University of Essex

Department of Mathematics

MA981- DISSERTATION

Credit Fraud Detection Using Data Science

Srikanth Reddy Gudibandi

2004445

A thesis submitted for the degree of MSc in Data Science

Supervisor: Dr Jesus Martinez Garcia

September 3, 2021

Colchester, United Kingdom

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 History of Banking

Banking is one of the oldest institutes which was established hundreds of years ago. These establishments have been recorded those dates to 2000 BC in countries like India, Greece, Roman Empire, etc [1]. 'Banca Monte Dei Paschi di Senia' is the oldest Italian bank which is still fully functional in 2021. It was mainly the merchants based in temples, who provided loans to the grain carrying farmers and the trades in their respective empires during the Palaeolithic age. Before the banking system was established, it was mainly the barter system which existed where two or more parties agree to trade in exchange of goods, services, or both.

During the medieval ages which was around 11th century, Italy was the centre where traders from all around the Europe would come to trade their goods with the fellow traders and people. Different traders from different countries had different currencies which use to create a problem during exchanges. These traders used to sit on a bench which is also known as 'Banco' in Italian language where they use to exchange and count their money hence, the name Bank came into existence. There was a dire need for a common currency and Gold was made as the main currency for all. The reason for gold becoming the most important currency is because of its durability, non-corrosive nature, and limited supply.

Over the years it became insanely difficult for the wealthy traders to store their assets, as there was always a risk of losing that gold during the ocean travels, theft, etc. To tackle this issue, traders started to partner up with the goldsmiths. Usually, goldsmiths kept their gold safe in a safety vault which was known to be very safe. Traders started giving all their gold to these goldsmiths who would keep their gold safe over a small fee for the service. This is how the concept of 'Safe Deposits' came into existence. This system became so successful that all the goldsmiths started getting huge deposits of gold from the merchants and traders. For all the gold that the traders deposited, goldsmiths issued them with receipts certifying the quality of the gold. These receipts were not transferable, only the trader who deposited the gold could come and collect the gold.

Over the period, goldsmiths came up with a wonderful plan of lending the money to the ordinary people, institutes, etc, on behalf of the trader's gold. This was done by issuing the borrower with a certificate which eventually developed into a modern-day banking practice. By lending the money, goldsmiths were able to make interest, and this was a very profitable idea. Even the traders were able to get a good share of the interest. The borrowers found it convenient to carry a piece of receipt rather than a pouch of gold which was heavy and risky to carry around. These paper receipts were very much respected that people adopted them very quickly and this became the primary form of money in those days.

There were some issues with this system just like every system does. Fake gold coins started to circulate as a few traders started to trade gold alloys as pure gold and the weights of the coins were not equal.

This could be termed fraud. In order to tackle this fraud, reputed goldsmiths used to forge their symbols on the coins which weighed a lot of value in the open market since the common people could trust that coin.

Over the years, goldsmiths started lending receipts beyond what they were capable of. This worked very well for a while, but it didn't take much time for this idea to crash. For example, if a goldsmith has a gold reserve worth 1 million pounds of his own and 5 million of the traders, he should only be able to lend receipts worth 6 million. A lot of times greedy goldsmiths lend receipts that were way above their gold reserve's value. In such cases, a lot of receipts started floating across the economy and when the trader brought their receipt to get their share of gold and did not get any gold in return, that led to a massive panic among the people. In such situations, everyone would start bringing their receipts to cash their gold out which doesn't exist. The name of such issue is called 'Bank Run'. Bank Run happens even in the 21st century. However, to avoid such events happening again, these banks are institutionalized, and they run under a central bank which monitors them. The rules and regulations may vary in different countries.

Bank of England became the first banking institute to issue currency notes from the year 1695. These notes were short-lived notes which were hand-written. However, they were replaced by standard printing notes which were issued after the year 1745. A wide variety of notes were released in the open market where they ranged from 20 to 1000 denominations.

1.2 Modern Day Banking

The banking system was first brought into existence to make life easy, however, nowadays it has become extremely complex. In this 21st century, banks are in the risk management business [2]. Nowadays, all the banks across the world take money from their customers and clients in the form of printed notes, digital money, assets, gold, etc, and provide them some interest. This interest usually differs from bank to bank and country. In the UK, the interest rate is typically around 3% whereas in India this is around 6-8%. Banks cannot make a profit only by taking money and giving interest, they must lend the money as well which usually enables them to make big margins. Banks in the UK charge more than 6% for educational loans and banks in India charge more than 10%. The interest rate is typically higher when the banks lend the money compared to the event when the bank takes the money from their customers, it is a calculated risk. There are a lot of occasions when the borrower defaults on the debt, and the bank must bear all the losses. In recent times, banks have been providing credit in a lot of different ways, they are in the form of personal loans, home loans, education loans, auto loans, gold loans, etc. A lot of jobs have been created in the last 30 years across the globe especially due to the rise of globalization and the internet. A lot of world economies have been kicking off at an unprecedented rate, India is a good example. People have been making a lot of purchases of goods and

services especially online. Online retailers like Amazon, Walmart, Tesco, Asda, Alibaba, etc are great examples. Banks use their marketing strategies to make the customers buy some products that they don't even need [3]. We have built an economy that 70% depends on consumer spending. Banks provide interesting offers to the customers to make them buy a few products at discounted prices through credit. For example, Bank of America may provide 10% off on a laptop if bought on Amazon through a credit card. A customer may get tempted to buy that product because of its discounted price but they end up getting into debt that they must clear within a timeframe.

A Credit card has become a very important part of people's life recently. It is easy to carry and store. In the UK, 62.8 million credit cards have been issued to the residents as of January 2021. It is very important for a person to spend wisely to avoid the credit card debt trap. A lot of people become bankrupt because of the bad choices they make and a series of events that trigger this cause.

People are living with a lot of debt over their shoulders. According to The Money Charity, the average debt per household in the UK including the mortgage was 62000 by May 2021. Average credit card debt averaged to be around 2000 per adult [4]. This data suggests that Banks would be able to make a massive profit in the form of interest from their customers. However, there is a problem, making profits is only possible if the customer pays back. There are a lot of instances where customers default on their loans and credit card bills. On the other hand, online frauds have increased exponentially over the years. This is a major issue in almost every country on this planet. Every year banks lose millions of pounds over poor loans, debts, and frauds.

1.3 Fraud Process- Credit Card

Digital money has become very convenient for people to do transactions. The only thing a customer needs is a plastic card which could be either a credit or a debit card. Almost all the banks in the UK provide credit cards to their customers. A credit card model is very simple, a bank gives credit to their customers at the start of the month. The credit value varies across the customers, also known as Credit Limit, which depends upon their credit score. This credit can be used anywhere in society. They can be used to purchase a wide variety of goods and services. A person with a credit card can use it in places like Tesco and Sainsbury to buy daily household items and vegetables and the person can also use his/her card to buy an iPhone from an Apple store. Now that a person has used his/her credit for the month, they should pay the credit back to the bank within a certain time limit. By doing so people can avoid paying the interest for using the bank's credit. A lot of people with stable jobs use credit cards in their daily life and set any auto-debit for the credit card bills. By doing so, they not only avoid paying interest but also increase their credit score. On the other hand, people who do not have enough money to pay their credit card bills, end up paying only the interest of the principal amount. For example, a person could have £1000 credit card debt that needs to be paid before the end of September. If he/she

doesn't have a £1000, they might push this bill to the next month by paying only the interest for the principal amount for September. Usually, this interest is around 10-12% of the principal amount. This is how banks make profits on the credit card business. They usually want people who only pay a part of their debt along with interest. In the United States, big financial giants like Barclays made $3.1 billion in total over credit interest in the year 2019 [7]. They have 16.3 million accounts, and they made an average of $180.5 per account in the form of credit card interest fees.

Credit Card has become the primary need for individuals over the past few years. However, it has its severe downfalls too. There are a lot of ways how banks lose their money. Defaulting could be blamed as the main cause and fraud could be termed as a secondary cause. There is a difference between default and fraud. A default could be defined as a situation when a customer cannot afford to pay the bills because of financial hardships, etc. A customer can also intentionally not pay which could be termed as fraud.

Generally, fraud could be termed as an event where the customer loses money unknowingly. A credit card fraud could be described as an event where a credit card has been used in the fraud process.

People have adopted the online purchasing habit over the last decade. It has been observed that the majority of the frauds happened during online purchases. The main reason could be the lack of fraud knowledge among the users. A smart conn man can easily trick a certain section of the society by acquiring their card details easily. This section usually comprises vulnerable old and uneducated people.

A rise of almost a 2000% increase in credit card frauds has been observed over the past 16 years. In 2020 alone, card frauds that sum up to £575 million have been identified according to UK Finance [5]. It is not possible for the banks to refund the customers in all cases. Often customers end up bearing the loss made by the fraudsters. In the year 2017-2018, people in the UK who became fraud victims have lost an average of £800.

There are mainly three types of credit card fraud [6]:

1. Lost/ Stolen Card
2. Card Skimming
3. Phishing, etc.

A lost card is the easiest way to lose money to the fraudster. The card can be used by the fraudster to make payments anywhere. They can use the card in Walmart to purchase products, etc. Most of the time, lost credit card details are sold to fraudsters in other countries. It can be sometimes observed that fraud transactions happen from countries where the victim has never been to. Hence, it is very important to inform the bank regarding the lost card to avoid ending up in such trouble. Banks usually blocks the card and flag it as a fraud transaction/attempt.

Card skimming is a very clever and advanced technique where a fraudster records the card details of the victim with the help of a small electronic device. Such type of frauds generally happens in places like restaurants and gas stations. The fraudster could take the user's card for swiping purposes and secretly skim the details without the owner knowing about it. It is very important to avoid giving the card to other people for payment purposes.

The third type of fraud is called Phishing. These are the most common frauds that happen everywhere. The fraudster usually pretends to be a member of the credit card customer support, and they try to obtain a vast amount of information from the customer to log in to their device and make unauthorized payments. Usually, the fraudster tries to get information like date of birth, address, secret questions, etc to get the user's online bank account/ credit card access. Once these details are compromised, it only takes a split amount of time for the fraud to occur and there is very little that the victim could do about it.

The most common types of fraud examples could be IRS fraud, Microsoft antivirus fraud, and online car insurance fraud. These frauds come under the Phishing category. The customer tries to represent himself/ herself as the customer care operative who usually have little information about their target victims. These fraudsters try to gain sensitive information from the customers by asking them tricky questions.

1.4 Data Science in Fraud Detection

Identifying fraud is a very difficult task. Organizations spend a great amount of time and millions of pounds to create strategies to tackle fraud activities. It is not always possible to stop fraud at certain times. However, data science has shown proven results that tackle this issue with better and accurate results. Techniques like clustering analysis and supervised learning algorithms can find patterns in the fraud database and can easily classify fraudulent transactions so that immediate actions can be taken. The fraud detection process can never be completely automated using data science. It requires a hybrid model where a data science model will classify a set of transactions into fraud/non-fraud and the fraud team can investigate a few sensitive cases manually.

1.5 Objective

This dissertation report will explain the working functionality of the data science algorithms and statistical techniques that will solve the credit card fraud detection problem. The data science lifecycle consists of a lot of steps. The majority of these steps will be implemented and explained in this report. After a thorough preliminary analysis of the credit fraud dataset, a certain number of data science

algorithms will be used to classify the transactions into fraud and safe categories. This report will mainly focus on the working functionality of different data sampling techniques.

The primary objective of this project is to perform a comparative study on how different data science algorithms perform on various data samples that have been created using sampling techniques versus the data that has not been sampled at all. Various evaluation metrics will be used to assess the performance of the statistical models.

2. Problem Statement

It has been observed over the past few years that credit card frauds have skyrocketed. Banks are investing a lot of resources in terms of creating strategies and hiring a lot of data scientists in their teams to restrict such frauds from happening.

2.1 Rule Based Learning

During the early stages of fraud detection, financial institutes usually were dependent on rule-based learning approaches [8]. Subject matter expertise, who are usually fraud investigators, had a good understanding of the fraud patterns. They used to write certain rules which were extremely complex. Every credit card transaction had to pass through these rules. These rules were some of the attributes which explains the characteristics of the transactions. The attributes explain a lot of information like account number, transaction amount, frequency, location, etc. If a transaction doesn't pass through these rules, it is flagged as a fraud. The most common and important rules are the one that explains the fraudulent characteristics of the transaction. These are usually location, frequency, amount, merchant, time, device, etc.

People usually have a certain spending characteristic; their spending pattern remains the same most of the time. These rule-based learners observe these patterns and learn from them. When a fraudster gets the access to victim's card details, they tend to spend money very rapidly at a high frequency. Their only aim is to get as much money as they could before the victim informs the bank. At times, the victims do not notice these unauthorized transactions on time, and they lose all the money. In such scenarios, rule-based learners work brilliantly. These learners observe these anomalies and blocks these transactions immediately. Fraud investigators judge the nature of such transactions and let the victim know about them. By doing so, both the bank and the customers do not lose any money.

Sometimes a user may spend too much money at a rapid speed. This could happen during online sale seasons during Christmas. Unfortunately, rule-based learners often misclassify them as an anomaly, and they are stopped by the system [10]. This could be a serious issue as this leads to a bad customers experience. For example, a customer who usually doesn't spend much can make a big purchase on Amazon.com during a flash sale on Christmas to buy a laptop which is at discounted price. Since laptops are expensive, the customer's transaction amount would be huge, and the bank's system may block it and the customers could end up not getting the product. Often, the rules are not well defined by the fraud analysts which could lead to situations where fraud may occur, and the bank doesn't even know until the victim lets them know.

Creating a data science solution is the only possible solution for this problem. Data science algorithms can learn most of the patters from the available data and it reduces the human interventions up to a great extent.

2.2 Class Imbalance Issue

The majority of the financial institutes that issues a credit card to their customers, get an overwhelming amount of data which is transaction information. Unfortunately, the fraud cases are often very less if compared to that of non-fraud cases. The fraud cases comprise mostly less than 1% of the entire transaction dataset. The class distribution is extremely skewed toward the non-fraud cases. It becomes very difficult for the data science algorithms to train on such data unless some robust sampling techniques are applied.

2.3 Curse of Dimensionality

Financial institutes generally collect a lot of information about their customers and their transactions. This information is stored in the relational databases in the form of columns. Training machine learning models on such data with a lot of columns could be a challenging task. Using the right features is very important in such scenarios. A good feature selection technique is needed to overcome this issue.

Getting fast results is extremely important in the financial sector. When a person with a credit card makes a transaction, the machine learning algorithm should be able to classify the transaction correctly within a fraction of seconds. Algorithms like SVM could take more time to perform this operation [11,12]. It is very important to choose an algorithm that does not take much time to classify.

In this essay, all the shortcomings addressed above will be discussed in detail and a better alternative will be implemented. The class imbalance issue will be addressed using a wide variety of data sampling techniques. A set of different classification algorithms will be used for training the models using the sampled data and the model's performance will be evaluated.

3. Literature Survey

A paper by K. R. Seeja and Masoumeh Zareapoor [13] discusses the use of Fraud Miner in credit card fraud detection. This algorithm creates fraud and non-fraud transaction patterns from the training data during the training phase. This is achieved with the help of the frequent itemset technique. The incoming transactions are matched with the patterns created by the Fraud Miner algorithms and are labeled appropriately.

Research carried by V. Dheepa and R. Dhanapal [14] demonstrates the use of behavior-based techniques using SVM with different paraments. Four different kernel tricks have been used to train the SVM models and it was found out that the RBF kernel produced the best results. The model produced a high true positive rate with a very low false-positive rate.

A study carried out by Asha RB, Suresh Kumar 2021 demonstrates the use of SVM, KNN, and Artificial Neural Network. ANN model outperformed the other two models in terms of accuracy. However, it doesn't discuss much about sampling techniques used.

M. Sathyapriya and Dr. V. Thiagarasu [15] discuss the idea of using clustering techniques and hidden Markov models to solve the fraud detection problem. Initially, a K-means clustering is performed on the expenditure database of the customers and transaction amounts are clustered on a customer basis. Finally, HMM models are applied to the customer spending category them into three levels (low, medium, and high). During testing, anomaly detection was done based on the expenditure levels.

4. Data Science Lifecycle: An overview

4.1 Data Science Overview

Data science is a field where data is thoroughly studied to get meaningful insights. It comes under a group of multiple fields which includes data analysis, artificial learning, statistics, etc. All these fields combine to form a tool that can extract information from a raw data source that could be very beneficial for the organizations to make business decisions. Big data is a term that is often used today. It can be defined as a data repository that contains a huge amount of data which could be in a structured, semi-structured or mostly unstructured form. Due to the rise of the internet, 90% of the digital data that has been generated was in the last two years. There are three phases of big data [16], Phase 1 lasted from 1970-200. During this phase, researchers invented a lot of new algorithms which we currently use in a day-to-day world. A data science algorithm cannot work unless it can be fetched with a lot of data for training. Unfortunately, until the late 2000s, there wasn't much data available. The second phase lasted from 2000-2010. This was the phase when organizations started getting data due to the increase in usage of the internet. Techniques like social media analytics, sentiment analysis were carried out. A lot of new organizations like Netflix and Facebook were built on the concept of customer analytics. The third or present phase stared in 2010 and this was the turning point for the data science field. Big organizations have Peta-bites of data in their data warehouses waiting to be processed. Organizations can make a fortune if they can successfully mine the information from the available data.

There is terminology named 4V's in big data.

1. Volume
2. Variety
3. Velocity
4. Veracity

Volume defines the size of the data that an organization has with them. Internet and retail organizations like Google, YouTube, Amazon have hundreds of peta-bites of data with them. Variety is defined as the type of data available, it could be semi-structured, unstructured, etc. Velocity is the speed at which an organization receives the data. Usually internet companies Google, Facebook receive a lot of data at a rapid pace. Veracity is the trustworthiness of the data. Often the data is unstructured and unfit for analysis. Drawing conclusions from such data could lead to disastrous results, however, this can be solved by performing proper data analysis. Data Science can be further categorized into mainly two components, supervised and unsupervised learning methods. The usage of these methods depends upon the type of problem we want to solve.

Supervised learning is one of the most common learning techniques in the field of data. This approach is used when an algorithm learns the patterns from the data using a class label. Usually, a dataset contains variables of two categories, X and Y. The variable X describes the data, and the variable Y is the outcome of the X. For example, a fraud dataset may contain variables like User ID, Location, Amount, Frequency, etc which comes under variable X category, the class variable Y has values that could be in a binary form, 0 and 1.

Supervised learning techniques can be further divided into 2 categories, Classification, and Regression. Classification techniques are used when the class label Y has discrete values, for example, True/False, 0/1, etc. The Regression technique should be used when the class label contains continuous values, for example, Salary, Temperature, etc.

4.2 Data Science Lifecycle

There are various steps in a data science lifecycle. Each step has various methodologies. This lifecycle mostly follows a waterfall model [17] with a combination of spiral model [18] at certain stages. Since the lifecycle follows the waterfall method, it is very important to complete each step with optimum attention before proceeding to the next one. These steps could include business understanding, data collection, data preparation, exploratory data analysis and modelling. The modelling and evaluation steps follow the spiral approach since it requires a lot of tuning. Model training often faces the bias and variance issues hence it is mandatory to retrain the models over and over again.

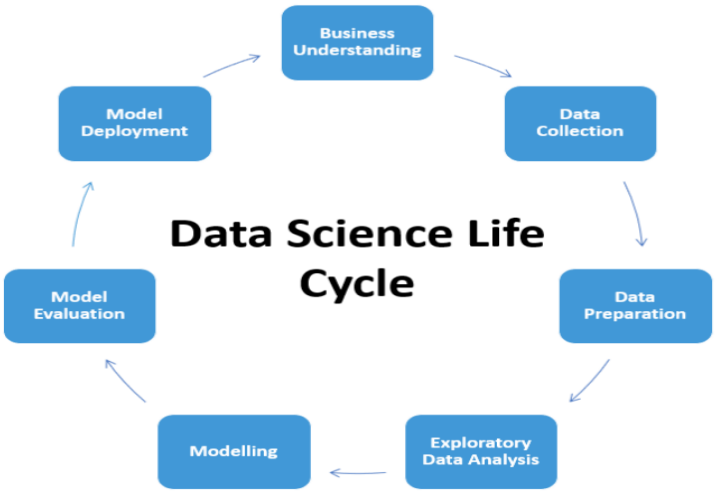

Figure 1: Data Science Lifecycle [19]

In this section, we will discuss every step in the entire data science lifecycle and relate it with the objective of this dissertation, which is building a data science system to detect fraud transactions.

4.2.1 Data Collection

The first step in building a data science project is data collection. This is perhaps the most important step as the entire project depends on what kind of data we have. Organizations collect the required data from different sources. This mainly depends on the nature of the organization's work. Product-based companies like Google, Facebook, YouTube, etc get their data directly from customers through their websites. Service-based organizations like Accenture, Capgemini, etc get the data from their clients. Asking for the relevant data for the project is a second important step. A data science algorithm cannot learn properly if any important data is missing. For example, retail companies like Amazon need information like Consumer Price Index to perform discount offer campaigns. It would be impossible for any organization to run without knowing about their customer spending habits, hence, getting the right data is very important.

Usually, data is stored in the data warehouses that could be in places like the cloud, physical hard drives, etc. The data is often stored in formats like the CSV (comma-separated value) files, JSON (Javascript object notation), etc. The data used in this project has been collected from Kaggle.com. Kaggle is one of the biggest repositories available on the internet for data science professionals. It contains numerous datasets in the form of projects where data science learners and professionals can compete by solving the problems. The dataset that has been used in this project has been taken from the challenge named "Credit Card Fraud Detection". This data consists of transaction information of the European credit card holders from the year 2013. This data is in CSV format and the size of this dataset is approximately 144 Mega Bites.

4.2.2 Data Description and Analysis

After the completion of the data collection process, the data analysis phase starts. To carry out any data analysis, there is a need for an appropriate platform and a programming language. A platform could be an editor or a tool. For example, data analysis or cleaning could be carried out using tools like Weka, Tableau, Power BI, Jupyter Notebook, RStudio, etc. A wide variety of programming languages like Python, R, SAS could be used to carry out the data analysis.

In this project, Jupyter Notebook and Python programming language have been used to carry out the data analysis and model training. A set of different python data analysis packages like Pandas, Numpy, Matplotlib, etc have been used for the analysis.

The dataset that has been used in this project consists of 284,807 transaction records and 30 features. Out of these 30 features, 28 have been feature transformed using PCA (Principal Component Analysis) to maintain customer privacy. Principal Component Analysis is a dimensionality reduction technique [20]. A machine-learning algorithm could get overwhelmed in the presence of a high number of features. Generally, financial data contains a lot of features that could cause problems while training

algorithms like SVMs which has a high training cost. PCA tries to reduce the number of features in the data by preserving most of the information and making the models more interpretable. This is done with the help of principal components. The user may select 'k' most important features, which is 28 in this dataset. PCA also transforms the name of the feature and replaces them with (V1, V2, …Vk).

Though the dataset contains approximately 285,000 data points, only 492 comes under the fraud category. The further analysis describes that non-fraud transactions comprise of almost 99.83% of the entire data and the rest are fraud cases, which is just 0.172%. This is an interesting observation since the data needs to be sampled in further steps before training.

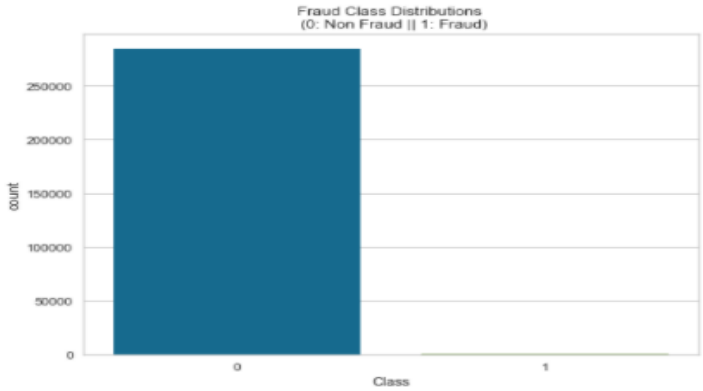

Figure 2: Class Distribution Plot

4.2.2.1 Missing Value Analysis

A missing value is a phenomenon when a dataset contains blank values up to various levels. These missing values are represented by the value 'NaN'. There could be a lot of reasons behind a value being missing. One of the most common reasons is the data collection techniques. Sometimes the data is collected through questionnaires and the customers tend to skip answering a few questions which are not mandatory and can be left blank. For example, a customer filling a supermarket's google form answers all the mandatory questions and skips certain optional questions like the favourite product category, location, pin code, etc. In such a scenario, the organization planning to perform customer analytics will have missing values and they cannot be avoided. Missing values can also occur because of bad data collection practices. In the industry, bad programming practices and a set of rules can cause the missing value issue in the database.

Understanding the nature of missing value is very important. There are two ways to understand this problem [22,23]:

1. Missing Completely at Random
2. Missing not at Random

First, missing completely at random could be described as a scenario where the missing data doesn't have any relation with other features in the dataset. The values that are missing are completely random. For example, a person may feel frustrated during filling a questionnaire because of personal reasons and leave it halfway through. Second, missing not at random could be described as a scenario where there is a relation between the missing value and the other features in the dataset. For example, a database contains five columns related to income over five decades. A person who was born in 2000 will have his/her salary section bank for the decade 1990-2000.

Missing value analysis has been carried out on the fraud dataset and it has been found out that the data doesn't contain any missing values. This is a big advantage since having missing values in the data can be problematic during the training phase. Unfortunately, there are not any algorithms present that could take data as input that contain missing value. These values need to be imputed which can be done using a variety of imputation techniques. Mean, Median, Mode, SMOTE, and Regression are the most used techniques [21]. If a feature consists of a lot of missing values, it can also be dropped.

4.2.2.2 Outlier Detection

Outliers can be defined as data that are completely different from the rest of the points. These points contain extreme variations. For example, in an employee dataset, the average salaries of people could be around £50000. However, people like Bill Gates, Jeff Bezos could have salaries in millions. In this case, Jeff Bezos and Bill Gates can be termed as outliers. There are a lot of methods to detect outliers [26], for example, plotting boxplots, finding Z-Score, Interquartile range, distribution plots, etc.

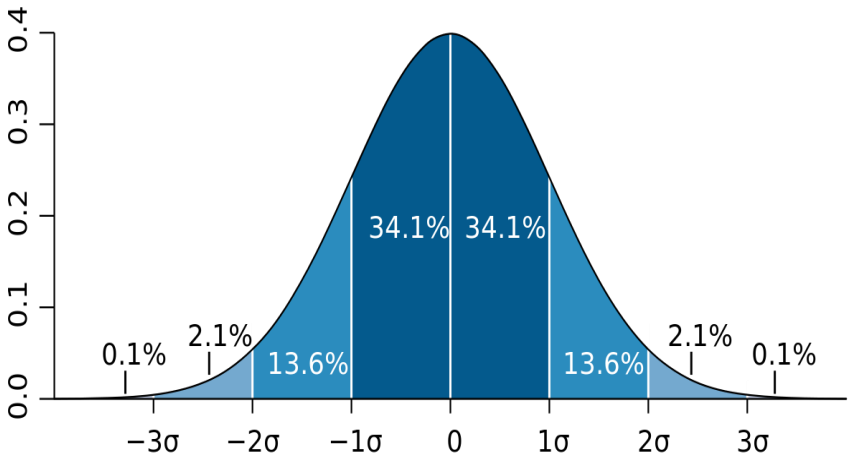

Figure 3: Normal Distribution [24]

From the gaussian distribution in Figure 3, it can be understood that any point that lies after the 2nd standard deviation on both positive and negative axis can be termed as an outlier. Outliers can lead to

disastrous results if left untreated before training a predictive model. Outliers can have a very adverse effect on Supervised learning algorithms like Linear Regression. A simple linear Regression tries to fit the perfect line in 2-dimensional space. In the presence of an outlier, the regression line could bend towards the outliers.

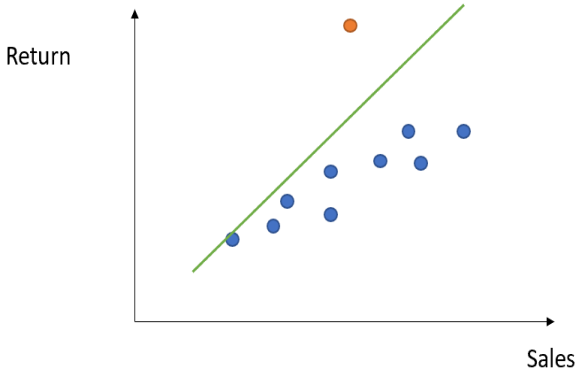

Figure 4: Linear Regression [47]

From Figure 4, it can be observed how the outlier disturbed the regression line from passing through the blue set of points to bending towards it. Hence it is very important to treat these outliers.

There are a lot of ways through which the outliers could be treated. One of the most used techniques is by performing a feature transformation. Feature transformation is a part of feature engineering [25] hacks where a feature is transformed using mathematical formulae. A log transformation works most of the time if the data doesn't contain any negative values since the log is undefined for negative values. Outliers are often removed from the dataset if they comprise a very small percent of the entire data and the training sample size is very big. However, this technique may not work in the case of small datasets since removing the outliers can remove a lot of unexplained variances.

In this project, IQR (Inter Quartile Range) technique has been used to detect the outliers from the dataset. IQR is a metric that captures 50% of the most similar data between a certain range. For example, in a list of 7 consecutive numbers starting from 1, IQR can be calculated using the following method:

Step1- Order the numbers: 1,2,3,4,5,6,7.

Step2- Find the median: 1,2,3,**4**,5,6,7.

Step3- Define Quarter1 and Quarer3: (1,**2**,3), **4**, (5,**6**,7).

Step4- Interquartile Range: Q3-Q4= 6-2 = 4.

A point can be termed as an outlier if its value is below the difference between the Q1 and 150% value of the interquartile range or the value is above the sum of the Q3 and 150% value of the interquartile range.

Outlier = Q1 - 1.5 * IQR

Outlier = Q3 + 1.5 * IQR

In the fraud dataset used in this dissertation project, all the 30 variables in the dataset contained outliers. Outlier treatment has not been performed since a lot of data points are outliers. In this project, a lot of sampling techniques were used before training the models, a few of those techniques handled the outlier issue.

4.2.3 Exploratory Data Analysis

Exploratory data analysis can be defined as a practice that involves a thorough investigation of data to find critical information which could be extremely useful in the later stages of the projects. An organization can learn a lot about its customers in the EDA phase. Every problem does not have the need of a statistical model to solve the objective, a smile exploratory data analysis can yield fruitful results.

EDA can be divided into two categories, descriptive statistics, and inferential statistics. Descriptive statistics use numbers to explains the data. The focus is to describe the data which is already known. For example, mean, median, mode, standard deviation, etc are great examples of Descriptive Statistics. On the other hand, Inferential statistics describe the sample, and it tries to generalize it to the entire population. Hypothesis testing, Confidence Intervals, ANOVA, etc are a few examples.

The data has been PCA transformed, and we do not know the names of the features. Hence, it wouldn't be wise to perform EDA on the fraud dataset as we may get limited information from it.

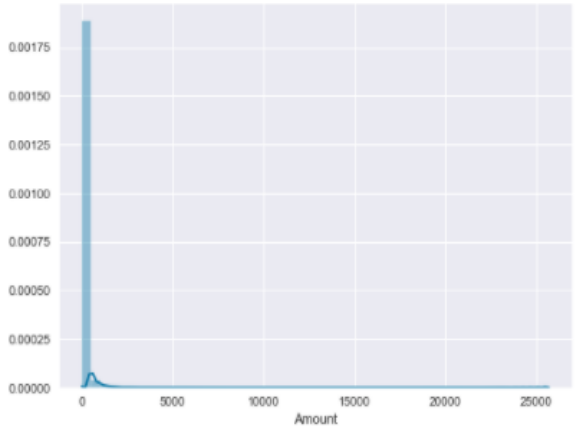

Figure 5: Amount Distribution

From Figure 5 it can be observed that the Amount variable is extremely positively skewed. The majority of the Amount lies around £0-200. There are a few transactions that are as high as £25691.

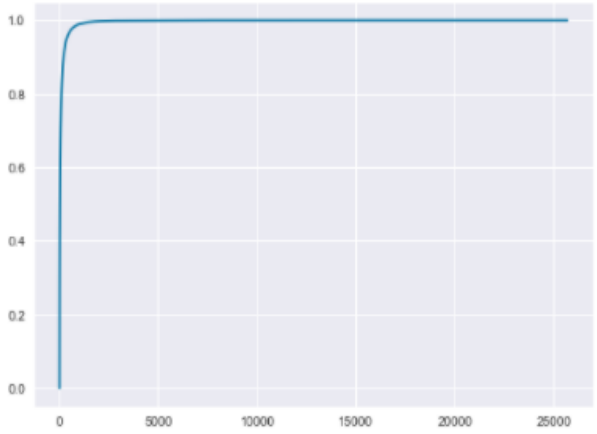

Figure 6: Cumulative Distributive Function

From Fig 6, it can be observed that almost 80% of the data contains the 'Amount' which is less than £200.

-

5. Methods

The analysis phase has been completed and the relevant information has been noted. In this section, most of the techniques that are needed to be done before training a model will be carried out. All the tools and techniques will be explained in detail. This section will include discussions about Sampling Techniques, Feature Selection, Model Training, etc.

Before proceeding to any further steps, it is very important to divide the data into train and test sections. It is a very important step since all the analysis should be carried out on the training set and the test set should be left untouched. The reason behind leaving the test set out of the further analysis is to avoid overfitting of the machine learning models. The test set can be considered as the future data that is unknown, and the aim should be to build a system that can work well on the unknown data.

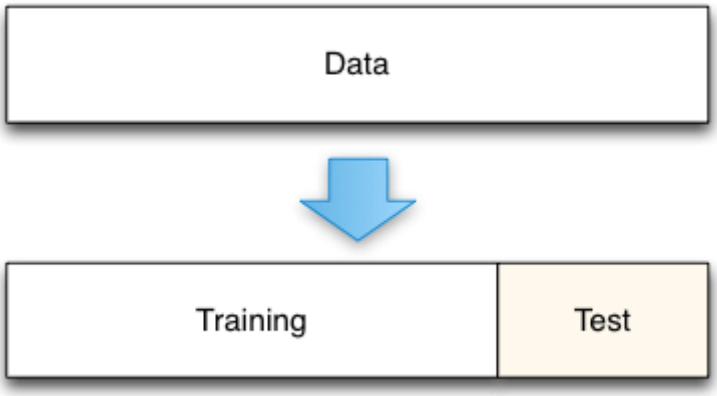

Figure 7: Train Test Split [48]

In this project, the data has been split into 2 parts, training, and testing. The train set contains 70% of the transaction records and the test set contains the remaining 30% of the records of the entire dataset. After this split, the train set contains 345 fraud instances, and the remaining are non-fraud transactions.

5.1 Sampling Methods

Machine learning algorithms do not produce accurate results when the dataset contains imbalanced data classes. Imbalanced classes can be defined as a situation where the data points from one class are more than that of the other class with a great margin. There is no rule of thumb to decide if a data is imbalanced, however, a dataset with an imbalance degree of 1:10 can be called an imbalanced dataset. Generally, most machine learning algorithms are designed in such a way that they only work when the datasets contain data of an equal number of classes. It is very important to check the distribution of the classes before training a classification model to satisfy the assumption of class equality. Algorithms like Decision Tree are biased towards the data from the majority class [28], it considers points from the minority class as outliers, and they are ignored. There is always a high chance of misclassification in

such models since they are trained on the majority class and classifying a minority class could be difficult. This is because the model doesn't train well on the minority class.

Just because the dataset is imbalanced does not mean that a good classifier cannot be built. There are a few factors that affect the complexity of the classifier identifying the rare events. The first factor is having a small sample size. A machine learning model is useless if the data needed for training is small. In a small data sample, a classifier finds it difficult to find the patters not just in the minority class but also in the majority class [30, 31]. The second factor is class separability, the more separable the data, the better the model can perform. Even in the cases where the imbalance ratio is extremely high, but the class is easily linearly separable, a good classifier can be built [32].

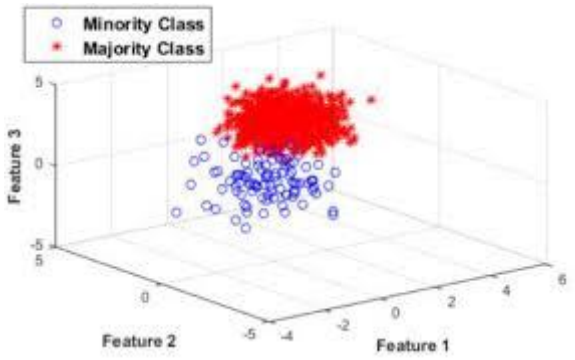

Figure 8: Linearly Separable Classes [32]

From figure 8, it can be observed that a simple clustering algorithm can classify the minority class with ease since the data is perfectly linearly separable. The third factor is known as a within-class imbalance. This is a problem where a class, mainly minority class, is distributed among a lot of sub-clusters and those clusters may not contain the same number of examples. Such a problem increases the complexity of the data, and the model finds it difficult to define the rules of separation.

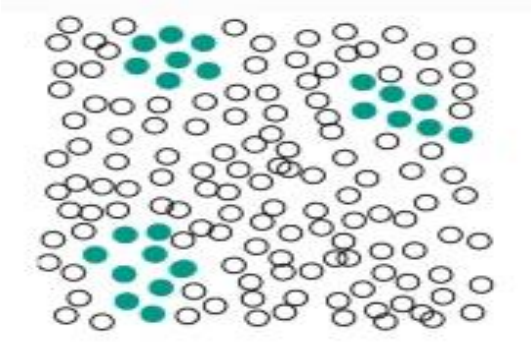

Figure 9: Sub-Clusters of Minority Class [49]

From Figure 9, it can be observed that a dataset has 2 classes where the minority class has 3 clusters. In such a case, the complexity of the minority increases hence difficult to classify correctly.

To deal with all the imbalance problems, data sampling techniques can be very effective. Data Sampling can be defined as a technique where part of a data is selected to perform an action. There are a lot of sampling techniques [29], for example:

1. Under Sampling
2. Over Sampling
3. Hybrid Sampling, etc.

In this project, it has been observed that the fraud transaction comprises only 0.17% of the entire dataset. This is one of the best examples of the class imbalance issue. This issue is not just observed in the fraud domain but also in applications like network intrusion detection, medical diagnosis, etc. Churn Rate detection is one of the best examples. With such an extremely skewed class distribution, algorithms like Decision Tree, Random Forest, and Logistic Regression will not work. Hence, sampling techniques will be used to create data samples so that the algorithms can be trained on the sampled data. It is mandatory to create the samples using the training set which contains 70% of the entire data.

5.1.1 Data Under Sampling

Under-sampling methods refers to the process of reducing the data from the majority class until a specific criterion is met. Most of the time, the criterion is to make the imbalance degree to 1. For example, if a dataset contains 1000 data points where the imbalance degree is 1:10, the under-sampled dataset will contain 200 points if the imbalance ratio is set to 1.

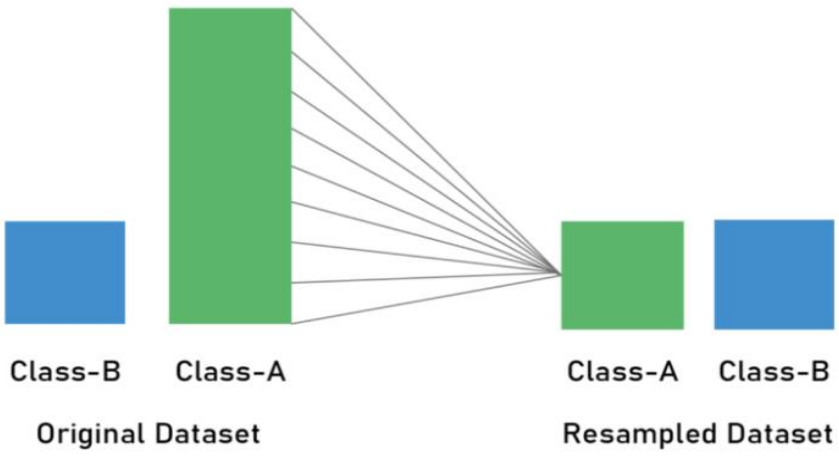

Figure 10: Under Sampling Method [33]

There are a lot of under sampling techniques and those techniques come under two groups, fixed under sampling and cleaning groups. The fixed under sampling method tries to equal the data points from both classes. A few techniques under this category are Random under sampling, NearMiss under sampling, Instance Hardness under sampling, etc. Cleaning under sampling are the method where the

sample points are removed from the majority class based on a certain criterion that varies among the different algorithms.

In this dissertation, a total of five under sampling methods with their detailed working proof are discussed.

5.1.1.1 Random Under Sampling

The first under sampling technique is Random under sampling which comes under the data-level approach. In this technique, data from the majority class is removed on a completely random basis until a certain balancing ratio is achieved. Choosing the appropriate balancing ratio depends upon the use case or domain. All the points from the minority class are preserved. This technique does not assume anything about the data during majority class elimination hence it is known as a Naïve technique. This technique is suitable to use when the data from the minority class is big enough to train a statistical model. If the data has a good mix of both classes, the models can focus on them equally. However, it has a few limitations too. By removing the data points from the majority class, a lot of crucial data is usually lost which could have a lot of variance/ patterns that help the data science models during the training process. Imbalanced-learn is the python package that has been used to carry out the sampling operations.

In the credit card dataset, a random sampling technique has been applied on the training set which contains 70% of the entire transactions. Since the training set contains 345 fraud instances, the size of the dataset becomes 690 after the training data is fitted with the random under sample method. The fitted data has an equal ratio of fraud and non-fraud instances since the balancing ratio is 1. A random seed of 0 is used to generate reproducible results. The sampling strategy is set to 'auto' so that the Imbalanced-learn can evaluate the target and determine the class with fewer observations as the minority and under sample from the majority class. The parameter 'replacement' is set to False to avoid the under sampled data points being added with replacement. Since the fraud instances are very less, it would be better to have unique data points from the majority class instead of having them with replacement.

5.1.1.2 Condensed Nearest Neighbours

The second under sampling technique is CNN [35]. This technique focuses on retaining points that are located at the class boundary. In the first step of the CNN technique, all the data from the minority class is selected and a single data point named 'P' from a majority class is selected at random and added into a group named 'S'. The second step comprises of training a KNN [34] model using sample S. The third step involves testing a new data point 'X', which is from the majority class, on the KNN model trained on the sample 'S'. If the model correctly predicts the class of the point X, point P is removed from the

sample S and a new point from the majority class is added. However, if the model misclassifies the X's class, the point P is permanently added to the sample S and a new point from the majority class is added and the same process is continued till no more misclassification is observed. By the end of this process, a subset of data points will be available with different classes closely located to each other. The aim is to classify all the points from the majority class using a KNN model that has been trained on the data containing all the minority classes and a few points from the majority class that are located close to the boundary. This method will eliminate all the points from the majority class that are located far away from the boundary. A different predictive model can be used to make predictions by training on this CNN sample so that it can yield robust results by looking at the patterns that contain hard instances. However, this technique may fail at certain times when the classes completely overlap with each other. In such scenarios, the majority of the algorithms cannot work. Hence, the effectiveness of this under sampling technique depends entirely upon the type of the dataset.

In the credit card dataset, CNN sampling is done with the sampling strategy set to auto so that only the majority class observations are under sampled, and the number of neighbours is set to the value of 1. In each iteration, only 1 nearest neighbour from the majority class has been tested. A sample size of 2167 transaction instances has been produced.

5.1.1.3 Tomek Links

Tomek links can be described as a pair of data points from different classes that are very close to each other and located at the boundary. In this technique, the Tomek links are considered as noise points and

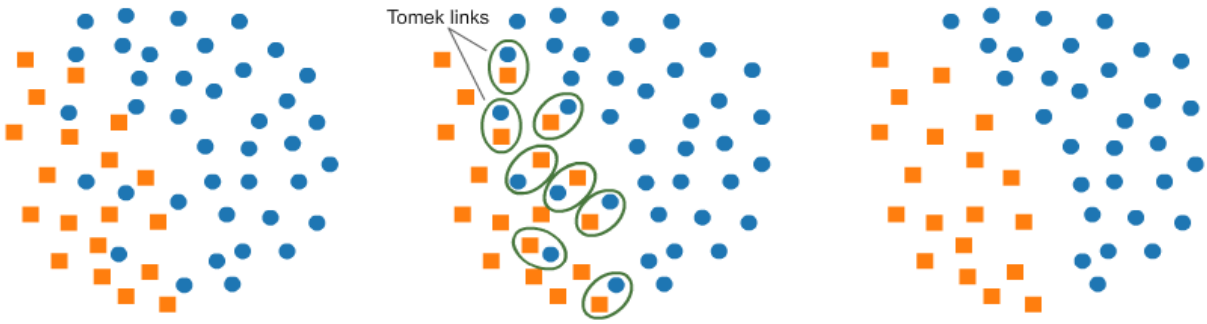

Figure 11: Tomek Links [50]

they are removed from the dataset so that the classes will have a good separation. This technique comes under the cleaning procedure.

There are certain limitations in removing Tomek links. If these links are removed, the information about the data that lies close to the boundary is lost. Hence it is better to preserve the minority points and eliminate the points from the majority class. From Figure 11, it can be observed that only the blue points from the majority class have been eliminated and the points from the minority class have been

preserved. In this way, good class separability is achieved. Additionally, the Tomek link sample consists of a lot of noise data from the majority class. This technique is more suitable in situations where only a few points lie at the boundary.

In the credit card fraud dataset, the Tomek link technique is applied where the sampling method is set to 'auto' instead of 'all'. Setting the sampling method to 'auto' removed the data that belonged to the majority class and the minority class points have been preserved which is very important since the size of fraud transactions is already very less.

5.1.1.4 One Sided Selection

One sided selection technique [36] will solve the limitations of the resampling techniques like CNN and Tomek Links. CNN doesn't work if the classes overlap with each other and the Tomek links technique removes a lot of data points that are located at the boundary which are considered as useful data.

One sided selection follows the same strategy that is followed by the CNN technique, which is to preserve all the points from the minority class and the data points from the majority class that are close to the boundary. This technique follows one additional step where it removes the Tomek Links from the sample created by the CNN technique. This technique is a combination of both CNN and Tomek link sampling methods. The outcome of this technique will have a sample where all the data from the minority class and the data from the majority class that is closely located to the boundary, and these points will also be linearly separable, unlike the CNN technique's outcome. The size of the sample completely depends upon the distribution of the data.

In the credit card fraud dataset, the one-sided technique has been applied where the sampling strategy has been set to 'auto' so that only the data points from the majority class are under sampled keeping all the data from the minority class. The data points in the Tomek links from the minority class are preserved. The sample outcome of the data consists of 199015 observations out of 199364 observations from the training dataset. It can be inferred that 349 points that have been removed are either noise or the points from the majority class that belongs to Tomek links.

5.1.1.5 Edited Nearest Neighbour

Edited nearest neighbours [37] is a sampling technique where the data from the majority class that are close to the border, near the opposite class, are removed to improve the separation of the classes. In this technique, a KNN model is trained on all the data as part of the first step. The second step involves finding K-nearest neighbours for the data points from the majority class and typically the value of K is

3. In the third step, if all the neighbours are found to be from the same class, the data point is retained else removed. This results in having a final dataset that has good class separability.

In the credit card fraud dataset, the data has been resampled using the edited nearest neighbour technique where the sampling is done from the majority class and the data is trained on 3-nearest neighbours. The data from the majority is removed where all its neighbours do not belong to the same class. Out of 199364 data points from the training set, 199044 data points have been retained. A total of 320 data points has been eliminated since they lie on the border near the minority classes.

5.1.2 Data Over Sampling

The second sampling method that has been used in this dissertation is Over Sampling. In this method, the data from the minority class from an imbalanced data set is increased until the desired balancing ratio is achieved. The aim is to create synthetic points so that the count of the minority class observations come close to that of the majority class. For example, oversampling a dataset containing 1000 data points with a balancing ratio of 1:10 will produce a dataset with 1800 data points if the balancing ratio for the sampling is set to 1. By doing so, the oversampled data can be used to train a machine leaning algorithm which has the data equality assumption. There are two oversampling practices, sample extraction and sample generation.

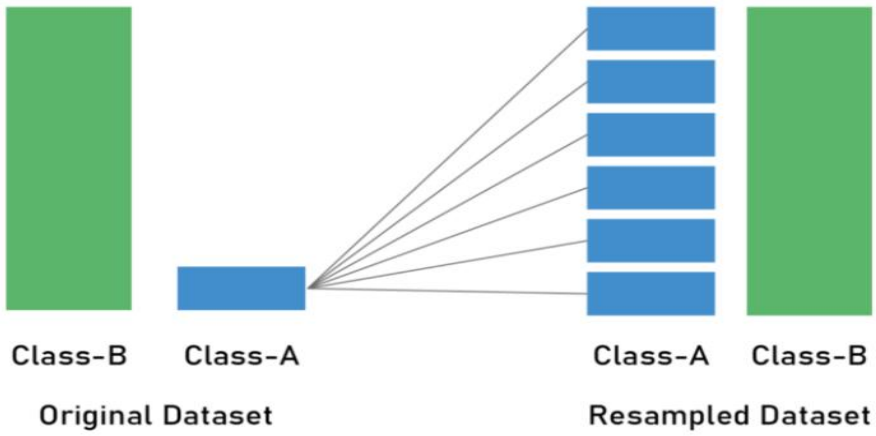

Figure 12: Over-Sampling Method [33]

In sample extraction, data from the minority class is randomly over sampled until the desired balancing ratio is achieved. Random oversampling is the only technique that comes under this category.

In sample generation, data from the minority class is over sampled using synthetic points. These points are created using the minority class data points; however, the synthetic points are not the same as the

original data, but they follow a certain similarity with the original points to a good extent. SMOTE and ADASYN technique comes under the sample generation category.

In this dissertation, two over sampling techniques have been applied.

5.1.2.1 Random Over Sampling

Random over sampling is a technique where the data from the minority class is duplicated until the desired balancing ratio is achieved. This technique has a few serious limitations. Since the data is duplicated, a machine learning model can easily overfit. To avoid such a situation from happening, the balancing ratio can be reduced.

In this dissertation, a random sampling technique has been applied to the training dataset. The data has been over sampled from the minority class with a balancing ratio of 1. A sample containing 3,98,038 data points have been created because of the random over sampling technique where equal number of fraud and non-fraud instances are present.

5.1.2.2 SMOTE

SMOTE [38] stands for synthetic minority over sampling technique. It is one of the most used techniques in the real world. In this technique, data from the minority class is over sampled using artificial or synthetic data that is created using instances from the minority class itself. The artificial points are extremely similar to the minority class data but not exactly the same.

Figure 13: SMOTE Method [33]

This technique avoids the creation of duplicate data hence the chances of overfitting a predictive model remains low if compared to that of the random over sampling technique. In the Fig 13, the red points are the synthetic data that have been created with the help of the minority class, green points.

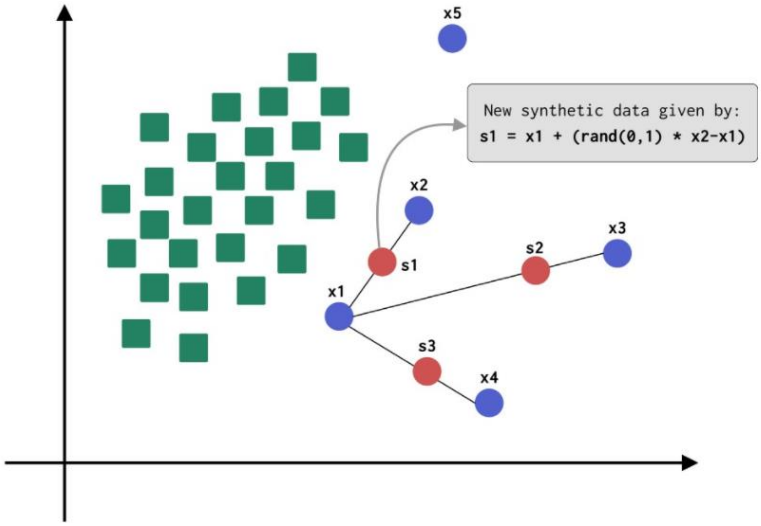

Figure 14: Synthetic Points Creation [39]

There are a total of 5 steps that need to be followed to create synthetic data points.

1. The first step in creating synthetic points involves choosing all the data points from the minority class.
2. The second step involves training 5-nearest neighbours on the minority class data.
3. Determined the balancing ratio.
4. Select data points from the minority class and its 5 neighbours.
5. Create synthetic points between the selected data points and the distance between its neighbours.

$$S1= x1 + rand (0,1) * (x2-x1)$$

This technique has a few limitations. If the data points from the minority class are very few and have a few outliers, the SMOTE technique can create a lot of synthetic data on those outliers. This could lead to creating a very bad data sample and a predictive model would end up training on the noise data.

In the credit card fraud dataset, SMOTE technique has been applied on the trainset and the sampling strategy has been set to 'auto'. Hence, only the data points from the minority class have been resampled using the 5-KNN method. A total of 198,674 synthetic data points has been created in the minority class.

5.1.3 Hybrid Data Sampling

Sampling techniques like under and over sampling have a few limitations. A lot of under sampling techniques either remove the data from the boundary or remove the data away from the boundary. This leads to a loss of valuable information in both cases. However, this method has the advantage that it removes the outliers which can be also referred as noise data. Over sampling techniques work well if the data from the minority class is good enough to create synthetic points. However, if the minority sample is less and it contains outliers, techniques like SMOTE can create artificial points based on noise.

Hybrid data sampling is a technique that overcomes the limitations of both the data over and under sampling methods. It combines a set of both methods and creates a data sample that choose the best data instances.

In this project, we have taken a combination of SMOTE with Tomek links and Edited Nearest Neighbours which could overcome the issues of the individual sampling techniques and could sample the best data.

5.3 Predictive Models

Predictive models [40] can be defined as an intelligent system that has the capability to predict future events based on the events that happened in the past. These models are a combination of statistical method which enables them to make smart predictions. There are a lot of predictive models which have different use cases. The most common models are Linear Regression, which comes under the regression category. Logistic Regression, Naïve Bayes, KNN, SVM, Decision Tree, Random Forest, Artificial Neural Network, etc are a few examples that come under the classification category. In this report, only 3 algorithms have been used which are Logistic Regression, Decision Tree and Random Forest.

The training dataset has been sampled using a lot of sampling techniques. In this section, the predictive model building will be discussed in detail.

5.3.1 Model Training

The first step that comes in the model building process is model training. A data science model needs to learn from the past data so that it could identify the patterns and trends from them and predict future events.

This section will focus on describing the predictive models used for classifying credit card transactions. A detailed analysis of all the steps used in this process will be discussed which includes model description, tuning the algorithms, feature selection/importance, etc.

5.3.1.1 Decision Tree

A decision tree is a supervised learning algorithm that can solve both classification and regression problems. It is a tree-based learner that makes decisions based on certain conditions by creating axis parallel boundaries. There are 2 types of decisions trees. If the dataset contains categorical variables, a categorical decision tree is built and if a dataset contains continuous values, a continuous decision tree is built. A continuous-valued dataset needs to be converted into a discrete dataset since the decision tree assumes that all the features are categorical. The decision tree is one of the few statistical models which can be trained even with very limited data size.

A decision tree's structure comprises a set of nodes and branches where the root node is considered the most important node. The rest of the nodes are created as a result of the splitting of the root node. From Figure 15, the structure of the decision tree could be understood.

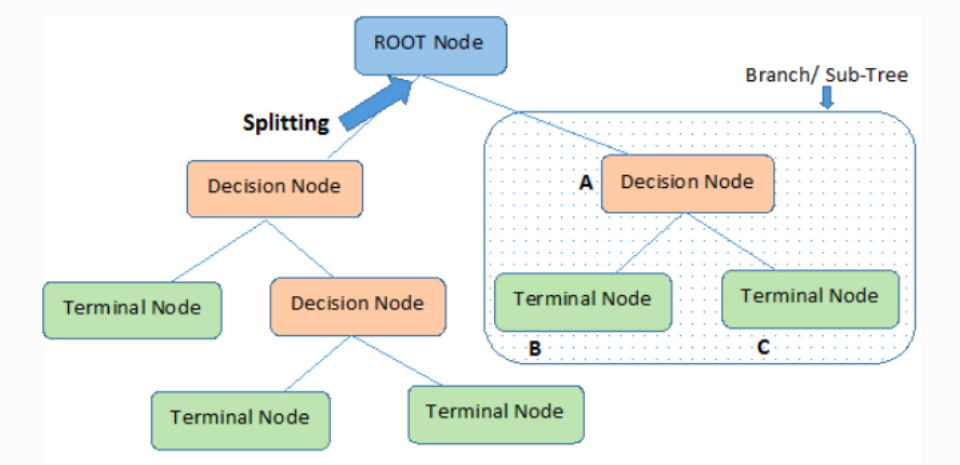

Figure 15: Decision Tree Structure [42]

The blue node is the root node from which the tree is derived. The green nodes are the terminal nodes, also known as leaf nodes. The decisions are made from the outcome of leaf nodes.

The first step in building a decision tree is by calculating the Entropy [41] of the dataset. Entropy is the measure of randomness [42] in the data. Entropy ranges from 0 to 1 and a dataset that has a perfect class balance has an entropy value of 0.5.

$$Entropy = -\Sigma P(x) \log P(x)$$

The next step is calculating the Information Gain. Information gain is the amount of variance a feature can explain in the data about the target. The feature having the highest Information Gain becomes the

splitting node. The most important feature in the dataset has the highest Information Gain and it becomes the root node. Decision trees are mostly used because of their interpretability, they are very easy to understand. It also explains a lot about the important features of the dataset which could be very crucial for an organization to carry out business decisions. Decision trees are often used for feature selection. Theoretically, a higher number of features describes a lot of information about the data, however, it is practically impossible to train such data since it takes a lot of computational power in the processing stage. Organizations mostly use cloud-based computation to solve machine learning problems that run on a credit-based system. More credits are consumed in the case of higher computation hence using all the features at any stage in the data science lifecycle is not a cost-efficient approach.

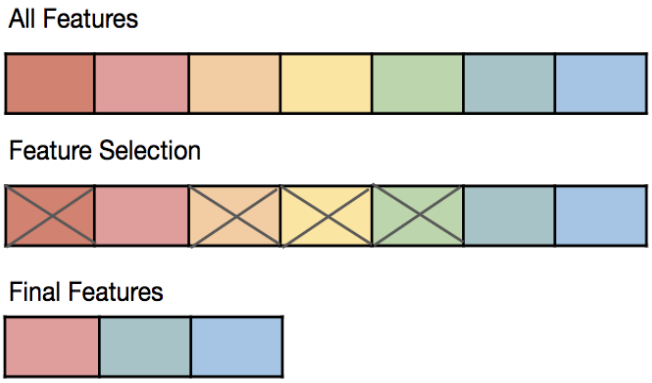

Figure 16: Feature Selection [51]

From the Fig 17, it can be observed that V14 is the most important feature of the credit card dataset.

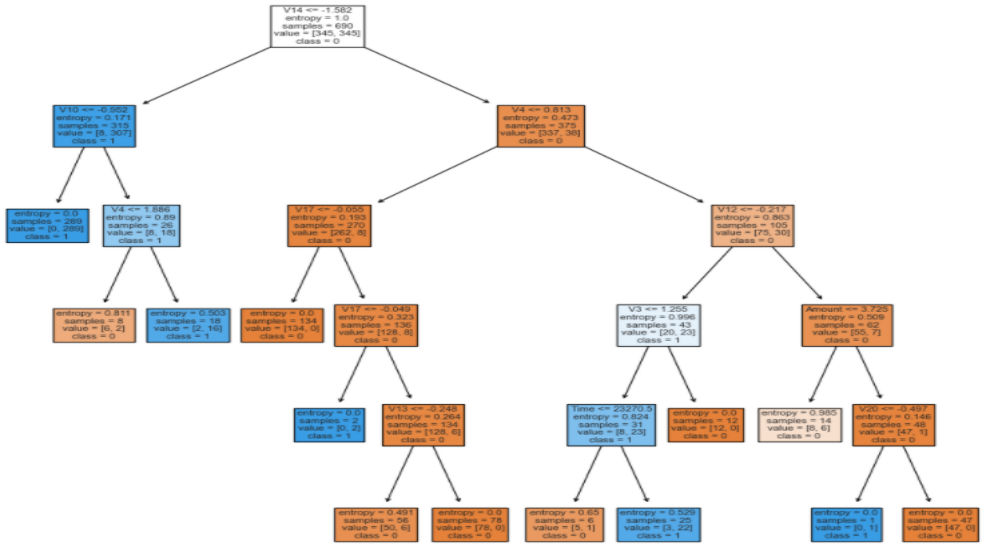

Figure 17: Feature Selection using Decision Tree

Pruning a decision tree is very important since shallow decision trees usually overfit. Leaf nodes or terminal nodes usually have very few data points in a decision tree that is not pruned. If those points

contain outliers, the model would be making decisions based on those outliers. Hence, it is mandatory to either remove the outliers before fitting the model or prune accordingly.

Random Search cross-validation [43] is one of the techniques that helps machine learning algorithms avoid overfitting. It enables the models to learn the data using different parameters. Finally, the most optimum parameters are used to train the learners. The parameters that are used in this dissertation project are:

1. Maximum Depth- the depth of the decision tree: [1,2,3,4,5,6,7,8,9]
2. Minimum Sample Split- the minimum number of samples at a node after which splitting should be stopped: [10,13,15,18,20,21,22,23,24,25,28,31,35,50]
3. Criterion: [Gini, Entropy]

Random search lets the algorithm perform cross-validation on the data during the training phase. K-fold cross-validation [44] is one of the famous techniques that can be used to perform this operation. A k-fold cross-validation can be defined as an optimization technique that helps machine learning models to train on k-folds of training data which tells the goodness of a model. In this process, first, the training data is divided into k-folds, then the model is trained on (k-1) chunks and tested on the last chunk. Each fold is trained and tested iteratively and the error on the test set is averaged. The average error on the test dataset describes the performance of the trained models.

Figure 18: K-fold Cross Validation [52]

From Figure 18, it could be understood that the data is divided into 10-folds and each fold has been divided into 10 chunks. In each iteration, the model will train on 9-chunks and test on only one chunk. The term E in the diagram is the average test error of the algorithm that has been obtained by averaging the test error at each stage.

5.3.1.2 Random Forest

Random Forest is an ensemble technique where multiple decision trees are built to solve a problem. This algorithm can solve both classification as well as regression problems. An ensemble [45] can be described as a technique where multiple machine learning algorithms like Logistic Regression, SVM, Naïve Bayes, KNN, etc are trained to make a prediction. The outcome of the ensemble algorithms is mostly made on a majority vote basis. An ensemble learning has 2 methods, bagging, and boosting. Bagging or bootstrap aggregation can be described as a technique where multiple week learners are trained, and the outcome is achieved by majority voting. Boosting is a technique where the algorithms are trained on the errors made by the previous learners, hence robust models are created.

In the random forest technique, the training data is initially sampled into k-subsets with replacement, and 'k' decision trees are trained on them. These samples are called bootstrap samples since they are created with replacement. In a classification dataset, the outcome of the random forest is decided based on the outcome of the decision trees, which is the majority count.

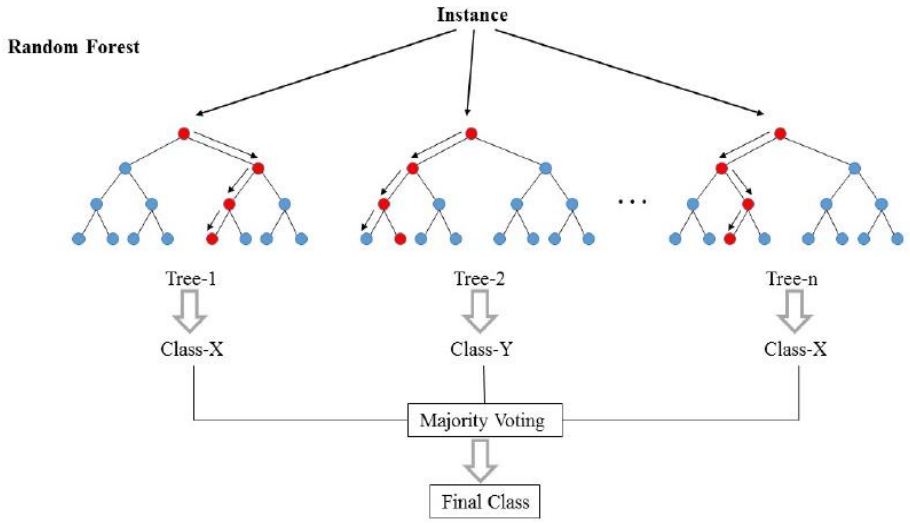

Figure 19: Random Forest [53]

From Fig 19, it can be observed that the majority of the decision trees have predicted X as the outcome, hence the output of Random Forest is X.

Since the random forest models are built on a lot of decision trees, they are less subjected to noise and predictions at random. A general decision tree may produce good/bad results based on the data available in the training set. If the training sets contain a sample of data that does not have any hard instance, a decision tree may build a model which produces a great training accuracy but perform very badly on the test set. Hence, bootstrap aggregation is used to solve this problem. The chance of randomness in the training data reduces during training the decision tree models.

There are a few limitations to this technique. Random forests produce great predictions; however, they are computationally expensive to build and are less interpretable. However, this issue can be solved if the model training is done using appropriate parameters using a random search cross validation technique. The parameters that have been used in this project are as follows:

1. N-estimators- number of decision trees to be built: [5,10,15,20,25,30,35,40,45,50,55,60]
2. Maximum Depth- the depth of the decision tree: [1,2,3,4,5,6,7,8,9]
3. Minimum Sample Split- the minimum number of samples at a node after which splitting should be stopped: [10,13,15,18,20,21,22,23,24,25,28,31,35,50]
4. Criterion: [Gini, Entropy]

5.3.1.3 Logistic Regression

Logistic regression is a classification algorithm which follows a probabilistic approach where the algorithm assigns a probability score to each observation instead of giving a binary outcome. A logistic regression tries to build a hyper-plane that divided the two classes in the dataset. A plane (π) can be represented by,

$$\omega^T x + b = 0$$

The ω is the normal and b is the intercept. The intercept would be zero if the plane (π) passes through the origin. The logistic regression tries to solve the following optimization problem where it tries to

$$\omega^* = \arg max[\ \sum_{i=1}^{N} y_i \omega^T x_i]$$

maximise the margin so that misclassification rate remains less. However, the above logistic regression equation is prone to outliers. To make this equation outlier proof, a sigmoid function can be added to the data point X using the squashing technique.

$$\sigma(x) = \frac{1}{1 + e^x}$$

The logistic algorithm produces a probability score for each data point that ranges from 0 to1. Scikit-learn frame has a 0.5 value as the threshold for classifying the data points.

Standardizing the data before training the logistic regression is very important. By standardizing the data, the scaling effect can be avoided. For example, the same temperature in Celsius and Fahrenheit

will be interpreted differently by the logistic regression model since the Fahrenheit scale has a higher value than that of Celsius.

$$Standardised\ value(x') = \frac{Original\ value(x) - Sample\ Mean(u)}{Sample\ Standard\ Deviation(\sigma)}$$

Data standardizing is not mandatory for algorithms like Decision tree and Random Forest.

A regularization term like L1 and L2 can be applied to the logistic equation to avoid the model from overfitting. The hyperparameter used in this project are as follows:

1. C: [50,10,1,0.1,0.05,0.01,0.005,0.001,0.0005,0.0001]
2. Penalty-the regularization terms: [L1, L2]

Advantages of using a logistic regression include the interpretability of the model. Though there are a few algorithms that perform better than logistic regression, for example, boosting algorithms, a lot of financial institutes still use logit models because they are easy to interpret unlike boosting algorithms.

5.3.2 Model Testing and Evaluation

Model testing is one of the most important phases in the entire data science lifecycle. A trained model needs to be thoroughly tested before making any conclusions about the data. The trade-off between bias and variance can be evaluated only after testing the trained model. Data training and testing is an iterative approach, a model needs to be retrained when it has high variance. High variance can be described as a scenario when the trained model performs well on the training set but poorly on the test set. A model having a low bias and high variance is useless since it cannot accurately predict future data.

Determining the performance on the test set can be done by using a few evaluation metrics like Accuracy, Precision, Recall, F1 Score, etc. A confusion matrix provides a good overview of the model's performance.

A confusion matrix is an evaluation technique that explains the model's performance on the test dataset. In a binary classification task, a confusion matrix is a 2*2 matrix that provides the relation between the actual data and the predicted data in a numerical form. There are 4 important evaluation metrics that can be calculated using the confusion matrix which are TP (True Positives), FP (False Positives), TN (True Negatives), FN (False Negatives). Metrics like Accuracy, Precision, Recall and F1 score can be calculated using this confusion matrix.

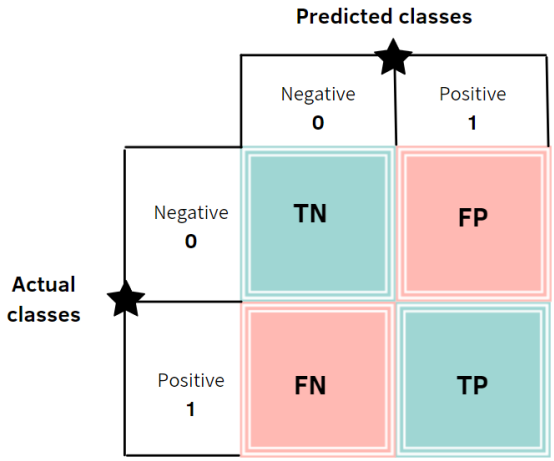

Figure 20: Confusion Matrix [46]

A model predicts the class of the test dataset, and the test dataset contains actual class values. True positives are the number of correctly classified positive points, True negatives are the number of points that are correctly classified as negatives. False positives are the number of observations that are classified as positive but are negative, and False Negatives are the observations that are positive but predicted as negative. False-positive can also be called a Type 1 error and false-negative can be called a Type 2 error.

Accuracy is an evaluation metric that is most used. This is the ratio of correctly classified instances and the total instances.

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

Accuracy is not the best metric to evaluate a model's performance if the data is imbalanced. For example, if the model predicts all the data points from a test set as positives where the balancing ratio of the test set is 1:20, the accuracy would still be 95%.

Precision can be described as the ratio of true positive observation and the model's predicted points that are positive. Precision can be used as an evaluation metric in scenarios where a system's scope in making false-positive decisions should be less. For example, a nuclear launch detection system should never misclassify an event like a nuclear attack that is not happening.

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

Recall is true positivity rate, also known as, the ratio of true positive observations and the actual positive observations. Recall should be used in fraud detection systems since the system's wrongly classifying a fraud transaction as genuine could be very dangerous.

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

The F1 score is the harmonic mean of Precision and Recall. F1 score has a range from 0 to 1. More the F1 score, better the model's performance.

$$F1\ Score = 2 * \frac{Precision + Recall}{Precisio * Recall}$$

In this dissertation, all the evaluation metrics discussed above are used to test the trained model's performance. Since the fraud dataset is extremely imbalanced data, accuracy has not been used as the evaluation metric during scoring the models. Hyperparameter tuning has been carried out using Random Search CV. During this process, a five-fold cross validation technique has been applied and the models have been scored using 'Recall' as the evaluation metric on the test dataset. Hence all the three trained models, which are Decision Tree, Random Forest, and Logistic Regression, have the capability to handle minority class is very good.

The reason behind choosing 'Recall' as the evaluation metric is because the aim of the project is to correctly identify the fraud instances. The credit card business cannot run if a fraudulent transaction is misclassified as non-fraud. However, it is acceptable for a machine learning algorithm to flag a genuine transaction as a fraud up to a certain extent since those transactions can be immediately investigated by the fraud analysts and cleared off. Hence, having a high recall is desired in a credit card fraud scenario even at a cost of having a low precision.

6. Result Analysis

6.1 Decision Tree

6.1.1 Without Sampling

Initially the credit card data was divided into a 70-30 split, and they were named as train and test dataset. The train dataset contains 199019 genuine and 345 fraud instances. A decision tree was trained on this data using a set of various optimization parameters. Same hyperparameter values have been used for all the decision tree models that have been built in this project. This is because, a comparative study can be done only if same metrics are used in all the experiments.

The best parameters obtained for the decision tree are as follows:

1. Maximum Depth: 3
2. Minimum Sample Split: 28
3. Criterion: Entropy

The decision tree model that was trained on these optimal parameters was tested on the test set which comprises of 30% of the entire fraud dataset which was not used in any part of the analysis. The test set contains 146 fraud instances.

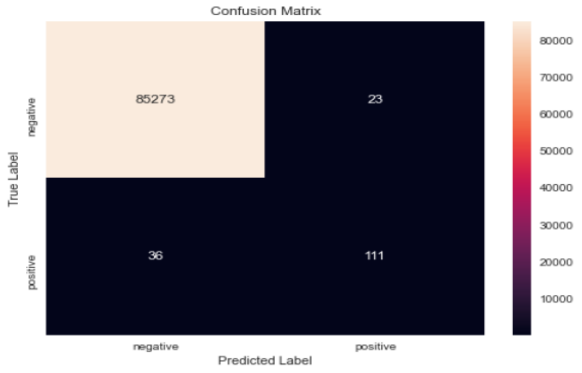

Figure 21: Confusion Matrix- Decision Tree

From the confusion matrix in the Figure 21, metrics like accuracy, recall, precision and f1 score can be calculated. It can be observed that the accuracy is extremely high, however, the recall score is low. The evaluation scores for this model are as follows:

1. Accuracy: 99.931%
2. Precision: 0.828
3. Recall: 0.755
4. F1 Score:0.79

The performance of the decision tree in this scenario is fairly good. It has a good balance of the precision, recall and F1 scores. Accuracy is extremely high, but it should be neglected since the data is imbalanced.

6.1.2 Random Under Sampling

The data points from the majority class of the train data have been under sampled to achieve a balancing ratio of 1. Hence the size of the random under sampled data is 690 that comprises of equal fraud and non-fraud data points. This sample has been used for training the decision tree model.

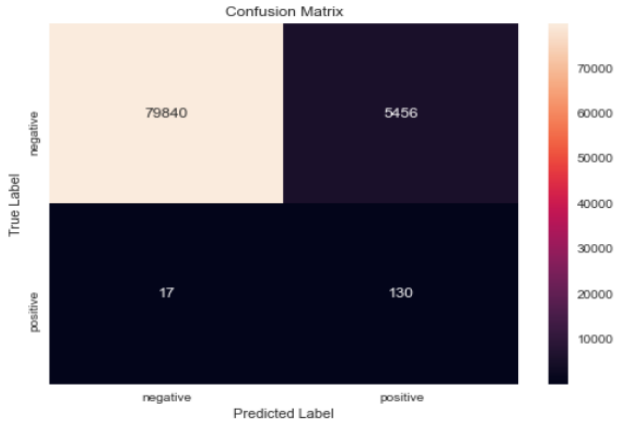

Figure 22: Random Under Sample- Decision Tree

From the Fig 22, it can be observed that the decision tree that has been trained on the random under sampled data has a more focus on reducing the risk of wrongly classifying the fraud instances as genuine transaction. This can be verified by looking at the Recall score which is 0.884. However, the precision is extremely low, 0.023. The model is classifying a lot of non-fraud transactions as fraud. Because of this issue, the F1 score is 0.045, which also very low.

This model may not be feasible to use in reality because of its disproportionate precision and recall values.

6.1.3 Condensed Nearest Neighbours Sampling

The condensed nearest neighbour sampling technique extracted a total of 4334 transaction instances from the training data and this sample contains equal number of classes. From the confusion matrix, it can be inferred that the decision tree model has a good Recall score of 0.85, however it doesn't have a good precision, which is 0.077. It classifies almost all the transactions as fraud instances.

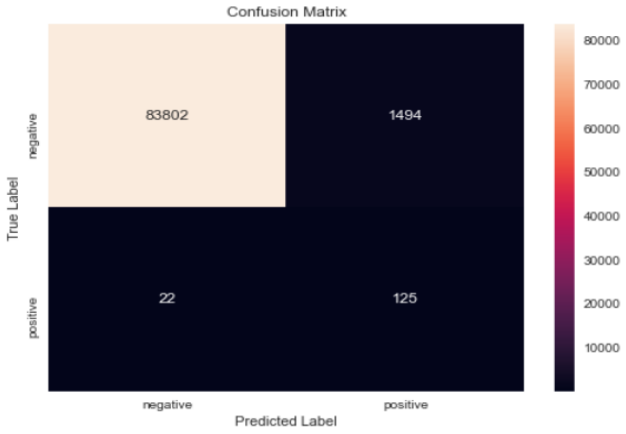

Figure 23: CNN- Decision Tree

6.1.4 Tomek Links Under Sampling

Tomek links from the majority class of the train dataset has been removed and the final sample consisting of 199304 instances has been created. The decision tree trained on these instances has produced very good results. Though the recall score is 0.76, which is not bad, a high precision and f1 score is achieved having values 0.855 and 0.806 respectively.

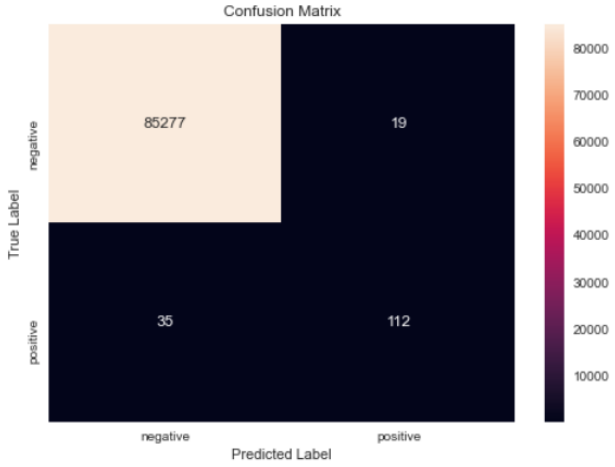

Figure 24: Tomek Link- Decision Tree

6.1.5 One Sided Selection Under Sampling

The one-sided selection has produced same results which were produced by the Tomek link under sampling technique. However, 199015 training instances were used by the decision tree and the training time is also less if compared to that of Tomek link sample.

6.1.6 Edited Nearest Neighbours Under Sample

Edited nearest neighbour sampling technique sampled a total of 199044 transaction records from the train dataset. The decision tree trained on the samples created by Tomek links, one-sided selection and edited nearest neighbour has produced same results. However, the size of data used for training and the training time were different.

6.1.7 Random Over Sampling

The random oversampling technique sampled a total of 398038 instances from the train dataset. The decision tree that has been trained on this data produced poor results. Though the model has a good recall score, however, it has a terrible precision too. It is classifying almost a lot of genuine data points as fraud transactions.

Figure 25: Random Over Sampling- Decision Tree

6.1.8 SMOTE Over Sampling

Decision tree trained on SMOTE sample did not produce great results. A recall value of 0.871 was achieved and a very low precision was noted.

6.1.9 Hybrid Sampling- SMOTE-ENN & SMOTE-Tomek

Decision tree built on the samples created by both the hybrid techniques did not produce good results. Usually, hybrid sampling techniques are known to sample the best instances from the dataset which could be easily trained. However, decision tree did not provide the best results. The model had a high recall score and a very low precision.

6.2 Random Forest

6.2.1 Without Sampling

The optimal parameter used to train the random forest were kept same during multiple training phases using different samples. The parameters are as follows:

1. N-estimators: 15
2. Minimum sample split: 22
3. Maximum depth: 9
4. Criterion: Entropy

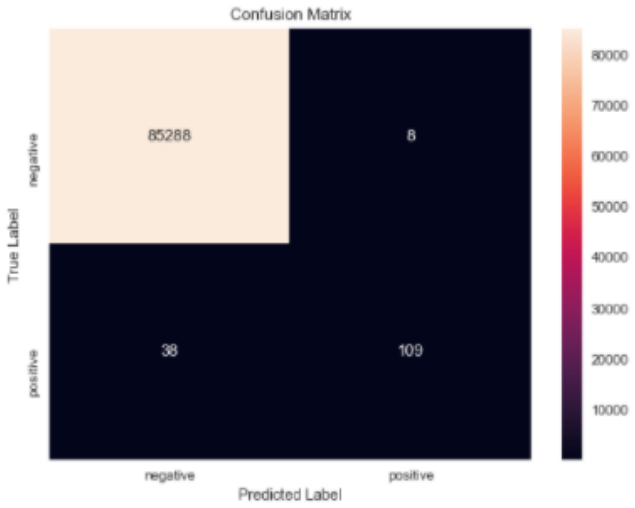

Figure 26: Random Forest- Without Sampling

It can be observed from Figure 26 that the Random Forest classifier has performed very well on the training data. The precision is 0.93 which explains how good the model is correctly classifying the non-fraud instances. The recall score, which is 0.741, is not as good if compared to the precision but still acceptable, nonetheless.

6.2.2 Random Under Sampling

The random forest model did not perform well on the random under sampled data. It produced a good recall score but a very bad precision and F1 score.

6.2.3 Condensed Nearest Neighbour Sampling

The random forest model produced decent results after training on the sample created by CNN method.

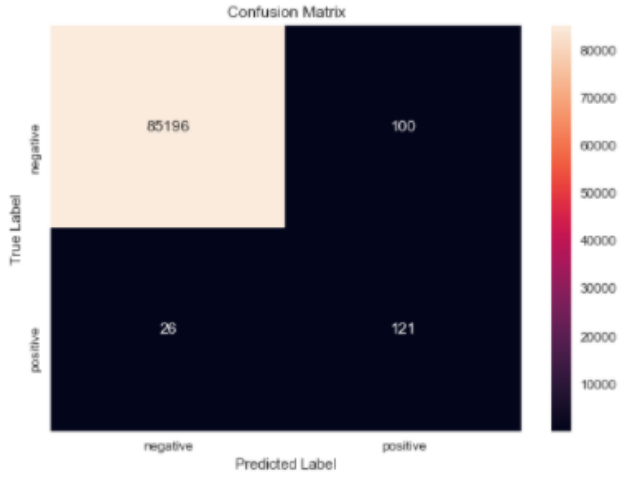

Figure 27: Random Forest on CNN Sampling

From the Figure 27 it could be inferred that a good recall and a decent precision score has been produced.

6.3 Logistic Regression

6.3.1 Without Sampling

The optimal parameter used to train the logistic regression were kept same during multiple training phases using different samples. The parameters are as follows:

1. C: 0.05
2. Penalty: L2

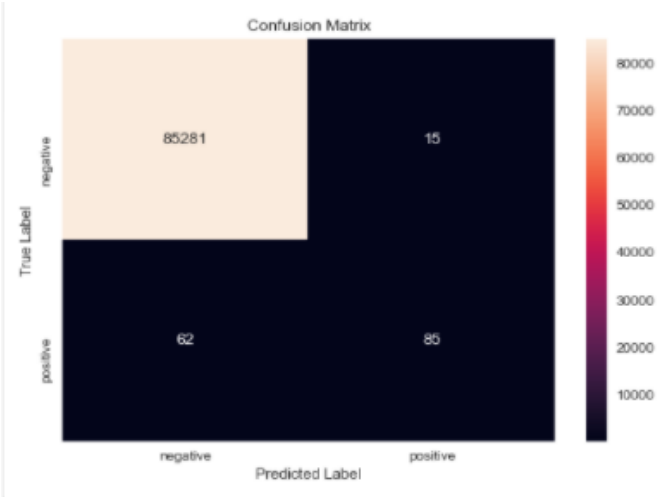

Figure 28: Logistic Regression - Without Sampling

From the Figure 28, it could be inferred that the Logistic Regression trained using the training dataset has produced bad results. A Recall score of 0.578 has been achieved, however, a good Precision score has been produced.

6.3.2 Random Under Sampling

The Logistic Regression model trained on the data created using Random under sampling approach has produced bad results. A very high Recall score has been noted, however, a very bad Precision score was achieved.

Figure 29: Logistic Regression- Random Under Sampling

6.3.3 Tomek Links Under Sampling

From Fig 30, it can be observed that the Logistic Regression trained on the Tomek Link sample has

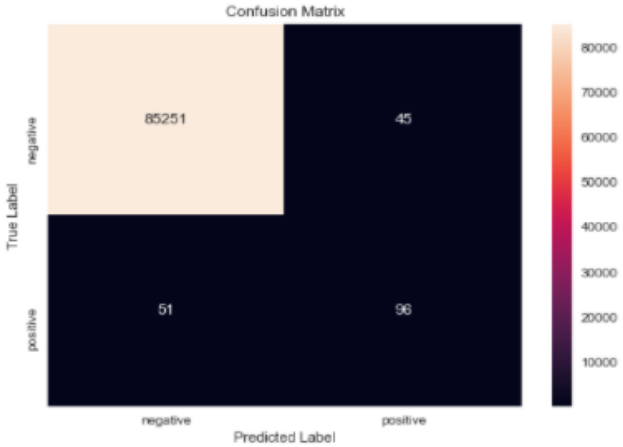

Figure 30: Logistic Regression -Tomek Link

provided decent results. A good balance between precision, recall and F1 score can be observe. Generally, a model providing consistent results in across all the metrics are desired.

6.4 Result Summary

This section will give an overview about the performance of the machine learning models which were trained using different types of data samples. A detailed overview of the model's performance can be seen in the Table 1.

Algorithms	Sampling Techniques	Accuracy	Precision	Recall	F1 Score
Decision Tree	Without Sampling	99.93%	0.828	0.755	0.79
	Random Under Sampling	93.59%	0.023	0.884	0.045
	CNN	98.23%	0.077	0.85	0.142
	TOMEK Links	99.94%	0.855	0.762	0.806
	One Sided Selection	99.94%	0.855	0.762	0.806
	Edited Nearest Neighbours	99.94%	0.855	0.762	0.806
	Random Over Sampling	93.31%	0.022	0.884	0.043
	SMOTE	97.58%	0.059	0.871	0.11
	SMOTE-ENN	96.65%	0.041	0.823	0.078
	SMOTE-Tomek	96.57%	0.04	0.823	0.076
Random Forest	Without Sampling	99.95%	0.932	0.741	0.826
	Random Under Sampling	96.18%	0.038	0.884	0.074
	CNN	99.85%	0.548	0.823	0.658
	TOMEK Links	99.95%	0.964	0.721	0.825
	One Sided Selection	99.94%	0.946	0.714	0.814
	Edited Nearest Neighbours	99.94%	0.929	0.714	0.808
	Random Over Sampling	99.92%	0.76	0.796	0.777
	SMOTE	99.85%	0.543	0.823	0.65
	SMOTE-ENN	99.71%	0.351	0.83	0.493
	SMOTE-Tomek	99.83%	0.502	0.83	0.626
Logistic Regression	Without Sampling	99.91%	0.85	0.578	0.688
	Random Under Sampling	93.74%	0.025	0.932	0.049
	CNN	98.94%	0.121	0.823	0.212
	TOMEK Links	99.89%	0.681	0.653	0.667
	One Sided Selection	99.90%	0.695	0.728	0.711
	Edited Nearest Neighbours	99.89%	0.676	0.667	0.671
	Random Over Sampling	96.45%	0.043	0.921	0.081
	SMOTE	98.43%	0.086	0.85	0.157
	SMOTE-ENN	97.51%	0.058	0.891	0.11
	SMOTE-Tomek	97.53%	0.059	0.891	0.111

Table 1: Algorithm's Performance Summary

From Table 1, it can be observed that all the algorithms have an extremely high Accuracy score, especially Random Forest. Accuracy has not been considered to make interpretation about the predictive models in this report. A good accuracy is meaningless since the dataset is highly imbalanced. It can be observed that all the models had a very good Recall scores in most of the occasions. This is because of the parameters that have been used to train these models. These parameters were obtained as a result of performing Random Search cross-validation on the models where Recall was considered as the scoring metric. Hence, all the models performed good in correctly identifying the fraud transactions.

Generally, a machine learning model is desired that can produce consistent Recall and Precision scores. However, in the scenario of fraud, it is acceptable to have a model that has a great Recall with decent Precision. The decision tree has produced good results when trained with data that has been sampled using Tomek Links, One Sided Selection and Edited Nearest Neighbours resampling strategies. Random Forest produced great results when trained with data that was resampled using Condensed Nearest Neighbours technique. The model misclassified only 26 fraudulent transactions which led to having a Recall score of 0.823. Moreover, it flagged only 100 genuine transactions as fraud which is very reasonable. Generally, such models are desired in the credit card industries, where flagging genuine transactions is accepted up to a certain extent.

It can be observed that Logistic Regression is the only model that had the most consistent results when trained with data that was resampled using Tomek Links, One Sided Selection and Edited Nearest Neighbours techniques. The Precision, Recall and F1 scores are not very high, however, the scores are extremely similar. Such models can not only classify the fraud transactions but also reduce flagging lesser genuine transaction as fraud. Random Forest had better Recall values at a big cost of Precision.

7. Conclusion

The objective of this dissertation was to solve the credit card fraud detection problem using predictive models. Various analyses were done at each stage of the data science lifecycle. The data was initially explored using the exploratory data analysis and relevant information was noted. Various checks like outlier detection, etc were carried out.

Training a predictive model in a scenario where the data is imbalanced is not always a straightforward process. The data must be sampled using various sampling techniques so that the machine learning algorithm can find the best patterns. In this dissertation project, a total of 10 data sampling techniques has been used to sample the data for training the model. Every sampling technique had a different approach, few techniques tried to retain the data, which was present near the border, also known as hyperplane, and others tried to remove the data that was present near the hyperplane. The sampled data was used to train 3 different machine learning models.

From the Table 1, it can be understood that all the three algorithms, Decision Tree, Random Forest, and Logistic Regression, performed better when they were trained only on the training dataset which was not obtained using any specialized sampling techniques. Random Forest outperformed the Logistic Regression and Decision in terms of Precision and Recall scores.

Making a comparative analysis on the performance of all the 3 machine learning algorithms could be a tricky task in this scenario. This is because all the 3 algorithms were trained on different data sampled each time.

If the Recall scores are analysed, the Random Forest Classifier performed better than Logistic Regression on almost each instance when the models were trained using the data that was produced using under sampling techniques. However, the Logistic Regression had a better recall score than that of Random Forest when it was trained on the data that was produced using Over Sampling and Hybrid techniques. The problem in this scenario is that all these samples contain different data points.

Comparing the models that were trained on the same dataset could be meaningful. In such case, it can be said that the Decision Tree is the model that performed better than the other two models when trained on samples created by Tomek Links, Edited Nearest Neighbours and One-Sided Selection.

8. References

[1] History of Banking, Retrieved from Wikipedia: https://en.wikipedia.org/wiki/History_of_banking

[2] Kurzgesagt- In a Nutshell, "Banking Explained-Money and Credit", Online video clip. YouTube, 12 Mar, 2015.Web. 25 July 2021.

[3] Amanj Mohamed Ahmed, Consumer Behaviour toward the Use of Credit Cards: The Empirical Evidence from Iraq, Shirkah Journal of Economics and Business, 5(1) pp: 53-69, 2020.

[4] The Money Statistics, (August 2021). Retrieved from The Money Charity Website: https://themoneycharity.org.uk/money-statistics/

[5] Lexix Nexix Risk Solution, "Fraud- THE Facts", The definitive overview of Payment Industry Fraud, https://www.ukfinance.org.uk/system/files/Fraud%20The%20Facts%202021-%20FINAL.pdf

[6] Brean Horne. (25 Jul 2021), 4 examples of credit card fraud. Retrieved from My Wallet Hero website: https://www.fool.co.uk/mywallethero/credit-cards/learn/4-examples-of-credit-card-fraud/

[7] Joe Resendiz. (28 Jan 2021), How Credit Card Companies Make and Earn Money, Retrieved from Value Penguin Website: https://www.valuepenguin.com/how-do-credit-card-companies-make-money

[8] Rule Based Fraud Detection, Retrieved from Fraud.net Website: https://fraud.net/d/rules-based-fraud-detection/

[9] Igor Mektreovic, Mladan Karan, Damin Pintar, Credit Card Fraud Detection in Card-Not-Present Transactions: Where to Invest, MDPI.com, Appl Sci, *11*(15), 2021

[10] Kevin J. Leonard, The development of a rule-based expert system model for a fraud alert in consumer credit., European Journal of Operational Research, Volume:80, PP:350-356, 1995.

[11] Jakub Nalepa, Michael Kawaulok, Selecting training sets for support vector machine: a review, International Science and Engineering Journal, 52, PP:857-900, 2019.

[12] Theodoros Evgeniuo, Massimiliano Pontil, Support Vector Machines: Theory and Applications, Machine Learning and Its Applications, Advance Lectures, 2001.

[13] K. R. Seeja and Masoumeh Zareapoor, FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining, The Scientific World Journal, 2014.

[14] V. Dheepa and R. Dhanapal, Behavior Based Credit Card Fraud Detection Using Support Vector Machines, Ictact Journal on Soft Computing, ISSN: 2229-6956, VOLUME: 02, ISSUE: 04, JULY 2012.

[15] M. Sathyapriya, Dr. V. Thiagarasu, A Cluster Based Approach for Credit Card Fraud Detection System using Hmm with the Implementation of Big Data Technology, International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 14, Number 2 (2019) pp. 393-396.

[16] A short history of Big Data, Retrieved from Enterprise Big Data Framework Website: https://www.bigdataframework.org/short-history-of-big-data/

[17] Adetokunbo Adenowo, Basirat A Adenowo, Software Engineering Methodologies: A review of the Waterfall Model and Object- Oriented Approach, International Journal of Scientific and Engineering Research, 4(7), pp:427-434, 2020.

[18] Jiujiu Yu, Research Process on Software Development Model, IOP Conference Series: Materials Science and Engineering, Volume 394, Issue 3, 2018.

[19] Priya. (2 Oct 2020), Complete Lifecycle of a Data Science/Machine Learning Project, Retrieved from Start it up Website: https://medium.com/swlh/complete-life-cycle-of-a-data-science-machine-learning-project-13df81bbd8eb

[20] Sidharth Mishra, Uttam Sarkar, Subhash Taraphder, Principal Component Analysis, International Journal of Livestock, DOI:10.5455/ijlr.20170415115235, 2017.

[21] Hyun Kang, The prevention and handling of the missing data, Korean Journal of Anaesthesiology, doi: 10.4097/kjae.2013.64.5.402, 64(5): 402–406, 2013

[22] Liang Jin, Yingtao Bi, Chengi Hu, A comparative study of evaluating missing value imputation methods in lable-free proteomics, Scientific Report, Volume 11, 2021

[23] Runmin Wei, Jingye Wang, Mingming Su, Missing Value Imputation Approach for Mass Spectrometry-based Metavolomics Data, Scientific Report, Volume 8, 2018.

[24] Retrieved from Wikipedia Website: https://en.wikipedia.org/wiki/Standard_deviation

[25] Fatemeh Nargesian, Horst Samulowitz, Learning Feature Engineering for Classification, Conference: Twenty-Six International Joint Conference on Artificial Intelligence, DOI:10.24963/ijcai.2017/352, 2017.

[26] Pawel D. Domanski, Study on Statistical Outlier Detection and Labelling, International Journal of Automation and Computing, Volume 17, pp: 788-811, 2020.

[27] Naveen Venkat, The Curse of Dimensionality: Inside Out, DOI:10.13140/RG.2.2.29631.36006, 2018.

[28] Guest Blog. (17 March 2017), Imbalanced DATA: How to handle Imbalanced Classification Problems, Retrieved from Analytics Vidhya Website: https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

[29] Batrosz Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence, 5, pp:221-232, 2016.

[30] KP Suresh, S Chandrashekhara, Sample size estimation and power analysis for clinical research studies, Journal of Human Reproductive Sciences, 5(1), pp:7-13, 2012.

[31] Andrew Wong, Mohammed S. Kamel, Classification of Imbalanced Data: a review, International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 2011.

[32] Pasapitch Chujai, Kittipong Chomboon, Kedkarn Chaiyakhan, Kittisak Kerdprasop, Nittaya Kerdprasop, A Cluster Based Classification of Imbalanced Data with Overlapping Regions Between Classes, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2017 Vol I, IMECS 2017, March 15 - 17, 2017.

[33] Younes Charafaoui. (6 Dec 2019), Resampling to Properly Handle Imbalanced DATASETS IN Machine Learning, Retrieved from Heartbeat Website: https://heartbeat.fritz.ai/resampling-to-properly-handle-imbalanced-datasets-in-machine-learning-64d82c16ceaa

[34] Zhongheng Zhand, Introduction to machine learning: k-nearest neighbours, Annals of Translation Medicine, 4(11), 2016.

[35] Chien-Hsing Chou, Bo-Han Kuo, Fu zhi Chang, The Generalized Condensed Nearest Neighbour Rule as a Data Reduction Method, 18th International Conference on Pattern Recognition, PP:20-24, 2006.

[36] Gustavo Enrique Batista, Andre de Carvalho, Maria-Carolina Monard, Applying Onse-Sided Selection to Unbalanced Datasets, Mexican International Conference on Artificial Intelligence, 11-14, 2000.

[37] Roberto Alejo, Jose Martinez, Rosa Maria, Edited Nearest Rule for Improving Neural Networks Classification, Conference: 7th International Symposium on Neural Networks, 2010.

[38] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, pp: 321–357, 2002.

[39] Fernando Lopez. (1 Mar 2021), SMOTE: Synthetic Data Augmentation for Tabular Data, Retrieved from Towards Data Science Website: https://towardsdatascience.com/smote-synthetic-data-augmentation-for-tabular-data-1ce28090debc

[40] Vaibhav Kumar, M.L, Predictive Analytics: A Review of Trends and Techniques,

International Journal of Computer Applications, 182(1), pp:31-37, 2018.

[41] Yan-yan Song, Decision tree methods: applications for classification and prediction, Shanghai Archives of Psychiatry, 27(2), pp:130-135, 2015.

[42] Nagesh Singh Chauhan. (Jan 2020), Decision Tree Algorithm, Retrieved from KD Nuggets Website: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

[43] James Bergstra, Yoshua Bengio, Random Search for Hyper-Parameter Optimization, Journal of Machine Learning, 13(2012), 281-305.

[44] Rory P.Bunker, Fadi Thabtah, A machine learning framework for sport result prediction, Applied Computing and Informatics, Volum 15, Issue 1, pp: 27-33, 2019.

[45] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang, an improved ensemble learning approach for the prediction of heart disease risk, Informatics in Medicine Unlocked, https://doi.org/10.1016/j.imu.2020.100402, Volume 20, 2020.

[46] Eugenia Anello. (23 Feb 2021), How to Evaluate you model using the Confusion Matrix, Retrieved from Towards AI Website: https://towardsai.net/p/data-science/how-to-evaluate-you-model-using-the-confusion-matrix

[47] Valentina Alto. (20 Aug 2019), Detecting outliers with PyOD, Retrieved from Towards Data Science Website: https://towardsdatascience.com/detecting-and-modeling-outliers-with-pyod-d40590a96488

[48] Brian Mwandau, Investigating Keystroke Dynamics as a Twi-Factor Biometric Security, Research Gate, 2018.

[49] Dariusz Brzezinski, Leandro L. Minku, The impact of data difficulty factors on a classification of imbalanced and concept drifting data streams, Knowledge and Information Systems, 63, 1429-1469, 2021.

[50] Rafael Alencar. (15 Nov 2017), Resampling strategies for imbalanced datasets, Retrieved from Kaggle website: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets.

[51] Dor Amir, Feature Selection: Beyond feature Importance, Retrieved from KD Nuggests Website: https://www.kdnuggets.com/2019/10/feature-selection-beyond-feature-importance.html

[52] Johar M. Ashfaque, Introduction to Support Vector Machine and Kernel Methods, Project: Ideas in Machine Learning, 2019.

[53] Random Forest, Retrieved from Wikipedia Website: https://en.wikipedia.org/wiki/Randomforest.