**ASSIGNMENT 1A & 1B:**

**Loan default prediction and investment strategies in online lending.**

**PART A**

**1.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? (Not more than 1.5 pages, single spaced, 11 pt font. Please cite your sources).**

**Ans:**

Peer-to-peer lending (P2P), sometimes called “social” or “crowd” lending is the practice of [lending](https://en.wikipedia.org/wiki/Loan) money to individuals or businesses through online services that match lenders with borrowers, cutting out the financial institution as the middleman. [Websites that facilitate P2P lending](https://www.investopedia.com/articles/investing/092315/7-best-peertopeer-lending-websites.asp) have greatly increased its adoption as an alternative method of financing. Each website sets the rates and terms and enables the transaction. Most sites have a wide range of interest rates based on the creditworthiness of the applicant. One can register as a borrower or lender on any platform after undergoing a verification process by furnishing relevant details. Once registered, investors can reach out to listed borrowers and vice versa.

The rate of interest typically ranges from 10% to 28% and the loan tenure may range from 3 months to 36 months. Both the lender and borrower get benefit from the P2P business model as the lenders get to earn higher returns from their investments and borrowers get to access quick loans at lower cost. The key growth strategies of Peer-to-Peer Lending market players include product portfolio expansion, mergers & acquisitions, agreements, geographical expansion, and collaborations.P2P Lending platforms earn money from the various fees they charge like origination fee which is a one-time fee of 3-6% of the loan amount based on the credit rating, transaction fee, pre-processing fee, interest rate etc. from the borrowers.

P2P lending platforms have many options like reinvestment where investors earn their income from the loans, they invest that get credited to their escrow account on the platform every month. They have the option to withdraw these amounts or reinvest them back in loans listed on the platform. By activating reinvestment, investors ensure that their monthly earnings automatically get reinvested in the same products or plans that they have selected and continue to generate returns for them reducing their time and effort. It provides automated investment options which reduce the time and effort required in building a portfolio, instead of spending time in studying and selecting each borrower profile, investors can choose to add funds to auto invest and select the various parameters which match their investment strategy. The algorithm automatically builds the portfolio by matching the investment objectives with borrower profiles listed on the platform. The latest, most efficient and least time taking method of investing in P2P lending is when lots of investors pool their monies into a single portfolio to achieve efficiency in portfolio building and management. The pool uses data science and artificial intelligence (AI) to build and manage a portfolio that has the potential to deliver high and stable returns. Once investors add their investment amount and authorized the platform to disburse it, the job is done. The platform’s algorithm will disburse the pool money into a diverse mix of loans and loan products who as per it have the repaying capacity to provide high aggregate returns. These options attract the investors/lenders towards P2P lending platforms.

The global peer to peer (P2P) lending market size was valued at $67.93 billion in 2019, and is projected to reach $558.91 billion by 2027, growing at a CAGR of 29.7% from 2020 to 2027. In September 2021, P2P loan interest rates were as low as 5.99%.P2P platforms like upstart have a starting interest rate of 3.22% with a minimum credit score which is not in case of banks and that’s the reasons P2P platforms have emerged so good over the years. The peer-to-peer lending industry in the US started in February 2006 with the launch of [Prosper Marketplace](https://en.wikipedia.org/wiki/Prosper_Marketplace), followed by [LendingClub](https://en.wikipedia.org/wiki/LendingClub" \o "LendingClub). In 2013, LendingClub was the largest peer-to-peer lender in US based upon issued loan volume and revenue, followed by Prosper. LendingClub was also the largest peer-to-peer lending platform worldwide. The interest rates ranged from 5.6–35.8%, depending on the loan term and borrower rating. The default rates varied from about 1.5% to 10% for the riskier borrowers. LendingClub abandoned the peer-to-peer lending model in the fall of 2020 and acquired a bank to focus on other ventures of its business.

LendingClub offers the unique combination of being a well-capitalized, publicly traded U.S. corporation (LendingClub Corporation), and a national bank (LendingClub Bank), along with a digital consumer lending arm that has originated over $63 billion in loans since inception. Today they have partnered with dozens of asset managers, hedge funds and other investment vehicles as well as over 60 U.S. banks who have previously purchased personal loans. Their business model allows bank partners to save all loan marketing and origination costs, increasing projected return even further.

They have multiple platforms that allows investors to purchase loans on a passive or active basis, in bulk or as individual loans, and as whole loans. This provides flexibility to accommodate and meet their partner’s needs. Lastly, they can serve banks of all sizes, from those wishing to purchase a small amount of loans to several hundreds of millions every month. In addition to offering personal loans across the full credit spectrum, LendingClub offers auto refinancing loans and purchase finance loans (including medical, dental, and educational financing). Thus, Lendingclub’s strong customer loyalty, engagement, ease for the lenders and borrowers to purchase and invests loans has made them different and superior from other P2P lending platforms.

CITATIONS:

1.Peer-to-Peer (P2P) Lending By [JULIA KAGAN](https://www.investopedia.com/contributors/53409/) (Updated May 11, 2020) - <https://www.investopedia.com/terms/p/peer-to-peer-lending.asp>

2.How To Earn Passive Income With P2P Lending by [Rajat Gandhi](https://www.forbes.com/advisor/in/author/rajat-gandhi/) - <https://www.forbes.com/advisor/in/personal-finance/how-to-earn-passive-income-with-p2p-lending/#:~:text=Lenders%20earn%20their%20income%20from,loans%20listed%20on%20the%20platform.&text=Thereafter%2C%20they%20do%20not%20have,more%20time%20investing%20those%20funds>

3. “Peer to Peer (P2P) Lending Market – global opportunity analysis & industry forecast, 2020-2027.” Allied Market Research, <https://www.alliedmarketresearch.com/peer-to-peer-lending-market>.

<https://www.lendingclub.com/investing/peer-to-peer>

<https://www.lendingclub.com/investing/institutional/overview>

<https://www.lendingclub.com/loans/personal-loans>

4. Peer-to-peer lending - <https://en.wikipedia.org/wiki/Peer-to-peer_lending#United_States>

5.What is peer-to-peer lending? Here are 5 things to know -   
<https://economictimes.indiatimes.com/wealth/borrow/what-is-peer-to-peer-lending-here-are-5-things-to-know/articleshow/85921771.cms>

6.LendingClub - <https://en.wikipedia.org/wiki/LendingClub>

7.Best Peer-to-Peer Lending by [ALLISON BETHEL](https://www.investopedia.com/allison-bethel-5191094) - <https://www.investopedia.com/articles/investing/092315/7-best-peertopeer-lending-websites.asp>

**2. Your team’s ultimate goal is to help a client determine whether s/he should invest in p2p loans. What is the final decision that you will help the client make? What is the objective, and how will you evaluate ‘better’ vs ‘worse’ decisions? What is the goal of predictive models for this? What will be the potential target variables?**

**Ans:**

P2P lending is a platform through which people can invest and enjoy higher returns, higher than even high-yielding savings account or other deposits. In this lending system, the investors deposit their money that will be loaned out through the platform to potential borrowers. The lender is then able to view all the loan requests by different borrowers, assess the potential risks of investing in a particular loan and approve and provide the full or partial loan amount for that loan.

Our goal here is to make this assessing process easier for the investors. In order to help the clients or lenders choose the right loan to invest on, we would create a predictive model displaying the various comparisons such as high-risk vs low-risk loans, high return vs low return loans, etc. We would aim at building a model which would help lenders distinguish between a good vs bad investment so that they can invest their money wisely and generate better returns.

The lenders could be private individuals and/or industrial investors. Based on their financial status, the definition for ‘better’ or ‘worse’ decision would vary. If he/she wants to go for a low-risk investment, then higher grade loans would be a better choice since it would have less chances of defaults at a lower interest rate, thereby not giving higher returns for the lender. But if the lender is not concerned about the risk and is only looking for good returns, then he/she should opt for loans with lower grades as they have higher interest rates along with higher risk of defaulting. Also, high-risk loans with low or no security may have higher interest as compared to low-risk loans with good security and low interest. Therefore, the goal of our predictive model would be to display all these combinations of possibilities so that the lender can make a right choice for them. The target variables that we will involve in our study would be the loan status (fully paid vs charged off), loan grades and subgrades, interest rates, loan amount, annual income, employment length, credits history, loan amount to annual income, etc.

Sources:

[Best Peer-to-Peer Lending Websites of February 2022 (investopedia.com)](https://www.investopedia.com/articles/investing/092315/7-best-peertopeer-lending-websites.asp)

**3. Data exploration**

**a) Take a look at the data attributes. How would you categorize these attributes, in broad terms, considering what they pertain to? What are attribute types - which are numeric, categorical, and date variables? What do you think will be the important attributes to consider for your decision task? Which attributes do you think will help determined performance?**

**Ans:**

We can categorize the attributes based on what information they provide us. For example, Lending Club (LC) uses grades and subgrades to assess each borrower. They have 7 grades, and each grade has 5 subgrades. LC looks at the various factors based on the information provided by the borrower to assign these grades. Loan interest rate is another such variable which LC calculates for each loan grade. Similarly, loan amount, term, purpose of the loan, application type, disbursement method are the various attributes that tell us more about the loan. LC captures plenty of information about the borrowers so that lenders can learn about each borrower in depth before investing in their loan requirement. Some such borrower characteristics are employment details, annual income, housing situation, etc. Also, a few variables also tell us about the loan’s performance like the loan status, total payment, recoveries, number of charge-offs within 12 months, etc. The credit history is calculated based on number of enquires, number of past due incidences of delinquency, past due delinquency amount, number of open credit lines for a borrower, etc. Therefore, the attributes can be divided in the following five categories:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Borrower Characteristics** | **Loan Characteristics** | **Platform Characteristics** | **Loan Performance** | **Credit History Characteristics** |
| emp\_title | loan\_amnt | Id | loan\_status | delinq\_2yrs |
| emp\_length | funded\_amnt | member\_id | total\_pymnt | inq\_last\_6mths |
| home\_ownership | funded\_amnt\_inv | int\_rate | total\_pymnt\_inv | mths\_since\_last\_delinq |
| annual\_inc | Term | Grade | total\_rec\_prncp | mths\_since\_last\_record |
| zip\_code | issue\_d | sub\_grade | total\_rec\_int | open\_acc |
| addr\_state | pymnt\_plan | verification\_status | total\_rec\_late\_fee | pub\_rec |
| Dti | Desc |  | last\_pymnt\_d | revol\_bal |
| annual\_inc\_joint | Purpose |  | last\_pymnt\_amnt | revol\_util |
| dti\_joint | disbursement\_method |  | chargeoff\_within\_12\_mths | total\_acc |
| tot\_cur\_bal |  |  |  |  |

The variables belong to either numeric, character, logical or date data type. We have converted the categorical variables to factors and date variables to date type. Below is the categorization of all the variables based on their data type:

|  |  |  |  |
| --- | --- | --- | --- |
| **Numeric** | **Date** | **Factor** | **Logical** |
| loan\_amnt | last\_pymnt\_d | application\_type | id |
| funded\_amnt | last\_credit\_pull\_d | initial\_list\_status | member\_id |
| funded\_amnt\_inv | earliest\_cr\_line | Title | url |
| int\_rate | next\_pymnt\_d | Purpose | desc |
| Installment | issue\_d | loan\_status | annual\_inc\_joint |
| annual\_inc |  | verification\_status | dti\_joint |
| Dti |  | home\_ownership | verification\_status\_joint |
| delinq\_2yrs |  | emp\_length | revol\_bal\_joint |
| inq\_last\_6mths |  | Grade | sec\_app\_earliest\_cr\_line |
| mths\_since\_last\_delinq |  | sub\_grade | sec\_app\_inq\_last\_6mths |
| mths\_since\_last\_record |  | Term | sec\_app\_mort\_acc |
| open\_acc |  | debt\_settlement\_flag | sec\_app\_open\_acc |
| pub\_rec |  | hardship\_flag | sec\_app\_revol\_util |
| revol\_bal |  | disbursement\_method | sec\_app\_open\_act\_il |
| revol\_util |  | pymnt\_plan | sec\_app\_num\_rev\_accts |
| total\_acc |  | addr\_state | sec\_app\_chargeoff\_within\_12\_mths |
| out\_prncp |  | emp\_title | sec\_app\_collections\_12\_mths\_ex\_med |
| out\_prncp\_inv |  | zip\_code | sec\_app\_mths\_since\_last\_major\_derog |
| total\_pymnt |  | hardship\_flag | hardship\_type |
| total\_pymnt\_inv |  | disbursement\_method | hardship\_reason |
| total\_rec\_prncp |  | pymnt\_plan | hardship\_status |
| total\_rec\_int |  | addr\_state | deferral\_term |
| total\_rec\_late\_fee |  | emp\_title | hardship\_amount |
| Recoveries |  | zip\_code | hardship\_start\_date |
| collection\_recovery\_fee |  |  | hardship\_end\_date |
| last\_pymnt\_amnt |  |  | payment\_plan\_start\_date |
| collections\_12\_mths\_ex\_med |  |  | hardship\_length |
| mths\_since\_last\_major\_derog |  |  | hardship\_dpd |
| policy\_code |  |  | hardship\_loan\_status |
| acc\_now\_delinq |  |  | orig\_projected\_additional\_accrued\_interest |
| tot\_coll\_amt |  |  | hardship\_payoff\_balance\_amount |
| tot\_cur\_bal |  |  | hardship\_last\_payment\_amount |
| open\_acc\_6m |  |  | debt\_settlement\_flag\_date |
| open\_act\_il |  |  | settlement\_status |
| open\_il\_12m |  |  | settlement\_date |
| open\_il\_24m |  |  | settlement\_amount |
| mths\_since\_rcnt\_il |  |  | settlement\_percentage |
| total\_bal\_il |  |  | settlement\_term |
| il\_util |  |  |  |
| open\_rv\_12m |  |  |  |
| open\_rv\_24m |  |  |  |
| max\_bal\_bc |  |  |  |
| all\_util |  |  |  |
| total\_rev\_hi\_lim |  |  |  |
| inq\_fi |  |  |  |
| total\_cu\_tl |  |  |  |
| inq\_last\_12m |  |  |  |
| acc\_open\_past\_24mths |  |  |  |
| avg\_cur\_bal |  |  |  |
| bc\_open\_to\_buy |  |  |  |
| bc\_util |  |  |  |
| chargeoff\_within\_12\_mths |  |  |  |
| delinq\_amnt |  |  |  |
| mo\_sin\_old\_il\_acct |  |  |  |
| mo\_sin\_old\_rev\_tl\_op |  |  |  |
| mo\_sin\_rcnt\_rev\_tl\_op |  |  |  |
| mo\_sin\_rcnt\_tl |  |  |  |
| mort\_acc |  |  |  |
| mths\_since\_recent\_bc |  |  |  |
| mths\_since\_recent\_bc\_dlq |  |  |  |
| mths\_since\_recent\_inq |  |  |  |
| mths\_since\_recent\_revol\_delinq |  |  |  |
| 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\_120dpd\_2m |  |  |  |
| num\_tl\_30dpd |  |  |  |
| num\_tl\_90g\_dpd\_24m |  |  |  |
| num\_tl\_op\_past\_12m |  |  |  |
| pct\_tl\_nvr\_dlq |  |  |  |
| percent\_bc\_gt\_75 |  |  |  |
| pub\_rec\_bankruptcies |  |  |  |
| tax\_liens |  |  |  |
| tot\_hi\_cred\_lim |  |  |  |
| total\_bal\_ex\_mort |  |  |  |
| total\_bc\_limit |  |  |  |
| total\_il\_high\_credit\_limit |  |  |  |

The attributes belonging to the Loan like the loan amount, purpose, term, funded amount, loan status etc., and Platform attributes like the grade, subgrade, interest rate, etc. could be considered as important attributes for our analysis.

The loan performance attributes like the loan status, payments received to date, principal received to date, outstanding principal amount, term, interest rate, number of charge-offs within 12 months, recoveries, total loan amount, funded amount are few of the attributes which could help us determine the performance of a loan.

**b) How will you calculate performance (returns) from a loan? There are multiple ways for calculating this. Outline two ways to calculate returns based on the data attributes; what are their advantages and disadvantages.**

**Ans:**

There are multiple approaches to calculate the performance or returns from a loan. We could use the below formula to calculate the annualized actual return with using the loan term as 3 years (or 36 months), if we assume that the amount paid back is not going to be reinvested till the term ends. This means that our investment is locked up until the term of the loan. This is a pessimistic approach to avoid risks. But if the loans are paid back early, then that money could be reinvested in other loans and returns could be gained.

**((Total Payment – Funded amount)/ Funded amount) \* 12/36 \*100**

Suppose if we calculate the actual return using the loan term as the actual term of the loan rather than 3 years (or 36 months), then we can use the below formula. In this case, we could reinvest the money from the early paid back loans into other loans and gain returns. But by doing so, we are being overly optimistic in thinking that the loans we initially invest in will not default. For example, if they default in the first month itself then the loss is 100% and the annualized return calculated using the below formula will show 1200% loss.

**((Total Payment – Funded amount)/ Funded amount) \* 12/Actual Loan Term \*100**

So, the ideal way would be to calculate the annualized return based on the funded amount and the total payments received. If the difference between the funded amount and the total payment received is greater than 0, i.e., it is fully paid, then use the formula with the actual loan term other if the difference between the funded amount and the total payment received is less than 0, i.e., Charged off, then use the formula with 3 years (or 36 months) as the loan term.

**IF Fully Paid, then ((Total Payment – Funded amount)/ Funded amount) \* 12/Actual Loan Term \*100**

**IF Charged Off, then ((Total Payment – Funded amount)/ Funded amount) \* 12/36 \*100**

**c) Examine the attributes which you think will be useful in your analyses and modeling. Obtain data descriptions and develop some plots to visualize the data. Summarize your observations (you answer should be more than just the figures and plots – what is the ‘story’ from your initial observations)?**

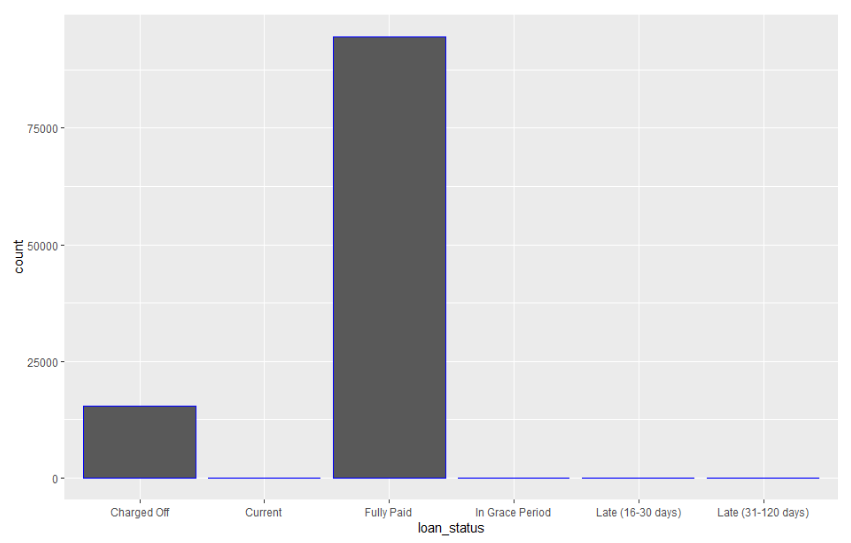
**Ans:** The attributes which will be useful for the analyses and modelling includes the following:

|  |  |
| --- | --- |
| **ATTRIBUTES** | **DATA DESCRIPTIONS** |
| loan\_amnt | Loan\_amnt is an important attribute as it tells us how much amount is given to borrower and if credit department reduces the loan amount that will be reflected in the loan\_amnt. |
| funded\_amnt | Funded\_amnt tells us about the amount that is committed to that loan. |
| funded\_amnt\_inv term | This amount talks about the amount funded by the investors/lenders for a loan. |
| int\_rate | This attribute helps us to analyze the interest rate over a period of time, and the rates for a particular grade and sub grade. |
| Installment | The monthly payment owed by the borrower if the loan originates. |
| Grade | Grade helps to analyse which loan an investor should invest it |
| sub\_grade | Grade has further bifurcations called subgrades which gives us more detail information on which loan an investor should invest under a particular grade. |
| home\_ownership | The home ownership status provided by the borrower during registration. Values included RENT, OWN, MORTGAGE, OTHER. This will certainly help the lenders to know about borrowers home status |
| annual\_inc | The self-reported annual income provided by the borrower during registration will let the lenders to know the information about annual income. |
| verification\_status | Indicates if income was verified by LC, not verified, or if the income source was verified |
| loan\_status | Current status of the loan. Helps to us analyze the counts of loan which are fully paid or charged off. |
| Purpose | A category provided by the borrower for the loan request. This attribute explains the purpose of the loan provided by the borrower. With this information there are chances that a lender can easily lend loan to a particular borrower according to their necessity with more or less interest rates. |
| Dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| mths\_since\_last\_delinq | The past-due amount owed for the accounts on which the borrower is now delinquent. |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| total\_pymnt | Payments received to date for total amount funded. |
| total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| total\_rec\_prncp | Principal received to date |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |

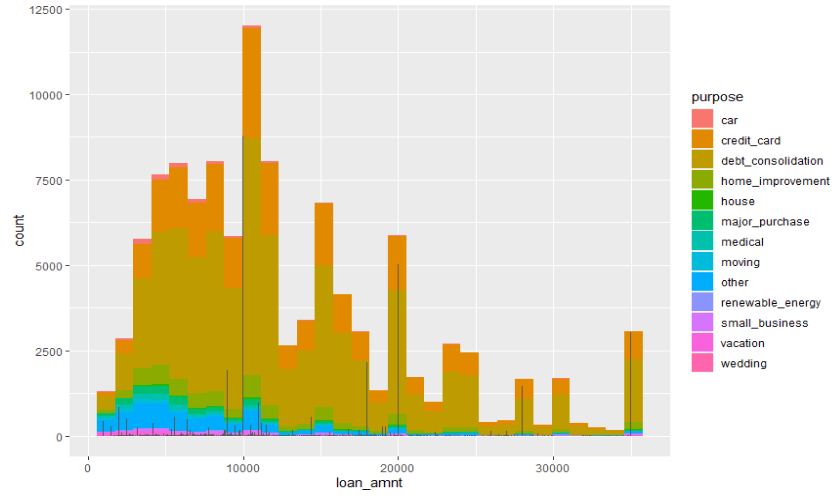
Here the below table shows the count of each loan status.



From the results, majority of the customers come under Fully Paid category and 15,377 come under Charged Off category (Fig 3.3.1).

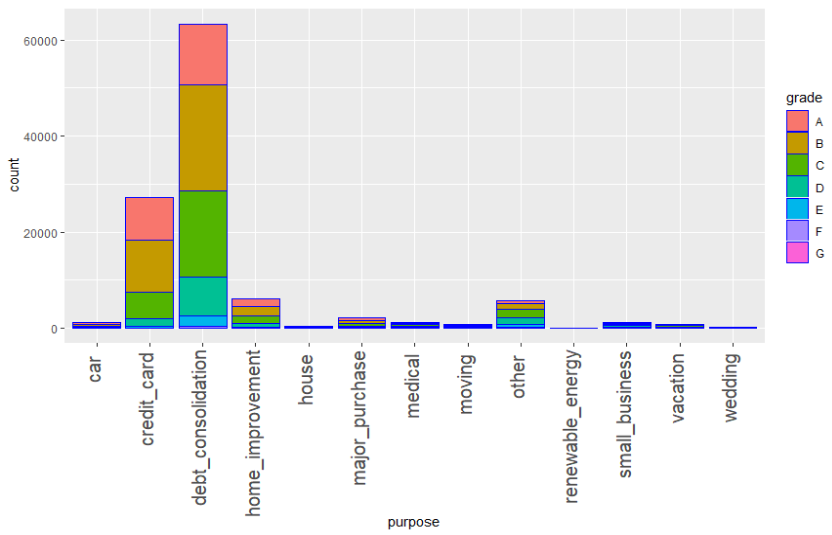


**Fig 3.3.1:** Loan status counts



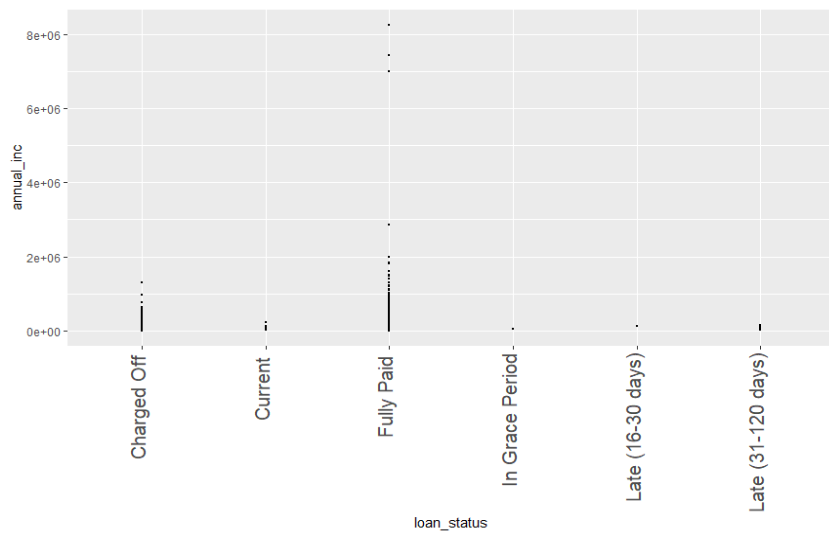
**Fig 3.3.2:** Loan amount vs purpose

Here the graph shows that maximum loan amounts are issued for the purpose of debt consolidation followed by credit card (Fig 3.3.2).

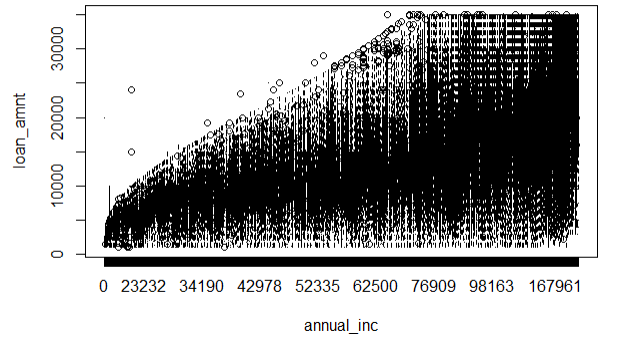


**Fig 3.3.3:** Loan issued for various purposes and their grades

The above graph shows number of people taking loans for different purpose and their grades. From the graph it is observed that most of them are going for grade A and grade B loans (Fig 3.3.3).



**Fig: 3.3.4:** Loan status vs Annual income

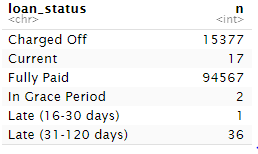


**Fig: 3.3.5:** Annual income vs loan amount

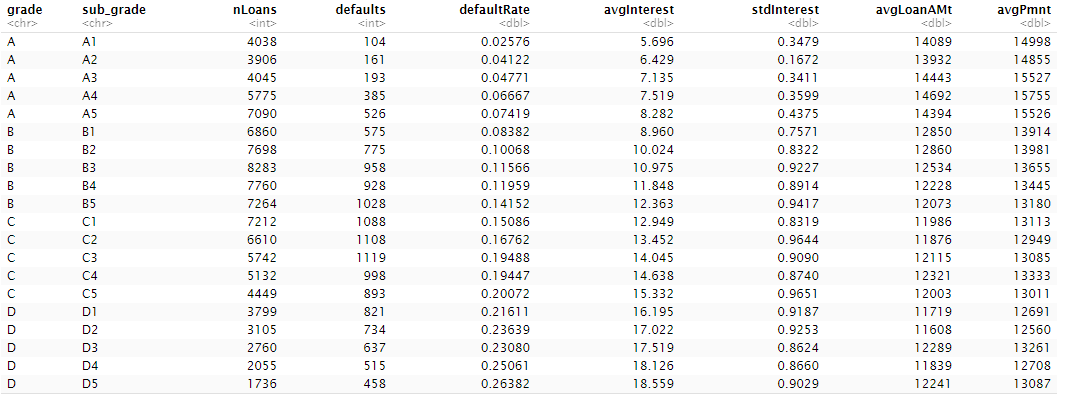
The graph shows that with increase in annual income, loan amount also increases (Fig 3.3.4 and fog 3.3.5). So, borrowers with high annual income, the credit department sanctions a loan of high amount and so the borrowers can fully pay off the loans within the given period of term.

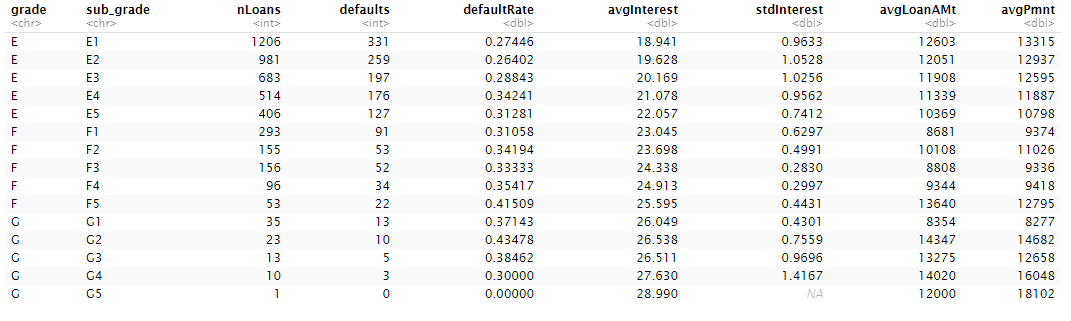
**d) i) What are the values for loan\_status? Are there values other than “fully paid”, “charged off”? We want to restrict attention to “fully paid” and “charged off” loans, so, other values should be removed. What is the proportion of defaults (‘charged off’ vs ‘fully paid’ loans) in the data? How does default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why?**

**Ans:** The different values for the loan statuses are as follows:



The table consists of values other than “fully paid” and “charged off”. As we are only interested in fully paid and charged off loans we need to remove other values. The proportion of Charged off loans are 15377 and for fully paid are 94567.



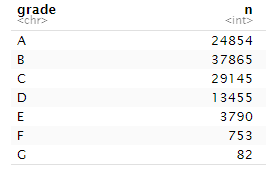


**Fig: 3.4.1.1:** Loan grade and subgrade level details

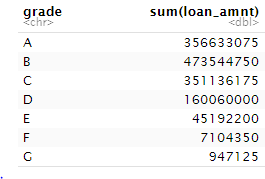
From the above Fig. 3.4.1.1, it is evident that as the loan grade decreases the average interest rates increases with maximum interest rates at grade F and G loans. And as the loan grade goes down from A to G default rate increases, this could be because, with such high interest rates, people are not able to repay the loans. It can be seen that in loan grade G and subgrade G5, the default rate is zero though interest rate is maximum that’s because the average payment is high (18102) and so the default count is zero.

**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?**

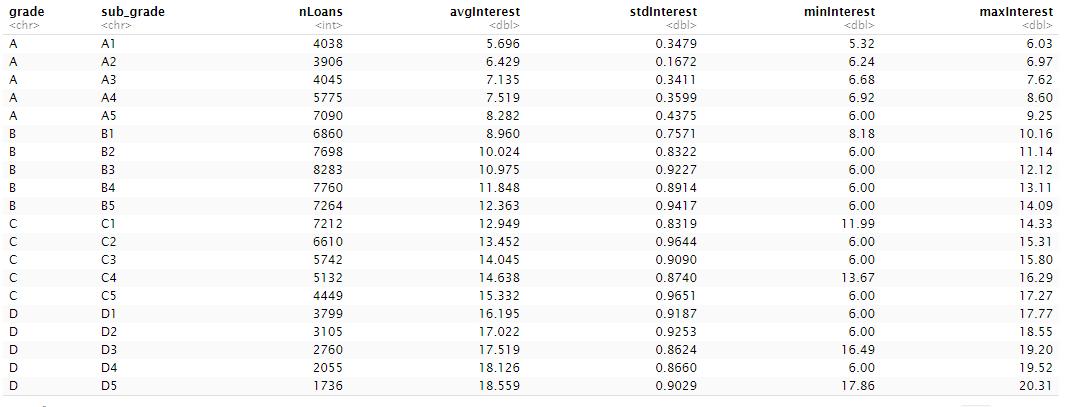
**Ans:** The loan distribution between the grades is as shown below:

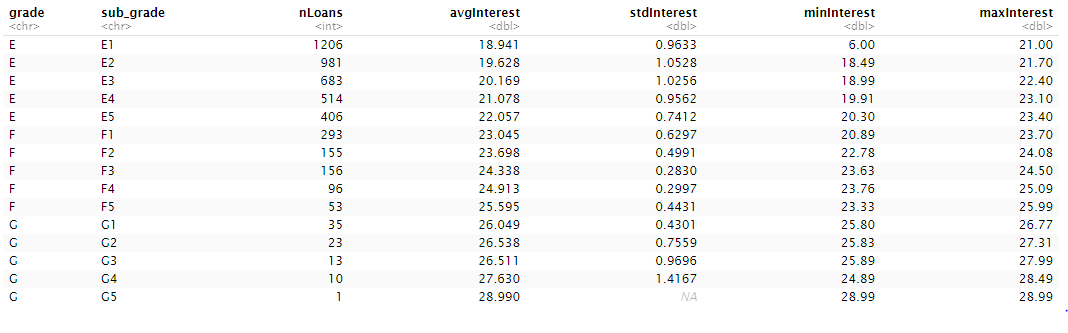


Grades A through D have the maximum number of loans. Since their average interest rate is also low, borrowers can make timely payments and repay the loan and so their default rates are also low.



The above table is evidence that loan amounts also vary with grade. The maximum loan amount is for Loan grade B which is 473544750 while loan grade G has the least loan amount 947125.





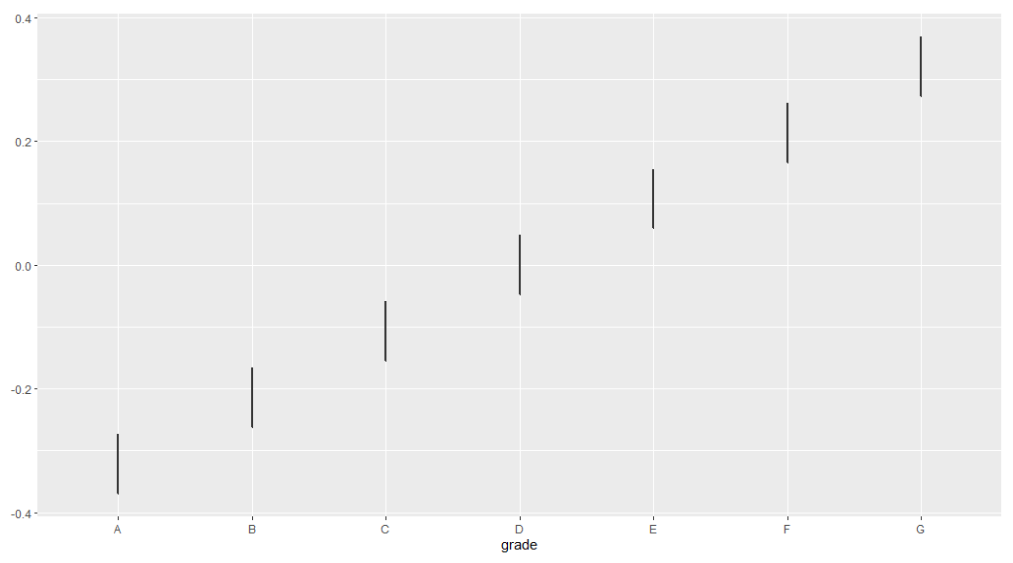
**Fig: 3.4.2.1:** Variables by loan grade and subgrades

From Fig: 3.4.2.1, interest rates for loans also vary with grade and sub grade. Here in the table loan grade A seems to be having least interest rate as compared to other loan grades. With each sub grade there is a steady increase in the average interest rate. The subgrades A1-A5 have interest rates in the range of 5.6% to 8.5% whereas loan grade G has sub grades G1-G5 with interest rates in the range 26%-29% which is very high.

The interest rates of the loans in the subgrades E2, E3 and G4 are almost 1.5 times deviated from the average interest rates. Further the table shows minimum and maximum interest rates for each sub grade e.g., A1 sub grade has an interest rate in the range of 5.3% to 6.3% which is the least while G5 has a range of 28.99% to 28.99%. It shows that each sub grade would be having interest rate in the respective range and not out of the range. This would help borrowers to decide on which loans to go for and help lenders to decide on which loans to invest by looking at the statistics of loans fully paid and charged off with the parameters like the above.

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

**Ans:** The term of a loan can be estimated by the date of the last payment made. However, there may be cases where the loans are paid back early. In those cases, the actual term of a loan can be calculated as the difference between the issued date of the loan and the last payment date.



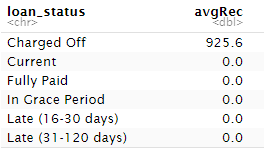
**Fig: 3.4.3.1:** Loan grade vs actual term

The average actual term increases with decreasing levels of grades (Fig 3.4.3.1). This can be true as the interest rate increases as the loan grade decreases and it might be difficult for the borrowers to pay back the loan at such high interest rates causing them to default and not paying back the loan in time. When the interest rates are lower, i.e., in higher loan grades, the loans are paid back earlier than their actual issued term.

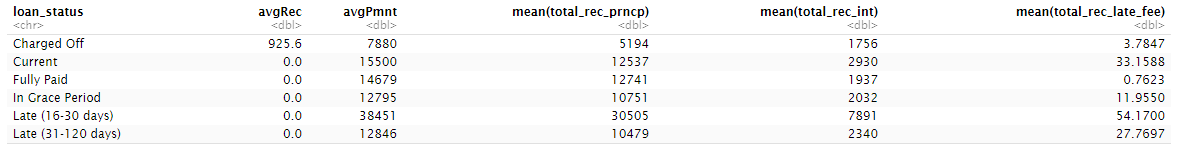
**iv) What is 'recoveries'? Can we assume that recoveries are only for Charged\_off loans? The data has multiple attributes on recoveries – what is the total amount of recoveries? For charged-off loans, does total\_pymnt include recoveries ?**

**Ans**:

The recoveries or recovery rate is the estimated percent of a loan or obligation that it will still be repaid to creditors in the event of a default or bankruptcy. From the table below with the loan status and average recovery rate recoveries are there only for the charged off loans. If a charged-off loan is sold to a third party or funds are recovered on a previously charged off loan, investors will receive a pro rata share of the sales proceeds or recovery amount, respectively, less any fees.



The average recovery rate for charged off loans is 925.6 and for rest of the loans it is found to be zero.



**Fig: 3.4.4.1:** Recovery details for different loan statuses

'Recoveries' has the total of recoveries on principal, on interest, and late-fees. Charged off loans have a total recovery principal of 51.94%, total recovery interest of 17.56% and total recovery late fee of 3.78%.

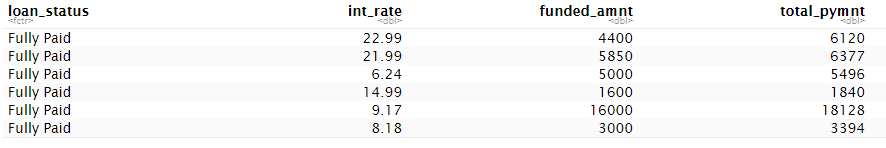
**v) Calculate the annual return. Show how you calculate the percentage annual return. Is there any return from loans which are ‘charged off’? Explain. How does return from charged -off loans vary by loan grade? Compare the average return values with the average interest\_rate on loans – do you notice 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?**

Ans:

Loans have interest rates, which is the percentage of amount charged over the loaned amount by the lender to a borrower on an annual basis. It means the rate of returns for a lender. However, we cannot totally rely on a loan’s interest rate to calculate the return from it, and this has been explained below.

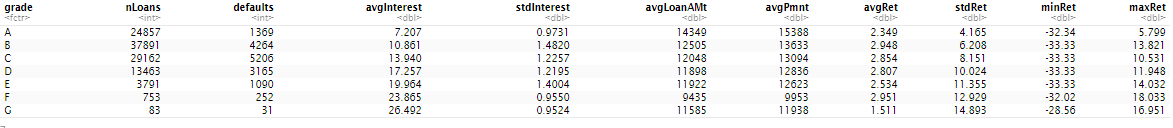
If we look at the below table (Table3.4.5.1), it shows the interest rates, loaned amount and total payments for a few fully paid loans for 3 years. By just looking at the interest rate on the funded amount, we can say that the total payment received is much lower than what was expected. So then let’s say that the actual return should be the difference between the funded amount and the total payments. And based on above hypothesis, we can come up with the below formula for calculating annualized return (in %) as:

**((Total Payment – Funded amount)/ Funded amount) \* 12/36 \*100**

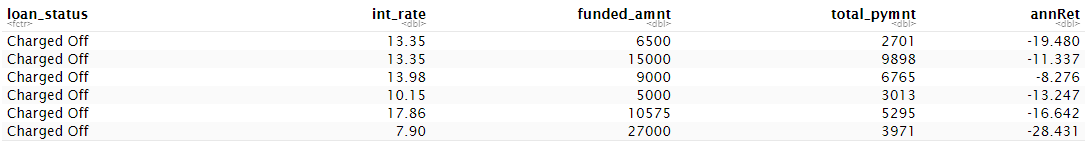


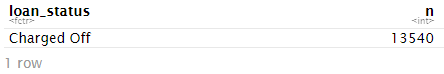
**Table 3.4.5.1** Loan details for Fully Paid loans

Now let’s look at the summary of the loans along with their average annualized returns (Table3.4.5.2). By comparing the values in the average interest rate (avgInterest) and maximum returns (maxRet), we can say that the actual return is much less than expected. Although the average interest rates show a higher value, there are some negative values in the minimum returns (minRet) columns. These negative values could be from Charged Off loans since in that case we know that there could be a loss of money instead of gain. To reconfirm this, lets filter out the data where the annualized return is less than 0 or negative (Table 3.4.5.3 a) and group these loans based on their loan statuses, we see that they all belong to the Charged Off loans (Table 3.4.5.3 b).



**Table 3.4.5.2:** Loan details of Loans

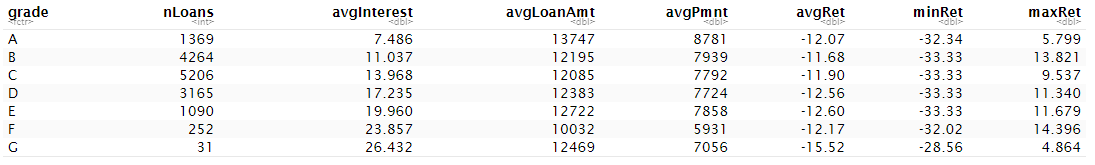
(a)



(b)

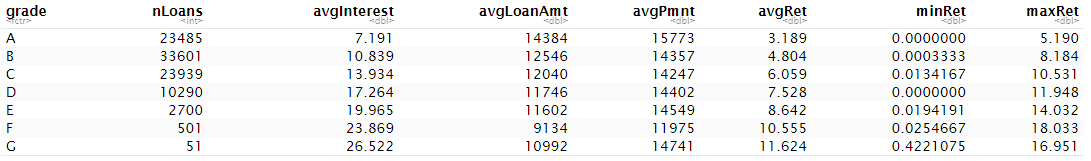
**Table 3.4.5.3**: (a) Loans where the annualized returns are less than 0 (b) All the loans from (a) are Charged off loans

Let’s look more closely at the annualized returns from Charged Off loans (Table 3.4.5.4). We already saw that the Charged off loans have negative minimum returns as their funded amount are not paid back fully and defaulted. But if we look at their maximum returns, there is not a complete loss of money. Some are even better as compared to the fully paid back loans(Table 3.4.5.5).



**Table 3.4.5.4:** Charged Off Loans

Now let’s look at the annualized returns from the Fully Paid loans (Table3.4.5.5). The average returns earned by the Fully Paid loans are higher compared to the Charged off loans we examined above (Table3.4.5.4). With lower grades, the average interest rates increase along with the average actual returns. However, there is still a gap between the average interest rates and average actual returns calculated. Also, the maximum returns are much less compared to the average interest rate. This could be because some borrowers may be paying off their loans before their actual term of 3 years (as per our assumption).



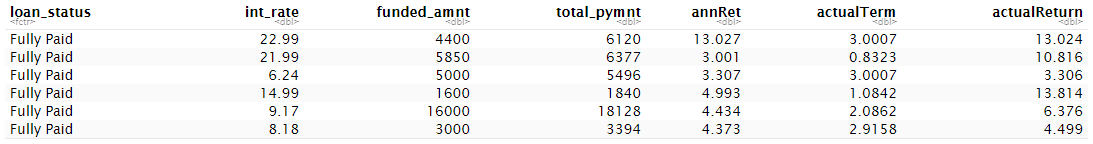
**Table 3.4.5.5:** Fully Paid loans

So, the problem really is when the borrowers pay off their loans earlier than the term date. To solve this let’s fix our formula for annualized returns. Instead of taking the 3 years, lets calculate the actual loan term i.e., the duration between the last payment date and the issue date of the loan. This would give us the actual time taken to fully pay back a loan. This will fix the fully paid back loans, but can we apply the same logic to Charged off loans? The answer is no, since the Charged off loans are never paid back fully. So, for these loans, we can let the actual term be 3 years (36 months) as before. The improved formula will be:

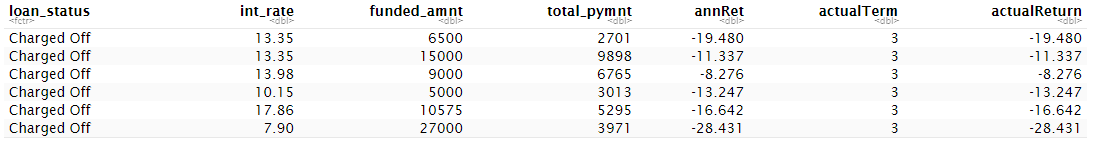
**IF Fully Paid, then ((Total Payment – Funded amount)/ Funded amount) \* 12/Actual Loan Term \*100**

**IF Charged Off, then ((Total Payment – Funded amount)/ Funded amount) \* 12/36 \*100**

As per the new formula, the actual return(actualReturn) is much higher than the annualized return(annRet) we had calculated before in case of Fully Paid loans (Table3.4.5.6 a). However, the annualized return and the actual return is the same in case of the Charged Off loans since we did not change the formula in case of the Charged off loans (Table3.4.5.6 b).



(a)



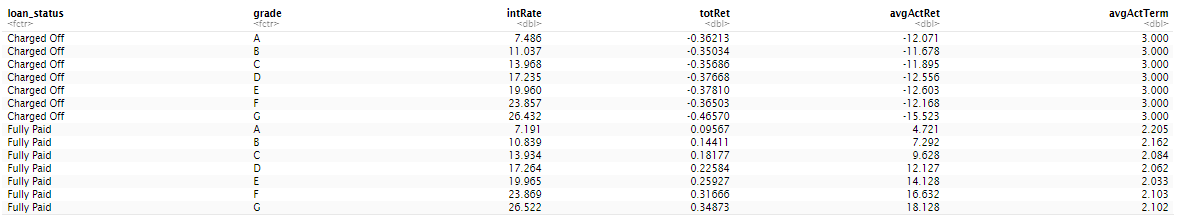
(b)

**Table 3.4.5.6:** New Annualized returns and actual term for (a) Fully paid loans (b) Charged Off loans

If we look at the summarized figures for both types of loans (Table3.4.5.7), then we can see that the new annualized return(avgActRet) gives us better and more realistic values than before.



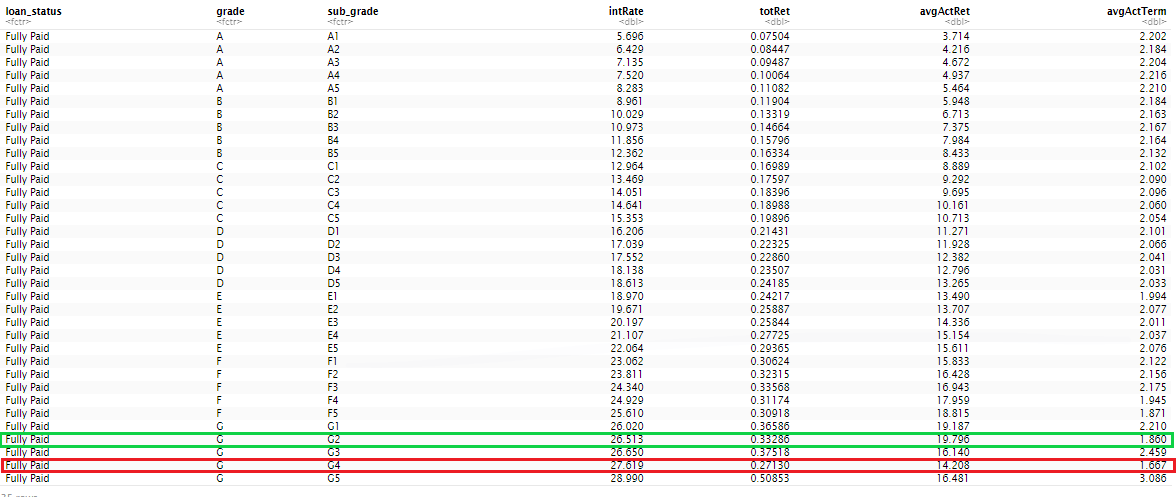
**Table 3.4.5.7:** Summarized values for Fully Paid and Charged Off loans



**Table 3.4.5.8:** Summarized values as per loan grade and loan status

In the above summary of loans as per their grade and status (Table3.4.5.8), we can see that for Charged off loans the total returns are in negative since the money invested is not paid back fully. However, for Fully paid back loans, with lower grades the interest rates increase and so is its average actual returns. Also, some of the lower grade loans are returned much faster than the higher-grade ones as per the values in the average actual term(avgAcctTerm) column. Therefore, looking at this data we could say that lower grade fully paid loans are better to invest in since they have higher returns. But the caveat here is they may also have higher risk of defaulting.

Let’s dig in further in case of Fully paid back loans and look at their distribution at sub grade levels as well (Table 3.4.5.9). The interest rate still increases as we move down the grade and subgrade levels. The loans at grade G have highest average actual returns and interest rates as compared to other grades. The loans with subgrades G2 and G4 are paid back much earlier than most of the other grade loans. Further, subgrade G4 has lower average actual returns than G2. G2, in fact has the highest actual return as compared to all grades even though its interest rate is less than a few other loans in the same grade category as G. Therefore, as per our data exploration we believe G2 is a better loan grade to invest in for fully paid back loans, even with a risk of defaulting.

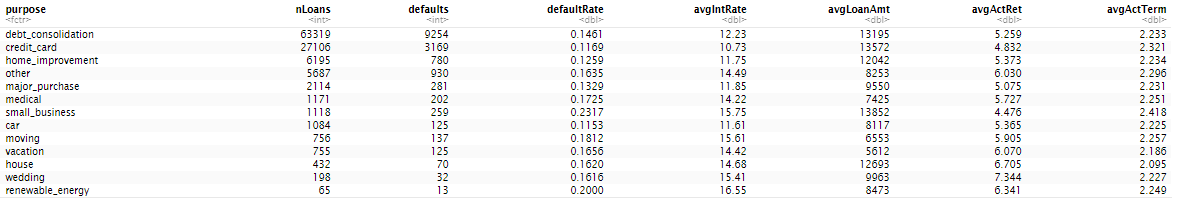


**Table 3.4.5.9:** Summarized values for Fully paid loans grouped by grade and subgrade

**vi)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?**

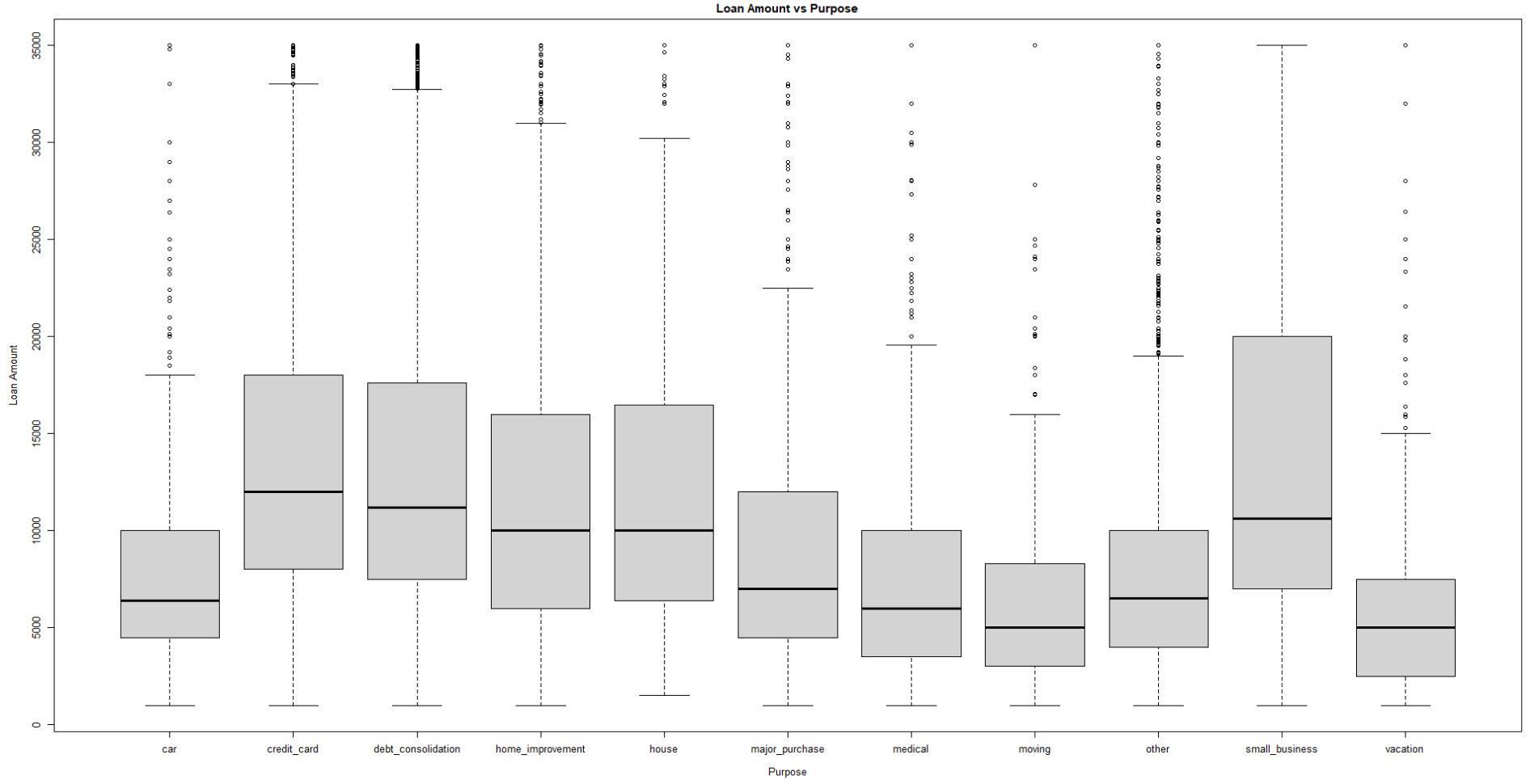
**Ans:**

The loans were borrowed for debt consolidation, credit card payments, home improvements, medical purposes, business purposes, car, moving, vacations, etc. The summary of the loans based on their purposes have been exhibited by the table 3.6.1. The maximum number of loans are issued for debt consolidation, followed by credit card and then home improvements.



**Table 3.4.6.1** Summary of loans as per their purposes

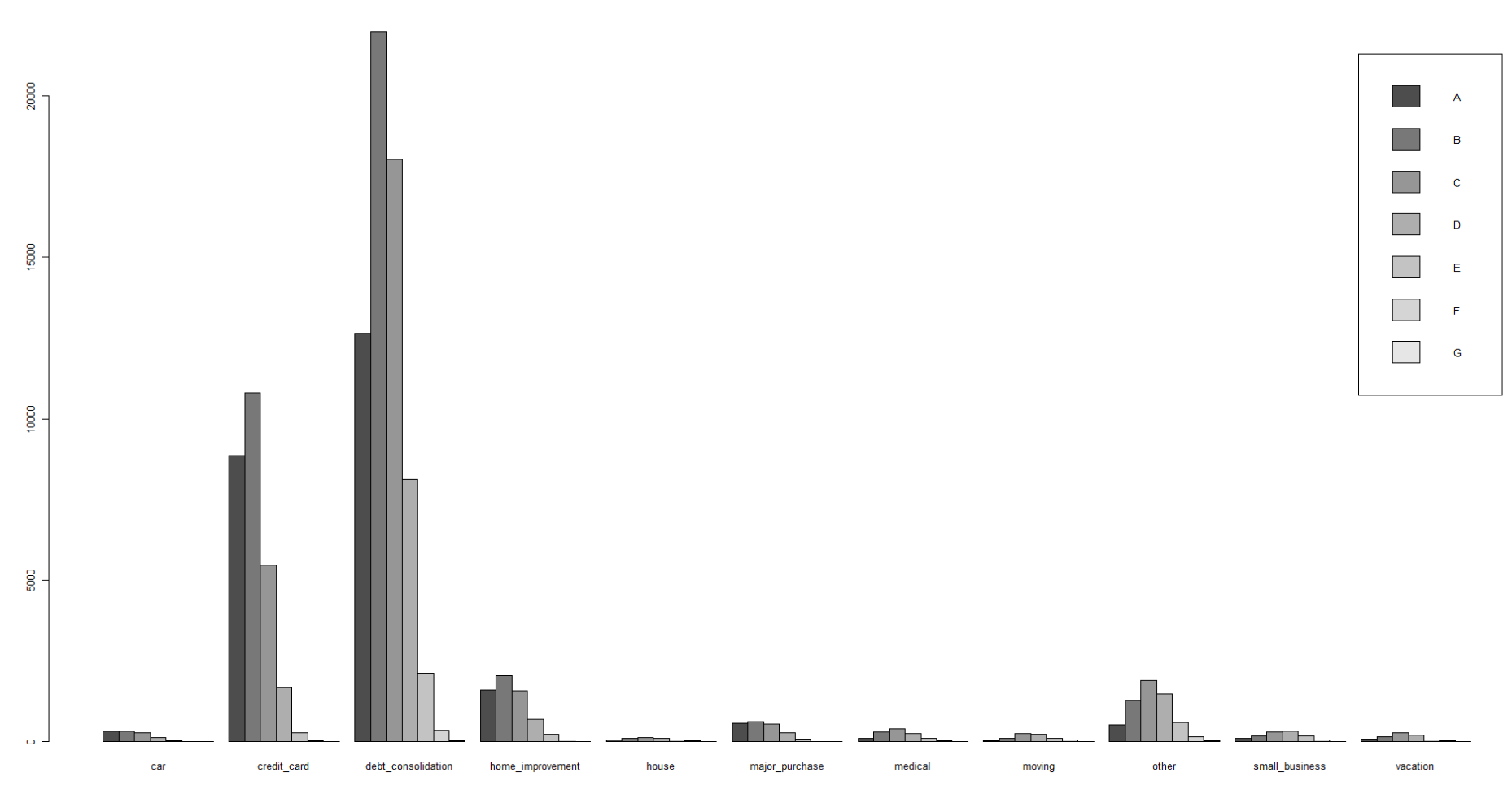
The loan amounts do vary by their purposes as seen in the boxplot below (Fig 3.4.6.1). The small businesses have the highest average loan amount issued for them, followed by credit card and then debt consolidation.



**Fig 3.4.6.1** Loan Amount vs purpose

From Table 3.4.6.1, we can see that the defaults also vary by the purpose. As the number of loans issued for debt consolidation purpose is highest, the number of defaults is also high for debt consolidation. Therefore, there is a direct correlation between the number of loans and defaults. However, the default rate is the maximum for the small businesses as the loan amount for it is the highest.

The grades also vary by the purposes as evident from the below bar plot (Fig 3.4.6.2). In most of the purposes, the maximum number lies in the grade B or C category.

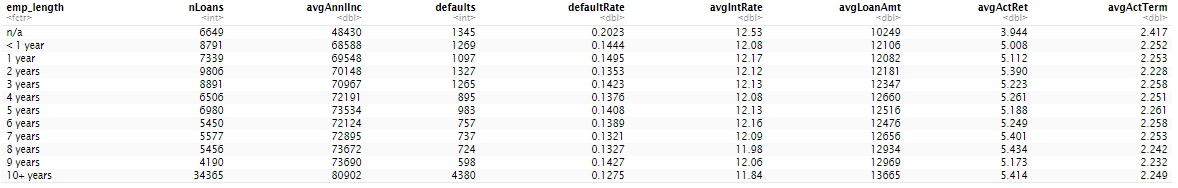


**Fig 3.4.6.2:** Loan purposes vs loan grades

**vii) 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.**

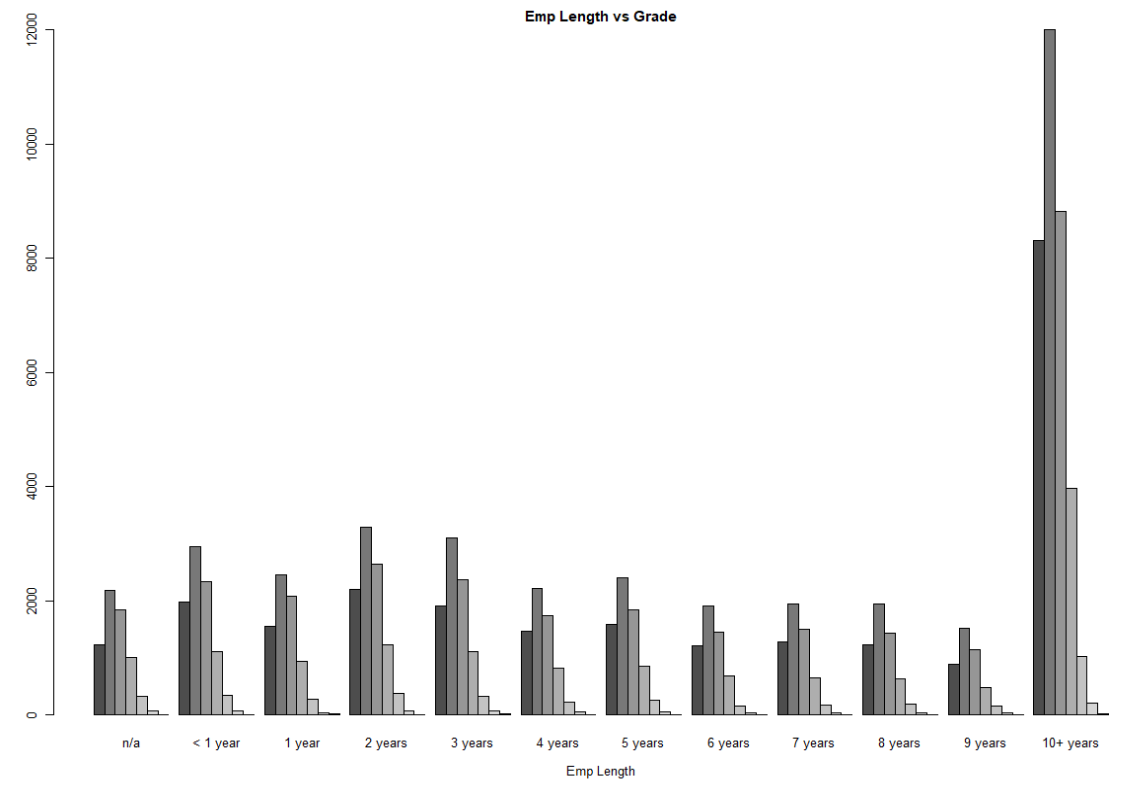
**Ans:**

The table below (Table 3.4.7.1) displays the relation between the borrower characteristics like the employment length and annual income to the loan attributes like loan amount, actual return, defaults, interest rates, term etc. The maximum number of loans are taken by the borrowers who are employed for more than 10 years. As their annual income is the highest, they have the lowest default rate. The interest rate decreases between 1 to 4 years of employment length of borrowers and is again the lowest for 10+ years of employed borrowers and they are most likely to take a loan. The average loan amount also increases with the employment length and is the highest for the 10+ years employed borrowers. The average annual return is almost the same for all the borrowers with minimal differences. And so is the case for the actual term of the loan.



**Table 3.4.7.1:** Borrower characteristics

If we look at the bar plot below (Fig 3.4.7.1), we see that in every category of employment length of the borrowers, the most number of loans belong to the grade B.



**Fig 3.4.7.1:** Emp length vs loan grade

When we compare the loan statuses and the emp length, the borrowers with more than 10 years of employment have the maximum number of loans in case of both the statuses, Fully Paid and Charged Off.

As we saw in our previous question, the loan purpose “debt consolidation” has the highest number of the loans issued and most of these accounts to the borrowers with the employment length greater than 10 years.

**viii) 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:**

1. propSatisBCAccts: Proportion of satisfactory bankcard accounts to the total bankcard accounts. This basically shows the accounts in good standing for a borrower.
2. borrHistory: Length of borrower's history with LC which is the difference between the month the borrower's earliest reported credit line was opened and the month the loan was funded. It shows a borrower’s ability to repay debts and is directly proportional to the default rate.
3. ratioOpenAccount: It is the ratio of the number of open credit lines in the borrower's credit file to the total number of credit lines. This gives an idea of the number of active accounts of the borrower.

**(d2) Summarize your conclusions and main themes from your analyses above.**

**Ans:**

From the above data exploration, we can conclude that majority of the loans are in Fully Paid status (around 86%) while the rest is in Charged Off status (14%). The average returns earned on the Fully paid back loans are higher than that in case of Charged off loans. However, this doesn't always mean that investing on the Charged Off loans would always lead to a loss of investment as there are few Charged off loans which have higher returns than that compared to Fully paid ones.

We also see that the average interest rates increase as the loan grade decreases. The higher interest rates may be since the lower grade loan categories have higher risk at defaulting and LC may charge higher in order to minimize the loss incurred to the investors. The medium grade loans could be considered as a safer bet for investors as they have moderately better returns and lower risk of being charged off.

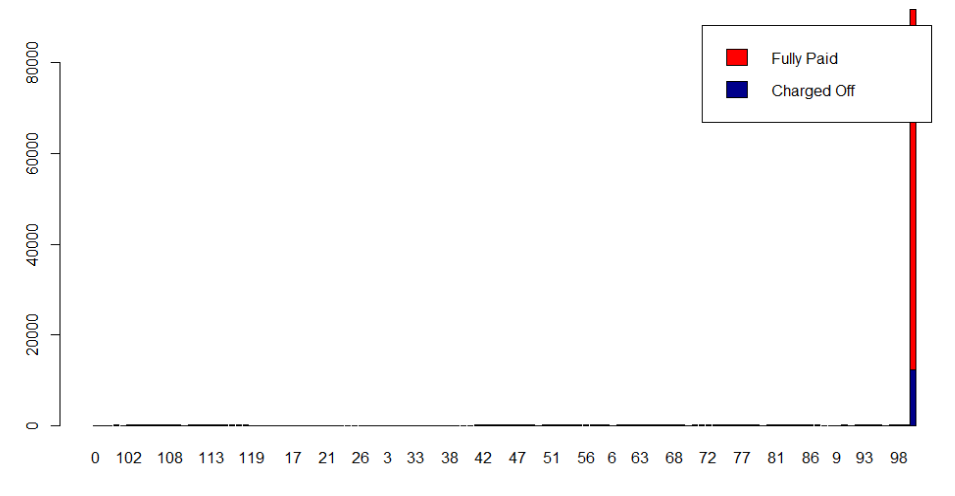
The small businesses, credit card and debt consolidation have the maximum amount of loan amount issued for each purpose, the default rate is also the highest in these cases as the borrowers may find it difficult to pay back such high amounts on time. The mid-grade loans have the most amount of these loan purposes.

Therefore, to conclude there are higher returns in case of the lower grade loans but with higher risk of defaulting and lower risk with lower returns in case of higher-grade loans. The mid-grade loans could be considered safe for better profits and lower risk.

**e) 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?**

**Ans:**

A missing value is one whose value is unknown. Missing values are represented by the NA symbol. NA is a special value whose properties are different from other values. In this dataset there are many missing values present. Attributes like “id”, “member”, “url” which is URL for the LC page with listing data, “desc” which is Loan description provided by the borrower have all the values as NA. These attributes are not so anyway useful for the analysis of this dataset. We can remove those attributes or values. Generally missing data creates imbalanced observations, cause biased estimates, and in extreme cases, can even lead to invalid conclusions. So, it is always advisable to remove them or substitute them with a column mean/median or mean of nearest neighbours or moving average.



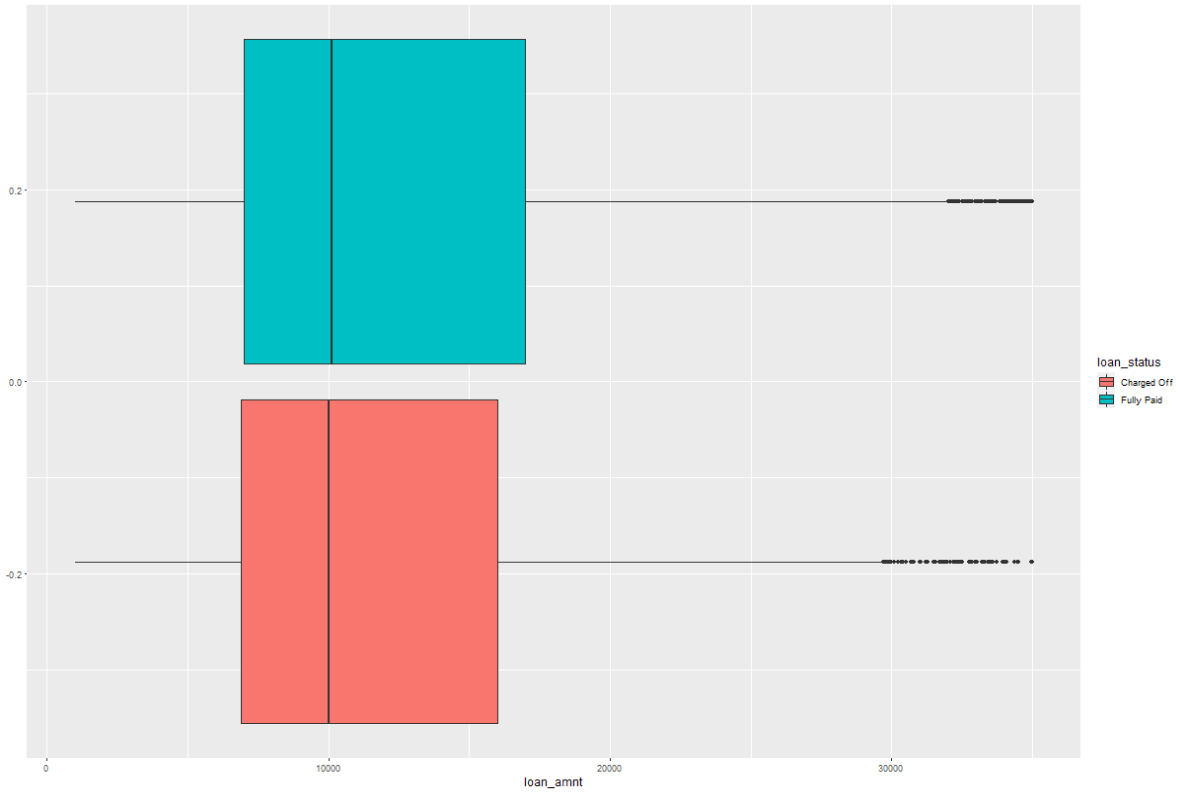
**Fig: 3.5.1:** Values at different loan statuses

The graph (Fig. 3.5.1) shows the counts by loan status at different values of the variable. Variable like monthsSinceLastDeliquency may have no value for someone who has not yet had a delinquency; The sensible value to replace the missing values in this case could be mean/median of the whole column or find a set of records which is similar to this attribute and which does not have missing values or zero if one really requires this attribute for analyses else the attribute can be ignored. Here in this dataset the few missing values in a column are replaced by the column median values.

**f) Consider potential outliers. Explain how you identify outliers – i.e. what specific analyses you use (eg. summary(), histograms, boxplots,…). Describe how a boxplot identifies outliers. Would you use this approach here (or, should outliers be determined based on data specifics and application context --leading question If you do choose to remove outliers, explain what you do, and how this affects your data.**

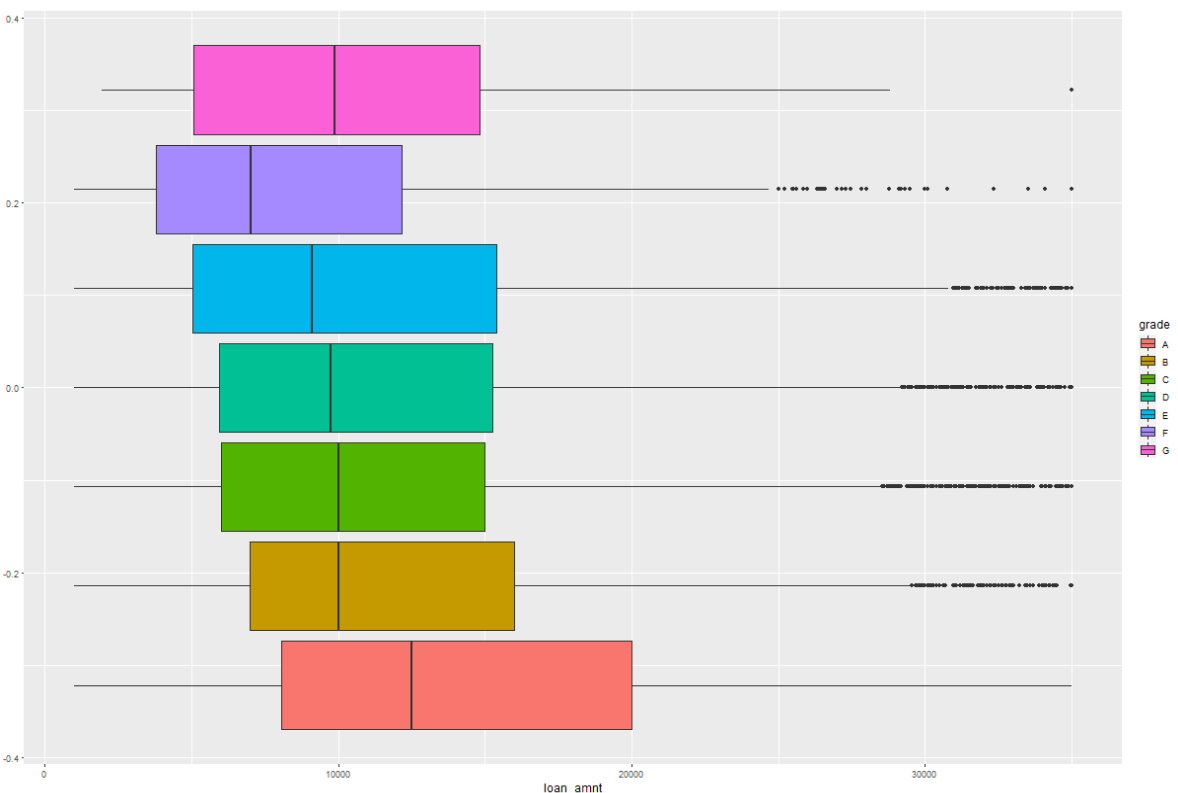
**Ans:**

An outlier is a data point which is distant from the rest of the data points. A box plot can be used to detect outliers within a data set. In a box plot, the data points which are outside the [25%Q-1.5\*IRQ, 75%Q + 1.5\*IRQ] range are considered as the outliers and marked in bold. However, it may be problematic if we remove all these data points considering them as outliers as removing too many may result in very few data samples or we may even lose out on important data points.



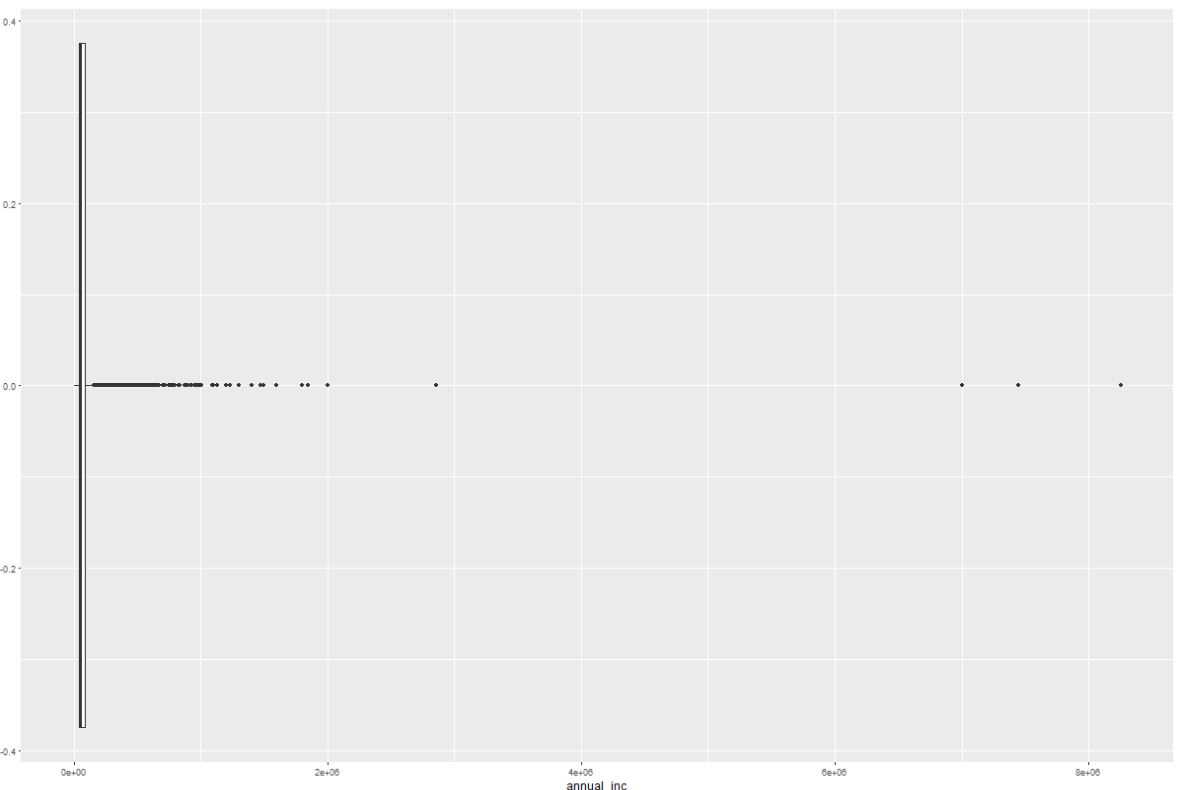
**Fig 3.6.1:**  Loam amount vs loan grades

If we look at the above boxplot(Fig3.6.1) we can see that when considering the loan amounts there are few outliers in both the loan statues.



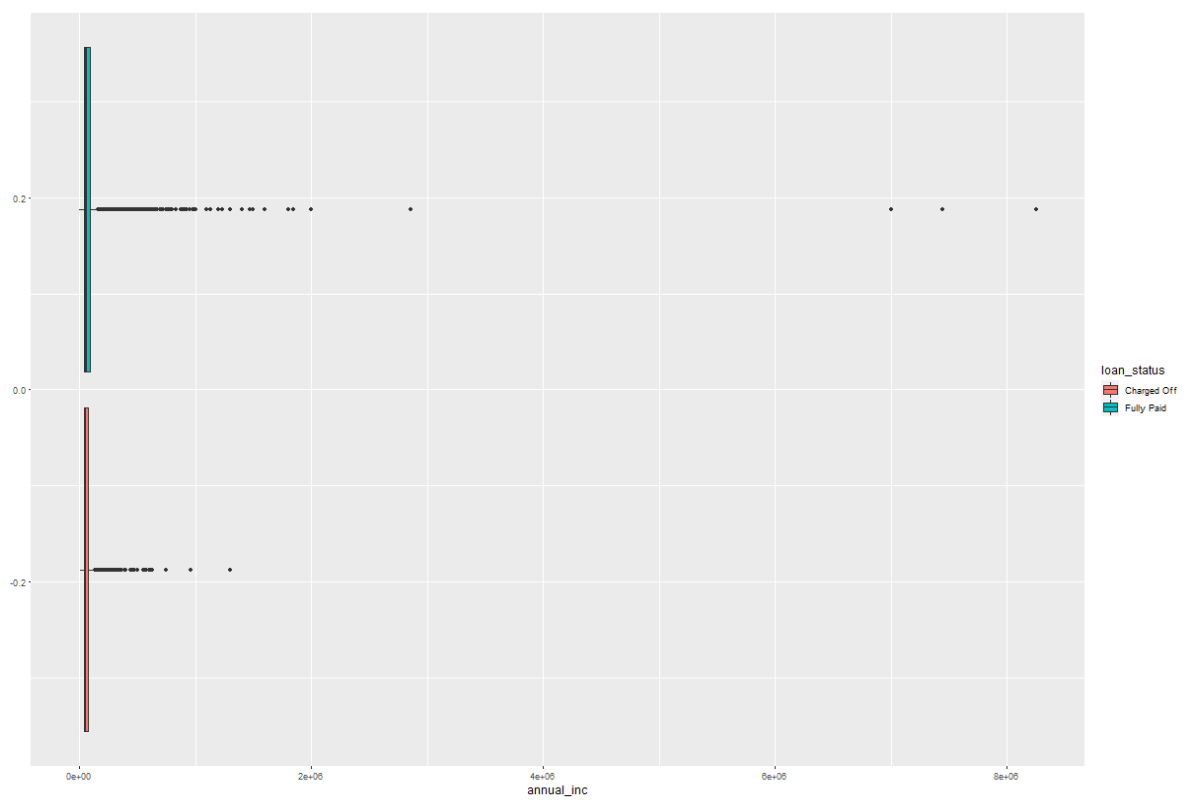
**Fig 3.6.2:** Loan amount vs grade

From fig 3.6.2, it is evident that there are few outliers in almost every grade of the loan.

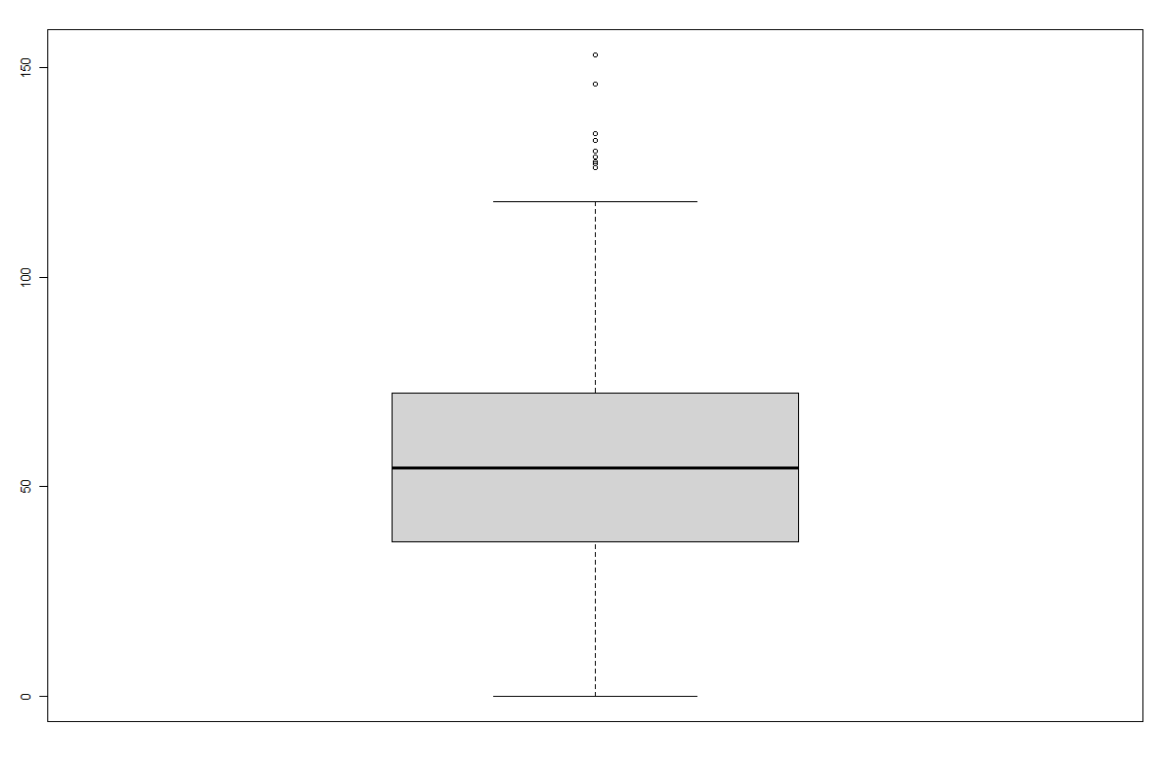


**Fig 3.6.3:** Annual Income

If we look at the annual incomes, we see a few higher income values. Now to check if these values really are outliers and needs removal, lets consider the box plot of the annual income and their statuses. (Fig 3.6.4) From the figure below, it looks like very high-income values are from fully paid off loans and so they can be excluded from our dataset as we already know that higher paid borrowers can afford to repay the loan and there might not be any chance of defaulting. So we could filter out the data set by removing the annual incomes greater than $1.5M as the number of outliers are also less in this case.



**Fig 3.6.4:** Annual Income vs loan status



**Fig 3.6.3:** Boxplot for revol\_util

Let’s look at the boxplot for revol\_util column (Fig 3.6.5). There are 9 outliers in case of the revolving line utilization rate. After analyzing these values we can conclude that we can remove these values safely.

**4. 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 for leakage considerations, and explain why?**

**Ans:**

When the training data carries information which may not be available at the time of applying the model to new data, then it is called as data leakage. Variables such as funded\_amnt\_inv, term, emp\_title, pymnt\_plan, title, zip\_code, collection\_recovery\_fee, address\_state, policy\_code, application\_type, disbursement\_method et.al are removed from this dataset as they are not required for the analysis and certainly, they will lead to data leakage. Here we also need to drop variables like last\_pymnt\_d, last\_pyment\_amnt, next\_pymnt\_d, playment\_plan\_start\_date et.al as they are not required before the issue of loan. These variables might also be potential for a data leakage. It is very important to utilize the necessary attributes for analyses as using the wrong or unwanted attributes can lead to false analyses or prediction of data.

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

**Ans:**

When considering a single variable model to predict the values of loan\_status, we could use area under the ROC curve (AUC) as the measure for evaluating the accuracy of our model. We will consider the threshold as 0.5 and all the variables which have AUC > 0.5 can be used as potential predictor variable. However, some of the variables like actual return, actual term, recoveries, total payments, principal received to date, interest received to date, payment plan, hardship flag can contribute to data leakage and must be removed from our analyses. The below are the variables which can be considered as good predictors:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| loan\_amnt | The listed amount of the loan applied for by the borrower |
| int\_rate | Interest Rate on the loan |
| Installment | The monthly payment owed by the borrower if the loan originates. |
| Grade | LC assigned loan grade |
| sub\_grade | LC assigned loan subgrade |
| home\_ownership | The home ownership status provided by the borrower during registration or obtained from the credit report. |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| verification\_status | Indicates if income was verified by LC, not verified, or if the income source was verified |
| Purpose | A category provided by the borrower for the loan request. |
| Dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |
| num\_accts\_ever\_120\_pd | Number of accounts ever 120 or more days past due |
| num\_actv\_bc\_tl | Number of currently active bankcard accounts |
| num\_actv\_rev\_tl | Number of currently active revolving trades |
| num\_op\_rev\_tl | Number of open revolving accounts |
| num\_rev\_tl\_bal\_gt\_0 | Number of revolving trades with balance >0 |
| num\_sats | Number of satisfactory accounts |
| num\_tl\_90g\_dpd\_24m | Number of accounts 90 or more days past due in last 24 months |
| num\_tl\_op\_past\_12m | Number of accounts opened in past 12 months |
| tot\_hi\_cred\_lim | Total high credit/credit limit |
| total\_bal\_ex\_mort | Total credit balance excluding mortgage |
| total\_bc\_limit | Total bankcard high credit/credit limit |
| total\_il\_high\_credit\_limit | Total installment high credit/credit limit |

The loan’s purpose could also be considered as a good predictor variable. However, since it is an unordered categorical variable, we performed the Chi square test and got the chi squared value of 290 along with a much less p-value, concluding that the two variables (loan status and purpose) are in fact dependent.

**PART B**

**6. Develop decision tree models to predict default.**

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

**Ans:** To develop a predictive model, it is highly important to split the data into appropriate proportions in order to get training and validation data sets. We split the data so that we can use the Training data to develop the model and the Validation data set to implement the model and evaluate its performance on unseen data. We have divided the data into half i.e, 50% as training data and 50% as test data.

**b) Train decision tree models (use both rpart, c50) Remember - if the model performance looks “too” good, it may be due to leakage – make sure you check to ensure that none of the variables used in modeling have leakage problems. Look at variable importance in the models – any leakage causing variables will typically be among the most important. In building decision tree models, what parameters do you experiment with, and what performance do you obtain (on training and validation sets)? Clearly tabulate performance for different parameter settings, and briefly describe your findings. 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.**

**Model 1:**

We have built a decision tree with minsplit as 30 and cp=0.0001 along with 50% of the training data as the input for this model.

The Root Node Error we get is : 7691/54963 = 0.14.

*Confusion matrix and evaluation measure: (50%* ***Training data****; minsplit = 30)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 45012 | 6947 |
| Charged Off | 2266 | 738 |

Accuracy : 0.832

95% CI : (0.829, 0.835)

No Information Rate : 0.86

P-Value [Acc > NIR] : 1

Sensitivity : 0.952

Specificity : 0.096

*Confusion matrix and evaluation measure: (50%* ***Pruned Training data****; minsplit = 30; cp = 0.00015)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 46700 | 6296 |
| Charged Off | 572 | 1395 |

Accuracy : 0.875

95% CI : (0.872, 0.878)

No Information Rate : 0.86

P-Value [Acc > NIR] : <2e-16

Sensitivity : 0.988

Specificity : 0.181

*Confusion matrix and evaluation measure: (50%* ***Test data****; minsplit = 30)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 45012 | 6947 |
| Charged Off | 2266 | 738 |

Accuracy : 0.832

95% CI : (0.829, 0.835)

No Information Rate : 0.86

P-Value [Acc > NIR] : 1

Sensitivity : 0.952

Specificity : 0.096

*Confusion matrix and evaluation measure: (50%* ***Pruned Test data****; minsplit = 30; cp = 0.00015)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 45806 | 7130 |
| Charged Off | 1472 | 555 |

Accuracy : 0.843

95% CI : (0.84, 0.847)

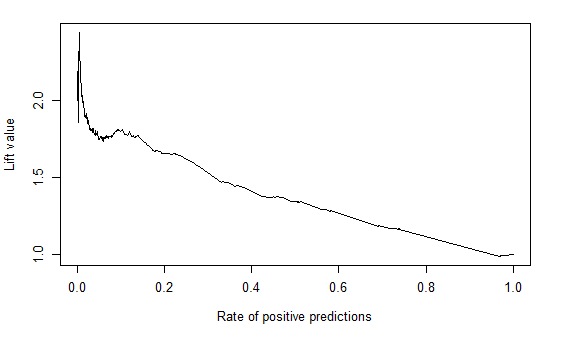
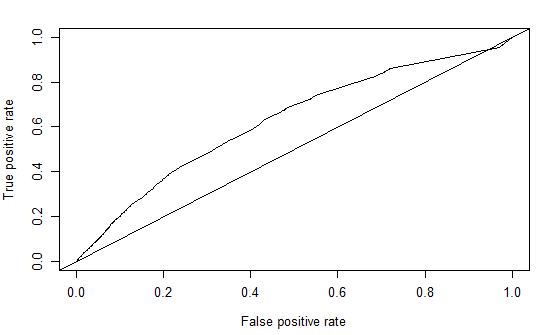
No Information Rate : 0.86

P-Value [Acc > NIR] : 1

Sensitivity : 0.9689

Specificity : 0.0722

From the above measures we can see that the performance of the decision tree is improved after pruning the data in terms of Accuracy, Sensitivity and Specificity of the model in both training data as well as test data.



**Fig:** Left: ROC curve for Model1 (y values = 0.6269) Right: Lift Curve for Model1

**Model 2:**

We have built a decision tree with minsplit as 70 and cp=0.0001 along with 50% of the training data as the input for this model.

The Root Node Error we get is : 7691/54963 = 0.14.

*Confusion matrix and evaluation measure: (50%* ***Training data****; minsplit = 70)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 46777 | 6921 |
| Charged Off | 495 | 770 |

Accuracy : 0.865

95% CI : (0.862, 0.868)

No Information Rate : 0.86

P-Value [Acc > NIR] : 0.000349

Sensitivity : 0.990

Specificity : 0.100

*Confusion matrix and evaluation measure: (50%* ***Pruned******Training data****; minsplit = 70; cp = 0.00015)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 46814 | 6976 |
| Charged Off | 458 | 715 |

Accuracy : 0.865

95% CI : (0.862, 0.868)

No Information Rate : 0.86

P-Value [Acc > NIR] : 0.00077

Sensitivity : 0.990

Specificity : 0.093

*Confusion matrix and evaluation measure: (50%* ***Test data****; minsplit = 70)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 46374 | 7312 |
| Charged Off | 904 | 373 |

Accuracy : 0.851

95% CI : (0.848, 0.853)

No Information Rate : 0.86

P-Value [Acc > NIR] : 1

Sensitivity : 0.9809

Specificity : 0.0485

*Confusion matrix and evaluation measure: (50%* ***Pruned******Test data****; minsplit = 70; cp = 0.00015)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 46436 | 7331 |
| Charged Off | 842 | 354 |

Accuracy : 0.851

95% CI : (0.848, 0.854)

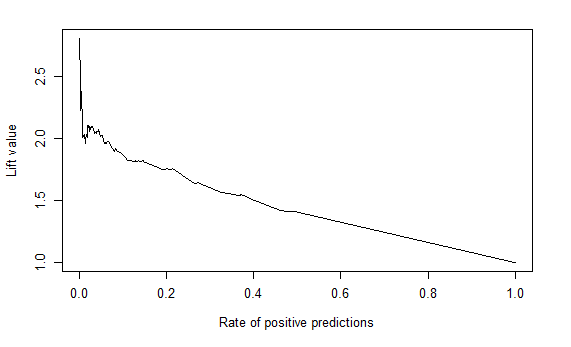
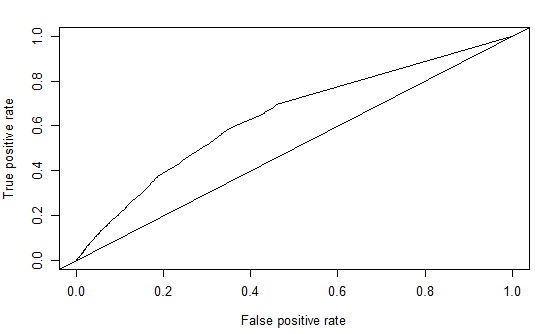
No Information Rate : 0.86

P-Value [Acc > NIR] : 1

Sensitivity : 0.9822

Specificity : 0.0461

When we compare the decision tree with its pruned version, there is not much difference in the evaluation measures.



**Fig:** Left: ROC curve for Model2(y value = 0.6419) Right: Lift curve for Model2

**C5.0 :**

When we look at the distribution of Fully paid and Charged off loans, we see that there are approximately 6 times more Fully paid loans than the Charged off loans and therefore to balance the data for the decision tree we have used weights as 6 for Charged off loans and 1 for Fully Paid loans.

*Confusion matrix and evaluation measure: (50%* ***Training data****; minCases = 30)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 30119 | 1518 |
| Charged Off | 17153 | 6173 |

Accuracy : 0.66

95% CI : (0.656, 0.664)

No Information Rate : 0.86

P-Value [Acc > NIR] : 1

Sensitivity : 0.637

Specificity : 0.803

*Confusion matrix and evaluation measure: (50%* ***Test data****; minsplit = 30)*

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | Fully Paid | Charged Off |
| Fully Paid | 28342 | 3306 |
| Charged Off | 18936 | 4379 |

Accuracy : 0.595

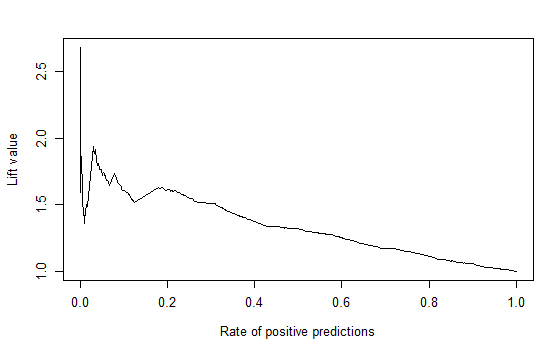
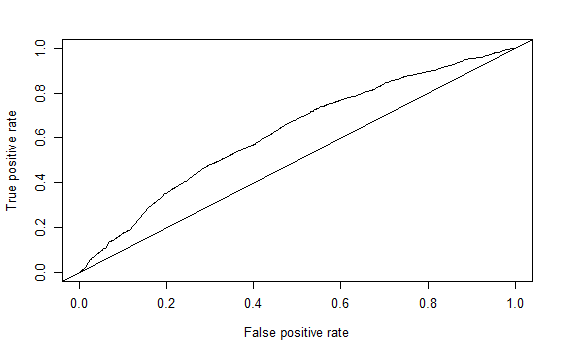
95% CI : (0.591, 0.599)

No Information Rate : 0.86

P-Value [Acc > NIR] : 1

Sensitivity : 0.599

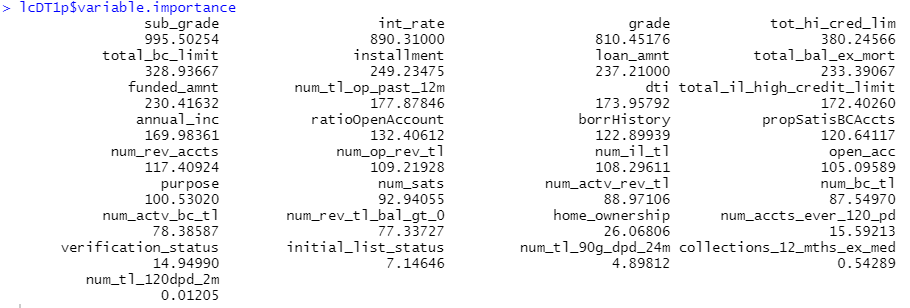
Specificity : 0.570



**Fig:** Left: ROC curve for C5.0 (y value = 0.6238) Right: Lift curve for C5.0

**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 6 Actual 5 above? Briefly describe how variable importance is obtained (the process used in your best decision tree – note that the approach is not the same for rpart and C50).**

**Ans.**  According to above analysis of the different models and their evaluation measures, we can see that Model1 with minimum split as 30 and cp=0.0001 is better than others in terms of Accuracy, Sensitivity and Specificity. It performs well when the tree is pruned on both the training as well as the test data. There is no overfitting of data and the model seems to work well with on the test data with better accuracy. The variable importance given by the model is as shown below:



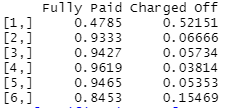
Loan sub grade seems to be the most important variable, followed by interest rate and then grade. This is in sync with our previous analysis of the effects of these variables on the loan status. Hence, by all means, the Model 1 (minsplit = 30) is better than the others.

**7. 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 the previous question. Do you find the importance of variables to be similar/different? Which model would you prefer, and why?**

**Ans:**

Random forest is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It **aggregates the votes from different decision trees** to decide the final class of the test object. In the given dataset, after performing decision trees, we are estimating the performance of data using random forest to find which is better. Ranger function is used here to perform random forest in the given dataset as it is well known for its faster computations. The random forest analyses are done here using two models rfmodel1 and rfmodel2. In both the models the numbers of trees are changed. The rfmodel1 has 200 trees defined while the rfmodel2 has 500 trees defined.

In **rfmodel1** performance evaluation we have looked for the parameters like accuracy, precision, specificity, ROC and AUC curve. The predicted scores on the training data:



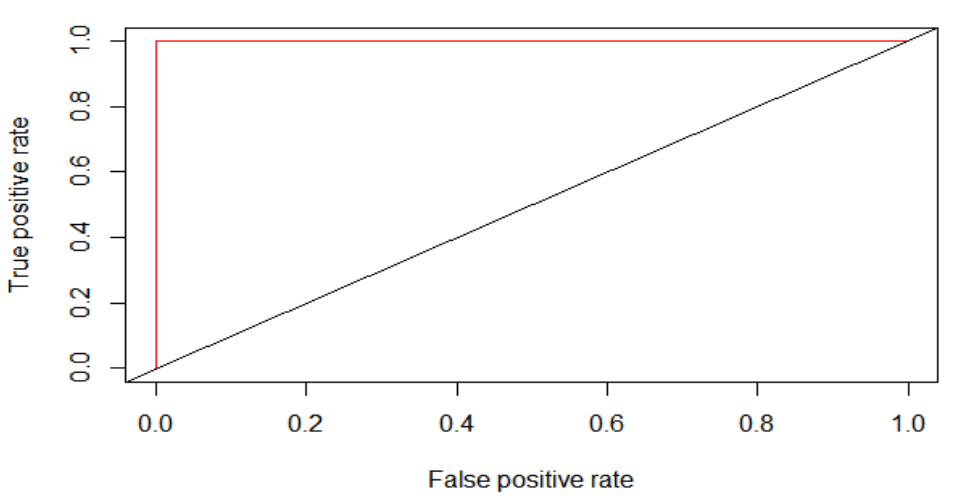
With specific threshold = 0.7, for “Fully paid” loans the classification performance in the training data and test data are as follows:

Training data: Test data:

Using confusion matrix, Accuracy = 0.9999 Accuracy = 0.8318

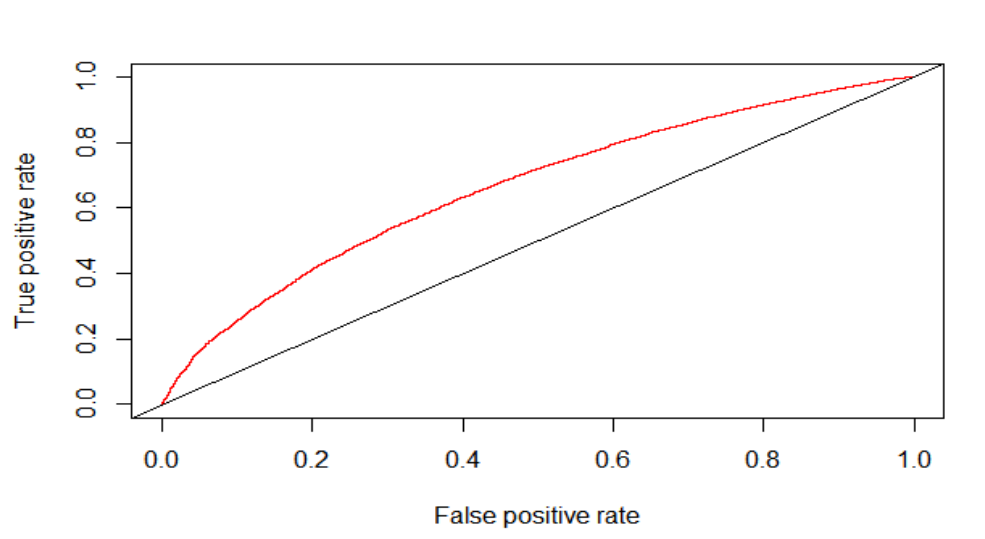
ROC curve on training data:



**Fig. 7.1:** ROC curve on training data for rfModel1

The above ROC curve seems to be overfit as it looks so ideal in case of training data. Though the fit might look excellent but will produce high complexity in case of test data. The AUC value for training data comes to 0.660912

ROC curve on test data:

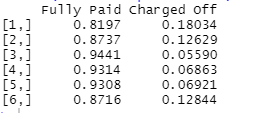


**Fig. 7.2:** ROC curve on test data for rfModel1

The above graph shows ROC curve for testing data. Here the AUC value comes to 0.661213.

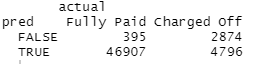
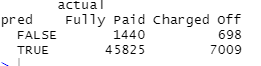
From the above metrics it looks like the rfModel1 is an overfit and therefore, we need to check for another model by changing the parameters, that could give us better results as compared to rfmodel1.

In **rfmodel2** performance evaluation, we have increased the number of trees = 500 and added two new parameters viz minimum node size which is user defined = 50 and maximum depth of the tree = 15. The predicted scores on the training data:



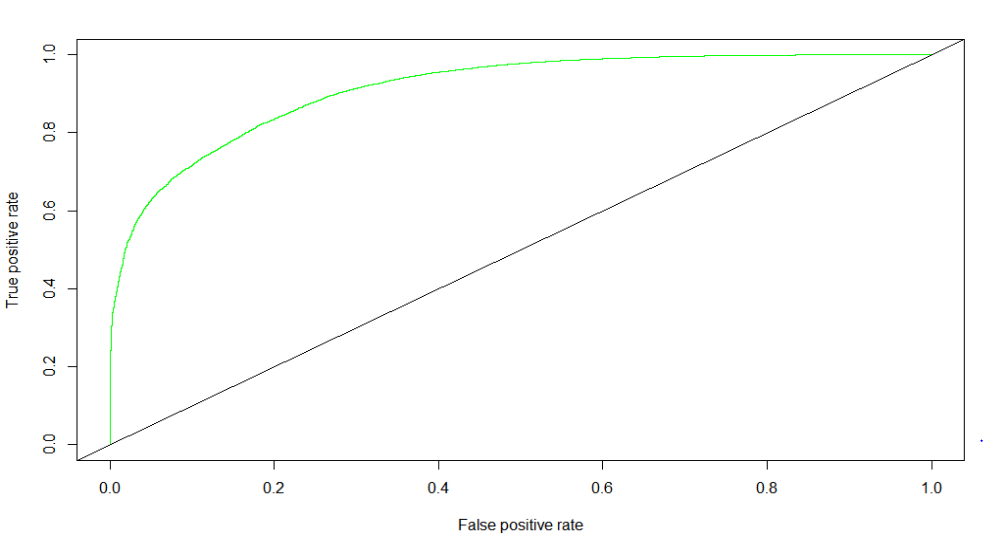
Here we have kept the same specific threshold = 0.7, the classification performance in the training data and test data are as follows:

Training data Test data

The accuracies still look good as compared to rfmodel1 but we need to check for other parameters as well like ROC and AUC before coming to the conclusion.

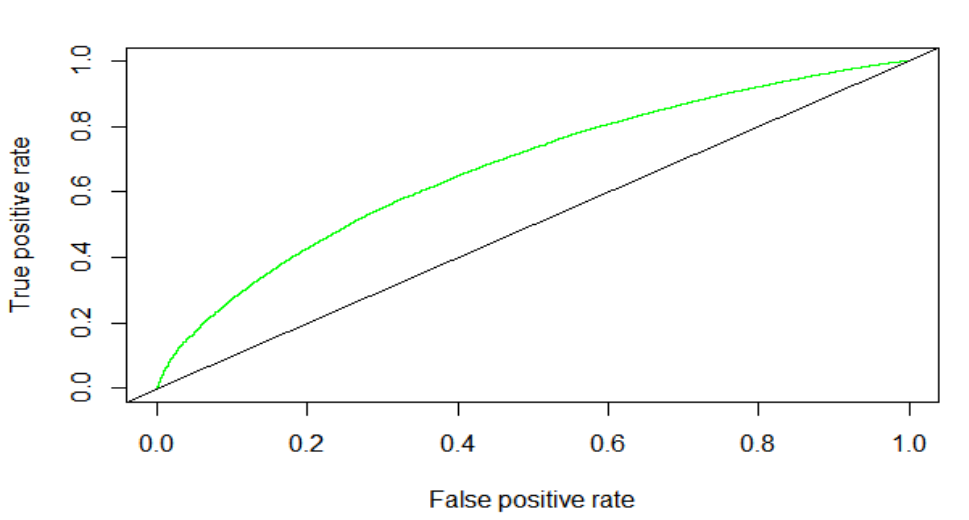
ROC curve on training data:



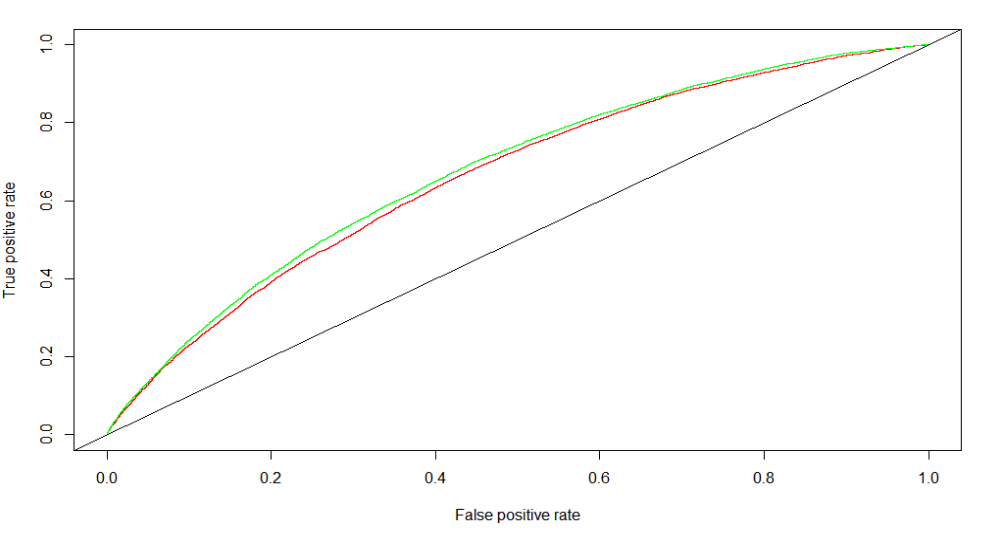
**Fig. 7.3:** ROC curve on training data for rfModel2

The above ROC curve seems to be better as compared to rfmodel1. It looks good to be implemented on any dataset. The AUC value comes to 0.912884.

ROC Curve on test data:



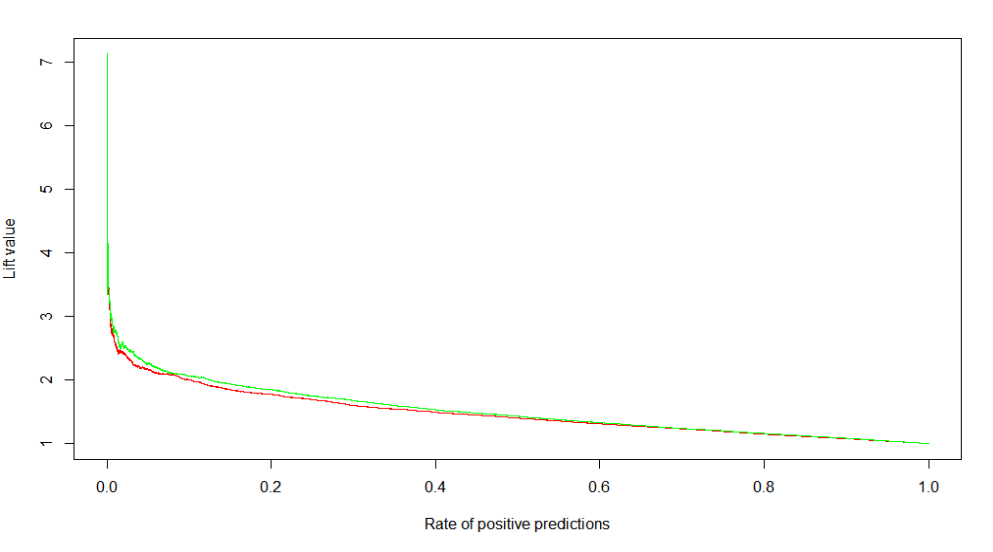
**Fig. 7.4:** ROC curve on test data for rfModel2



**Fig. 7.5:** Comparing ROC Curves for both rfModel1(red) and rfModel2(green) on unseen data

The graph shows that model2(green) is slightly higher than model1(red).

The Lift curve shows the curves for analyzing the proportion of true positive data instances in relation to the classifier's threshold or the number of instances that we classify as positive.

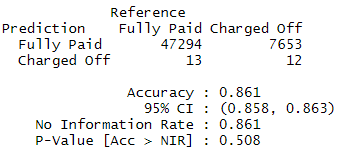
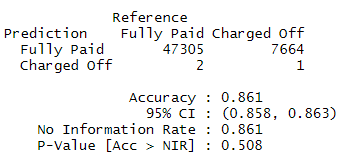


**Fig 7.6:** Lift curves for rfModel1(red) and rfModel2(green) on unseen data

For model2(green) the lift curve looks better than for model1(red). Therefore, the rate of positive predictors is better in case of model2.

Below is the summary of the confusion matrix of both the models:

Confusion matrix for rfModel1 with 200 trees Confusion matrix for rfModel2 with 500 trees

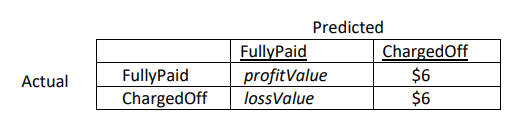
Thus comparing both the models of random forest i.e rfmodel1 and rfmodel2, it can be concluded that on the basis of accuracy and ROC, AUC parameters, rfmodel1 though looks excellent and better than rfmodel2 but rfmodel1 is prone to overftting which is not, in case of rfmodel2. Rfmodel1 can create higher complex relationships in test data or unseen data which is technically of not much use. So rfmodel2 is the suitable model for the given dataset.

**8) 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, suppose the average int\_rate in the data is 12%; so after 3 years, the $100 will be worth (100 + 3\*12) = 136, i.e a profit of $36. Now, is 12% 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 should 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:**



**Ans:**

The average interest rate for fully paid loans comes to 11.75. Therefore, we can take the profit value on Fully Paid loans = (100+3\*11.8) = 135.4. So, PROFIT VALUE = 35.

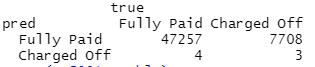
The average interest rate for charged off loans comes to 13.86. Therefore, we can take the profit value on Charged off loans = (100+3\*13.86) = 141.58. So, LOSS VALUE = 42.

Confusion matrix for profit/loss amounts:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | ***Predicted*** | |
|  |  | ***FULLY PAID*** | ***CHARGED OFF*** |
| ***Actual*** | ***FULLY PAID*** | 35 | $6 |
| ***CHARGED OFF*** | -42 | $6 |

Calculating profit from the decision tree model and random forest model:

For Decision Tree (pruned DT model 1): For Random Forest(rfmodel2):

Profit: 1314211 Profit: 1330301

The profit obtained from decision tree is 1314211 whereas the profit obtained from random forest is 1330301. This shows that profit obtained from random forest model(rfmodel2) is more as compared to decision tree and hence random forest model should be preferred over the decision tree.

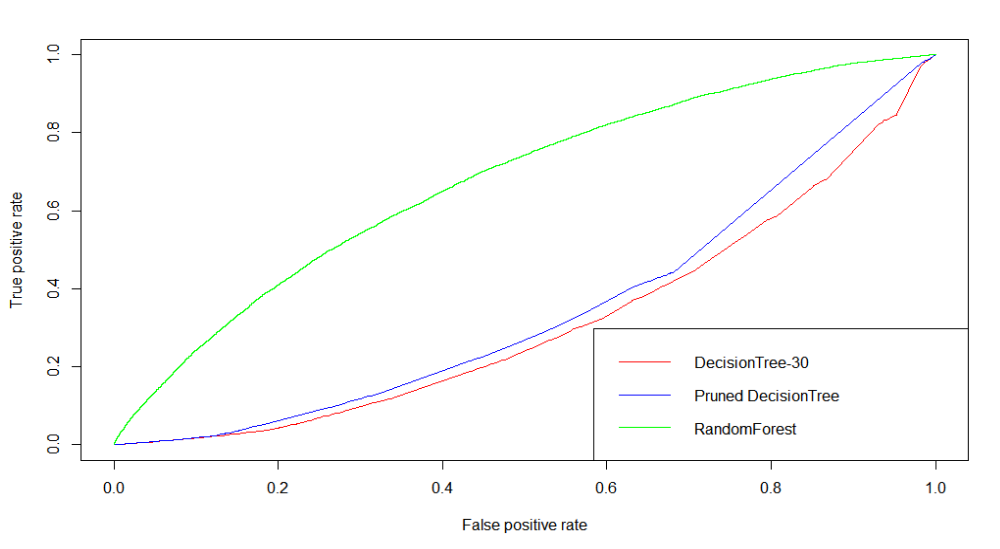
1. **Compare the performance of your models from Questions 6, 7 above based on this. Note that the confusion matrix depends on the classification threshold/cutoff you use. Evaluate different thresholds and analyze performance. Which model do you think will be best, and why?**

**Ans:**

Decision trees are much easier to interpret and understand. Since a random forest combines multiple decision trees, it becomes more difficult to interpret. Random Forest is suitable for situations when we have a large dataset, and interpretability is not a major concern. It depends on one’s requirements. If one has time constraints, on a model, better to go with a decision tree. However, stability and reliable predictions are in the basket of random forests. In an ideal world, where we need to reduce both bias-related and variance-related errors. This issue is well-addressed by random forests. They are so powerful because of their capability to reduce overfitting without massively increasing error due to bias, and hence produce usable results.

In the lending club dataset, on performing decision tree model(pruned DT model 1) we got an accuracy of 0.843 on the test data, whereas on performing random forest model (rfmodel2) we got an accuracy of 0.86 which is slightly more as compared to decision tree model and hence this is one of the reasons to conclude the fact that performance of random forest model with number of trees = 500 would be better for this dataset.

Below graph shows the ROC Curve for decision tree, pruned decision tree and random forest, the random forest curve looks good as compared to the other two curves.



**Fig. 8.1:** ROC curve for both models of DT and rfModel2

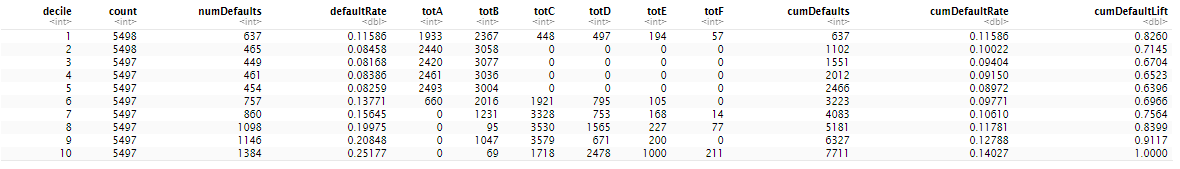
1. **Another approach to determining the optimal threshold for implementing the model is to directly consider how the model will be used – you can order the loans in descending order of prob(fully-paid). Then, you can consider starting with the loans which are most likely to be fully-paid and go down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

**Ans:**

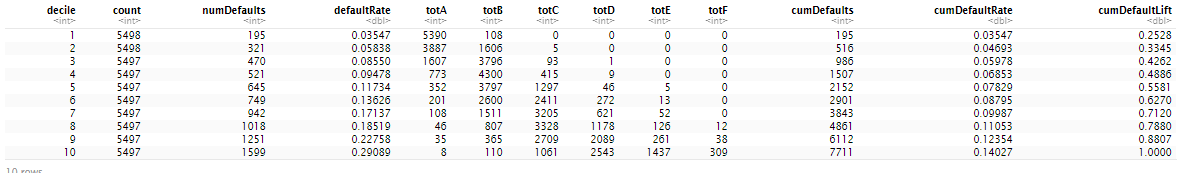
Performance with profit- loss for Decision tree, the maximum profit obtained from decision tree model is 1330280.

Performance with profit- loss for random forest, the maximum profit obtained from Random forest model is 1330476.

**Decile lift performance on defaults– Decision tree model (pruned DT model 1 test data)**

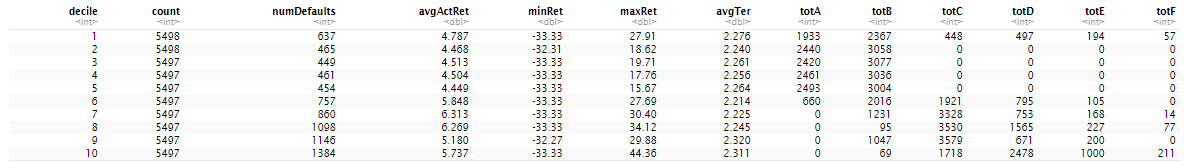


**Decile lift performance on defaults – random forest model (rfmodel2 test data)**

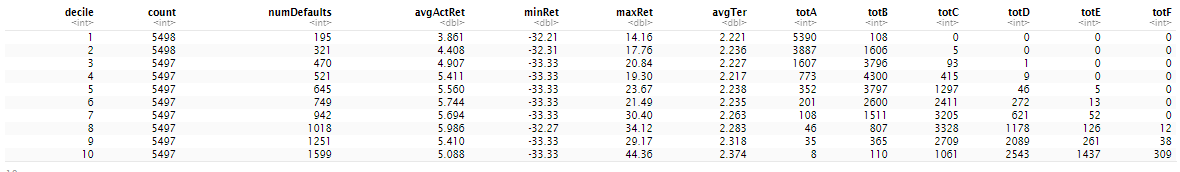


Here since we are finding the performance of Fully paid loans, the number of defaults column in both the models has the lowest number which means here is lowest probability of defaults in the top deciles for both the cases. However, in case of random forest model the number of defaults is less than observed in decision tree model.

**Returns Performance by decile – Decision tree (pruned DT model 1 test data)**



**Returns Performance by decile – Random Forest (rfmodel2 test data)**

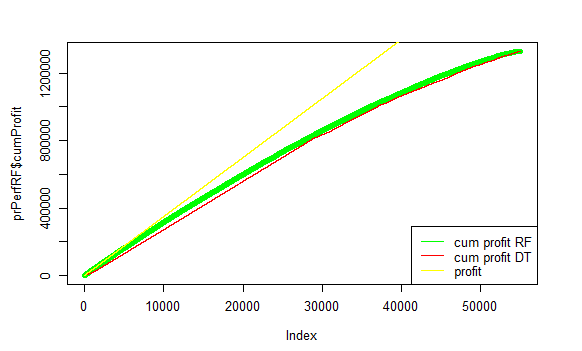


In returns performance by decile, when we look at the returns from the Fully paid loans in both the model using the decile lifts, the top decile has the greatest number of fully paid loans and hence the number of defaults is less. In random forest model, the average actual return is not very high in the top decile. It is not very high since most of these loans belong to the loan grades A and B and as we already saw in our previous analysis, these have the least interest rates and are also paid back early. The average actual return is highest in the 8th decile. It is the same even in the case of the decision tree model, hence this can be considered as the threshold.

In top deciles the random forest models have a greater number of loans from grade A and B. Therefore, the model uses these grades for prediction of status. This is true because even the variable importance of grades is also high for these models. So, if we want to invest in the Fully paid loans from the top decile since there are less chances of the defaults as compared to the rest of the deciles. However, the average actual return of these is low and this can be because of the reason that they are paid back early. Also, here we can see that as we go down the deciles, the avg return goes up until the 8th decile.

Therefore, random forest is a better model. Thus Accuracy, ROC and AUC parameter and cost base performance wise, it can be concluded that random forest model is the best suitable model for this given lending club dataset.

Below is the comparison of our above models with that of investing in the safe CD’s. It is evident that after around 40000 indices, the slope of the cumulative profit tends to flatten and therefore we can conclude that we can invest in loans greater than this index value.



**Fig. 8.2.1:** Comparison of different models to investing in safe CD’s