FINANCIAL FRAUD DETECTION USING LOGISTIC ALGORITHM

Kushal TM[1], Ravindra V[2], Chaithanya R[3], Rahmathulla Khan H[4], Prof. Vidyasagar[5]

[1]*Electronics and Communication Engineering, Reva University, India*

ABSTRACT

This paper presents the Gaussian Naïve Bayes for Fraud Detection of the credit card. The aim is to boost the accuracy and enhance the flexibleness of the algorithm. The first target of this project is to perform an investigation on the credit cards fraud detection dataset utilizing ML procedures and distinguish the deceitful exchanges from the given dataset. Diverse examining methods are executed to handle the category irregularity exchange issue and arrangement of ML calculations like Logistic Regression and Gaussian naïve Bayes are going to be actualized on the dataset, and the outcomes are going to be accounted for with estimating the data credited utilizing proposed calculations with more exactness and adequacy. During this paper, fraud classification using ML algorithms is proposed. This technique uses logistic regression to make the classifier stop frauds in credit card transactions. To handle unwanted data and to make sure a high degree of detection accuracy, a pre-processing step is used.

Keywords: Machine Learning, Confusion matrix, Data analysis, Gaussian Naïve Bayes, Logistic Regression, Decision Tree.

I. INTRODUCTION

Machine learning is that the technology which is nothing but the applying of AI i.e., artificial intelligence gives computer systems the potential to find out and improve from previous cases and past experiences. Machine learning focuses on developing codes or algorithms or computer programs that can access and analyzing given data and so later use that analyzed data to be told on their own. The procedure of learning begins with some observations or previous data, like examples, past experiences, or instructions, to seek out and find out hidden patterns within the dataset and take better decisive actions within the future supported the examples and data that we offer. The last word objective of machine learning is to provide computers the ability to earn automatically i.e., learn by themselves with no unnecessary human assistance so scrutinize the data and take decisions accordingly.

Fraudulent transactions are a relatively rare occurrence among all transactions. We must always also be sure to not classify an outsized number of genuine transactions as a fraudulent transaction, because it would waste lots of resources to research transactions flagged as fraudulent transactions, i.e., the number of false-positive must be less.

First, select the best features among a sizable number of features. We cannot build a decent model. When there is a sizable number of features. We should simplest model possible. The second step is to separate the dataset into training data and testing data. Because after we train the model. We must always ensure testing data does not affect the training process. This allows us to gauge the model on unseen data. Here we decide Gaussian Naïve Bayes model to classify transactions as genuine or fraudulent.

We train the model using data set with a process called Cross-Validation, it is used to improve the performance of the model over train and test split of a dataset. Once a model is trained, we used the trained model on a testing dataset to evaluate the performance of the model.

II. RELATED WORK

Some Algorithm methods have already been proposed and tested. Fraudulent activities are causing major loss, which motivated researchers to seek out an answer that may detect and forestall frauds.

Standard algorithms which incorporate Logistic Regression, Decision Tree. Naïve Bayes and a mixture of certain classifiers were used, which led to the high recall of over 92% on a dataset, European dataset was also used, and a comparison was made between the models supported LR, GNB, and DT. Among the three models, DT proved to be the most effective, with an accuracy of 99.92%, followed by GNB with 98.35%, and LR with an accuracy of 96.45%.

III. MATERIALS AND METHODS

1.1 Dataset

In this paper the credit card Fraud Detection dataset was used, this dataset contains transactions, occurred in two days, made in September 2013 by European cardholders. The dataset contains 31 numerical features. Since several the input variables contain financial information, the Feature "Amount" is that the amount of the transactions made by credit card. Feature "Class" represents the label and takes only 2 values: value 1 just in case of fraud transaction and 0 otherwise. Feature "Time" shows the time between the first transaction and therefore every other transaction within the dataset. The dataset used for the analysis of credit card detection during this paper contains data from European credit cardholders consisting of rows of transactions made by credit cards. The entire number of transactions captured was 500,000 and therefore the number of features captured was 320. Data pre-processing was done to drop the missing values. Principal component analysis was done to work out the foremost relevant features. The results of data pre-processing yielded 284,807 records and 31 most prominent features were chosen. The features included 28 masked features which are intentionally masked by the data source, and ‘time’ and ‘amount’ of the transaction.

1.2 Preprocessing

1.2.1 Data Analysis

The dataset which has been selected and used holds the records of European cardholders who made transactions using their credit cards within September 2013. This dataset holds the record of transactions that were made within two days and total transactions made within two days are 284,807 transactions from which 492 transactions were found as fraudulent which makes the dataset highly imbalanced, more oriented because the positive class i.e., fraud transactions are 0.172% out of total transactions. And therefore, the dataset is in CSV format i.e., in an exceedingly format where the data values are separated by commas.

1.2.2 Data Cleaning

Data cleaning is additionally called and referred to as Data cleansing because during this process the incorrect and corrupted records from the dataset or a record-set or a table are identified and corrected i.e., removed and this process focuses on the identification of incorrect, irrelevant, inaccurate, or incomplete parts of the data then modification of that specific part by replacing it with some different value or completely deleting the dirty data .

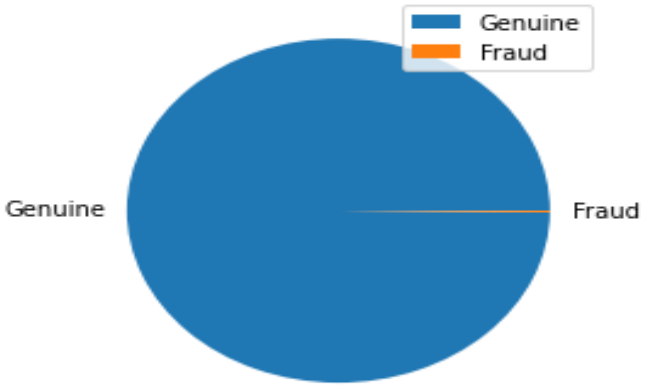


Figure 1: Classification of transactions.

The experiment system environment is Windows 10 OS, and therefore the software is Spyder Anaconda, a scientific python development environment, which is an element of the Anaconda platform. Used libraries include Numpy, Pandas, Matplotlib, Sklearn, and Seaborn.

1.3 Experiment

Logistic regression is one of the foremost popular classification algorithms in machine learning. Logistic regression uses Sigmoid Function. The logistic regression model describes the relationship between predictors that may be continuous, binary, and categorical. Variable quantity is binary. Supported some predictors we predict whether something will happen or not. We estimate the probability of belonging to every category for a given set of predictors.

A Decision Tree is an algorithm that may be employed in both classification and regression problems. It consists of many decision trees. This algorithm gives better results when there is a higher number of a tree within the forest and preventing the model from over fitting. Each decision tree in the forest gives some results. These results are merged to urge more accurate and stable prediction.

1.4 Implementing Naïve Bayes

The Naïve Bayes machine learning classifier tries to predict a category that is thought of as outcome class supported probabilities, and conditional probabilities of its occurrence from the training data. This type of learning is extremely efficient, fast, and high in accuracy for real-world scenarios, and this learning type is understood as supervised learning.

The initial step for the Naïve Bayes classification algorithm is the Bayes theorem for conditional probability,

P (H/E) = (P (H)*P (E/H)) / P (E)

Where

P (H) = probability a hypothesis is true (before any event).

P (E/H) = probability of seeing the event if the hypothesis is true.

P (E) = probability of seeing the event.

P (H/E) = probability a hypothesis is true given some event.

1.5 Confusion Matrix

The confusion matrix summarizes the performance of the algorithm. The idea of what's algorithm doing correct and what is doing incorrectly are often understood from it. Confusion matrix rows represent predicted class, while rows represent an actual class.

Types of Confusion Matrix namely:

1.5.1 True Positive (TP)

• The actual value matches by predicted value.

• The actual value was positive and therefore the model predicted a positive value.

1.5.2 True Negative (TN)

• The actual value matches by predicted value.

• The actual value was negative and therefore the model predicted a negative value.

1.5.3. False Positive (FP)

• The predicted value was falsely predicted

• The actual value was negative, but the model predicted a positive value

1.5.4. False Negative (FN)

• The predicted value was falsely predicted

• The actual value was positive, but the model predicted a negative value

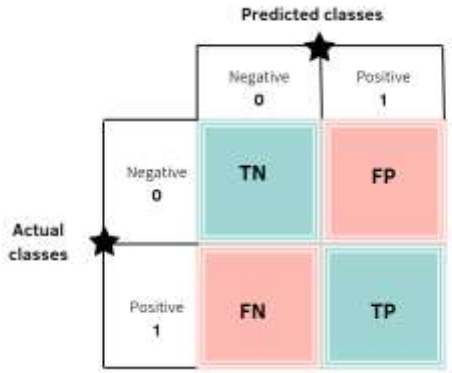


Figure 2: Confusion Matrix

1.6 Precision, Recall and F1_Score

Precision: the ratio of the number of true positives divided by the number of true positives plus the number of false positives.

Precision = True Positive / (True Positive + False Positive).

Recall the ratio of the number of true positives divided by the number of true positives plus the number of false negatives.

Recall = True Positive / (True Positive + False Negative).

F1_SCORE: F1_score is defined as an average of Precision and Recall. Therefore, this score takes both false positives and false negatives into consideration. Basically, it is not as easy to know as accuracy, but F1_score is typically more useful than accuracy. Accuracy works best if false positives and false negatives have similar costs. If the value of false positives and false negatives are very different, it is better to seem Precision and Recall. During this paper, F1_score is 0.9663.

F1_Score = (2*(Precision * Recall)) / (Precision + Recall).

IV. FEATURE SELECTION USING SELECTKBEST LIBRARY

Feature selection could be a technique where we elect those features in our data that contribute most to the target variable. In other words, we decide the most effective predictors for the target variable.

In this paper, selecting the features using SelectKBest module Sklearn.feature_selection library, classes within the Sklearn.feature_selection module are often used for feature selection/dimensionality reduction on sample

sets, either to enhance estimator's accuracy scores or to boost their performance on very high-dimensional datasets.

1.1 Advantages of SelectKBest Module is

- Reduces Over fitting: Less redundant data means less possibility of creating decisions supported redundant data/noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: fewer data implies that algorithms train faster.

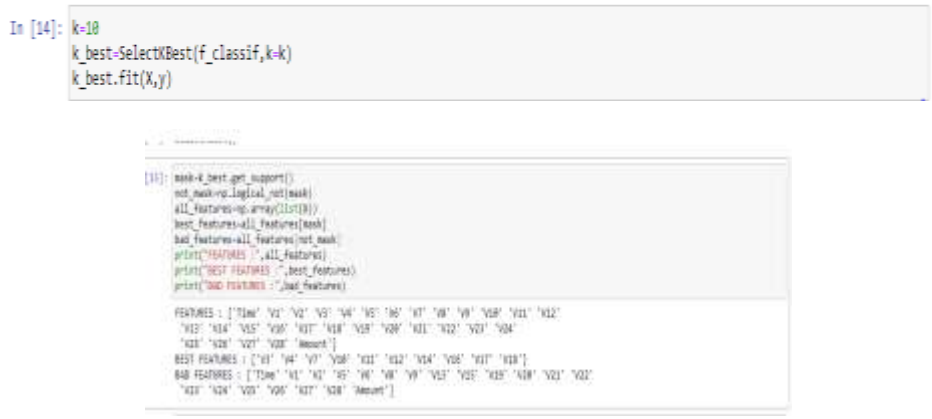


Figure 3: Selecting Features Using SelectKbest Library

1.2 Features from Dataset are:

Features: [‘time’ ‘v1’ ‘v2’ ‘v3’ ‘v4’ ‘v5’ ‘v6’ ‘v7’ ‘v8’ ‘v9’ ‘v10’ ‘v11’ ‘v12'

'v13’ ‘v14’ ‘v15’ ‘v16’ ‘v17’ ‘v18’ ‘v19’ ‘v20’ ‘v21’ ‘v22’ ‘v23’ ‘v24'

'v25’ ‘v26’ ‘v27’ ‘v28’ ‘amount’]

1.3 Best Features from Dataset are:

Best features: ['v3’ ‘v4’ ‘v7’ ‘v10’ ‘v11’ ‘v12’ ‘v14’ ‘v16’ ‘v17’ ‘v18'], if both fraud and genuine transactions have different shapes, we can easily distinguish whether it is fraud or genuine transaction.

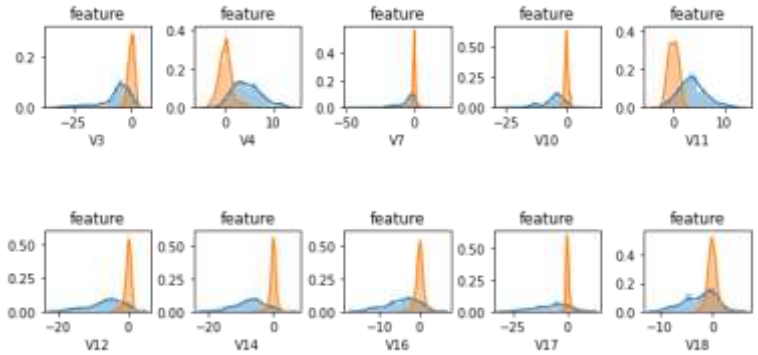


Figure 4: Classification of best features from dataset.

1.4 Bad features from the Dataset are:

Bad features: ['time' 'v1' 'v2' 'v5' 'v6' 'v8' 'v9' 'v13' 'v15' 'v19' 'v20' 'v21' 'v22' 'v23' 'v24' 'v25' 'v26' ' v27' 'v28' 'amount']

If both fraud and genuine transactions lie on top of every other, we cannot be able to distinguish whether it is fraud or genuine transaction.

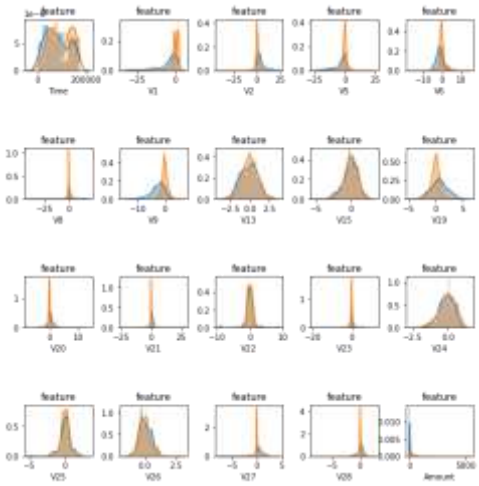


Figure 5: classification of bad features from dataset.

V. RESULTS AND DISCUSSIONS

The most used metrics for determining the results of machine learning algorithms are accuracy, recall, and precision. All the mentioned metrics will be calculated from a Confusion matrix. Since the test set consists of 20% of the entire dataset, the total sum of samples is 56962.

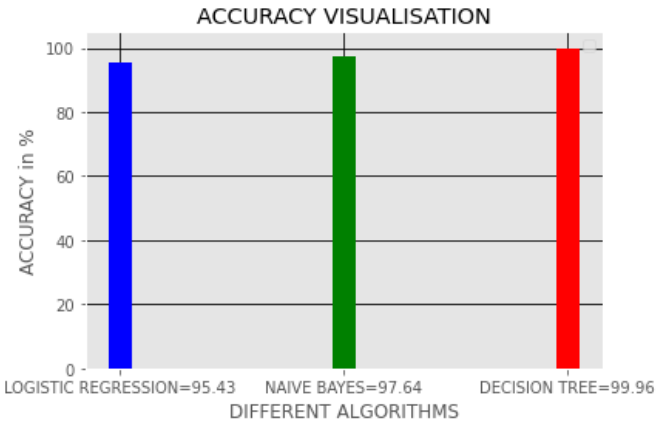
From 56962 samples, 56375 samples classify as True positive,508 sample classifies as True negative,13 samples classify as False positive,60 samples classify as False-negative.

1.1 ALGORITHM ANALYSIS TABLE

#	ALGORITHM	ACCURACY	F1 SCORE
1	LOGISTIC REGRESSION	95.43%	96.63
2	NAIVE BAYES	97.63%	10.83
3	DECISION TREE	99.96%	94

Figure 9: Algorithm analysis table.

1.2 ACCURACY VISUALIZATION FOR GIVEN ALGORITHMS



1.3 F. F1 SCORE VISUALIZATION FOR GIVEN ALGORITHMS

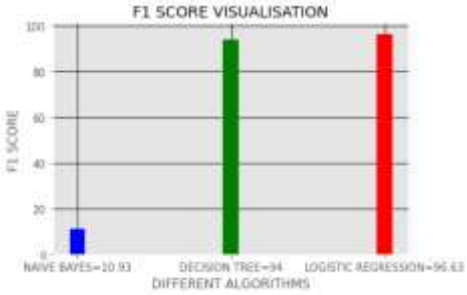


Figure 11: F1-Score visualization.

F1-score visualization for logistic regression, Gaussian naïve Bayes, and decision tree algorithm. From table 9, the Logistic Regression algorithm gives the most effective results because it is the highest f1-score.

VI. CONCLUSION

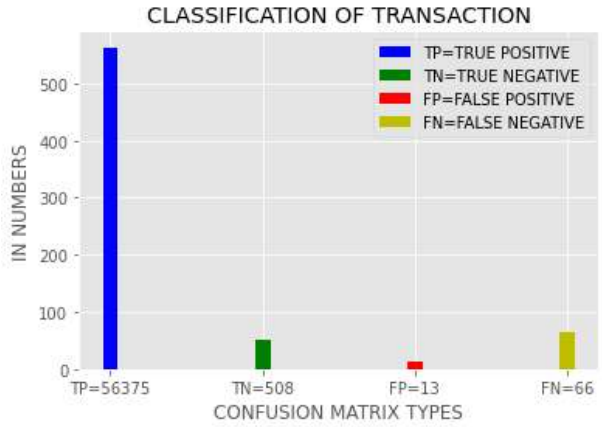


Figure 12: transactions classified as genuine or fraud.

The main goal of this paper was to match certain machine learning algorithms for the detection of fraudulent transactions. The logistic Regression algorithm gives the simplest results because its highest f1-score i.e., best classifies whether transactions are fraud or not. This was established using different metrics, like recall, accuracy, f1-score, and precision. For this type of problem, it is important to possess recall with high value.

Feature selection and balancing of the dataset have shown to be extremely important in achieving significant results.

VII. REFERENCES

[1] AS. Akila and U. Srinivasulu Reddy, "Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection," Journal of Computational Science, vol. 27, pp. 247–254, Jul. 2018, DOI: 10.1016/j.jocs.2018.06.009.

[2] Y. Lucas et al., "Towards automated feature engineering for credit card fraud detection using multiperspective HMMs," Future Generation Computer Systems, vol. 102, pp. 393–402, Jan. 2020, DOI: 10.1016/j.future.2019.08.029.

[3] Chan, P.K., Fan, W., Prodromidis, A.L., & Stolfo, S.J. (1999) Distributed data mining in credit card fraud detection, IEEE Intelligent Systems, and their Applications. Vol. 14, No. 16, pp. 67 – 74.

[4] J. O. Awoyemi, A. O. Adentumbi, S. A. Oluwadare, "Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis", Computing Networking and Informatics (ICCNI), 2017 International Conference on pp. 1-9. IEEE.

[5] Z. Kazemi, H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions", Knowledge-Based Engineering and Innovation (KBEI), 2017 IEEE 4th International Conference on pp. 630-633. IEEE.

[6] S. Dhankhad, B. Far, E. A. Mohammed, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study", 2018 IEEE International Conference on Information Reuse and Integration (IRI) pp. 122-125. IEEE.

[7] ANALYSIS OF HANDWRITING CHARACTERISTICS IN RELATION TO CHILD ABUSE, Vishav Jyoti Bajaj, Dr. Renu Verma, International Journal Of Advance Research In Science And Engineering http://www.ijarse.com IJARSE, Volume No. 10, Issue No. 01, January 2021 ISSN-2319-8354(E).

[8] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai, S. Pan, "Credit card fraud detection based on whale algorithm optimized BP neural network", 2018 13th International Conference on Computer Science & Education (ICCSE) pp. 1-4. IEEE.

[9] F. Ogwueleka, "Data mining application in credit card fraud detection system," Journal of Engineering Science and Technology, Vol. 6, No. 3, 2011.

[10] R. Patidar and L. Sharma, "Credit Card Fraud Detection Using Neural Network," International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-1, IssueNCAI2011, and June 2011.

[11] J. Dara and L. Gundemoni, "Credit Card Security and E-Payment," 2006 [15] P. Chan, W. Fan, Prodromidis and Salvatore, "Distributed Data Mining in Credit Card Fraud Detection," IEEE December 1999.