Presented by:

Anudnya Ware

Gifty Reji

Soham Pawar

Vaishnavi Mayekar

**Acquisition Scorecard**

Developing an Acquisition Scorecard for a Bank which predicts whether or not to give a Loan to a Customer.

Contents

[BACKGROUND 2](#_Toc78970546)

[Reason for modelling Scorecard 5](#_Toc78970547)

[Benefits of scorecard system 5](#_Toc78970548)

[Objectives 6](#_Toc78970549)

[Ideal Nature of the Data 8](#_Toc78970550)

[Limitations 12](#_Toc78970551)

[Methodology 13](#_Toc78970552)

[Deliverables 15](#_Toc78970553)

[Model Accuracy 15](#_Toc78970554)

[Lift Table 15](#_Toc78970555)

[Waterfall Analysis 15](#_Toc78970556)

[ROC Curve 15](#_Toc78970557)

[Trend Interpretation 15](#_Toc78970558)

[Conclusion 16](#_Toc78970559)

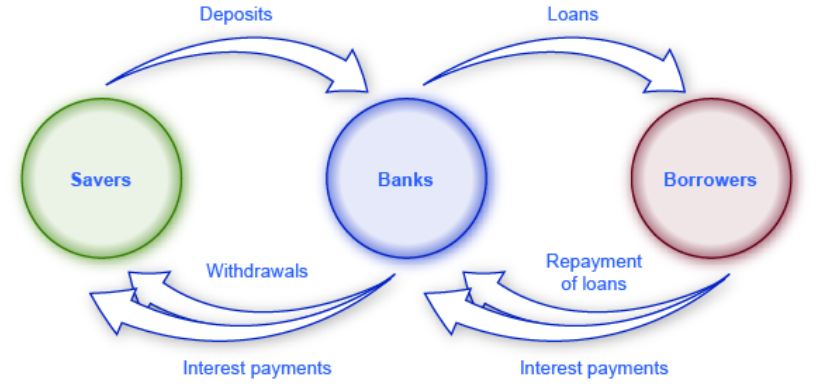
# BACKGROUND

Banks play an important role in the economy by offering services for people to save, to finance businesses who wish to invest and expand. These loans and business investments are necessary for enabling economic growth.

People borrow corporate and non-corporate loans from banks. Hence, banks act as a support for most industrialists, small businessmen, and even salaried individuals. In return, the bank makes a profit by payments of interest rate that is applicable on these loans.

*How does a bank provide loan and earn profit at the same time?*

It’s a loop! A bank has several customers who either own a savings account, a current account, or both. The customer deposits the money into the account for which the bank pays them some interest. At the same time bank provides loans to the borrowers and in return charges interest on the base amount of the loan. This is how the money flows.

  
When the borrower is in default and has not paid the monthly principal and interest repayments for a specified period, such loans are called non-performing loans (NPL).

Before the concept of Credit Scorecard surfaced, lending decisions were in the hands of loan officers or underwriters who worked at banks. Customers had to come down to the bank and discuss their need for a loan with the loan officer, and the latter would take a call on whether or not the customer is worthy of a loan based on their meeting.

This approach was centred on individual judgement, hence not being the fairest way to carry out banking decisions. This invited room for error in judgement as well as discrimination in terms of race, gender, etc.

The need to overcome the negative outcomes of such subjective lending decisions paved the way for the Credit Scorecards. Credit Scorecard is a risking tool used to assess the level of risk associated with the customers applying for a loan. Over time Scorecard system were adapted into the world of finance. Soon enough, leading credit bureaus like Experian, CIBIL, etc formulated Credit Reports which incorporated the Credit Scorecard concept.

Credit Information Bureau (India) Limited (CIBIL) is a credit bureau engaged in maintaining the records of all the credit-related activities of companies as well as individuals, including credit cards and loans. It provides Credit Information Report of a customer to the banks to quickly and efficiently filter the loan applications which they receive in the course of their business.

Ultimately, the need for bank to have an acquisition scorecard is to predict the possibility that one might or might not be a good customer was generated.

Acquisition scorecard can be created for predicting the customer’s behaviour on the basis of CIBIL score, KYC, debit score, credit score, internal performance and previous loan instalment history.

Credit scoring in India provides a useful framework for considering the social and ethical consequences of algorithmic decision-making more broadly.

India highlights how in an emerging economy with relatively weak institutions and low financial literacy, credit scoring through alternate data creates the possibility for rapid progress in financial inclusion but under weaker consumer protection standards.

In the past five years, the banks have started using this tech element that is much faster, paperless, and user-friendly.

* India's ICICI Bank is using satellite imaging to assess the creditworthiness of farmers applying for loans. ICICI has worked on further scoring models to create indices at the district level, village level as well as individual land to provide an estimate of the past and future agriculture income, the timing of harvest, and sources of income, to deliver detailed inputs to credit assessments.
* Bank of Baroda NSE, Union Bank of India NSE, and Syndicate Bank have taken the first steps in transparently segregating retail loans into their versions of prime and subprime risk exposure, using third-party credit scores of potential borrowers to offer them different home-financing rates.
* TransUnion CIBIL recently launched a new algorithm to assess first-time creditors, or new to credit (NTC) borrowers as they are known in banking parlance called CreditVision NTC Score which will help the bank assess an NTC consumer’s eligibility.

A credit invisible is someone having no credit activity for more than 2 years or no credit exposure on his/her name. Fundamentally this is not a bad situation from banks to be in. Some banks may reject the application, while a few banks or NBFCs might still decide to lend based on a credit appraisal or on other factors including the salary/income proofs, educational background, employer etc. of the prospective borrower.

In addition to the scorecard for new borrowers from various credit agencies, banks have started using various surrogate parameters such as:

1. Balance in Savings Bank Account.
2. No cheque bounces
3. Begin with small secured loan
4. Employment status

To decide whether or not to give a loan, a customer is categorized a customer as good or bad on the basis of these factors-

**Credit History**

Bank will want to review the customer's personal credit history to check for the inaccuracies in payments that will directly affect the approval of the loan application. In the case of credit history, the bureau helps the bank by providing a detailed Credit Information Report (CIR), assigning a score to the customer on the basis of their performance.

**Cash Flow History/ Debit History**

Debit history is considered to understand how the money flows in and out of the customer’s account. The amount that is saved and then the expenditure of the customer on a monthly basis is analyzed. This acts as an important feature when the customer is applying for a personal loan.

**Providing Collateral to Secure a Loan**

When it comes to obtaining a secured loan, providing collateral is a must. To a bank, collateral is simply defined as property that secures a loan or other debt, so that the bank may seize that property, if customer fail to make proper payments on the loan.

When lenders demand collateral for a secured loan, they are seeking to minimize the risks of extending credit. In order to ensure that the particular collateral provides appropriate security, the lender will want to match the type of collateral with the loan being made. To further limit their risks, banks usually discount the value of the collateral so that they are not extending 100 percent of the collateral's highest market value.

**Income tax returns and a business plan**

Bank will want to know how an applicant plans to use the money and have a strong ability to repay. They may require a solid business plan that details the purpose of the loan and how the customer expect it to increase profits. A detailed business plan is needed when the applicant wants a loan to fund his start-up, or an already-existing company. Similarly, non-salaried people have to submit income tax returns. These documents, will help the lender to understand your business, the nature and extent of existing borrowings, profitability of the business and quantum of own investment. These documents, will also help the lender to understand your saving habits.

**Character**

Banks look for customers that have paid their previous bills on time, have good relationships with other vendors, and are current with the Indian Revenue Service (IRS). Bank will want to conduct business with and make loans to people they can trust to act in good faith at all times – in good and bad. They will also want to talk with previous business contacts to determine how you and your business have conducted business in the past.

## Reason for modelling Scorecard

As and when more and more customers are coming in, the underwriter would get piled up with a lot of applications and the customer would typically want a quick loan. So therefore, at this stage we use Application Scorecard, completely data driven backed by statistical understanding which can identify which customers to be given a loan or not. This way, a good number of customers can skip the underwriting stage and can directly go into operation.

Computation of Scorecard System are not intended to replace underwriters or loan officers or other commercial staff, but rather to complement and facilitate their work by supporting assessment of willingness to pay.

## Benefits of scorecard system

Improves accuracy of credit decisions by minimizing rejection of credit to creditworthy applicants and maximizing rejection of high-risk applicants.

Scorecard enables the bank to Automate the application decision processes which reduces the time to approve or reject the loan application.

The scorecard yields a numeric score provided for each applicant with higher scores corresponding to lower levels of estimated risk.

It facilitates the ability of financial institutes to make accurate, consistent, fact-based decisions. Therefore, Accuracy of Acquisition Scorecard is higher as the prediction is only for the specific population.

Scorecard system reduces operating expenditure and increases the profit margin. Thus, boosting the economy of the bank and increase in customer base.

With the promise of fast and automatic decisions, bank reply on having automatic and fast credit risk models to assess risk. We will be introducing a robust model used in the process of creating an acquisition scorecard.

# Objectives

The project aims to develop an Acquisition scorecard that will understand the various aspects required to approve a loan and predict the probability of the customer to be good.

In order to achieve the target, we need,

* To quantify the creditworthiness of customers by assessing the risk.
* To reduce turnaround times for customers for processing applications through automated decision making.
* To identify the percentage of population that can be directly approved or rejected without going through the underwriting stage.

**To quantify the creditworthiness of customers by assessing the risk.**

Creditworthiness of a customer is determined by the capacity to pay back credit obligations in a timely manner.

* BAD customers: The applicants who have defaulted or are at a 90-day delinquency.
* GOOD customers: Applicants who have completed all payments successfully. Also considering those who have completed 90% / 95% of their payments as well.
* Indeterminate customers: Clients who are still in the early stages of loan repayment and could potentially turn to be good or bad. They are usually not considered for the model building process.

Based on the proportion of good applicants to bad applicants at each group level, this method measures the “strength” of grouping for differentiating good and bad customer.

What variables determines the categorisation of the applicant into good or bad?

It is believed that there is no optimal number of variables that should be used in building scoring models. In turn, the selection of the variables in building scorecard depends on the data providers (Internal, external, bureaus). For each attribute, the scorecard assigns a number of points which contribute to an overall credit score.

**To reduce turnaround times for customers for processing applications through automated decision making.**

Turnaround time is the time from the application of a loan by a customer till the final disbursal of the loan amount.

Due to heavy competition, the opportunities for new businesses open for short time periods. Hence for assessing customer's profile, the turnaround time should be reduced which helps to avoid losing good customers.

Sometimes even with the collateral, banks face loss-making business. Therefore, selecting credible clients and keeping out the potentially delinquent ones becomes the primary objective of the lender.

This calls for building an automated risk scorecard that can process thousands of clients within a few minutes and distinguishes a “good” client from a “bad” one.

**To identify the percentage of population that can be directly approved or rejected without going through the underwriting stage.**

Underwriting is the process through which an individual or institution takes on financial risk for a fee. By measuring the probability of default, we can identify the percentage of the population that can be directly approved or rejected without going through the underwriting stage and can directly go into operation.

One of the successful and transparent way to achieve this is by applying logistic regression function for calculating the probability to default.

In general terms, the model will be tested by splitting the data into two parts. 20% of data is training data which will give us training error and 80% of testing data which will give us testing error.

A model can be created by using machine learning algorithms like Random forest, XGBoost or Logistic Regression which learns from the training dataset to distinguish between “good” and “bad” on the testing dataset.

# Ideal Nature of the Data

Acquisition scorecard models are used for making approval or reject decisions. Successful performance of a scorecard depends mostly on the data used during the development stage.

When a customer walks in for a loan, it can be either an internal customer or external customer and for each such customers, we are supposed to study information such as Income, Previous History, Investment Details, Spending Habits, demographics, social media and so on.

Considering different Data sources (we will be handling secondary dataset):

Banks collect Data mostly in 3 different forms

***Data bank has:*** Data which is provided by RBI and Credit Bureaus such as CIBIL, Bank collects data such as credit history, credit scores from these organization. Also, If it is an Internal customer, The banks may already have their credit and debit card spending history.

***Data collected directly from Customer:*** Bank can collect enormous amount of data like withdrawals of ATM, online payments, KYC etc

***Data Collected from Third Party:*** Bank gather information from third party organizations such as Facebook and Google. Client’s social media helps to understand about the tastes, preferences, and needs of their customers. Various telecom service providers also help banks to check their billing history, their locations, etc.

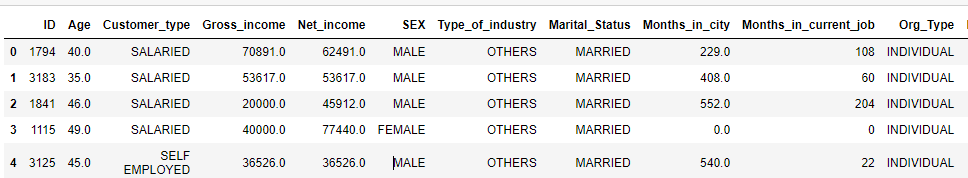
Acquisitions scorecards are used only for loan approvals. There are different types of loans that the bank can avail such as home loan, Property Loan, Education Loan, etc. and for each such products or product segments, we can implement different Acquisition Scorecards. This is because each product is dynamically different from each other and have a different set of policy rules.

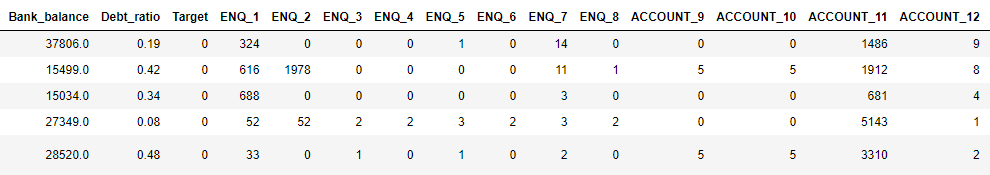
Data which we have received consisted of **3980** observations with **35** variables. Some of them were numeric as well as categorical. It also had some missing values which were replaced by mean or mode depending on the data type.

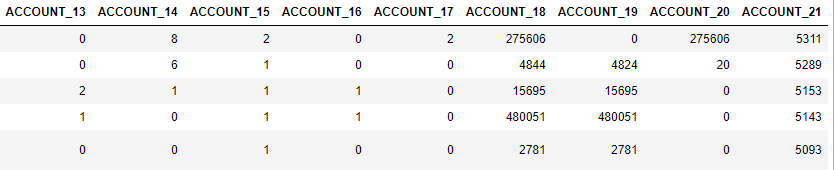
The names of the 35 variables are as follows:

|  |  |  |
| --- | --- | --- |
| Sr.no | Name | Description |
| 1 | ID | Customer ID |
| 2 | Age | Age of the Customer |
| 3 | Customer\_type | Salaried / Self Employed |
| 4 | Gross\_income | Gross income declared by Customer |
| 5 | Net\_income | Net income declared by Customer |
| 6 | SEX | Gender |
| 7 | Type\_of\_industry | Type of industry |
| 8 | Marital\_Status | Marital Status declared by Customer |
| 9 | Months\_in\_city | Months in city |
| 10 | Months\_in\_current\_job | Months in current job |
| 11 | Org\_Type | Org Type declared by Customer |
| 12 | Bank\_balance | Bank balance declared by Customer |
| 13 | Debt\_ratio | Debt ratio declared by Customer |
| 14 | Target | Default or not. |
| 15 | ENQ\_1 | Days since last enquiry was made with any other bank |
| 16 | ENQ\_2 | Days since last Home loan enquiry made with any other bank |
| 17 | ENQ\_3 | Total enquiries in last 3 months |
| 18 | ENQ\_4 | Total home loan enquiries in last 3 months |
| 19 | ENQ\_5 | Total enquiries in last 12 months |
| 20 | ENQ\_6 | Total home loan enquiries in last 12 months |
| 21 | ENQ\_7 | Total enquiries |
| 22 | ENQ\_8 | Total home loan enquires |
| 23 | ACCOUNT\_9 | Number of defaults in last 3 months |
| 24 | ACCOUNT\_10 | Number of default in last 12 months |
| 25 | ACCOUNT\_11 | Days since last account was open |
| 26 | ACCOUNT\_12 | Total number of loans with other bank ( live + closed ) |
| 27 | ACCOUNT\_13 | Total number of home loans with other bank ( live + closed ) |
| 28 | ACCOUNT\_14 | Total number of unsecured loans with other bank ( live + closed ) |
| 29 | ACCOUNT\_15 | Total number of live loans |
| 30 | ACCOUNT\_16 | Total number of live home loans |
| 31 | ACCOUNT\_17 | Total number of live unsecured loans |
| 32 | ACCOUNT\_18 | Total outstanding amount |
| 33 | ACCOUNT\_19 | Total secured outstanding amount |
| 34 | ACCOUNT\_20 | Total unsecured outstanding amount |
| 35 | ACCOUNT\_21 | Average number of days the individual is in debt |

Here is the glimpse of the data







As we can see Variables Having Numerical Data types are Age (of customer), Bank Balance (of Customer), Income (of Customer), Enquires it is further divided into 3 months and 12 months enquires, Account, No of Default with respect to current and Previous Loans.

As we can see Variables Having Categorical Data type are Customer Type (Salaried, Self-Employed, Non-Earning), Marital Status, SEX, Industry (Agriculture, Banking, Cement) etc.

And the important variable is the Target Variable having two values 0 and 1 where 0 denotes non-Defaulter and 1 denotes Defaulter.

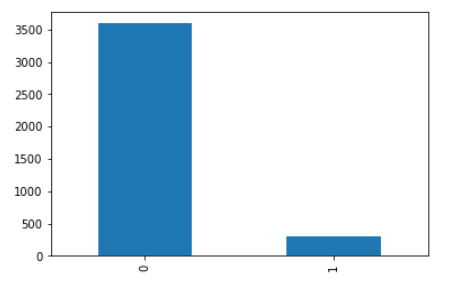
We have also performed some pre-processing on the data such as dropping duplicate values and unwanted columns such as *Months in city and Months in current job*.

There were 6 columns that had missing records and the bank balance column had more than 15% of the data missing. We will be handling the missing data

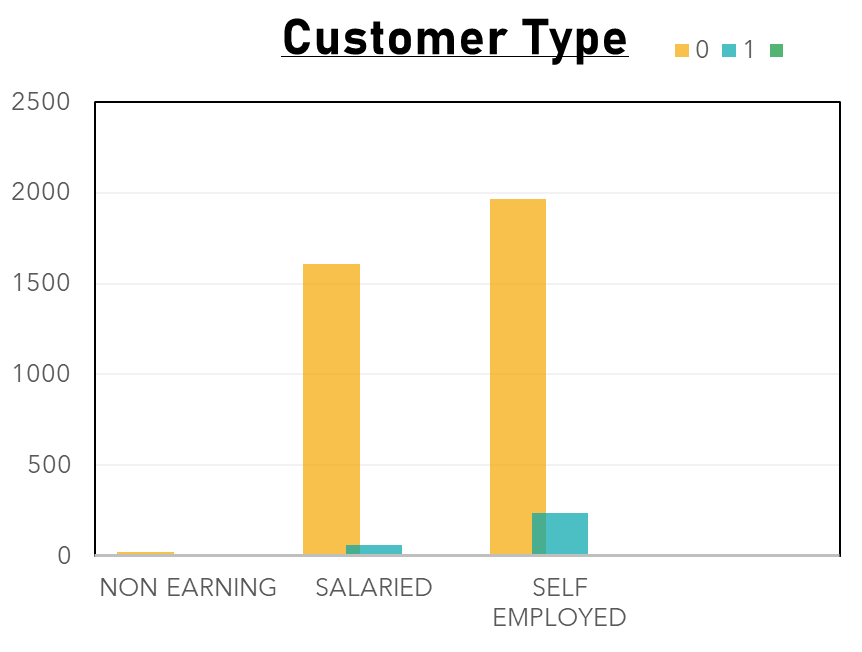
* by replacing the null values for Age with mean
* for Bank balance with -999
* for Type of industry with Mode.
* For categorical variables such as SEX and Marital Status we replaced missing values with NOT SPECIFIED.

We derived 30 more variables from the existing set to get a clearer picture of the data which are

1. Income/Age
2. Tot\_ENQ\_12M/Tot\_ENQ
3. Tot\_ENQ\_3M/Tot\_ENQ
4. Tot\_Home\_ENQ\_12M/Tot\_Home\_ENQ
5. Tot\_Home\_ENQ\_3M/Tot\_Home\_ENQ
6. Tot\_ENQ\_3M/12M
7. Tot\_Home\_ENQ\_3M/12M
8. Tot\_Home\_ENQ\_3M/Tot\_ENQ\_3M
9. Tot\_Home\_ENQ\_12M/Tot\_ENQ\_12M
10. Tot\_ENQ/Working\_Age
11. Tot\_Home\_ENQ\_3M/Tot\_ENQ
12. Tot\_Home\_ENQ\_12M/Tot\_ENQ
13. Defaults\_3M/12M
14. Tot\_UNSEC/Tot\_Live\_Loan
15. Months\_in\_Debt/Tot\_Live\_Loans
16. o/s\_Amt\_SEC/Tot\_o/s
17. o/s\_Amt\_UNSEC/Tot\_o/s
18. Tot\_Live/Tot\_(L+C)
19. Tot\_Home\_Live/Tot\_Home\_(L+C)
20. Tot\_UNSEC\_Live/Tot\_UNSEC\_(L+C)
21. Tot\_Home\_Live/Tot\_(L+C)
22. Tot\_UNSEC\_Live/Tot\_(L+C)
23. Tot\_ACC/Working\_Age
24. Tot\_no\_Home(L+C)/Tot\_Loan\_Other\_Bank
25. Tot\_no\_UNSEC(L+C)/Tot\_Loan\_Other\_Bank']
26. Closed\_Tot/Tot\_(L+C)
27. Closed\_Home/Tot\_(L+C)
28. Closed\_Home/Tot\_Home\_(L+C)
29. Closed\_UNSEC/Tot\_(L+C)
30. Closed\_UNSEC/Tot\_UNSEC\_(L+C)

**Target Count**

As we can see the count of Target Variable where Default rate as per the data is **7.6%** and non-Default rate is **92.34%**



As we had three types of customers, we rejected the non-Earning type as its influence was very low and the count of non-earning customers were only 23 among the overall data of 3980.

# Limitations

The first issue that needs to be addressed while building an Acquisition Model is the quality of data.

Data quality issues mainly concern with:

* Accuracy: User input errors or errors in software.
* Consistency: Relating to situations where multiple data sources are used and due to lack of standardisation, where data items may conflict with each other.
* Completeness of data: Refers to the degree to which all data in a data set is available. A measure of data completeness is the percentage of missing data entries.

A decision tree may be more appropriate for cases where there is significant missing data, or where the relationship between characteristics and targets is nonlinear. Sufficiency in the quantity of data is required to ensure construction of good quality and robust scorecard models, so the composition of credit scoring data set is homogeneous across lenders and regions.

Some of the major limitations from the data are as follows:

***Insufficient Data:*** The data that we have received had comparatively a smaller number of observations. Also, it lacks some of the major features such as Loan Amount, Tenure, Interest Rate, Funded Amount, Payment Plan etc.

***Segmentation:*** We tried segmenting on Customer type by considering Salaried and Self Employed but default distribution was way too low and also the model accuracy which we were getting was not up to the mark hence we avoided Segmentation.

***No visibility in how default request is calculated:*** As discussed earlier, Target variable has two values 0 and 1 where 0 denotes non-Defaulter and 1 denotes Defaulter but there is no clarification in how the customer is defined as Defaulter or not.

***No accessibility to PSI data:*** There are typically four types of data: Train, OOS (Out of sample), OOT (Out of time) and PSI (Population stability index) PSI is a metric to measure how much a variable has shifted in distribution between two samples over time. It is widely used for monitoring changes in the characteristics of a population and for diagnosing possible problems in model performance. This data is essentially important while studying the significant changes in the population distribution.

# Methodology

Typically, it includes 3 segments, first being Dimensionality reduction, the train and test split of the dataset and the modelling where we used Logistic regression.

## Dimensionality Reduction

As we know that more input features often make a predictive modelling task more challenging, which we also call as the curse of dimensionality. Therefore, dimensionality reduction refers to techniques that reduce the number of input variables in a dataset.

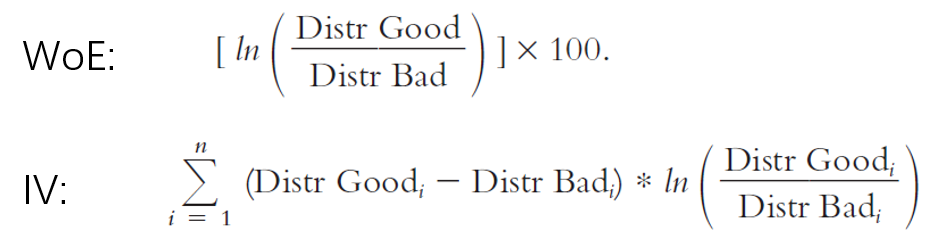
There are two ways to work with dimensionality reduction i.e., feature selection and feature extraction.

We are using Feature selection methods:

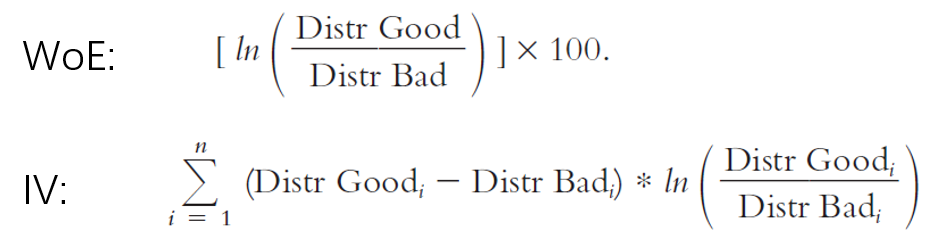
***Random Forest***: Works by creating multiple random trees and classifies the variables on the basis of their characteristics

***IV and WOE:*** It helps to understand the predictive power of a variable & its value ranges from 0 to 1.

*Weight of Evidence (WoE):*After data exploration and cleaning, there is a need to transform all the independent variables (like age, income etc.) using the weight of evidence (WoE) method. WoE is used to find the predictive power of an independent variable in relation to the dependent variable (target variable). Based on the proportion of good applicants to bad applicants at each group level, this method measures the “strength” of grouping for differentiating good and bad risk, and attempts to find a monotonic relationship between the independent variables and the target variable.



*Information Value (IV):*Information value is used to determine which independent variables have more influence on the dependent (target) variable. Classically this serves as variable ranking method and allows us to perform feature selection, which is less computationally demanding as other methods.



***Varclus***: It is used to perform variable clustering (varclus) with a hierarchical structure. It helps us select a variable by clustering them together on the basis of their characteristics.

***Lasso regression:*** Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model.

Using these concepts, we reduced our dimensionality from 62 variables to 10 variables.

## Train and Test split:

The training set the largest corpus of your dataset that you reserve for training your model. After training, inference on these images will be taken with a grain of salt, since the model has already had a chance to look at and memorize the correct output.

After all of the training experiments have concluded, it is important to hold back a completely separate stronghold of data - the test set. We can run evaluation metrics on the test set at the very end of the project, to get a sense of how well our model will do in production.

Before running the model, it is important to split the data into Train and Test. Ideally it is suggested to split it into 80:20 or 70:30. We have chosen to split the data into 70 and 30.

## Logistic Regression:

While making an Acquisition Scorecard Model, Accuracy Matters but Interpretation matters more.

Logistic regression is a model for binary classification predictive modelling which refers to those classification problems that have two class labels, e.g., true/false or 0/1. The logistic regression model outperforms many state-of-the-art machine learning algorithms and are still so widely used for its relatively easy to perform reality checks.

The logistic model formula computes the probability of the selected response as a function of the values of the predictor variables. If a predictor variable is categorical variable with two values, then one of the values is assigned the value 1 and the other is assigned the value 0. In our case, the scorecard will help us determine whether to approve or reject an application; 0 being the application approved and 1 being application is rejected.

The parameters of a logistic regression model can be estimated by the probabilistic framework called Maximum Likelihood Estimation. Under this framework, a probability distribution for the target variable (class label) must be assumed and then a likelihood function defined that calculates the probability of observing the outcome given the input data and the model. This function can then be optimized to find the set of parameters that results in the largest sum likelihood over the training dataset.

Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.

**Reasons why we use Logistic Regression for an Acquisition Scorecard model**

Here we try to explain methods for improving the accuracy of Acquisition scoring models.

There exist models that are intuitively hard to interpret such as

artificial neural networks (ANN) or complicated composite methods using ensemble

learning. Despite this fact, the model of choice for many financial institutes developing

scoring models is the relatively simple logistic regression.

1. **Auditing:** As according to the RBI audit, banks have to justify why the applicants are rejected/accepted. This is been studied on Monthly/Quarterly basis. Hence Logistic Regression model do assign weightage to the variables which provide a justification for acceptance or rejection.
2. **Weightage:** Logistic algorithm gives the weightage of the variables whereas other machine learning algorithms such as XG Boost, Random Forest, SVM, etc. gives the strength of the variables which is not enough to justify the cause of approval or rejection of loan
3. **Interpretation:** An acquisition scorecard needs to be interpretable which is very important as the banks have to justify why a particular customer was approved or rejected for a loan. Interpretability of a model also depends on the credit analyst’s knowledge and its format which is determined by adopted analytical techniques such as Logistic Regression. Logistic Regression are white box models that allow the user to understand the underlying reasons why the model signals a customer to be a defaulter.
4. **Deployability:** Implementation platforms (i.e., whether the application-processing system is able to implement a particular type of scorecard). For example, a neural network model may be ideal, but unusable, if the application-processing system is unable to implement it. The real-time requirements of the scorecard are not so high, and the model is usually updated once in 3 months in the bank.

***Execution Flow:***

1. By using Random Forest algorithm, we got these top ten variables.
2. Using IV and WOE, we selected 39 variables by giving the condition of predictive power greater than 0.02.
3. Later on, we used clustering method on our data and found around 32 variables. Then we took all the common variables from the above 3 methods and got a list of 16 variables.
4. We bifurcated our data into Train data and test data which had 70% and 30% of our data respectively.
5. After applying Lasso Regression on these datasets, we got a final list of 10 variables. All of these variables were numeric except, customer\_type\_salaried is a categorical variable.
6. We used logistic regression for modelling as it has high interpretability and accuracy. We created a Logistic regression model using logit function by undertaking these 10 variables and calculated the probability of default for each variable with respect to delinquency of the applicant which means that calculating the probability of default for each and every customer.

**Scorecard Scaling**

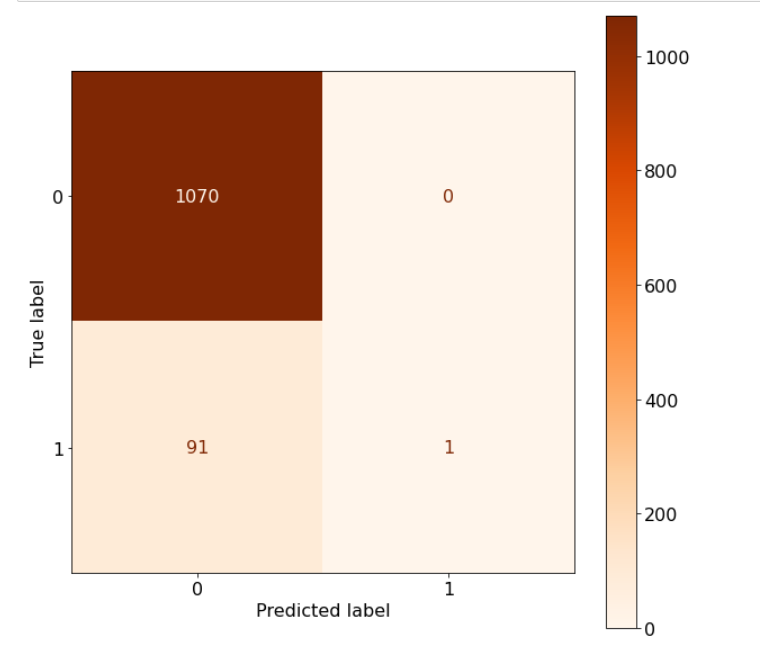
Scaling a Scorecard refers to making the scorecard conform to a particular range of scores and the regression coefficients are used to scale the scorecard. Logistic regression models are linear models, in that the logit-transformed prediction probability is a linear function of the predictor variable values.

**Reject Inference**

Performance of rejected applications is not readily available, and hence the development of a model based on both rejected and approved applications is a challenge. Reject Inference is a method of improving the quality of the scorecard based on the use of data contained in rejected loan applications.

# Deliverables

## Model Accuracy



92.08 %

0.09 %

After we fit the Logistic Regression on Train data, our Accuracy came up to 92.43% i.e., No. of classifications our model correctly predicted by total number of predictions. However, the accuracy for test data went up to 92.17% slightly lower than the train.

Here we have calculated the confusion matrix for our model. A confusion matrix is a performance measure used to evaluate Accuracy, Sensitivity and Specifications. As we can see there is 92% of them as non-defaulters and only 0.09% of all true defaulters. It is strange that among 1162 data points, there is only 1 defaulter. This is because the performance measure completely depends on the Target variable which is binary. So, the output of our algorithm will be a probability of default for each customer.

If the Probability of default is above 0.5, it is considered as Defaulter and below 0.5 as non-Defaulter. Ideally, we expect that higher the Probability of default, more defaulting customers. But our probability scale for test data ranges from 0 to 0.6 which can hardly capture true defaulters above cut-off value 0.5. So, to evaluate the accurate cut-off point, we make use of KS-statistic and Lift Table.

## Lift Table

**Lift Table for Train**



**Lift Table for Test**



By logit function, we evaluated the predicted probabilities ranging from 0 to 1 and break them into 10 equal bins. After that, we sorted our Bad rate for each bin in descending order.

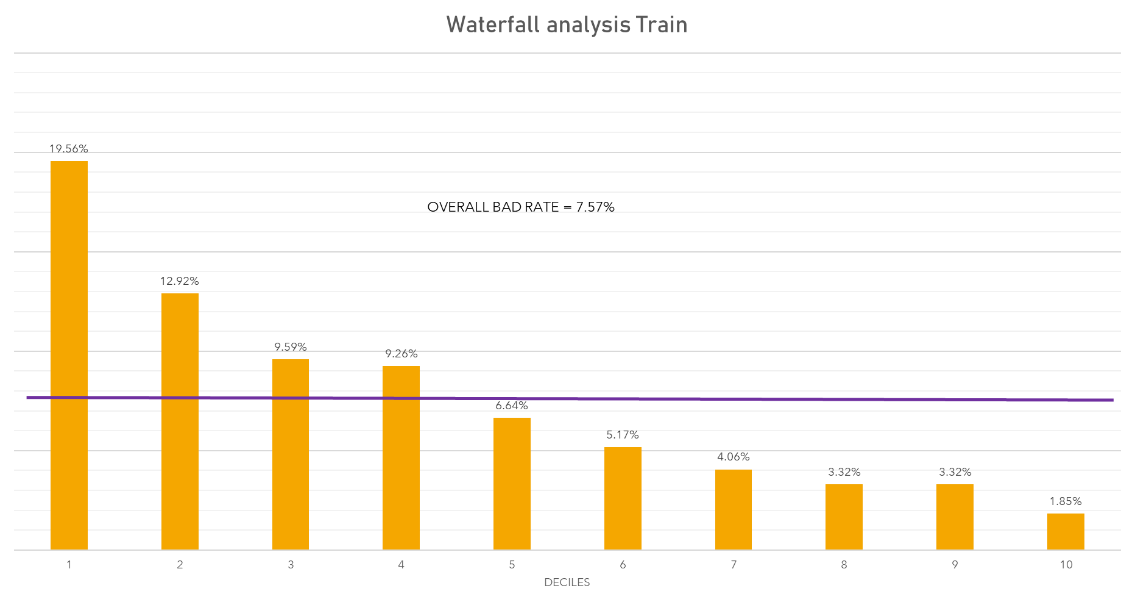
AUC here is nothing but the area covered by each bin and KS being a measure of degree of separation between good and bad distributions.

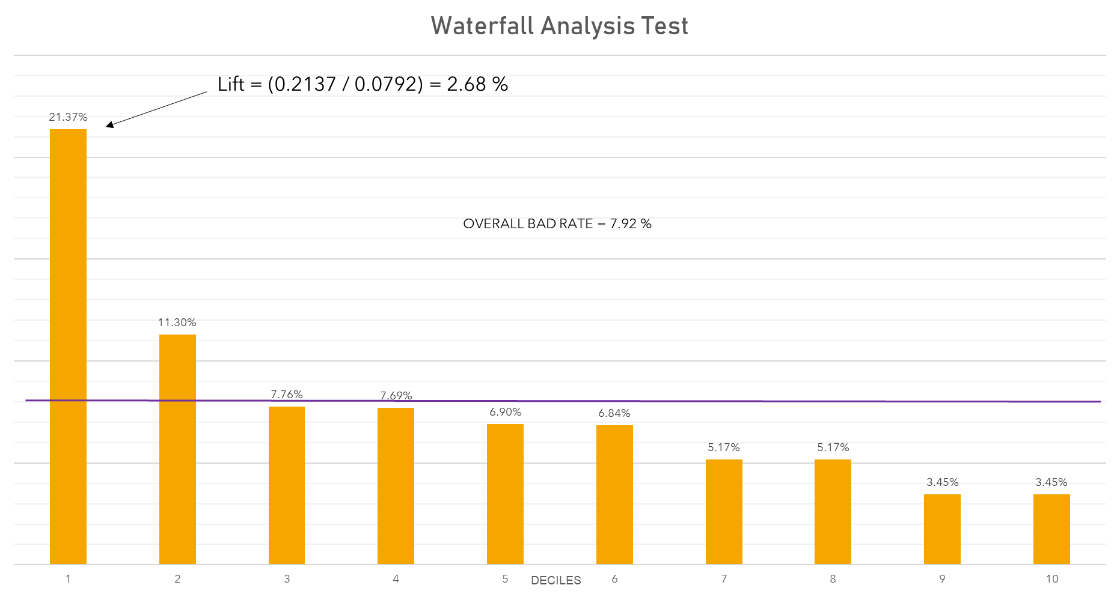
KS Statistic will fall between 0% to 100% and higher the value better the model is at separating the defaulters from non-Defaulters.

Here, the Lift table of Train data for the corresponding Maximum KS value depicts that 67.8% of defaulters have been covered in top 40% of data.

Whereas, in test data, we were likely to capture 41.3% of defaulters in top 20% of data for corresponding Max KS value.

## Waterfall Analysis

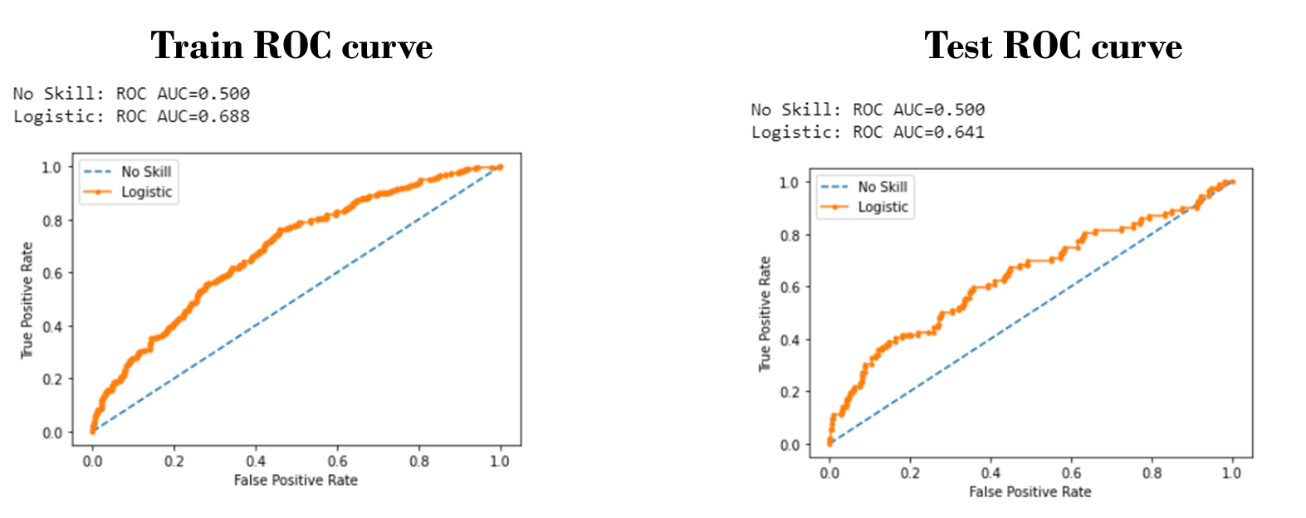




Let's understand the lift from waterfall Analysis on train as well as test data. For the 1st decile, the default rate being 13.56%. when compared with overall default rate of 7.57% given the lift of 2.58%.

This illustrates that all of the applicants in decile 1, 2, 3, 4 have a higher default rate predictive model. Similarly for the Test data, for 1st decile, the default rate being 21.37% when being compared with overall 7.92% gave us the lift of 2.698%. Therefore, all the applicants in decile 1, 2 have a higher default rate.

## ROC Curve



ROC curves are drawn using Default vs non-Default Rates. So, it exactly tells us how our model has predicted the defaulters correctly. As the curve from the graph sticks to the y-axis with no or very little Default Rate. Banks would want the FP rate i.e., the Default rate to be really low. Bank wouldn’t want to lose a genuine applicant just because our algorithm was too aggressive.

On the other hand, if our classifier is predicting whether someone is prone to be a defaulter, Bank might be okay to allow these higher risky applicants just to make sure that they don’t miss any non-Defaulters.

So here the area under the ROC curve for train data is slightly higher than Test data.

## Trend Interpretation

As all the selected variables except the salaried customer type were numerical, we binned our data on the basis of the default rate trend followed by train data and further by using those bins we computed the default rate on Test data.

1) Tot\_ENQ by Working\_Age is the first variable that we selected. We conclude that the default rate goes on increasing as the enquiries increase when the working age is minor or trivial.

2) Days Since Last Account was Open

If the applicant has opened his credit account a year ago is less likely to default whereas the applicant who has opened his/her account in last 2 or 3 months, they are much more riskier .

3) Average number of days in debt

If applicants already have debts that need to be repaid and he/she has already been in debt for almost 2 years, they have higher probability of defaulting.

4) Total Home Loan Enquiry in Last 12 months

Home Loans being a high value investment i.e., high EMI and if applicant already has debt that need to be repaid, then the loan application might be rejected. Here, if a person applies for credit multiple times in a very short period that means he/she is not able to manage finances properly and is always looking for credit as they over spend.

5) Outstanding unsecured amount by total outstanding amount

Unsecured debt refers to loans that are not backed by collateral. Hence, unsecured loans are considered riskier for the banks and the one with higher unsecured outstanding amount are more prone to default.

# Conclusion

Acquisitions Scorecards should be monitored regularly. Detection of defaulters is identified in early stage because of the Acquisition Scorecard. The scorecard has improved the way of understanding the overall accuracy of the application selection process. The final features were selected based on Enquiries, Outstanding amount and Customer Type (Salaried/ Self-employed). The model that we created is overfitted. Population of bad is 67.8% In Train whereas it is 41.3% in Test, KS cut off for train is 0.3 and for Test is 0.23. Model will be a good help for credit analyst to focus more on distinctive decision.

# References

|  |  |
| --- | --- |
| [1] | S. Liu, “Credit Scoring — Scorecard Development Process,” [Online]. Available: https://medium.com/@yanhuiliu104/credit-scoring-scorecard-development-process-8554c3492b2b. |
| [2] | Y. H. H. Q. a. Leonardo Gambacorta, “How do ml and non-traditional data affect credit scoring,” [Online]. Available: https://www.bis.org/publ/work834.pdf. |
| [3] | [Online]. Available: https://blog.roboflow.com/train-test-split/. |
| [4] | [Online]. Available: https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050910X00023/1-s2.0-S1877050910002796/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEAIaCXVzLWVhc3QtMSJGMEQCIAKCDtFwuZjDmcajFbnuFxvgyR9HI0oeU0NooxSFhG3LAiBhX6TMbSfqvBd9aCpDqz0S1s3KL3J1dEdNojU7z%2B0iky. |
| [5] | T. |. K. Property, “Intro to Credit Scorecard,” [Online]. Available: https://towardsdatascience.com/intro-to-credit-scorecard-9afeaaa3725f. |
| [6] | CIBIL, [Online]. Available: https://www.cibil.com/loan-approval-process. |