

Challenges in Credit Card Fraud Detection using Bayesian Optimized Random Forest Classifier

DISSERTATION

Submitted in partial fulfilment of the requirements of the
MTech Data Science and Engineering Degree programme

By

Rajesh PK

2018AB04080

Under the supervision of

Jyjesh TS

Deputy Director (Authentication)

Dissertation work carried out at

Unique Identification Authority of India, Technology Centre

Bengaluru

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

January, 2021

S2-19_DSECLZG628T- DISSERTATION

**Challenges in Credit Card Fraud Detection using Bayesian Optimized
Random Forest Classifier**

Submitted in partial fulfilment of the requirements of the
MTech Data Science and Engineering Degree programme

By

Rajesh PK

2018AB04080

Under the supervision of

Jyjesh TS

Deputy Director (Authentication)

Dissertation work carried out at

Unique Identification Authority of India, Technology Centre

Bengaluru

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

January, 2021

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation titled "***Challenges in Credit Card Fraud Detection using Bayesian Optimized Random Forest Classifier***" is submitted by **Mr. Rajesh P K**, ID No. **2018AB04080** in partial fulfilment of the requirements of S2-19_DSECLZG628T Dissertation, embodies the work done by him under my supervision.



(Signature of the Supervisor)

Place : Bengaluru

Date : 07.01.2021

Jyesh T S

Deputy Director (Authentication)

**Unique Identification Authority of
India, Technology Centre,
Bengaluru- 560092**

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
SECOND SEMESTER 2020-21
S2-19_DSECLZG628T – DISSERTATION

Dissertation Title : **CHALLENGES IN CREDIT CARD FRAUD DETECTION USING BAYESIAN
OPTIMIZED RANDOM FOREST CLASSIFIER**

Name of Supervisor : **JYJESH T S**

Name of Student : **RAJESH P K**

ID No. of Student : **2018AB04080**

ABSTRACT

With technological and economic advances, which facilitated the communication process and increased purchasing power, credit card transactions have become the main payment method in national and international retail. In this regard, the increase in the number of credit card transactions is crucial for generating more opportunities for fraudsters to produce new forms of fraud, which results in great losses for the financial system. This fact arouses the interest of fraudsters. The card market sees fraud as operating costs, which are passed on to consumers and society in general. Still, the high volume of transactions and the need to fight fraud open space for the application of Artificial Intelligence & Machine Learning techniques.

This research work highlights the importance and challenges faced while using auto monitoring to detect credit card fraud to prevent different risks to our assets. Automated Learning techniques have proven to be the solution to supervised learning. This work identifies techniques such as Random Forests and Bayesian optimized Random Forest Classifiers as the best techniques according to related works. This work focused on carrying out the entire process that a project like this addresses, that is to say, features engineering, preprocessing the data, among others. Binary particle swarm optimization algorithm is used for selection of features from the open dataset provided by Kaggle[66]. Sensitivity, precision, f-score and accuracy are used as a performance evaluation tool to find the best parameterization of both the techniques.

Keywords: Bayesian Optimization, BPSO, Credit Card, Data Mining, Random Forest Classifier, etc.

ACKNOWLEDGEMENTS

“In a day, when you don't come across any problems - you can be sure that you are travelling in a wrong path”

— Swami Vivekananda

Foremost, I would like to express my deepest gratitude to my advisor and supervisor for this dissertation, **Mr. Jyesh T S**. He as an advisor provided me the guidance and support all throughout the dissertation and allowed me to explore on my own.

It gives me immense pleasure in acknowledging the guidance and support provided by my mentor **Dr. P M Saravanan** and **Mr. Anup Kumar** (Dy. Director General, UIDAI Technology Centre, Bengaluru) throughout this project.

A very special thanks to **Prof. Dr. Jagadish S Kallimani** whose continuous feedback and guidance has helped in setting pace and direction to this project and attain its desired conclusion.

Finally, I feel great reverence for all my family members and the Almighty, for their blessings and for being a constant source of encouragement.

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	ii
TABLE OF CONTENTS	iii
LIST OF FIGURES	v
LIST OF TABLES	vi
CHAPTER 1	1
INTRODUCTION	1
1.1 Overview	2
1.2 Motivation	5
1.3 Objective	7
1.4 Outline of Dissertation	8
CHAPTER 2	10
LITERATURE REVIEW AND PROBLEM DOMAIN	10
2.1 Literature Review	11
2.2 Problem Statement	15
2.3 Difficulty of the Problem	16
2.3.1 Unbalanced Classes	16
2.3.2 Lack of Real Data	17
2.3.3 Dynamics of Fraud	17
2.4 Solution	17
CHAPTER 3	19
THEORETICAL FRAMEWORK FOR CREDIT CARD FRAUD	19
3.1 Introduction	20
3.2 History	20
3.3 Credit Card Operation	22
3.3.1 Credit Card Agents	22
3.3.2 Credit Card Brands	23
3.3.3 Trends for the Future	24
3.3.4 Credit Card Structure	26
3.3.5 Step by Step Transaction	27
3.3.6 Internet Transactions (Electronic Commerce or E-commerce)	28
3.4 Credit Card Fraud	29
3.4.1 Definition	29
3.4.2 Types of Fraud	31

3.4.3 Methods to Commit Fraud	33
3.4.4 Some Types of Attacks	34
3.4.5 Costs and Fraud Cycle	35
3.4.6 Measures to Prevent Credit Card Fraud	37
3.5 Fraud Prevention and Detection Methods	37
3.5.1 Definition	37
3.5.2 Detection Systems	38
3.5.3 Statistical Methods	40
3.5.3.1 Unsupervised Methods	40
3.5.3.2 Supervised Methods	41
3.5.3.3 Rules-based Systems	41
3.5.3.4 Scoring Models	42
3.5.3.5 Traditional Statistical Techniques for Fraud Detection	43
3.5.4 E-Commerce Transaction Fraud	45
CHAPTER 4	47
PROPOSED METHODOLOGY	47
4.1 Dataset Description	48
4.2 Pre-Processing	49
4.3 Feature Selection using Binary Particle Swarm Optimization	49
4.4 Classification Algorithms	50
4.4.1 Random Forest Classifier	50
4.4.2 Bayesian Optimization of Random Forest Classifier	52
CHAPTER 5	53
SIMULATION RESULTS	53
5.1 Evaluation Parameters	54
5.2 Simulation Results	56
CHAPTER 6	59
CONCLUSION AND FUTURE SCOPE	59
6.1 Conclusion	60
6.2 Future Scope	60
REFERENCES	61
APPENDICES	69
CHECKLIST	70

LIST OF FIGURES

Figure No.	Figure Name	Page No.
1.1	Current total card fraud rate by country	6
3.1	Credit card structure	26
3.2	Transaction flow; Source- MasterCard	28
4.1	Flow diagram of proposed research work	48
5.1	Objective function model	56
5.2	Minimum objective vs. number of function evaluations	56
5.3	Confusion matrix plot for random forest classifier based credit card fraud detection	57
5.4	Confusion matrix plot for Bayesian optimized random forest classifier based credit card fraud detection	58

LIST OF TABLES

Table No.	Table Name	Page No.
3.1	Industry identification	26
3.2	Brand identification	27
3.3	Comparison of statistical techniques for detecting credit card fraud	43

CHAPTER 1

INTRODUCTION

1.1 Overview

Cards, be they credit or debit cards, are means of payment highly used in commerce in general. Even so, there is room for them to be used even more. These factors arouse the interest of fraudsters. In a first analysis, when a fraud is successful, it could be said that the companies involved bear the costs, but, in fact, this operational loss makes prices more expensive for the final consumer. Hence, combating fraud in this payment method brings benefits to the whole society. This task, due to its complexity and size, has to be done in a concomitantly effective and efficient way. Therefore, there is an opportunity to apply machine learning techniques.

Cards, also known as electronic means of payment, are an important form of payment, whether in the world, in India, in face-to-face transactions or in which the holder is not physically in front of the terminal where the transaction is carried out. The use of this means of payment grows at substantial rates year after year. There are several reasons for this growth: for those who use the card as a form of payment, it is practical and, under certain conditions, credit is instant, expenses are concentrated on the invoice and cards are an accepted instrument in many establishments; for those who receive payments with cards, the risk of default is considerably low, sales control is facilitated and there are many people willing to pay for their purchases with them.

However, a fundamental requirement for this payment system is security. Because it is a means of payment, that is, it is closely linked to the financial data of users, if fraud is very frequent and involves many costs, the card will lose its appeal for use by people. In addition, the greater the financial volumes involved in fraud, the greater the financial losses of the companies participating in that market, which may make the maintenance of this payment system unfeasible.

This high scenario and increasing volume of transactions and the need to prevent and detect fraud creates the opportunity to apply machine learning techniques to combat card fraud, as the high number of transactions prevents each of them from being analyzed by a human resource. Thus, these techniques can be applied so that transactions, on the verge of happening, are classified between fraudulent and legitimate, preventing the execution of those identified as fraudulent.

In the developing global world, people's specific needs have increased in importance over time. Now, concepts such as time, speed and security are indispensable for people. These reasons have been brought to the forefront in the concepts of banking and finance, and its value and scope have constantly increased. Such importance of the concept of money has brought to mind the formula money equals bank. The development of the two structures is directly proportional in the global world. With the development of the internet system and the virtual world engulfing humanity, it has entered the virtual world under the name of internet banking in banks. Banks that have switched to online banking have increased the fraud threats they have already faced, thanks to this virtual structure. With this development, banks now provide their customers with faster, higher quality and user-friendly systems, in other words, while making life easier, it has brought certain security threats. Security threats in general [1];

- Interception of credit card information.
- Fraud through call centers.
- Payment systems scam.

Many threats are now valid and important for banks.

Banks have had to use technological developments that create the same threats in order to eliminate or prevent these incoming threats. These are [2]:

- Mail applications
- Passwords
- Security questions
- Artificial intelligence algorithms
- Instant SMS

However, no matter what is done, no matter how many fraudulent methods are detected and how many prevention practices are, banks have now entered a mutual and continuous development phase. At this stage, fraudsters have to create new threats and banks have to take new measures [3].

The created categories are associated with customer shopping. Associating is not only in the form of matching, but the customer's habits are determined by adding the amount of shopping to the system. Location-based shopping can be used. In other words, it can be added to the system at shopping locations.

The customer, whose credit card has been copied or stolen, will be added to the system so that the information remains in the system, and the system will be detected if there are any violations and information will be provided about the customer's risk group. Prevention studies can also be conducted on the determined risk group. These prevention actions can be [4];

- SMS sending by phone to the customer.
- Notification by e-mail, confirmation code.
- Suspend the account until you are sure the security of the account.

Technology and globalization have had significant advances, which has led some financial institutions to continue operating have had the need to invest additional capital. Firstly, to be able to provide their clients with a better service and keep their products attractive, and secondly, that in order to survive they have had to join others, since by themselves if they do not have sufficient capital, they can hardly remain competitive. One of the products that financial institutions offer their customers is the credit card. Product that for years has reached a fairly high level of acceptance and presence in the market, achieving the objective that credit card brands pursue, both in Guatemala and worldwide, such as becoming the universal means of payment, which carries with it the risk of credit card fraud.

This situation derives the importance of the credit card issue being documented and made public so that both the issuing entity and the user can collaborate to minimize the risk of fraud that exists in the use of the credit card. In recent years, this problem has cost banks issuing credit cards large losses, because they have been concerned about increasing their presence in the market, neglecting the part of inherent risk that exists from the moment of issuance of the plastics credit card, without reviewing or establishing safety procedures and controls.

1.2 Motivation

Globalization has made the use of the internet very necessary, we currently carry out daily activities such as online shopping. As is well known, this is one of the means that has the most dangers for our economic well-being, something very seen is cybercriminals, who rob us of our money deposited in a bank account of a credit or debit card.

This event, which is commonly known as a crime, is called fraud, which aims to impersonate the identity when it is presented to a financial entity and is validated by another person.

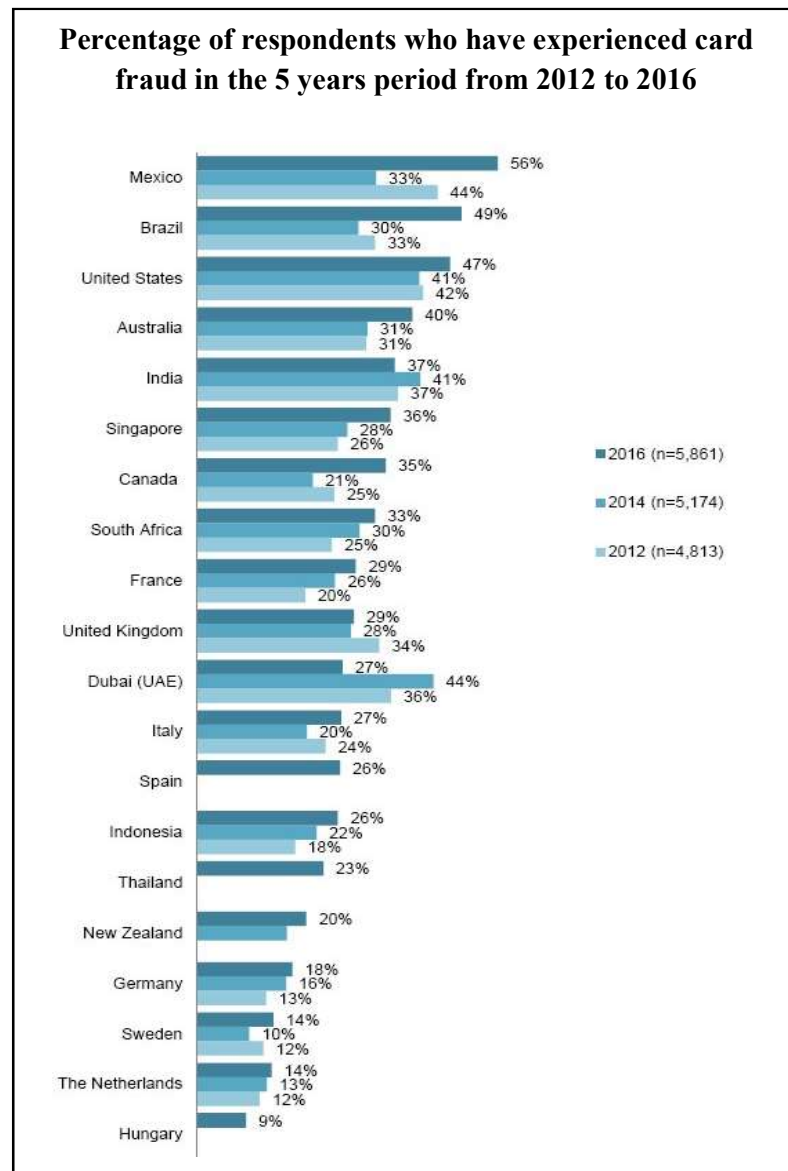


Figure 1.1: Current total card fraud rate by country [5]

However, at present these attacks are committed online, being more common and dangerous, because sensitive user data can be obtained through different techniques against the user's security. Today we can find many of these vulnerabilities at our fingertips. Also, some users do not have sufficient knowledge to safeguard their data, making access to these criminals more accessible.

Therefore, it is very attractive to commit this type of fraud by cybercriminals. Many bank users and identities have suffered numerous losses. That is, a study shows in Figure 1.1 that in the years 2012, 2014 and 2016 in a total of 20 countries, the

percentage of respondents who have experienced credit and debit card fraud is observed [5].

We observe that, in countries such as Mexico, there is a great increase in these years from 44% in 2012 to 56% in 2016, also in Brazil from 33% in 2012 to 49% in 2016.

Analyzing the study more carefully, we can confirm that there is a great loss of money despite the fact that it has already begun to combat this problem of detecting this large amount of fraud. However, it is not being as accurate as desired because the problem of false positives is still incurred.

Due to this problem it is that in the last advances data with certain user behavior has been incorporated. Thanks to this, deviations from these characteristics can be detected and a certain degree of certainty can be confirmed that it is fraud.

This area of detection of credit card fraud is really studied. There are machine learning techniques that are divided into supervised and unsupervised. In the first case, you need a set of transactions based on fraud or legitimate, while the unsupervised, an unusual transaction should be classified as fraudulent or not. In this work, the use of Bayesian optimization in random forest classifier technique is proposed, which provides the best results.

Although fraud detection has some research problems due to the unavailability of real-world data on which researchers can experiment, many fraud detection models have been developed with good results.

Some successful applications of various machine learning techniques can be found, with [6], [7], [8], [9], [10], and [11] being the most powerful and recent techniques in related works.

1.3 Objective

This research work has following objectives:

- To identify the challenges in an automated credit card fraud detection framework

- To develop an automatic credit card fraud detection framework using ***Bayesian Optimization based Random Forest classifier***.
- To develop a feature selection technique based on ***Binary Particle Swarm Optimization*** for achieving reduced training overhead of Kaggle dataset.
- To obtain the simulation results in terms of evaluation parameters; ***sensitivity, precision, f-score and accuracy***.

1.4 Outline of Dissertation

Chapter-1	Introduction: This chapter includes the general overview of research area, motivation behind the work and the objective of the research.
Chapter-2	Literature Review and Problem Domain: It is the most important part of the dissertation work, with the help of various researches and published articles, anyone can easily find out pros and cons of the work which is already carried out by various researchers and scholars. A brief review of this research; article and papers related with credit card fraud detection are included in this part. This chapter also focuses on the problem statement associated with the current research work along with its solution.
Chapter-3	Theoretical Framework for Credit Card Fraud: This chapter presents a detailed explanation about history and operation of credit card, fraud prevention and detection methods.
Chapter-4	Proposed Methodology: After successful formulation of problems in the previous chapters, this chapter provides solution of the problem with brief description of dataset used in this research work and an overall details description regarding the implementation of credit card fraud detection using Bayesian optimized random forest classifier.

Chapter-5	Simulation Results: Simulation of the algorithms and result analysis plays an important role to justify that current research performs better than the existing approach. This chapter starts with the evaluation parameters used in the research work, then the simulation results of proposed methodology are analyzed.
Chapter-6	Conclusion and Future Scope: Finally, in this chapter all the research work is summarizes, which include all the contribution and provides some criticisms and perspectives regarding this dissertation work.

CHAPTER 2

LITERATURE REVIEW AND PROBLEM DOMAIN

2.1 Literature Review

In the initial stages of the investigation, a bibliographic search is carried out to understand what the state of the art in the area is. In this process, we find that open access publications are few and almost always made by university researchers, since companies in the field do not want to disclose their procedures. This reserve is also seen in the handling of the data: there are several articles that cite the difficulties they present to find bases and that consequently resort to artificial or public access data. Even in the cases that use real data, the bases have a considerably less pronounced imbalance than ours, or a much smaller number of transactions, so we have less confidence that the published results can be extrapolated to our context [12], [13], [14], [15] and [16].

A credit card is a plastic document with a security band and a chip issued by a financial or commercial entity to purchase goods and services with a 30-day deferred payment method. Credit cards originated in 1914 in the United States when the Western Union company granted them to its most select and exclusive clientele with the purpose of assuring users a preferential attention in all branches of the company and, in addition, providing them with the possibility of a deferred payment. The bank credit card modality was born in 1951 by Franklin National Bank of Long Island, New York. It identifies the customer's current account number and credit line. Sandoval (1991) [17]

Credit cards are part of the global commercial growth of the economies of emerging and developed countries through their traditional channels and in line with the world; Its exponential growth in online money transfer systems has contributed to the expansion of electronic commerce and to a greater number of consumers buying and selling goods and services in which India is no exception due to its technological infrastructure in mobile networks and Internet.

The fraud as an action contrary to truth and rectitude, which harms the person or entity against whom it is committed, this leads to financial losses and legal problems. Today in the 21st century, with computational and scientific technological advances in statistics and data mining tools, fraud can be detected and predicted before it is committed, there are some platforms that offer financial entities the fraud detection

service, some of them uses machine learning data analytics for essential analytical processing of artificial intelligence to manage the transactional fraud detection and payment monitoring needs of an organization.

Fraud against business establishments and people is the result of improper handling of corporate or personal information, there are several modalities and techniques used by cybercriminals, in one of them they usurp your information through malicious emails called phishing where a malicious program it appropriates personal and corporate information, obtaining the keys to access bank accounts, since these malicious emails recreate critical portals of financial entities where the user enters their data and passwords.

The task of detecting fraud is not an easy issue to solve, taking into account the multiple modalities and rapid evolution that this issue has had nowadays, financial entities worldwide use the science of statistics with tools of data mining and machine learning to recognize patterns of fraudulent behavior, for this, most of the current detection systems offer two types of alerts: alert by probabilistic expiration and by compliance with rules, in the first type of alert predictive models are almost always used for a score expiration, for the second case filters based on SQL command sentences are used [18].

Data mining and machine learning techniques use efficient probabilistic models such as: generalized regression models, artificial neural networks, decision trees and Bayesian optimization networks to determine and predict fraud with a probability or score, they use an autonomous learning system For the recognition of patterns and trends based on historical facts, the data of transactions made by customers are used to determine the patterns, these allow to quickly identify circumstances outside the daily behavior of a customer that may be indications of fraud [19].

The association rules look for the possible relationships in a data set to obtain patterns of fraudulent behavior existing between the presence of an item and a certain set of transactions [20]. An item is a set of binary attributes, and can be labeled as fraud with response where the value of one (1) identifies the fraudulent items and the value zero (0) the non-fraudulent ones.

Methodologies to detect fraud are essential if we want to identify scammers once fraud prevention has failed, statistics and machine learning provide correlated and effective information for fraud detection, they are widely applicable and with great success to detect fraudulent activities such as money laundering, credit card fraud, electronic commerce, telecommunications fraud and private networks; on the contrary, fraud detection involves identifying the fraud as quickly as possible once it has been perpetrated [21] [22].

A meta-classifier is the combination of several models that can be of the same or different types, in order to improve the precision of their predictions. Stacked models consist of the combination of classifier models of different types of learning algorithms from the same data set [23].

Neural networks and classification trees have proven to be a very powerful data mining tool because of their efficient methods in their predictions [24]. In [25], authors explained that its purpose is to compare the performance of the models by combining different algorithms through the stacking or stacking method, the stacking method consists of the construction of multiple models of different types and his learning method assesses how best to combine the predictions of the primary models.

There are various algorithms and systems developed from past to present to detect credit card frauds. In order for the techniques to be developed to prevent credit card fraud to be efficient and effective, it is necessary to examine the credit card fraud methods that have been experienced and seen from the past to the present. In line with the information obtained as a result of the researches, more confident steps will be taken in detecting credit card fraudsters. One of the purposes of the systems developed to detect fraud is to reduce or even reset the material and moral losses caused by fraud. Supervised approaches treat the problem as one of classification into two classes, and as such, in most cases it is approached in the traditional way: processing the data, building characteristics and training a classifier. However, we do not observe that there is a classification algorithm that is preferred over the rest: the most usual is that it is tested with linear regressions, SVM, decision trees, neural

networks and / or Random Forest to stay with the one with the best performance (generally it is one of the last two [13], [14] and [26]).

A technique that is repeated in a large number of articles is the use of accumulators (or transaction aggregation) as a method of extracting characteristics [15], [27], [28] and [29]. In [30] and [31] a similar technique is proposed that would allow adding information (accumulating) with temporal variables, which have the particularity of presenting a metric circular.

In [29], a feature extraction method based on network analysis is proposed: by constructing a graph that relates the cards to the businesses in which it operates, it is proposed to execute a propagation algorithm to finally obtain exposure scores of the merchant, cardholder and transaction that can be used as characteristics.

Another family of methods would be those that think of the problem as one of anomaly detection, taking into account that fraud is rare and it is expected that they deviate from normal behavior. In this sense, the approximations focus on modeling a typical behavior and determining the distance of a transaction from it. Very diverse approaches are used, such as artificial immune systems [32], clustering [12] [33], Principal Component Analysis (PCA) [34] or outlier detectors based on entropy [33]. In some cases, supervised and unsupervised approaches are combined to obtain the financial model [29] [35].

The problem generated by class imbalance is treated in several different ways. In some cases, a sub-sampling of legitimate transactions is performed [35], while in others techniques such as SMOTE [36] are used to generate artificial fraud [13]. In [30] and [31] the results are presented in a smaller subset but with a higher proportion of fraud, which could be thought of as another way to reduce the imbalance (although it is not specified how the remaining group is treated).

Another approach that mitigates the imbalance without modifying the data is to incorporate cost matrices in the classification algorithms [16]. This also has the added utility of biasing classifiers to focus on the more expensive scams. In these cases, the

performance of the models is also evaluated with metrics adapted to take into account the monetary losses avoided and not only the amount of fraud detected [14] [37].

Artificial neural networks [38] [39], support vector machines [40] [41], decision trees [42], random forest classifier [11] and statistical methods [17], which are among data mining techniques, are used in systems that have been put forward for fraud detection in line with the researches from past to present. Although each technique has its own characteristics, it is possible to describe in a single sentence what all these techniques generally used do. In these techniques used to prevent fraud, the spending habits of credit card users are analyzed and divided into many different categories such as persons, profile types and transaction types according to the analysis results. Later, the system evaluates the data according to the criteria in the transaction mechanism established in its structure and creates an output. Necessary actions are performed by interpreting the resulting outputs. In another method performed for detecting credit card fraud, fuzzy logic system using the optimum threshold value and containing fraud calculation information was used [43]. In this system, which is another branch of artificial intelligence, the outputs show the possibility that the user may become a fraudulent, and the reason why this opinion was reached can be explained.

2.2 Problem Statement

Credit card fraud is a serious and growing problem by the day. With the increase in these as a means of payment, due to the rapid increase in online sales and the change in people's behaviour when paying, there has been a greater exposure to transactional fraud.

In addition, over the years and the evolution of methods to detect fraud, people who commit fraud have also evolved their practices to avoid detection, so fraud is dynamic, it is always changing. Consequently, fraud detection methods need to be constantly improved.

Within the behaviour of fraud, we want to find invariants or patterns that allow predicting fraud at the transactional level. This refers to finding strange purchasing behaviours based on the customer's transactional history and data from the incoming

transaction. In this research, an efficient detection model will be applied to the available data to achieve this objective.

There are two edges that make this problem interesting. First, from a conceptual point of view, it is a classification problem with ambiguity, two very similar transactions can one be fraudulent and the other not. It is not clear how to solve it, as there are no laws that describe the behaviour of customers, as this is a social behaviour. Second, from an economic point of view, it is important to prevent fraud and stop losing money for this concept, because in a normal transaction the profit is a percentage of the sale, while in a fraudulent transaction it is the almost total loss. In addition, as a consequence, the client is protected and gives greater security when transacting with the card.

2.3 Difficulty of the Problem

There are several difficulties associated with detecting fraud, such as unbalanced classes (skewed data distributions), scarcity of real data, among others.

2.3.1 Unbalanced Classes

The higher class, in this case normal transactions, is far superior in number to the lower class (fraud). This is natural, since many normal transactions are needed to finance a fraudulent transaction, otherwise a trade is not sustainable over time. The fraud rate in this study is less than 0.1%, which cannot be disclosed due to confidentiality issues. To handle unbalanced classes, methods such as under-sampling [28] and oversampling [44] are proposed, which provide a way to choose an appropriate sample of cases to build the models. The term under-sampling refers to decreasing the proportion of the larger class, choosing a certain number of cases from this class until reaching a more balanced distribution with respect to the selected cases of the smaller class. The term oversampling consists in increasing the proportion of the smaller class, repeating the cases of this class until reaching the desired distribution with respect to the selected cases of the larger class. Another consequence of having such a class imbalance is that some performance measures for the models are not very useful, for example it is possible to have a global precision

(percentage of correctly classified cases) close to 100% classifying all fraudulent transactions as normal.

2.3.2 Lack of Real Data

Most publications on detecting credit card fraud talk about the lack of real data. Out of those reviewed for this report, a few have actual data [17], [28], [45], [46], [47], [48], [49] and the rest use synthetic or simulated data or survey data. The lack of data is due to the fact that these are sensitive and confidential, since they correspond to the transactions of large companies and banks, where there is great competition. For this reason, in publications that do use real data, they have publication restrictions, for example on the number of transactions, amounts, variable names, etc. These situations make it more difficult to search for references in the literature. Out of the publications on detecting credit card fraud and using real data, only in [28] mention the variables used.

2.3.3 Dynamics of Fraud

Another problem is that fraud patterns change over time. When measures are taken to prevent fraud, soon after the people who commit fraud look for alternative ways to evade the controls, then there is a cycle between those who do fraud and those who wish to detect or prevent it, there is the "cat game and the mouse". This makes detecting fraud patterns very difficult. Furthermore, the fraud patterns depend on the type of transaction, since, for example, items such as telephone recharges and air tickets have different behaviour (amounts, periodicity, etc.). Coping with the change over time can take several months of transactions and also focus on only one type of transaction, to address the heterogeneity of behaviour. The market is also always changing with the creation of new businesses. This is another increasingly important source of noise. For example, in new businesses, there is no transactional history with which to compare and detect strange buying patterns.

2.4 Solution

Within the scope of this study, it will be tied to determine the best model required to prevent credit card fraud by using machine learning methods, which are among the

data mining techniques, in the most effective and efficient way. The purpose of this research work is to examine existing models against credit card frauds that may occur in such an intense environment of credit card use and to develop a more effective and efficient model method in addition to these. It is to try to reduce credit card fraud as much as possible by successfully applying this model method against a real problem in the future.

CHAPTER 3

THEORETICAL FRAMEWORK FOR CREDIT CARD FRAUD

3.1 Introduction

In general, the card market can be classified as a two-sided market, in other words, its characteristic is the existence of a platform that organizes and allows the meeting of two different groups of end consumers. In the card market, the groups of end consumers identified are the network of cardholder consumers and the network of merchants. Note that from the point of view of the credit card system manager, merchants are also consumers.

This structure is characterized by the presence of network externalities, which can be defined as follows: the value (utility) given by the final consumer in one of the groups depends on the number of existing consumers in the other group, and the number of opportunities to carry out intergroup transactions using the same platform. Thus, the platform is responsible for creating conditions for consumers on both sides to meet and carry out as many transactions as possible [50].

For the card payment system to become successful, a minimum number of cardholders must be attracted so that the platform is sufficiently attractive to commercial establishments, and the reciprocal argument is also valid, that is, there must be , also, a minimum number of commercial establishments to be attractive to cardholders.

In order for the two-sided market to be considered efficient, each group of consumers must pay the cost of the platform, discounted (or added) from the externalities, that is, each group internalizes (through the price) the network externalities generated in the card market.

Therefore, the two-sided market in which the card segment is inserted is defined by the interdependence between two markets through a platform in which a stimulating action in one of these markets has direct consequences for the other market.

3.2 History

Credit cards have increasingly become an essential financial instrument in people's lives, replacing checks, banknotes and coins as a means of payment. With a credit card

you can buy almost everything around you. This includes shopping at the supermarket, purchasing tickets to cinemas and games, filling up automobiles, and even purchasing more valuable products, such as cars.

With credit limits set by card issuers that often reach more than double their monthly income, cardholders carry out their purchases without the concern of knowing how much money they have in their wallets, in addition to the possibility of splitting purchases for higher amounts. One of the advantages of this payment method for merchants is the fact that they receive these amounts directly into the bank account, avoiding the handling of checks and cash, and, on the other hand, for the cardholder is the fact that expenses are concentrated in one place and the participation of loyalty programs, for example. The ease and speed of making a credit card transaction has contributed to the expansion of this payment method in several branches of activity.

According to information from the Reserve Bank of India, credit cards are not real money, they simply register the consumer's payment intention with their signature or password and other checks. In turn, the consumer will have to pay the card expenses on a predetermined date, by means of automatic debit, cash or check. In this way, the credit card is an immediate form of credit.

The following are the main historical landmarks related to the credit card market in the world, from its invention to the present day:

- **1887** - Edward Bellamy invents the concept of shopping using a card in his utopian novel Looking Backward;
- **1914** - In the United States, many retailers start issuing cards to their wealthiest customers in an attempt to build customer loyalty and increase sales of their most expensive products;
- **1928** - Retailers begin to issue a type of card that was embedded in a piece of metal;
- **1934** - American Airlines creates the Air Travel card for its customers;
- **1938** - Western Union begins to issue a charge card (Charge card 1) to frequent customers;
- **1941** - 50% of American Airlines revenue comes from the Air Travel card;

- **1948** - The Air Travel card becomes the first internationally accepted expense card;
- **1950** - The Diners Club card is created. Originally, it would be used to pay bills in restaurants, and later it would become a universal card that allowed the holder to buy services and goods in a variety of establishments;
- **1959** - The concept of revolving balance is introduced by Master Charge;
- **1966** - American Express creates its own credit card. Bank of America also decides to create its own brand called BankAmericard;
- **1970s** - Paper-based credit card systems are replaced by electronic systems;
- **1976** - BankAmericard changes its name to Visa in order to develop an international image;
- **1980** - Master Charge changes its name to Mastercard;
- **1986** - Department store Sears, a large American retailer, launches the Discover card;
- **1990s** - Credit card companies begin to offer rewards and incentives to encourage credit card use;
- **2005** - Paypal is founded, allowing you to receive and make payments over the internet with a credit card;
- **2010** - The company Square is created, allowing users to accept credit cards through cell phones;
- **2011** - The Google Wallet mobile payment system is launched, which creates the virtual credit card on mobile devices.

In India, the Central Bank of India launched the first credit card in 1980, then Andhra Bank in the same year, both under the Visa brand.

3.3 Credit Card Operation

In this section we present the agents involved in the credit card market, the structure of a card, the flow of a transaction, in addition to the security standards used in this market.

3.3.1 Credit Card Agents

Credit cards currently have five agents involved in their operation: bearer (card holder), merchant, acquirer, banner (card brand or association) and issuer.

The following is a brief description of these agents [51]:

- **Bearer:** It the individual who has the card, who is responsible for starting the system when deciding to pay for his purchases with his credit card.
- **Establishment:** Any company, natural person or accredited legal entity that accepts the credit card through specific equipment.
- **Acquirer:** The function of the acquiring company is to accredit, supervise and pass on the purchase amounts to establishments that accept credit cards, in addition to being responsible for the implementation and maintenance of the models, called POS (point of sales), and the programs of capture of transactions.
- **Issuer:** In general, issuers are banks, responsible for the distribution of credit cards to their customers through the approval of credit risk by each institution's own policies. Until 2005, in India, most credit cards belonged to issuers independent from banks. The issuers 'main revenues come from customers' revolving financing, annuities and insurance or services added to the credit card product.
- **Brand:** Responsible for defining the policy rules (relationship between issuers and acquirers), operations of the global communications network, institutional marketing executions and research and development of new technologies and services. The main sources of revenue for the brands are fines imposed on customers for non-compliance with rules and deadlines and a percentage of the fee charged to establishments per transaction, known as MDR (Merchant Discount Rate). Examples of brands are: Visa, MasterCard, Amex and Hipercard.

3.3.2 Credit Card Brands

The following are the main credit card brands in the market, they are:

- **Visa:** It is an association of 21,000 financial institutions worldwide, which issue the Visa card. There are 1.3 billion Visa cards in circulation, accepted at more than 24 million merchants in more than 150 countries. In 2005, the volume of transactions generated by these cards was 3 trillion dollars.
- **Mastercard:** It has more than 25,000 issuing partners in the world. There are approximately 720 million Mastercard cards in circulation, accepted at 32 million commercial establishments and in more than 210 countries. In 2005, the volume of transactions generated by Mastercard cards was approximately 1.2 trillion dollars.
- **American Express:** It is an institution founded in 1850, but the first card was issued only in 1958. The approximately 57 million Amex cards in circulation are accepted in more than 200 countries. In 2005, Amex cards generated around 150 billion dollars in transactions.

Other card brands spread around the world, with greater or lesser concentration, depending on the region, are: Diners Club, JCB (Japanese Credit Bureau), Discover and Solo. In India, there are also some national brands with good publicity, among which we can mention the Yatra SBI Credit Card, IndianOil Citi Platinum Card, HDFC Freedom Credit Card and HDFC Bank Diners Club Black Card etc.

3.3.3 Trends for the Future

In this subsection we list some of the main trends for the card market observed around the world:

- **Mobile Payment:** the acceptance of credit card payments through smartphones and tablets tend to increase due to the expansion of these devices.
- **Approximation Card:** also known as a contactless card, it adds the functionality of carrying out transactions when the card is approached at the merchant's terminal. Using RFID (Radio Frequency IDentification) technology or NFC (Near Field Communication). This technology seeks to optimize the time of the payment process, which would eliminate long lines in establishments with large flows of people, such as restaurants and coffee shops.

- **Smart Card:** also known as a display card, this card gathers chip and dynamic password (token) technologies, which change after a certain period of time. This technology may allow greater security in non-essential transactions with the use of the token to authenticate the cardholder.
- **Dynamic Card:** a single physical card that has a chip that controls several numbers of cards. There will be a small keyboard to access multiple cards safely.

3.3.4 Credit Card Structure

The credit card number is usually 16 digits and is described as part 1 of ISO / IEC 7812 published by the International Organization for Standardization (ISO). Credit card brands such as Visa and Mastercard have adopted the following card structure:

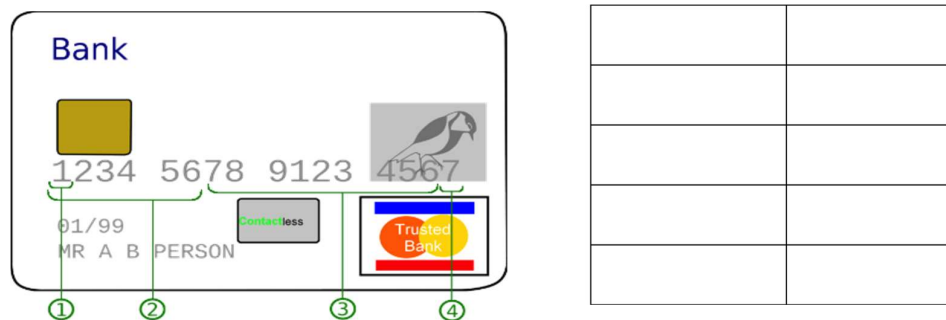


Figure 3.1: Credit card structure

The first digit of the credit card number is the industry identifier MII (Major Industry Identifier), which represents the category of the entity that issued the credit card to the holder. Different MIIs represent the following industries:

Table 3.1: Industry identification

MII digit	Category
0	ISO / TC 68 and assignments from other industries
1	Airlines
2	Airlines and other industry assignments
3	Travel and entertainment (such as American Express and Diners Club)
4	Financial and banking (Visa)
5	Financial and banking (MasterCard)
6	Merchandizing and banking (Discover)
7	Petroleum
8	Telecommunications and other industry assignments

9	National allocation
---	---------------------

The first six digits are known as the Issuer Identification Number (IIN) or Bank Identification Number (BIN). They identify the financial institution that issued the card to the holder. We can say that some of the most popular brands are identified by the first digit, such as Mastercard and Visa, digits 5 and 4, respectively (Table 3.2).

Table 3.2: Brand identification

Brand	Identifier	Number of digits on the card
American Express	34XXXX, 37XXXX	15
Visa	4XXXXX	13, 16
Mastercard	5XXXXX	16
Discover	6011XX	16
Diners Club/Carte Blanche	300xxx-305xxx, 36xxxx, 38xxxx	14

In addition, a curious fact about the credit card is its geometric shape. The card has a rectangle shape, very close to the gold rectangle. In the gold rectangle, dividing the width by the height, the result will be the golden ratio.

3.3.5 Step by Step Transaction

A transaction is a purchase that begins with the cardholder. The following flow briefly describes the card transaction process:

- **Start of the transaction:** The cardholder purchases goods or services from an establishment, which then sends the transaction information to the acquirer;
- **Authentication:** The establishment, through the acquirers' systems, captures the transaction, a fee for the service is deducted from the transaction amount;

- **Submission:** The acquirer submits the transaction, which passes through the banner systems beforehand, to the card issuer. The transaction can be approved, denied or referred by the issuer through credit and fraud analysis;
- **Payment to the merchant:** the card issuer pays the acquirer who, in turn, makes the payment to the merchant;
- **Payment to the issuer:** finally, the cardholder pays to the issuer the amounts corresponding to the goods or services he has purchased from the establishment.

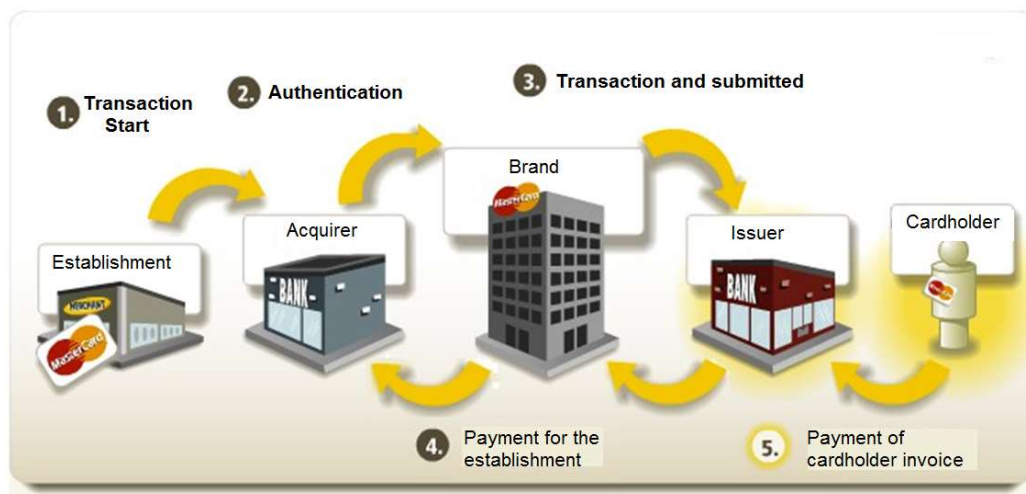


Figure 3.2: Transaction flow; Source- Mastercard

There are many details of credit card authorization systems that incorporate sophisticated communication techniques, parallel algorithms, encryption and algorithms to prevent intruders. These systems seek to be the safest against intrusion compared to existing banking systems. Another important feature of the authorization system is the high availability, which through some business intelligence, such as STAND-IN (authorization of the transaction by the brand when the issuer system is not available), keeps the system available for much longer in relation to other banking systems [51].

3.3.6 Internet Transactions (Electronic Commerce or E-commerce)

The idea of paying for goods and services electronically is not a new one. Since the 1970s and early 1980s, a variety of structures have been proposed to allow payment

to be made through a computer network. After a period of exponential growth, it is estimated that more than 2.7 billion people have access to the internet in 2013, this represents 39% of the world population that year.

The electronic payment system started in late 1996 and the first part of 1997, with a huge variety of payment methods. Some of them were launched on the market, however, failed to reach a large population.

With the growing interest in electronic commerce, electronic payment techniques have increased more and more. Credit card payment is the most popular form probably because of its simplicity and convenience for the buyer. The user simply enters the card details (usually card number, validity, name of the holder and security code), the merchant validates the information and, with the approval of the card issuer, sends the goods or provides the requested service.

3.4 Credit Card Fraud

3.4.1 Definition

The Houaiss dictionary of the Portuguese language defines fraud as "any cunning, deceptive, bad-faith act, with the intention of harming or deceiving others, or of failing to fulfill a certain duty; deception".

In the world of credit cards, in a general and simple way, fraud can be characterized by a transaction that is not recognized by the legitimate credit card holder.

When financial institutions lose money due to fraudulent credit card transactions, their holders pay partially and indirectly for it through higher interest rates, reduced benefits and annuities. Thus, the interest of financial institutions to start developing fraud detection systems and processes is aroused [48].

The ease of access to credit cards and the increased use of them have attracted criminals interested in illicit gains arising from fraudulent transactions. One of the biggest attractions is to win large amounts of money in a short time without being exposed to big risks. This is because criminals are rarely discovered and imprisoned for a long time in current Indian legislation for this type of crime.

Fraudsters, in general, are well organized and always look for the easiest and cheapest way to obtain advantages. According to Hand et al. [52], fraud is a profitable, stable and very well organized and managed “business”.

One question that emerges is: why do people commit fraud? There is no single reason behind the fraud and any explanation needs to take several factors into account. Looking from the fraudster's point of view, it is necessary to take into account the following points according to Doody et al. [53]:

- Motivation of potential fraudsters;
- Conditions for understanding / rationalizing crime;
- Opportunities to commit the crime;
- Identification of the target to commit the fraud;
- Technical skill of the fraudster;
- Expected and actual risk of discovery after the fraud was carried out;
- Expectations of consequences of the discovery (including non-criminal consequences, such as: loss of job and stigma by family and friends).

A model that brings together a series of these aspects is the Fraud Triangle. This model is based on the premise that fraud can result from a combination of three factors: motivation, opportunity and rationalization, described below:

- **Motivation:** It is typically based on any greed or need. Greed has been shown to be the main cause of fraud events and other causes cited were debt and gambling problems.
- Many people are faced with the opportunity to commit fraud, but only a minority of the greedy and needy commit it. It also adds up how much people fear and are frightened by the consequences of taking risks to execute a given fraud. Some people with good principles and goals can fall into bad company and develop a taste for quick money, which guides them on the path of fraud. Others are tempted only when they face difficult times in their lives.
- **Opportunity:** In terms of opportunity, fraud is more likely in companies where there is a weak internal control system, lack of security over company assets,

absence of detection and punishment systems, as well as unclear policies regarding respect to acceptable behavior. Research shows that some employees are totally honest, some are totally dishonest, but that many are seduced by the opportunity to commit fraud.

- **Rationalization:** Many people obey the law because they believe in it and / or are afraid of being humiliated or rejected by people if they are caught in a fraudulent act. However, some people try to rationalize fraudulent actions like:
- **Necessary:** especially when done against the company that operates;
- **Harmless:** because the victim is big enough to absorb the impact;
- **Justified:** because the victim “deserved it” or because the fraudster believed he was being treated badly.

One of the most effective ways to tackle the problem of fraud is to adopt methods that will decrease opportunities for the fraudster. Motivation and rationalization are aspects and personal choices of the individual and more difficult to combat in a comprehensive and efficient manner. However, if the company has a strong culture and clear values, this can become a useful prevention tool.

The performance of a fraudster can be classified into one of the following categories:

- **Pre-planned fraudsters:** they start from the beginning with the intention of committing fraud. These can be short-term fraudsters, like many who use stolen credit cards or fake ID numbers. Or they can be long-term, like fraudsters who run complex money laundering schemes.
- **Intermediate Fraudsters:** start out being honest and become fraudsters when times get tough or when some kind of event occurs in their lives, such as irritation at not being chosen for promotion or the need to pay for medical care for a family member .
- **Indebted fraudsters:** they simply continue to negotiate, even when, objectively, they are unable to pay their debts. This type of fraudster profile can be seen both in ordinary traders, as well as in large entrepreneurs and entrepreneurs.

3.4.2 Types of Fraud

Among the types of fraud committed in credit card transactions, the following are observed most frequently:

- **Account hacking:** involves criminals who obtained personal information from customers, such as account numbers and passwords. These criminals try to take possession of customers' bank accounts or credit card numbers. The fraudster, when impersonating the real customer, usually requests payments, loans and other banking products that are available, as well as changing registration information to request credit cards;
- **Lost or stolen card:** In this type of fraud, fraudsters or people with bad intentions seek to carry out transactions by posing as the real cardholders, who have their cards lost or stolen;
- **Counterfeiting (Skimming):** The counterfeiting or cloning of a card happens when a card is created, with the information of another card that already exists, without the authorization of the card issuer. The information for creating this "clone" card is often obtained through the magnetic track of the real card;
- **E-commerce / MOTO (Mail Order Telephone Order):** it is a type of fraud carried out via the internet, telephone, and fax or by letter, where the card is not physically present. Most of the time, to carry out this type of fraud, the card data was obtained physically by copying the card's magnetic track or leaking information from previous transactions that were stored by establishments that operate in electronic commerce.
- **Loss:** happens when credit cards are stolen in the process of sending the card issuer to the holder. The risks for this type of fraud involve delivery companies known as couriers and the national postal system;
- **Identity theft:** it is in most cases related to organized criminal groups that have information from legitimate and legitimate people. They make use of such information to obtain financial products and earnings through them. Basically, it is the illegal use of third party information to carry out fraudulent transactions. In some situations, information is stolen through sophisticated

computational methods, such as using Trojan horses and phishing, as well as through social engineering.

3.4.3 Methods to Commit Fraud

The methods for committing fraud are renewed or created by fraudsters at all times. However, there are some methods that are repeated and are better known, such as those presented below:

- **Cloning:** The magnetic stripe information on the card is copied through devices known as skimmers. After that, another card is produced with the same information.
- **Dumpster Diving:** Fraudsters look in trash cans, or other types of material dispensers, for personal information that can be used for fraud.
- **Correspondence Theft:** Fraudsters search mailboxes, steal postmen, or solicit courier employees to obtain new credit cards and invoices.
- **Internal Fraud:** A dishonest employee with access to personal information, payroll or number of accounts can trade this type of information with third parties to carry out fraud.
- **Impostors:** People who use other people's identities (proven by false documents) to open accounts, apply for credit cards and loans.
- **Online Data:** Fraudsters access information available in public databases, social networks, credit bureaus, among other sources of information.
- **Malicious programs:** There are several types of malicious programs, however, all seek to steal personal and financial information stored on personal computers. Among them, the following stand out: the computer virus (replicated between computers), Trojan horse (installed with other programs, but captures information from the keyboard and can control the computer) and spyware (gathers information from the hacked computer) and transmits to an external entity without having control of the computer).
- **Phishing:** Criminals obtain personal data through e-mails and fake websites from companies known in the market, such as financial institutions and e-

commerce. Usually, the victim receives an email from a financial institution asking him to fill in some personal information on a fake website.

- **Access to personal documents at home:** Identity theft occurs through people who have access to personal information about the homeowner and its residents.
- **Wallet / Purse Theft:** It is usual for wallets and purses to contain credit cards and identification documents, and fraud can occur when they are stolen by criminals.
- **Hacking:** Criminals have the ability to break into e-commerce systems and databases, credit card processors, payment service providers, among other institutions, to obtain personal information from consumers and financial transactions.
- **Social Engineering:** Criminals contact clients of financial institutions posing as employees of the same. Most of the time, they use customer information posted on social media to make them believe that they are really in touch with the financial institution. After this convincing step, customers end up informing the number of credit cards, bank account numbers, passwords and other types of information.

3.4.4 Some Types of Attacks

Carding is a term used for a process that checks the validity and security code of a stolen card. The thief presents the card information on a website that processes and validates the information in real time. If the card is successfully processed, the thief knows that the card is still valid. Often, the fraudster seeks to test these cards in electronic shops that sell services, games or low-value products, in order to avoid drawing attention from the issuer's prevention and detection systems and not compromising the card limit.

BIN Attack: credit cards are produced in several BIN categories. When the issuer does not use random card number generation, that is, it produces card numbers sequentially, the fraudster can obtain a real card number and generate valid card numbers that were produced in the sequence.

3.4.5 Costs and Fraud Cycle

In the challenge of finding the balance between avoiding fraudulent transactions and inconveniencing the legitimate cardholder as little as possible, there are tangible and intangible costs, which are listed below [54]:

Tangible Costs of Fraud (Financial)

- Financial loss due to card misuse;
- Cost of investigation and arrest of the fraudster;
- Card reissue and delivery;
- Customer service calls;
- Exchange (requests and disputes with the brands): In some cases it is possible to recover the financial loss, but this process has a cost;
- Cost of the referred transactions: It is a tariff paid to the brand for each denied transaction, due to suspicion of fraud, when it exceeds a certain cut-off point;
- Card cancellation cost: Customers cancel their cards and migrate to the competition;
- Costs of protection bulletin with the brands: Cost for inclusion and maintenance of card numbers in the “negative list” maintained by the brands;
- Potential reduction in revenue;
- Reduction of market share: Reduction of the size of the institution and loss of position for some competitor.

Intangible costs of fraud (non-financial)

- Customer dissatisfaction;
- Feeling of violation and vulnerability in relation to the company;
- Threat to the business and the people connected to it (employees and third parties);
- Loss of brand and brand loyalty;
- Opportunity Cost: The investor (partner) of the institution could be investing their money in something less risky from the point of view of fraud, and may obtain a higher return on investment (ROI).

In addition, the fraud cycle can be described in eight stages:

1. **Intimidation:** Characterized by actions designed to inhibit or discourage the fraudster before the fraud is executed;
2. **Prevention:** comprises activities that make the execution of fraud more difficult, hardening the defenses against fraudsters. This includes the personal identification of numbers for credit cards, security systems for transactions over the internet and the use of personal passwords for accessing accounts in the system, both via computers and by telephone. These methods are not perfect and it is necessary to establish a middle ground between the expenses involved in the management of fraud and the inconvenience it can generate for the client;
3. **Detection:** set of actions and activities that are used to identify fraud. These detection methods are used when fraud prevention methods fail. In practice, fraud detection is used continuously, ignoring prevention;
4. **Measures:** taking measures that prevent the occurrence of losses or their continuity and / or prevent a fraudster from continuing to defraud or terminate his fraud activity;
5. **Analysis:** the factors that lead fraudsters to commit fraud are identified and studied, through statistical modeling;
6. **Policy:** set of activities that intend to create, evaluate, communicate and help in the implementation of policies to reduce the incidence of fraud;
7. **Investigation:** involves obtaining sufficient evidence and information to reduce or completely inhibit fraudulent activities, in order to recover resources or obtain their refund;
8. **Prosecution:** at this stage, the need for legal support to convict criminals is clear.

Through a balanced interaction between the stages mentioned above, very efficient results can be achieved. Each institution must find the best balance for its business. These stages interact with each other dynamically and are not necessarily executed in the order previously described.

3.4.6 Measures to Prevent Credit Card Fraud

In view of all this risk scenario in credit card transactions, there are some preventive measures that contribute to prevent fraud from taking place or to discourage fraudsters' actions. These measures are presented below:

- Check on the monthly invoice that all transactions are recognized. Report unrecognized transactions to the card issuer;
- Never send payment information by email, such as password and card number. Information that travels on the Internet can be read by third parties;
- When providing payment information on the internet, look for a lock icon. This signal indicates that the information traffic will be secured using data encryption methods;
- Protect the information on your computer: Use antivirus software, spyware filters and e-mails, as well as e-wall programs to keep the information on your computer safe;
- Do not respond to emails or phone calls that request personal information, such as passwords, credit card numbers or account numbers;
- Never divulge your password to anyone. No employee of financial institutions and establishments should ask, verbally, for their password;
- When possible, always keep your eyes on your card at the time of the transaction, in order to prevent card data from being copied / cloned;
- If your card is stolen or lost, report it immediately to the financial institution that issued it;
- Destroy all personal information, such as offers for new credit cards, ATM receipts and invoices before throwing them away.

3.5 Fraud Prevention and Detection Methods

3.5.1 Definition

In general, fraud detection is a forecasting problem. Its objective is to maximize the correct forecast and maintain incorrect predictions at an acceptable level of costs [55].

In order to understand fraud prevention and detection methods, we must first distinguish what is considered prevention and detection [56].

Fraud prevention consists of a set of measures and processes to prevent fraud from occurring. These measures and processes consist of actions such as including watermarks and holograms in documents, dynamic passwords, passwords for credit cards, among other measures. There is certainly no perfect method. However, the cost and inconvenience to the customer must be weighed against the effectiveness of the method.

On the other hand, fraud detection involves actions to detect fraud quickly once it has happened. It is defined as the process of identifying fraudulent transactions in the set of all transactions. In other words, the detection process is the process of classifying transactions into two classes: legitimate transactions and fraudulent transactions.

Prevention and detection are closely related, since fraud detection comes into play once prevention processes fail. It is a process in continuous development, because whenever a detection method is known or discovered by criminals, they will adapt their strategies to try other types of fraud. Many of them will not be aware of the fraud detection methods that have been successful in the past, and will look for new strategies that lead to unidentifiable fraud for a while.

One of the major problems associated with the development of new methods of fraud prevention and detection is the lack of literature that provides experimental results and real data for researchers to carry out new experiments. This is because the fraud prevention and detection process is associated with sensitive financial information from customers, which must be kept in a safe and private environment. In addition, it is not recommended to describe fraud prevention and detection techniques in great detail in the public domain, as this gives criminals the information they need to avoid detection [57].

3.5.2 Detection Systems

One of the difficulties with fraud detection is that many legitimate cases usually need to be analyzed to identify a fraudulent transaction, which generates a considerable

cost for this analysis. In practice, a balance must be struck, often an economic assessment, between the cost of detecting fraud and the savings to be made to detect it.

This leads us to a more general conclusion: fraud can be reduced to a very low level, but only because of a high level of effort and cost.

As noted by authors of [48], to produce good results, an efficient fraud detection system must be able to:

- Work with unbalanced distributions, as only a small percentage of transactions are fraudulent: To solve this problem, a set of training data is often divided into parts whose distribution is less unbalanced [58];
- Control noise: This means controlling the presence of data errors, such as incorrect dates and values. One way to solve this problem is to structure a data cleaning process [59];
- Adapt to new types of fraud: A known method of fraud becomes less efficient when it becomes better known, which produces new attempts to carry out the fraud;
- Generate good indicators to evaluate the system's performance in classifying legitimate and fraudulent transactions;
- Take into account the cost of detecting a transaction with fraudulent behavior and the cost of avoiding it.
- On the contrary, the following reasons have created difficulties for the construction of an effective fraud detection system [60]:
- Financial institutions do not share fraud information for reasons of competition and legal reasons;
- The transaction databases are huge and the number of records is growing rapidly, which requires a scalable and very well structured system;
- Real-time analysis is highly desirable for updating models when a new fraud event is detected;
- Easy distribution of models to keep the detection process updated;

- Data overlap is another type of problem. Many good transactions may resemble fraudulent transactions, and vice versa.

Processing this volume of data in search of fraudulent transactions requires more than statistical models, requires fast and efficient algorithms, and data mining techniques are also relevant. This volume of data indicates the potential value of fraud detection: if 0.1% or 10 basis points (basis points) of 100 million transactions are fraudulent, and the average transaction value is \$100, then it is estimated that the company will lose \$10 million due to fraud.

Another way of thinking is as follows: if a financial institution has 2% revenue from each transaction, to disburse \$100 in fraud, it will have to have a transaction in the amount of \$5,000 to equal the amount of the fraud. Not to mention the loss of operating costs involved in the transaction.

3.5.3 Statistical Methods

The statistical methods frequently used for fraud detection can be classified as unsupervised or supervised [56]. We detail the following two types.

3.5.3.1 Unsupervised Methods

Unsupervised methods seek information from accounts, customers, and card numbers, among others, that behave differently than normal, and that are often referred to as anomalies or outliers and characterized as a basic non-standard form of observation.

Tools for data quality control can be used, but the detection of accidental errors is a problem quite different from the detection of deliberately falsified data or data that accurately describe a fraudulent pattern.

The analysis process consists of modeling a reference distribution, which represents normal behavior, and then an attempt to detect observations that show deviations from that distribution. Digit analysis, using Benford's law, is an example of such a method [61].

3.5.3.2 Supervised Methods

In supervised methods, samples of fraudulent and legitimate cases are used to build models that allow assigning new observations to one of the two classes. This certainly requires discriminating a set of variables that can correctly classify, among these classes, the original data used to build the models. In addition, it can only be used to detect fraud that has previously occurred in a given period of time.

Traditional statistical methods, such as discriminant analysis and decision tree, have proven to be effective tools for fraud detection [62] [63].

However, more powerful tools, such as neural networks and binary logistic regression, have been widely used [64] [65].

In the set of supervised methods, two of them have traditionally been adopted by most financial institutions to detect fraud: systems based on rules and scoring models.

3.5.3.3 Rules-based Systems

Rules-based systems are supervised learning algorithms that produce classifiers using rules as follows:

If (certain condition is true) Then (do certain action).

One of the advantages of this method is the easy configuration and agility to place new rules in a production environment. However, it requires frequent and expert updates for the development of rules, in addition to reflecting a limited standard.

This strategy can be used to segment customers and products, focusing on segments with a higher risk of fraud, such as cards with very high limits. Below, we list positive and negative points associated with rule-based systems.

Positive Points:

- Dynamic update (usually a new rule enters the production environment in a few seconds);
- Ease of development and deployment;

- Control;
- Low cost and speed.

Negative Points:

- Requires frequent updating;
- Large volume of rules;
- Need for experts to develop rules;
- Reflects a limited pattern;
- Difficult to understand the relationship between the rules and duplication of rules.

3.5.3.4 Scoring Models

Scoring models use statistical techniques to return a score (score) for a given transaction. Generally, the higher the score, the more likely (suspicious) that a transaction is fraudulent.

The score can be computed for each transaction in the database and used in the prevention process to approve, deny or refer a transaction, as well as be used in rules-based systems in combination with other variables. With the score value, the highest scoring cases can be prioritized in the process of investigating the transaction.

At this point, cost issues are considered, as it is very expensive to conduct a detailed investigation of all cases. An investigation should focus on cases that are most suspected of fraud.

In general, binary logistic regression models and neural networks are used to generate this score. On a scale of 0 to 100, or 0 to 1000, the probability of a transaction being fraudulent based on characteristics such as time of the transaction, industry, value, among other variables, is measured. Cut-off points are adopted to adapt the capacity to handle the volume of cases / alerts generated in work queues. Below, we highlight positive and negative points associated with the scoring models.

Positive Points:

- Used by the entire financial industry due to its effectiveness in the decision;
- Ideal for large volumes of transactions in which the decision needs to be made quickly;
- Covers profiles of individual behaviors.

Negative Points:

- Does not keep up with recent fraud trends;
- Depending on the model, it may not reflect characteristics of local fraud;
- High cost;
- No control for quick changes;
- It can generate results that cannot be explained.

Although scoring models show good effectiveness in detecting fraud, operate with large volumes of transactions and cover profiles of individual behaviors, the issue that emerges is the frequency with which these models are updated in order to keep up with recent fraud trends. . In this study we will work on the problem of updating these models in order to reflect the characteristics of fraud over time.

3.5.3.5 Traditional Statistical Techniques for Fraud Detection

Some statistical techniques, such as neural networks, decision tree, regression, among others, have been shown to be effective in solving the classification problem that exists in the fraud detection process. In the work of Al-Khatib [55], there is a summary of the advantages and disadvantages of the traditional statistical techniques used to detect credit card fraud, which are presented in Table 3.3 below.

Table 3.3: Comparison of statistical techniques for detecting credit card fraud [55]

Technique	Advantage	Disadvantage
Neural networks	Effective for handling noise data (a set of data that is apparently inconsistent with the rest of the existing data) for identifying patterns, solving complex problems and processing new	Slight explanatory capacity, less efficient in the processing of large data sets, difficult to configure and operate, sensitive to the data format, different data representations can produce

	instances. The code can be generated to be used in real-time, highly accurate, portable, fast systems and has better performance than other techniques.	different results. In addition, it only works with numeric data with values between 0 and 1; non-numeric data needs to be converted and normalized.
Decision tree	Scalable, high forecasting accuracy, easy to use, easy to explain the results, easy to interpret the rules created by the tree and easy to implement in applications.	It is not easy to deal with continuous data, difficult to deal with missing data, over-fitting problems can occur, the size of the data and attributes to be used for the division in the construction of the tree can impact performance.
Rule-Based Systems	Easy to modify, easy to develop, build and implement. It has a high degree of accuracy, is easy to explain and performs well. Rules from other techniques, such as neural networks and decision tree, can be extracted, modified and stored in the system.	Poor in dealing with missing data or unexpected data values.
Past Case-Based Systems	Useful in a domain that has a large number of examples, has the ability to work with incomplete or noisy data, effective, flexible, easy to update and maintain, can be used in a hybrid approach.	You may suffer from the problem of incomplete or noisy data.
Genetic algorithms	It works well with noisy data and is easy to integrate with other systems. Often combined with other techniques to increase their performance.	Requires in-depth technical knowledge to install and operate.
Inductive Logic Programming	It has powerful modeling language that can model complex relationships.	It has low predictive power, is very sensitive to noise and its performance deteriorates rapidly in the presence of outliers.
Regression	Easy to understand, easy to build and used with two classification classes.	Poor with noise or outliers, not applicable to complex applications, does

	Logistic regression may be more accurate than learning techniques for small data sets.	not work well with non-numeric data, accuracy is good, but not high.
--	--	--

In the modeling process, it is worth mentioning that the following factors can significantly impact the performance of statistical techniques:

- **Incorrect Data:** the lack of accuracy in the data and inconsistency directly impact the predictive power of the technique used;
- **Missing Data:** can cause problems in the process of classifying relevant variables for the model. There are techniques that are used to treat this type of problem;
- **Scalable System:** generally large databases are used in the modeling process. Even using fast-running algorithms, computers with scalable systems are necessary to maintain the process in the long run;
- **Unbalanced Distribution:** data for detecting fraud is often unbalanced, which requires appropriate techniques to deal with this characteristic.

3.5.4 E-Commerce Transaction Fraud

When the credit card is not present at the point of sale, a set of different challenges is created when it comes to fraud. In this scenario, the problem of authenticating transactions emerges, that is, the ability to verify whether the buyer is actually the cardholder or whether the purchase was authorized by the same.

Purchase orders for products, services or subscriptions flow anonymously from a computer, located somewhere in the world, and the online store cannot authenticate with 100% certainty that the card number shown is associated with the buyer's information. At this point, there are some solutions offered by the market, such as 3D Secure and monitoring systems.

3D Secure is an XML-based protocol designed to be an additional layer of security for credit and debit transactions on the Internet. It is a service developed by the brands and offered to the establishments through the purchasers. From the moment the

merchant contracts this service, the burden of the fraudulent transaction becomes the card issuer. However, the establishment's approval rate tends to decrease and, consequently, the number of sales decreases. This is because the card issuer is taking the risk of the transaction and tends to evaluate 3D secure transactions more carefully.

The monitoring systems, with internal or external development of the establishment, seek, in general, to analyze the characteristics of the buyer's order, computer data, purchase history, among other variables, to decide whether or not to proceed with the approval of the order purchase. In this scenario, the merchant is responsible for any fraudulent transactions that it may suffer and approval rates are maintained at high levels. It is the opposite of the scenario observed when using the 3D Secure service. Issuers tend to monitor these transactions with less priority just to maintain acceptable levels of fraud for the brands.

Another fact is fraudulent transactions on the internet that are classified as "friendly fraud". A transaction can legitimately be initiated by the credit card holder, however, when confronted by someone, often family members, about the purchase of a product or service, the cardholder denies knowledge of the transaction and reports it as being a fraud. This type of situation often happens when shopping for services like pornography.

CHAPTER 4

PROPOSED METHODOLOGY

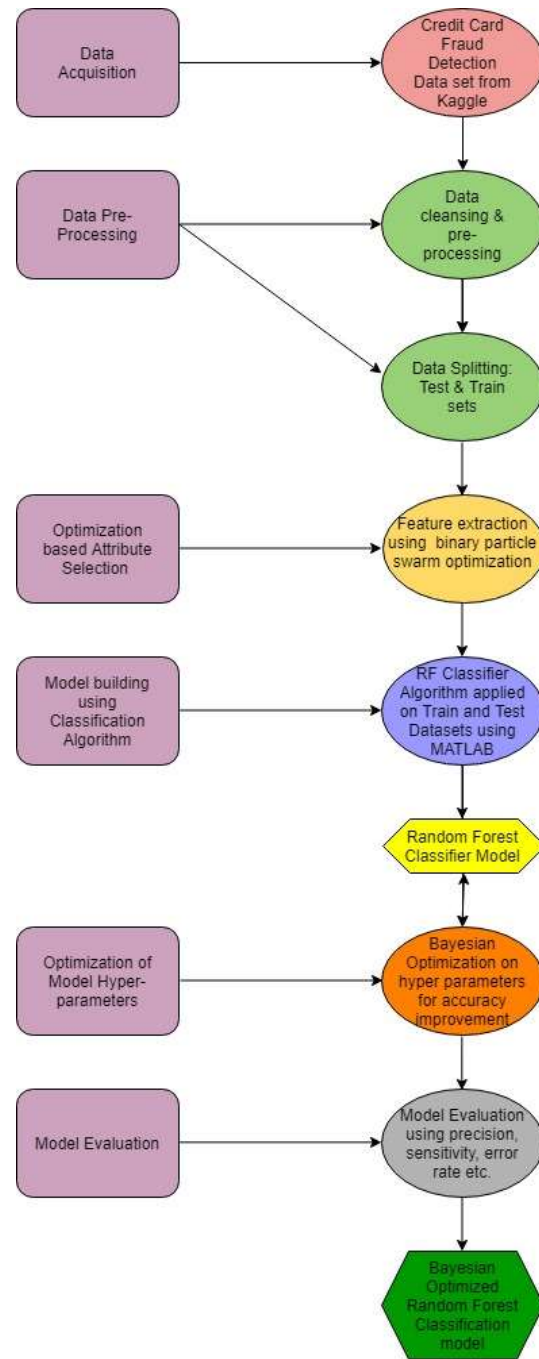


Figure 4.1: Flow diagram of proposed research work

4.1 Dataset Description

Kaggle dataset [66] includes credit card transactions made in September 2013 by European cardholders. This database shows actions completed in two days and 492

out of 284,807 transactions. Inaccurate database, positive type (frauds) accounts for 0.172% of all transactions [66].

4.2 Pre-Processing

- Create a categorical response variable where:

1= fraud activity

0 = normal or non-fraudulent activity

- Time attribute removal from the categorical class

4.3 Feature Selection using Binary Particle Swarm Optimization

Particle swarm optimization is an intuitive method. It is an advanced algorithm with the inclusion of some herd behavior in nature. The particle swarm optimization algorithm that occurs by observing the behavior of birds, fish, and bees is population-based [67].

Binary Particle swarm optimization was introduced by Kennedy and Eberhart [67]. If the elements of the problem can be sorted or grouped, binary particle swarm optimization can be used to solve such discrete problems [68]. Research on many optimization problems, such as path problems or scheduling, takes place in a discrete space.

For the research area $S = \{0,1\}^D$, the fitness function f maximizes i.e. $(\max f(x))$. The i^{th} particle in D dimension is defined as:

$$X_i = (x_{i1}, x_{i2}, \dots, x_{id})^T, x_{id} \in \{0,1\}, d = 1,2, \dots, D \quad (4.1)$$

The velocity vector in D dimension can be represented as:

$$V_i = (v_{i1}, v_{i2}, \dots, v_{id})^T, v_{id} \in [-V_{max}, V_{max}], d = 1,2, \dots, D \quad (4.2)$$

Where V_{max} is the maximum velocity vector.

The previous best position can be given as [68]:

$$p_i = (p_{i1}, p_{i2}, \dots, p_{id})^T, p_{id} \in \{0,1\}, d = 1,2, \dots, D \quad (4.3)$$

Definitions according to the given notation can be given as:

Equation of Velocity:

$$v_{id} = v_{id} + c_1 rand_1(p_{id} + x_{id}) + c_2 rand_2(p_{gd} - x_{id}) \quad (4.4)$$

Equation of position:

$$X_{id} = \begin{cases} 1 & \text{if } U(0,1) < \text{sigm}(v) \\ 0 & \text{otherwise} \end{cases}, d = 1,2, \dots, D; i = 1,2, \dots, N \quad (4.5)$$

Transfer function:

$$\text{sigm}(v_{id}) = \frac{1}{1 + \exp(-\lambda v_{id})} \quad (4.6)$$

g : index of the best performing particle

p_{gd} : best part

N : the width of the fortification

c_1, c_2 : social and cognitive component constants

$rand_1, rand_2$: $U(0,1)$ random numbers

$\text{sigm}(v_{id})$: sigmoid transform function

4.4 Classification Algorithms

4.4.1 Random Forest Classifier

Random forests were introduced by Breiman (2001) by the following very general definition [69]:

A Random Forest is a classifier comprising a set of elementary classifiers of the decision tree type, noted:

$$\{h(x, \Theta_k), k = 1, \dots, L\} \quad (4.7)$$

Let $(\hat{h}(\Theta_1), \dots, \hat{h}(\Theta_q))$ a collection of tree predictors, with $\Theta_1, \dots, \Theta_q$ random variables independent of \mathcal{L}_n . The predictor of random forests \hat{h}_{RF} is obtained by aggregating this collection of random trees as follows:

$\hat{h}_{RF}(x) = \frac{1}{q} \sum_{l=1}^q \hat{h}(x, \Theta_l)$ Average of individual tree predictions in regression.

$\hat{h}_{RF}(x) = \arg \max_{1 \leq k \leq K} \sum_{l=1}^q 1_{\hat{h}(x, \Theta_l)=k}$ Majority vote among individual predictions trees in classification.

The term random forest comes from the fact that individual predictors are, here, explicitly predictors per tree, and that each tree depends on an additional random variable (that is, in addition to \mathcal{L}_n).

The classification process followed by Random Forest consists of:

- Assignment to a node if it is terminal, deciding whether a node will be labeled as a leaf or it will carry a test.
- If the node is not terminal, then we have to select a test to assign it.
- If the node is terminal, then we must give it a class.

The general algorithm for decision trees is as follows:

Input: sample S

Initialize the current tree to the empty tree;

The root designates the current node

Repeat

See if the current node is terminal

If the node is terminal then

Assign it a class

If not

Select a test and generate as many new child nodes as there are answers to this test

End if

Explore another node if there is one

Until a decision tree A is obtained

Exit: decision tree A.

4.4.2 Bayesian Optimization of Random Forest Classifier

The algorithm calculates conditional probabilities. Based on the information observed, it calculates the probability of unobserved data. For example, according to the symptoms of a patient, the probabilities of the different pathologies compatible with his symptoms are calculated. We can also calculate the probability of unobserved symptoms, and deduce the most interesting complementary examinations.

A direction for Bayesian optimization is to optimize continuous and mixed (discrete and continuous) variables in solving problems with various types of data. The main objective of using Bayesian optimization here is to find the suitable value for each parameter of random forest algorithm. There are at least three important practical choices that we need to consider: the covariance functions, selection of its hyper parameters and the acquisition functions. A default choice of covariance function is to use squared exponential kernel. Automatic relevance determination (ARD) Matern 5/2 kernel is used for the same [70] [71].

$$K_{M5}(x, x') = \theta_0 \left(1 + \sqrt{5r^2(x, x')} + \frac{5}{3}r^2(x, x') \right) \exp \left\{ -\sqrt{5r^2(x, x')} \right\} \quad (4.8)$$

The above kernel function itself has few parameters that needs to be managed (such as covariance amplitude θ_0 and the observation noise ν). It can be done by marginalize over hyper parameters and compute the integrated acquisition function.

CHAPTER 5

SIMULATION RESULTS

5.1 Evaluation Parameters

Once you have built your model, the most important question that arises is how good is your model? Therefore, evaluating your model is the most important task in the project that outlines how good your predictions are.

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data whose actual values are known.

True Positive: These are the correctly predicted positive values, which means that the value of the actual class is yes and the value of the predicted class is also yes. For example, if the real value of the class indicates that the passenger survived and the expected class tells the same.

True Negative: These are the correctly predicted negative values, which means that the value of the actual class is no and the value of the predicted class is also no. For example, if the actual class says that this passenger did not survive, and the predicted class says the same.

False Positive and False Negative: These values occur when their actual class contradicts the predicted class.

- **False Positive:** When the real class is no and the predicted class is yes. For example, if the actual class says this passenger did not survive, but the predicted class tells you that this passenger will survive.
- **False Negative:** When the real class is yes, but the predicted class is no. For example, if the real value of the class indicates that the passenger survived and the predicted class indicates that the passenger will die'.

Now we can calculate the accuracy, precision, recall, and F-score.

Accuracy: Accuracy is the most intuitive measure of performance and is simply a relationship between the correctly predicted observation and the total observations. One may think that if we have high accuracy, then our model is the best. Yes, accuracy is a great measure, but only when symmetric data sets are available in which the false

positive and false negative values are nearly the same. Therefore, you have to look at other parameters to evaluate the performance of your model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.1)$$

Precision: The precision is the relationship between the correctly predicted positive observations and the total predicted positive observations. The question that this metric response is from all the passengers who were labeled as survivors, how many actually survived? High accuracy is related to a low false positive rate.

$$Precision = \frac{TP}{TP+FP} \quad (5.2)$$

Recall: Recall is the relationship between correctly predicted positive observations and all observations in the real class, yes. The answer to the remembering question is: Of all the passengers who actually survived, how many did we tag?

$$Recall = \frac{TP}{TP+F} \quad (5.3)$$

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

$$F - Score = \frac{2TP}{2TP+FP+FN} \quad (5.4)$$

5.2 Simulation Results

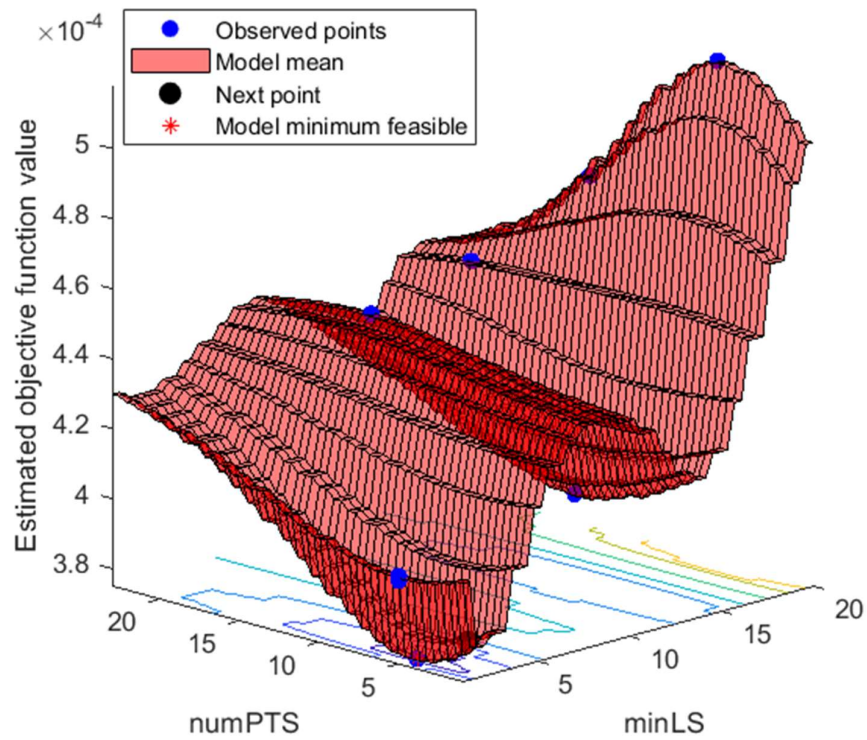


Figure 5.1: Objective function model

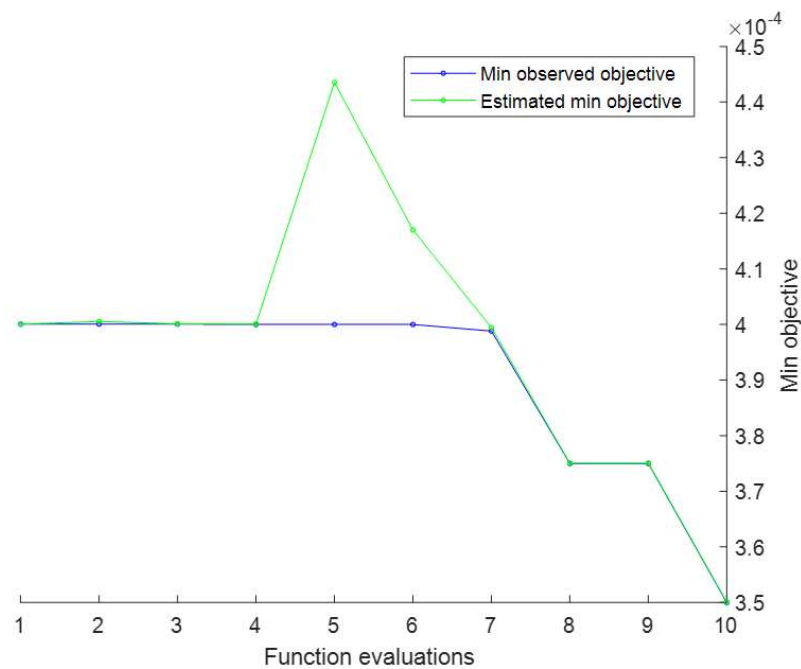


Figure 5.2: Minimum objective vs. number of function evaluations

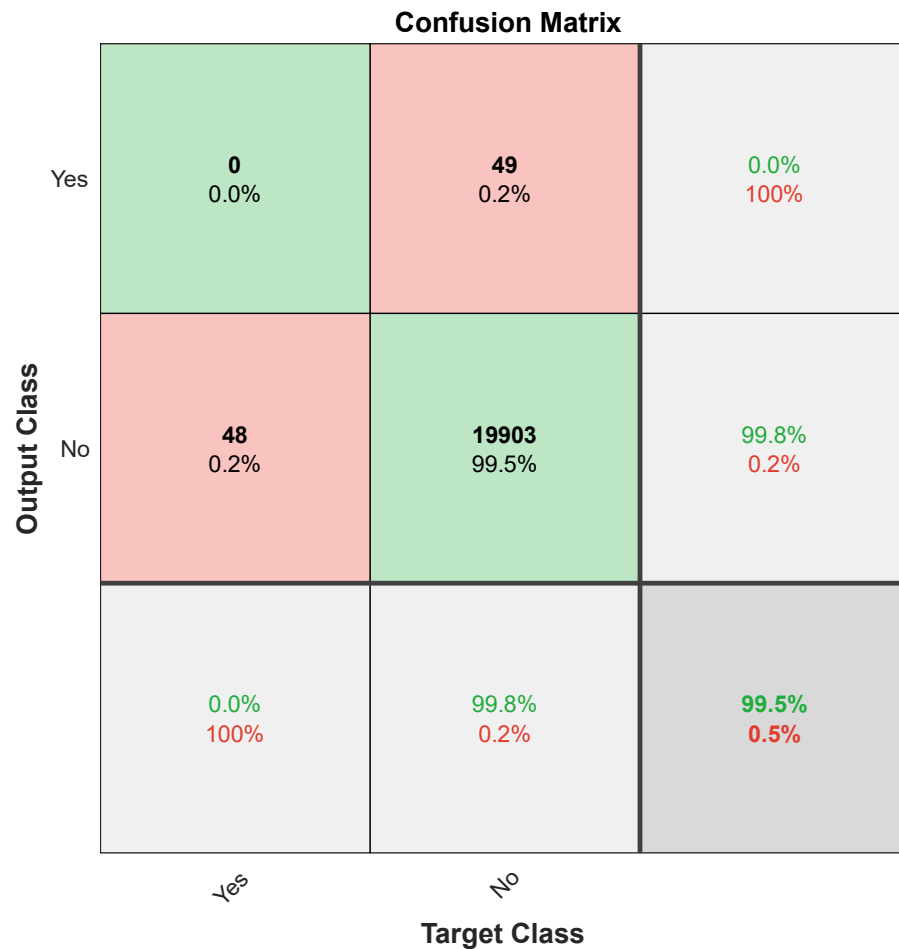


Figure 5.3: Confusion matrix plot for random forest classifier based credit card fraud detection

Here, TP=0, TN=19903, FP=49 and FN=48

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} = \frac{0+19903}{0+19903+49+48} = 99.5\%$$

$$Precision = \frac{TP}{TP+FP} = \frac{0}{0+49} = 0\%$$

$$Recall \text{ or } Sensitivity = \frac{TP}{TP+FN} = \frac{0}{0+48} = 0\%$$

$$F - Score = \frac{2TP}{2TP+FP+FN} = \frac{2 \times 0}{2 \times 0 + 49 + 48} = 0\%$$

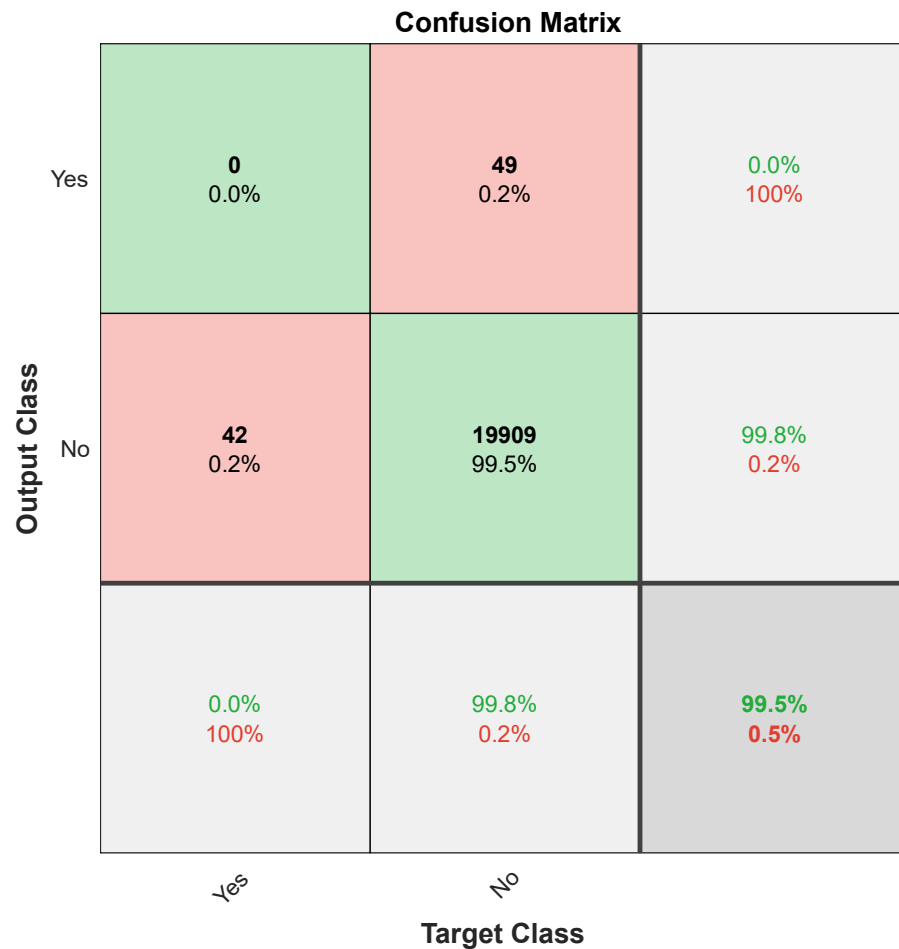


Figure 5.4: Confusion matrix plot for Bayesian optimized random forest classifier based credit card fraud detection

Here, TP=0, TN=19909, FP=49 and FN=42

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{0+199}{0+19909+49+42} = 99.5\%$$

$$Precision = \frac{TP}{TP+FP} = \frac{0}{0+49} = 0\%$$

$$Recall \text{ or } Sensitivity = \frac{TP}{TP+FN} = \frac{0}{0+42} = 0\%$$

$$F - Score = \frac{2TP}{2TP+FP+FN} = \frac{2 \times 0}{2 \times 0 + 49 + 42} = 0\%$$

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

The research for this dissertation is focused on the application of automatic learning techniques in the detection of credit card fraud: a difficult problem, motivated in the financial industry and highly applicable. There were abundant data available, but limited in terms of the number and diversity of variables.

The investigation began with a critical analysis of the state of the art in fraud detection and the previous modeling procedure, which concluded with the proposal of several improvements. Then, the possibility of using approaches related to the detection of using supervised learning was explored to construct a useful score for the detection. Following this path, an important contribution of this work was produced: the development of a Bayesian optimized random forest classifier based credit card fraud detection. The notion of Random forest was also introduced. With the accuracy of 99.5%, it was found that the proposed approach is useful for fraud detection and to achieve good discrimination between classes.

6.2 Future Scope

Throughout the work, it was possible to identify some work fronts that can be addressed in the future:

- Sampling: There are several alternatives to perform data sampling (stratified, not sensitive to cost, among others). Thus, it would be interesting to explore these different alternatives and compare the results obtained.
- Dynamic Modeling in Time: This work did not use the time variable in its context, that is, transactions were considered equally, regardless of the date of their occurrence. Experts know that the time factor is important in the domain. Therefore, this information could be introduced in the modeling and, for example, weigh the rules in some way that takes the time factor into account.

REFERENCES

- [1] Bhatla, Tej Paul, Vikram Prabhu, and Amit Dua. "Understanding credit card frauds." *Cards business review* 1, no. 6 (2003): 1-15.
- [2] Newman, Graeme R. *Check and card fraud*. US Department of Justice, Office of Community Oriented Policing Services, 2003.
- [3] Khan, M. A. A., A. A. S. Qureshi, and M. Farooqui. "Double Security of RFID Credit Cards." *International Journal of Computer Sciences and Engineering* 5, no. 5 (2017): 42-46.
- [4] Ogata, Hisao, Tomoyoshi Ishikawa, Norichika Miyamoto, and Tsutomu Matsumoto. "An ATM security measure for smart card transactions to prevent unauthorized cash withdrawal." *IEICE Transactions on Information and Systems* 102, no. 3 (2019): 559-567.
- [5] Knieff, Ben. *Global Consumer Card Fraud: Where Card Fraud Is Coming From*. Technical Report, Aite Group, Boston, USA. URL: <https://www.aitegroup.com/report/2016-global-consumer-card-fraud-where-card-fraud-coming>, 2016.
- [6] Demla, Nancy, and A. Aggarwal. "Credit card fraud detection using svm and reduction of false alarms." *International Journal of Innovations in Engineering and Technology (IJJET)* 7, no. 2 (2016): 176-182.
- [7] Mishra, Chandrahas, D. Lal Gupta, and Raghuraj Singh. "Credit Card Fraud Identification Using Artificial Neural Networks." *International Journal of Computer Systems* 4, no. 07 (2017).
- [8] Sohony, Ishan, Rameshwar Pratap, and Ullas Nambiar. "Ensemble learning for credit card fraud detection." In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pp. 289-294. 2018.
- [9] Manlangit, Sylvester, Sami Azam, Bharanidharan Shanmugam, Krishnan Kannoorpatti, Mirjam Jonkman, and Arasu Balasubramaniam. "An efficient method for detecting fraudulent transactions using classification algorithms on an anonymized credit card data set." In *International Conference on*

Intelligent Systems Design and Applications, pp. 418-429. Springer, Cham, 2017.

- [10] Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." *Procedia computer science* 48, no. 2015 (2015): 679-685.
- [11] Xuan, Shiyang, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang. "Random forest for credit card fraud detection." In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 1-6. IEEE, 2018.
- [12] Lei, John Zhong, and Ali A. Ghorbani. "Improved competitive learning neural networks for network intrusion and fraud detection." *Neurocomputing* 75, no. 1 (2012): 135-145.
- [13] Dal Pozzolo, Andrea, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. "Learned lessons in credit card fraud detection from a practitioner perspective." *Expert systems with applications* 41, no. 10 (2014): 4915-4928.
- [14] Bahnsen, Alejandro Correa, Aleksandar Stojanovic, Djamila Aouada, and Björn Ottersten. "Improving credit card fraud detection with calibrated probabilities." In *Proceedings of the 2014 SIAM international conference on data mining*, pp. 677-685. Society for Industrial and Applied Mathematics, 2014.
- [15] Jha, Sanjeev, Montserrat Guillen, and J. Christopher Westland. "Employing transaction aggregation strategy to detect credit card fraud." *Expert systems with applications* 39, no. 16 (2012): 12650-12657.
- [16] Mahmoudi, Nader, and Ekrem Duman. "Detecting credit card fraud by modified Fisher discriminant analysis." *Expert Systems with Applications* 42, no. 5 (2015): 2510-2516.
- [17] Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis." In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pp. 1-9. IEEE, 2017.

- [18] Yee, Ong Shu, Saravanan Sagadevan, and Nurul Hashimah Ahamed Hassain Malim. "Credit card fraud detection using machine learning as data mining technique." *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10, no. 1-4 (2018): 23-27.
- [19] Carneiro, Nuno, Goncalo Figueira, and Miguel Costa. "A data mining based system for credit-card fraud detection in e-tail." *Decision Support Systems* 95 (2017): 91-101.
- [20] Sánchez, Daniel, M. A. Vila, L. Cerda, and José-Maria Serrano. "Association rules applied to credit card fraud detection." *Expert systems with applications* 36, no. 2 (2009): 3630-3640.
- [21] Lopez-Rojas, Edgar Alonso, and Stefan Axelsson. "A review of computer simulation for fraud detection research in financial datasets." In *2016 Future Technologies Conference (FTC)*, pp. 932-935. IEEE, 2016.
- [22] Ahuja, Mini Singh, and Lovepreet Singh. "Online fraud detection-a review." *International Research Journal of Engineering and Technology* 4, no. 7 (2017): 2509-2515.
- [23] Taware, Rutuja. "Credit Card Fraud Detection Using Meta-classifiers Consisting of Semi-supervised and Supervised Algorithms." In *Advanced Computing Technologies and Applications*, pp. 503-511. Springer, Singapore, 2020.
- [24] Gulati, Aman, Prakash Dubey, C. MdFuzail, Jasmine Norman, and R. Mangayarkarasi. "Credit card fraud detection using neural network and geolocation." In *IOP Conference Series: Materials Science and Engineering*, vol. 263, no. 4, p. 042039. 2017.
- [25] Bahnsen, Alejandro Correa, Sergio Villegas, Djamila Aouada, Björn Ottersten, A. M. Correa, and S. Villegas. "Fraud detection by stacking cost-sensitive decision trees." *Data Science for Cyber-Security* (2017).
- [26] West, Jarrod, and Maumita Bhattacharya. "Intelligent Financial Fraud Detection: A Comprehensive." *Computers & Security*, vol. 57, pp. 47–66. 2016.

- [27] Whitrow, Christopher, David J. Hand, Piotr Juszczak, David Weston, and Niall M. Adams. "Transaction aggregation as a strategy for credit card fraud detection." *Data mining and knowledge discovery* 18, no. 1 (2009): 30-55.
- [28] Bhattacharyya, Siddhartha, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. "Data mining for credit card fraud: A comparative study." *Decision Support Systems* 50, no. 3 (2011): 602-613.
- [29] Van Vlasselaer, Véronique, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions." *Decision Support Systems* 75 (2015): 38-48.
- [30] Bahnsen, Alejandro Correa, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. "Detecting credit card fraud using periodic features." In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 208-213. IEEE, 2015.
- [31] Bahnsen, Alejandro Correa, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. "Feature engineering strategies for credit card fraud detection." *Expert Systems with Applications* 51 (2016): 134-142.
- [32] Halvaiee, Neda Soltani, and Mohammad Kazem Akbari. "A novel model for credit card fraud detection using Artificial Immune Systems." *Applied soft computing* 24 (2014): 40-49.
- [33] Daneshpazhouh, Armin, and Ashkan Sami. "Entropy-based outlier detection using semi-supervised approach with few positive examples." *Pattern Recognition Letters* 49 (2014): 77-84.
- [34] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41, no. 3 (2009): 1-58.
- [35] Duman, Ekrem, and M. Hamdi Ozelik. "Detecting credit card fraud by genetic algorithm and scatter search." *Expert Systems with Applications* 38, no. 10 (2011): 13057-13063.
- [36] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

- [37] Sahin, Yusuf, Serol Bulkan, and Ekrem Duman. "A cost-sensitive decision tree approach for fraud detection." *Expert Systems with Applications* 40, no. 15 (2013): 5916-5923.
- [38] Fu, K., Cheng, D., Tu, Y. and Zhang, L., 2016, October. "Credit card fraud detection using convolutional neural networks." In *International Conference on Neural Information Processing* (pp. 483-490). Springer, Cham.
- [39] Dighe, D., Patil, S. and Kokate, S., 2018, August. "Detection of credit card fraud transactions using machine learning algorithms and neural networks: A comparative study." In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.
- [40] Dheepa, V. and Dhanapal, R., 2012. "Behavior based credit card fraud detection using support vector machines." *ICTACT Journal on Soft computing*, 2(07), p.2012.
- [41] Chen, R.C., Chiu, M.L., Huang, Y.L. and Chen, L.T., 2004, August. "Detecting credit card fraud by using questionnaire-responded transaction model based on support vector machines." In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 800-806). Springer, Berlin, Heidelberg.
- [42] Gaikwad, J.R., Deshmane, A.B., Somavanshi, H.V., Patil, S.V. and Badgujar, R.A., 2014. "Credit Card Fraud Detection using Decision Tree Induction Algorithm." *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 4(6).
- [43] Razooqi, T., Khurana, P., Raahemifar, K. and Abhari, A., 2016, April. "Credit card fraud detection using fuzzy logic and neural network." In *Proceedings of the 19th Communications & Networking Symposium* (pp. 1-5).
- [44] Phua, C., Alahakoon, D. and Lee, V., 2004. "Minority report in fraud detection: classification of skewed data." *Acm sigkdd explorations newsletter*, 6(1), pp.50-59.
- [45] Bentley, P.J., Kim, J., Jung, G.H. and Choi, J.U., 2000, October. "Fuzzy darwinian detection of credit card fraud." In *the 14th Annual Fall Symposium of the Korean Information Processing Society* (Vol. 14).

- [46] Chiu, C.C. and Tsai, C.Y., 2004, March. "A web services-based collaborative scheme for credit card fraud detection." In *IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004. EEE'04. 2004* (pp. 177-181). IEEE.
- [47] Fan, W., 2004, August. "Systematic data selection to mine concept-drifting data streams." In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 128-137).
- [48] Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B., 2002, January. "Credit card fraud detection using Bayesian and neural networks." In *Proceedings of the 1st international naiso congress on neuro fuzzy technologies* (pp. 261-270).
- [49] Wheeler, R. and Aitken, S., 2000. "Multiple algorithms for fraud detection." In *Applications and Innovations in Intelligent Systems VII* (pp. 219-231). Springer, London.
- [50] Rochet, Jean-Charles, and Jean Tirole. "Two-sided markets: an overview." *Institut d'Economie Industrielle working paper* (2004).
- [51] Shen, Aihua, Rencheng Tong, and Yaochen Deng. "Application of classification models on credit card fraud detection." In *2007 International conference on service systems and service management*, pp. 1-4. IEEE, 2007.
- [52] Hand, David J. "Pattern detection and discovery." In *Pattern detection and discovery*, pp. 1-12. Springer, Berlin, Heidelberg, 2002.
- [53] Doody, Helenne. *Fraud risk management: A guide to good practice*. CIMA, 2009.
- [54] Dal Pozzolo, Andrea, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. "Credit card fraud detection: a realistic modeling and a novel learning strategy." *IEEE transactions on neural networks and learning systems* 29, no. 8 (2017): 3784-3797.
- [55] Al-Khatib, Adnan. "Electronic payment fraud detection techniques." *World of Computer Science and Information Technology Journal (WCSIT)* 2, no. 4 (2012): 137-141.
- [56] Bolton, Richard J., and David J. Hand. "Statistical fraud detection: A review." *Statistical science* (2002): 235-249.

- [57] Leonard, Kevin J. "Detecting credit card fraud using expert systems." *Computers & industrial engineering* 25, no. 1-4 (1993): 103-106.
- [58] Chan, P. K., and S. J. Stolfo. "Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection [Internet]. Palo Alto (CA): Association for the Advancement of Artificial Intelligence; 1998 [cited 2017 Dec 20]."
- [59] Fawcett, Tom, and Foster Provost. "Adaptive fraud detection." *Data mining and knowledge discovery* 1, no. 3 (1997): 291-316.
- [60] Stolfo, Salvatore, David W. Fan, Wenke Lee, Andreas Prodromidis, and P. Chan. "Credit card fraud detection using meta-learning: Issues and initial results." In *AAAI-97 Workshop on Fraud Detection and Risk Management*, pp. 83-90. 1997.
- [61] Hill, Theodore P. "A statistical derivation of the significant-digit law." *Statistical science* 10, no. 4 (1995): 354-363.
- [62] Hand, David J. "Discrimination and classification." *diel* (1981).
- [63] McLachlan, Geoffrey J. *Discriminant analysis and statistical pattern recognition*. Vol. 544. John Wiley & Sons, 2004.
- [64] Ripley, Brian D. *Pattern recognition and neural networks*. Cambridge university press, 2007.
- [65] Fukunaga, Keinosuke. *Introduction to statistical pattern recognition*. Elsevier, 2013.
- [66] Credit Card Fraud Detection Dataset. Online available at: <https://www.kaggle.com/mlg-ulb/creditcardfraud>
- [67] Kenedy, J., and R. C. Eberhart. "A discrete binary version of the particle swarm optimization." *Computational cybernatics and simulation* 5 (1997): 4104-4108.
- [68] Khanesar, Mojtaba Ahmadi, Mohammad Teshnehlab, and Mahdi Aliyari Shoorehdeli. "A novel binary particle swarm optimization." In *2007 Mediterranean Conference on Control & Automation*, pp. 1-6. IEEE, 2007.
- [69] Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5-32.

- [70] Pelikan, M., Goldberg, D.E. and Cantú-Paz, E., 1999, July. BOA: The Bayesian optimization algorithm. In *Proceedings of the genetic and evolutionary computation conference GECCO-99* (Vol. 1, pp. 525-532).
- [71] Snoek, J., Larochelle, H. and Adams, R.P., 2012. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951-2959).

APPENDICES

List of Tables

Table No.	Table Name	Page No.
3.1	Industry identification	26
3.2	Brand identification	27
3.3	Comparison of statistical techniques for detecting credit card fraud	43


List of Figures

Figure No.	Figure Name	Page No.
1.1	Current total card fraud rate by country	6
3.1	Credit card structure	26
3.2	Transaction flow; Source- MasterCard	28
4.1	Flow diagram of proposed research work	48
5.1	Objective function model	56
5.2	Minimum objective vs. number of function evaluations	56
5.3	Confusion matrix plot for random forest classifier based credit card fraud detection	57
5.4	Confusion matrix plot for Bayesian optimized random forest classifier based credit card fraud detection	58

CHECK LIST

- | | | |
|------|--|------------------|
| a) | Is the Cover page in proper format? | Y / N |
| b) | Is the Title page in proper format? | Y / N |
| c) | Is the Certificate from the Supervisor in proper format? Has it been signed? | Y / N |
| d) | Is Abstract included in the Report? Is it properly written? | Y / N |
| e) | Does the Table of Contents page include chapter page numbers? | Y / N |
| f) | Does the Report contain a summary of the literature survey? | Y / N |
| i. | Are the Pages numbered properly? | Y / N |
| ii. | Are the Figures numbered properly? | Y / N |
| iii. | Are the Tables numbered properly? | Y / N |
| iv. | Are the Captions for the Figures and Tables proper? | Y / N |
| v. | Are the Appendices numbered? | Y / N |
| g) | Does the Report have Conclusion / Recommendations of the work? | Y / N |
| h) | Are References/Bibliography given in the Report? | Y / N |
| i) | Have the References been cited in the Report? | Y / N |
| j) | Is the citation of References / Bibliography in proper format? | Y / N |

 RAJESH.P.K.

Verified

7/1/2021
JYJESH.P.S