Unsupervised Anomaly Detection Using Machine Learning

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology

in

Electronics & Communication

(Communication)

By

Preksha Patel

(18MECC10)

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2020

Unsupervised Anomaly Detection Using Machine Learning

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology

in

Electronics & Communication

(Communication)

By

Preksha Patel

(18MECC10)

Guided By

Internal Guide:

Dr. Yogesh Trivedi,

Professor, EC Department

Nirma University, Ahmedabad.

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

INSTITUTE OF TECHNOLOGY, NIRMA UNIVERSITY

AHMEDABAD-382481

May 2020

Declaration

This is to certify that

1. The thesis comprises my original work towards the degree of Master of Technology in Communication at Nirma University and has not been submitted elsewhere for a degree.

2. Due acknowledgment has been made in the text to all other material used.

- Preksha Patel

18MECC10.

Certificate

This is to certify that the Major Project entitled **"Unsupervised Anomaly Detection Using Machine Learning"** submitted by **Preksha Patel (Roll No: 18MECC10)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Electronics & Communication (Communication) of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this Major Project part-I, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Date: Place: Ahmedabad.

Dr. Yogesh Trivedi
Internal Guide & Program Coordinator M.Tech-EC (Communication),
EC Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Dhaval Pujara
Head, EC Department
Institute of Technology,
Nirma University, Ahmedabad.

Dr. R. N. Patel
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Preksha Patel**, Roll. No. **18MECC10**, give undertaking that the Major Project entitled "**Unsupervised Anomaly Detection Using Machine Learning**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Electronics & Communication (Communication)** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

_____

Endorsed by

Dr. Yogesh Trivedi.

(Signature of Guide)

Acknowledgements

I take this opportunity to my express my profound gratitude and regards to my internship project guide **Dr. Yogesh N. Trivedi** for his extreme great support, exemplary guidance, monitoring and constant encouragement.

I thank to my parents, faculty members, friends and colleagues for their constant support and encouragement during this project work.

- Preksha Patel

18MECC10

Abstract

In current era, Internet has become the essential component of life. The Internet is one of the most influential innovations in recent history. Though most people use the Internet for productive purposes, some use it as malicious intent. The Internet and the computers connected to it increasingly become more enticing targets of attacks, As the Internet links more users together and computers are more prevalent in our daily lives. Computer security often focuses on preventing attacks using usually authentication, filtering, and encryption techniques, but another important facet is detecting attacks once the preventive measures are breached.

Prevention and detection complement each other to provide a more secure environment. Fraud detection methods are continuously developed to defend criminals in adapting to their strategies. Fraud detection techniques quickly identify frauds. Here, for credit card frauds, clustering approach is used. Data is generated randomly and then for detecting the transaction, K-means clustering algorithm is used. Clusters are formed to detect fraud in transaction which are low, high, risky and high risky. K-means algorithm is simple and efficient for credit card fraud detection. Clustering and related techniques have been used to locate anamolies in a dataset. The algorithms are implemented in Python language using the Spyder software provided by the Anaconda Distribution Platform. The results show that Isolation Forest algorithm for unsupervised anomaly detection have better accuracy compared to K-Means algorithm. As Credit card has the power to purchase the things, its frauds also increased.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

According to Fraud Benchmark Report by cyber source 83% of North American businesses conduct manual reviews, and on an average, they review 29% of orders manually. Involvement of humans gives insights about fraud patterns and genuine customer behavior. Over 90% of online fraud detection platforms use transaction rules to direct suspicious transactions through to human review. Now-a-days almost every important work of business depends on the computer network rather than manual analyzing. Presence of the outliers can create serious issues, so detection process is used. The Anomaly Detection also have the capability of identifying the threats as they arise and allows to react Immediately on the problem before it affects process.

1.2 Problem Statement

Earlier the top priority was the big data collection. Business leaders needs to find innovative ways to collect as much information about customers and operations as possible. Now this goal has been accomplished, a new problem has arisen. There is enough data available to optimize user experience, network performance, business operations, and more. There is an overwhelming amount of different metrics and systems to track, making it increasingly difficult to evaluate business patterns and, more importantly, deviations. For this anomaly detection plays such a critical role in the modern, efficient enterprise.

Anomalies in your business can be either positive or negative. In both cases, you need an efficient way to pinpoint the precise business incidents causing changes in data patterns.

1.3 Approach

To address this problem, anomaly detection typically uses models based only on non-anomalous data. In this approach, a model is constructed for the normal state of equipment. To detect anomalies, the degree of deviation between observed data and the normal state is then computed using this model. Here is a technique to identify the unusual patterns which do not comply as expected called outliers. This overview covers several methods of detecting anomalies,and also how to build a detector in Python using simple moving average (SMA) or low-pass filter. Unsupervised algorithms apply a statistical test to determine if a specific data point is an anomaly. A system based on this kind of anomaly detection technique is able to detect ones which have never been seen before.

1.4 Scope of Work

Anomaly detection algorithm will provide a scope for accurate and consistency results and application for complex systems and abnormal situations. Anomaly detection plays an important role in network security. Experimental work on anomaly detection in networking is also covered in this work.

1.5 Outline of thesis

- **Chapter 2 Literature Survey:**
  It starts with the workflow and Literature Studies, followed by the basics of Machine learning with its factors and categories.

- **Chapter 3 Software Design**
  This chapter discusses the detailed Machine Learning, its types and its use to detect anomalies. It also explains the clustering algorithms as here the unsupervised

learning is used.

- **Chapter 4 D7 Usecase Creator:**

  This chapter shows the detailed study about anomaly detection or outliers detection, followed by its categories, techniques, challenges. Also, it explains about the fraud detection and preventions with its types and detection anomalies.

- **Chapter 5 Simulation:**

  In this chapter, we will deal with Python coding and shows the algorithms used for detection using an example for unsupervised learning and compare the result analysis of algorithms.

Chapter 2

Literature Survey

Anomaly Detection is an important and dynamic research area in various field. This survey tries to provide a basic and structured overview of anomaly detection. The survey discusses the application domain where anomaly detection techniques have been applied and developed. Fraud attempts institutions have seen a drastic increase in recent years. Despite efforts on the part of the affected, hundreds of millions of dollars are lost to fraud every year. An important early step in fraud detection is to identify factors that can lead to fraud.

2.1 Workflow Of Literature Survey

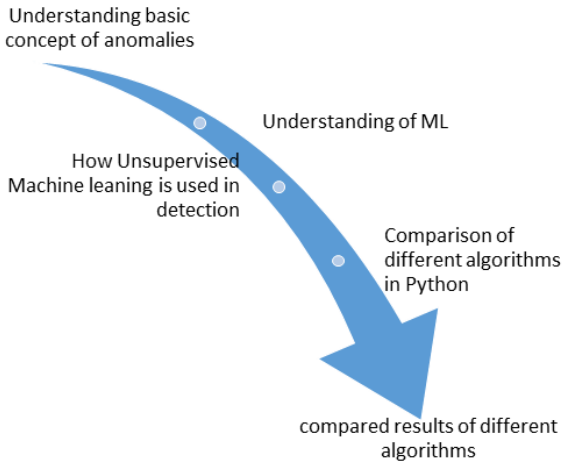

Figure 2.1: Flowchart of Literature Survey

2.2 Literature Study

A book on **Machine learning by Ethem Alpaydin et. al.** offers a concise overview of the subject for the general reader, describing its evolution, explaining important learning algorithms, and presenting example applications.

A book on **Machine learning by Saikat Dutt et. al.** introduces the concepts of machine learning in an easy to read and understandable manner. After a basic introduction to the concepts of machine learning, it expands deep into the details of supervised and unsupervised learning algorithms. Numerous worked-out problems help for easy understanding.

Andrew Ng, a professor at Stanford University delivers an 11-week course on Machine Learning in Coursera. The course starts with an introduction to machine learning, various parameters in machine learning and then goes in detail into various supervised and unsupervised algorithms. The algorithms covered are Linear Regression, Logistic Regression, Neural networks, Support Vector Machines, Principal Component Analysis. Some practical problems covered spam detection, image recognition, clustering and building recommender systems. This course includes programming assignments to be done using Octave or MATLAB.

Machine Learning Classification **Bootcamp** in Python is an online course available in Udemy, which focuses on building practical projects Machine Learning using Python and Scikit Learn library. This course provides knowledge and hands-on experience in machine learning classification algorithms such as Isolation Forest and Visualisation, K-means, DBSCAN. The course contains practical hands-on python coding projects for practise.

KAI LI et. al. supported by the educational department of the Hebei province. discusses about clustering algorithm based on models. Framework of clustering for objects of model is presented. As its application, a method for improving diversity of ensemble learning for neural networks is investigated. The relations between the number of cluster

in clustering analysis, the size of ensemble learning, and performance of ensemble learning are studied by experiments.

Ramyashree. K et. al. discusses about one of the concept of machine learning that is data mining, that is used for credit card fraud. This concept is enhanced to online learning models. Use of internet instruction to allow the quick awareness of credit card fraud, the proposed system is help to detect and before prevent the fraudulent transaction and activities, so to decrease the unit of dropping in economic industry.

2.3 Machine Learning

The three factors which explain the importance of machine learning are – Speed, Scale, Efficiency.

Speed: Machine learning can evaluate huge numbers of transactions in real time. It is continuously analyzing and processing new data.

Scale: The cost of maintaining the fraud detection system multiplies as customer base increases. The ML model can pick out the differences and similarities between multiple behaviors.

Efficiency: Contrary to humans, machines can perform repetitive tasks. ML can often be more effective than humans at detecting subtle or non-intuitive patterns to help identify fraudulent transactions.

2.3.1 How to detect Fraud using Machine Learning?

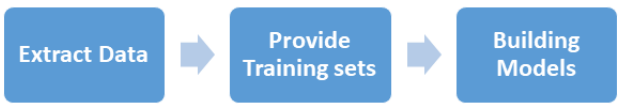

Figure 2.2: Categories of ML

Machine learning can be summed up into four different categories:

- Regression: predicting a value.
- Classification: predicting a type or category.
- Clustering: finding the hidden structure of data.
- Dimensionality Reduction: getting the set of principal values in data.

Chapter 3

Machine Learning

Machine learning is a subset of Artificial Intelligence, and the key difference is the learning. With machine learning, we are able to give a computer a large amount of information and it can learn how to make decisions about the data, similar to a way that a human does.

3.1 Factors for Machine Learning

The three factors which explain the importance of machine learning are – Speed, Scale, Efficiency.

Speed: People create ad hoc rules to determine which types of orders to accept or reject. This process is time-consuming and involves manual interaction. As the velocity of commerce is increasing, it's very important to have a quicker solution to detect fraud. We want results fast.

Scale: The cost of maintaining the fraud detection system multiplies as the customer base increases. The ML model can pick out the differences and similarities between multiple behaviors. There is a risk in scaling at a fast pace. If there is an undetected fraud in the training data machine learning will train the system to ignore that type of fraud in the future.

Machine learning has four common classes of applications:

- Classification,
- Predicting next value,
- Anomaly detection, &
- Discovering structure.

Classification: It is a process of placing each individual from the population under study in many classes. This is identified as independent variables. It helps analysts to use measurements of an object to identify the category to which that object belongs.

Prediction: It refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when you're trying to forecast the likelihood of a particular outcome.

Anomaly: The detection is the process of identifying unexpected items or events in datasets which differ from the norm. It is often applied on unlabelled data known as unsupervised anomaly detection. It is applicable in domains like intrusion detection, event detection in sensor networks, fraud detection, fault detection, detecting ecosystem disturbances and system health monitoring. To remove anomalous data from the dataset in pre-processing, it is often used.

It has to basic assumptions, first one is it occurs very rarely in the data and second one is their features differs from normal samples undoubtedly. It is referred to the identification of items or events that do not conform to an expected pattern or to other items present in a dataset. Machine learning algorithms have the ability to learn from data and make predictions based on that data.

Discovering structure: This will get the set of principal variables in data and analyse it.

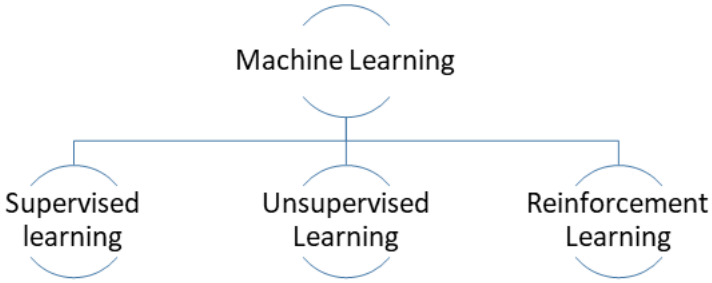

Figure 3.1: Types of Machine Learning

3.1.1 Types of Machine Learning

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

3.1.2 Supervised Learning

Supervised Learning involves direct supervision of the operation. In case, the developer labels sample data corpus and set strict boundaries upon which the algorithm operates. Primarily the scope of data is scaled and predictions of unavailable, future or unseen data based on labeled sample data are made.
It includes two major processes:

- Classification
- Regression

Classification: The process where input data is labeled based on prior samples and manually trains the algorithm to recognize types of objects and categorize them. The system learns how to differentiate types of information, perform an optical character, image, or binary recognition. When the outputs are restricted to a limited set of values, it is used.

Regression: The process of identifying patterns and calculating the predictions of continuous outcomes. The system learns the numbers, their values, grouping. When the outputs may have any numerical value within a range, it is used.

3.1.3 Unsupervised Learning

It is very much the opposite of supervised learning that features no labels. Here, the algorithm would be fed a lot of data and tools are provided to understand the properties of the data. It can learn to group, cluster, and/or organize the data so other intelligent algorithms can make sense of the newly organized data. In supervised machine learning results are known and need to sort out the data, while in unsupervised machine learning algorithms the desired results are unknown and yet to be defined. Unsupervised learning feeds on unlabelled data.

The unsupervised machine learning algorithm is used for:

- exploring the information structure;
- extracting valuable insights;
- detecting patterns;
- implementing this into its operation to increase efficiency.

It describes information by sifting through it and making sense of it. This algorithms apply the following techniques:

Clustering: An analytical data used to separate into specified groups based on their internal patterns without prior knowledge of group credentials that are by similar to individual data objects and dissimilar from the rest.

Dimensionality reduction: The incoming data consists of lots of noise that is removed using dimensionality reduction while distilling the relevant information.

The widely used algorithms are:

- k-means clustering;
- PCA (Principal Component Analysis);
- Association rule.

3.1.4 Reinforcement Learning

One of three basic machine learning paradigms, which is very behavior-driven. This has influences from the fields of neuroscience and psychology. It is an area that is concerned with software agents that how to take actions in order to maximize some notion of cumulative reward. It is all about developing a self-sustained system throughout contiguous sequences of tries and fails, improves itself based on the combination labeled data and interactions with the incoming data. It uses the technique called exploration/exploitation. The mechanics are simple - the action takes place, the consequences are observed, and the next action considers the results of the first action.

Most reinforcement learning algorithms include:

- Q-Learning;
- Temporal Difference (TD);
- Monte-Carlo Tree Search (MCTS);
- Asynchronous Actor-Critic Agents (A3C).

It is used to amplify and adjust natural language processing (NLP) and dialogue generation for chatbots to:

- mimic the style of an input message
- develop more engaging, informative kinds of responses
- find relevant responses according to the user reaction.

3.2 Clustering

Clustering is an unsupervised learning technique that aims at grouping a set of objects into clusters. The objects in the same clusters should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in other clusters. It aims to group a collection of patterns into clusters based on similarity. For comparing various data items, clustering technique uses a similarity function. It is one of the most common exploratory data analysis techniques that is used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data. It is the main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Clustering using distance functions, called distance-based clustering, which is a popular technique to cluster the objects and has given good results[1].

One can identify the following scheme while analyzing cluster:

1. Determine data representation
2. Preprocess data
3. Cluster data
4. Validate results
5. Iterate This is an iterative process.

A third formulation phrases it in terms of density. It can be described as regions of high density and are separated from other by low density.

3.2.1 Clustering for Anomaly Detection

The main aim is to learn the normal mode in the data already available (train) and then using to point out if is anomalous or not when new data is provided (test).

3.2.2 General steps (after general pre-processing)

1. Select the best model according to your data.

2. Fit the model to the training data, this step can vary on complexity depending on the chosen models, some hyper-parameter tuning should be done at this point.

3. Compare new data with the results of the model and determine if an anomaly or not, the way to perform this classification depends highly on the mode, some are based on distance, while others use probabilities.

4. If data evolution over time, the model should be retrained after a while, in order to learn new behavior.

Chapter 4

Anomaly Detection

Applications in business are from intrusion detection (identifies strange patterns in network traffic that could signal a hack) to system health monitoring (spotting a malignant tumour in an MRI scan), and from credit card fraud transactions to fault detection in operating environments. Anomaly detection is the identification of rare items, events or observations which raise suspicious by differing significantly from the majority of the data.
[2].

- It is similar to noise removal and novelty detection.
- Novelty detection is identifying an unobserved pattern in new observations that are not included in training data, for instance, a sudden interest in a new channel on YouTube during Christmas.
- Noise removal (NR) is the process of removing noise from an otherwise meaningful signal.

4.1 Anomaly Detection Techniques

4.1.1 Simple Statistical Methods

It is used to identify irregularities in data that is to flag the data points that deviate from common statistical properties of distribution. It is one that deviates by a certain

standard deviation from the mean. Traversing mean over time-series data is neither trivial nor static. You would need a rolling window to compute the average across the data points which is technically called a rolling average or a moving average. It's intended to smooth short-term fluctuations and highlight long-term ones. An n-period simple moving average can also be defined as a "low pass filter."

4.1.2 Challenges with Simple Statistical Methods

In simple use cases, the low pass filter identifies anomalies and there are certain situations where this technique won't work, like:

- As the boundary between normal and abnormal behavior is often not precise, the data contains noise which might be similar to abnormal behavior.
- As malicious adversaries constantly adapt themselves, the definition of abnormal or normal may frequently change. So the threshold based on a moving average may not always apply.
- To identify the change in seasonality involves sophisticated methods that uses patterns, such as decomposing the data into multiple trends.

4.2 Fraud Prevention

Fraud prevention works proactively, that stop frauds prior it occurs such as watermarks, passwords, holographs on banknotes. Credit Cards can get skimmed and Passwords can get stolen, so fraud detection comes here.

4.3 Fraud Detection

Once fraud prevention has failed, fraud detection works reactively. If a fraudster knows a fraud detection system is in a place one can figure out new ways to work around it, so this need to be continuously developed. While creating this system one has to be careful that new fraudsters can use new as well as old techniques tried by other fraudsters earlier. Under the hood it is the classification of data that can be done with several

approaches of the types supervised, reinforced or unsupervised. Statistical tools used for detection are varied due to different sizes and data types. But the observed data is compared with expected values. Statistical fraud detection methods can generally be divided into supervised and unsupervised methods. Methods used are from traditional statistical classification methods such as discriminant analysis and also powerful tools such as artificial neural networks which have proven successful. Rule-based methods have been applied extensively within supervised fraud detection which can be explained as a series of "if-then-else" cases. Unsupervised fraud detection does not need examples of fraudulent and non-fraudulent behavior as it works by finding entries that are dissimilar to the rest of the dataset.

4.4 Detection of Anomaly

To classify the point as anomalous or not, the following is considered:

1. Calculate the distance from the new points to all the core points that actually defining the clusters and find closest neighbor inside a cluster.

2. Compare the distance to the closest neighbor inside a cluster with eps, as this is the limit between two points to be considered neighbors with our test data.

3. If the distance is larger than eps the point is labeled as anomalous as it has no neighbors in the clusters.

Chapter 5

Algorithms

5.1 Isolation Forest and Visualisation Algorithm

An anomaly detection framework that is able to pre-processing, training, predicting data transactions in real time is built. The adapted model using isolation forest as follows: When new transactions arrive, the iForest algorithm is occupied in order to assign a score to the observation that will clarify whether the transaction is fraudulent or not. Anomaly score close to 0 is considered normal, value 1 is considered as an anomaly [2].

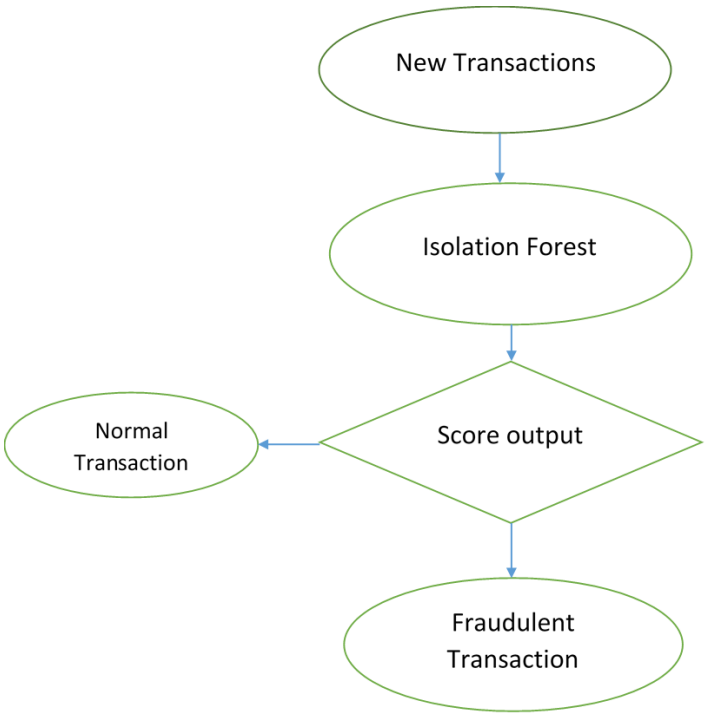

Figure 5.1: Workflow of Credit Card fraud Detection

Anomaly detection using Isolation Forest has two stages:

- **Training stage:** builds IForest, we start to multiple isolation trees using subsamples of the training set.
- **Testing Stage:** passes each data point through isolation tree to calculate the average number of splits across all the trees that isolate observation in order to obtain an anomaly score for each instance [3].

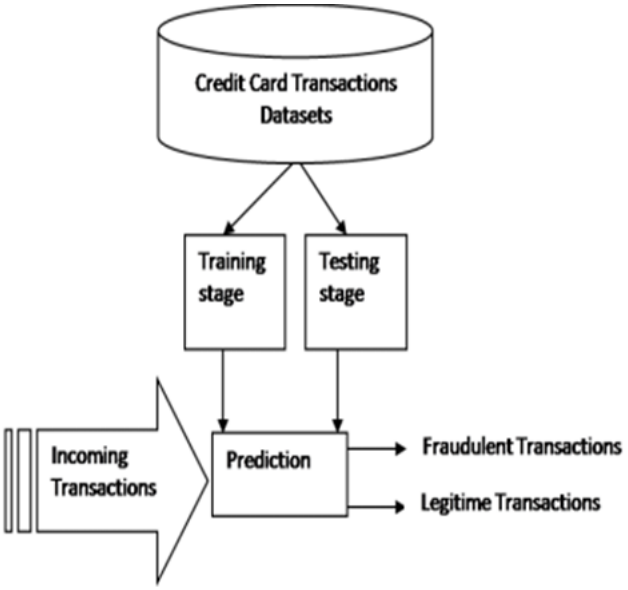

Figure 5.2: Credit Card fraud detection Model

```
the Model used is Isolation Forest
The accuracy is  0.9978933323970366
The precision is 0.375
The recall is 0.336734693877551
The F1-Score is 0.3548387096774193
The Matthews correlation coefficient is0.3543008067850027
```

Figure 5.3: Results

5.2 K-means clustering algorithm

The algorithm is composed of the following steps:

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated [4].

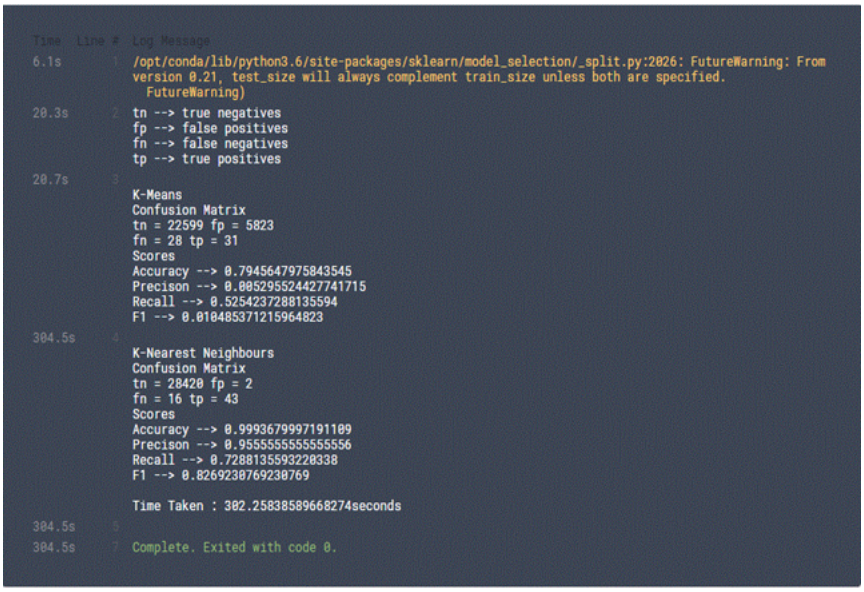

Figure 5.4: K-Means Algorithm Output

5.3 DBSCAN Algorithm

It is a density-based clustering non-parametric algorithm: There is a set of points in some space which groups closely packed together points with many close neighbours that consider it as outliers lie alone in low-density regions.

5.3.1 Parameters:

It requires two parameters:

1. **Eps:** It defines the neighborhood around a data point, i.e. if the distance between two points is lower or equal to eps then they are considered as neighbors.

2. **MinPts:** Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. MinPts ¿= D+1; Minimum Value must be chosen at least 3.

DBSCAN can be used with any distance function. The distance function (dist) can therefore be seen as an additional parameter.
The DBSCAN algorithm can be abstracted into the following steps:

1. Find the points in the ϵ (eps) neighborhood of every point, and identify the core points with more than minPts neighbors.

2. Find the connected components of core points on the neighbor graph, ignoring all non-core points.

3. Assign each non-core point to a nearby cluster if the cluster is an ϵ (eps) neighbor, otherwise assign it to noise.

In the Apriori data there is no requirement to specify the data that is as opposed to k-means. It also finds arbitrarily shaped clusters. It has a notion of noise, and is robust to outliers. It requires just two parameters and is mostly insensitive to the ordering of the points in the database [5].

5.3.2 Flowchart

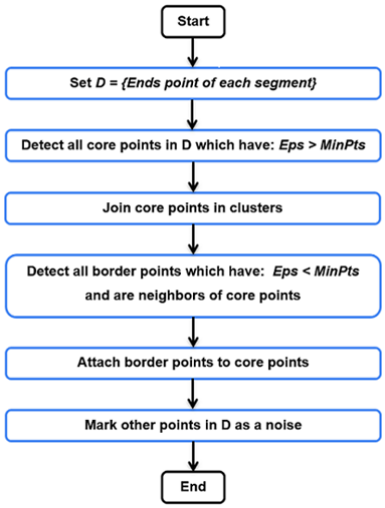

Figure 5.5: Flowchart

5.3.3 Types of Data Points

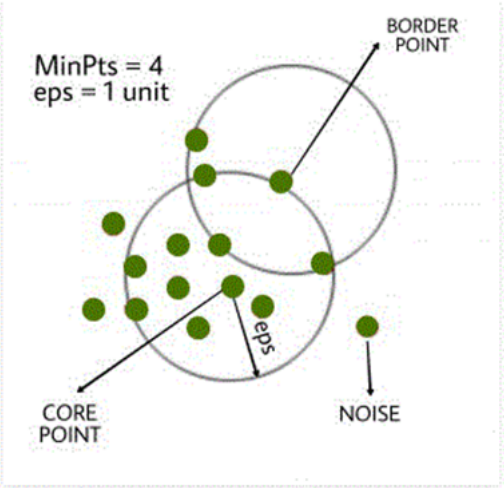

Figure 5.6: Types of Data Points

- **Core Point:** Points more than MinPts points within eps.
- **Border Point:** A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.
- **Noise or outlier:** A point which is not a core point or border point.

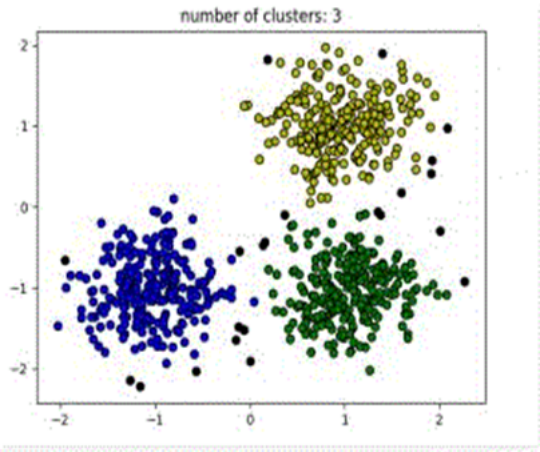

Figure 5.7: Results

5.4 SOM

It is a process that maps the input patterns in a high-dimensional vector space to a low-dimensional (typically 2-D) output space, the feature map, so that the nodes in the neighborhood of this map respond similarly to a group of similar input patterns.

A type of ANN which is directed using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction.

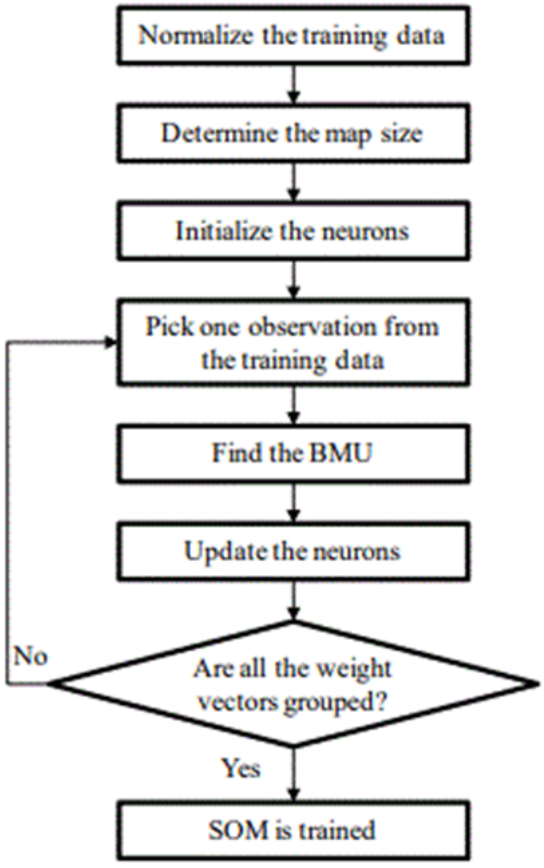

Figure 5.8: Flowchart

Algorithm:

1. Each node's weights are initialized.

2. A vector is chosen at random from the set of training data.

3. Every node is examined to calculate which one's weights are most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU).

4. Then the neighbourhood of the BMU is calculated. The amount of neighbors decreases over time.

5. The winning weight is rewarded with becoming more like the sample vector. The closer a node is to the BMU, the more its weights get altered and the farther away the neighbor is from the BMU, the less it learns [6].

6. Repeat step 2 for N iterations

5.5 Confusion Matrix

The columns of the matrix represent the predictions, and the rows represent the actual class. Correct predictions always lie on the diagonal of the matrix. The general structure of confusion matrix is given below

$$ConfusionMatrix \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

where, True Positives (TP) indicate the number of instances of the minority that were correctly predicted, True Negatives (TN) indicate the number of instances of the majority that were correctly predicted. False Positives (FP) indicate the number of instances of the majority that were incorrectly predicted as minority class instances and False Negatives (FN) indicate the number of the minority that were incorrectly predicted as majority class instances. The confusion matrix gives a better outlook than accuracy.

Chapter 6

Conclusion

6.1 Conclusion

Due to the advancement in the technology, there is increase in the number of frauds. It is not possible to secure or detect the real time data manually. There will be the requirement of Machines for detecting and securing big amount of data. Thus the detection of anomalies using Machine Learning techniques have become one of the reliable approaches to counter this illegal activity.

The four Algorithms used for detecting the anomalies have different base for their application. K-Means algorithm is simplest one but DBSCAN algorithm is for real time data compared to K-means. The goal of self-organizing map is to cause different parts of the network to respond similarly to certain input patterns. While the Isolation Forest Algorithm has better accuracy rate compared to all other of around 99%.

Bibliography

[1] K. Li and L. Cui, "Study of clustering algorithm based on model data," in *2007 International Conference on Machine Learning and Cybernetics*, vol. 7, pp. 3961–3964, 2007.

[2] A. Krishnan., "Anomaly detection with isolation forest & visualization." Available at https://towardsdatascience.com/anomaly-detection-with-isolation-forest-visualization-23cd75c281e2 (2019/28/11).

[3] S. Ounacer, H. Ait el Bour, Y. Oubrahim, M. Ghoumari, and M. Azzouazi, "Using isolation forest in anomaly detection: the case of credit card transactions," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 6, p. 394, 11 2018.

[4] I. Dabbura, "K-means clustering: Algorithm, applications, evaluation methods, and drawbacks." Available at https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e64 (2019/28/11).

[5] K. Salton., "How dbscan works and why should we use it?." Available at https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80 (2020/19/02).

[6] A. Ralhan., "Self organizing maps." Available at https://towardsdatascience.com/self-organizing-maps-ff5853a118d4 (2020/18/03).