IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF CREDIT CARD DEFAULTERS USING MACHINE LEARNING

[1] Akshay A. Bardiya, [2] Dr. P. A. Tijare

[1] PG Scholar, [2] Professor

[01] Department of Computer Science and Engineering,

[1] Sipna College of Engineering and Technology, Amravati, Maharashtra, India

Abstract: Now a day's online transactions have emerged as a critical and vital part of our lives. It is important for the credit issuers companies or financial sectors to keep track on the users so that they can easily assign a credit card for them and need to maintain the records of the customers and on the basis of their previous history need to analysis whether that user is capable to pay their next month bill or not. If they don't track and don't keep watch on the customer's activities, then it may impact on their business, and they may lose much. For avoiding such situation, we are establishing a system by using different machine learning algorithm so that credit card assigner companies can easily capable to issue credit card to that customers which clears our previous records and have good records based on their previous transactions or activates. In this process, we have focused on analyzing and preprocessing data sets as well as the finding prediction accuracy from the dataset by using different machine learning algorithms such Naive Byes, Logistic Regression and so on.

Index Terms **– Credit Card, Machine Learning.**

I. INTRODUCTION

Credit Card Defaulters can be defined as a scenario where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used. Due to much use of E- Commerce, there has been a tremendous use of credit cards for online shopping, online businesses, online transaction and so on which led to High amount of frauds related to credit cards. In the sector of digitalization, there is a need to identify credit card frauds is necessary. Fraud detection or credit card defaulters involves monitoring and analyzing the behavior of various users in order to estimate detect or avoid undesirable behavior. In order to identify credit card defaulter's detection effectively, we need to understand the various technologies, algorithms and types involved in detecting credit card frauds. Algorithm can differentiate transactions which are defaulters or not. Find defaulter, they need to passed dataset and knowledge of fraudulent transaction. They analyze the dataset and classify all transactions. Credit Card Defaulter detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behavior, which consist of fraud, intrusion, and defaulting. Machine learning algorithms are employed to analyses all the authorized transactions and report the suspicious ones.

As increasingly more purchasers depend upon the credit score card to pay their ordinary purchases in online and bodily retail store, the quantity of issued credit score playing cards and the overpowering quantity of credit score card debt via way of means of the cardholders have swiftly increased. Therefore, maximum economic establishments need to address the troubles of credit score card default further to the credit score card fraud which include credit score card dump. Both the credit score card verification carried out to the cardholders and the default hazard control after card issued are critical to the destiny achievement of maximum economic establishments

II. LITERATURE REVIEW

Prediction of Credit card default requires the use of various machine learning techniques. Some of the work done by various researchers is summarized below:

Real-time Credit Card Fraud/scam Detection Using Machine Learning. [1] This paper centers around four principle fraud events in certifiable transactions. Every fraud is tended to strategy is chosen through an assessment. Significant key territory which we discourse in our venture is constant credit card scam identification.

Ref. [2] proposes model for providing the measures for loss probabilities as well as the evaluation of credit risk. The data used for this purpose consists of account level data from six different banks. Three different machine learning techniques including the decision trees, random forests and logistic regression are evaluated in the proposed model. A credit card amount that is not recovered for a period of more than 90 days is considered as non-recoverable or a default.

E-commerce industry is growing rapidly and this leads to the increased usage of credit card payments for online purchases. In this paper investigation of the performance of logistic regression, random forest and decision tree for credit card fraud detection is carried out. The dataset for credit card fraud detection is gathered from kaggle and this dataset consists of over 2, 84,808 credit card transaction data of a European bank. Fraud transactions are considered as positive class and the genuine transactions are considered as negative class. Dataset consists of imbalanced 0.172% of fraud transitions and the reaming transactions are genuine. Performance is evaluated based on accuracy, sensitivity, error rate and specificity. [3]

[4] In this paper the author has used a case sensitive method which is based on Bays maximum risk and then it is presented using proposed cost measure. The dataset is based on the real life transaction data obtained from a European company and maintaining the confidentiality of the personal data. The accuracy of the algorithm used is 50%. The main significance of this paper is to reduce the cost. The result obtained was 23%.

Credit Card scam identification- Machine Learning methods [5] Credit Card Scam identification database was utilized in an analysis. Since the database was profoundly non balanced, destroyed strategy was utilized in over sampling. Later on, highlight determination was done and database was part in two sections, preparing data and testing data. The techniques utilized for the investigation were Logistic Regression, Random Forest, and Naive Bayes with Multilayer Perception.

Credit Card Fraud/scam Detection using Deep Learning [6] this paper is tied in with developing a credit card scam identification framework utilizing Deep Learning Neural Networks. Regardless of whether or not the Neural Network is prepared above an extensive variety of emphases, that isn't always effectively accurate to categorize the data as fraud or valid because of skewness of the database. We make use of two sampling systems: Under-Sampling, from lessening number of legitimate perceptions and Over-Sampling, in which the fraud class perception is copied. Detection of Credit Card Fraud/scam Transactions Using Machine Learning Algorithms and Neural Networks [7] Credit card fraud coming about because of abuse for the framework is characterized like burglary or abuse of someone's credit card data that is utilized for individual increases unescorted by the consent of the owner of card. For identifying these scams, this is essential for checking the use examples for a client by the previous transaction. Contrasting the utilization example and present day transaction, we could categorize this like one or the other scam or a real transaction. In this research, the procedures utilized are KNN, Naïve Bayes, Logistic Regression, Chebyshev Functional Link Artificial Neural Network (CFLANN), Multi-Layer Perceptron and Decision Trees.

This paper checks and investigates the performance of Random Forest, SVM, logistic regression and Decision tree on a highly skewed credit card fraud data. The dataset was gathered by a European cardholders consisting of about 2, 84,786 transactions. The result obtained was 97.7% accuracy by Logistic regression, 97.5% by SVM and 98.6% precise accuracy obtained by Random Forest. [8]

In this paper one of the best data mining algorithm called machine learning algorithm was introduced, which was used to recognize the credit card fraud. A half bread grouping framework with exception recognition was utilized in order to differentiate between misrepresentations of internet recreations. The framework obtained online calculations with factual data in order to distinguish various extraction types. This framework attained extreme location rate at 98% along with 0.1% fault rate. [9]

This paper discusses about supervisor based classification using Bayesian network classifiers such as Naïve Bayes, K2, Tree Augmented Naïve Bayes (TAN, logistics and J48 classifiers. The datasets are pre-processed by using normalization and principal component analysis. Two datasets were used dummy dataset which represented the characteristics of the credit card data and newly generated dataset using data normalization and principal component analysis technique. All these classifiers achieved over 95% accuracy. [10]

III. PROBLEM STATEMENT

Credit card default are increasing heavily because of fraud financial loss increasing drastically. Every year due to fraud, Billions of amounts loss. To predict the default in credit card, there is lack of research. Taking this point in mind we are developing a system which can predict and find the defaulter person, so that financial sector may secure from this fraud.

IV. OBJECTIVES

The objective of our paper is to offer a system or tool to detect 100% of the fraudulent transactions while minimizing the fraud or defaulter's classifications. In this process we filter the records which we got from the financial sector or banking data and on the filter data we train the data and apply to test phase. Here we are using multiple algorithms to find the best out of best accuracy result from them and whichever gives us a best accuracy prediction result we will use them in our system. Our Objective is to determine whether a person defaults the credit card payment for the next month.

V. PROPOSED SYSTEM

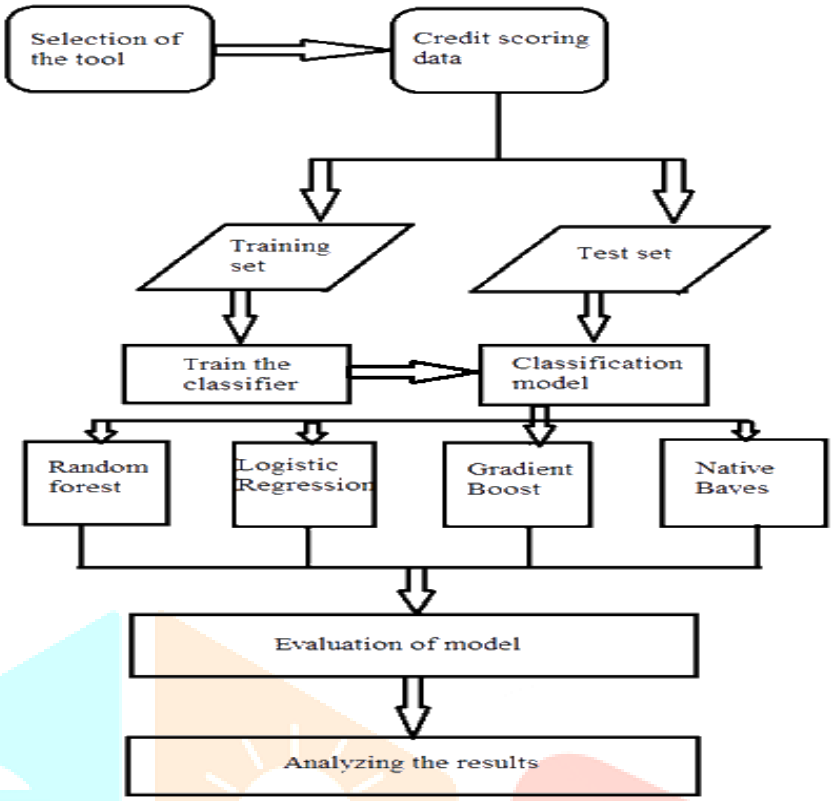

Fig 1: System Flow Diagram

In propose system, we propose four different algorithms to find the best accuracy in predicting the default candidate for the next month. The credit card data set is divided into two parts- the training set as well as the test set. The classification model is trained using the training set and the remaining observations passes to next level to perform the prediction task using different techniques.

A brief introduction of these techniques is as follows:

Random Forest: Random forest is a type of supervised learning algorithm getting to know set of rules this is used broadly in Classification and Regression problems. It builds decision trees on distinct samples and takes their majority vote for classification and average in case of regression. In Random Forest model, Random means each tree is only trained on a random subset of samples drawn from the training set (with repetition) and possibly a random subset of features and Forest because there are several trees.

Logistic Regression: Logistic regression model attempts to pick out the correlation between the dependent and the independent variables. It is basically used for binary classification where the target variable is binary and one or more independent variables can be continuous or binary. This model uses the logistic function to identify or to track the probability of the output with respect to the input. The classification is applied such that a threshold is provided, and all the probability values greater than a certain threshold are assigned one class and the values less than the threshold are assigned the other class.

Gradient Boost: The main purpose of Gradient Boost model is to help weak prediction models becomes stronger. It works by building one tree at a time, and correct errors made by previously tree. It can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier).

Naive Bayes: Naive Bayes classifier is one of the supervised learning algorithms that is primarily based totally on Bayes theorem and makes use of the probabilistic features. It assumes the independence among all the features for a particular model. Naive Bayes classifier is a simplest method to integrate as it assumes conditional independence. The probabilities of conditional method are identified for the attributes and classification is performed such that the class with the most probable hypothesis or maximum a posteriori (MAP) is assigned to the given element.

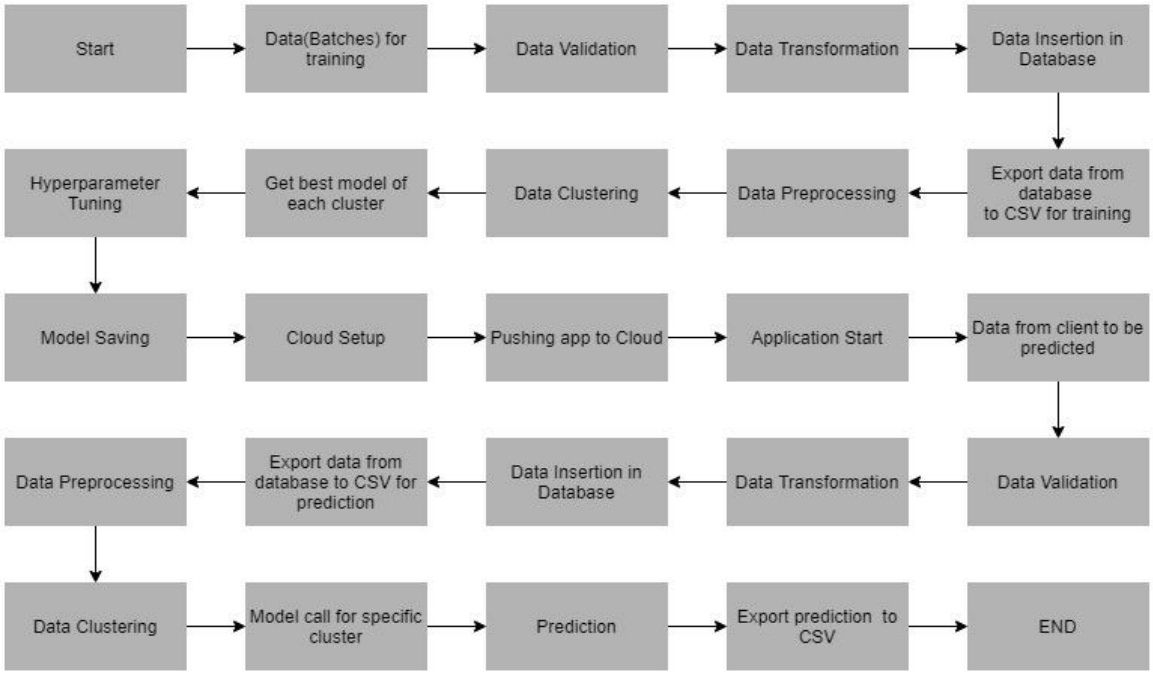

Fig 2: Architecture Flow Diagram

VI. DATA REQUIREMENT AND IMPLEMENTATION

There are 24 Features we are using in this system:

1. LIMIT_BAL: It provides limit of the credit card for a person. It provides continuous values
2. GENDER: It is divided into 2 categories: 1 = male; 2 = female
3. EDUCATION: It contains categorical data: 1 = graduate school; 2 = university; 3 = high; 4 = others
4. MARRIAGE: Categorical: 1 = married; 2 = single; 3 = others
5. AGE: Continuous values i.e. it provides numerical data
6. PAY_0 to PAY_6: History of past payment. We tracked the past monthly payment records
7. BILL_AMT1 to BILL_AMT6: Amount of bill statements.
8. PAY_AMT1 to PAY_AMT6: Amount of previous payments.
9. Default payment next month: yes = 1; no = 0

We are using 30,000 records from the kaggle.
Below figure shows the relationship between two variables; it is generated using heatmap function of machine learning Heatmap is also used in finding the correlation between different sets of attributes.

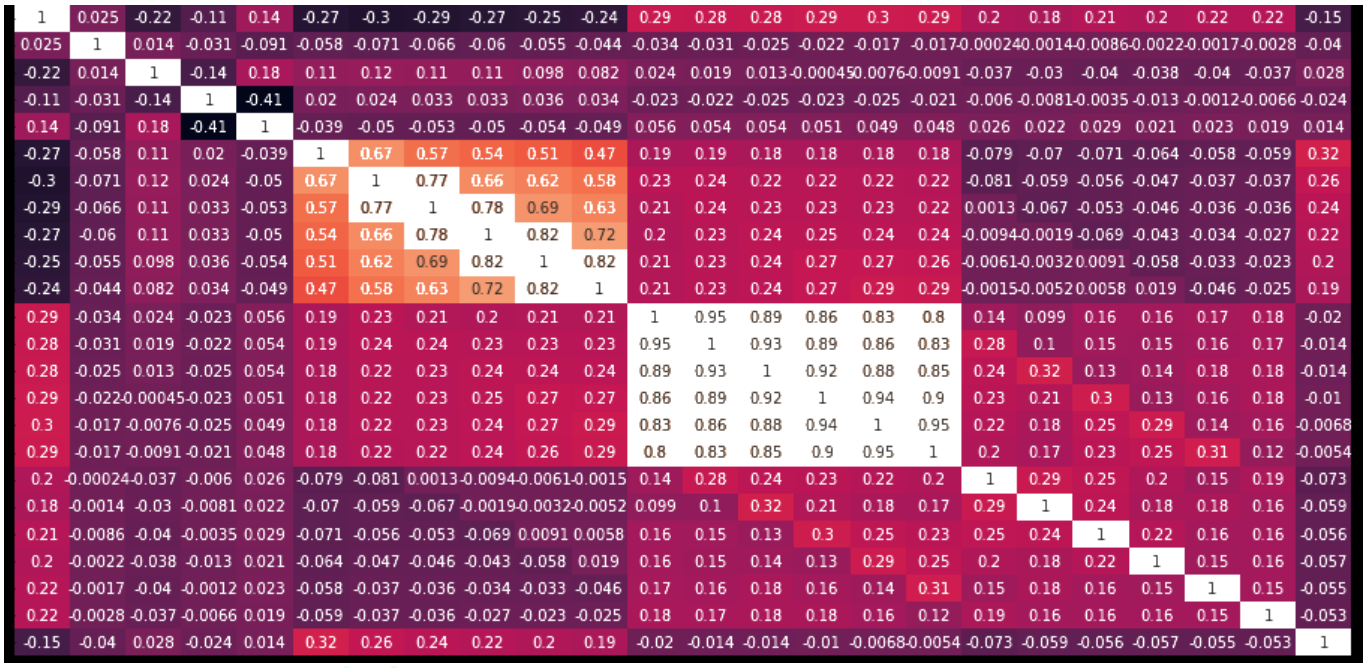

Fig 3: Showing Relationship between two variables

Data Validation:

In this step we perform different sets of validation on the given set of training files.

1. **Name Validation:** We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name given in the schema file to use for validation. After validating the pattern in the name, we check for the length of date in the file name as well as the length of time in the file name. If all the values are as per requirement, we move such files to "Good_Data_Folder"(Working data) else we move such files to "Bad_Data_Folder."(Raw/actual data).

2. **Number of Columns:** We validate the number of columns present in the file, and if it doesn't match with the value given in the schema file, then the file is moved to "Bad_Data_Folder.".

3. **Name of Columns:** The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad_Data_Folder".

4. **The datatype of columns:** The datatype of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad_Data_Folder.

After validating the data, we have train and test the data on x-axis and y-axis using following commands:

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=42)
```

```
from sklearn.preprocessing import StandardScaler
train_scaler=StandardScaler()
test_scaler=StandardScaler()
```

Once the data is train and test we have to apply different algorithms to find the accuracy of the prediction for that we use:

The accuracy result when applying different algorithms are:

1. **Naive Byes:** Naive Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems. It predicts on the basis of the probability of an object. Examples: Spam Filtration, Sentimental analysis and classifying articles.

Bayes theorem:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Where, P (A | B) is probability of hypothesis

P (B | A) is Probability of the evidence given that the probability of a hypothesis is true.

P (A) is the probability of hypothesis before observing the evidence.

P (B) is the probability of evidence.

```
from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
```

```
pred_y=gnb.fit(scaled_train_df,y_train).predict(scaled_test_df)
```

```
pred_y
```

```
array([1, 0, 0, ..., 0, 0, 0], dtype=int64)
```

```
from sklearn.metrics import accuracy_score
```

```
ac=accuracy_score(y_test, pred_y)
```

```
ac
```

Here we find out the accuracy of the naive byes algorithm is 66 percentage.

2. **Logistic Regression:** Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
   In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

```
# import the class
from sklearn.linear_model import LogisticRegression

# instantiate the model (using the default parameters)
logreg = LogisticRegression()

# fit the model with data
logreg.fit(X_train,y_train)

#
y_pred=logreg.predict(X_test)
```

```
# import the metrics class
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
```

```
array([[4694,    2],
       [1304,    0]], dtype=int64)
```

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.7823333333333333
Precision: 0.0
Recall: 0.0
```

Here we find out the accuracy of the logistic regression algorithm is 78 percentage.

VII. CONCLUSION

Credit card fraud cases are increasing day by day and it is one of the major concerns in financial service sectors. This occurs when no proper security measures are taken into consideration are. In this paper an attempt is made to identify the number of fraudulent transactions in a particular dataset by using various machine learning algorithms such as local outlier factor and isolation forest method. Only a part of dataset was used in order to speed up the computational process.

VIII. ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my **Dr. P. A. Tijare** who has in the literal sense, guided and supervised me. I am indebted with a deep sense of gratitude for the constant inspiration and valuable guidance throughout the work.

IX. REFERENCES

[1] Anuruddha Thennakoon; Chee Bhagyani; Sasitha Premadasa; Shalitha Mihiranga; Nuwan Kuruwitaarachchi 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence);10-11 Jan. 2019.

[2] Zhou H, Lan Y, Soh Y, Huang G and Zhang R (2012), "Credit risk evaluation with extreme learning machine", IEEE International Conference on Systems, Man, and Cybernetic(SMC), Seoul, 2012, pp. 1064-1069.

[3] M. BM and H. Mohapatra, “Human centric software engineering,” International Journal of Innovations & Advancement in Computer Science (IJIACS), vol. 4, no. 7, pp. 86-95, 2015.

[4] H. Mohapatra, C Programming: Practice, Vols. ISBN: 1726820874, 9781726820875, Kindle, 2018.

[5] Dejan Varmedja; Mirjana Karanovic; Srdjan Sladojevic; Marko Arsenovic; Andras Anderla; 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) 20-22 March 2019

[6] Pranali Shenvi; Neel Samant; Shubham Kumar; Vaishali Kulkarni; 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) 29-31 March 2019

[7] Deepti Dighe; Sneha Patil; Shrikant Kokate; 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)16-18 Aug. 2018

[8] H. Mohapatra and A. Rath, Advancing generation Z employability through new forms of learning: quality assurance and recognition of alternative credentials, ResearchGate, 2020.

[9] H. Mohapatra and A. Rath, Fundamentals of software engineering: Designed to provide an insight into the software engineering concepts, BPB, 2020.

[10] V. Ande and H. Mohapatra, “SSO mechanism in distributed environment,” International Journal of Innovations & Advancement in Computer Science, vol. 4, no. 6, pp. 133-136, 2015.