Università
Ca'Foscari
Venezia

Master's Degree

Data Analytics for
Business and
Society

Final Thesis

Fraud detection using machine learning and the effectiveness of different algorithms

Supervisor
Prof. Raffaele Pesenti

Graduand
Kanan Mammadli
Matriculation number:
888195

Academic Year
2022 / 2023

Abstract

Fraud has grown to be a significant issue as technology is developed in the banking, insurance, energy, and nearly every other area. current technology like artificial intelligence should be able to identify fraud in the real-world cases. This thesis defines the many types of fraud and their detection methods. At the same time, real-world case analysis was used to evaluate various models and prioritize machine learning preventative techniques. This paper examines the current state of machine learning applications, particularly in the financial sector, and analyzes the critical issue of credit card fraud. It showed the best approach to detect fraud in financial data carried out using the Python programming language in the last chapter, while the theoretical foundations of models were covered in first two chapters.

Keywords: Fraud Detection, Machine Learning, Algorithms, Banking, Technology

Table of Contents

Introduction

In order to make our lives easier, we use technology to conduct everything from shopping to business, and banking, and many of our hobbies are constantly evolving. At the point of transferring solutions for all aspects of life to digital, not only do numerous institutions but also individuals focus on digitally compatible business fields and hobbies. Here we might have some problems like fraud, cyber-attacks, and some more dangers.

Fraud is a serious business risk that should be recognized and moderated in time. According to analysis, a typical company loses its 5% of annual revenue because of Fraud. Data Science provides us with multiple techniques in order to prevent these frauds by analyzing mass amounts of data and finding anomaly patterns. Currently, machine learning is the leading tool that is used by various industries including banking, e-commerce, finance, healthcare and so on.

There are several advantages of using machine learning techniques to identify fraud. For example, it might take months or impossible to find suspicious patterns or behaviors if this detection is conducted by humans or in other traditional ways.

This thesis is going to explore fraud detection deeper, its coverage, technology, and algorithms that try to detect and how effective is machine learning in various domains. At the same time, research will investigate the supervised and unsupervised learning models' usage. in the Final Chapter, there will be quantitative analysis and practical solutions with a help of Python programming language which is going to help to explore the Fraud Detection topic deeper from a numerical point of view and together with qualitative analysis, helps to provide valuable insights related to our topic. Moreover, this thesis discusses the data collection methods and preprocessing techniques used to prepare the data for prediction. I hope this practical solution is going to be used in real-world scenarios.

Overall, this study aims to support ongoing efforts to reduce fraud risk in a variety of industries and help consumers to protect themselves from financial harm.

Chapter 1. The problem of Fraud Detection

1.1 Overview of fraud detection and prevention methods

Fraud detection is vital piece of each and every computerized business. It would be smart to implement it in any industry, its significance also demonstrates the widespread impact that fraud has on businesses today. The Oxford Dictionary characterize Fraud as follow: "*Wrongful or criminal deception intended to result in financial or personal gain*". In this chapter meaning of fraud and its types, basic techniques to detect fraud, introduction to machine learning will be covered.

Fraud detection and fraud prevention are two essential components of nearly every effective strategy to combat fraud. The ability to identify or discover fraudulent activities is referred to as fraud detection, while the measures that can be taken to avoid or reduce fraud are referred to as fraud prevention. There is a distinct distinction between the two: The first is an ex-post (based on actual results rather than forecasts) approach, while the second is an ex-ante (based on forecasts rather than actual results) one. However, two ways have one shared objective, *fraud reduction*.

Classical approach for Fraud detection is expert based approach which means fraud generally identified by fraud analyst who has business and domain understanding. This expert-based approach let analyst to use his/her intuition and experience to identify fraud. However, this manual investigation is going only in suspicious cases. For instance, a customer indicated that his card was charged even though he did not conduct any transaction. By the time Fraudsters developed new methods and ways to deceive Banks, Businesses and Organizations and it became hard to catch frauded actions.[1]

1.1.1 Types of Fraud

Fraud is broad term and there are various kind of frauds for non-financial or financial sectors. According to Jonh Marshall Bank there are several types of Fraud which divided also some sub-types. Identifying and dividing Frauds into several types helps companies to concentrate and focus on more the type that harm the business more. For example, Federal Trade Commission

[1] Bart Baesens, Véronique Van Vlasselaer, Wouter Verbeke. (2015) Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques., Wiley, pp 44.

Imposter listed the top 10 Fraud scenarios in last year and Imposter Scams climbed to the top of list with $616.0 Million damage.[2] Therefore, it is important to differentiate types:[3]

1.1.1.1 Bank Fraud

Bank Fraud is a type of Fraud that by obtaining of private information of someone help to access financial institutions. However, Bank Fraud itself divided into several parts, since there are different kind of methods Fraudsters use:

• Loan Fraud: It has been hundreds of years that thieves uses people's personal information to get loans from Banks. Especially nowadays Banks offers online Banking services and sometimes there are some security vulnerabilities and weaknesses that thieve can steal information. However, the most cases are related to small payday loan companies, because they have much lesser requirement to give loan.

• Account takeover: This approach mainly used by criminals where they use another person's account to make illegal purchase and withdrawals. Generally, these accounts are linked to credit cards. Fraudsters reach these accounts either online or just simple stealing.

• Wire Transfer Fraud: One of the common ways to stole money is wiring which include fake invoices, falsified e-mail traps, etc. According to Federal Trade Commission Wire transfers damaged about almost $0.5 billion in 2021 and this amount is bigger than total amount of credit cards and debit cards fraud.[4]

•Charge Back Fraud: This Banking Fraud is also known as Friendly Fraud. Friendly fraud is the point at which a consumer endeavors to start a chargeback interaction under fraudulent cases. Customers purchases good and services on online and claim his charge back while keeping the good on the hand. In this case sometimes Bankes are forced to give money back since it is sometimes hard to identify these issues on time. According to estimation, Chargeback fraud will be damaged about $25 Billion by 2025 with 41% growing-rate. Sometimes customers are innocent

[2] Federal Trade Commission. (2021). Consumer Sentinel Network: Protecting consumers in the digital age (p. 8).
[3] "Types of Fraud", John Marshall Bank, n.d., available: www.johnmarshallbank.com/resources/security-center/types-of-fraud/.
[4] Federal Trade Commission. (2021). Consumer Sentinel Network: Protecting consumers in the digital age (p. 11).

when they ask their money for the first time, and then they know it is easy to get refund and start to request it frequently. [5]

1.1.2. Phishing

Phishing attacks are fake emails, texts, phone calls, or websites that try to trick people into downloading malware, sharing sensitive information (like Social Security and credit card numbers, bank account numbers, and login credentials), or doing other things that put them or their organizations at risk of cybercrime. Phishing is mainly focused on touchy data, for example, Federal Retirement Numbers, Individual Recognizable proof Numbers and Tax Codes, PIN codes, etc.[6]

1.1.1.3 E-commerce

The selling and buying of goods, as well as the transmission of data or payments, over an electronic network, is called e-commerce. E-commerce is directly linked to internet where customers can surf on web pages to buy or order new thing for consuming. Then E-commerce companies charge these customers via online transaction for each good and service that they provide. When number of online shopping is increasing, the number of frauds is increasing at the same time. Fraudsters use different kind of techniques on internet to steal customers' private information or bank details. [7]

1.2 Process and Common methods used in fraud detection

Nowadays there are four accepted steps from data to insight in Fraud Analytics.[8] Companies turning data-driven approach to combat with Fraud and this four steps Fraud Analytics approach is accepted by globally. Fraud Analytics is collection and analyzing the data from different sources. Data sources are various, people may use financials transactions, customer records, etc. depends on Fraud that they combat.

[5] Liu, D., Lee, J.-H. (2021). CFLedger: Preventing chargeback fraud with blockchain. Journal of Systems and Software. Available: ScieneDirect.com
[6] IBM. What is phishing? Available: www.ibm.com/topics/phishing
[7] Atiqah S. Mat Taupit, N. Azizan. (2023) The Planning Process of the Online Transaction Fraud Detection Using Backlogging on an E-Commerce Website. Available: researchgate.com
[8] Delena D. Spann. (2014) Fraud Analytics: Strategies and Methods for Detection and Prevention. Wiley

1. Data identification. Picking right data might be essential of analysis since picking up wrong data won't allow to see what you are looking for.

2. Forensic data collection. When doing fraud analytics as part of an investigation, it is critical that you adhere to well established forensic preservation requirements. These criteria include preserving the data's chain of existence and confirming the reliability of the information to guarantee that all transactions have been documented and that no manipulation has occurred.

3. Data normalization and structuring. All information gathered will need to be organized and normalized in order for it to be connected, whether it came from internal sources or other sources. Some will be unstructured, like text-heavy data, while others will be organized, like data coming from databases. You won't be able to gain the fullest understanding of the data you have collected until it has been normalized and formatted.

4. **Data analysis.** As a last step Data Scientist should decide the best method to analyze data. Analysis should not only data-preparation and modelling, but also should contain great visuals. Additionally, models that are going to be used should be much more advanced in order to detect anomalies or patterns that couldn't be identified previously.

There are various methods to detect Fraud in Analytics part and effectiveness of these methods relies on sort of Fraud that companies face. Moreover, sometimes using union of these methods give extensive way to deal with Fraud. We will talk about detailed in next chapter, however here are the well-known common methods with brief explanation:

1. Rule-Based Approach

Rule-Based Approach is one of the possible systems to detect fraud and it work with simple pre-defined rules that was set. This system alerts when certain patterns are not same with pre-defined ones. Rule-Based systems are basically made with "*IF (condition) - ELSE (consequen*t)" statements or the complex ones by Data Science experts. [9] Customized parameters could be established with differential analysis to identify anomalies or potential frauds which was not seen

[9] Michaela Baumann. (2021) Improving a Rule-based Fraud Detection Systems with Classification Based on Association Rule

in trained data. Moreover, these methods perform well where Fraud's signs are clear and these signs can set as a various rule.

There are some drawbacks of this method as well. Scholars argue that this method is time consuming and analysis should be very precious since it is hard to identify all potential threats. There need huge support and maintenance for these systems, because new fraud types need dynamic, refreshed rules and system must take into consideration not only current or emerging but also all possible threats. When rules are too strict it may cause too much False Positives or it would be really hard to detect much severe frauds. Additionally, the system's efficiency declines severe when the more data the system must process. In order to avoid some of drawback the rules should be regularly updated.[10]

2. Anomaly Detection.

Imagine you are credit card auditor in the banking sector, and in order to protect customer of bank you need to take into consideration the transactions are not common. For example, sometimes some huge amount of spendings from student's card or spending money 4,000 km away from hometown. In these cases, customers generally confirm that payments made by them. These kinds of payments can be seen as a anomalies or outliers.

Every customer has own spending behavior and after a while banks or credit card companies get used to know customer's spending pattern. However, when these cards were stolen, spending behavior change immediately.

Anomaly Detection is also known as Outlier detection is one of the common methods that help to identify observation or group of observations that not similar to rest. This observation named as outliers and this method wide range accepted in Healthcare, Banking, Sensor and Video network surveillance, etc.

[10] Yufeng K., Chang-Tien Lu, Sirirat S., Yo-Ping H. (2004) Survey of Fraud Detection Techniques. International Conference on Networking, Sensing & Control.

With basic graphical representation in Figure 1.1 it is shown that most observations obey Gaussian Distribution (Normal Distribution), however opposite to rest, observations in circle R are completely different.

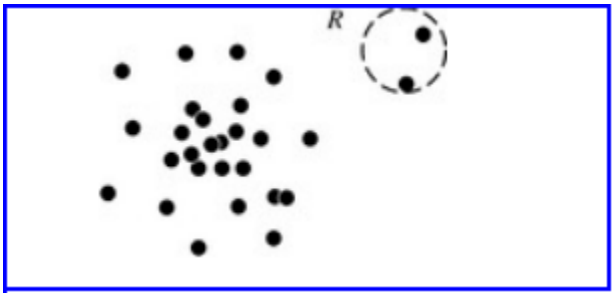

Figure 1.1. Observations in circle "R" are outliers

It is vital to remember that outliers and noisy data are not one and the exact phenomenon. There are approaches to identifying noisy data, which can frequently be a random mistake in a measured data set. Consider that for instance, client buying habits create random variables and that noisy data is produced when a customer purchases two identical items at once. However, it would be difficult to determine outliers if data scientists also took into account this noisy data. Therefore, certain machine learning approaches exist to remove noisy data in data-preprocessing process.

This approach has some shortcomings, including the use of distance measurement. Determining the distance ratio and the connection model to represent data items is crucial for anomaly detection. Different fields require various measurements. For instance, a little variation in health data science is sufficient to detect anomalies, but in customer-related sectors like banking, it will not produce sufficient significant results. Furthermore, as was already noted, noisy data may make things difficult and reduce the accuracy of results. [11]

3. Machine Learning.

Even if 2nd Chapter will be about Machine Learning techniques for Fraud detection and they work, it would be useful to start give an introduction here. Machine Learning, in simple terms, is part of Artificial Intelligence to allow computer to learn by itself. This indicates that the system is not programmed to carry out a particular task and instead learns, improvises, and adjusts on its own through experience. The Machine Learning models can obtain, preprocess and give

[11] Jiawei Han, Micheline Kamber, Jian Pei. (2011) Data Mining: Concepts and Techniques Third Edition. Morgan Kaufmann. Pp 444-450.

meaningful result by own based on target of project. Machine learning may be divided into three categories based on human supervision while training: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

• Supervised Learning: Algorithms that belong this type are trying to find or predict "label" based on rest of data features. It means, supervised model finds patterns in labeled data, and then use this pattern to make prediction in new data. Additionally, Supervised Learning is also divided into two parts: Classification and Regression. [12]

- *Classification* is group of algorithms that correctly categorize test data into specific groups, like distinguishing pears from bananas. For example, as a real case classification is used to identify spam mails in another folder of mailbox. SVM (Support Vector Machines), Linear Classifier, Decision Tree, Random Forest , KNN are all classification techniques.
- *Regression* - algorithms are aimed to predict numerical values based on the rest of the data features. It is generally about the relationship between numerical dependent and independent variables (they could be categorical and numeric). For example, the average grade of students could be predicted based on study time, sleep and other factors. In real life, businesses use it to predict sales revenue with the help of various factors, such as employee turnover, product price and even outside factors.

• Unsupervised Learning: This approach's algorithms are designed to examine and cluster "unlabeled" data, in contrast to supervised learning. These algorithms look for connections and patterns between data points. Unsupervised learning, for instance, may be applied to market segmentation to identify comparable clients based on their purchase patterns. It can be classified in to 3 categories because of usage: Clustering, association and dimensionality reduction.

- *Clustering* algorithms group data points based on similarities and specially distance. Clustering works by selecting the closest two data points by different distance calculations methods (Euclidian, Manhattan etc.) and then algorithm group rest of data point based on similarities. Generally, there are some accepted clustering methods: [13]

[12] Rudolph Russell. (2018). Machine Learning, Step-by-Step Guide to Implement Machine Learning Algorithms with Python. Pp 13-18.
[13] Julianna Delua, IBM Analytics, (2021). Supervised vs. Unsupervised Learning: What's the Difference?

1. K- means works by minimizing the sum of squared distances between each data point and the centroid of its assigned cluster, it divides the input data into K clusters.

2. Hierarchical clusters works by either dividing larger clusters into smaller ones (divisive) or smaller clusters are merging together (agglomerative).

3. DBSCAN is density-based clustering algorithm gather data points based on their density in input data space and when they are close enough, they are grouped in same cluster.

- *Dimensionality reduction* is techniques to minimize the number of input variables in dataset. When there are too many variables, it is becoming a bit complicated to analyze large dataset. Fewer input selection generates easy structured Machine Learning model, which bring degree of freedom and when there is too much degree of freedom, model is going to eventually overfit. Overfitting is always huge risk for future data. There are several ways to reduce dimensionality, such as Feature Selection, Matrix Factorization, Manifold Learning.[14]

- *Association* is A form of unsupervised learning called association analysis finds common patterns, connections, or links between elements in a group of items. Market basket research usually use association analysis to determine which goods are most frequently purchased in tandem. When association analysis is completed, the results are frequently presented as association rules, such as "if a customer buys product A, then they are likely to buy product B."

• Reinforcement Learning: This strategy may represent the most recent machine learning research trend. One of the issues in reinforcement learning is learning how to map situations to behaviors in order to maximize a numerical reward signal. These issues are closed-loop in nature and need experimentation on the learner's part to ascertain which activities will lead to the greatest rewards. In contrast to other types of machine learning, the learner is not given instructions. The likelihood that decisions will affect future circumstances and rewards, as well as the present benefit, makes these sorts of problems more challenging.

In short, Reinforcement learning is the process of figuring out how to maximize a numerical reward signal, and the actions taken by the learning system affect the inputs it receives in the future. The agent's abilities to detect the environment's state, act in a way that influences it, and

[14] Jason Brownlee. (2020) Introduction to Dimensionality Reduction for Machine Learning. Available: machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

have a goal or objectives connected to the environment are all necessary. In contrast to reinforcement learning, which is distinct, a system learns from a training set of labelled examples given by an external supervisor. The agent in reinforcement learning must draw their knowledge from their own experiences since it is usually impracticable to find instances of desired behavior that are both accurate and representative of all circumstances.

One of the useful examples could be a skilled chess player's makes a move on the game. The choice is influenced by planning, which includes imagining potential reactions and retorts, as well as quick, gut-level assessments of the merits of various stances and movements.[15]

Overall, Machine Learning is beneficial in financial sector, especially for Credit card frauds. Every single transaction that happens on a daily basis needs to be examined to prevent fraud. When new data is introduced, the system can test it automatically and provide a result since machine learning models are trained on historical data. Train data should only include legitimate transactions since during testing, a system may quickly identify fraud or other suspect behavior. After all, any questionable activity will be looked into independently.[16]

4. Deep Learning.

The current fad in artificial intelligence is deep learning, which achieves results with greater accuracy than machine learning in most cases thanks to the use of numerous hidden layers. The computational capabilities of deep learning enable this level of accuracy. Obviously, to identify extortion a huge number of perceptions ought to be investigated and prepared. Traditional machine learning algorithms frequently lack the capacity to process massive amounts of data as a result. Moreover, a few calculations had a top with a careful measure of prepared information and when the information size is expanded, the exhibition stays stable. Then again, this isn't true with profound realizing, where execution continues to improve as information sum develops. Deep neural networks, convolution neural networks, autoencoders, generative adversarial networks, and

[15] Richard S. Sutton, Andrew G. Barto. (2014, 2015). Reinforcement Learning: An Introduction. The MIT Press
[16] S. Khatri, A. Arora., A. P. Agrawal. (2021) Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison. Available: IEEE Xplore - ieeexplore.ieee.org

recurrent neural networks are just a few of the many subfields that fall under the umbrella of deep learning.[17]

• Neural Network: Structure and working mechanism of algorithm inspired by human brain. In the middle of the input and output layers there are several hidden layers. It is composed of layers of nodes, sometimes known as "neurons," that are interconnected and analyze input data before creating output data. It is going to be useful to explain neural networks in fraud detection scenario.

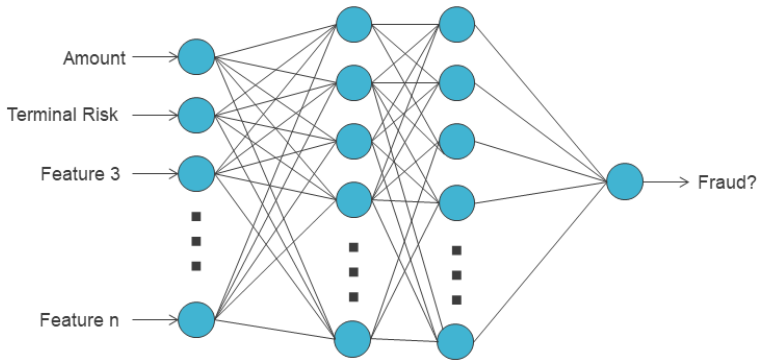

Figure 1.2. A network is made up of a series of layers, each layer comprising neurons that accept the output values of the layer before them as inputs.

Let's say the goal is to uncover credit card transaction fraud. In the first step, neural networks receive card transaction features as parameters for the input layer of the model, such as amount, terminal risk, etc. The likelihood that a transaction is fraudulent or not is finally shown by the output layer as a single neuron (Figure 1.2). The practitioner has an opportunity to which activation methods, extra layers, and hyperparameters they choose. The basic feed-forward neural network is the artificial neuron. It consists of an activation function named $\boldsymbol{\sigma}$ (sigmoid, ReLU, tanh) and a linear collection of

$$h = \sigma\left(\sum_{i=1}^{n} w_i * x_i\right)$$

activation variables as inputs. A list of $\boldsymbol{n}$ input variables, $\boldsymbol{x_i}$ may be used to determine the output, where they all try to determine $\boldsymbol{h}$:

The preferred algorithm for training feedforward neural networks is known as backpropagation. For all samples in the dataset, the same two crucial procedures are repeated: First, the forward pass

[17] T. Sejnowski. (2018)."The Deep Learning Revolution", MIT Press.

processes all the layers and sets up the input neurons with the sample's feature values to compute the predicted output.

The reverse pass is then started, in which an optimizer-like gradient descent reduces a cost function that measures the discrepancy between the anticipated output and the result obtained using the ground truth. Layer by layer, starting with the output layer and moving backwards to the input layer, the optimizer modifies the weights. [18]

• Convolutional Neural Network: CNN algorithm is typically used for image and visual data; however, it can also be applied to text data due to their excellent adaptability. A methodology for mining fraud patterns in credit card transactions based on convolutional neural networks (CNN) has been suggested by K. Fu. [19] They turned each transaction in the dataset into a feature matrix in order to reveal the underlying relationships in the time series. A better fraud detection model has been produced by combining the cost-based sampling technique with the unbalanced sample datasets.

Convolutional Neural Network consist of 3 main layers: convolutional layer, pooling layer, and fully connected layer. The convolution and pooling layers typically handle feature extraction, while the fully linked third layer handles the mapping of the extracted features to the final outcome, such as classification. *The convolutional layer* is the core part of CNN, is where the majority of feature extraction is done using linear and nonlinear processes like convolution operation and activation function. *The pooling layer* is doing a down-examining activity that reduces the element maps' in-plane dimensionality, establishing an interpretation invariance with minor alterations and shrinking the size of progressive learnable boundaries. The feature maps from the last convolution or pooling layer are connected to one or more fully connected layers, which are also known as dense layers. That time a one-dimensional array of integers is produced. At this layer, a learnable weight links every result to every single input. Fully connected layers translate the network's final result, such as the probabilities for each class in the classification problem, to the features that

[18] Yann-Aël Le Borgne, Wissam Siblini, Bertrand Lebichot, Gianluca Bontempi. (2021). Reproducible Machine Learning for Credit Card Fraud Detection- Practical Handbook.
[19] K. Fu, D. Cheng, Y. Tu, and L. Zhang. (2016) "Credit Card Fraud Detection Using Convolutional Neural Networks," Neural Information Processing Lecture Notes in Computer Science, pp. 483–490.

were extracted by the convolution layers, down-sampled by the pooling layer, and then down-sampled again by the convolution layers.[20]

1.3 Challenges and Limitations in Fraud Detection

As mentioned above Fraud Detection and its techniques are relatively new, so some challenges and obstacles are emerged recently. Of course, some challenges have been coming from the last decades and still actual (e.g., Data Quality). However, most challenges are coming by advancing technologies. Fraudsters are continuously working to improve techniques and methods. Nowadays, there are some common challenges for Scientist to detect frauds.

1.3.1 Data Quality Issues

Information can emerge out of a few sources, including informal communities, online business, medical care, and others, in organized, unstructured, and semi-organized designs. The absolute initial step is to manage this information concerning cleaning, eradicating rehashed information, pressing information, etc. The last kind of information is presently ready for stockpiling in distributed databases, NoSQL databases, etc. After that, data is broken down by implementing a different of methods, such as statistical analysis, Artificial Intelligence and data mining.[21]

One example is a novel strategy that combines sampling and data completeness to improve the quality of data used to detect credit card fraud and overcome limitations. Filling the dataset's sparse matrix with spectral regularization as a first step will reduce the impact of missing data. An over-sampling strategy should also be used to correct the imbalance between positive and negative samples and ensure that the sample percentage in the dataset remains constant. Then, on the dataset, evaluate how well spectral regularization methods outperform conventional matrix completion methods and methods for getting rid of missing values. Nevertheless, compare the various sample ratio approaches by employing over- and under-sampling. Experiments show that the credit card fraud detection dataset can benefit significantly from the suggested spectral regularization and oversampling techniques.

[20] Muhammad L. G., Anazida Zainal, Mohamad Nizam Kassim. (2022). A Convolutional Neural Network Model for Credit Card Fraud Detection. International Conference on Data Science and Its Applications.
[21] Vinaya Keskar a, Jyoti Yadav b, Ajay Kumar (2022). Perspective of anomaly detection in big data for data quality improvement. Available: Science Direct.

Data Imbalance issue comes due to the inability of traditional classifiers to correctly differentiate the poorly represented classes, some classes in the data are underrepresented. There are two common methods to handle imbalanced data. Under sampling for majority type, oversampling for minority type. A common sample-based method known as one-sided selection can potentially eliminate most of the training samples, including noise, boundaries, and redundancies. However, these procedures typically only solve a minority of situations, so they may not be very useful if the majority/minority ratio is too high. Training with multiple classifiers is one of two additional methods of under sampling. These methods reduce prediction variance by using a large number of subclassifies rather than dealing directly with noise or boundary data.

Oversampling is the exact opposite of undersampling. Some samples are inserted or duplicated to reduce data imbalance. The oversampling approach assumes that the neighborhood of positive instances and the instances between two positive examples are both positive. However, similar hypotheses may be factually based.[22]

1.3.2 Large Amount of Data.

The volume of data significantly affects the effectiveness and efficiency of fraud detection systems. For instance, with credit card detection systems, processing time and also complexity are decreasing when there are fewer transactions. The consumption habits of cardholders, which have a high correlation with characteristics, are determined by credit card attributes. There are more than 20 feature of in credit card as an attribute such as Overdue, Number of cards holding, card usage repetition, etc. To handle these amount of feature, use of data reduction approach applied, and this approach known as dimensionality reduction and numerosity reduction.

A nonparametric method called the "Numerosity reduction approach" is used to collect credit card transactions and identify fraudulent transactions by using these aggregations to model and estimate customer purchasing behavior prior to each transaction. In contrast, a Principal Component Analysis-based dimensionality reduction strategy was used in credit card fraud detection systems to reduce the credit card training dataset's dimension. In addition, Sherly and Nedunchezhian selected relevant characteristics for the model using an embedded method that is carried out as part of the Iterative Dichotomiser 3 decision tree. The wrapper strategy, developed

[22] Rongrong Jing, Hu Tian, Yidi Li, Xing Wei Zhang, Xiaolong Zheng ,Zhu Zhang , Daniel Zeng. (2019). Improving the Data Quality for Credit Card Fraud Detection.

by Paasch and Lei, makes use of a genetic algorithm for global search to heuristically locate subsets of characteristics while this is taking place.[23]

The attributes of interest to fraud detection models in payment platforms may be found in many types of reference graphs with various properties. This intricacy causes trouble in creating successful graph resolution. Due to its structural properties, graph neural networks offer the potential for parallelism, making it potentially feasible to execute batch calculations from many graphs.

Lately, methods to enable operation on many graph types have recently been presented. For example, Cross-channel fraud involves managing numerous graphs with distinct properties at once while also keeping to real-time response standards. These graphs may be combined into a single structure using several techniques, and label propagation can be carried out. Cross-channel fraud presents several practical issues, but no clear solution has yet been identified to handle them.

Due to their enormous size, large graphs, like those employed in financial crime and fraud detection applications, present particular difficulties. In payment systems, transaction graphs might encompass tens of millions of users, and their interconnectedness increases the complexity. It is quite difficult to comprehend and draw conclusions from these huge graphs or their subgraphs. To find financial crime, it's essential to recognize the differences in features across various sub-graphs. For instance, a clique of nodes in the graph may represent a network of fraud and illegal money laundering. Techniques to identify and characterize the differences in diverse graph areas have recently been presented. These methods are especially helpful for examining and resolving alerts on sizable graphs that call for the interpretation of several sub-graphs.

1.3.3 Dynamic Fraud patterns.

Dynamic fraud patterns are the ever-changing methods fraudsters use to commit crimes. As fraudsters create new strategies to target weaknesses, these patterns change over time. For fraud detection systems, seeing dynamic fraud patterns creates significant hardness as they must keep

[23]Aisha Abdallah n , Mohd Aizaini Maarof, Anazida Zainal. (2016). Fraud detection system: A survey

up with the latest trends and fraud patterns. Financial frauds could be the best example in this context and there are some techniques that attacks become adaptive regarding to defense systems.

• Bypassing Detection Shield: Rule-based methods for identification of fraudulent activity are currently losing their efficiency because of how easily fraudsters can predict static rules and find solutions for each step. Since fraudsters frequently use cutting-edge tactics to evade the use of machine learning detection, supervised learning systems also confront significant difficulties. For instance, when machine learning-based detection systems reject fraudulent payment transactions from a financial crime organization, new strategies are frequently used right away.

• Changes in Fraud Schemes and Channel Usage: As stated earlier, the structures of financial crimes have significantly changed from straightforward single-channel instances to intricate multichannel situations. Fraudulent have also been using modern techniques that aren't often connected to related fraud forms. For instance, according to industry studies, money laundering operations have been increasingly using new channels like digital currency (Bitcoin, ADA, Ethereum) transactions.

• Rapidly Altering Strategies: Using mobile devices and making peer-to-peer payments is in target of criminals. These payments are really type of fraud, because these frauds display quick tactical adjustments that brings difficulties for supervised learning-based approaches. For these application scenarios, anomaly detection and adaptive methods have been investigated and used. In the long run, adversarial and adaptive strategies offer an extremely feasible resolution pathway since they most properly represent the application environment.

• Robustness: The whole robustness of defining resolution is greatly impacted by adversarial strategies. Fraudulent groups may frequently respond to defensive actions quickly and on a large scale. A rising problem is their greater use of automation and technical advancements. The robustness of ML-based detection techniques may be significantly impacted by these properties in real-world systems. [24]

[23] Eren Kurshan, Hongda Shen. Graph Computing for Financial Crime and Fraud Detection: Trends, Challenges and Outlook. 3.6 / 3.8 . Available: arxiv.org

1.3.4. Noisy Data.

Noisy Data is dataset that contain illegible, noisy, and include duplicated or old entries. As a result, many scientists choose to perform a preprocessing phase on the dataset prior to developing their model in order to clean it up and turn it into an appropriate shape. Cleaning, integrating, and lowering the data's dimensionality are the three processes that make up data preparation, as was previously indicated. The goal of data integration is to handle issues related to missing values, outliers, and inconsistent data while retaining data (often divergent) that originates from different resources within a single dataset. Trying to minimize the total amount of dimensions in highly dimensional data is known as reduction of dimensionality. Noisy data can have a significant negative impact on the identification of fraud, particularly in terms of accuracy. It has also one of significant impact on variance in the dataset. The key distinction between outliers and noise means that the earlier type isn't in the structure's best interest and may have been brought on, for example, by a mistake made by a person. It's crucial to spot outliers in a system since they might be interesting abnormalities with important implications. It also draws attention to the fact that systems may label fraudulent activity as noise. This emphasizes how important it is to do a preprocessing step before training a model. This is due to the fact that noise may be randomly generated or purposefully created by scammers. Therefore, in order to successfully identify and deal with cases of fraud, Fraud Detection Systems must be able to distinguish between noise and real anomalies.[25]

1.3.5. False Positives.

Sometimes legal transactions and activities are wrongly labeled as fraudulent in fraud detection, which raises the number of false positives. Calculating the False Positives Rate is as follows:

$$FPR = FP / (FP + TN)$$[26]

FP stands for "false positives", whereas TN stands for "true negatives."

[25] Kasra Babaei, ZhiYuan Chen, Tomas Maul. (2020). A Study of Fraud Types, Challenges and Detection Approaches in Telecommunication. Pp 255.
[26] Google Developers. Machine Learning Crash Course. False Positive and True Negatives.

Limiting the rate of false positives while applying data science approaches is one of the most difficult issues. During the investigation of fraud in large amounts of data or transactions, there are too many anomalies appear. Any considerable number of exceptions, nevertheless, points to internal control problems. Further strategies may be applied to draw attention towards the transactions with greater threats rather than cycling through all the exceptions. Rerunning the tests with an alternative or more stringent criteria isn't complicated. Particular sets of transactions might be isolated for analysis if several of the anomalies share similar traits. Auditors and researchers are more willing to use these time-saving strategies since testing is simple to apply, particularly when they are fully automated. It is reasonable to employ a comprehensive or complete method for performing any data-driven test which could disclose red flags of fraud due to the simplicity of use as well as the speed of data analysis software. After evaluating the preliminary findings, the tests may be changed with relatively less effort.[27]

False Positive problem is still actual in the industry and it has 11-16% rates. Only one out of every five transactions reported as fraudulent are actually fraudulent. According to some experts, this large radius of false positive rates may cause greater damage or expense than actual fraud. The majority of businesses have implemented a multi-step procedure which incorporates the work of human scientists with artificial intelligence systems in order to reduce issues. In order to identify possible fraudulent activities, this procedure often begins with AI model that generates a risk rating and combines it with scientist criteria. The consequent warnings appear in a 24-hour monitoring centre, where human beings review and classify them. This procedure has a chance to lower the FPR by 5%, but this reduction requires substantial (and expensive) human participation. Even if these procedures are applied, a lot of False Positive will still remain.

Automation of Feature Engineering. Let's imagine we are trying to investigate Credit Card Fraud. Having a mass amount of data about transactions or other activity during the Fraud Detection process helps to create more potential features. There are several efficient ways, for example, manually selecting and producing the features which is a costly and time-consuming process. Therefore, scientists developed an automatic way to easily create a thorough collection of attributes that describe the usage habits for an individual account or payment. According to scientists if this method will be applied, analysts could focus on ML algorithms and results. DFS

[27] Sunder Gee. (2015). Fraud and Fraud Detection. Wiley Press. Pp 325.

is presented with relationships and objects that represent the data's structural nature. Researchers could create features in two different methods based on the various characteristics that are gathered throughout each transaction:

- By relying just on transactional data: Every registered activity is described by a number of properties from which we can derive a plenty of features. The majority of features are straightforward inquiries with binary responses, such as "Have you ever been in Baku branch?". By applying one-hot encoding to transform categorical values, these variables are produced. Furthermore, the transaction's whole numerical component is accepted in its current state.

- By combining historical data: Since each transaction is connected to a certain card, analysts will have access to all historical or previous data. By combining this data, we may create variables and these features are primarily numerical; one example is "How much do he/she spends daily?". The difficulty in obtaining these variables comes from the circumstance that a researcher is able to utilize aggregates created about the card using transactions that occurred prior to time η when producing variables that characterize a transaction during the moment. In addition to that, this method is operationally costly both while training the model and using these variables.[28]

[28] Roy Wedge, James Max Kanter, Kalyan Veeramachaneni, Santiago Moral Rubio, Sergio Iglesias Perez. Solving the "false positives" problem in fraud prediction. Ch. 6-10. LIDS, MIT, Cambridge, Banco Bilbao Vizcaya Argentaria (BBVA).

Chapter 2. Machine Learning Techniques for Fraud Detection.

As was previously noted, machine learning makes use of computer capacity to identify hidden or dynamic patterns even in datasets with an extensive amount of data. Machine learning algorithms are able to distinguish between legitimate and fraudulent transactions with the use of previous data training. One major benefit of machine learning is its capacity to adapt and adjust strategies as fraud risks evolve. Since everyday new fraudulent methods are emerging, machine leaning models could be updated and improved in order to be effective.

Furthermore, the fraud detection business has been receiving new ideas thanks to newer technologies as ChapGPT. This evolutionist tool supports companies in proactively fighting off attacks from fraudsters. Certainly, we will go deeper into this subject.

Machine learning algorithms' definition and scope of use are already mentioned in first chapter, however in this chapter we will go deeper and try to mentioned everything about using machine leaning in fraud detection.

2.1. A new era of artificial intelligence – ChatGPT and Fraud detection.

Recent advancements in natural language processing have been made possible by the growth of Large Language Models (LLMs), especially GPT-3 (2020), PaLM (2022), and ChatGPT (OpenAI, 2022), which enable far better outcomes for related tasks like language understanding (2022), question-answering, and communication systems. [29] The advantages of such systems may be seen in a variety of fields, include medicine, education, banking etc. These systems have demonstrated exceptional ability in interpreting and producing humanized communication. These models are always getting more accurate and adaptable thanks to their capacity to derive information from enormous volumes of data, opening the door to novel and intriguing possibilities. At the same time thoughts have been raised about these systems' possibility of being used maliciously as a result of their widespread adoption. The employment of huge language models to pose as human users and take part in illegal activities like fraud, and pestering represent a few of the biggest concerns. As an example, cybercriminals may employ ChatGPT agents to take over every single customer service platform for different businesses, like online shopping, aviation, and

[29] Hong Wang , Xuan Luo , Weizhi Wang , Xifeng Yan.(2023). Detecting ChatGPT Imposters with A Single Question. University of California Santa Barbara , Xi'an Jiaotong University. Pp 1-12.

financial institutions. Additionally, machine-generated speech might potentially take control of telephone systems such as 911 using the assistance of text-to-speech (TTS) technology, resulting in serious societal emergencies. Both trustworthiness and reliability of online relationships might be seriously damaged by these kinds of assaults, which could also impact internet service companies and their customers.

2.1.1 Information Security Risks from ChatGPT.

As it was mentioned above this AI has huge threats for several industries. But first of all, we need to understand how do these systems work. It is intended to allow users to input in natural speech while receiving responses in understandable speech. According to the circumstances, ChatGPT evaluates the person's data and responds. To understand person's data, the system employs ML and rules-based techniques. After processing the data, the system generates an answer in accordance with the set of rules.[30] The person using it is then supplied the reply in a format that they recognize. The ability of ChatGPT to interpret input from natural languages and provide replies that are suited to the user's particular requirements and circumstances serves as one of its key advantages; this makes it simpler for users to communicate with the bots and decreases the likelihood that they will get an unsuitable answer. The system may also be used to give people experiences that are more customized to their needs, such as relevant suggestions or individualized support.

Regardless of all of these expected advantages of ChatGPT, there has been shown that the platform draws hackers. The system may potentially continue to develop a channel through which attackers may quickly launch cyberattacks. In accordance to Check Point Researchers (CPR)[31], there are a number of cases when hackers have admitted to using ChatGPT from OpenAI to aid their criminal acts. Given that one of the practices on the Dark Web is the trade in of viruses, a report by CPR demonstrates the possibility of expansion of the Dark Web Marketplace. A piece of code written by fraudsters that use the application programming interface (API) as an external resource serves as proof of this. The purpose of this algorithm was to obtain immediate and up-to-date cryptocurrency, particularly Bitcoin, Monero, and Ethereum. Although ChatGPT is still in its

[30] J. Robinson, 2023. "The cost of science a look at the ethical implications of ChatGPT" .
[31] Dash, Bibhu, and Pawankumar Sharma. (2023) "Are ChatGPT and Deepfake Algorithms Endangering the Cybersecurity Industry? A Review." International Journal of Engineering and Applied Sciences, 10(1).

early phases of development, there is a good chance that, if poorly protected, this platform might increase security issues. In conclusion, AI may have a harmful effect on cybersecurity.

2.1.2 Advantages of ChatGPT on Fraud detection and Information Security

ChatGPT can act as a defense engine against online threats. By altering the algorithms which underpin fraudulent activity and behavior, attacks can be stopped. It can be customized for various business kinds for the purpose to educate staff members about cybersecurity and reduce the effectiveness of phishing attacks. To protect corporate data, ChatGPT can help with bug fixing and fraud detection.[32] Additionally, we can say that, ChatGPT is able to be applied as an extra dimension for fraud protection. ChatGPT quickly recognizes possibly fraudulent transactions or behavior through evaluating trends, user habits, and financial information. Whenever AI notices suspect account access, weird purchases tendencies, or additional conduct that differs from the client's typical financial habits, it is able to send notifications. ChatGPT is able to adjust and improve its prevention skills by ongoing learning, and keeping ahead with new fraud patterns and methods.[33] ChatGPT may help in detecting and preventing with the help of Natural Language Processing to find abnormalities in emails. Additionally, ChatGPT may assess if the language used is comparable with the person's prior styles of writing by comparing the body text of emails to earlier emails written by a particular person. It facilitates in preventing fraudsters from trying to defeat the detection of fraud by speaking an identical language as real individuals.

2.2 Theoretical foundations of machine learning for fraud detection

The application of machine learning teaches computers how to deal with data more effectively. Occasionally, even after examining the data, we are unable to evaluate or extrapolate the information and here we are using machine learning algorithms. The availability of a large number of information has increased the need for machine learning. Machine learning is used across numerous sectors to retrieve valuable information and the aim is to get this information from the dataset. Many studies have focused on how to make robots learn on their own despite the fact that they must be manually programmed. Many statisticians and scientists use a variety of

[32] Pawankumar Sharma, Bibhu Dash. (2023). Impact of Big Data Analytics and ChatGPT on Cybersecurity. International Conference on Computing and Communication Systems (I3CS).
[33] David Peterson. (2023). Transforming Banking with ChatGPT: Enhanced Chat, Fraud Detection, Personalized Recommendations, and Risk Management. First National Bankers Bancshares Inc.

techniques to solve this challenge, which involves handling enormous amounts of data.The researchers want to emphasize the fact that there's no single algorithm that works optimally in all situations since machine learning applies a variety of methods to handle data-related issues. The sort of algorithm used will vary depending on the type of issue you're trying to answer, how many variables there are, what kind of model will work most effectively, and other factors. In the following chart (Figure 2.1) you can find the most popular machine learning algorithms which we have mentioned in first chapter, but we will discuss all details in second chapter.[34]

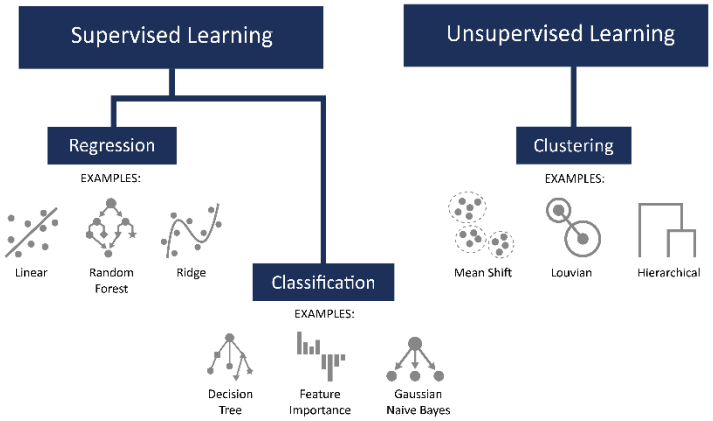

Figure 2.2.1 Machine Learning techniques

2.2.1 Supervised Learning

We have already given a definition for Supervised and Unsupervised learnings by couple of sentences in first chapter. Not it is time to how do this learning and its algorithms work and help to identify Fraud detection.

Supervised learning is using trained data in order to get proper result. Trained datasets contain information about the past and algorithms are using historical data to get results and modify it until getting the best result. The loss function serves as an indicator of the algorithm's reliability, and iterations are made till the error rate is suitably reduced.

[34] Batta Mahesh. (2020). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR).

2.2.1.1 Linear Regression.

Linear regression is supervised learning algorithm, specifically a regression model. It try to identify relationship between dependent and independent variables. In context of Machine learning it is defined as – the continues dependent variable ***y*** is identified by independent single or more ***x*** explanatory variables. Regression analysis's main objective is to forecast a continuous objective factor. [35]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + e$$

βo represents here the intercept of regression line, while *β1 and β2* are constants. In linear regression works with as all other supervised learning algorithms: train labeled data in order find unlabeled data with testing.

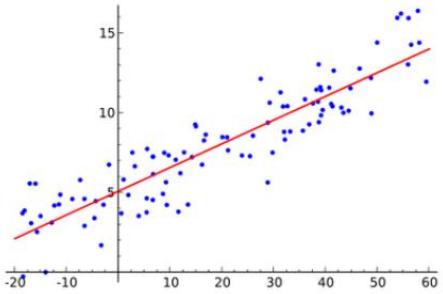

Figure 2.2.2. Visual Representation of Linear Regression.

For example, Figure 2.2.2 illustrates the process of linear regression and here the final model is represented by red line. Bleu points here the model's training data and by minimizing loss function model try to fit training dataset as accurately as possible. By minimizing the coefficient of a selected loss function, that measures the difference between the algorithm's projected values and the actual labels of the training data points, the fitting process is accomplished. When this optimizing is over, with the help of model it is possible to predict *y* which were unknown.

[35] Vladimir Nasteski. An overview of the supervised machine learning methods. Faculty of Information and Communication Technologies, Partizanska bb.

When used for identifying fraudulent activity, linear regression has several restrictions. It is also assumed by several researchers[36] that linear relationship, could fail to determine the complicated and chaotic trends frequently observed in fraud data. Additionally, susceptible to anomalies, linear regression's effectiveness may be impacted. It might be difficult for linear regression to effectively categorize the minority (fraudulent) class in the presence of imbalanced data, which is prevalent in fraud detection. Furthermore, although linear regression approach is generally understandable, they could not give thorough insights into the intricate relationships between factors in fraud detection. Consequently, more sophisticated approaches could be more suited for successfully identifying fraud trends. (see Figure 2.2.3)

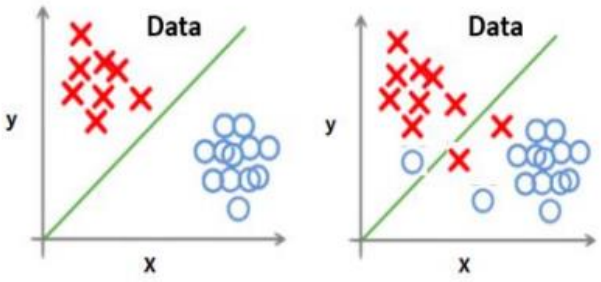

Figure 2.2.3. Linear Regression in imbalanced dataset

2.2.1.2 Logistic Regression.

Logistic regression is supervised learning algorithm that formulated to predict binary outcomes. It means this classification model is suitable for problems where dependent variables are binary like “yes” or “no”, “true” or “false” etc. The main calculation behind of the logistic regression is sigmoid function where it transforms linear regression equation to probability between 0 and 1.

Sigmoid function is S shaped and the illustration (Figure 2.2.4) demonstrates that σ (0) =0.5. Whenever $z>0$, the function's value gets closer to 1 and becomes the 1 category as z rises. The function value gets closer to 0 and z becomes 0 class, satisfying the conditions of the categorization above function. Following are some examples of how logistic regression classifies

[36] Bao, Y., Yang, S., Raghavan, V.V., & Srinivasan, A. (2017). Fraud detection in online advertising: A data-driven approach. Decision Support Systems, 100, 59-70.

data: By allocating regression coefficients (*w0, w1,..., wn*) to each variable, the characteristics of the data under investigation, indicated as (*x0, x1, x2,..., xn*), may be summarized. The sigmoid function, which can be written as follows, is then used to combine the obtained values:[37]

$$z = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

Where:

$$z = W^t\, x$$

Last formula uses the column vector x and the row vector w to represent the input data for the regression. The characteristics or variables of the data under analysis are contained in the column vector x, and the related regression coefficients are represented by the row vector w. As a result, when result is less than 0.5, it means regiression predict result as 0 and when it is bigger than 0.5, it means classifier take category 1.

Of course using of differnet algorithms in both supervised and unsupervised learning is depend on their performance in specific tasks and various of evaluation metrics are using in order to decide if model fits well or not and these evaluation metrics topic is going to be discussed in next sub-chapter. However,there are some common advantages of using Logisctic regression in Fraud Detection without taking into consederation one specific task. First of all Compared to linear regression, logistic regression is simpler to use and more effective when training. Then there is no presumption regarding the distribution of classes inside the input space in logistic regression. Additionally, it is a flexible technique that can manage a variety of data distributions without putting restrictions on the data and it may be quickly expanded to include additional categories. Last but not least the technique is quite effective in categorizing unidentified data. [38] In some of fraud cases such as finding anomalies in transactions , sometimes they show linear patterns and logistic regression might be useful like identifiying unusual high transaction amounts.

[37] Xiaonan Zou, Zhewen Tian ,Yong Hu, Kaiyuan Shen.(2019). Logistic Regression Model Optimization and Case Analysis. IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). pp 136-137.

[38] Hala Z Alenzi, Nojood O Aljehane. (2020). Fraud Detection in Credit Cards using Logistic Regression. International Journal of Advanced Computer Science and Applications. pp 549.

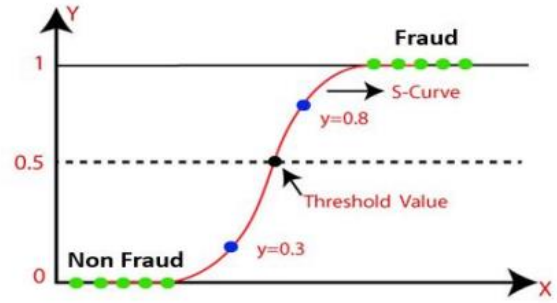

Figure 2.2.4. Sigmoid Function

For example, let have an example, for credit card fraud detection. After traning logistic regression model and all evalution metrics were applied, model is ready to detect fraud. When new credit card card transaction occur, model takes transaction's variables as training data and based on treshold that was chosen, model detect whether it is fraudulent activity or not. (See 2.2.4)

Tianyou Wang[39] suggest in her research Logistic regression may effectively used in credit card fraud detection case. Even if study used imbalance dataset, the great majority of observations are normal observations, that could attain excellent accuracy yet lack a great recall rate, hence logistic regression likely to assess most observations as normal observations. An enhanced method built on a random oversampling technique is called synthetic minority oversampling technology. Currently, this strategy is widely used to deal with data imbalances and is well-accepted in academia and business.

Label	Precision (Unbalanced/Balanced)	Recall (Unbalanced/Balanced)
0	1/0.94	1/0.99
1	0.85/0.99	0.54/0.93

Figure 2.2.5 Model performance before and after SMOTE

Research claimed several times that, SMOTE method should be applied to Logistic regression in order to balance dataset, instead of using traditional methods such as oversampling and under sampling. In order to prove SMOTE method's effectiveness, there were a comparison of the models where first model trained without SMOTE and for the second one SMOTE is

[39] Tianyou Wing, Zhao Yucheng. 2022. Credit Card Fraud Detection using Logistic Regression. BDİCN.

applied. Figure shows below that, 2nd model are capable to detect fraud better than 1st model since precision and recall to detect fraud increase 0.99 and 0.93 respectively. With the balanced data, the model could detect both labels accurately.

2.2.1.3 Support Vector Machines (SVM)

Support Vector Machines, an excellent supervised machine learning method, is utilized to address classification and regression issues. It performs particularly effectively if there isn't any path that could be used to divide the information into categories since this indicates that the data is unable to be separated into groups linearly. The fundamental idea behind Support Vector Machines is to identify a decision boundary or hyperplane that is capable of dividing the data into various categories. The model finds the support vectors—also known as the most important data values in the dataset—in order to perform the goal. The main goal of the Support Vector Machines method is to build a hyperplane which successfully divides the information being processed, and these support vectors are crucial to achieving this goal. Based on the closest values in the available data, the decision boundary is established using the support vectors. The fundamental objective of hyperplane-based algorithms is to minimize the margin between the actual data points and the predicted decision boundary. By optimizing the margin, these algorithms aim to achieve a better fit and generalization performance for the given dataset. In conclusion, the Support Vector Machine (SVM) is a potent method that discovers many uses in the field of machine learning, especially for problems like face recognition, fraud detection, sorting of texts.[40] Next formula expresses the decision function, which is an important part of this algorithm. The equation effectively expresses the fundamental ideas behind SVMs and makes it possible to apply them to classification applications:

$$f(x) = sng(x, w) + b$$

This formula is using to calculate the line dividing the data into two distinct categories. The input vector x in the SVM decision function formula is made up of weights and a constant term, indicated by the symbol b, which together help to determine the location and direction of the decision border. The SVM method optimizes the decision function to successfully classify new

[40] Sushant Kumbhar, Jaykishan Pandey, Ashish Lade, Prof. A. B. Ghandat. (2023). Support Vector Machine based Credit Card Fraud Detection. International Journal of Engineering Research & Technology (IJERT). Available : www.ijert.org.

examples by varying these weights and the bias term. To properly classify new instances based on their characteristics and their location in relation to the decision boundary, the SVM algorithm learns the optimal values of the weight vector, denoted as *w*, and the constant term, denoted as *b*, during the training phase. The SVM technique seeks to maximize the gap among the two categories by determining the optimum values of w and b. The algorithm's long-term objective is to widen the gap among the two groups.

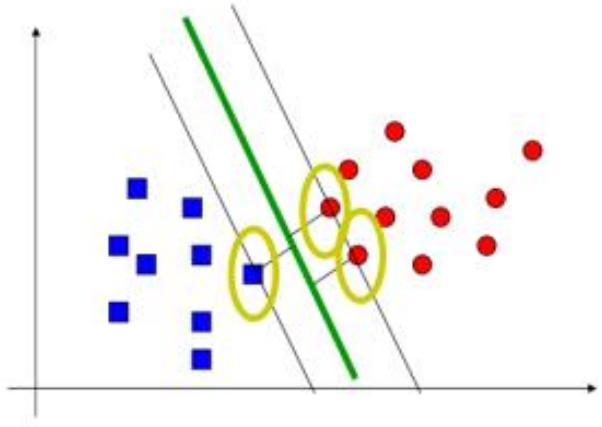

Figure 2.2.4 Hyper-Plane and how it does decide for two classes.

To identify the ideal hyperplane which divides the two categories, the SVM method calculates a parameter named the Margin, which is the gap across the two hyperplanes. The two groups are better categorized the wider the gap. The ideal decision boundary is represented by $H: y = w.\ x + b = 0$, and the SVM approach seeks to increase the margin in order to find it. Additionally, it makes use of the two hyperplanes $H1:\ y = w.\ x + b = +1$ and $H2: y = w.\ x + b = 1$. The margin is represented as 2 divided by the norm of the vector w, which is denoted as $\|w\|$. The boundary gets "soft" if the data is not completely separable, i.e., when there are multiple cases of overlap within each of the groups. This raises the possibility of categorization mistakes. The SVM algorithm seeks to enhance the margin in order to minimize these mistakes. The method seeks the optimum decision boundary that divides the two categories with the fewest mistakes through increasing the margin.

In terms of use of SVM on Fraud detection shows great performance, especially when dataset is imbalanced and skewed. The recent study conducted by several researchers[41] proves that SVM performs very good even it was compared to traditional techniques for fraud detection such as Naive Bayes, Decision Tree, KNN, and Logistic Regression. There are several reasons behind it. SVM can solve non-linearity problem by using kernel trick which bring data high-dimensional space. This issue could be problem for traditional machine learning algorithms. Additionally, it has huge robustness for outliers, since hyperplane maximize the margin between groups. Moreover, since the margin maximized as mentioned above, provide SVM model to find global optimum, and avoid the stuck in local minimum. Last but not least, naturally SVM has regularization parameter, it avoids overfitting problem. There all reasons, bring to SVM huge advantages. The research focused on credit card fraud and they balanced by downsampling the positive class and upsampling the negative class. The performance evaluated with the help of accuracy, specificity and sensitivity metrics. The results shows that all three metrics were higher than the above-mentioned traditional techniques. The study highlighted effectiveness of SVM algorithm in fraud detection, especially in imbalanced datasets.

2.2.1.4 Naïve Baye's classifier.

Another traditional supervised classification technique is Naïve Bayes which is based on Bayes formula that is mentioned in equation below. It is a hypothesis that is predicated on two tenets. The initial contribution of each feature in a particular entry that has to be bifurcated is equal. Furthermore, every single provided characteristic is statistically independent from one another, which means that we cannot infer anything regarding an attribute from the data that it presents. This could occasionally not be the case, and in these circumstances, the Bayes rule is applied to determine whether the claim is genuine or not. For example, the projected class would have the highest likelihood. [42]

This supervised learning classifier applies Bayes theorem that indicated below. The formula's naive Bayes classifier can determine the likelihood that the input data correspond to a

[41] Sheo Kumar, Vinit Kumar Gunjan, Mohd Dilshad Ansari,and Rashmi Pathak.(2022). Credit Card Fraud Detection Using Support Vector Machine. Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications.

[42] M. J. Madhuryaa , H. L. Gururaj a , B. C. Soundaryaa , K. P. Vidyashree a , A. B. Rajendraa.(2022) Exploratory analysis of credit card fraud detection using machine learning techniques. Global Transitions Proceedings.

certain category, denoted by A, by examining the values (input data) of a given collection of elements or parameters, denoted as B:

$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)}$$

The likelihood that incoming data will fall into each of the currently recognized categories must be calculated in order to classify it, and the category to which it most likely goes is chosen. The category a that has the greatest likelihood must thus be identified as given next formula, as bi represents one of the n observable features/predictors.

The likelihood that input data will be classified into each of the currently recognized classes must be calculated in order to classify it, and the category to which it most likely goes is selected. The category a that has the greatest likelihood therefore has to be identified as given formula below, where b_i is one of the n observed classifier: [43]

$$a = argmax_a\, Pab_1 \ldots\ldots b_n$$

As all variables are assumed to be independent by a naive Bayes classifier, just a little amount of training data is required to calculate the classification-related variables. Categories would stand in for faults or a collection of problems that a model may generate in the framework of fault diagnostics, and the classifier would stand in for the signs that the model is now exhibiting. Despite the fact that naive Bayes classifiers are simple and quick to develop, treating the predictors as independent variants is occasionally considered as a drawback of the algorithm because in the majority of real fraud detection situations, the signs could be interdependent.

There are some studies shows that Naïve Bayes is really powerful algorithm specially in real-world scenarios. Sai Kiran[44] and other researchers assume that Naïve bayes is this method is highly effective, quick, and accurate for situations of fraud detection. Even though Naïve Bayes is a relatively simple algorithm to implement. For example, Amit Gupta [45] conducted research about implementation of some machine learning algorithms in financial fraud detection. He used big data

[43] Gilberto Francisco Martha de Souza, Renan Favarão da Silva etc. (2022). Reliability Analysis and Asset Management of Engineering Systems. 6.3.1.1 Naïve Bayes classifier- pp
[44] Sai Kiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar etc. (2018). Credit card fraud detection using Naïve Bayes model based and KNN classifier. International Journal of Advance Research, Ideas and Innovations in Technology.
[45] Amit Gupta, Lohani, Mahesh Manchanda. (2021). Financial fraud detection using naive bayes algorithm in highly imbalance data set

that contain credit card transactions which is also imbalanced dataset with more than 30 features. According to training data, fraudulent activity was only 0.18% of total transactions. Training testing division implemented 70-30 proportion in respectively. After data pre-processing and train-test division 3 different model are applied. In comparison to existing algorithms, the naïve bayes model provides improved accuracy in credit card fraud detection. The primary method in the suggested work is the Naive Bayes algorithm, which classifies the dataset quite effectively. We can see how each method performed in terms of AUC ,Classification Accuracy, Precision, and Recall with the aid of Figure below.

Model	AUC	CA	F1	Precision	Recall
SVM	64.9	99.9%	59.7	56.4	63.4
Random Forest	92.8	100%	84.2	93.4	76.7
Naive Bayes	96.3	99.9%	73.6	67.9	80.4
Logistic Regression	91.1	99.9%	68.2	70.6	65.9

Figure 2.2.5 Comparison of Naïve bayes and all other applied algorithms

Moreover, Naïve Bayes cover more space in ROC curve, which is obvious that the Nave Bayes algorithm performs well in terms of labeling fraudulent transactions as such.

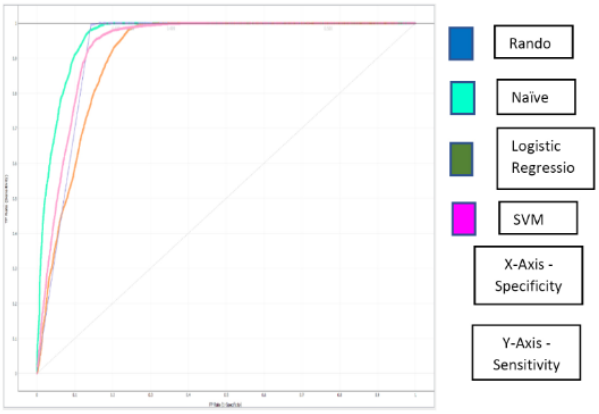

Figure 2.2.6 Comparison of Naïve bayes and all other applied algorithms via ROC curve

2.2.1.5 Decision Tree

Decision Tree is one of the most popular machine learning models not only for fraud detection but also for other problems. The model is constructed in the shape of a tree, as suggested by its name in which tree structure resembles a diagram, where each internal node indicates a test on a feature, every branch signals the test's result, and each leaf node (or terminal stores a class

label.[46] The root node (the amount in the figure), as can be seen in figure 2.2.6's simple decision tree graph, is the highest node in a tree. It is important to mention that the partition is learnt and shown as a tree in a hierarchical manner. The decision tree in the graph represents simple fraud detection algorithm. This is just simple example, and the logic works like: For example, if the amount of credit card payment is less than $500 and there were 3 or more payment were made in 1 day, and region code is 10267, thus "no fraud" will be the answer.[47]

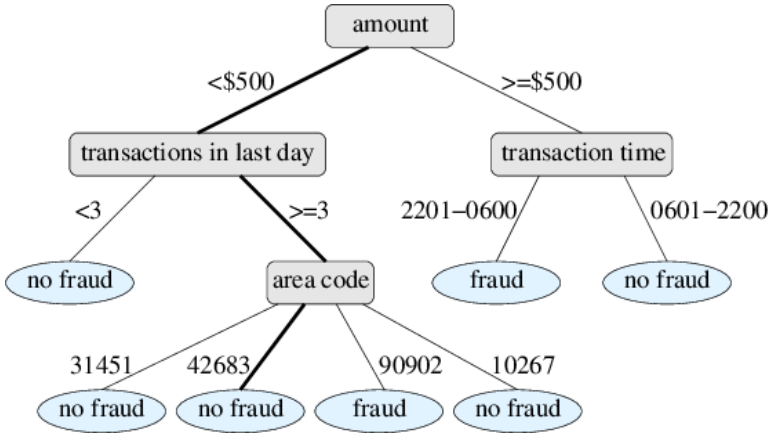

Figure 2.2.6. Illustrative decision tree for credit card

There are well-known methods for generating these decision trees from the trained data such as "ID3" and "C4.5". As it is mentioned above, the recursive splitting of nodes continues until a termination condition is satisfied. Splits often relate to a variable's value or an interval within where an actual values variable may fall (See Figure 2.2.6). The appropriate split for each node is normally calculated and selected as the most suitable one to divide a node's data into "most homogeneous" sets. This "homogeneity" is decided by the help of information gain or gain ration when the task is classification.

Usually, partitions are selected greedily according to a 1-step anticipation. Multi-step anticipation is frequently disregarded due to the increased computing expense, even if it could result in superior splits. However, a greedy method frequently results in branches with "informative" characteristics near the top, dividing the data towards more homogenous areas that

[46] Jiawei Han, Micheline Kamber, Jian Pei. (2011). Data Mining: Concepts and Techniques Third Edition. Morgan Kaufmann is an imprint of Elsevier.
[47] Shivaram Kalyan Krishnan, Deepthi Singh, Ravi Kant. (2014) On Building Decision Trees from Large-scale Data in Applications of On-line Advertising. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 669--678, ACM.

other characteristics may later turn discriminative. The trees grows and the nodes are spilling based on the homogeneity of the data's features and this splitting are going until most of the nodes became similar. These similar nodes decide the classification task. In case of regression analysis, tree take a mean of the nodes give result based on the that information.

A predictive decision support system called a tree maps out potential results based on various data. To create class models and using decision trees, many well-known classifiers are available. These classifications build a decision tree throughout the pruning process and remove any unnecessary subtrees to increase precision and prevent overfitting. The decision tree comprises two phases and relies on the detection of credit card fraud. Employing the available training data, a decision tree is first constructed, and then decision rules are used to categorize coming transactions. The input data for the decision tree includes labelled classes assigned to it, which include authentic or fraudulent. The model checks every amount and use variables to decide if there are fraudulent transaction or not. Every transaction is related to decision criteria of the tree. Depends on the transaction results, every transaction goes appropriate label. Otherwise, it alert higher risk, and then it would be easy to detect fraud. The method employs MLPC to regulate the trees' evolution and trim them. Recursive partitioning is carried out using the entropy value and a few chosen properties s. The process is finished when a set of requirements is satisfied.[48]

The decision tree approach offers the advantages of not requiring feature scaling, becoming resilient to anomalies, and dynamically addressing missing data. It solves classification and Fraud detection issues effectively and trains more quickly. [49] Additionally, it is generally accepted that decision tree models are computationally cheap.

However, of course there are some disadvantages of applying decision tree on fraud detection analysis. Specially, in order to detect fraud by decision tree, it require large amount of transactions (for credit card analysis) or training data. Then, using this large or training this large amount of data can case overfitting. [50]

[48] Aditya Joshi, Anuj Singh, Shikha Chauhan, Anupama Sharma. Decision Tree Algorithm for Credit Card Fraud Detection.
[49] Jonathan Kwaku Afriyie a , Kassim Tawiah a,b , Wilhelmina Adoma Pels a , Sandra Addai-Henne. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. (2023). Decision Analytics Journal
[50] Neli Kalcheva , Maya Todorova, Ginka Marinova. (2020). Naive Bayes Classifier, Decision Tree and Adaboost Ensemble Algorithm – Advantages and Disadvantages .

Attribute Selection Measures.

In the context of Decision tree, attribute selection is performing in order to split dataset (D) into smaller and homogenies partitions based on their categories. The purpose of this split is to have pure partitions that eventually they all have same category. This method guides how the nodes should be divided and splitting of the nodes depends on the score that attribute got during training. There are 3 main attribute selection techniques: information gain, gain ratio, and Gini index.

Information Gain. As we mentioned above the splitting attribute is determined to have the maximum information gain. The information expected to categorize the tuples in the resultant partitions is reduced by this attribute. Suppose D, the training set partition, consist of a data set of tuples with class labels. Let's say that the class label attribute, C_i *(for* $i = 1..., m$), has *m* different values that define *m* different categories. Next, the anticipated data required to categorize a tuple in D is provided by:[51]

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i).$$

Info(D) is the usual quantity of information required to determine the classification label of a tuple in D, commonly referred to as the entropy of D, and p_i is the non-zero likelihood that a given tuple in D belongs to the category C_i and is calculated by $|C_{iD}|/|D|$.

Let's say we need to divide the tuples in D based on an attribute A that has v different values, such as $\{a_1, a_2 \ldots a_v\}$. Following that, the anticipated data needed to categorize the tuple from D according to characteristic A is:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

The weight of the *j* th partition is represented by the phrase $|Dj| / |D|$. The anticipated information needed to categorize a tuple from *D* according to *A*'s partitioning is known as $Info_A^{(D)}$.

[51] Jiawei Han, Micheline Kamber, Jian Pei. (2011). Data Mining: Concepts and Techniques Third Edition. Morgan Kaufmann is an imprint of Elsevier.pp-279.

The term "information gain" refers to the differences between the initial information required and the current requirement.

$$Gain(A) = Info\ (D) - Info_A^{(D)}$$

The splitting attribute has been determined by which of attribute A's information gains is higher.

Gain Ratio. The information gain metric favors choosing qualities with many possible values, which indicates that it is biased in favor of tests with multiple results. Consider a property like product_ID, which serves as an individual identification number as an illustration. Here might be around the same number of partitions as their values, every one containing a single tuple, with a separating on product_ID. Every partition is pure, therefore the details needed to categorize data set D and using this method of partitioning could be $Info_{product_{ID}}(D) = 0$. As a result, segmentation on this property yields the most information, and it is obvious that such segmentation is insufficient for categorization. C4.5, the replacement for ID3, provides an enhancement to information gain known as the gain ratio as a way to address this partiality. Applying a "split information" metric defined as formula that written below, it normalizes information gain in some way.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right).$$

By dividing the data set, D, into v groups and assigning the outcomes of a test on parameter A to each group, the value indicated the potential information that may be obtained. Notably, it considers the percentage of tuples that include each result in relation to the total number of tuples in D that contain each result. It is distinct from information gain, that evaluates the categorization of newly obtained information on the basis of identical partitioning.

$$\text{Error(D)} = \text{GainRatio(A)} = Gain(\text{A}) \,/\, SplitInfoA^{(D)}$$

The splitting attribute is chosen based on its maximal gain ratio. But take note that the ratio gets unsteady when the split information gets closer to 0. To prevent this, a condition is imposed

that states that the information gain of the assessment chosen must be substantial—at least as substantial as the average gain across all tests reviewed.

Gini Index. Following a split across a certain property, the GINI index assesses the purity of a specific category. The purest split improves the sets that come from the separation in terms of purity. GINI is described as follows if *L* is a dataset with *j* distinct category labels:[52]

$$\mathrm{GINI(L)} = 1 - \sum_{i=1}^{j} \mathrm{p_i}^2$$

The relative frequency of category *i* in *L* is denoted by p_i . The following formula is used to determine GINI when the training set is divided into two groups, *L1* and *L2*, each of size *N1* and *N2*:

$$\mathrm{GINI_A(L)} = \frac{\mathrm{N_1}}{\mathrm{N}} \mathrm{GINI(L_1)} + \frac{\mathrm{N_2}}{\mathrm{N}} \mathrm{GINI(L_2)}$$

And then Gini is calculated as: $\Delta\mathrm{GINI(A)} = \mathrm{GINI(L)} - \mathrm{GINI_A L}$.

2.2.1.6 K-nearest Neighbor

K-nearest neighbor technique is widely used by scientist and researchers in order to find fraud detection and develop the detectors. Additionally, KNN demonstrated how effectively performs in supervised learning-based credit card fraud detection systems. According to the KNN systems, the newly generated issue will be categorized using this manner. The following three elements affect KNN's outcomes: [53]

1) The measurement of distance applied to identify the closest neighbors.

2) The K-nearest neighbor distance rule, which applies to the categorization.

3) The quantity of neighbors taken into account while categorizing the fresh values.

---

[52] Suryakanthi Tangirala. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. (IJACSA) International Journal of Advanced Computer Science and Applications

[53] N.Malini, Dr.M.Pushpa. (2017). Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection. 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics.

K-nearest neighbor delivers excellent accuracy rates without employing the prior distributional assumptions when we explore several supervised statistical pattern recognition-based approaches for fraud detection. For example, if credit card fraud detection applying KNN method, every receiving transaction will have its closest point to the new receiving transaction determined. The separation between two data values could be calculated in a variety of ways. Euclidean distance is employed to determine continuous characteristics, while an easy-matching coefficient is applied in the case of categorical characteristics. In case of multivariate data, the distance is determined for each characteristic separately and then merged. With the use of more accurate distance measures, the K-nearest neighbor technique may be improved. This strategy requires feeding either truthful and fraudulent instances for the purpose to train the data. This process is quick and produces few false alarms.

In order to understand how does algorithm work, let's have an example and discuss about Figure 2.2.7. Purpose of the classification to decide whether classification result is "minus" or "plus". As was previously indicated, a different number of neighbors may have a different outcome. Consider the case where the algorithm chooses one neighbor. As can be seen from the graph, "plus" is the value that is closest to the query point. If the number of neighbors were increased to 2 there would be two separate signs, and the result would no longer be accurate.

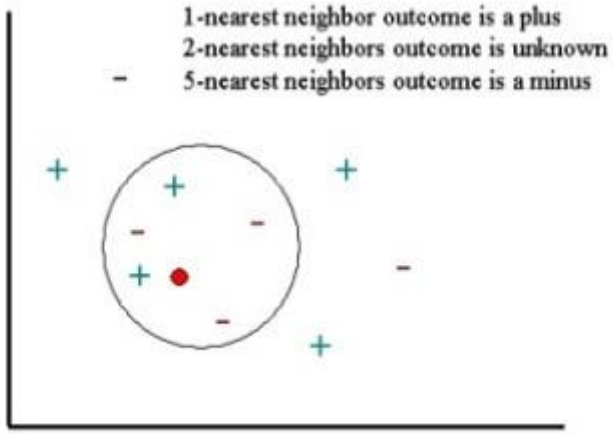

Figure 2.2.7. How does KNN work?

Next, the number of neighbors is raised to 5, and the response will be "minus" in 3 instances. As seen by the example, the outcome might vary depending on the number of neighbors. Finding the optimal number of neighbors is crucial.

KNN performs well as a fraud detection system, especially with a large dataset it works better than most of the well-known algorithms and accuracy of it is good enough to conduct this kind of analysis. Because, it is really effective in term of noisy data, since higher number of K values are useful in minimizing the impact of noisy data. It is accepted that for anomaly detection systems KNN works very quickly and effectively. In the KNN procedure, we categorize any incoming transaction by finding the closest point to the new transaction. If the next-door neighbor is fraudulent, the transaction will be shown to be positive alert. [54]

2.2.1.7. Random Forest

Due to the algorithmic simplicity and adaptability of decision tree systems to various data characteristic kinds, they have become extremely common in machine learning. The single-tree approach, nevertheless, could be susceptible to certain data and is simple to overfit. Ensemble approaches, which are better than solo classifiers, may address these issues by synthesizing a collection of separate choices. A variety of tree-based approaches are combined to create a random forest. Here each tree is reliant on a different unique random data whereas every tree has an identical frequency. The capability of a random forest can be affected by the association between various trees as well as the overall power of every single tree. The effectiveness of a random forest increases with the power of each individual tree and decreases with the association between distinct trees. Despite the fact that training data may contain certain cases that were incorrectly categorized, random forest nevertheless remains resistant to noisy values and anomalies.[55]

The Main benefit of applying random forest technique is that it could be used in either classification and regression problems. It is the most popular algorithm which offers more

[54] Rahul Powar, Rohan Dawkhar, Pratichi.(2020). CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

[55] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang. (2018). Random Forest for Credit Card Fraud Detection. IEEE

accuracy than any algorithms now existing. The study about Credit Card Fraud detection systems shows that [56] accuracy range of random forest is between 90-95%.

There are, however, certain difficulties as well. Because it includes multiple decision trees, the random forest algorithm is slower and more computationally expensive than the standard model. Additionally, because random forest needs more training data, it takes greater capacity. Furthermore, compared to decision trees, random forests can occasionally be challenging to comprehend.[57]

Researchers [58] conducted analysis and compered Random Forest with other 3 popular algorithm and try to find fraud in financial transactions. The methodology for analysis were same for all algorithms. The dataset was divided into 5 parts, first 4 for the training and rest for testing. In the testing phase cross-validation was used. Initially, the model was built with all features and got the error. After that some variables are exluded and only variables chosen which bring lowest error rate. (see Figure 2.2.8)

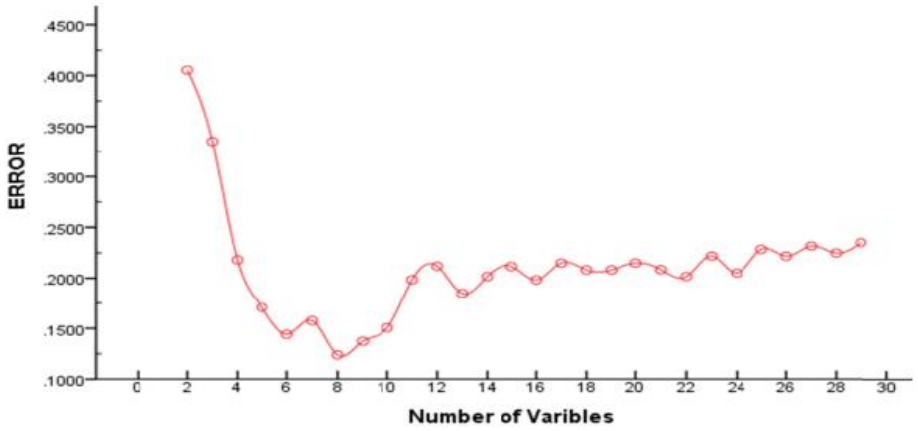

Figure 2.2.8 Random Forest five-fold across-test

Figure above proves that model with 8 independent variables performs better and bring more accuracy. Additionally in order to increase accuracy study showed the best parameter optimize and number of trees is decided to be not more than 500. The other algorithms also

[56] M.Suresh Kumar, V.Soundarya , S.Kavitha ,E.S.Keerthika , E.Aswini. (2019). CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM
[57] IBM. (2020) What is random forest?. Available: www.ibm.com
[58] Chengwei Liu , Yixiang Chan , Syed Hasnain Alam Kazmi1 & Hao Fu.(2015). Financial Fraud Detection Model: Based on Random Forest

developed and optimized. As a result, Random Forest performed better than Logistic regression, KNN and single Decision Tree. Thus, study compared two parametric and two non-parametric algorithms and results shows that classification algorithms have huge advantages over other algorithms. Essentially, Random Forest has the greatest effectiveness among practically all the algorithms. Furthermore, it handles more irregular data fields effectively by dispelling the notion that data should be normal. It can take a lot of high-dimensional data analysis and a difficult case of co-linear over-fitting to show up. Additionally, it can determine the relative relevance of each feature and effectively remove those that are not crucial. At last, we it decides which factors to combine to create the most effective models. The study proved that, non-parametric models in terms of accuracy are better than parametric models. Fitting non-normal data (generally, financial data doesn't follow normal distribution) with normal parameter models bring lowest recognition efficient, however parameter models successfully find the frauds. In Conclusion, non-parametric models show great results in financial fraud detection.

2.2.2. Unsupervised Learning

As it is mentioned under the "common fraud detection techniques" sub-chapter, supervised learning is a traditional method to combat fraud and anomaly detection. Unsupervised learning algorithms have been explored recently, and the impressive results indicate that these techniques should also be applied. Unsupervised learning is a common technique used by financial organizations nowadays to identify unexpected trends in transaction history and payment history, or by energy providers to identify unusual consumption of energy, etc. Unsupervised learning techniques come in a variety of forms, and we will learn about the most often used ones in this article.

2.2.2.1 Clustering Algorithms

Identifying grouping clusters in a data collection may be done using a wide range of approaches collectively referred to as clustering. Clustering we want to divide them into clearly defined categories in order to ensure the values in the same group are relatively identical to one another, whilst the values in various categories are very dissimilar from one another. Clustering means trying to divide these up into various categories such that the values within every single group are extremely comparable to one another, whereas the values in other categories are distinct from one another.

K-means. It is a very straightforward and efficient method for dividing a data collection into K unique, separate groups. Basically, *"K"* number of clusters is decided by the human as a first step and after that each data point will be included one of these clusters. According to the K-means clustering theory, successful clustering is a type where the variance inside of the group is as limited as feasible. Algorithm try to minimize distance between cluster center (centroid) and each observation inside of the cluster. There are several ways to measure of this distance, but the most common method is *"Euclidean distance".*[59]

In order to assess effectiveness of algorithm the method, the criterion function's error is squared and then summed. The grouping element is absent from a particular set of data X that only has a descriptive property. Find the first group by determining the *X* first clustering midway is " $\{C_1, C_2 \ldots\ldots\ C_k\}$", the time following the end of every type of subgroup equals $\{X_1, X \ldots\ldots\ X_k\}$", according to $C_i\ \varepsilon\ X_i$ *(i = 1, 2... k).* There are a total of n_1, $n_2 \ldots n_k$ observations in every cluster subgroup, and every cluster subgroup's average value is$\{\ m_1, m_2 \ldots m_k\}$; comparing these values to the original cluster middle point, which is cl' $\{\ c_1, c_2 \ldots m_k\}$.The clusters centre will be fixed if the two assessments don't vary. If not, give the centre an average amount and carry on with the subsequent iteration. The formula will be as: [60]

$$E = \sum_{i=1}^{k} \sum_{p \in X_i} \|p - c_i\|^2$$

Analysis is going to be over if function is not changing significantly anymore, if not iteration is continuing to minimize the change. K-means is really powerful algorithm in fraud and anomaly detection specially when there is need for quick insight. It helps to find hidden patterns. However, when there is unbalanced dataset, k-means may have some struggle.

K-means could be used to find financial fraud, let's take a look how it could be applied credit card fraud detection. [61] In the example, there are categorical and numerical variables available. First of all, one hot encoding method was applied in order to get equal weight for

[59] Gareth James , Daniela Witten , Trevor Hastie Robert Tibshirani.(2013) An Introduction to Statistical Learning with Applications in R. Springer.

[60] Guojun Shi, Bingkun Gaol , Li Zhang.(2013). The Optimized K-means Algorithms for Improving Randomly-initialed Midpoints. 3 2nd International Conference on Measurement, Information and Control.

[61] Navin Kasa, Andrew Dahbura, Charishma Ravoori, Stephen Adams. (2019) Improving Credit Card Fraud Detection by Profiling and Clustering Accounts. University of Virginia

categorical variables. For the numerical variables min-max scaling applied for the same reason The variation of numerical characteristics is reduced by the min max scaling strategy to a range between 0 and 1. We have already discussed how does K-means works and distance metrics. And in order to decide the ideal number of 'K' plot is visualized where it shows sum of squared errors versus number of clusters. (See Figure 2.2.9) The perfect situation is one in which the total of clusters changes decreases dramatically up to an appropriate number of clusters ("K"), beyond which it becomes unimportant.

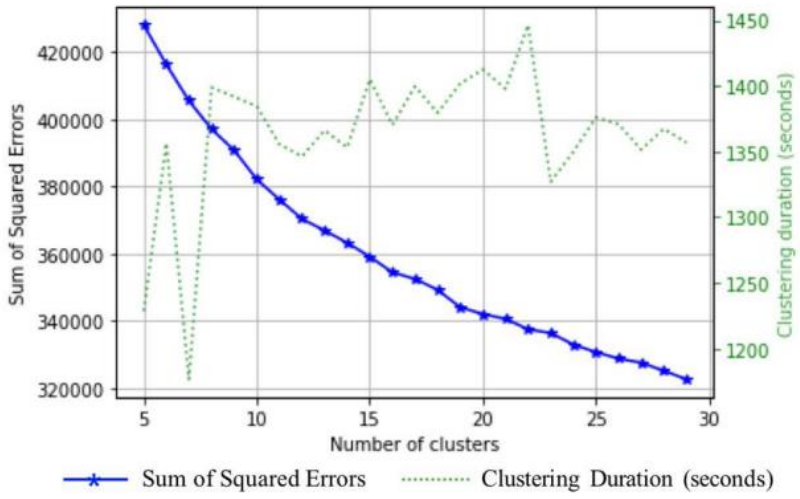

Figure 2.2.9 Sum of squared errors by number of clusters.

After calculation it is decided to continue with 10 clusters, because the SSE curve down steeper until 10 and after 10 clusters amount of time that required for training is increasing which is computationally expensive. So, making tradeoff in 10 clusters will make a difference. The accounts in each cluster were examined, and certain client clusters were discovered. Further research must be done on these consumer categories to determine how the bank could enhance its variable set or cluster allocations. This should enable poor clusters to see an improvement in the effectiveness of the model. As mentioned before, new customer clusters like "Low Spender", "Frequent spender"," Loyal Spender "and 7 more groups were found. The fraud percentages were significantly different among groups in each sample, even when the average AUC doesn't vary. Certain groups with a greater fraud rate nevertheless observe an apparent boost in model AUC outcomes in comparison to the basic model. (see Figure 2.2.10)

FRAUD RATES ACROSS CUSTOMER GROUPS

Customer Groups	Average Fraud Rate	Standard Deviation
Global Spenders	6.36%	0.79%
Diverse Spenders	1.21%	0.11%
Low Spenders	2.17%	0.80%
Frequent Spenders	1.64%	0.33%
Big Spenders	4.74%	0.08%
Loyal Spenders	2.20%	0.74%
General	1.54%	0.30%
Local Spenders	1.34%	0.05%
Big and Infrequent Spenders	2.95%	0.08%
Glocal Spenders	2.93%	1.10%

Figure 2.2.10.

Then, for each cluster, Random Forest and XGboost models were used. However, there isn't a significant difference seen when we compare the AUC performance of the base model to the AUC of each cluster to which the abovementioned methods are applied. However, several clustering performs very well like "Big Spenders" cluster shows 0.003 AUC increase, which is important to detect high-value fraud transactions. Overall, K-means help to understand customer groups and their behavior, and suggest really powerful insight for fraud detection.

Hierarchical Clustering. There are two subcategories of hierarchical clustering are agglomerative and divisive hierarchical clustering.[62] This subcategory is determined by if the algorithm bottom-up or top-down. An inclusive cluster appears at the highest point of the multilayered series of divisions created by hierarchical approaches, while singleton clusters of distinct items appear at the bottom. Every stage could be thought of as joining two clusters from the level below or dividing one from the level above. A dendrogram is an illustration of the outcome of a hierarchical clustering method. The problem with hierarchical approaches is that whenever a merging phase or a division stage, it is not possible reverse it.

A bottom-up technique is used in the *agglomerative hierarchical* technique. Usually, technique begins by allowing every item to establish a distinct cluster before repeatedly merging smaller clusters to bigger ones unless all of the items are in one group or when particular end requirements are met. The hierarchy's root is the one group. In the combining process, it identifies the two groups which are most identical to one another (based on certain metric) and joins together

[62] Yogita Rani[1] and Harish Rohil. A study of Hierarchical Clustering Algorithm. International Journal of Information and Computation Technology.

to create a one group. An agglomerative technique needs a maximum of *n* rounds since two groups are combined every iteration, and every group includes a minimum of one item.

A top-down approach is used in a procedure called divisive hierarchical clustering. All items are first gathered into a single group, serving as the hierarchy's base. The root is then split down into multiple tiny sub-groups, which are subsequently iteratively divided into smaller groups. When every group at the ground level is significantly coherent—either containing just one item or the items found inside the group are significantly close to one another—the splitting procedure is repeated. [63]

This method is not commonly used in fraud detection, but it is still used in anomaly detection in order to discover unusual patterns. Primary reason of not using this method is that it is computationally expensive and since fraud pattern is changing over time, it is really hard to detect with computationally expensive. It is generally accepted that this clustering model is used for feature extraction, anomaly detection and granular insights.

2.2.2.2 Auto-encoder

A combination of encode and decode stages are used by an auto-encoder (AE) to train it to relate feed to outcome. In order to rebuild the parameters, the encoder is mappings through the data entered to a hidden layer and the decoder is mapping through the hidden layers to the outcome layer. The information that is entered is represented in low-dimensional as well as irregular ways within the hidden layers. As seen below, the Auto-encoder is created:[64]

$$\hat{X} = D(E(X))$$

Here, X denotes the data that was entered, E denotes an encoder, D denotes a decoder, where X is the raw data that has been rebuilt. The algorithm's goal has to as closely imitate the distribution of X as it can. It may be thought of as an answer to the aforementioned optimization issues in specifics:

$$min_{D,E} \|X - D(E(X))\|$$

[63] Jiawei Han, Micheline Kamber, Jian Pei.2015. DATA MINING: CONCEPTS AND TECHNIQUES/ 3RD EDITION. Morgan Kaufmann is an imprint of Elsevier

[64] Xuetong Niu, Li Wang, Xulei Yang. (2019). A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised

Here we have two condition is frequently used as || • ||. A complex auto-encoder with numerous layers—multiple pairings of encoders and decoders—allows for the modeling of complicated X distributions. The benefit of technique over gradient descent is that it uses a dynamic average of the momentum, which enables the Adam method to employ larger and more powerful iterations while not requiring for fine-tuning[65]. This method's expensive calculation is its biggest drawback.

2.2.2.3 Principal Component Analysis

Principal Component Analysis is an effective approach that enables us to acquire a comprehensive understanding of correlations between many credit card transaction variables with just a few computations. Its adaptability could be shown in the reality that it is capable of working on very big data sets, regardless of their features or dimensions, which is crucial for this issue. A statistical technique called PCA converts variants that are correlated into uncorrelated ones. This approach seeks to express purchases specified by various characteristics (purchase amount, range, etc.) in the current situation in a more compact space than the first one, losing the smallest quantity of data feasible.

There are frequently insufficient fraud instances to practice on in order to detect fraudulent transactions. But you can have a lot of instances of successful transactions. The solution is provided by the PCA-Based Anomaly Detection component, which examines the characteristics at hand to decide what makes a "normal" class. After that, the component uses distance measures to find instances that are anomalies. Using existing unbalanced data to train a model is possible using this method.[66]

The research shows that[67] PCA performs very well in terms of fraud detection. Having reasonable accuracy, Principal Component Analysis immediately categorizes the transactions and has the ability to spot newly emerging fraudulent behavior. Principal Component Analysis proves more versatile while also providing a comprehensive perspective of the relationships between many qualities. However, the possibility of achieving a "local" as opposed to a "global" optimum

[65] Y. Bengio,(2012) "Practical recommendations for gradientbased training of deep architectures," Neural networks: Tricks of the trade, pp. 437-478.
[66] Microsoft Learning.(2021) PCA-Based Anomaly Detection component.
[67] Maria R. Lepoivre, Chloé O. Avanzini, Guillaume Bignon, Loïc Legendre, and Aristide K. Piwele.(2016). Credit Card Fraud Detection with Unsupervised Algorithms. Journal of Advances in Information Technology Vol. 7, No. 1

still exists. This risk might be decreased by repeatedly running the "k means" procedure with various beginning clusters, although doing so would lengthen its runtime.

2.3 Evaluation Metrics for the Machine Learning algorithms

One of the crucial phases in creating a successful machine learning model is evaluating its performance. Various metrics—also referred to as performance metrics or evaluation indicators—are applied to assess the effectiveness or quality of the model. These indicators of performance enable us to evaluate how effectively the model handled the supplied data. By adjusting the hyper-parameters, it could make the model function better. Metrics for performance assist measure how well a machine learning technique applies on new or previously unexplored data. Here we are going to explain the most common evaluation metrics such as Classification Accuracy, Confusion Matrix, Area under Curve, F1 Score, Mean Absolute Error, Mean Squared Error.

When machine learning model is developing data is divided into two parts, called training and test sets. The ML system's mistakes may be analyzed using a portion of the training set, and the system's parameters are able to be tuned in response. The learning performance and efficiency of the model is calculated this testing phase. The k-fold cross validation approach is applied when data is limited. It involves splitting the data into *k* folds, using the first *k-1* folds for training, while the last fold, the 1-fold, for assessments. The folds change till each fold have been trained and tested on the rest of the k-1 folds, after which a mean is acquired. The most common cross validation method is using 10-fold.

Precision and recall are both criteria employed to assess how well a retrieval system is working. According to formula that is written below, *precision* is calculated as the entire number of accurate cases divided by all other cases. *Recall* is calculated using second formula is written below, which divides the total number of accurate cases by the number of right occurrences retrieved. Examples of cases include individual words or entire documents that were pulled from a corpus of documents. However, first of all confusion matrix should be explained in order to understand abovementioned evaluation metrics.

		Predicted annotation	
		Positive	Negative
Gold annotation	Positive	True positive (tp)	False negative (fn)
	Negative	False positive (fp)	True negative (tn)

Figure 2.3.1 Confusion Matrix

Confusion matrix is a chart that shows how values are expected and actual. Now it is time to share formulas for abovementioned evaluation metrics. The terms "*precision*" and "*recall*" are defined here:

$$\text{Precision :P} = \frac{TP}{TP+FP}$$

$$\text{Recall: R} = \frac{TP}{TP+FN}$$

The harmonic average of the model's precision and recall is known as the F-score, which is a method of combining the model's precision and recall. β represents weight function here:

$$\text{F-score: } F_\beta = (1+\beta^2)*\frac{P*R}{\beta^2*P+R}$$

$$\text{F-score: } F_1 = 2*\frac{P*R}{P+R}$$

During the computation, Precision makes use of every document obtained. There can be a chance to simplify the computation once there are many documents by employing precision at the threshold amount. After there is a confusion matrix and total number of transactions in the dataset, we can have Sensitivity and Specificity. The percentage of negatives which are accurately classified as negative or as lacking the ailment is known as specificity. The percentage of negatives that are correctly detected (example: the percentage of non-fraud transactions which have been correctly recognized as not having the ailment) is measured by sensitivity, which is similar as recall. Another metric called *accuracy* is the percentage of accurate instances—both positive and negative—that are found among all the results. [68]

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

[68] H. Dalianis, Clinical Text Mining. (2018). The Author(s). Chapter 6.

Precision and inverse precision are weighted arithmetic means that makeup accuracy. These evaluation metrics are crucial part of the machine learning. Without them it is impossible to choose the best model for the specific situation. In next chapter, we will have quantitative test, where we are going to use these metrics for model selection.

Chapter 3. Quantitative Analysis Fraud Detection.

3.1 Research and methodology

This section provides an explanation of how the different machine learning models might be effective and efficient by using real case scenario. Quantitative analysis plays vital role here because we are going to dive in Fraud detection deeply and try to investigate unusual patterns, anomalies and specially evaluation metrics of some important algorithms that we have already explained in second chapter. The analysis will cover credit card fraud which is the most common fraud type in last decade. In order to prevent customers from being charged products that they haven't bought, banks must be able to identify fraudulent transactions. Analysis will provide deep insight about dataset and results with help of different graphical representations. The comparison of different machine learning models in fraud detection and to find the most effective model for the credit card fraud detection are key objectives of the research.

Dataset. The dataset contains credit card transactions made in September 2013 by cardholders across Europe and it was gotten from Kaggle. In a dataset with about 280,000 transactions, there are about 500 fraudulent transactions. Because of the dataset's extreme imbalance, it is challenging to identify fraudulent transactions. In the dataset, "Class" is the target variable, while all other variables are principal components found using principal component analysis aside from "Amount" and "Time". Fraudulent transactions are indicated by a "1"; otherwise, transactions in the "Class" category will be "0." [69]

Methodology and Research Questions. The analysis follows an approach that is generally accepted: first, data pre-processing and cleaning will be applied. Following the exploratory data analysis (EDA), a model will be built and evaluated. In order to interpret all of the results of the study, we will choose the best approaches. All analysis will be conducted with a help of Python programming language and the machine learning algorithms like logistic regression, decision tree, SVM, random forest will be used, and of course the effectiveness of each model will be discussed. In order to increase evaluation metrics coefficients, some methods are going to be applied.

Research Questions:

1. What are the key patterns and characteristics of fraudulent transactions?

[69] Dataset: Credit Card Fraud Detection. available: kaggle.com

2. Which metrics should be used for credit card fraud detection?
3. What is the effectiveness and performance of the different machine learning models?

3.2. Model development and evaluation

In this phase of research, the machine learning models are going be built and shown in order to answer research questions. The dataset and it's first characteristics were already mentioned above, however here we are going to depth and explain dataset's inner problems and features. After pre-processing and EDA part, some modelling part come alongside.

Pre-processing and EDA. The dataset is quite clear in terms of missing values and NULL variables. However, it was identified that there are some number of duplicated observations and definitely, it should be handled. But most important problem of the dataset is the balancing. Since we have only 492 frauds over 280,000 observations, it is only 0.17% of dataset. Which makes the dataset imbalanced.

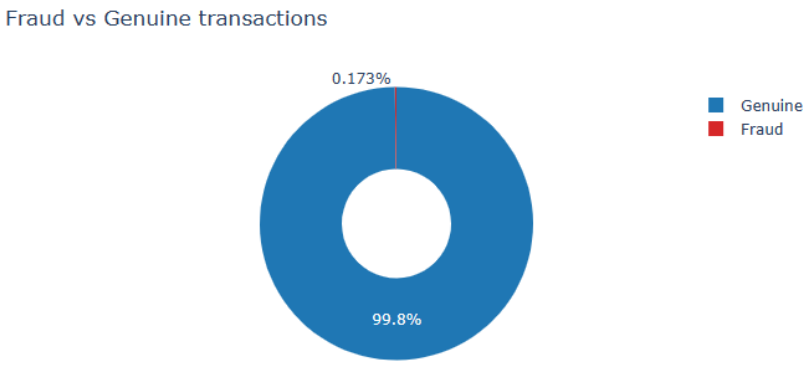

Figure 3.1 The imbalanced dataset

With the blind guess can identify 98% of Genuine transactions, so imbalanced dataset somehow dangerous for modelling. Let's keep it in mind, and before modelling some methods (Such as SMOTE) will be applied to balance it. Moreover, before starting to analyze data, 100 genuine and fraudulent transactions (for each class) are randomly selected and saved as a new data frame. This selected data frame is deleted from all further analysis until the final testing of the model.

At the same time, we tried to identify outliers and increase the accuracy for models that we will have later on. But when outliers are removed, it seems that the dataset lose half of fraud observation which is quite huge number and then we decide to skip dropping the outliers.

In order to get insight about feature importance and feature correlations with target variables, it is important to visualize correlation matrix and other graphs. It is clear that some variables such as V17, V14 and 10 more variables are somehow correlated with Class.

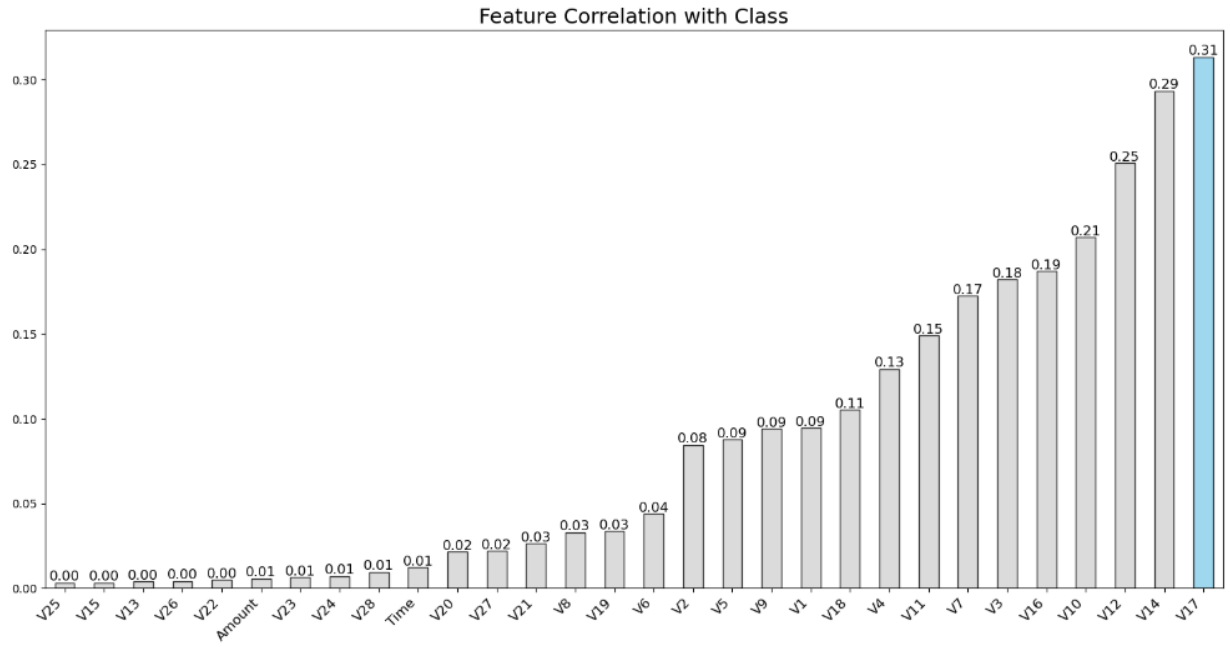

Figure 3.2 Correlation with Class.

We also did correlation matrix (see figure 3.2 and 3.2.1) in order find some relationship between not only with target variable but also, with other variables. Strong correlation might indicate potential patterns and relationship. It was seen in the correlation matrix that, beside target variable, there are variables such as amount and time which have some important correlations, and this information might be helpful in feature selection. Interestingly, it is seen in correlation matrix that was visualized during analysis, the most correlated features are negatively correlated with class. For example, the score of V17 is -0.31, the score of V14 is -0.29 and the score of V12 is -0.25.

In order to get known the dataset more in depth, the clustering technique might be useful for the analysis. However, before moving on there, the all variables should be normalized which is good practice to ensure the stability, efficiency, and effectiveness of the model later. In clustering the most important 12 variables are selected for creating clusters based on correlation matrix that shows the relationship between class and rest of the variables. But for modelling, only the most important 9 variables are selected. (See figure 3.2 and 3.2.1)

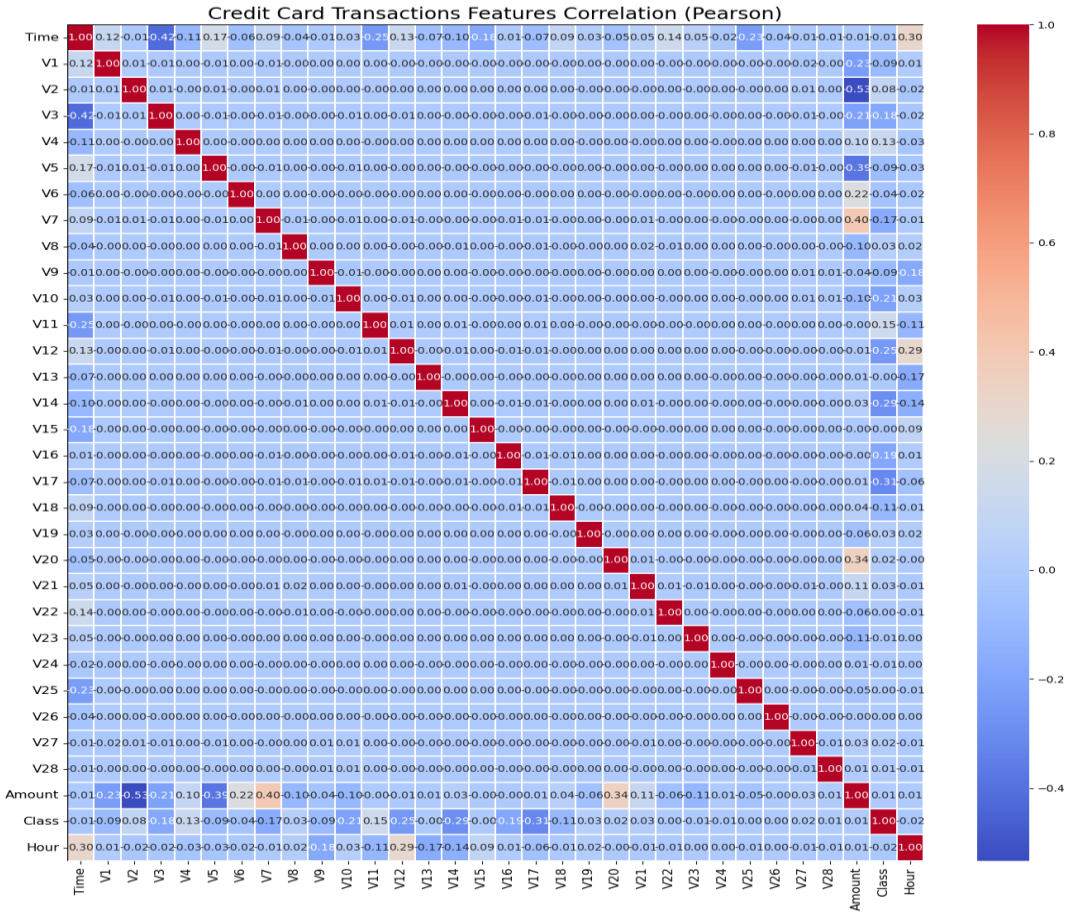

Figure 3.2.1 Correlation matrix

K-Means clustering is applied and number of clusters are defined by Elbow method, which is one of the well-known techniques. The figure 3.3 shows that there is significant reduction in WCSS occurs between 4 and 5 clusters. This could be the "elbow point" and that is why we have chosen 4 as a cluster number. As a results, we get 4 different clusters based on Fraud class. In 3rd and 4th clusters share almost 80% of the Frauds from dataset with 330 and 100 fraudulent transactions. The 1st and 2nd clusters contain only 5 and 38 respectively. (See Figure 3.4).

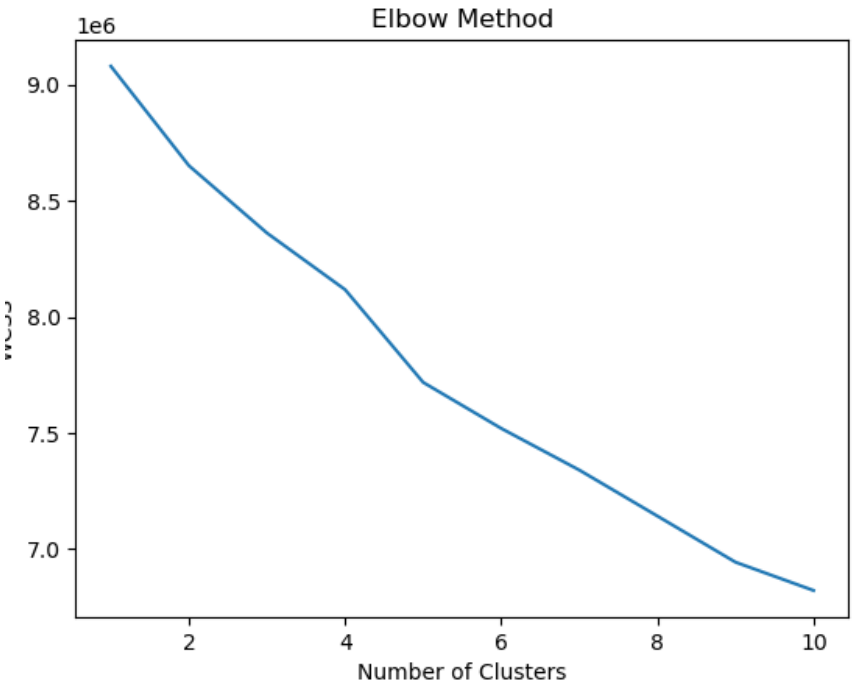

Figure 3.3 Method to define number of clusters in K-means

It means that, it is also possible to build model based on clusters instead of train whole dataset. Since we have applied only 12 variables for the clustering, we can group somehow the fraudulent transactions in certain clusters. It is very useful technique for segmentation of the frauds and exploratory analysis. At the same time, it is possible to take one cluster which contains most of the frauds and try to build model based on the data that already grouped. Clustering may also show and uncover the patterns in data that might be hidden. In conclusion, instead of using clusters for modelling in our case, it would be better if there will be ANOVA score for each variable and compare each model based on training data where features differ based on correlation matrix or ANOVA test. With a help of this comparison, it would be easy to decide whether correlation matrix or ANOVA score perform better in feature selection.

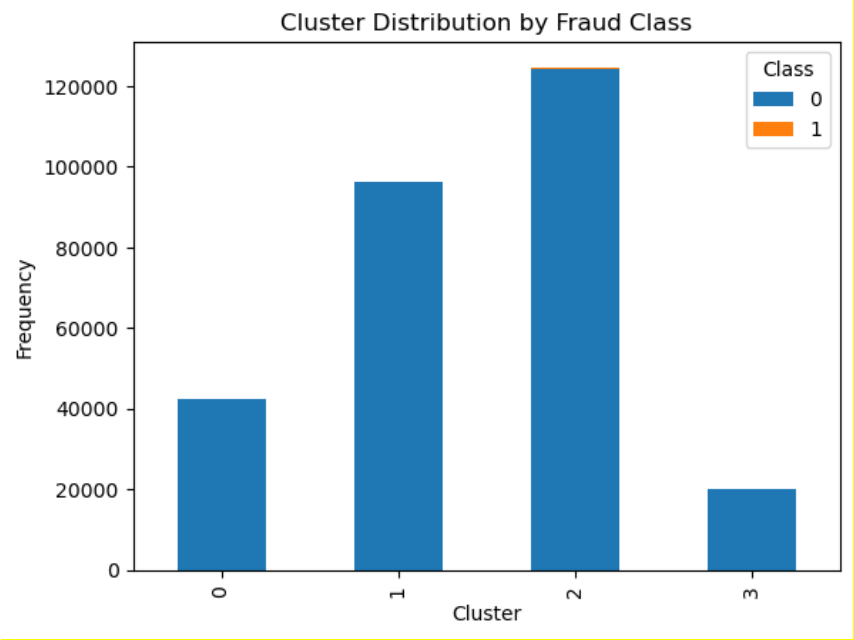

Figure 3.4. Clusters by Fraud Class

Before moving to modelling part, there are still some steps left. First of all, there are still two features called – "Amount" and "time" left where the standardizing hasn't applied. We applied robust scaler function in order to aid in improving the performance of models by facilitating the models' ability to understand patterns and correlations in the data without being unduly impacted by the features' different sizes.

Lastly, ANOVA score is calculated for each variable and visualized in order to see better understanding (See Figure 3.5). ANOVA is one of the common methods for feature selection. This procedure can assist in determining which variables have a major impact on the target variable and may help model perform better. Variables with high F-statistics and low p-values regarded as significant predictors. [70]

[70] Tianyi Zhao , Yingzhe Zheng , Zhe Wu. (2023) Feature selection-based machine learning modeling for distributed model predictive control of nonlinear processes. Computers & Chemical Engineering.

The relevance of that characteristic in relation to the target variable is higher the higher the ANOVA score. We will exclude features with values lower than 50 based on the plot above. In this instance, we'll build two models using characteristics from the correlation plot and ANOVA score.

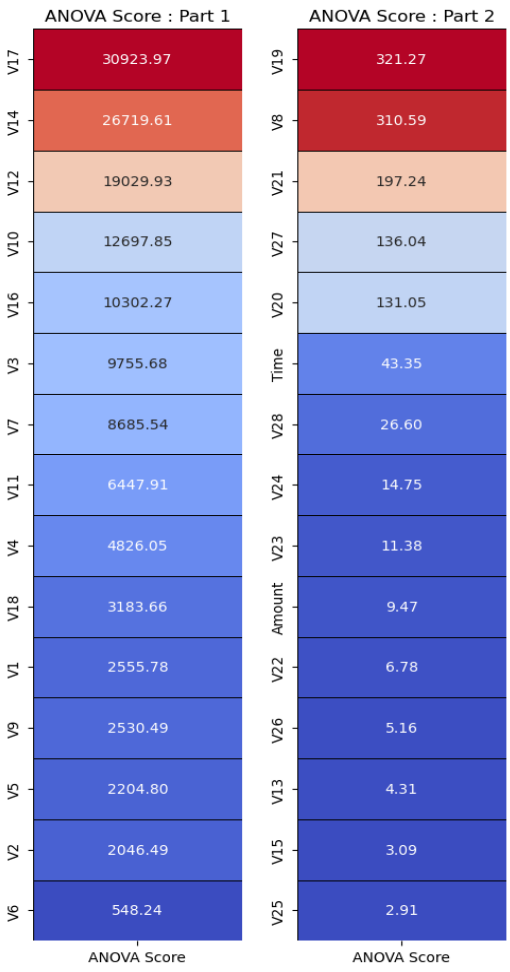

Figure 3.5 ANOVA score for each feature

Since the data is so imbalanced as it was mentioned above, it would be irrelevant if we start modelling without balancing it. It is already explained in 2nd chapter that if there is imbalanced dataset, there are some methods such as oversampling, undersampling and SMOTE methods to balance dataset in order to get right accuracy. [71] Here SMOTE method is applied, for the minority class, SMOTE creates synthetic samples, which can aid in capturing the fundamental distribution of the data. Better generalization and model performance may result from this. When the minority class is dispersed thinly, it can be very effective. After applying this method, Fraud cases reached to roughly 2365 whereas, non-fraud cases decreased to 4370.

Model development and evaluation. Here different models such as logistic regression, SVM, decision tree and random forest based on features that were obtained and decided by correlation matrix and ANOVA score. Dataset divided into train and test 20% and 80% proportions respectively.

Logistic Regression. First model is very well-known classification method and it is fed by training data that was decided by correlation matrix. The features 'V3', 'V4', 'V7', 'V10', 'V11', 'V12', 'V14', 'V16', 'V17' are base of the first model. As it can be seen from Figure 3.6, first model performed well, and overall accuracy is around 95%. However, it is important to keep it in mind

[71] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall ,W. Philip Kegelmeyer. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research.

	precision	recall	f1-score	support
0	0.94	0.98	0.96	760
1	0.96	0.87	0.92	359
accuracy			0.95	1119
macro avg	0.95	0.93	0.94	1119
weighted avg	0.95	0.95	0.95	1119

Figure 3.6 Logistic Regression with correlation matrix-based features

that accuracy not always shows best results, especially cases like fraud detection. So, precision[72] , recall and f1-score might be useful. Here precision ratio shows correctly predicted positive values to the total predicted positives. In our case it represents percentage of transactions which model correctly identify fraudulent transactions that are actually fraudulent. It shows us here, model 99% identify fraudulent transaction which is perfect result. Recall is representing the percentage of correctly predicted positive observations to the total actual positives. In fraud detection, it demonstrates the model's capacity to successfully detect all fraudulent transactions. So, it is not high as precision, but relatively good result for logistic regression. F1-score is just harmonic mean of precision and recall. Result shows that, model's ability to detect non-fraudulent transaction is slightly higher than detection fraud by seeing F1-score. Additionally, ROC-AUC curve (See figure 3.7) for the first model shows that logistic regression performs well with features that selected based on correlation matrix. With the 98% of AUC indicates that the model can effectively differentiate positive and negative observations and makes classification process more precise. When AUC close to 1, it means that the model successfully distinguishes between the classes. An overall assessment of performance across all potential categorization criteria is provided by AUC. AUC may be seen as the likelihood that the model values a randomly chosen positive example higher than a randomly chosen negative example.[73] And here we could say that, it is pretty high.

[72] Pasi Fränti a, Radu Mariescu-Istodor b. (2023) Soft precision and recall. Pattern Recognition Letters
[73] Google Developer. Machine Learning course. Available: developers.google.com/machine-learning

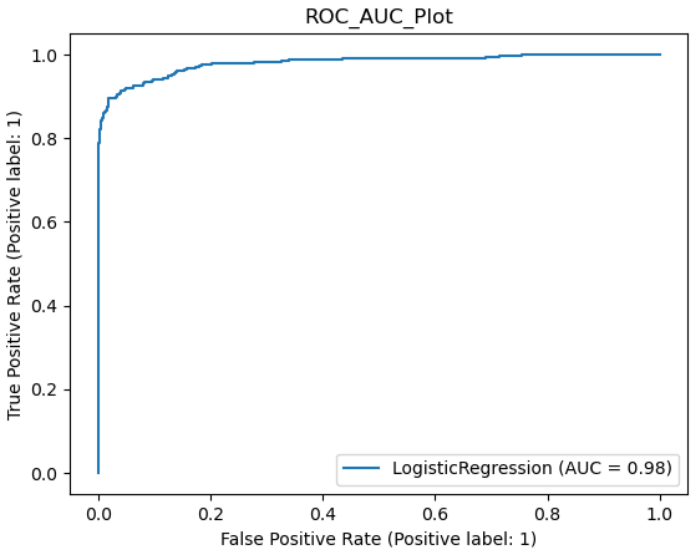

Figure 3.7 ROC-AUC plot for first Logistic Regression

Confusion matrix shows True positive rate which is important to catch fraudulent values and our model almost identify all fraudulent transactions. At the same time True Negative which represents non-fraudulent transactions are identified very well when we compare it with False Negative where it is less than 5% (See 3.8).

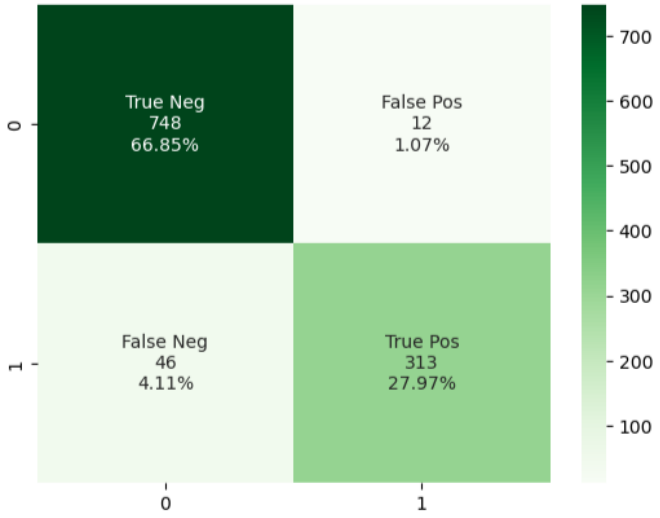

Figure 3.8 Confusion Matrix for Logistic regression.

Let's move on the model with features that selected based on ANOVA score and build logistic model. Logistic model with these features performed well and got the higher accuracy (100%) in all evaluation metrics (See 3.9)

	precision	recall	f1-score	support
0	1.00	1.00	1.00	961
1	1.00	1.00	1.00	458
accuracy			1.00	1419
macro avg	1.00	1.00	1.00	1419
weighted avg	1.00	1.00	1.00	1419

Figure 3.9 Logistic Regression model with training features-based ANOVA score

It is clear that these training that got from ANOVA score plot brought the analysis the overfitting. In order to avoid this overfitting, we may increase our sample size or just simply drop feature selection method with ANOVA method. Even though we had included Ridge Regression (l2 penalty) in the logistic model, the outcomes were not significantly affected. As a result, we chose characteristics that are more than 100 rather than 50 as our next approach, improving the feature selection method using ANOVA. As a result, features like "scaled_Time," "V27," "V21," "V8," "V19," and "V6" are no longer present. Overfitting was finally avoided throughout the testing phase, and the results were comparable to those of the primary logistic model.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	1437
1	0.97	0.88	0.93	692
accuracy			0.95	2129
macro avg	0.96	0.94	0.95	2129
weighted avg	0.95	0.95	0.95	2129

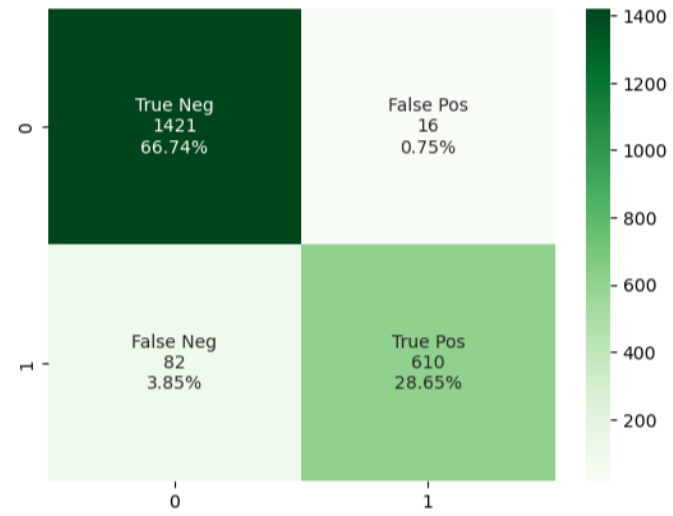

Figure 3.9.1 Logistic Regression model with training features-based ANOVA score (Overfitting avoided)

Support Vector Machines (SVM). It is interesting that SVM is able to handle imbalanced dataset by special class weight adjustment or cost-sensitive learning. Since here the dataset has already been balanced by SMOTE method, we just try to apply same features that were already used in logistic model. As we have done already, correlation matrix features were used first.

	precision	recall	f1-score	support
0	0.93	0.99	0.96	760
1	0.97	0.85	0.91	359
accuracy			0.94	1119
macro avg	0.95	0.92	0.93	1119
weighted avg	0.95	0.94	0.94	1119

Figure 3.10 SVM with features based on correlation matrix

It is clear that, this model is also performing very well, but all instances are slightly lower that logistic regression. However, it still shows great result to detecting fraudulent transaction with 0.97 Precision.

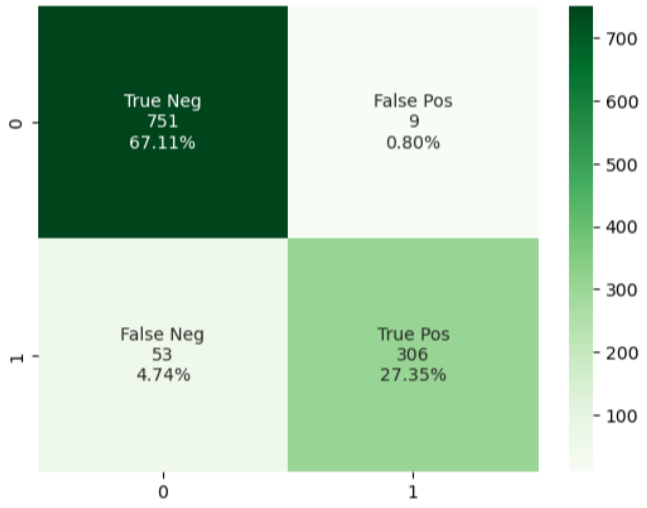

Figure 3.11 SVM confusion matrix with correlated based features

Confusion matrix of SVM model shows relatively good result to logistic one, with slightly difference. Model almost catches all fraudulent transaction with 0.8 False Positive rate where only

9 transactions belong. Hence True Positive rate is almost 1% less than what logistic regression model shows.

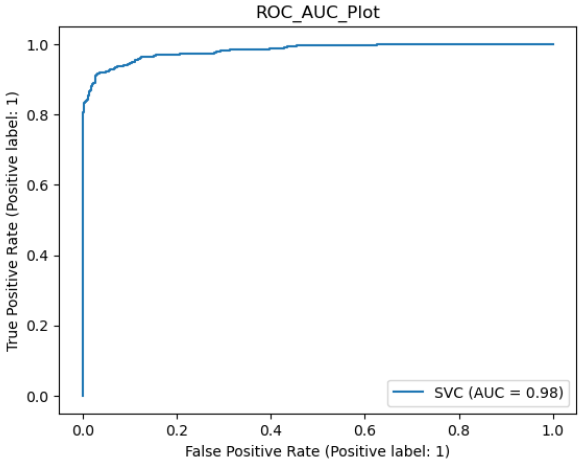

Figure 3.12 ROC-AUC plot for first SVM model

As soon as we obtain the ROC-AUC curve, which illustrating the trade-off between True Positive and False Positive rates for the SVM model, we can validate that the model is enough to detect fraudulent transactions with just a minor variation from the logistic model.

However, the features for ANOVA testing will bring overfitting to the SVM model as well, so here we make sure to that testing should be carried out only correlation matrix selected features in order to avoid overfitting.

Decision Classifier and Random Forest Classifier. The decision tree is one of the most effective models for detecting fraud, as was already noted, and it does an excellent job of doing so in this instance as well. It performs about the same as logistic but somewhat better than SVM model, according to testing results.

	precision	recall	f1-score	support
0	0.96	0.98	0.97	760
1	0.95	0.91	0.93	359
accuracy			0.95	1119
macro avg	0.95	0.94	0.95	1119
weighted avg	0.95	0.95	0.95	1119

Figure 3.13 Decision tree model with features based on correlation matrix

As it can be seen from Figure 3.13, precision is about to 0.95 where indicate model is able to detect almost all fraudulent transactions. In confusion matrix, we make sure that model is better

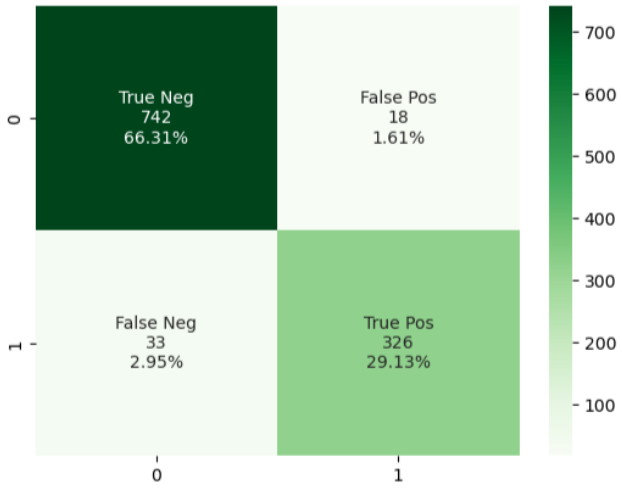

Figure 3.14 Decision tree confusion matrix with correlated based features

perform than SVM since higher true negative and true positive rate, and lower false negative and false positive rates.

Random Forest perform the best in our fraud detection case since it has higher indicators in all evaluation criteria.

	precision	recall	f1-score	support
0	0.94	0.99	0.96	760
1	0.99	0.86	0.92	359
accuracy			0.95	1119
macro avg	0.96	0.92	0.94	1119
weighted avg	0.95	0.95	0.95	1119

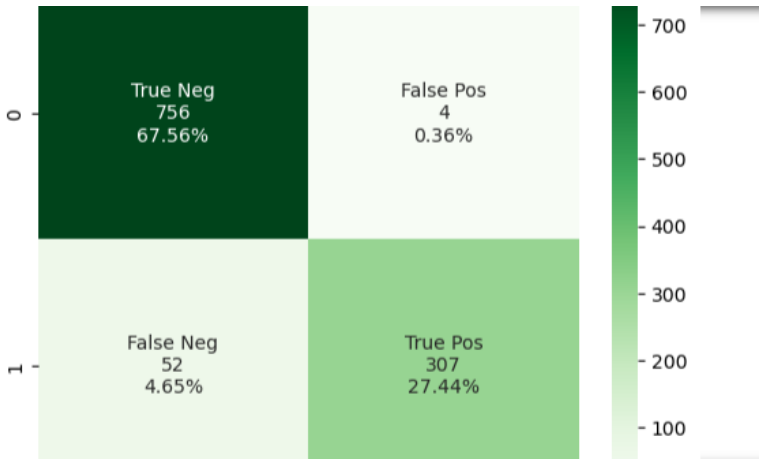

Figure 3.15 Random Forest model's evaluation metrics

In order to effectively identify fraud, it is crucial to reduce false positive rate, which are actual frauds that the model misses. When we examine the same training data with other models, Random Forest shows the maximum precision and recall, thus it could be accepted as the best model. It has lowest false positive rate (0.36%) and it trigger to continue with this model.

3.3 Results and Conclusion

This paper compares and explains several fraud detection techniques and models. Particular attention in the study was paid to credit card fraud, which costs the financial system billions of dollars and seriously harms both banks and people. Therefore, the dataset for credit card fraud detection was employed in the last chapter's quantitative study of machine learning models. Decision tree, random forest, logistic regression, support vector machines were applied to training dataset with more than 280,000 transactions. After the preprocessing and EDA part, it is possible to answer to first research question that set in the beginning of 3rd chapter.

Q1: What are the key patterns and characteristics of fraudulent transactions? First of all, credit card fraud dataset is very imbalanced dataset and before the applying the models some techniques in order to get accurate result and model. There are some reasons behind this imbalance. Firstly, fraudulent transaction is relatively rare case compared to genuine transaction that happens millions of times in a day. So, it has lower than 1% coverage in dataset.

Secondly, there are some features that has specific patterns in fraudulent cases. When the box-plot visualized, it is clearly seen that features such as "V18", "Amount of money', "V16" and some other variables have slightly different distribution range. Even if the amounts are similar,

they span a wide variety of financial transactions, unlike fraudulent instances, which aim to steal substantial sums of money. So, with a help of these features it might help to get better accuracy.

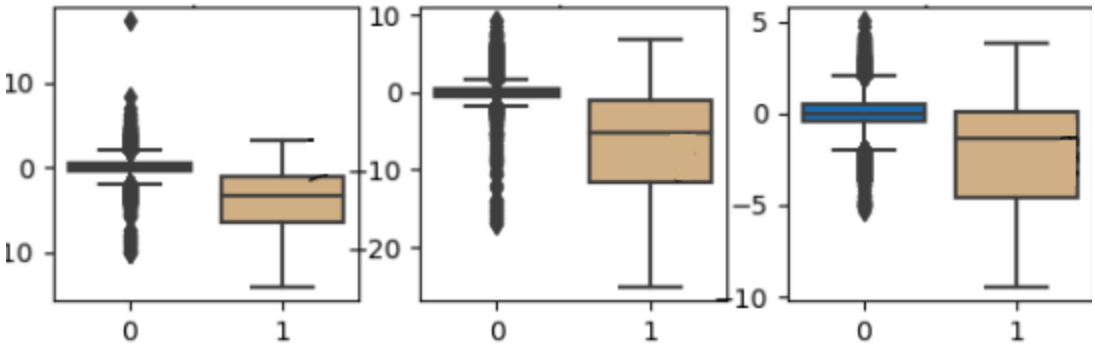

Figure 3.3.1 Box Plot for feature V16, V17, V18.

Moreover, clustering could be used to group the dataset and there is an example in the research, where KNN groups the dataset into 4 parts and it might also help to get better accuracy when model fed by the single group of data.

Q2: Which metrics should be used for credit card fraud detection? As it was already mentioned during research there are different evaluation metrics that are used by scientist. In this particular imbalanced fraud detection case, there are some limitations and recommendations about evaluation metrics. Accuracy is most useful evaluation metrics in general, however in this case random guess might bring us 98% of accuracy. As a result, false positive rates were attempted to be kept to a minimum throughout research, and precision was considered to be one of the most crucial criteria. Obviously, the ROC-AUC curve, the confusion matrix, and other assessment metrics were also shown and used in the study, but precision is somewhat more crucial in order to prevent missing fraudulent transactions.

Q3: What is the effectiveness and performance of the different machine learning models? This question covers almost the aim of the study and the title. Every machine learning model has unique power and strengths in specific topic and here 4 different algorithms were tested. However, last opinion based on testing is not enough. Therefore a-priori testing was examined with the data which was excluded from the main dataset before analysis. The outcomes of the testing are fairly similar to those of regular testing in the modelling phase. Here, the random forest and logistic regression both outperformed the other evaluated models. With a precision rate of 1, random forest identified all fraudulent transactions. Even if random forest does not perform as well as other

models in identifying legitimate transactions, it is still the best option since we must choose the model with the lowest false positive rate.

Model	Precison	Recall	F1-Score	Accuracy
Logistic Regression	0.98	0.89	0.93	0.94
SVM	0.99	0.89	0.94	0.94
Desicion tree	0.95	0.89	0.92	0.92
Random Forest	1	0.86	0.92	0.93

Figure 3.3.2 Final comparisons of models

Each model works effectively in this study, however logistic regression and random forest definitely provide somewhat superior outcomes. However, it would be difficult to determine which model is the best at detecting fraud in transactions if we just used these measures. As a result of the evaluation of the confusion matrix for each model, it has been determined that random forest has the lowest rate of false positives when compared to logistic regression and the other models.

In conclusion, this paper defined the all kind of frauds and their types, then suggest some modern machine learning models as a solution. This models' theoretical backgrounds and used cases in fraud detection were explained. Lastly, in order to test different algorithms, the most used 4 models has been applied in one of the greatest fraud threats in last decades – credit card frauds. The research has demonstrated that models has slightly same performance, but random forest model emerged as a most effective among other algorithms that we have tested. However, the modelling part was not that easy, and there were some challenges such as imbalance dataset which is common in fraud detection. Nevertheless, it was possible to bring valuable insights from the dataset and visualize the relationship between the variables. Comparing different variables for each single model help us to understand the importance of feature selection as well. This paper will not only contribute to the financial security field, but also help to deal with imbalance dataset problems in real-world scenarios.