BEHAVIOUR BASED CREDIT CARD FRAUD DETECTION ANALYSIS BY EMPLOYING MACHINE LEARNING ALGORITHMS

Govind Prasad Buddha GRADXS PhD Research Scholar registered with Computer Science ,Liutebm University. under supervision of GRADXS Research Mentor supervisor
Dr. NAGAMALLESWARA RAO. Department of Computer Science and Engineering, SDES, Hyderabad.

ABSTRACT

The expansion of online purchases is a direct outcome of the rise of e-commerce, which in turn has spawned a wide range of safety concerns. Even while there are always potential security risks when using any human-created technology, online or electronic payment systems provide the highest level of protection possible. Since e-payments are so convenient, they are attractive to both merchants and consumers. As a result, there is an urgent need for efficient tools to counteract the risks posed by the internet. The use of credit and debit cards is one of the most prevalent forms of electronic commerce that is vulnerable to these dangers. Stopping credit card fraud has consistently ranked towards the top of banks' lists of security concerns. There are numerous methods used in credit card fraud. The ability to identify and stop such scams is one of today's most pressing concerns. Due to the rise in online fraud, scientists have turned to a variety of machine learning techniques to help them identify and analyse instances of online theft. The primary goal of this work is to provide a brand-new fraud detection approach for Streaming Transaction Data by analysing customers' past financial dealings and drawing conclusions about their habits based on what they learn. Methods that rely on rules-based approaches or classic point solutions are out of date. For banks and other financial services institutions, the time and money spent by legal and compliance departments attempting to overcome these roadblocks is prohibitive. Advanced analytics, as well as AI and ML capabilities, free up fraud and compliance teams to focus on the most difficult cases. Complex algorithms powered by ML can lessen the need for manual investigation; when used in tandem with rules, this approach to fraud detection has considerable advantages over rule-based systems alone. In this paper, we compare and contrast the various machine learning algorithms that are currently in use for credit card fraud detection.

Key words: Credit card, Fraud, Machine Learning.
DOI Number: 10.14704/nq.2022.20.11.NQ66190 **NeuroQuantology 2022; 20(11): 1953-1962**

1. INTRODUCTION

Since the 1970s, people have been able to make purchases with their computers instead of writing checks. A wide variety of payment systems were developed. Following the advent of the internet and its subsequent widespread adoption, its user base has grown exponentially around the globe. In the late 1990s, electronic forms of payment began to gain widespread acceptance. The concept of cashless transactions was developed around this time. Electronic commerce or cashless transactions relied on this. Also at this time, e-commerce began to emerge. As a result, scientists began examining the procedure, and numerous scholarly articles were published about it. Research into this field began to flourish as it became clear that there was substantial business interest in the process as well.

There has been a rise in the use of credit card transactions in recent years. You can use it for both in-store and online purchases. Nowadays, credit card payments are not only common but also highly practical. In view of the increasing volume of fraudulent financial transactions, it is critical to pinpoint the most efficient fraud detection model.

The proliferation of advanced technologies has coincided with a meteoric rise in fraudulent activity [1]. Furthermore, today's information explosion [2] is unprecedented in human history. Analysis of the cardholder's spending habits is a promising method for

spotting the scam. The key to spotting a fraudster is learning to spot the suspicious one [3]. If there is a change in spending habits, it is investigated further as a possible red flag. In this research, we use support vector machines and a behavior-based approach to detect fraud.

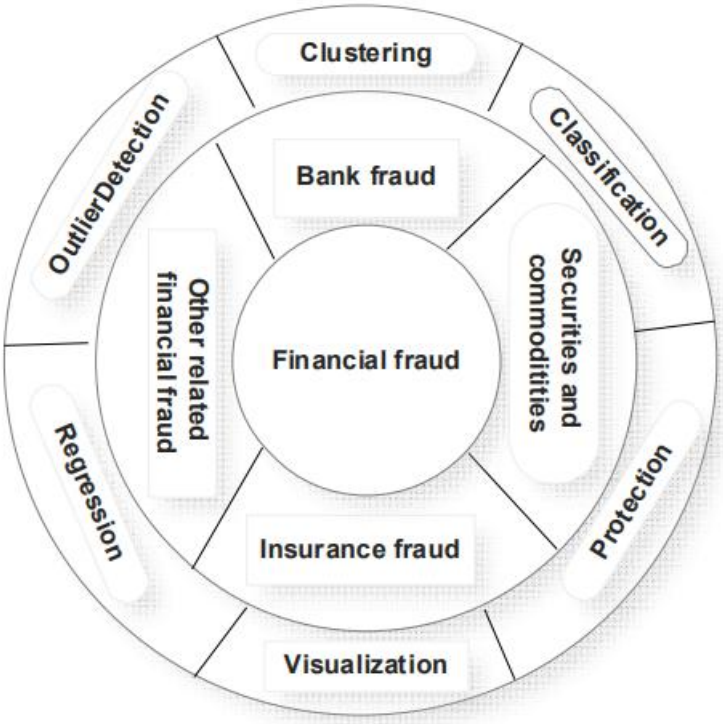

Figure 1. Conceptual Framework for Classification of Frauds

The conceptual framework for fraud classification is depicted in Figure 1. The term "distance" is used in the context of comparing data points in order to identify outliers that are extremely dissimilar to the norm. Some on-card transactions are now safer due to the widespread adoption of EMV cards by banks; these are smart cards that store their data on integrated circuits rather than on magnetic stripes. However, card-not-present fraud is still a significant problem. The risk of credit card fraud persists even then. There are a plethora of machine learning methods that can be used to fix this issue.

2. LITERATURER REVIEW

Multiple, cutting-edge fraud detection technologies, including neural networks, data mining, and distributed data mining, are described together with novel research methodologies in the subject of credit card fraud detection. Many other techniques can also be used to detect credit card fraud. According to the research we conducted, there are numerous alternatives to Machine Learning for detecting credit card fraud. Credit card fraud investigations employ both Machine Learning and Deep Learning techniques.

In 2019, Yashvi Jain, NamrataTiwari, Shripriya Dubey, and Sarika jain investigated various approaches to detecting credit card fraud [1]. These approaches included: support vector machines (SVM), artificial neural networks (ANN), Bayesian Networks (BN), Hidden Markov Model (HMM), K-Nearest Neighbors (KNN), Fuzzy Logic systems (FLS), and Decision Trees. Based on their research, they concluded that the k-nearest neighbour method, decision trees, and the support vector machine approach all had similar levels of accuracy. The algorithms with the lowest accuracy are Fuzzy Logic and Logistic Regression. High rates of detention are provided by neural networks, naïve bayes,

eISSN 1303-5150

fuzzy systems, and KNN. For a reliable medium-level detection rate, try the Logistic Regression, SVM decision trees. In terms of overall performance, ANN and Naive Bayesian Networks are the two best methods to use. Training one of these is quite costly.

Criminals have moved their attention to CNP transactions as the security of chip cards has been strengthened, as reported in 2017 [2] by the US Payments Forum. In Fig. 2, we can see the total number of CNP fraud reports for each year.

Even though Supervised and Semi-Supervised machine learning algorithms are widely used for fraud detection [3], our aim is to overcome the following three primary challenges when working with datasets linked to card frauds: (1) a large gap between classes, (2) the presence of both labelled and unlabelled samples, and (3) the necessity of handling a large number of transactions simultaneously.

Supervised machine learning [4] methods including Decision Trees, Naive Bayes Classification, Least Squares Regression, Logistic Regression, and Support Vector Machines can be used to identify fraudulent transactions in streaming data. Training the behavioural characteristics of typical and atypical financial dealings is accomplished using two techniques falling under random forests [5]. They use a CART- and a random forest-based methodology, respectively. While random forest does well on small sets of data, it can run into issues when the sets are unbalanced. The aim of the upcoming effort is to address the issue described above. There is room for improvement in the random forest algorithm.

Using highly imbalanced credit card fraud data, researchers compare the efficacy of different meta-classifiers and meta-learning strategies. These strategies include logistic regression, K-nearest neighbour, and naive bayes. Some fraud scenarios may not be detectable even if supervised learning techniques are utilised. An anomaly detection model built on top of a deep Auto-encoder and a restricted Boltzmann machine (RBM) [2]. Also, a hybrid approach combining Adaboost and Majority Voting has been devised [6].

3. FRAUD DETECTION

3.1 FRAUD DETECTION PROBLEM

The legal definition of fraud is dishonest misrepresentation. The goal of fraud could be anything from stealing merchandise without paying to draining bank accounts of their money. The two options for dealing with fraud are detection and prevention. When it comes to preventing fraud, preventative measures are taken. If measures to avoid fraud are unsuccessful, detecting it becomes a priority. Detecting fraud entails spotting actions that don't add up and are likely fraudulent. Many methods exist for identifying and preventing fraudulent activities. The fundamental objective is to ensure the legitimacy of monetary dealings and improve the precision of related predictions.

There are two main categories of credit card fraud: in-person and remote. The fraud begins with either the physical theft of a credit card or the unauthorised disclosure of account information to a third party, such as a cardholder's name, address, or phone number during a valid purchase. Using a stolen credit card in an in-person scam. Someone without permission takes the actual card. Afterward, he made purchases with the fraudulent card. Up until the card's expiration date, he can use it to make purchases. A huge loss occurs for the user and the financial institutions involved if the user is unaware of the theft [16]. When committing online fraud, only the card details are required to make a purchase. In this case, all the fraudster needs are the card data. Identity theft refers to the fraudulent use of a person's personal information. This form of fraud typically involves dealings conducted

eISSN 1303-5150

over the telephone or the internet. If a fraudster uses a card number that was seen or stolen, the real cardholder usually doesn't respond.

3.2 CHALLENGING ISSUES IN FRAUD DETECTION

The data sets are significantly biassed and severely imbalanced. Generally speaking, legitimate business makes up the bulk of all transactions. Instances of fraud are uncommon. For this reason, identifying fraud is challenging. Consideration of the fraudulent transaction as legitimate would result in significant financial loss if it were to occur.

Extremely large datasets with great dimensionality. Managing this huge data effectively is not a simple task. In order to handle such a massive data set, a scalable machine learning system is required. To protect the user's privacy, among other things, we don't disclose the actual data. Typically, the cost of a false positive is substantial for these kinds of detections. Costs associated with misclassification can be reduced with effective action.

3.3 BEHAVIOR BASED MODEL

The analysis of the user's financial behaviour is the central idea in fraud detection. A red flag is raised whenever a person's spending habits deviate significantly from their norm. Plus, this is being taken into consideration. Spending habits are quite individual. Credit card fraud can be detected using the cardholder's historical spending patterns, which is an exciting development. If a fraud detection model is said to be behaviorally based, it means that the data it employs come, either directly or indirectly, from the cardholder's transactional activity. One person's extravagant habits could be completely at odds with another's. Currently available fraud detection technologies rely heavily on analysing patterns of behaviour to pinpoint instances of shady financial dealings. Customers' typical behaviours, including transaction size, billing address, and other details, are gleaned from their spending habits. Changes between the billing and shipping addresses, substantial purchases made in an unfamiliar location, etc., are all indicators that something is amiss and should raise suspicions. The same way, out-of-the-ordinary actions are viewed with suspicion and investigated thoroughly.

4. COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHM

There are a total of five methods employed in this investigation: the support vector machine (SVM), logistic regression (LR), naive Bayes (NB), the K-Nearest Neighbor classifier (KNN), and random forest (RF). Grid Search was used to determine the optimal settings for this technique, which improved our model's accuracy after extensive testing.

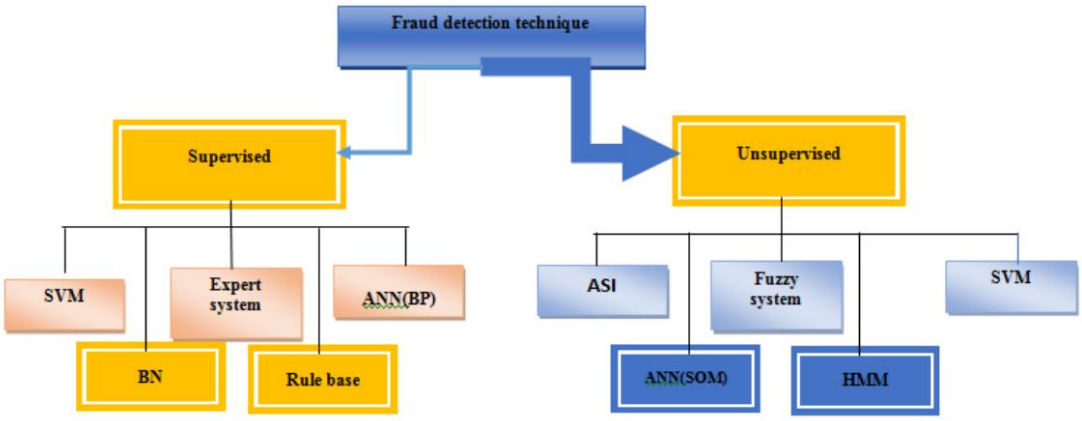

Figure 2: Machine Learning Techniques for Credit Card Fraud Detection

eISSN 1303-5150

SVM: We opted for the SVM approach because it performs well on non-linear classification-based tasks, can accommodate for non-uniform data structures, and has a low risk of model overfitting.

Logistic Regression: Using it on data with associated qualities yields the best results. It makes minimal use of computing resources. Due to its simplicity of implementation, it can serve as a baseline against which other algorithms are compared. In most cases, it provides the most insightful results of any classification strategy.

Gaussian NB (Naïve Bayes): The approach is conditional probability-based, thus it works well with current data. It has the potential to result in a well-formed recommendation system. To use it, all you need is a big data set. Conditional probabilities are determined using a formula. What is that formula?

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

Where P (A|B) = Posterior Probability, P(B|A) = Prior Probability, P(A) = Likelihood, P(A) = Evidence. This allows for a Probabilistic Prediction to be made with a reduced training dataset. It's flexible enough to work with both continuous and discrete information.

K-Neighbor classifier: In terms of dealing with noisy data, it performs admirably. It's a memory-based method that allows for the simultaneous usage of multiple categorization kinds (Binary class and Multi class) with minimal additional work. Classification and regression are two more applicable applications. Choosing initial parameters is challenging but eventually converges on the initial parameter.

The Random Forest: There is no need to transform or rescale the data in this method. It has useful applications in the areas of Classification and Regression. In order to achieve a positive outcome, the algorithm splits the data into distinct feature-based trees, each of which has high variance and low bias. It trains the model quickly, is simple to deploy, and is robust enough to deal with significant feature loss and errors in the data set.

5. A NOVEL APPROACH TO CREDIT CARD FRAUD DETECTION MODEL

5.1 Introduction

An unsupervised approach is at the heart of a new approach of spotting fraudulent charges on credit cards. It is generally agreed that unsupervised approaches are the best for fraud detection since the methods by which frauds occur have the property of vastly varied from the aforementioned categories of frauds. This technique has a higher propensity for uncovering previously unseen frauds [95] since it employs an unsupervised methodology. Rather than relying on a training base, which assumes that each instance of fraud would be similar to the last, it is preferable to implement a system that operates autonomously in the detection process. It is possible that labelled examples of both typical and out-of-the-ordinary data will not be readily available for use in the supervised models. Now that it's a disciplined process, uncovering fraudulent behaviour follows a strict schedule. This means that trends in fraudulent activity can never be accurately predicted. Since the researcher doesn't have access to the training dataset for a certain type of fraud occurrence, it's difficult for them to detect fraud.

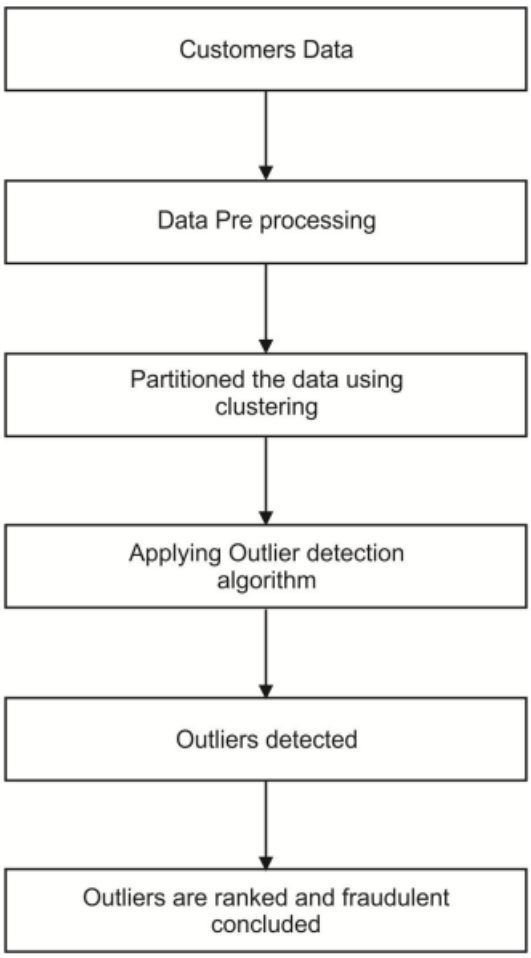

Figure 3. Outline of the Fraud Detection Process

Figure 3 shows an overview of the fraud detection procedure.

5.2 Data Pre Processing

The initial step involves transforming the unprocessed transaction data into a format that can be read by the Clustering algorithm. In most cases, the raw data is useless and must be cleaned up before it can be used for analysis. This necessitates the use of a pre-processing phase. In order to employ the clustering and outlier identification techniques, a single value from an attribute is all that is required. This number can be gleaned from the raw data itself, or it can be calculated using a subset of the current attributes. The user is responsible for providing their own preference here.

5.3 Detection Process using Clustering

The K-Means Clustering method is used to accomplish the clustering. Before the K-Means algorithm can begin its work, it needs to know how many individual groups should be made. It processes only one attribute value at a time, whether that value is direct or derived. The sequence of the items is not important. You can put the components in any order you like. Cluster centres are first chosen at random. These are the very first purchases recorded in the database. For each cycle, the transaction is examined to determine which cluster centre is most relevant. After this is done, each newly created cluster is evaluated independently, and its new cluster centre is determined by using the data points within it. For each iteration of the transactions, the closest cluster centre match is looked for in each

eISSN 1303-5150

transaction. This is done until no change occurs in the cluster's centre during successive repetitions. The completion of the cluster processing has been reached. In most cases, the starting point for a clustering method is a single parameter.

5.4 Algorithm

1. Initialize the number of clusters to be formed (n)
2. Set random n transactions to be the cluster center
3. For each transaction t
a. Find the cluster center(c) that is closest to t
b. Set c as the cluster center for t
4. End for
5. For each cluster cc
a. Using the contents of the cluster, find its center (c1)
6. End for
7. If c and c1 are the same
a. Clustering completed
b. Exit
8. Else
a. Goto step 3

The system produces a set of clusters once the clustering procedure is finished. When all possible data is available, it is inevitable that the cluster groups will include both clusters that show typical behaviour and those that show unusual behaviour. Since anomalies rarely occur in a system, it stands to reason that regular data clusters would be more denser than their abnormal counterparts. But that isn't enough to prove the data is legitimate. There is a strong chance that the end user will make a minor change to their typical spending habits on rare occasions. The unusual nature of these deals prevents us from categorising them as such. As a result, the transaction needs additional processing to verify its authenticity. The outlier detection algorithm handles this process.

6. RESULTS AND ANALYSIS

Table 1 Sample Confusion Matrix Set for SVM

TP	FP	TN	FN	FPR	TPR
33	25	30	12	0.454545	0.733333
127	187	102	84	0.647059	0.601896
411	187	275	160	0.404762	0.71979
623	285	408	184	0.411255	0.771995
749	392	551	308	0.415695	0.708609
867	448	819	366	0.353591	0.703163
1034	522	1093	351	0.32322	0.74657

The TPR and FPR readings are shown in Table 1, and the ROC for the readings is shown in Figure 4.

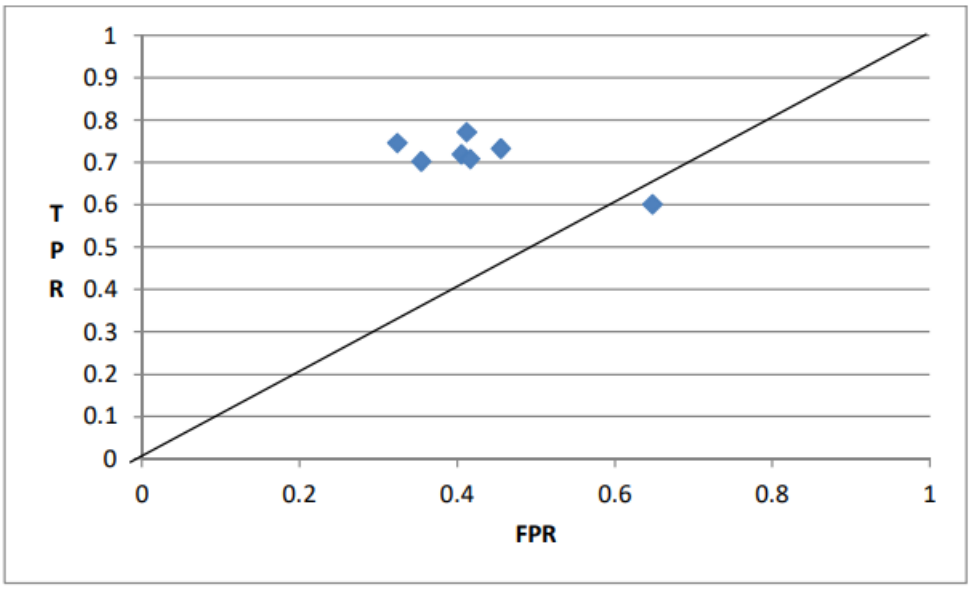

Figure 4. ROC – SVM

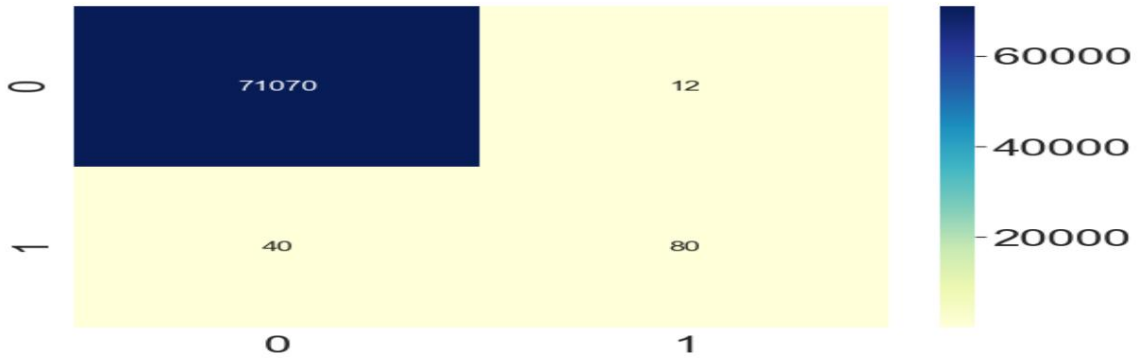

Figure 5: confusion Matrix of Logistic Regression and Naïve Bayes

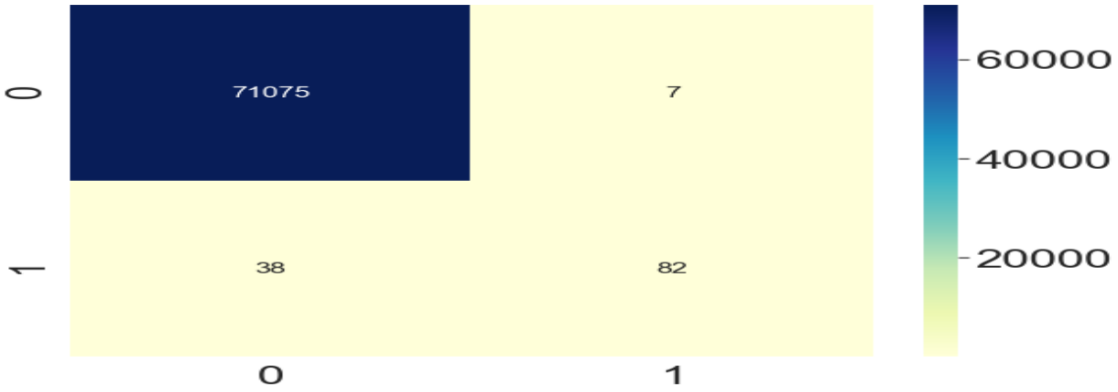

Figure 6: confusion matrix for K nearest neighbor

eISSN 1303-5150

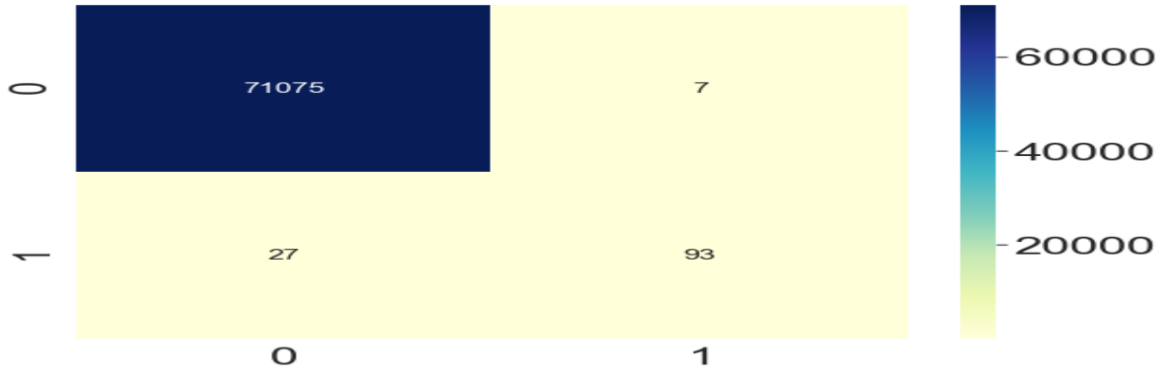

Figure 7: confusion matrix for Random Forest

CONCLUSION

In this study, we evaluate several machine learning methods on a dataset created specifically for the purpose of identifying credit card fraud. There are a total of five algorithms deployed here, including SVM, Naive Bayes, Logistic Regression, K-Nearest Neighbors, and Random Forest. In which Random forest and then KNN provide the highest-scoring results. The MCC is a useful metric for gauging an algorithm's efficiency, with a perfect score of 1 and a range of -1 to 1. It can be enhanced by combining it with other algorithms and by adding new technology to achieve more accurate results in identifying credit card fraud. By spotting fraudulent activity early on, we can cut down on it and save money. It will lessen the amount of money scammers lose through credit card fraud.

REFERENCES

[1] F. Ghobadi and M. Rohani, "Cost sensitive modeling of credit card fraud using neural network strategy," 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS), Tehran, 2016, pp. 1-5. doi: 10.1109/ICSPIS.2016.7869880

[2] https://www.ftc.gov/news-events/press-releases/2019/02/imposter-scams-top-complaints-made-ftc-2018

[3]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.

[4]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, 1 July 2018, doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.

[5]Xuan, Shiyang, et al. "Random Forest for Credit Card Fraud Detection." 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, doi:10.1109/icnsc.2018.8361343.

[6]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.

[7]. SamanehSorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi,A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective.

[8]. Kuldeep Randhawa, Chu Kiong Loo, ManjeevanSeera, Chee Peng Lim, Ashoke K. Nandi,Credit Card Fraud Detection Using AdaBoost and Majority Voting, Published in: IEEE Access on 15 February 2018, vol. no.6, pp. 14277 – 14283.

[9]. N. Sivakumar, Dr.R. Balasubramanian, Fraud Detection in Credit Card Transactions: Classification, Risks and Prevention Techniques, Published in (IJCSIT) International Journal of Computer Science and Information

eISSN 1303-5150

Technologies, vol no. 6 (2), 2015, pp. 1379-1386.

[10]. Sai Kiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya, Maheshwar Sharma,Credit card fraud detection using Naïve Bayes model based and KNN classifier, Published in: International Journal of Advance Research, Ideas and Innovations in Technology, Issue no. 3, vol.no. 4, 2018, pp.44-47.

[11]. Rishi Banerjee, Gabriela Boural, Steven Chen, Mehal Kashyap, Sonia Purohit, Jacob Battipagali,Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection, 2018.

[12]. Kaithekuzhical Leena Kurien and Dr. Ajeet Chikkamannur,Detection and prediction of credit card fraud transactions using machine learning, Published in: International Journal of engineering science & research technology (IJESRT), 2019, pp. 199-208.

[13]. Md.Akster Hossain and Mohammed Nazim Uddin,A Differentiate Analysis for Credit Card Fraud Detection, Published in: Int. Conf. on Innovations in Science, Engineering and Technology (ICISET), 27- 28 October 2018 (IEEE), pp. 328-333.

[14]. Suresh K Shirgave, Chetan J. Awati, Rashmi More, Sonam S. Patil,A Review on Credit Card Fraud Detection Using Machine Learning, Published in International journal of Science and Technology Research, vol.no.8, Issue no. 10, October 2019, pp. 1217-1220.

[15]. S P Mani raj, Aditya Saini, Swarna Deep Sarkar Shadab Ahmed,Credit Card Fraud Detection using Machine Learning and Data Science, Published in: International Journal of Engineering Research & Technology (IJERT), vol.no.8, Issue no. 09, September 2019, pp.110-115.