iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5452

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Credit Card Customer Predicting using Machine Learning

K. SriLaxmi[1], N. Divya Sri[2], P. Lakshmi Durga Bhavani[3], A. Vidya Rani[4], S. Hameeda Fatima[5], A. Tanuj[6]
[1]*Assistant Professor,* [2, 3, 4, 5, 6]*Department of Information Technology, Sreenidhi Institute Of Science and Technology*

Abstract: ***Credit cards are an excellent way to make payments and used widely all over the world. Banks don't approve all the credit card applications. There are multiple factors and criteria for approval of a credit card including income, demographics, credit bureau etc. of the customer. Since past few years, the credit card providers had been experiencing a credit loss. This project aims to help a company to help find genuine customers using predictive models. Furthermore, by using the credit card provider's historical data, we can identify key factors for credit risk, build strategies to minimize the risk and evaluate the financial gain that the business may derive from the models. The methodology used in this project is "Logistic Regression," and model results from "random forest" can be evaluated by Model Evaluation. We are likely to obtain the right credit customer via that.***

Keywords: Credit Card, Logistic Regression, Random Forest

I. INTRODUCTION

The credit card emerged into the standard payment and credit card theft often grows rapidly. As we learn, of many popular mining methods used to identify credit fraud, such as Hidden motel, Fuzzy logic, etc. This study uses logistic regression and random forest that involves techniques for seeking the best answer to the problem and extracting the outcome of the fraudulent transaction implicitly. It is an important factor that a credit card issuing company must be incurred with a good fraud detection system to avoid these frauds which decreases the profits for the company.

Credit cards help to increase the purchasing power of the customer who otherwise, cannot afford it. Through a credit card it becomes easy to earn a huge sum of amount and hence the credit card frauds. A fraudster requires just a personal pin of the credit holder to use the card fraudulently. There may be a person/company who charges more from your credit card when compared to the promised amount without the user's recognition.

As a result, not only the customers but the credit card companies also face losses and hence it is also their responsibility to control these credit card frauds through introducing various techniques. Later, among the transactions, the fraudulent and genuine transactions are identified to detect the fraud.

We are solving above case study using logistic regression and random forest analysis techniques in R programming language. We build models using these to determine the significant factors contributing for credit card approval

A. Regression Analysis

It helps you to examine different independent and dependent variable at a time where the calculating of the data is easy. It is defined as the model that defines a output based on their past data that can eventually end up calculating between the reversible variable to the irreversible variable, it generally defines about the comparison between the two obtained values.

B. Logistic Regression

This is the model for calculating an event that happens or the probability of certain class that can end up being loss or fail at some extent of time. It can be calculated based on the datasets on obtaining a required solution of the occurring problem in today's world.

C. Random Forest

Random forests or decision forests for classifying the forming of trees at training time and the outputting of the classes is defined as in the form of trees and sub trees where the calculation of the decision based algorithm is easy and efficient for any user. The values thus obtained are calculated by defining the construction of the model.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

II. LITERATURE

With the increase in the fraudulent measures during the transactions, detecting them has always been a challenge. Auto encoder[1] which has two stage models establishes the required result of these models .A group of transactions can be trained and used in anticipating the upcoming transactions However, coping the future can be made easy. By constructing a personalizedQrt model on collecting the data using surveys and with the useof SVM [2] classifies the transactions. By using this personalized model they detect the fraud in the new transactions with most accuracy and build three different models for abnormal predictions.

The evaluation [3] of using the data processing techniques that eventually helps detecting the fraudulent methods. Modern techniques with the artificial intelligence and machine learning have been introduced which specifies about the mining method. Fraud transaction can be analyzed by performing diagnostic experts system with sets of data. Datamining and neural network algorithm [4] related techniques are used to minimize the number of false alarms and fraud detections.

Different cardinal components are used by Radial basis function network (RBFN) methodology[5] for analyzing the fraud score to minimize the future frauds in credit card transactions with the help of past research information of credit card.an other study [6] aims to define the used methodology and technologies held in Malaysian Banks. This work has allowed use of instruments and methods in data mining like gradual learning schemes with a view to improving classification methods and different supports to overloading their capability.

III.PROPOSEDSYSTEM

In the proposed system using R is aimed to help a company to pick out proper clients to use of forecast technique.Also, by using the past data of the credit card provider, we will determine elements impact credit card management, create techniques to moderate the acquisition hazard and check the financial advantage that company can get from the models.

The algorithm that has been used in this project and performance of the model can be analysed through Model Evaluation. Through this, we can acquire the right credit customer.

A. Credit Card Customer Prediction Algorithm

CRISP-DM seems to be a robust, six-phase, data mining methodology. This is a gradual technique which gives systematic approach for those processes of data mining. These 6 stages may be performed through either manner, however it will involve daily backtracking towards previous steps and repeating of acts every so often.

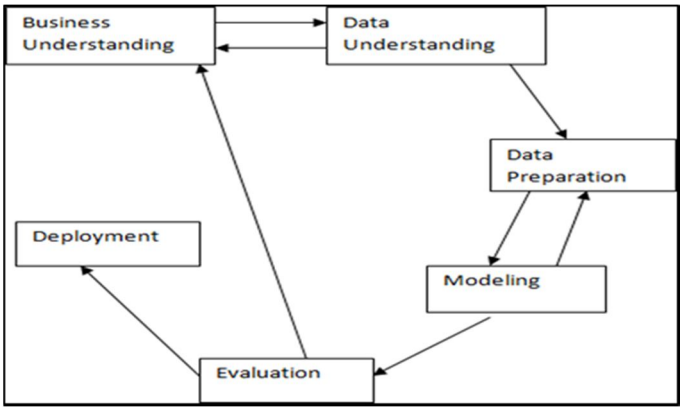


Fig 1: Architecture of Proposed System

The basic cross-industry framework for data mining, recognized as CRISP-DM, is such a free software configuration which defines frequent strategies often used experts in data mining. It is the most commonly used pattern for analytics

The six phases involved:

1) *Business Understanding:* During this phase, the corporate goals that have been set as well as the necessary elements are discovered which will contribute to achieving the purpose. Understand the challenge objectives and specifications from a corporate entity viewpoint further and convert a certain understanding it into definition of data mining difficulties and a preparatory diagram intended to reach the goals.
2) *Data Understanding:* In this phase, It might well gather the complete information and fill in the tool statistics (even when use

of a software). The records are mentioned well with source of their records, position, how they are obtained and if any problem has been experienced. To check its completeness the data is visualized and queried. The information is supplied in two distinctive documents in this segment. The file Demographic information is acquired through the records supplied by way of the candidates at a moment of savings to an application of card . It consists of purchaser related statistics like age, gender, salary,status of marraige etc. Credit Bureau data file consists of the records taken from credit bureau businesses which carries statistics associated to customer's savings history, delinquency details, balance details etc.

3) *Data Preparation:* This includes of choosing a suitable information, data cleansing, setting up elements through records, combining information through more than one database. Software analysis is the most critical segment of detection fashions, since its information comprises of inconsistencies, bugs, inconsistencies that you need to clean up in advance. These statistics collected through more than one resources initially it is combined and next cleanses at an entire information accrued is now not appropriate for modelling usage. A data consisting of special numeric values have no value, since they no longer make a significant contribution to anticipating modelling. Sectors including several missing attributes would like to be excluded too.
4) *Modelling:* In this Phase, determining of the data processing approach which includes decision-tree, produce check layout besides evaluating the data chosen, build models through the set of data as well as evaluate the designed framework of specialists to study the outcome. . Select and observe a range of modelling techniques, and calibrate device parameters to most efficient values. Typically, there are quite a few strategies for the identical data mining trouble kind.A few techniques on data shape have special necessities. Therefore, it is frequently important to move into the teaching statistics section once again.
5) *Evaluation:* One such stage must decide to what extent the ensuing model meets the required standards of the market sector. Assessment can be carried out with the help of testing the model on actual applications. Where the technique is checked for any repeatable errors or steps. The attributes are identified for process classification

B. Exploratory Data Analysis (EDA)

This method of interpreting and visualizing the data in order to obtain a deeper understanding and insight into the data. There are different steps involved in EDA, but the following are the basic steps a data analyst may take when conducting EDA:

1) Firstly process the data
2) Analysing categorical variables
3) Analysing numerical variables
4) Analysing numerical and categorical at the same time

Basic EDA key points conversion:

a) Data types.
b) Outliers.
c) Missing values.
d) Distributions (numerically and graphically) for both, numerical and categorical variables.

C. Weight of Evidence (WOE)

It explains a connection of an independent variable's predictive ability with respect to the dependent variable. Since it originated from the credit scoring industry, it is widely known as a measure of the distinction between good and bad clients. "Bad customers" shall refer to all consumers who defaulted on a loan. Thus, "Better Customers" applies to all consumers whoever paid their loans.

WOE calculation =In (Distribution of goods/Distribution of bads)

Distribution of Goods - % of Good Customers in a group.

Distribution of Bads - % of Bad Customers in a particular group.

ln - Natural Log.

Since they have individual experiences apart from credit risk, several people do not acknowledge the terms goods / bads. It is necessary to understand the concept of WOE in terms of events and not-events. It is determined by taking the normal logarithm of division of percentage of non-events (log to base e) and percentage of event.

WOE formula=In (% of non-events/%of events)

WOE- Calculation steps

1) Data is divided into parts (depending on the distribution) for a continuous variable
2) Number of events in each group and of non-events (bin) are to be calculated
3) The % of events and non-events in each group should also be calculated.
4) By taking natural log of the percentage of non-events and percentage of events divided calculate WOE.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

D. Datasets

We have been given two sets of data i.e.

1) *Demographic Data:* Demographic data is received from the facts supply with the aid of the candidates at the time of deposit card application. which consists of data associated to the consumer such as age , gender, salary , married state etc.
2) *Credit Bureau Data:* Credit Bureau data is the data taken from credit bureau agencies which contains information related to customers credit history, delinquency details, balance details etc. Information contains expected variable where the results tag that indicates that the user defaulted after obtaining a credit card.

E. Algorithm

The algorithm follows as stated below according to the steps.

Here we discuss pseudo code of Credit card Customer prediction, which is implemented by Crisp method using logistic regression and random forest.

1) *Step1:* Import of packages
2) *Step2:* Data cleaning and merging of data sets
3) *Step3:* Finding the continuous variables and categorical variables
4) *Step4:* Exploratory data Analysis
5) *Step5:* WOE and Information value
6) *Step6:* Dividing into train and test datasets
7) *Step7:* Model Evaluation
8) *Step8:* Plot graph

IV. PERFORMANCE ANALYSIS

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 for themodels. We checked which model is giving best predicted probability of default. The three main characteristics we considered in model evaluationare

A. Discriminatory power
B. Accuracy
C. Stability of the model

Final result we got for Logistic Regression are as

1) Accuracy: 0.61
2) Sensitivity: 0.63
3) Specificity: 0.61

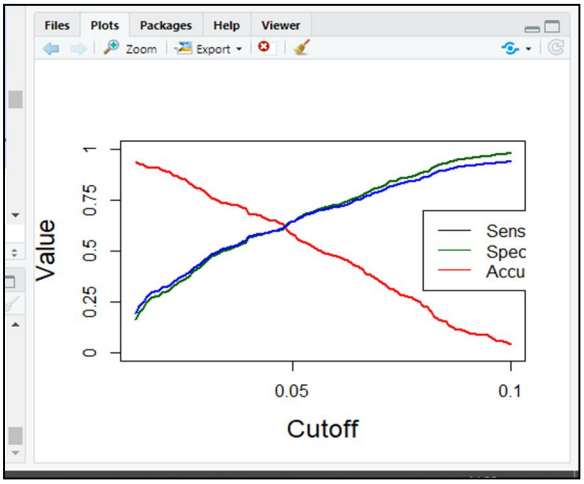


Fig 2: Observational Resultunder Logistic Regression

In Random Forest, we divided the master data into train and test datasets and build trees for the train dataset. In this the modelling is done in the form of trees. Here, we can observe the values from the graph and the final result which we got as:
Sensitivity =around 63%
Specificity=around 60%
Accuracy=around 60%

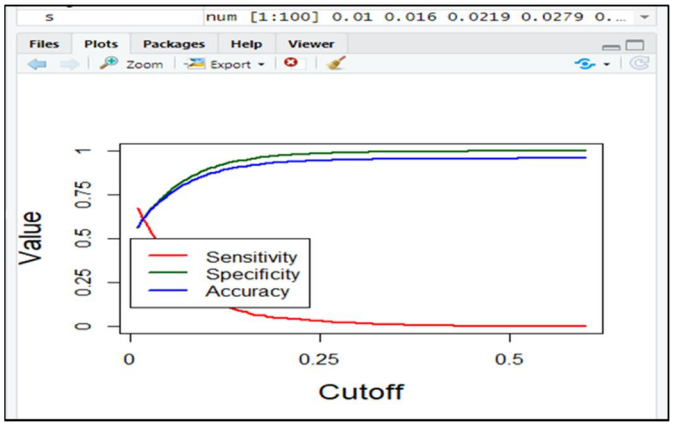


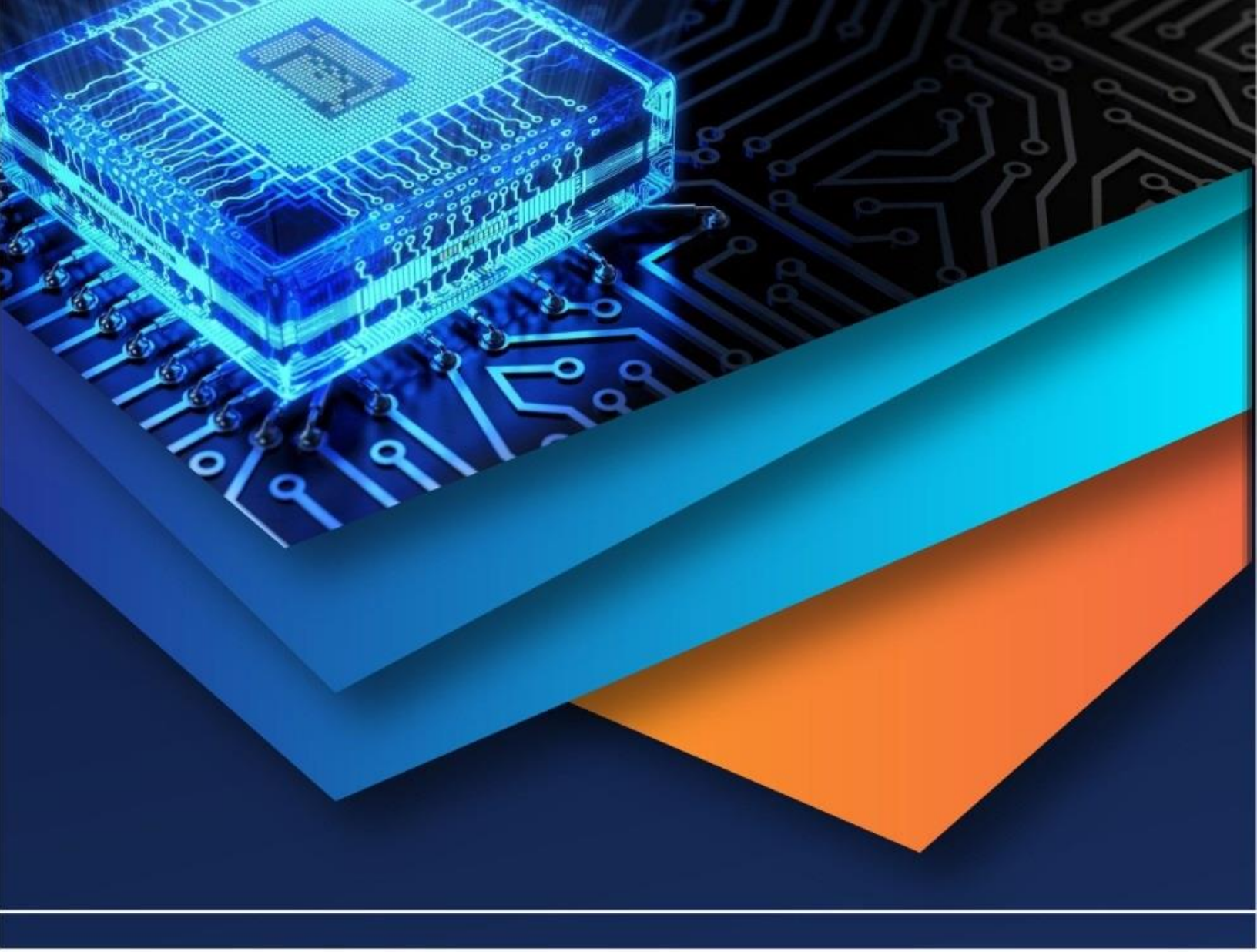
Fig 3: Observational Result under Random Forest

V. CONCLUSION AND FUTURESCOPE

Credit cards will be issued to the customers based on various factors such as demographic data, credit bureau data, and CIBIL score. Utilization of credit cards are becoming frequent in each and every industry in your day to day life, various credit card providers are getting thousands of credit card applicants every year. But in the past few years, they are experiencing an increase in credit loss. This paper specifies about the at most considering of reaching the highest point with an outstanding accuracy when the entire procedure is taken care we reach an accuracy rate of about 61%where the obtained result is considered at a certain point of the required data the later stage can include of bothering with the other specialized methods for the accurate accuracy.Finally we can conclude that No. of times 30.DPD or worse in last 12 months,Avgas.CC Utilization in last 12 months, No. of Inquiries in last 12 months excluding home auto loans, Outstanding. Balance are the significant factors that contribute for the credit card approval and the model is showing 61% Accuracy, Sensitivity 63% and Specificity61% in logistic regression analysis and in random forest it is showing Sensitivity around 63% Specificity around 60%, Accuracy around 60%.

REFERENCES

[1] Kou, Y., Lu, C.-T., Sirwongwattana, S., Huang, Y.-P.: Survey of fraud detection techniques. In: Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control, Taipei, Taiwan (2004)
[2] Chen, R., Chiu, M., Huang, Y., Chen, L.: Detecting credit card fraud by using questionnaire responded transaction model based on SVMs. In: Proceedings of IDEAL2004 (2004)
[3] Sahin, Y., Duman, E.: An overview of business domains where fraud can take place, and a survey of various fraud detection techniques. In: Proceedings of the 1st International Symposium on Computing in Science and Engineering, Aydin, Turkey (2010)
[4] Brause, R., Langsdorf, T., Hepp, M.: Neural data mining for credit card fraud detection. In: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (1999)
[5] Hanagandi, V., Dhar, A., Buescher, K.: Density-Based Clustering and Radial Basis Function Modeling to Generate Credit Card Fraud Scores. In: Proceedings of the IEEE/IAFE 1996 Conference (1996)
[6] ingKock Sheng, and Teh Ying Wah, "A comparative study of data mining techniques in predicting consumers' credit card risk in banks", African Journal of Business Management Vol. 5 (20), pp. 8307-8312, 16 September, 2011



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY