2/16/2020

IDS-572 – Assignment-1

Lending Club – Case Study

Ritu Gangwal
UIN – 670646774

Vivek Kumar
UIN - 670460685

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)

We always have people looking for extra sum of money to grow their business or to upgrade their lifestyle or just for indulgence and we have people looking forward to invest their surplus income and get higher returns. While Banks fulfills this gap in a much-regulated way but there is very high interest rate for borrowers and very low returns for investors due to high operational costs, however Lending Club aims to bridge this gap with minimal operational cost with the help of technology and data-analytics

To be able to turn out this thought to practice, Lending Club created a business model and because of its well thought business model, they have been able to issue loans worth $50 Billion till date. The Business Model of Lending Club takes into consideration more than 100 parameters to evaluate credit worthiness of a borrower. Borrowers having higher credit worthiness are given loans at substantially low rate (which traditional lenders are not able to match) and borrowers with low credit worthiness (fico range should be more than 660) also get loans but at a bit higher rate (traditional lenders will not issue loans to such customers), so it becomes a win-win situation for borrowers with all kind of credit portfolio. On top of that, it also allows loans to be closed early without any penalty. Borrowers from all states except Ohio can apply for loans at Lending Club.

Lending Club is not only accommodative for borrowers but is flexible for investors as well, it gives them liberty to choose from the type of loans they wish to invest in. It then depends on their risk appetite that which kind of customers they want to lend to, higher return with higher risks and lower return with lower risk. For this lending club assigns grade (A to G) to borrowers basis their credit worthiness. Grade A being the highest grade with lowest interest rate and G being lowest grade with highest interest rate. Investors can also minimize their risk by diversifying their investments with amount as low as $25. Investors from 31 states can invest in any kind of loans at Lending Club.

The primary stakeholders for Lending Club are Lenders and Investors. While Lending Club was started with an idea of peer to peer lending but with the scale of business it has also attracted institutional investors and hedge funds etc. Lending Club makes money by charging a onetime origination fee in the range of 1.1%-5.0% of the loan amount from the borrower and a 1% of the loan amount as the Lending Club fee from the Investor. In addition to that, in case a borrower delays the payment for more than 15 days then Lending Club charges them 5% of the unpaid amount or $15 (whichever is higher).

Sources:

1. https://www.lendingclub.com/loans/personal-loans/rates-fees
2. https://www.lendingclub.com/investing/peer-to-peer
3. https://en.wikipedia.org/wiki/LendingClub

2. Data exploration

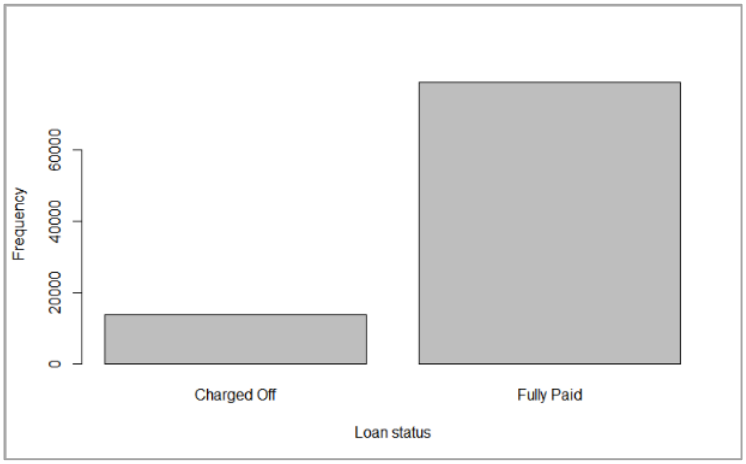
a. Some questions to consider

(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?

The proportion of "Charged Off" vs "Fully Paid" is 14.7 : 85.3. This means approximately 15% of the total no. of loans are defaulted and rest 85% are fully paid. Total no. of charged off loans = 13652 and fully paid = 78972. Also, the data has only two type of loan_status i.e. 'Charged Off' and 'Fully Paid'.

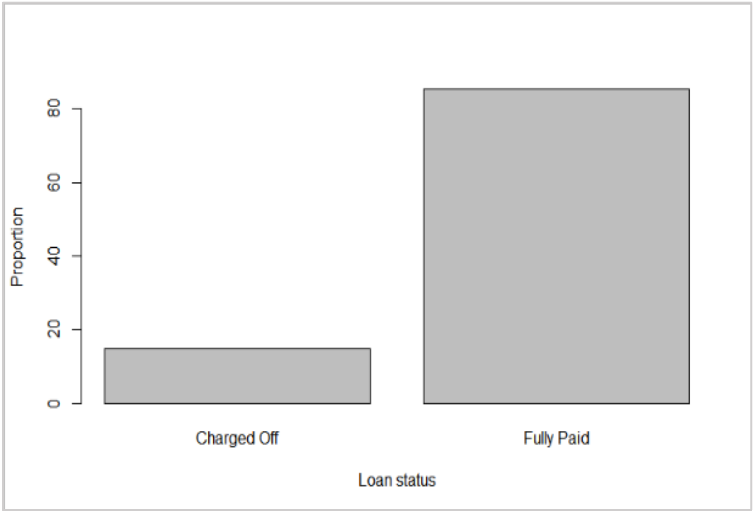
Below is the graph showing the proportion of each type of loans:

- By number:



Charged-Off	Fully Paid
13652	78972

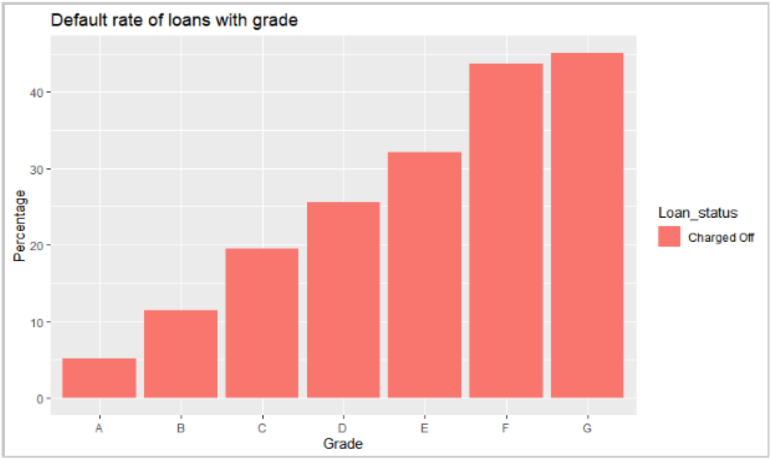
By proportion:

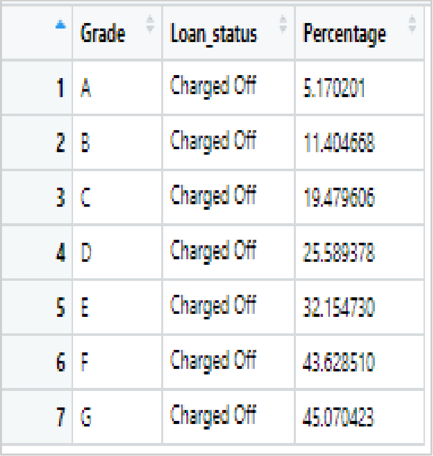


	Var1	Freq
1	Charged-Off	14.73916
2	Fully Paid	85.26084

The default rate and loan grade (risk-factor) has a correlation which follows a pattern similar to a linear relationship i.e. with increase in risk factor (Loan Grade A has lowest risk and G has highest risk), the rate of default increases. We have mentioned the relationship as similar to a linear relationship because Loan Grade is a categorical variable. The correlation between default rate and loan-grade is quite evident from the graph shown below, out of the total loans issued for loan-grade A (lowest-risk) just 5.17% of the loans have been defaulted whereas this ratio for loan-grade G is more than 45%.

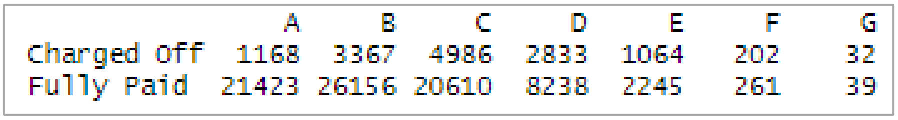
➢ <u>Percentage of Charged-Off Loans for each Loan-Grade:</u>



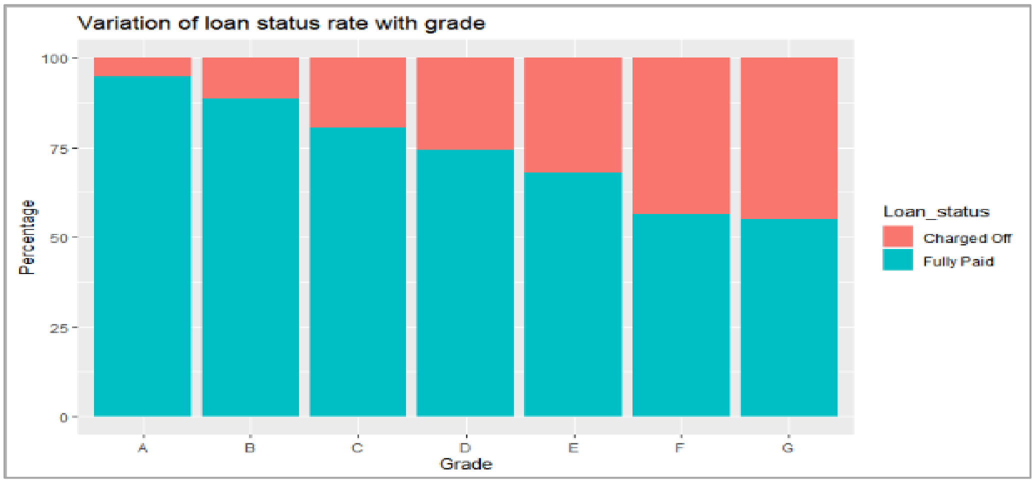


	Grade	Loan_status	Percentage
1	A	Charged Off	5.170201
2	B	Charged Off	11.404668
3	C	Charged Off	19.479606
4	D	Charged Off	25.589378
5	E	Charged Off	32.154730
6	F	Charged Off	43.628510
7	G	Charged Off	45.070423

➢ <u>Split of ‘Charged Off’ and ‘Fully Paid’ cases for each Loan-Grade</u>



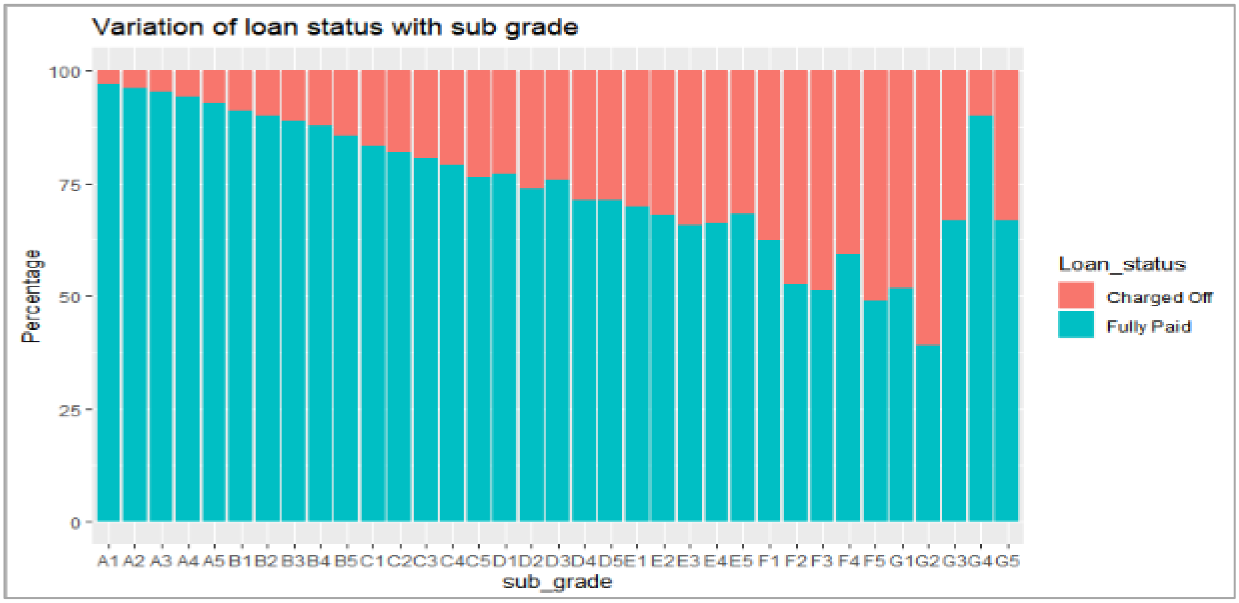
```
                  A     B     C    D    E   F  G
Charged Off   1168  3367  4986 2833 1064 202 32
Fully Paid   21423 26156 20610 8238 2245 261 39
```

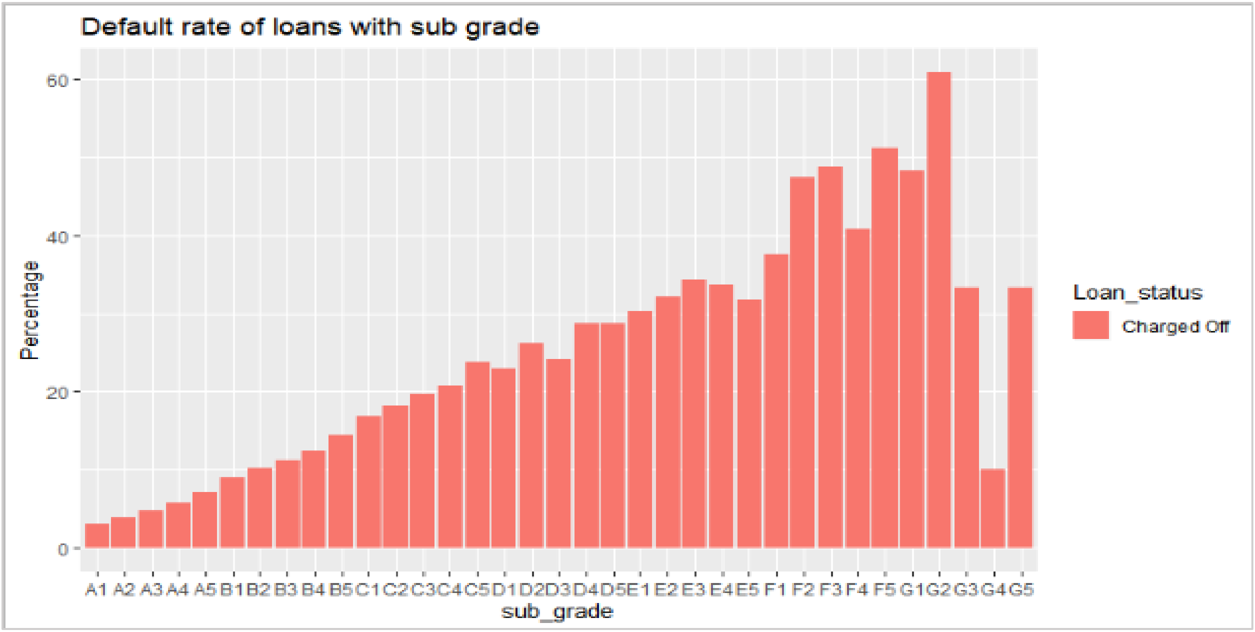


The correaltion of default rate and sub-grade is same as what is there between default rate and loan-grade. As the business model has segmented cases basis their risk factors into loan-grade, the same logic has been applied to segment it futher into subgrade. For example, risk-factor of A is lesser than B, C, D…

G, similarly risk factor of A1 is lesser than A2, A3…. A5 and eventually B1, B2…..G5. Hence the correlation between default rate and sub-grade is a finer version of correlation between default rate and loan-grade.

We expected the same because a sub-grade will exhibit behaviour of its loan grade and in addition to that will have finer classification amongst sub-grades of the parent loan grade. However the trend goes abrupt for sub-grades G3, G4 and G5 because of less number of people investing in those loans.

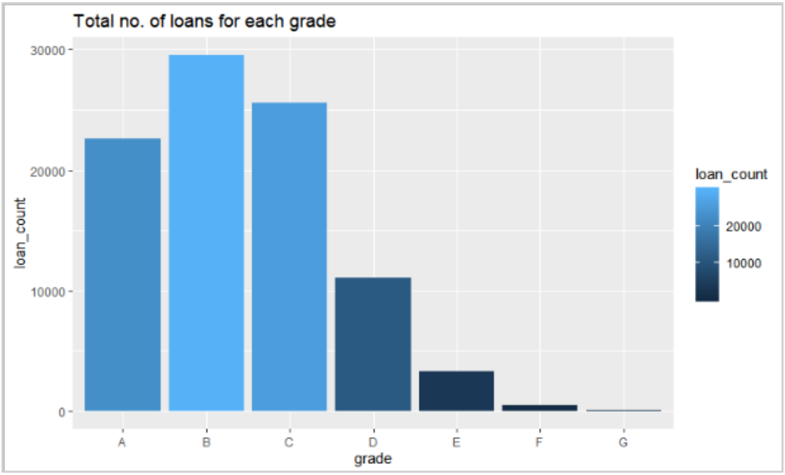


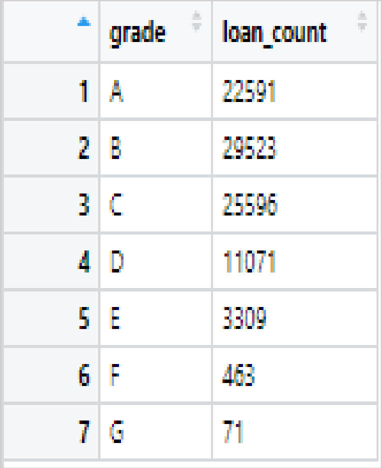


The pattern in these graphs shows robustness of the business model and algorithm used for bucketing loans into loan-grades and further into sub-grades.

(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?

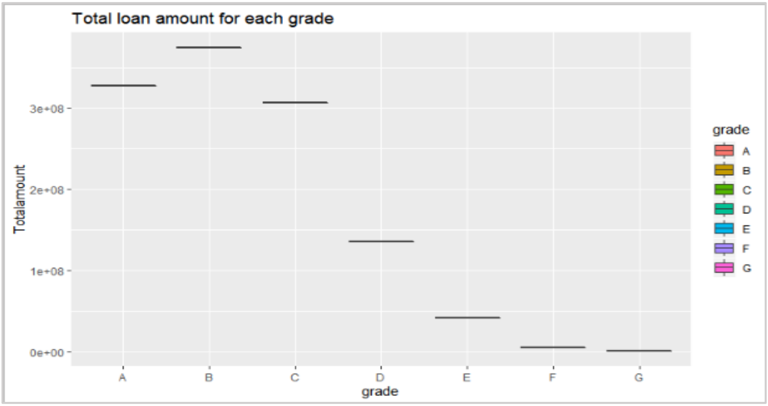
The graph below shows the count of loans for each loan grade. As loan grade is a measure of risk-factor and credit worthiness of the customer and as per the business model of Lending Club, loans with lower risk-factor (e.g. loan grade A) have lower interest rate and loans with higher risk factor (e.g. loan grade G) has higher interest rate. The graph below shows behavior of investors, as most of the investor tends to maximize the return keeping the risk minimal or as per their appetite, hence we have count of loans maximum for loan-grade B where the return is better than loan-grade A (lowest risk factor) and risk is lesser than loan grade C, D,…..G (highest risk factor). The count of loans decreases as we move from loan grade B towards loan grade G, the reason for the same is that it is difficult to find investors willing to invest money with such high risk-factor.

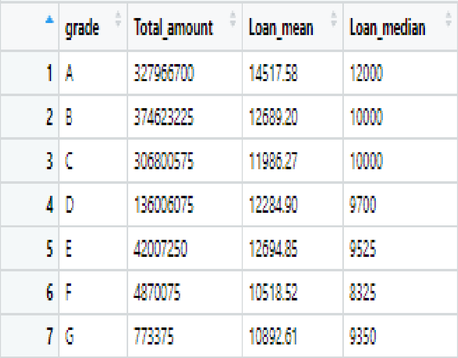




	grade	loan_count
1	A	22591
2	B	29523
3	C	25596
4	D	11071
5	E	3309
6	F	463
7	G	71

The total loan amount also varies linearly with grade similar to the loan count – loan grade correlation i.e. The total amount of loan for each loan grade decreases as we move from loan grade B to loan grade G beacuase the risk factor increases in that direction and it wouldn't be wise to invest a lot in high risk loans. However for loan grade A, the total loan amount is slightly lesser than loan grade B (3.91% of the total loan amount), it can be attributed to the fact that count of loans in loan grade B is higher than of loan grade A, reason being trade-off between risk and return.

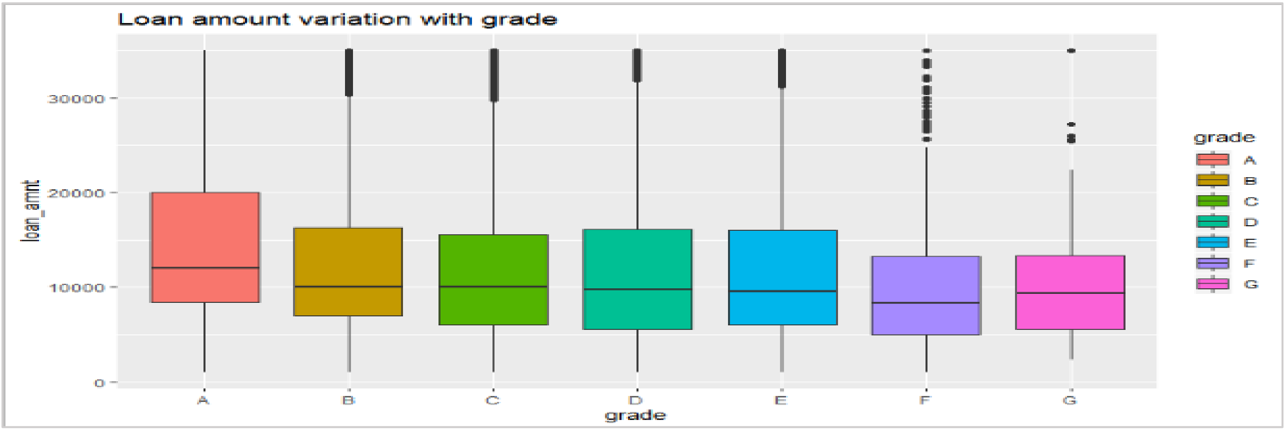




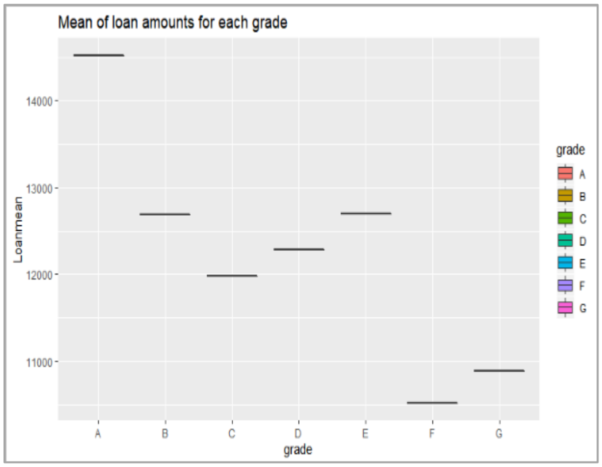
	grade	Total_amount	Loan_mean	Loan_median
1	A	327966700	14517.58	12000
2	B	374623225	12689.20	10000
3	C	306800575	11986.27	10000
4	D	136006075	12284.90	9700
5	E	42007250	12694.85	9525
6	F	4870075	10518.52	8325
7	G	773375	10892.61	9350

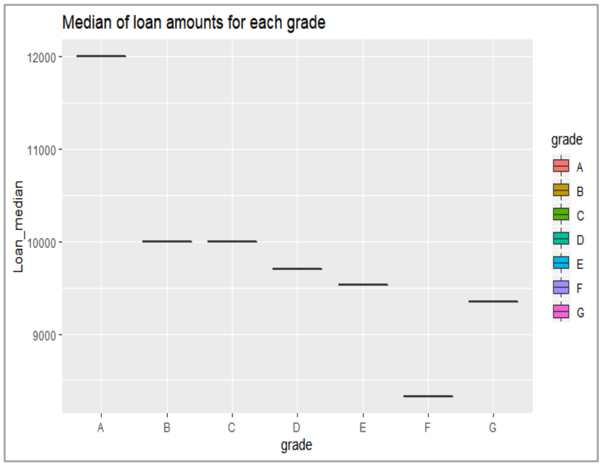
The graph and table above show the total amount invested in each grade of loans. The amount invested in Grade B is around $374 million on contrary to grade G that has only $773 thousand. Hence, we can conclude that there is a linear relationship between grade and loan amount. As the grade decreases, the invested loan amount also decreases.

However, for loans amounts in each category has a direct correlation with the risk factor i.e. the highest amount of loans are with loan grade A and the lowest with loan grade G, the reason being same as that of loan count and loan amount. This is quiet obvious as well because those who choose to invest in low risk factor loan-grade would prefer to put in maximum amount and those who choose to invest in high risk factor loan would prefer to keep the invested amount as less as possible.The box plot below of loan amount of each loan grade depicts the same.

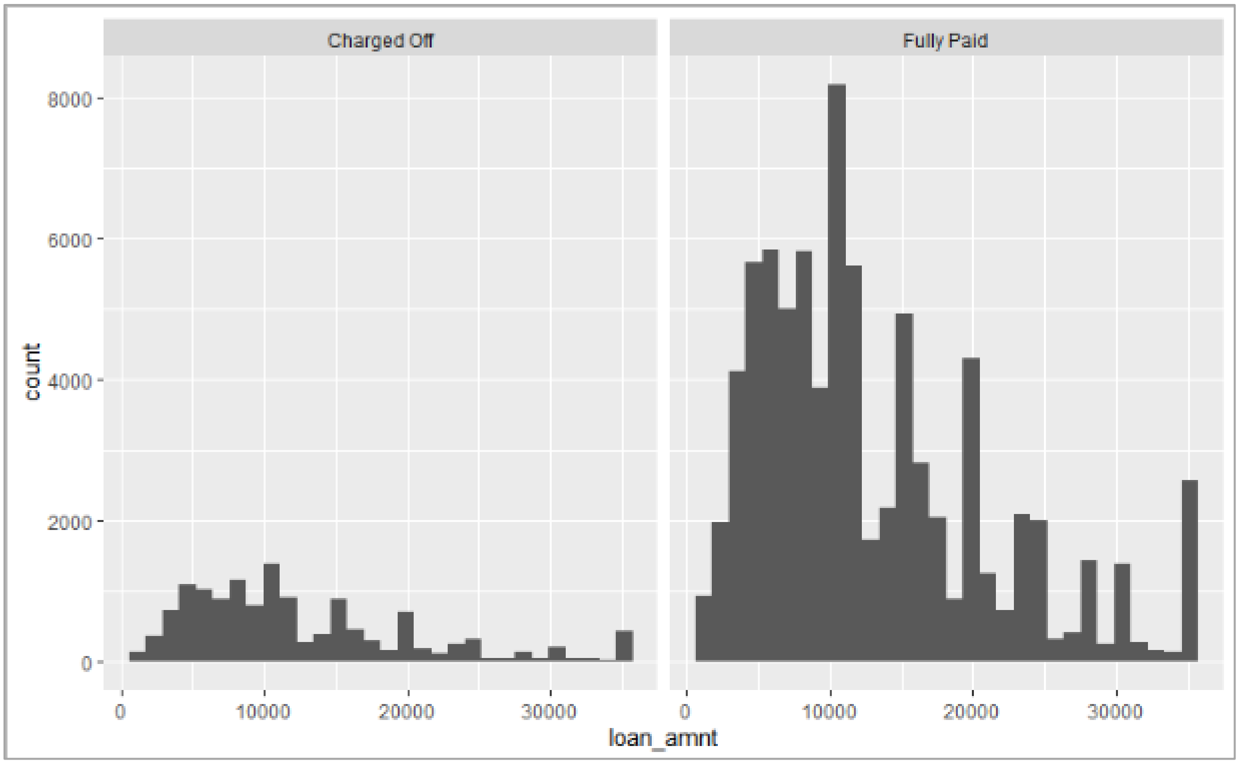


The mean of loan amounts of each grade has a similar pattern with a little deviation, the reason for it is the distribution of loan amount across each loan grade, some loan grades have more count of high amount loan and others have more count of low amount loan. There are no outliers impacting the mean as the min ($1000) and max($35000) of loan amount is same for all loan grade.

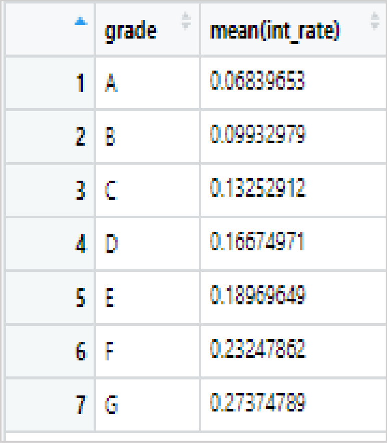




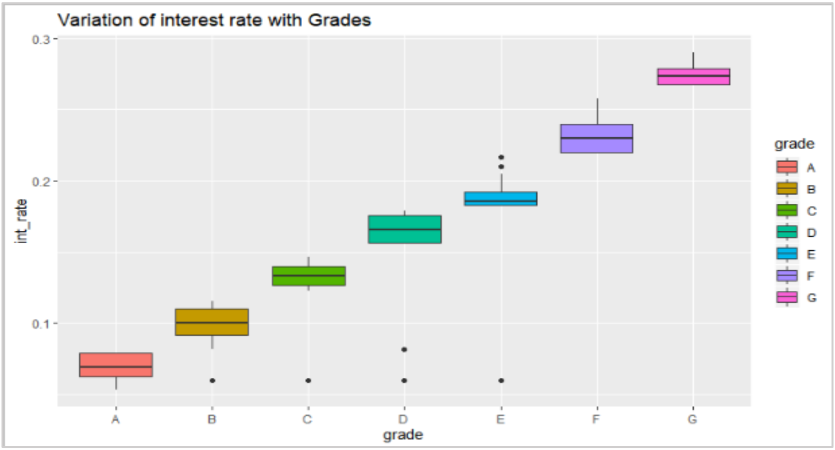
The graph below shows comparison of loan count vs loan amount split across loan status. Charged off loans have count of loans decreasing as the loan amount increases (depicts good health of the company), whereas for fully paid loans the count of loan has a spikes at few places showing the most of borrowes and investors prefers to choose loan amounts like $10000, $15000, $20000 and likewise.



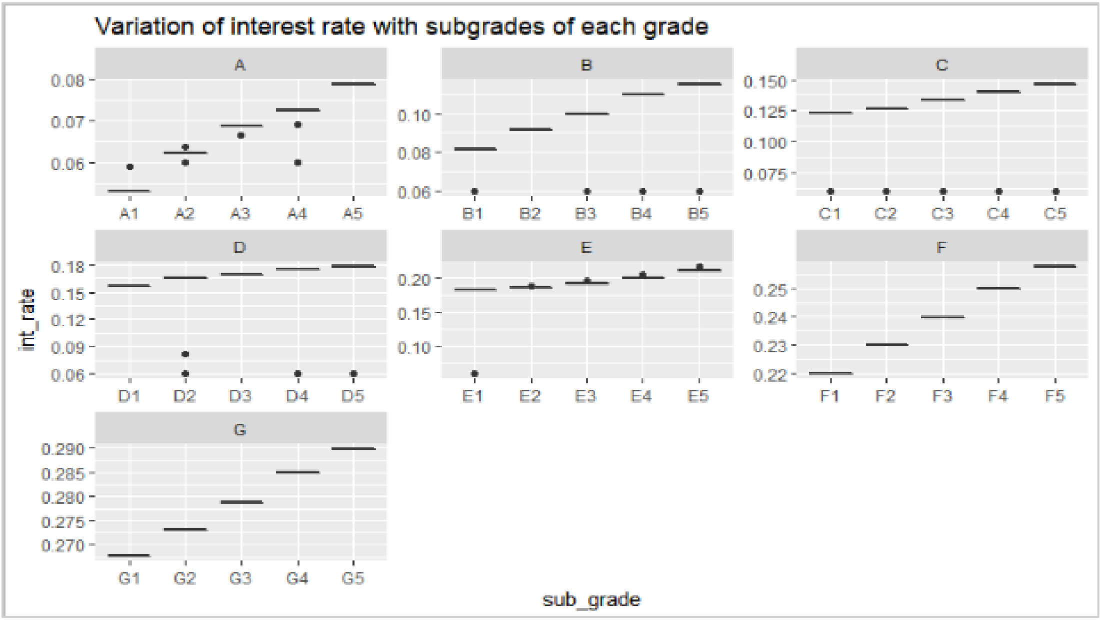
The interest rate varies with loan grade and sub-grade in a linear fashion. Basically, the purpose of assigning loan grades and sub grades is to classify loan basis their risk factor and thus determine the rate of interest for each case. The graph below shows the business model of lending club where loan grades and sub grades are used to assign interest rate to each case. Loans with low risk factors are assigned a lower interest rate and it increases as the risk factor increases.



	grade	mean(int_rate)
1	A	0.06839653
2	B	0.09932979
3	C	0.13252912
4	D	0.16674971
5	E	0.18969649
6	F	0.23247862
7	G	0.27374789



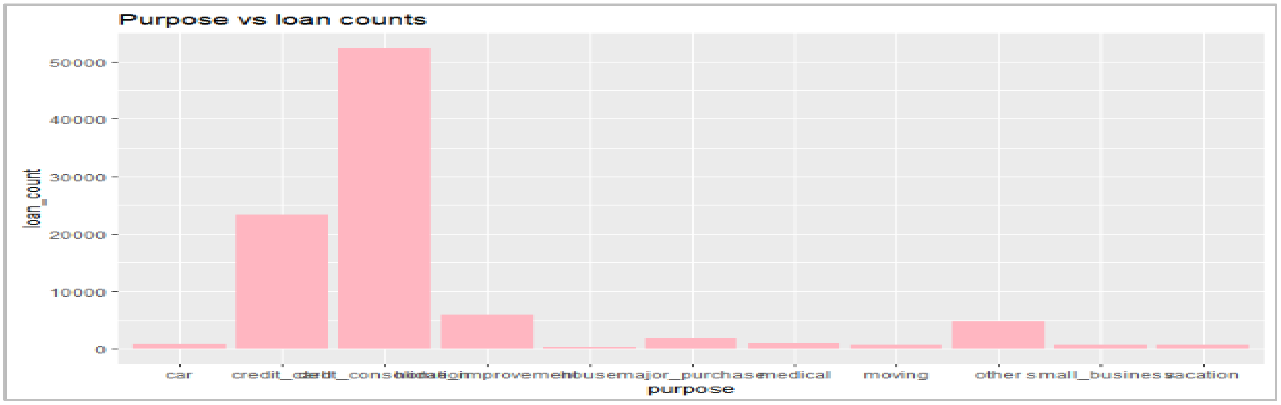
The correlation between sub-grade and interest rate is same as the correlation between loan grade and interest rate. This similarity explains the further segmentation of loan grade into sub grades so that it becomes easier to assign different and justified interest rate to two almost look alike cases. Just like for loan grade the interest rate increases with increase in risk factor i.e. interest rate of A is lesser than that of B, C,…G. Similarly interest rate for A1 is lesser than that of A2, A3.. A5 and eventually lesser than B1, B2… G5.



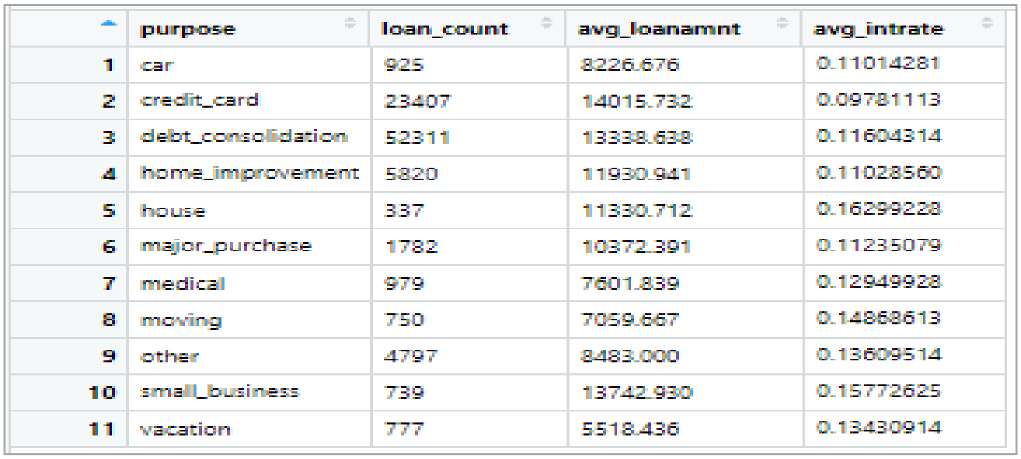
(iii).What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose? And within grade? Do defaults vary by purpose?

The below graph shows the total no. of loans borrowed for each purpose. For simplicity, we have merged "renewable energy" and "wedding" purposes with "others". We can clearly see that maximum people are borrowing money for **debt consolidation** and secondly for **credit cards** with house loans being the minimum.

- Count of loans plotted against purpose for which it is taken

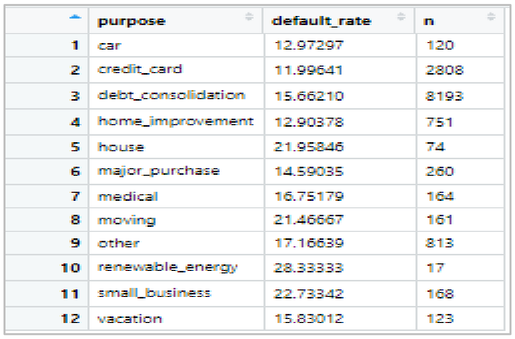


The below table shows the **count** of loan and it's **mean** for each purpose. Mean of loans also follows the pattern same as loan count i.e. among all the purposes mean of loan amount is maximum for credit_card and debt_consolidation.

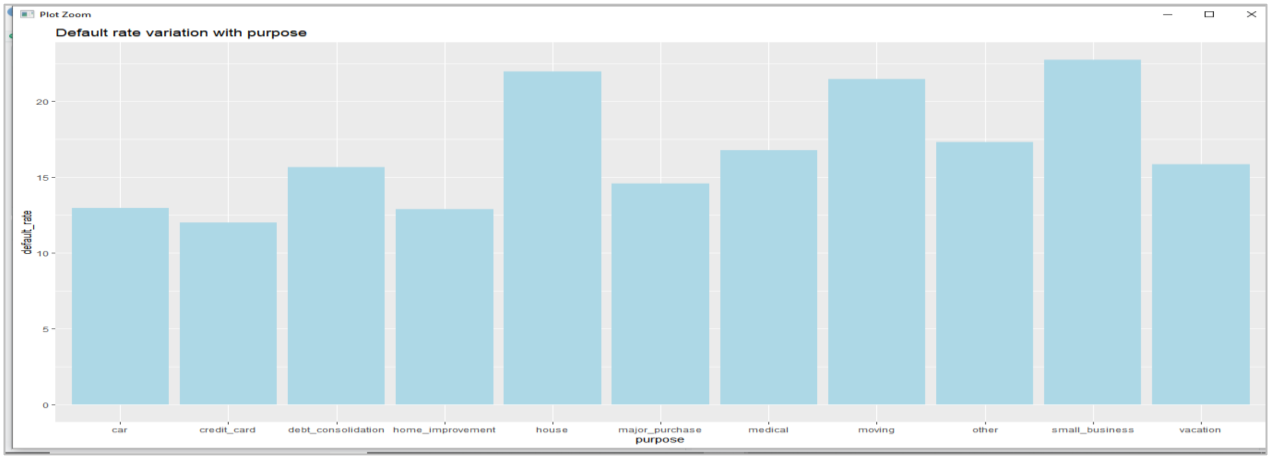


	purpose	loan_count	avg_loanamnt	avg_intrate
1	car	925	8226.676	0.11014281
2	credit_card	23407	14015.732	0.09781113
3	debt_consolidation	52311	13338.638	0.11604314
4	home_improvement	5820	11930.941	0.11028560
5	house	337	11330.712	0.16299228
6	major_purchase	1782	10372.391	0.11235079
7	medical	979	7601.839	0.12949928
8	moving	750	7059.667	0.14868613
9	other	4797	8483.000	0.13609514
10	small_business	739	13742.930	0.15772625
11	vacation	777	5518.436	0.13430914

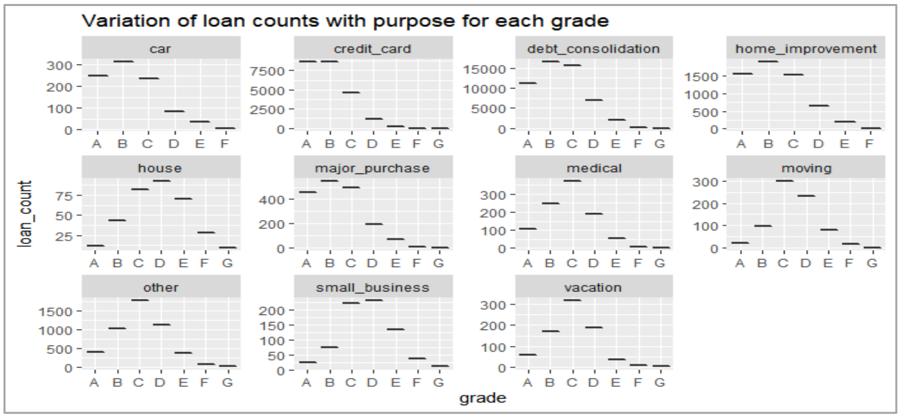
Just like most of the loans (~ 80%) are issued for credit_card and debt_consolidation, similar is the pattern when we plot the defaulted against purpose. This pattern is kind of obvious as the purpose for which maximum number of loans are issued will have defaults in that very proportion, had it been any pattern other than this, then Lending Club would have assigned interest rate basis purpose and not on risk factor. The table and graph below shows variation of default rate for that very category. The default rate per category seems to similar for all purposes but the total count of defaulted loans will shoot up for credit_card and debt_consolidation given the high total loan count for these purposes.

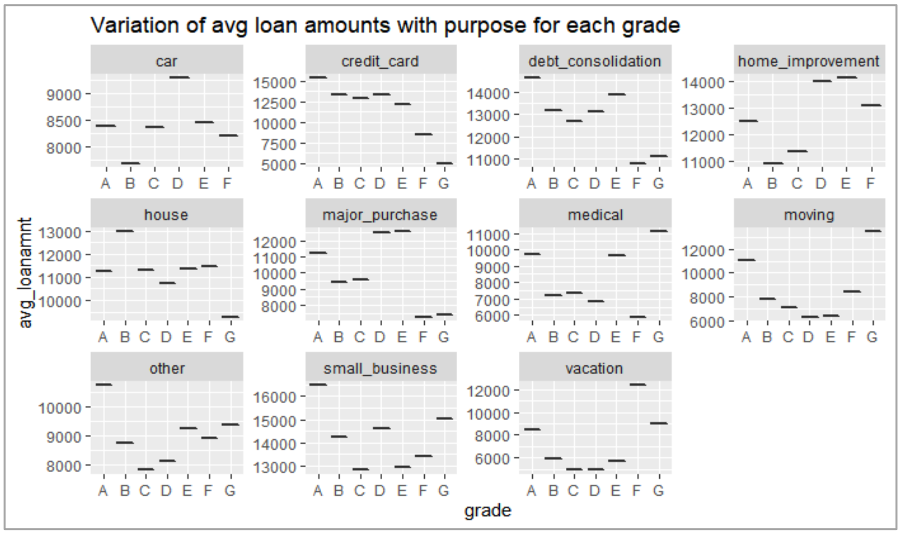


	purpose	default_rate	n
1	car	12.97297	120
2	credit_card	11.99641	2808
3	debt_consolidation	15.66210	8193
4	home_improvement	12.90378	751
5	house	21.95846	74
6	major_purchase	14.59035	260
7	medical	16.75179	164
8	moving	21.46667	161
9	other	17.16639	813
10	renewable_energy	28.33333	17
11	small_business	22.73342	168
12	vacation	15.83012	123

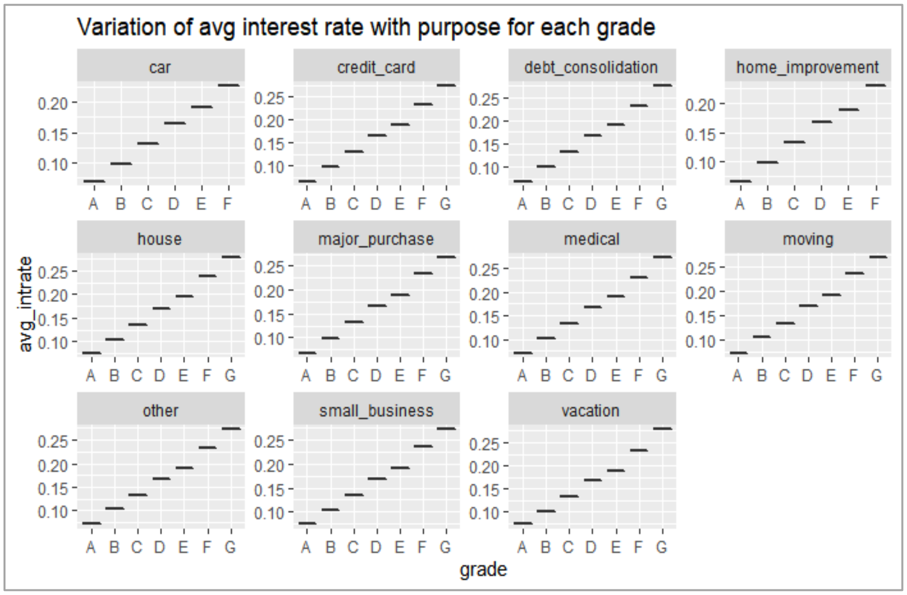


The variation of purpose for each loan grade has a similar pattern like for overall loans. The reason for all these similarities is credit_card and debt_consolidation having more than 80% share of the total loans as per its classification on purpose.


Variation of loan counts with purpose for each grade
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
small_business
vacation
loan_count
grade
A B C D E F G


Variation of avg loan amounts with purpose for each grade
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
small_business
vacation
avg_loanamnt
grade
A B C D E F G

-


Variation of avg interest rate with purpose for each grade
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
small_business
vacation
avg_intrate
grade
0.10
0.15
0.20
0.25
A B C D E F G

(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? 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?

Solution- The difference of total payment and funded amount will give us the profit for each loan.

This difference when divided by funded amount will give us the return of that loan.

As term for all loans is 36 months, so to be able to calculate Annual Return of each loan, we would need to pro-rate it to 12 months i.e. multiply the total return by 12/36.

Multiplication of annual return with 100 will yield the Annual Return Percentage.

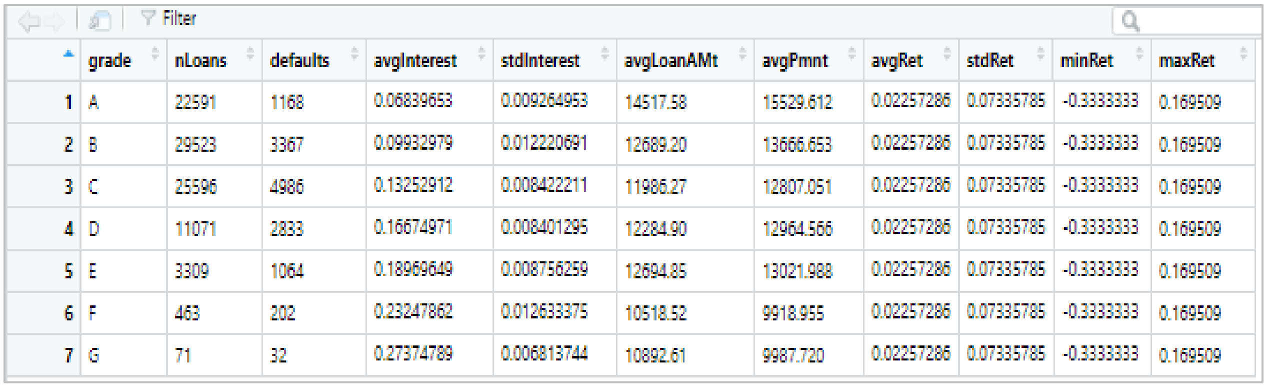
Annual Return = (Funded amount – total payment)/funded amount *12/36 * 100

Using the above approach, we can calculate the annual return percentage for all the loans. Mean of all these values will give us Average Annual Return Percentage, which happens to be **2.26%** for given data.

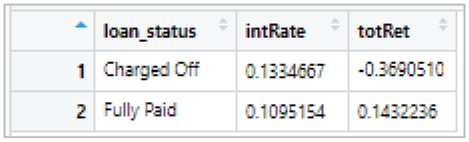
Similarly, Average Interest Rate can be calculated by taking mean of the of the interest rate. To convert the average interest rate to Average Interest Rate Percentage, we need to multiply the earlier by 100.

The Average Interest Rate Percentage can be calculated by multiplying the average interest rate by 100, which for this case turns out to be **11.3%**

The tables below shows average interest rate and average annual return percentage for loans on the basis of grade and on basis of loan_status. The negative average return rate for Charged-off loans is expected as for most of the cases, total payment would not have been completed.



	grade	nLoans	defaults	avgInterest	stdInterest	avgLoanAMt	avgPmnt	avgRet	stdRet	minRet	maxRet
1	A	22591	1168	0.06839653	0.009264953	14517.58	15529.612	0.02257286	0.07335785	-0.3333333	0.169509
2	B	29523	3367	0.09932979	0.012220691	12689.20	13666.653	0.02257286	0.07335785	-0.3333333	0.169509
3	C	25596	4986	0.13252912	0.008422211	11986.27	12807.051	0.02257286	0.07335785	-0.3333333	0.169509
4	D	11071	2833	0.16674971	0.008401295	12284.90	12964.566	0.02257286	0.07335785	-0.3333333	0.169509
5	E	3309	1064	0.18969649	0.008756259	12694.85	13021.988	0.02257286	0.07335785	-0.3333333	0.169509
6	F	463	202	0.23247862	0.012633375	10518.52	9918.955	0.02257286	0.07335785	-0.3333333	0.169509
7	G	71	32	0.27374789	0.006813744	10892.61	9987.720	0.02257286	0.07335785	-0.3333333	0.169509



	loan_status	intRate	totRet
1	Charged Off	0.1334667	-0.3690510
2	Fully Paid	0.1095154	0.1432236

If we want to invest in a loan basis this data exploration then we would prefer loan grade A, as it has lowest risk factor and provides more secure return.

While installments of all loans is 36 months, but as per the given data some loans were closed before the expected period. For this, we need to calculate actual-term of the loan, which can be calculated by the difference of last payment date and loan issue date.

This in-turn will also impact the annual return and annualized percentage return i.e.

Actual Annual Return = ((Total Payment – Funded Amount)/Funded Amount)/Actual Term

The above formula will give the **Actual Annual Return Percentage** when multiplied by 100, which can be used to calculate the average of actual annual return percentage, which happens to be **4.57%** for the given data.

Note – Introduction of Actual_Term for this data increases the total number of variables to 151.

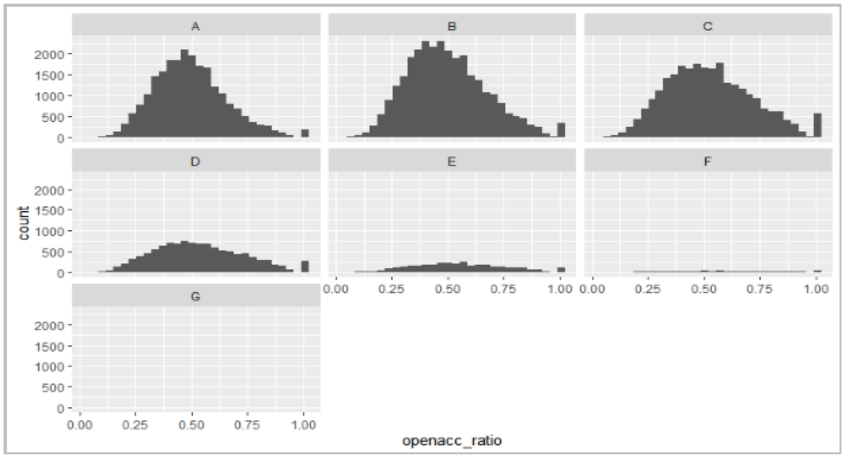
(v) Generate some new derived attributes which you think may be useful for predicting default., and explain what these are?

Some columns in the given data are dependent on each other, these variables can be clubbed to form a derived variable and make the decision making simpler. We were able to figure out introduce few derived variables which will replace the dependent variables in the data set and simplify the decision-making process.

a. Proportion of Satisfactory bankcard accounts – Number of Satisfactory Bankcard Account is a subset of Number of Bankcard Account, so these two can be clubbed together to form a derived variable i.e. PropSatisBankcardAccts = 0, if num_bc_tl=0, else (num_bc_sats /num_bc_tl)

b. Ratio of total open accounts to total accounts – Number of Total Open Account is a subset of Number of Total Accounts, so these two can be merged to form a derived variable i.e. openacc_ratio = open_acc/total_acc (In the given data total_acc != 0)

   The graph below shows comparison of openacc_ratio with all loan grades, this ratio has a similar pattern for all loan grades but the ratio decreases as the risk-factor increases, we can draw an inference from this observation that borrowers with higher openacc_ratio tend to have a better risk profile i.e. lesser risk factor.

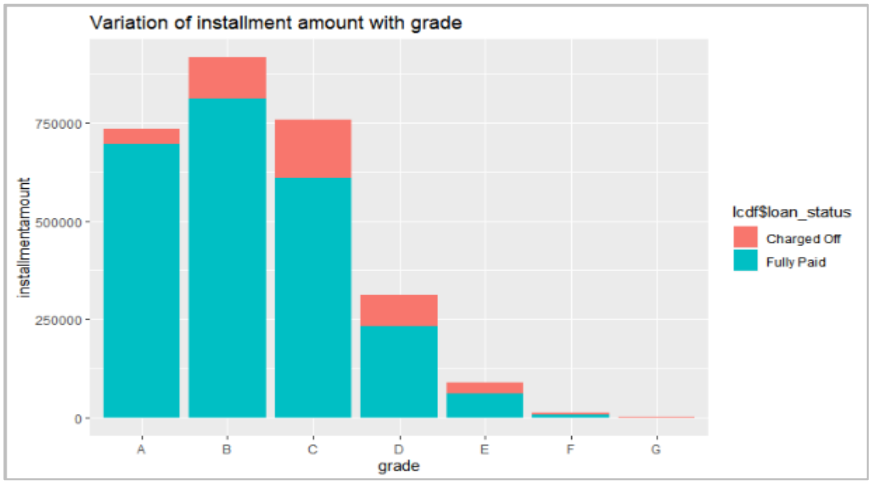


c. Ratio of Funded Amount Invested to Loan Amount – Funded Amount Invested can either be equal to or lesser than Loan Amount and this relationship can be used to form another derived variable i.e. the percentage amount that an investor has committed to the loan borrower percent_committed = funded_amnt_inv/loan_amnt

d. Ratio of Total Current Balance of All Accounts to Number of Open Credit Lines in the borrower's credit file – The former parameter is a subset of the later and this relationship can be used to form another derived product curbal_open_acc = tot_cur_bal/open_acc

e. Ratio of funded amount to installment – The ratio of funded amount with installment gives the duration in which the total amount will get paid
i.e. installmentamount = funded_amnt/installment

The graph below shows variation of derived variable of funded amount & installment with loan_grade, the pattern is similar to that of sum of funded amount, thus it leaves us with one less variable and easier decision making.



2(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?

There are missing values in the given data. Once we removed the variables with all records as 'NA'(49 variables), we were left with **108** (157-49) variables out of total 157 (150 + ActualTerm + 6 derived) variables. After removing variables with more than 60% of missing data(list of those missing columns is mentioned below), we are left with 98 variables.

Variables with all 'NA' **Count of variables removed = 49**	id, member_id, url, desc, next_pymnt_d, annual_inc_joint, dti_joint, verification_status_joint, open_acc_6m, open_act_il, open_il_12m, open_il_24m, mths_since_rcnt_il, total_bal_il, il_util, open_rv_12m, open_rv_24m, max_bal_bc, all_util, inq_fi, total_cu_tl, inq_last_12m, revol_bal_joint, sec_app_fico_range_low, sec_app_fico_range_high, sec_app_earliest_cr_line, sec_app_inq_last_6mths, sec_app_mort_acc, sec_app_open_acc, sec_app_revol_util, sec_app_open_act_il, sec_app_num_rev_accts, sec_app_chargeoff_within_12_mths, sec_app_collections_12_mths_ex_med, sec_app_mths_since_last_major_derog, hardship_type, hardship_reason, hardship_status, deferral_term, hardship_amount, hardship_start_date, hardship_end_date, payment_plan_start_date, hardship_length, hardship_dpd, hardship_loan_status, orig_projected_additional_accrued_interest, hardship_payoff_balance_amount, hardship_last_payment_amount

Number of variables left = 151+6-49=108

Columns with more than 60% 'NA'	mths_since_last_record, mths_since_last_major_derog, mths_since_recent_bc_dlq, mths_since_recent_revol_delinq, debt_settlement_flag_date, settlement_status, settlement_date, settlement_amount, settlement_percentage, settlement_term

Number of variables left = 108-10=98

The below list of 12 variables are the ones which has missing values. We have replaced the missing variables for these columns either with mean, median or zero basis nature of the attributes.

Missing values replaced with mean, median or zero for these variables	emp_title, mths_since_last_delinq, revol_util, last_pymnt_d, bc_open_to_buy, bc_util, mo_sin_old_il_acct, mths_since_recent_bc, mths_since_recent_inq, num_tl_120dpd_2m, percent_bc_gt_75, actualTerm

The table below indicates logic used for replacement of NA for the above variables:

S.No.	Column Name	Logic Used for replacement of Missing Values
1	mths_since_last_delinq	This column has 48% missing values which is because of no delinquency, so we can replace it by max value (170) or higher, we will experiment this replacement in lcx dataset
2	revol_util	Replaced by median
3	bc_open_to_buy	Replaced by median
4	bc_util	For this column, mean of data is 63 whereas third and max quartile is 84.5 and 202 respectively which indicates presence of outlier, hence replaced by median
5	mo_sin_old_il_acct	This column is for months since oldest bank account was opened, hence makes sense to replace with zero

6	mths_since_recent_bc	This column is for months since recent bankcard account was opened, replacing missing values with zero for it as well
7	#percent_bc_gt_75	As percentage of missing values for this column is just 1%, so we can replace the missing values by median for it
8	num_tl_120dpd_2m	Replacing with zero
9	mths_since_recent_inq	This column gives us months since recent inquiry was done, missing values indicate that this person may not have applied for a loan before, hence replacing NA with Zero
10	actualTerm	Replacing by Zero
11	emp_title & last_payment_d	Removed these two columns

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.

Variables removed to avoid data leakage Count of variables removed = **28**	funded_amnt_inv, term, pymnt_plan, title, zip_code, addr_state, open_acc, initial_list_status, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_amnt, last_credit_pull_d, last_fico_range_high, last_fico_range_low, collections_12_mths_ex_med, policy_code, application_type, hardship_flag, debt_settlement_flag, no.ofinstallments, actualTerm

Variables removed due to more than 80% correlation Count of variables removed = **19**	open_acc, num_sats, num_op_rev_tl, num_rev_accts, total_bc_limit, num_bc_sats, num_actv_rev_tl, tot_hi_cred_lim, tot_cur_bal, total_rev_hi_lim, total_bal_ex_mort, loan_amnt, funded_amnt, avg_cur_bal, fico_range_low, mo_sin_old_rev_tl_op, revol_util, bc_util, num_tl_30dpd

4.Develop decision tree models to predict default.

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

Firstly, we consider 50:50 split for our data. Here we have taken two seed values to determine the consistency of our results. We have used rpart method to develop the below comparisons:

Seed Value	Accuracy (Training)	Accuracy (Test)	Difference
2204	90.36%	85.15%	5.21%
70	94.44%	80.07%	14.37%

Also, we have performed the same seed values 70:30 split.

Seed Value	Accuracy (Training)	Accuracy (Test)	Difference
2204	85.21%	85.37%	-0.16%
70	85.31%	85.13%	0.18%

The 50:50 split model is relatively unstable, because when we changed the seed value in training data, there is a change in the differences of accuracies. Whereas in 70:30 split model is more stable with mode average difference of 0.15%.

Also, it is always recommended to have greater values in training set in order to capture all the information and the aspects of the variables. By taking 70% as our training set, there is a high probability that we have captured almost all detailed information in our training set which is useful in making a good model.

On the other hand, if we take 50:50 split, we won't have much confidence to capture all the desired observations. Also, with smaller training set, model will simply replicate the training examples rather than generalizing the results. This might result in capturing the noise of training set and resulting in over fitting.

As we can see in the below above by taking 50:50 split, we are getting accuracy of training data as 94.44% which is remarkably high and is a case of overfitting.

Hence considering the concerns that the lesser training data increases the parameter estimate variance and with lesser testing data, higher variance is expected in our performance statistics, we arrive at the conclusion to have the 70:30 data split.

(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?

After arriving at the conclusion that we will go forward with 70:30 split, we will be running our decision tree models by both rpart and C50.

RPART:

Firstly, we train our decision tree models using rpart on the below defined parameters:

Input Parameters

1) Minsplit - It is the minimum number of observations that must exist in a node for a split to be attempted
2) CP is the "Complexity Parameter" of the tree.
3) Max Depth – The number of nodes from the root down to the furthest leaf node
4) Split - Information gain or Gini
5) Threshold – Probability of predicting each output class i.e. 'Charged-Off' or 'Fully Paid' for this data-set
6) Oversampling – Increasing the sampling frequency by repeating some values a few times. This is done as the actual data is highly imbalanced.
7) Pruning – Reduction in the size of decision tree by removing sections of the tree which provides lesser power to classify instances

Parameters to be considered during Performance Evaluation:

In order to estimate the best model, we have taken into consideration the values mentioned below:

1. Accuracy – It measures the overall accuracy of model classification
2. Sensitivity or Recall **–** It is a measure of true positive rate i.e. For "yes" how often it predicts "Yes".
3. Specificity **–** It is measure of true negative rate i.e. For "No", how often it predicts "No"
4. ROC – Receiver Operating Characteristics Curve
5. AUC – Area Under Curve
6. LIFT Curve – The lift curve is plot of lift versus the portion of observations

We evaluated the performance of the models using the confusion matrix, ROC curve and AUC values. The confusion matrix gives us the number of correct and incorrect predictions by the classification model in comparison to the actual target values in the data, which will help us understand the number of false positives and true positives. By this we can observe the accuracy of the model, which will assist us in comparing the performance of various models as well.

Cases	Parameters	Training data				Test data			
		Accuracy	Specificity	Sensitivity	AUC	Accuracy	Specificity	Sensitivity	AUC
rpart model 1	Method = Information gain min split =10 cp =0.0001 max depth =15	88.33%	98.64%	29.03%	0.82143	82.41%	94.73%	9.97%	0.626
rpart model 2	Method = Gini min split = 10 cp = 0.0001 max depth = 15	89.56%	98.51%	38.06%	0.81847	81.62%	93.61%	11.13%	0.612
rpart model 3	Method = Information gain min split = 30 cp = 0.0004 max depth = 10	85.25%	99.90%	1.05%	0.66082	85.41%	99.83%	5.94%	0.658
rpart model 4	Method = Information gain min split=10 cp = 0.0001 max depth = 15 Threshold = 0.25	84.38%	90.27%	50.53%	0.704	77.57%	86.08%	27.87%	0.569
rpart model 5	Method = Information gain min split=10 cp = 0.0001 max depth = 15 Threshold = 0.25 Oversampling done	75.50%	61.94%	92.22%	0.771	56.72%	56.57%	57.65%	0.571
rpart model 6	Method = Information gain min split=10 cp = 0.0001 max depth = 15 Threshold = 0.25 Oversampling done Pruned at cp = 0.0001027	73.96%	60.56%	90.48%	0.755	57.13%	56.20%	62.60%	0.601
rpart model 7	Method = Information gain min split=10 cp = 0.0003 max depth = 10 Threshold = 0.4 Oversampling done Pruned at min cp	67.40%	63.39%	72.34%	0.679	62.26%	62.16%	62.82%	0.667

EXPLAINATION:

Comparison between models	Parameters compared	Reason
1 and 2	We have tried running model 1 with "information" split and model 2 with "gini" split, keeping cp, minsplit and max depth as same.	We have observed that accuracy, specificity and AUC are better with "information". Hence, we select model 1.
1 and 3	Here, we have tried increasing the cp and minsplit values and decreasing the maxdepth value	We see both the models work approximately same, but results in a very low sensitivity.
4	We have included the threshold value of 0.25	We do observe that the sensitivity of the test data increases
5	Along with the threshold, we have done oversampling of the training data. This is because the actual data is highly imbalanced/ skewed with 85% of fully paid and only 15% of charged off loans	We observe there is drastic increase in sensitivity, but accuracy and AUC values decreases.
6 and 7	In order to have a optimal model, we have pruned the tree at optimal CP value.	After applying pruning with threshold and oversampling, we see the final model 7 have decent AUC, accuracy and sensitivity.

Accuracy alone is not a good measure of performance on unbalanced classes. Hence, it's usually better to look at the confusion matrix to better understand the model and look at metrics other than accuracy such as the sensitivity or AUC.

C50:

When we train our model to C50, we took the below input parameters to decide on the best model:

- subset - optional expression saying that only a subset of the rows of the data should be used in the fit. The model defaults this parameter to FALSE, meaning no attempted groupings will be evaluated during the tree growing stage
- Control factor(cf) - A number in (0, 1) for the confidence factor.
- mincases - an integer for the smallest number of samples that must be put in at least two of the splits
- trials - an integer specifying the number of boosting iterations. A value of one indicates that a single model is used.

On doing the modelling though C50 based on the mentioned input parameters, we got the below observations:

Model	Input Parameters					Training Data		Test Data	
	subset	control factor	mincases	trials	oversampling	Accuracy	Sensitivity	Accuracy	Sensitivity
C50_model1	Default(F)	Default	Default	Default	No	85.58%	3.50%	85.30%	1.20%
C50_model2	T	0.5	5	3	No	85.45%	2.29%	85.39%	0.80%
C50_model3	T	0.5	10	5	No	85.79%	8.30%	84.91%	4.15%
C50_model4	T	0.8	10	10	No	86.05%	7.60%	85.17%	2.80%
C50_model5	T	0.25	2	5	Yes	99.99%	99.99%	80.21%	16.46%

We clearly observe that though the accuracy in all the cases is quite high, the sensitivity is extremely low which can't be accepted for a good model. When we tried with oversampling, the sensitivity is improved a bit, but the training data is highly overfitted and hence the model captures all noises of the training data.

(c) Identify the best tree model. Why do you consider it best? Describe this model – in terms of size, variable importance. Do you find the different decision tree models to differ on variable importance?

Best model – rpart model 7: On comparing all the observations of rpart and C50, we have concluded that model 7 of rpart with information gain method and 70:30 split as our best model. This is because this models correctly trains our training data without capturing much noise, hence no over fitting. We have also pruned our decision tree and did oversampling in order to predict the "charged off" category correctly. We can see that the final model gives us good accuracy, AUC and sensitivity on the test data i.e. all greater than 60%.

Cases	Parameters	Training data				Test data			
		Accuracy	Specificity	Sensitivity	AUC	Accuracy	Specificity	Sensitivity	AUC
rpart model 7 with 70:30 split	Method = Information gain min split=10 cp = 0.0003 max depth = 10 Threshold = 0.4 Oversampling done Pruned at min cp	67.40%	63.39%	72.34%	0.679	62.26%	62.16%	62.82%	0.667

Performance evaluation on test data-

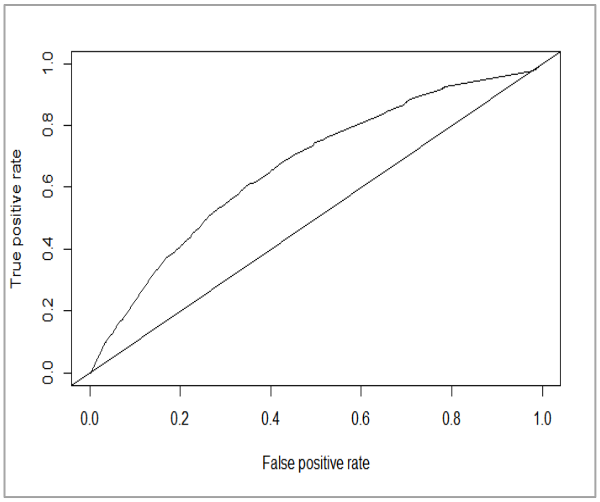
Confusion matrix for the rpart model 7 on test data is explained below:

	References	
Prediction	Fully Paid	Charged Off
Fully Paid	14912	1519
Charged Off	8835	2521

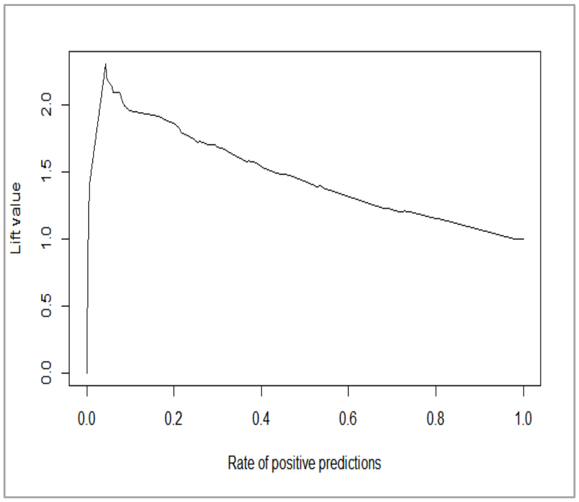
Inference: Our Decision tree on descriptive model data predicts the following

- 2521 defaulters predicted as defaulters correctly
- 1519 defaulters predicted as fully paid incorrectly
- 8835 fully paid predicted as defaulters incorrectly and
- 14912 fully paid predicted as fully paid correctly

ROC Curve of test data:



Lift Curve of test data:



Since ROC curve is in the N-W side of the ROC space, it means the model has better performance i.e. higher true positive rate and lower false positive rate.

Below is the list of important variables for rpart model 7 and c50 model, these variables are listed basis their rank for both the types of model:

Rank	Rpart	C50
1	int_rate	int_rate
2	sub_grade	installment
3	grade	grade
4	fico_range_high	sub_grade
5	bc_open_to_buy	num_accts_ever_120_pd
6	purpose	fico_range_high

S.No.	Common & Unique Variables	Variable_Name
1.	Top Ranked variables common to Rpart & C50	int_rate, sub_grade, grade & fico_range_high
2.	Top Ranked variables unique to Rpart	bc_open_to_buy & purpose
3.	Top Ranked variables unique to C50	Installment & num_accts_ever_120_pd

The above table clearly depicts the reason for both the model to perform differently, out of the top 6 ranked variables for both the models 2 are unique to each model (**S.No. 2 & 3**).

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?

The parameters we experimented with in the random forest are

- the number of trees and
- mtry(No. of variables to choose at every split)

These two are the key parameters for random forest model apart from the common parameters for tree-based models e.g. depth, child node etc. We have also tried with the oversampled data. The default value of mtry is sqrt(p) where p is the number of variables. The number of variables in our model is 48 and hence we tried mtry = 6 as well and found the performance to be similar.
Below are the various observations:

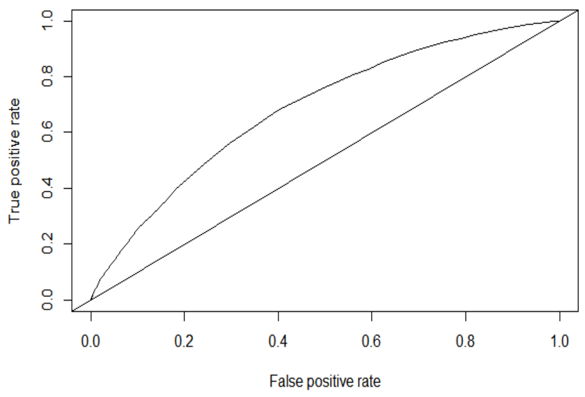
			Test data				
Models	No. of trees	mtry	Oversampling	Accuracy	Specificity	Sensitivity	AUC
rfmodel_1	100	Default	No	85.43%	99.79%	1.20%	0.504
rfmodel_2	50	6	No	85.37%	99.62%	1.60%	0.506
rfmodel_3	30	Default	No	85.23%	99.43%	1.80%	0.517
rfmodel_4	100	Default	Yes	85.18%	99.03%	3.76%	0.54
rfmodel_5 with threshold = 0.25	100	Default	Yes	68.35%	70.44%	56.06%	0.686

We select rf model 5 as best model as it gives good accuracy and sensitivity.

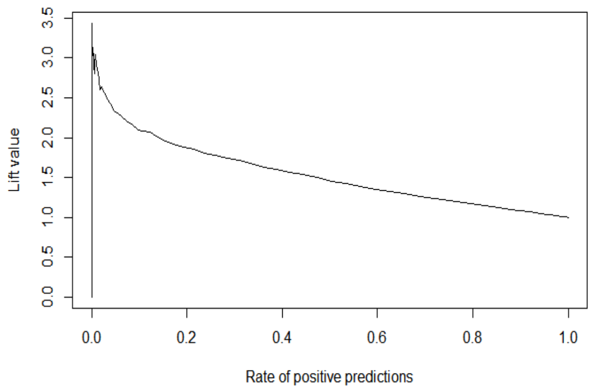
Confusion Matrix of test data:

	References	
Prediction	Fully Paid	Charged Off
Fully Paid	16727	1775
Charged Off	702	2265

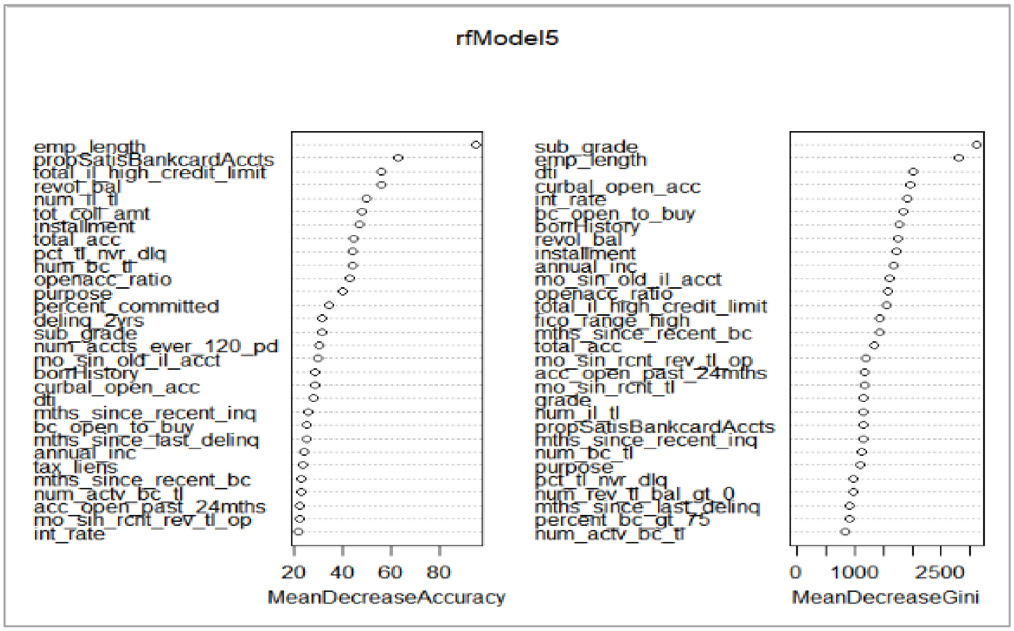
ROC for test data:



Lift curve for test data:



Variable importance graph for random forest model 5:



Comparison between rpart model 7 and rf model 5:

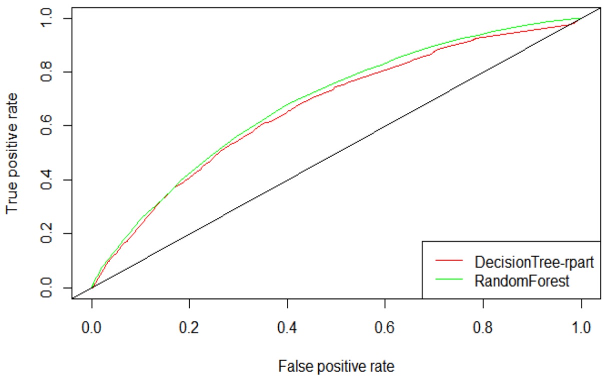
We chose random forest model 5 over rpart model 7 as it gives higher accuracy, AUC & sensitivity.

Models	Accuracy	Specificity	Sensitivity	AUC	Conclusion
rfmodel_5 with threshold = 0.25	68.35%	70.44%	56.06%	0.686	Better than rpart
rpart model 7	62.26%	62.16%	62.82%	0.667	

We would prefer Random Forest over Decision Trees because the idea behind Random Forest is bagging which will help to reduce bias and increase the variance.

There is difference in the variable importance of both the models. The top variables considered for preparing the rpart decision tree are sub_grade, int_rate, purpose, grade and fico_range_high whereas for the random forest model the top variables are emp_length, probStatisbankcardaccts, total_il_high_credit_limit, revol_bal and num_il_tl which shows a difference in the variable importance for the random forest and the rpart decision tree model.

Below is the comparison of ROC curves for both the models:



Random forest gives importance to weak predictors as well by randomly sub-setting the variables at each split. Random Forest can catch interactions from weak predictors as well. The performance of Random Forest is significantly higher than our rpart decision tree model for the reasons stated above. Also, as it can be seen in the plot, Random Forest is quick and more accurate in achieving better true positive rate (sensitivity), which can be beneficial for us as our main aim is to predict the "defaults".

6.The purpose of the model is to help make investment decisions on loans. How will you evaluate the models on this business objective?

In order to evaluate the models on business objective of making investment decisions on loans, calculating expected profit on the test dataset will be the correct approach.

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?

On calculating the actual annual return by formula:

Actual annual return rate = (total payment – funded amount)/ funded amount) * (1/actual Term)

Average return rate for fully paid for 3 yrs = Mean (actual annual returns for fully paid) * 3 = 7.53% * 3 = **22.51%**

Average return rate for charged off for 3yrs = Mean (actual annual returns for charged off) * 3 = -12.31% * 3 = **-36.90%**

So on an average, we can earn a total return of 22.51% on principle amount after 3 years of investing in Fully Paid loans. The potential loss from a charged off loan for 3 years is 36.90%.

Therefore, if we will invest in Fully Paid, total return would be $122.51 and if we invest in Charged off loan, potential loss can be $36.9 and we will only get $63.1 as return.

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.

The average interest rate in the LC dataset is 11.2% thus gain would be $133.6 but as seen above that the average return on investment $122.51 which is less than $133.6, because a loan can be paid back early (the actual loan duration can be less than 3 years).

Hence, the actual average return was calculated as $122.51. Therefore, interest rate won't be a correct indicator of actual returns on loan. Instead, we should use total payment amount, total funded amount and actual loan term to calculate the actual returns.

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.

For a loan which is charged off, there is on average 36.9% loss and not 100% loss. This is because there are some loans that have partial returned amount which accounts for reducing the loss. We can expect 63% of the invested amount to recover.

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:

(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?

In order to calculate the total profit and evaluate the performance of our models, we first have to calculate the actual returns for fully paid and for charged off.

Avg actual return for fully paid = mean (total payment amount of fully paid – funded amount of fully paid) = $ 1822.147

Avg actual return for charged off = mean (total payment amount of charged off – funded amount of charged off) = -$4607.713

RPART MODEL 7:

1. Threshold = 0.3

Confusion Matrix of test data:

	References	
Prediction	Fully Paid	Charged Off
Fully Paid	8659	695
Charged Off	15088	3345

In terms of profit and loss:

	References	
Prediction	Fully Paid	Charged Off
Fully Paid	$ 15,777,972.00	$ (3,202,360.00)
Charged Off	$ 27,492,555.00	$ (15,412,799.00)

Hence, total benefit – total loss = $495,855.2
Now, calculating in a similar manner for 0.4 and 0.5 thresholds and for random forest model 5 for all 3 different thresholds:

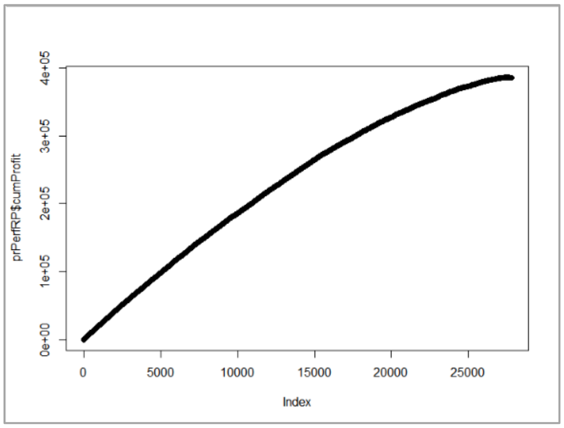
Thresholds	rpart model 7	RF model 5
0.3	$ 495,855.20	$ 23,632,444.00
0.4	$ 15,690,116.00	$ 25,659,751.00
0.5	$ 19,437,969.00	$ 25,228,020.00

After calculating total calculated profit and loss from rpart model and random forest models at different Classification Thresholds of 0.3, 0.4 and 0.5; it can be concluded that random forest model performed well with a total profit of $ 25,659,751.00 at 40% threshold.

(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).

Rpart model 7:

Similar analysis was done on rpart model 7. After sorting the data on probabilities for fully-paid loan status on testing dataset for rpart model 7, the values for cumulative profits is shown in figure below. The corresponding probability/cutoff threshold of fully-paid for the maximum cumulative profit value for our model is coming out to be 0.3 i.e. 30% threshold.

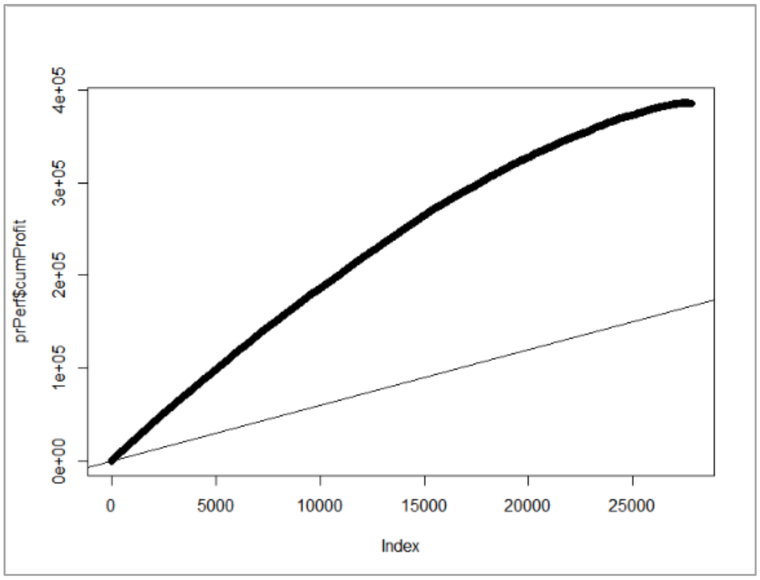


Random Forest model 5:

After sorting the data on probabilities for fully-paid loan status on testing dataset for random forest model 5, the values for cumulative profits is shown in figure below. As it can be observed from the plot that value of profit increases up to a certain value and then declines afterwards.

The corresponding probability of fully-paid for the maximum cumulative profit value should be the cutoff threshold for our model. From the analysis the threshold is **0.48 with the total profit of $ 386507.9.** While training the same model, the threshold was taken as 0.25.

Also, on comparing this model with that of investing in safe CDs, we found the below comparison graph:



It can be easily interpreted that the cumulative profits from random forest are exponentially higher than that of safe CDs. The highest profit obtained by random forest is at TH = 0.48 and for CDs, it is 0.26

Threshold	Random forest profit	CDs profit
0.48 - maximum for RF	$ 386,507.90	$ 165,162.00
0.26 - maximum for CDs	$ 385,469.00	$ 166,722.00

In both the cases, random forest proved to be the better model. Now, comparing this finding with that of part (a), we got the best model at TH = 0.4 and this is in sync with the findings in (b).

Hence, RF model 5 with threshold 0.48 is the best model for our Lending club data set.