International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 12, Issue 5 [May. 2023] PP: 69-73

Credit Card Fraud Detection Using Machine Learning

[1]Dr.Ramkumar J, [2]Mr.Dhanush B
[1]Professor, Department of Computer Science,
Dr.N.G.P Arts and Science College,
Coimbatore-48,Tamil Nadu, India
[2]II Msc Computer Science, Dr.N.G.P Arts and Science College,
Coimbatore-48,Tamil Nadu, India.

ABSTRACT:
One of the primary purposes of banking is to manage credit risk. Risk is categorised by banks based on their characteristics. Despite the proliferation of algorithms, there is still a problem to be solved. The acquired results from cluster analysis and artificial neural networks related to fraud detection demonstrated attribute clustering and neural inputs can be decreased. Non-existence, data normalisation is applied before cluster analysis. The significance of the work is in developing a cost-cutting algorithm. The algorithm employed was Minimal Bayesian-Risk, and the outcome was 23%. (MBR). Random Forest Algorithm is employed for classification and regression in the suggested system. Because it corrects the bad behaviour of overgrazing its training datasets, a random forest has an advantage over a decision tree. It is discovered to offer a good generalisation error estimate that is resistant to overfitting. Credit card datasets are gathered for training datasets in credit card fraud detection, whereas user credit card queries are gathered for testing datasets. The Random Forest Algorithm is used to assess the datasets and current datasets after the categorization phase. The accuracy gained by Random Forest when optimization is completed is 99.9%.

KEYWORDS: *Machine Learning techniques, Credit Card, Data Analysis*

Date of Submission: 28-04-2023 Date of acceptance: 07-05-2023

I. INTRODUCTION:

People's main concern with data mining in recent years has been the model used to detect credit card fraud. The traditional data mining algorithms are not immediately applicable to our topic because it is handled as a classification problem. The goal of this project is to suggest a supervised learning algorithm-based system for detecting credit card fraud. Aiming to produce better solutions over time, supervised algorithms are evolutionary algorithms.

The most often used form of payment is a credit card. Identity theft and fraud are on the rise as the number of people using credit cards globally rises. Only the card information—card number, expiration date, security code, etc.is needed to make a virtual card purchase. Typically, these purchases are made over the phone or the Internet. All one needs to do to conduct fraud on these kinds of purchases is to know the card information. Credit cards are the most popular payment option for internet transactions. Credit card information should be kept secret. Credit card privacy information shouldn't be affected. Examples of ways to steal credit card information include phishing websites, lost or stolen credit cards, phone credit cards, the theft of card information, intercepted cards, etc. For safety concerns, it is best to steer clear of the aforementioned activities. Online fraud simply requires the card information and takes place remotely. A manual signature, PIN, or card imprint are not required at the time of transaction[1].

Usually, the legitimate cardholder is not aware that their card information has been viewed or stolen. Examining each card's spending patterns and looking for any departures from "normal" spending patterns is the simplest technique to identify this type of fraud. The best way to reduce successful credit card fraud is to identify it by looking at recent purchases of cardholder data. Because the data sets are not available and the results are not disclosed. Two sorts of data can be utilised to find instances of fraud: logged data and user behaviour. For the time being, fraud detection methods include data mining, analytics, and artificial intelligence.

II. LITERATURE SURVEY:

Multiple Supervised and Semi-Supervised machine learning techniques are used for fraud detection [8], but we aim is to overcome three main challenges with card frauds related dataset i.e., strong class imbalance, the inclusion of labelled and unlabelled samples, and to increase the ability to process a large number of transactions.

Different Supervised machine learning algorithms [3] like Decision Trees, Naive Bayes Classification, Least Squares Regression, Logistic Regression and SVM are used to detect fraudulent transactions in real-time datasets.

Two methods under random forests [6] are used to train the behavioural features of normal and abnormal transactions. They are Random-tree-based random forest and CART-based. Even though random forest obtains good results on small set data, there are still some problems in case of imbalanced data. The future work will focus on solving the above-mentioned problem. The algorithm of the random forest itself should be improved.

Performance of Logistic Regression, K-Nearest Neighbour, and Naïve Bayes are analysed on highly skewed credit card fraud data where Research is carried out on examining meta-classifiers and meta-learning approaches in handling highly imbalanced credit card fraud data[10].

Through supervised learning methods can be used there may fail at certain cases of detecting the fraud cases. A model of deep Auto-encoder and restricted Boltzmann machine (RBM) [2] that can construct normal transactions to find anomalies from normal patterns. Not only that a hybrid method is developed with a combination of Adaboost and Majority Voting methods [4].

MODULE DESCRIPTION:

The model generation subsequently occurs, followed by analysis of the results. Each step of the model is discussed in detail in subsequent sections.

Data acquisition:

The card fraud detection dataset obtained from the Kaggle. It contains 31 features and 7973 records.

	Time	V1	V2	V3	V4	V5	V6	V7
0	0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599
1	0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803
2	1	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461
3	1	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609
4	2	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941
...	...	...	...	...	...	...	...	...
7968	10980	1.284388	-0.013181	0.646174	0.198985	-0.568675	-0.526121	-0.448235
7969	10981	1.190428	-0.122329	0.954945	0.267101	-0.971026	-0.652279	-0.612992
7970	10981	-0.725175	0.298202	1.824761	-2.587170	0.283605	-0.016617	0.153659
7971	10981	1.226153	-0.129645	0.735197	0.142752	-0.703245	-0.349641	-0.612641
7972	10981	1.145381	-0.059349	0.968088	0.267891	-0.822582	-0.597727	-0.450197

7973 rows × 31 columns

Data pre processing:

Credit card fraud detection dataset is first loaded and then data cleaning and finding missing values was performed on all records.The dataset contains complete information.

Splitting dataset:

The splitting of the dataset in the ratios of training and testing set in percentile

Selected algorithm for implementing:

Following classification algorithms are then applied on the pre-processed dataset.

a)Supervised Learning:

In this technique, both the input and output are known ahead of time. This is known as supervised learning because it learns from a training data set and builds a model from it, which then predicts results when applied to new data. Supervised learning techniques include Logistic Regression, Naive Bayes algorithm and Random Forest.

Logistic Regression:

Logistic Regression is a popular means of supervised learning which is used to estimate outcomes such as win/loss, positive/negative etc[5]. It makes use of a sigmoid function whose value lies in between 0 and 1.

Accuracy Score

```
[ ] # accuracy on training data
    from sklearn.metrics import accuracy_score
    X_train_prediction = model.predict(X_train)
    training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[ ] print('Accuracy on Training data : ', training_data_accuracy)

    Accuracy on Training data :  0.9992472713586752

    # accuracy on test data
    X_test_prediction = model.predict(X_test)
    test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[ ] print('Accuracy score on Test Data : ', test_data_accuracy)

    Accuracy score on Test Data :  1.0
```

Naive Bayes:

This Naive Bayes classifier is based on simplest Bayesian network models. This classifier is highly scalable requiring a number of parameters in a problem[6]. It is based on Bayes theorem on conditional probability and the attributes however are assumed to be independent of each other.

Naive Bayes

```
[ ] from sklearn.naive_bayes import GaussianNB

[ ] nb = GaussianNB()
    nb.fit(X_train,Y_train)
    acc = nb.score(X_test,Y_test)*100
    print("Accuracy of Naive Bayes: {:.2f}%".format(acc))

    Accuracy of Naive Bayes: 98.75%
```

Random Forest:

Random Forest is an ensemble method which relies on averaging a lot of decision trees and is used for classification and regression. Unlike decision trees this method is less prone to overfitting[7]. Its goal is to reduce variance. Although there is a small increase in bias and some loss of interpretability the overall performance is boosted.

Random Forest

```
    from sklearn.ensemble import RandomForestClassifier

[ ] rfc = RandomForestClassifier()
    rfc.fit(X_train,Y_train)

    RandomForestClassifier
    RandomForestClassifier()

[ ] y_pred = rfc.predict(X_test)

[ ] print("The model used is Random Forest classifier")
    acc= accuracy_score(Y_test,y_pred)
    print("The accuracy is  {}".format(acc))

    The model used is Random Forest classifier
    The accuracy is  1.0
```

b)Unsupervised Learning:

When we have only input data and no corresponding output variable, we call it unsupervised learning. Unsupervised learning's main task is to automatically create class labels. The association between the data can be discovered using unsupervised learning methods to see if they can be grouped together. Clusters are the name for this type of group. Cluster analyses is another term for unsupervised learning. Unsupervised learning techniques include K Means Clustering.

K-Means:

K-means uses the initial cluster centers to group similar objects to any one of them and thus form arbitrary shapes called clusters. The parameters that are required are the value of K and the initial choice of cluster centers for the K clusters. The shapes of the clusters highly depend on the initial choice of cluster[9].

```
[ ] kmeans_predicted_test_labels=kmeans.predict(test_features)
    kmeans_accuracy_score=accuracy_score(test_labels,kmeans_predicted_test_labels)

    print("Accuracy -->",kmeans_accuracy_score)

    Accuracy --> 0.21105527638190955
```

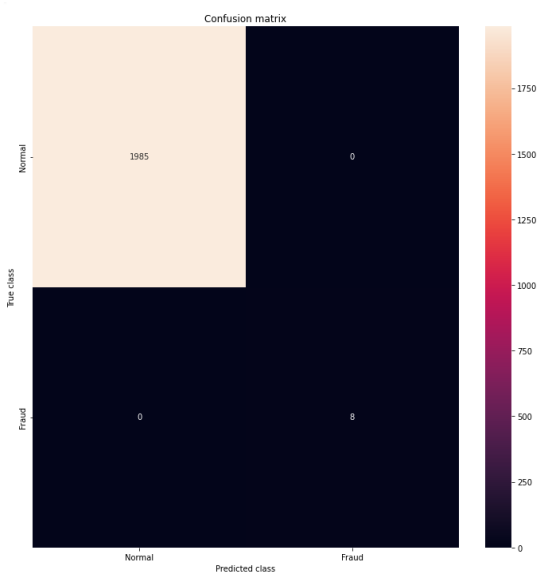
Confusion matrix:

The confusion matrix offers further information about a predictive model's performance, as well as which classes are correctly and mistakenly predicted, and what kinds of mistakes are being produced. A two-class classification issue with negative and positive classes has the simplest confusion matrix. Each cell in the table in this kind of confusion matrix has a distinct and understandable name. Accuracy is the percentage of correctly classified instances. It is one of the most widely used classification performance metrics.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total Number of predictions}}$$

Or for binary classification models. The accuracy can be defined as:

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

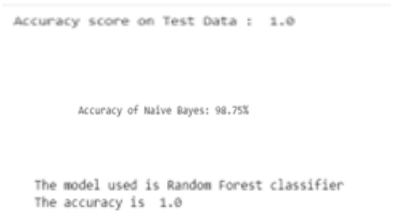


SPECIFICITY:

Computing specificity is done by dividing the total number of negatives by the number of correctly predicted negative outcomes (B)(D).

Specificity=B/D.

The performance of the system to find credit card fraud is reported using accuracy, error rate, sensitivity, and specificity. Three machine learning methods are created in this article to identify credit card system fraud. 30% of the dataset is utilised for testing and validation, while the remaining 70% is used to train the algorithms. accuracy, error rate, sensitivity, and specificity are utilized to evaluate three methods for various variables. Results for Logistics Regression, Naive Bayes, Random Forest accuracy level are 1.0, 98.75% and 1.0. The comparisons' outcomes show that the Logistics Regression and Random Forest methodology performs better than the Naive Bayes.



III. DISCUSSION:

This is because it simply requires one straightforward mathematical operation (the sum operation) to identify the classes and provide the desired result. In terms of reaction time, the Naive Bayes classifier comes in second. This is due to the fact that this classifier must carry out more mathematical operations in order to determine the distances between the new value and the centres of each cluster, which takes more time. The performance of the Naive Bayes classifier is the worst when compared to the other classifiers.

IV. CONCLUSION:

The best results are produced by the logistic regression-based classifier (accuracy = 1.0 sensitivity = 99%, and error rate = 0.1%). The algorithm satisfies this criterion. The technique not only addresses the issue of non-unique outcomes, but it also has broad applicability to other problem kinds. Our approach is better able to handle issues with both uniform and non-uniform distribution of data points. When using our algorithm, in order to improve the k-means clustering technique by removing one of its shortcomings. However, there is still more effort to be done to improve the k-means algorithm. K-Means can only be used with numerical data. But in everyday life, we come across situations that include both numerical and categorical data values. Therefore, further research may be done to adapt the k-means algorithm to mixed data type of data.

REFERENCES:

[1]. Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A., 2017, October. Credit card fraud detection using machine learning techniques: A co parative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-9). IEEE.

[2]. Pumsirirat, A. and Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).

[3]. Mohammed, Emad, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1July2018,doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.

[4]. Randhawa, Kuldeep, et al. "Credit Card Fraud Detection Using AdaBoost and Majority Voting." IEEE Access, vol. 6, 2018, pp. 14277–14284., doi:10.1109/access.2018.2806420.

[5]. Ng, A. Y., and Jordan, M. I., (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 2,841-848.

[6]. Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B., (2002). Credit card fraud detection using Bayesian and neural networks. Proceeding International NAISO Congress on Neuro Fuzzy Technologies.

[7]. Shen, A., Tong, R., & Deng, Y. (2007). Application of classification models on credit card fraud detection. In Service Systems and Service Management, 2007 International Conference on (pp. 1-4). IEEE

[8]. Melo-Acosta, German E., et al. "Fraud Detection in Big Data Using Supervised and Semi-Supervised Learning Techniques." 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 2017, doi:10.1109/colcomcon.2017.8088206.

[9]. Adewumi, A.O. and Akinyelu, A.A., 2017. A survey of machine-learning and nature-inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management, 8(2), pp.937-953.

[10]. Bhatla, T.P.; Prabhu, V.; and Dua, A. (2003). Understanding credit card frauds. Crads Business Review# 2003-1, Tata Consultancy Services.