MITIGATION OF BANKING AND FINANCIAL SERVICES INVOLVED RISKS

V. Nikhileswar [1], Sunil Bhutada [2], Mekala Sreenivas [3]

[1]M. Tech Student, IT Department, Sreenidhi Institute of Science & Technology, Yamnampet, Hyderabad, Email: nikhil.niki205@gmail.com

[2]Professor, IT Department, Sreenidhi Institute of Science & Technology, Yamnampet, Hyderabad, Email: sunilb@sreenidhi.edu.in

[3]Associate Professor, IT Department, Sreenidhi Institute of Science & Technology, Yamnampet, Hyderabad Email: msreenivas@sreenidhi.edu.in

Abstract: Nowadays there are numerous risks associated with bank loans, especially with banks reducing their capital losses. Risk estimation and measurement of defaults are critical. Banks retain large quantities of data relating to consumer conduct on which they are unable to draw a decision whether a claimant may or may not be defaulting. Descriptive analytics can let a company know what has been in the past, leveraging the stored data to provide you with past analysis. Past actions that help them make statistical decisions using empirical evidence need to be learned. Data Mining is a capable field of data processing aimed at collecting valuable information from a large number of complex data sets. In this paper, you can help to classify the correct client by using statistical models. Using historical data from the bank's client, you need to assess the factors influencing credit risk, establish measures to reduce the acquisition risk, and analyze the project's financial worth. For the prototype data set for estimation, the final model is used and the experimental findings show the usefulness of the constructed model.

Keyword: Credit Risk, Data Mining, Predicting, R

I.INTRODUCTION

The goal of credit risk detection has been various banks, and they have been working tirelessly to reduce the risks of credit. It is believed that the risk of credit is the chance that a consumer will not pay back an advance. The procedure for finding out if a borrower can default at a later stage is credit risk identification. This procedure allows banks to reduce the likelihood of hazards and to measure progress. A gauge of the default likelihood (PD) of the candidate would be the result of this credit risk identification. In this way, it is imperative to create a model that considers the different sections of the candidate and provides a sign of the default of the candidate. Previously, a lot of such work was done, but the use of the highlights remembered for the R bundle was not investigated. R Package is an excellent measured and data mining platform that can accommodate any measure of ordered as well as unstructured knowledge and reliably deliver outcomes, introducing both material and practical forms. This encourages the boss to accurately anticipate and decipher the results. The aim of this work is to propose an application for information mining that uses R to PD for fresh bank advance candidates. Numerous errors are included in the information used for investigation, such as missing characteristics, exceptions and abnormalities.

II. LITERATURE REVIEW

The author presents an efficient prescient model in [1] to foresee exact customers who have applied for bank credits. The final objective of predicting the important properties of trustworthiness is to apply the Decision Tree. To coordinate client advance requests, this test model may not be used. The structure suggested in [2] was assembled to conjecture the status of

credits using monetary segment information. This model uses three structure equations, such as J48, Bayes Net and gullible Bayes, for instance. The model is executed using Weka and approved. Based on accuracy, the best J48 calculation was picked. In [3], a serious Multidimensional Probability Grouping Equation is known for determining applicants for awful credits. In this exploration, the main and optional level risk investigations are used and the Affiliation Rule is executed to forestall duplication. The methodology proposed is to gauge with more remarkable precision and takes less time than previous models.

As a classifier in [4], a choice tree model has been used and an element choice is used for the extraction of traits. Using Weka, the model was tried. Two models of FICO evaluations are proposed in the work in [5], using information mining procedures to assist Jordanian business bank loaning choices. Results show that the strategic relapse model has performed better than the useful spiral base model, considering the accuracy rate. Inquiry [6] on the construction of numerous non-parametric credit scoring models. These depend on the technique of the multi layer perceptron. Research measures its efficacy against other models that use traditional linear discriminatory analysis and quadratic discriminatory analysis techniques. The results indicate that the neural network architecture is outperforming the other three strategies.

The study in [7] compares credit-scoring models based on a vector machine that were designed in Wide and Narrow using default meanings. Models constructed from the Wide default definition have been shown to outperform models created from the Narrow default meaning. For example, Choice Tree, Arbitrary Backwoods, Boosting, Bayes arrangement, Sacking calculation and various methods used in money-related information investigation, Bank advance defaults risk assessment, type of score and distinctive information mining procedures have been contemplated in [8].

The aim of the review [9] is to assemble an unmistakable endurance model to test the chance of default and to provide trial evidence using the Italian financial framework. The work in [10] screens the appropriateness of the coordinated model to a collection of information taken from Indian banks as an example. The model provides a number of methods for logistic regression, Radial Basis Neural Network, Multilayer Perceptron Model, Decision Tree and Vector Machine Backing. It thus measures the adequacy of such FICO evaluation systems.

III. METHODOLOGY

Credit risk recognition that identifies data has become more relevant for firms to lend their customers on the basis of their legitimacy. To this end, the methodology focusing on true assessments is currently the most searched for banking framework methodology that needs the bank chief's endorsement. The Probability of Default (PD) is the most reliable and broadly used proportion of the FICO evaluation. The default is the one who is not likely to repay the credit measure or who may have more than 90 days to pay the advance. In this way, the PD's assurance is a significant advance for the FICO evaluation. Clients who are looking for credit from a bank.

This paper introduces a PD of the knowledge collection in the R Package using the necessary data methodological methodologies. It will be ideal if you were to use one of the financial information indexes with 71295 documents and 31 characteristics for this purpose, the information used for the plan and testing of this model would be taken from the UCI storehouse. In R programming, the mathematical data architecture is stacked and a lot of knowledge readiness steps are done. It is used prior to the counterpart to build an order model. The dataset we have used has no data that is incomplete. Nonetheless, there is a continuing possibility that the information set includes a few lost or ascribed information that could be supplemented with actual information produced using available genuine information. To perform different ascriptions, the closest k neighbour calculation is used for this power. This is done with the DMwR bundle's trouble-examination feature. Prior to this step, the mathematical highlights will be simplified.

The dataset has various features that describe the reputation of consumers looking for different kinds of loans. There may be exceptions to qualities for these properties that do not fit within the ordinary information scope. It is also important that

the abnormalities be removed before the dataset is used for further demonstration. Using the degree () of capability, the anomaly discovery for quantitative highlights is carried out. The boxplot technique is used for anomaly detection for numerical an entity and is performed using the daisy() skill of the bunch package. In any event, the numeric information must be normalized into a space of [0, 1] before that.

The heaviness of the structure of evidence (misfortune) is used in the outer placement. This is achieved using the DMwR bundle's Abnormalities Ranking () functionality. Those out of sight are disregarded until the distant information is placed, and the majority of the information focuses are stacked with invalid characteristics.

Before the classification model is constructed, data anomalies, such as the imbalanced dataset, must be balanced. This problem has a variety of real-time datasets and thus needs to be fixed for better performance. However, historically, this procedure demands that the test dataset be broken into separate combining testing and test datasets (i.e. 70 percent of the data from the training dataset and thirty percent of the data from the test dataset). The testing dataset using the SMOTE() package DMWR function will now perform the balancing step.

Next, using the dataset planning, the similarity between the different credits should be checked if there is any surplus data talked of using two characteristics. This is completed using the circle bundle's plotcorr() functionality. The fascinating highlights will be placed at that point and the quantity of extraordinarily positioned highlights will be chosen for the model structure in view of a particular edge.

For the characterization measurements, the subsequent knowledge index with a reduced number of highlights is currently prepared for use. Recognizable data is one of the intelligence investigation techniques that project the class's marks. With the use of chosen trees, identifying facts can be completed is one of multiple systems and one of the most reasonable for the difficult to select. The arrangement is carried out in two phases: I use the class names of the preparation dataset to construct the model of the selection tree, and (ii) this model is used to anticipate the class marks of the evaluation dataset on the test dataset. For the initial stage, the rpart() feature of the rpart module will be used. Anticipate() is used to render a corresponding progression. The following expectation is then attempted against the class names of the first evaluation dataset in order to evaluate the model's accuracy.

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.

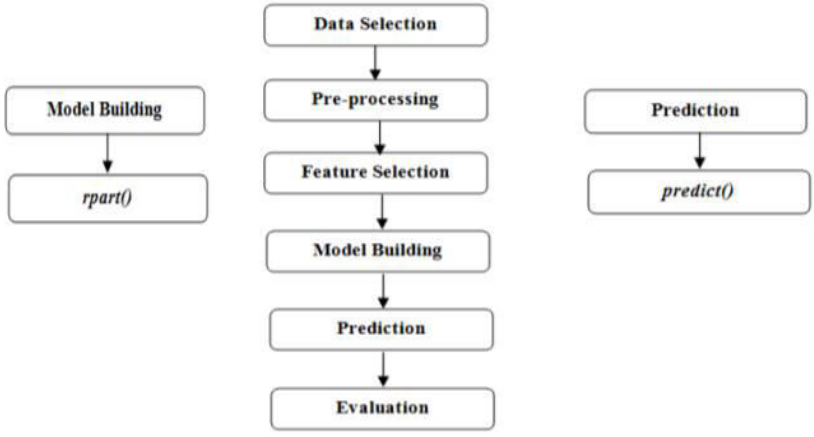

Fig 1: Steps of Crisp-Dm Framework

3.1 STEPS OF CRISP-DM FRAME WORK:

1. Business understanding: The first phase of CRISP-DM is the first phase of CRISP-DM, Market Understanding the Perception of the Organization. This procedure is intended to define the business priorities of the Bank for the purpose of this report. Identification of crime with a track record of fraud is the proposed target. The need to gather data in such a manner to have a better understanding of certain transactions that can lead to theft should also be taken into consideration. A successful evaluation of the current situation in banks is therefore very significant, particularly with regard to the damages incurred by fraud involving clients and the bank itself. The evaluation will verify if these risks have been reduced after the model has been applied.

2. Data Understanding: Data Comprehension is the second step of CRISP-DM. Collecting the initial data and producing a description of this data, as well as verifying its accuracy, is important. That is where the fraud history of the bank is synthesized, with the requisite characteristics such as the fraud date, the number of frauds, the types of fraud, etc.

3. Data preparation: The next step is to prepare the data for importation into fraud detecting applications, so this is the method of data processing. We are planning data for use in logistic regression in our case study. It is the stage in which measured fields are found, external records are integrated, data is cleaned properly and attributes are categorized as irrelevant, categorical and numerical.

4. Modeling: This method uses data modeling techniques that have been developed to select, scan and use an efficient modeling technique, such as neural networks, in the stage of data preparation. In our case study, they use decision trees, using a database to plan, validate and review bank frauds.

5. Evaluation: In this stage, a checking protocol is performed to decide if we have used the correct method of data mining and check that the knowledge actually reflects the definition understood in the market comprehension phase. If more processes are to be modeled, the system returns to the Consumer Comprehension step and reiterates the whole loop.

6. Deployment: A verification protocol is undertaken at this stage to establish if we have used the proper data mining methodology and check that the information actually reflects the facts established in the company understanding process. If more programmes are to be modeled, the procedure returns to the Market Awareness step and reiterates the whole method. As shown below, the processing steps in this modeling technique are given.

Phase 1: Collection of data

Phase 2: Pre-processing data

- Step 2.1: Identification of outsiders
- Step 2.2: Outlier rankings
- Step 2.3: Elimination outside
- Step 2.4: Imputation reduction
- Step 2.5: 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	

Fig 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 Utilization 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	-

Fig 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.

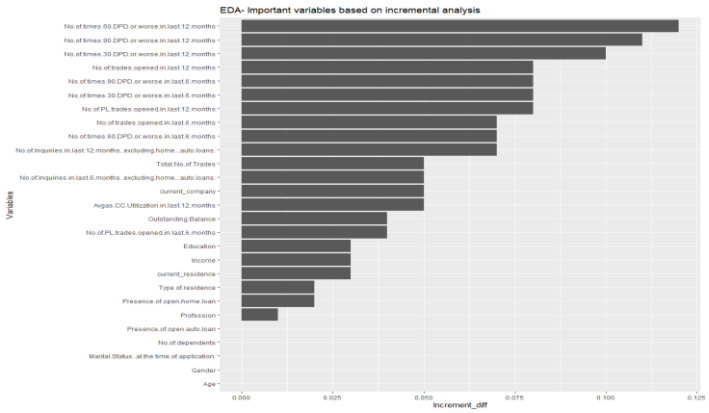

Fig 4: Important Variables Based On Incremental Analysis

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

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

Fig 5: Woe Calculation

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

Fig 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 dichotomized 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

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

When k is the number of independent variables,

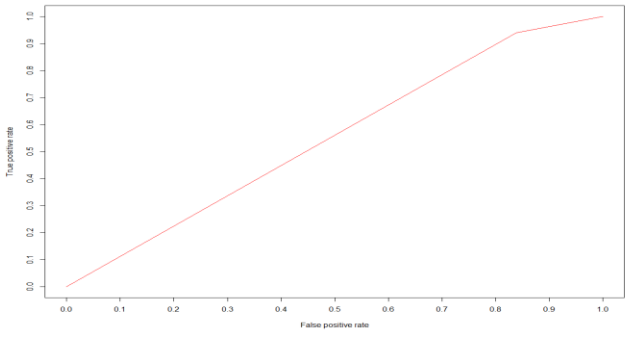

Fig 7: Default Prediction Rate on Logistic Regression

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.

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 labeled as 'good' or 'bad'.

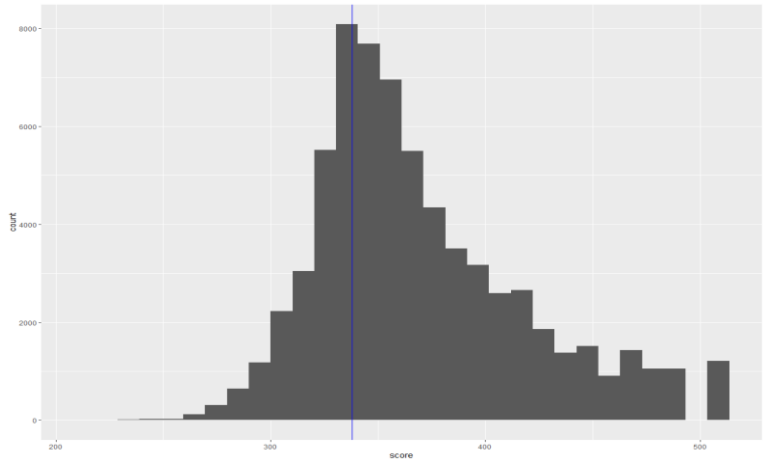

Fig 8: Application Score Card

- 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. CONCLUSION

The goal of this thesis was to use machine learning algorithms based on data from the large financial field of the banking sector to develop predictive models for credit scoring. Such caution should be taken when developing credit rating models to ensure the precision of the formula and its resulting applicability. Precautions in sampling, consistent definition of criteria for the classification of good and bad clients and treatment of variables in the database prior to the application of the techniques were the steps taken in this analysis, with the aim of optimizing performance and eliminating 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: Neural scoring 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.