UNIVERSITY OF SOUTHAMPTON

Mathematical Science

Identification of Unusual Patterns in Product Returns: An Unsupervised Learning Approach to Fraud Detection

by

[redacted]

This Dissertation is submitted in part-fulfilment of the requirements for the degree of MSc in Data and Decision Analytics

September 2022

Student Number: 33105693

ERGO Number: 76372

Table of Contents

Table of Tables

Table of Figures

ABSTRACT

Fraud detection and prevention is one of the most crucial aspects of the retail industry. Product returns presents one of the major challenges in present day retail, and as e-tailing continues to grow towards becoming the most popular consumer option to purchase goods, fraud methods are becoming more sophisticated. Retailers use their flexibility in return policy as one of the major tools towards attracting and pleasing the consumer, but this often poses the issue of return fraud largely affecting profits. As flexible return policies and customer satisfaction are highly correlated, it is important to determine a way that guarantees that losses are not incurred due to high number of fraudulent returns. This problem continues to pose a concern both strategically and economically, and accordingly, the investment of time and resources into developing a model that accounts for factors that identify fraudulent transactions could be crucial in the process of balancing between customer satisfaction and maximization of profits. In this era where technology is dominating, machine learning and big data provide the tools necessary to manage such challenges. Many studies have proposed methods to detect fraud through supervised machine learning models, and while that performs well, the lack of fraud-labelled data in the retail industry makes it difficult to develop a supervised learning model.

Therefore, we propose two unsupervised learning models based on the Isolation Forest and one-class Support Vector Machine algorithms to detect unusual patterns and outliers in transaction data. These models aim to detect a group of outliers, apply artificial labels of fraud to these outliers, and train a different dataset to check for further patterns or indicators of fraud. The model will be implemented on transaction data grouped by different consumers, ranging from August 8th, 2021, through August 7th, 2022, containing data on over 49 million consumers. We conclude that certain features of transaction data and consumer transaction history could be significant indicators of fraud.

Contribution to your discipline

Research on product returns and fraud detection has increased throughout the years, as the scope of the problem increases. The aim of this paper is to consider an approach to fraud detection using unsupervised learning techniques, as literature suggests there is a considerable lack of fraud-labelled data in the retail industry. We aim to propose two unsupervised learning models that can create a distinction between normal behaviour and suspicious or fraudulent behaviour in transaction data. This could prove vital in the process of fraud prevention, an area that is continuously developing in retail.

Declaration Of Authorship

I, [redacted]

declare that this dissertation and the work presented in it are my own and has been generated by me as the result of my own original research.

I am happy for the Dissertation PDF and supporting materials to be made freely available online and that there are no copyright or other reasons prohibiting this.

I confirm that:

1. This work was done wholly or mainly while in candidature for a degree at this University;
2. I submit this dissertation in accordance with the requirements for the above programme of study;
3. I hereby give my consent to the University making this dissertation available for consultation by other students and staff within the University;
4. Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
5. Where I have consulted the published work of others, this is always clearly attributed;
6. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work;
7. I have acknowledged all main sources of help;
8. Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
9. Where this dissertation is submitted in multiple formats and copies (e.g. printed and digital PDF) I confirm that all copies were produced from the same master document and are identical;
10. None of this work has been published before submission;

Signed:[redacted]

Date: [The date]

Acknowledgements

I would first like to acknowledge the incredible program at the University of Southampton for the great opportunity provided for me to pursue my Masters in the field that I enjoy the most. I am eternally grateful to Dr. Steffen Bayer, Dr. Regina Frei and Chief Data Scientist at Appriss Retail David Speights for the opportunity to work under their supervision and guidance which has developed my personality and skills more than I could have imagined. I am also grateful the rest of the Data Science team for their continuous support and guidance.

This positive experience has been something that I will hold at high standards, and I believe working alongside these incredible academics has prepared me for what comes next in my career.

Chapter 1 Introduction

With the development and advancement of larger e-commerce channels, product returns have drastically increased in importance to retailers. The increasing importance of technology, in addition to the necessity of online businesses throughout the Covid-19 pandemic, meant that a smoother transition into the availability of online retailing has become vital. According to a study by Statista on online retail sales in the United Kingdom, the value of online retail sales in the UK was approximated to be close to £120 billion in 2021, which represents around a £50 billion increase from sales in 2018, before the pandemic (*UK Online retail sales 2012-2021*, 2022). Another study by the US Department of Commerce shows an overall increase of approximately 56% in US ecommerce sales from the first quarter of 2020 to the first quarter of 2022, summarised in Figure 1. With larger sales, and flexible return policies due to the complexity of online sales, comes a great responsibility towards managing product returns (Bao, Hilary and Ke, 2022). Flexible product return policies are considered one of the valuable tools towards customer satisfaction, which in turn could conflict with profit optimization. According to (Frei, Jack and Brown, 2020), a high quality customer service is perceived to be the basis of a seamless shopping experience and driving sales.

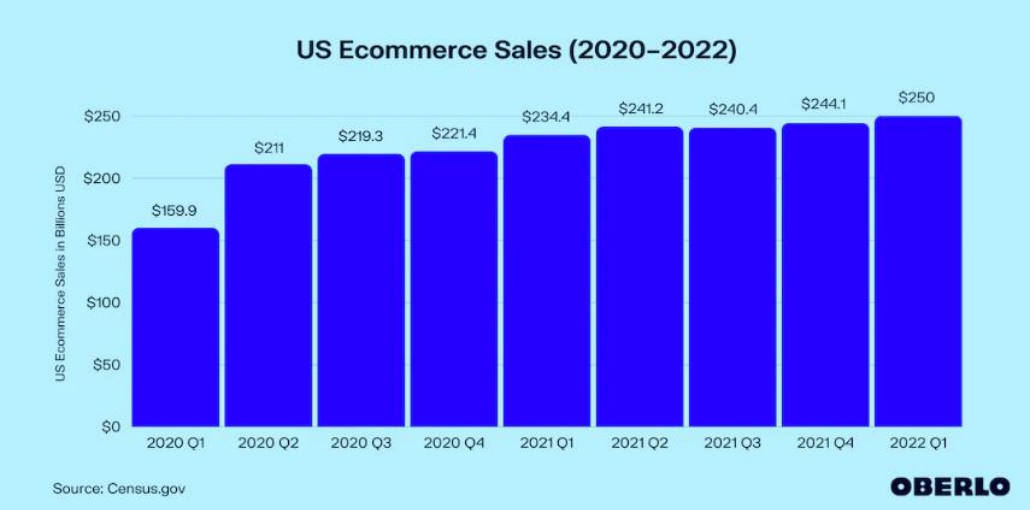

Figure 1: US Ecommerce Sales (2020-2022)

While these flexible return policies provide the necessary customer experience to enhance sales, it also poses a large issue that could be detrimental to a business if remained undealth with, fraudulent returns. Returns is a complex issue to deal with, as each product requires a different process, which could incur a large cost on retailers. Return fraud and abuse make this a largely more complex issue, with return fraud estimated to cost US retailers $23 billion per year (Abbey,

Ketzenberg and Metters, 2018). As a result, developing an appropriate fraud detection method has become necessary not just to keep profits high, but also to survive. Changing product return policies would be a step towards minimizing fraud, however, it would imminently affect the business as many legitimate customers would be affected by that change. Therefore, a fraud detection model that can keep these flexible product return policies while providing protection against fraud would lead to a significant increase in profit with little to no effect on customer satisfaction. This model could be developed based on the consumer transaction history, with emphasis on suspicious or unusual behaviours. These behaviours could be captured through the study of transaction data variables such as total amount returned, total amount purchased, total quantity of purchases and returns, in addition to consumer purchase patterns captured by looking at stores visited, receipted and non-receipted returns ratio and return behaviour. Ultimately, a model would continuously update a suspicious or fraudulent score for each consumer and would set a flag in real time on any user that is categorized as fraudulent above a certain threshold score to be set based on the algorithm used. A decision to warn, block or allow the consumer to continue with their purchase is then taken based on company policy and how strict the company decides to be.

Existing models to detect fraud largely depend on the availability of labelled data. Whether this deals with credit card fraud, retail fraud, or other types of fraud, a simple yet effective way of identifying fraudsters is through training a machine learning model to pick up on the behaviours that correspond to fraudulent transactions. However, fraud in retail is often subjective, where a consumer is considered fraudulent based on suspicious behaviour, and not on a clear instance of fraudulent activity. Most retailers develop on idea of what would be considered as fraud but cannot put a certain label to it as it is not very common to act against fraud where it is solely based on observed behaviours, and therefore fraud is retail is not labelled definitively. Consequently, it is important consider both supervised and unsupervised learning approaches to fraud detection, with an understanding to which learning technique is most applicable to our data.

Therefore, we propose a fraud detection model that aims to identify unusual behaviours and patterns in transaction data based on two unsupervised learning algorithms: Isolation Forest and one-class SVM. These models aim to extract features that are predictive of unusual or suspicious behaviour through isolation of a cluster from the rest of the data, in addition to applying artificial fraud labels to this cluster to be tested on a different dataset. This is done to ensure that the model will be updated to include anomalies in datasets that are distributed differently, in addition to protection against classifying a ‘normal’ data point as an anomaly. The implementations of these models raise some questions that we will aim to answer through our analysis of the results.

1) What are the variables that could be predictive fraud in transaction data?
2) Amongst the considered unsupervised learning models, how are outliers selected differently?

To answer these questions, we look at transaction data from a high value retailer in the United States through the Appriss Retail database in Greenplum. By studying the different variables available in transaction data, we can create secondary variables to arrive at a final dataset that is used to create the model, which will then output a set of anomalies. This will help us understand the variables most likely to predict fraud, in addition to understanding the relationship between these anomalies and their corresponding consumer behaviour, leading to a final model that provides a clear distinction between 'normal' behaviour and outlier or anomaly behaviour.

The following chapters in this paper will look at the following. Chapter 2 will look at a background study on product returns and return fraud. Chapter 3 presents a literature review of machine learning techniques in the retail industry, with explanation of the different models used and discussion of advantages and disadvantages to each of these models. Chapter 4 presents the methodology to arrive at a final model, including details on methods and steps applied to preparing the data, in addition to the choice of parameters for each model. Chapter 5 presents a description of the data used, with key statistics for each of the models and performance measurement. Chapter 6 presents the results of each of the models with a subjective discussion of the findings and their implications. Chapter 7 concludes the paper with reference to the research questions proposed, in addition to a discussion of the limitations and the potential of further research into similar topics.

Chapter 2 Product Returns – A Background Study

In this chapter, we will start by going through a general review of product returns, the policies and implications of reverse logistics, and an outline of the issues involved with the process of product returns. We then discuss fraud detection in the context of product returns in retail, with emphasis on the economic and strategic importance of investment into fraud detection resources and solutions. Finally, we provide a conclusion of the background research conducted, with an introduction to the direction we decided to take in our project and how our contribution can provide a different dimension to fraud detection solutions.

2.1 Product Returns

The concept of product returns is a critical issue to any retailer, as it forms a large aspect of the supply chain and has a significant effect on business performance. The variability in demand, quality of product, price and period means retailers must have a thorough understanding of the nature of product returns, the patterns associated with customer behaviour and the size of the profit/loss that their return policies are bringing to their business. Products can be purchased and returned online, in store or a combination of both, which adds to the complexity of identifying these patterns. While having a flexible product return policy is crucial to customer satisfaction, a priority for most retailers, it is important to manage product return policies in a way that prevents drain on profits. With the rise of online sales, product returns have increased, a natural result of purchasing a product with higher uncertainty than in store purchases. Digital technology continues to rise and dominate in today's world, and with the covid-19 pandemic, the popularity of remote and hybrid jobs, and the ease of internet access, digitalization has become necessary for every competing business. As the internet continues to grow to be a focal point in sales, retailers are now working towards an omnichannel model, with a top priority centred towards great customer experience (Frei, Jack and Brown, 2020). The flexibility of product returns is one major contributors towards customer satisfaction, in addition to its large impact on customer loyalty and their long-term value (Mollenkopf, Frankel and Russo, 2011). Figure 2 demonstrates the growth in online sales between 2012 and 2021 in the United States, with a clear spike in online sales in 2020, largely credited to the covid-19 pandemic. This is further confirmed with an observed growth in online sales of 43.7% between the first and second quarter of 2020, where the covid-19 pandemic effectively started (Young, 2022). With the growth of online sales comes growth of product returns, as the process involved in product returns is continuously optimised for consumer satisfaction. Many retailers tend to underestimate the problem of product returns, a major factor in retailers that are struggling to survive (Frei, Jack and Brown, 2020). A study by (Speights, 2013) of 10 retail clients of Appriss Retail found that all 10 clients underestimated their return rates.

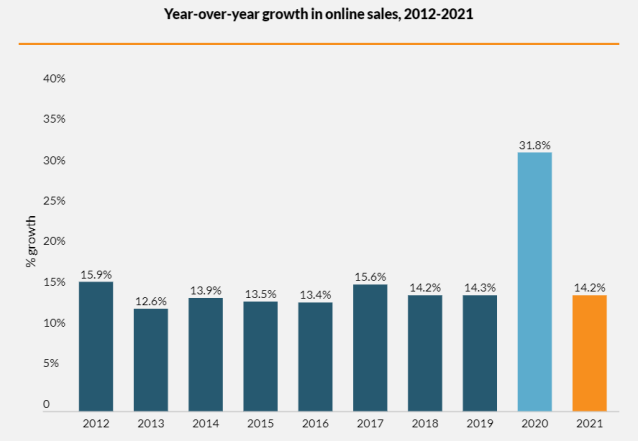

Figure 2: % Growth in online sales, 2012-2021.

Large consumer returns mean that this increase in sales is not always profitable for retailers, as reverse logistics incurs costs that could counter profit. Reverse logistics involves a complex

process that includes the decision to resell or resend, repair or scrap, and supply chain processes involving the customer, manufacturer and retailer (Akhilesh and Swapnil, 2017). The return of a product from a customer's perspective can be as simple as dropping off a product to a nearby store, but for retailers, it involves a complicated and often inefficient process (Potdar and Rogers, 2010). Within the process of a return, items may be damaged, stolen, or involve a long delay, which can often lead to some products being deemed unfit to return to sales and are auctioned or sold for a much lower price than their market value (Frei, Jack and Brown, 2020). Although flexible return policies are one of the major aspects leading to customer satisfaction, they require additional resources and incur additional costs in the supply chain (Anderson, Hansen and Simester, 2009). These costs include repairing, repackaging and sorting of items, which require a significant amount of resources (Abbey, Ketzenberg and Metters, 2018). As reverse logistics is a necessary aspect of the supply chain and has become increasingly more important with the popularity of ecommerce, retailers now must develop a different insight into how this can be used to their benefit. This includes developing an understanding of product return volume and the expected workload in reverse logistics (Kranz, Urbanke and Kolbe, 2015). Reverse logistics represents a process that is more complex to handle, as it could involve additional inspections and checks, sorting products, managing possible repair and refurbishment costs, and involves low spatial utilization (Robertson, Hamilton and Jap, 2020). However, this allows reverse logistics to be treated as an opportunity, as it plays an important role in the supply chain network, and offers a second chance to profitability (Potdar and Rogers, 2010).

The implications of product returns, and how they are often written off as miscellaneous losses meant that retailers now pay more attention to understanding the scope of the issue (Ambilkar *et al.*, 2022). Consumer returns amounted to 351$ billion in 2017, a number high enough to rank second on the Fortune 500 if it represented an independent company's worth (Abbey, Ketzenberg and Metters, 2018). These numbers are countered with higher sales numbers, but very often contradict the forecasted profits based on sales, which can lead to retailers considering a change to their return policies and flexibility. According to (Abbey, Ketzenberg and Metters, 2018), L.L. Bean had an estimate of worthless returns costing the company 30% of its annual profits. This led the company to establishing a new policy that sets a limit on the period in which product returns are allowed. These return restrictions are not unique to L.L. Bean, as multiple companies such as Costco and Best Buy have set similar return restrictions. Product returns is not as issue that only affects the consumer and retailer. Product waste and product transportation means there is also an impact on the environment. (Abbey, Ketzenberg and Metters, 2018) emphasizes the issue of return abuse, labelling it as one of the most common causes of profit drainage, amounting to

approximately 23$ billion per year in loss of profit. This is represented in many ways, from fraudulent returns, to buying expensive items for single use and then returning them, which is commonly referred to as retailer borrowing. This is made possible by extremely lenient return policies, allowing returns abuse through returning items for full price after purchasing on sale, or buying items on credit cards with travel rewards and getting a return in cash.

Many factors affect the implications of returns on retailers, including legitimacy, processing, how resalable an item is, and the effect of return on customer satisfaction (Robertson, Hamilton and Jap, 2020). Technology and data have promoted the opportunity to get creative with return policies and decisions, in addition to understanding the nature of returns and what can be considered fraudulent returns. Some facilitators of fraudulent returns found by King et al. (2007) are experience of successful fraud, positive attitude towards fraudulent returns, consumer awareness of return policies. The implications of returns also differ based on the channel used, as online and in store returns have different processing steps and therefore present different costs to the retailer. This is also something that must be considered by retailers in their analysis of their return rates and the scope of the problem, as understanding where fraud is most prominent could be important in determining the most suitable solution for the problem.

The dilemma of allowing product returns leniently while preserving profits is largely resulting from flexible return policies being one of the main attractive features for customers, especially in online retail. Retailers assume that allowing lenient return policies encourages future sales and attracts loyal customers (Robertson, Hamilton and Jap, 2020). Top retailers continue to allow free product returns, even though the numbers are significantly rising. However, this could prove costly to a business if it is not continuously monitored. According to (Ram, 2016), return rates are approximated to be 20-40%, with figures going up to as high as 70%. (Speights, 2013) suggests that a consumer with return rates greater than 20% to 30% can lead to a negative impact on operating profits. These high return rates can be extremely costly, as certain products require heavy resources before they are ready to sell again and might have little to no profit in selling again due to these costly resources. According to (Frei, Jack and Brown, 2020), a strategy that improves the rate of return by 5% can have an impact as large as an additional 200 basis points in net margin. This can only motivate further research into product returns, the factors affecting return rate and possible strategies that could ensure a balance between losses and customer experience. Although returns are significantly affecting both large and small businesses, it is still reported that only 32% of top retailers quantify the full cost of returns Multiple factors impact the size of the loss incurred by a company due to a return, including time spent processing returns, determining whether an item can be resold or not, and administrative expenses. (Speights, 2013)

The main challenge faced in reducing these numbers is the complexity of devising product return policies and the attachment of customers to lenient returns. The difference between product returns being favourable for a good customer experience and unfavourable due to its leniency is very slim, and so retailers must construct a clear and exhaustive strategy to deal with product returns. Further research and understanding of how customers view product return behaviour and what they consider acceptable could serve as a starting point for retailers in constructing return policies (Robertson, Hamilton and Jap, 2020). Return policies are a significant factor that affect the market for every retailer, and so they can be viewed as a marketing strategy that is a major factor in profitability and customer satisfaction. Different companies have set different return policies, based on their business and profit priorities. Firms such as Warby Parker have integrated return activity as a crucial part of the sales. Other firms such as Zappos and Nordstrom encourage customers to return with easy return policies as they believe it encourages future customer purchases (Robertson, Hamilton and Jap, 2020). Understanding the product return patterns and devising return policy strategies accordingly could be a powerful tool to strategically manage demand uncertainty and shift the process of dealing with unsold goods to the supplies (Robertson, Hamilton and Jap, 2020).

2.2 Return Fraud

The rise of e-commerce has opened the door for fraudsters to get more creative. Criminals are continually searching for weakness in fraud detection and fraud prevention systems (Malphrus, 2009). The use of electronic transactions continues to rise every year, as e-commerce becomes the more popular option, supported by high scale retailers such as Amazon and eBay (Jha, Sivasankari and Venugopal, 2020). Before online retail was popular and the process of product returns was as simple as dropping off a package, returns involved a more inconvenient process for the customer, with much more room for identification. In today's digitally transforming world, dealing with fraud has become much more complex. Return fraud can be described as consumers returning goods to retailers with the knowledge that the returns is against company policies (Shih *et al.*, 2021). Fraud has become more prominent in online retail. Online fraud is an area that has gained significant interest in research, with first party fraud increasingly becoming a challenge to businesses, with significant resources being allocated to dealing with this issue (Amasiatu and Shah, 2018). The popularity of e-commerce means credit cards are also in use, which opens another opportunity for fraud. (Jha, Sivasankari and Venugopal, 2020) states that credit card related fraud occurs 3-4 times more than face-to-face transactions, which is in line with the witnessed increase in fraudulent transactions as retail becomes more popular online. Online marketplaces have become a more

popular option amongst consumers as they provide more selection opportunities, more competitive prices, larger availability of inventory, and are generally more convenient (Renjith, 2018). The rise of commerce conducted over the internet represents a greater risk in criminal activity. The mass transactions conducted over the internet means criminals can illegally gain access to personal and financial information of consumers, to be used by or sold to fraudsters (Malphrus, 2009).

Fraud is most popular in the retail and financial industries, due to the nature of the industries and sensitivity of the information used in conducting business. Most fraud occurrences in the e-commerce industry include stolen financial information, in addition to fraudulent return of products (Renjith, 2018). Fraudulent activities pose a significant challenge to multiple industries such as retail, banking, and public sector establishments (Jha, Sivasankari and Venugopal, 2020). According to (Speights, 2013), approximately 8% of returns in North America are fraudulent, a number that is high enough to indicate a serious issue. (Lopez-Rojas, 2015) reports that return fraud is estimated to cost US retailers approximately 9$ billion every year. (Shih *et al.*, 2021) also report that return fraud causes at least $220 million in losses. Retailers are continuously and widely affected by fraud, as fraud detection and prevention continues to present a crucial challenge to overcome (Jha, Sivasankari and Venugopal, 2020). There is no lack of understanding or awareness of the problem, as research suggests that 82% of large market retailers are aware of return fraud and the extent of the problem (King, Dennis and McHendry, 2007)**.** Industry experts including payments system providers, financial institutions, law enforcement organizations suggest that although payments fraud is not considered a crisis, organizations must continue to adapt and work towards limiting criminal activity (Malphrus, 2009). The importance of fraud detection is now recognised worldwide, as both EU and US have introduced mandates for use of fraud detection as a minimum-security requirement for financial services (European Central Bank, 2013).

Fraud detection and prevention can be different for different industries and retailers but is certainly based on detection of patterns and characteristics of fraudsters. Since most fraud incidents have certain characteristics relating to environmental conditions and consumer behaviour, fraud detection can benefit from a deeper understanding of these characteristics (Mustika, Nenda and Ramadhan, 2021). These characteristics are often unique to each industry, and further highlight a trend or pattern that identifies the motivation behind committing fraud. (Amasiatu and Shah, 2018) highlights the importance of deterrence, as fraud is widely considered as opportunistic, in addition to the cost of fraud investigation and underlying motivation behind committing fraud. Some research suggests that consumers who commit or engage in first party fraud cite reasons such as economic distress, low self-esteem and revenge motives (Amasiatu and Shah, 2018). Consumers tend to have a certain view or orientation towards product returns, and will often have different

ethical outlooks on the issue of return abuse (Wachter *et al.*, 2012). These reasons and motives cannot be identified by data, but many fraudsters tend to be repeat offenders, something that can be captured by a model that identifies these unusual behaviours. The ease of committing fraud based on flexible return policies increases the availability of fraud opportunities. These opportunities decrease as more resources and attention is raised towards anti-fraud policies, as the likelihood of detection and risk of committing fraud increases. Ethical dependence on the consumer does not always pay dividends, as committing fraud is not always considered wrong or illegal by the consumer when they cannot directly see the harm they are causing. This phenomenon is referred to as rationalisation, where first party fraud offenders use rationalisation to either argue that their behaviour does not hurt anyone (and is therefore excusable), that their victim/retailer deserves the wrongdoing, or that circumstantial pressures outside their control led to their behaviour (Harris and Daunt, 2011).

One of the biggest challenges in facing return fraud is the importance of existing return policies for most retailers (Petersen and Kumar, 2009). The nature of the retail industry means that flexible return policies will promote higher sales, and therefore higher profits. (Bower and Maxham, 2012) states that return policies, amongst several other factors, are essential to consumers during the purchase decision Customer experience is one of the cornerstones of the retail industry, and hence it is prioritised in most situations. This presents a dilemma to retailers, as fraud prevention can cause a conflict with customer experience. One of the debates of altering return policies lies in the effect on consumers and whether it may drive some consumers away. (Speights, 2013) analysed shopping patterns of consumers before and after a return denial and found that the shopping patterns of about 40% of consumers were not affected, and that net sales of all denied consumers eventually return to their level before denial. The use of technology is one way of managing this conflict, as it provides the ability to track consumer return behaviour, as denying suspicious returns is an essential step into managing the issue of fraud while maintaining customer satisfaction. However, fraud detection does not stop at simply classifying a transaction or a consumer as fraudulent. Further investigation is necessary, as some transactions may have fraudulent patterns but are otherwise legitimate. Investigating a transaction includes obtaining sufficient evidence to stop first party fraud, and to provide support for sanctioning those that commit fraud (Amasiatu and Shah, 2018). Most fraud detection and prevention solutions considered by top retailers prioritise the minimization of rejecting legitimate consumers over accepting fraudsters, as the business cost of rejecting returns from legitimate customers is higher than that of accepting a fraudulent return. As a result, retailers are generally hesitant in rejecting or acting against fraudsters, as these decisions are not always based on definitive and conclusive evidence, and therefore resort to considering fraud as a

miscellaneous cost. This stands in the way of fraud prevention, as the penalty for committing fraud is considered minor. (Amasiatu and Shah, 2018) argues that retailers should do more than just consider fraud to be cost of doing business. Further escalation of action against fraud can include rejecting fraudulent claims, blocking offenders from future purchases, and reporting to law enforcement where appropriate. Additionally, fraudulent patterns are continuously changing, and so retailers must adapt these changes to their existing fraud prevention solutions. (Amasiatu and Shah, 2018) highlights the importance of monitoring the anti-fraud framework implemented, as feedback from assessment of losses can be used to improve the anti-fraud techniques in use, and therefore better assess the performance of the overall fraud strategy.

Return fraud is viewed as a significant challenge for retailed due to its continuously evolving nature, in addition to the rise of ecommerce opening more innovative ways of abusing return policies (Bower and Maxham, 2012). Understanding the fraud and abusive return schemes is an essential step in formulating a defence against these abusers (Speights, 2013). Although some returns are not fraudulent or illegal based on flexible return policies, these returns can nevertheless be disadvantageous to retailers. While it is important to distinguish between fraud and non-fraud, it is likely more difficult to have binary classification, as different customers inhibit different behaviours. (John, Shah and Kartha, 2020) classifies returners into three categories based on their behaviours: fraudulent, over-ordered, and legitimate. Finding behavioural patterns that could classify a customer into one of these categories is beneficial to retailers, as this could serve as a basis for a decision of action or no action against a customer. Those who order in bulk knowing that they can return items flexibly know that this can be profitable under certain conditions. (Speights, 2013) discusses multiple types of return fraud, including: Renting/Wardrobing, which is the act of buying something for limited use, such as a dress for an occasion, and then returning it for a full refund. This accounts for just over 50% of return fraud occurrences. Another fraud type considered is price arbitrage, which is the act of purchasing two similar items with different retail prices and repackaging the cheaper item in the expensive items box to return for a full refund. These are common in retail and exhibit certain behaviours that can be identified through detailed analysis of customer behaviour. Therefore, it is crucial that transaction data is thoroughly analysed to identify these patterns and find a way to distinguish between a profitable and non-profitable customer. This is made possible through data analytics. While traditional fraud detection systems use machine learning technologies to identify fraud, it is important to realise the importance of big data and the trends that lie within these large datasets to identify fraudulent behaviours (Jha, Sivasankari and Venugopal, 2020). Data mining and machine learning can provide techniques that are useful in interpreting data and gathering useful insights (Ribeiro, Oliveira and Gama, 2016). The evolution of

big data and its constant advancements provides a platform that makes it possible to perform analysis on historic data to understand consumer and seller behaviour to identify potential instances of fraud, and therefore make conclusions on profitable actions and decisions (Renjith, 2018). One of the great attributes of big data analytics is its ability to discover useful patterns from large data sets by systemically gather, organizing and assessing the unstructured data. (Jha, Sivasankari and Venugopal, 2020). The popularity of ecommerce meant that customers are more easily identified, and large amounts of transaction data is available. A machine learning model that can extract useful features in predicting fraud or identifying anomalies, which can be used to reduce the damage caused by product returns, in addition to altering policies and regulations relating to returns and fraud.

Chapter 3 Machine Learning Techniques – A Literature Review

To manage the issue of return fraud, many companies rely on their own data to create models that provide insights on indicators of fraud and how company policies affect the general return and fraud rates (Ülkü, Dailey and Yayla-Küllü, 2013). To be able to formulate a return policy that will not lead to unpleasant return experiences, and at the same time, not be vulnerable to abusers, it is important to understand the consumer and their previous behaviours, a process made possible by digital footprint analysis (Shih et al., 2021). The omni-channel world is actively integrating the consideration of interactions between the customer, brand and retailer, in addition to data on orders, returns and exchanges (Shih et al., 2021). For example, Amazon uses its own data to cancel accounts that show excessive return occurrences (Robertson, Hamilton and Jap, 2020). In an e-commerce environment, a simple rule-based expert system can be effectful in fraud prevention. However, this is often based on human analysis and hence can be biased, which motivates the inclusion of a computer model that is able to filter transactions based on non-manually created rules. This is where machine learning excels, as it is based on training a model to identify patterns. Machine learning exploits the study of algorithms that learn from data, as these algorithms work on models based on a set of inputs and use them to train a model to output a prediction of a certain variable or decision (Ribeiro, Oliveira and Gama, 2016). This is one of the most popular areas used in fraud frameworks, however, retailers and data experts continue to innovate to find alternatives that can suit various levels of data availability. As resorting to stricter return policies does not guarantee a positive sacrifice in terms of fraud prevention and customer satisfaction, a fraud

prevention model can achieve lower fraud rates and lower return rates, numbers that are desirable for business performance. Return rates highly correlate with operating margins, as (Speights, 2013) suggests a 1% decrease in return rates can lead to an improvement in operating margins as high as 6%.

Existing literature suggests various ways that retailers can deal with fraud. As the size of the ecommerce market grows, traditional methods for fraud identification and prevention becomes inefficient as it involves a lot of training and resources (Shih *et al.*, 2021). (Frei, Jack and Brown, 2020) identifies some main vulnerabilities in current product return systems, including poor data management, faulty IT systems, treating returns as an asset without further monitoring or evaluation, lack of thorough understanding of consumer behaviour and complexity of the return process. Using data collected on consumer return behaviour, it is possible to develop statistical models that can identify common patterns of fraud and abuse to be used for fraud detection (Speights, 2013). (Malphrus, 2009) suggests using information on the account application for fraud detection. This includes identity proofing, credit card fraud detection, verification of age and telephone numbers, or telephone-based user verification. This is now extended as a cloud-based solution that uses multi factor authentication as an identity verification technique. Machine learning and data mining techniques such as classification, prediction, clustering and regression are used to identify a pattern within a large dataset (Burkart and Huber, 2021). This provides an effective approach to analyse unusual patterns in large transaction data, which could provide insights into what the key predictors of fraud are (Jha, Sivasankari and Venugopal, 2020).

Using machine learning in fraud detection can be split into two categories: Supervised and Unsupervised learning. Supervised learning approaches are important in terms of identifying fraud but are often biased towards a certain type or pattern of fraud. As technology and online retail becomes more dominant, fraudsters continuously find ways to innovate in terms of committing fraud and staying anonymous(Petersen and Kumar, 2009). Due to the small percentage of fraudulent transactions or returns, these models are built on highly imbalanced datasets, with methods such Support Vector Machines, Logistic Regression and Decision Trees used for modelling ADDIN ZOTERO_ITEM CSL_CITATION {"citationID":"BJz8bvEi","properties":{"formattedCitation":"(Potdar and Rogers, 2010)","plainCitation":"(Potdar and Rogers, 2010)","noteIndex":0},"citationItems":[{"id":27,"uris":["http://zotero.org/users/9743652/items/9V39S3B9"],"itemData":{"id":27,"type":"paper-conference","abstract":"One important aspect of reverse logistics is to have a correct and timely estimation of return flow of material. Improved forecast accuracy can lead to a better decision making in strategic, tactical and operational areas of

the organization. Very little research has been done about the forecasting aspect of reverse logistics. For higher forecast accuracy, more robust method is required. The methodology presented here is based on the return reason codes (RC). The incoming returns are split into different categories using return reason codes. These reason codes are further analysed to forecast returns. The computation part of this model uses a combination of two approaches namely extreme point approach and central tendency approach. Both the approaches are used separately for separate types of reason codes and then results are added together. The extreme point approach is based upon data envelopment analysis (DEA) as a first step combined with a linear regression while central tendency approach uses a moving average. For certain type of returns, DEA evaluates relative ranks of the products using single input and multiple outputs. Once this is completed, linear regression defines a correlation between relative rank (predictor variable) and return quantity (response variable). For the remaining type of returns we use a moving average of percent returns to estimate the central tendency. Thus, by combining two approaches for different types of return reason codes, we have developed a model that can be used to forecast product returns for the consumer (Potdar and Rogers, 2010). These methods result in a machine that predicts whether a transaction will be fraudulent before it occurs but is likely to be biased due to the imbalanced data. This means that any fraud detection model that would be implemented must have a recommendation and not a definitive action on a transaction, as it would be problematic if non-fraudsters are commonly denied. These models are assessed through accuracy measures, such as false negative (Incorrectly classifying a fraudulent transaction as non-fraudulent) and false positive (Incorrectly classifying a non-fraudulent transaction as fraud) percentages. Since definitive action on potential fraudulent transactions is rare, most retailers that implement a fraud detection solution do not have labelled data, as it is difficult to completely decide to allocate transactions into a binary classification process. For this reason, unsupervised learning approaches are extremely important in the process of fraud detection and prevention.

Unsupervised learning is a machine learning algorithm that learns patterns about the data through analysis of multiple features. These patterns could then be used to classify data, whether that is binary classification or categorizing the data according to different patterns. Methods such as clustering, isolation forests, and one-class support vector machines are used to classify data into a ‘normal’ cluster and one or more ‘unusual behaviour’ clusters. In terms of fraud detection, these algorithms take features about a transaction and the history of the customer to classify this customer into one of those clusters. In terms of interpretation of the model, the conventional performance measures are not available as there is nothing to measure against. This means that more thorough analysis of the resulting clusters is required, to classify which clusters are simply a different class of

consumers, and which clusters present suspicious behaviour. This is often different for each retailer and involves a team of consultants or fraud experts to identify relevant clusters. To implement an unsupervised learning model, we must have extensive analysis of the data to extract certain features that would help set parameters to ensure that the resulting clusters will have the size and properties required to be identified as fraud. As a result, this makes the process of developing such model extremely sensitive to the preliminary data analysis conducted and increases the importance of understanding which features are likely to be fraud indicators.

(Caldeira *et al.*, 2012) describes detection and evaluation of fraud through supervised and unsupervised strategies. Supervised strategies use Support Vector Machines (SVM), Decision Trees, and Bayesian networks to examine labelled data to identify fraudulent transactions. Unsupervised strategies analyse data to detect behavioural changes of a user, or to detect unusual transactions and anomalies. Clustering and anomaly detection are popular methods used for unsupervised learning. Supervised learning is useful only when data is labelled or classified. The output variable must be categorical or numerical, and popular algorithms such as logistic regression, decision trees, random forests and support vector machines are used to train a model to classify a data entry or in this case, a transaction (Mustika, Nenda and Ramadhan, 2021). Unsupervised learning works on data that has no labels, such as basic transaction data, to identify patterns and potentially apply artificial labels to be used for model training and optimisation.

In the next section, we will discuss existing literature on both supervised and unsupervised learning methods in fraud detection, to better understand the cases in which each of these machine learning types can be used, in addition to the advantages and disadvantages of these models.

3.1 Supervised Learning

In simpler terms, supervised machine learning refers to classification, where instances or sets of data are given a label based on characteristics or known facts. Classification involves building a model based on a training set made up of a multivariate dataset and corresponding class labels (Bhavsar and Ganatra, 2012). The resulting model is then considered to have learned the data and its associated labels and is used to predict the class label of a test dataset, and performance measures are computed accordingly. However, the requirement of labelled data acts as a significant limitation. When it comes to topics that have a definitive way of classification, such as classifying a patient as diabetic or non-diabetic based on medical testing, this data can be easily available. In terms of fraud, this is often much more challenging. If a company has a set of transactions that are

confirmed to be fraudulent, supervised learning can be used to design and implement a fraud detection model. The data is used as a training set, and new data that have no definitive outcome of fraud is used to predict based on the supervised learning model (Ribeiro, Oliveira and Gama, 2016). These models are created using multiple analytical methods, including multiple regression, classification models including random forests, SVM and shrinkage methods (Abbey, Ketzenberg and Metters, 2018). These methods are used to devise a fraud detection system, which can be assessed by looking at the percentage of unprofitable consumers to whom returns are denied compared to profitable consumers that make returns. (Speights, 2013). We review different implementations of supervised learning methods in fraud detection, with emphasis on retail and product returns fraud.

Table 1: Supervised Learning Techniques

Author	***Method***	***Description***	***Outcome***
(Sahin and Duman, 2011)	Support Vector Machine and Decision Trees	• Compared SVM based and decision tree-based fraud detection systems. • The model divides the dataset into three groups with different fraudulent to non-fraudulent transaction ratios. These datasets are then split into training and testing data.	• Decision tree models were found to outperform SVM models.
(Zareapoor and Shamsolmoali, 2015)	Ensemble Trees	• Compared the use of SVM, Naïve Bayes, and K-Nearest-Neighbour methods to bagging ensemble classifiers based on decision trees.	• Bagging ensemble classifiers led to better fraud detection rate and false alarm rate. • Model capable of handling imbalanced data.
(Mahmoudi and Duman, 2015)	Fischer Discriminant Analysis	• Proposed a model based on Fischers discriminant function, with the addition of a weight function responsible for classifying transactions with higher financial cost implications, as they have a larger effect on profit and loss margins. • The function assigns higher weight to cards with higher spending limits. • Feature selection performed using decision trees.	• Results obtained were positive, with clear indication that a Fischers Discriminant Analysis based model is capable of transaction classification.
(Khan, Akhtar and Qureshi, 2014)	Artificial Neural Networks	• Proposed a model that uses the simulated annealing technique in addition to neural networks to develop a fraud detection model. • Simulated annealing was used to control the parameters involved in the neural network and generate random weights for all connections on the neural network.	• The model was deemed successful, resulting in a classification accuracy of 89.6%.
(Ishfaq, Raja and Rao, 2016)	Logistic Regression and Poisson Modelling	• Studied the impact of product and sales attributes on product returns using logistic regression, as the response variable indicating whether a product is returned or not is binary.	• Sales price was not significant in determining whether a product will be returned.

		• The analysis of the most significant attributes affecting returns is then used to develop an efficient sales return process. • Study focuses on the interaction effect between scarcity of a product and its lowest price guarantee. • Study was further expanded to introduce a Poisson based model used to model the process of sales and returns.	• Product scarcity provided a higher likelihood of a product being returned.
(Gadal and Mokhtar, 2017)	K-Means Clustering	• Model uses K-Means Clustering and Sequential Minimal Optimization (SMO) to construct an anomaly detection system used for online network anomalies.	• The hybrid approach of using k-means clustering and SMO proved to have a significant increase of around 25% in detection accuracy.
(Mustika, Nenda and Ramadhan, 2021)	Randoms Forests	• Studied the performance of different classification algorithms including K-Nearest-Neighbour, Logistic Regression, SVM, Decision Trees and Random forests on fraud-labelled transaction data.	• The study found that the random forest algorithm was the most successful at detecting fraud based on different characteristics in the data.

(Abbey, Ketzenberg and Metters, 2018) conducted research to understand the signs of a profitable customer by studying which variables from transaction data and personal information can be indicators of a profitable customer. They used multiple supervised learning methods to combine and find common patterns in identifying profitable customers, in a study on a retailer that operates brick-and-mortar properties. Out of a total of more than 75 million transactions identifying 1 million customers, they found supervised learning models to have 99.96% accuracy in classifying customers into profitable and non-profitable after 5 transactions, with an increase of 0.02% for customers that conduct more than 10 transactions. The study found that personal information such as age and income were irrelevant in classification, while number of customer purchases, total customer refunds, amount of current refund, number of purchase categories, average time to return, value of average item returned, and return frequency were most significant in identifying profitable customers. The study also looked to classify customers into legitimate returners, non-returners and abusive returners, and found that legitimate returners were the most profitable group, while a small number of abusive returners accounted for large losses, amounting to

approximately 60$ million annually. The strategy by (Abbey, Ketzenberg and Metters, 2018) suggests using such analysis to selectively decide when to impose stricter return restrictions, which could help protect returns while maintaining customer satisfaction.

The growth of ecommerce has been one of the main culprits in increasing retail fraud (Whitehead, 2021). One of the earliest findings of cybercrime came through credit card fraud, as it became a target for most fraudsters with the increasing popularity of ecommerce. Credit card fraud is one of the most common fraud types in online retailing, and according to (Adewumi and Akinyelu, 2017), credit card users are at increased risk of falling victim to fraud. To minimize credit card fraud risk, companies use authentication methods, including validation of phone numbers, physical address, secret question and answer (Wong *et al.*, 2012). Multiple supervised learning methods are used to counter credit card fraud in retail. Techniques based on machine learning, neural networks and data mining can be used to detect credit card specific fraud (Jha, Sivasankari and Venugopal, 2020). Summarised below is an outline of existing implementations of these supervised learning methods.

3.2 Unsupervised Learning

Although fraud is a considerable issue in retail, it is still rare compared to the mass purchases. Furthermore, classifying a transaction or a person as fraudulent is very often subjective, as it is not very common for definitive action to be taken against fraudsters, especially when the losses are not significant. For this reason, fraud-labelled data is scarce in retail. However, those who commit fraud still have unique characteristics compared to legitimate users. Classification analysis can be used as a tool to detect fraud based on these unique characteristics. This can be extended into a machine learning model that can be applied to classify a transaction as fraud or not fraud before it is executed (Mustika, Nenda and Ramadhan, 2021). An unsupervised learning model uses data to analyse accounts, customers, suppliers, and financial records to spot unusual behaviours and suspicious outputs (Ribeiro, Oliveira and Gama, 2016). Such model is constructed based on customer transaction history, allowing for action before the return occurs, which can protect loyal customers from stricter return policies. As with any other machine learning model, the output of these models presents the likelihood of a transaction being fraudulent within a specific certainty or accuracy (Ribeiro, Oliveira and Gama, 2016). After that, it becomes a business decision on what the action should be based on this likelihood output.

To identify these suspicious behaviours, there must be an understanding of what is considered ‘normal’ customer behaviours. It is important to understand what is considered standard or normal

consumer behaviour, as it provides a basis to find anomalies and aid the detection of unusual or malicious behaviour (Lopez-Rojas, 2015). Therefore, outlier detection plays a key role in unsupervised learning algorithms. Outlier detection simply refers to identifying patterns within the data that do not match the expected behaviour or what is considered usual behaviour (Ribeiro, Oliveira and Gama, 2016). (Chandola, Banerjee and Kumar, 2009) states that all datasets contain outliers, which can be classified into three types:

1) Global: Comparing an individual observation to other observations.
2) Contextual: Outliers defined depending on context. (For example, an outlier in terms of purchase amount needs context, as one transaction could contain a significantly more expensive item compared to others.)
3) Collective: A subset of observations is different relative to the overall dataset.

The main issue with unsupervised learning is the dependence on human analysis. Common techniques in fraud detection and prevention models involve a retailer approving a transaction using some decision process, often human created. Companies such as the retail equation work with retailer data to assess their returns based on a risk score, a score that is found from shopping and return history and is used to determine the acceptance or denial of a return (Robertson, Hamilton and Jap, 2020). To counter this human effect, various data analysis tools can be used to detect and envision outliers in large datasets. In practice, outlier detection works to simply classify pattern observations into one or more classes. This is useful in the context of fraud detection, as fraud detection systems use prediction algorithms to classify pattern observations. One way these fraud detection systems use outlier detection is through spending history of a user, as an observed deviation in the normal spending history can represent suspicious behaviour. Boxplots is one of the simpler data analysis tools used for this purpose, as individual boxplots can be used on each variable in a dataset to determine outliers (Ribeiro, Oliveira and Gama, 2016). However, fraud detection is more complex than detecting outliers within one variable in the data, as the aim is to combine these variables to find global outliers. This is where unsupervised learning thrives, as it works on training models on complex data to find unusual patterns and behaviours. This is achieved through algorithms such as Regression Trees, Principal Component Analysis, One-off SVM, and Isolation Forests. We next review different implementations of unsupervised learning methods in fraud detection.

Table 2: Unsupervised Learning Techniques

Author	***Method***	***Description***	***Findings/Limitations***
(Balasupramanian, Ephrem and Al-Barwani, 2017)	Principal Component Analysis	• Proposed a model that uses transaction data to extract significant attributes using principal component analysis. • These significant features are used to train the model to identify fraudulent transactions.	This puts an emphasis on user transaction behaviour being similar, which could give rise to problems such as new user transactions and unusual but non-fraudulent transactions.
(Ribeiro, Oliveira and Gama, 2016)	Regression Tree (CART)	• Studied transactions of returned items over a period of 3 months. • Data containing information on the store in which the return occurred, the region, the period, product type, number of items and average value of items returned. • The study uses a regression tree algorithm to obtain partitions of the full dataset, as this can detect outliers on partitions of the dataset instead of individual variables. • The CART algorithm splits the data subgroups in a way that decreases the variance of the target variable at every split, aiming to obtain partitions of the full dataset to detect outliers.	Study found multivariate analysis to be most useful in terms of extracting global outliers and revealing in depth details on fraudulent transactions.
(Khan, Pathan and Ahmed, 2014)	Hidden Markov Model	• Proposed a Hidden Markov Model (HMM) to model a sequence of credit card transactions, and then use the K-Clustering algorithm to divide transactions into high, medium, and low clusters. • Incoming transactions would then be compared to past 10 transactions performed by the user and authorised only if there is a match.	Get Result

		• Otherwise, transaction is terminated, and the IP address is saved to be traced by the HMM.	
(Carminati *et al.*, 2015)	Decision Support System	• Developed an online fraud detection system that builds models for different customer behaviour on existing transaction data. • The model stores a weight of the anomaly of each transaction, and then uses both unsupervised and semi-supervised techniques to search for other users with comparable patterns.	Get Result
(Renjith, 2018)	Support Vector Machine	• Proposed a Support Vector Machine based algorithm that can be used to build a model that can classify new transactions to fraudulent or non-fraudulent categories. • Feature extraction can be used to transform seller and transaction information into distinctive features, which can then be used to find patterns which can help classify individual transactions. T • The final fraud detection model involves multiple processes, including reputation data, inputs from fraud experts, outputs from set rules engine and prediction based on SVM classification.	

(Altendorf *et al.*, 2005)	Randoms Forests	• Proposed a decision process with an aim to improve the quality of fraud detection while decreasing the need for human analysis. • The random forest algorithm was used to create an ensemble of decision tree classifiers from a dataset containing transaction data in addition to geolocation and network information. • Characteristics found to be associated with fraudulent orders include users who tend to be transient geographically, use non-standard internet connections, have unique purchase pattern timings, and purchase small items that are easily resalable. • These characteristics were deduced through the importance of features such as billing/shipping address, product codes, and number of recent orders.	

Each of the discussed models have unique characteristics that make it advantageous to use in the context of fraud detection, but as with any unsupervised learning method, the performance assessment of the model is mostly subjective. The theme with all these models is the introduction of a method to separate customers based on their behaviours, as this is an effective method of observing suspicious behaviour, and therefore identify fraud. (Jha, Sivasankari and Venugopal, 2020) suggests cluster analysis as a powerful tool to identify fraudulent transactions, as it is a powerful statistical technique used to group customers that exhibit similar behaviours. A heuristic approach can also be used to find patterns and anomalies in big data, which can be used as a base towards a fraud prevention strategy.

Chapter 4 Methodology

Chapter 5

The importance of fraud detection being a binary process presents a challenge to these clustering algorithms, as transaction data can have different patterns for different classes of customers that do not necessarily suggest fraud. For example, a clustering algorithm may suggest five different clusters that exhibit different behaviours, which defeats the aim of splitting the data based on patterns to identify fraud. Accordingly, we consider the Isolation Forest algorithm, which can classify the data into outliers/anomalies and normal data, as it could give a clearer indication of what could be classified as fraud. This algorithm is based on automatic outlier detection, where the data is fed into a model for predictions on whether a data point represents an outlier/anomaly compared to the rest of the dataset. We next present an overview of the models suggested, with a brief discussion of its main advantages and disadvantages. We will then walk through the data pre-processing phase of the project, with description of variables used in addition to relevance to the model.

Choosing the features to feed into the model for training and testing purposes plays a key role in fraud detection (Gadal and Mokhtar, 2017). In other implementations of the isolation forest algorithm, it is simply used to detect outliers in a set of data. In the context of fraud detection, it is important to only consider variables that have been determined to be potential predictors or indicators of fraud. Therefore, based on the literature review and expert input on the potential indicators of fraud, we applied the following steps to arrive at a final dataset to be used for modelling.

1) **Selecting relevant variables:** By assessment of the different tables provided on transaction data in the database, we choose a combination of 16 variables that we believe can have an impact on identification of fraud.
2) **Creating secondary variables:** We manipulate the selected 16 variables to create 8 secondary variables that combine all user transactions and group these values by user.
3) **Data cleaning and Pre-processing:** Apply data cleaning and pre-processing methods to remove any duplicate or incomplete data. This also includes applying a Bayesian smoothing approach to some computations to reduce bias in the model outcome.

These steps were applied on transaction data ranging from the 8th of August 2021 till the 7th of August 2022, identifying 49.3 million consumers that have at least one transaction in that range.

After the implementation of these steps, we arrive at a final dataset containing the following variables for each consumer:

Table 3: Dataset Description

Variable	***Description***
Dollar Return Rate	Sum of Dollars Returned over Sum of Dollars Purchased. A Bayesian based smoothing adjustment is applied to this value to account for the different purchase and return amounts.
Total Number of Returns	Count of Returns.
Total Dollars Returned	Count of Dollars Returned.
Total Dollars per Return	The total dollars returned over the total number of returns.
Number of Unique Stores Visited	Count of the number of unique stores visited. An online store is only counted once regardless of distribution centre, while physical stores are counted based on their unique store number.
Non-Receipted Return Rate	The number of non-receipted returns over the total number of returns.
Count of Boro	Count of transaction bought and returned online.
Count of Boris	Count of transactions bought online and returned in store.

These variables will be explained in more detail in Chapter 5. By applying the different algorithms on this dataset, we can examine the different values of these variables based on the outputted outliers, and therefore determine what values of these thresholds indicates fraud in transaction data. We next present an overview of the models suggested, with a brief discussion of their main advantages and disadvantages. We will then walk through the data preparation phase of the project, with further description of variables used in addition to descriptive statistics on the final dataset used.

5.1 The Isolation Forest Algorithm

The Isolation Forest algorithm is an unsupervised learning algorithm based on the idea of random forests. Data is sub-sampled, and a decision tree is created based on random selection of values of randomly selected variables in the dataset (Hariri, Kind and Brunner, 2021). An anomaly indicator is assigned based on the aggregated lengths of the tree branches, with a point that travels deeper into the tree being less likely to be an anomaly. Sub-sampling is useful in the context of isolating anomalies, as large amounts of data make it more likely for normal instances of data being closer to anomalies, increasing the chances of misclassifying an anomaly as normal. Figure 2 highlights this distinction, as the sub-sample shows a clear gap between anomalies (red dots) and normal data, while the original sample presents anomalies in a way that can blend in with normal data. Repeatedly splitting the tree based on different sub-samples provides a reliable way of isolating these anomalies with minimal misclassification.

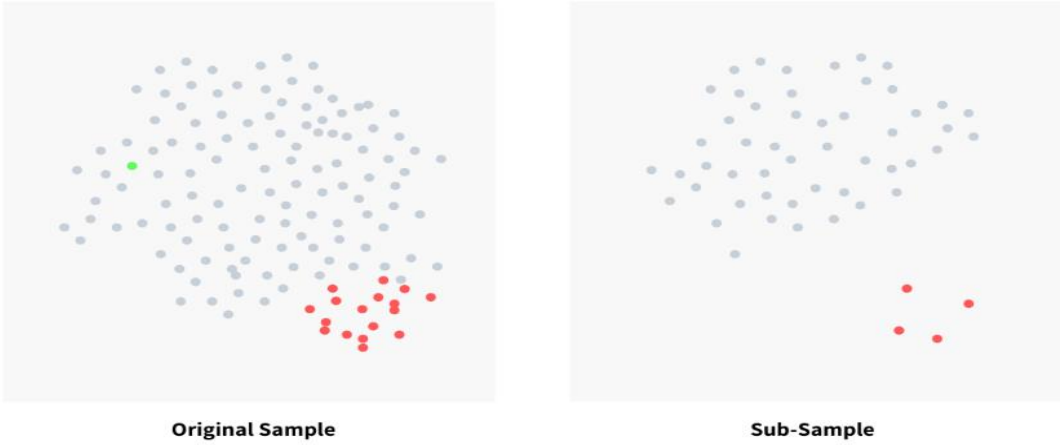

Figure 3: Outlier detection in Original Sample data vs Sub-Sample data

The algorithm begins by selecting a feature from the data, and then randomly selects a split value in the range of this feature. This is done repeatedly until all features are selected and traversed, then repeated for all samples. The anomaly score is then given to each sample based on how many splits the tree goes through, which is then fed through to a decision function that classifies the sample as normal (automatic algorithm assigns 1) or an anomaly (automatic algorithm assigns -1). Figure 3 shows a small sample tree, where grey points represent a split decision value, and yellow points represent the samples.

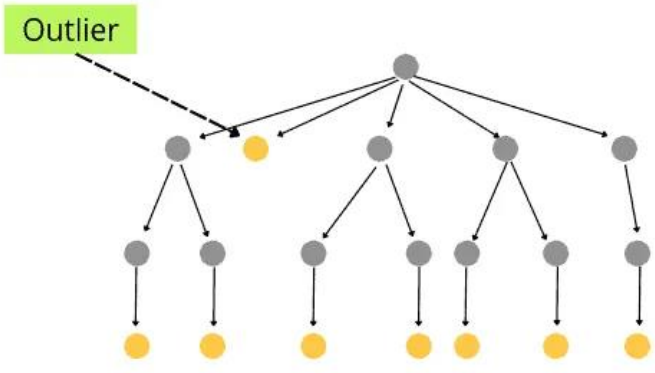

Figure 4: Decision Tree to Isolate Outlier Instances

The sample identified at the first split in the tree is labelled by the algorithm as an outlier while the rest of the samples that required the full length of the tree to be identified are likely to be inliers.

The isolation forest algorithm works on varied sizes of data with different dimensions. However, it is important to consider the parameters used to implement this algorithm, and how these parameters can be selected based on the context of the data provided. The table below summarizes the description of each of the parameters.

Table 4: Isolation Forest Model Parameters

Parameter	Description
Number of Estimators	Refers to the number of base estimators in the ensemble, which simply refers to the number of trees built in the forest.
Max Samples	Refers to the number of samples drawn to train each base estimator.
Max Features	Refers to the number of features used to train each base estimator.
Contamination	Expected percentage of outliers in the dataset.

To implement the model, we need to understand how the selection of model parameters can affect the outcome of our model. The isolation forest algorithm is sensitive to the contamination parameter, which must be selected carefully to avoid overestimating or underestimating the number of fraud cases. According to a study by the National Retail Federation (NRF) in 2021, fraudulent returns are estimated to be approximately 10% of total returns. Since our study is looking to identify fraudulent consumers and not individual fraudulent transactions, we need to use these estimates of fraud returns, in addition to expert input and research outlook on fraud in retail, to identify appropriate contamination parameters to use for our model. Accordingly, and based on an estimate of a final 3.5% of returners being fraudulent, we choose the contamination parameter for our model to reflect a final output that represents 3.5% of the total number of returners. Each consumer in the dataset will be classified as an anomaly or a normal data point. The ration of classifications can be increased or decreased based on how strict a retailer wants the system to be, and what the standards are for acting against fraud. These parameters will be discussed in more detail in Chapter 5, where we present an overview of the implementation of the different models.

5.2 One-Class Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm used mainly for classification problems (Alam, 2020). In the context of supervised learning, the algorithm simply learns the patterns of different variables to assign labels to objects (Noble, 2006). In essence, the SVM algorithm works by separating one class of observations from others in a multi-dimensional space (Alam, 2020). This is straight forward when a dataset has two or more classes, as the algorithm simply looks to maximize a mathematical function with respect to the data used, to arrive at classification for one or more clusters of data (Noble, 2006). In the case of unsupervised learning, an SVM based algorithm can be used to train a model to learn what is normal and what is not. This is referred to as the One-Class Support Vector Machine (One-Class SVM), where all data come from a single class, or all data is not assigned to a class initially. The One-Class SVM algorithm attempts to find a decision boundary or threshold in which the largest separation is achieved between the data points and the origin (Amer, Goldstein and Abdennadher, 2013). The model creates a boundary line or plane depending on the dimensions of the data and will flag any points outside this boundary as outliers or anomalies. Figure 5 demonstrates a simple one-dimensional example of the One-Class SVM. The blue circle represents a boundary in which normal data points fall into, and any data point that falls outside this boundary (the red points) represents an anomaly (Minogue, 2020).

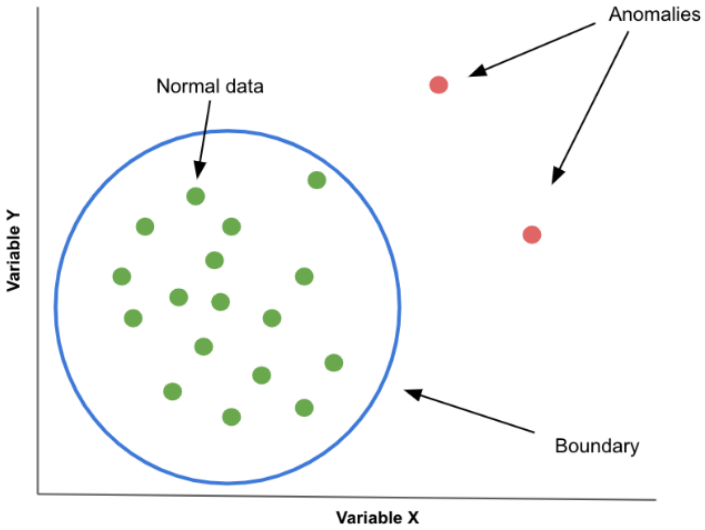

Figure 5: One-Class SVM Example

Using the one-class SVM to create a decision boundary is certainly useful in identifying fraud. The main issue with identifying fraud with confirmed cases of fraud to train the data on is the dependence on subjective or human decision boundary settings. The one-class SVM algorithm can help set a decision boundary by analysing multi-dimensional data and determining what categories of each variable fall outside the decision boundary. Much like the Isolation Forest algorithm, the output decision boundary and outliers detected by the One-Class SVM are highly dependent on setting the different parameters that the model takes. The table below provides a summary of each of these parameters.

Table 5: One-Class Support Vector Machine Model Parameters

Parameter	Description
Kernel	Refers to the kernel type to be used in the algorithm.
Gamma	Refers to kernel coefficient to be used, which defines the possible shape of the decision boundary.
Nu	Refers to the expected percentage of outliers in the dataset.

Similar to the Isolation Forest algorithm implementation, these parameters can be manipulated based on how strict the retailer decides to be on defining a fraudulent consumer. The parameter selection process and implementation of the model will be discussed further in Chapter 5 of this paper.

Chapter 6 Data Preparation and Implementation

After an initial study of the data available, and an understanding of the different techniques used in fraud detection, it is important to prepare the dataset in a way that maximizes the potential of the model. The process of selecting the relevant variables, cleaning the data, identifying bias points and inaccurate data, in addition to finalizing and grouping the data by consumer is a challenge aspect of the project that requires careful consideration. The integrity and accuracy of the final dataset used is vital in determining the usefulness of the model outcome, in addition to understanding how the different variables contribute to the possibility of fraud detection. In this chapter, we will go through the process of selecting and finalizing the dataset in more detail, in addition to how the dataset was prepared for the different models. Finally, we will discuss the implementation of the two proposed models, with emphasis on parameter manipulation.

6.1 Data Preparation and Description

The ability to collect data for individual consumers through every transaction they make is vital in understanding consumer shopping behaviours, trends and patterns. This is important not only in fraud analysis, but also in commercial strategy and business processes. Typical transaction data includes an identifier for the consumer, date, store, transaction amount and type, product type,

in addition to multiple indicators of what this transaction represents. When looking into fraud detection and prevention, we must first understand what type of data we have, and how this data can indicate fraud.

We consider transaction data ranging from the 8th of August 2021 till the 7th of August 2022, as discussed in Chapter 4. This data includes approximately 489 million transactions marked in that period, with two identifiers for every consumer. We select variables from these transactions based on what we believe is relevant to our fraud detection model. This belief is developed through expert input and what research has suggested. These variables include date, transaction amount, transaction type, in addition to 20 other variables. From these variables, we derive the variables presented in Table 1 Chapter 4 for each consumer that has at least one transaction in the considered period. Below is a summary of these variables, their importance in the context of fraud detection, and a basic description of how they are computed.

Table 6: Dataset Variable Description

Variable	***Description and Computation***	***Fraud Indicator***
Dollar Return Rate	Sum of Dollars Returned over Sum of Dollars Purchased. A Bayesian based smoothing adjustment is applied to this value to account for the different purchase and return amounts.	A high return rate could be an indicator of fraudulent behaviour.
Total Number of Returns	Count of every return transaction.	A high number of returns could be an indicator of fraudulent behaviour (Gașpar, 2022).
Total Number of Purchases	Count of every purchase transaction.	A high purchase amount is needed for context when considering the total number of returns (Gașpar, 2022).
Total Dollars Returned	The sum of all return amounts.	A high amount of dollars returned could be an indicator of fraudulent behaviour.

Total Dollars per Return	The total dollars returned over the total number of returns.	A high dollar per return value could indicate fraudulent behaviour (Gașpar, 2022).
Number of Unique Stores Visited	Count of the number of unique stores visited. An online store is only counted once regardless of distribution centre, while physical stores are counted based on their unique store number.	Many stores visited could be considered suspicious behaviour, especially if it is above the number of stores in the city/state of the consumer.
Non-Receipted Return Rate	The number of non-receipted returns over the total number of returns.	A high non-receipted return rate is a major indicator of fraudulent behaviour (Team, 2021).
Count of Boris	Count of transactions bought online and returned in store.	Many items bought online and returned in store can be an indicator of suspicious behaviour.

We derive these variables for every consumer by finding these values based on two identifiers for each consumer. We then arrive at a dataset including what we can consider the transaction history of each consumer, with only the variables relevant to fraud identification. This results in a dataset identifying approximately 62 million consumers. However, due to the linking system in the original database, some duplicates are created. Therefore, we drop all duplicates to ensure that our final dataset has unique values for both consumer identifiers. This means that to ensure data accuracy for our model, we take a sample of data from the original dataset. This results in a dataset containing 49,996,658 consumers with their corresponding 8 variables representing their transaction history.

Since we do not have labelled data, the typical performance measures are not applicable in unsupervised learning methods. Therefore, to answer the question of possible predictors of fraud, we need to understand the normal or expected values of each of these parameters, to compare with the anomalies in the data and understand which variable significantly changes for these anomalies. We first start by looking at data for the different channels. The different channels we consider are:

1) Omnichannel: Contains shoppers that have transactions both in-store and online.
2) Ecommerce: Contain shoppers that only have transactions in ecommerce (online).
3) In-Store: Contains shoppers that only have transactions in-store.

Although our final implementation is targeting omnichannel returns, it would be beneficial to understand the size of each channel, in addition to the typical values of each of these parameters for the different channels. Table 7 looks at different descriptive statistics of different variables for our sample of data. We observe that for the retailer we are considering, in-store returns still present the highest total value of returns, however, the average value of a return for omnichannel is significantly higher than other channels, with the highest for one consumer being $19,399.

Table 7: Data Descriptive Statistics

	Total Dollars Returned	Total Quantity of Returns	Maximum Dollars Returned by 1 Consumer	Maximum Dollars Per Return by 1 Consumer
Omnichannel	$1,089,066,513	14.2M	$125,789	$19,399
Ecommerce	$686,997,606	9.1M	$131,574	$7,499
In-Store	$4,416,516,966	12.5M	$115,175	$11,999

We then consider the total values for some of our derived variables, in addition to further descriptive statistics for each of the channels in Table 8. We find that the return rate is highest for in-store purchases for our sample of data, while the overall return rate for this retailer is 8.22%, which lies within the reported value by various research articles in retail. We also consider the return rate values for returners only, as we are interested in identifying return fraud and therefore, we are more interested in returner behaviour rather than the overall consumer behaviour. We find that the return rate for returners only significantly increases for all channels, with highest return rates seen in In-Store and Ecommerce shoppers, therefore emphasizing the need to assess both channels. We also find that non-receipted returns for returners is highest in omnichannel and In-Store consumers, which is natural as non-receipted returns are not possible in ecommerce.

Table 8: Variable Descriptive Statistics

	Total Consumers	Total Returners	Dollar Return Rate	Average Return Rate (Returners Only)	Average Non-Receipted Return Rate (Returners Only)	Average Number of Returns (Per User Returners Only)
Omnichannel	50.0M	8.2M	8.2%	32.6%	4.0%	7.8
Ecommerce	24.2M	0.61M	2.0%	50.6%	0.0%	3.2
In-Store	50.4M	7.3M	10.3%	54.5%	4.5%	6.7

We then look at the distributions for different variables, to develop an idea of where fraud is more likely to occur, and what percentage of consumers fall into categories where these numbers are high enough to raise suspicions. Figure 6 shows the distribution for omnichannel returns, where we can clearly see that most consumers have 0 returns, and as the number of returns increases, the percentage of consumers decreases. This trend is similar for both in-store and ecommerce consumers, as shown in Figure 7 and Figure 8.

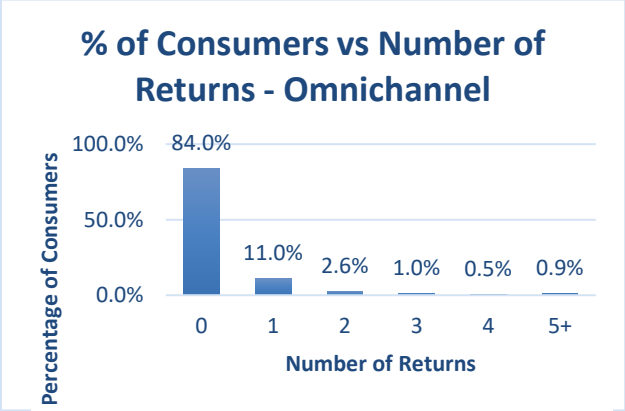

Figure 6: Omnichannel Distribution of Returns

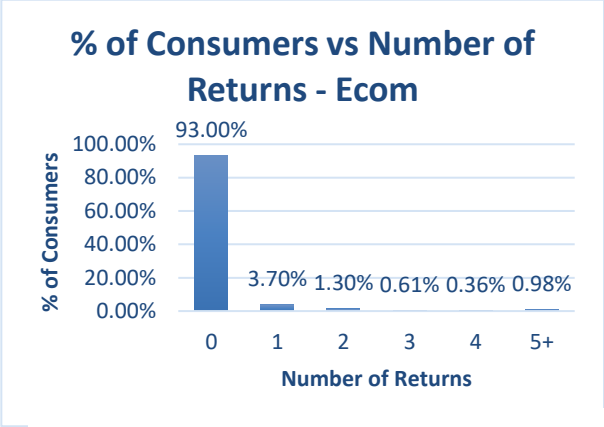

Figure 7: Ecommerce Distribution of Returns

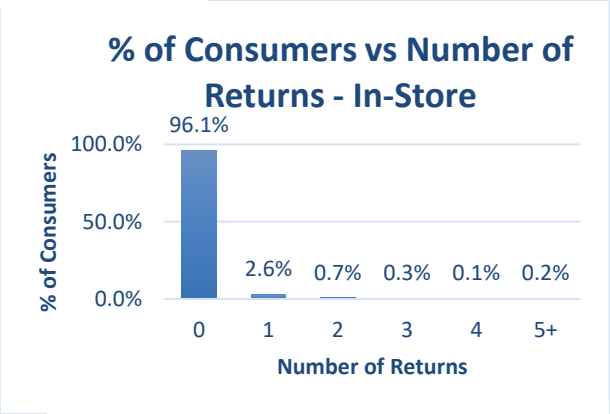

Figure 8: In-Store Distribution of Returns

We also notice that for all three channels, the percentage of consumers that have more than 5 returns represents 1% or less of the total, making it a clear target for us as a threshold value where fraud can occur.

We consider the distribution for the number of stores visited, to understand how many stores the average consumer visits, in addition to where most consumers fall within the distribution. Figure 9 and Figure 10 show the distribution for Omnichannel and In-Store consumers respectively, as ecommerce shoppers are shopping in one store only.

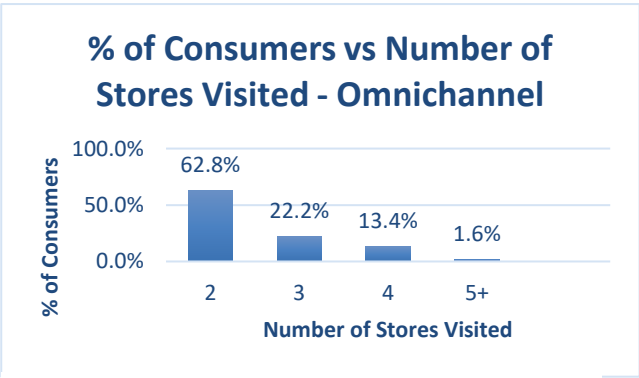

Figure 10: Omnichannel distribution of Number of Stores

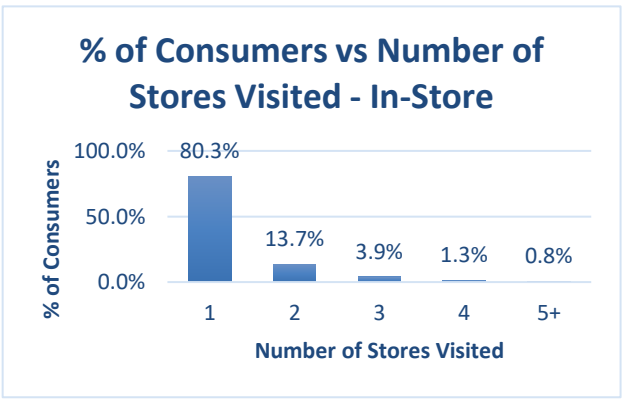

Figure 9: In-Store distribution of Number of Stores

We also observe that most consumers visit one store only for In-Store shopping, and two stores for omnichannel. The percentages decrease as the number of stores visited increases, with a threshold of 5 or more stores visited representing the minority of consumers.

As for dollar return rate, the average value of 8.2% is low since most consumers do not have any returns, and hence a return rate of 0 is most common (87% of all consumers). We demonstrate this by visualizing the distribution of the dollar return rate by percentage of consumers for omnichannel consumers in Figure 11.

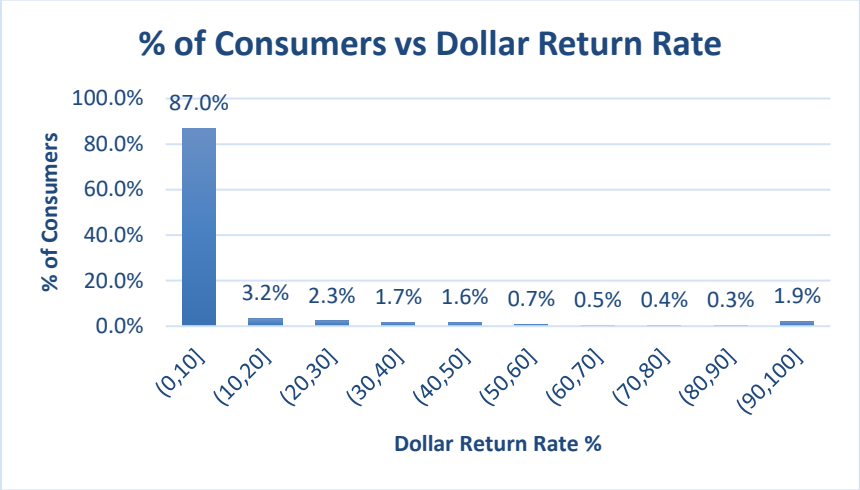

Figure 11: Return Rate Distribution - Full Data

Since we are looking for indicators of fraud, we are only interested in finding outliers in returners data, and so we consider two cases:

1) **Returners Data only:** This simply takes all consumers that have at least one return. We consider this to find the average dollar return rate and other parameters for any consumer with a return, regardless of the number of returns.
2) **Returners with 5 or more returns:** We use this dataset to feed into our models. This is done based on the data exploration and preliminary analysis conducted, as we found consumers with 5 or more returns to represent a large minority in the data. Furthermore, a retailer is unlikely to act against a consumer with a low number of returns, and therefore, we would only be interested in outliers that have unusual behaviour above a threshold of total returns.

Figure 12 and Figure 13 demonstrate the dollar return rate for cases 1 and 2 respectively. We can see that for returners data, the percentage of consumers decreases as the dollar return rate increases, signalling to a possibility of more anomalies found for higher return rates. This

is also demonstrated in Figure 13, as around 50% of consumers have a return rate between 10% and 30%, while return rates higher than 70% represent a minority of consumers. We cannot base our anomaly detection model based on this parameter only, however, this shows a clear indication of where anomalies will be found, and how dollar return rate could be a strong predictor of fraudulent behaviour.

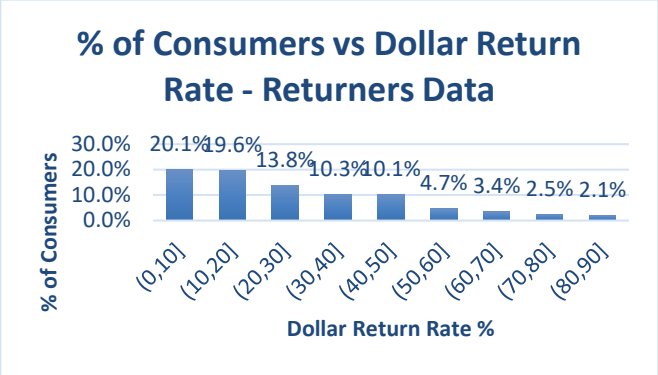

Figure 12: Return Rate Distribution - Returners Data

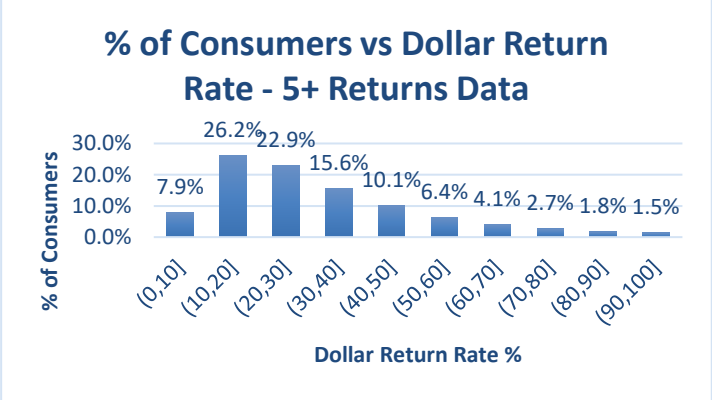

Figure 13: Return Rate Distribution – 5+ Returns Data

The total dollars returned also plays a key role in determining the possibility of a consumer falling within a minority of the data. Figure 14, Figure 15 and Figure 16 represent the distribution of consumers by total dollars returned in the full data, returners data, and 5 or more returns data respectively. Figure 14 shows that only 0.1% consumers have total dollars returned over $1,000. Similarly, only 0.9% have total dollars returned over $1,000 for returners only. For consumers with 5 or more returns, the total dollars return naturally increases as the number of returns considered is higher than other datasets, but we still find that most consumers are within the $250 to $1,000 range. This gives us an idea of what is likely to be classified as an anomaly, as the expected value for these parameters is found through these distributions.

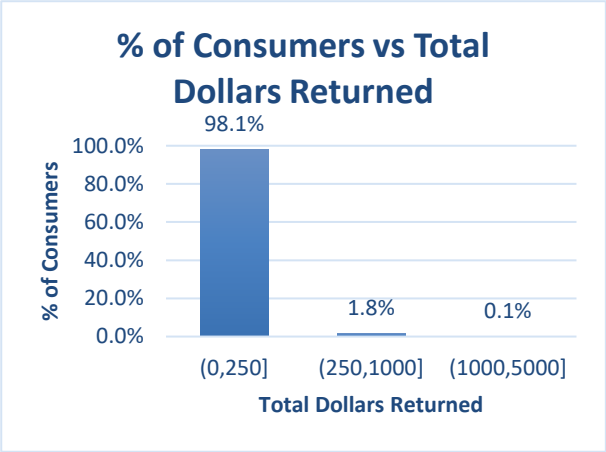

Figure 14: Dollars Returned Distribution - Full Data

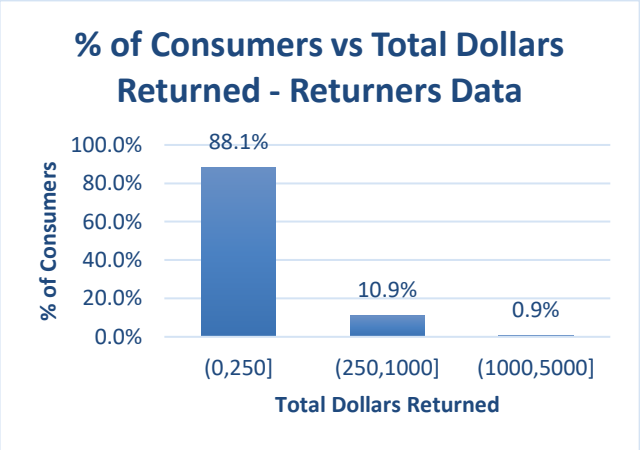

Figure 15: Dollars Returned Distribution – Returners Data

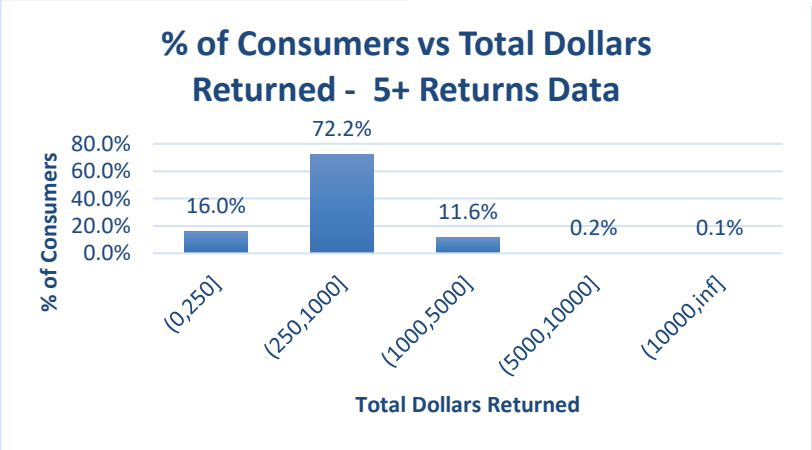

Figure 16: Dollars Returned Distribution – 5+ Returns Data

By completing the data exploration of the different datasets generated, we have developed an idea of what fraud patterns can look like, and what values are more likely to present a fraudulent consumer. We next discuss the implementation of this data on the two different models, with a discussion on how parameters were selected based on these descriptive statistics observed.

6.2 Implementation

The implementation of the two different proposed models is similar as both look to locate anomalies within the dataset. The isolation forest algorithm does this based on the idea of random forests, while the one-class SVM does this based on the idea of support vector machines. Therefore, since these models acts similarly, we can feed both models the same

dataset containing consumers with 5 or more returns. Both models exhibit general advantages and disadvantages in the context of anomaly detection. Table 9 summarises this with emphasis on how this affects our model and aim to identify fraud through anomaly detection.

Table 9: Model Comparison - Advantages and Disadvantages

	Advantages	Disadvantages
Isolation Forest	• Low time complexity. • Can provide quantitative description of anomalies. • Can handle high-dimensional data with irrelevant attributes.	• Less accurate in detecting local anomaly points (Gao *et al.*, 2019). • Largely dependent on the contamination parameter, hence a prior must be known on the data.
One-Class Support Vector Machine	• Works well for a clear margin of separation between normal behaviour and anomalous behaviour. • Effective in high dimensional space. • Memory efficient	• High computational complexity. • Does not work well for data that has overlapping classes.

We base our choice of model parameters on our dataset. Since our desired output is approximately 3.5% of all returners are fraudulent, and based on the assumption that fraud can be found in returners with over 5 returns, we choose the parameters for each of the models as follows:

1) Isolation Forest

Table 10: Isolation Forest Parameter Selection

Parameter	Value	Justification
Number of Estimators	100	Keep default value since it has little to no effect on our model output.
Max Samples	256	Keep default value since it has little to no effect on our model output.
Max Features	1	We assume that each of our variables can have a contribution to fraud detection, and hence we need to train based on 1 feature.
Contamination	0.05	This would output a total of 21,288 consumers as fraudulent, equivalent to 3.5% of all returners, which is the desired output.

2) One-Class Support Vector Machine

Table 11: One-Class Support Vector Machine Parameter Selection

Parameter	Value	Justification
Kernel	Radial Basis Function (RBF)	Default Kernel function used for One-Class SVM
Gamma	Reciprocal of the number of features	Optimal value for the RBF kernel function.
Nu	0.05	This would output a total of 21,288 consumers as fraudulent, equivalent to 3.5% of all returners, which is the desired output.

Both models will be implemented on the dataset containing consumers with 5 or more returns. We will next discuss results for both models to find out the shortcomings of each model in terms of identifying anomalies relating to fraud detection, in addition to how successful each model is at answering our research questions.

Chapter 7 Results

Based on our research questions, our emphasis is on understanding how these models can be used to understand potentially fraudulent patterns in consumer transaction data, in addition to understanding the weight of the different features on anomaly classification. To answer these questions, we consider the output of our models as new independent datasets to analyse the different distributions and characteristics of each of the features, in addition to some summary data on the value of consumers identified through these models. We will first present these for each model, and then choose the model that we find to be most compatible with our learned definition of fraud to present 5 unique cases of consumers identified as anomalies through different patterns.

The fact that we are dealing with an unsupervised learning method makes model analysis more challenging. Since there is no basis to consider accuracy measures, it is important to compare the output of our model to the original dataset provided, and to compare the different output classes.

Therefore, the anomalies are labelled by the model as -1, while non-anomalies are labelled with 1. To emphasize the size of the problem, and to demonstrate the effect of these unusual returns, we analyse the numbers for all anomalies compared to the original dataset containing all consumers with returns. Furthermore, we consider the top 1% anomalies, to show how only the slightest action against the most suspicious consumers can lead to a large change in profits and a greater understanding of fraud in transaction data.

7.1 Isolation Forest Model Results

Table 12: Isolation Forest Model Results

Feature	Top 1% Anomalies	Total Anomalies	Total Non-Anomalies	5 or more Returns Data
Total Dollars Returned	$5,100,714	$38,774,372	$218,211,852	$256,986,225
Total Number of Returns	17,810	336,750	3,024,497	3,361,247
Average Dollars Per Return	$286	$115	$72	$76
Average Number of Returns Per Consumer	84.0	22.6	7.4	7.8
Average Return Rate	79.0%	73.0%	31.1%	32.6%

Average Non-Receipted Return Rate	2.9%	9.2%	3.8%	4.0%
Average Number of Stores Visited	6.6	5.3	3.8	3.8

Table 12 shows these values for the different categories analysed. By observing the different features, we can clearly see how the return rate significantly increases for anomalies. This shows an approximate 47% increase in return rate compared to the full data, and an approximate 48% change in return rate between anomalies and non-anomalies. The non-receipted rate also shows an increase in total anomalies, however, there is no increase in the top 1% anomalies, meaning that is not as predictive or anomalies or fraud as the return rate. We also consider the top 1% of anomalies, which represents only 212 returners out of a total 426,388 returners. These 212 consumers account for just over $5 million in returns, representing around 2% of the total dollars returned. Furthermore, although anomalies only represent 3.5% of all returners with 5 or more returns based on our fraud percentage prediction, the total dollars returned by these anomalies represent approximately 15% of all dollars returned. This is also true for the average number of returns per consumer, as we can see a substantial increase between anomalies and non-anomalies, with top 1% anomalies averaging over 10 times the number of returns of non-anomalies. The number of stores visited also realizes an increase for anomalies, with just under a 50% increase in number of stores visited for the top 1%.

We next present five different cases showing the ability of the model to identify suspicious behaviour through different features.

7.1.1 Potentially Fraudulent Cases

Table 13: Isolation Forest Fraud Cases

Case Number	Dollar Return Rate	Number of Purchases	Number of Returns	Number of Stores Visited	Total Dollars Returned	Dollars per Return	Non-Receipted Return Rate
Case 1	99.9%	22	21	11	$13,139	$625	4.8%
Case 2	105.6%	44	38	15	$93,393	$2,458	2.6%
Case 3	82.5%	176	101	9	$10,487	$104	4.0%
Case 4	88.3%	80	60	26	$15,882	$265	71.7%
Case 5	69.8%	32	11	4	$119,648	$10,877	9.1%

The cases in Table 13 represent five different anomalies out of the total 212 anomalies detected. We choose these five cases as they represent significant value of returns, in addition to different features indicating fraud. We summarise the indicators for each case below.

1) **Case 1:** This case presents a consumer with a large number of returns and an approximately unit return rate. This consumer could be classified into the wardrobing category of fraud, or a return abuser, as they seem to be returning everything they are purchasing. Although the non-receipted rate is around the average for all returners, the model identifies this case as anomaly through a large return rate and a large average dollars per return.
2) **Case 2:** This case presents a consumer with a return rate over 100%. This could happen due to transactions outside of the period considered, or due discounted returns and non-receipted returns. Although their non-receipted return rate is well below the average, this consumer also seems to be returning most of their purchases, with a significantly high amount for dollar per return.
3) **Case 3:** This case presents a unique consumer in terms of purchase and return quantities. While their dollars per return and non-receipted rate are within the range of normal behaviour, their number of returns, in addition to total dollars returned and return rate present unusual behaviour.
4) **Case 4:** This case simply presents the models ability to identify cases through different features. While the return amounts and quantities are high, this case presents an abnormally large non-receipted rate, a clear indicator of fraudulent returns. This also shows a large number of stores visited, a very unusual aspect of a consumer's shopping behaviours as the typical consumer visits an average of 4 stores.
5) **Case 5:** This case also presents the models diversity in identifying anomalies. While the return rate and non-receipted rate are not substantially higher than the returners average, this case presents an abnormally high dollar per return value of $10,877, which is well over 100 times larger than the average dollar per return for a returner.

7.2 One-Class Support Vector Machine Model Results

Table 14: One-Class SVM Model Results

Feature	Top 1% Anomalies	Total Anomalies	Total Non-Anomalies	5 or more Returns Data
Total Dollars Returned	$4,906,134	$33,215,574	$223,771,651	$256,986,225
Total Number of Returns	15,117	293,892	3,067,355	3,361,247
Average Dollars Per Return	$324	$113	$73	$76
Average Number of Returns Per Consumer	83.0	22.1	7.6	7.8
Average Return Rate	77.2%	69.1%	33.4%	32.6%
Average Non-Receipted Return Rate	3.7%	7.4%	3.1%	4.0%
Average Number of Stores Visited	8.2	5.8	3.1	3.8

Table 13 shows the output of the One-Class Support Vector Machine model, with statistics of the different features for the total anomalies, top 1% anomalies, total non-anomalies and the full dataset. We find that the one-class SVM performs almost identically to the isolation forest model, with values increasing and decreasing in parallel. One of the few differences we find is that one-class SVM finds a higher number of stores visited for anomalies compared to the isolation forest model. The total dollars returned is also similar for both models, however, the one-class SVM shows a $5 million decrease in total dollars returned for anomalies, meaning that the isolation forest algorithm provides more weight for the value of returns than the one-class SVM.

On investigation of cases for the one-class SVM, we find that the top 1% anomalies have an approximate 93% similarity between the two models, with 197 cases in common between the top 1% of both models. This decreases to 84% when considering all anomalies, however, these values are considered high enough to conclude that the performance of these models is similar, and the anomaly detection for both models works in a way that identifies most consumers that represent a major value of returns. The one-class SVM model is also able to identify different cases of potentially fraudulent consumers through different features, with all five cases presented for the isolation forest model identified by the one-class SVM model.

Chapter 8 Conclusion

In this chapter, we first start by summarising our research and results. We then discuss the limitations of our work, with a discussion on which direction future fraud detection research could be leading.

8.1 Summary

We have proposed an unsupervised learning-based approach to fraud detection and prevention. While fraud detection framework generally looks at confirmed cases of fraud, and therefore can extract patterns related to fraudulent behaviour, it is important to develop an idea of how to implement a fraud prevention solution in cases where data is not labelled, through analysis of consumer behaviour. The mass amounts of data available on consumer shopping behaviour in retail opens plenty of opportunities for data analysis to find useful insights and patterns that could drive business performance. In the context of fraud detection, most fraudulent consumers are left unidentified due to the lack of careful attention to the details in these data, in addition to flexible return policies that are not closely monitored and evaluated. The delegation of fraud

identification and prevention to fraud experts required massive resources and time-consuming analysis, which would likely lead to only the rarest cases of fraud to be identified. By implementing an unsupervised learning model, we can evaluate consumer behaviour in real-time, with immediate reporting on blocking, warning or allowing a consumer to continue with their purchases. Through analysis of our model results, in addition to research in retail fraud and expert inputs, we can answer our research questions as follows:

1) **What are the variables that could be predictive fraud in transaction data?**

There exist multiple indicators of suspicious or fraudulent behaviour through transaction data. We proposed 8 different derived variables to identify some potentially indicators of fraud. Summarised below are our findings on these variables:

Table 15: Model Variables and Fraud Indication

Variable	***Fraud Indication***
Dollar Return Rate	By analysis of output anomalies in our data, we find that the risk of unusual or fraudulent behaviour increases as the dollar return rate increases.
Total Number of Returns	We find that the total number of returns presents an important indicator to consumers that are likely to commit fraud.
Total Number of Purchases	We find that the total purchases alone have no correlation with suspicious behaviour, but grouped with number of returns, can be an indicator of fraud.
Total Dollars Returned	Total Dollars Returned is a significant indicator of fraud and is also one of the variables that must be closely monitored to identify consumers with negative effects on profit through high reverse logistics costs.
Total Dollars per Return	This is also important in identification of fraud, as high dollars per return can suggest unusual consumer behaviour.
Number of Unique Stores Visited	This also presents one of the unusual indicators of fraud. While this variable alone cannot be used to determine fraud, a consumer with a high number of stores visited is certainly worth further investigation.
Non-Receipted Return Rate	Even though non-receipted returns are one of the major indicators of fraud, we find that it does not bear as big of a weight as other variables, however, it still acts as an important indicator of fraudulent behaviour.

Count of Boris	We find that the count of transactions bought online and returned in store has no direct implications on fraud through our models. However, further investigation into cases with high Boris value could lead to a different conclusion.

2) **Amongst the considered unsupervised learning models, how are outliers selected differently?**

As discussed in Chapter 6.2, both models select cases in a similar way. This strengthens our case for identifying fraud through these variables. However, testing other models on the same data to examine the different outliers found by other models could be beneficial in enhancing the model and allowing for a more diverse identification of fraud.

8.1.1 Recommendations

Based on our model outputs and research conducted, we recommend the following to combat fraud in retail:

1) Identify the existing top 1% consumers classified as fraudulent and conduct further investigation into each of their cases. Assuming these cases can be confirmed as fraudulent, this could save the retailer over $5 million in returns, in addition to the cost of reverse logistics incurred through these returns.
2) Implement a model that works in real time to identify a fraudulent consumer before their transaction goes through. This would mean implementing an approval process based on the features presented, with approval, rejection or warning on a return transaction issued based on consumer transaction history.

8.2 Limitations

We summarise the limitations of our research and models as follows:

1) The major limitation of our proposed method of identification of fraud is the same limitation that occurs for any unsupervised learning model; it lacks quantitative performance measures. This means that any analysis conducted on results will lack a definitive outcome on how the model is performing, and part of the analysis remains subjective. While this cannot be solved with the available dataset, the possibility of implementing our dataset on different models, and sharing our results with multiple

experts can be an approach towards obtaining a more confident result through unsupervised learning.

2) Most research on fraud detection framework looks at symptoms of fraud in retail, rather than confirmed cases of fraud. This makes most available research similar in terms of identifying different types of fraud, their behavioural patterns and their effect on business performance. This means that there is a subjective view of fraud through most retailers, that biases that identification of these patters as fraudulent.
3) The contamination parameter in the isolation forest algorithm and the nu parameter in the one-class SVM algorithm are both based on an estimate of fraud in the data, and both play a significant role in the model outcome. Therefore, the accuracy of these estimates has a vital impact on what is identified as fraud, and therefore a more accurate estimate is needed to present these results.

8.3 Future Research

Based on the limitations faced, in addition to research conducted, we propose the following areas of research as vital to the process of developing a greater view of fraud detection in retail:

1) Considering retailers that can label their data as fraudulent through business processes that identify and act against fraud. This could lead to labelled datasets on transaction data in retail, which could verify the results of unsupervised learning implementations of fraud detection, in addition to allowing for the implementation of supervised learning approaches, resulting in quantitative performance measures for all models.
2) Conduct further research into how these model parameters can be selected for unique characteristics of features in different datasets. Studying the effect of the number of features on the isolation forest model output, in addition to the kernel used in the one-class SVM model, could lead to a better understanding of how these decision boundaries are set, and how they can be altered in a way that maximizes the value of identified fraud.

List of References

Abbey, J., Ketzenberg, M. and Metters, R. (2018) 'A More Profitable Approach to Product Returns', *MIT Sloan Management Review*, 60(1), pp. 1–6.

Adewumi, A.O. and Akinyelu, A.A. (2017) 'A survey of machine-learning and nature-inspired based credit card fraud detection techniques', *International Journal of System Assurance Engineering and Management*, 8(2), pp. 937–953. Available at: https://doi.org/10.1007/s13198-016-0551-y.

Akhilesh, K. and Swapnil, S. (2017) *Supply Chain Management Strategies and Risk Assessment in Retail Environments*. IGI Global.

Alam, M. (2020) *Support Vector Machine (SVM) for Anomaly Detection*, *Medium*. Available at: https://towardsdatascience.com/support-vector-machine-svm-for-anomaly-detection-73a8d676c331 (Accessed: 14 September 2022).

Altendorf, E. *et al.* (2005) 'Fraud Detection for Online Retail using Random Forests'.

Amasiatu, C.V. and Shah, M.H. (2018) 'First party fraud management: framework for the retail industry', *International Journal of Retail & Distribution Management*, 46(4), pp. 350–363. Available at: https://doi.org/10.1108/IJRDM-10-2016-0185.

Ambilkar, P. *et al.* (2022) 'Product returns management: a comprehensive review and future research agenda', *International Journal of Production Research*, 60(12), pp. 3920–3944. Available at: https://doi.org/10.1080/00207543.2021.1933645.

Amer, M., Goldstein, M. and Abdennadher, S. (2013) 'Enhancing one-class support vector machines for unsupervised anomaly detection', in *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description - ODD '13*. *the ACM SIGKDD Workshop*, Chicago, Illinois: ACM Press, pp. 8–15. Available at: https://doi.org/10.1145/2500853.2500857.

Anderson, E.T., Hansen, K. and Simester, D. (2009) 'The Option Value of Returns: Theory and Empirical Evidence', *Marketing Science*, 28(3), pp. 405–423. Available at: https://doi.org/10.1287/mksc.1080.0430.

Balasupramanian, N., Ephrem, B.G. and Al-Barwani, I.S. (2017) 'User pattern based online fraud detection and prevention using big data analytics and self-organizing maps', in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. *2017*

International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), pp. 691–694. Available at: https://doi.org/10.1109/ICICICT1.2017.8342647.

Bank, E.C. (2013) *ECB releases final Recommendations for the security of internet payments and starts public consultation on payment account access services*, *European Central Bank*. Available at: https://www.ecb.europa.eu/press/pr/date/2013/html/pr130131_1.en.html (Accessed: 17 August 2022).

Bao, Y., Hilary, G. and Ke, B. (2022) 'Artificial Intelligence and Fraud Detection', in V. Babich, J.R. Birge, and G. Hilary (eds) *Innovative Technology at the Interface of Finance and Operations: Volume I*. Cham: Springer International Publishing (Springer Series in Supply Chain Management), pp. 223–247. Available at: https://doi.org/10.1007/978-3-030-75729-8_8.

Bhavsar, H. and Ganatra, A. (2016) 'A Comparative Study of Training Algorithms for Supervised Machine Learning'.

Bower, A.B. and Maxham, J.G. (2012) 'Return Shipping Policies of Online Retailers: Normative Assumptions and the Long-Term Consequences of Fee and Free Returns', *Journal of Marketing*, 76(5), pp. 110–124. Available at: https://doi.org/10.1509/jm.10.0419.

Burkart, N. and Huber, M.F. (2021) 'A Survey on the Explainability of Supervised Machine Learning', *Journal of Artificial Intelligence Research*, 70, pp. 245–317. Available at: https://doi.org/10.1613/jair.1.12228.

Caldeira, E. *et al.* (2012) 'Characterizing and Evaluating Fraud in Electronic Transactions', in *2012 Eighth Latin American Web Congress*. *2012 Eighth Latin American Web Congress*, pp. 115–122. Available at: https://doi.org/10.1109/LA-WEB.2012.16.

Carminati, M. *et al.* (2015) 'BankSealer: A decision support system for online banking fraud analysis and investigation', *Computers & Security*, 53, pp. 175–186. Available at: https://doi.org/10.1016/j.cose.2015.04.002.

Chandola, V., Banerjee, A. and Kumar, V. (2009) 'Anomaly detection: A survey', *ACM Computing Surveys*, 41(3), p. 15:1-15:58. Available at: https://doi.org/10.1145/1541880.1541882.

Frei, R., Jack, L. and Brown, S. (2020) 'Product returns: a growing problem for business, society and environment', *International Journal of Operations & Production Management*, 40(10), pp. 1613–1621. Available at: https://doi.org/10.1108/IJOPM-02-2020-0083.

Gadal, S.M.A.M. and Mokhtar, R.A. (2017) ‘Anomaly detection approach using hybrid algorithm of data mining technique’, in *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*. *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, pp. 1–6. Available at: https://doi.org/10.1109/ICCCCEE.2017.7867661.

Gao, R. *et al.* (2019) ‘Research and Improvement of Isolation Forest in Detection of Local Anomaly Points’, *Journal of Physics: Conference Series*, 1237(5), p. 052023. Available at: https://doi.org/10.1088/1742-6596/1237/5/052023.

Gașpar, D. (2022) *Return Fraud Signs & Best Practices on How To Avoid Fraudulent Returns, WeSupply | Labs*. Available at: https://wesupplylabs.com/return-fraud-signs-best-practices-on-how-to-avoid-fraudulent-returns/ (Accessed: 14 September 2022).

Hariri, S., Kind, M.C. and Brunner, R.J. (2021) ‘Extended Isolation Forest’, *IEEE Transactions on Knowledge and Data Engineering*, 33(4), pp. 1479–1489. Available at: https://doi.org/10.1109/TKDE.2019.2947676.

Harris, L.C. and Daunt, K.L. (2011) ‘Deviant customer behaviour: A study of techniques of neutralisation’, *Journal of Marketing Management*, 27(7–8), pp. 834–853. Available at: https://doi.org/10.1080/0267257X.2010.498149.

Jha, B.K., Sivasankari, G.G. and Venugopal, K.R. (2020) ‘Fraud Detection and Prevention by using Big Data Analytics’, in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 267–274. Available at: https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00050.

John, S., Shah, B.J. and Kartha, P. (2020) ‘Refund fraud analytics for an online retail purchases’, *Journal of Business Analytics*, 3(1), pp. 56–66. Available at: https://doi.org/10.1080/2573234X.2020.1776164.

Khan, M.Z., Pathan, J.D. and Ahmed, A.H.E. (2014) ‘Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering’, 3(2), p. 4.

King, T., Dennis, C. and McHendry, J. (2007) ‘The management of deshopping and its effects on service: A mass market case study’, *International Journal of Retail & Distribution Management*, 35(9), pp. 720–733. Available at: https://doi.org/10.1108/09590550710773264.

Kranz, J., Urbanke, P. and Kolbe, L. (2015) *Predicting Product Returns in E-Commerce: The Contribution of Mahalanobis Feature Extraction*.

Lopez-Rojas, E.A. (2015) 'Using the RetSim simulator for fraud detection research', *International Journal of Simulation and Process Modelling*, 10(2). Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-12930 (Accessed: 26 June 2022).

Malphrus, S. (2009) 'Perspectives on Retail Payments Fraud'. Rochester, NY. Available at: https://papers.ssrn.com/abstract=1341233 (Accessed: 26 June 2022).

Minogue, P. (2020) *Anomaly detection - can it help in contact centre management?*, *EdgeTier*. Available at: https://www.edgetier.com/blog/anomaly-detection-can-it-help-in-contact-centre-management/ (Accessed: 14 September 2022).

Mollenkopf, D.A., Frankel, R. and Russo, I. (2011) 'Creating value through returns management: Exploring the marketing-operations interface', *Journal of Operations Management*, 29(5), pp. 391–403. Available at: https://doi.org/10.1016/j.jom.2010.11.004.

Mustika, N.I., Nenda, B. and Ramadhan, D. (2021) 'Machine Learning Algorithms in Fraud Detection: Case Study on Retail Consumer Financing Company', *Asia Pacific Fraud Journal*, 6(2), pp. 213–221. Available at: https://doi.org/10.21532/apfjournal.v6i2.216.

Noble, W.S. (2006) 'What is a support vector machine?', *Nature Biotechnology*, 24(12), pp. 1565–1567. Available at: https://doi.org/10.1038/nbt1206-1565.

Petersen, J.A. and Kumar, V. (2009) 'Are Product Returns a Necessary Evil? Antecedents and Consequences', *Journal of Marketing*, 73(3), pp. 35–51. Available at: https://doi.org/10.1509/jmkg.73.3.035.

Potdar, A. and Rogers, J. (2010) 'Methodology to forecast product returns for the consumer electronics industry', in *PICMET 2010 TECHNOLOGY MANAGEMENT FOR GLOBAL ECONOMIC GROWTH*. *PICMET 2010 TECHNOLOGY MANAGEMENT FOR GLOBAL ECONOMIC GROWTH*, pp. 1–11.

Renjith, S. (2018) 'Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach', *International Journal of Engineering Trends and Technology*, 57(1), pp. 48–53. Available at: https://doi.org/10.14445/22315381/IJETT-V57P210.

Ribeiro, R.P., Oliveira, R. and Gama, J. (2016) 'Detection of Fraud Symptoms in the Retail Industry', in M. Montes y Gómez et al. (eds) *Advances in Artificial Intelligence - IBERAMIA 2016*.

Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 189–200. Available at: https://doi.org/10.1007/978-3-319-47955-2_16.

Robertson, T.S., Hamilton, R. and Jap, S.D. (2020) 'Many (Un)happy Returns? The Changing Nature of Retail Product Returns and Future Research Directions', *Journal of Retailing*, 96(2), pp. 172–177. Available at: https://doi.org/10.1016/j.jretai.2020.04.001.

Shih, D.-H. *et al.* (2021) 'Preventing Return Fraud in Reverse Logistics—A Case Study of ESPRES Solution by Ethereum', *Journal of Theoretical and Applied Electronic Commerce Research*, 16(6), pp. 2170–2191. Available at: https://doi.org/10.3390/jtaer16060121.

Speights, D. (2013) 'Return Fraud and Abuse: How to Protect Profits', p. 6.

Team, I. (2021) *Breaking Down Return Fraud and Ways To Prevent It, Intellicheck*. Available at: https://intellicheck.com/breaking-down-return-fraud-and-ways-to-prevent-it/ (Accessed: 14 September 2022).

UK Online retail sales 2012-2021 (2021) *Statista*. Available at: https://www.statista.com/statistics/315506/online-retail-sales-in-the-united-kingdom/ (Accessed: 2 August 2022).

Ülkü, M.A., Dailey, L.C. and Yayla-Küllü, H.M. (2013) 'Serving Fraudulent Consumers? The Impact of Return Policies on Retailer's Profitability', *Service Science*, 5(4), pp. 296–309. Available at: https://doi.org/10.1287/serv.2013.0051.

Wachter, K. *et al.* (2012) 'Exploring consumer orientation toward returns: unethical dimensions', *Business Ethics: A European Review*, 21(1), pp. 115–128. Available at: https://doi.org/10.1111/j.1467-8608.2011.01639.x.

Whitehead, E. (2021) 'Why e-commerce attracts fraud', *Computer Fraud & Security*, 2021(10), pp. 6–7. Available at: https://doi.org/10.1016/S1361-3723(21)00106-8.

Young |, J. (2021) *US ecommerce grows 14.2% in 2021*, *Digital Commerce 360*. Available at: https://www.digitalcommerce360.com/article/us-ecommerce-sales/ (Accessed: 9 August 2022).