# Credit Analysis in Banking Industry

[1]Aniesha Razdan, [2]Akshay Naik, [3]Mayank Pathak, [4]Ankush Hutke

[1,2,3]UG Student, [4]Assistant Professor, Rajiv Gandhi Institute of Technology, Mumbai, India,

[1] aniesha.razdan11@gmail.com, [2] akshaynaik299@gmail.com, [3] mayank74pathak@gmail.com, [4]ankush.hutke@mctrgit.ac.in

**Abstract: Our economy depends on data which is everywhere, in every section, in every country. We produce and convert it into useable data. Information allows us to enhance the business processes and provide our customers and partners with the best quality standards, services and products.**

**The phenomenal rise in information, and problems encountered while dealing with huge amount of data, make it necessary for an organization to introduce a technique that can overcome these problems and provide an effective solution. Every year the banking organizations, generate enormous amount of valuable data from their customers and their transactions. These valuable data need to be saved and analyzed effectively using big data analytic techniques so as to get the necessary and useful insights for the banking sector.**

**Big Data Analytics (BDA) provides a better consumer experience with better data management creating transparency, collecting more accurate and detailed performance data, setting up controlling experiments, segmenting populations to customize actions, and replacing/supporting human decisions making with automated algorithms.**

**The primary focus of our proposed work will be on identifying the issues that the banking sector faces in decision making while granting loans to customers, in detail and providing an optimal solution using the Big Data approach and tools like Hadoop, HDFS, Spark etc.**

***Keywords — Big Data Analytics, Big data, Hadoop, HDFS, Spark***

## I. Introduction

Enhancement in the economic development leads to a huge rise in the requirement of personal loans for customers as the behavior of the borrowers have uncertainty and fuzzy nature. As for the banking industry, banks provide finances to household and the companies by lending loan. One should pay focus towards the riskiness in lending loans to the customers. For both lenders and borrowers, credit risk is a major task, which directly or indirectly affects the reliability of the banks.

Every minute the banking organizations, generate enormous amount of valuable data from their customers and their transactions. These elixir data need to be saved and analyzed effectively using big data analytic techniques so as to get the necessary insights for the banking organizations. In today's market trend, analyzing large amount of data sets comprising of variety of data is of high attention to discover knowledge, market tendencies, customer likings and other business insights. With the capital investment enhancement of many enterprises, credit risk assessment in commercial banks is a crucial area which can have an important impact on stability of bank operation. Therefore, discriminating good borrowers from bad ones with high accuracy is of critical importance to them.

With effective classification among clients, the commercial banks can split customers into different levels by their information, and so the banks can decide whether to provide a loan to their clients according to the classification.

Before approving or rejecting a particular retail loan application, the credit division of a bank thoroughly evaluates that loan application. Loan evaluation refers to utilization of different types of techniques to support automatic decisions that have been utilized so far. Among the models used for this purpose, the decision making process relies on large numbers of historical data spanning over many years of providing credit and decision variables, being statistically analyzed and expressed in crisp values. However, because of information being ambiguous, incomplete, uncertain and imprecise, such approaches cannot help in modeling the way human experts make their decisions about the creditworthiness of the customer.

The banking industry is among many industries which have huge and useful data about their customers but very few banks are utilizing this set of information to enhance the

DOI : 10.35291/2454-9150.2020.0366

customer experience and using the data information to prevent intruder. The challenge is not about dealing with trillions of bytes of streaming data, it is about getting initial step with a quantitative approach so that you can drive value from your data, whatever length that data is. They are very well aware of the fact that if the data can be used effectively they can fulfill the needs of customer accurately.

Big Data promises huge impact on the banking and financial services which is why Big Data Analytics is in high demand and becoming essential in making the business decision and providing the biggest edge over the competitors as well as an ocean of opportunities out there for skilled professionals.

Major Scopes of Big Data in Banking and Finance industry in the present and near future include Customer Segmentation, Fraud Detection, Offering Personalized Services, Risk Management, Addressing Compliance Requirements etc. Big data analytics is an emerging trend, particularly in banking and finance and this analytics technology is expected to help the banking industry grow. Banks are taking the big step towards the new need of dealing with the continuous rise in generation of data, with the integration of big data technologies and applications. They have prepared themselves to take in and process large volumes of data created and collected for the future growth of the banking and finance industry, as well as many other business organizations.

## II. LITERATURE SURVEY

A lot of research works have been done on ways to enhance decision making for evaluation credit risks while granting loan to potential customers.

Shweta Yadav et al. [1], have measured the credit worthiness of the bank and to calculate the bank performance, they have collected the data and used the big data approach that is Hadoop which is used for analyzing the bank data by performing analytical tests on the data set of the bank based on various parameters so as to know its performance. The analysis has been done so as to know is it worth for a borrower and investors to buy or purchase loan also showing the worthiness of the Hadoop and its ecosystem. Analytical result of the bank loan which reveals the performance and the credit risk of the banks show that there is good opportunity for the borrowers to buy and purchase loan and the investors to invest so as to get higher interest rate with fewer risk but with greater returns.

Chen et al. [2], have studied patterns of default in big data of the banks considering that banks are often worried about the status of repayment because the behaviors of every borrower are always fuzzy and uncertain and ambiguity and credit risk of borrowers must be realized when they review the applications of personal consumer loans. The method of self-organizing mapping (SOM) was applied to split the borrowers into different groups. The result showed 5 groups of borrowers and the 5 degrees of default. Finally, this study comes to know characteristics of borrowers in each groups, and provides the appropriate strategy to each single groups to lower the credit risk in personal consumer loans.

S. Mammadli [3], has proposed a fuzzy logic model for retail loan evaluation. Before approving or denying, especially a retail loan, the credit division of a bank evaluates the loan application. The fuzzy model consists of five input variables such as “income”, “credit history”, “employment”, “character”, and “collateral condition” and single output variable which signals credit standing. Whether loan applicant’s credit standing shall be considered as “low”, “medium” or “high” depends on the control of membership for the linguistic terms of fuzzy output.

C. Xiaojie et al, [4], have utilized a proposed algorithm called simultaneous clustering and attribute discrimination (SCAD) and performs clustering and feature weighting simultaneously. First, the algorithm has been analyzed in detail through a series of compare experiments, confirming this algorithm to have the high clustering precision. Finally, the algorithm is relevant in the analyzing of the bank loan repaid information that can efficiently search the weightage association of the main factors in loan information and realize potential customer.

V. Kumar et al [5], have aimed to analyze the credit risk involved in peer-to-peer (P2P) lending system of “Lending Club” Company. The P2P system allows investors to get worthy higher return on investment as compared to bank deposit, but it comes with a risk of the loan and interest not being repaid. Ensemble machine learning algorithms and preprocessing techniques are used to explore, analyze and determine the factors which play main role in predicting the credit risk involved. A loan is considered “good” if it's repaid with interest and on time. The algorithms are optimized to favor the potential good loans whilst establish defaults or risky credits.

Xudong Lin et al [6], explore the factors affecting the loan willingness of bank and implement some specific measures, which are conducive to solve financing difficult problem for various organizations. They have firstly put forward four crucial factors of which affecting bank's loan willingness, namely target profit, non-performing loan ratio, value customer ratio, business promotion ratio set by the bank itself, then based on likelihood theory, they have measured these factors, and established a simulation model to calculate loan willingness. As a result, with the change of reference point of four main effective factors, and the change of the upper bound and lower bound of four factors, the bank's loan willingness will alter accordingly.

A. Calis et al [7] aimed to smaller the rate of risk to minimum in decision making via analysis of existing personal loan customers and estimate potential customers’

DOI : 10.35291/2454-9150.2020.0366

payment outcomes with k-means method, one of the clustering techniques and decision trees method.

Roszbach et al [8] have proposed a model with a variable censoring threshold and sample selection effects is for the decision to provide a loan or not and the survival of granted loans. The model is shown to be an intended result tool to separate applicants with short survival times from those with long survivals, the bank´s loan provision process is shown to be inefficient. Loans are granted in a way that disagreement with both default risk diminish and survival time maximization. There is thus no trade-off between higher default risk and higher return in the protocol of banks.

Fernando A. F. Ferreira et al [9] provide a study that aims to create a system that integrates cognitive maps and the measuring attractiveness using a categorical based evaluation technique. The multiple criteria expert system delivers results that prove that this approach allows the credit risk evaluation process to be more transparent and informed.

Sudhamathy G. [10] aims to design a model and prototype the same using a data set available in the UCI repository, which is pre-processed and made ready to provide efficient predictions. The idea is to use a decision tree based classification model that uses the functions available in the R Package. The parameters contributing to the prediction and credit risk assessment are identified, and whether the loan applicant is a defaulter or not is predicted.

## III. Data Set

The dataset used has been extracted from Kaggle containing 10,000 entries. This data set contains random data designed manually for enabling users to perform computations on it and draw inferences from it post its detailed analysis. It contains total 18 attributes out of which credit score, annual income and age attributes have been selected for credit analysis.

| Attribute Name | Attribute Description | Attribute Type |
|---|---|---|
| Age | Between 30 and 50 (most preferred) | Numerical |
| Credit Score | Credit score type (1: 800 (best score), 2: between 600 and 800 (good score), 3: less than 600 (application rejected)) | Numerical |
| Annual Income | Higher your income, more will be your scope of borrowing. | Numerical |

**Table 1: Dataset Attributes**

This dataset was chosen as it provides apt financial information of customers that fulfills the requirement for the analysis to be carried out. As access to actual financial dataset from banks cannot be permitted, this dataset provides the basis for performing analysis using Big Data Analytics.

## IV. EXISTING SYSTEM

The existing system follows a step by step by procedure that takes a lot of time, effort and involves a lot of paperwork. It becomes troublesome for the loan applicants to visit the bank a numerous time to fulfill the formalities and requirements till the application is verified and approved.

Also in some cases, because of information being ambiguous, incomplete, uncertain and imprecise, such approaches cannot help in modeling the way human experts make their decisions about the creditworthiness of the applicant.

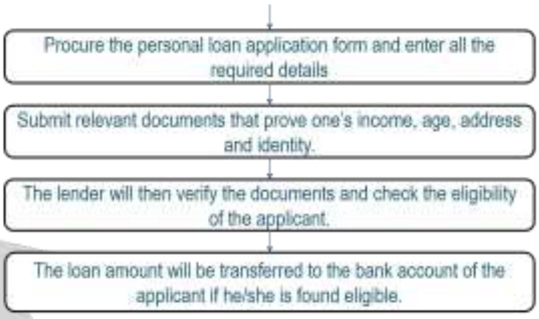
Procure the personal loan application form and enter all the required details
Submit relevant documents that prove one's income, age, address and identity.
The lender will then verify the documents and check the eligibility of the applicant.
The loan amount will be transferred to the bank account of the applicant if he/she is found eligible.

**Fig 1: Step by step procedure currently followed**

The above mentioned step-by-step procedure exhibits the long, time taking process of existing systems that is carried out to grant loan to customers. It takes minimum 7 days for the bank to verify all the credentials of the customer, check his/her eligibility, and evaluate credit risks to ensure that the loan is granted to potential customers with safety assured.

Thus, the main motivation behind this project is to simplify and fasten the process of granting loans to potential customers after validating them thoroughly as well as evaluation of the credit risk involved before sanctioning the loans.

## V. RESEARCH METHODOLOGY

The proposed work is used for identifying potential customer after evaluating credit risks for granting loans. Various components of the Hadoop ecosystem have been used to fulfill different stages of the verification process of the loan application. A dataset from Kaggle, world's largest data science community has been used for the implementation. The output given by the model is the decision, whether to grant loan to a particular customer or not.

### 4.1 Process Flow

**Step 1: Initial Phase: Input raw customer data**

The customer data of thousands and lakhs of customers associated with the bank is subjected to a cleaning process that includes its pre-processing to transform the raw data into standard and consolidated form. The pre-processing techniques include:

- Selection of relevant data
- Cleaning to remove errors and inconsistencies
- Integration of data from multiple sources
- Reduction to have a reduced representation of data
- Transformation

**Step 2: Application of Big Data Analytics techniques**

Using Big Data Analytics techniques, like data mining, regression analysis etc., the data is analyzed in detail to achieve the following and form a training data set.

- Segmentation of customers and Profiling
- Customer Feedback and Analysis
- Fraud Management and Prevention
- Analysis of customer spending patterns
- Risk Assessment etc.

**Step 3: Comparison of Customer data with training data set**

The training data set will act as a standard against which every customer data and application will be tested, compared to draw conclusions and make decisions effectively.

The analysis will help identify the nature of the customer and transactions made by him/her.

**Step 4: Analysis**

The application of Data Analytics technique will help in achieving a detailed analysis of customer data and fulfill the objectives:

- Efficient risk management that helps detect errors and frauds
- Correct understanding of customers and their separation from risky ones
- To achieve High precision and accuracy leading to proper data distribution
- Identification of the connection between data captured and possible results
- Transformation of business processes and conducts to identify business opportunities and potential threats.

### 4.2 Parameters considered for analysis of customer data:

- Credit Score
- Age
- Annual Income

### 4.3 Conditions for granting loan:

```
1. If Credit_Score > 700, then
       grant loan.
2. If Credit_Score < 600, then
       reject application.
3. If Credit_Score >= 600 and <=700, then
       check Annual_Income and age.
       if Annual_Income > 4,00,000 and age < 55,
       then grant loan.
```

## VI. RESULTS AND DISCUSSIONS

In this section, we explain the working of our proposed system illustrated with the help of screenshots shown below.

### 5.1 Beginning with HUE

Hue is a web interface that enables user to browse, query and visualize data. Hue takes the best querying experience with the most intelligent autocompletes, query sharing, result charting and download for any database and helps users find the correct data among thousands of databases and self-document it. The user enters the Hadoop Framework through Hue, where the dataset to be used will be uploaded.

**Fig 2: Hadoop User Interface**

### 5.2 Data stored in HDFS

The Hadoop Distributed File System (HDFS) is the key data storage system used by Hadoop applications. It is a java

DOI : 10.35291/2454-9150.2020.0366

based file system that provides expandable, fault tolerance, reliable and cost efficient data storage for Big data. and runs on commodity hardware. The dataset uploaded has to be stored in HDFS for processing.

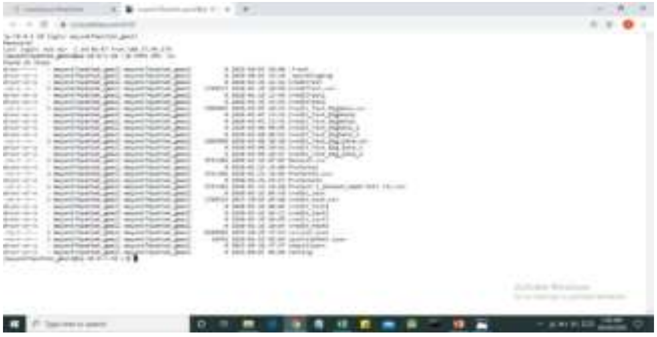

**Fig 3: Hadoop Distributed File System**

### 5.3 Spark User Interface

Spark is a lightning-fast cluster computing technology, designed for enhance computations which include interactive queries and stream processing, it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The processing starts in Spark using Python programming.

**Fig 4: Spark UI**

### 5.4 Data cleaning process by creating Resilient Distributed Dataset

Resilient Distributed Datasets (RDD) is a foundation in data structure of Spark. It is an unchangeable distributed collection of objects. Each dataset in RDD is split into logical partitions, which may be computed on different nodes of the cluster. RDDs can hold any type of Python, Java, or Scala objects, including user-defined classes.

Formally, an RDD is a read-only, distributed collection of records. RDDs can be generated through deterministic operations on either data on stable storage or other RDDs. RDD is a fault-tolerant collection of elements that can be control on in parallel.

**Fig 5: Data Cleaning Process**

### 5.5 Hadoop Output: Spark execution using file uploaded on HDFS and view of output in part-00000 file.

The file uploaded on HDFS can be processed using Spark. The previously used RDD creates logical partitions of the dataset. The output files are by default named part-yyyyy where: yyyyy is the task number (zero based).

**Fig 6: Hadoop Output**

### 5.6 Loading data into Spark DataFrame

Spark DataFrame is a distributed collection of data organized into named columns that provides operations to filter, group, or compute aggregates, and can be used with Spark SQL. DataFrames can be constructed from existing RDDs, structured data files, tables in Hive, or external databases. It eases the processing of tabular data and enables structured data manipulation.

**Fig 7: Loading Data into Spark DataFrame**

### 5.7 Schema of DataFrame in tree format

A schema gives the description of the structure of data. It can be implicit and inferred at runtime, or explicit and

DOI : 10.35291/2454-9150.2020.0366

known at compile time. The following showcases schema of DataFrame in tree format.

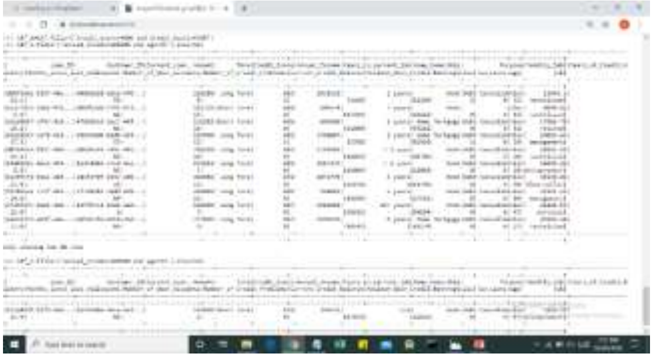

Fig 8: Schema of DataFrame in tree format

### 5.8 Filtering rows from DataFrame based on condition: 'Credit_Score>700'

The dataset can be manipulated, and selected data can be extracted using appropriate queries. Here, the rows from DataFrame are filtered on the basis of the given condition.

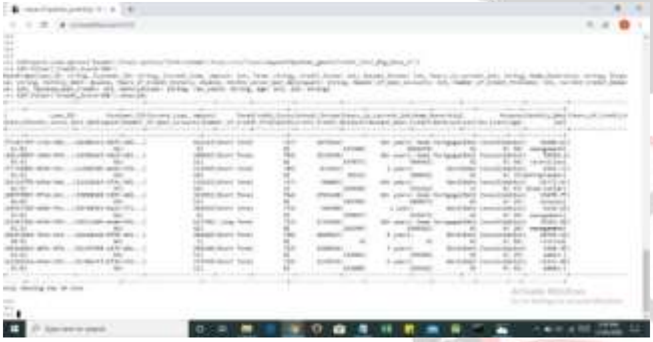

Fig 9: Filtering rows from DataFrame

### 5.9 Filtering rows from DataFrame based on condition: 'Credit_Score<600'

Fig 10: Filtering rows from DataFrame

### 5.10 Filtering rows from DataFrame based on condition: 'Credit_Score>=600 and Credit_Score<=700'

Fig 11: Filtering rows from DataFrame

### 5.11 Filtering rows from DataFrame based on condition: 'Annual_Income>400000 and age<55' for credit score between 600 to 700

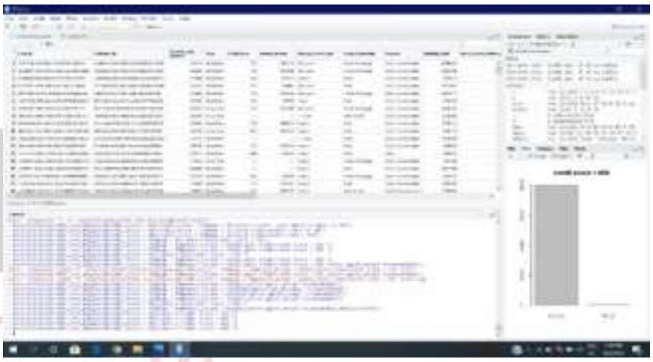

Fig 12: Filtering rows from DataFrame

### 5.12 Visualization

R is a language and environment for statistical analyses based on computing and graphics. R is also extremely flexible and easy to use when it comes to creating visualizations. One of its capabilities is to produce benchmark quality plots with minimum codes.

It offers a set of essential part of the functions and libraries to build visualizations and present data. For visualization of the output produced and to analyze the trends and patterns, R Studio is used.

In the following images, data that has been filtered from the DataFrames on the basis of given conditions has been visualized using R.

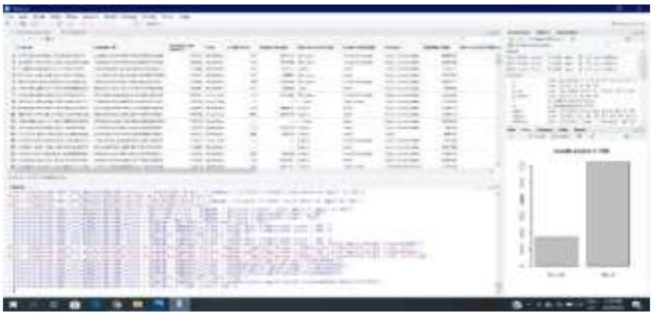

Fig 13: Credit_Score>700

DOI : 10.35291/2454-9150.2020.0366

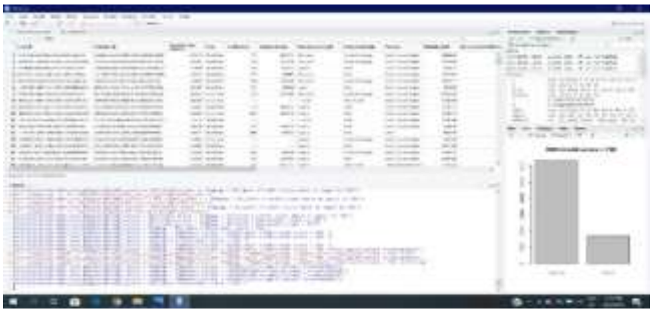

**Fig 14: Annual_Income>400000 and age<55 (For credit score between 600 to 700)**

From the experimental results, it was found that customers with credit score>600, and of age in between 30 to 50 with high annual income are considered the most by banks while granting loans to the applicants. These customers are categorized as safe and potential customers after evaluating the credit risks involved with respect to each one of them.

## VII. Conclusion

In this paper, a financial dataset depicting data entries of users has been taken that involves information regarding their loan application. After knowing the issues that the banking sector faces in handling huge amount of data, we chose Big Data Analytics to store, process and analyze the huge amount of data generated particularly when customers apply for loan. To simplify the whole validation and verification process of loan applicants and identify potential candidates, to whom bank can grant loans with minimum credit risk, we proposed our Loan Evaluation model that would overcome drawbacks of the manual process and be beneficial for the customers.

Our proposed work takes into consideration the essential parameters considered by banks for loan purposes including credit score, age and annual income. It verifies the eligibility of the customer based on these parameters and evaluates the credit risks involved while granting loans. This way, it fastens the whole loan application and decision making process by categorizing customers into safe and potential ones to whom the bank can give a nod for granting loan.

## References

[1] Shweta Yadav and Sanjeev Thakur, 2017, "Bank Loan Analysis using Customer Usage Data:A Big Data Approach Using Hadoop", 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)

[2] Chen, Qiao-ling, and Jian-bang Lin, November 2015, "Integrating of business intelligence and CRM in banks: An empirical study of SOM applied in personal customer loans," Fuzzy Theory and Its Applications (iFUZZY), 2015 International Conference on. IEEE.

[3] S.Mammadli, December 2016, "Fuzzy Logic Based Loan Evaluation System," Procedia Computer Science, pp.495-499.

[4] C. Xiaojie and D.Huailin, W.Qingfeng, July 2009, "SCAD algorithm and its application in analysis of bank loan," Computer Science & Education,. ICCSE'09. 4th International Conference, pp. 1934-1939.

[5] V. Kumar, Natarajan.S, Keerthana, Chinmayi K.M. and Lakshmi Lupu, September 2012, "Credit Risk Analysis in Peer-to-Peer Lending System" Knowledge Engineering and Applications (ICKEA), IEEE International Conference on IEEE, pp. 193-196.

[6] Xudong Lin, Lin Cheng, and W.Mao, Z.Qiu, June 2015, "Research on measuring the bank's loan willingness based on prospect theory," Service Systems and Service Management (ICSSSM), 12th International Conference on IEEE ,p.p 1-5.

[7] A.Calis, A.Boyaci, and K.Baynal,March 2015, "Data mining application in banking sector with clustering and classification methods, " Industrial Engineering and Operations Management (IEOM),International Conference on IEEE , p.p 1-8.

[8] Roszbach, Kasper, 1998, "Bank Lending Policy, Credit Scoring and the Survival of Loans," SSE/EFI Working Paper Series in Economics and Finance 261, Stockholm School of Economics.

[9] Fernando A. F. Ferreira, Ieva Meidutė-Kavaliauskienė, Edmundas K. Zavadskas, Marjan S. Jalali and Sandra M. J, January 2019, "A Judgment-Based Risk Assessment Framework for Consumer Loans," International Journal of Information Technology & Decision Making (IJITDM), World Scientific Publishing Co. Pte. Ltd., vol. 18(01), pages 7-33.

[10] Sudhamathy G., October 2016, "Credit Risk Analysis and Prediction Modelling of Banks using R", International Journal of Engineering and Technology 8(5):1954-1966.

DOI : 10.35291/2454-9150.2020.0366