Credit Card Fraud Detection using Machine Learning Techniques

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY
IN

SOFTWARE ENGINEERING

Submitted by:

Anumeha Garg

2K20/SWE/05

Under the supervision of

Prof. Ruchika Malhotra

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY 2022

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

CANDIDATE'S DECLARATION

I, Anumeha Garg, Roll No. 2K20/SWE/05 student of M. Tech (Software Engineering), hereby declare that the project Dissertation titled "Credit Card Fraud Detection using Machine Learning Techniques" which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Anumeha

Place: Delhi

Date:

Anumeha Garg

2K20/SWE/05

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Credit Card Fraud Detection using Machine Learning Techniques"** which is submitted by Anumeha Garg, 2K20/SWE/05 Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

Prof. Ruchika Malhotra

SUPERVISOR

ABSTRACT

Over the previous few decades, the increasing credit card fraud cases has always been a major source of concern. This situation is because of the widespread of new technologies, particularly the growing popularity of online banking transactions. However, to recognize scam tendencies it takes computational strength and complexity in designing and creating the pattern matching rule basis. The major purpose is to identify methods and strategies that have significant influence on fraud detection, with a focus on existing research work. Support Vector Machines (SVMs), naive Bayesian, Artificial Neural Networks (ANNs), Decision Tree, K-Nearest Neighbor (k-NN) and Frequent Pattern Mining algorithms are all studied and compared for detecting suspicious transactions.

Ensemble models like bagging and clustering have been utilized in conjunction with an algorithmic technique. Boosting has been performed to a dataset of 284807 transactions that is significantly skewed. To name a few only 492 of the total transactions have been flagged as suspicious. Models of prediction, such as the logistic as well as XGBoost when used with various resampling approaches have yielded. It has been used to determine a transaction is genuine or fraudulent. The model's performance is assessed using the following criteria: recall, precision, f1-score, precision-recall (PR) curve, and receiver operating characteristics (ROC) curves.

ACKNOWLEDGMENT

The success of this project requires the assistance and input of numerous people and the organization. I am grateful to everyone who helped in shaping the result of the project.

I express my sincere thanks to **Prof. Ruchika Malhotra**, my project guide, for providing me with the opportunity to undertake this project under her guidance. Her constant support and encouragement have made me realize that it is the process of learning which weighs more than the result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work helped me throughout with newideas, provided information necessary and pushed me to complete the work.

I also thank all my fellow students and my family for their continued support.

ANUMEHA GARG

2K20/SWE/05

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The scammers are more interested in exploiting and manipulating transactions between credit cards and electronic companies. This is due to several flaws in the current detection systems, as well as a significant element corporate carelessness. False statements about income, overstatements or claims for inflated deductions, money laundering, and swindle transactions are just a few examples of the numerous sorts of fraud that may be committed. Using rule-based systems, fraud knowledge is stored and manipulated such that it can be interpreted in a meaningful way.

An AI methodology known as rule-based reasoning is one of the prevalent methods for storing and retrieving knowledge from a computer system, and it employs rules as a representation of that knowledge. A semantic reasoner infers information based on the interaction between input and the rule base and user interface that handle the incoming data flow and the outgoing flow of prediction results are standard components of a rule-based system. Rule-based systems infer information contained in its rule base and emulate human experts' thinking while addressing a knowledge-intensive challenge, such as a coding difficulty. The use of rule-based systems in fraud detection and prediction has become increasingly common in recent years.

Rule-based systems are accurate in identifying and forecasting fraud, but they demand enormous computational capacity for specific domain pattern matching. The rule base is extremely difficult to update and modify. Contradictions may arise

when a company introduces new knowledge to uncover new fraud trends. Furthermore, fraudsters are extremely adaptable and will always find a way around preventive measures if given enough time. Simply by bypassing simple pattern matching or rule-based detection methods they can trick systems to believe they are dealing with a legitimate buyer and seller. Fraud detection will be inaccurate because of this problem.

Rule-based systems, in addition to their flaws, lack analytical and predictive skills. Having the ability to do these operations on received data will help to better identify future fraud cases. Thus, recent breakthroughs in fraud detection using data mining and machine learning approaches have been extensively discovered. To better detect and anticipate frauds, new tools enable correlation analysis of fraud data as they fall under the category of artificial intelligence, data mining and machine learning. The tools can shift through massive data to find patterns and, in turn, uncover previously undiscovered information. It is possible to utilize a variety of machine learning and data mining approaches to not only detect suspicious transactions, but to also anticipate the suspicious rate at which transaction data will transact over time. In this research, ANNs, SVMs, k-Nearest Neighbor (k-NN), Bayesian Classification, Frequent Pattern (FP), Decision Tree (DT) Mining algorithms are examined for their capabilities and performance in detecting fraud.

Distinguishing between phoney and real financial data is an important part of these systems since it allows fraudsters to be exposed and the effect of their operations is reduced [15]. Data mining and machine learning techniques are examined and analyzed as a means of overcoming the difficulties in detecting fraud activities.

To verify a transaction's validity, these strategies examine the transaction history and analyze data thoroughly. We hope to provide an analysis of machine learning and data mining approaches for evaluating suspicious behavior, identify the data sources and features based on the fraud detection and prediction studies that have been conducted.

1.1 CREDIT CARD FRAUD DETECTION PROCESS

The transactions validity is initially confirmed at the terminal point, as shown in Figure 1. Certain conditions are confirmed at the terminal point like sufficient balance, valid PIN (Personal Identification Number), and the transactions are filtered. The prediction model categorizes all the transactions as real or fraudulent. Each suspicious warning is investigated by the investigators, who in return provide feedback to prediction model which enhances the model performance. This is only about the prediction model.

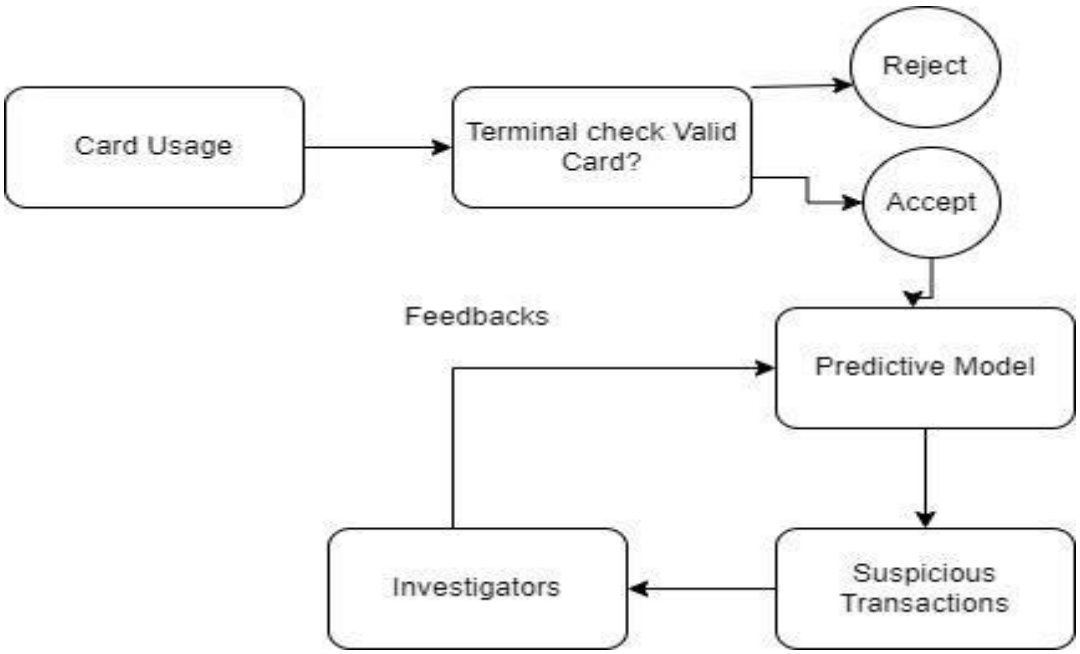

Figure 1.1 Credit Card Fraud Detection Process

Fraud detection systems are more complicated than they appear to be. In practice, the practitioner must determine which classification technique to apply (decision trees or logistic regression) as well as how one should cope with the problem of class imbalance (Suspicious cases are exceedingly less in contrast to valid ones). Detection of fraud is not simply a problem because of the disparity between the rich

and poor. Due to a lack of transaction data, many machine learning algorithms fail in the classification job because of the overlap between the real and fraudulent classes.

An actual fraud detection scenario involves a model that uses artificial intelligence to identify suspicious transactions and send an alert to the appropriate authorities when one of those transactions is determined to be either authentic or fraudulent. The fraud detection system is improved by investigators who investigates and give their findings back to the system. As a result, only few transactions get certified timely by investigators through this approach. When few feedbacks are supplied to the predictive model, it is often less accurate.

Because financial institutions seldom release consumer data owing to privacy concerns, it is extremely difficult to find the true financial datasets. An important problem in fraud detection system is overcoming this obstacle.

1.2 OBJECTIVE

The main purpose of this thesis is to do predictive analysis on a credit card transaction dataset using machine learning techniques to identify suspicious transactions.

The idea is to employ prediction algorithms to figure out whether a transaction is legitimate or not. Numerous sampling procedures will be employed to solve the class imbalance problem, and several machine learning algorithms such as random forest, logistic regression and XGBoost will be applied to the dataset and the results will be reported.

1.3 CHALLENGES IN FRAUD DETECTION

In real-world, a credit card fraud detection model predicts the nature (genuine or suspicious) and sends an alert to the investigators for the suspect transaction. Then investigators conduct a second investigation and provide input to the fraud detection system to improve its efficiency. However, this process can be time consuming for investigators, resulting in few transactions being validated on time. In

this situation, the predictive model receives only a few feedbacks which results in less accurate model.

Another problem in classification process is the overlap of authentic and fraudulent classes due to limited transaction records. Under these conditions, machine learning algorithms perform poorly.

As financial institutions rarely release client data to the public because of confidentiality concerns, true financial datasets are difficult to get. This is the most difficult problem in fraud detection research.

CHAPTER 2

PRIOR WORK AND PRELIMINARIES

2.1 RELATED WORK

The study helps the scientific community better transfer their findings into the real world. This will show that only minor real-world advances have been achieved, resulting in limited benefits to the general population. It is vital to shed light on these earlier efforts to help us understand how science progresses throughout time. This examination is not only devoted to discussing the historical contexts but also to look at the most popular (or best) methods.

The published research in the field of credit card fraud detection has been undertaken. Rather of relying on secondary citations, this study uses real-world measures and expands and consolidates data from prior studies into a single evaluation that serves as a benchmark for the industry. IEEE Xplore Digital Library and scholar were used most frequently to search for a broad collection of papers using phrases like "fraud detection," "credit card fraud," as well as more specific search terms. Additional references might be found in literature surveys, broad subject descriptions, and novels, however these sources were omitted from the study itself. There are several studies in this survey that deal with the issue of payment fraud detection using artificial intelligence and machine learning techniques. Their innovation, publication date, methodologies and algorithmic findings and implementations are assessed.

There has been an extensive literature review of fraud detection strategies from 1991 to 2002 by Yufeng et al. (2004), which includes approaches for detecting credit card fraud. In Phua et al. (2010), the years 1994–2004 are covered in a study of

data mining algorithms for fraud detection in general. General approaches for detecting credit card fraud are discussed by Sethi and Gera (2014), including an overview of the most prevalent fraud vectors. The prior techniques are discursively summarized in a brief literature review in Ryman-Tubb (2011). Financial fraud detection patents from 1998 to 2013 may be found in Danenas (2015). Ahmed et al. provides a study of anomaly detection approaches that includes fraud detection (2016). In Adewumi and Akinyelu, a brief assessment of fraud detection strategies from 2005 to 2015, with machine learning emphasis is provided (2016). For example, payment card fraud, telephony scams, insurance claims, and online auction fraud are all discussed by Abdallah et al. (2016) in their review. There are other surveys out there, but this one stands out because it is thorough and uses a consistent assessment that are based on the actual demands of the payment card business.

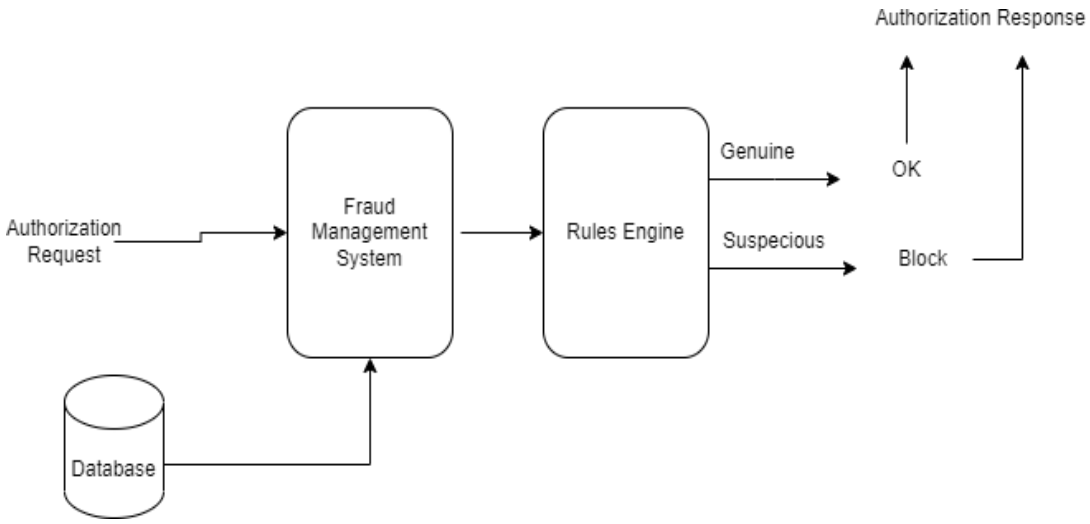

Figure 2.1 Real Time FPS

For this study, the term "expert system" is used to refer to artificial intelligence that is based on any symbolic human representation of knowledge. The expert system is one of the oldest forms of contemporary artificial intelligence, with early implementations beginning in the 1970s (Feigenbaum, 1977). To tackle real-world issues that standard software could not handle, expert systems were created. Symbolic rules are generally employed to encode the knowledge of human specialists.

The information in this database may then be used to derive a conclusion using logical reasoning. Recently, newer systems have been able to derive rules directly from datasets using ML approaches like inductive learning. An essential industrial demand is the creation of a reliable and transparent solution that leads to enhanced human comprehension through expert systems.

As noted by Shao et al. (1995), expert systems are widely used in the payments sector, and this is still the case today. Fraud specialists write rules that encapsulates their expertise in fraud vectors. According to Dazeley (2006), this is a time-consuming and expensive process that necessitates the involvement of subject matter specialists. Small, well-written regulations are easy to follow and comprehend since they are clearly laid out. The fixed rules must be modified in response to changes in the external environment (1.4.8). Many issues, such as detecting fraud, may necessitate adjustments to account for new forms of payment and criminal activity. It's tough to keep track of and understand a big set of rules. When it expands, the impact of fresh regulations becomes more difficult to assess since it requires more computational power.

An early expert system for detecting credit card fraud was proposed by Leonard (1993), using data from a Canadian bank totaling 12,709 transactions. Even for that decade, the RGF of 21 in the Canadian Bank sample is exceptionally minimal. The human team would have had to analyze 611 alerts (Af) to discover a single fraud in an implausible 496 k AlertD. Vatta et al. (2009) suggest an even more advanced expert system that incorporates game theory and an expert system depicted. The fraudster and the FMS are the two participants in game theory. Both parties are vying for the same thing: to maximize their own gains to win. Because of this, fraudsters employ a variety of methods, and the FMS must catch them all in the earliest stages of execution to minimize losses. Criminals have been shown to continue using a stolen CHD until it is banned. Low-value transactions at low-risk locations are typically used as part of a standard fraud procedure to avoid being flagged by the Fraud Management System (FMS). They believe that the FMS works as an "opponent" in the eyes of the public. According to a Nash Equilibrium, the criminal.v. FMS is playing (Rosenthal, 1973). As a result, the FMS should adjust by anticipating the next "move" made by the offender. Only if real-world input is incorporated into the system, it can be used to

make many "moves" where the criminal alters their behavior to avoid being stopped. To increase the chance of "winning," the FMS uses machine learning to alter its belief based on the information it learns with each transaction. Synthetic dataset findings were presented as the number of fraudulent transactions accurate improved across nine rounds, from 45 percent to 70 percent. According to the results shown in Fig. 6, the FMS progresses by adjusting its approach every step-in order to properly detect more fraudulent transactions.

Using fuzzy criteria, HaratiNik et al. (2012) ranks the worst in this benchmark. Terms, such as "high," "average," and "low," are defined as fuzzy rules. A membership function, often a Gaussian function, is used to assign these terms to numerical values. These terms are combined to produce the rule's output, which is normally in the range [0,1]. There is a TPR of 91.6% but a weak FPR of 77.5%, and the AlertD is 4.4 m when employing Tier-1 volumes. This is far worse than a coin toss. In this benchmark, Correia et al. (2015) describe how a manual trial-and-error process was used to develop a set of 14 rules based on known fraud vectors.

The proposed approach is put into practice using the open-source software package PROTON (IBM, 2015). For each written rule, Event Processing Agents (EPAs) is used to set up Event Processing Network (EPN). Each field was accompanied with a derived PDF, allowing the EPN to provide a certainty output for each new transaction. Since the value of the fields may be inaccurate, an uncertainty measure was utilized. For each transaction, the total of PDFs in each field will be used to calculate a certainty score.

Morgan and Sonquist go into great depth on DTs, which are graphical representations of decision trees (1963). Inductive learning is used to develop a DT, which is a classifier that divides classes into mutually exclusive subgroups at each node along the tree's branching path (Quinlan, 1986). When using this strategy, it is possible to look at DTs similarly as expert systems. From the base of the tree to the leaf of categorization, these are like English and easy to understand. A DT can be induced from training vectors using well-known algorithms, such as Quinlan (2007) and Cohen (2007). A neural network beats DT learning when performance and generalization are significant, according to previous research (Fisher and McKusick, 1989). Decision tree is not likely a part of real-world payment card. However, the study shows that this

is not always the case. Overfitting concerns in prior decision tree algorithms have been addressed by merging several subsets of the training.

It was employed in the early work (Breiman, 1996) where a random pick from the training dataset was used to produce each DT. Characteristics are taken from an unknown subset of all available features and a threshold is chosen based on an information gain requirement in newer approaches (Geurts et al., 2006).

2.2 PRELIMINARIES

2.2.1 MACHINE LEARNING

Machine learning, in the broadest sense, is a discipline of artificial intelligence that involves automatic learning using algorithmic models. Machine learning is distinct from old computing methods, in which a system must be explicitly coded to answer a problem. Datasets are given to artificial intelligence (AI) so that based on its training data, it can learn new patterns and predict unknown consequences. ML has huge number of applications. It is used in weather forecasting, spam filtering, fake news detection and similar functions.

2.2.2 CLASSIFICATION

In machine learning, classification task is to determine the class label for a given data item. Credit card fraud detection could be used as an example of a categorization issue. The major goal here is to evaluate whether a transaction is authentic or a scam. Multi-label classification has data samples (not mutually exclusive), and an individual label for each data sample. Binary classification has two output labels (e.g., categorizing a transaction as legitimate or fake), multi-class classification has more than two output labels (e.g., classifying flower images as Jasmine, Olivia, or Sunflower). This study investigates a binary classification problem in which the output is either genuine or fake.

2.2.3 BOOTSTRAPPING

Bootstrapping is critical concept in bagging algorithms. Bootstrapping is the process of randomly picking the training data and replacing it. Each bootstrap sample is chosen in a way that each sample has unique attributes, as shown in Figure 2.2. By training on these instances, models can get a better grasp of the data and then use that information to create more accurate predictions.

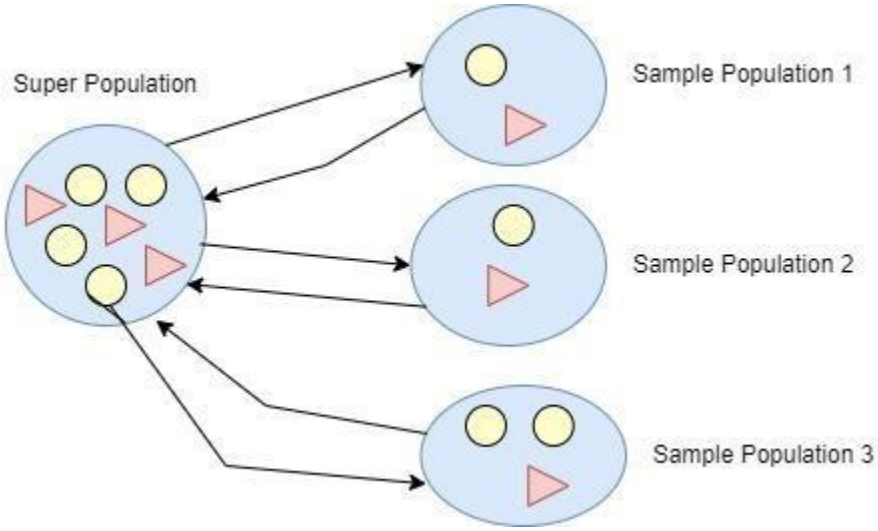

Figure 2.2 Bootstrapping

2.2.4 LOGISTIC REGRESSION

Logistic regression is widely used machine learning algorithm for classification. Although as the name implies logistic regression is not a regression algorithm.

Logistic regression was named after the origin in linear regression, a ML technique widely used in regression analysis. The possibility of a result for each class is stated as the probability with that class's outcome. A linear regression model predicts real-valued outputs using an input variable and weights. In this example, consider x to be the single independent variable and y to be the dependent variable. As a result, the linear regression hypothesis is $y = a0 + a1\ x$.

2.2.5 RANDOM FOREST

Random forest can be used for regression and classification problems. To put it another way, it is a bagging extension. The bagging strategy is combination of multiple weak pupils. Before we go into the inner workings of the random forest, let us study up on decision tree.

Decision tree for regression is used to solve classification problems. A node that represents an attribute (e.g., the weather be sunny, runny, or overcast tomorrow) is used to express if weather will be sunny, overcast, or rainy tomorrow. Each branch in tree reflects a distinct test result, with the leaf nodes representing the results. This approach works by subdividing training set into several subsamples.

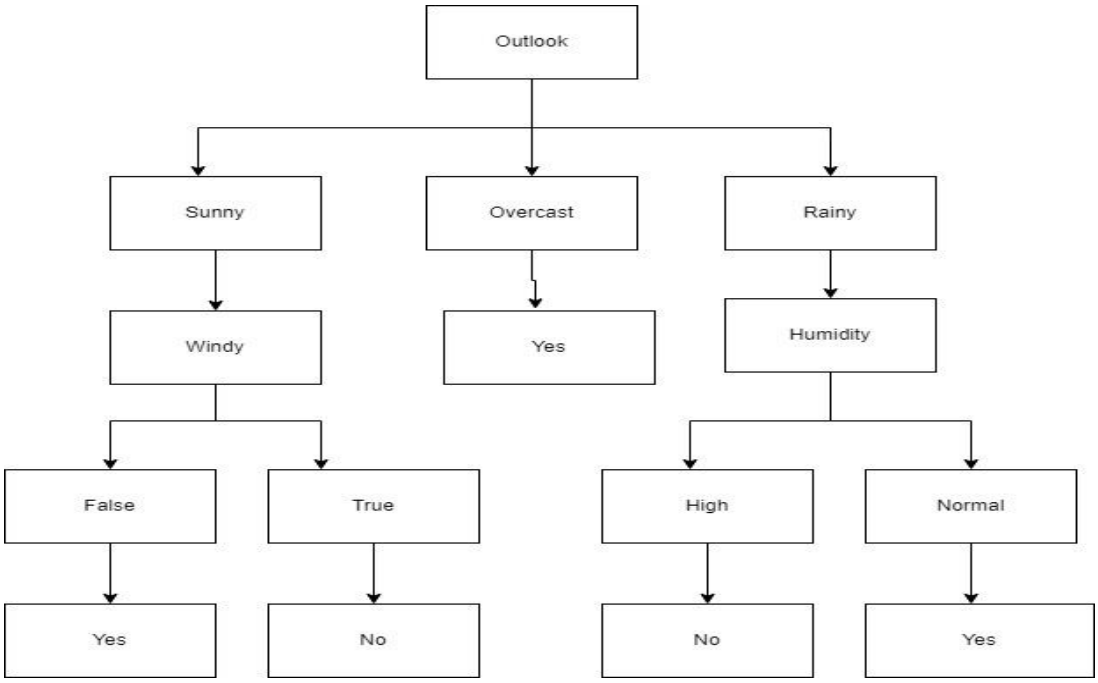

Figure 2.3 Decision Tree

2.2.6 XGBOOST

eXtreme Gradient Boosting (XGBoost) is improved version of gradient boosting. Gradient boosting refers to boosting techniques that combines weak learners into powerful ones. It produces poor students during the learning process. This method starts with weak learner predicting the class label, followed by the loss calculation. Based on the loss, it creates a new weak learner who is then trained on the remaining

faults. This cycle can go on indefinitely. Gradient boosting is the name given to this technique for solving the optimization problem known as gradient descent. While Ada boost uses weights to assign more importance to incorrectly classified data, this is another way to ensure that the mistakenly classified data is assigned to the next weak learner.

As previously stated, Xgboost is more advanced variant of the gradient boosting approach. This approach uses decision trees as weak learners, overcomes many of the limitations of the standard gradient boosting strategy. The regularization of gradient boosting is lower than that of XGBoost. Overfitting is reduced as a result. XGboost is substantially more efficient than standard gradient boosting thanks to its parallel processing. As it is built in, XGBoost has no problem with missing data. Gradient Boosting stops splitting when it finds a negative loss in splitting, but xgboost keeps splitting till the maximum depth. The cross-validation mechanism of XGBoost makes the number of boosting rounds easy. The hyperparameters should be modified to get best possible results from xgboost algorithm. It is the most used assessment measure in predictive analysis because of simplicity and capacity to compute metrics like accuracy and recall. A NxN matrix represents the total performance of a model where N is number of class labels in classification task.

CHAPTER 3

METHODOLOGY

When compared to the customer's prior transactions, card transactions are always new. Concept drift difficulties are a particularly tough challenge to solve in the actual world because of this unfamiliarity. It is possible to describe concept drift as a variable that varies over time and in unexpected ways. A lot of data is out of whack because of these factors. One of the most important goals of our work is to find a solution to the problem of Concept Drift in a real-world context. Any time a transaction is made, its essential characteristics are listed in Table 1.

Table 3.1: Raw features of credit card transactions

Attribute name	Description
Transaction id	Identification number of a transaction
Cardholder id	Unique Identification number given to the cardholder
Amount	Amount transferred or credited in a particular transaction by the customer
Time	Details like time and date, to identify when the transaction was made
Label	To specify whether the transaction is genuine or fraudulent

During the month of September 2013, a cardholder completed two transactions, which are included in this dataset. In a total of 284,807 transactions, there are 492 fraudulent transactions, or 0.172 percent. This data set is skewed in favor of one side. Since a transaction's specifics are made public, to protect client anonymity, most of the dataset's characteristics have been reworked. Using Principal Component Analysis

(PCA). There are PCA-applied characteristics (V1, V2...V28), as well as the rest (i.e., "time"). Table 2 shows that 'amount' and 'class' are non-PCA applied features.

Table 3.2 Attributes of dataset

S. No.	Feature	Description
1.	Time	Time in seconds to specify the elapses between the current transaction and first transaction.
2.	Amount	Transaction amount
3.	Class	0 - not fraud 1 – fraud

3.1 PROPOSED SOLUTION

- As a first step, we apply a clustering algorithm to separate the cardholders into three distinct groups depending on transaction amounts (high/medium vs low/medium).
- A sliding-window approach is used to organize transactions into distinct groups, i.e., extract certain elements from the window to identify cardholder behavior patterns. There are features such as a maximum and minimum transaction amount, as well as an average transaction amount and even the time elapsed.
- All transactions are executed in the same order in which they were received, except for those that have already been withdrawn from the window. There are algorithms in [1] that allude to the Sliding-Window-based approach of aggregate calculation.
- After pre-processing, we use the cardholders' behavior patterns to train classifiers and extract fraud characteristics from each group. Even if we apply classifiers to the dataset, the classifiers do not perform well because of the dataset's imbalance.

3.2 BLOCK DIAGRAM OF PROPOSED MODEL

Data-driven model and learning to rank technique are the major emphasis of our Fraud-Detection System (FDS). It also examines how recent supervised samples are presented in the form of alert feedback interactions. Figure 5 depicts the proposed system's block diagram.

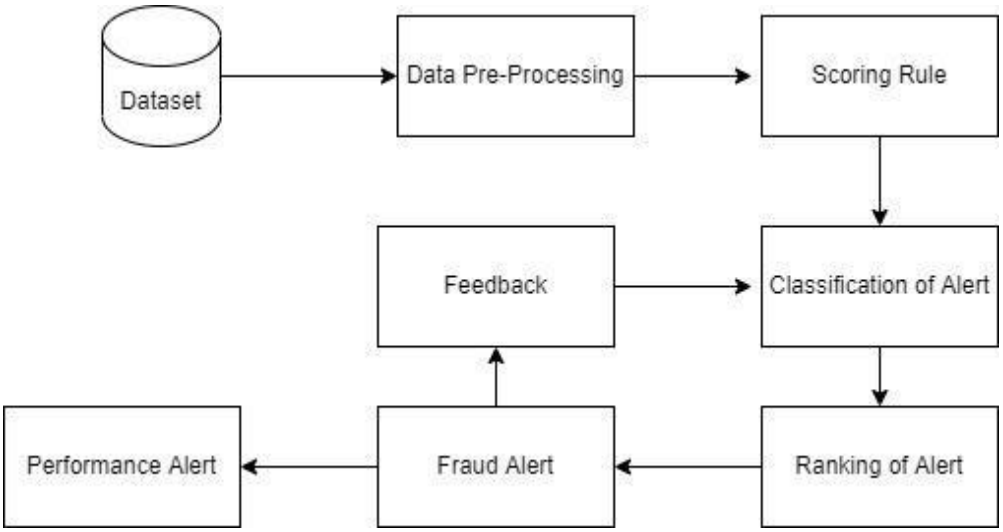

Figure 3.1 Block Diagram

CHAPTER 4

RESULT AND ANALYSIS

4.1 DATASET PROCESSING

Due to the large number of negative (majority) class instances outnumbering the number of positive (minority) class instances, we can conclude that our two datasets are severely unbalanced. Dataset-1 shows that in Dataset-1, frauds account for less than 0.171 percent of all transactions. By creating synthetic cases of minority fraud, we can improve the classification performance of the most interesting class by using an advanced oversampling technique called Synthetic Minority Oversampling Technique (SMOTE).

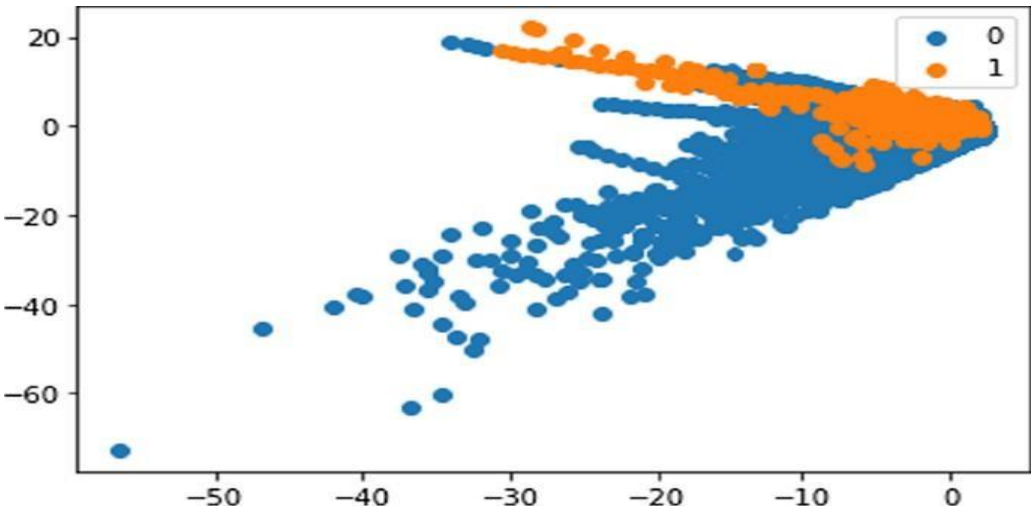

Figure 4.1 Dataset before SMOTE transformation

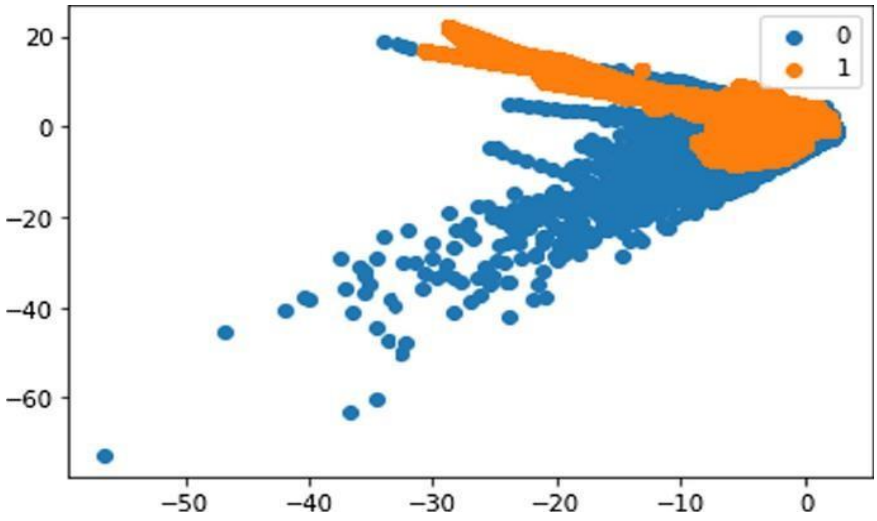

Figure 4.2 Dataset after SMOTE transformation

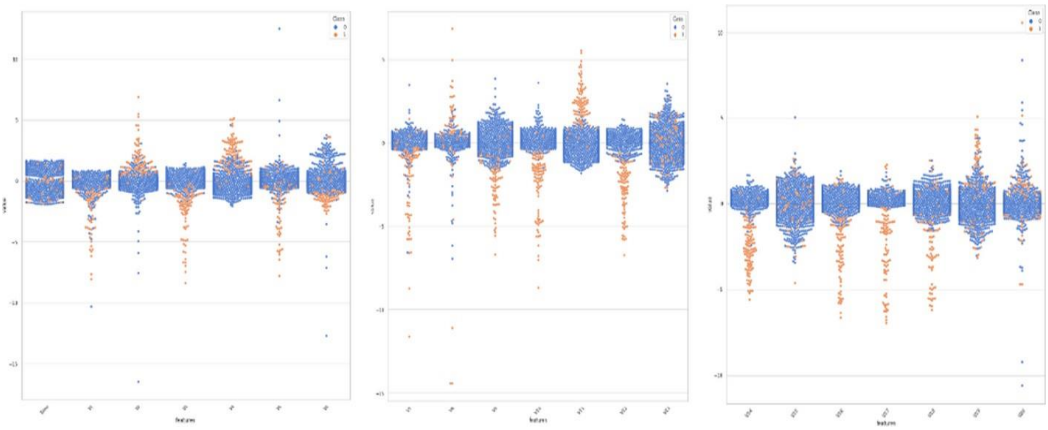

Figure 4.3 Swarm intelligent plot

4.2 IMPLEMENTATION AND EVALUATION METRICS

To replicate the cyclical nature of credit card transactions, we use Long Short Term Memory networks. There are many advantages to the LSTM's hidden state architecture, including the ability to connect neural network nodes across time.

When inputs are spread in time, the model can discern temporal relationships between events by retaining information from previous inputs. LSTM is a suitable model for sequential data points where the occurrence of one event may depend on the presence of numerous other events that occurred earlier in the time.

4.3 ANALYSIS

Table 3 compares the performance of several learning algorithms for detecting credit card fraud.

Comparing their precision, accuracy, and specificity is the basis for this discussion.

Table: 4.1 Machine learning approaches

Classifiers	Accuracy	Precision	Specificity
Random Forest	0.961	0.996	0.987
Logistics Regression	0.947	0.996	0.979
KNN	0.942	0.41	0.971
SVM	0.938	0.782	0.984
Decision Tree	0.908	0.911	0.912
Naïve Bayes	0.937	0.504	0.9741

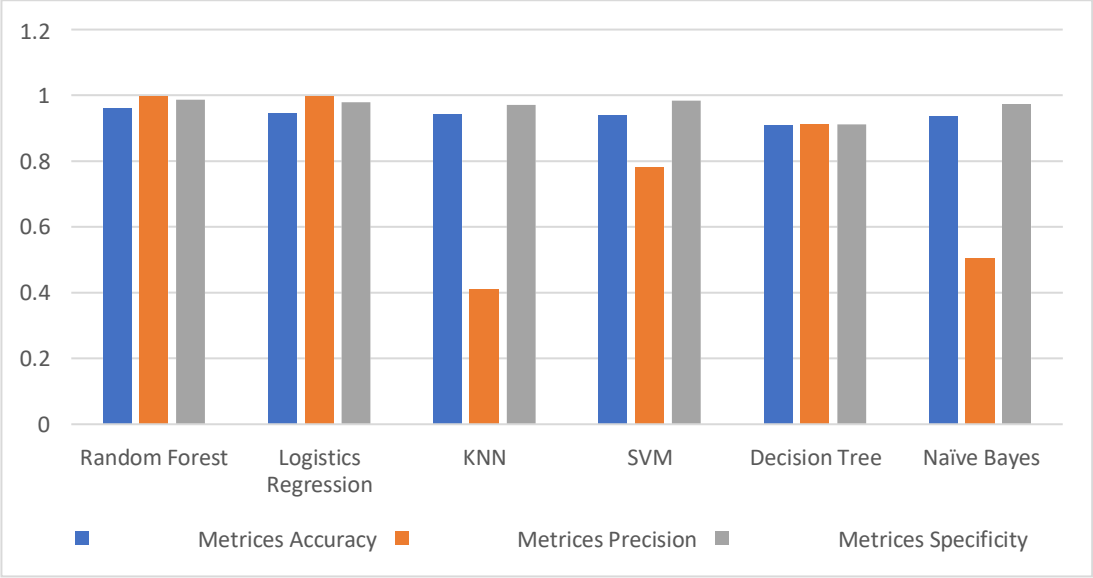

Figure 4.4 Accuracy and precision on different methodologies

Table 1 shows that random forest accuracy is superior to the other learning methods by a wide margin. In Fig. 9, we can see that Random Forest has the best accuracy, precision, and specificity, followed by Logistic regression and SVM. As a result, for bigger sets of training data, the suggested system employing random forest would perform better.

Credit cards are used for a multitude of purposes in today's modern culture. There has also been an increase in credit card transaction fraud over the last few years. Illegal credit card transactions result in enormous financial losses every year. Fraud may take many shapes and sizes, and it's not always easy to detect. As a result, credit card fraud detection difficulties must be addressed. In addition, fraudsters are finding new ways to perpetrate fraud as technology advances. ML techniques will be used to create a system for credit card transaction fraud detection that will present investigators with modest but reliable fraud warnings to solve this issue.

Following problems is addressed in the literature:

- Use feedbacks and delayed samples to train the model to recognize alerts and sum up their likelihood to do so
- The implementation of ML techniques to deal with the issue of idea drifts and class imbalance.
- Improve alert accuracy by using a learning-to-rank technique
- Introduce measures of performance that are relevant in the real world of FDS

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

In this work, a variety of machine learning algorithms have been explored to detect credit card fraud. Performances of these procedures are evaluated using accuracy and precision, as well as specificity and accuracy. To determine whether the alert is genuine or not, we have used the supervised learning technique Random Forest. Feedback and a delayed supervised sample will be used to train this classifier. To identify alerts, it will begin by aggregating each probability. The next step was to suggest a learning-to-rank approach, in which each warning is ranked according to its importance. The class imbalance and idea drift issues can be solved by using the method proposed here. Semi-supervised learning approaches will be used to classify alerts in FDS in the future.

To overcome the problem of under sampling in detection of suspicious transactions, swarm intelligence was used to select the relevant features, the Uniform Manifold Approximation and Projection (UMAP) method was used to reduce dataset dimensionality, and finally, the Synthetic Minority Oversampling Technique (SMOTE) was used to overcome the problem of under sampling. We have created a model that can detect fraudulent transactions by identifying patterns in client behavior.

We tested our algorithm on two other credit card datasets to confirm our findings and discovered that it can effectively identify fraudulent transactions. Furthermore, the performance of our model is superior to that of more current ones.

REFERENCES

[1] Jalinus, N., Nabawi, R. A., & Mardin, A. (2017). The seven steps of project based learning model to enhance productive competences of vocational students. *Advances in Social Science, Education and Humanities Research, 102*, 251-256.

[2] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems*, *29*(8), 3784-3797.

[3] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March). Random forest for credit card fraud detection. In *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)* (pp. 1-6). IEEE.

[4] Sahin, Y., & Duman, E. (2011, June). Detecting credit card fraud by ANN and logistic regression. In *2011 international symposium on innovations in intelligent systems and applications* (pp. 315-319). IEEE.

[5] Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916-5923.

[6] Sahin, Y., & Duman, E. (2010, March). Detecting credit card fraud by decision trees and support vector machines. In *World Congress on Engineering 2012. July 4-6, 2012. London, UK.* (Vol. 2188, pp. 442-447). International Association of Engineers.

[7] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015, July). Credit card fraud detection and concept-drift adaptation with delayed supervised information. In *2015 international joint conference on Neural networks (IJCNN)* (pp. 1-8). IEEE.

[8] Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by modified Fisher discriminant analysis. *Expert Systems with Applications*, *42*(5), 2510-2516.

[9] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2015, December). Detecting credit card fraud using periodic features. In *2015 IEEE 14th*

international conference on machine learning and applications (ICMLA) (pp. 208-213). IEEE.

[10] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019, January). Real-time credit card fraud detection using machine learning. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 488-493). IEEE.

[11] Wang, S., Minku, L. L., & Yao, X. (2014). Resampling-based ensemble methods for online class imbalance learning. *IEEE Transactions on Knowledge and Data Engineering*, *27*(5), 1356-1368.

[12] Behera, T. K., & Panigrahi, S. (2015, May). Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In *2015 second international conference on advances in computing and communication engineering* (pp. 494-499). IEEE.

[13] Carminati, M., Caron, R., Maggi, F., Epifani, I., & Zanero, S. (2015). BankSealer: A decision support system for online banking fraud analysis and investigation. *computers & security*, *53*, 175-186.

[14] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*, 235-255.