A Fraudulent Transaction Prediction in Credit Card by Using Novel LGBA over LR Algorithms

Raghavendra Reddy P and Sivanesh Kumar A

Abstract: The Research is to predict the accuracy of credit card fraudulent transactions using the Light Gradient Booster algorithm. Novel Light Gradient Booster with sample size = 10 then Logistic Regression of sample size = 10 were executed for estimating accuracy rate of credit card fraudulent transactions. The sigmoid functions used in Light Gradient maps the values between 0 and 1. Light Gradient Booster Algorithm has a Maximum accuracy (91.6%) while comparing to the performance of Logistic Regression (81.4%). There was a statistical significance difference between Light Gradient Booster and Logistic Regression with p = 0.0001 (p<0.05) based on 2-tailed analysis. Light Gradient algorithm helps in predicting with better accuracy percentage of credit card fraudulent transaction than Logistic regression.

Keywords: Machine Learning, Novel Light Gradient Booster, Logistic Regression, Fraudulent Transactions, Fraudulent Dataset.

1. Introduction

The objective of the work is to predict the accuracy rate of credit card fraudulent transactions using the Novel Light Gradient Booster algorithm. Detecting credit card fraud is an inclusive term for various scenarios especially using a payment gateway [1]. The purpose of using a credit card is to acquire goods and services, or to make payment which is controlled by the fraudster's environment [2]. It is actually used to prevent criminal actions effectively, as a result, account information is leaked. namely Skimming and counterfoil [3]. Due to the risk of losing their reputation and the client loyalty, credit card issuers should consider implementing advanced credit card fraud protection and detection systems, depending on the scenario [4]. It is purposely used in banking sectors especially credit card transaction division and loan transaction division for identifying customer financial transactions. Based on information about each cardholder's behaviour, machine learning-based solutions can constantly enhance the accuracy of fraud protection [5]. Online Credit Card Fraud detection plays a vital role in many industries including banking, financial sectors and government agencies [11].
In this research work, credit cards formulated with datasets especially non-anonymized features with sequences of classified data [12]. The most important scenarios for predicting credit card fraudulent transactions are formulated with various factors to enhance the accuracy percentage of predicting fraudulent credit card transactions using machine learning Novel Light gradient Booster algorithm compared with Logistic Regression [13].

2. Materials and Methods

The research work was done at Image Processing Laboratory, Department of Computer Science and Engineering, Saveetha School of Engineering, SIMATS. A total of two groups were used in this study. Group one is the Light Gradient Booster method and Group two is Logistic Regression. Light Gradient and Logistic Regression was iterated with a sample size of 10, with 95% confidence interval, $\alpha$-0.05 and G- power of 80%. Real time dataset used was collected from kaggle.com. The dataset consists of 31 attributes and 2848 rows [7]. After collection of the dataset and data about data was uploaded and preprocessed with respect to data scenario. All null values and missing values have been classified into two segments namely cleaning and cleansing the unstructured data. Class is a responsive variable and the values assigned for class with viewpoints as 0 and 1. Attained data sets short of worthless values and missing values are well skillled for appraising the ML algorithm [8].

2.1. Light Gradient Booster Algorithm

Light Gradient Booster algorithm is a ML that is implemented with respected to Light gradient booster algorithm. Basically, the light Gradient Booster method splits tree leaf-wise as incompatible to other methods that increase tree level-wise based on the sequential approach from the user. It selects a leaf with determined delta loss to grow for random check. It demonstrates a gradient boosting framework based on decision trees to improve model efficiency and memory utilization [9]. Gradient boosting is one of the machine learning techniques for classification and regression problems that generates a prediction model from an ensemble of weak prediction models, usually decision trees. Light Gradient Booster algorithm as shown in equation (1).

\[
\text{Loss} = \text{MSE} = \sum(y_i - y_{ip})^2 \quad (1)
\]

where, $y_i = i$ th target value
$y_{ip} = i$ th prediction
$L(y_i - y_{ip}) = \text{Loss function}$

2.2. Logistic Regression Algorithm

Logistic Regression is classified as a controlled classification regression algorithm. The link among dependent and independent variables is demonstrated using logistic regression. Regression solves the binary classification problem [10]. In this proposed work credit card fraudulent transactions classify as 0 or 1 and fraud or Non fraud.
Over-all calculation for Logistic Regression as shown in the equation (2). The Logistic Regression algorithm is exposed in Figure1.

\[ Y = \frac{e^{(b_0 + b_1 * X)}}{1 + e^{(b_0 + b_1 * X)}} \]  

Wherever, \( Y \) is predicted output

\( b_0 \) is Bias term

\( b_1 \) is coefficient for input value

Novel Sigmoid function used in Logistic Regression will predict the probability range within the limit of 0 and 1. To predict the values options to the sigmoid function was adopted as shown in equation (3). This is implemented to map the original value between 1 and 0.

\[ S(z) = \frac{1}{1 + e^{-z}} \]  

Where, \( S(z) \) is output among 1 and 0 value

\( Z \) - input of this function

\( e \) - base-natural log

Accuracy for Light Gradient Booster Algorithm and Logistic Regression was calculated founded on calculation (4)

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \]  

Where TP- No.of factual positive categorized by the model

TN- No.of factual negative categorized by the model

FP- No.of False Positive categorized by the model

FN-No.of False Negative categorized by the model

Hardware Arrangement was an Intel i5 processor with 8GB RAM. The X64 based processor with HDD machine with 917GB are used and OS used is 64 bit. Windows 10 Operating system is used and the google colab was the tool used for implementing python. Initially the dataset collected has been preprocessed by removing all null values after preprocessing the dataset in order. Randomly it is then spitted into two sets namely 25% testing set and 75% training set. Light Gradient method was implemented by analyzing the test and train data in the preprocessed datasets which is stored in the repository and it helps to predict the accuracy percentage of credit card fraudulent transactions in financial groups

2.3. Statistical Analysis

SPSS Framework has been used to conduct a sample T-test. The independent variables are Class, Amount and Time. The dependent variable is the accuracy percentage of credit card fraudulent transactions.
3. Result

Light Gradient Booster and Logistic Regression are running at various times in google colabs with 10 sample size and performance is measured. It’s found in Table 1,

Table 1. Predicted Accuracy of Credit card Fraudulent Transactions of Light Gradient Booster algorithm is 91.6% and Logistic Regression is 81.4%. The accuracy varies from different sample sizes.

<table>
<thead>
<tr>
<th>SI.NO</th>
<th>Random state</th>
<th>Light GBM</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>94.50</td>
<td>85.60</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>95.00</td>
<td>84.12</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>92.00</td>
<td>79.12</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>93.00</td>
<td>82.45</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>92.61</td>
<td>81.65</td>
</tr>
<tr>
<td>6</td>
<td>600</td>
<td>89.25</td>
<td>78.58</td>
</tr>
<tr>
<td>7</td>
<td>700</td>
<td>91.60</td>
<td>81.40</td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>90.11</td>
<td>80.14</td>
</tr>
<tr>
<td>9</td>
<td>900</td>
<td>88.80</td>
<td>84.26</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>90.26</td>
<td>81.47</td>
</tr>
</tbody>
</table>

Mean accuracy of Light gradient Booster was 91.6% and Logistic Regression was 81.4%. In addition to analysis of the experiment the work was statistically significant, analyzed using the SPSS (Statistical Package for the Social Sciences). For both Light Gradient Booster and Logistic Regression Algorithm the Standard Deviation, Mean, and Standard Error Mean shows in Table 2, the Light Gradient Booster algorithm is (91.6%) is more than Logistic Regression (81.4%) Error mean for Light Gradient Booster algorithm is 0.67 and for Logistic Regression is 0.86. Loss percentage of Light gradient Booster is 8.3 is improved than Logistic regression 18.5.

Table 2. Group Statistics for Mean of Light Gradient Booster algorithm and Logistic Regression.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>N</th>
<th>Mean</th>
<th>Std_deviation</th>
<th>Std_Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Light GBM</td>
<td>91.6780</td>
<td>2.12263</td>
<td>.67123</td>
</tr>
<tr>
<td></td>
<td>Log_ regression</td>
<td>81.4540</td>
<td>2.73170</td>
<td>.86384</td>
</tr>
<tr>
<td>Loss</td>
<td>Light GBM</td>
<td>8.3080</td>
<td>2.12089</td>
<td>.67068</td>
</tr>
<tr>
<td></td>
<td>Log_ regression</td>
<td>18.5460</td>
<td>2.73170</td>
<td>.86384</td>
</tr>
</tbody>
</table>

The Mean of Light Gradient (2.12) is improved than the Logistic Regression with a SD of (2.73). Light Gradient produces a significant difference than Logistic Regression with 0.0001 (p<0.05) based on 2-tailed analysis and with 95% confidence interval as shown in Table 3. In Fig. 1 it shows the bar chart comparison of two algorithms with accuracy of 91.6% and Loss of 8.3% for Light Gradient algorithm.
Table 3. Independent Sample T-test results

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>Sig.</th>
<th>t</th>
<th>df</th>
<th>Sig.(2-tailed)</th>
<th>Mean Difference</th>
<th>Std. Error Differences</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>.525</td>
<td>.478</td>
<td>9.346</td>
<td>18</td>
<td>0.0001</td>
<td>10.22</td>
<td>1.0939</td>
<td>7.9256</td>
<td>12.522</td>
</tr>
<tr>
<td>Equal Variances assumed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal Variances not assumed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Loss</strong></td>
<td>.519</td>
<td>.481</td>
<td>-9.346</td>
<td>18</td>
<td>0.0001</td>
<td>-10.23</td>
<td>1.0939</td>
<td>-12.535</td>
<td>-7.9403</td>
</tr>
<tr>
<td>Equal Variances assumed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal Variances not assumed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Comparison of Light Gradient and Logistic Regression in terms of Mean accuracy.
Mean accuracy of Light Gradient Booster (91.6%) is better than Logistic Regression (81.4%). Loss percentage of Light Gradient is (8.3) is healthier than Logistic Regression (18.5). X-axis: Light Gradient Booster vs Logistic Regression Algorithm, Y-axis: detection of accuracy mean. That is with error of 95% C.I, +/- 2SD.

4. Discussion

In this research work Light Gradient Booster and Logistic Regression was analysed for predicting the accuracy percentage of credit card fraudulent transactions. It is observed that Light Gradient Booster achieved better accuracy (91.6%) compared to Logistic Regression (81.4%). Sigmoid role plans the prediction to the possibilities of credit card fraudulent transactions based on class and amount attributes of the dataset, It is sequenced to improving the accuracy percentage with significance of 0.0001 (p<0.005), Finally the result shows evidence that there is a significant modification between Light Gradient and Logistic Regression algorithms. Features that affect the accuracy ratio of credit card fraudulent transactions are V1,V2,V3 to V28, the principal components which are used to obtain PCA [14]. Class and amount are features that are mainly focused to improve accuracy ratio of credit card fraudulent transactions in various financial environments [15]. Light Gradient Booster Algorithm got better accuracy compared with previous research articles. In real world applications for public credit card transactions and all online transactions the proposed work can be applied. The limitation in proposed work is the original features attributes are not provided, the attributes provided for PCA are not clearly defined [16],[17]. If the Fraudulent dataset could provide transaction time and number of transactions, accuracy can be improved further. In future, by implementing bootstrapping and resampling techniques, we can increase the prediction percentage of fraudulent credit card transactions [18].

5. Conclusion

It is identified that the Proposed model proves Light Gradient Booster Algorithm has a better accuracy of 91.6% with maximum of 25 iterations with respect to SSPS tool and its sigmoid variations, when compared to Logistic Regression with accuracy of 81.4% for predicting credit card fraudulent transactions. In near future the same can be improved in advanced by formulating the repository of the credit card data base, which may increase the accuracy of finding fraudulent transactions.

References


