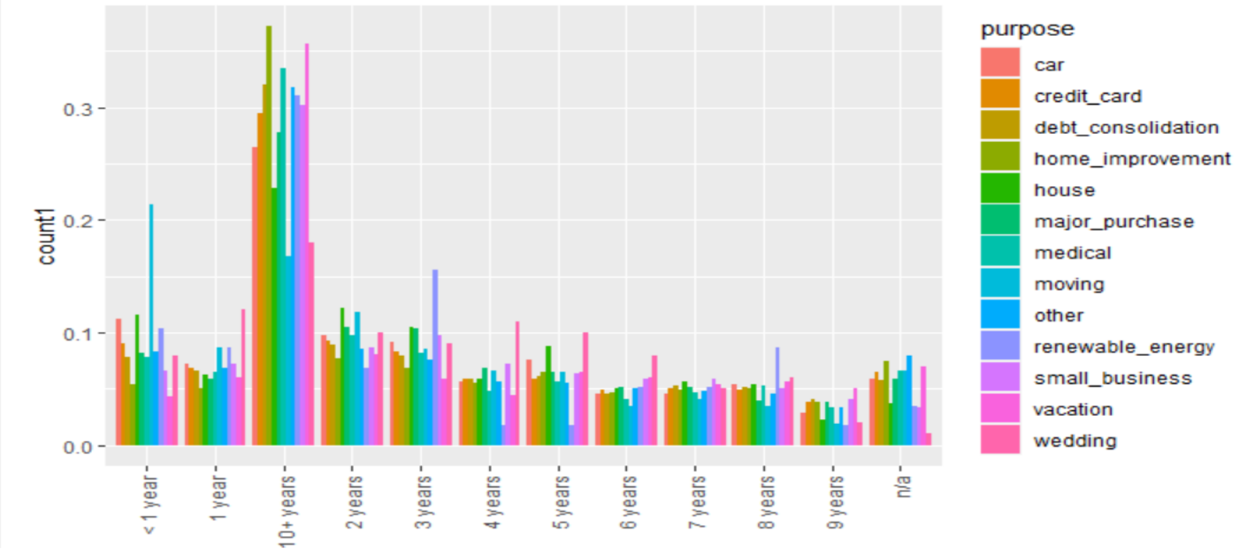
Ans:

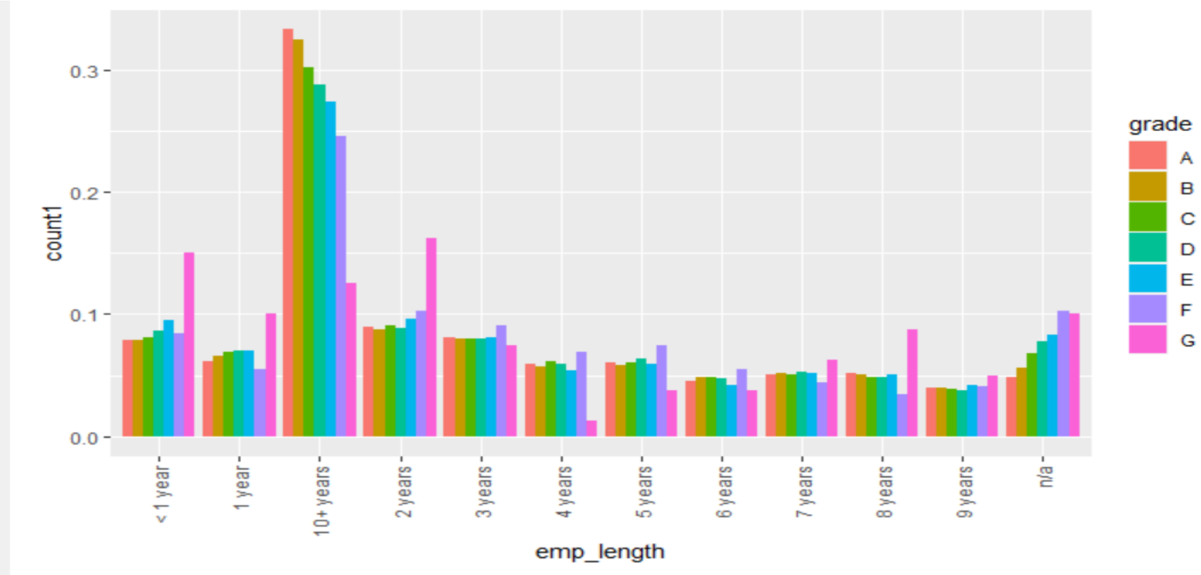
Relations between employment_length and Loan Amount: Average loan amount is almost same for all employment lengths

Relations between employment_length and grade: Lenders in their early careers seem to have lowest grade in contrast to the people with 10+ years experience have the best grade

Relations between employment_length and purpose: People with all 10+ years experience tend to take more loans for vacation and those with 4,5,6, years experience tend to take loans for wedding.

Relations between employment_length and actual return: Average Actual Return is approximately same for all employment lengths.





(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 analyses as in the questions above (as reasonable based on the derived variables).

Ans: One of the derived attributes is Total_Loss_to_inv – it consists of loss to the investors when the loan is charged off. It is derived by deducting funded_amnt from the total_pymnt.

Another one is Percent_of_rev_acc which is obtained by dividing num_op_rev_tl by the num_rev_acc. It depicts the open Credit Loans.

Third is Proportion of amount received (prop_amt_rec) which is obtained by summing up total_rec_int, total_rec_late_fee, total_rec_prncp and then dividing by funded-amnt. It is more than 1 for fully paid off loan and less than 1 for charged off loan.

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

Ans:

From the above analyses, we observed following points

1. Majority of loans are paid off (86%), which may show that their algorithm to identify viable candidates for loans is working
2. Loan grade and subgrade is an aggregated measure of the risk profile of a borrower.
3. Number of loans and average loan amount increases as loan grade becomes better (I.e., from G to A).

4. Average Default Rate increases as loan grade becomes worse (I.e., from A to G). This shows that loan grade is an effective predictor for defaulters.
5. Average interest rate and average returns increase as the loan grade becomes worse. This again becomes intuitive as loan grade becomes worse, the risk of a loan increases.

6. For fully paid loans, the median of actual term is less than 3 years. The actual term of 3 years is seen in the 75th percentile of borrowers.
7. Majority of loans are taken under ‘credit card’ and ‘debt consolidation’. Purpose can be a useful indicator of defaults.
8. Loans under ‘house’, ‘medical’, ‘moving’, ‘other’ and ‘renewable’ have more than overall default rate (14%). They also have higher proportion of loans with grade G
9. People in their career have least grade and people 10+years experience have the best grade.

(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 he 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?

Ans: 33 out of 145 columns are all missing. And 36 columns have some missing values. Columns which have more than 95% missing values can be excluded as we cannot just replace all of the values.

Types of Handling the Missing values:

1. For the variables with data margins is small and cannot be zero, we can prevent information loss by replacing the missing value with Median. Here missing values in number of revolving accounts and Months since most recent 90 day or worst rating can be replaced with median because 99% and 76% are filled and none of them are zero.

2. For Variables which have values only when an event occurs, the missing values can be safely assumed as zero because missing values may have caused due to the non occurance event. some such variables are Months since last inquiry, Months since recent Bankcard.

3. Some missing values cannot be replaced. but we cannot leave them as the outputs in some forms cannot show the existance of missing values and causes misinterpretation of the data. So we fill them with a fixed charecter or string to denote the missing value. here we used the string "missing"

Some missing values in columns like job title, purpose, employment duration cannot be handled. Missing values in num_rev_accts, num_tl_120dpd_2m, revol_util can be filled with the median value of the all other accounts. And Missing values in columns like pct_tl_nvr_dlq, mtths_since_recent_bc, mths_since_recent_inq can be replaced with zeros as it can be the possibility of no actions from the account handler.

We can see that some columns have same proportions of missing values. The values in those columns are related to their saperate installment and revolving accounts. They do not contribute any extra value to data model. We can extract that data and use for different purpose.

Col_name	na_proportion
num_rev_accts m	0.00001
avg_cur_bal	0.00002
last_credit_pull_d	0.00004
title	0.00012
pct_tl_nvr_dlq 0	0.00016

revol_util		0.00041
last_pymnt_d		0.00064
last_pymnt_d_new		0.00064
mths_since_recent_bc	0	0.00911
		0.00964
bc_open_to_buy		
percent_bc_gt_75		0.01034
bc_util		0.01044
mo_sin_old_il_acct		0.03620
num_tl_120dpd_2m	m	0.03824
emp_title		0.06705
mths_since_recent_inq	0	0.10612
mths_since_last_delinq	0	0.49919
mths_since_recent_revol_delinq		0.64746
mths_since_last_major_derog		0.71995
mths_since_recent_bc_dlq		0.74329
mths_since_last_record		0.82423
open_acc_6m		0.97313
open_act_il		0.97313
open_il_12m		0.97313
open_il_24m		0.97313
total_bal_il		0.97313
open_rv_12m		0.97313
open_rv_24m		0.97313
max_bal_bc		0.97313
all_util		0.97313
inq_fi		0.97313
total_cu_tl		0.97313
inq_last_12m		0.97313
mths_since_rcnt_il		0.97393
il_util		0.97694
settlement_term		0.99535
hardship_dpd		0.99955

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 will you exclude from the model.

Ans:

Variables not available during Loan Application

- loanStatus- The Status of the Loan.
- Total_pymnt - Total Amount Paid.
- Total_pymnt_inv - Payments received to date for portion of total amount funded by investors.

- Total_rec_late_fee - Late fees received to date.
- Recoveries - post charge off gross recovery.
- Collection_recovery_fee - post charge off collection fee.
- Last_pymnt_d - Last month payment was received.
- Last_pymnt_amnt - Last total payment amount received.
- Collections_12_mths_ex_med - Number of collections in 12 months excluding medical collections.
- Mths_since_last_major_derog - Months since most recent 90-day or worse rating.

Since these variables are not available before the Loan is Funded, including these variables can entirely destroy the purpose of Prediction. These variables will basically act as outliers in our dataset. Since these variables were known after the loan was funded, it can cause our prediction to be overly optimistic.

Variables which have been updated during Loan Period

- inqLast6Mths - The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
- mthsSinceLastDelinq - The number of months since the borrower's last delinquency.
- mthsSinceLastRecord - The number of months since the last public record.
- revolBal - Total credit revolving balance
- revolUtil - Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
- totalAcc - The total number of credit lines currently in the borrower's credit file
- initialListStatus - The initial listing status of the loan. Possible values are – W, F.
- Last_credit_pull_d - The most recent month LC pulled credit for this loan.
- Tot_cur_bal - Total current balance of all accounts.
- Open_acc_6m - Number of open trades in last 6 months
- Open_act_il - Number of currently active installment trades
- Open_il_12m - Number of installment accounts opened in past 12 months
- Open_il_24m - Number of installment accounts opened in past 24 months
- Mths_since_rcnt_il - Months since most recent installment accounts opened
- Total_bal_il - Total current balance of all installment accounts
- Il_util - Ratio of total current balance to high credit/credit limit on all install acct
- Open_rv_12m - Number of revolving trades opened in past 12 months
- Open_rv_24m - Number of revolving trades opened in past 24 months
- Max_bal_bc - Maximum current balance owed on all revolving accounts.
- Inq_last_12m - Number of credit inquiries in past 12 months
- AccOpenPast24Mths - Number of trades opened in past 24 months.
- Avg_cur_bal - Average current balance of all accounts.
- BcOpenToBuy - Total open to buy on revolving bankcards.
- Bc_util - Ratio of total current balance to high credit/credit limit for all bankcard accounts.
- Chargeoff_within_12_mths - Number of charge-offs within 12 months.

Since these variables are been changed constantly with the time as the loan duration proceeds, including these can cause discrepancies in our prediction model.

4. Do a univariate analyses 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).

Ans:

We have used AUC measure to identify the relationship between predictor (independent variable) (numeric variable) and dependent variable.

annualRet gives 0.98 AUC

And for the following variable the AUC is more 0.75

total_rec_prncp
last_pymnt_amnt
total_pymnt_inv
total_pymnt

But these variables have the issue of data leakage as they were generated after the loan was granted/funded or closed.

Rest of the variables' AUC lies around 0.5

PART B

Classification tree:

```
rpart(formula = loan_status ~ ., data = nDataTrn, method = "class",
    parms = list(loss = lossmatrix), control = rpart.control(cp = 0))
```

Variables actually used in tree construction:

addr_state, avg_cur_bal, bc_open_to_buy, bc_util,dti, earliest_cr_line, emp_length,emp_title, int_rate, loan_amnt, mo_sin_old_il_acct, mo_sin_old_rev_tl_op, mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl,

```
mths_since_last_major_derog, num_accts_ever_120_pd, num_il_tl, num_sats, purpose, revol_bal,
revol_util, sub_grade, title, tot_cur_bal, tot_hi_cred_lim, total_acc, total_bal_ex_mort, total_bc_limit,
total_il_high_credit_limit, zip_code

Root node error: 38504/70000 = 0.55006

n= 70000


           CP nsplit rel-error  xerror      xstd
1   4.4310e-01      0  1.000000 0.25000 0.0023664
2   3.8957e-02      1  0.556903 1.55929 0.0106260
3   3.3347e-02      3  0.478989 1.43159 0.0102390
4   1.1116e-02      4  0.445642 1.06846 0.0088829
5   1.1038e-02      7  0.412009 1.04311 0.0087660
6   9.2977e-03      8  0.400971 1.08571 0.0089526
7   9.2068e-03      9  0.391674 1.10560 0.0090408
8   8.5705e-03     11  0.373260 1.10222 0.0090261
9   8.0511e-03     12  0.364689 1.07106 0.0088931
10  6.3045e-03     13  0.356638 1.11638 0.0090896
11  4.9086e-03     17  0.331420 1.19668 0.0094162
12  4.3892e-03     18  0.326512 1.15217 0.0092325
13  3.7139e-03     21  0.311318 1.16339 0.0092780
14  3.4542e-03     22  0.307604 1.16442 0.0092833
15  2.8049e-03     24  0.300696 1.15069 0.0092291
16  2.6837e-03     28  0.286723 1.14552 0.0092073
17  2.6621e-03     34  0.268959 1.14292 0.0091964
18  2.5841e-03     36  0.263635 1.13518 0.0091643
19  2.5712e-03     38  0.258467 1.13536 0.0091656
20  2.5452e-03     42  0.248156 1.13648 0.0091700
21  2.5322e-03     45  0.239196 1.13596 0.0091676
22  2.3894e-03     47  0.234132 1.12547 0.0091234
23  2.3374e-03     48  0.231742 1.12053 0.0091033
24  2.2595e-03     50  0.227067 1.11786 0.0090926
```

25	2.2335e-03	51	0.224808	1.10900	0.0090549
26	2.0777e-03	52	0.222574	1.10734	0.0090483
27	1.9998e-03	53	0.220497	1.11103	0.0090623
28	1.9478e-03	54	0.218497	1.09952	0.0090130
29	1.8440e-03	56	0.214601	1.09651	0.0090014
30	1.7271e-03	57	0.212757	1.08511	0.0089528
31	1.6622e-03	60	0.206706	1.08680	0.0089600
32	1.6362e-03	61	0.205044	1.08145	0.0089366
33	1.5583e-03	63	0.201771	1.07339	0.0089025
34	1.3765e-03	64	0.200213	1.05519	0.0088185
35	1.3116e-03	65	0.198836	1.05602	0.0088201
36	1.2726e-03	67	0.196213	1.05571	0.0088179
37	1.2466e-03	71	0.191123	1.05620	0.0088201
38	1.1817e-03	72	0.189876	1.05586	0.0088178
39	1.1557e-03	74	0.187513	1.05088	0.0087956
40	1.1427e-03	76	0.185202	1.04989	0.0087907
41	1.0908e-03	79	0.181773	1.04254	0.0087573
42	1.0778e-03	80	0.180683	1.03766	0.0087357
43	1.0648e-03	82	0.178527	1.03719	0.0087335
44	1.0562e-03	83	0.177462	1.03542	0.0087255
45	1.0518e-03	86	0.174294	1.03542	0.0087255
46	1.0389e-03	90	0.169515	1.03262	0.0087131
47	9.7392e-04	92	0.167437	1.02031	0.0086577
48	9.6094e-04	96	0.162347	1.01792	0.0086469
49	9.1765e-04	98	0.160425	1.01441	0.0086306
50	9.0900e-04	105	0.149465	1.00935	0.0086076
51	8.9601e-04	109	0.145829	1.00969	0.0086089
52	8.9168e-04	111	0.144037	1.00813	0.0086021
53	8.7004e-04	118	0.137492	1.00813	0.0086021
54	8.4407e-04	120	0.135752	1.00475	0.0085862

55	8.0511e-04	122	0.134064	1.00117	0.0085698
56	7.2720e-04	123	0.133259	0.98496	0.0084938
57	7.0123e-04	128	0.129623	0.97000	0.0084232
58	6.7525e-04	132	0.126818	0.96439	0.0083959
59	6.4928e-04	138	0.122740	0.96042	0.0083774
60	6.2331e-04	142	0.120143	0.95774	0.0083646
61	6.1465e-04	145	0.118273	0.95193	0.0083355
62	6.1033e-04	148	0.116429	0.95123	0.0083322
63	5.9734e-04	150	0.115209	0.94865	0.0083192
64	5.8435e-04	152	0.114014	0.94629	0.0083077
65	5.7137e-04	156	0.111677	0.94629	0.0083077
66	5.4540e-04	159	0.109963	0.93949	0.0082734
67	5.1943e-04	167	0.105599	0.93318	0.0082428
68	4.9346e-04	173	0.102171	0.92720	0.0082136
69	4.6748e-04	176	0.100691	0.91918	0.0081737
70	4.5017e-04	186	0.095834	0.91248	0.0081396
71	4.4151e-04	189	0.094484	0.90980	0.0081252
72	4.1554e-04	196	0.091393	0.90515	0.0081017
73	3.8957e-04	200	0.089497	0.89884	0.0080696
74	3.7658e-04	206	0.087108	0.89682	0.0080590
75	3.6360e-04	208	0.086355	0.88910	0.0080205
76	3.5061e-04	213	0.084537	0.88580	0.0080032
77	3.3763e-04	223	0.081031	0.87731	0.0079592
78	3.3113e-04	232	0.077992	0.87713	0.0079582
79	3.1166e-04	241	0.074901	0.87375	0.0079407
80	2.9867e-04	254	0.070850	0.86422	0.0078914
81	2.8568e-04	258	0.069655	0.86113	0.0078752
82	2.7270e-04	262	0.068512	0.85583	0.0078464
83	2.5971e-04	273	0.065058	0.85152	0.0078233
84	2.4673e-04	283	0.062357	0.84482	0.0077873

```
85  2.4240e-04    292  0.060124 0.84129 0.0077686
86  2.3374e-04    295  0.059396 0.84077 0.0077660
87  2.2508e-04    313  0.055189 0.83615 0.0077408
88  2.2076e-04    316  0.054514 0.83594 0.0077398
89  2.0777e-04    320  0.053631 0.83106 0.0077129
90  1.9478e-04    339  0.049553 0.82140 0.0076598
91  1.8180e-04    343  0.048774 0.82137 0.0076593
92  1.6881e-04    361  0.045476 0.81869 0.0076435
93  1.5583e-04    367  0.044463 0.81282 0.0076117
94  1.4284e-04    382  0.042100 0.80841 0.0075874
95  1.2986e-04    390  0.040957 0.80496 0.0075681
96  1.1687e-04    403  0.039243 0.80111 0.0075470
97  1.0389e-04    415  0.037840 0.79522 0.0075129
98  9.0900e-05    437  0.035555 0.79348 0.0075029
99  8.6571e-05    443  0.035009 0.79280 0.0074989
100 7.7914e-05    449  0.034490 0.79342 0.0075018
101 6.0600e-05    465  0.033243 0.79290 0.0074990
102 5.1943e-05    468  0.033062 0.78901 0.0074775
103 3.8957e-05    481  0.032386 0.78901 0.0074775
104 2.5971e-05    483  0.032308 0.78911 0.0074776
105 1.7314e-05    496  0.031971 0.78768 0.0074686
106 1.2986e-05    499  0.031919 0.78768 0.0074686
107 0.0000e+00    503  0.031867 0.78750 0.0074674
```

We usually choose cp of the smallest tree whose xerror is within 1 standard deviation of the least xerror.

Here least xerror is 0.78750 and it’s standard deviation is 0.0074674 so required cp of tree with xerror<0.7949674. Cp~0.0001

5. Develop decision tree models to predict default.

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

(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?

(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? Briefly describe how variable importance is obtained (the process used in decision trees).

Ans:

(a) We have split the 70 percent of the dataset into training and the remaining 30 percent for validation.Since Training is an important part of our Model, we keep the training data to 70% of our dataset. It is important to train our model on the maximum data available and proportionally try to segregate validation dataset so that the model has enough data to test on.

(b) The total of the decrease in error when a variable is split is used to compute Variable Importance. The variable Importance is then divided by the highest Variable Importance Values, yielding relative Importance values.

For the Decision Tree Model, first we dropped the Variables with 100% Missing Values, then split the Data into training and validation in the ratio of 70:30. Then developed Decision tree on Training Data which gave the Confusion Table shown below

#Confusion Table for training data (cp=0.0001)

Pred	Charged Off	Fully Paid
Charged Off	9562	1113
Fully Paid	64	59261

Accuracy on training Data: 0.9859857

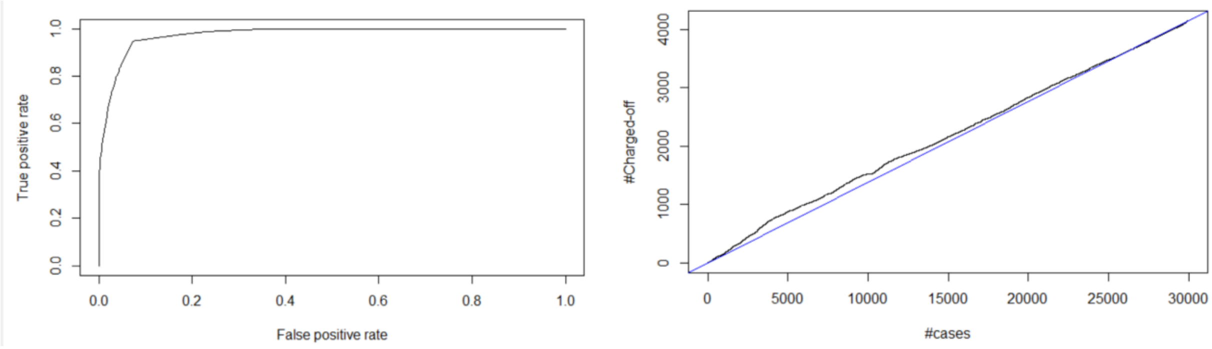
(c)Variable Importance is the ranking of each Variable based on the contribution of the Predictors that make up the model. This process helps us in segregating certain variables which contribute nothing and instead add up to the processing time. Top 6 Variable Importance:

emp_title	zip_code	earliest_cr_line	emp_length	title	addr_state
21760.090	6677.241	5834.066	2276.530	1933.014	1669.878

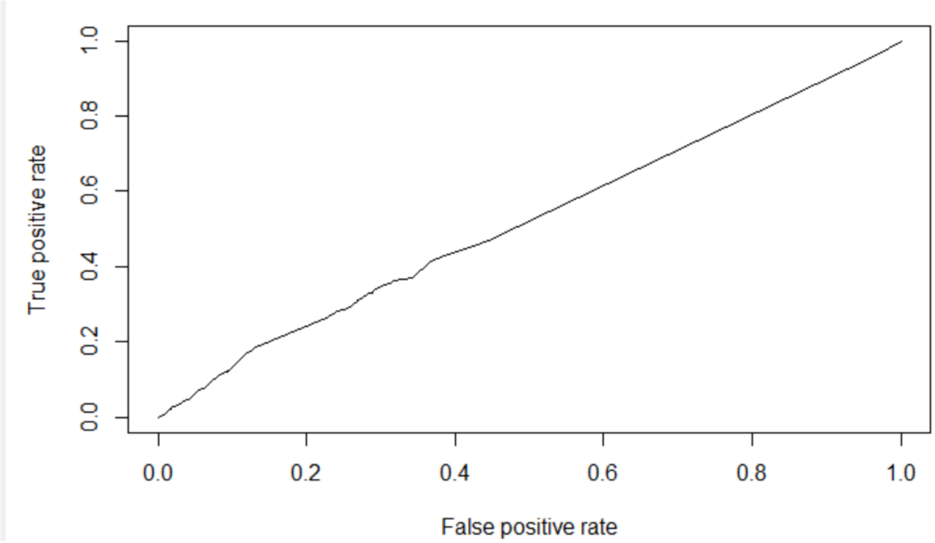
Confusion Table for testing data (cp=0.001)

Pred	Charged Off	Fully Paid
Charged Off	747	3272
Fully Paid	3412	22569

Accuracy on test data: 0.7772

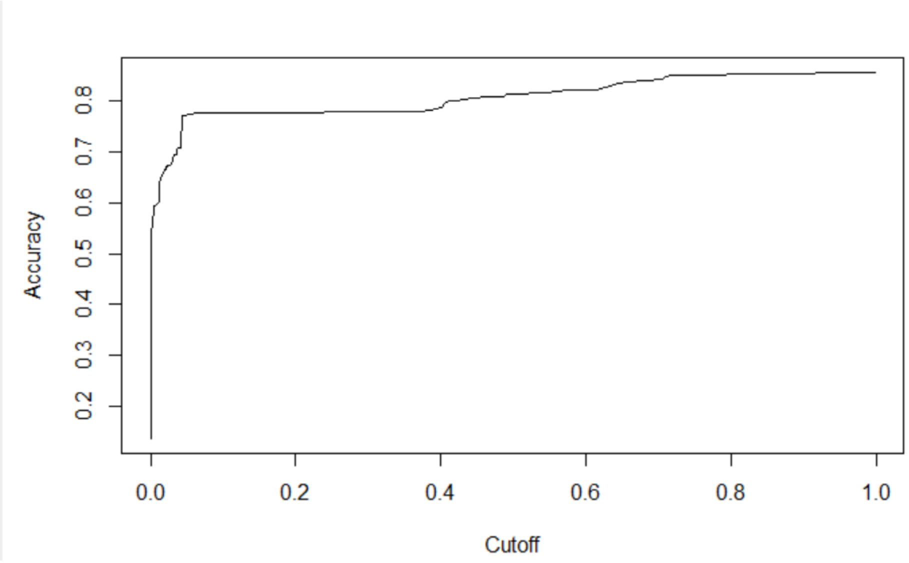


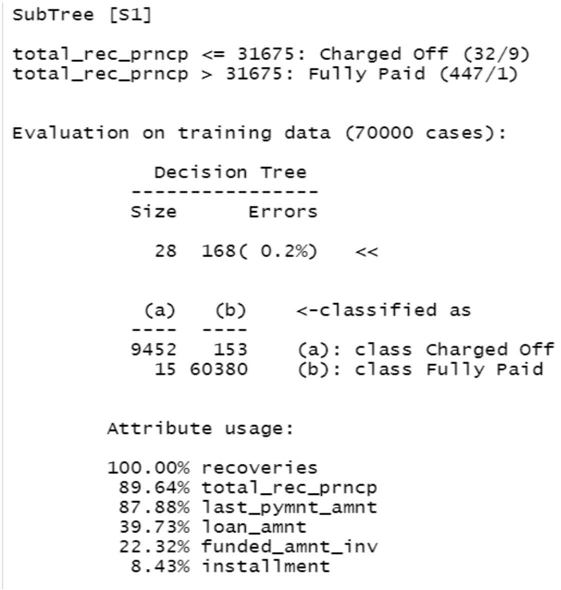
Lift Curve for training data and test Data



ROC for test Data, AUC= 0.5213665

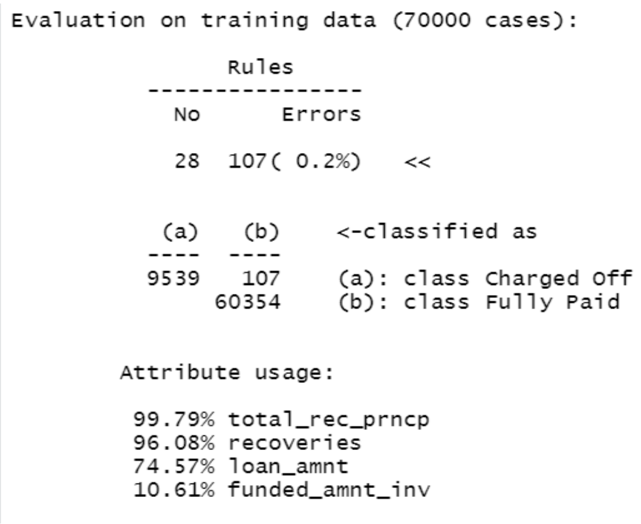
Accuracy Performance of Decision Tree Model:





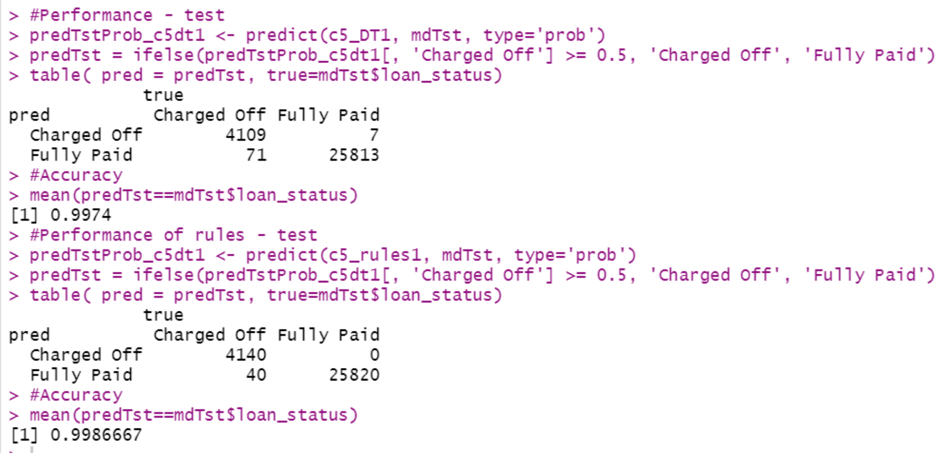
```
SubTree [S1]

total_rec_prncp <= 31675: Charged Off (32/9)
total_rec_prncp > 31675: Fully Paid (447/1)


Evaluation on training data (70000 cases):

            Decision Tree
          ----------------
          Size      Errors

            28  168( 0.2%)   <<


           (a)   (b)    <-classified as
          ----  ----
          9452   153    (a): class Charged Off
            15 60380    (b): class Fully Paid


        Attribute usage:

        100.00% recoveries
         89.64% total_rec_prncp
         87.88% last_pymnt_amnt
         39.73% loan_amnt
         22.32% funded_amnt_inv
          8.43% installment
```

Evaluation on Training Data using C50 (a)



```
Evaluation on training data (70000 cases):

                Rules
          ----------------
            No      Errors

            28   107( 0.2%)   <<


           (a)   (b)    <-classified as
          ----  ----
          9539   107    (a): class Charged Off
               60354    (b): class Fully Paid


        Attribute usage:

         99.79% total_rec_prncp
         96.08% recoveries
         74.57% loan_amnt
         10.61% funded_amnt_inv
```

Evaluation on Training Data (b)



```
> #Performance - test
> predTstProb_c5dt1 <- predict(c5_DT1, mdTst, type='prob')
> predTst = ifelse(predTstProb_c5dt1[, 'Charged Off'] >= 0.5, 'Charged Off', 'Fully Paid')
> table( pred = predTst, true=mdTst$loan_status)
             true
pred          Charged Off Fully Paid
  Charged Off        4109          7
  Fully Paid           71      25813
> #Accuracy
> mean(predTst==mdTst$loan_status)
[1] 0.9974
> #Performance of rules - test
> predTstProb_c5dt1 <- predict(c5_rules1, mdTst, type='prob')
> predTst = ifelse(predTstProb_c5dt1[, 'Charged Off'] >= 0.5, 'Charged Off', 'Fully Paid')
> table( pred = predTst, true=mdTst$loan_status)
             true
pred          Charged Off Fully Paid
  Charged Off        4140          0
  Fully Paid           40      25820
> #Accuracy
> mean(predTst==mdTst$loan_status)
[1] 0.9986667
```

Accuracy on Training and Testing using C50

It created decision tree of Level 1 .

CF was varied from 0.25 to 0.5

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

Ans:

The out of bag error of the forest was almost same for 60:40, 70:30, 80:20 partitions of training and testing data which are 0.1163, 0.1158, 0.1178 respectively. So, The Random Forest Model was build using Ranger with 70% training data keeping aside 30% of data for testing to avoid imbalances in the sampling. The target node size of the trees was taken as 10 with 200 trees in it. The accuracies of training and testing data are 96.8% and 86.1%.

The confusion table of the model with the testing data is:

	True		
Predicted		Charged Off	Fully Paid
	Charged Off	12	31
	Fully Paid	4147	25810

From the confusion Table:

Error= 0.139

Accuracy=0.86

Precision= 0.00406

Recall= 0.369

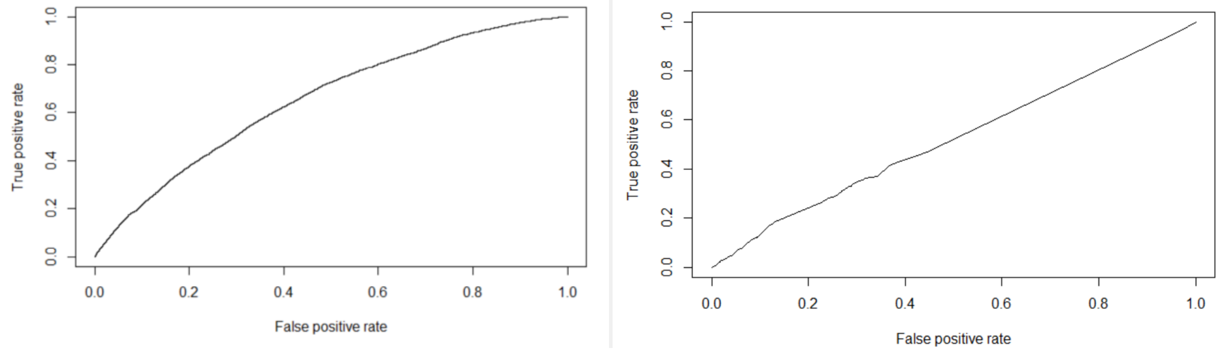
F-Score: 0.00401

The confusion table of the model with the test data is:

The variable importance of the two models (Decision Tree and random forest) are totally different. The following are the top 5 higher important variables.

Decision Tree Model	Random Forest model
Employee Title	Sub Grade
Zip Code	Total high credit/credit limit
Earliest Credit	Interest Rate
Employment Length	Annual Income of lender
Title of the loan	Total revolving high credit/credit limit

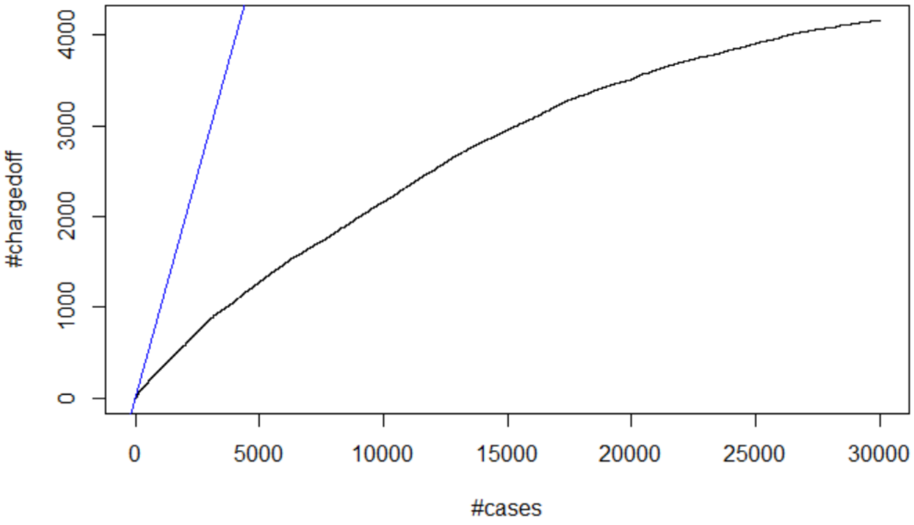
The ROC curve of the test data for Random Forest Model and Decision Tree Model Respectively



Accuracies of Decision Tree and Random Forest are 77.7% and 86.1% respectively on test data. The Area under the test data is 0.6586389. The higher the Area under the ROC curve, the better the model fits the data. From the above results, Random Forest model is better than Decision Tree model. The ROC of Decision Tree Model is almost along diagonal of the graph which means the model was poorly fitting the data.

From Accuracies, ROC curves, Lift curves, AUC we can say that Random Forest Model is better than Decision tree model. Although the accuracies of the models seem high; the ROC curves say otherwise. The ROC curve is not as convex as we imagined it to be from accuracy. It is because the data set is not uniform and the has 86% of 'Fully Paid' with only 14% of 'Charged Off'. Even though random sampling was done to avoid any bias, the proportion was exceptionally large

Lift curve of the test data:



7. **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 must be charged off ?**

Ans:

a.

loan_status	AvgInt	AvgActInt	avgTerm
Charged Off	13.85188	-11.962532	3.00000
Fully Paid	11.71333	8.020985	2.12855

From the above table,

Profit:

For fully paid loans, average term is 2.13 years and actual interest rate is 8.02 percent.

Suppose the investor invests these returns in CD, they earn a fixed return of 2%

ProfitVal= 8.02*2.13 + 2*0.87

= 18.82%

LossVal = -11.96*3

= -35.88%

Considering these values with following loan amounts:

pred_status	loan_status total	LoanAmnt	AvgLoanAmnt
Charged Off	Charged Off	86,600	7216.667
Charged Off	Fully Paid	441,525	14242.742
Fully Paid	Charged Off	50,635,500	12210.152
Fully Paid	Fully Paid	329,900,050	12781.869

Confusion Matrix for RF Test data:

	True		
Predicted		Charged Off	Fully Paid
	Charged Off	12	31
	Fully Paid	4147	25810

Average loan amount in each category: loan_amnt/total_loans

	True		
Predicted		Charged Off	Fully Paid
	Charged Off	7,216.67	14,242.74
	Fully Paid	12210.15	12781.87

Profit/Loss interest rate

	True		
Predicted		Charged Off	Fully Paid
	Charged Off	6%	6%
	Fully Paid	-35.88%	18.82%

Total Profit/Loss Matrix

	True		
Predicted		Charged Off	Fully Paid
	Charged Off	5,196	26.491.36
	Fully Paid	-18,168,007	62,087,245.5

Total Profit from the model = 43,950,925

b.

Total profit from Random Forest = 635552.5