Credit Risk Model

PD Estimation for a Housing Loan Portfolio

Internship under NIBM Vittarth

Tharun Kumar Gajula (120)

• Executive Summary:

There has been a growing demand in the housing sector which in turn increases the demand for housing loans which in reality forms a majority percentage of commercial banks' retail loan portfolios. So, there is an obvious need for developing proper credit risk assessment tools for housing loan customers which is the main focus of this project.

This project focuses on developing a credit scoring model for a housing loan portfolio of a bank. It is a type of loan evaluation technique to assess the creditworthiness of the customers using past data with the help of logistic regression which creates an equation that contains the probability of default as output. In this, when found significant variables data is inserted into the regression model, it provides us with a probability of default estimate of any test customer data or any other fresh customer data. The data consisted of demographic, financial, and other relevant data. First, we got the raw data of customers, which need to be analyzed and pick out the important parameters which can be useful in determining the default behavior of the customer. Next, the data is sorted, transformed, and cleaned so that, it is fit for running the logistic regression model. Divided the data into train and test data.

We have used MS Office Excel and STATA as tools for developing the model. After developing the model significant variables that can predict the default probability, we obtained the threshold for PD by using the KS statistic technique and then applied the limit to the train data separated. Then a confusion matrix technique is used to validate the model by finding the parameters like Prediction percentage, Recall percentage, and Accuracy Percentage. Then the model is used on the test data and the confusion matrix is drawn and again the same parameters are obtained that show the predictive power of the model. Also, the ROC curve is generated to validate the model and hence the power of the model. When the credit risk scoring model developed to attain the target accuracy, predictive power, and discriminatory power, banks can use the model for the fresh housing loan portfolio to assess the credit risk and minimize the losses.

- **Introduction:**

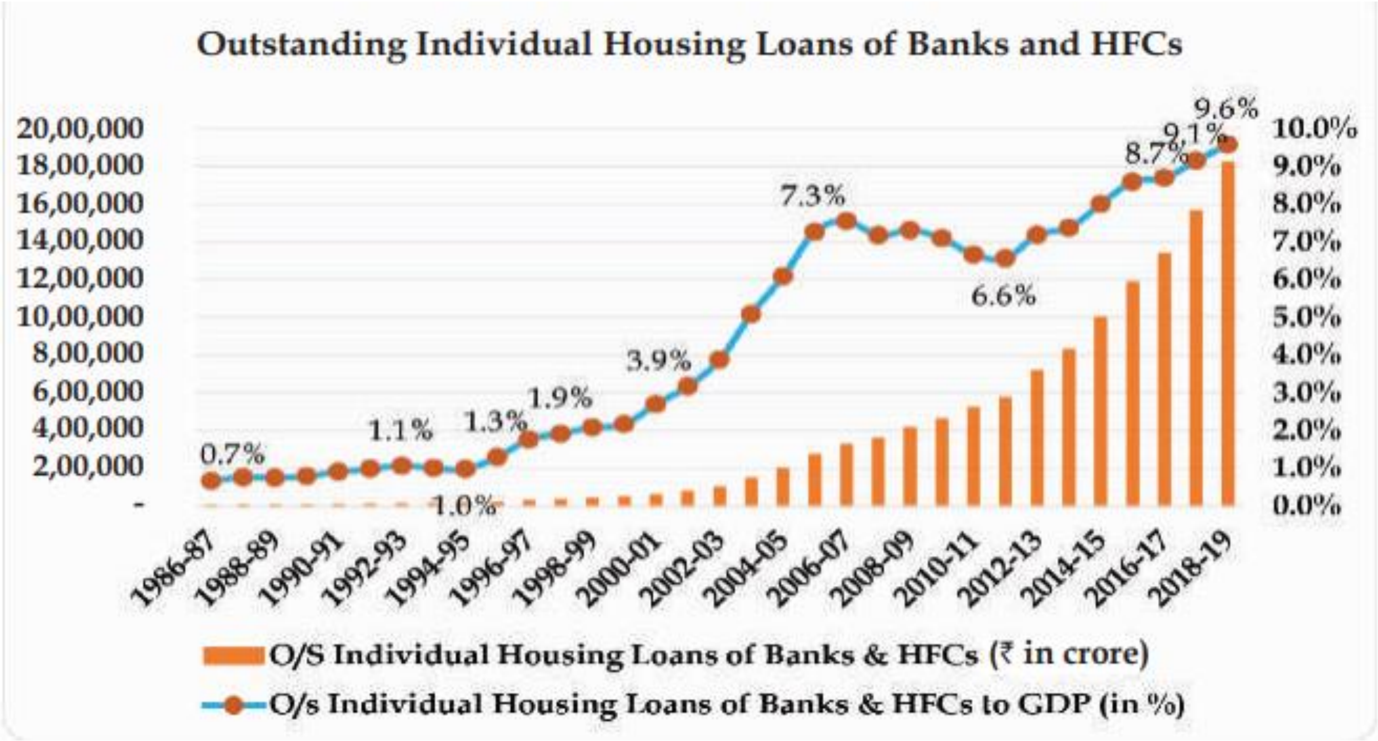
  The motivation for this project is to explore the default levels among the individual housing loan customers and also to explore the tools that can be deployed during lending practice by the banks to ultimately reduce the number of non-performing assets of the Indian commercial banks.

- **Current Scenario of Housing Finance Sector in Brief:**

  Outstanding housing loans as a percentage of GDP is a good measure of housing finance growth. This percentage has increased from 6.6% in 2010-11 to 9.6% in 2018-19. However, there was moderation in growth observed during 2018-19 because of the economic backdrop of the banking sector. The year 2018-19 has witnessed turmoil and skewed market sentiments in the NBFC and HFC sector leading to liquidity crunch and thereby limiting the growth. This increased HFCs cost of funds.

  This government has shown importance to this sector through its support in various schemes and policies, unlike previous governments. NHB also increased its refinancing limits in this 2018-19 year. Housing is the fourth largest contributor to Indian GDP and the sector has the potential to become the engine of domestic growth for the Indian economy in the coming years.

  In the backdrop of robust demand for affordable housing, measures announced by the RBI and Government of India, the housing finance sector continues to be sustainable on stable lines and we expect the housing finance sector to play a pivotal role in the growth story of the country.



- **Importance and relevance of the topic:**

Banking Business comes with various risk factors and it is inevitable. There are various risks like strategic risk, compliance risk, credit risk, cybersecurity risk, market risk, liquidity risk, Interest rate risk, operational risk, etc. Credit risk is of paramount importance for commercial banks which affects the profitability of the bank significantly. Credit risk understood simply as risk of borrowers failing to pay the dues in time which in turn results in losses to the bank. So banks give huge importance for credit risk management which is in turn more internalized in the recent periods so that it is customized to specifically for them and leverage the benefits of its efficiency.

As already mentioned, the housing finance industry's contribution, it is crucial for banks to assess the creditworthiness of housing loan customers.

Hence this relevance and importance is the motivation behind the topic of the project.

Benefits of scoring the borrower:

1. Facilitates risk management as already mentioned
2. Quantifies risk as a percentage chance (like Probability of Default) that something bad will happen.
3. The risk evaluation will be quick, systematic, unbiased and consistent

4. Factors linking risk and borrowers can be quantified
5. Limit sanctioning
6. Margin calculation
7. Risk-based pricing
8. Loan portfolio monitoring
9. Estimation of expected loss(EL) and unexpected loss(UL) which are used in the analysis of the adequacy of loan loss reserves or capital
10. Measuring potential concentrations in the loan portfolio
11. Credit allocation decisions and hence management of credit portfolio
12. Meeting regulatory requirements

• Literature review

With government support in the form of various measures, policies and other initiatives, the overall share of outstanding individual house loans of Banks and Housing Finance Companies to GDP (at market price) has increased from 6.6% in 2010-11 to 9.6% in 2018-19. Also, retail housing loan has been the largest asset class as it forms roughly around 60% in the total retail advances of HFCs, NBFCs, and SCBs (NHB Annual Report, 2018-19). So, it is important to put more reliable efforts in assessing the credit risk for this asset class.

The precision with which, all these institutions, evaluate the credit risk of loans (Housing loans in this case) affects not only the profitability but also the extent to which applications that would have been profitable but rejected (82 Fed. Res. Bull. 621, 1996).

Lowering the cost of serving a low-income consumer and at the same time increasing the service quality and customer satisfaction help to grow the retail house loan portfolio for a bank. Developing proper risk assessment tools like credit risk scorecards will have a significant impact on reliable portfolio growth, hence impacting profitability, service quality, and customer satisfaction positively (Maria Fernandez Vidal and Fernando Barbon, CGAP, 2019).

The trend is that banks are moving away from outsourcing the development of scorecards to building internally on their own. Following are some important factors which led to the widespread usage of scorecards and decision by banks to build them in-house. Increased regulation, ease of access to sizable and reliable data, creating value and boosting profitability, improved customer experience, efficiency and process improvement, availability of greater educational material and training……etc. Especially, banks that have opted to comply with Foundation or Advanced Internal Ratings Based Approach of BASEL II were required to generate Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD) internally. While Basel II implementation is ongoing in many countries, those banks gain most which sees it as not just a mandatory regulation exercise to follow but as a set of best practices leading to a path that can make their internal processes better (Naeem Siddiqi, Intelligent Credit Scoring).

Along with being a tool to assess levels of risk, risk scoring is also been effectively been applied to other areas such as streamlining the decision-making process, reducing turnaround time, economic and regulatory capital allocation, forecasting…..etc. Therefore, credit scoring provides lenders with consistent and objective decision making based on empirically derived information (Naeem Siddiqi, Intelligent Credit Scoring)

In a research paper by Robert, Raphael, Calem, and Canner, they described nicely about default in case of mortgage loans. A loan is in default as soon as the borrower misses the payment. There can four possible situations in case of a mortgage loan:

- A lender has been forced to foreclose a mortgage to gain title to the property securing the loan.
- The borrower chooses to give the lender title to the property in case of a foreclosure scenario.
- The borrower sells the home and makes less money than full payment on the mortgage obligation.
- The lender agrees to renegotiate or modify the terms of the loan and forgives some or all of the delinquent principal and interest payments (basically, Restructuring of the loan)

Based on different practices in the lending industry, not all the above situations are consistently considered as default by lenders. It depends on the lender, foreclosure process length, and other things following the situation.

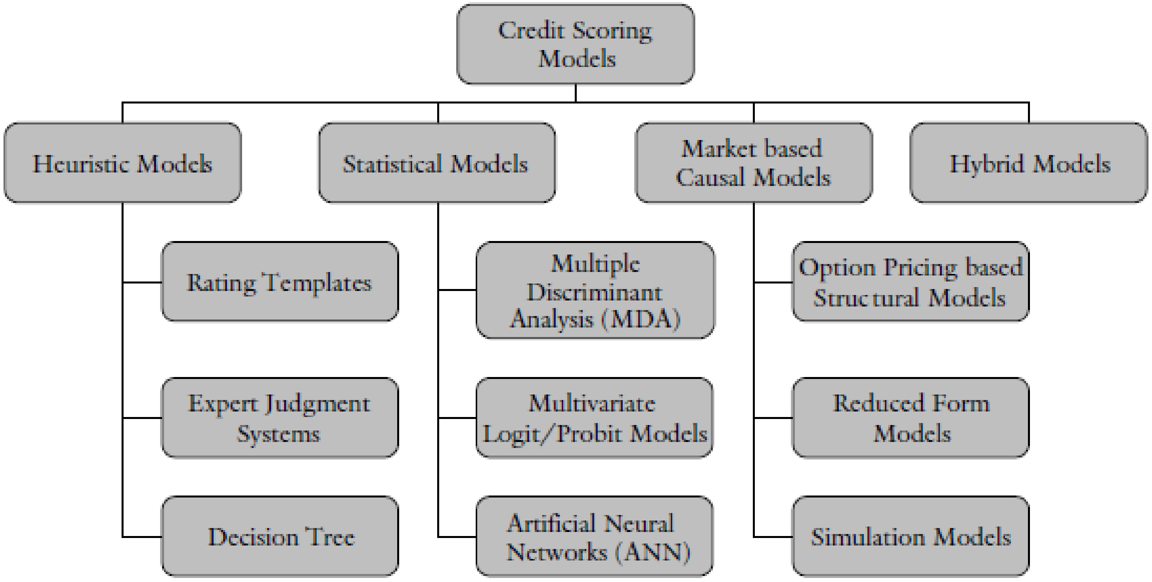
Credit Scoring is the set of decision-making models and techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrower to the lenders. Credit scoring has used the data on consumer behavior for the first time so it can be declared as the grandfather of data mining. Firstly, a lender should take two decisions in the credit approval process; one is whether to give loan to a fresh borrower; the technique that used to make this judgment is application credit scoring and, other, whether to increase the credit limits of the existing debtors; the techniques that assist the second decision are called behavioral credit scoring (Thomas, Edelman, and Crook).

Here, a proper credit scoring model can be used to assess the creditworthiness of a loan applicant by estimating the probability of default using statistical tools, based on past data. The Decision-making process for credit scoring can be subjective or statistical. Subjective scoring relies on qualitative judgment produced by experts. Statistical scoring relies on loan portfolio quantitative past data to forecast risk as a probability. Statistical models are especially useful in the case of banks with large volumes of credit assessments with relatively lower loan amounts, like retail loans for consumers and small businesses (Maria Fernandez Vidal and Fernando Barbon, CGAP, 2019).

Statistical scoring cannot replace the loan officers because ultimately the credit analysts must make the credit decision and these scoring techniques can act as a helpful guide. Statistical scoring reminds the credit manager of the elements of risks that they have ignored (Schreiner, 2002).

In this project, we will develop a credit scoring model using a technique called Logistic Regression which estimates the Probability of Default of the housing loan customers based on historic housing loan consumer data. Then PD can be used by the credit analysts or loan officers for decision making in lending to the customers.

Methodology



Credit Scoring Models
Heuristic Models
Statistical Models
Market based Causal Models
Hybrid Models
Rating Templates
Expert Judgment Systems
Decision Tree
Multiple Discriminant Analysis (MDA)
Multivariate Logit/Probit Models
Artificial Neural Networks (ANN)
Option Pricing based Structural Models
Reduced Form Models
Simulation Models

Source: OeNB (2004). Guidelines on Credit Risk Management-Rating Models and Validation. Oesterreichische National Bank (OeNB), Vienna, Australia.

In this project, we used the logit PD prediction model which is a statistical credit scoring model, in which logistic regression technique is applied. Statistical Softwares like SPSS, STATA, and SAS can be used as tools in this process. STATA along with MS Excel is used in the project. The logit model technique investigates the relationship between binary (default or not default) and explanatory variables (risk factors). The dependent variable is the natural log of odd ratio – default to non-default.

$$\log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- **Sampling techniques:**

  Random sampling technique has been used for dividing data into two samples
  Training sample: on which model developed
  Test sample: on which model is validated
  They are considered in 80:20 ratio

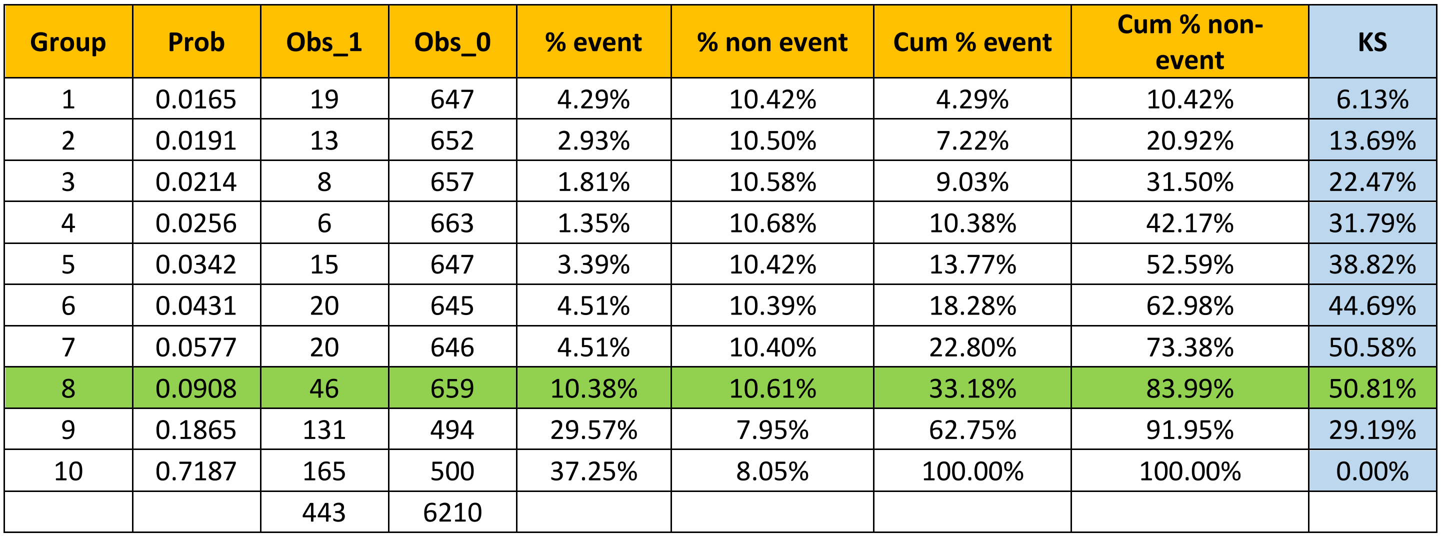
- **Process:**

- Raw data contained the following demographic, financial parameters, etc. It is sorted, transformed, cleaned, and made ready for further process.
- Binary Dependent Variable: If due payment date for the customer is greater than 90 days default (assigned '1') or else solvent (assigned '0').
- Some variables are transformed using the binning approach. For some, logarithmic values are considered and for others, categorical variables are considered.
- Missing values are treated accordingly by eliminating or replacing with mean values or other methods like binning. Extreme values (Outliers) are either eliminated or treated with the binning approach.
- Information value(IV) and Weight of Evidence(WOE) techniques used for the variable selection.
- IV and WOE are calculated for the sorted variables and selection of variables is done based on those variables for the final modeling run.
- Finally, a set of independent variables is finalized in one of the three forms – Numerical, categorical, or ordinal.
- Summary statistics of the data is prepared using STATA.
- Correlation with each other and also with the dependent variable are checked which if present can disturb the modeling process. A correlation matrix is prepared using STATA.
- Logistic Regression is performed in STATA
- Model performance is checked using various techniques like Kolmogorov Smirnov Statistic(KS), Receiving operating curve, and Confusion Matrix.

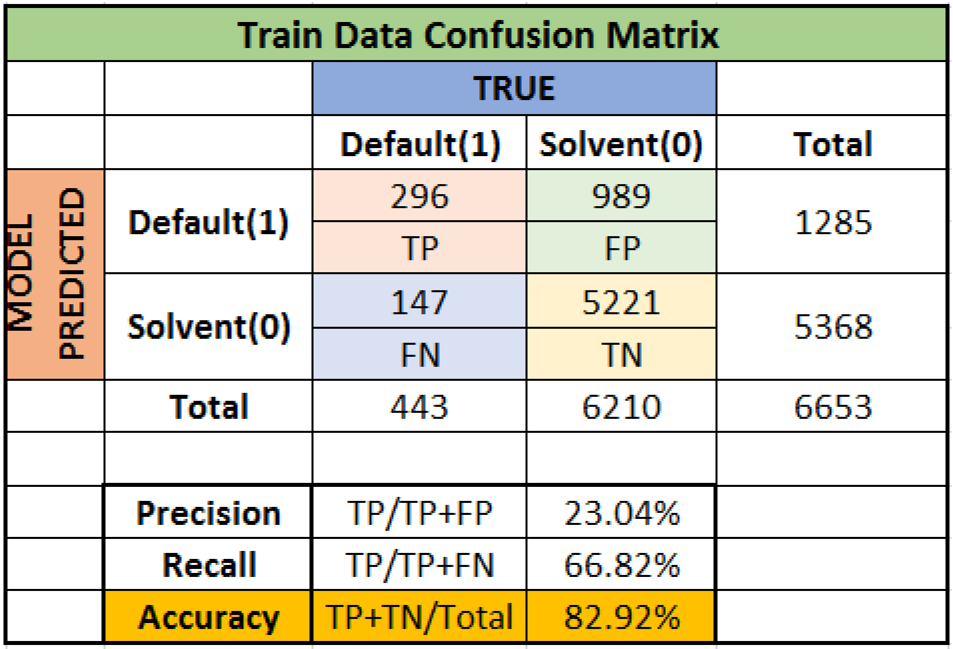
Variable	Description
ACID	Account ID
ACCT_OPN_DATE	Date on which loan account was opened
ACCT_CLS_DATE	Date on which loan account was closed
STATE_CODE	State to which the borrower belongs
DATE_OF_BIRTH	Date of birth of borrower
CUST_SEX	Gender of the customer
CUST_FIRST_ACCT_DATE	First time when borrower opened account with bank
OCCP_CODE_DESC	Occupation of borrower
ANNUAL_INCOME	Annual income of borrower
SRC_INCOME	Source of income
NET_WORTH	Net worth of borrower
DMD_OVDU_DATE	EMI due date (i.e. date on which installment is supposed to pay)
MAX_LAST_ADJ_DATE	Date on which installment was paid in actual
SANCT_LIM	Amount of loan sanctioned
SEC_VALUE	Value of security with bank for the loan
SHDL_NUM	No. of times the EMI schedule has been changed
FLOW_START_DATE	Date on which first installment is to be paid
LR_FREQ_TYPE	Frequency of payment of installment (monthly, quarterly, half yearly etc.)
NUM_OF_FLOWS	No. of EMI or installments to be paid
FLOW_AMT	EMI value i.e. amount of single instalment
NUM_OF_EMI_PAID	No. of installments already paid till date
NEXT_DMD_DATE	Next due date for the instalment

	Variable	Description	Variable type considered	Dependent or Independent
1	**DEFAULT**	DEFAULT Variable (0 - Solvent, 1 - Default)	Categorical	Dependent
2	**LTV**	Loan Sanctioned Amount to Security Value Amount	Numerical	Independent
3	**MORATM**	Time between account opening date and flow start date	Ordinal	Independent
4	**GENDER**	Gender of the customer (Male = 0, Female = 1)	Categorical	Independent
5	**C_STAT**	First account opened for loan purpose = 0, Otherwise = 1	Categorical	Independent
6	**ANN_INC**	Annual Income	Ordinal	Independent
7	**NETWORTH**	Networth of the customer	Ordinal	Independent
8	**STATE**	State to which customer belongs to	Categorical	Independent
9	**TENURE**	EMI or Flow Amount	Ordinal	Independent
10	**LN_EMI**	Logarithm of flow amount	Numerical	Independent
11	**AGE**	Age of customer taken as ordinal variable	Ordinal	Independent
12	**LN_AMNT**	Logarithm of sanctioned loan amount	Numerical	Independent

VARIABLE	IV Value	
GENDER	0.0020	Useless
C_STAT	0.0706	Weak
LTV	0.0023	Useless
ANN_INC	0.4627	Strong
NETWORTH	0.0724	Weak
AGE	0.0217	Weak
EMI	0.9697	Suspicious
MORATRM	0.1698	Medium
TENURE	0.2206	Medium
STATE	0.5098	Strong
IV Category	**Strength of Predictor**	
<0.02	Useless	
0.02-0.1	Weak	
0.1-0.3	Medium	
0.3-0.5	Strong	
>0.5	Suspicious	

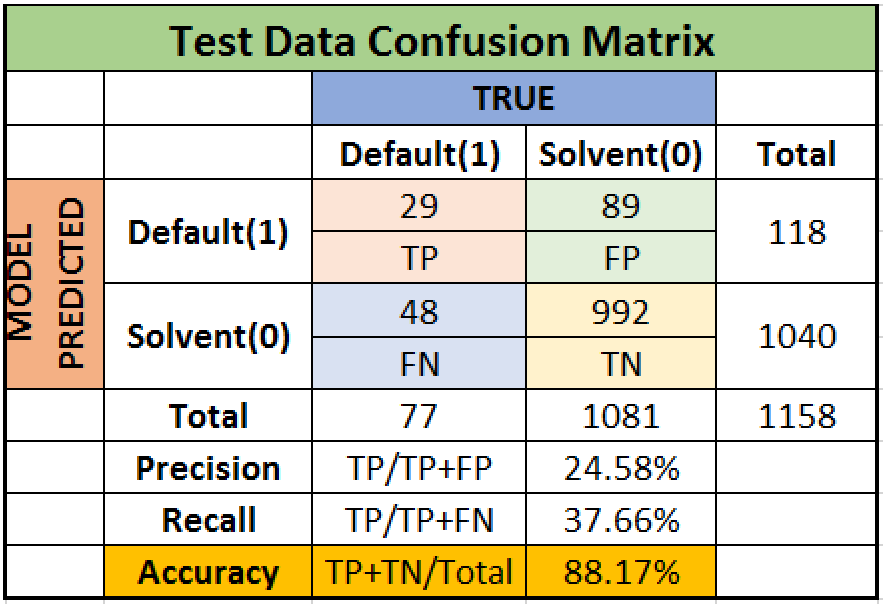


Group	Prob	Obs_1	Obs_0	% event	% non event	Cum % event	Cum % non-event	KS
1	0.0165	19	647	4.29%	10.42%	4.29%	10.42%	6.13%
2	0.0191	13	652	2.93%	10.50%	7.22%	20.92%	13.69%
3	0.0214	8	657	1.81%	10.58%	9.03%	31.50%	22.47%
4	0.0256	6	663	1.35%	10.68%	10.38%	42.17%	31.79%
5	0.0342	15	647	3.39%	10.42%	13.77%	52.59%	38.82%
6	0.0431	20	645	4.51%	10.39%	18.28%	62.98%	44.69%
7	0.0577	20	646	4.51%	10.40%	22.80%	73.38%	50.58%
8	0.0908	46	659	10.38%	10.61%	33.18%	83.99%	50.81%
9	0.1865	131	494	29.57%	7.95%	62.75%	91.95%	29.19%
10	0.7187	165	500	37.25%	8.05%	100.00%	100.00%	0.00%
		443	6210					



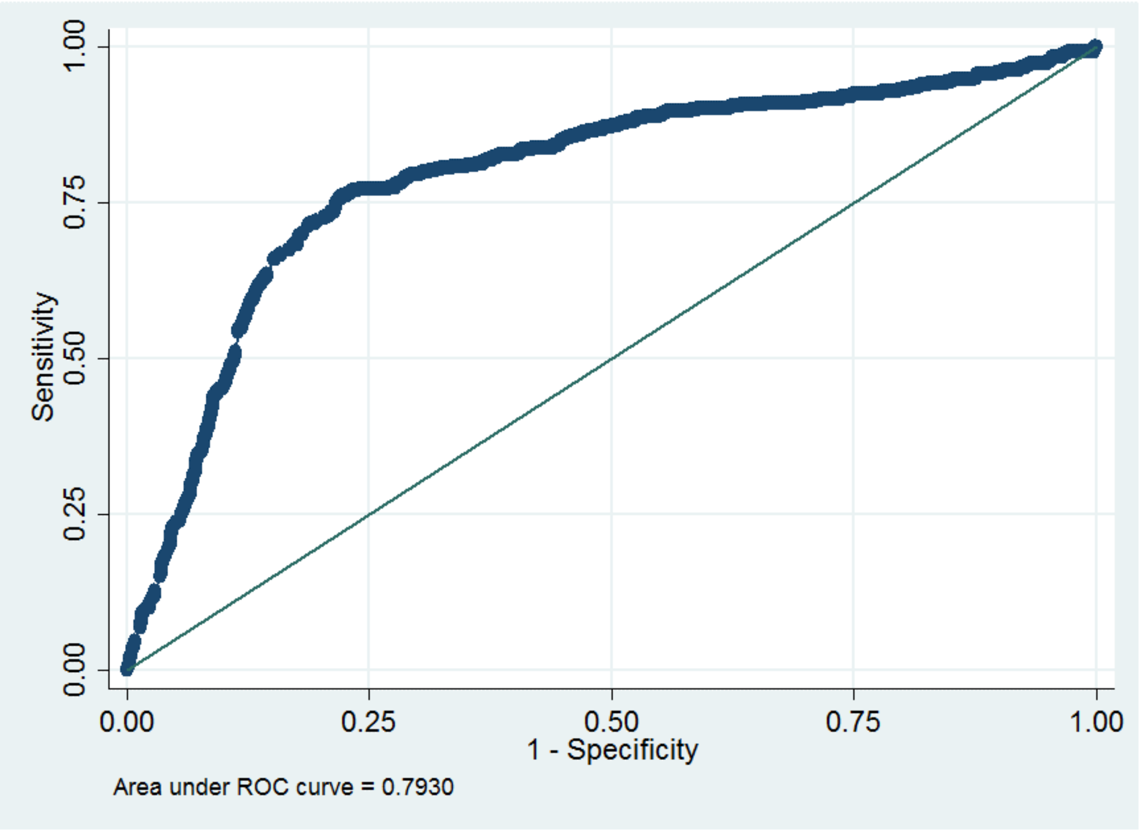
Train Data Confusion Matrix

		TRUE		
		Default(1)	Solvent(0)	Total
MODEL PREDICTED	Default(1)	296 TP	989 FP	1285
	Solvent(0)	147 FN	5221 TN	5368
	Total	443	6210	6653
	Precision	TP/TP+FP	23.04%	
	Recall	TP/TP+FN	66.82%	
	Accuracy	TP+TN/Total	82.92%	



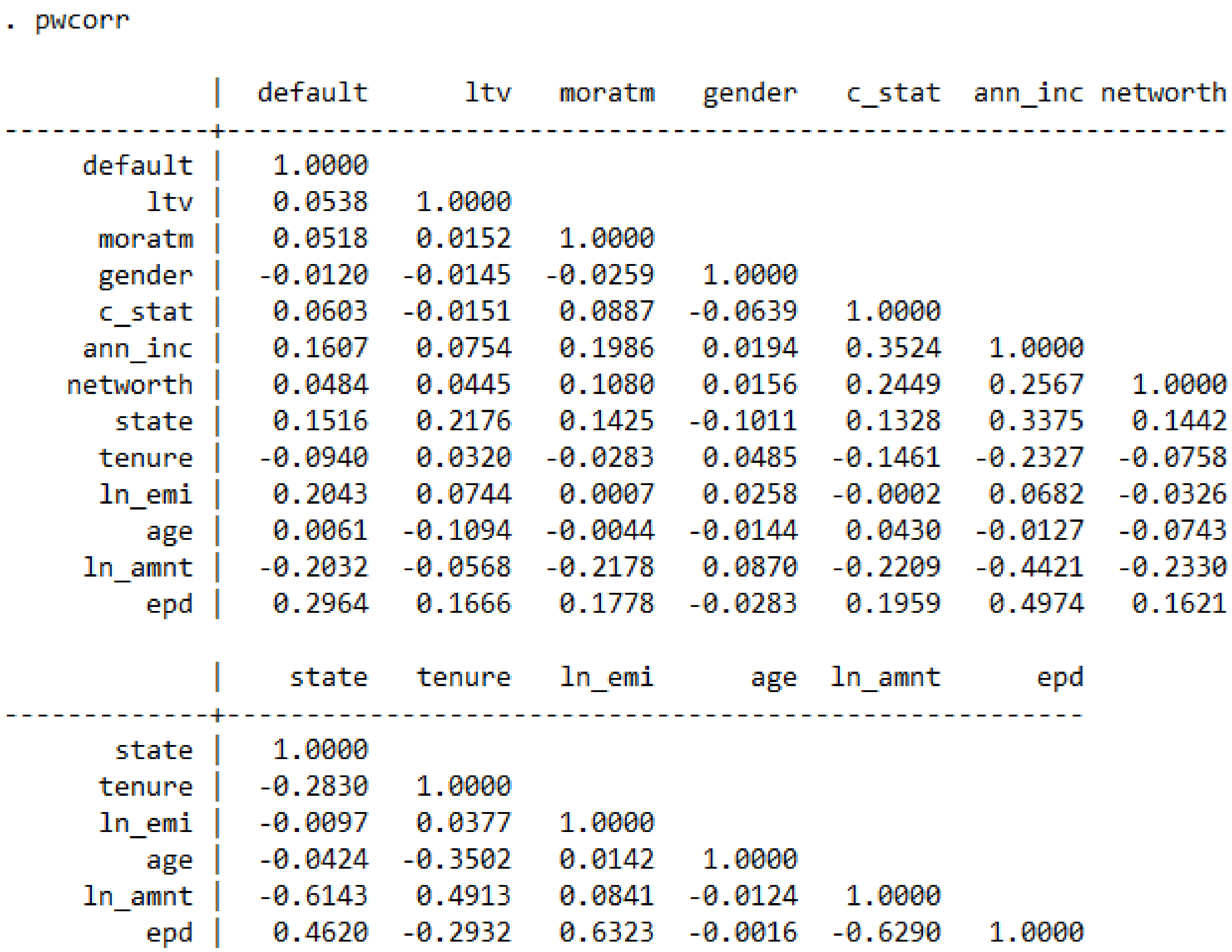
Test Data Confusion Matrix

		TRUE		
		Default(1)	Solvent(0)	Total
MODEL PREDICTED	Default(1)	29 TP	89 FP	118
	Solvent(0)	48 FN	992 TN	1040
	Total	77	1081	1158
	Precision	TP/TP+FP	24.58%	
	Recall	TP/TP+FN	37.66%	
	Accuracy	TP+TN/Total	88.17%	



Area under ROC curve = 0.7930

Correlation Matrix

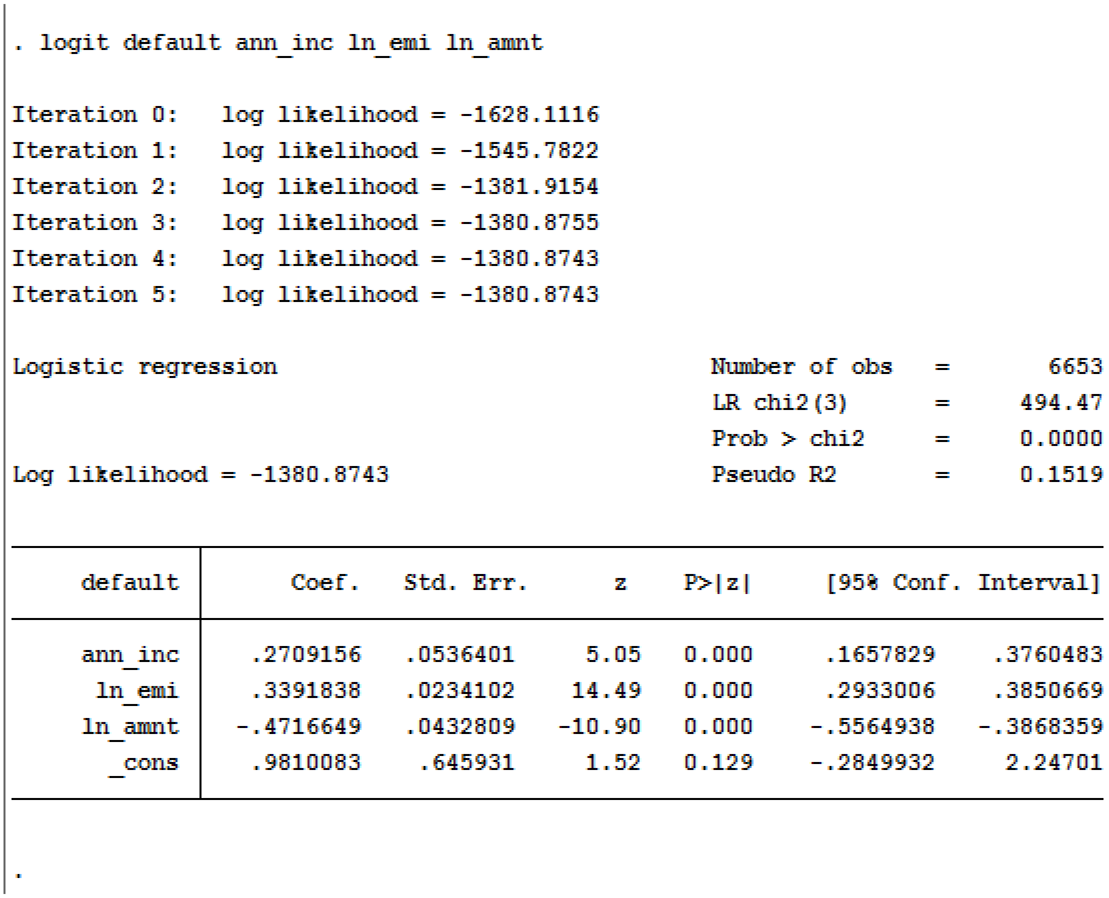


```
. pwcorr

             |  default      ltv   moratm   gender   c_stat  ann_inc networth
-------------+---------------------------------------------------------------
     default |   1.0000 
         ltv |   0.0538   1.0000 
      moratm |   0.0518   0.0152   1.0000 
      gender |  -0.0120  -0.0145  -0.0259   1.0000 
      c_stat |   0.0603  -0.0151   0.0887  -0.0639   1.0000 
     ann_inc |   0.1607   0.0754   0.1986   0.0194   0.3524   1.0000 
    networth |   0.0484   0.0445   0.1080   0.0156   0.2449   0.2567   1.0000 
       state |   0.1516   0.2176   0.1425  -0.1011   0.1328   0.3375   0.1442 
      tenure |  -0.0940   0.0320  -0.0283   0.0485  -0.1461  -0.2327  -0.0758 
      ln_emi |   0.2043   0.0744   0.0007   0.0258  -0.0002   0.0682  -0.0326 
         age |   0.0061  -0.1094  -0.0044  -0.0144   0.0430  -0.0127  -0.0743 
     ln_amnt |  -0.2032  -0.0568  -0.2178   0.0870  -0.2209  -0.4421  -0.2330 
         epd |   0.2964   0.1666   0.1778  -0.0283   0.1959   0.4974   0.1621 

             |    state   tenure   ln_emi      age  ln_amnt      epd
-------------+------------------------------------------------------
       state |   1.0000 
      tenure |  -0.2830   1.0000 
      ln_emi |  -0.0097   0.0377   1.0000 
         age |  -0.0424  -0.3502   0.0142   1.0000 
     ln_amnt |  -0.6143   0.4913   0.0841  -0.0124   1.0000 
         epd |   0.4620  -0.2932   0.6323  -0.0016  -0.6290   1.0000 
```

Logistic Regression Model Output



```
. logit default ann_inc ln_emi ln_amnt

Iteration 0:   log likelihood = -1628.1116
Iteration 1:   log likelihood = -1545.7822
Iteration 2:   log likelihood = -1381.9154
Iteration 3:   log likelihood = -1380.8755
Iteration 4:   log likelihood = -1380.8743
Iteration 5:   log likelihood = -1380.8743

Logistic regression                             Number of obs     =      6653
                                                LR chi2(3)        =    494.47
                                                Prob > chi2       =    0.0000
Log likelihood = -1380.8743                     Pseudo R2         =    0.1519

------------------------------------------------------------------------------
     default |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ann_inc |   .2709156   .0536401     5.05   0.000     .1657829    .3760483
      ln_emi |   .3391838   .0234102    14.49   0.000     .2933006    .3850669
     ln_amnt |  -.4716649   .0432809   -10.90   0.000    -.5564938   -.3868359
       _cons |   .9810083    .645931     1.52   0.129    -.2849932     2.24701
------------------------------------------------------------------------------

.
```

Variable	Coefficient
ann_inc	0.2709156
ln_emi	0.3391838
ln_amnt	-0.4716649

Learnings and conclusion:

Learnings and findings: Data is very important in developing credit scoring models.
Important considerations about data:

1. Sufficient data size. Sufficient data size makes the model more reliable. At first, in this project, there were 7499 observations for the considered variables and performed the entire process. Then additional data of 1189 observations were included and again the entire process carried out. After including the additional data, the obtained model was better than the previous one.
2. Sufficient proportion of default observations in the data. The effect of this proportion was also observed after adding additional data which contained many default observations. Consequently, the model was better after that.
3. The Binning approach is very important in transforming the variables, which if done properly can increase the explanatory power of the independent variables. It was useful in cases of missing data, outliers(extreme values). It makes sure the model is less disturbed by such cases. Also, it is very important to follow a logical trend while binning the variables.
4. Sufficient test data also important to test the power of the model.
5. Validating the model using various techniques is very important to know the power of our model. It quantifies the power of the model by using various parameters and can be used to modify the model to develop a better one.

The logistic regression model developed from the data showed around 83 % accuracy on training data and 88 % accuracy on test data, ROC of 80 % .This can be improved by considering more different customer attributes, more data, proper data engineering techniques, etc.

Scope for further research:

For further research projects, advanced credit scoring techniques can be explored like genetic algorithms, fuzzy discriminant analysis, and neural networks. For making a model, a more generalized and accurate one, large data of borrowers is recommended. Variables types of new data are available in this age of big data and considering such data and exploring advanced techniques to evaluate, can increase the power of the model. More data of rejected applications from the bank is also of high importance.

- ***REFERENCES:***

1. *Managing Portfolio Credit Risk by Arindam Bandyopadhyay*
2. *Intelligent Credit Scoring Models by Naeem Siddiqi*
3. *Recent developments in consumer credit risk assessment – European Journal by Jonathan Crook, David Edelman and Lyn Thomas*
4. *Risk management model – Euro Journal by Vasanthi Peter and Raja Peter*
5. *Forecasting creditworthiness of individual borrowers – International Journal by Asia Samreen and Fahreen Batul Zaidi*
6. *NHB Annual Report 2018-19*
7. *NIBM Vittarth data*