T.R

ÜSKÜDAR UNIVERSITY

FACULTY OF ENGINEERING AND NATURAL SCIENCES

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING ALGORITHM AND ANALYSIS BASED ON TIME SERIES DATA

GRADUATION THESIS

MUHAMMAD QASIM RAZA

DEPARTMENT OF CYBER SECURITY

Supervisor

Dr. Rowanda D. AHMED

İSTANBUL- 2022

T.R

ÜSKÜDAR UNIVERSITY

FACULTY OF ENGINEERING AND NATURAL SCIENCES

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING ALGORITHM AND ANALYSIS BASED ON TIME SERIES DATA

GRADUATION THESIS

MUHAMMAD QASIM RAZA

DEPARTMENT OF CYBER SECURITY

Supervisor

Dr. Rowanda D. AHMED

İSTANBUL- 2022

Muhammad Qasim Raza, student of Üsküdar University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering student ID 204306916 successfully defended the thesis/dissertation entitled “Credit Card Fraud Detection Using Machine Learning Algorithm and Analysis Based on Time Series Data”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor:

Asst.Prof Dr. Rowanda D. AHMED

Üsküdar University

Jury Members:

- **Asst.Prof.Dr. Muhammad Ilyas**

  Altınbaş University

- **Asst.Prof.Dr. Ahmet ŞENOL**

  Üsküdar University

Date of Submission : JULY 2022

Date of Defense : JULY 2022

2022

FORM OF DECLARATION

I hereby certify that the contents which was used in experiment, content and results are original and have not been submitted in whole or in part for consideration for any other degree or qualification at this or any other university, except when specific references to the work of others are made. Except as stated in the text and acknowledgements, this dissertation is my own work and contains nothing that is the result of collaboration with others. I hereby certify that the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification at this or any other university. Except as stated in the text and acknowledgements, this dissertation is my own work and contains nothing that is the result of collaboration with others.

01/07/2022

Muhammad Qasim Raza

ACKNOWLEDGEMENT

By the grace of ALLAH ALMIGHTY, prayers of my parents and utmost support of my professor I was finally able to complete my thesis. I hope that the people who made this research possible are blessed and have great success in their life.

This project was impossible without the support of my advisor **Dr. Rowanda D. AHMED.** I am very thankful for her guidance and support.

ABSTARCT

Purpose of this research is to find and automate the process of finding frauds in Credit card frauds. Due to involvement of confidential information in the whole process this problem is very different in its nature. Our research will address this issue by detecting the frauds from previous credit card information's and their fraud ratio. In our work we selected a dataset which have the information of amount of transaction and time of the transaction. Based on these features of data we will find out the accuracy of fraud detections with different ML algorithm as well as we will also analyze the data based on time stamps.

Machine Learning proved to be a very effective system in prediction matters. Many problems have been solved by using Machine Learning techniques. In our research a new way was proposed which help in solving the problem of credit card frauds. Based on previous data of client and its activity with its credit card enable us to predict about the future transaction, that will help the system to avoid fraudulent transactions.

Research involving the hybrid method of machine learning algorithms and time stamps will be a method which can provide a solution to solve this problem.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In today’s world, technology is grooming day by day. Everything can be accessed and can be used through online. Payment’s method is increasing, within increase in payment methods frauds and hacks are also on their verge. You can do your shopping’s, order your food online or you can arrange your vacations. One must pay for the things purchased or any bookings and the process follows is to pay by your credit card. Users give their details online i.e., credit card details, personal details to do their transactions. This information can be stollen and can be used without user permission which is illegal and a fraud to avoid these types of transactions we come up with a method to overcome credit card frauds.

Machine learning and Artificial Intelligence proved to be very effective in these types of works. Many algorithms are trained on the previous data and applied on the new transactions which results almost perfectly with minimal errors. Machine learning is one of the most innovative technologies, which contains different type of algorithms. ML algorithms have widely been implemented in various fields to overcome the modern problems. ML algorithms work based on data behaviors, types, and values, through which it can predict about the futuristic data and flaws which are present in system.

ML algorithms like Decision tree, Random Forest, Support vector machines have been implemented in many cases to predict the behavior of data and results have been created through them which are very useful and effective.

Fraudulent Transactions

Fraudulent transactions are the transactions which are based on frauds or identity theft is involved in it. As in case of credit card fraudulent transactions, these transactions are transactions which are not being performed by the original person. Some thieves or hacker stole their information, based on that information they are performing transaction e.g.

Someone stole the credit card information which includes credit card number, expiry date, CVV number and address of the real owner of credit card. This is the information which can be passed to any website as a payment method. If this information is not being passed by the original owner, then this transaction be called as fraudulent transaction. If we summarize the fraudulent transaction, we will say that it is a transaction which is not being performed by original owner or without the consent of credit card owner.

This research thesis aims to investigate fraudulent transactions which are made without the user permission or a fake transaction using our ML algorithms. Often auto bots are used to match the numbers of credit cards and find a correct match for the numbers and PIN codes. Our research will focus on the details of the information and to find out whether transaction being done is a real transaction or a fake transaction. We will do our analysis on basis of data provided e.g., Transaction time, time elapsed during transaction, amount of transaction etc. and find out irregular activities during the transactions to find out our results

1.1 Research Objectives

The study aims to investigate about the fraudulent and non-fraudulent transaction. Our focus is to predict the behavior of previous data of transaction done in a year, based on which we can find out the fraudulent transaction in a real time and avoid these transactions for a better customer service and security. The main hypothesis of research work is drafted as follow:

"Credit Card Fraud Detection Using Machine Learning and their Comparison and analysis based on time series data."

In this work, we are using a data of transaction done in September 2013 by European customers. This data contains the information of each transaction done within specific 24 hours. Confidential information will be kept hidden for client privacy. Other than that, we have a time series and amount of each transaction, based on which we will compile our results.

CHAPTER 02

LITERATURE REVIEW

Credit card fraud is one of the modern problems, which is growing more and more today. As the problem rises many researchers are trying to find the solution of this modern-day problem. Different techniques were used to encounter this problem. Many techniques were implemented but the results are not much satisfied in the manual detection of the frauds. So, automated detection systems were used to solve the problems and stops these frauds. In automated automation techniques which shows much accuracy are algorithms of Data Mining and Machine or Deep Learning. Some of the previous work in last years will be discussed here how automated system results with different kinds of algorithms.

(Save et al., 2017) Suggest the methods of machine learning to encounter this problem. In their research they used a hybrid approach to detect the frauds in the ongoing transaction process. Decision tree and Luhn's and Hunt's algorithm are implemented on the previous data to get results. Using decision tree, the match the details of customer and the transaction details. If the details matched with the previous detail's transaction be considered as genuine. Based on this detail's clusters have been made of previous transactions. New transaction data is matched with the previous cluster if details match then it should be considered as real one and if its mismatch with the cluster it automatically marked as a fraudulent transaction. Implementing Luhn's and Hunt's algorithm to validate the card details with the transaction details as well as the address of the customer to validate transaction. Luhn's and Hunt's algorithms work based on Bayes algorithm in which the card number are added and sum of odd and even number is matched with total sum of card numbers. This technique makes it difficult for the fraudsters to do use wrong information for the transaction. Researchers propose that by using this technique frauds can be minimized and there is very low chance of fake alarms as in every transaction data is being matched with the details of the customer and with the cluster which has been made based on previous data.

(Vidyavardhaka College of Engineering et al., n.d.) Mrs. Vimala Devi. J, De. Kavitha have worked on this problem and find the classification algorithm method to be more feasible in this kind of problem. They use different algorithms of machine learning such as classification and regression algorithms which include decision tree, random forest, SVM etc. to generate their results for their classification algorithms and the compare all the results and the accuracy of each algorithm are compared to get better results.

Classification algorithms classifies between the genuine and fraudulent transaction based on the previous data of the transaction and predict the results of the upcoming transactions. Algorithms of Machine learning are used to classify in between the transactions, accuracy of the results is maintained. Other than that, behaviours and pattern of transactions are tracked based on the customer's detail's as how long have been he using this card, and which are the most transactions being done and what's the pattern of these transactions. By tracking this data and behaviours of these transaction it made it easier to track the activity of customer transaction and to classify in between the fraudulent transactions and the genuine transactions. Whenever there is any malicious activity occur it automatically differs from our classification, and it can be checked more or denied from the final approval. Noisy data and outliers' techniques are also used to remove the outliers from the data to get better results. It eradicates all the NAN values or outliers which does not fit with the general characteristics. Inconsistent data is also handled, so that it doesn't disturb our results and effects the accuracy of the results. All the results were visualized and concluded according to that.

An Analysis of different techniques are presented. Every method has its own procedure and methods to provide results. After the results were visualized their accuracy is compared with each other and Decision tree algorithm provides the results with more accuracy than the others as the data used is not in a good balance so, they suggest the oversampling of data to get better results.

(Xuan et al., 2018) Shiyang Xuan, Guanjun Liu**,** uses the techniques to classify data and make prediction on that. They suggested to use the previous data of transaction to classify them and make prediction according to that. By applying the random forest in two different ways can provide better results in predicting the fraudulent transactions. Behaviours classified from the previous transactions are very much helpful to find the fraudulent transactions as they mismatch with the classifier data. In this research they use random forest in two different ways to get better results in the accuracy.

Experiment is performed on the data of an Ecommerce company locate in China. People use different techniques to steal the data of the people credit cares main techniques which are used to get information is Trojan and phishing which allows fraudsters to steal the data of the cards of other peoples. Firstly, in this research all the data present in dataset has been analysed whether it's a normal data or a fraud transaction this data was later used to make predictions about the future transactions. Random forest is used to classify the data but this time they implemented random tree twice with the different base classifier and then compared the results and analysed them according to needs. Performance of the results of the random forests are compared on the same subset of the data. Results show that random forest can provide us good results if the data we are using is limited. Future transactions are predicted upon the classifiers and subset of the data any malicious activity can be found easily. CART-based random forest differentiates between the attributes of the decision tree to get accuracy.

Researchers concluded that random forest is providing good results but when the dataset or the data become large then if will affect in the results and accuracy. So, the random forest can be good on the limited amount of data. Imbalance in data also cause problems which should be dealt with to get better results. There is a lot to work on the data to get better results using random forest.

(Sadgali et al., 2020) Imane Sadgali, Nawal Sael**,** carried out their research in Casablanca to introduce an adaptive model for detecting fraudulent transaction done from credit card. They introduced the model which have high accuracy and can provide better performance. A multi-level framework is introduced which involves the data of the banking security aspects, customer data and history of the data of customer who is using the credit card. Combining all the data with some useful algorithms can avoid the credit card frauds and protect the customer from being robbed. This adaptive approach has high accuracy in predicting the fraud transactions.

In their research, a Hybrid model description is introduced which is divide into four different layers which are Authentication layer, Behavioural layer, Smart layer and the background processing layer which works differently on different columns of the data and produce results which will be helpful in stopping the fraud transactions.

Authentication layer is a usual security layer which will verify the customer data entered for the transaction process and based on the previous data submitted to the bank and the available in dataset it has been compared and allowed if it matches. After that behavioural layer come into action. Feature selection techniques are used as well as association rules of fuzzy logic are implemented to analyse the feature and the behaviour of previous transaction with the ongoing transaction. Algorithms used in this technique is FAR which is feasible and can provide good results. Smart layers distinguish between the normal and fraudulent transactions, using the transaction data, Support Vector machine algorithm SVM as well as GRU algorithms predicts the data with the classifications and provide results whether transaction being performed is a normal one or malicious. Background layer works parallel with all the layers which main purpose is to train the multiframework model and establish the rules of prediction for the future transaction whenever new transaction is popped all the process starts after verifying the data transaction will be allowed and background layer train itself whenever new transaction is performed which make the process strong and can deliver good results.

This adaptive model seems to be more significant as it is dealing with all the factors involved in the transactions and it is training itself with all the new incoming data. In this whole process we can judge human behaviours, transactional behaviours and, we can easily deal with the imbalanced data and can avoid data overflows. A good detection tool can be created based on this approach as it covers all the factors involved in performing the transaction.

(Dornadula & Geetha, 2019) Vaishnavi Nath, Geetha suggests that the details of the customers and their previous transactions are analysed according to their transaction they are clustered into different groups and based on these clusters different methods and techniques are used to mark and avoid the fraudulent transactions. Clustering is most common techniques which is widely used in machine learning methods to classify data of different types and applying algorithms. These clusters are very helpful in predicting the future data. Here they used different techniques along with the clusters to generate results up to highest accuracy.

Details of the customer and their transactions are classified into different categories, from where they purchased the most and how the normal transaction work and what's their difference with the fraudulent transactions. Sliding window strategy is applied on the different groups of clusters and behaviour of the transaction are analysed. In slide window algorithm a window is formed according to need and that window slide the data which it needs and leave the data which is not useful the slides data is used for making predictions or generating results. Create a new transaction group and line up the data in these groups and extract some dependent variable based on their behaviours during the transaction. In this condition, algorithm is trained in such a way so that it can take useful data which effects the transaction type and can predict results. Best results which are obtained through this technique are matched with the new transaction which can easily allow us to find the fraudulent transactions, it also trains our system in such a way that whenever new transaction is performed it can easily find the fraudulent transaction. By using different groups and the change in the values of data based on behaviour changes, all the activity is recorded how the variable change and when. Which will provide such a data or details, which will help us to predict the future transaction. If a new transaction outlies with the cluster groups or have a distinct change in comparison to the dependent variables can easily be marked as a fraudulent transaction. To balance the dataset SMOTE technique was used to get better results.

In this hybrid group approach, we can allow our system to group the data of customers and the transactions, by applying different classifiers on different groups we can record the changes into the data. Predicting based on this data will be more efficient as all the factors related to the transaction are taken in consideration as well as the imbalance data. Balanced set have a better performance with the classifiers.

(Mittal & Tyagi, 2019) Sangeet Mittal, Shivani Tyagi worked on both sides of the Machine learning and generate results. They use both the methods of supervised learning and unsupervised learning to overcome the problem of credit card frauds. Highly imbalanced data is use in the experiment to review the behaviour of the data and to record the problem of outliers that how outliers effect results and accuracy of the problem.

Unsupervised learning process is based on formation of clusters. Identical data rows which belong to same class are used to detect the outliers in the data. K-nearest neighbour algorithm is implemented on the data. Outliers detect the data which is outside the cluster which helps to find out the fraudulent transactions. Based on imbalanced dataset, data in the clusters for dependent variables help us to distinguish in between the normal the fraudulent transactions. Data handling which can be done in unsupervised learning provide very much good result from the supervised learning.

Results are also generated from the algorithms of the supervised learning, some of techniques and algorithms which are considered in the research are logistic regression, KNN, Decision tree, Deep Forest etc., Results were generated and then compared the accuracy of both the methods.

Supervised learning and unsupervised learning both methods are implemented and recorded the results. Which is explained in the research. Main things which are recorded during the comparison are unsupervised handles the data skew better than the supervised learning. NAN’s value in the result table is also resampled by using different variables, classifiers or by adding new features in the data. KNN in unsupervised learning models have the best results with the accuracy of 0.99 approximately.

(Patil et al., 2015) Snehal Patil, Harshada S introduces an advanced methodology, which uses a cost-effective approach which focus on the classification method. The problem rises with the classification are addressed in the paper. By splitting the node on different terminal, the method handles the misclassification of data and compare the data with traditional classification methods. To achieve this goal, they introduced a query method to protect from the fraudsters. Queries were asked on each stage of transaction which minimize the threats of fraudulent transactions. Decision tree is implemented for the classification with other algorithms.

Machine learning known method of classification have huge impact in solving this problem but as in different research result generated varies from data to data and method to method. In large amount of data, it became almost impossible to classify data and choose its dependent variables and attributes to consider for the clusters. In this research, decision tree will make the decide on each node and transfer the controller to next node. Decisions were made based on attributes selected, details of the transaction and behaviours. Previous actions and behaviour of the transaction are keep tracked and whenever new transaction is made decision will take place based on all the considered elements. Queries will be asked on each node which will be compared to the previous data. Asking queries on each node will make it difficult for the fraudulent to crack the decision tree, as if queries were unmatched, false alarm will be raised. Decision tree in this case is firstly trained on the data present in the data set, after that whenever transaction is being performed it will automatically train the algorithm. If record is maintained the algorithm become more and more powerful to perform decisions. In this whole process both the methods are kept in consideration.

This method allows the user to avoid fraudulent transactions easily as it involves both the techniques of authorization, and it is predicting on each node of the decision tree which seems to be feasible to avoid frauds. This system seems to be more accountable as its automated behaviour doesn’t let anyone to do a transaction with a fake identity or some phishing methods. Monitoring is one of the main elements that can be done with this procedure by the banks and big organizations who want to track the record as the system allows to record the queries asked and their answers which will establish a strong decision tree that results in predicting up to high accuracy.

(Yu et al., 2020) Xia Ohan Yu, Yiyang Dong**,** taking into the consideration of the previous research and the problem rises in this kind of research. Xia Ohan proposed an automated system to prevent the credit card frauds an in manual system it become more difficult to find the fraudulent transaction as every time when the transaction is performed, system cannot do a whole process. Beside machine learning methods and algorithms, they introduced a deep learning mechanism to solve this problem. Data skew problems which rises are also keep in view to avoid the miscalculations in the results.

Whole process is divided into different section. Firstly, about the previous studies and deep learning methods. In second section five layers method is introduced which explain all the layers and their importance in process. After that main data processing techniques are used to organize data. In the last hardware capabilities are discussed as well as the experiments on different models and accuracy like Auc Roc. Deep learning method works like human being brain as neural networks are trained in a way to predict things and make decision and keep track of data. Techniques of log transform and focal loss are used to make the previous data more useful and effective during the process. Data skew problem arises in the previous research which disturb the whole data set, in this case using log transform data is skewed to avoid distortions in the data as well as to make data in more straight manner other data processing techniques are used like hot encoding and standard normalization because whenever you are dealing with this type of data, there are many attribute which are to be addressed during the classifications or using the predictions algorithms, as these data mis transformations disturbs the results a lot. Variables involved in the process are sorted and distinguished between dependent and independent variables. NAN values are removed from the data, if NAN values appear on the dependent variable, then it’s handled by feature engineering technique. GPU card was suggested and is used because to compute a large amount of data if the hardware is of less capacity, it will disturb the results and accuracy of results. After all these data processing methods deep, neural network predicts about the transaction results and train itself according to the results and every time when there is new incoming data it will be handled through data processing system, which makes the data in a good manner. Results generated from that data have high accuracy.

The whole process of predicting the credit card frauds and the problem arises, related to the previous research are briefly explained in the article by keeping in view all the components involved like data processing which often disturb the accuracy of result, Hardware compatibility to process large amount of data and removing all the garbage data can improve results and accuracy. All the factors are discussed which can have impact on the results.

(Mahmoudi & Duman, 2015) Nader Mahmoudi, Ekrem Duman, worked on the modern Fisher discrimination technique to overcome the problem of credit card frauds. Supervised learning algorithm solve this problem up to a good extent but to achieve this algorithm become too complex to handle. Similar function can be used which are less complex and more efficient in terms of results and accuracy, Linear discrimination function is one of them. Its classifiers are less complex. In this technique Fischer discrimination modified function is designed in such a wat that it can handle the problem of credit card frauds.

When we observe the work done through the supervised learning, we came to a result that sometimes it overfits the data which will cause problem in predicting good results. In modern Fischer technique past data is analysed and based on which important instance that occurs during the data analyzation is considered all other data is neglected and based on these instances' predictions are made. In linear discrimination analysis input of data is divided into different regions of decisions. Weight of each instance and the decision made on each instance is recorded to have good results. Data which is used during the transactions. Dataset is divided into three different parts to apply a three-fold cross validation scheme. After that, different algorithms were implemented on the selected data which are Decision tree, classification, and regression tree.

Results of the experiment are analysed and visualized in a tabular form to obtain a comprehensive result from the experiment. Other than that same dataset is also analysed by the traditional methods of decision tree, Fischer discrimination and Artificial neural networks etc. to get the results and compare it with the modified techniques. In comparison to FDA the modified FDA seems to be more profitable. Accuracy we get from the Modified FDA is more than all the other algorithms. Research suggests that by applying this method on a real-life data we can obtain results with minimize the fraud transaction up to 0.001%. Modification techniques applied during the experiment in which weight of each instance and variable is defined is an important consideration during formulation of results. They proposed that many other linear functions can also be implemented to achieve good accurate results and to avoid false alarm which create dissatisfaction among the credit card user which can cost more problems and money.

(West & Bhattacharya, 2016)A comprehensive review of the previous work on intelligent frauds was done by Jarrod West & Maumita Bhattacharya**,** they analysed that the manual techniques which are used to solve this kind of problems cost us time and accuracy in terms of results. In modern day world, there are many techniques which can be used and have an automated response to the things which will be more helpful in detecting credit card frauds, in this comprehensive review, all the methods of data mining are kept in consideration primarily focusing on the computational intelligence.

Directly using the techniques and algorithms of machine learning or data mining are not enough to predict about the frauds happening in a new transaction and cannot be caught in a real time transaction. Useful framework applied can reduce fraudulent transaction. Hybrid techniques by using different framework on different levels can reduce fraudulent transactions Reviewing 2% of the transaction can help us and could possibly reduce 1% of the fraudulent transaction from whole data. Statistical and computational techniques are used to analyze the previous data and generate results based on which it will help algorithm to train itself and avoid further fraudulent transactions. Large number of techniques exist to avoid and eradicate the problem of frauds. Main techniques which are used in the previous period are:

- Bayesian Belief Network
- Neural Networks
- Logistic Models
- Support Vector Machine (SVM)
- Text Mining
- Decision and Forest Tree
- Artificial immune system
- Self-organizing Map etc.

All the above method are studied during the research all of them have their own results and accuracy which differs with each other up to some extent, dependent variable consideration is the main cause of difference in results and classification method effect the results a lot. But there are still much more required to be done to achieve this target So that fraudster cannot be allowed to do fraudulent transactions.

Further research in different fraud methods can lead to a general framework which can be implemented on any kind of system to make it secure and hybrid type which is not much dependent on the behavioral activities. Every fraud action in different organizations have different activity to address it like that within every organization will be very cost effective and to solve problem become impossible. New adaptive techniques like neural network and support vector mechanism are highly effective to the evolving methods of frauds.

(Shirodkar et al., n.d.) Nikita Shirodkar, Partikesh Kumar considers artificial neural networks to be more effective in finding the credit card frauds because this system trains itself based on the neural networks and the history of customers. These behavioural activities seem to be useful in training the algorithm's new method of genetic algorithm is introduced in this research to record the activity and behaviours of the customer which can later make the algorithm stronger and more let it to predict up to 100% accuracy. With the detection of frauds this research also focusses on minimizing the risk of identifying legal transaction as illegal. To achieve this artificial neural network are used along with the genetic algorithms.

Artificial neural networks are a neural network which contains different algorithms. It works like a mechanism of digital logic gate and make decisions based on OR functions. If both inputs are ON it will give us the result ON and if the one of them is ON then the results generated are also ON and if any case both is OFF then the result will be negative, based on this methodology it will decide. Customer behaviours and actions are considered in this to make good results. Genetic algorithm used in this mechanism record and study the previous activity of the customer who is performing the action which make it easy for the neural network to decide as it trained itself in the basis of previous data which give a lot of information about the customer and the activity of customer. A Random initial population is created, on each step new population is created based on the previous one. Neural networks recognize similar patterns, predictive value based on these patterns. But in this hybrid approach using ANN with the genetic algorithm can produce results up to 99.83% accuracy.

An Artificial neural network consists of a group which relate to each other. These groups are formed of different artificial neurons. Neural networks recognize similar pattern, predict values or event based upon the associative memory of the pattern. There are different ways to detect the fraudulent transaction, we concluded that every method and technique has its own advantages and disadvantages, there is still a window of improvement in fraud detection techniques.

(**Zeager et al., 2017)** Mary Zaeger, Aksheetha Sridhar, have conducted their research in University of Virginia and introduce a method of Adversarial Learning in the problems of frauds and its detection systems. Many different methods were introduced which can overcome this problem, but we explain the methods and the thing which motivates the fraudster on this activity A game theoretical learning system allow us to model the fraudsters techniques, strategies, and approach to a fake transaction.

For a fraud detection mechanism, a logistic regression technique is used to identify the best approach of the fraudsters. Which we can find by studying the previous data and the activities, behaviour of the transactions. Which will allow us to improve our classifiers. Datasets was split into sub datasets into training, testing, and validation sets. Testing data sets will test the model to check the confusion occurring during the transactions. Using a gaussian mixture model we will distribute our data population and distinguish it with the other data present in the data set.

SMOTE technique is used to handle the oversampled data present in the datasets and on the basis of that best frauds are generated from the datasets. Modelling and splitting them give us transaction differences on which data can be identified. Basis of the above discussed we summarize the technique like following methods are involve during the experiment.

- Dataset splitting

- Testing Models
- Gaussian Mixture Model
- SMOTE Algorithm
- Modelling of Data

Ten rounds of the data were chosen, and the experiments were performed on each round and results were generated and visualized. ROC curves tested on the validation sets, activity of the curve was recorded and analysed the difference between the curves. Modelling adversaries are one of the possible strategies to retrain models that proved to be in good performance then the static model. Separation between AUC score of the adversarial learning model and static fraud model increase although the difference may seem small. Use of GMM in determining a best strategy is a useful way in finding the new transaction optimization. SMOTE results are proved to be very helpful in the technique. Overall, these two contributions provided tools able to mimic an adversary learning and thought process. Further research work may prove to be very effective with addition to the new elements, could possibly improve the results. New methods can also be used with the existing method to get better performance.

Many other researchers work on this problem to find solutions using Machine Learning some of them are mentioned in tabular form.

Authors	**Paper**	**Year**	**Method Used**
Joy Iong-Zong Chen, Kong-Long Lai(**Chen & Lai, 2021)**	Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert	2021	Deep Convolution Neural Network (DCNN)
Fabrizio Carcillo, Yann-Aël Le Borgne,	Combining unsupervised and	2021	Assumption of fraudulent patterns and predict about the

Olivier Caelen, Yacine Kessaci **(Carcillo et al., 2021)**	supervised learning in credit card fraud detection		future transaction using Machine Learning Algorithms
Andrea Dal Pozzolo, Olivier Caelen, Serge Waterschoot **(Dal Pozzolo et al., 2014)**	Learned lessons in credit card fraud detection from a practitioner perspective	2014	Make prediction by considering the abnormalities of data like unbalanced ness, non-stationarity and assessment. Using Machine Learning
Tzu-Hsuan Lin, Jehn-Ruey Jiang**(Lin & Jiang, 2021)**	Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest	2021	Data resampling techniques SMOTE, adaptive synthetic and TOMEK link are applied on dataset for predictions
Asha RB, Suresh Kumar KR **(RB & KR, 2021)**	Credit card fraud detection using artificial neural network	2021	Multiple ML algorithms in parallel to differentiation of previous supervised and Deep Learning techniques
D. Tanouz, R Raja Subramanian, D. Eswar and co.**(Tanouz et al., 2021)**	Credit Card Fraud Detection Using Machine Learning	2021	Machine Learning techniques for prediction and focused on precision and accuracy of results using f1 score, Roc-auc score and confusion matrix

CHAPTER 03

METHODOLOGY

3.1 Theoretical Framework

3.1.1 Data Mining

Data mining is a method that is used as a rear for KDD. One could even say that data mining is often called a piece of the KDD- process, which entails the application of approach and algorithm is picked out and implemented on the data set. Hence, it's an important thing when it comes to discovery in databases process. Data mining could take any way of data and implying analysis algorithms so that you can show patterns or ways within the data set and utilizing these structures to categorize the data into distinct classes. It comprises several disquisition fields analogous as database systems, statistics, and ways of recognition. Data booby-trapping tasks are chosen on the information that algorithms have about the classes in the set.

Managed knowledge involves all the works in which the algorithm has access to affair and input values. Only values that are allowed to be used in algorithms are the input values that are defined as the external information, analogous particularity are Meta data and values, on the other hand affair values are the specific labels of the class particularly. It shows that structure of the data is commonly used, and uses of those programs are to assign new data to the right classes.

In distinction to managed and known knowledge, unmanaged knowledge contains those tasks that have no access to affair values and that is why they try to find structures inside the data by creating classes on their own. Because the proposed task is a double type of error, concerning the two known classes' professional paraphrase and automated paraphrase, this thing will concentrate on supervised knowledge styles. There are two things that make the core of data mining: discovery and verification. The two things work opposite each other when going through a thesis. On one hand, verification focuses on proving the thesis of the user and on the other hand, discovery challenges this thesis and finds new loopholes. Discovery is further divided into description and prophecy.

Description entails finding the right patterns so that you can articulate the data in a comprehensive manner and use an accessible format. Prophecy involves trying to figure out the future problems of data gauging from the patterns. You can further divide group prophecy into two groups: type and regression tasks. Type tasks involve fixed labels, and every data record contains a fixed label as its class particularity value. On the other hand, regression tasks contain continuous values as affair. Regression tasks focus on the algorithms that aim to prophesize if a given specialized documentation has been paraphrased properly, formally, or automatically. Thus, the issue entails two fixed labels such as professional paraphrase and automated paraphrase. These fixed labels belong to the two sectors of data mining: discovery and prophecy.

3.1.2 Machine Learning

In the info mining step, it is necessary to decide which approach to take in order to cope better with the task. This approach can be decided based on some victimization learning methods. One of the crucial points that distinguishes a human from a laptop is that a human manually solves a problem while a laptop does not. Humans tackle problems by making appropriate corrections and by finding new approaches to acknowledge the problem. On the other hand, computer programs don’t have the ability to deduce the results of their tasks and are not able to alter their way.

Machine learning aims to improve this shortcoming of machines by creating laptop programs that are able to find the flaws in their tasks and improve their performances by using their previous experiences and available knowledge. In 1952, A. Samuel became the first person to create a self-learning program that became higher at enjoying the sport checkers with the number of games. In 1967, the first pattern recognition program found patterns in data by comparing the new data to popular data and finding similarities between them.

Since 1990s, machine learning has been used in data fields of data processing, accommodative package systems, and as text associated learning fields. For example: A bug that collects data regarding the purchasers of an e-commerce business and makes higher customized advertisements out of those items data that has ability to amass new ideas and comes closer to AI.

Moreover, machine learning systems are commonly categorized by judging their learning strategies that typically known by the quantity of logical thinking a PC can do.

- **Rote Learning** describes the strategy that every ancient laptop program uses. These ancient laptop programs do not perform any reasonably logical thinking and the information is directly informed by the programmer, since the applying is unable get any results or demodulations from the available information.

- **Learning from Instruction** comprises every program that can convert information from the input language to an enclosed language. The information required to make this change possible is provided by the programmer himself, but it is still easier than learning without having proper idea about it.

- In distinction to Learning from Instruction, **Learning by Analogy** aims to advance new skills that are quite like the skills that are already there. This is made possible by acting transformations on famous information. It opens doors to new functionalities, that were not known to the first bug and hence requires a lot of inference.

- **Learning from Examples** is the popular method as it provides versatility and allows laptop programs to develop altogether new skills that are not similar to the existing skills. It also allows innovation and construction of not known structures and patterns in information. Practically learning may be a way that is usually used in classification and data processing tasks to prophesize the category label of latest data entries which they can deduce by learning from famous examples. While this is underway, the planned analysis queries can be tackled with algorithms and strategies that belong to the present category.

Here are some of the common machine learning techniques:

3.1.3 Decision Tree

Decision Tree is a way that concentrates on things which can be understudied, simply apprehensible illustration type and is very commonly learning ways. Decision Trees use digital audiotape sets that carries with it attribute vectors, that successively has a collection of classification attributes explaining the vector and a category attribute assignment the information entry to a precise class. A Decision Tree is constructed by repeatedly ripping the data attack the attribute that separates the data moreover as doable into the various existing categories till a certain stop criterion is reached. The illustration type allows users to induce a fast summary of the data, since call Trees will simply be visualized in a very tree structured format that is simple to know for humans.

Among the primary algorithms regarding call Tree coaching were the repetitive Dichotomiser three (ID3) and its successor the C4.5 algorithm, each developed by Ross Quinlan in 1986 and 1993. These algorithms laid down the basis of innovation and shaped the idea for several additional developments.

Decision trees are directed trees, which are used as a choice support tool. They represent decision rules and illustrate consecutive decisions. In Decision Trees, nodes will be separated into the foundation node, inner nodes, Associate in Nursing finish nodes, conjointly known as leaves. The foundation node represents the beginning of the choice support method and has no incoming edges. The inner nodes have precisely one incoming edge and have a minimum of two outgoing edges. They contain a check supported an attribute of the information set.

For example, such a test may ask: “Is the client older than thirty-five for the attribute age?” Leaf nodes carries with it a solution to the decision problem that is mostly delineate by a category prediction. As Associate in Nursing example, a choice drawback may be the question whether a client in a web look can build a procurement or not, with the category predictions being affirmative and no. Leaf nodes don't have any outgoing and precisely one incoming edge. Edges represent the choice taken from the previous node. Given a node n, all following nodes that are separated by exactly one edge to n are called youngsters of n, whereas n is named parent of all its kid nodes.

3.1.4 Artificial Neural Networks

Neural networks and deep gaining knowledge of presently offer several the maximum dependable picture recognition, speech recognition, and herbal language processing answers available. However, it wasn’t constantly that way. One of the earliest and best coaching philosophies for synthetic intelligence become marginally successful. It advised that loading the most quantity of facts right into an effective pc after which maximizing the instructions used to recognize the statistics need to deliver that pc the cap potential to “think.” It became an easy concept, and it became genuinely well worth a try.

This type of computer training is based on rigid, pre-existing rules that are meticulously drafted by engineers (if it happens, answer this way; if it happens, answer it). this way). It is not thinking. It's like a habitual, uncontrolled reaction. An artificial neural network is a computer structure loosely modelled on the neural network of the human mind. Although they are no longer effective, they work relatively well. The mind learns from what it goes through, and so do these structures. The artificial neural network analyses tasks using evaluation patterns.

For example, even as getting to know picture recognition, neural networks in schooling could learn how to perceive photographs containing puppies with the aid of using analysing pattern photographs which have been tagged with “canine” or “no canine” labels after which use the ones effects to find and perceive puppies in new photographs. These neural networks begin from zero, without an information approximately canine traits, including tails, ears, and fur. The structures broaden their very own know-how of applicable traits primarily based totally at the getting to know cloth being processed. (The human mind doesn’t begin from zero. Room for a bit evolution?)

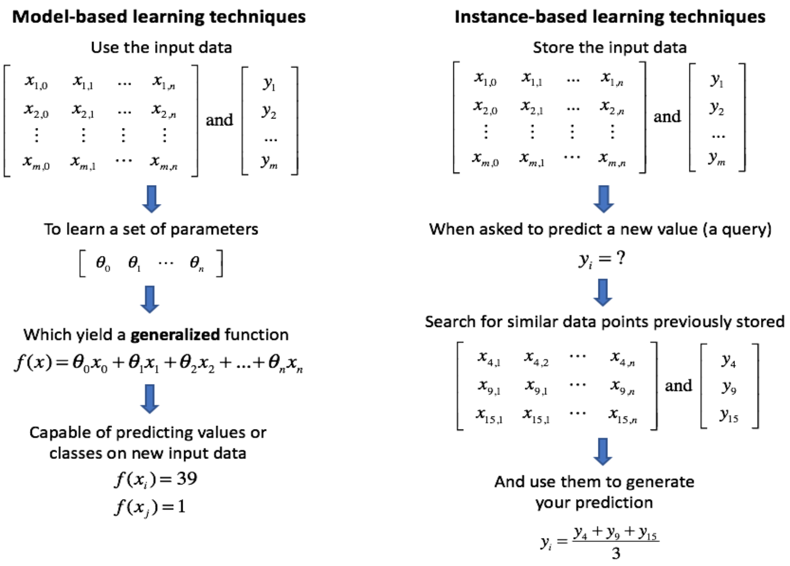
One giant gain of neural networks is their cap capacity to examine in nonlinear ways. This manner they`ve the cap capacity to perceive competencies in a photo that are not obvious. For example, whilst identifying oranges, neural networks might also additionally need to

identify some in direct daylight hours and others with within the coloration on a tree, or they'll spot a bowl of oranges on a shelf in an image with a one-of-a-type subject. This cap capacity is the result of an activation layer designed to attention at the useful facts with within the identification procedure.

A synthetic neural community makes use of a group of linked nodes known as synthetic neurons – a simplistic imitation of organic neurons. The connections are variations of synapses and function while a synthetic neuron transmits a sign from one to another. The synthetic neuron that gets the sign can procedure it after which sign synthetic neurons linked to it.

3.1.5 K-Nearest Neighbours (KNN)

Many Machine Learning strategies contain constructing a version this is able to represent the information after which locating the foremost parameters for the version to decrease error. K-nearest neighbours, however, is an instance of instance-primarily based totally studying in which we as a substitute clearly keep the schooling information and it is used for making predictions.



Model-based learning techniques
Use the input data
$x_{1,0}$ $x_{1,1}$ ⋯ $x_{1,n}$
$x_{2,0}$ $x_{2,1}$ ⋯ $x_{2,n}$
$x_{m,0}$ $x_{m,1}$ ⋯ $x_{m,n}$
and
y_1 y_2 … y_m
To learn a set of parameters
$[\theta_0 \ \theta_1 \ \cdots \ \theta_n]$
Which yield a **generalized** function
$f(x)=\theta_0x_0+\theta_1x_1+\theta_2x_2+\ldots+\theta_nx_n$
Capable of predicting values or classes on new input data
$f(x_i)=39$
$f(x_j)=1$
Instance-based learning techniques
Store the input data
$x_{1,0}$ $x_{1,1}$ ⋯ $x_{1,n}$
$x_{2,0}$ $x_{2,1}$ ⋯ $x_{2,n}$
$x_{m,0}$ $x_{m,1}$ ⋯ $x_{m,n}$
and
y_1 y_2 … y_m
When asked to predict a new value (a query)
$y_i=?$
Search for similar data points previously stored
$x_{4,1}$ $x_{4,2}$ ⋯ $x_{4,n}$
$x_{9,1}$ $x_{9,1}$ ⋯ $x_{9,n}$
$x_{15,1}$ $x_{15,1}$ ⋯ $x_{15,n}$
and
y_4 y_9 y_{15}
And use them to generate your prediction
$y_i=\frac{y_4+y_9+y_{15}}{3}$

FIGURE 1 REPRESENTATION OF KNN ALGORITHM

In general, instance-primarily totally based strategies inclusive of k-nearest neighbours are lazy inexperienced persons, in comparison to version-primarily totally based strategies which might be keen inexperienced persons. A lazy technique will only "analyse" from the information (to make a prediction) while a brand-new question is made at the same time as a keen learner will analyse from the information proper away and construct a generalized version able to predicting any value. Thus, someone who is procrastinator and inexperienced person is rapid to educate and very slow to question, at the same time as keen inexperienced persons are slower to educate however could predict new things very fast.

K-nearest neighbours are primarily based totally on the belief of your information, that close by factors have comparable values. Thus, what is an awesome manner to be expecting a brand-new information point? Just search for comparable observations to your schooling information and expand your quality guess. Particularly, use the k-nearest observations to formulate your prediction. For categorization, this prediction is a few degrees of the common value of the encircling pals. For classification, this prediction is a result of a few votes casting mechanism of the encircling neighbors.

3.1.6 Logistic Regression

Logistic regression is a supervised studying set of policies notably used for classification. We use logistic regression to anticipate a binary outcome (1/ 0, Yes/ No, True/False) given a tough and rapid of independent variables. To represent binary/explicit outcomes, we use dummy variables. Logistic regression makes use of an equation as its representation, very just like linear regression. In fact, logistic regression isn`t lots one-of-a-kind from linear regression, besides we match a sigmoid characteristic withinside the linear regression equation. The logistic function, moreover, referred to as the sigmoid function have become superior via statisticians to give an explanation for homes of population growth in ecology, developing short and maxing out at the carrying capacity of the environment. It`s an S-fashioned curve that would take any real-valued amount and map it proper right into a fee amongst 0 and 1, but in no way exactly on the one`s limits.

1 / (1 + e^-value)

Where e is the bottom of the herbal logarithms (Euler's quantity or the EXP () characteristic for your spreadsheet) and price is the real numerical price which you need to transform. Below is a plot of the numbers among -five and five converted into the variety zero and 1 the use of the logistic characteristic.

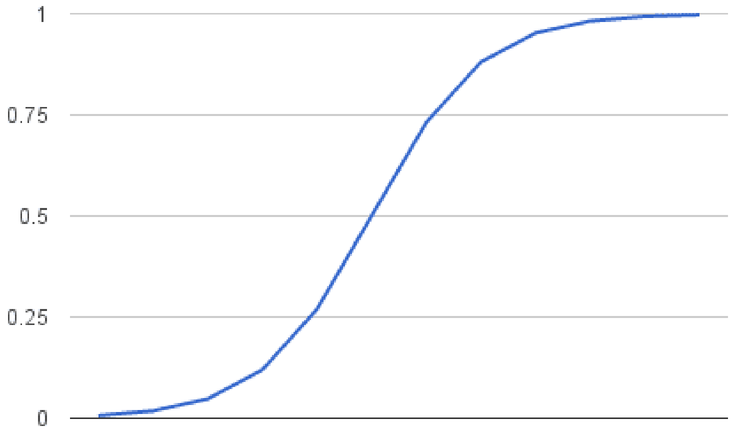


Figure 2 Logistic Regression Graph

3.1.7 Random Forest

Random Forest is a powerful machine learning algorithm that could be used for different tasks including regression and classification. This is an aggregate method, which means that a random forest model consists of many small decision trees called assessors, each of them give its own estimate. Random forest models combine assessors' estimates to produce very correct estimates.

The problem with standard decision tree classifiers is that they tend to over fit the training set. Random forest ensembles allow random forests to compensate for this and collect nicely into invisible data, including missing data. Random forests are also suitable for working with large arrays of high-dimensional data and heterogeneous feature types (e.g., when one column is categorical, and the other is numeric).

Usually, random forests are very suitable for classification problems, but not the best option for regression problems. Unlike linear regression, each forest regress is not capable of making predictions beyond the reach of its training data.

Decision tree in its non-ensemble form has existed in some forms since around 1950s, though no one has been able to name a researcher as the inventor and may have been discovered more than once. Earliest articles to mention or give any reference about decision trees was a 1959 article by a British statistician on algorithms for constructing decision trees for the classification of biological organisms.

In the coming decades, decision trees were gradually refined by the statistical community. The most famous algorithms are C4.5, ID3 and CART. But neither approach involves the ensemble method, nor the decision tree is still prone to over-assembly, as the Boston case is mentioned above as for an example.

In 1995, Hong Kong-based American researcher Ho Ting-Kam made the first random forest algorithm and during this time he was working at Bell Labs in New Jersey. It uses the random subspace method to reduce correlations between ratters, exposing each ratter to a subset of the full feature set, but still using the full training set to train each ratter.

In 2001, American computer scientist Leo Braiman (one of the actual inventors of the CART algorithm) refined the Ho ensemble algorithm to introduce boot packing or aggregation. Braiman’s idea was to take a sub-sampling of the original training kit to train each ratter.

3.1.7 Support Vector Machine (SVM)

Supporting Vector Machines belongs to the field of controlled learning methods and hence need to be labelled, known data to classify new invisible data. The main method or way to classification data, starting with trying to create a function that separates the data points into their individual points. Label it with (a) as few errors as possible or (b) as wide a field as possible. This is because larger space results in less next to the split feature Errors because the labels are different from each other. Therefore, the field around the split function is used as additional parameters to assess the quality of the separation.

Formally, Support Vector Machines produce one or more than one hyper plane in Associating in n-dimensional space. The primary tries within the method of cacophonous the info is always, to undertake to linearly separate the data into the corresponding labels. Example given, for a task of predicting the chance of a sale of a client in an internet search uses a knowledge set with n data points, wherever every datum consists of a label y ∈ and an attribute vector ~x containing the data values for that very session. The Support Vector Machine currently tries to find a function that separates all data points (~x, y) with y = affirmative from all data points of the shape (~x, y) with y = no. If the info is totally divisible in linear fashion, the ensuing function will be accustomed classify future events. Steinwart and Christmann mention 2 main concerns, regarding this approach.

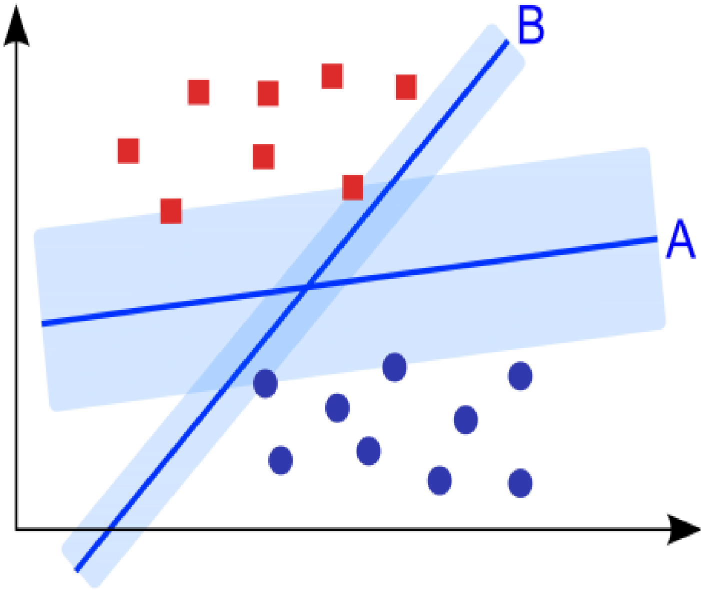


FIGURE 3 SUPPORT VECTOR MACHINE (SVM) GRAPH

The data may not be well linearly divisible or not linearly divisible at each, which is frequently the case for real world data. In the illustration given above, it can be for illustration that two guests are acting fully analogous in an online shop, with only one of them making a purchase. This would affect in not divisible data, due to the same trait vector having different markers. The alternate concern is the possibility of over fitting the SVM. To avoid this, data has to be pre-processed to identify noise and accept some misclassifications. Else, the delicacy values of the SVM will be defective and affect in further incorrect bracket for unborn events.

3.1.8 XG Boost Classifier

XGBoost is associate algorithm. Also, it's recently been dominating applied machine learning. XGBoost is an implementation of gradient boosted call trees. Although, it had been designed for speed and performance. Basically, it's a sort of package library. That you just will transfer and install on your machine. Then got to access it from a range of interfaces.

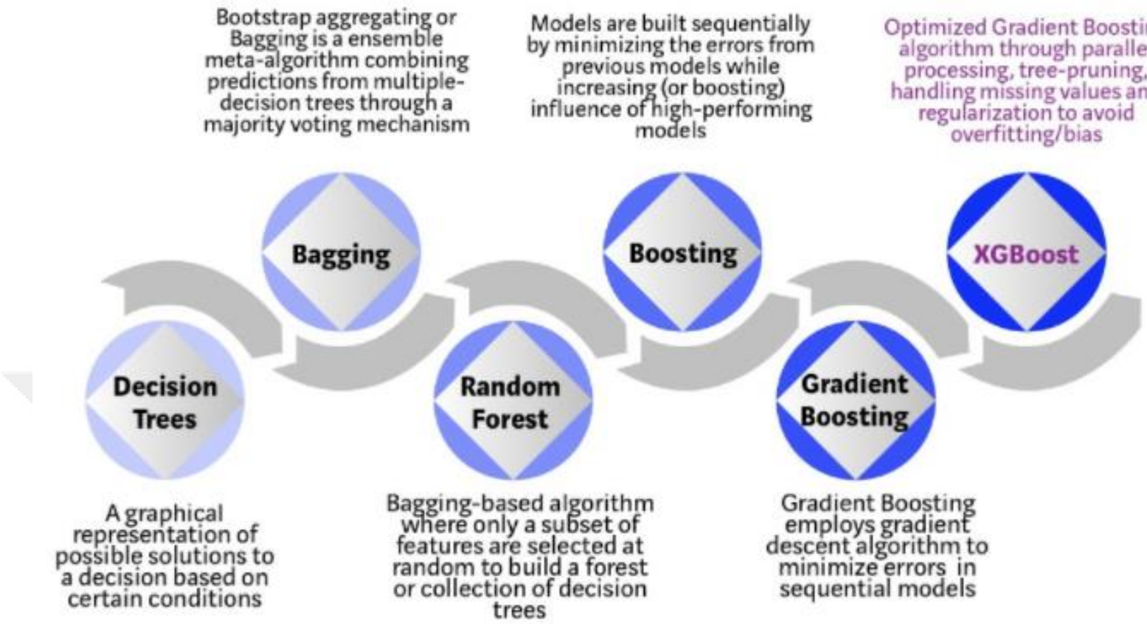

Bootstrap aggregating or Bagging is a ensemble meta-algorithm combining predictions from multiple-decision trees through a majority voting mechanism
Models are built sequentially by minimizing the errors from previous models while increasing (or boosting) influence of high-performing models
Optimized Gradient Boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting/bias
Bagging
Boosting
XGBoost
Decision Trees
Random Forest
Gradient Boosting
A graphical representation of possible solutions to a decision based on certain conditions
Bagging-based algorithm where only a subset of features are selected at random to build a forest or collection of decision trees
Gradient Boosting employs gradient descent algorithm to minimize errors in sequential models

FIGURE 4 EVOLUTION OF XGBOOST ALGORITHM FROM DECISION TREE (KRISHNAN ET AL., 2021)

XGBoost mechanically delivers feature relevancy evaluations supported a trained prognosticative model. When constructing a boosting tree, it retrieves feature importance ratings for every attribute. The feature importance provides a score that reflects however helpful each feature was within the model' building of the improved call trees. In scikit-learn, feature importance scores are often accustomed choose features. Will be accomplished by utilizing the SelectFromModel category, which settle for a model and should flip a dataset into a set with chosen features. This class can accept a pre-trained model, one that has been trained on the total coaching dataset. It can then decide that options to pick by applying a threshold. This threshold is used after you use the transform() technique on the SelectFromModel instance to select constant features on the coaching and check datasets consistently. By mistreatment scikit-permutation every feature within the model is shuffled at random and therefore the distinction in performance will be computed.

The algorithm learns the way to handle missing values by treating the non-presence as a missing value. Once the non-presence corresponds to a user nominative value, the algorithm may be applied by enumerating solely consistent solutions. All scantiness patterns are handled uniformly by XGBoost. This sparsely is employed to form computation quality proportional to the amount of non-missing parts within the input.

3.2 Modelling Credit Card Fraud Detection Using Machine Learning

3.2.1 Collection Data

The dataset used to generate the models comprised 284,807 transactions. This means that each row in the dataset referred to a unique transaction which was made on specific time and amount.   The complete dataset is a subset of data that was scraped from the freely available datasets on Kaggle.com.

(https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud).

The dataset contains transactions made by customers in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

When we study about the details of a problem which contains confidential information. It became more difficult to analyze the variables and the dependency of a data on each other. In case of credit card frauds. Dataset we found also contains the values which are private and haven’t shared due to confidentiality issues.

Dataset we are using contain 284,807 rows and 30 columns. Each row specifies about the detail of one single transaction and the values which are being associated with it. When we studied about the columns, we have distinguished our independent and dependent features based on that we will perform our experiment. For each transaction we have values present in different columns. These values are labeled as time of transaction, amount of transaction and there are some confidential values which are unknown to us marked as V1, V2, V3……… V28. Our independent variables are time and amount of the transaction other variable are dependent and their values are changing in each transaction. We will consider these features as confidential and dependent, to record the changes in these features we will perform some test to find out whether these features are important in result prediction or not.

These variables contain just mathematical info factors which are the consequence of a PCA change. Sadly, because of privacy issues, we can't give the first highlights and more foundation data about the information. Highlights V1, V2, … V28 are the foremost parts acquired with PCA, the main elements which have not been changed with PCA are 'Time' and 'Sum'. Highlight 'Time' contains the seconds slipped by between every exchange and the principal exchange in the dataset. The element 'Sum' is the exchange Amount, this component can be utilized for instance dependent expense touchy learning. Include 'Class' is the reaction variable, and it takes esteem 1 in the event of misrepresentation and 0 in any case.

3.2.2 Analyzing the Dataset

Data Analysis is characterized as a course of cleaning, changing, and demonstrating information to find valuable data for experiments. The reason for Data Analysis is to separate valuable data from information and taking the choice considering the information. Dataset we are using contains a lot of confidential data, we must identify the variables based on which we can make our machine learning algorithms to give us accuracy on predictions. Confidential data is also present in dataset, but it was not exposed, it was just highlighted in a numerical form, we will also consider that data and check the distortion of data related with the fraudulent and normal transactions, through which we can understand whether the details of customers effect the results or accuracy of results. Firstly, we will identify our dependent and independent variables based on which experiments are performed, **Time elapsed during each transaction** and **number of transactions** are the independent variables we are going to test their behaviour and make predictions according to change in behaviours. Data present in each row shows the second elapsed during transaction has been formatted in term of the data and time, Data has been analysed in three different ways which are as following

1. Descriptive Analysis
2. Visualization of Dataset
3. Graph plotting

Each of the above methods has been performed on data to check out the behaviour of data and the indifferences between the data.

3.2.3 Machine Learning Methods

The most widely used machine learning algorithms in the previous studies are linear modelling, Random Forest, and Decision Tree algorithms In this subsection, Decision Trees (DT), K-Nearest Neighbor, Logistic Regression, Support Vector Machine, Random Forest (RF) and XG Boost Classifier will be explained as the main techniques used in our experiments.

Framework Used for Experiments:

We have performed all the experiment in Jupyter Notebook. Python is the programming language we have used to derive results for our experiments. We have used this framework because it provides a very interactive environment and tools to perform actions. In python we worked with Pandas which is built on python libraries. Matplotlib is used for the visualization and graphs plotting whereas NumPy is used for the mathematical operations.

We then compared results of different algorithms and find out the best of them having at most accuracy in finding fraudulent transactions.

- **Decision Tree**

A DT is a tree-like model of decisions and their possible outcomes. It includes the concepts of nodes, branches, and leaves (or terminal nodes). A node is a point where a decision must be made, extending into branches. Each branch represents a possible alternative or course of action. The main advantage of using DT is their high interpretability, however, usage of only one Decision Tree is not sufficient.

The entropy measures the homogeneity of a sample (output variable) in the tree using the frequency table of each attribute (features). Information gain is based on the decrease in entropy after the dataset is split into the attributes.

Algorithm is tested and trained on basis of our variables (Time elapsed during transaction and Number of transactions) to find out the accuracy of results.

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision.

- **K-Nearest Neighbor**

The k-Nearest Neighbours algorithm or KNN for short is a very simple technique**.** The entire training dataset is stored. During prediction, the k-most similar records to a new record from the training dataset are then located. From these neighbours, a summarized prediction is made. This process involves three steps which are:

1. Calculate the distance
2. Get Nearest Neighbor
3. Make predictions

- **Logistic Regression**

Logistic regression is a linear classifier, so you’ll use a linear function $\mathbf{(x) = b_0 + b_1x_1 + \cdots + b_rx_r,}$ also called the logit. The variables b_0, b_1, …, b_r are the estimators of the regression coefficients, which are also called the predicted weights or just coefficients.

- **Support Vector Machine (SVM)**

SVM offers very high accuracy compared to other classifiers such as logistic regression, and decision trees. It is known for its kernel trick to handle nonlinear input spaces. We will pass our training and testing set to the SVM function to get our results

- **Random Forest**

Random Forests (RFs) are ensembles of DT which operate by constructing a multitude of such trees while training and letting them vote for the class of the instances in the test set. The main principle behind this ensemble approach is that a group of “weak learners” can come together to form a “strong learner”.

- **XG Boost Classifier**

The classifier models can be added until all the items in the training dataset is predicted correctly, or a maximum number of classifier models are added. The optimal maximum number of classifier models to train can be determined using hyper parameter tuning. At gradient boosting when classifying the model, the loss function is reduced by gradient descent. Technically, the loss function is considered an error, i.e., the difference between the expected amount and the actual amount. Of course, the smaller the error, the better the machine learning model.

3.2.4 Time Series trend analysis

- **Kwiatkowski–Phillips–Schmidt–Shin (*KPSS*)**

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) check figures out if a statistic is stationary around a mean or linear trend or is non-stationary because of a unit root. A stationary time series is one wherever applied math properties just like the mean and variance are constant over time.

- The null hypothesis for the test is that the info is stationary.
- The alternate hypothesis for the test is that the data isn't stationary.

This is a crucial distinction since its potential for a time series to be non-stationary, haven't any unit root nonetheless been trend stationary. In each unit root and trend-stationary processes, the mean is often growing or decreasing over time; however, within the presence of a shock, trend-stationary processes are mean-reverting (i.e., transitory, the time series can converge once more towards the growing mean, that wasn't suffering from the shock) whereas unit-root processes have a permanent impact on the mean (i.e., no convergence over time). A major disadvantage for the KPSS check is that it's a high rate of kind I errors (it tends to reject the null hypothesis too often). If AN attempt tries are created to regulate these errors (by having larger p-values), then that negatively impacts the test’s power.

A method to affect the potential for prime kind I errors is to mix the KPSS with an ADF test. If the result from each tests suggests that the statistic in stationary.

- **ADF (Augmented Dickey-Fuller) test**

Named for American statisticians David Dickey and Wayne Fuller, who developed the test in 1979, the Dickey-Fuller test is employed to work out whether a unit root (a feature that may cause problems in applied mathematics inference) is gift in an autoregressive model. The formula is acceptable for trending statistic like quality prices. It’s the only approach to check for a unit root, however most economic and money times series have a lot of sophisticated and dynamic structure than what will be captured by an easy autoregressive model, that is wherever the increased Dickey-Fuller test comes into play. Augmented Dickey Fuller look at (ADF Test) could be a common applied math test won’t to test whether or not a given statistic is stationary or not. It’s one amongst the foremost usually used statistical test once it involves analyzing the stationary of a series.

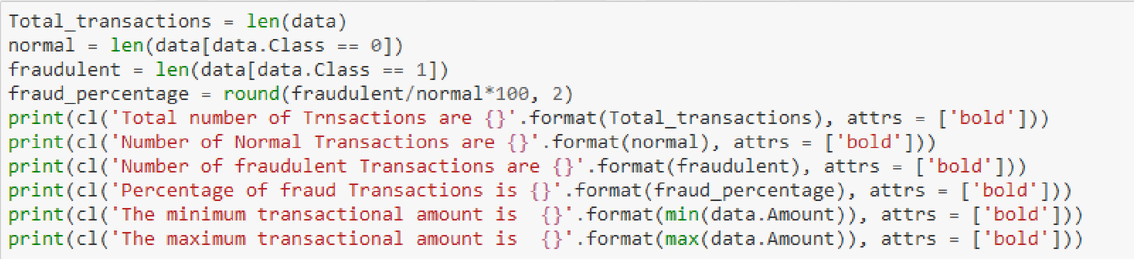
The augmented Dickey-Fuller statistic used in the ADF test is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root. Of course, this is only at some level of confidence. That is to say that if the ADF test statistic is positive, one can automatically decide not to reject the null hypothesis of a unit root. In one example, with three lags, a value of -3.17 constituted rejection at the p-value of.10

CHAPTER 4

RESULTS

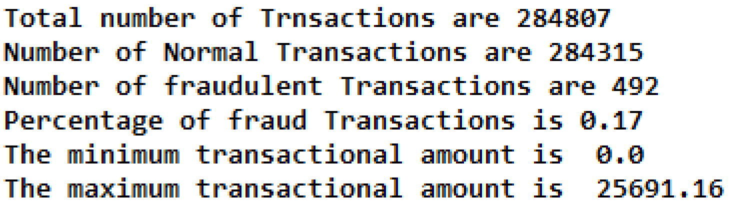
4.1 Descriptive Analysis

Descriptive statistics are utilized to depict or sum up the attributes of an example or informational collection, like a variable's mean, standard deviation, or recurrence. Inferential insights can assist us with understanding the aggregate properties of the components of an information test. Different functions have been performed on the existing dataset to find the details of the dataset in python, and results were generated.



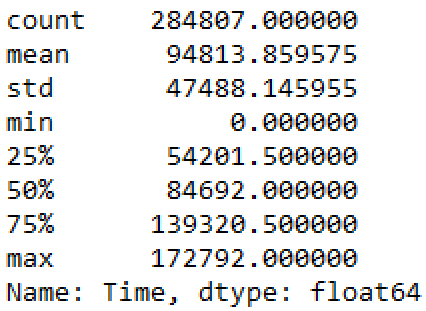
```
Total_transactions = len(data)
normal = len(data[data.Class == 0])
fraudulent = len(data[data.Class == 1])
fraud_percentage = round(fraudulent/normal*100, 2)
print(cl('Total number of Trnsactions are {}'.format(Total_transactions), attrs = ['bold']))
print(cl('Number of Normal Transactions are {}'.format(normal), attrs = ['bold']))
print(cl('Number of fraudulent Transactions are {}'.format(fraudulent), attrs = ['bold']))
print(cl('Percentage of fraud Transactions is {}'.format(fraud_percentage), attrs = ['bold']))
print(cl('The minimum transactional amount is  {}'.format(min(data.Amount)), attrs = ['bold']))
print(cl('The maximum transactional amount is  {}'.format(max(data.Amount)), attrs = ['bold']))
```

By using the length () in python we have generated the following results



```
Total number of Trnsactions are 284807
Number of Normal Transactions are 284315
Number of fraudulent Transactions are 492
Percentage of fraud Transactions is 0.17
The minimum transactional amount is  0.0
The maximum transactional amount is  25691.16
```

Now as the methods used in the analysis, we will find out the variable mean, standard deviation, range etc. using the python function of **describe ().** Following results are generated.



```
count    284807.000000
mean      94813.859575
std       47488.145955
min           0.000000
25%       54201.500000
50%       84692.000000
75%      139320.500000
max      172792.000000
Name: Time, dtype: float64
```

4.2 Visualization of Dataset

After identifying our dependent and independent variables. Their behaviour had been visualized, Average time elapsed during each transaction has been divided into per hours, per 30 minutes, per 15 minutes and per minutes and their behaviour is visualized.

4.2.1 Per Hour

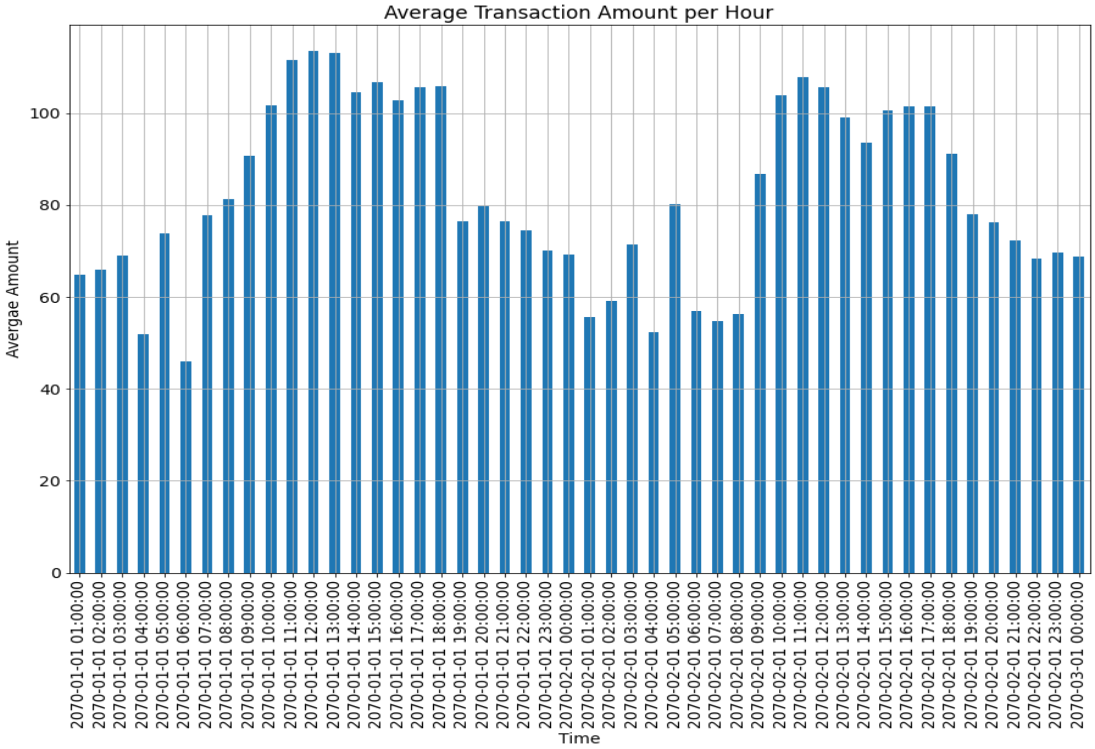


FIGURE 5: AVERAGE TRANSACTION AMOUNT PER HOUR

4.2.2 Per 30 Minutes

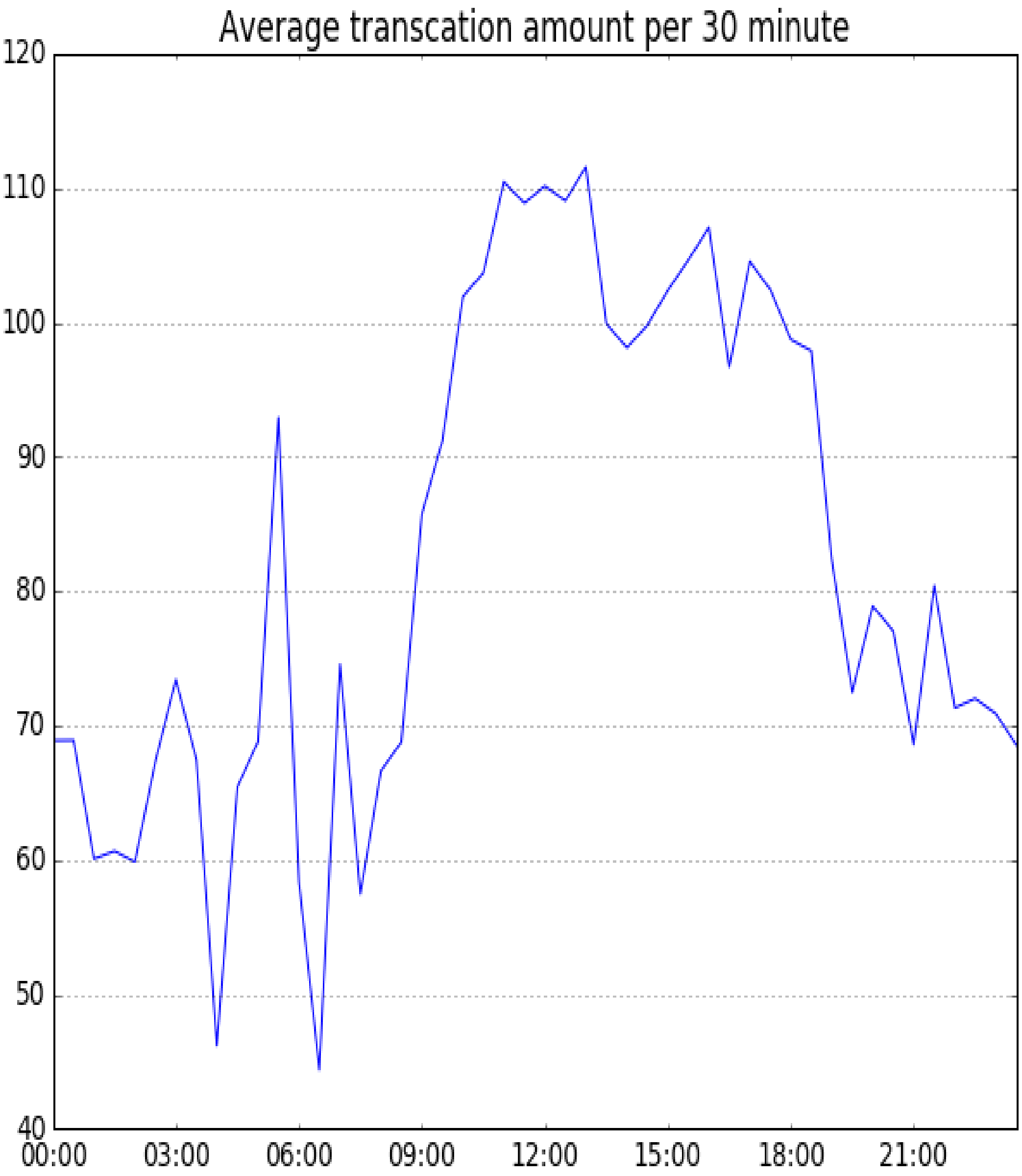


FIGURE 6: AVERAGE TRANSACTION AMOUNT EVERY 30 MINUTES

4.2.3 Per 15 Minutes

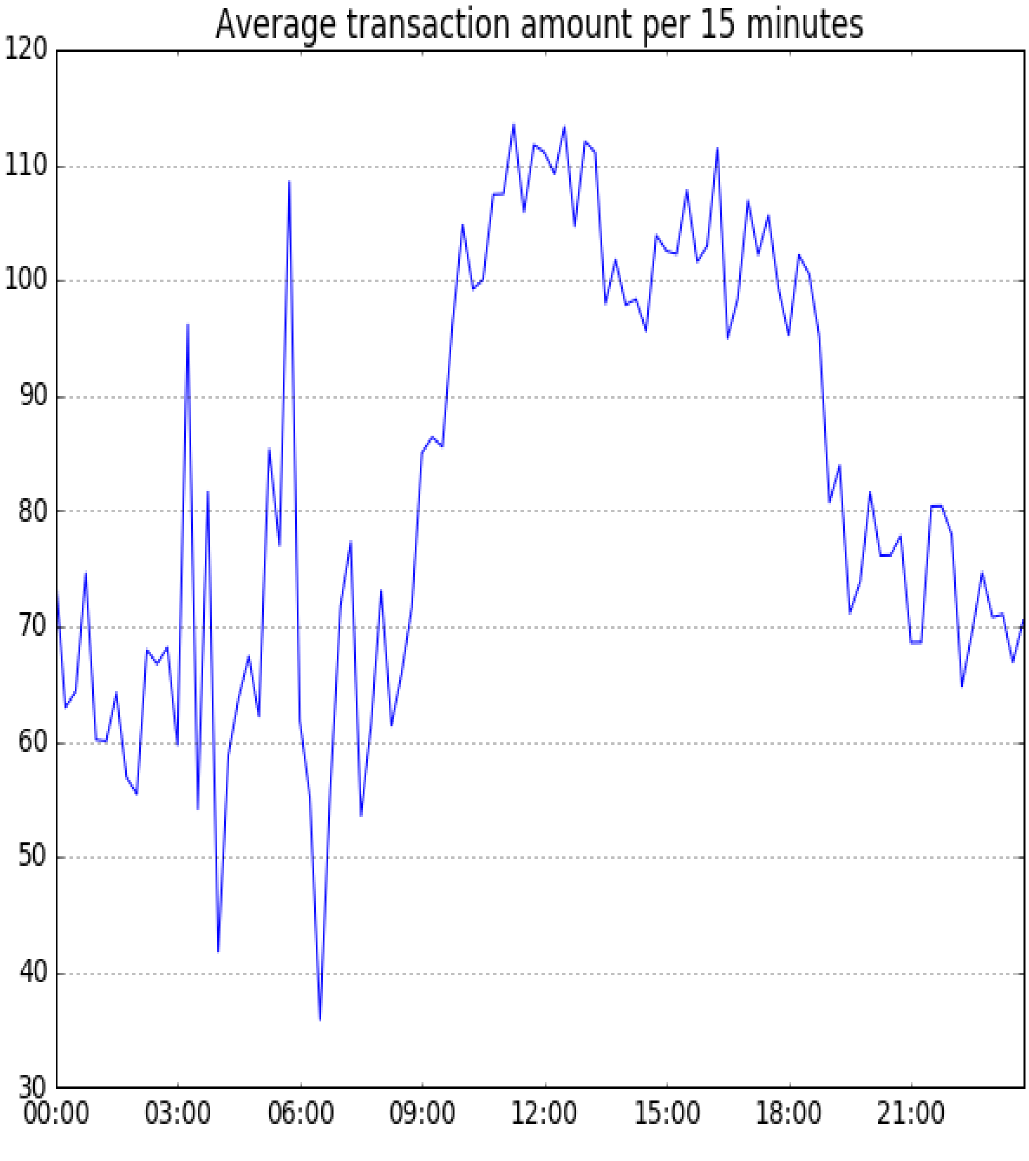


FIGURE 7: AVERAGE TRANSACTION AMOUNT EVERY 15 MINUTES IN AN HOUR

4.2.4 Per minute

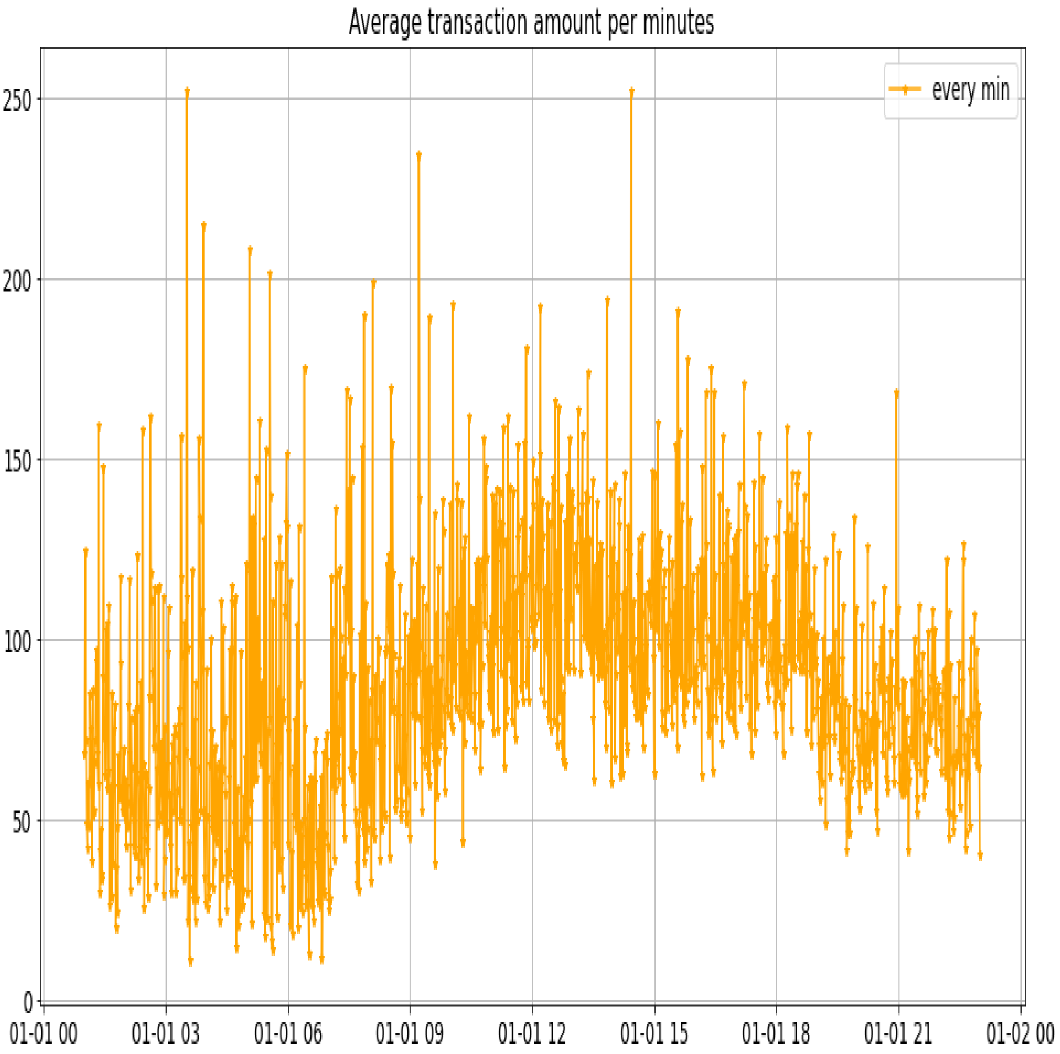


FIGURE 8: OVERVIEW OF TRANSACTION AMOUNT EVERY MINUTE OF THE HOUR

By visualizing our data based on time and the time elapsed we can easily see the spikes generated during the transaction which will help us to identify our fraudulent transaction and prediction of future transaction.

4.3 Overview of transactional behavior on basis of time

In the previous step we have visualized all the transaction on the basis of time, by using the technique of graph plotting, we will merge all our data into one graph which will show us all the transactions and their behaviours across 24 hours of time. The peaks identified in the trend represent large amounts and help to identify the fraudulent transactions.

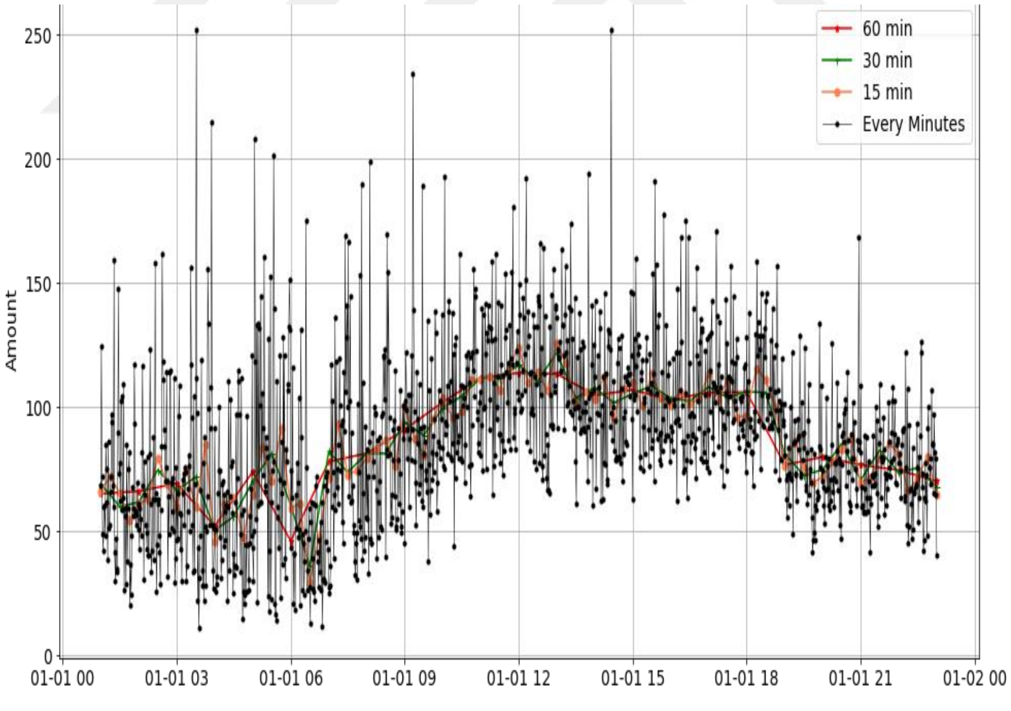


Figure 9 Credit Card Transaction Amount Across 24 Hours

4.4 Principal components (Fraudulent/Normal Transaction)

The dataset consists of 28 principal component features which represent the nature of the transactions. We further performed the PCA analysis on these features to visualize the clusters for fraudulent and non-fraudulent transactions. The first two principal components explain more than 80% of the variance in the data and clearly separate the two clusters.

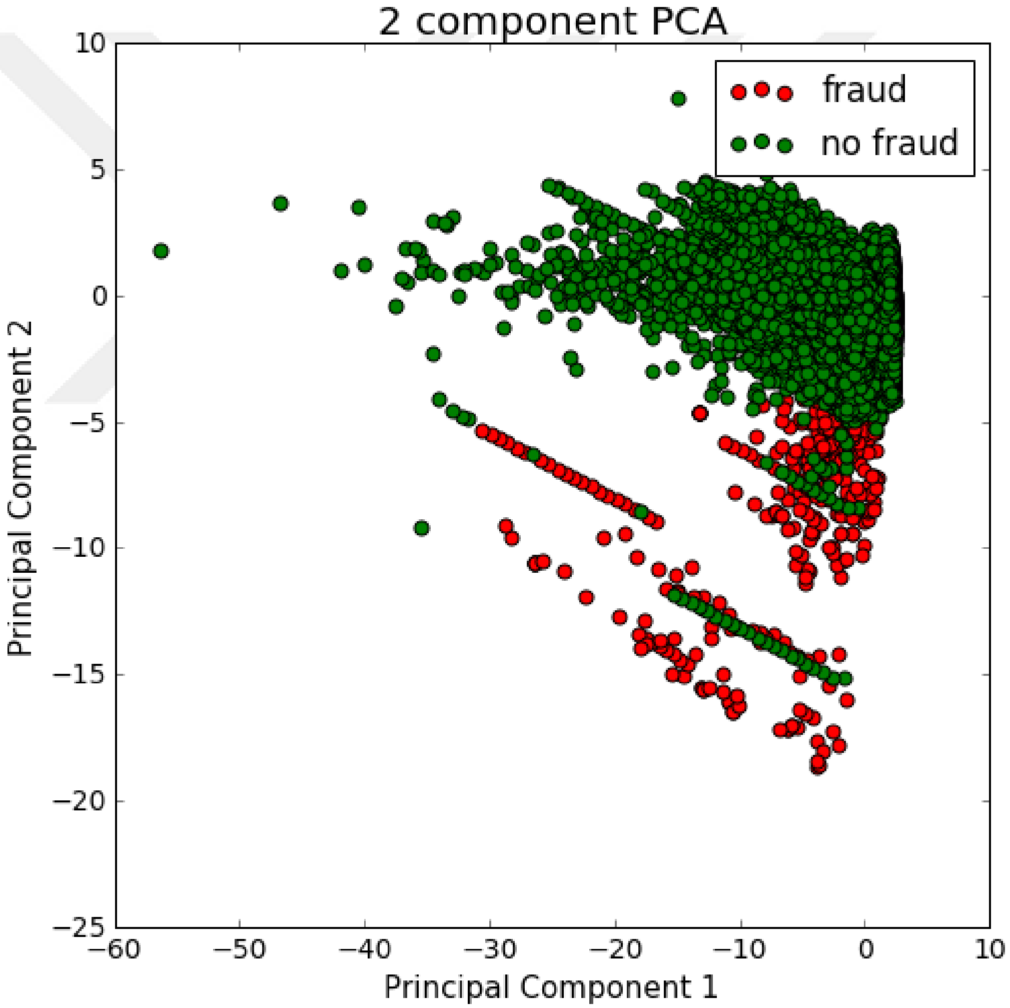
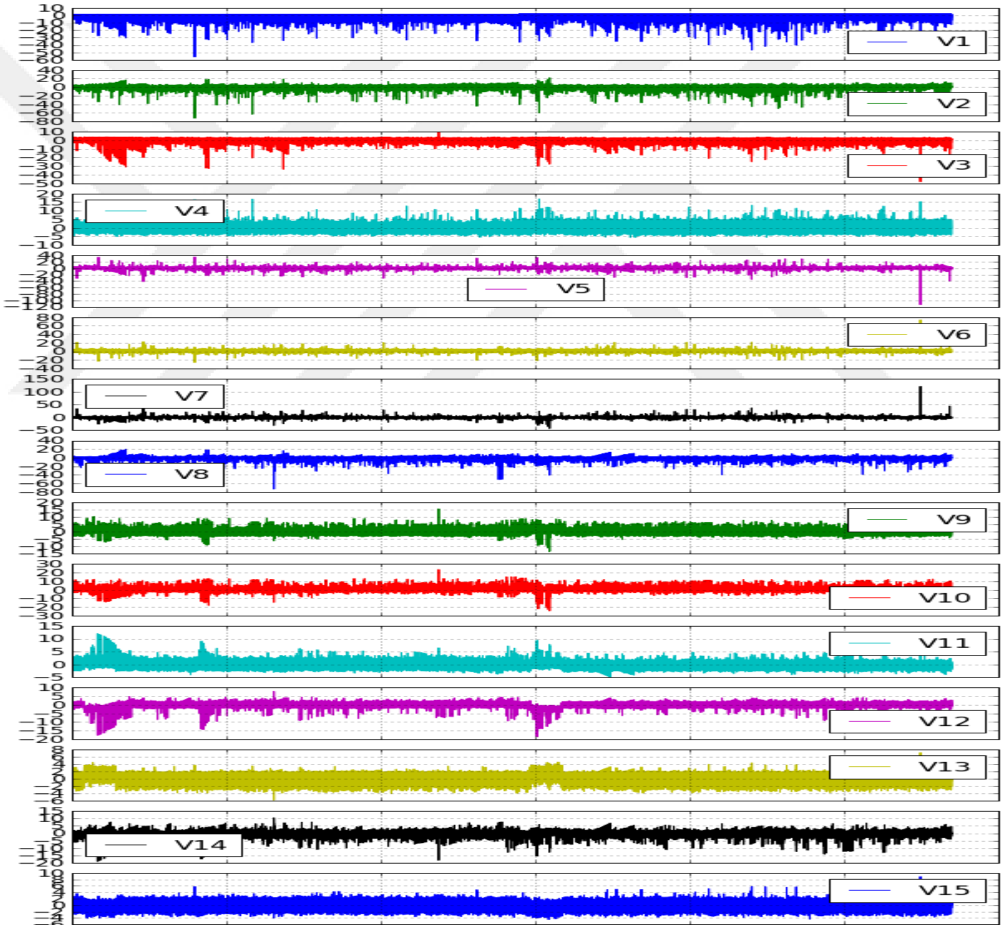


FIGURE 10 PRINCIPAL COMPONENT'S ANALYSIS

4.5 Overview of 28 principal component features present in the dataset

After observing the fraudulent & non-fraudulent clusters for each transaction, we have plotted the trend line for each feature. The overview of features represents the peaks of transactions which can be identification of frauds.

We have numerical values of data for these components in our dataset. We have plotted the with X-axis being constant 0 and on Y-axis we have plotted the value which is present in a specific cell. Below is result we got from these values.



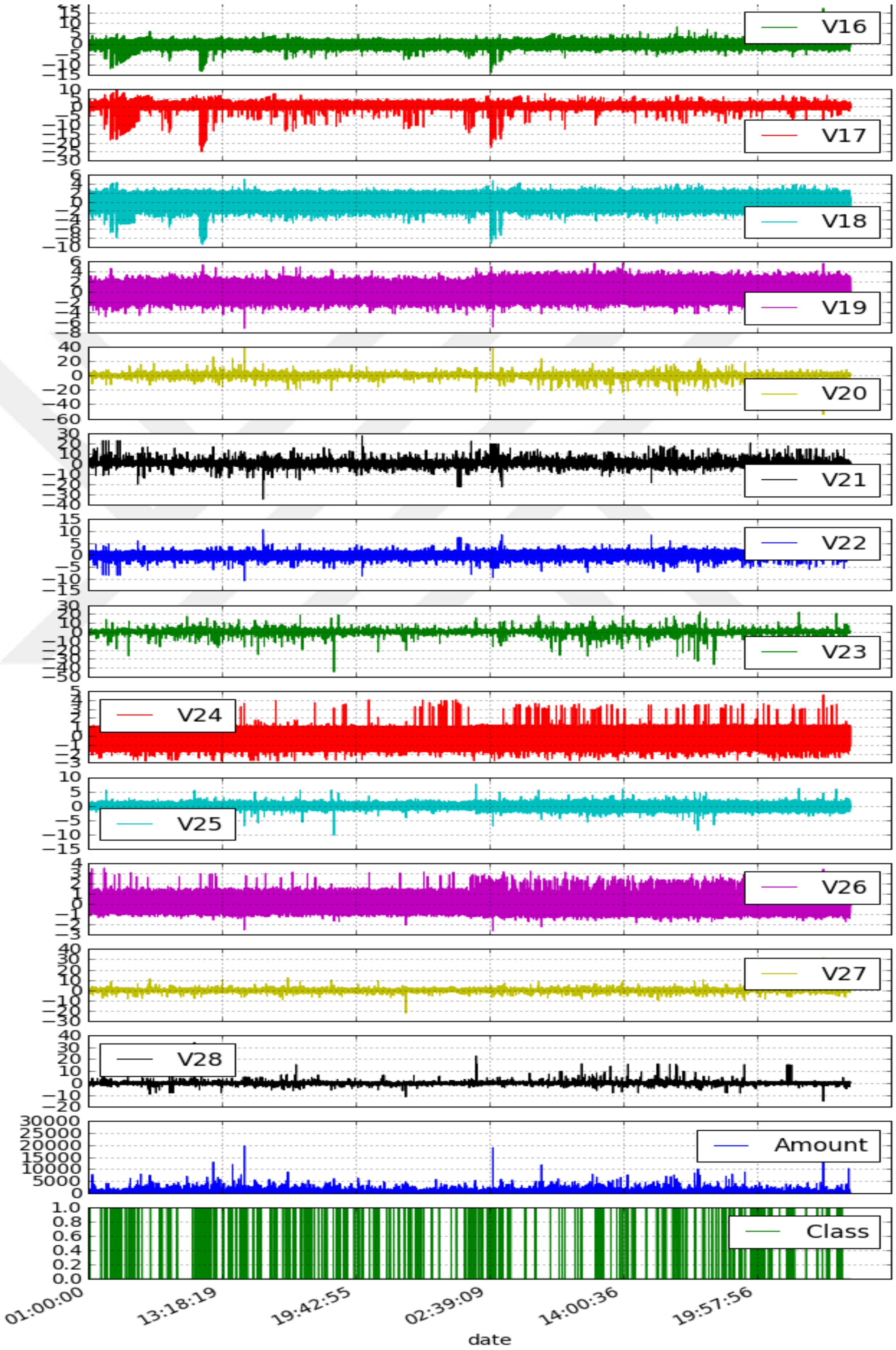


FIGURE 11 INDEPENDENT PRINCIPAL COMPONENT'S AND THEIR BEHAVIORS ANALYSIS

4.6 Correlation between dependent and independent variables

The dataset does not contain the real features of each transaction, but the 28 principal components, which are descriptive of each feature. In our analysis, and predictive models, we have used these 28 principal components are independent features and ‘transaction amount’ as the dependent feature. Prior to running the classification algorithms, we have also performed correlation analysis of these independent features. This analysis helped us to check the multi-collinearity between independent features.

We have calculated the Pearson correlation between all the independent features. If any of the independent features correlate with each other (i.e., Pearson correlation >0.5) than that feature should be removed from the analysis. However, we have observed that all the features have zero correlation, as shown in the heat map below.

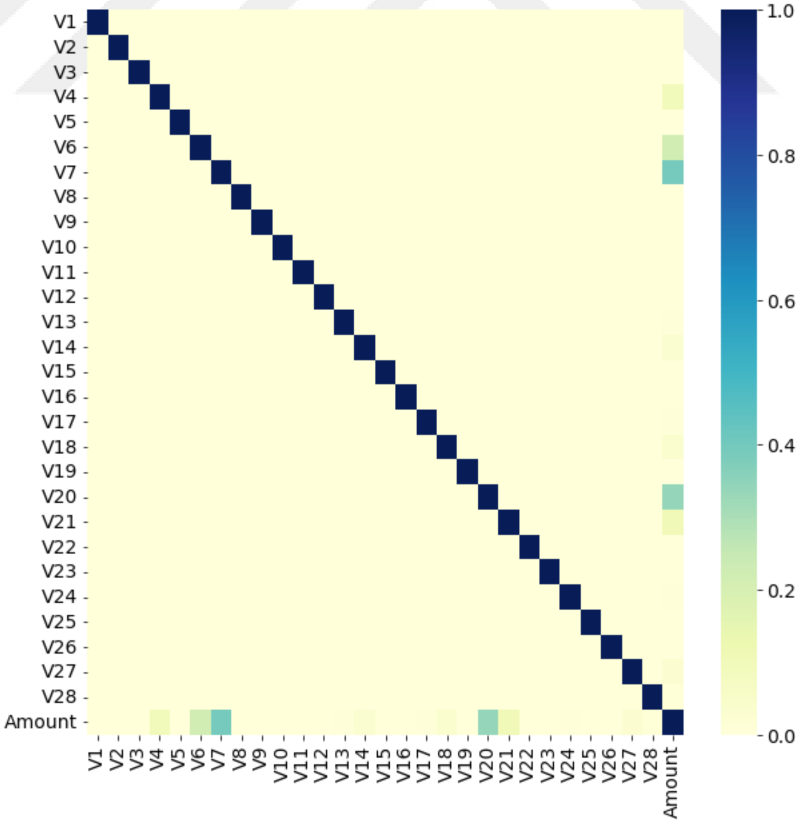


FIGURE 12: HEATMAP OF CORRELATION MATRIX BETWEEN INDEPENDENT AND DEPENDENT FEATURES USING PEARSON CORRELATION

Additionally, we have also performed the correlation analysis of each independent feature with transaction amount (dependent feature). It was observed that four features (V6, V7, V20, V21) have weak correlation (Pearson correlation = 0.25) with the transaction amount. We further explored this correlation by plotting the trend of these four features against the transaction amount using scatter plot. These plots show a weak increasing trend and positive correlation of V6, V7, V20, V21 with transaction amount.

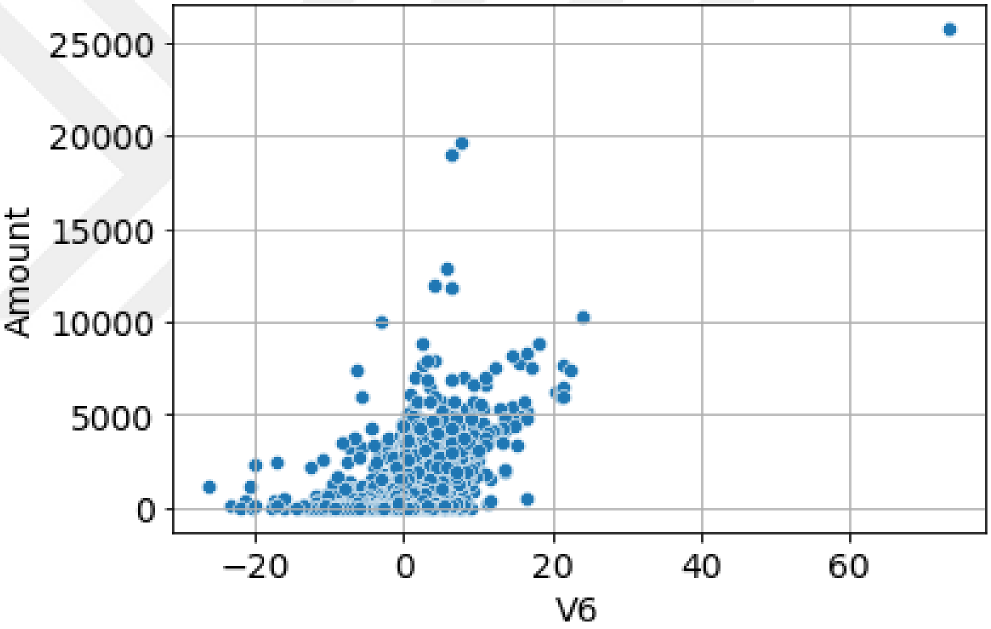


FIGURE 13: SCATTER PLOT BETWEEN 6TH PRINCIPAL COMPONENT (V6) FEATURE AND TRANSACTION AMOUNT

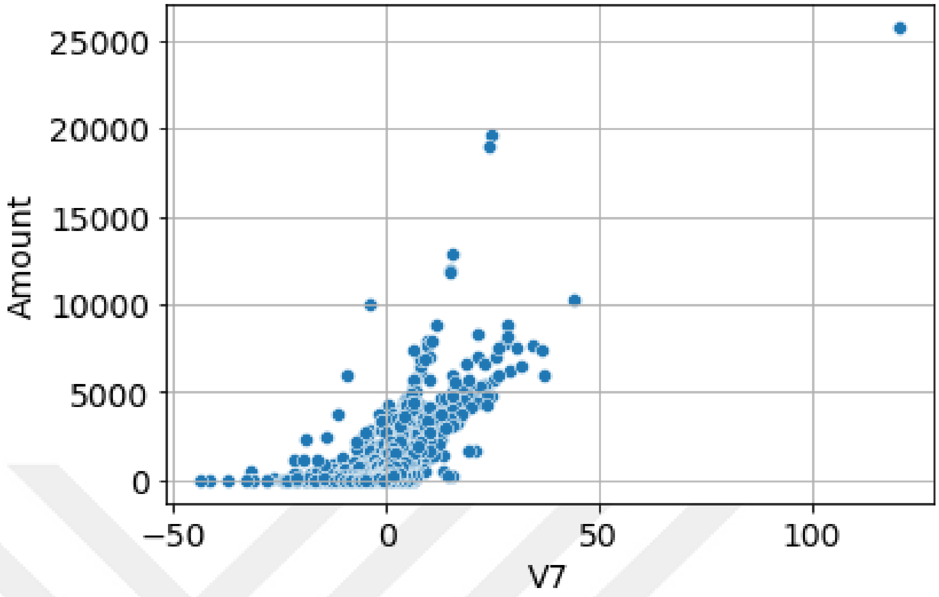


FIGURE 14: SCATTER PLOT BETWEEN 7TH PRINCIPAL COMPONENT (V7) FEATURE AND TRANSACTION AMOUNT

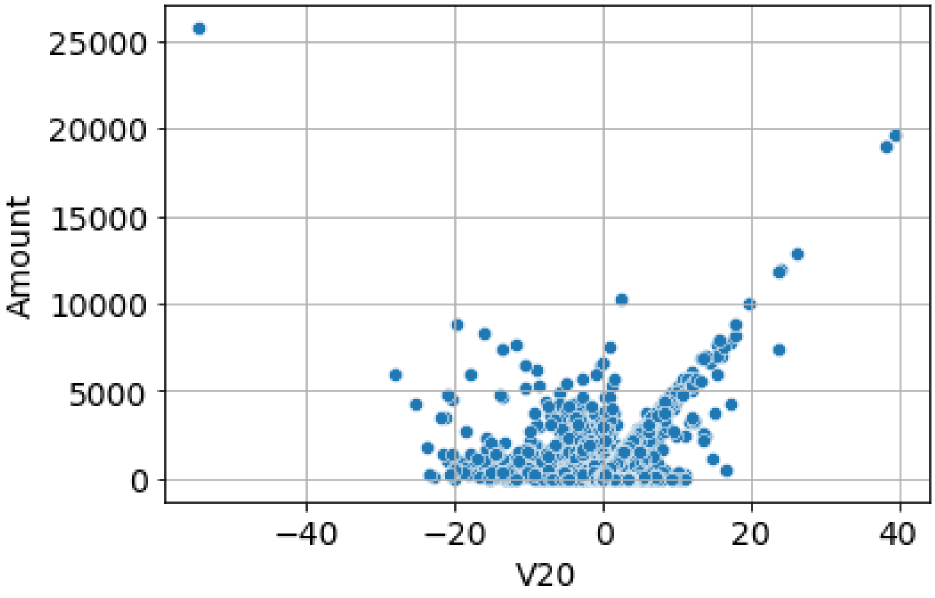


FIGURE 15: SCATTER PLOT BETWEEN 20TH PRINCIPAL COMPONENT (V20) FEATURE AND TRANSACTION AMOUNT

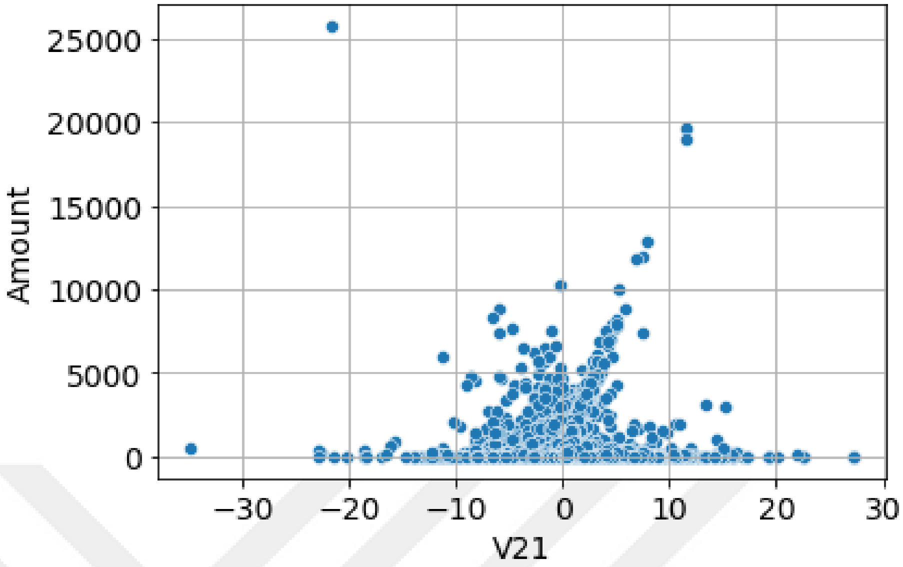


FIGURE 16: SCATTER PLOT BETWEEN 21ST PRINCIPAL COMPONENT (V21) FEATURE AND TRANSACTION AMOUNT

4.7 Comparative analysis of different machine learning classification algorithms

As explained in the methodology section that we have experimented with 6 different classification algorithms to predict the fraudulent transactions from non-fraud transactions. The F1 score and accuracy was measured for each classification algorithm using the 30% of the unknown test dataset. To select the best classification algorithm, we compared these metrics and mentioned in the table below

Classification Algorithm	**Accuracy**	**F1 score**
Decision Trees	0.99	0.99
K nearest neighbours	0.98	0.102
Logistic regression	0.99	0.66
Support Vector Machines	0.99	0.00002
Random Forest	0.99	0.75
XGBoost Classifier	0.99	0.85

F1 Score

Accuracy measures the results based on positive value we got from our results e.g., in our case our accuracy represents the values of fraudulent transactions detected from a dataset over a total number of transactions. Whereas F1 score provide the balance and precision value based on the negative results we got from our experiment. F1 scores and accuracy are opposite to each other as F1 scores can be calculated based on true negative results during the experiments. F1 scores are mostly used when we are doing our experiment with highly imbalanced data as it considers the true negative values and provides precise results.

F1 scores can never be 0. The results we got of F1 score is which is almost 0.00002 is due to the highly imbalanced data in our data set due to which classification became too much difficult. Due to these bad classifiers results seems to be 0 for F1 score.

Based on the accuracy and F1 measures for different algorithms, we have observed that XGBoost performed best to identify the fraudulent transactions from the non-fraud transactions. Hence, we will select this algorithm for any further analysis.

4.8 Time Series Analysis

The main purpose of this research study is to predict the fraudulent transactions in the given credit card data. We have addressed this problem by machine learning algorithms and identified that XGBoost classifier is the best predictive model for this data. Additionally, we have also performed a time series analysis of the data since the credit card data is time indexed.

Time indexed data is valuable because of the time stamp information associated with each observation. This means that one can identify if the value of one data point is statistically dependent on the other data point in another time. This helps in identifying a fundamental assumption that data is statistically independent. For this reason, we have performed two statistical tests on this data to identify whether the time series is stationary or seasonal, meaning is there is a dependency between data points occurring at different time stamps. If there is a dependency, we should observe that time series have certain trends, and it is not stationary.

According to KPSS and ADF test (explained in methodology section), we have identified there is no specific seasonal pattern present in this credit card transaction dataset. The time series is also stationary, as shown in the figure below. The analysis was performed by using a moving average series with a window size of 13. In the figure below, it is quite evident that the original series and the moving average series overlap with each other. This means that there is no specific peak or trend in this data

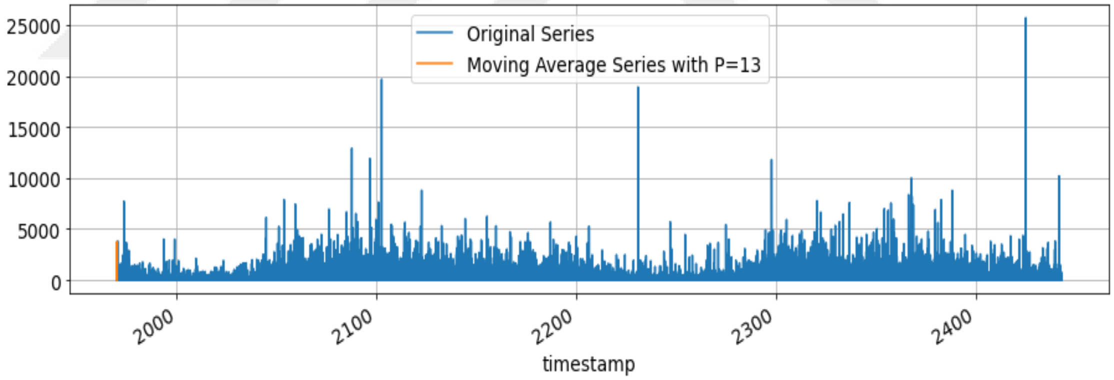


FIGURE 17: STATISTICAL ANALYSIS PERFORMED ON CREDIT CARD TRANSACTION DATA REVEALS THAT THE TIME SERIES IS STATIONARY AND NOT SEASONAL

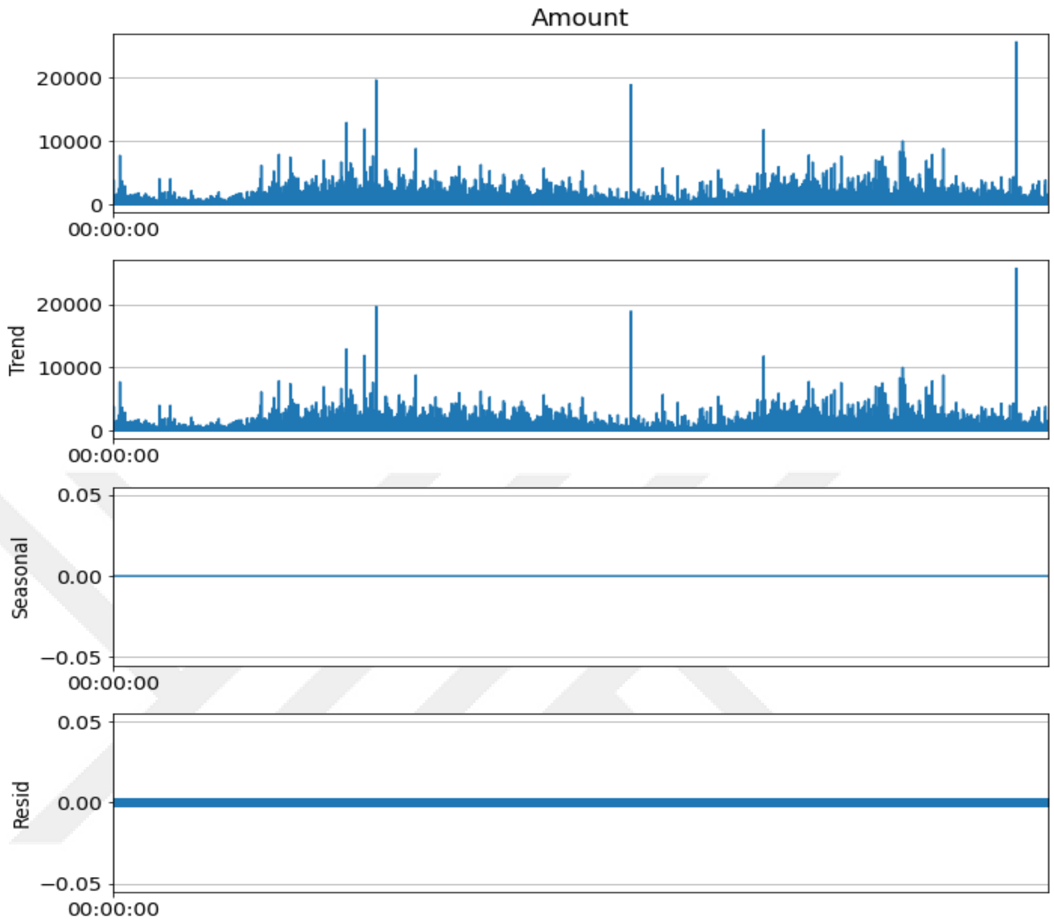


FIGURE 18: TIME SERIES ANALYSIS OF CREDIT CARD TRANSACTION DATA OVER A PERIOD OF 48HR.

THE TOP PANEL SHOWS THE TRANSACTION AMOUNT OVER A PERIOD OF 24-48 HRS. SECOND PANEL SHOW CASES THE 'TREND' IN THE TIME SERIES. AS NOTICED, NO TREND WAS OBSERVED IN THIS DATA. THIRD& FOURTH PANEL IN THE FIGURE REPRESENT THE SEASONALITY AND RESIDUAL OF THE TIME SERIES RESPECTIVELY. AS MENTIONED IN TEXT, THERE WAS NO SEASONALITY OBSERVED IN THIS DATA

Lastly, we performed experiments to calculate the autocorrelation between different data points in this data. Auto-correlation refers to the degree of similarity between A) a given time series, and B) a lagged version of itself, over C) successive time intervals. Auto correlation measures the relationship between variable’s present value and any past value. Usually, this type of analysis is very helpful in stock market finance analysis. We have observed no auto correlation between transaction amounts. The fraudulent transactions are very different from non-fraudulent transactions; hence no auto correlation can be observed in this type of data. This could also be observed from the stationary time series.

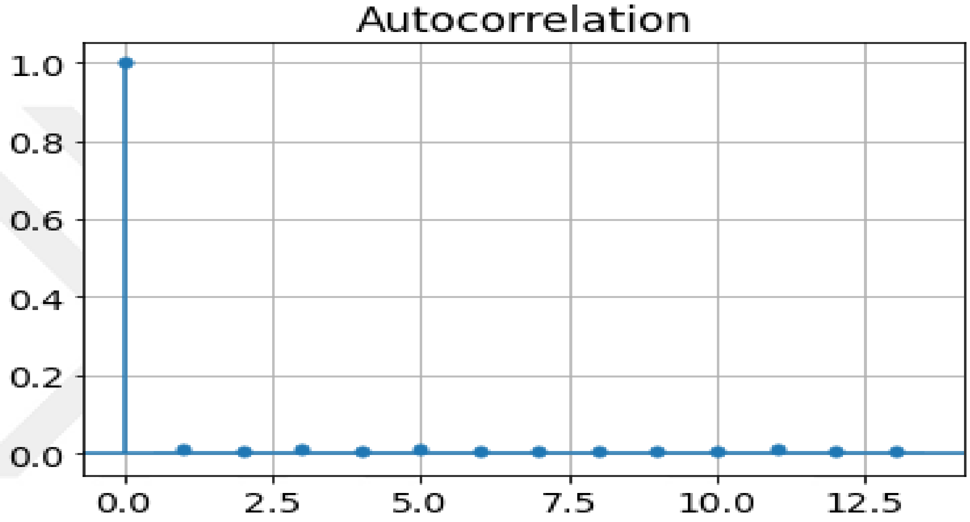


Figure 19: Autocorrelation between transaction amounts. The experiment shows that there is no autocorrelation between amounts. This could be observed from a stationary time series

Hence, overall, the time series analysis for credit card transactions did not show case any pattern or trend which could be leveraged for predicting fraudulent transactions. Each amount is variable, and it is highly unlikely to expect a specific trend from this data. Hence to predict any anomalies, one should either rely on the classification algorithms as experimented in this research study or use the anomaly detection methods with moving average on each datapoint.

CHAPTER 05

CONCLUSION

In this era of technology, Machine learning seems to be a divinity in solving complex problems. As it offers us to predict, analyse and make decisions on our data according to our needs. ML algorithms are so strong and efficient that they consider every aspect of data and analyse the behaviours in different condition. So far, we have concluded this thesis on following which are below explained.

- **Datasets and their confidential nature.**
- **Machine learning prediction algorithms and results.**
- **Time Series and Importance of time-based transactions.**
- **Fraudulent transaction and real time behaviors.**

5.1 Datasets and their confidential nature

Datasets present in case of credit card is very difficult to analyse because most of the information we got from these datasets are confidential which makes it difficult to analyse and derive results from that. For a good result it would be necessary that all the information of the transactions to be analysed. Algorithms we have used in this research can be implemented in a way that it can analyse the real time transactions and predict about the frauds and flags the anomalies in the transactions. But when we compare it with some other dataset the behaviour of data will be changed, variables will be different. In every case we must define a model for that.

5.2 ML prediction algorithm and their results

Machine learning algorithms performs very efficiently in prediction. We have analysed our data with our different ML algorithms. Result generated are very satisfied and when we see it in a long approach these results can be effective in real time. Accuracy and F1 score of each algorithm are discussed briefly in Results chapter. Our research suggest that Machine learning is a suitable method to predict for frauds and it can perform better when the confidential information is present in real time. Hybrid model can be established which will intake the data in real time and instantly give results for that.

5.3 Time Series and Importance of time-based transactions

We managed to find out a dataset which have a time series. Each transaction is analysed with respect to time. This factor has its own importance in this case. If we develop a system which predicts the transactional behavior in real time, time series will be very helpful and effective. We can create intervals of times, through which we can find out how much transaction are performed on regular basis. Some out of order transaction can be analysed thoroughly, decisions can be made on that whether it’s a normal transaction or a fraudulent transaction.

5.4 Fraudulent transaction and real time behavior’s

Different research has been performed in the last decade but still this research cannot be implemented in real time, main problem that arouse is the lack of information, because in different organization different type of data is being used for a payment method of transaction. Model we have created is a good model in case of dataset analyzed but when we see it in other aspects it would not give results for any other organizational data. It is a kind of Model which can be implemented as real time where the time elapsed and amount of transaction is available. Moreover, there are some dependent variables which are analyzed in the research, resulting their effects on the transactions.

We suggest a hybrid model to overcome this problem. But for making that kind of model all the values are data would be known, so that when it is implemented in a real time there would not be any problems.

There is a lot more which can be done to avoid fraudulent transaction. As increasing in technology many new methods are now being introduced on daily basis which works for every payment. This problem should be solved as early as possible.

Future Work

Results we got from our experiments are very satisfying and analyzing them based on the time series data will enable us to have result up to 100% accuracy. In future we will extend our work more vastly by working in a dataset which contains all the values and history of transaction as well as the history if customers spending which will help us to build a model which can be implemented in a real life and can predict about the transaction whether it’s a normal or fraudulent transaction. Automated systems can be installed in banks based on these system fraudulent transactions can be minimized.

REFERENCES

Carcillo, F., le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

Chen, J. I.-Z., & Lai, K.-L. (2021). Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert. *Journal of Artificial Intelligence and Capsule Networks*, *3*(2), 101–112. https://doi.org/10.36548/jaicn.2021.2.003

Dal Pozzolo, A., Caelen, O., le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026

Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

Krishnan, S., Neyaz, A., & Liu, Q. (2021). *IoT Network Attack Detection using Supervised Machine Learning*.

Lin, T. H., & Jiang, J. R. (2021). Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics*, *9*(21). https://doi.org/10.3390/math9212683

Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by Modified Fisher Discriminant Analysis. *Expert Systems with Applications*, *42*(5), 2510–2516. https://doi.org/10.1016/j.eswa.2014.10.037

Mittal, S., & Tyagi, S. (2019). Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 320–324. https://doi.org/10.1109/CONFLUENCE.2019.8776925

Patil, S., Somavanshi, H., Gaikwad, J., Deshmane, A., & Badgujar, R. (2015). International Journal of Computer Science and Mobile Computing Credit Card Fraud Detection Using Decision Tree Induction Algorithm. In *International Journal of Computer Science and Mobile Computing* (Vol. 4). www.ijcsmc.com

RB, A., & KR, S. K. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35–41. https://doi.org/10.1016/j.gltp.2021.01.006

Sadgali, I., Sael, N., & Benabbou, F. (2020). Adaptive Model for Credit Card Fraud Detection. *International Journal of Interactive Mobile Technologies (IJIM), 14*(03), 54. https://doi.org/10.3991/ijim.v14i03.11763

Save, P., Tiwarekar, P., N., K., & Mahyavanshi, N. (2017). A Novel Idea for Credit Card Fraud Detection using Decision Tree. *International Journal of Computer Applications*, *161*(13), 6–9. https://doi.org/10.5120/ijca2017913413

Shirodkar, N., Shet Mandrekar, R., Sakhalkar, R., Chaman Kumar, K. M., & Aswale, S. (n.d.). *Credit Card Fraud Detection Techniques*. https://www.researchgate.net/publication/351686482

Tanouz, D., Subramanian, R. R., Eswar, D., Reddy, G. V. P., Kumar, A. R., & Praneeth, C. H. V. N. M. (2021). Credit card fraud detection using machine learning. *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*, 967–972. https://doi.org/10.1109/ICICCS51141.2021.9432308

Vidyavardhaka College of Engineering, Institute of Electrical and Electronics Engineers. Banglore Section., & Institute of Electrical and Electronics Engineers. (n.d.). *International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC) - 2017 : 8-9, September 2017*.

West, J., & Bhattacharya, M. (2016). Some Experimental Issues in Financial Fraud Mining. *Procedia Computer Science*, *80*, 1734–1744. https://doi.org/10.1016/j.procs.2016.05.515

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). Random forest for credit card fraud detection. *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, 1–6. https://doi.org/10.1109/ICNSC.2018.8361343

Yu, X., Li, X., Dong, Y., & Zheng, R. (2020). A Deep Neural Network Algorithm for Detecting Credit Card Fraud. *2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 181–183. https://doi.org/10.1109/ICBAIE49996.2020.00045

Zeager, M. F., Sridhar, A., Fogal, N., Adams, S., Brown, D. E., & Beling, P. A. (2017). Adversarial learning in credit card fraud detection. *2017 Systems and Information Engineering Design Symposium (SIEDS)*, 112–116. https://doi.org/10.1109/SIEDS.2017.7937699