MITIGATING THE RISKS INVOLVED IN BANKING AND FINANCIAL SERVICES

V.Nikhileswar
M. Tech 2 nd year, IT Department Sreenidhi Institue of Science & Technology Yamnampet, Hyderabad

Sunil Bhutada
Professor, IT Department Sreenidhi Institute of Science & Technology Yamnampet, Hyderabad

Mekala Sreenivas
Associate Professor, IT Department Sreenidhi Institute of Science & Technology Yamnampet, Hyderabad

Abstract- There are different risks associated with bank loans nowadays, especially for banks to reduce their capital loss. Risk analysis and default assessment becomes crucial. Banks hold huge volumes of customer behavior-related data from which they are unable to reach the verdict when an applicant may or may not be defaulting. Descriptive analytics will help an organization know what has happened in the past, giving you past analysis using the stored data. It is necessary to know the past actions which help them to make statistical decisions using historical data. Data Mining is a capable area of data analysis which aims to extract useful knowledge from a tremendous amount of complex data sets. In this paper, Using predictive models you will help to identify the right customer. You need to evaluate the variables impacting credit risk using historical evidence from the customer of the bank, build actions to reduce the acquisition danger, and analyze the financial value of the project. The final model is used with the test data set for estimation, and the experimental results prove the utility of the designed model.

Keyword-Credit Risk, Data Mining, Predicting, R

I. INTRODUCTION

Numerous banks have been the target of credit risk identification, and they have been working relentlessly to decrease the dangers of credit. The danger of credit is assumed to be the possibility that a customer won't pay back an advance. Credit risk identification is the procedure for finding out if a borrower can default at a later stage. This procedure permits banks to lessen likely dangers and can build the measure of advances. The result of this credit risk identification would be a gauge of the candidate's default likelihood (PD). It is in this manner imperative to make a model which considers the various parts of the candidate and which gives a sign of the candidate's defaulting. A ton of such work has been done previously, yet the utilization of the highlights remembered for the R bundle has not been investigated. R bundle is an exceptional measurable and information mining device that can deal with any measure of organized just as unstructured information and produce results effectively, introducing brings about both content and realistic ways. This encourages the chief to precisely foresee and decipher the outcomes. The objective of this work is to propose an information mining application that utilizes R to PD for new bank advance candidates. The information utilized for investigation incorporates numerous errors, for example, missing qualities, exceptions and abnormalities.

II.LITERATURE REVIEW

In [1] the author presents an effective prescient model for foreseeing exact clients who have applied for bank credits. decision Tree is applied with the end goal of foreseeing the trustworthiness important properties. This test model may not be utilized to coordinate client advance requests. The structure proposed in [2] was assembled utilizing information from the monetary segment to conjecture the status of credits. This model uses three calculations for arrangement, for example J48, Bayes Net and gullible Bayes. The model is executed and approved utilizing Weka. The best J48 calculation was picked based on accuracy. A serious Multidimensional Likelihood Grouping Calculation is acquainted in [3] with decide candidates for terrible credits. The primary and optional level danger investigations are utilized in this exploration, and Affiliation Rule is executed to forestall duplication. The

methodology proposed is gauging with more noteworthy accuracy and takes less time than past models. A choice tree model has been utilized as a classifier in [4], and an element choice is utilized for the extraction of traits. The model was tried utilizing Weka. The work in[5] proposes two models of FICO assessments, utilizing information mining procedures to help Jordanian business bank loaning choices. Results, considering the accuracy rate Show that the strategic relapse model has performed better than the useful spiral base model. The investigation at [6] building numerous non-parametric models for credit scoring. These depend on the multilayer perceptron technique. A research measures their effectiveness against other models that apply traditional linear discriminatory analysis, and techniques of quadratic discriminating analysis. The findings indicate the architecture of the neural network is outperforming the other three strategies. The research in [7] contrasts credit-scoring models based on a vector machine that were designed using default meanings in Wide and Narrow. It has been shown that models constructed from default Wide definition will outperform models created from default Narrow meaning. Bank advance defaults risk assessment, kind of score and distinctive information mining procedures, for example, Choice Tree, Arbitrary Backwoods, Boosting, Bayes arrangement, Sacking calculation and different methods utilized in money related information investigation have been contemplated in[8]. An objective of the examination [9] is to assemble an unmistakable endurance model to test default chance and give trial proof utilizing the Italian financial framework. The work in [10] screens the appropriateness of the coordinated model to an example informational collection taken from indian banks. The model is a variety of logistic regression, Radial Basis Neural Network, Multilayer Perceptron Model, decision Tree and Vector Machine Backing strategies. So it gauges the adequacy of such FICO assessment systems.

III.METHODOLOGY

Credit risk identificaion which distinguishes proof has become more crucial for businesses to lend their customers based on there validity. Towards this end, the methodology centres around true evaluations are presently the most searched after methodology by banking framework that need the endorsement of the bank chief. The most exact and broadly utilized proportion of the FICO assessment is the Likelihood of Default (PD). Defaulter is the individual who is probably not going to reimburse the measure of the credit or who will have over 90 days to pay the advance. The assurance of the PD is in this way an essential advance for the FICO assessment. Clients who are searching for a bank credit.

This paper presents a PD of the informational collection utilizing the suitable information methodological methodologies in the R Bundle. The information utilized for the plan and testing of this model will be taken from the UCI storehouse, it would be ideal if you For this reason, one of the financial informational indexes with 71295 records and 31 characteristics is utilized. The mathematical information design is stacked into the R programming and a lot of information readiness steps is executed. It is utilized to develop an order model before the equivalent. The dataset which we have chosen doesn't have any missing data. Nonetheless, continuously, there is a danger that the informational collection contains a few missing or ascribed information that should be supplanted by real information produced using the genuine information accessible. For this capacity, the closest k neighbor calculation is utilized to play out various ascriptions. This is accomplished with the trouble examination capacity of the DMwR bundle. The mathematical highlights are streamlined before this move.

The dataset has numerous properties that characterize the believability of customers searching for various sort of loans. Qualities for these properties may have exceptions that don't fit inside the ordinary scope of information. It is subsequently important to eliminate the anomalies before the dataset is utilized for additional demonstrating. The anomaly discovery for quantitative highlights is performed utilizing the level () of the capacity. For numeric an element, the boxplot strategy is utilized for anomaly recognition and is executed utilizing the daisy () capacity of the bunch bundle. In any case, before that, the numeric information must be standardized into a space of [0, 1]. The heaviness of the proof (misfortune) structure is utilized in the outward positioning. This is practiced utilizing the Anomalies rating () capacity of the DMwR bundle. After the distant information is positioned, those out of range are disregarded and the rest of the information focuses are stacked with invalid qualities.

Data anomalies, such as the imbalanced dataset, must be balanced before the classification model is built. A variety of real-time datasets has this problem and thus need to be fixed for better performance. Previously, however, this process needs that the test dataset be divided into different mixing training and test datasets (i.e. training dataset 70 percent of the data and thirty percent of the data will be the test dataset). The balancing stage will now be performed by training dataset using the SMOTE() package DMWR feature.

Next, utilizing the preparation of the dataset, the correlation between's the various credits should be verified whether there is any excess data spoke to utilizing two qualities. This is finished utilizing the plotcorr () capacity of the circle bundle. The interesting highlights will at that point be positioned and, in view of a specific edge, the quantity of exceptionally positioned highlights will be picked for the model structure.

The subsequent informational index with a decreased number of highlights is presently prepared for use by the characterization calculations. Recognizable proof is one of the strategies for information investigation that predicts the marks of the class. Distinguishing proof can be completed with the utilization of choice trees is one of numerous structures and one of the most reasonable for the difficult picked. Arrangement is acted in two phases – I utilize the class names of the preparation dataset to build the choice tree model and (ii) this model will be utilized on the test dataset to anticipate the class marks of the test dataset. The rpart() capacity of the rpart parcel will be utilized for the initial step. Anticipate() is utilized to play out a subsequent advance. The accompanying expectation is then tried against the first test dataset class names so as to survey the exactness of the model.

CRISP-DM FRAMEWORK

The CRISP- DM Methodology (Cross Industry Standard Process For Data Mining) (CRISP- DM, 2007) Was Used To Build A Classification Model. The model identifies the different stages in implementing a data mining project, as described below.

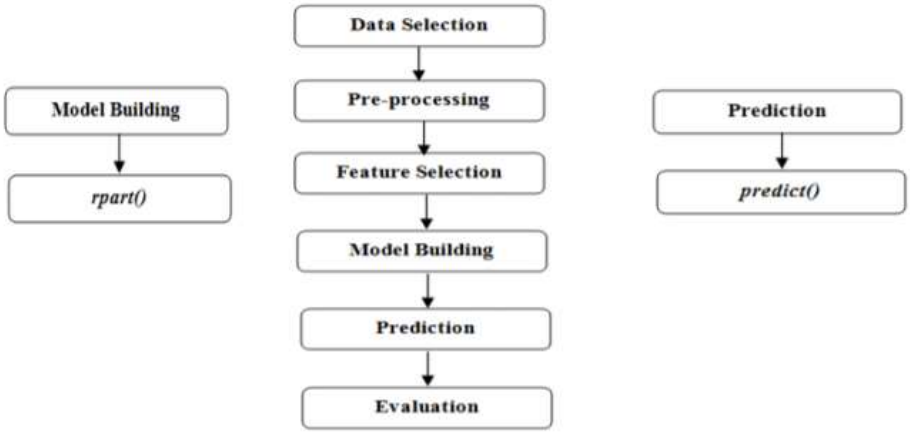

Figure 1: Steps Of Crisp-Dm Framework

3.1 STEPS OF CRISP-DM FRAME WORK

1. Business understanding: The first phase of the CRISP-DM is the Business Understanding the Company Perception is the first step of CRISP-DM. for the sake of this paper this process is intended to identify the Bank's business objectives. The suggested goal is the identification of Fraud with a track record of fraud. It should also be mindful of the need to collect data in such a way that Gain a better understanding of those transactions which can lead to fraud. A successful assessment of the current situation in the banks is also very relevant, especially with regard to losses due to fraud is affecting the customers, and the bank itself. After the model has been implemented the Evaluation will check that these risks have been minimised.
2. Data Understanding: The second phase of the CRISP-DM is the Data Understanding. It is important to collect the initial data and generate a summary of this data, as well as a check of its consistency. It is where the bank's fraud history is synthesized, with the necessary attributes like the date of the fraud, the number of frauds, the forms of fraud, etc.
3. Data preparation: The next step is to prepare the data for import into software for fraud detection, so this is the Data Preparation process. In our case study we prepare data for use in the logistic regression. It is the phase of finding calculated fields, incorporating external databases, performing a good cleaning of data and classifying attributes as irrelevant, categorical and numerical.
4. Modeling: This process uses data modeling techniques that were developed in the step of data preparation to pick, seek and use an effective modeling technique, such as neural networks. They use the decision trees in our case study, using a database to prepare, verify and check bank frauds.
5. Evaluation: In this step, a testing procedure is conducted to determine whether we have used the right data mining method and verify that the data is actually representing the concept that was known in the process of business

understanding. The method returns to the Market Comprehension step and reiterates the whole cycle if further processes are to be modelled

6. Deployment: During this step, a testing procedure is conducted to determine that we have used the right data mining technique and verifies that the data is actually representing the truth that was known in the process of business comprehension. The method returns to the Business Understanding step and reiterates the entire cycle if further systems are to be modelled. The processing steps in this modelling methodology are shown as below.

Phase 1: Collection of data
Phase 2: Pre-processing data
Step2.1: Identification of outsiders
Step2.2: Outlier rankings
Step2.3: Elimination outside
Step2.4: Imputation reduction
Step2.5: Separated Training & Test Dataset
Phase 3: Exploratory Data Analysis
Phase 4: Model Building
Phase 5: Application Scorecard

Phase 1: Data Selection

Demographic data: This is obtained from the information provided by the applicants at the time of the request for a credit card. It includes customer-level information on age, gender, income, marital status, etc.

Credit bureau data: It is taken from the credit bureau and includes factors such as 'number of times in the last 3/6/12 months 30 DPD or worse,' 'outstanding balance,' 'number of transactions, etc. After the data set is selected and understood, it is loaded into the R programme using the code below. With the name credit data the dataset is loaded into R.

Phase 2: Pre-Processing data

- Removed duplicates in the data by checking on application id of customers.
- Merged both Demographic and Credit Bureau data on the key application id, created a master file. Check whether predicted variable has any missing values.
- As there are some records exists with predicted variable missing, which indicates that those applicants have been rejected. We sub setted the data which has predicted variable missing and kept it aside for further usages in model evaluation. After sub setting we removed those sub setted data from master data set.
- Then we have checked for any missing values in predictor variables. We have found some variables which are having values missed.
- As we have decided to do missing value imputation using WOE, we performed EDA as the next step to assess which are important predictor variables.

Column Name	Missing Data	Erroneous Data
Age	-	20 wrong data ranging from -3 to 0
Gender	2 rows doesn't have any value	-
Marital Status	6 rows doesn't have any value	-
Number of Dependents	3 rows doesn't have any value	-
Income	-	81 rows have income less than 0.
Education	119 rows doesn't have any value	-
Profession	14 rows doesn't have any value	-
Type of Residence	8 rows doesn't have any value	-
No of months in current residence	-	-
No of months in current company	-	-
Performance Tag	1425 rows doesn't have any value	

Figure 2: Data Quality Issues In Demographic Data

Column Name	Missing Data	Erroneous Data
No of times 90 DPD or worse in last 6 months	-	-
No of times 60 DPD or worse in last 6 months	-	-
No of times 30 DPD or worse in last 6 months	-	-
No of times 90 DPD or worse in last 12 months	-	-
No of times 60 DPD or worse in last 12 months	-	-
No of times 30 DPD or worse in last 12 months	-	-
Avgas CC Utilisation in last 12 months	1058 rows doesn't have any value	
No of trades opened in last 6 months	1 row doesn't have any value	-
No of trades opened in last 12 months	-	-
No of PL trades opened in last 6 months	-	-
No of PL trades opened in last 12 months	-	-
No of Inquiries in last 6 months (excluding home & auto loans)	-	-
No of Inquiries in last 12 months (excluding home & auto loans)	-	-
Presence of open home loan	272 rows doesn't have any value	-
	272 rows doesn't have any value	-

Figure 3: Data Quality Issues In Credit Bureau Data

Phase 3: Exploratory Data Analysis

INCREMENTAL ANALYASIS

Let's understanding the incremental gain within the levels of categorical variables
Creating a data frame which contains the incremental values

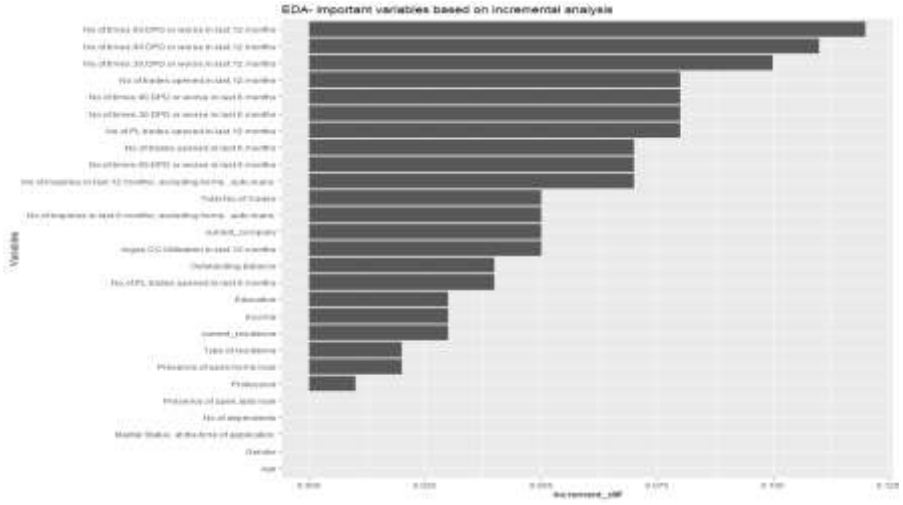

Figure 4: Important Variables Based On Incremental Analysis

Weight Of Evidence

- The method used above is a good but crude way to understand the importance of variables
- For default prediction. We want to quantify the importance of each predictor variable.
- In other words, we want to find the 'information value' of each variable.
- The weight of evidence (WOE) shows the predictive power of an independent variable in relation to the dependent variable.
- Woe is a measure of how well a variable separates the good customers from the bad ones. So, woe= ln(Distribution of good/Distribution of Bads)
- Information value and WOE calculation

$$WOE = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right)$$

Figure 5: Woe Calculation

$$IV = \sum (\%\ \text{of non-events} - \%\ \text{of events}) * WOE$$

Figure 6: Information Value Formula

Phase 4: Model Building

- As demographic data is merged in the main dataset we sub setted only demographic data from merged set and used for model building.
- We did all the data preparation activities on demographic data. We removed data where variables having missing values as the number of records are less than 2%.
- Checked for any outliers and replaced outliers with recent quantile values.
- Then we go on building model by dividing data into train and test data sets.
- After building model we have checked how good our model is performing on test data and also analyzed what are the important variables our model has given. Next step we used the entire merged data which has missing value imputation in our WOE analysis to build model and predict default.As we have different models to try out, we have chosen logistic regression model to start with.
- We converted all categorical variables to dummy variables.
- We divided data into train and test and built model using train data set.
- We did check P-Value and VIF for variable importance and correlation factors and removed the variables which are of less importance and high correlation factor.
- In the final model we are left with the variables which are of highly important in predicting the default of an applicant.

LOGISTIC REGRESSION

Logistic regression is the most commonly used technique in the market for credit scoring model development (ROSA, 2000; OHTOSHI, 2003).

The dependent variable in logistic regression analysis is normally a binary variable. (nominal or ordinal), and the independent variables may be either categorical (as long as they are dichotomised after transformation) or continuous. The Logistic Regression model is a special case of the Generalized Linear Templates (DOBSON 1990; PAULA 2002). The characteristic function of the model is given by the

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta 0 + \beta 1.\, x1 \qquad (1)$$

Most important of these are:

1. Linearity in the Variables explanatory.
2. A lack of interactions between explanatory variables.

In terms of nature, the Logistic Regression model is simplest. We can consider it has one explanatory variable for its simplicity then we can rewrite it as

Now suppose x is a Boolean variable, then we get two equations:

$$\ln\left(\frac{p(1)}{1-p(1)}\right) = \beta 0 + \beta 1. \qquad (2)$$

$$\ln\left(\frac{p(0)}{1-p(0)}\right) = \beta 0 \qquad (3)$$

This two together lead to neat result:

$$\frac{(0)}{1-(0)} = e^{\beta 1}\,\frac{p(1)}{1-p(1)} \qquad (4)$$

This is the interpretation of the co-efficient β1. It describes the change in the default probability if the change in variables is exactly 1 unit.

So finally the simple logistic equation can be represented by

$$P(\text{Loan Status = default or 1}) = \frac{1}{1+e^{-(\beta 0+\beta 1 x+\cdots+\beta k\, xk)}} \qquad (5)$$

When k is the number of independent variables,

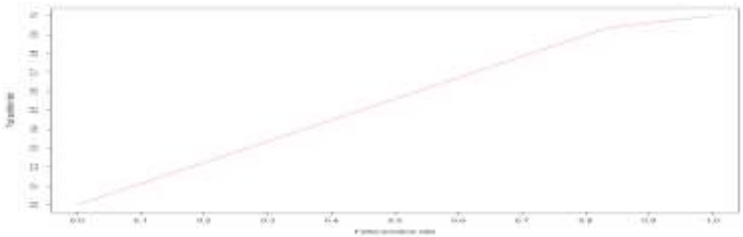

Figure 7: Defaut Prediction Rate On Logistic Regression

MODEL EVALUATION

- From the models that we have built we conducted tests on the models which gives more default prediction rate.
- For this we followed different approaches respective to the models.
- We have used rejected population which we kept aside to assess the model performance.
- We have checked Accuracy, Sensitivity and Specificity of all the models.
- We checked which model is giving best predicted probability of default.
- We found that logistic regression model is predicting the likelihood of default.
- We have chosen that logistic regression model is best for our data.

Phase 5: Application Score Card

Here we will build application scorecard for each applicant by using the odds obtained for each applicant.

$$\text{Log (odds)} = \sum \beta i x i$$

Once we get the odds of we will sort applicants from high to low odds. Then we will scale these odds for getting scores between 200 to 900. After the scores are calculated we can decide some threshold on which and applicant will be labelled as ‘good’ or ‘bad’.

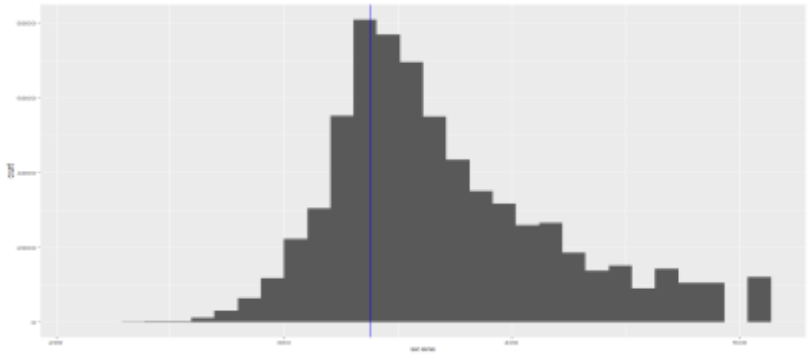

Figure 8: Application Score Card

- Cut-off: 338 is the baseline for providing credit card to the customers
- 70% of defaulters correctly identified. Average score of rejected population is less than the average score of approved* population
- Total rejected applications by bank: 1423
- Identified correctly at cut-off score by model: 1006

IV.CONCLUSIONS

The objective of this study was to establish predictive models for credit scoring using machine learning algorithms b ased on data from a broad banking sector financial.

When designing the models of credit rating such care should be taken to ensure the model's accuracy and its subsequ ent applicability. The steps taken in this study were precautions in sampling, consistent specification of parameters f or the classification of good and poor clients, and treatment of variables in the database prior to implementation of th e techniques, with the goal of maximizing results and reducing errors.

REFERENCES

1.M. Sudhakar, and C.V.K. Reddy, "Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5(3), pp. 705-718, 2016.

2.J. H. Aboobyda, and M.A. Tarig, "Developing Prediction Model Of Loan Risk In Banks Using Data Mining", Machine Learning and Applications: An International Journal (MLAIJ), vol. 3(1), pp. 1–9, 2016.

3.K. Kavitha, "Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp. 162–166, 2016

4. Z. Somayyeh, and M. Abdolkarim, "Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran", Jurnal UMP Social Sciences and Technology Management, vol. 3(2), pp. 307–316, 2015

5. A.B. Hussain, and F.K.E. Shorouq, "Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach", Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014

6. A. Blanco, R. Mejias, J. Lara, and S. Rayo, "Credit scoring models for the microfinance industry using neural networks: evidence from Peru", Expert Systems with Applications, vol. 40, pp. 356–364, 2013.

7. T. Harris, "Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions", Expert Systems with Applications, vol. 40, pp. 4404–4413, 2013.

8. A. Abhijit, and P.M. Chawan, "Study of Data Mining Techniques used for Financial Data Analysis", International Journal of Engineering Science and Innovative Technology, vol. 2(3), pp. 503-509, 2013.

9. D. Adnan, and D. Dzenana, "Data Mining Techniques for Credit Risk Assessment Task", in Proceedings of the 4th International Conference on Applied Informatics and Computing Theory (AICT '13), 2013, p. 105-110.

10. G. Francesca, "A Discrete-Time Hazard Model for Loans: Some Evidence from Italian Banking System", American Journal of Applied Sciences, 9(9), pp. 1337–1346, 2012.