LENDING CLUB LOAN ANALYSIS

Descriptive and Predictive Analytics of Dataset

By

Vedant Kashyap – vkashya2@depaul.edu

Fadairo, Temitope - tfadairo@depaul.edu

Dammu, Yasaswini Niharika - ydammu@depaul.edu

LENDING CLUB LOAN ANALYSIS AND RISK ASSESSMENT

Motivation

Lending Club Loan dataset is a big data set with a lot of variables. These variables are, to name a few, annual income, interest rate on loan, term of loan, loan amount, installments paid by customers, loan purpose, individual or joint loan etc. These variables were easy to understand and simple to work with. We were also able to find a file that explained all the variables in brief which motivated us even further to choose this data set.

After checking the data set, we were motivated to find if the data could help us find the mean of annual incomes of customers, maximum interest rates that were given out, why were the loans given out, predicting default rate of a customer and loan performance given by organization.

It helps to determine those people that are likely to default and those that are not. It gives a true picture of their behavioral pattern when they are faced with a test of integrity rather than a security pledge.

It helps to know how alternative loan sources reach the unbanked and contribute to financial inclusion in the economy.

Objective

Lending Club Loan Analysis and Risk Assessment is an interesting topic as it gives insight into how individuals and businesses behave when they get loans.

The questions that need to be answered by the data set can help us and Lending Club to understand what type of customers they currently have, what is the interest rates they are providing and if it is affecting the repayment of loan by customers. Data set will also help us recognize if a customer is loan worthy or no and at what interest rate should they be given loan at.

Some the questions we will get an answer for are –

1. What is the maximum annual income of customers?
2. What type of customers does Lending Club have currently?
3. What is the purpose of loan taken by customers?
4. What is the mean of interest rates given out by Lending Club?
5. What is the correlation between financial performance variables?
6. Is there a difference in the population between the different groups of the independent variables with respect to the dependent variables?
7. To Identify the Null Hypothesis and the Alternative hypothesis using data set. Null Hypothesis will be no risk is lending to customers. At different significant levels.
8. What is the chance of a customer defaulting on the loan?
9. What is the overall performance of the loan portfolio?

Answers

We used a mixture of SAS and R to answer our questions mentioned above.

We used SAS main for descriptive analysis and finding simple regression. R will be used for predictive analysis.

SAS

We used descriptive statistics to get insight into the dataset to determine the pattern, distribution, mean, and variability. We employed procedures in SAS such as **proc mean**, **proc summary**, **proc univariate**, **proc anova**, **proc corr**, **proc reg** and **proc sgplot**.

Using these procedures in SAS, we will be able to identify –

1. mean of interest rates,
2. maximum annual income of customers,
3. Outliers in data if any
4. Difference between 2 or more groups.
5. Correlation between each variable
6. Plot output on histograms and bar charts.
7. Find out simple regression between 2 variables.

We used random forest and neuron network model to predict the customers that are likely to default and those that are not.

The result shows that over half of the customers are likely to default; this is further buttressed by the neuron network result. The lending club management should take immediate steps to minimize risk by tightening the lending policy such that high-risk individuals are charged higher interest rates to match the possible loss that may occur if there is default.

They should also diversify risk amongst different loan categories such that there is no concentration risk on a particular loan segment or individual/business.

Data and Empirical methodology

About the data

Lending Club Loan data consists of loan data from 2015. It has 421094 observations spread across 77 variables. These variables range from borrowers' personal information and reason for borrowing along with their past dues or deficiencies.

The data information include:

addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
desc	Loan description provided by the borrower
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.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Status of the loan
member_id	A unique LC assigned Id for the borrower member.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
next_pymnt_d	Next scheduled payment date

open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
total_acc	The total number of credit lines currently 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_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
open_acc_6m	Number of open trades in last 6 months
open_il_6m	Number of currently active installment trades
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
mths_since_rcnt_il	Months since most recent installment accounts opened
total_bal_il	Total current balance of all installment accounts
il_util	Ratio of total current balance to high credit/credit limit on all install acct
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
max_bal_bc	Maximum current balance owed on all revolving accounts
all_util	Balance to credit limit on all trades
total_rev_hi_lim	Total revolving high credit/credit limit
total_cu_tl	Number of finance trades
inq_last_12m	Number of credit inquiries in past 12 months
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
tot_coll_amt	Total collection amounts ever owed

While using the data for descriptive and predictive analytics, there was a need to change the variables into different types of variables such as a numeric and categorical variable. Changing these variables helped us correctly finding out the analytics of the data.

Some of the variables that were used were.

Separting numerical value from text

```
e_length = input(compress (emp_length, , 'kd'),?? best32.);

loan_term_months = input(compress (term, , 'kd'),?? best32.);
```

Creating Categorical variables for loan status

```
if loan_status = "Default" then status = 1; else  status = 0;
```

Creating categorical variables in R

```
lc$status <- ifelse(lc$loan_status == "Default",1,0)

lc$grade_i <- ifelse(lc$grade %in% c("A", "B"), 1,
              ifelse(lc$grade %in% c("C", "D","E", "F", "G"),0,NA))

lc$verification <- ifelse(lc$verification_status == "Not Verified",1,0)
```

Creating Factor Variables

```
train1$pred_status <- as.factor(train1$status)

test1$pred_status <- as.factor(test1$status)
```

Statistics of Data

As mentioned above, we did descriptive analytics using SAS programing language and got some interesting results.

1. Using **Proc Means**, we were able to find out that
   a. Maximum Loan amount given out was $35000.
   b. Minimum Interest Rates given out were 5.32%.
   c. Average annual income of customers is approximately around $77038.
   d. Maximum delinquency days in past 2yrs is 30days.

The SAS System

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
loan_amnt	150000	15252.53	8574.63	1000.00	35000.00
int_rate	150000	12.5826131	4.3147250	5.3200000	28.9900000
annual_inc	150000	77038.08	71897.57	1770.00	9500000.00
annual_inc_joint	170	105235.88	46174.67	28000.00	270000.00
dti	150000	19.1331740	9.1554922	0	1092.52
delinq_2yrs	150000	0.3455733	0.9175391	0	30.0000000
mths_since_last_delinq	77404	34.0060979	21.9897723	0	176.0000000
open_acc	150000	11.9513200	5.6334171	1.0000000	74.0000000

2. Using **Proc Summary** and classifying data for Default customers, we were able to find that
    a. Out of 150000 customer observation we took, 13 customers had defaulted on their loans.
    b. Annual income means of customers who defaulted on their loans was lower than other loan borrowers.
    c. This data also shows that customers who are defaulting on their loans have higher mean interest rates compared to others.
    d. Customers defaulting on loans are mostly individuals since there are no values in annual_inc_join

The SAS System

The SUMMARY Procedure

status	N Obs	Variable	Mean	Median	Std Dev	Minimum	Maximum	5th Pctl	95th Pctl
0	149861	annual_inc	77046.89	65000.00	71921.09	1770.00	9500000.00	29000.00	157000.00
		annual_inc_joint	105235.88	97425.00	46174.67	28000.00	270000.00	46122.00	190000.00
		dti	19.1300153	18.5500000	9.1556447	0	1092.52	5.6400000	34.3100000
		funded_amnt	15252.97	14000.00	8575.21	1000.00	35000.00	4000.00	33600.00
		int_rate	12.5791359	12.2900000	4.3133097	5.3200000	28.9900000	6.3900000	19.9900000
		last_pymnt_amnt	1102.07	392.8100000	3483.66	0	36257.59	61.0700000	2700.00
		mths_since_last_delinq	34.0035557	30.0000000	21.9900153	0	176.0000000	5.0000000	74.0000000
1	139	annual_inc	67542.42	55000.00	37988.10	15000.00	202000.00	25000.00	150000.00
		annual_inc_joint	.	.	.	.	.	.	.
		dti	22.5387050	22.6700000	8.3489476	3.3800000	39.0600000	8.5200000	37.3000000
		funded_amnt	14781.65	13625.00	7941.27	1000.00	35000.00	3000.00	32800.00
		int_rate	16.3314388	16.5500000	4.2241438	6.3900000	25.9900000	9.9900000	24.5000000
		last_pymnt_amnt	436.4887770	406.2300000	256.2851439	0	1835.94	86.2500000	886.1100000
		mths_since_last_delinq	37.1269841	35.0000000	21.6359111	1.0000000	81.0000000	8.0000000	72.0000000

home_ownership	N Obs	Variable	Mean	Median	Std Dev	Minimum	Maximum	5th Pctl	95th Pctl
MORTGAGE	74147	annual_inc	87015.35	75000.00	78720.31	1770.00	9500000.00	35000.00	175000.00
		annual_inc_joint	115437.95	104900.00	46453.71	29448.00	270000.00	55000.00	195000.00
		dti	19.0732376	18.5200000	9.4628959	0	1092.52	5.8500000	33.9500000
		funded_amnt	16827.63	15000.00	8863.58	1000.00	35000.00	4475.00	35000.00
		int_rate	12.2948149	12.2900000	4.3620393	5.3200000	28.9900000	6.2400000	19.9900000
		last_pymnt_amnt	1237.58	446.3100000	3856.27	0	36257.59	68.9900000	3978.20
		mths_since_last_delinq	33.0582197	29.0000000	22.1720796	0	171.0000000	4.0000000	73.0000000
		delinq_2yrs	0.3889975	0	0.9606193	0	26.0000000	0	2.0000000
OWN	16226	annual_inc	71944.67	60000.00	58364.34	7000.00	2100000.00	24000.00	150000.00
		annual_inc_joint	87017.77	78374.36	32176.58	37000.00	148000.00	37000.00	148000.00
		dti	19.7611580	19.3700000	9.0947448	0	55.1000000	5.3800000	35.2800000
		funded_amnt	14867.21	13000.00	8496.84	1000.00	35000.00	3600.00	32000.00
		int_rate	12.6252841	12.2900000	4.3485494	5.3200000	28.9900000	6.2400000	19.9900000
		last_pymnt_amnt	1137.84	386.7650000	3602.46	0	36058.71	43.6700000	4018.18
		mths_since_last_delinq	33.9143683	30.0000000	22.0555903	0	176.0000000	4.0000000	74.0000000
		delinq_2yrs	0.3332922	0	0.8870079	0	22.0000000	0	2.0000000
RENT	59627	annual_inc	66017.25	55815.00	64179.06	5000.00	8900060.00	25000.00	135000.00
		annual_inc_joint	85063.65	76900.50	41238.15	28000.00	260000.00	44000.00	155000.00
		dti	19.0368155	18.4100000	8.7686575	0	72.3000000	5.4700000	34.4500000
		funded_amnt	13398.73	12000.00	7814.11	1000.00	35000.00	3200.00	30000.00
		int_rate	12.9288822	12.6900000	4.2194587	5.3200000	28.9900000	6.8900000	19.9900000
		last_pymnt_amnt	922.2865148	350.2100000	2902.35	0	35964.28	52.7600000	1347.63
		mths_since_last_delinq	35.3612666	32.0000000	21.6416955	0	152.0000000	5.0000000	74.0000000
		delinq_2yrs	0.2949167	0	0.8667938	0	30.0000000	0	2.0000000

This proc summary data tells us about the type of home ownership the customer had when he applied for a loan.

Installment Amount VS Application Type

The UNIVARIATE Procedure
Variable: loan_amnt

Moments			
N	150000	Sum Weights	150000
Mean	15252.5292	Sum Observations	2287879375
Std Deviation	8574.63445	Variance	73524355.9
Skewness	0.625679	Kurtosis	-0.3788953
Uncorrected SS	4.59245E13	Corrected SS	1.10286E13
Coeff Variation	56.2177876	Std Error Mean	22.1396109

Basic Statistical Measures			
Location		Variability	
Mean	15252.53	Std Deviation	8575
Median	14000.00	Variance	73524356
Mode	10000.00	Range	34000
		Interquartile Range	11500

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	688.9249	Pr > |t|	<.0001
Sign	M	75000	Pr >= |M|	<.0001
Signed Rank	S	5.625E9	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	35000
99%	35000
95%	33600
90%	28000
75% Q3	20000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1000	149904	35000	149893
1000	149774	35000	149903
1000	149468	35000	149922
1000	149411	35000	149948
1000	149344	35000	149962

This explains the Proc Univariate for loan amount. Here, in the output we can see key features such as mean, standard deviation, variance, standard error, coefficient of variance, quantile values at different levels of data.

9 Variables:	annual_inc annual_inc_joint dti funded_amnt int_rate last_pymnt_amnt mths_since_last_delinq delinq_2yrs installment

Variances and Covariances Covariance / Row Var Variance / Col Var Variance / DF	annual_inc	annual_inc_joint	dti	funded_amnt	int_rate	last_pymnt_amnt	mths_since_last_delinq	delinq_2yrs	installment
annual_inc	5169261007 5169261007 5169261007 149999	879042013 1054392389 2132099956 169	-115072 5169261007 84 149999	191944437 5169261007 73524356 149999	-29349 5169261007 19 149999	15296593 5169261007 12125116 149999	-69664 4639583670 484 77403	2542 5169261007 1 149999	5328704 5169261007 60173 149999
annual_inc_joint	879042013 2132099956 1054392389 169	2132099956 2132099956 2132099956 169	68580 2132099956 7658 169	194456644 2132099956 79733493 169	-9666 2132099956 17 169	1151130 2132099956 97453 169	-103119 1947459411 455 92	1093 2132099956 1 169	4651505 2132099956 65826 169
dti	-115072 84 5169261007 149999	68580 7658 2132099956 169	84 84 84 149999	1423 84 73524356 149999	8 84 19 149999	-958 84 12125116 149999	2 90 484 77403	-0 84 1 149999	14 84 60173 149999
funded_amnt	191944437 73524356 5169261007 149999	194456644 79733493 2132099956 169	1423 73524356 84 149999	73524356 73524356 73524356 149999	5069 73524356 19 149999	5001765 73524356 12125116 149999	-6177 71675133 484 77403	-80 73524356 1 149999	1979998 73524356 60173 149999
int_rate	-29349 19 5169261007 149999	-9666 17 2132099956 169	8 19 84 149999	5069 19 73524356 149999	19 19 19 149999	1156 19 12125116 149999	-1 18 484 77403	0 19 1 149999	127 19 60173 149999
last_pymnt_amnt	15296593 12125116 5169261007 149999	1151130 97453 2132099956 169	-958 12125116 84 149999	5001765 12125116 73524356 149999	1156 12125116 19 149999	12125116 12125116 12125116 149999	462 11475175 484 77403	-38 12125116 1 149999	148832 12125116 60173 149999
mths_since_last_delinq	-69664 484 4639583670 77403	-103119 455 1947459411 92	2 484 90 77403	-6177 484 71675133 77403	-1 484 18 77403	462 484 11475175 77403	484 484 484 77403	-14 484 1 77403	-195 484 59782 77403
delinq_2yrs	2542 1 5169261007 149999	1093 1 2132099956 169	-0 1 84 149999	-80 1 73524356 149999	0 1 19 149999	-38 1 12125116 149999	-14 1 484 77403	1 1 1 149999	-0 1 60173 149999
installment	5328704 60173 5169261007 149999	4651505 65826 2132099956 169	14 60173 84 149999	1979998 60173 73524356 149999	127 60173 19 149999	148832 60173 12125116 149999	-195 59782 484 77403	-0 60173 1 149999	60173 60173 60173 149999

This table shows the relationship between variance and covariance of different variables used in descriptive analysis.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	annual_inc	annual_inc_joint	dti	funded_amnt	int_rate	last_pymnt_amnt	mths_since_last_delinq	delinq_2yrs	installment
annual_inc	1.00000 150000	0.58628 <.0001 170	-0.17481 <.0001 150000	0.31135 <.0001 150000	-0.09461 <.0001 150000	0.06110 <.0001 150000	-0.04651 <.0001 77404	0.03854 <.0001 150000	0.30214 <.0001 150000
annual_inc_joint	0.58628 <.0001 170	1.00000 170	0.01697 0.8261 170	0.47163 <.0001 170	-0.05029 0.5148 170	0.07986 0.3006 170	-0.10949 0.2961 93	0.02973 0.7004 170	0.39264 <.0001 170
dti	-0.17481 <.0001 150000	0.01697 0.8261 170	1.00000 150000	0.01813 <.0001 150000	0.19704 <.0001 150000	-0.03005 <.0001 150000	0.00784 0.0292 77404	-0.02077 <.0001 150000	0.00619 0.0164 150000
funded_amnt	0.31135 <.0001 150000	0.47163 <.0001 170	0.01813 <.0001 150000	1.00000 150000	0.13700 <.0001 150000	0.16752 <.0001 150000	-0.03318 <.0001 77404	-0.01015 <.0001 150000	0.94135 <.0001 150000
int_rate	-0.09461 <.0001 150000	-0.05029 0.5148 170	0.19704 <.0001 150000	0.13700 <.0001 150000	1.00000 150000	0.07695 <.0001 150000	-0.01531 <.0001 77404	0.04317 <.0001 150000	0.12043 <.0001 150000
last_pymnt_amnt	0.06110 <.0001 150000	0.07986 0.3006 170	-0.03005 <.0001 150000	0.16752 <.0001 150000	0.07695 <.0001 150000	1.00000 150000	0.00620 0.0847 77404	-0.01194 <.0001 150000	0.17424 <.0001 150000
mths_since_last_delinq	-0.04651 <.0001 77404	-0.10949 0.2961 93	0.00784 0.0292 77404	-0.03318 <.0001 77404	-0.01531 <.0001 77404	0.00620 0.0847 77404	1.00000 77404	-0.55800 <.0001 77404	-0.03633 <.0001 77404
delinq_2yrs	0.03854 <.0001 150000	0.02973 0.7004 170	-0.02077 <.0001 150000	-0.01015 <.0001 150000	0.04317 <.0001 150000	-0.01194 <.0001 150000	-0.55800 <.0001 77404	1.00000 150000	-0.00213 0.4089 150000
installment	0.30214 <.0001 150000	0.39264 <.0001 170	0.00619 0.0164 150000	0.94135 <.0001 150000	0.12043 <.0001 150000	0.17424 <.0001 150000	-0.03633 <.0001 77404	-0.00213 0.4089 150000	1.00000 150000

The table represents the relationship between correlation coefficients between the variables we have used to define our outputs. In this table, values close to 1 have a positive correlation with the corresponding variables. If values is close to 0, then they have a negative correlation with the corresponding variable.

Installment Amount VS Application Type

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=3 Maxiter=1

Initial Seeds			
Cluster	annual_inc	int_rate	loan_amnt
1	4695472.000	8.180	3000.000
2	9500000.000	7.890	24000.000
3	1770.000	17.860	6550.000

Criterion Based on Final Seeds = 31473.6

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	12	576007	1880012		3	2938074
2	3	185632	366753		1	6118658
3	149985	30756.0	1823340		1	2938074

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
annual_inc	71898	53277	0.450918	0.821221
int_rate	4.31472	4.31471	0.000020	0.000020
loan_amnt	8575	8575	0.000002	0.000002
OVER-ALL	41804	31155	0.444594	0.800485

Pseudo F Statistic = 60035.21

Approximate Expected Over-All R-Squared = 0.87643

Cubic Clustering Criterion = -481.502

Cluster Means			
Cluster	annual_inc	int_rate	loan_amnt
1	3014695.250	10.799	13893.750
2	9133353.333	10.130	15183.333
3	76621.898	12.583	15252.639

Cluster Standard Deviations			
Cluster	annual_inc	int_rate	loan_amnt
1	997607.5428	3.3997	11414.9629
2	321433.2505	4.7721	7638.7717
3	52576.2729	4.3148	8574.4520

Using proc fastclus, we clustered out data and found out that optimal number of clusters was 3. We used Non-Hierarchical K-Means method, since we have a very last data set and we took 150,000 observations for our analysis. K-Means is best suited for very large data sets.

Loan Amount to Interest Rate for Different Ownership type

The ANOVA Procedure

Dependent Variable: status

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	109	0.3671804	0.0033686	3.65	<.0001
Error	149890	138.5040130	0.0009240		
Corrected Total	149999	138.8711933			

R-Square	Coeff Var	Root MSE	status Mean
0.002644	3280.358	0.030398	0.000927

Source	DF	Anova SS	Mean Square	F Value	Pr > F
int_rate	109	0.36718036	0.00336863	3.65	<.0001

The ANOVA Procedure

Bonferroni (Dunn) t Tests for status

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	149890
Error Mean Square	0.000924
Critical Value of t	4.45641

Comparisons significant at the 0.05 level are indicated by ***.				
int_rate Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
25.89 - 24.5	0.0106324	-0.0082362	0.0295010	
25.89 - 25.83	0.0158983	-0.0068582	0.0386547	
25.89 - 25.99	0.0193108	-0.0020254	0.0406469	
25.89 - 18.54	0.0207662	0.0028009	0.0387316	***
25.89 - 19.52	0.0212817	0.0039696	0.0385937	***
25.89 - 25.57	0.0226090	0.0027419	0.0424762	***
25.89 - 17.14	0.0232134	0.0056697	0.0407572	***
25.89 - 22.99	0.0236314	0.0065703	0.0406926	***
25.89 - 21.67	0.0240064	0.0061328	0.0418800	***
25.89 - 15.99	0.0243782	0.0070387	0.0417177	***
25.89 - 18.84	0.0249475	0.0077087	0.0421863	***
25.89 - 12.99	0.0250354	0.0080076	0.0420632	***

Using Proc Anova analysis, we were able to compare the means of interest rates and status of a loan. We were able to find the F-values and P values of the 2 variables.

Estimation Equations

The estimating equation represents the relationship between the status of a loan application (whether it defaulted or not) and various predictor variables such as annual income, debt-to-income ratio, employment length, loan amount, loan grade, number of inquiries in the last 6 months, total number of credit accounts, number of public records, revolving line utilization rate, and number of delinquencies in the past 2 years.

For simple regression we use regression equation

$$E(y|x) = \beta_0 + \beta_1 x$$

- $E(y|x)$ = expected value of y for a given value of x
- β_0 = y-intercept of the regression line
- β_1 = slope

- The graph of the simple linear regression equation is a straight line.

Here, out simple regression equation, the one that we are using for regression is

Yhat(annual_income) = 96874 + (-1576.45)*(interest rate)

For multiple regression, we use regression equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

- y = dependent variable
- $x_1, x_2, \ldots, x_q$ = independent variables
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ = parameters
- ε = error term (accounts for the variability in y that cannot be explained by the linear effect of the q independent variables.)

status = -0.0010205090310417*(Intercept) + 6.98671033345974e-09*annual_inc + -1.24068870190142e-07*dti + -0.00237330559019757*emp_length< 1 year + 0.00195975599225035*emp_length1 year + -0.000414507932757675*emp_length10+ years + -0.00231477108931*emp_length2 years + -0.00220330612953037*emp_length3 years + 0.000230303574054304*emp_length4 years + 0.000211050676656935*emp_length5 years + -0.00247457118482226*emp_length6 years + -0.00220999027802783*emp_length7 years + -0.0023273731096 2239*emp_length8 years + 0.0012382618184884*emp_length9 years + -9.48604039470427e-08*loan_amnt + 0.0023409310168 1438*grade_i + -0.000508773676386702*inq_last_6mths + -3.37835479909979e-06*total_acc + -0.000378869452968356*pub_rec + 1.56793765161329e-05*revol_util + -0.000148618786954317*delinq_2yrs

Each coefficient represents the impact of 1its corresponding predictor variable on the target variable "status". For example, a positive coefficient indicates that an increase in the predictor variable will lead to an increase in the predicted status, while a negative coefficient indicates the opposite. The magnitude of the coefficient represents the strength of this relationship.

Coefficients:

annual_inc: This coefficient indicates the change in the status of the loan for a one-unit increase in the annual income of the borrower.

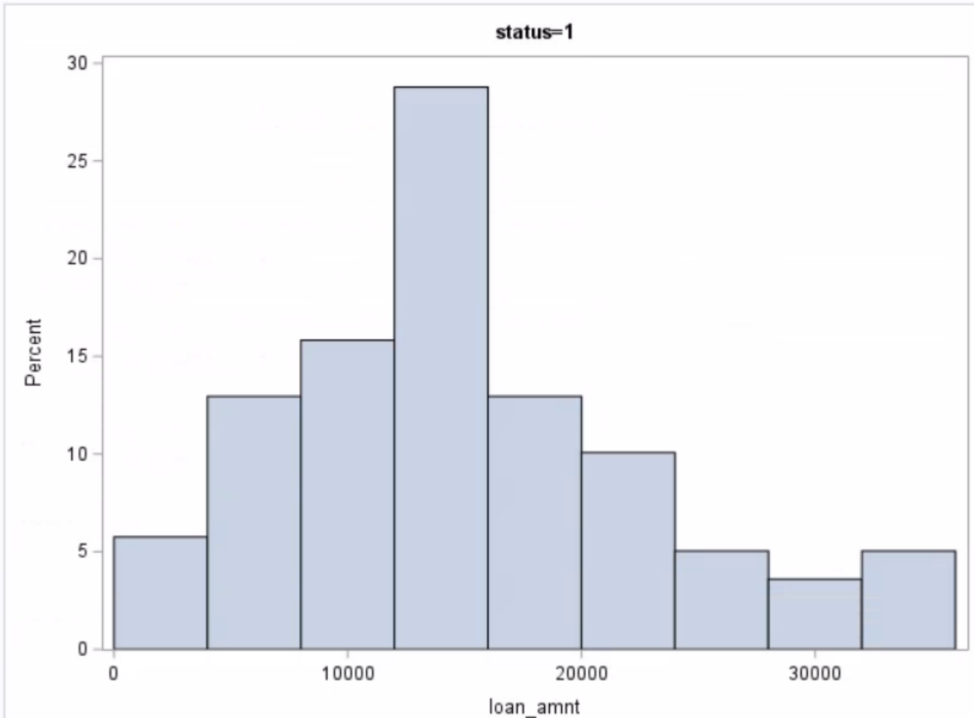
loan_amnt: This coefficient, which is like annual income, shows how the loan status changes when a borrower increases the loan amount by one unit.

inq_last_6mths, total_acc, pub_rec, revol_util, delinq_2yrs: Each of these coefficients represents the change in loan status for a one-unit increase in the respective variable.

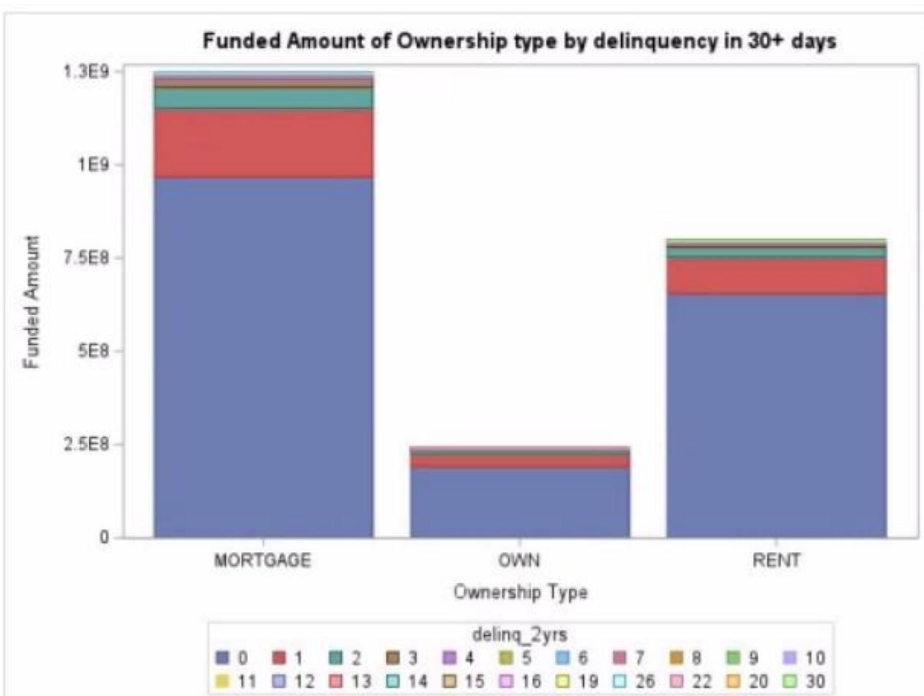
Based on the values of predictor variables, the regression model may be applied to predict the outcome of future loan applications. Making informed decisions and evaluating the risk involved with new loan applications could both benefit from this.

Descriptive Analysis Results

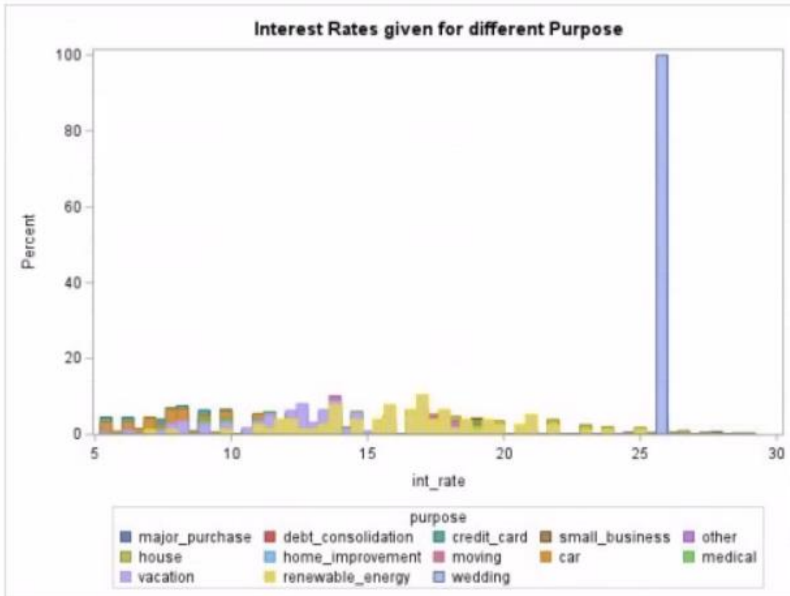
1. Using **Proc sgplot,** we were able to plot histograms showing different results of descriptive analysis of the data set.



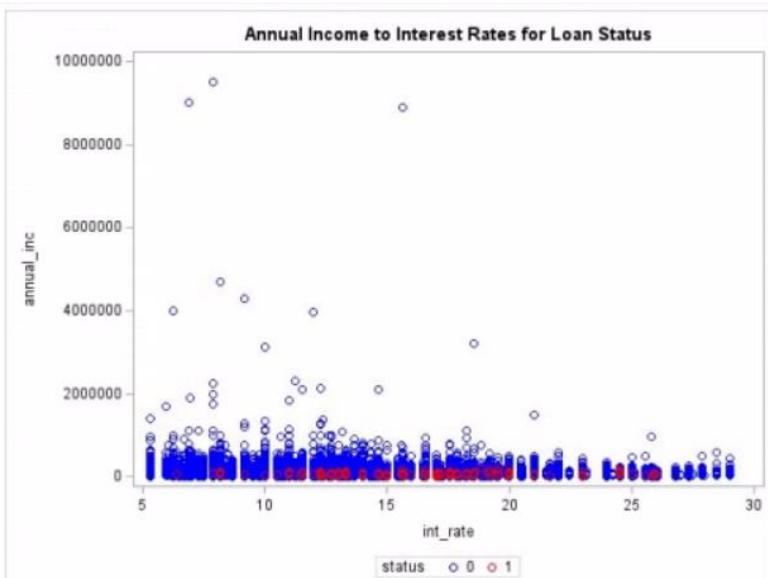
This histogram shows us that lending club's loan default amount by customers is mostly left skewed with maximum number of customers borrowing amount between $8000-15000



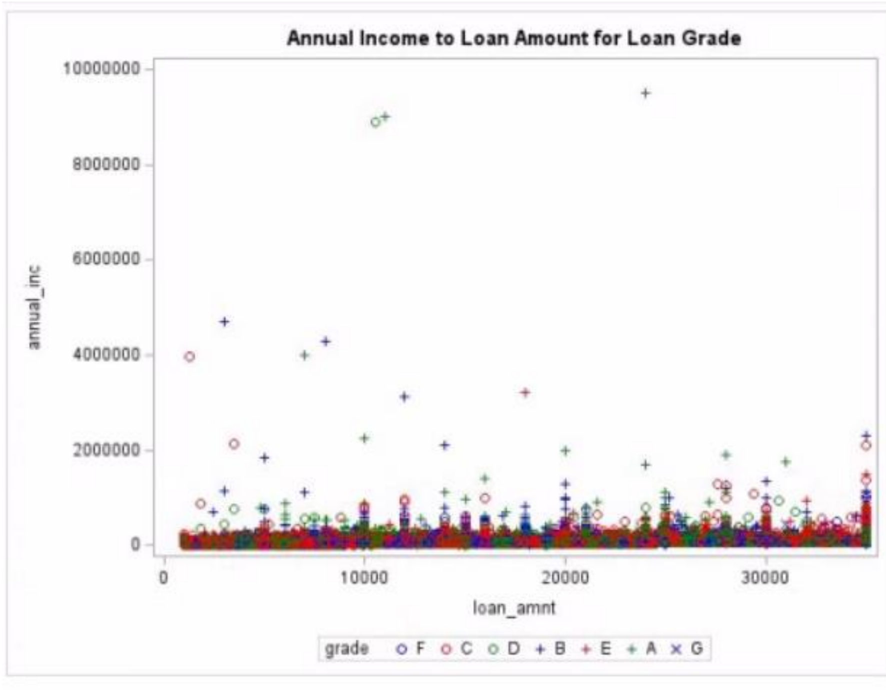
This graph shows us different types of homeowners and their funded amount and how many days of delinquency have they had in last 2years



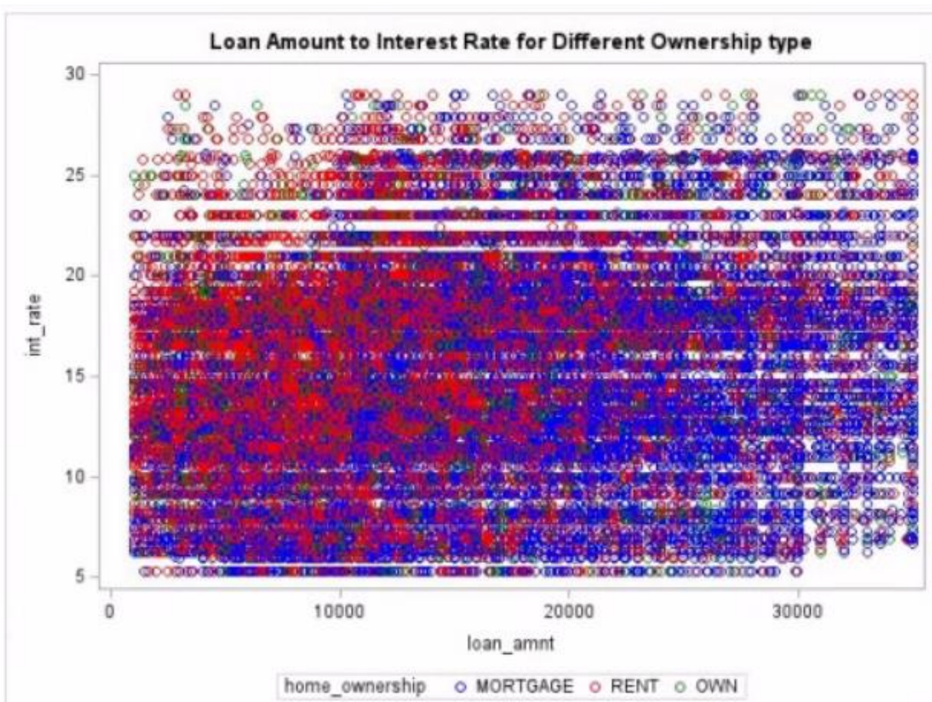
From this graph we can see that maximum percentage of loan given out were for wedding(personal loan) and it has the highest interest rate.



K-Means cluster analysis of Annual Income VS Interest rates for default status.



K-Means cluster analysis of Annual Income VS Loan Amount for loan grade type(A>B>C>D...)



K-Means cluster analysis of Loan Amount VS Interest Rates for different home ownerships

Model 1 : Simple Regression Analysis

```
Residual standard error: 0.4967 on 66498 degrees of freedom
Multiple R-squared:  0.004834,  Adjusted R-squared:  0.00482
F-statistic:   323 on 1 and 66498 DF,  p-value: < 2.2e-16
```

Model 1 RMSE: 1.614606e-17
RMSE measures the average magnitude of the errors or residuals between predicted values and actual observed values.

From the above output of simple regression analysis, we find out that our multiple R-squared value : 0.004834 and Adjusted R-squared value: 0.00482 are not significant. It shows a relatively low Adjusted R-squared value (0.00482), indicating limited prediction value.

Model 2: Multiple Regression Analysis

```
Residual standard error: 0.4705 on 66480 degrees of freedom
Multiple R-squared:  0.1073,    Adjusted R-squared:  0.107
F-statistic: 420.6 on 19 and 66480 DF,  p-value: < 2.2e-16
```

Model 2 RMSE: 3.916554e-18

Model 2, which has multiple predictors, demonstrates a higher Adjusted R-squared value (0.107), suggesting improved explanatory power compared to Model 1. However, the model's performance might still be limited, as indicated by the modest Adjusted R-squared value.

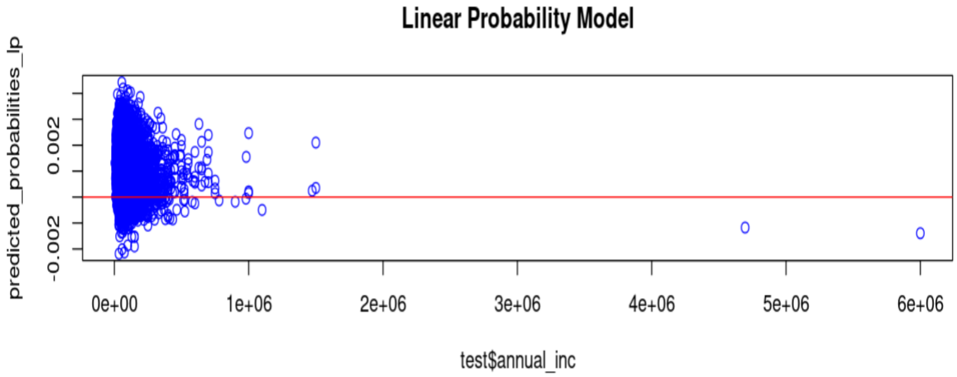
Model 3: Multiple Regression Analysis Using Stepwise Method

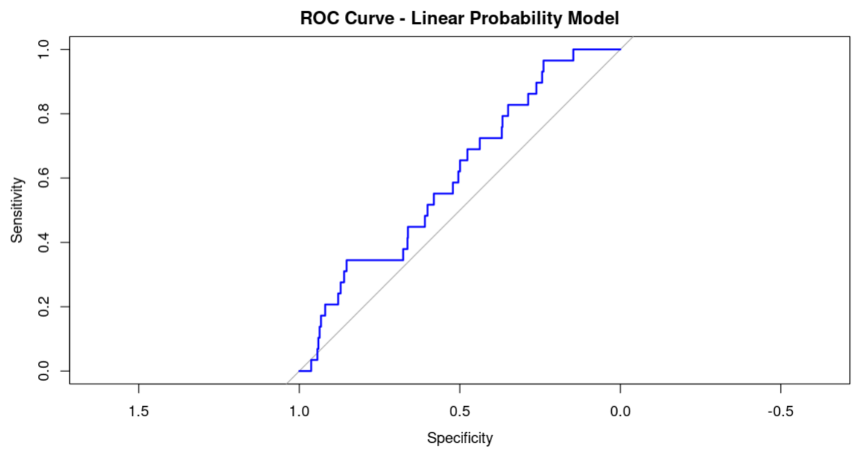
```
Residual standard error: 0.4705 on 66480 degrees of freedom
Multiple R-squared:  0.1073,    Adjusted R-squared:  0.107
F-statistic: 420.6 on 19 and 66480 DF,  p-value: < 2.2e-16
```

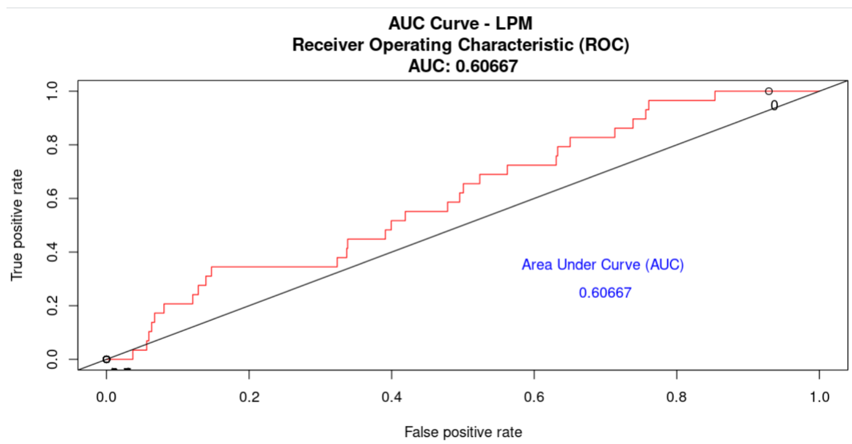
Model 3 RMSE: 3.916554e-18

Model 3 which is a stepwise regression, selects the most relevant predictors and offers a balance between model complexity and performance. It provides insights into the significant predictors while potentially improving model interpretability and generalization.

Model 4: Linear Probability Model

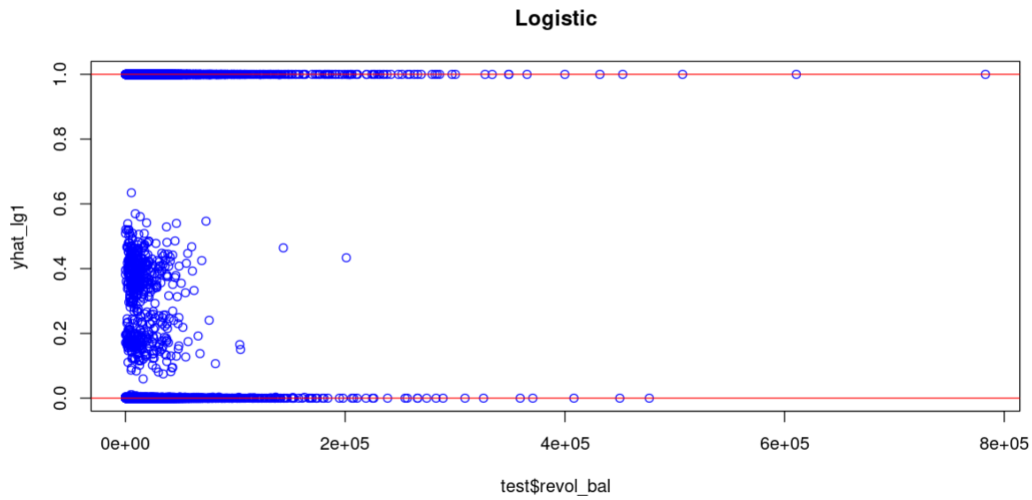


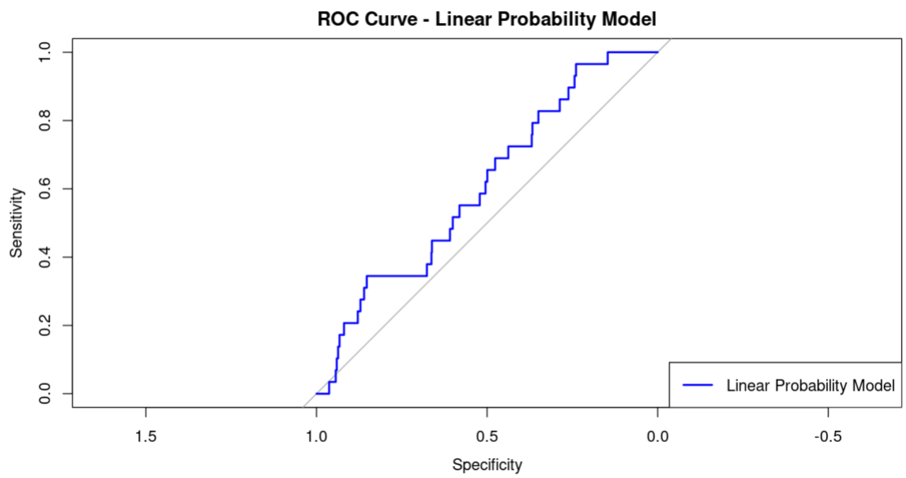


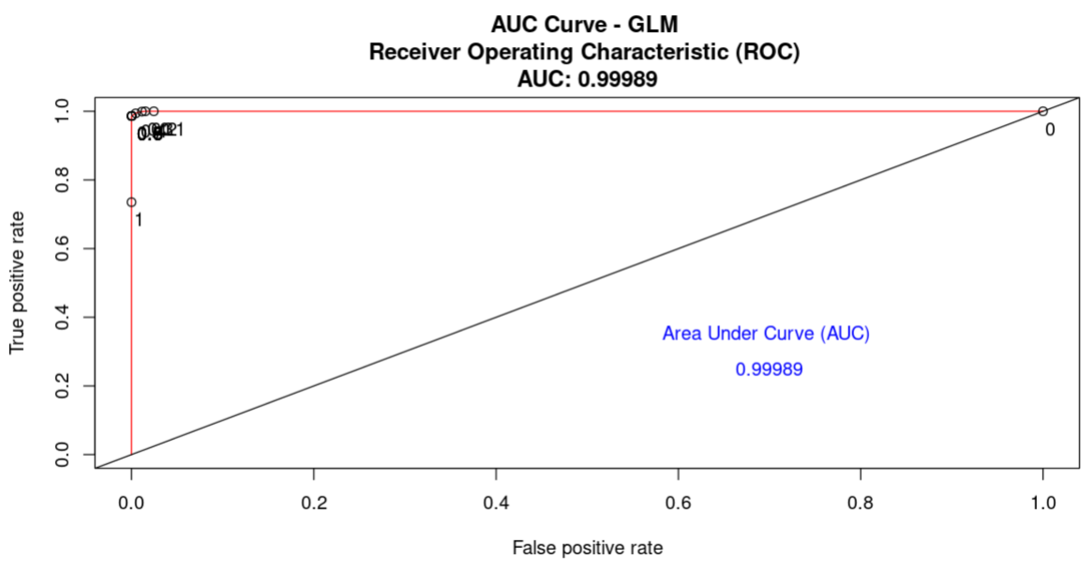


Model 4, utilizing a linear probability model and evaluating the AUC, focuses on the discriminatory ability of the model. The AUC value of 0.6534 suggests moderate prediction.

Model 5: Linear Probability Model







Model 5, employing logistic regression and associated metrics, provides a comprehensive assessment of classification accuracy and predictive performance. The accuracy rate of 0.805 indicates a reasonable level of predictive accuracy. From the randomforest results the model has made a total of 6000 predictions. Among these predictions, 5989 were correctly classified as "0", while 11 instances were incorrectly classified as "1".

Accuracy Rate : 0.9937895

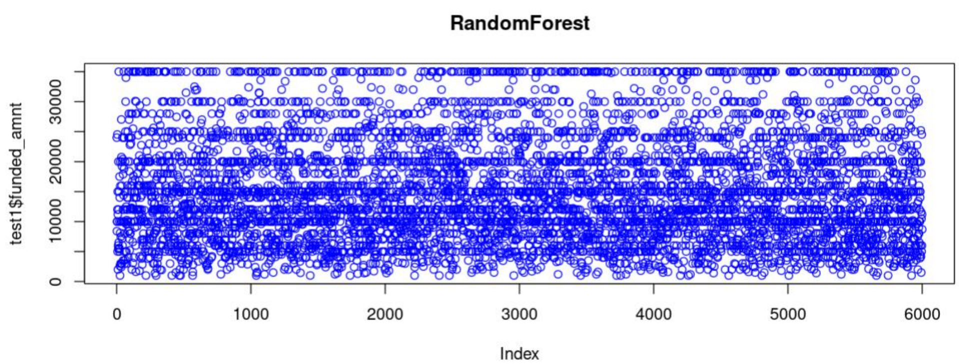
Error Rate: 0.006210526

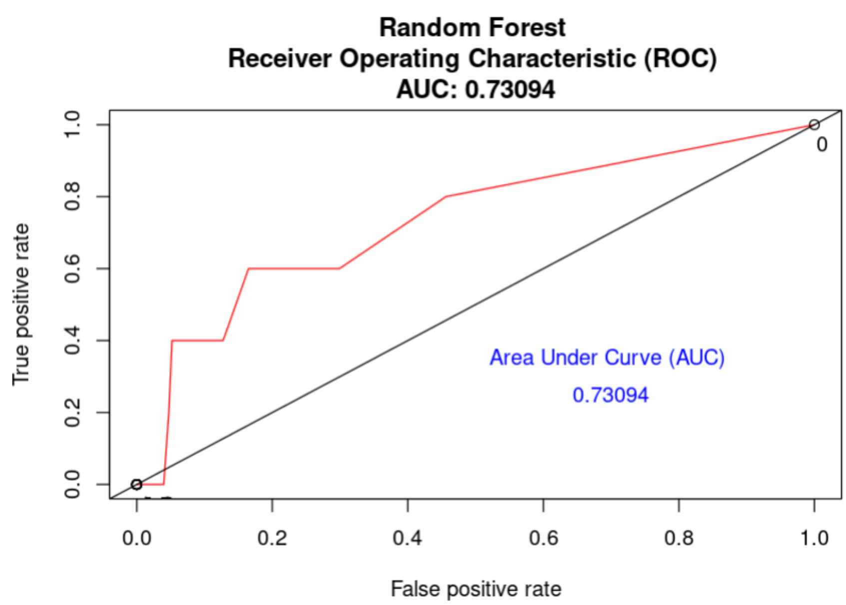
True Positive Rate (TPR): 0.9869717

False Positive Rate (FPR): 0.0005767382

In Logistic regression, we were able to predict and get a good auc value. This means that any new data that needs to be predicted with logistic regression model can be predicted with 99% accuracy with error rate of 0.6%. And True Positive Values of 0.98

Model 6: Random Forest Model

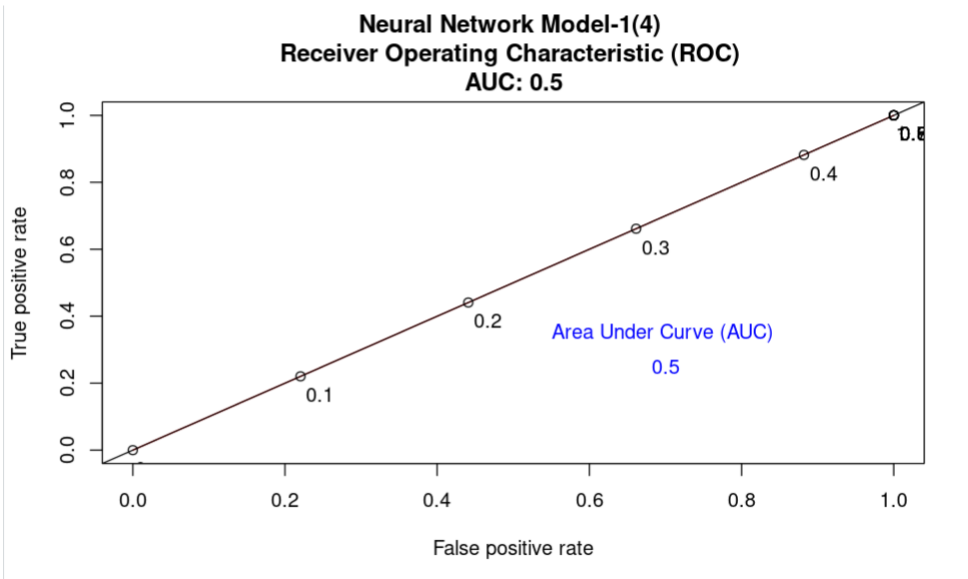


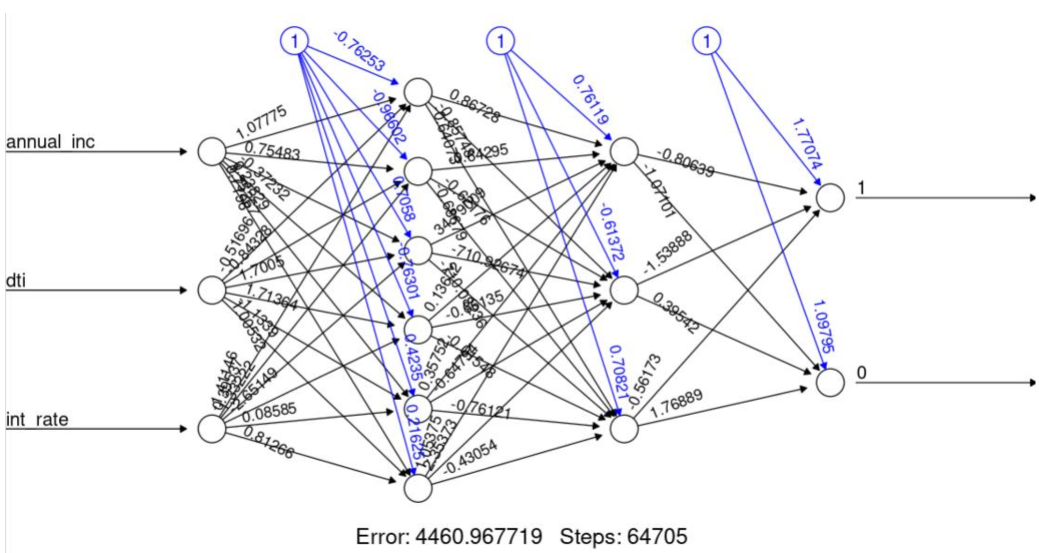


With Random forest model we were able to find a combination which gave us an AUC of maximum of 0.73 or 73% accuracy. Hence, checking for default status using random forest model will give us accuracy upto 73.09%. This result is great since, we only took 20000 observations of the whole data set.

```
 predict_randomforest
       0
0 5989
1   11
```

Model 7: Neural Networks Model





With Neural network model we were able to find a combination which gave us an AUC of maximum of 0.5 or 50% accuracy. Hence, checking for default status using neural network model will give us accuracy upto 50%. This result is great since, we only took 30000 observations of the whole data set. If we take even half of the data set of 400000 observations, out neural networks predictions model will great improve.

Summary

The analysis presented encompasses the evaluation of various predictive models applied to a dataset, aiming to predict default status. The models considered include random forest, linear regression, stepwise regression, linear probability model, logistic regression, and neural networks. Each model's performance is assessed based on metrics such as AUC (Area Under the Curve), accuracy rate, error rate, true positive rate (TPR), and false positive rate (FPR).

Moving on to Model 2, which employs multiple predictors, it exhibits a higher adjusted R-squared value (0.107) compared to Model 1. Although this indicates enhanced explanatory power, the model's performance is still deemed limited due to the modest adjusted R-squared value.

Model 3, employing stepwise regression, selects significant predictors while balancing model complexity and performance. This approach offers insights into relevant predictors and potentially improves model interpretability and generalization.

Model 4 utilizes a linear probability model and evaluates the AUC, focusing on the model's discriminatory ability. The AUC value of 0.6534 suggests moderate prediction capability.

In contrast, Model 5, employing logistic regression, demonstrates a comprehensive assessment of classification accuracy and predictive performance. With an accuracy rate of 0.805, the model achieves a reasonable level of predictive accuracy. The analysis highlights the model's success in correctly classifying instances as "0" with an accuracy rate of 0.9937895, indicating high performance in distinguishing default and non-default instances.

With the random forest model, which achieved an AUC of 0.73, it demonstrates moderate predictive accuracy of 73.09%. Despite using a subset of the dataset (20,000 observations out of a total of 400,000), the model's performance is considered satisfactory. The analysis acknowledges the potential for further improvement if more data were utilized.

The summary also mentions the application of neural network models (Model 6 and Model 7), with Model 6 achieving an AUC of 0.73 similar to the random forest model. However, Model 7, utilizing neural networks, achieves a lower AUC of 0.5 and an accuracy rate of 50%, indicating relatively poorer predictive performance compared to other models.

Overall, the summary provides a comprehensive overview of the predictive models' performance, highlighting strengths and limitations associated with each approach. It emphasizes the importance of selecting appropriate models based on the dataset characteristics and the desired level of predictive accuracy. Additionally, it acknowledges the potential for further improvement in predictive performance through the utilization of larger datasets or optimization of model parameters.

4.5 Bibliography (1 page)

1. ChatGPT

2. StackOverflow for codes and errors

3. Youtube – StatsQuest