"Accuracy Study of Random Forest and Support Vector Machine for Fraudulent Credit Card Transactions "

Ms. Shweta Anil Kanojia[1], Dr. R.N. Jugele[2]

[1]Research Scholar, Department of Computer Science,
Science College, Nagpur.
[2]Professor, Department of Computer Science,
Science College, Nagpur

Abstract- Over time, e-commerce and digital transactions have become ingrained in every person's life and during the pandemic, an exponential growth has been noted in both cases. People are now more inclined to avoid carrying cash and instead use a variety of payment gateways, including debit cards, credit cards, net-banking, wallets, and UPI, to name a few. These notable developments in online transactions have resulted in an exponential rise in the number of credit card holders. But as the use of the online transactions is increasing, Frauds are also increasing day by day. Financial institutions are increasingly concerned about fraud because it costs them a lot of money and they are constantly under pressure to control the increasing risk that their clients are shown to. Using data analytics tools to quickly spot any fraud is essential for effective fraud management. This paper proposes a machine learning (ML) based credit card fraud detection engine using ML classifiers: Random Forest and Support Vector Machine. To validate the performance, the proposed credit card fraud detection engine is evaluated using a dataset generated from European cardholders. This study looks at how machine learning approaches, specifically Random Forest (RF) and Support Vector Machine (SVM), can improve the precision and effectiveness of fraud detection systems for credit card fraud. This paper also aims to suggest the best machine learning algorithm out of Random Forest and Support Vector Machine for Credit Card Fraud Detection.

Keywords: Credit Card Fraud Detection, Machine learning Algorithms, Fraud Detection, Random Forest , Support Vector Machine.

Introduction- The amount of credit card fraud is rising daily. It is possible to perform credit card fraud in both offline and online transactions. While virtual cards are needed for online transactions, physical cards are needed for offline transactions in order to conduct fraudulent or criminal activity. Therefore, these credit card fraud activities may result in multiple fraudulent transactions that the actual users are unaware of. [1]. To carry out transactions, fraudsters seek sensitive data, including bank account numbers, credit card numbers, and other user information [2]. When doing transactions locally, fraudsters must take control of the user's credit card; when conducting transactions online, they must take control of the user's identity and login credentials [3]. As a result, credit card fraud has emerged as a major concern in today's technological environment, significantly preventing bank operations. Machine learning-based credit card identification involves training the algorithm based previous data and improving it through experience. Later, the system may detect fraud in future transactions. If fraud is detected, the activity can be instantly rejected and the cardholder and bank alerted accordingly [8]. Sensitive data is lost as a result of multiple fraudulent transactions that are difficult for both the user and the banking authorities to detect [4]. Various models are used to detect fraud transactions based on transaction behavior. These techniques fall into two main categories, such as supervised learning and unsupervised learning algorithm. The aim of this paper is to detect the accuracy of fradulent transactions by using Random Forest and Support Vector Machine Algorithm.

Random Forest:

One of the machine learning methods that is most frequently used in practical applications and deployed models is the random forest technique [1].It is an Supervised Machine Learning technique. Classification and Regression are the applications of Random Forest [8]. Random Forest is an ensemble technique as it consists of several decision trees [8]. The accuracy increases as the number of trees increases [2]. It is a mixture of several classifiers [3]. A random subset of each partition's characteristics is measured by constructing each tree with a random subset of the data set. Individual trees become more volatile as a result of this variability, lowering the likelihood of overfitting and improving overall prediction performance [5].

Support Vector Machine:

SVM, a linear model, is an efficient text algorithm. This is a supervised machine learning algorithm. Classification and regression problems can be solved using the linear model Support Vector Machine (Classifier) [1]. With large margins, the SVM distinguishes between positive and negative examples. The SVM gave superior results for fraud detection than naïve bayes [4]. Based on support vectors, training points are divided into two groups using a decision surface. To effectively categorize future data, the SVM algorithm builds a hyperplane that divides a number of qualities into categories. Support vectors are the extreme points that are chosen when using this method to construct a hyperplane [6].

Random Forest	**Support Vector Machine**
As random forests can show the importance of a trait, they are usually simpler to understand.	SVMs can be more difficult to understand particularly when they include complicated kernels.
Large datasets usually scale better with random forests	Large datasets can make SVMs computationally costly and slow, especially when non-linear kernels are used.
SVMs are more sensitive to feature scaling.	Random forests are more resistant to feature scaling.
Random forests are relatively robust to noise in the data due to the averaging process across multiple trees	By adjusting the regularization parameter, SVMs can find a balance between maximizing the margin and allowing some misclassifications, which helps prevent overfitting.

Table 1: Strength of Random Forest and Support Vector Machine [7]

Experimental Results:

Evaluation Criteria

The accuracy score, classification report, F1-score, confusion matrix, and other parameters are used to assess the performance of the classification algorithms. The following are a few important definitions:

- **True positive (TP)-** It is an outcome in which the model accurately predicts the positive class.
- **False positive (FP)-** It occurs when the positive class is predicted wrongly by the model.
- **True negative (TN)-** It is an outcome in which the model accurately predicts the negative class.
- **False negative (FN)-** It is an outcome in which the model predicts the negative class inaccurately.

Accuracy- It is calculated by dividing the number of correct predictions to the total number of predictions.

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

Precision- It is defined as the number of true positives divided by the total number of positive predictions (true positives plus false positives).

$$\textbf{Precision} =\frac{TP}{TP+FP}$$

Recall - It is computed by dividing the number of true positives by the number of positive cases.

$$\textbf{Recall} =\frac{TP}{TP+FN}$$

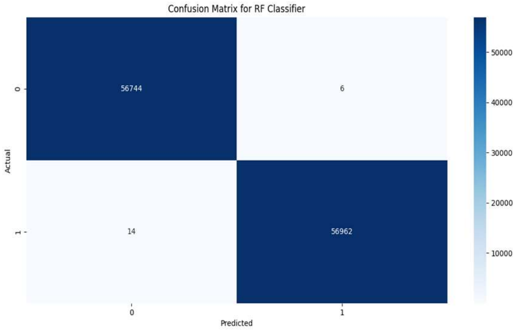


Fig 1: Confusion Matrix for RF classifier

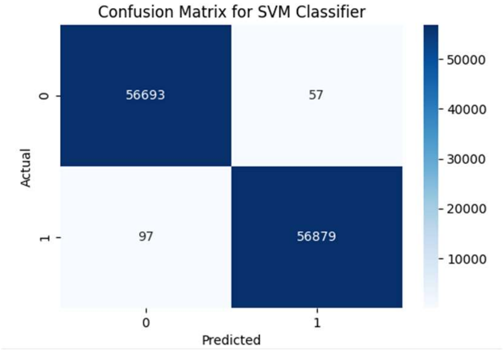
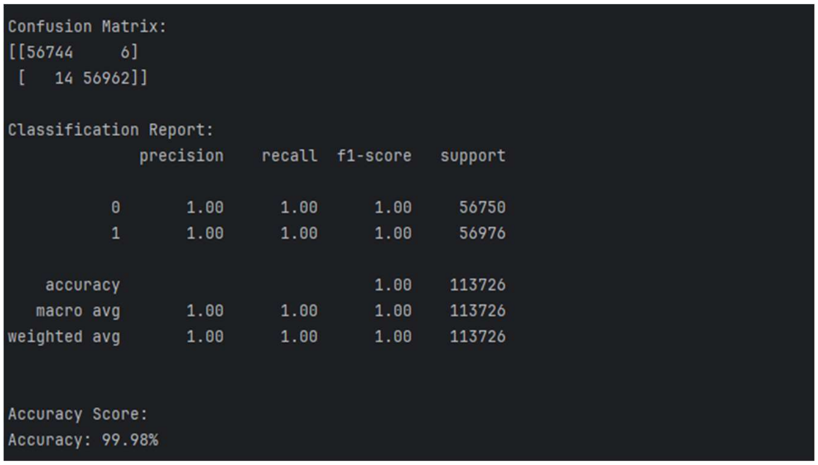


Fig 2: Confusion Matrix for SVM classifier

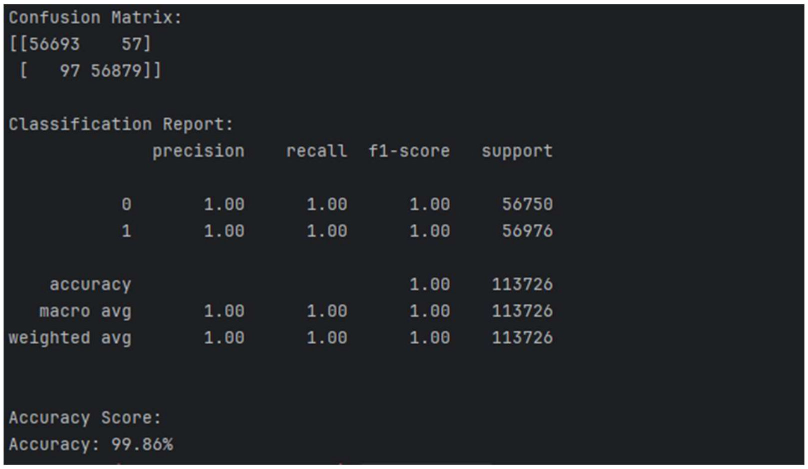
Results and Discussion:

Dataset Summary: The dataset used is from Kaggle and includes transactions of European credit card customers. The dataset was saved as a CSV file. The collection contains 550,000 transactions.



```
Confusion Matrix:
[[56744     6]
 [   14 56962]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56750
           1       1.00      1.00      1.00     56976

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726


Accuracy Score:
Accuracy: 99.98%
```

Fig 3: Output of Random Forest



```
Confusion Matrix:
[[56693    57]
 [   97 56879]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56750
           1       1.00      1.00      1.00     56976

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726


Accuracy Score:
Accuracy: 99.86%
```

Fig 4: Output of Support Vector Machine

	Accuracy	Precision	Recall(Sensitivity)
Random Forest	0.9998	0.9997	0.9998
Support Vector Machine	0.9986	0.9982	0.9989

Table 2: Classification reports of models

Conclusion: Advances in electronic financial transaction technology and simple payment mechanisms have increased the risk of fraudulent payments due to simplified authentication processes.Credit card fraud can take various forms, such as theft, identity theft, card counterfeiting, and information theft. Card information theft is a result of phishing, pharming and card information leaks. The issue of credit card theft has drawn a lot of attention in recent years. This research paper's conclusion summarizes the information collected by comparing and contrasting two machine learning models with regard to credit card fraud detection. This study has demonstrated the distinct capabilities and advantages of Machine Learning Algorithms- Random Forest and Support Vector Machine, in identifying fraudulent transactions. The study shows that both the algorithm has high accuracy but Random Forest has the highest accuracy of 0.9998 as compared to Support Vector Machine. Also, the execution time of Support Vector Machine is comparatively more than that of Random Forest.

References:

[1] A Sreni Chowdary ,Cherapalli Bavitha , Earagaraju Mounisha, Chatna Reethika, A Comparison Study of Fraud Detection in Usage of Credit Cards using Machine Learning, Mr. P.Yogendra Prasad , Proceedings of the 7th International Conference on Trends in Electronics and Informatics (ICOEI 2023) IEEE Xplore Part Number: CFP23J32-ART; ISBN: 979-8-3503-9728-4

[2] Uqba Jabeen ,Dr. Karan Singh, Satvik Vats , Credit Card Fraud Detection Scheme using Machine Learning and Synthetic Minority Oversampling Technique (SMOTE), Proceedings of the 5th International Conference on Inventive Research in Computing Applications (ICIRCA 2023) IEEE Xplore Part Number: CFP23N67-ART; ISBN: 979-8-3503-2142-5

[3] Indrani Vejalla ,Sai Preethi Battula, Kartheek Kalluri, Hemantha Kumar Kalluri, Credit Card Fraud Detection Using Machine Learning Techniques, 023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS) | 979-8-3503-1071-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/PCEMS58491.2023.10136040

[4] Fawaz Khaled Alarfaj , Iqra Malik, Hikmat Ullah Khan , Naif Almusallam , Muhammad Ramzan And Muzamil Ahmed, Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms.

[5] M.Suresh Kumar, V.Soundarya , S.Kavitha , E.S.Keerthika and E.Aswini , Credit Card Fraud Detection Using Random Forest Algorithm , 978-1-5386-9371-1/19/$31.00 c 2019 IEEE

[6] Prateeksha M.S , B. Naga Swetha and Manjula Patil , Credit Card Fraud Detection Using Machine-Learning , DOI: 10.21474/IJAR01/16824

[7] https://www.geeksforgeeks.org/ml-credit-card-fraud-detection

[8] Alavikunhu Panthakkan, Najiya Valappil , Majida Appathil, Seema Verma, Wathiq Mansoor, and Hussain Al-Ahmad, Performance Comparision of Credit Card Fraud Detection using Machine Learning , 022 5th International Conference on Signal Processing and Information Security (ICSPIS) | 978-1-6654-9265-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSPIS57063.2022.10002517