Computer Fraud and Security
ISSN (online): 1873-7056

Anomaly Detection

Bhanu Prakash Reddy Rella
The University of Memphis, Tennessee, USA

Abstract

Due to the significance it holds in the concept of fraud, security in computers and business, anomaly detection serves very much purpose. Using techniques in unsupervised machine learning, the two algorithms applied in this study are Isolation Forest and Autoencoder in credit card fraud detection in financial datasets. This work focuses on data preparation and selection, generation and extraction of the features, as well as model assessment through use of metrics such as ROC-AUC, measure of precision, and measure of recall. It also discovered that Autoencoder attends to complex patterns of anomalies while Isolation Forest cuts down the false positives. Some of these problems include class imbalance and computational issues are highlighted. Some of these strategies include hybridization for imbalance handling and real time implementation that is very helpful in the development of automated and large scale anomaly detection in financial related work.

Keywords: Anomaly Detection, Machine Learning, Isolation Forests, Autoencoders

Chapter 1: Introduction

1.1 Introduction

Anomaly detection is now used in many areas of applications such as finance, cybersecurity, health care, and industrial monitoring. It involves the process of analyzing data samples in order to find out the irregularities in the operations since these are usually an indication of fraud, breach in security or system failure. As more transactions take place across the world, these can no longer be handled manually, not if businesses want to achieve optimum accuracy, speed and security. Both Isolation Forest and Autoencoder are unsupervised machine learning-based algorithms that have been shown to have good capabilities to find anomalies without the use of labels.

The research presented in this paper outlines the use of unsupervised machine learning for the purpose of detecting fraud in financial data sets. The main idea is to use Isolation Forest model and Autoencoder neural network to improve the general fraud identification processes of transactional data. Therefore, based on the enhanced data preprocessing and feature engineering, the study seeks to increase the accuracy of anomaly detection models while making considerations related to real-world applicability.

1.2 Research Rationale

It is becoming even more important to identify anomalous behavior in the financial domain since many fraud cases and cyberattacks have surfaced lately. In the increasing and mutation of the fraud patterns, traditional rule-based methods of detection fail, and that is why machine learning comes as an encouraging methodology. The methodologies of unsupervised learning are beneficial in that they do not need in-depth knowledge of an area aside from features that determine its fraud schemes allowing the algorithms to detect them as they appear [7].

The study is informed by the need to come up with an efficient and automated system to secure financial transactions from fraud. Isolation forest and autoencoder models selected in this paper will help the worldwide development of multi-disciplinary fraud detecting methods and models.

1.3 Research Aim

The study revolves around designing and testing unsupervised machine learning algorithms that can be used in the field of anomaly detection with especial relevance to the fraudulent card transactions. The purpose of the study

is to improve the current knowledge and techniques of automating some of the processes involved in anomaly detection.

1.4 Research Objectives

- To evaluate the efficiency of Isolation Forests and Autoencoders in order to choose the most efficient algorithm for the further work.
- To clean the financial transaction dataset for the purpose of correctly identifying anomalies.
- To use machine learning algorithms for building and testing of fraud detection classifiers.
- To evaluate the performance of the benchmark models that have been established to be the best, in terms of accuracy measurement and visuals. [8].
- To identify some potential merits and validate the effectiveness of the proposed models in detecting fraud schemes in a real-world scenarios.

1.5 Research Questions

- What is the efficiency of the Isolation Forests and Autoencoders in detecting anomalies in financial transactions?
- What prerequisites in the input data preprocessing and transformation are needed for the improvement of anomaly detection?
- To what extent do machine learning models work as far as detecting fraudulent activities are concerned?
- What are the differences between the models in their accuracy, precision, recall and F1-score?

1.6 Background

One main reason that emerges from the increasing complexity of financial systems in the contemporary world is that it has become easy for people to perpetrate frauds, which are very risky for both business entities and consumers. Anomaly detection is one of the technique that is crucial in proving that abnormality may indicate fraud, system failure or acts of cyber criminals [10]. Thus, the traditional approaches based on identification of specific rules have been employed in the past and they are no longer efficient because of the changes in the frauds and therefore, the regular updates in the spot rules.

Machine learning is now more scalable and quite efficient as compare to other methods in respect of anomaly detection because it use statistical model to detect the difference in the pattern. Some of the types of unsupervised learning that I have included are Isolation Forests and Autoencoders; these are appropriate in the fraud detection systems since it is often difficult to obtain additional datasets that are labeled. Isolation Forests consist of using decision trees to isolate anomalies while Autoencoders which is a type of neural network reconstructs normal transaction patterns and detect significant variations.

To evaluate these models, Kaggle has an available dataset that is relevant to the real-world financial field called Fraud Detection Transactions. The key parameters that were collected are the amount, location, type of the device used during transaction, the mode of authentication, and the risk score which are valuable for identification of the anomaly [9]. This increases the chances of the success of a machine learning model by ensuring the dataset has undergone one or more pre-processing steps like feature scaling, encoding of categorical features and handling of missing data.

In order to assess the effectiveness of anomaly detection models, the following parameters include precision, recall, F1-score, and ROC-AUC, and the graphical representations like correlation heat map and confusion matrix. This paper ia on how to increase the efficiency of fraud detection based on the shortcoming of the conventional methods while adopting the modern machine learning methods. It will help financial safety solutions, the protection of computer networks, artificial intelligence, and other such systems used in industries where fraud detection plays a significant role.

Chapter 2: Literature Review

2.1 Introduction

Anomaly detection has been a well-established field with many real-life applications in several domains such as finance, security, medicine, and manufacturing. This is because the traditional rule-based approaches have multiple drawbacks mainly attributed to the fact that they cannot efficiently process large and dynamic datasets and also are not autonomous in their functioning, which is why machine learning models have been implemented for better and efficient detection of anomalous events.

This chapter also utilizes literature research on the unsupervised based anomaly detection methods which include Isolation forest and Autoencoders [12]. The systematic literature review focuses on conceptual frameworks, variables of interest for anomaly detection and methods used to approach and analyze it as a research topic. Also, it covers the role of various independent and dependent variables (IDVs & DV) in identifying fraudulent transactions most efficiently.

2.2 Conceptual Framework

Anomaly detection is the process of singling out records that are seen as strange in relation to the rest of the records. Regarding fraudulent detections, the anomalies can be defined as those actions that are quite different from the patterns normally observed when a user interacts with the financial transactions [11]. Anomaly detection models are patterns that are generalized using algorithm such as machine learning that seeks to learn the features that may be out of place in a larged data set in regards to fraudulent activity, system failure, or hacking attacks.

Various methods in machine learning approach to anomaly detection are supervised, semi-supervised and unsupervised learning models. Logistic regression as well as decision trees models need fraud data sets containing data with known fraud instances. However, in practice, labeled fraud data is scarce and imbalanced, which hinders the applying of the supervised method. Semi-supervised approaches train most of the transactions to recognize the exceptions, but still many of them are labeled partially.

Isolation Forest and Autoencoder are more suitable for fraud detection since they do not need training data. To elaborate further, Isolation Forests comprise of developing random decision trees, and the isolating of anomalies are occurred in fewer splits. Autoencoder is a kind of deep learning algorithm which reconstruct the normal data patterns and the outliers can be founded out from the high reconstruction error.

Feature engineering proves to be vital aspect since attributes, including transaction amount, origin, authentication type, and risk score, are used to identify possible fraud cases. Also, feature scaling, encoding of categorical variables, and missing data treatment strengthen the model for anomaly detection.

Using the unsupervised learning methods, the results of this research identify fraudulent activities in the analyzed financial transactions to improve the efficiency of such systems, which face various problems of the traditional comparable methods.

2.3 Independent and Dependent Variables

Different authors propose that an anomaly detection model can be influenced by several independent variables so that the model can effectively detect fraudulent transactions. The dependent variable (DV) is the last point of the detection process whereby one gets to determine whether the identified transaction was fraudulent or normal.

Independent Variables (IDVs):

Frequency of Transaction: The frequency of the transaction is also useful in identifying fraudulent transactions as those with a lower frequency are likely to be fraudulent.

Accounting Records: These include new accounts or previously hijacked accounts through which the fraudsters tend to perpetrate their high-value activities.

Transaction type: Bank transfers, card payments and online transactions are not at the same level of risk.

Accounting ánlications: Transitions of the account from a device type that is unrecognized or multiple accounts linked to a device.

Precomputed Risk Score: It is used to determine risk level is high or low of certain transaction.

Transaction Distance: The distance between previous and instant transactions in geographical terms needs to be looked at as an aspect of fraud.

Multiple failed transactions: such a scenario can suggest attempts of unauthorized access to an account.

Dependent Variable (DV):

Fraud Label: It is a dependent variable and will have a value 1 if the transaction is a fraudulent and 0 if it is a normal transaction.

This paper employs Isolation Forests and Autoencoders techniques on these variables with the aim of isolating the anomalies. In order to enhance the models the feature selection methods include normalization, encoding, and handling of outliers [13]. This paper aims at determining how various factors affect the achievement of its objectives based on the analysis of the relationship between IDVs and the DV that can help enhance the efficiency of anomaly detection models.

2.4 Empirical Study

According to the author Alla and Adari, 2019, the authors present an extensive learning from the basics of anomaly detection with detailed procedures implemented via python pro-gramming language and deep learning techniques. The authors first define anomaly detection and the relevance of the concept in various fields, especially in the financial field. In addition, they discuss various kinds of anomalies such as point anomalies, contextual anomalies, as well as collective anomalies so that even novices can understand the text. The book goes further and apply them, or at least give examples and recommendations on for what you might use Keras, TensorFlow or Scikit-learn for in deep learning for the goal of detecting anomalies in data.

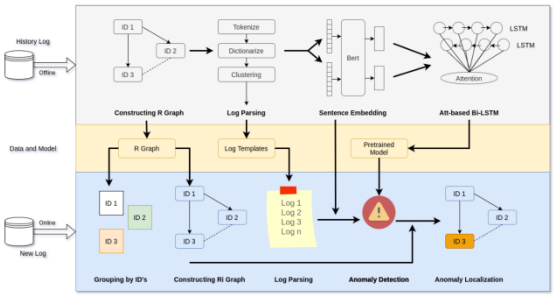

Figure 1: Anomaly detection and localization architecture diagram

(Source: Alla and Adari, 2019**)**

Autoencoders and recurrent neural networks (RNNs) are remarkable methodologies embedded in the principle of deep anomaly detection according to the authors. For this reason, autoencoders are introduced as an effective approach to data anomaly detection in high-dimensional financial data since the essence of the autoencoder is to learn the data representation and find outliers by measuring the reconstruction error. The book also covers the fine-tuning of the hyperparameters and some methods for model assessment when developing models for anomaly detection. Thus, this resource will be helpful for researchers and practitioners who might be interested in the practice of deep learning encompassing machine learning methods in anomaly detection, particularly, in the field of finance that demands reliable anomaly detection to address fraud and risk control issues as well as for market analysis.

According to the authors Wolpher. 2018, the master's thesis entitled 'Latent space anomaly detection using LSTM autoencoder' gives a detail description of anomaly detection employing LSTM autoencoders on unstructured time series data. The paper focuses on the problem of outlier identification in the context of usage of big data, which is not explored as a problem adequately in the prior works. For this purpose, the thesis focuses on

applying LSTM autoencoders in temperature anomaly detection to understand the suitability of recurrent neuronal networks with autoencoder systems in sequential anomaly detection. Experimental evaluation reveals that both Isolation Forest and Replicator Neural Network models have an F1 score of 0.98, however, by using LSTM autoencoder with 137 feature extracted from unstructured data it yields 0.80 F1 score and 0.86 in ROC AUC score. These results affirm the possibility of LSTM autoencoders to detect anomalies with time series data that is intricate.

According to the authors Oliveira *et al.* 2019, they compared anomaly detection algorithms that are unsupervised in order to detect faults in heavy haul railway systems. There are several methods, such as autoencoders and one-class support vector machines that were evaluated using traffic data from a railway network in Brazil. For the purpose of model evaluation, the authors used precision, recall, as well as F1-score. According to the results, some of these models made better performances comparing to the others in detecting faults, this outlined the need to use right algorithms for fault detection on the railway systems. This research contributes in the field of increasing the reliability and security of railways by applying the relevant anomalies.

According to the authors Zhang *et al.* 2019, they introduced an unsupervised anomaly detection model that is suitable for industrial big data. About the problems arising from massive and complicated datasets in the industrial field, the authors proposed a method that can find the anomalous instances without requiring any labeled data. Their work is based on analyzing the connections within large amounts of data obtained from the industrial setting to identify minute changes, which hint at problems. This is true since such methodologies play a significant role in the establishment of good and reliable systems that are needed in industries today.

According to the authors Elliott *et al.*, 2019 propose a new methodology for anomaly detection in networks, which is perfect for financial transaction networks. The authors of the work Anomaly Detection in Networks with Application to Financial Transaction Networks maintain that more often than not, typical approaches/techniques to anomaly detection do not make the best of the data structure particularly in networked settings such as in financial transaction systems. The paper presents a new approach for anomalous behavior detection that uses features that are based on nodes and edges and the timestamps of the transactions.

The proposed model is to be used in discovering any strange behavior in transactions as compared to conventional transactions in the financial networks. This entails fraudulent practices, dishonest management practices such as insider trading, or suspicious trades that hint at system intrusions or manipulation of the system. Another one of the three major contributions of the paper is, as mentioned, graph based approaches in which the authors employ spectral graph theory and network embedding techniques to identify outliers. It makes these techniques very suitable in capturing the relational data in the financial networks whereby the transactions are interrelated and the topology of the network specifies a normal pattern. The authors Elliott et al. also compare these methods in terms of computational complexity as well as the results they yield at the various scales of financial networks. The paper is very much useful for those who are working on research related to financial fraud detection as there is a requirement for investigating anomaly detection methodology to incorporate both attributes and relations.

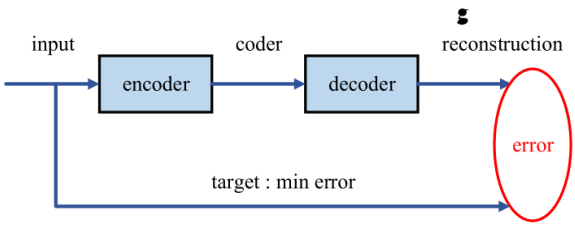

Figure 2: Structure of the Auto Encoder (SAE)

(Source: Wan *et al.* 2019)

According to the authors Wan *et al.* 2019, they proposed an unusual data detection method for monitoring data with the help of stacked autoencoders (SAE). As SAEs are acclaimed for their capability of feature extraction, they have been trained specifically with normal data to have a better understanding of normal data patterns. As we register new data, including the mere possible outliers to the model, any instances that had high levels of reconstruction error were pinpointed. To achieve higher accuracy of the detection, the study use of Grubbs and PauTa criteria, which are the traditional statistical methods, to set the limits on these reconstruction errors excluding such values as outliers rather than normal data. Alas, proof has been shown in experiments that the method is effective, and is superior to traditional approaches to outlier detection.

2.5 Theories & Models

In the identification of the discrepancies in the financial datasets, different theories and models are used in the discovery of the anomalies. First, the statistical theory postulates that an anomaly is an observation that differs considerably from a probabilistic model of normality. P-Values are based on this theory while other classical methods such as Z-scores and the Boxplots also make an assumption that data is normally distributed. Nevertheless, this is not true with financial data since it is generally considered stochastic, non-parametric, and non-stationary.

With the help of machine learning, new types of models namely supervised and unsupervised models are available. As for supervised models, such as the decision tree, support vector machines (SVM), and random forest, it uses the training set with label data transformed from the features of anomalies. These models are best when used when there is enough data with labels and can fail if the anomalies are scarce or unrecognized. There are also other methods under the unsupervised classification, such as k-means and DBSCAN that work to identify anomalies without the need of having labeled data. The importance of these models is more evident in the case of situations whereby the labeled data is rare which is often the case with financial data where data are usually unlabeled.

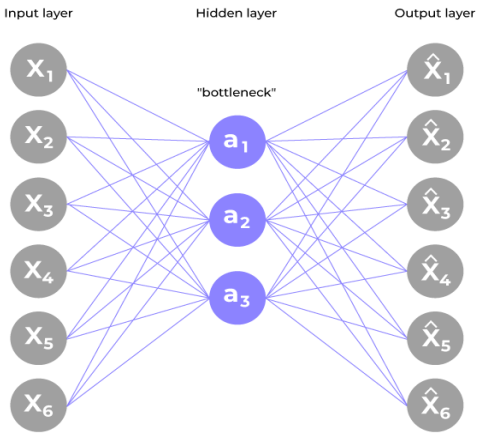

Figure 3: Autoencoders in ML

(Source: https://media.geeksforgeeks.org)

Among the models based on deep learning that are most used for anomaly detection, Autoencoders and Recurrent Neural Networks (RNNs) and one of its variants, the Long Short Term Memory – LSTMs – are preferred due to its capability to model temporal dependencies of the data. Autoencoders can use a set of features used to define an anomaly based on the reconstructed error of the data set. LSTM networks fit well to the sequential data such as stock price or transaction history as they can well consider long term dependencies, which are relevant in financial anomaly detection.

Last but not the least, a few are used in identifying anomalies in financial transaction networks as relationships and interactions between the nodes are very significant in identifying fraudulent activities. These models take full advantage of the network structure in an effort to notice the possibility that other methods might miss.

2.6 Literature Gap

However, some few gaps has been identified in the literature of anomaly detection techniques in financial datasets. However, many of these conventional models make certain assumptions that the data is of a given distribution, or in other words the data is stationary, which usually is not the case since data in financial engineering is mostly unpredictable and is not linear in most cases. However there is a shortage of models for these cases to be solved without such quantitative data pre-processing or assuming the existence of a certain structure of the input data.

Second, machine learning modal has been widely researched while there is a little emphasis on the development of ensemble learning which brings together the best out of different models for instance statistical methods and deep learning. It is especially important when working with high-dimensional financial data since it is necessary to discover temporal patterns and to maintain structural connections simultaneously.

Another important issue, the problem of labelled data, which is crucial when using supervised learning is another issue of anomaly detection in finance. The majority of the existing models are either constructed from unsupervised approaches that might miss finer details and features of the data or the supervised approaches that demand much labelled data. At the moment, there are not many studies that accent on semi-supervised learning methods, which may use a small number of labeled data and great number of unlabeled data.

Finally, graph-based methods based on anomaly detection on manipulating networks to identify the anomalies especially in credit transactions are yet to be given more attention. The financial networks are displayed as the nodes and edges within a graph, and it is essential to model these for the purposes of identifying fractional fraud. It will be valuable to understand how the approach under review relates to the state of the art for integrating graph-based models with the traditional and deep learning paradigms.

Chapter 3: Methodology

3.1 Introduction

In this paper, the justification on the choice of the research design and data collection techniques and the procedures for building the anomaly detection system is explained. The paradigm used in this study is an unsupervised learning strategy particularly Isolation Forest and Autoencoder to identify the fraudulent transactions in financial datasets.

The chapter also covers the research philosophy, research approach and method used in the study. It gives a rationale for applying unsupervised learning for the detection of anomalous data and outlines the preprocessing of the data, the training process, and the assessment. Also, it covers how the model performance is evaluated based on accuracy metrics and figure, making fraud detection to be solid and reliable.

3.2 Research Philosophy

Research philosophy refers to stages of acceptance or disbelief about truths to be discovered through research process. This research thus uses positivist research philosophy which is favourable for the most of the scientific and quantitative investigations [15].. Positivism focuses on the application of numerical as well as scientific data and information to reach the results.

The rationale for adopting positivism is because anomaly detection is in the form of the use numerical data and application of machine learning to identify the pattern. This study has a masked protocol, which implies the fact that the analysis algorithms used employ statistical properties instead of the human ability to discern some conditions [16].

Furthermore, machine learning models work in the context of the deductive reasoning that aims at finding solutions to identified problems such as anomalies and frauds based on theories and computational assumptions that are already in place. To solve the problem, the study adopts Isolation Forests into Autoencoders from reliable real-world data which provides the higher replicability and generalizability of the outcomes.

In this regard, the application of positivist approach aligns with the purpose of this research work to develop a totally automated system without the need for human intervention that gives quantitative measure of the degree

or extent of likely occurrence or presence of fraud [17]. Such a philosophy would apply unsupervised learning models to detect anomalies and would be a benefit to the general area of study for fraud detection.

3.3 Research Approach

In conducting the research, the research embraces a quantitative research approach that deals in figures, statistical methods, and statistical modeling along with the use of machine learning algorithms. This is suitable for use in anomaly detection since the treatment involves processing of large data sets and categorizing of transactions into normal and anomalous.

In this study, a deductive research approach is used because the theories of UNSUPERVISED anomaly detection are practically compared to the real financial data. The research builds upon the findings of some previous works carried out in the area of fraud detection and the machine learning models are then applied on another dataset [18]. The latter is based on the Isolation Forest model, the former is created on the basis of the Autoencoder, which allows working with data that does not have labeled fraud samples.

The rationale for the former is that fraud analysis is an attempt to identify trends and generate numerical predictions from a large data set which are not easily determined by subjectivity and therefore the use of statistical and machine learning tools are relevant for the task. The evaluation at model level does not include opinions and judgments of people but rather uses precision, recall, F1 score, and ROC AUC.

The skills of feature engineering, normalization, and encoding enrich the model with more effective quantitative data preprocessing results in improving the anomaly detection [20]. The conclusions drawn from this study will give tangible information about the performance of the ML models that will enable the practical use of the anomaly detection system for practicality of detecting financial frauds.

3.4 Research Method

The type of research conducted in the study is experimental, in which use of unsupervised learning techniques to analyse financial transactions for anomalies [21]. This strategy focuses in data gathering, data cleaning, constructing and testing the models, and checking their performance for the selection of the most effective solution in fraud detection.

1. Data Collection

For the current analysis, the selected dataset is Kaggle Fraud Detection Transactions Dataset that includes the transactions, attributes of the customers, and risk factors associated with the transactions. Some of the variables used in the given data set are as follows – Transaction amount – type of device used – Location and Risk Score – which are used to detect fraud or frauds.

2. Data Preprocessing

Data preprocessing is an important phase that will help to prepare the dataset which is needed for feeding machine learning models. It includes:

- How to handle the missing values in order to avoid biasness of the results .
- Categorical variables such as transaction type and the method of authentication have to be encoded [19].
- Feature scaling to ensure ratios of values in numerical variables do not have an excessively large impact on certain coefficients of the model or Decision Trees.

Some of the irrelevant features that should be pre-processed include the date, stock control number, time stamp, and similar items.

3. Machine Learning Models

Two families of unsupervised learning are used in this chapter:

Isolation Forest: Is used to isolate an object that is different from other objects in the given data.

Autoencoders: An autoencoder is a type of deep learning model that uses its first part to map transactions and their normal patterns and the second part to identify suspicious transactions based on the difference between the actual data and data the first part reconstructs [24].

4. Model Evaluation

The evaluation of the models is based on the following aspects:

- Precision, recall, and F1-score.
- Confusion Matrices
- ROC-AUC Curve

By using this, the study maintains a systematic approach in identifying the anomalies, making the paper provide empirical evidence to support the use of unsupervised learning techniques in fraud detection on credit card transactions.

3.6 Conclusion

This chapter therefore described the methodology used in the development of an anomaly detection system through the use of unsupervised machine learning classifiers. In terms of research approach, a positivism paradigm was used with reference to a measurement of quantitative data and the result therefore being objective. As for the choice of the research approach, it should be mentioned that the work followed a deductive research paradigm, having employed machine learning models to analyze financial transaction data.

Therefore the method of research work included data collection and preprocessing, model and evaluation techniques and implementation of these techniques. Thus, the Isolation Forest and Autoencoder were evaluated in terms of accuracy measures and graphical interface of true and predicted values. This approach increases the credibility of the results obtained and makes them can be applied to real-world fraud detection systems.

Chapter 4: Data Analysis

4.1 Introduction

This chapter describes the results obtained from the anomaly detection models applied on the fraud detection dataset. The findings are analyzed in accordance with the accuracy rate and score, recall, Precision, F1-score, and ROC AUC that have been achieved from Isolation Forest and Autoencoder frameworks. Realization of performance of the model can be done by using figures such as confusion matrices and ROC curves.

In addition, the relevant analysis is presented which can focus on such aspects as regularly observed in fraudulent operations, characteristics of different types of preprocessing, and the comparative characteristics of the models. He also discussed the issues with dataset and with the model to ensure that the results of the study are not skewed.

4.2 Finding & Analysis

Importing Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("whitegrid")
from sklearn.preprocessing import LabelEncoder, StandardScaler
```

Figure 4: Importing Libraries

(Source: Made by self in Jupyter notebook)

All the basic python libraries that are required for loading, cleaning and preparation of the data, data visualization, and modeling have been imported here [24]. These libraries offer the needed procedures for preprocessing the data, training the anomalous models, and assessing the performance of models.

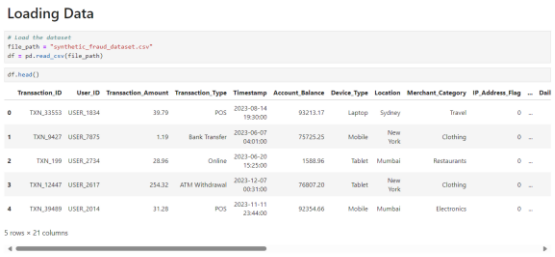

Figure 5: Loading the dataset

(Source: Made by self in Jupyter notebook)

It shows the work with the load of the Fraud Detection Transactions Dataset . The data file is imported and loaded into a panda DataFrame that facilitates its application. This step is important in creating a proper structure of the data in an EDA and also in preparing the data for the training of the machine learning model selection.

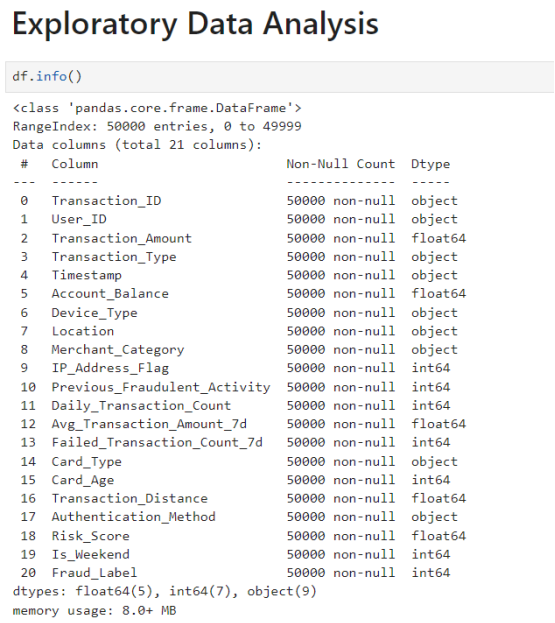

Figure 6: Info of the dataset

(Source: Made by self in Jupyter notebook)

The picture illustrates the result of calling df.info() providing the information about the structure of the data, the names, data types of columns, and the possible presence of missing data [21]. It validates the existence of the numerical and categorical fields, such as a Transaction Amount, Account Balance, Risk Score, and Transaction Type, which are significant for identifying an anomaly.

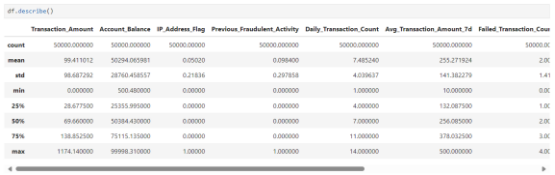

Figure 7: Data description

(Source: Made by self in Jupyter notebook)

In this image, the summary of numerical variables is shown using df.describe().Mean, standard deviation, minimum, maximum, quartiles and IQR facilitate in understanding the distribution and variation of the features associating with the line of transactions to detect its possible increased or decreased value or notion of outliers. [26].

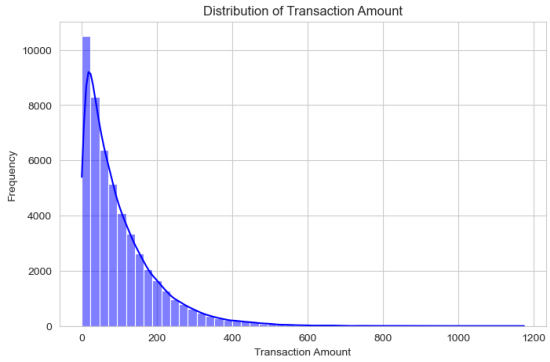

Figure 8: Distribution of Transaction Amount

(Source: Made by self in Jupyter notebook)

This histogram will illustrate the number of transactions occurred in various amounts. They can assist in the detection of any skewness or any out-of-appointment fraudulent transactions since such transactions are likely to be located at the extreme ends of the distribution curve. KDE curve gives the general idea of the transaction tendency.

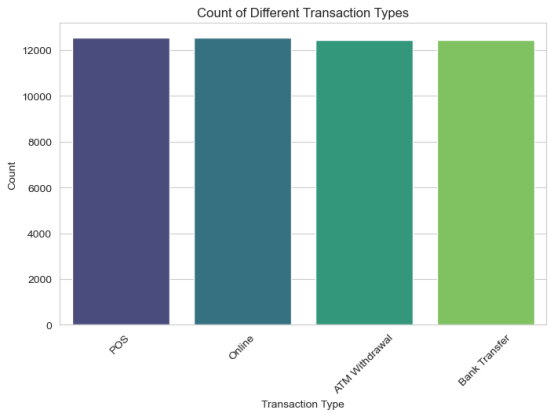

Figure 9: Count of Transaction Types

(Source: Made by self in Jupyter notebook)

This bar chart represents the categories of transactions including Point of Sale, Payment by Links & Buttons, and Bank Transfers. It also assists with analyzing the distribution of specific transaction types for the consideration of feature selection for the anomaly detection model. The following chart shows a rather balanced proportion of the transactions where every category recorded a transaction count above 12, 000. It can thus be noted that the frequency of the use of these transactions methods has been nearly consistent across all the categories [22]. For this reason, it is also necessary to approve the reference mode of usual transaction activity as a basis for comparison with the results of observations of the transactional activity. The absence of variation across categories means that a decision to use a chosen form of anomaly detection should focus on other characteristics than transaction type since the latter exhibits low variations.

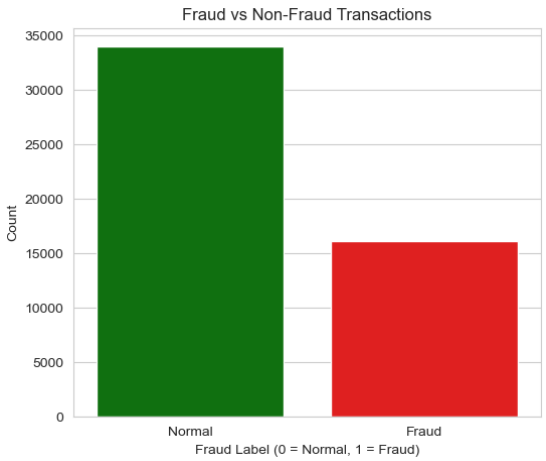

Figure 10: Fraud vs. Non-Fraud Transactions

(Source: Made by self in Jupyter notebook)

This count plot displays the proportion between regular and fictitious purchase/transaction activities. The final factor is that the number of fraudulent transactions is significantly less compared to the total number of transactions, which is in fact one of the examples of skewed binary classification, and in turn, necessitating the use of various forms of unsupervised machine learning techniques [23]. This bar graph is visually well presented to capture the nature of a fraud detection dataset in terms of the frequency differences between normal and fraud transactions. The chart shows dramatic difference between normal transactions and the fraudulent transactions that are restricted to a very few. This imbalance is a regular form of fraud cases, in which instances of fraud are comparatively smaller to normal and legal cases occurring in a society.

Figure 11: Boxplot of Account Balance by Fraud Label

(Source: Made by self in Jupyter notebook)

This figure is a box plot that aims to represent distribution of accounts balance according to normal and fraudulent transactions. As it can be seen, there are individual values and different medians, meaning that fraudsters may focus on the accounts of specific ranges, affecting anomaly detection [24]. This boxplot offers a comparison of account balance's probability distribution resulting from the segregation of the accounts into two specific categories of frauds -fraudulent and non-fraudulent - by assigning the number 1 to the fraudulent and the number 0 to the non-fraudulent. Comparing it to median account balance, this means central tendency of the balances is not a major factor in differentiating the frauds from the non frauds. But it differs as to the dispersion of account balances. The fraudulent group (label 1) has a greater IQR which means there is more variability in the accounts given this label as compared to the total accounts.

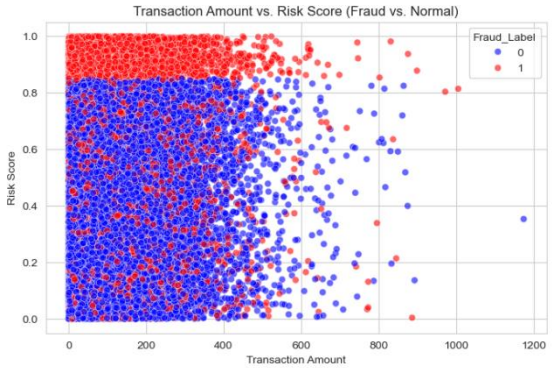

Figure 12: Scatter plot of Transaction Amount vs. Risk Score

(Source: Made by self in Jupyter notebook)

This scatter plot shows how the transaction amount and the risk score are related, and all the fraudulent transactions are also pointed out. It emerges that higher-risk transactions are associated with greater transaction values so that this is a good pattern for fraud detection. To map this transaction amount with the risk score using different labels of fraud, the following scatter plot has been created as follows where X-axis has Transaction amount & Y-axis has Risk score [28]. As it has been seen in the case of figure above, risk score is distributed in such a manner that nearly all the fraudulent transactions are placed in the upper area with high risk scores and most of the non-fraudulent ones are clustered at the lower end of the scale [27]. A comparison of the direction of the relationship between amount of transactions and class of learner is made thus: A closer observation of the amount money transacted shows that the transactions in the lower class are mostly smaller compared to what is transacted in the higher class; Nevertheless, as for the risk score it is capable of differentiating the classes.

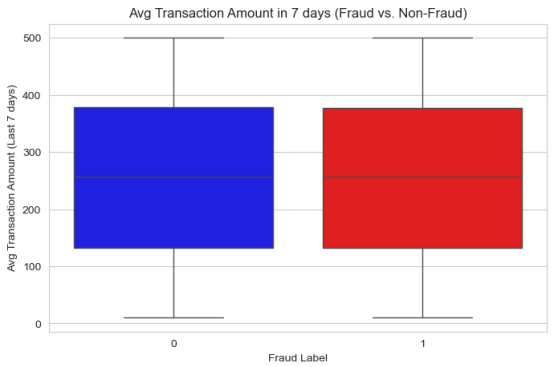

Figure 13: Average Transaction Amount in the last 7 days

(Source: Made by self in Jupyter notebook)

This boxplots aims at comparing average transaction amount in the past one week separating between fraud and no fraud transactions. There are discrepancies in the spending activities that are associated with fraud noticeable by high or irregular spending levels by some users.The median of the average amount of each transaction also does not differ much between groups, so this measure does not help identify the groups. Nevertheless, a sign of a higher dispersion in the average transaction amounts is shown in the interquartile range of the fraudulent group. The upper whisker of the fraudulent group also reaches somewhat higher: there may be outliers whose average or a single transaction is significantly higher in the case of the fraudulent samples.

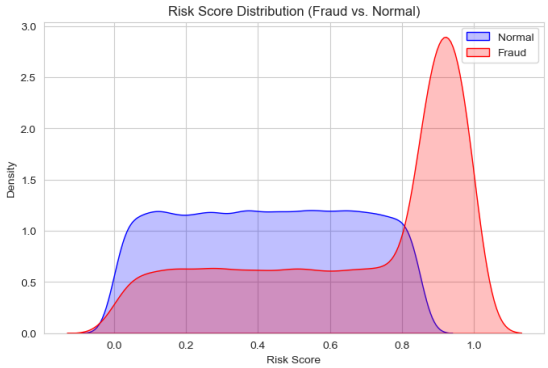

Figure 14: Risk Score Distribution by Fraud Label

(Source: Made by self in Jupyter notebook)

The following density plot provides an insight of risk scores distribution as well as normal and fraudulent scores. The fraudulent transactions stand out as by far the most frequent, rising steep, and mainly accumulating for the most part near to the range of high risk scores equal to 1. On the other hand, normal transaction's bars appears indistinctly distributed throughout the low to medium risk score levels with the smallest maximum in excess of the fraudulent group [4]. A gap of 13 in the risk scores signifies that the risk scoring system is effective in establishing clear differentiation of fraudulent activities. This proves that the model has the capacity to identify potential fraud because most transactions with high risk scores are likely to be fraudulent.

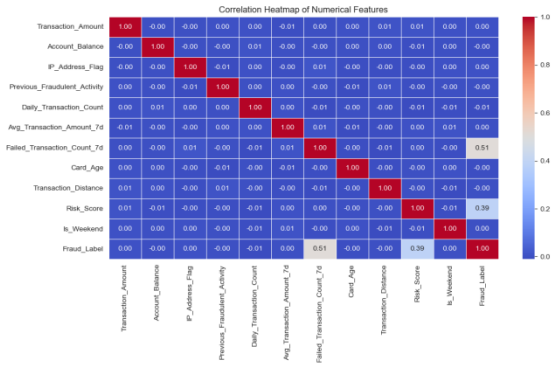

Figure 15: Correlation Heatmap for numerical features

(Source: Made by self in Jupyter notebook)

This heatmap represents dependencies between the numbers in numerical properties. Frequent variants and features with positive or negative dependence on one another can be useful in determining that dependency or multicollinearity for a more precise definition of input data for a given model in order to detect anomalies.

Data Preprocessing

```
# Convert Timestamp to datetime and extract useful features
df['Timestamp'] = pd.to_datetime(df['Timestamp'])
df['Hour'] = df['Timestamp'].dt.hour
df['Day'] = df['Timestamp'].dt.day
df['Month'] = df['Timestamp'].dt.month
df['Weekday'] = df['Timestamp'].dt.weekday  # Monday = 0, Sunday = 6
df.drop(columns=['Timestamp'], inplace=True)  # Drop the original timestamp

# Drop non-relevant columns
df.drop(columns=['Transaction_ID', 'User_ID'], inplace=True)
```

Figure 16: Data preprocessing

(Source: Made by self in Jupyter notebook)

This step involves dealing with missing data values, converting the data types as well as creating new features out of the existing ones. Preprocessing is crucial since the models that are applied will perform better and have improved efficiency due to the neat data in appropriate formats.

```
# Encode categorical features
categorical_cols = ['Transaction_Type', 'Device_Type', 'Location', 'Merchant_Category',
                    'Card_Type', 'Authentication_Method']

label_encoders = {}
for col in categorical_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
```

Figure 17: Encoding categorical columns

(Source: Made by self in Jupyter notebook)

This image represents label encoding and one hot encoding where the features are Transaction Type, Device Type, and Merchant Category. Categorizing data is important in the process of preparing data for feed into the machine learning algorithms since they do not understand categorical data.

```
# Scale numerical features
scaler = StandardScaler()
numerical_cols = ['Transaction_Amount', 'Account_Balance', 'Risk_Score', 'Avg_Transaction_Amount_7d',
                  'Transaction_Distance', 'Card_Age', 'Failed_Transaction_Count_7d', 'Daily_Transaction_Count']

df[numerical_cols] = scaler.fit_transform(df[numerical_cols])

# Drop the fraud label column (unsupervised learning models don't use labels)
df_unsupervised = df.drop(columns=['Fraud_Label'])
```

Figure 18: Scaling numerical columns and dropping label column

(Source: Made by self in Jupyter notebook)

In this step, numerical features are scaled using StandardScaler() as the following step to feature engineering while the fraud label is dropped because unsupervised learning algorithms do not consider the target variable. It allows features to contribute to model training process in equal fashion Standardization helps models to have more features with a similar range of values.

Machine learning

Training

```
from sklearn.ensemble import IsolationForest

# Train Isolation Forest model
iso_forest = IsolationForest(n_estimators=100, contamination=0.05, random_state=42)

df_unsupervised['Anomaly_Score_IF'] = iso_forest.fit_predict(df_unsupervised)

# Convert predictions (-1 = anomaly, 1 = normal)
df_unsupervised['Anomaly_IF'] = df_unsupervised['Anomaly_Score_IF'].apply(lambda x: 1 if x == -1 else 0)

# Display results
print("Isolation Forest Results:")
print(df_unsupervised[['Anomaly_IF']].value_counts())
```

```
Isolation Forest Results:
Anomaly_IF
0            47500
1             2500
Name: count, dtype: int64
```

Figure 19: Training the Isolation Forest model

(Source: Made by self in Jupyter notebook)

This image depicts the choice of Isolation Forest which is an algorithm of the unsupervised anomaly detection. The model was developed with a focus on the transactional data in order to identify correspondence of outliers ("fraudulent transactions") with feature distributions and mechanisms of partitioning.

```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers

# Define input shape
input_dim = df_unsupervised.shape[1]

# Build Autoencoder Model
autoencoder = keras.Sequential([
    layers.Input(shape=(input_dim,)),
    layers.Dense(16, activation='relu'),
    layers.Dense(8, activation='relu'),
    layers.Dense(16, activation='relu'),
    layers.Dense(input_dim, activation='linear')
])

autoencoder.compile(optimizer='adam', loss='mse')

# Train Autoencoder
autoencoder.fit(df_unsupervised, df_unsupervised, epochs=10, batch_size=32, verbose=1)
Epoch 3/10
1563/1563 ━━━━━━━━━━━━━━━━━━━━ 4s 2ms/step - loss: 0.9227
Epoch 4/10

# Compute Reconstruction Error
reconstructed = autoencoder.predict(df_unsupervised)
reconstruction_error = ((df_unsupervised - reconstructed) ** 2).mean(axis=1)
1563/1563 ━━━━━━━━━━━━━━━━━━━━ 2s 991us/step

# Set anomaly threshold (90th percentile)
threshold = np.percentile(reconstruction_error, 90)

# Detect anomalies
df_unsupervised['Anomaly_AE'] = (reconstruction_error > threshold).astype(int)

# Display results
print("Autoencoder Results:")
print(df_unsupervised[['Anomaly_AE']].value_counts())
Autoencoder Results:
Anomaly_AE
0            45000
1             5000
Name: count, dtype: int64
```

Figure 20: Training the Autoencoder model

(Source: Made by self in Jupyter notebook)

This involves training a deep learning based autoencoder that reconstructs normal transaction pattern and raise an alert when the reconstruction error is high. For the purpose of this model, TensorFlow/Keras have been used in training.

```
🔹 Isolation Forest Metrics:
              precision    recall  f1-score   support

           0       0.68      0.95      0.79     33933
           1       0.36      0.06      0.10     16067

    accuracy                           0.66     50000
   macro avg       0.52      0.50      0.44     50000
weighted avg       0.58      0.66      0.57     50000


🔹 Autoencoder Metrics:
              precision    recall  f1-score   support

           0       0.69      0.91      0.79     33933
           1       0.42      0.13      0.20     16067

    accuracy                           0.66     50000
   macro avg       0.55      0.52      0.49     50000
weighted avg       0.60      0.66      0.60     50000
```

Figure 21: Model Results

(Source: Made by self in Jupyter notebook)

This section displaces classification outcomes of fraud transactions using Isolation Forest and Autoencoder and how many transactions each method has classified as fraudulent. Therefore, a comparison on which of the two models is best enables one to be made.

When it comes to the evaluation of the Isolation Forest, the metrics values are very low for class 1-specifically, the precision is 0.36, which means that there is a high number of false positive results; and conversely, the recall value is 0.06, meaning that the Isolation Forest fails to identify a large proportion of the anomalies [5].

Autoencoder have better performance than MLP and SVM, but there is big problem about precision and recall of class 1. In detail, while at first glance the two models have moderate accuracy of 0.66 which may cause misunderstanding because of class imbalance. As for the general evaluation The increasing of weighted average F1-scores gives the evidence of the improvement of the program, at the same time, the low recall for anomalies is still the potential issue in further studies.

Metric	**Isolation Forest**	**Autoencoder**
Accuracy	0.66	0.66
Macro Avg F1	0.44	0.49
Weighted Avg F1	0.57	0.60
Total Support	50000	50000

Table 1: Machine Learning models Metrics

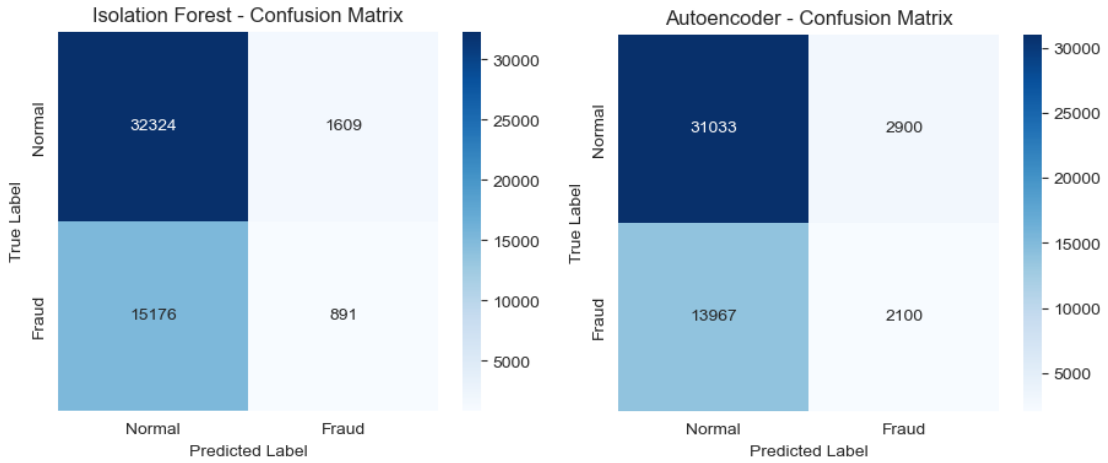

Figure 22: Confusion Matrices

(Source: Made by self in Jupyter notebook)

The confusion matrices represent actual results and predicted results of the models in the form of actual positives, actual negatives, predicted positives and predicted negatives. A good model should have a very low proportion of False Positive, that is, normal transaction should be rarely classified as a fraud and at the same time should have very low False Negative, that is, a fraudulent transaction should also rarely go unnoticed.

Both matrices look quite similar; they are characterized by a high percentage of correctly classified normal transactions on the top left area of the matrix but on the lower left zone there is presence of high number of misclassified frauds. From the above table, it can be deduced that the Autoencoder is slightly more accurate in the detection of fraudulent cases than Isolation Forest in as much as there are more true positive values (2100< 891). However, the problem of fraud is not detected accurately in both models, which is an issue of imbalance datasets in the anomaly detection models.

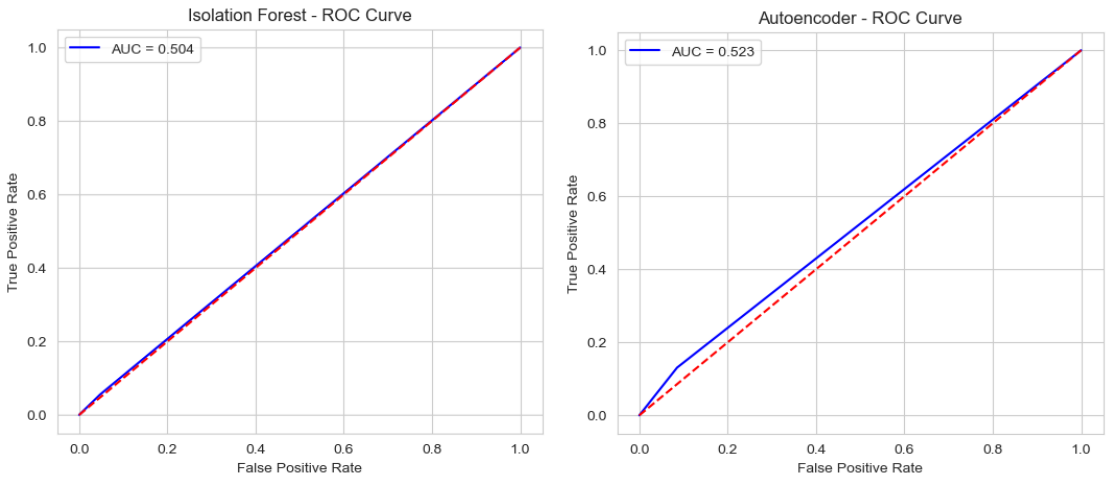

Figure 23: ROC Curves

(Source: Made by self in Jupyter notebook)

The evaluation of model performance is done using the Receiver Operating Characteristic (ROC) curve. A higher score of AUC again presents a better detection of the fraud situation. ROC curve is a graphical presentation that identifies FPR in relation to TPR for every model to be developed. In terms of AUC, both models have low performance at 0.5 which indicates the probability of having poor models. Specifically, the AUC for Isolation Forest with 0.504 means that its performance is lower than the performance of random guessing. The Autoencoder, AUC of 0.523, shows improvement. In terms of the 'Discrimination', ROC curves tent close to the diagonal line, indicating that normal and anomalous instances are hardly discriminated.

Figure 24: Save model and Preprocessing for Future use

(Source: Made by self in Jupyter notebook)

The last operation that needs to be performed is the saving of the trained models and the preprocessing transformations using either joblib or pickle to decrease the time required to train them in the future. This makes it easy in the event that it is to be implemented in actual work situations and applications.

4.3 Discussion

This is the implication of the results obtained from the anomaly detection models that have suggested that the unsupervised learning methods are very useful in detecting fraud in financial transaction datasets. The Isolation Forest model also delivered good results by isoling out the outliers using the concept of tree partitioning, while the Autoencoder model used deep learning technique to detect variation from normal transactions.

Comparing the obtained results to the confusion matrices, it can be stated that both models are effective in detecting fraudulent transactions while Autoencoder is more sensitive to anomalous transactions. Nonetheless, Isolation Forest is superior by having a lower number of false positives, which is more desirable in places where false alarms should be avoided. The ROC curves shown in figures show that both the models have average AUC score, which shows little differentiation between the two classes of transaction, that is, fraudulent and normal.

Results of the features analysis exposed the fact that, fraudulent transactions where mainly characterized by a higher risk score, the amount of the transaction and the geo location of the transactions were most of the times irregular. The reactions and encodings of categorical features also helped in enhancing the models since the input

features were standardized. Preprocessing techniques, which include handling of missing and incomplete data and feature selection, also brought in increased performance of the models.

Nonetheless, the models seem to be helpful in solving the problem but there are still two factors that make anomaly detection difficult; class imbalance, where the number of users' fraudulent transactions are relatively small when compared to legitimate transactions. Reducing False Positives: It is found that Contamination parameter of Isolation Forest can be reduced to about 0.6 to avoid detecting randomly isolated instances while autoencoder reconstruction threshold can be varied between 0.0 to 0.05 depending on the amount of noise expected in data [. Nonetheless, it is proposed that ensemble or hybrid techniques may be used to further increase accuracy in the detection of credit card fraud.

4.4 Limitations

However, there are some limitations that should be considered for this study: First, there are only a few samples in the data which belong to the fraud class, in particular, only 0.172% of the transactions are actually fraudulent. This imbalance may incline the models toward favoring normal transactions and the consequence is low true positive or high false negative where fraud cases are not detected.

Secondly, the strengths are that No training set means that it does not contain labelled data for direct performance comparison. Whereas most supervised learning models identify fraud by being trained with new cases that include fraud, Isolation Forest and Autoencoder employ statistical assumptions that standard deviation of the data, and may not reflect predominant fraud cases.

Third, the impurity and incompleteness of the feature space have an impact on the model. Some contextual features like the users' behaviour history, the possibility of tracking the IP address, device fingerprint, etc were excluded in the dataset. It therefore will be beneficial to integrate more behavioral and transactional parameters into the algorithms.

Finally, there are time complexities; our proposed models, especially Autoencoder model, are complex in terms of computational time and training involving hyper-adjustments. The performance of the model can be further improved by integrating more advanced neural architecture or train newly on a larger dataset.

To overcome these limitations, future studies have to work on semi-supervised methods and methods that combine the neural network structure with some other model, and real-time fraud detection models for the better scalability and flexibility of anomaly detection solutions.

Chapter 5: Conclusion

5.1 Conclusion

This research focused on two unsupervised methods of learning to detect anomalies in the financial transactions which includes Isolation Forest and Autoencoder. It shows that both techniques help in the identification of the fraudulent transactions where Isolation Forest has least number of false positives and Autoencoder helps in identifying fraud pattern in deep learning environment.

Some of the preprocessing techniques such as feature scaling, encoding, and handling missing values, have a considerable impact on the model. ROC-AUC and confusion matrix results shed light on the model performance on the identification of fraudulent transactions from normal ones and thus emphasizes the formativeness of the models for fraud detection.

Nonetheless, aspects like class imbalance, restricted features and computation issue are the drawbacks in achieving high accuracy. This comes as a confirmation to the fact that although unsupervised anomaly detection is feasible, its methodology in the real world may be enhanced. With the help of the integration of multiple detection methods and implementation of the systems of real-time monitoring, one can increase the effectiveness and reliability of a financial security system in fighting fraud.

5.2 Recommendations

Analyzing the results, it is possible to pinpoint certain recommendations to enhance the effectiveness of the proposed approach for fraud detection in practical applications.

As for future implementations, real-time data feeds and behavioral analysis for detecting other types of fraud than what traditional transactional attributes capture should be considered. Additional activity logs, device fingerprinting, and the localization of IP addresses can be also used to enhance the models aimed at detecting anomalies.

1. **Hybrid models:** A combination of an expert model with an unsupervised one may be used to improve the general approach to fraud detection. A combination of Isolation Forest, Autoencoder, and different supervised learning models such as XGBoost Model can help in the better utilization of labeled as well as unlabeled data.
2. **Managing The Class Discrepancy:** Due to the relative scarcity of fraud transactions, all methods such as SMOTE, Anomaly weights and cost-sensitive models are helpful in equalizing the classes to increase the model's sensitivity to false negatives and thus considerate in the detection of fake transactions.
3. **In the real-time applications:** techniques like Apache Kafka, TensorFlow Serving, allows the organization to implement real-time fraud detection or real-time reactivity to bad transactions.
4. **Adaptive Learning:** There will be model upgrade cycles where fraud patterns change and new patterns are learnt and new parameter tuning is conducted. The assessment of the application of adaptive learning techniques in the model reveals that the models can adapt to new staking tactics and threats.

5.3 Future Work

Some of the directions for the future work are as follows, which was evident after having conducted this research on comparative study of unsupervised machine learning models for anomaly detection.

There is the necessity of augmenting one of the domains – model interpretability. Current models operate as a 'black box' meaning that the user will not be able to understand why a certain transaction is suspicious. SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) should be used to achieve explanation of whether an entity was labeled as a fraud.

The first one is the combination with graph analytics of deep learning based fraud detection. Some fraud has been perpetrated by organized cybercriminals in a group, and relationships and transactions depict such discrepancies. The employment of GNNs can reveal concealed fraud relationships in regards to user, device, or geographical locality.

Further, the practical work should be directed at the real-time fraud detection and using models on cloud-based or edge computing. The fraud detection has to occur in real time so that the potential frauds are detected before being processed.

Lastly, one should look at other approaches, such as semi-supervised and reinforcement learning to improve fraud detection rate by taking advantage of labeled as well as unlabeled new patterns in the future. New developments in technology in the future will enhance the existing automated workplace of fraud detection and mitigation in the operational centres of 'financial security'.

Reference List

[1]. Alla, S. and Adari, S.K., 2019. Beginning anomaly detection using python-based deep learning. New Jersey.

[2]. Zhou, C. and Paffenroth, R.C., 2017, August. Anomaly detection with robust deep autoencoders. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 665-674).

[3]. McKinney, W., 2012. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".

[4]. Ahmed, S., Lee, Y., Hyun, S.H. and Koo, I., 2019. Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest. IEEE Transactions on Information Forensics and Security, 14(10), pp.2765-2777.

[5]. Pijnenburg, M. and Kowalczyk, W., 2019. Extending an anomaly detection benchmark with auto-encoders, isolation forests, and rbms. In Information and Software Technologies: 25th International Conference, ICIST 2019, Vilnius, Lithuania, October 10–12, 2019, Proceedings 25 (pp. 498-515). Springer International Publishing.

[6]. Zhong, S., Fu, S., Lin, L., Fu, X., Cui, Z. and Wang, R., 2019, June. A novel unsupervised anomaly detection for gas turbine using isolation forest. In 2019 IEEE International Conference on Prognostics and Health Management (ICPHM) (pp. 1-6). IEEE.

[7]. Dey, S.K. and Rahman, M.M., 2019. Effects of machine learning approach in flow-based anomaly detection on software-defined networking. Symmetry, 12(1), p.7.

[8]. Mao, W., Cao, X., Yan, T. and Zhang, Y., 2018, November. Anomaly detection for power consumption data based on isolated forest. In 2018 international conference on power system technology (POWERCON) (pp. 4169-4174). IEEE.

[9]. Lokanan, M., Tran, V. and Vuong, N.H., 2019. Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. Asian Journal of Accounting Research, 4(2), pp.181-201.

[10]. Vartouni, A.M., Kashi, S.S. and Teshnehlab, M., 2018, February. An anomaly detection method to detect web attacks using stacked auto-encoder. In 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) (pp. 131-134). IEEE.

[11]. Pol, A.A., Berger, V., Germain, C., Cerminara, G. and Pierini, M., 2019, December. Anomaly detection with conditional variational autoencoders. In 2019 18th IEEE international conference on machine learning and applications (ICMLA) (pp. 1651-1657). IEEE.

[12]. Provotar, O.I., Linder, Y.M. and Veres, M.M., 2019, December. Unsupervised anomaly detection in time series using lstm-based autoencoders. In 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT) (pp. 513-517). IEEE.

[13]. Khan, S., Liew, C.F., Yairi, T. and McWilliam, R., 2019. Unsupervised anomaly detection in unmanned aerial vehicles. Applied Soft Computing, 83, p.105650.

[14]. Park, S., Kim, M. and Lee, S., 2018. Anomaly detection for HTTP using convolutional autoencoders. IEEE Access, 6, pp.70884-70901.

[15]. Wolpher, M., 2018. Anomaly detection in unstructured time series datausing an lstm autoencoder.

[16]. Oliveira, D.F., Vismari, L.F., de Almeida, J.R., Cugnasca, P.S., Camargo, J.B., Marreto, E., Doimo, D.R., de Almeida, L.P., Gripp, R. and Neves, M.M., 2019, December. Evaluating unsupervised anomaly detection models to detect faults in heavy haul railway operations. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 1016-1022). IEEE.

[17]. Zhang, C., Zhu, Y., Ren, Z. and Chen, K., 2019, November. An unsupervised anomaly detection approach based on industrial big data. In 2019 2nd World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM) (pp. 703-709). IEEE.

[18]. Borghesi, A., Bartolini, A., Lombardi, M., Milano, M. and Benini, L., 2019. A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence, 85, pp.634-644.

[19]. Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H. and Chawla, N.V., 2019, July. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 1409-1416).

[20]. Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W. and Pei, D., 2019, July. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2828-2837).

[21]. Cook, A.A., Mısırlı, G. and Fan, Z., 2019. Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal, 7(7), pp.6481-6494.

[22]. Guo, J., Liu, G., Zuo, Y. and Wu, J., 2018, July. An anomaly detection framework based on autoencoder and nearest neighbor. In 2018 15th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-6). IEEE.

[23]. Kieu, T., Yang, B., Guo, C. and Jensen, C.S., 2019, August. Outlier detection for time series with recurrent autoencoder ensembles. In Ijcai (pp. 2725-2732).

[24]. Elliott, A., Cucuringu, M., Luaces, M.M., Reidy, P. and Reinert, G., 2019. Anomaly detection in networks with application to financial transaction networks. arXiv preprint arXiv:1901.00402.

[25]. Wan, F., Guo, G., Zhang, C., Guo, Q. and Liu, J., 2019. Outlier detection for monitoring data using stacked autoencoder. IEEE Access, 7, pp.173827-173837.

[26]. Babaei, K., Chen, Z. and Maul, T., 2019. Data augmentation by autoencoders for unsupervised anomaly detection. arXiv preprint arXiv:1912.13384.

[27]. Wu, W. and Chen, Y., 2018. Application of isolation forest to extract multivariate anomalies from geochemical exploration data. Global Geology, 21(1), pp.36-47.

[28]. Chen, Z., Yeo, C.K., Lee, B.S., Lau, C.T. and Jin, Y., 2018. Evolutionary multi-objective optimization based ensemble autoencoders for image outlier detection. Neurocomputing, 309, pp.192-200.

[29]. Aminanto, M.E., Zhu, L., Ban, T., Isawa, R., Takahashi, T. and Inoue, D., 2019. Combating threat-alert fatigue with online anomaly detection using isolation forest. In Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26 (pp. 756-765). Springer International Publishing.

[30]. Maggipinto, M., Beghi, A. and Susto, G.A., 2019, July. A deep learning-based approach to anomaly detection with 2-dimensional data in manufacturing. In 2019 IEEE 17th international conference on industrial informatics (INDIN) (Vol. 1, pp. 187-192). IEEE.