**Symbiosis Centre for Management and Human Resource Development**
**[Constituent of Symbiosis International (Deemed University)]**
(Established under Section3 of the UGC Act 1956,bynotificationNo.F.9-12/2001– U.3 of the Government of India)

# Credit Scoring Model

By

Vatsal Aima

Under the Guidance of
Mr. Gautam Mahesh, AVP – Products and Analytics

In (Partial) fulfillment
of requirements for the degree of
Masters of Business Administration.

DATE: 08/06/2018

**CERTIFICATE**
**(This needs to be scanned and attached in the report)**

This is to certify that Mr. *Vatsal Aima* has successfully completed the summer internship project titled *Credit Scoring Model*, in the *Data Science* department of *EPIMONEY PRIVATE LIMITED*. It is an independent research work done under my supervision during *April – May 2018*. It is being submitted to the Symbiosis Centre for Management and Human Resource Development, Constituent of Symbiosis International (Deemed University) in partial fulfillment for the award of the Degree of Master of Business Administration.

(Gautam Mahesh)

Signature of Project Guide 30th May, 2018

Designation AVP - Products & Analytics

Company Stamp

EpiMoney Private Limited
Mumbai

**Acknowledgement**

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance I would not forget to thank them.

I am grateful for the opportunity provided by Epimoney Private Limited to be a part of their company. I'm thankful to co-founders Mr Manish Lunia, Mr Abhishek Kothari, Mr Ritesh Jain and Mr Deepak Jain for accommodating me as intern.

I respect and thank Mr Abhishek Kothari for providing me an opportunity to do the project work which enhance my knowledge of data analytics and its application in SME lending business. I am thankful to Ms Prathama for providing such a nice support, although she had busy schedule managing the corporate affairs.

I owe deep gratitude to our project guide Mr Gautam Mahesh who took keen interest on our project work and guided us all along, till completion of our project work by providing all necessary information for developing a good system.

I would not forget to remember my project partner Ms Neha Yegnamurthy for her contribution and timely support till the completion of our project work.

Vatsal Aima,

MBA Business Analytics,

SCMHRD, Pune.

## Table of Contents

**Abstract**

Credit scoring model plays a fundamental role in the risk management practice at most banks. They are used to quantify credit risk at counterparty or transaction level in the different phases of the credit cycle (e.g. application, behavioral, collection models). The credit score empowers users to make quick decisions or even to automate decisions and this is extremely desirable when banks are dealing with large volumes of clients and relatively small margin of profits at individual transaction level (i.e. consumer lending, but increasingly also small business lending). In this project, a credit scoring model has been developed to analyze the company's lending behaviour towards various customer segments based on their historical data and compare this model with the existing one. Moreover, we analyze the key steps of the credit scoring model's lifecycle (i.e. assessment, implementation, validation and deployment) highlighting the main requirement imposed by company for model development. From the results, it is found that our model (Challenger Model) reduces the Type 1 error as well as supersede the accuracy of company's model (Proprietary Model) in order to determine whether the loan should be approved or rejected based on the parameters that customer provides while applying for loan.

## Company Profile

FlexiLoans Technologies Pvt. Ltd. operates a technology based financing platform that provides small business owners and consumers with access to loans from its affiliate NBFCs. Its platform provides access to working capital loans, including business expansion and seasonal inventory loans; discounting loans, such as PO and invoice discounting loans and other business loans, which include equipment financing. The company also provides a mobile version of its platform through its application. FlexiLoans offers four types of collateral free loans to help SMEs and individuals fund their businesses. These are Flexi-Term, Flexi Vendor Financing, Flexi – Merchant Advance, Flexi-Line. FlexiLoans Technologies Pvt. Ltd. was founded by Mr. Manish Lunia, Mr. Ritesh Jain, Mr. Deepak Jain, and Mr. Abhishek Kothari It is incorporated in 2016 and is headquartered in Mumbai, India.

## Research Methodology

### Credit Scoring

Credit scoring is a statistical analysis performed by lenders and financial institutions to access a person's creditworthiness, using their information such as age, previous loan history, income etc. Lenders use credit scoring, among other factors to decide whether to extend a loan to a customer. It is the first formal approach to the problem of assessing credit risk in a scientific and automated way, as the volume of loan applications has been growing.

A credit scoring model is a tool build using credit scoring methods which is used in decision making process of accepting or rejecting loan applications and gives an estimate of the probability of default.

There's no standard credit scoring model followed by a company. Models vary from one company to another depending on their product offerings and target customers. Within each company there may be multiple models for different types of loans and customers. The attributes used in credit score modelling can also vary.

### CIBIL Score

A credit score is a number that represents the results of a credit scoring model. In India, most lending companies use CIBIL (Credit Information Bureau India Ltd.) score as an important criterion for their decision making. CIBIL is India's first Credit Information Company that collects and maintains monthly reports of individuals' loan and credit card payment history from banks and financial institutions. A credit score, in other words the CIBIL score, is generated based on these reports.

The CIBIL Score is typically three-digit numeric figure that ranges between 300 and 900 for a record that is older than 6 months. A higher score indicates higher credit worthiness and thus higher chances of getting loans. A record less than 6 months old has a CIBIL score ranging between 1 and 5.

## Objective

In this project, objective was the development of “credit scoring model” which anticipate what is the likelihood that customer’s request will be accepted/qualified or rejected/disqualified on the basis of information that he/she provides while applying for loan. In order to accomplish this objective, the work was divided into two phases. In each phase, a predictive model was designed to accomplish certain goal of objective.

The purpose of the model developed in Phase I is to indicate the probability that a loan is rejected or approved once the loan application and documents are verified by the CRM (Customer Relationship Management) team. In total 14 attributes were used for analysis and model development. On these attributes three models were developed for the sake of comparison and based on statistical parameters, selection of best model is done. For model development, three algorithms were used namely Naïve Bayes, RPART and Random Forest.

The purpose of the model developed in Phase II is to indicate the probability that a loan passes the bureau record criteria based on the information in CIBIL records. This model was created as a challenger model against the model currently used by FlexiLoans (Proprietary model). For model development, four algorithms were used namely Naïve Bayes, RPART, C5.0 and Random Forest.

## Tools Used

- Microsoft Excel
- JetBrains PyCharm Community Edition
- R Studio
- MongoDB Compass Community
- MySQL Workbench

## Data Preparation

- Data Extraction

Data was present in two types of database:

- ❖ Relational Database Management System (RDBMS) – These consisted of normalized tables. SQL queries were used to extract the data from MySQL database.
- ❖ MongoDB – A NoSQL database program which uses JSON-like documents to store data, i.e. data is present in hierarchical form. The data available as JSON structure was then read into python in the form of nested dictionaries and lists. This was then programmatically converted into a relational structure i.e. a dataset.

- Data Manipulation
  - ❖ Missing Data Treatment

Several fields had missing data which required an extensive amount of treatment. Given the heterogonous nature of SMEs, it was a challenge to figure out most common attributes that could be used for modelling. Even these common attributes had enough missing data to work on.

A combination of several approaches was used to overcome this issue. For certain categorical fields, assumptions were made based on business intuition depending on the values of other attributes for those records. Some data was imputed using the MICE package available in R. Attributes such as CIBIL score for which neither assumptions can be made, nor imputations done, were the records omitted.

- Outlier Treatment

Categorical outliers i.e. levels of a categorical variable that occurred only a few times (less than 10) were either changed to something closer or omitted. Numerical outliers, which particularly was a challenge with respect to monthly sales, were treated mostly based on business intuition considering the values of other attributes of the corresponding rows.

- Creation of new fields

As date fields cannot be used directly in modelling, they were converted into attributes such as age, vintage etc. Some categorical variables had too many levels as compared to the number of records. Many of these levels had to be combined to reduce that total number of levels.

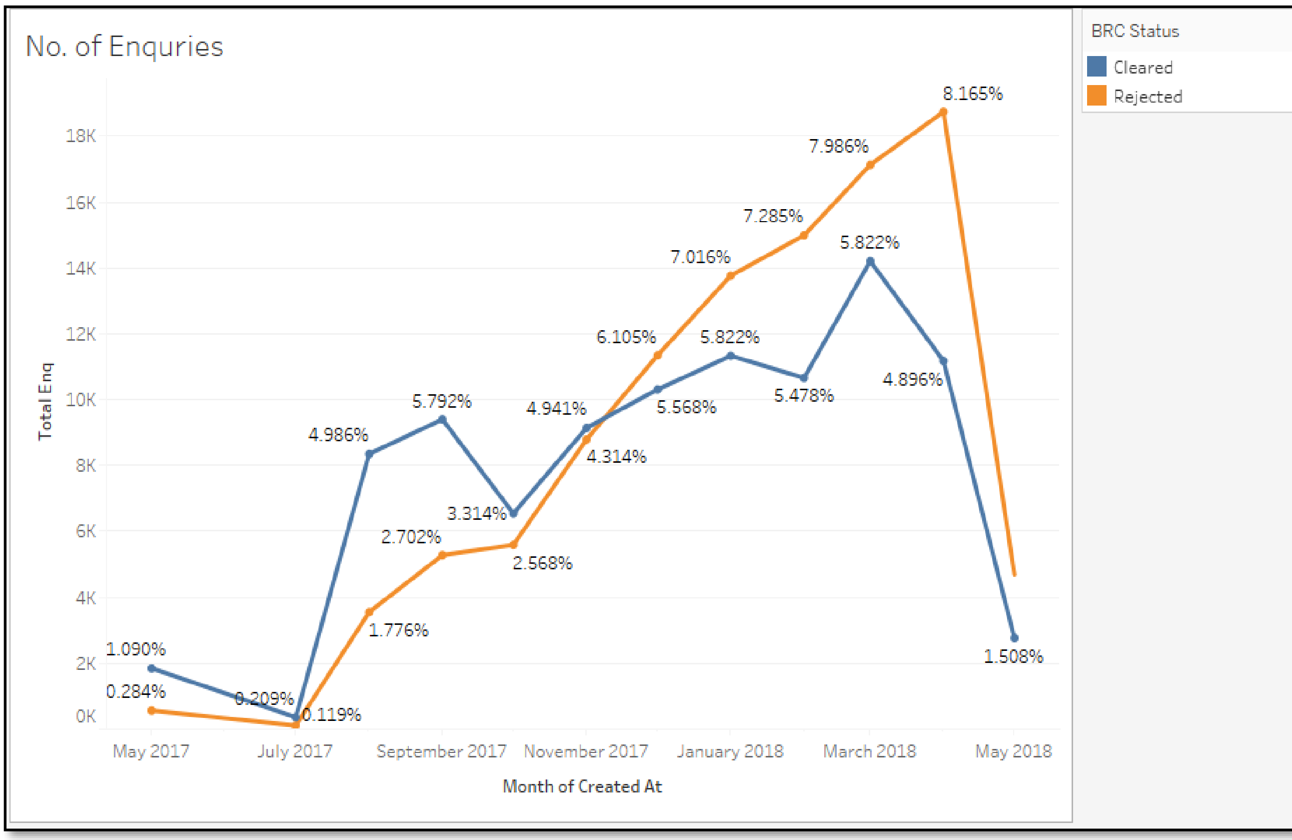
- Splitting data into Training and Test sets

Data was split into training and test set in the ratio of 80:20. Since the percentage of approvals is only about 20%, it was important to take care of the balance of data. The data was split in such a way that the ratio of approvals in training and test set were maintained. This minimizes the difference between accuracy and balanced accuracy.

## Results and Analysis

Visual Descriptive Analysis (Tableau Dashboard)

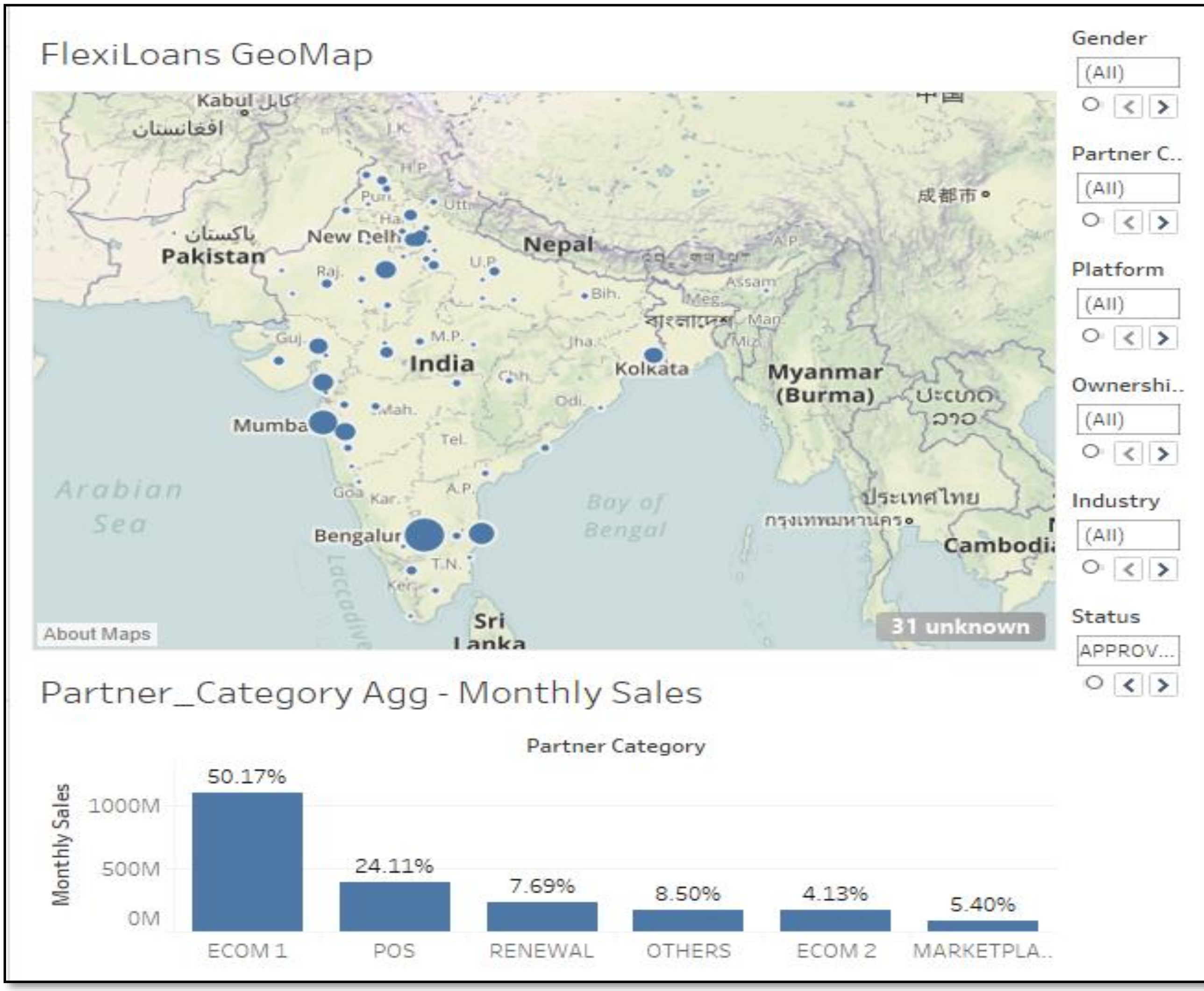
Before developing a credit scoring model, I did the descriptive analysis (visual) in order to find the key requirements and parameters for predictive model. Below are some snapshots from Tableau Dashboard.




*Number of loan queries generated on monthly basis (from May 2017 - 2018)*

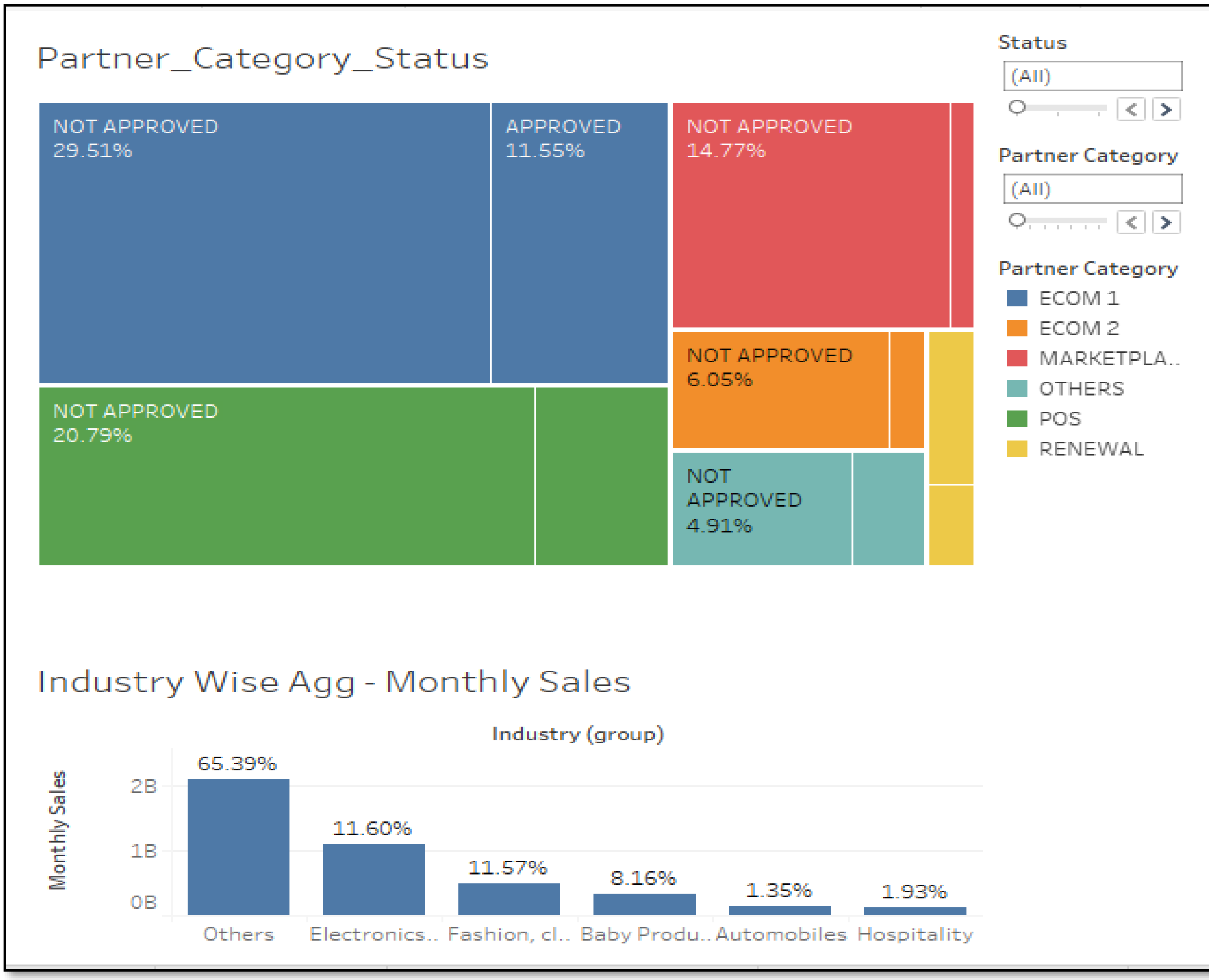
From the above, we can observe that has rejected number of loan requests because of either in suficiant details provided by customers or other reasons. Due to this huge gap, an opportunity arises where a significant number of request which were rejected should actually be approved.

For this reason, a new model was suggested which can identify this opportunity and predit more accuratly to reduce this gap.




*FlexiLoans potential market segment and their monthly sales*

The above figure shows the geographical map which identifies the potential market segment and their aggregated monthly sales. It gives the overall view of customer segment which are essential from company's perspective. Here, each segment i.e., ECOM 1, POS, RENEWAL, OTHERS, ECOM 2, MARKETPLACE etc. represents a group of customers based on specific criteria.




*Industry – wise monthly sales of potential market segment*

The above figure shows the relationship between loan status and customer category on tree map. It also depicts industry-wise monthly sales from customer segment. It gives the idea about industry segment for which FlexiLoans can target new customers. The "Others" represents some small and unknown category.

## Statistical Terminologies Used to Compare Models

Here I've used the sensitivity, specificity, accuracy, balanced accuracy, misclassification, Kappa, Receiver Operating Characteristic i.e., ROC and Area under the curve i.e., AUC in order to determine goodness of fit for each model. Below are the definitions of some terminologies:

- Confusion matrix:

  It is a table that describe the performance of a classification model (or "classifier") on a dataset for which the true values are known.

  For Example:

| **Actual** | **Predicted** | |
|---|---|---|
| | **Event** | **No Event** |
| **Event** | A | B |
| **No Event** | C | D |

- True positives: These are cases in which the predicted value for the model and the actual value will be same and both are TRUE. In above e.g.,, its *A*
- True negatives: These are cases in which the predicted value for the model is TRUE but the actual value will be FALSE. In above e.g.,, its *B*
- False positives: These are cases in which the predicted value for the model is but the FALSE actual value will be TRUE. In above e.g.,, its *C*
- False negatives: These are cases in which the predicted value for the model and the actual value will be same and both are FALSE. In above e.g.,, its *D*

- Sensitivity (*True Positive Rate*): It measures the proportion of actual positives that are correctly identified. It can be calculated as *A/(A + B)*
- Specificity (*True Negative Rate*): It measures the proportion of actual negatives that are correctly identified. 1 - Specificity is also known as *False Positive Rate*. It can be calculated as *D/(D + C)*
- Balanced Accuracy: It's the average of the proportion corrects of each class individually. The purpose of calculating balanced accuracy when test set is not balanced i.e., test set don't have same number of examples in each class. It can be calculated as *(Sensitivity + Specificity)/2*
- Accuracy: Rate of prediction values and actual values to be same

  *(A + D)/ (A + B + C + D)*
- Misclassification Rate: Rate of prediction values and actual values to be not same

  *(B +C)/ (A + B + C + D)*

➢ Kappa value:

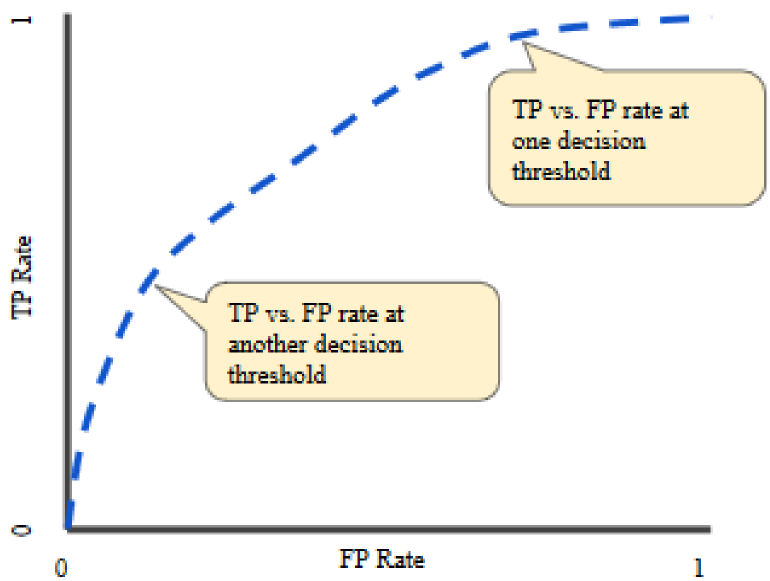
Cohen's kappa coefficient (k) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be more robust measure that simple percentage calculation, as k takes account into the possibility of the agreement occurring by chance. It can be calculated as

$K = (P_0 - P_e)/ (1 - P_e)$

Where $P_0$ is the relative observed agreement among events (identical to accuracy) and $P_e$ is the hypothetical probability of chance agreement using the observed data to calculate the probabilities of each observer randomly seeing each category.

- Area under the Curve (AUC):

It provides an aggregate measure of performance across all possible classification thresholds. AUC is the probability that the model ranks a random positive more highly than a random negative. It can be calculated as *True Positive (TP) Rate / False Positive (FP) Rate.*




TP vs. FP rate at different classification thresholds.

## Phase I - Credit analysis on Generic parameters

*Algorithms Applied For Model Development*

1. *Naive Bayes*
2. *RPART*
3. *Random Forest*

*Model Comparison Based on following Statistics*

| **Measure** | **Random Forest** | **RPART** | **Naive Bayes** |
|---|---|---|---|
| **Accuracy** | **94.49%** | 78.77% | 77.31% |
| **Sensitivity** | **95.55%** | 80.47% | 83.15% |
| **Specificity** | **91.67%** | 70.48% | 60.62% |
| **AUC** | **98.3%** | 79.3% | 81.2% |

*Best Model: Random Forest*

*Variable Importance*

| **Variables (Masked)** | **Approved** | **Not Approved** | **Mean Decrease Accuracy** | **Mean Decrease Gini** |
|---|---|---|---|---|
| M1X1 | 38.07296775 | 13.09469672 | 37.40644485 | 83.47608631 |
| M1X2 | 29.47064278 | 31.23958178 | 41.10397105 | 59.65327319 |
| M1X3 | 4.655789483 | 5.518999347 | 7.166516743 | 14.36397608 |
| M1X4 | 3.228139557 | 5.771792708 | 6.670467537 | 107.0607148 |
| M1X5 | 4.630389097 | 9.401412689 | 10.49865003 | 122.9276225 |
| M1X6 | 20.55947579 | 17.91472886 | 26.64974819 | 137.7701013 |
| M1X7 | 57.60907296 | 32.86545811 | 59.21096295 | 191.4301658 |
| M1X8 | 8.989513861 | 1.927072998 | 7.754954902 | 31.04763586 |
| M1X9 | 30.84486488 | 11.21628062 | 27.31688124 | 85.78271636 |
| M1X10 | 6.943449861 | 10.59368279 | 12.64262061 | 72.23100944 |
| M1X11 | 2.57951253 | 9.042835896 | 8.192539258 | 22.5615018 |
| M1X12 | 11.24227585 | -2.537053298 | 5.334259169 | 11.42126477 |
| M1X13 | 13.51343959 | 6.193460812 | 13.3534071 | 16.80318456 |

**Note**: *The above variable notation have been given below:*

*M2X1 – 12*: Model 1 Variable 1 to 12

*Confusion Matrix*

| **Actual** | **Predicted** | |
|---|---|---|
| | **Approved** | **Not Approved** |
| **Approved** | 154 | 20 |
| **Not Approved** | 14 | 429 |

*Statistics*

| | | |
|---|---|---|
| **Accuracy** | 94.49% | 95% CI (92.38%, 96.15%) |
| **Kappa** | 86.25% | |
| **Balanced Accuracy** | 93.61% | |

*Performance of Model with different Testsets*

RF Accuracy: 96.59 % for testset 1

RF Accuracy: 94.16 % for testset 2

RF Accuracy: 95.94 % for testset 3

RF Accuracy: 94.64 % for testset 4

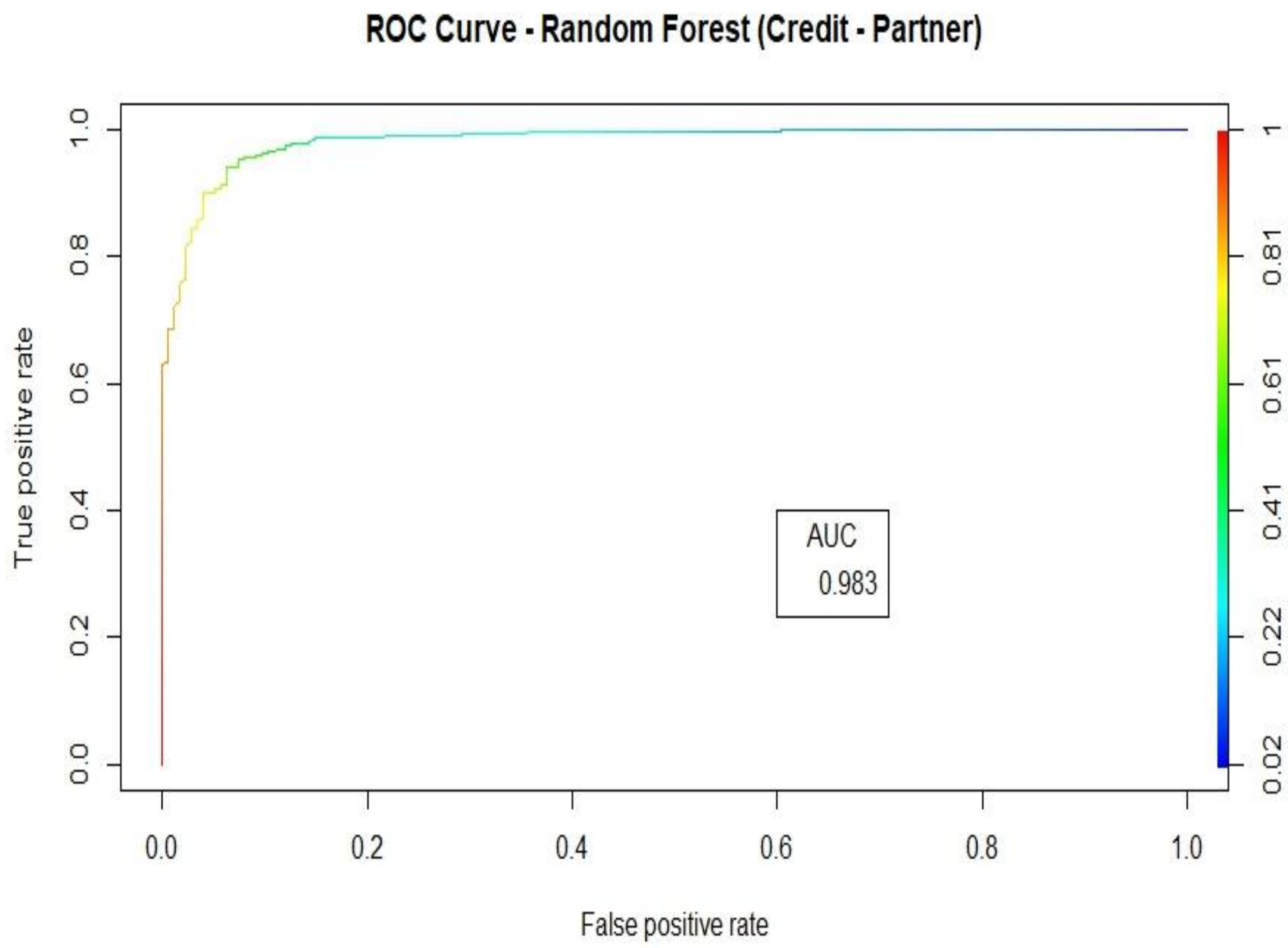
RF Accuracy: 93.67 % for testset 5

RF Accuracy: 95.13 % for testset 6

RF Accuracy: 94.81 % for testset 7

Aggregate Accuracy:  94.99 % of all 7 testsets

*ROC – Area Under The Curve (Graphical Representation)*



Phase II - Credit Analysis based on CIBIL parameters

*Algorithms Applied For Model Development*

1. *Naive Bayes*
2. *RPART*
3. *Random Forest*
4. *C5.0*

*Model Comparison Based on Statistics*

| **Measure** | **Random Forest** | **C5.0** | **RPART** | **Naive Bayes** |
|---|---|---|---|---|
| **Accuracy** | **97.69%** | **93.96%** | **90.52%** | **68.36%** |
| **Sensitivity** | **97.93%** | 93.07% | 87.38% | 61.96% |
| **Specificity** | **97.44%** | 94.90% | 94.36% | 92.81% |
| **AUC** | **99.5%** | 98.2% | 91% | 92% |

*Best Model: Random Forest*

*Variable Importance*

| Variables (Masked) | Cleared | Rejected | Mean Decrease Accuracy | Mean Decrease Gini |
|---|---|---|---|---|
| M2X1 | 54.77911071 | 51.41020219 | 69.67903283 | 491.746465 |
| M2X2 | 30.64063833 | 15.71708302 | 36.89584405 | 124.7453753 |
| M2X3 | 22.79503053 | 16.42599108 | 32.43745573 | 106.7909191 |
| M2X4 | 27.53887714 | 25.468335 | 30.59116767 | 519.6297098 |
| M2X5 | 28.25943943 | 40.97537394 | 39.38499906 | 700.2351248 |
| M2X6 | 22.66377761 | 8.674784221 | 24.37186163 | 85.3793625 |
| M2X7 | 29.23167936 | 14.12744423 | 34.79134769 | 115.6318838 |
| M2X8 | 27.98390693 | 10.93811359 | 32.08667433 | 79.87287365 |
| M2X9 | 25.22859847 | 5.763170385 | 26.04467122 | 94.37760416 |
| M2X10 | 40.02358248 | 10.65552612 | 40.34800919 | 95.93192453 |
| M2X11 | 27.75931673 | 0.859223881 | 27.17038807 | 84.92046736 |
| M2X12 | 23.71121078 | 11.55817333 | 28.07573717 | 67.13079092 |

**Note**: The above variable notation have been given below:

*M2X1 – 12*: Model 2 Variable 1 to 12

*Confusion Matrix*

| **Actual** | **Predicted** | |
|---|---|---|
| | **Cleared** | **Rejected** |
| **Cleared** | 614 | 48 |
| **Rejected** | 33 | 645 |

*Statistics*

| | | |
|---|---|---|
| **Accuracy** | 97.69% | 95% CI (96.73%, 98.42%) |
| **Kappa** | 95.37% | |
| **Balanced Accuracy** | 97.68% | |

*Performance of Model with different Testsets*

RF Accuracy:  96.94 % for testset 1

RF Accuracy:  96.27 % for testset 2

RF Accuracy:  96.42 % for testset 3

RF Accuracy:  95.59 % for testset 4

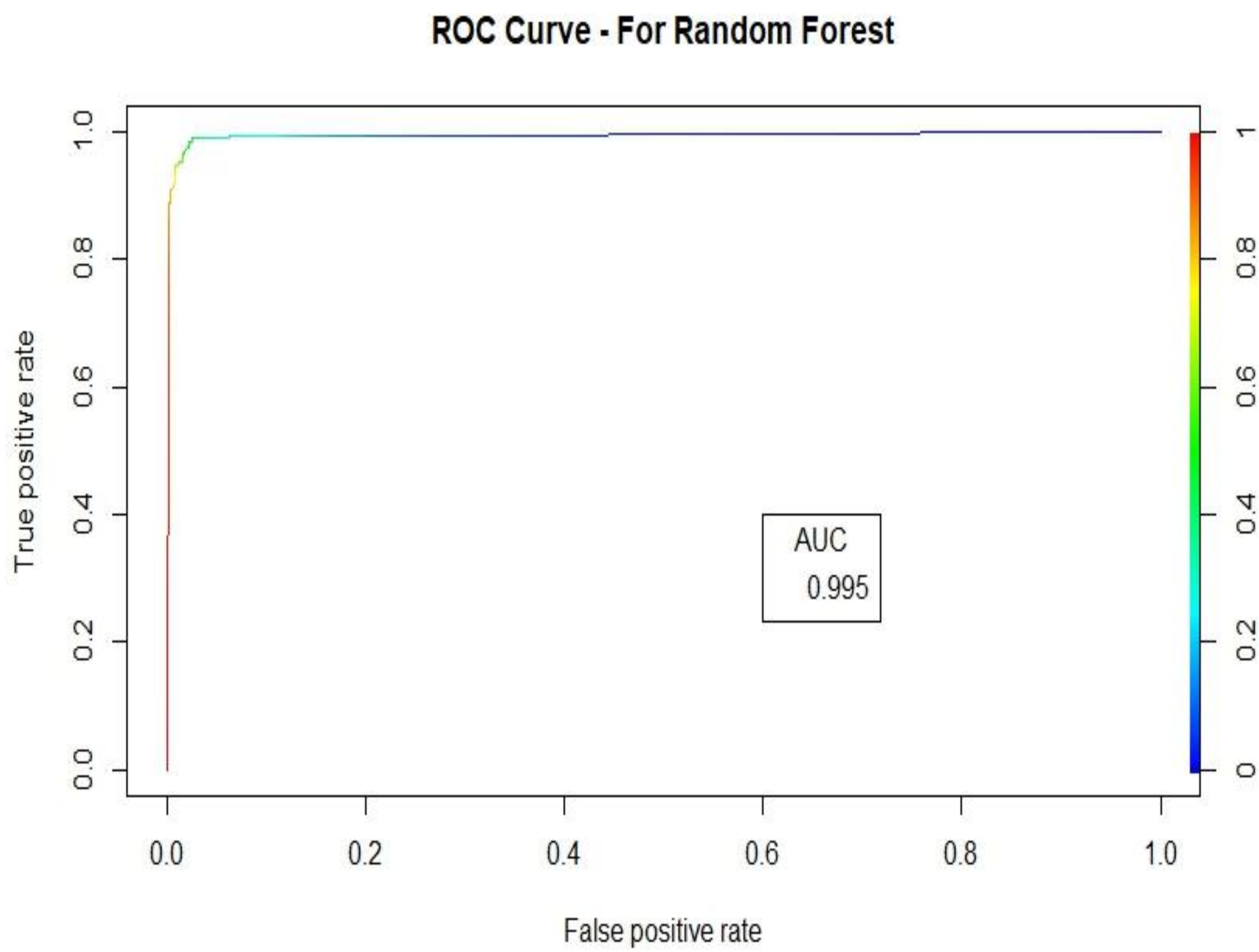
RF Accuracy:  96.56 % for testset 5

RF Accuracy:  95.82 % for testset 6

RF Accuracy:  95.52 % for testset 7

Aggregate Accuracy:  96.16 % of all 7 testsets

*ROC – Area Under the Curve (Graphical Representation)*



*Comparing Challenger Model with Proprietary Model*

Actual Vs Challenger Model

| **Confusion Matrix** | **Predicted** | | |
|---|---|---|---|
| **Actual** | **Cleared** | **Rejected** | **Grand Total** |
| Cleared | 82.62% | 1.07% | 83.69% |
| Rejected | 3.00% | 13.30% | 16.31% |
| **Grand Total** | **85.62%** | **14.38%** | **100.00%** |

Actual Vs Proprietary Model

| **Confusion Matrix** | **Predicted** | | |
|---|---|---|---|
| **Actual** | **Cleared** | **Rejected** | **Grand Total** |
| Cleared | 83.69% | 0.00% | 83.69% |
| Rejected | 12.88% | 3.43% | 16.31% |
| **Grand Total** | **96.57%** | **3.43%** | **100.00%** |

From the above confusion matrices, we can observer that in “Actual Vs Challenger Model” the misclassification rate is 4.07% whereas in “Actual Vs Proprietary Model” it is 12.88%. Also Type1 error (a statistics term used to refer to an error that is made in testing when a conclusive winner is declared although the test is actually inconclusive.) is lower i.e., 3% as compared with other one i.e., 12.88%. On comparing the proprietary model with the challenger model, the proprietary model gives an accuracy of **87.12%** while the challenger model gives an accuracy of **95.92%**. The challenger model is significantly better at predicting false rejections.

Proprietary Model Vs Challenger Model

| **Confusion Matrix** | **Predicted** | | |
|---|---|---|---|
| **Actual** | **Cleared** | **Rejected** | **Grand Total** |
| Cleared | 85.62% | 10.94% | 96.57% |
| Rejected | 0.00% | 3.43% | 3.43% |
| **Grand Total** | **85.62%** | **14.38%** | **100.00%** |

Here, Proprietary Model is being compared with Challenger Model so as to show the difference coherence i.e., agreement between the two classes is 89.05% where both muddled i.e., disagreement between the two is 10.94%.

**Discussion**

The goal of this project was to identify new parameters for lender rejection or approval criteria and rectify risky loan approvals in future with help of predictive model. Analysis of customer's financial behaviour also plays an important role in this regard for finding where exactly the customer face problems in his/her historical transactions.

The important part of this project was credit analysis on CIBIL parameters. Predictive analysis helped in understanding customer financial behaviour i.e. the model can inform about the customers financial stability based on financial parameters which falls under lender rejection category.

This will help the organization to a great extent in prioritizing the early rejections or easy in loan application process by knowing which customer is more likely to reject/accept for loan eligibility so that the application process can be expedited for him/her and make their experience smooth in loan disbursement process.

**Conclusion and Recommendations**

- Before model development, with the help of visual analysis, it was observed that from July 2017 to Sept 2017 there is significant raise of loan approval which could be cases of model inefficiency. Also verifying the integrity of data from their databases, it also found that such cases were present. For this reason a model was required to have correction on these issues.
- Apart from model development, from business perspective it was observed that in some geographical areas their influence (market capitalization) was lower for some prominent customers. It was recommended to do detailed market study and collect information from customers on this subject.
- The roadblock that most of the customers face was to provide bank statement in pdf format as it’s a requirement posed by company for its lending operations. Majority of the customers were unable to provide bank statement because it’s a tedious task for them to ask bank statement from their respective banks in specified format. As an alternative, they scanned their passbook from their mobile cameras and mailed to CRM team. For credit team it was more or less of manual work and such human interventions are prone to error. The process of such statements became a major bottleneck.
- It is recommended that there should be some alternate method for acquiring and processing the required documents from customers so that people who are not tech savvy, do not have to be much bothered for obtaining documents from their respective banks. In this matter, it can be tried to link customer’s bank account with the existing database after

due consent from customer so that the required documents can automatically be downloaded in the required format.

**Learning**

- Professional communication - This is a key factor for my learning because we were always accompanied with senior managers and project head of the organization so the professional way of communicating was very important and a key indicator for my behaviour.
- Technical skills – Improvement in technical knowledge and skill is required as I need to start a project from scratch i.e. understanding the data, data cleaning, make actionable insights from data, and finally presenting it to the project lead and suggesting him for making recommended changes. During internship, I learned many new things and hone my existing skills. Some of the learnings during internship are as follow:
    - **<u>Tableau:</u>** It provides instantaneous insights by transforming data into visually appealing, interactive interface called dashboards and stories. This process takes only seconds or minutes rather than months, and is achieved through the use of an easy to use drag-and-drop action. Much of the descriptive analysis were done on tableau and key findings were summarized on dashboards.
    - **<u>MongoDB (NoSQL Database):</u>** MongoDB Compass Community provides a visual console to easily administer MongoDB environments and gain better visibility into databases. All the NoSQL queries were executed on in order to fetch relevant data out of database for further operations.

- **<u>Python:</u>** JetBrains PyCharm Community Edition was used as an interface to perform python programming on MongoDB Database for executing real-time data extraction for model development process.
- **<u>MySQL (SQL Database):</u>** MySQL Workbench provides a visual console to easily administer MySQL environments and gain better visibility into databases. All the SQL queries were written on workbench interface.
- **<u>Classification Techniques (in R):</u>** Classification techniques like random forest, Rpart and C5.0 (in Phase II) were used along with Naïve Bayes. These new models provided an insight in how to increase the model accuracy through different methods.

➢ Smart work – My personal experience made me realize that hard work is not only the essential to carry on during internship but also being smart at work i.e. how I can make myself do better when compared to others in terms of ideas, duration of time, and how I work in a team with different people all around.

➢ Subject expert – I took help from different employees apart from mentor in order understand lending business in order to build efficient model. From initial phase of the project it was very mandatory to have a clean and clear way of approaching it.

➢ Time management – Throughout the period of my internship I was working with different employees and time of meeting were delayed because of their official work and I was asked to wait long till nights and discuss the issues with my works and share my ideology with them. I managed with my timings and was scheduling my time table instantly.

## Limitations

- The dataset extracted from the database contained lots of missing values as the sellers were reluctant to provide necessary information which results into removal of missing data. A significate data reduction occurred while cleaning process and less data points were available for analysis.
- Some important financial and generic parameters weren't captured or been missed while retrieving from customer at the time of request.

## References

https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=404519020

https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

http://mvpa.blogspot.com/2015/12/balanced-accuracy-what-and-why.html

https://en.wikipedia.org/wiki/Sensitivity_and_specificity