LendingClub

LOAN DEFAULT PREDICTION AND INVESTMENT STRATEGIES IN ONLINE LENDING

Done By
Pritha Ghosh
Tejaswi Cherukuri
Anoop Gopalam

IDS - 572 | Data Mining for Business-Assignment 1
Loan Default prediction and Investment Strategies in Online Lending

1) **Describe the business model for online lending platforms like Lending Club?**

The Lending Club operates on a standard Peer-to-Peer Lending business model, where the individual investors and borrowers are connected through an online lending platform with no third party involvement.

- **Lenders** (private individuals and/or institutional investors) invest excessive cash flow in loans on the platform and receives principal and interest in return.
- **The borrower** (a consumer or business) receives financing and pays interest on the loan amount in return.
- **The platform** is handling administration and attracts both borrowers and investors.

The below steps illustrates how P2P business models work:
STEP 1: Loan Picking and Loan Application
STEP 2: Negotiating the Loan Agreement
STEP 3: Transferring the Loan principal
STEP 4: Receiving the Loan Principal (minus any platform fees)
STEP 5: Repayment of Principal and Interests (including any platform fees)

Consider the stakeholders and their roles, and what advantages Lending Club offers.

- The primary stakeholders of LC are borrowers and lenders. The stakeholders behind Lending Club are most notably the investors, but also include the banking industry, consumers, national markets, and extends to employees, their families, and surrounding communities.
- Lending Club also has stakeholders within advertising marketing strategies by partnering with widely known and used organizations like Facebook and Credit Karma.
- The risk of investors is placed at the borrower, which means that if the borrower does not pay back the borrowed amount the investor might lose capital.

Advantages of lending club:

- They offer Business loans, Personal loans and auto refinance loans.
- Lending Club offers personal loans starting at $1,000 for borrowers who don't need a large sum.
- They offer 36 and 60 months term loans.
- Low, fixed rates and fixed monthly payments.
- No prepayment fees or penalties : LC won't charge a fee for making extra payments, even if you decide to pay off your loan early.
- Automatic payment withdrawals.
- Borrowers using Lending Club can take advantage of consolidating debt and paying off high interest credit cards with an average of 24 percent lower rates than traditional loans.
- Lending Club allows co-borrowers on its personal loans, which may be helpful if you believe you need an additional applicant with better finances to help you qualify.

What is the attraction for investors?

- Lending Club only accepts applications from borrowers with excellent credit(600+), positive returns for investors are virtually guaranteed.
- Till date, over $367 million in interest has been paid out to investors.
- Since facilitating its first loan in 2007, the company has doubled loan volume each year, investors earn attractive risk-adjusted returns.

How does the platform make money?

- Lending club collects fees from both borrowers and investors. It makes money through origination and service fees. Borrowers pay a one-time origination fee of 1.11% to 5% of the total loan amount, depending on the loan grade and term.
- Meanwhile, investors pay a service fee of 1% of the amount of any borrower payments received by the payment due date or during applicable grace periods, and a collection fee of up to 40% on all amounts collected on a delinquent loan to the extent any litigation has been initiated against the borrower, or up to 30% on all amounts collected on a delinquent loan in all cases not involving litigation.

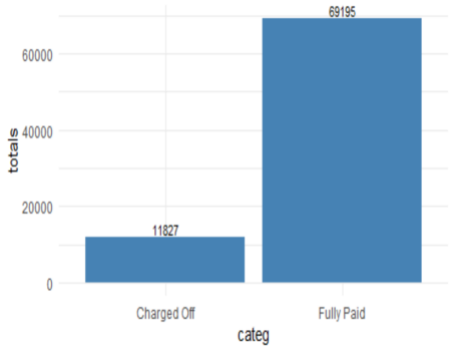
REFERENCES:

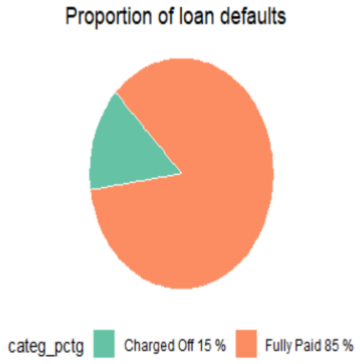
Reference 1 : - Lending Club Website
Reference 2:- P2P Lending Explained: Business Models, Definitions & Statistics
Reference 3:- Lending Club's Personal Loan Info Sheet
Reference 4:- Business Ethics Case Analyses on Lending Club

2a)

(i) What is the proportion of defaults ('charged off' vs 'fully paid' loans) in the data?

The number of charged off loans are 11827 and the number of fully paid loans are 69195.



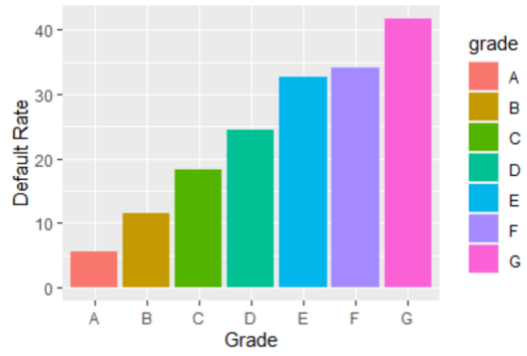


The pie chart depicts that the percentage of default/charged off loans is **15%.**

How does default rate vary with loan grade?

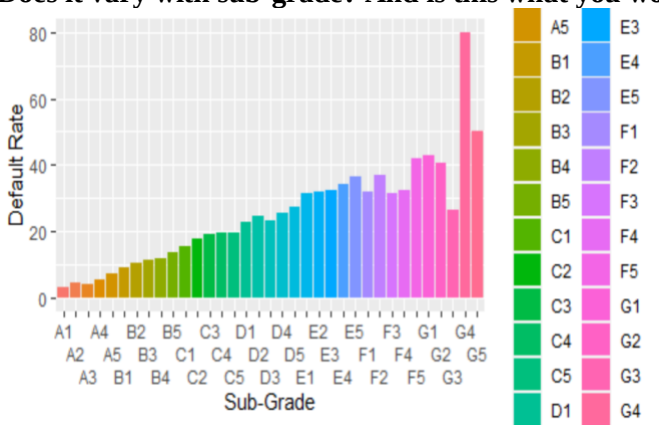
The table below depicts the rate of default for each loan grade.

grade	loan_status	freq	n	rate
A	Charged Off	1108	20402	5.43084
B	Charged Off	2682	23399	11.46203
C	Charged Off	4116	22577	18.23094
D	Charged Off	2647	10802	24.50472
E	Charged Off	1045	3191	32.74835
F	Charged Off	191	560	34.10714
G	Charged Off	38	91	41.75824



We can see from the data that the Default Rate increases as we move from Grade A to Grade G. This is in line with our expectations since the interest rate progressively increases as we move from Grade A to Grade G.

Does it vary with sub-grade? And is this what you would expect, and why?



sub_grade	loan_status	freq	n	rate
F4	Charged Off	25	77	32.467532
F5	Charged Off	23	55	41.818182
G1	Charged Off	18	42	42.857143
G2	Charged Off	11	27	40.740741
G3	Charged Off	4	15	26.666667
G4	Charged Off	4	5	80.000000
G5	Charged Off	1	2	50.000000

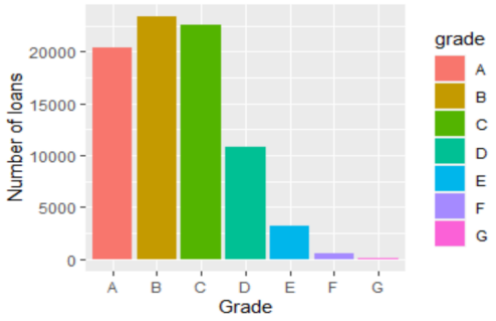
The above graph depicts the rate of default within each sub-grade. As we move A1 to G5, the general trend is that the default rate increases.

The most obvious disruption in the pattern is seen for Grade G, where the rate of default drops for G3, increases abruptly for G4 and then drops again for G5. We can attribute this to the total number of loans being very less in comparison to the other sub-grades.

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

A: **Number of loans by grade :**

grade	n
A	20402
B	23399
C	22577
D	10802
E	3191
F	560
G	91



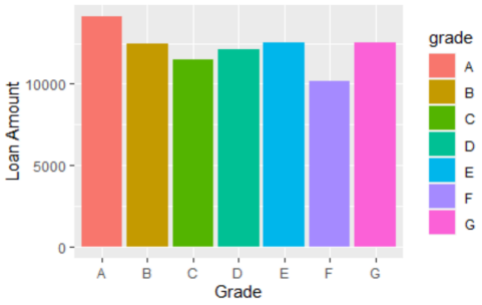
We can see from the data that Grade B has the maximum number of loans and Grade G has the minimum number of loans.

And do loan amounts vary by grade?

Loan amounts by grade:

The table below depicts the mean loan amounts by grade.

grade	loan_amount
A	14145.93
B	12458.19
C	11465.52
D	12150.06
E	12558.07
F	10168.62
G	12509.07

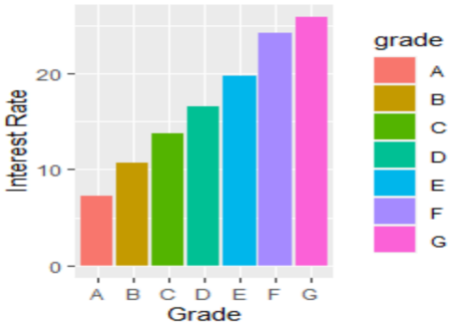


As depicted by the graph above, the mean loan amount is highest for Grade A and lowest for Grade F.

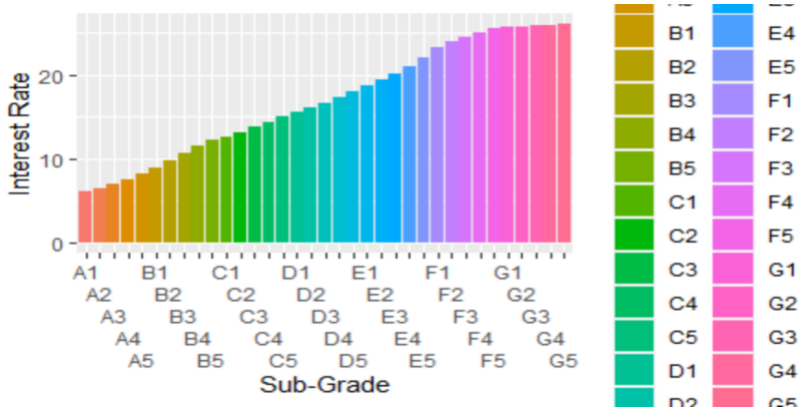
Does interest rate for loans vary with grade, subgrade? Look at the average, standard-deviation, min and max of interest rate by grade and subgrade. Is this what you expect, and why?

Mean interest rate by grade:

grade	int_rate
A	7.249607
B	10.711802
C	13.672434
D	16.531698
E	19.769035
F	24.119196
G	25.839890



The mean interest rate increases as we move up from Grade A to Grade G. This is as expected because the loans that are of higher risk of default are issued at higher interest rates. This conclusion is corroborated by our findings in the previous question, where we saw that as we move up from Grade A to Grade G, the rate of default on the loans increases.



The above graph depicts the interest rate by sub-grade. As we move from sub-grade A1 to G5, the average rate of interest increases. This is as expected as the risk attached to sub-grade A1 is considerably lower than that attached to sub-grade G5.

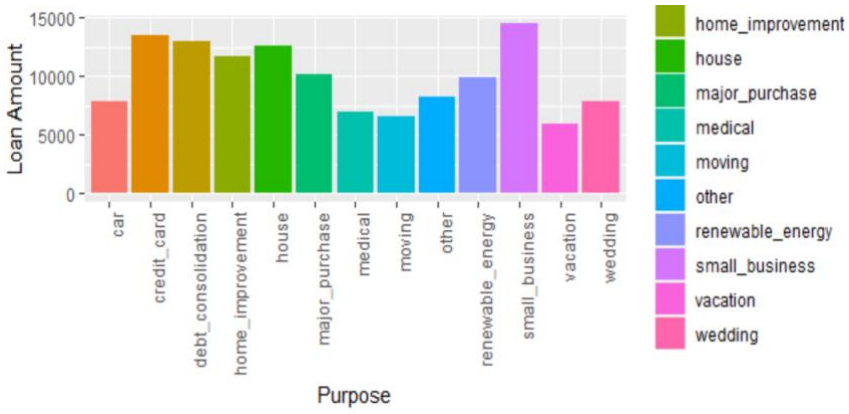
(iii)What are people borrowing money for (purpose)?

purpose	number_of_loans
car	719
credit_card	18780
debt_consolidation	48647
home_improvement	3942
house	254
major_purchase	1402
medical	900
moving	604
other	4455
renewable_energy	65
small_business	759
vacation	492
wedding	3

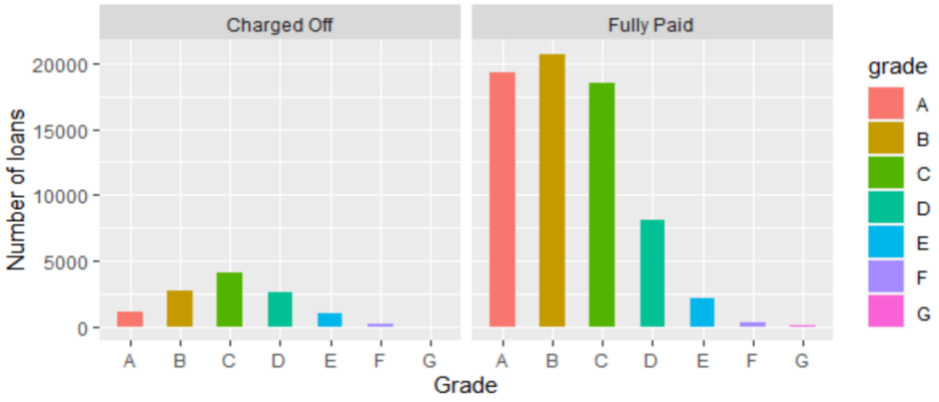
The purpose of borrowing loans ranges from a wide variety of reasons like Debt Consolidation, Credit Card, Home Improvement etc. There are 13 categories that are mentioned in the table above into which they can be grouped.

Examine how many loans, average amounts, etc. by purpose? And within grade? Do defaults vary by purpose?

purpose	avg_amt
small_business	14424.835
credit_card	13501.040
debt_consolidation	13007.644
house	12505.118
home_improvement	11706.925
major_purchase	10172.183
renewable_energy	9828.846
other	8270.831
wedding	7866.667
car	7819.645
medical	6980.722
moving	6542.384
vacation	5872.104

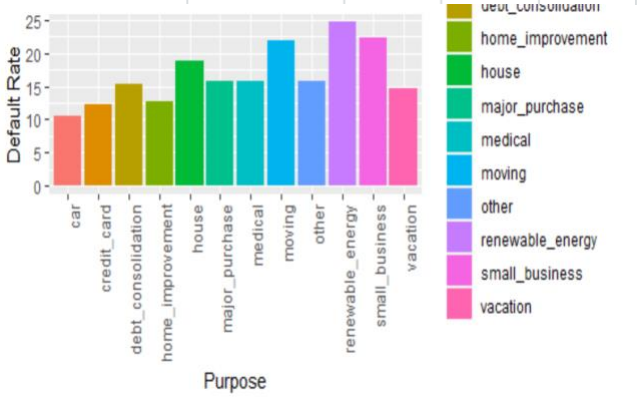


The maximum amount of loans are borrowed for Small Businesses, Credit Cards, Debt Consolidations and House categories.



For defaulted loans, the count is highest for Grade C and lowest for Grade G. On the other hand, for Fully Paid loans, the loan count is highest for Grade B and lowest for Grade G.

purpose	loan_status	freq	total	rate
renewable_energy	Charged Off	16	65	24.61538
small_business	Charged Off	169	759	22.26614
moving	Charged Off	132	604	21.85430
house	Charged Off	48	254	18.89764
other	Charged Off	702	4455	15.75758
major_purchase	Charged Off	220	1402	15.69187
medical	Charged Off	141	900	15.66667
debt_consolidation	Charged Off	7423	48647	15.25891
vacation	Charged Off	72	492	14.63415
home_improvement	Charged Off	503	3942	12.76002
credit_card	Charged Off	2326	18780	12.38552
car	Charged Off	75	719	10.43115



The default rates are highest for Renewable Energy, Small Businesses and Moving.

purpose	grade	Count	AvgLoanAmt	MedianLoanAmt	DefaultRate
house	A	8	20075.000	24500.0	25.000000
small_business	A	26	16153.846	15000.0	7.692308
major_purchase	A	361	10576.662	8000.0	6.094183
credit_card	A	7487	14682.573	12600.0	5.756645
other	A	338	10755.843	9550.0	5.621302
debt_consolidation	A	10824	14293.554	12000.0	5.358463
vacation	A	38	6747.368	5400.0	5.263158
medical	A	73	9565.753	7000.0	4.109589
home_improvement	A	1043	12712.464	10000.0	4.026846
car	A	194	8364.433	6550.0	2.577320
moving	A	8	10262.500	12500.0	0.000000
renewable_energy	A	2	5250.000	5250.0	0.000000

For Grade A, the highest default is for House and the lowest is for Moving and Renewable Energy.

purpose	grade	Count	AvgLoanAmt	MedianLoanAmt	DefaultRate
renewable_energy	B	6	15700.000	13100.0	16.666667
moving	B	63	7230.952	6000.0	14.285714
medical	B	199	6850.126	6000.0	13.065327
major_purchase	B	381	9174.344	7200.0	12.860892
credit_card	B	6105	13042.408	10450.0	12.137592
small_business	B	66	13750.758	12000.0	12.121212
debt_consolidation	B	14124	12893.451	11000.0	11.448598
home_improvement	B	1130	11113.119	8750.0	10.000000
other	B	980	8449.286	6500.0	9.285714
car	B	209	7449.282	6000.0	8.612440
vacation	B	106	6108.019	5000.0	8.490566
house	B	29	14268.966	12000.0	0.000000
wedding	B	1	6000.000	6000.0	0.000000

For Grade B, the highest default rate is for Renewable Energy and lowest for House and Wedding.

purpose	grade	Count	AvgLoanAmt	MedianLoanAmt	DefaultRate
credit_card	C	3779	12367.531	10000.0	20.031754
debt_consolidation	C	14419	12111.705	10000.0	18.378528
home_improvement	C	1059	10659.278	8000.0	18.035883
moving	C	240	7259.583	6000.0	17.500000
medical	C	371	6442.318	5000.0	17.250674
small_business	C	218	12921.789	10000.0	16.055046
other	C	1633	7431.996	5950.0	15.615432
major_purchase	C	377	9725.398	7200.0	15.384615
house	C	69	11396.377	10000.0	14.492754
vacation	C	196	4998.852	4500.0	13.775510
car	C	195	7662.821	6075.0	12.820513
renewable_energy	C	19	8584.211	6000.0	10.526316
wedding	C	2	8800.000	8800.0	0.000000

For Grade C, the highest default rate is for Credit Card and the lowest for Wedding.

purpose	grade	Count	AvgLoanAmt	MedianLoanAmt	DefaultRate
renewable_energy	D	24	9656.250	5562.5	33.333333
major_purchase	D	215	11490.116	8000.0	28.837209
credit_card	D	1135	12511.366	9875.0	27.753304
moving	D	205	5529.024	4200.0	25.365854
debt_consolidation	D	6972	13069.679	10100.0	25.172117
small_business	D	248	14098.085	10375.0	24.193548
home_improvement	D	520	12687.452	9875.0	21.730769
other	D	1021	8270.788	6000.0	19.588639
car	D	79	7806.646	6400.0	18.987342
vacation	D	117	6351.068	5000.0	18.803419
house	D	78	12170.513	8950.0	17.948718
medical	D	188	6594.415	5000.0	16.489362

For Grade D, the highest default rate is for Renewable Energy and the lowest for Medical.

purpose	grade	Count	AvgLoanAmt	MedianLoanAmt	DefaultRate
major_purchase	E	59	12369.915	8225.0	40.677966
debt_consolidation	E	1967	13423.144	10500.0	35.332994
small_business	E	142	16795.423	14737.5	33.098592
vacation	E	32	7837.500	6600.0	31.250000
credit_card	E	236	11425.636	9350.0	30.508475
house	E	43	15045.349	11500.0	30.232558
renewable_energy	E	10	10565.000	5125.0	30.000000
moving	E	66	5915.530	5000.0	27.272727
other	E	391	8962.340	7125.0	27.109974
car	E	30	8685.000	7125.0	26.666667
home_improvement	E	161	13704.348	9725.0	22.981366
medical	E	54	9133.333	8250.0	22.222222

For Grade E, the highest default rate is for Major Purchase and the lowest is for Medical.

purpose	grade	Count	AvgLoanAmt	MedianLoanAmt	DefaultRate
major_purchase	F	7	8535.714	8275.0	57.142857
moving	F	20	5456.250	5275.0	55.000000
renewable_energy	F	4	8418.750	9737.5	50.000000
debt_consolidation	F	308	11040.909	9212.5	36.363636
other	F	76	8861.513	6625.0	34.210526
medical	F	14	6326.786	4600.0	28.571429
small_business	F	44	14482.955	12225.0	25.000000
home_improvement	F	26	7087.500	6862.5	23.076923
credit_card	F	31	9408.065	7475.0	22.580645
car	F	9	6216.667	5600.0	22.222222
house	F	18	8251.389	8200.0	22.222222

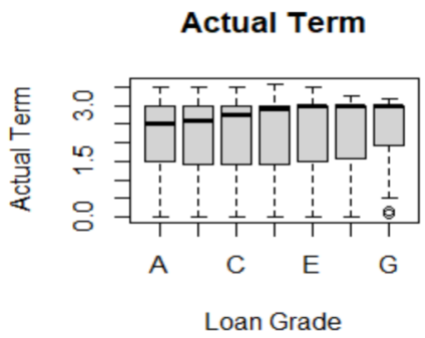
For Grade F, the highest default rate Is for Major Purchase and the lowest default rate is for Car and House.

purpose	grade	Count	AvgLoanAmt	MedianLoanAmt	DefaultRate
house	F	18	8251.389	8200.0	22.222222
medical	G	1	9550.000	9550.0	100.000000
car	G	3	5083.333	5000.0	66.666667
house	G	9	7863.889	7350.0	55.555556
major_purchase	G	2	10687.500	10687.5	50.000000
credit_card	G	7	10257.143	10000.0	42.857143
debt_consolidation	G	33	12059.091	10725.0	42.424242
small_business	G	15	19028.333	21075.0	40.000000
home_improvement	G	3	18483.333	16200.0	33.333333
other	G	16	10756.250	8087.5	31.250000
moving	G	2	19325.000	19325.0	0.000000

For Grade G, the highest default rate is for Medical and the lowest is for Moving.

(iv) 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).

For loans which are fully paid back, we can see that the actual term increases as we move from Grade A to Grade G. This is in line with our expectations, because Grade A is associated with the lowest risk of loan issued and Grade G is has the highest risk.

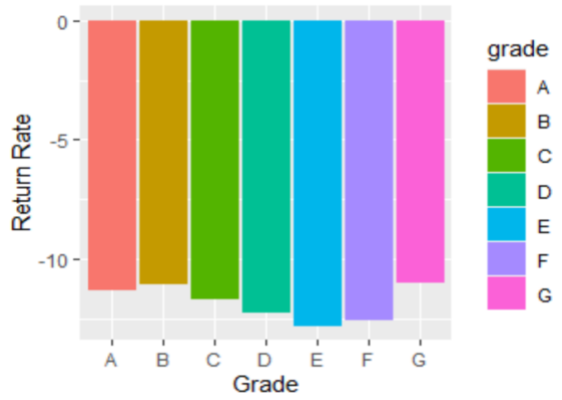


v) **Calculate the annual return. Show how you calculate the percentage annual return. Is there any return from loans which are 'charged off'? Explain. How does return from charged -off loans vary by loan grade? Compare the average return values with the average interest_rate on loans – do you notice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

A: The annual return can be calculated by the equation : ((Total Payment – Funded Amount)/Funded Amount)*100.

```
 loan_status intRate totRet avgActRet
 <chr>          <dbl>  <dbl>     <dbl>
 Charged Off     13.9 -0.351    -11.7
 Fully Paid      11.6  0.149      8.03
```

There is a negative return of -0.351 from loans which are "charged off". This means that for every dollar invested, there is a loss of 0.351 cents.

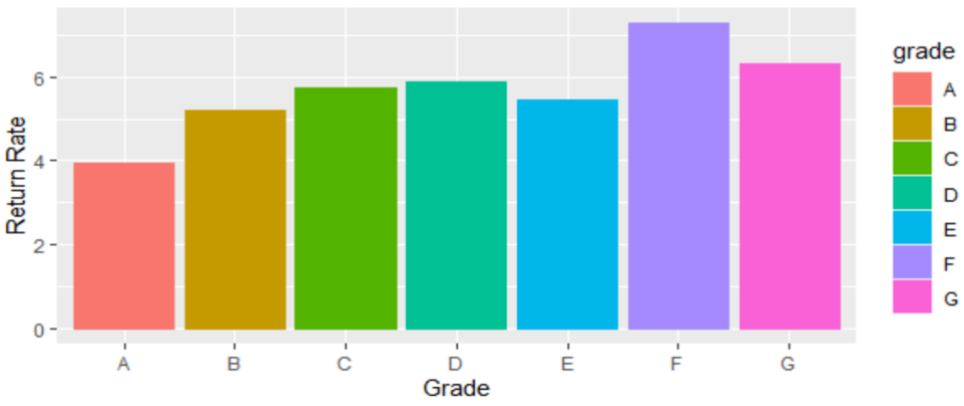


The Return Rate is negative for Charged Off loans across all the Grade types, indicating loss. The highest loss is indicated for Grade E.

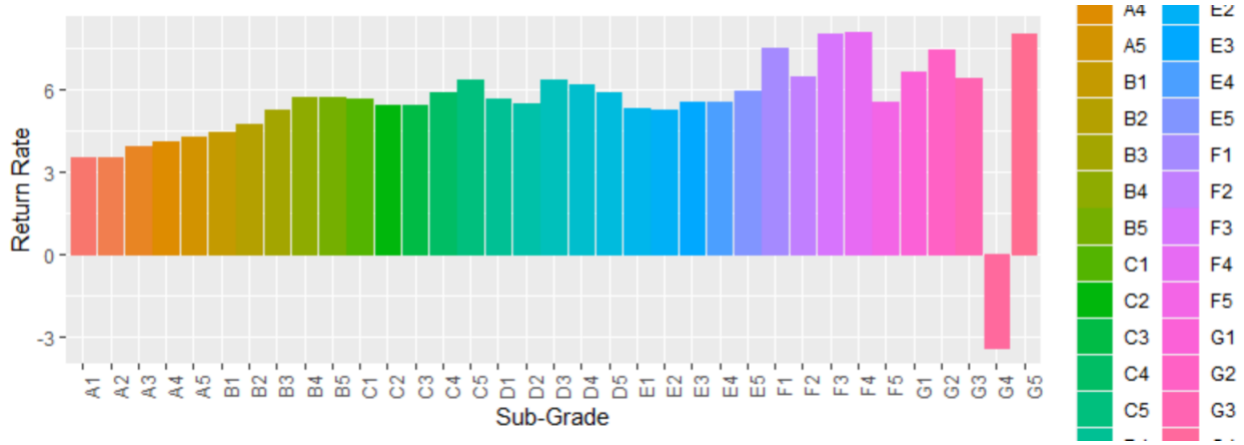
```
 actualReturn       int_rate
Min.   : 0.000   Min.   : 6.00
1st Qu.: 5.433   1st Qu.: 8.19
Median : 7.383   Median :11.67
Mean   : 8.031   Mean   :11.57
3rd Qu.: 9.934   3rd Qu.:13.98
Max.   :44.359   Max.   :26.06
```

We can see that the average interest rates and average rate of returns is quite comparable. However, when the interest rate is high, the rate of return is also high. This can be explained by the high rate of interest itself.

grade	mean(actualReturn)
A	3.936083
B	5.217587
C	5.729864
D	5.887068
E	5.446545
F	7.286130
G	6.332024



The average return rate is highest for Grade F and lowest for Grade A.
We can see that the average return rate increases progressively as we move from Grade A to Grade G, that is, the average return rate is higher for riskier loans.



The average return rate increases as we move from A1 to G5, that is, the average return rate is high with high risk loans.

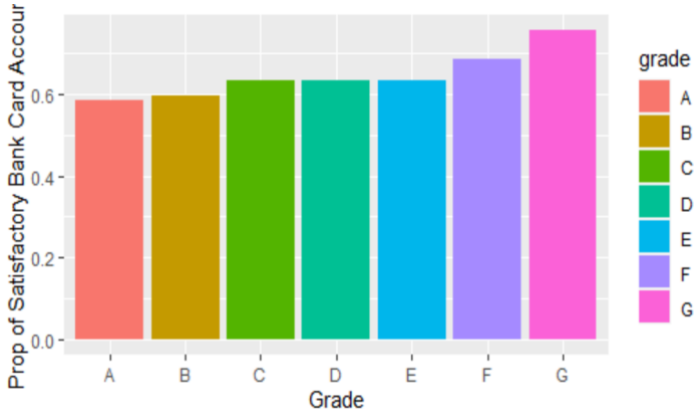
Based on the above analysis, I would invest in loans of Grade F as it would give me the highest returns.

(vi) Generate some (at least 3) new derived attributes which you think may be useful for predicting default., and explain what these are.

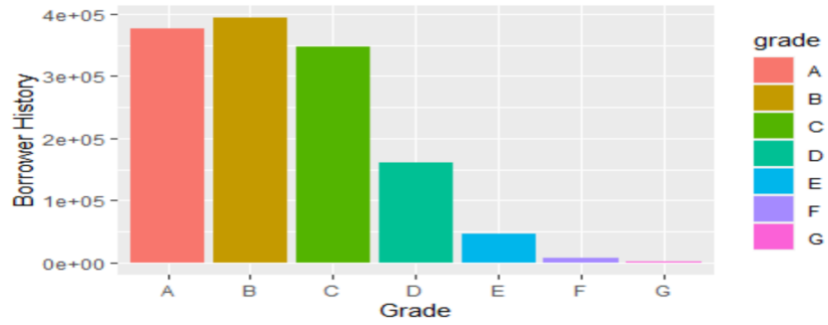
The newly derived attributes that may be useful for predicting default are :

Proportion of Satisfactory Bankcard Accounts : It is the ratio of the number of satisfactory bank card accounts to the total number of bankcard accounts.

grade	mean(propSatisBankcardAccts)
A	0.5843726
B	0.5969342
C	0.6345378
D	0.6338077
E	0.6352577
F	0.6872703
G	0.7572894

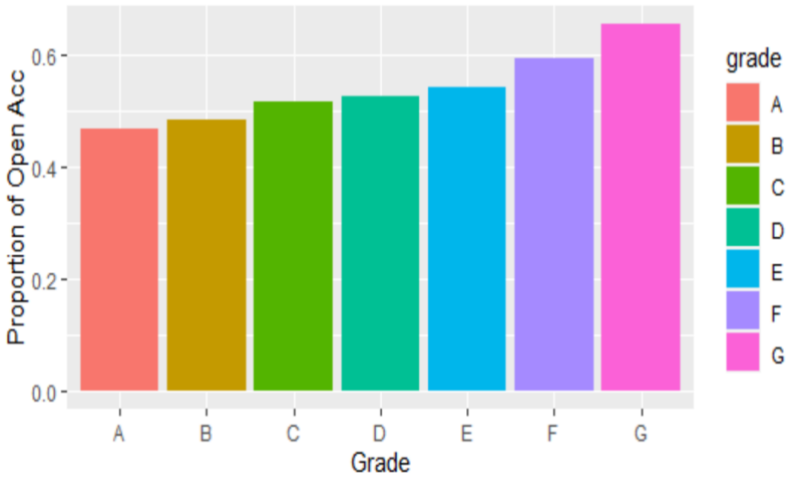


Borrower History : Time since earliest credit line was opened.



Ratio of Open accounts to Total accounts: This is the ratio of the number of open credit lines to the total number of credit lines in the borrower's file.

grade	open_acc_prop
A	0.4671925
B	0.4848766
C	0.5162364
D	0.5270647
E	0.5415535
F	0.5937703
G	0.6557227



2(b)

Are there missing values?

Yes there are 69 variables with missing values and the following table below details all the missing values:

S.No	Name of the variable	S.No	Name of the variable
1	id	36	mths_since_recent_revol_delinq
2	member_id	37	num_tl_120dpd_2m
3	emp_title	38	percent_bc_gt_75
4	url	39	revol_bal_joint
5	desc	40	sec_app_earliest_cr_line
6	mths_since_last_delinq	41	sec_app_inq_last_6mths
7	mths_since_last_record	42	sec_app_mort_acc
8	revol_util	43	sec_app_open_acc
9	last_pymnt_d	44	sec_app_revol_util
10	next_pymnt_d	45	sec_app_open_act_il
11	last_credit_pull_d	46	sec_app_num_rev_accts
12	mths_since_last_major_derog	47	sec_app_chargeoff_within_12_mths
13	annual_inc_joint	48	sec_app_collections_12_mths_ex_med
14	dti_joint	49	sec_app_mths_since_last_major_derog
15	verification_status_joint	50	hardship_type
16	open_acc_6m	51	hardship_reason

17	open_act_il	52	hardship_status
18	open_il_12m	53	deferral_term
19	open_il_24m	54	hardship_amount
20	mths_since_rcnt_il	55	hardship_start_date
21	total_bal_il	56	hardship_end_date
22	il_util	57	payment_plan_start_date
23	open_rv_12m	58	hardship_length
24	open_rv_24m	59	hardship_dpd
25	max_bal_bc	60	hardship_loan_status
26	all_util	61	orig_projected_additional_accrued_interest
27	inq_fi	62	hardship_payoff_balance_amount
28	total_cu_tl	63	hardship_last_payment_amount
29	inq_last_12m	64	debt_settlement_flag_date
30	bc_open_to_buy	65	settlement_status
31	bc_util	66	settlement_date
32	mo_sin_old_il_acct	67	settlement_amount
33	mths_since_recent_bc	68	settlement_percentage
34	mths_since_recent_bc_dlq	69	settlement_term
35	mths_since_recent_inq		

What is the proportion of missing values in different variables?

The proportion of missing values in the different variable are summarised in the below table:

S. No	Variable name	Proportion	S.No	Variable name	Proportion
1	id	1.0000000000	36	mths_since_recent_revol_delinq	0.6293398171
2	member_id	1.0000000000	37	num_tl_120dpd_2m	0.0257828024
3	emp_title	0.0630931958	38	percent_bc_gt_75	0.0124532540
4	url	1.0000000000	39	revol_bal_joint	1.0000000000
5	desc	1.0000000000	40	sec_app_earliest_cr_line	1.0000000000
6	mths_since_last_delinq	0.4782839441	41	sec_app_inq_last_6mths	1.0000000000
7	mths_since_last_record	0.8178541896	42	sec_app_mort_acc	1.0000000000
8	revol_util	0.0004196339	43	sec_app_open_acc	1.0000000000
9	last_pymnt_d	0.0005553978	44	sec_app_revol_util	1.0000000000
10	next_pymnt_d	1.0000000000	45	sec_app_open_act_il	1.0000000000
11	last_credit_pull_d	0.0001481061	46	sec_app_num_rev_accts	1.0000000000
12	mths_since_last_major_derog	0.7033434951	47	sec_app_chargeoff_within_12_mths	1.0000000000
13	annual_inc_joint	1.0000000000	48	sec_app_collections_12_mths_ex_med	1.0000000000
14	dti_joint	1.0000000000	49	sec_app_mths_since_last_major_derog	1.0000000000

15	verification_status_joint	1.0000000000	50	hardship_type	1.0000000000
16	open_acc_6m	1.0000000000	51	hardship_reason	1.0000000000
17	open_act_il	1.0000000000	52	hardship_status	1.0000000000
18	open_il_12m	1.0000000000	53	deferral_term	1.0000000000
19	open_il_24m	1.0000000000	54	hardship_amount	1.0000000000
20	mths_since_rcnt_il	1.0000000000	55	hardship_start_date	1.0000000000
21	total_bal_il	1.0000000000	56	hardship_end_date	1.0000000000
22	il_util	1.0000000000	57	payment_plan_start_date	1.0000000000
23	open_rv_12m	1.0000000000	58	hardship_length	1.0000000000
24	open_rv_24m	1.0000000000	59	hardship_dpd	0.9996791035
25	max_bal_bc	1.0000000000	60	hardship_loan_status	1.0000000000
26	all_util	1.0000000000	61	orig_projected_additional_accrued_interest	1.0000000000
27	inq_fi	1.0000000000	62	hardship_payoff_balance_amount	1.0000000000
28	total_cu_tl	1.0000000000	63	hardship_last_payment_amount	0.9999876578
29	inq_last_12m	1.0000000000	64	debt_settlement_flag_date	1.0000000000
30	bc_open_to_buy	0.0120583044	65	settlement_status	1.0000000000
31	bc_util	0.0126877553	66	settlement_date	1.0000000000
32	mo_sin_old_il_acct	0.0375325525	67	settlement_amount	1.0000000000
33	mths_since_recent_bc	0.0112066944	68	settlement_percentage	1.0000000000
34	mths_since_recent_bc_dlq	0.7287560322	69	settlement_term	0.9950137615
35	mths_since_recent_inq	0.1014773583			

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?

For the values stated below in the table , we have analysed and come to the conclusion that they have a very large proportion of missing values . Most of these values are also not relevant since we need data that will predict loan default and then decide which loans to invest in.

S.No	Name of the variable	S.No	Name of the variable
1	id	30	sec_app_mort_acc
2	member_id	31	sec_app_open_acc
3	url	32	sec_app_revol_util
4	desc	33	sec_app_open_act_il
5	mths_since_last_record	34	sec_app_num_rev_accts
6	next_pymnt_d	35	sec_app_chargeoff_within_12_mths
7	mths_since_last_major_derog	36	sec_app_collections_12_mths_ex_med

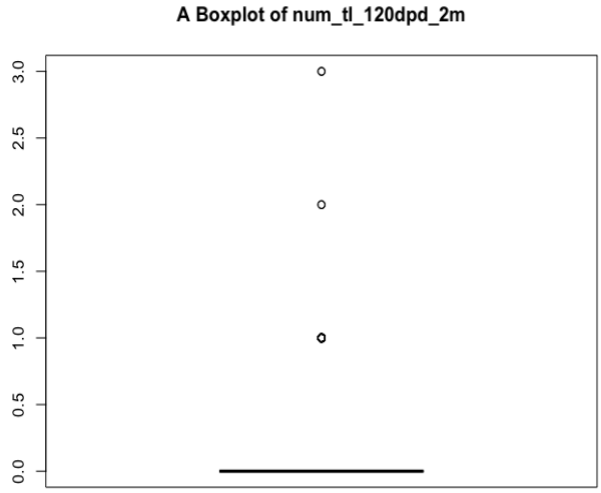
8	annual_inc_joint	37	sec_app_mths_since_last_major_derog
9	dti_joint	38	hardship_type
10	verification_status_joint	39	hardship_reason
11	open_acc_6m	40	hardship_status
12	open_act_il	41	deferral_term
13	open_il_12m	42	hardship_amount
14	open_il_24m	43	hardship_start_date
15	mths_since_rcnt_il	44	hardship_end_date
16	total_bal_il	45	payment_plan_start_date
17	il_util	46	hardship_length
18	open_rv_12m	47	hardship_dpd
19	open_rv_24m	48	hardship_loan_status
20	max_bal_bc	49	orig_projected_additional_accrued_interest
21	all_util	50	hardship_payoff_balance_amount
22	inq_fi	51	hardship_last_payment_amount
23	total_cu_tl	52	debt_settlement_flag_date
24	inq_last_12m	53	settlement_status
25	mths_since_recent_bc_dlq	54	settlement_date
26	mths_since_recent_revol_delinq	55	settlement_amount
27	revol_bal_joint	56	settlement_percentage
28	sec_app_earliest_cr_line	57	Settlement_term
29	sec_app_inq_last_6mths		

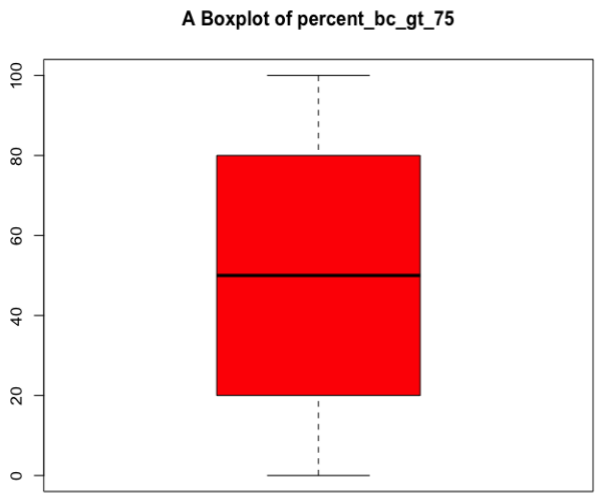
For variables with less than 60% missing values, we took the following action:

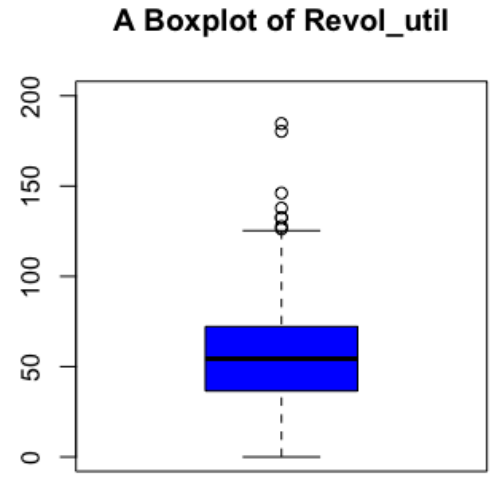
Variable	Action taken
mths_since_last_delinq	Replace with a high value, 500 in our case.
revol_util	Replace with median
bc_open_to_buy	Replace with median
bc_util	Replace with median
mo_sin_old_il_acct	Replace with 1000
mths_since_recent_bc	Replace with 1000
mths_since_recent_inq	Replace with 50
percent_bc_gt_75	Replace with median
num_tl_120dpd_2m	Replace with median

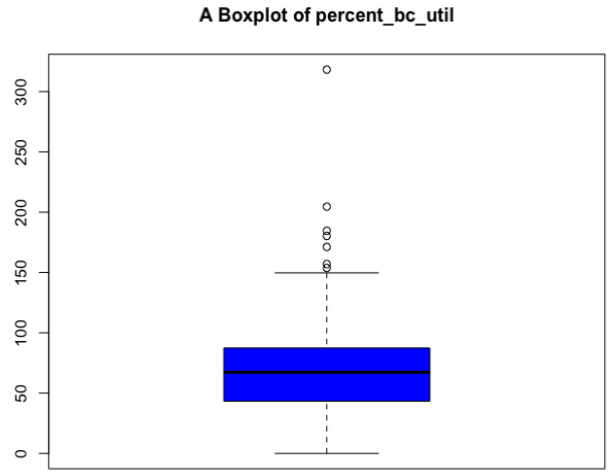
For the variables that we think would have a high chance of predicting defaults, we have replaced it with a value that is much higher than the maximum value of that variable available in our data.
For other variables, we have chosen to replace it with the median because unlike mean, median is resistant to the effect of outliers.

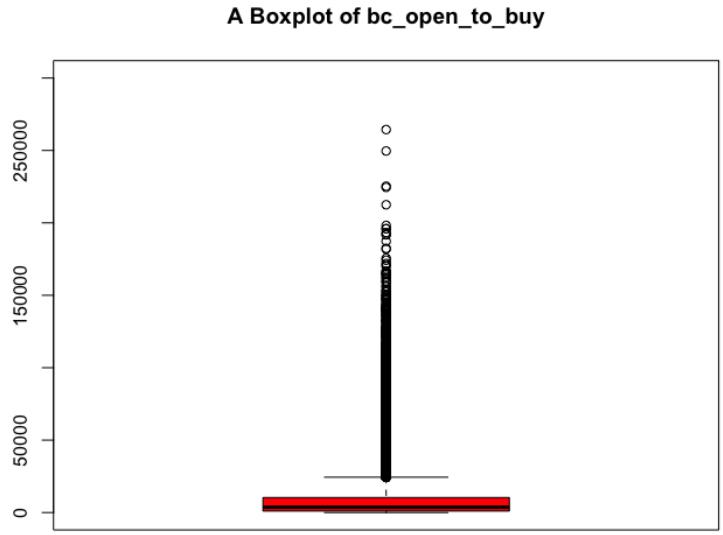
Given below are the boxplots for a few of the variables we have analysed for outliers.











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

The table below lists out some variables that have the potential for data leakage or is irrelevant to our model.

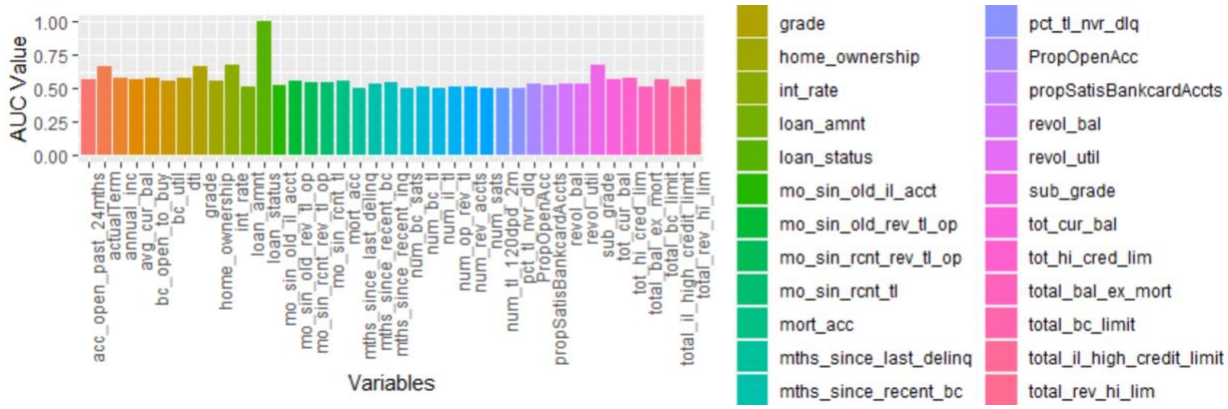
ATTRIBUTES REMOVED	REASONING	ATTRIBUTES REMOVED	REASONING
funded_amnt_inv	Not known before loan is funded	**debt_settlement_flag**	Causes data leakage
term	Not known before loan is funded	**annRet**	Updated after loan is funded
emp_title	Irrelevant to the prediction of defaults	**application_type**	Irrelevant to the prediction of defaults
actualReturn	Updated after loan is funded	**collection_recovery_fee**	Updated after loan is funded
pymnt_plan	Not known before loan is funded	**recoveries**	Updated after loan is funded
zip_code	Irrelevant to the prediction of defaults	**total_acc**	Since we already have the ratio of open accounts to total accounts in our model
addr_state	Irrelevant to the prediction of defaults	**initial_list_status**	Irrelevant to the prediction of defaults
out_prncp	Updated after loan is funded	**inq_last_6mths**	Can be updated with time
out_prncp_inv	Updated after loan is funded	**issue_d**	Updated after loan is funded
total_pymnt_inv	Updated after loan is funded	**last_pymnt_amnt**	Updated after loan is funded
total_rec_prncp	Updated after loan is funded	**last_pymnt_d**	Updated after loan is funded
total_rec_int	Updated after loan is funded	**open_acc**	Irrelevant to the prediction of defaults
last_credit_pull_d	Updated after loan is funded	**revol_bal**	Irrelevant to the prediction of defaults
policy_code	Not known before loan is funded	**revol_util**	Irrelevant to the prediction of defaults
disbursement_method	Updated after loan is funded	**total_acc**	Irrelevant to the prediction of defaults
hardship_flag	Causes data leakage	**total_pymnt**	Irrelevant to the prediction of defaults
Funded_amnt	Updated after loan is funded	**total_rec_late_fee**	Updated after loan is funded
tot_coll_amt	Updated after loan is funded	**open_acc**	Since we already have the ratio of open

			accounts to total accounts in our model

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

We are using Area Under the Curve (AUC) value as a measure of the importance of all the potential predictor values.

On setting the AUC threshold as 0.5, we get a list of 39 variables after dropping the variables which cause data leakage.

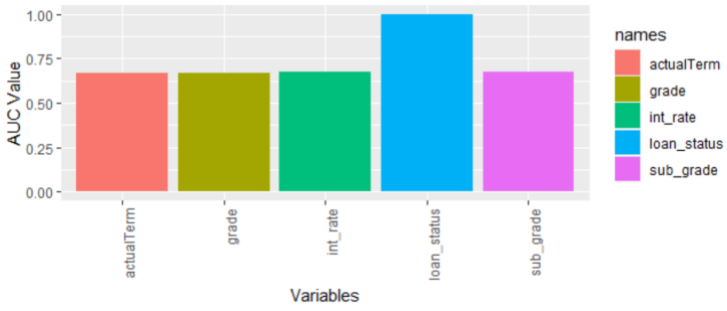


The top most predictor variables are :

sub_grade	0.6739424
int_rate	0.6736353
grade	0.6652935
actualTerm	0.6646336
bc_open_to_buy	0.5817614
annual_inc	0.5796404
dti	0.5792397
tot_hi_cred_lim	0.5773238
total_bc_limit	0.5746631
avg_cur_bal	0.5745592
acc_open_past_24mths	0.5728784
total_rev_hi_lim	0.5667577

On increasing the AUC threshold to 0.6, we are just left with 4 variables :

sub_grade	0.6739424
int_rate	0.6736353
grade	0.6652935
actualTerm	0.6646336


1.00
0.75
0.50
0.25
0.00
AUC Value
actualTerm
grade
int_rate
loan_status
sub_grade
Variables
names
actualTerm
grade
int_rate
loan_status
sub_grade