The University of Southampton

Academic Year (2021/2022)

Faculty Of Social Science

Southampton Business School

MSc Dissertation

Modelling Product Returns Behaviour at Level of Consumers

(ERGO reference number:76487)

(Student registration number:32865023)

Presented for MSc. Business Analytics and Finance

This project is entirely the original work of student registration number 32865023. Where material is obtained from published or unpublished works, this has been fully acknowledged by citation in the main text and inclusion in the list of references

Word Count:14866

Contents

Abstract

Due to the continuous increase of online sales, the number of product returns has also increased significantly, which has also generated corresponding return costs. For retailers, the factors affecting the return of products need to be clear, so that timely responses and measures should be taken. From the perspective of demographic characteristics, this study explores the impact of demographic characteristics, including purchasing power and average age, on product returns. In other words, at the customer level, what characteristics of customers are more inclined to return. In the process of data analysis, we use Logical Regression and Naive Bayes, which are suitable for binary classification problems, and Neural Networks that can perform well in various analyses. We use these three models to predict the two independent variables, purchasing power and average age, and the dependent variables (binary data: 1 means return, 0 means no return). The results suggest that these two features do have an impact on product return.

Keywords**:** Product Return, Demographic Characteristics, Customer, Machine Learning

1. Introduction

Nowadays, customers can buy any product anywhere as online sales grows increasingly in these years. From 2017 to 2021, online sales grew up from £295.9 billion to £465.4 billion in Europe (Online Shopping Behavior in The United Kingdom (UK),2021). The e-commerce market has increased rapidly the last couple of years. E-commerce players offer free shipping and hassle-free returns because of the high competition in e-commerce, and because of this, the number of product returns increased dramatically. Compared with offline shopping, online shopping cannot provide the real product so that when making a purchase decision, a customer cannot physically inspect, feel or touch a product to resolve issues associated with it. (Yan and Cao, 2017). In order to discover if a product is of interest to them, they may review the product description, view an image of the product, or rate the product on the website. Moreover, customers may be able to realize the value of a product and avoid the purchase of an unsatisfactory product as a result of product quality or product performance (Yoo and Seung, 2014). However, it is still a possibility to buy an unsuitable product to return it. Most retail businesses provide free delivery and multiple ways for returning items - including returning items bought online to a physical store without charge. In the Financial Times (Ram, 2016), it was reported that certain businesses reported up to 70% returns, which is higher than the typical returns rate of 20–40%.

Due to escalating rates of product returns, most firms have no choice but to incur these extra costs. Online purchasing has resulted in an increase in product returns, which show no signs of slowing down. There is no doubt that product returns lead to many operation and profit problems. For example, transaction costs, repair costs, supply chain management, inventory management and so on, such measures bringing to great management costs to enterprises. A growing number of studies have examined the costs of product returns and reverse logistics over the last 20 years. Existing research show costs of handling returns are much higher than that of delivering products even costs for returns are more than the cost to manufacture (Frei *et al*., 2020). According to another study by Cullen *et al.* (2013), even if a

company reported a return rate of 70%, each item sold in such a case is effectively causing the company to make a loss each time (Cullen *et al.,* 2013).

We must highlight that the various return routes make it challenging for companies to track products sold in stores or online, which is a big challenge to supply chain and inventory management and need a longer time to solve. There are many disciplinary approaches to the product returns problem, including customer behavior control, marketing, and advertising, purchasing, supply chain management, customer service, analytics, and strategic operations management, as well as circular economy, product design, material science, and waste management (Frei *et al.,* 2016).

In order to attract more loyal customers and provide competitive customer service, many online stores offer unconditional product returns and will not charge a fee for returning, which results in more and more product returns. However, even return fraud appears as a loose return policy. Return fraud is the act of buying goods with the intention of returning them, which is a form of informal or illegitimate borrowing. For example, some fashion blogger post outfit pictures on social media attracting many young girls to simulate their actions by illegitimate borrowing cloth and necessaries on the online stores; It has even been reported that some sports enthusiasts purchase large televisions to observe a sporting event over the summer holidays and then return them (Frei *et al.,* 2016). By abusing the return policy, customers can return a product they bought with the intention to return it at the time of purchase. In doing so, compared with abusive customers who just pay at little or no cost no matter what in aspects of physical, experiential, or financial, retailers have to cost much for this return action even disrupting their management system. For instance, According to estimates, retailers in the United States alone incur $5.6 billion in costs each year because of return abuse (Ketzenberg *et al.*,2020). Some retailers acknowledged such issues but struggled to avoid them due to the dilemma of rejecting improper returns while providing high levels of customer services (Jack *et al.*,2019). Some major retailers, like Victoria's Secret, manage their return process by using third-party customer profiling service (Ketzenberg *et al.*,2020). Returning merchandise requires the checkout clerk to scan the customer's ID along

with the original transaction receipt. A third-party service provider combines that information with the customer's transaction history to recommend whether to reject, warn, or accept the return (Speights and Rittman, 2015). A lack of academic studies supporting the implementation of better returns systems is present despite the recent increase in the academic literature on product returns.

On another side, increasing product returns leads to increased transportation and product waste, both of which affect the environment, including continual demand for products, placing strain on the Earth's resources, and resulting in issues like plastic waste in the food chain, and the possibility of catastrophic effects on marine and human life. Additionally, over the years, the widespread use of palm oil, wood, plastic, coconuts, and several other items over the years has led to the decimation of our rain forests, the elimination of biodiversity, the degradation of air quality, and health risks for humans (Baden and Frei, 2021). Deforestation, habitat loss, loss of biodiversity, pollution, congestion, and toxic waste are also environmental impacts of the production and transport of goods (Sala *et al*.,2019). After COVID-19 broke out in 2020, we were encouraged to consume to revive the economy, causing shops to close, high streets to empty, jobs to disappear, and GDP to plunge. Paradoxically, the more you buy, the more returns you will get, which will also cause more waste of resources and environmental pollution. There is no doubt that product returns not only have a lot of profit impact on retailers, but also pose a great challenge to environmental protection and sustainable development. This is a reason why we focus on product return so that more research can be applied to realize product return and customers' psychology and to provide effective suggestions to retailers to reduce costs and protect the environment.

In order to reduce product return, retailers have applied much measures and many research focuses on this specific. One of the most important is to identify who will return products and why customers return them. In this research, the goal is to address that confusion in terms of customer's product return behavior rather than in terms of product, retailer, and manufacturer. Specifically, this paper focus on the effects of customers' demographic characteristics such as age and purchasing power.

It is a diverse set of reasons that customers return products. However, it is very difficult for retailers to obtain data on demographic characteristics, so that it is limited to analyzing the impact of demographic characteristics on product returns. Our research is to identify the demographic characteristics that are manifest in customers who are more likely to return products, and explain which factor is most significant to product return. Therefore, we use the Logistic Regression model, Naïve Bayes and Neural Network to quantitative analyze the consumer personality of who is used to product return by using the data of second-hand electronic products in Europe and apply qualitative analysis to advise retailers on how to reduce return rate for decreasing cost of return and protecting the environment. Specifically, this paper attempt to achieve the goal of the study by considering the following research question:

1 In terms of demographic characteristics such as age and purchasing power , which factor might affect customer's return the products?

2 Which factor has greater influence?

To answer these questions, any such Machine Learning model will need to (1) identify demographic characteristics that might have an impact on product returns and(2) make classifications or predictions of customers over time with a certain level of accuracy.

In this section, we briefly introduce why we select this field to research and clarify what question we can solve and what technique will be used. In the next section, we give a detailed overview of the literature on product returns and identify areas that have not been studied. Meanwhile, we will review and summarize the research progress of product return and find out the current research gaps to help us better fill the gaps in this field. In the third section, the methodology adopted in this paper will be outlined. Including data selection, data cleaning, data preprocessing, model selection, model evaluation and model optimization. The purpose of this research was to explore the potential of Machine Learning, which can learn from data and make predictions. To create a robust and resilient return process, Machine Learning can be a successful technology that utilizes the enormous data capacity of the customers.

From the fourth to the sixth section, we summarize the analysis results and gain some valuable insights based on the data. We will extract the factors that affect the return behavior from the analysis results and put forward some basic solutions for retailers. However, in terms of actual creativity, it may be necessary to refer to specific issues. Furthermore, we explain the limitations of this research and critically think about the strengths and weaknesses of this research and the direction of future development for reference.

2. Literature Review

The literature review is divided into two parts, first of all, it summarizes the research on the influence of consumer return behavior so far, including various behaviors and factors that affect returns. Such as product quality, product price, lower than expected and so on. On the other hand, this part summarizes the existing research conclusions and future research directions from the return policy, return fraud, demographic characteristics, and product categories. For the problems that have been found, of course, corresponding solutions should be put forward. Consequently, we found out what methods retailers have taken to deal with and solve various problems arising from returns. The second part is to summarize the quantitative analysis methods that have been used in product returns so far. Suitable analysis tools and analysis results can be summarized according to previous studies. This is essential for us to use quantitative analysis methods in this study.

Product Returns refer to a process in which customers take previously purchased merchandise back to a retailer, from which they receive a refund, an exchange, or store credit as a result. Product return is critical not only in terms of retailers because of uncertainty related to price, demand, and quality of the product, but also in terms of suppliers, customers and the environment. First, the return will reduce profits tosome degree, which must influence the bargaining power of retailers in the whole industry. Second, product return will waste customers' energy and money so as to lose their customer for this brand and retailer. Third, the process of product return will generate many wastes of packages and transportation

which is a big challenge for our environment. Due to so many influences in different aspects, researchers focus on this field early and have got some achievements.

Reasons for product return can be categorized into 3 types in the product lifecycle, including manufacturing, distribution and customer returns. Manufacturing return refers to returns from surplus materials damage, scrap and so on. Distribution returns are initiated from external return sources, such as product recalls, damage return, wrong delivered the product, and stock adjustments. A product can already be returned from the manufacturer to the raw material producer. Of course, product returns might happen from retailers to their supplier (Ambilkar *et al*., 2021). This research, however, does not specifically focus on return flows from manufacturers, as they occur less frequently than returns from customers. Customers return, which this research focuses, indicate a return process from the end of the consumer, like product failure, lower product quality as expectated, unsuitable products, wrong delivered product, damaged package, and fraud return. Returns may occur at every stage of product sales.

2.1. Product Returns Behavior

2.1.1. Reasons For Product Returns

According to research (Wood, 2001), when deciding to purchase online, the decision involves two steps: either ordering or not ordering, which involves a high degree of uncertainty because the customer cannot assess the quality of the product (Bonifield *et al*., 2010). In the second phase, the customer decides whether to keep or return the items (Wood, 2001). Therefore, the customer will inspect all aspects of product information they can get. In an online purchase, the retailer describes product information on the website, like price, quality, appearance, and usage. Customers make a purchase decisions based on these information but cannot experience a real product before receiving it. A decision that whether they accept this product will be made when they receive the package (Teo and Yeong, 2003). Consequently, online retailers strive toward two goals to maximize sales: a high order

intention during the first stage and a low return intention during the second stage (Gelbrich,2017).

The simplest return behavior is when the product cannot fulfill needs. Petersen and Kumar (2009), 5% of all returns are because of defective products or incorrectly sold products. According to Pei and Paswan (2018), they suggested several reasons may lead customers to return a product if it is unsatisfactory, either internally (e.g., change of mind due to impulsiveness) or externally (e.g., the product does not meet expectations, perceived risk of keeping the product, negative attitude by social group). Therefore, in this research, Confirmation/Disconfirmation can be used to explain return behaviors that are caused by unsatisfactory purchases.

Quality and price of products are related to the possibility of return rate. Customer satisfaction is reduced when low-quality products and services are provided, which will result in product returns. Meanwhile, providing high-quality products and services is rewarded with higher selling prices (Kirmani and Rao, 2000). Another study finding (Jiang *et al*., 2005) indicates that customer satisfaction after delivery has a much higher influence on both overall customer satisfaction and repurchase intentions than satisfaction at checkout, and that price perception directly influences customer satisfaction and repurchase intentions. Meanwhile, A reduction in price (Anderson *et al*. 2009; Petersen and Kumar 2009) makes it less likely for products to be returned from a marketing-mix perspective. A decrease in customer demand can also be a result of an increase in price, especially when the consumer demand for a particular product is price-sensitive (Li *et al*., 2013). Recently study (Fan *et al.,*2022) indicates that the majority of returns even occur for other reasons, not defective products.

Even though The reasons for product return mentioned above, such as quality, failure to meet demand and express delivery problems, have been improved by improving quality and enhancing the role of online feedback, consumer product returns have been on the rise (D'Innocenzio, 2011), which suggests that dissatisfaction, product failure, or dishonest intent are not the only reasons for consumer product returns. Hence, the topic of product return is

focused persistently. At the same time, insight research (Petersen and Kummar, 2012) shows that taking into account the characteristics of products and customers, If expectations are high, there is a higher likelihood of purchasing and returning the product; if expectations are low, there is a lower likelihood of purchasing and returning the product. Similarly, some customers have buyer's remorse and change their minds after purchase and as a result, they return the products.

In terms of information related to products, Cuffie (2020) suggested that providing better descriptions of these characteristics could shrink the gap between customer expectations and the reality of the product offered. Additionally, providing customers with high-tech tools to bring them closer to the product would help decrease the number of returns. Minnema *et al*. (2016) suggested product reviews are related to product return. More specifically, a positive review of a product will not result in an exactly positive impact on a potential customer. Therefore, retailers shouldn't just encourage very satisfied customers to write reviews. Because excessive positive reviews increase the probability of purchase, the negative impact on the probability of product returns cannot be offset. By knowing how many other people have experienced the product, a buyer can be less uncertain about the product itself (Babić *et al.*,2016). Lower uncertainty then leads to lower return probability. In other words, review volume is expected to lower return probability. Researchers (Sahoo *et al.,*2018) found that unbiased online reviews improve consumer purchasing decisions, which reduces returns; biased reviews result in more returns. In the meantime, they observe that consumers are more likely to write negative reviews when they return products than if they don't return them. Another finding (Minnema *et al.,*2016) shows that there is an increased return rate for products whose displayed average rating is higher than their true rating when the displayed average rating is discontinuous.

It is especially important to consider the return policy when selling online since more than 70 percent of online consumers consider return policies when making purchase decisions (Su, 2009). A return policy that gives the consumer compensation for returns can boost consumer

demand and subsequently increase sales, resulting in an increase in returns and a higher cost of returns. Consumer responses to different return policies have been examined in prior studies. Despite lenient return policies, Wood (2001) finds that purchases increase without returns increasing. Using a direct sale model, Mukhopadhyay and Setoputro (2004) examine how pricing and return policies have an impact on the purchase and return decisions of customers. Return quantity is determined by a return policy, and product and service quality are not considered. More specifically in return policy detail, longer deadlines, according to Janakiraman and Ordónez (2012), result in consumers delaying or postponing their return decisions. It is possible for online retailers to implement a restrictive return policy as a method of reducing returns. Customers' return tendency is reduced by measures such as reshipment fees (Petersen and Kumar, 2009). However, it could deter customers from ordering in the first place since they anticipate costly reversals (Wood, 2001). Janakiraman *et al.* (2016) report that retailers avoid restrictive policies because of this side effect. Nonetheless, a lenient policy may result in an increase in returns, resulting in high costs for online retailers.

In terms of consuming experience, Bechwati and Siegal (2005) suggested consumers' involvement and choosing alternative products influence the likelihood of product returns. Unless a store provides high-quality products at competitive prices, it is more likely that products will be returned. As Lee and Yi (2017) demonstrate, gift-with-purchase promotions are associated with fewer returns of consumer products, hypothesizing that consumers are less likely to return products that are sold with free gifts. Walsh *et al.* (2016) illuminate product return rates are correlated with online retailers' reputations. A conclusion can be drawn from two experiments that show that reputation reduces return rates. The finding also shows that the strength of the relationship between reputation and product returns is influenced by shopping frequency. In the last step of the purchase process, one of the last opportunities for e-retailers to influence consumers' return behavior is through delivery package cues (Garretson and Burton, 2005). In comparison to what remains in consumers' memories (i.e., the stimuli displayed at the time of purchase), delivery packages probably carry clearer and fresher information. Finally, Zhou *et al*. (2018) explored the cognitive-emotional response

process of consumers after opening the package. According to this research, pleasure plays a crucial role in consumers' return choices.

On the other hand, Studies that focus on the characteristics of the consumers ordering the products are far fewer in number. Research on online shopping product returns (Cheung 2003; Chang *et al*., 2005; Cheung *et al*.,2005; Zhou *et al.,* 2007) has shown that demographic characteristics, such as gender, age, education, and income, influence the likelihood of the products being returned. The results indicate that these four variables are related to product return. Makkonen *et al.* (2021) focus on four demographic characteristics (i.e., gender, age, education, and income) as well as payment method preference. It is more likely to return a product when paying with a credit card. Meanwhile, among women, it is a greater probability of return frequency than among men, while the odds of return frequency decreased with age. Yan and Cao (2017) confirmed another point, an argument that the payment method influences the return of products as well. Due to the “buy-now-pay-later” mentality associated with credit cards, the researchers explain this finding with the fact that impulsive consumption behaviour is more likely to result, as well as a lower threshold of returns because there has been no exchange of money yet. For different categories, Clothing, and shoes, for example, are more likely to be returned by women and younger consumers (Deloitte, 2019), while consumer electronics are more likely to be returned by men and older consumers (Deloitte, 2019).

Table 2-1：Summary of Researches of Product Return

Name	Time	Primary Content
Petersen and Kumar	2009	About 5% of the goods are returned due to quality problems.

Pei and Paswan	2018	internal drivers (i.e., variety seeking, impulsiveness, perceived uniqueness, level of morality, and self-monitoring) and external drivers (i.e., product compatibility, returning cost, perceived risk, the complexity of procedure, and social group influence) are driven reasons to return products (Pei and Paswan, 2018.).
Kirmani and Rao	2000	low-quality products and services will result in product returns.
Anderson *et al.*	2009	A reduction in price makes it less likely for products to be returned from a marketing-mix perspective.
Bandi *et al.*	2018	When the price is expected to drop before purchase, customers tend to choose a payment method that is easy to return.
Fan *et al.*	2021	the majority of returns even occur for other reasons, not defective products.
D'Innocenzio	2011	dissatisfaction, product failure, or dishonest intent are not the only reasons for consumer product returns. However, the influence of other factors is not clearly stated.
Petersen and Kummar	2012	higher expectations should lead to higher purchase and return probabilities.
Khasawneh	2020	providing customers with high-tech tools to bring them closer to the product would help decrease the number of returns.
Minnema *et al.*	2016	product reviews are related to product return.
Babic *et al.*	2016	review volume is expected to lower the return probability.

Sahoo *et al.*	2018	unbiased online reviews improve consumer purchasing decisions, which reduces returns.
Minnema *et al.*	2016	products with higher average long-term ratings have a higher purchase and lower return probability.
Su	2009	Return compensation will lead to more product sales and more returns.
Wood	2001	A loose return policy will lead to an increase in product sales but will not result in an increase in products.
Janakiraman and Ordónez	2019	The return policy determines the quantity returned. And Longer deadlines, according to Janakiraman and Ordónez (2019), result in consumers delaying or postponing their return decisions.
Petersen and Kummar	2009	Customers' return tendency is reduced by measures such as reshipment fees.
Syrdal and Freling	2016	retailers avoid restrictive policies because of this side effect.
Bechwati and Siegal	2005	Substitutes will affect the possibility of customer return.
Lee and Yi	2017	Promotion products are less likely to be returned.
Albrecht *et al.*	2016	Product return rates are correlated with online retailers' reputations.
Garretson and Burton	2005	delivery package cues have an impact on product return. In other words, good packaging will reduce the possibility of customer return.

Makkonen *et al.*	2021	Women are more likely to return goods than men. And as the age decreases, the probability of return increases gradually.
Yan and Cao	2017	payment method influences the return of products as the probability of return is high under impulse consumption, and the payment method convenient for return is more favorable.

2.1.2. Return Fraud

Return fraud has attracted researchers' attention in recent years. Fraud return is a critical factor that results in product returns since the customer who fraud return plans to return it when experiencing the value of product. It is imperative to study consumers' product return behavior from an ethical perspective due to unethical behavior becoming an everyday matter in the workplace, marketplace, society, and even the academic scene (Craciun, 2006). As the first one to examine product returns from an ethical perspective, a study (Schmidt *et al*. 1999) use the term of “deshopping” and define it as the deliberate return of goods for reasons other than actual faults in the product. Using the term 'retail borrowing', Piron and Young (2000) study the effect of gender, income, and the economic status of the borrower on the behavior of retail lenders in order to identify unethical behavior. In Johnson and Rhee (2008), consumer traits, demographic characteristics, and social groups are studied in relation to merchandise borrowing, and the results show clear agreement with those reported by Piron and Young. The findings of Harris (2008) demonstrate a relationship between demographic factors such as age, sex, and level of education and psychographic factors such as the prior experience of fraudulent return and knowledge of returning rules and regulations. Resulting from this paper, it was found that fraud is more likely to occur among younger, female consumers with a lower levels of education, a conclusion that sexes, ages, and educational levels have an impact on fraud return. Consequently, in order to understand consumers' behavior relating to product returns, a comprehensive understanding is necessary, as well as

empirical data supporting the findings. This research (Harris, 2008) also demonstrates that there are eight psychographic factors linked to fraudulent returning tendency: past experience of fraudulent returns, level of public self-consciousness, knowledge of returning rules and regulations, consumer anomie, attitude toward complaining, social norms, consumption-related, thrill-seeking needs, and consumers' perceived impact on returning. To improve this condition, retailers are temptated to prevent return abuse by charging customers a return fee.

Table 2-2：Summary of Researches of Return Fraud

Name	Time	Primary Content
Craciun	2006	Use the term of “deshopping” and define it as the deliberate return of goods for reasons other than actual faults in the product.
Piron and Young	2000	Use the term “retail borrowing” to investigate unethical behavior and study gender effects and income effects on retail borrowing
Johnson and Rhee	2008	Results show highly consistent with Piron and Young
Harris	2008	Fraud is more likely to occur among younger, female consumers with lower levels of education. Furthermore, it has been found that fraud returning tendencies are influenced by eight psychographic factors as well.

In the study, we focus on demographic characteristics of product return. Specifically, we will explore whether the return rate will correspondingly change as the purchasing power and age change. Based on the purchasing power in each county and the average age in each town data in Germany, we can obtain some insights into the consumers’ habits and characteristics in Europe so that some measurements can be provided to the retail industry.

2.1.3. Return Policy and Actions to Reduce Return

Many firms have taken some measures to fit their product return management strategy and thus reduce the return rate. For example, the return window at Wal-Mart is 90 days, with some exceptions, and the return window at Dell is 21 days with a 15% restocking fee. As

another example, outdoor gear retailer REI has a "no-questions-asked" return policy, one of the most accommodating in the industry (Grind, 2013). Customers who abuse Best Buy's return policy are blacklisted and charged a 15% restocking fee (Boyle, 2006). Stock *et al*. (2002) points out that several companies are managing supply chains to simplify consumer returns in order to combat this problem. By outsourcing the return process to reverse logistics specialists, reducing costs by simplifying the return process, and redistributing returned merchandise, some profits can be salvaged. Even more, some companies implement more restrictive return policies like penalties. However, penalties imposed on customers who return a product can cause negative emotions such as regret, resulting in inaction (Bower and Maxham, 2012). Customers already feel negative emotions when a product doesn't meet their expectations. Further increases in this level may result from restrictive return handling. As a result of this disadvantage, a restrictive policy would seem highly unfeasible. Therefore, a way of rewarding instead of punishment is proposed (Gelbrich,2017). Keep rewards are defined as promotion strategies that offer an incentive to customers for keeping the ordered items (e.g., free shipping on their next purchase) while allowing lenient return policies. A high return rate of lenient policy is improved by adding a promotional component that may reduce the return tendency.

Except for adapting the return policy to reduce product return, some techniques to describe information about the product have been applied. Bechwati and Siegal (2005) mentioned that the information provided by retailers affects the ability of customers to adequately evaluate products before purchasing. Retailers have taken some technique to describe clearly about the product. For instance, to help customers make better decisions and to avoid return costs, retailers have invested in technologies like zoom features. Furthermore, Online Customer Reviews (OCRS) contribute to forming customer expectations before purchase (Chen and Xie, 2008), and may affect return rates. Furthermore, based on online review information, some retailers provide more information about products to customer. A study from De *et al*. (2013) examined the effectiveness of different online product inspection technologies in reducing product returns. On the basis of detailed information regarding how customers use

technology before purchase, they demonstrate that using an online zoom tool leads to fewer product returns and that using alternative images of a product leads to higher returns.

Table 2-3：Summary of Researches of Return Policy

Name	Time	Primary Content
Grind	2013	Wal-Mart is 90 days and the return window at Dell is 21 days with a 15% restocking fee. And outdoor gear retailer REI has a "no-questions-asked" return policy.
Boyle	2016	Customers who abuse Best Buy's return policy are blacklisted and charged a 15% restocking fee.
Stock, Speh, and Shear	2002	Several companies are managing supply chains to simplify consumer returns in order to combat this problem.
Bower and Maxham	2012	Penalties imposed on customers who return products may lead to negative emotions such as regret, resulting in more returns.
Gelbrich	2017	Rewarding(e.g., free shipping on their next purchase) can better reduce product returns than punishment.
Bechwati and Siegal	2005	The information provided by retailers is related to purchasing and product return.
Chen and Xie	2008	Online Customer Reviews (OCRS) can provide more information about products, thus resulting in reducing in product returns.
De *et al*.	2013	Using an online zoom tool is associated with fewer product returns.

2.2. Quantitative Analytics Model of Product Returns

Forecasting models are typically classified as qualitative forecasting models or quantitative Forecasting models (Zhou *et al*.,2016). The qualitative forecasting models are generally subjective and are mostly based on the opinions and judgments of experts. There is general use for such types of methods when there is little to no historical data available on which to base a forecast, or when there is very little data available. On the other hand, in quantitative

forecasting, the data available is used to make predictions about the future and a statistical association is presented between past and present values, based on the patterns in the data. In short, A subjective judgment is used for the former if historical data are unavailable, while a more practical approach is used for the latter. Quantitative forecasting methods include time series methods, such as moving averages and linear regressions, as well as measures of causality and econometric models, such as regression analysis or autoregressive moving averages with exogenous input and neural network (Salehzadeh *et al*.,2020).

According to Kumar and Yamaoka (2007), their research shows that dynamic regression models are a good choice when data are in a wide range variety. With the combination of DEA/linear regression and moving averages, Potdar and Rogers (2012) developed a model to forecast consumer electronics product returns using reason-codes. To forecast product returns, they are taking consumer behavior and turning it into meaningful data. In contrast, Alexandra *et al.* (2016) use Holt's and ARIMA methods to forecast the future returns of 36 products, which indicates a higher accuracy. Ma and Kim (2016) applied autoregressive statistical models to predict return quantity and time. There is also a recommendation here for the use of Gaussian distributions when counts are large (e.g., the data for reusable bottles presented in this study), as these distributions provide a good fit to the data. In this way, the total amount of returns within a certain time can be estimated, but individual return actions cannot be predicted.

Secondly, for the Machine Learning model, Using Mahalanobis feature extraction, Urbanke *et al.* (2015) predict return rates based on product features (e.g., brand, color, size), customer attributes (e.g., past return rates), and basket information, including platform, payment method, and the total number of items, an algorithm is applicable to product return business since most customer databases contain categorical data. However, Typically, the required information is available only after customers have completed their online shopping journey, so this method is not designed for customer-product level prediction. The information related to customers when they search on the online stores, such as what they like, consumption

level, and what products are in their purchase cart, is significant to analyze what personalities are inclined to return the product. Moreover, the historical purchase and return records for products in the past can be highly valuable sources of information but can be challenging to integrate in a principled way in order to predict future returns. The work by Zhu *et al.* (2018) focuses on modeling customer online shopping behaviors and predicting their return actions through the integration of the rich information that comes from the purchase and return history of products (e.g., return history, purchase-no-return behavior, and similarity between customers and products) by using HyGraph, which is a local random walk algorithm with a fixed running time based on the size of the output clusters, rather than the entire graph.

Zhou and Xie (2016) demonstrate that their study is the first model that has been developed to forecast the quantity, time, and probability of product return and remanufacturing by using the GERT stochastic network analysis technique. A data-driven model was developed by Cui *et al*. (2020) using detailed operational information on each product and information about the retailer to predict return volume by the retailer, product type and period. LASSO yields a predictive model achieving the best prediction accuracy for future return volume. Stacking and Vote algorithms from EML algorithms are used by Tüylü *et al.* (2022) to estimate product return rates, indicating that the EML algorithms can be used to predict product return rates.

In another research (Dzyabura *et al.*,2018), a Convolutional Neural Network and color histograms are employed to extract information from images using machine learning, then using this information in a gradient boosted regression tree prediction model. By incorporating visual characteristics into the model, the accuracy of predicting return rates is increased by an impressive 37% compared to models that do not include images. From Ketzenberg *et al.* (2020), several well-performing Machine Learning methods, such as Logistic Regression, Support Vector Machines, Random Forests, and Neural Networks are applied to measure return abuse, predictive models could achieve 99.6% accuracy.

According to previous research, Machine Learning methods such as Logistic Regression, Random Forest, Neural Network, Decision Tree, EML method and NLP have been applied

to data analysis in this field. In this study, other Machine Learning methods such as Bayesian, and other optimization methods are mainly used to predict.

Table 2-4: Summary of Data Analytics Methods of Product Return

Name	Time	Primary Content	Analytics Methods
Kumar and Yamaoka	2007	The Regression model is suited for a number of data.	Dynamic regression models
Potdar and Rogers	2012	The incoming returns are split into different categories using reason codes. The computation part of this model uses a combination of two approaches, namely the extreme point approach and the central tendency approach. The two approaches are employed separately for different types of reason codes, and then the combined results are analyzed.	Data envelopment analysis (DEA) as a first step combined with linear regression while the central tendency approach
Alexandra *et al*.	2016	To forecast the future returns of 36 products, which indicates a higher accuracy.	Holt's and ARIMA
Ma and Kim	2016	To predict return quantity and time.	Autoregressive statistical models
Urbanke *et al.*	2015	Predict return rates based on product features (e.g., brand, color, size), customer attributes (e.g., past return rates), and basket	The approach proposed in this study is to maximize the Mahalanobis distance of the observation from the expected value under the null hypothesis,

		information, including platform, payment method, and the total number of items.	thus minimizing the probability that the null hypothesis is true
Zhu *et al.*	2018	Incorporating the rich information in the purchase and return history of a product into a model for predicting customer online shopping behavior and predicting their return actions.	The Lograph Algorithm is based on random walks and its running time is determined by the size of the output cluster rather than the entire graph
Cui *et al.*	2020	Developed a data-driven model for predicting return volume at the retailer, product type and period levels.	Data-driven mode
Tüylü *et al*.	2022	To estimate product return rates and this model shows a good ability to predict.	Stacking and Vote algorithms from EML algorithms
Dzyabura *et al.*	2018	The accuracy of predicting return rates is increased by an impressive 37% compared to models that do not include images.	A convolutional neural network and color histograms are employed to extract information
Ketzenberg *et al.*	2020	To predict return abuse and predictive models could achieve 99.6% accuracy.	Logistic regression, support vector machines, random forests, and neural networks

2.3. Research Innovation

We have found that limited by the product return data, there are few relevant studies on the demographic characteristics of product returns. And the existing research focuses on North America and Asia. However, due to the influence of economic level, regional culture and education level, the consumption habits in different places are different. Compared with the conservative consumption habits in Asia and the free consumption concept in North America,

Europe has great uniqueness in consumption behavior. At the same time, Europe's economic level is relatively developed, and its culture is both traditional and open. Therefore, the impact of demographic characteristics on product returns may be different from that of other regions. In this study, we focus on the impact of demographic characteristics, including purchasing power and average age, on product returns in Europe. In addition, in terms of data analysis methods, statistical analysis methods, logistic regression and other basic research methods are used in the publications of product returns about demographic characteristics. Moreover, we not only use Logistic Regression in Machine Learning, a model that is suitable for binary classification data but also use Naive Bayesian classification and Neural Network classifier. These three analysis methods are more rational for the data applied in this study. The independent variable is continuous variables, and the dependent variable is binary data: 1 means the product has been returned, and 0 means the product has not been returned. According to the data analysis results, we can know the conclusion, that is, whether the return of products will change with the change in consumers' age and purchasing power. Furthermore, one of the goals of this study is to offer some advice to the retail industry in Europe according to our quantitative analysis and qualitative analysis. We hope that the conclusion we get is valuable and could be used for practical application.

3. Methodology

In this section, we follow the process of data mining, the data mining has been broken down into six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Specifically, subsection 1 is understanding a problem and the final goal. Subsection 2, data collection, introduces the data source and data type. Subsection 3 discusses data cleaning and data preprocessing: including processing missing values and outliers, data standardization, data discretization and features selection. Subsection 4 discusses model selection, including Logistic Regression, Naive Bayesian and Neural Networks, and model evaluation, including ROC and confusion matrix. In the last part, due to the dataset and model problems found in the previous part, the data and model are optimized.

We focus on the analysis of the third part of modeling and evaluation and the fourth part of model optimization, as this part is the focus of this paper. However, the performance of the model before optimization is limited, so the optimization process is significant.

3.1 Understanding A Problem and Final Goal

The return of products will be affected by many factors, such as the factors of the product itself, the factors of retailers, and even the problems of breakage during sales and transportation. At the same time, the return of products is also affected by the characteristics of consumers, such as income, gender, education, etc. In this study, we mainly explore the impact of demographic characteristics on product returns. Specifically, when the age or the purchasing power changes, how the frequency of consumer returns will change. Explore the relationship between these two features and product returns.

3.2 Data Collection

The original data comes from the sales data of a second-hand product sales website in Europe. The retailer mainly sells electronic products, including computers, mobile phones, televisions, and other products. The data content includes:

Table 3-1: Data Explanation

Name	Explanation
zip code	zip code of the consumer address
consumer ID	values corresponding to each customer are unique
product ID	values corresponding to each product are unique
invoice ID	corresponding to each product sold are unique values
payment methods	include PayPal, eBay payment, and Amazon payment
Sales channels	online and offline
Reasons for return	including product quality, price and change of mind.

Package type	DHL and Amazon Prime
the date of purchasing	from January 2020 to June 2022, which are the sales data after covid19 outbreak, so the impact of covid19 can be removed.

Although the data includes the return reason, there is no reason related to demographic characteristics in the return reason, so it cannot be used directly. Therefore, we matched the purchasing power of German counties and the average age of towns by zip code.

The original dataset needs to be matched with the purchasing power data and the average age data. We can conduct one-to-one accurate matching to counties through postal code, but some counties have high similarities in names, so we can't judge the specific location, a total of 854 (less than 1%). Similarly, there is a similar situation when matching the average age, and there are 7531 pieces of data (less than 5%) of the average age that cannot be identified.

3.3 Data preparation and preprocessing

3.3.1 Methods Introduction

3.3.1.1 Zscore Normalization

In data analysis, many independent variables that need to be analyzed are at the same level, so that different trend ranges of different units can be compared. Otherwise, it will cause difficulties in the analysis work and even affect the accuracy of later modeling.

This method normalizes the data based on the mean and standard deviation of the original data. Then the standard deviation is processed according to the following formula:

$$Z_{ij} = (X_{ij} - X_i) / S_i,$$

where: Z_{ij} is the value of the variable after standardization; X_{ij} is the actual variable value, S_i is the standard deviation (Saranya and Manikandan, 2013).

Then, reverse the sign before the indicator. The standardized variable value fluctuates around 0. A value greater than 0 indicates that it is above the average level, and a value less than 0 indicates that it is below the average level.

3.3.1.2 Weight of Evidence (WoE)

The data is discretized by WoE. Weights of evidence (WoE) measures the relative risk associated with an attribute category (Fan *et al*.,2011). The higher the weight of evidence (in favour of being good), the lower the risk for that category.

$WoE=\ln(pro\_good_{category}/pro\_bad_{category})$,

Where $pro\_good_{category}$=number of $good_{category}$/number of $goods_{total}$

$pro\_bad_{category}$= number of $bads_{category}$/number of $bads_{total}$

if $pro\_good_{category} > pro\_bad_{category}$, WoE >0

if $pro\_good_{catego} < pro\_bad_{category}$, WoE >0

3.3.1.3 Information Value

Information Value (IV) is a step to select feature, that is, among all variables, select the independent variable that is important to predict the dependent variable. IV is a measure of predictive power used to assess the appropriateness of the classification and select predictive variables (Wang *et al*., 2020).

$$IV = \Sigma((pro\_good_{category}\_pro\_bad_{category}) * WoE_{category})$$

Rule of thumb，

IV< 0.02: unpredictive; 0.02 – 0.1: weak; 0.1-0.3: medium; IV > 0.3: strong.

3.3.2 Implementation procedure

3.3.2.1 Data description

After data cleaning, There are 162663 observations and four variables, including Customer ID, Purchasing Power, Average Age, and Return Value. Variables are shown below，

Table 3-2: Features

Name	Explanation
Customer ID	Corresponding to each customer are unique values, object variables, dependent variables
Purchasing Power	The purchasing power of each county in Germany, numerical variable, dependent variables
Average Age	The average age of each town in Germany, numerical variable, dependent variables
Return Value	Whether a customer return product, 1 means return and 0 means do not return, binary variable, independent variables

Table 3-3: Data Description

	Purchasing Power	Avg Age	Return Value
count	161769	155332	162663
mean	24.881	44.283	0.098664
std	2.461	1.973	0.298211
min	19.901	27.900	0
25%	23.503	43.500	0
50%	23.694	43.600	0
75%	26.252	45.200	0
max	37.705	56.200	1

According to the data description, the purchasing power of each county in Germany is from 19.9k to 37.7k and the mean purchasing power of each county is 24.9k. The average age of each town is from 27.9 to 56.2 years old, but mostly around 43-45 years old since the 25 percentile to 75 percentile of average age lies in this area.

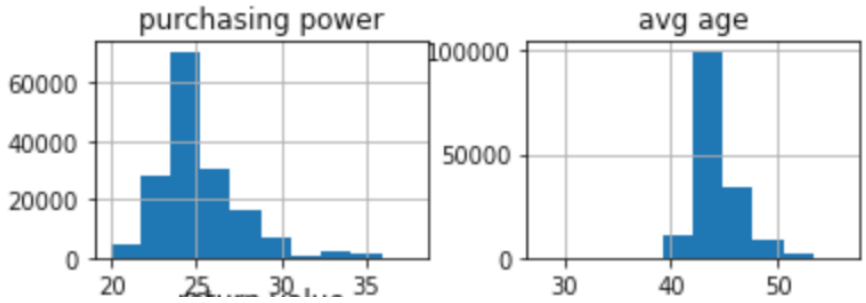

Figure 3-1: Histogram Plot of Variables

According to the histogram, purchasing power mostly lies on 24k, and the data shows the normal distribution. Meanwhile, the average age is mostly around 44, which suggests normal distribution as well.

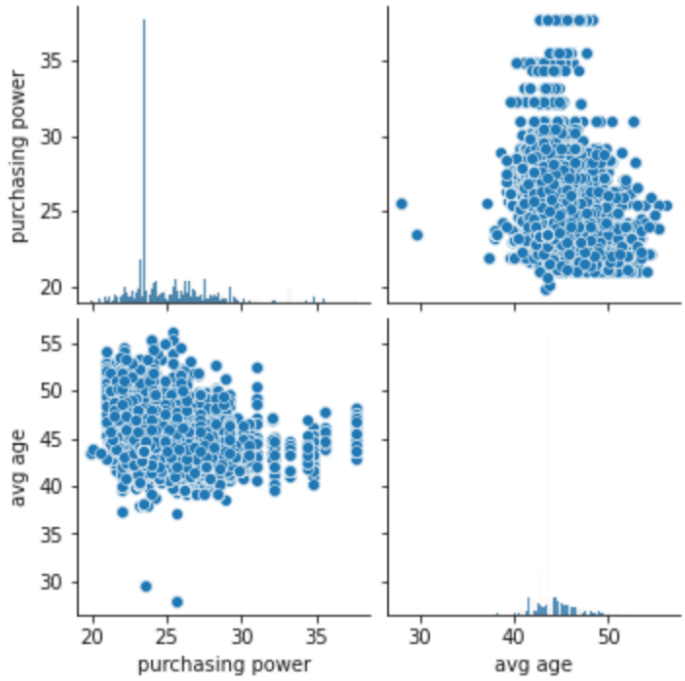

Figure 3-2: Pair Plot of Variables

The scatter chart can let us know whether there are patterns in the data; It can be seen from the figure that the scatter plot data of purchasing power and average age are unevenly distributed, which may mean that the variables are not related. Therefore, we can put them into the Machine Learning algorithm since many Machine Learning model is based on the assumption of uncorrelation between features.

The independent variables are regarded as inputs and the dependent variables are regarded as the outputs that depend on the inputs. By using Supervised Machine Learning algorithms, this paper will analyze a number of observations and try to mathematically express the dependence between inputs, purchasing power, and average age, and outputs, product return.

3.3.2.2 Data preprocessing

According to the table of data descriptions, it indicates that there are missing values in these data since the count number is less than 162663. There are 894 missing values in purchasing power and 7331missing value in average age. Meanwhile, there are only two independent variables, and the influence of each variable on the dependent variable is 50%, so the absence of any one variable will have a great influence on the dependent variable; In addition, the missing value of purchasing power in each county accounts for less than 1%, and the missing value of average age in each town is less than 5%. After deleting these missing values, the objectivity of the data and the correctness of the results will not lead to wrong analysis conclusions. The output results of the Machine Learning model can still reflect the real situation of the data. Therefore, the observation corresponding to the missing value can be removed when processing the missing value. Finally, the dataset includes 154840 observations after missing values are handled.

Table 3-4: Data Description After Dropping Missing Value

	Purchasing Power	Avg Age	Return Value
count	154840	154840	154840
mean	24.966	44.273	0.097
std	2.458	1.965	0.296
min	19.901	27.9	0
25%	23.503	43.5	0
50%	23.934	43.6	0
75%	26.326	45.1	0
max	37.705	56.2	1

According to the plot shown below, it is suggested that there are no obvious outliers as the figure shows an obvious bulge When the purchasing power is about 33, and there is a small bulge around 35, which means that the purchasing power of some counties is 33K and 35K, and there is no "long tail" after that, meaning that no obvious outliers.

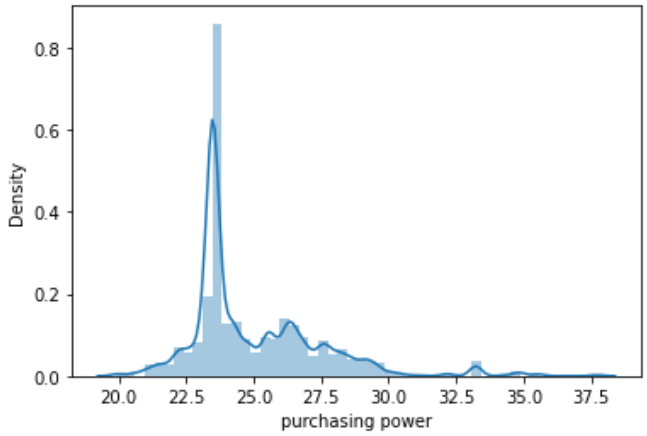

Figure 3-3: Distribution of Purchasing Power

Similarly, the figure shown below does not have a “long tail”, which indicates no obvious outlier so that the variable of average age does not need to be handled.

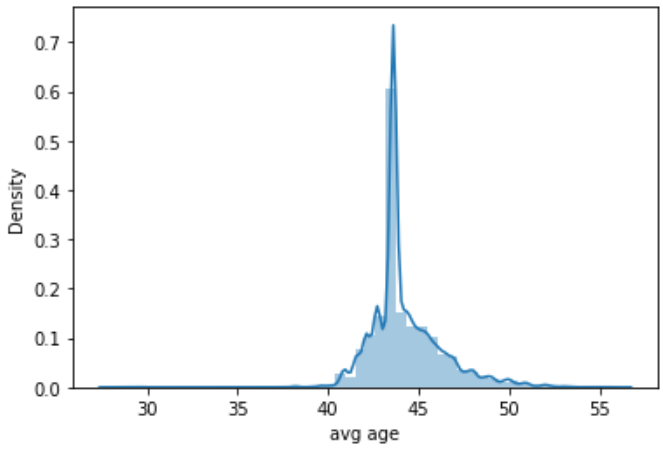

Figure 3-4: Distribution of Average Age

3.3.2.3 Data Normalization and Discretization

In this study, the average age of each town and the purchasing power of each county are not within the same order of magnitude and cannot be directly compared. After normalization, the data are mapped in the same interval, so that the variables can be compared directly.

Before WoE, the data needs to be coarse binning first. Data binning splits up the value range of continuous variables into separate intervals or bands, such as binning age data to 18-25,26-32,33-40, and so on. Meanwhile, binning data merge values of discrete variables into

a smaller number of categories. However, in this study, we just need to adjust continuous(numerical) variables since the average age of each town and purchasing power of each county are numerical variables. After manual adjustment, the binning of the purchasing power of each county and the average age of each town are adjusted to [0,1], and the trend of variables is monotonically increasing.

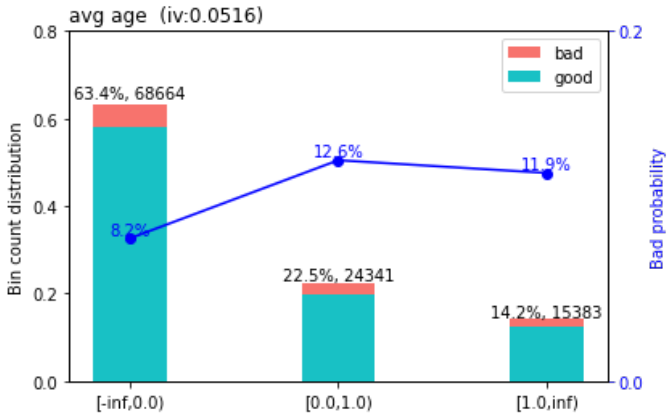

Figure 3-5: Data Binning of Average Age

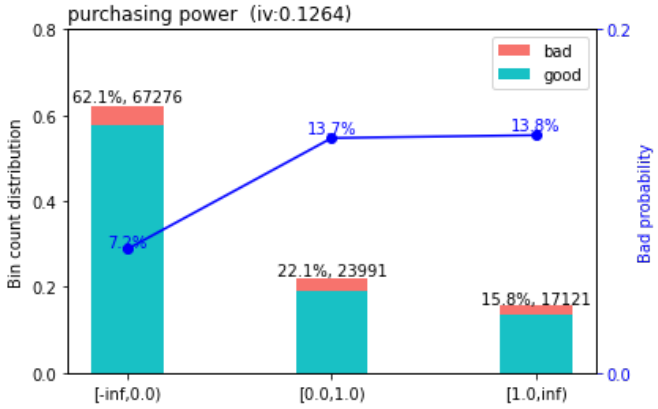

Figure 3-6: Data Binning of Purchasing Power

The WoE can indicate the prediction ability of the box for the dependent variable. If the WoE is positive and the value is larger, the probability of bad users of the box is higher; and if the negative value of the WoE is larger, the probability of good users of the box is higher. Therefore, the WoE has good business interpretability.

The general standard is that when the IV value is greater than 0.3, the variable has a strong prediction ability; When 0.1 < IV < 0.3, the prediction ability of this variable is general. When IV < 0.1, the predictive ability of this variable is weak.

The credit card dataset is split into two halves one training set and the other testing set. In this study, we chose the ratio of 7:3 for the training dataset over the testing dataset.

3.4 Modeling and Testing

In the dataset of product returns, there are two values for the classification of transactions which means that it is a binary classification problem where transactions are classified either as return (1) or non-return (0). After preprocessing the data, the classifiers are trained using the training data to evaluate the methods. In this study of classification techniques, We study several typically competitive, well-performing machine learning methods that include Logistic Regression, Naïve Bayes and Neural Networks

3.4.1 methods introduction

3.4.1.1 Logistic regression

Logistic regression is to fit the data by linear regression, and then use the logic function to predict the classification results. It is especially suitable for classification, when the dependent variable is dichotomized (0 / 1, true / false, yes / no), logistic regression will show a good fitting result. In this study, the dependent variable return value includes two cases: 1 and 0. 1 means that the product has been returned, and 0 means that the product has not been returned. Hilbe (2009) introduces Logistic regression models in detail.

Let y = 1 indicate the thing happens, y = 0 indicate that the thing does not happen, the probability of happing is P (y = 1) = P_1, the probability of not happening is P(y = 0) = P_0 = 1- P_1

Then odds= P_1/1- P_1= P_1/ P_0, Odds refers to the probability of happening compared with nonhappening

Logit=log(odds)=ln (P_1/ P_0)= $\beta_0+\beta_1X_1$

when the independent variable change by 1 unit, the odds change $e^{\beta 1}$ times.

3.4.1.2 Naïve Bayes Model

The Bayesian classification algorithm is the general name of a large class of classification algorithms and takes the probability that the sample may belong to a certain class as the classification basis. Naïve Bayes' Theorem is a mathematical formula used for calculating conditional probabilities. A conditional probability is a probability that an event will occur after another event has taken place. See more detail of the Naïve Bayes model in the book of Johnson(2022).

The formula is

$P(Y|X) = P(X|Y) * P(Y) / P(X)$

Where

P(Y|X) is called posterior probability, meaning how often A happens given that B happens

P(X|Y) represents how often X happens given that Y happens

P(Y) represents how likely Y is on its own

P(X) represents how likely X is on its own

If P (Y=1| X)>P (Y=0| X), class as happening; otherwise, class as not happening.

3.4.1.3 Neural Network

There is no explicit formula or algebraic expression for neural network regression. The process of training the Neural Network is an optimization problem, that is, to find out which parameters make the model work best. Chollet (2021) indicates that the observed input data is x, the output is y, and the predicted value of the model is f (x). When initializing the Neural Network model, the parameters of f (x) are random values. For the observation value x, the random parameters are randomly generated.

In this study, there are two classes, the Softmax function is,

$f(x)=1/\exp(\beta_1 X_1+\beta_0)$, which is common as an output function of probability estimations,

where β_1 is the coefficient of x1, β_0 is constant

This process is called forward propagation, and it is a process that the input (observation value) is calculated layer by layer (including linear calculation and nonlinear activation) to obtain the predicted value of the output value. To evaluate the accuracy of the predicted value, a loss function is required. A common loss function is the sum of squares of residuals loss = $loss = \sum i (\hat{y}_i - y_i)^2$.

The optimal parameters can be obtained by reducing the loss of the loss function. This process is called back propagation. In essence, the process of back propagation is to calculate the gradient value of each parameter for the loss function, and then adjust it according to the gradient value. Theoretically, convergence can be achieved quickly.

3.4.1.4 Confusion Matrix And Receiver Operating Characteristic Curve (ROC)

In prediction analysis, the confusion matrix is a two row and two column table composed of false positions, false negatives, true positions and true negatives. True positives are the cases that are predicted as positive and in reality, they are positive as well. True negative is the cases that are considered as negative in advance. False positives are cases that are expected to be positive but turn out to be. False negative is one that appears to be negative but is actually positive (Tze-Wey, 2003).

the value of ROC means: "the ROC is equal to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example（Hand and Till ,2001)." The value of ROC is on [0.5,1]. In the case of ROC > 0.5, the closer the value of ROC is to 1, the better the diagnostic effect.

3.4.2 Implementation procedure

3.4.2.1 Logistic regression

Let y = 1 indicate that the product is returned, y = 0 indicate that the product is not returned, the probability of product return is p (y = 1) = P_1, the probability of not returning is P(y = 0) = P_0 = 1- P_1, X_1 = purchasing power, X_2 = average age

Then odds= P_1/1- P_1= P_1/ P_0 Odds refers to the probability of product return compared with nonreturn Logit=log(odds)=ln(P_1/ P_0)= $\beta_0+\beta_1X_1+\beta_2X_2$

Since both X1 and X2 are continuous variables, when the independent variable changes by 1 unit, the odds change $e^{\beta 1}$ 和 $e^{\beta 2}$times.

3.4.2.2 Naïve Bayes model

According to the fundamental Naïve Bayes assumption, we assume that

No two features are dependent on each other, which means there is no correlation between each variable 'purchasing power', 'average age'. Meanwhile, according to table 3-2, it is evidently indicate that is true.

Each feature is equally influential, suggesting that each variable represents the same weight.

In this study,

$P(Y|X_1X_2) = P(X_1X_2|Y)*P(Y)/P(X_1X_2)=P(X_1|Y)P(X_2|Y) *P(Y)/P(X_1X_2)$

Where

$P(Y|X_1X_2)$ is a probability that product has been returned, given average purchasing power and average age of a place.

$P(X_1|Y)$ is a probability of a customer who has specific purchasing power return products

$P(X_2|Y)$ is a probability of a customer who is a specific age power return product

P(Y) is probability that product has been returned

$P(X_1X_2)$ is a probability of a customer who has a specific purchasing power and in a specific average age

If $P (Y=1| X_1X_2)>P (Y=0| X_1X_2)$, class as product return; otherwise, class as not product return

In this study, because the variable characteristics are binomial distribution, we chose Bernoulli Bayesian model to fit the data.

3.4.2.3 Neural Network

When initializing the Neural Network model, the parameters of f (x) are random values.

In this study, there are two classes, the softmax function is $f(x)=1/\exp(\beta_1X_1+\beta_2X_2+\beta_0)$, which is common as output function of probability estimations, where

β_1 is the coefficient of X_1, X_1 is independent variable of purchasing power in each county

β_2 is the coefficient of X_2, X_2 is independent variable of average age in each town

β_0 is constant

loss function is the sum of squares of residuals $loss = \sum i\,(\hat{y}_i - y_i)^2$.where $\hat{y}_i$ is the true value and y_i is predicted value (Janocha and Czarnecki, 2017).

3.5 Model optimization

3.5.1 Methods Introduction

3.5.1.1 Polynomial expansion

Polynomial expansion is the process of extending features to a polynomial space, which is represented by the n-degree combination of the original dimensions (James *et al*., 2015). Polynomial expansion improves the input variable to idempotent transformation, which helps to better reveal the important relationship between the input variable and the target variable. Sometimes these features can improve modeling performance, although at the cost of adding thousands or even millions of additional input variables. Polynomial features create new input features based on existing features. For example, if the dataset has an input feature x, the polynomial feature will be to add a new feature (column), where the value is calculated by squaring the value in X, for example, X^2. Ghaith and Li (2020) pointed out that the prediction performance of the model was good after the variables were processed by Polynomial Expansion.

3.5.1.2 Decision Tree Binning

The decision-making process of a decision tree is essentially a series of if / then statements, which make decisions through machine learning. decision trees have an important property: they are mutually exclusive and complete. This means that for each sample, there is and only one path from the root node to a leaf node. At the same time, because a series of conditional judgments are constantly made on the features, the decision tree can also be understood as the solution of the conditional probability of P (Yi| Xi). To construct the decision tree, the algorithm iterates all possible questions, finds the one with the largest amount of information

for the target variable, divides the data set into two parts, and repeats this process until the end. Therefore, In the first step, the decision tree needs to be divided into nodes. In the second Step: the condition for a node to stop dividing into leaf nodes is that all samples in the node belong to the same category. That is, nodes are "pure". Therefore, when selecting features for partitioning, the purer the nodes, the better the features. We use entropy to measure the "purity" of a group of samples. We recommend Rabuzin *et al*. (2014) International Journal of Data Warehousing and Mining to learn more about the discretization method.

For the discrete random variable X, its distribution is:

$$P(X=X_i) = p_i, i=1,2,3...k$$

Information entropy is

$$\mathrm{H(X)} = \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{pilnpi} \ , \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{pi} = 1$$

For binary classification,

$$\mathrm{p_1+p_2=1}$$

$$H=-(\ \mathrm{p_1} ln\ \mathrm{p_1} + \mathrm{p_2} ln \mathrm{p_2})$$

Then，

$$\frac{\partial \mathrm{H}}{\partial \mathrm{pi}} = -\mathrm{lnpi} - 1 - \frac{\mathrm{dpj}}{\mathrm{dpi}} \mathrm{lnpj} - \frac{\mathrm{dpj}}{\mathrm{dpi}} = \ln\left(\frac{1}{\mathrm{pi}} - 1\right)$$

When $\frac{1}{2} \leq p_i \leq 1$，$H'(pi) \leq 0$，$H$ decreasing gradually.

When $0 \leq p_i \leq \frac{1}{2}$, $H'(pi) \leq 0$，H increasing gradually.

The results of d When $p1=1/2$. Information entropy is maximum. For binary classification, the smaller the entropy is, the better the result is.

3.5.2 implementation procedure of optimization

3.5.2.1 feature transformation

According to the results of model analysis, among the three models, the ROC score is between 0.6 and 0.65, and that there is no big difference between different models. Therefore, it can be explained that there are some problems in the process of machine learning due to

the problems in the data itself. In reviewing the original data and data analysis, we found the following problems: 1) in the original data, only less than 10% of the return data, that is, the return value of 1 is less than 10%, and the return value of 0 is more than 90%. This phenomenon led to data imbalance. 2) There are too few features, which is not conducive to machine learning. In the data used in this study, there are only two characteristics, purchasing power and average age, and many models with excellent performance, such as neural networks, are suitable for data with a large amount of observations and more data features. In the process of optimizing the model, we first reduce the amount of data and increase the number of features to reduce overfitting. We only selected the data of the first half of 2022, with a total of 33543 observations.

Thus, we suppose that the independent variable, X_1=purchasing power, X_2=average age, and the additional variable by Polynomial expansion $X_3 = X_1 * X_2$, $X_4 = X_1 * X_1$, $X_5 = X_2 * X_2$, $X_6 = X_1 / X_2$

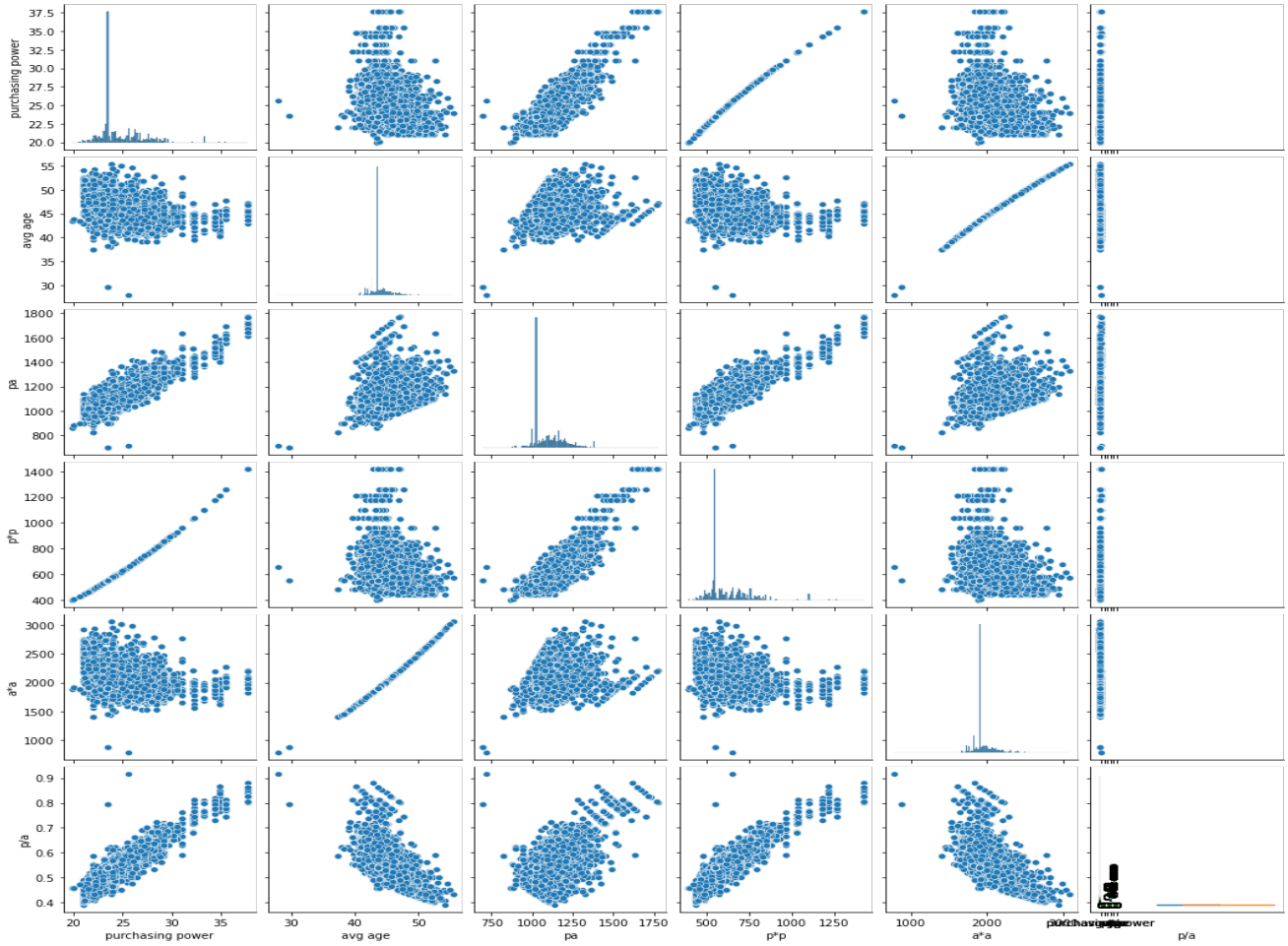

Figure 3-7：Scatter of Variables

There is no obvious relationship between purchasing power and average age but purchasing power and average age have a certain relationship with the newly added variables respectively, because the newly added variables are set based on purchasing power and average age.

3.5.2.2 Data Discretizing

When discretizing the data, we use decision tree binning. This method has two advantages: 1) pruning prevents overfitting. Pruning includes pre pruning and post pruning. The former controls the depth of the tree or the number of nodes by setting thresholds for continuous variables, and operates before the nodes are divided, thereby preventing overfitting. The latter is to examine non-leaf nodes from the bottom up. If replacing this internal node with a leaf node can improve the generalization ability of the decision tree, replace it. 2) In order to prevent the training decision tree from being too biased towards some categories due to too many samples in some categories of the training set. The algorithm calculates the weight by itself, and the category with a small sample size will have a higher weight.

Decision tree binning results are show below,

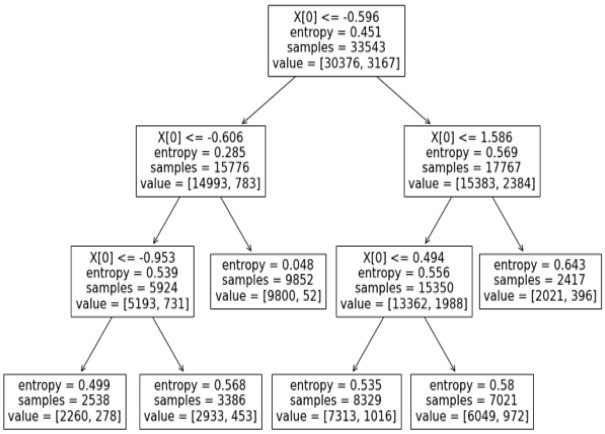

Figure 3-8: Entropy of X_1

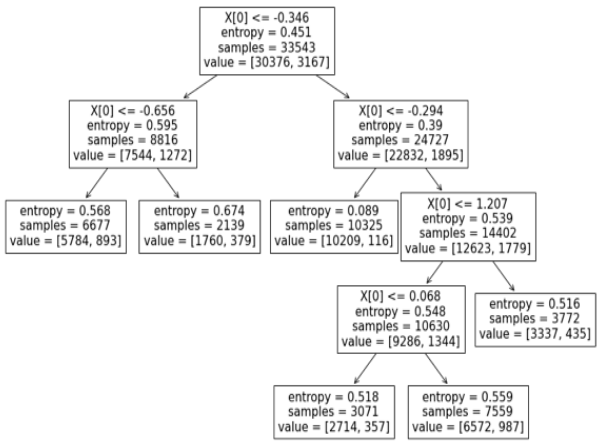

Figure 3-9:Entropy of X_2

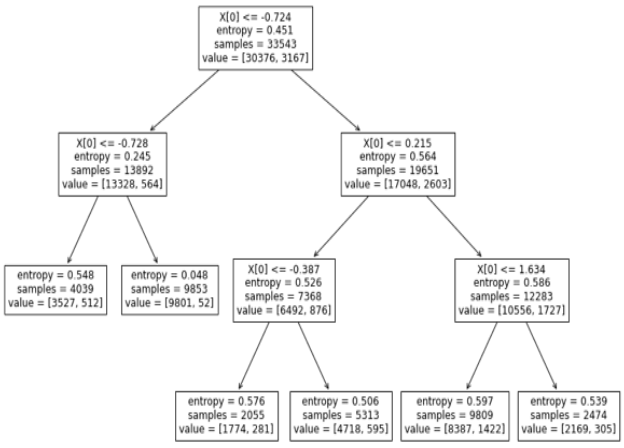

Figure 3-10: Entropy of X_3

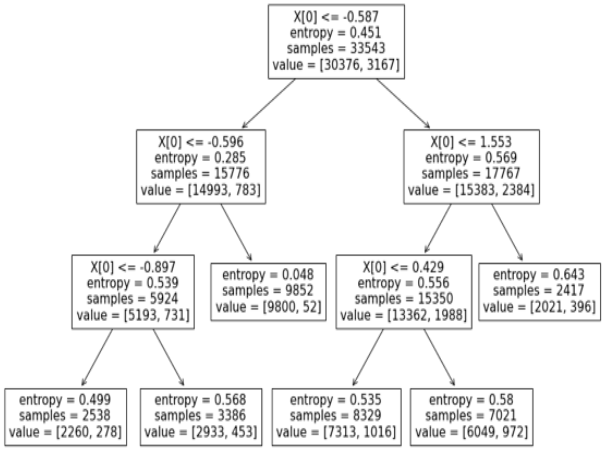

Figure 3-11: Entropy of X_4

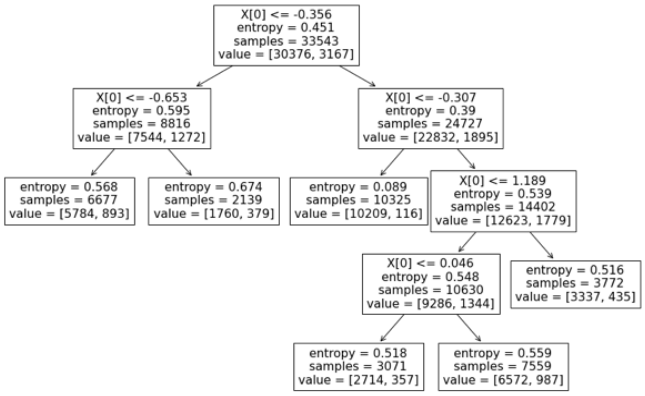

Figure 3-12: Entropy of X_5

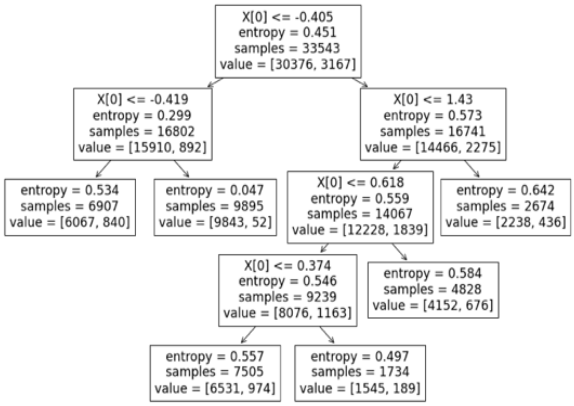

Figure 3-13: Entropy of X_6

In the optimization, we added dummy variables. Dummy variables are 0 or 1 to indicate whether a category appears. Because after the data is discretized, using only discrete numbers to represent categories may affect the accuracy of subsequent regression or classification. After deleting the missing values after discretization of the decision tree and adding dummy variables, there are 20185 observed values

3.5.2.3 Logistic Regression

Let y = 1 indicate that the product is returned, y = 0 indicate that the product is not returned, the probability of product return is p (y = 1) = P_1, the probability of not returning is P(y = 0) = P_0 = 1-P_1, X_1 = purchasing power, X_2 = average age， $X_3 = X_1 * X_2$, $X_4 = X_1 * X_1$, $X_5 = X_2 * X_2$, and $X_6 = X_1 / X_2$

Then odds= P_1/1- P_1= P_1/ P_0 Odds refers to the probability of product return compared with nonreturn

Since the Decision Tree Binning requires dummy variables. Therefore, in the process of discretization of variables, each variable is divided into different "boxes", and the number of variables is adjusted to 30 variables. Therefore, the six variables we assume cannot get the coefficients, but only the coefficients of the variables after discretization. Therefore, it is impossible to directly judge the specific impact of the two characteristics, purchasing power and average age, on product returns, but this model can be used to judge whether returns will occur at the customer level.

3.5.2.4 Naïve Bayes Model

The independent variable X_1, X_2, X_3, X_4 and X_5 dependent variable Y are consistent with those mentioned above，then

$P(Y|X_1X_2X_3X_4X_5) = P(X_1X_2X_3X_4X_5|Y) * P(Y)/P(X_1X_2X_3X_4X_5)$

$=P(X_1|Y)\ P(X_2|Y)P(X_3|Y)P(X_4|Y)\ P(X_5|Y)\ *P(Y)/P(X_1X_2\ X_3X_4\ X_5)$

Where

$P(Y|X_1X_2\ X_3X_4X_5\ X_6)$ is a probability that product has been returned, given average purchasing power and average age of a place and satisfy the other four additional variables.

$P(X_1|Y)$ is a probability of a customer who has specific purchasing power return products

$P(X_2|Y)$ is a probability of a customer who is a specific age power return product

$P(X_3|Y)$, $P(X_4|Y)$, $P(X_5|Y)$, $P(X_6|Y)$ is a probability of a customer who satisfies the other four additional variables respectively return products.

P(Y) is probability that product has been returned

$P(X_1X_2\ X_3X_4\ X_5\ X_6)$ is a probability of a customer who has a specific purchasing power and in a specific average age and satisfies the other four additional variables.

If $P(Y=1|\ X_1X_2X_3X_4\ X_5\ X_6)>P(Y=0|\ X_1X_2X_3X_4\ X_5\ X_6)$, class as product return; otherwise class as not product return.

3.5.2.5 Neural network

As this is a two-class problem, the last layer is adjusted to sigmoid as the activation function, which is more suitable than softmax function. Sigmoid function, that is $f(x)=1/(1+e^{-x})$. Neural networks are complex: the functions of each layer are different, and the results are obtained

after many iterations. Therefore, we cannot determine the relationship between independent variables and dependent variables through specific coefficients. Instead, a model can be established to judge whether customers will return goods from the perspective of customers.

4. Results

4.1 Results of Data Preprocessing

Based on the results, the table indicates that the minimum boxes of the average age of each town and the minimum boxes of the purchasing power of each county are negative, and the minimum boxes of the average age of each town are greater than the minimum boxes of the purchasing power of each county, indicating that there are more good users of the minimum boxes of the purchasing power of each county. Similarly, the maximum boxes of the average age of each town and the maximum boxes of the purchasing power of each county are both negative and positive, and the maximum boxes of the average age of each town are smaller than the maximum boxes of the purchasing power of each county, indicating that there are more bad users in the maximum boxes of the purchasing power of each county. Similarly, at 75percentile, there are more bad users in purchasing power of each county.

Table 4-1: Test Dataset Description After WoE

Index	Return Value	Avg Age_WoE	Purchasing Power_WoE
count	46452	46452	46452
mean	0.097	-0.020	-0.050
std	0.296	0.223	0.349
min	0	-0.189	-0.323
25%	0	-0.189	-0.323
50%	0	-0.189	-0.323
75%	0	0.230	0.390
max	1	0.297	0.404

Table 4-2: Train Dataset Description After Woe

Index	Return Value	Avg Age_WoE	Purchasing Power_WoE
count	108388	108388	108388
mean	0.097	-0.020	-0.050
std	0.296	0.223	0.349
min	0	-0.189	-0.323
25%	0	-0.189	-0.323
50%	0	-0.189	-0.323
75%	0	0.230	0.390
max	1	0.297	0.404

According to the result, IV of purchasing power is around 0.126, and IV of average age is 0.052, which suggests that purchasing power shows the medium predictable capability and average age is weak predictable. In other words, the influence of average age on product return is not obvious.

Table 4-3: information value of variables

variable	info_value
purchasing power_WoE	0.126382263
avg age_ WoE	0.051629219

However, according to business insights, it is generally believed that age is related to returns. And as there are few data features, removing the features may have an impact on the model effect, so we do not delete variables with IV values lower than 0.1.

4.2 Modeling Results

4.2.1 Logistic Regression

Logit=log(odds)=ln (P_1/ P_0)= 0.50989886+1.777191X_1+1.429618X_2

the coefficient of purchasing power is 1.777, and the coefficient of average age is 1.430. When X1 variable (purchasing power) is changed by one unit and other variables remain unchanged, the odds increase $e^{1.777191}$ times. Similarly, when the X2 variable (average age) is changed by one unit and other variables remain unchanged, the odds increase $e^{1.429618}$

time. Furthermore, the average age has a greater impact on the prediction of whether the product is returned.

Table 4-4: Parameter of Logistic Regression

Column	Coefficient
purchasing power_ WoE	1.777191
avg age_ WoE	1.429618
intercept	0.50989886

In order to understand the fitting effect of the model, we need to evaluate the logistic regression. Confusion matrix shows that TP is 0.64, FP is 0.36, FN is 0.46 and TN is 0.54. And classification accuracy is 0.59, classification error is 0.41, sensitivity is 0.58 and specificity is 0.60, meaning that this model has a not very strong performance and does not fit well with dataset.

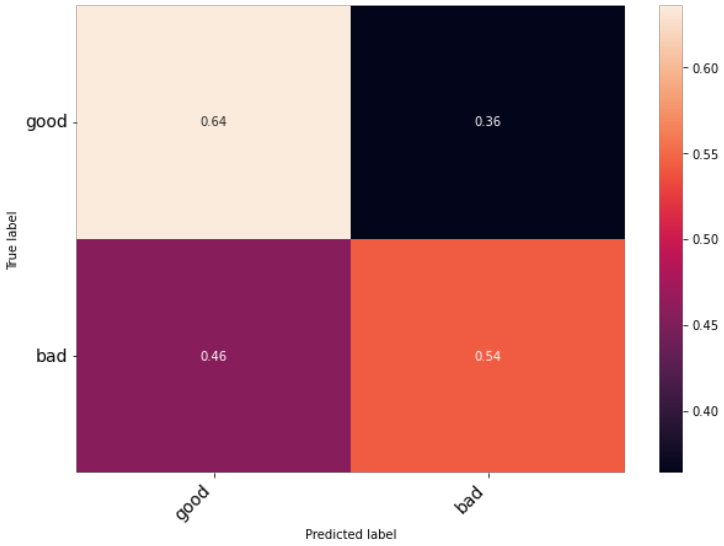

Figure 4-1: Confusion Matrix of Logistic Regression

It is generally considered that ROC exceeding 0.75 is acceptable. However, the ROC is 0.624, the result shows that this model does not fit very well with the data set. So, we need to optimize the model in the next subsection.

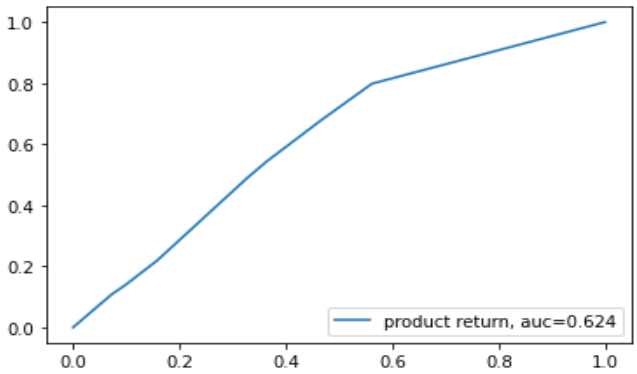

Figure 4-2：ROC of Logistic Regression

4.2.2 Results of naïve Bayes

Similarly, we evaluate Naïve Bayes models. ROC describes the relationship between TPR and FPR. The larger the area enclosed by ROC curve and coordinate graph boundary, the better the model; The ROC is 0.622, a result showing that the method has a low prediction ability, and this model does not fit well with the dataset.

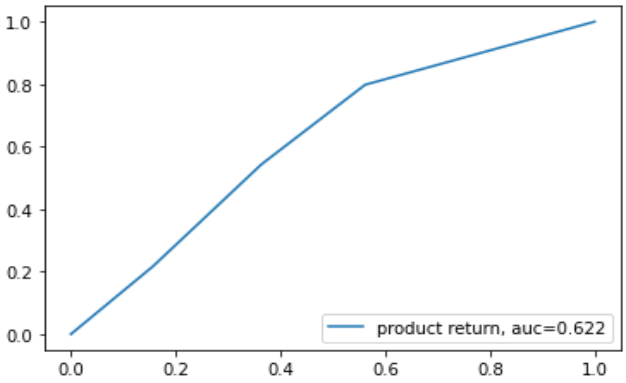

Figure 4-3：ROC of Naïve Bayes Classifier

4.2.3 Results of Neural Network Classifier

The ROC is 0.628, similarly with ROC of naïve bayes, a result shows that the method has not much good performance, and this model does not fit well with the dataset. Meanwhile, this model does not show better than logistic regression and naïve bayes.

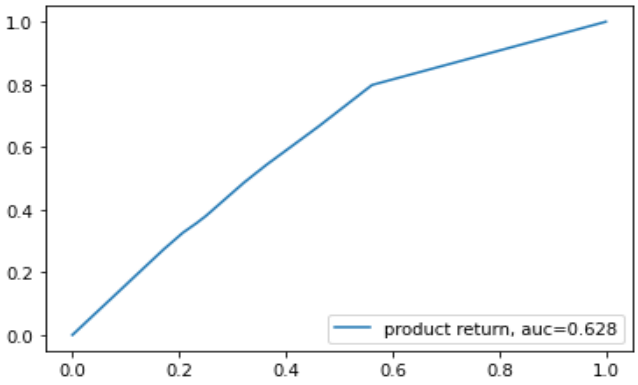

Figure 4-4：ROC of Neural Network Classifier

Although there are empirical rules and heuristics to determine the network structure, there are no known optimal decision rules (Brownlee, 2018). Therefore, we will improve the data in the preprocessing of the original data, rather than optimizing in the neural network model itself.

4.3 Model Optimization

4.3.1 Optimized Logistic Regression

After the optimization of the model, the ROC is 0.783, indicating that the fitting effect of this model has been greatly improved compared with that before the optimization, but better results will be obtained if better data sets are used.

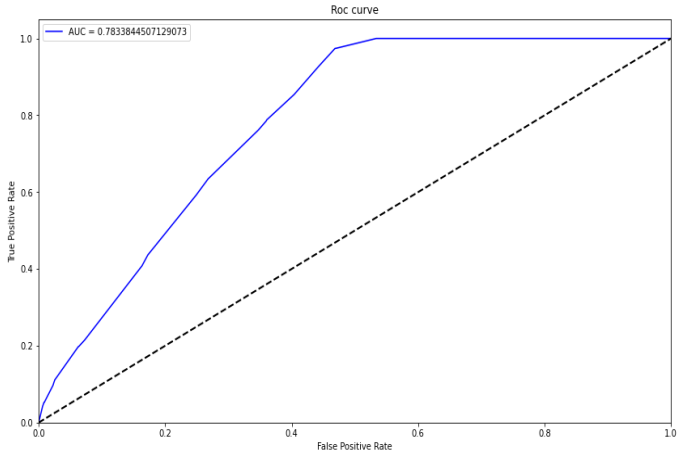

Figure 4-5：ROC of Optimized Logistic Regression

The accuracy is 0.9, indicating that the model has better prediction ability than before optimization. However, the generalization ability still has great limitations, that is to say, if the data set is changed to test the Machine Learning model, the results need to be discussed.

4.3.2 Optimized Naïve Bayes Model

Classification accuracy is 0.55, classification error is 0.45, sensitivity is 0.99 and specificity is 0.13, meaning that this model has a not good performance and there is around 45% possibility of the wrong prediction.

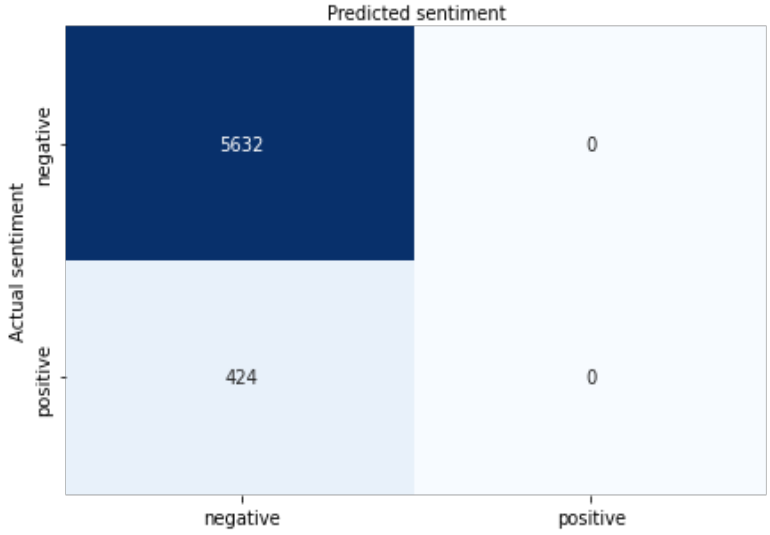

Figure 4-6：Confusion Matrix Of Naïve Bayes Classifier

Furthermore, the ROC is 0.743, the result shows that the dataset fit is better than the naïve Bayes model before optimization. However, the model still has the above-mentioned problems, including the imbalance of data caused by too few data features, and the generalization ability needs to be improved. Similarly，this model can be used to judge whether returns will occur at the customer level.

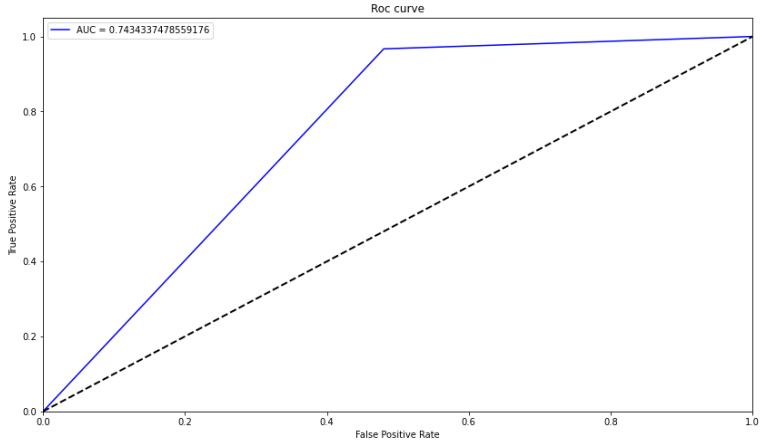

Figure 4-7：ROC of Naïve Bayes Classifier

4.3.3 Optimized Neural Network

classification accuracy is 0.94, classification error is 0.06, sensitivity is 0.94 and specificity is 0.71, meaning that this model has a relatively strong performance but there is only around 6% possibility of the wrong prediction.

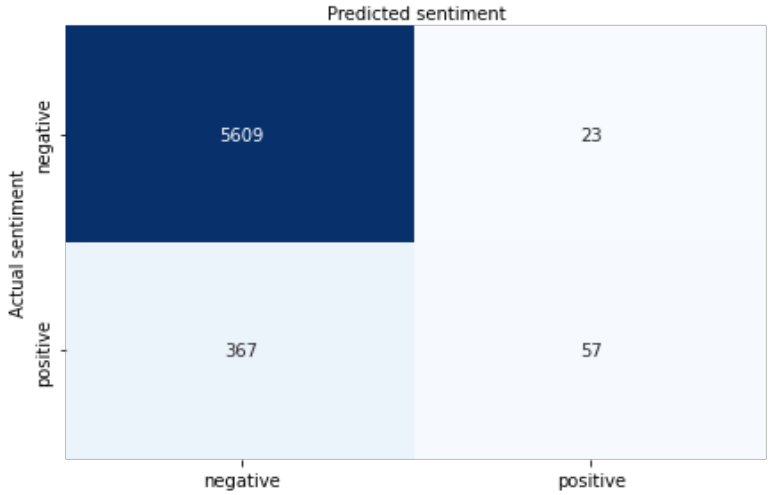

Figure 4-8：Confusion Matrix of Neural Network

The ROC is 0.801, indicating that compared with logistic regression and naïve bayes, the Neural Network model has relatively good predictability, but the defects still need to be solved in future research.

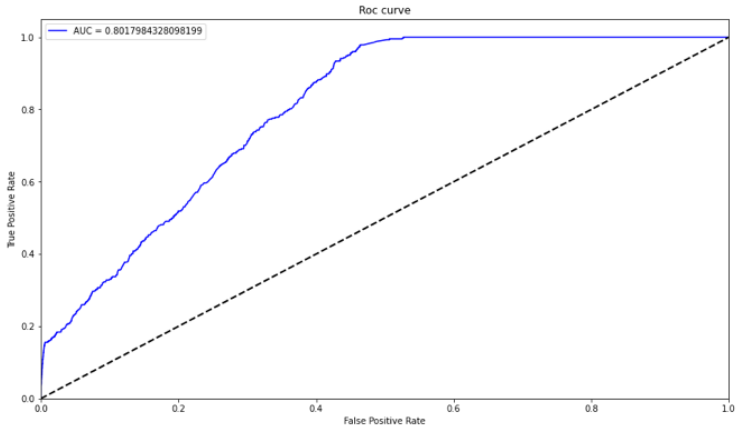

Figure 4-9：ROC Of Neural Network

4.3.4 Comparison of Predictive Methods

According to the evaluation results of each machine learning model, the Neural Network model has high accuracy and low errors, and the model has reached a relatively ideal state, that is, the prediction of unreturned products is unreturned, and the prediction of returned products is returned. The naive Bayesian model has low accuracy and high errors. The model

is not ideal, that is, it predicts that there are more returned goods and more returned goods. The accuracy of the logistic regression model is high, but the specificity cannot be predicted, which indicates that there are problems in prediction, and the confusion matrices FP and TN have defects.

Table 4-5: Performance of Three Model

	Logistic Regression	Naïve Bayes	Neural Network
Classification Accuracy	0.93	0.55	0.94
Classification Error	0.07	0.45	0.06
Sensitivity	0.93	1.00	0.94
Specificity		0.13	0.71
ROC	0.783	0.743	0.802

According to the value of ROC, it can be directly seen that the prediction effect of neural network model is better, and the accuracy of Naïve Bayes prediction is the lowest. Neural Network showed the optimal performance for all the data promotions as compared to Naïve Bayes and Logistic Regression.

The ROC results of the three models are not much different, which indicates that there are certain defects in the data itself, and the impact of these shortcomings cannot be avoided through different models. Although the prediction ability of the optimized model has been greatly improved, and the accuracy rate has reached a higher level. However, each index has its own bias, so even the evaluation results cannot fully explain the actual performance of the model.

5. Discussion

5.1 Discussion of Three Machine Learning Models

Among the three machine learning models selected, each machine learning algorithm has its unique advantages and is suitable for the data type used in this study. That's why we chose these models. For logistic regression, it performs well for simple datasets. In the datasets used in this study, there are only two independent variables and one binary dependent variable. However, this dataset is suited for logistic regression and also will not reduce the

performance of model. And Logistic regression is applicable when the dependent variable is a binary variable and the independent variables are categorical or continuous (Ershadi and Omidzadeh, 2018). 2) A Logistic Regression model is less likely to be over-fitted but it can be overfitted in high dimensional datasets.

In terms of Naïve Bayes, advantages include 1) Naïve Bayes is based on the independence assumption where training is very easy and fast. It requires considering each attribute in each class separately. It is a straightforward test involving the use of tables and calculating conditional probabilities according to a normal distribution. Even violating the independence assumption, the performance of Naive Bayes is comparable to the performance of most advanced classifiers (Gladence *et al*., 2015). 2) A naive Na|ve Bayes model has a higher asymptotic error than logistic regression, but naïve Bayes converges faster and approaches the higher error of logistic regression, which means that for an infinite training data set, logistic regression should be superior to the original Naïve Bayes because it has a lower error. However, due to the limited amount of data, Naïve Bayes may outperform regression since it requires less data to achieve optimal performance. (Witteveen *et al.,* 2018). 3) In order to obtain good results from Naive Bayes classifiers, it is necessary to collect a large number of records. In this dataset，we have more than 160k observations. (Aidaroos *et al.,* 2010).

In terms of neural network,1) In contrast to linear relation algorithms, neural network algorithms have the advantage of high precision, high precision, and high reliability when there is uncertainty regarding variable relationships and distribution forms between data or when complex systems cannot express the relationship between input and output data with general relation. 2) A small amount of knowledge about the problem is sufficient to achieve positive results, which is noteworthy due to the fact that this model does not require a great deal of specific information about variables (Bennett *et al.,* 2013). 3) self-adjusting ability to a given set of data (Sharghi *et al.*, 2018).even if the dataset has been handling not enough to fit an efficient model, neural network model will adjust predictable capacity automatically to get a higher accuracy.

5.2 The Limitations of This Study

This study has the following obvious limitations. First, the data analysis basis of the paper is based on a German retail company. However, in fact, the consumption habits of each country or region will be greatly different under the influence of various lifestyle. For example, Asian countries are used to a more conservative consumption habit. Therefore, their purchasing habits are more cautious so as to fewer product problems, and less frequency of products returns. On the contrary, in Europe, consumption habits are completely different, as consumption habits are more radical.

Second, the analysis is based on the data of second-hand electronic products, which suggests the kind of product is single. At the same time, the reasons for returning electronic products and consumables are certainly not same, and the impact of demographic characteristics on product returns is different from other products.

Third, the frequency of product returns may be the fact that if a person tends to return infrequently, the reasons for these rare returns may be related to some serious problems in the ordered product, such as failure or damage during delivery. On the contrary, if a person tends to return relatively frequently, the reason for the return is less likely to be related to the actual problems of the product, but more likely to be related to the mismatch between the product and personal needs, desires, or expectations. Therefore, the returned products can be classified, and the factors that really affect the product quality can be excluded for further analysis. In this study, since the product types cannot be matched with specific transaction, the influence of this reason cannot be excluded. In other words, if a second-hand product is of poor quality, the probability of returning is very high. Otherwise, the probability of returning is small.

Fourth, in our data, the average purchasing power and average age of the region are matched by the zip code of individuals who purchase and return goods. The purchasing power and age of each region are the average value of that region, which represents the relative status of the region, that is, the overall purchasing power of the region is high, but it does not rule out that individuals have low purchasing power. As for age, the data we use can only show that the region as a whole is younger or older, but it cannot show that individuals must be like

this. Therefore, the data itself is biased. Meanwhile, although we have considered the problems of the data itself as much as possible in the optimization, such as the overfitting problem caused by too many observations and too few features, which leads to poor generalization effect. However, from the optimized ROC, the fitting results and prediction accuracy of the data are only 0.75-0.8, indicating that the fitting and merging of the model have not reached the optimal state.

Fifth, the selected model has its own advantages and limitations In terms of logistic regression, it is quite sensitive to noise and overfitting. Especially, in this dataset there are Few independent variables, but many observations.

In terms of the Naïve Bayes classifier, 1) The Naive Bayes model is capable of generating classification bias, since the influence of these two attributes may be overvalued and the influence of other attributes may be undervalued (Aidaroos *et al*., 2010). 2) vanishing values can also be explained by combining several small probabilities together (e.g. 0.053). In this study, the average age and average purchasing power are matched by the zip code. However, in some regions, the product is purchased only twice, but the data set of the whole day is more than 160K, so the probability of occurrence is very low, thus vanishing value will happen.

In terms of the Neural Network, 1) it is impossible to establish a suitable model for the purposes of business decisions due to the lack of standard or fixed rules for governing the design and development of appropriate models (Abrahart *et al*., 2012). The inability to incorporate knowledge acquired from existing physical laws into ANNs also serves as a drawback to their use. 2) Overparametrization and overfitting problems are common problems among artificial neural networks (Adeyemo *et al*., 2018; Sayagavi and Charhate, 2017), especially in the absence of appropriate input selection and early stopping techniques.

To sum up, this study has certain limitations, which may lead to the analysis results not fully consistent with the actual situation. For example, in this study, we found that purchasing power and average age do not have a particularly large impact on product returns. Relatively speaking, age has a greater impact on product returns. However, if we use the data of other product types, such as clothes, cosmetics, and food., the data of other regions may deviate

from our results. In addition, we generally think that Logistic Regression and Naive Bayes are suitable for binary data, and Neural Networks often have better prediction results because they have no fixed formula mechanism, but in practical problems, it needs to be analyzed according to specific conditions. In future related research, more suitable models can be selected to predict based on more comprehensive data and other appropriate analysis methods, such as ABM decision making model.

6 Conclusion

6.1 Main conclusion

In this study, we mainly analyze the influence of purchasing power and average age on product return. Specifically, 1) do the two factors have an impact on product return? 2) Which factor has a greater impact on product returns? Purchasing power in essence reflects the disposable income of consumers; The influence of average age indicates whether there is a direct relationship between age and product return. For the first point, the study found that the two factors have an impact on product return at some level. In other words, changes in purchasing power will influence product returns. In summary, as the purchasing power and age increase will increase customers' willingness to return goods. This conclusion is inconsistent with the research conclusion of makkonen *et al*. (2021). Specifically, as the age increases, there is a tendency to reduce returns. makkonen *et al*. (2021) also explored the relationship between income and return but did not get a clear conclusion. For the second point, the influence of purchasing power on product return is greater than that of age. Among the coefficients of the logistic regression model, the coefficient of X_1 (purchasing power) is1.78, and the coefficient of X_2 (average age) is 1.43, indicating that purchasing power has a greater impact on the dependent variable Y.

The influence of age on product returns can be explained by the great differences in consumption habits among consumers of different ages. For older consumers, they have higher requirements for products, and they are not good at exploring products through online information, which may lead to more returns. On the contrary, for young people, the frequency of online shopping is high, and the value of consumption of young people is relatively low.

They have a greater tolerance for accepting secondary consumption and will choose less returns to avoid trouble. The impact of purchasing power on returns may be related to income level. For customer with strong purchasing power, the consumption level is usually high. Compared with customer with low purchasing power, if they also buy the same type of products, people with strong purchasing power are willing to spend more money on the same type of products, which will naturally make it easier to buy appropriate products, thus reducing the possibility of returns. On the contrary, consumers with low purchasing power may choose products with lower cost performance because they want to buy cheap products, which will make it easier to return the goods.

6.2 Suggestions For Reducing Product Returns

The significance of this study is that results can help retailers identify the demographic characteristics that are more prone to frequent returns and apply the analysis results to solve the problem of product returns, and ultimately reduce the return cost. Consequently, we give some suggestions based on the outcome found. For example, if there is an obvious correlation between product return and age, different return policies are used for specific customers. If the customer is older than 50 years old, a higher service fee is charged for return. For customers aged 30-50, a part of the service fee is charged for return. For customers aged 20-30, it can be returned for free. Customers can also be "credit rating" on the consumption platform. Customers who frequently return goods have low credit scores, and such customers will be charged a certain fee when returning goods. For different age groups, you can also send message prompts at the time of customer shopping checkout, such as "your return records are too many, which may affect your shopping experience on the shopping platform and implement a stricter return strategy. Please return with caution. Some studies have shown that appropriate prompts can play a deterrent role. Of course, a very important judgment of the retailers is what age group the customers belong to and what consumption level they have. These can also be calculated according to consumption records and combined with machine learning algorithms.

6.3 Suggestions For Future Research

We have explained the limitations of this study in detail in the previous part. In particular, in our data, the average purchasing power and average age of the region are matched by the zip code of individuals who purchase and return goods. There is a deviation between the average value and the real value of the actual individual, so there is a deviation in the data itself. If actual data, for example through questionnaire surveys, can be obtained in the future, the bias can be avoided, and analysis results will be more accurate. In addition, the data itself has too few features, too many observations, and a small proportion of data with a return value of 1 (indicating return). Although some reasonable methods can be used to adjust and achieve significant improvements, if we want to improve the generalization ability, apply it to more data, and get consistent conclusions and judgments, we still need to get more efficient dataset. Future research should focus on the diversity of data features that need to be improved and the imbalance of dependent variables so that more data features can be used to improve the generalization effect. Based on the characteristics of some models, we selected three models that we thought were suitable, but the results were not good enough. finally, if there is no problem with the data itself, researchers can try other more effective models to get a reasonable result and widely to deployment.

References

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin, A.Y., Solomatine, D.P., Toth, E. and Wilby, R.L. (2012) 'two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting'. *Progress in Physical Geography,* 36(4), pp.480-513. doi:10.1177/0309133312444943

Adeyemo, J., Oyebode, O. and Stretch, D. (2018) *River flow forecasting using an improved artificial neural network*. Available at : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266474 (Accessed: 10 July 2022).

Al-Aidaroos, K.M., Bakar, A.A. and Othman, Z. (2010) 'Naive Bayes variants in classification learning'. *In 2010 international conference on information retrieval & knowledge management (CAMP)*,pp. 276-281. IEEE. doi: 10.1109/INFRKM.2010.5466902.

Ambilkar, P., Dohale, V., Gunasekaran, A. and Bilolikar, V. (2022) 'Product returns management: a comprehensive review and future research agenda'. *International Journal of Production Research,* 60(12), pp.3920-3944. doi: 10.1080/00207543.2021.1933645

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. doi:10.1287/mksc.1080.0430

Babić Rosario, A., Sotgiu, F., De Valck, K. and Bijmolt, T.H. (2016) 'The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors'. *Journal of Marketing Research*, 53(3), pp.297-318. doi:10.1509/jmr.14.0380

Baden, D. and Frei, R. (2021) 'Product Returns: An Opportunity to Shift towards an Access-Based Economy?'. *Sustainability*, 14(1), p.410. doi:10.3390/su14010410

Bandi, C., Moreno, A., Ngwe, D. and Xu, Z. (2018) *Opportunistic returns and dynamic pricing: Empirical evidence from online retailing in emerging markets*. *Harvard business school marketing unit working paper,* (19-030), pp.19-030. Available at:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266474 (Accessed: 17 August 2022).

Bechwati, N.N. and Siegal, W.S. (2005) 'The impact of the prechoice process on product returns'. *Journal of Marketing Research*, 42(3), pp.358-367.doi:10.1509/jmkr.2005.42.3.358

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C. and Pierce, S.A.(2013) 'Characterising performance of environmental models'. *Environmental Modelling & Software,* 40, pp.1-20. doi:10.1016/j.envsoft.2012.09.011

Bonifield, C., Cole, C. and Schultz, R.L. (2010) 'Product returns on the internet: a case of mixed signals?'. *Journal of Business Research,* 63(9-10), pp.1058-1065. doi:10.1016/j.jbusres.2008.12.009

Bower, A.B. and Maxham III, 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. doi:10.1509/jm.10.0419

Boyle, M. (2006) *Best Buy's giant gamble*. Available at: https://xinkaishi.typepad.com/a_new_start/files/best_buy_fortune.pdf (Accessed: 19 August 2022).

Chang, M. K., Cheung, W., & Lai, V. S. (2005) 'Literature derived reference models for the adoption of online shopping'. *Information & Management,* 42(4), 543–559.doi:10.1016/j.im.2004.02.006.

Chen, Y. and Xie, J. (2008) 'Online consumer review: Word-of-mouth as a new element of marketing communication mix', *Management Science,* vol. 54, no. 3, pp. 477-491. doi:10.1287/mnsc.1070.0810

Cheung, C. M. K., Chan, G. W. W., & Limayem, M. (2005) 'A critical review of online consumer behavior: Empirical research'. *Journal of Electronic Commerce in Organizations,* 3(4), 1–19. doi:10.4018/jeco.2005100101

Chollet, F. (2021) *Deep learning with Python*. Simon and Schuster.

Chan, G., Cheung, C., Kwong, T., Limayem, M. and Zhu, L. (2003) *Online consumer behavior: a review and agenda for future research*. Available at:https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1112&context=bled2003 (Accessed: 9 September 2022).

Craciun, G.M. (2006) *Mood effects on ordinary unethical behavior*. University of South Carolina.

Cuffie, H.G., Najar, R.I. and Khasawneh, M.T. (2020) *Topic Modeling for Customer Returns Retail Data*. Available at:https://www.proquest.com/openview/403992bc950e55a4f9efd6b69fce7c74/1?pq-origsite=gscholar&cbl=51908 (Accessed: 9 September 2022).

Cui, H., Rajagopalan, S. and Ward, A.R. (2020) 'Predicting product return volume using machine learning methods'. *European Journal of Operational Research*, 281(3), pp.612-627. doi:10.1016/j.ejor.2019.05.046

Cullen, J., Tsamenyi, M., Bernon, M. and Gorst, J.(2013) 'Reverse logistics in the UK retail sector: A case study of the role of management accounting in driving organisational change'. *Management Accounting Research*, 24(3), pp.212-227. doi:10.1016/j.mar.2013.01.002

D'Innocenzio, A. and Beck, R. (2011) *Wal-Mart, humbled king of retail, plots rebound*. Available at: https://www.cbsnews.com/news/wal-mart-humbled-king-of-retail-plots-rebound-07-02-2011/ (Accessed: 11 September 2022).

De, P., Hu, Y. and Rahman, M.S. (2013) 'Product-oriented web technologies and product returns: An exploratory study'. *Information Systems Research*, 24(4), pp.998-1010.doi:10.1287/isre.2013.0487

Dzyabura, D., El Kihal, S. and Ibragimov, M.(2018) *Leveraging the power of images in predicting product return rates*. Available at: https://pages.stern.nyu.edu/~ddzyabur/index_files/DzyaburaElKihalIbragimov.pdf (Accessed: 11 September 2022).

Ershadi, M.J. and Omidzadeh, D. (2018) *Customer validation using hybrid logistic regression and credit scoring model: a case study*. Available at:https://en.irandoc.ac.ir/sites/fa/files/attach/article/q-asvol19no167december-2018p59-62.pdf (Accessed: 11 July 2022).

Fan, D., Cui, X.M., Yuan, D.B., Wang, J., Yang, J. and Wang, S. (2011) 'Weight of evidence method and its applications and development'. *Procedia Environmental Sciences*, 11, pp.1412-1418. doi:10.1016/j.proenv.2011.12.212

Fan, H., Khouja, M. and Zhou, J. (2022) 'Design of win-win return policies for online retailers'. *European Journal of Operational Research*, 301(2), pp.675-693. doi: 10.1016/j.ejor.2021.11.030

Fawcett, T. (2006) 'An introduction to ROC analysis'. *Pattern Recognition Letters*, 27(8), pp.861-874.doi: https://doi.org/10.1016/j.patrec.2005.10.010

Fontana, R., Luciano, E. and Semeraro, P. (2021) 'Model risk in credit risk', *Mathematical Finance*, 31(1), pp. 176–202. doi: 10.1111/mafi.12285.

Frei, R., Bines, A., Lothian, I. and Jack, L. (2016) 'Understanding reverse supply chains'. *International Journal of Supply Chain and Operations Resilience,* 2(3), pp.246-266 Available at:https://core.ac.uk/download/pdf/77049013.pdf (Accessed: 20 June 2022)

Frei, R., Jack, L. and Krzyzaniak, S.A.(2020) 'Sustainable reverse supply chains and circular economy in multichannel retail returns'. *Business Strategy and the Environment,* 29(5), pp.1925-1940. Available at:https://onlinelibrary.wiley.com/doi/pdf/10.1002/bse.2479 (Accessed: 20 June 2022)

Gareth, J., Daniela, W., Trevor, H. and Robert, T. (2013) *An introduction to statistical learning: with applications in R. Spinger*. Available at:https://dspace.agu.edu.vn/handle/agu_library/13322 (Accessed: 25June 2022)

Garretson, J.A. and Burton, S.(2005) 'The role of spokescharacters as advertisement and package cues in integrated marketing communications'. *Journal of Marketing*, 69(4), pp.118-132. doi:10.1509/jmkg.2005.69.4.118

Gelbrich, K., Gäthke, J. and Hübner, A. (2017) 'Rewarding customers who keep a product: How reinforcement affects customers' product return decision in online retailing'. *Psychology & marketing*, 34(9), pp.853-867. doi:10.1002/mar.21027

Gelbrich, K., Gäthke, J. and Hübner, A. (2017) 'Rewarding customers who keep a product: How reinforcement affects customers' product return decision in online retailing'. *Psychology & marketing*, 34(9), pp.853-867. doi:10.1002/mar.21027

Ghaith, M. and Li, Z. (2020) 'Propagation of parameter uncertainty in SWAT: A probabilistic forecasting method based on polynomial chaos expansion and machine learning'. *Journal of Hydrology,* 586, p.124854. doi: 10.1016/j.jhydrol.2020.124854

Gladence, L.M., Karthi, M. and Anu, V.M.(2015) 'A statistical comparison of logistic regression and different Bayes classification methods for machine learning'. *ARPN Journal of Engineering and Applied Sciences*, 10(14), pp.5947-5953. Available at: https://www.researchgate.net/profile/Mary-Gladence-L/publication/282921131_A_statistical_comparison_of_logistic_regression_and_different_bayes_classification_methods_for_machine_learning/links/570228d408aea6b7746a8689/A-statistical-comparison-of-logistic-regression-and-different-bayes-classification-methods-for-machine-learning.pdf (Accessed: 31June 2022).

Grind, K. (2013) 'Retailer REI ends era of many happy returns'. *Wall Street Journal,* 16. Available at:https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Grind%2C+2013&btnG= (Accessed: 31June 2022).

Harris, L.C.(2008) 'Fraudulent return proclivity: an empirical analysis'. *Journal of Retailing*, 84(4), pp.461-476. doi:10.1016/j.jretai.2008.09.003

Hilbe, J.M. (2009) *Logistic regression models*. Chapman and hall/CRC.

Jack, L., Frei, R. and Krzyzaniak, S.A.C. (2019) *Buy online, return to store: the challenges and opportunities of product returns in a multichannel environment*. Available at:https://eprints.soton.ac.uk/432483/ (Accessed: 29 June 2022).

Janakiraman, N. and Ordóñez, L. (2012) 'Effect of effort and deadlines on consumer product returns'. *Journal of Consumer Psychology*, 22(2), pp.260-271. doi:10.1016/j.jcps.2011.05.002

Jiang, P. and Rosenbloom, B.(2005) 'Customer intention to return online: price perception, attribute-level performance, and satisfaction unfolding over time'. *European journal of marketing,* 39(1/2), pp.150-174. doi:https://doi.org/10.1108/03090560510572061

Johnson, A.A., Ott, M.Q. and Dogucu, M.(2022) *Bayes Rules!: An Introduction to Applied Bayesian Modeling.* CRC Press.

Janocha, K. and Czarnecki, W.M.(2017) *On loss functions for deep neural networks in classification*. Available at: https://arxiv.org/abs/1702.05659 (Accessed: 29 July 2022).

Johnson, K.K. and Rhee, J.(2008) 'AN INVESTIGATION OF CONSUMER TRAITS AND THEIR RELATIONSHIP TO MERCHANDISE BORROWING WITH UNDERGRADUATES'. *Journal of Family & Consumer Sciences Education,* 26(1). Available at:http://www.natefacs.org/Pages/v26no1/v26n1k_johnson.pdf (Accessed: 16 July 2022).

Ketzenberg, M.E., Abbey, J.D., Heim, G.R. and Kumar, S.(2020) 'Assessing customer return behaviors through data analytics'. *Journal of Operations Management,* 66(6), pp.622-645. doi:10.1002/joom.1086

Kirmani, A. and Rao, A.R.(2000) 'No pain, no gain: A critical review of the literature on signaling unobservable product quality'. *Journal of marketing*, 64(2), pp.66-79. doi:10.1509/jmkg.64.2.66.18000

Kumar, S. and Yamaoka, T.(2007) 'System dynamics study of the Japanese automotive industry closed loop supply chain'. *Journal of Manufacturing Technology Management.* Available at: https://www.emerald.com/insight/content/doi/10.1108/17410380710722854/full/html (Accessed: 16 July 2022).

Lee, S. and Yi, Y.(2017) 'Seize the Deal, or Return It Losing Your Free Gift": The Effect of a Gift-With-Purchase Promotion on Product Return Intention'. *Psychology & Marketing*, 34(3), pp.249-263. doi: https://doi.org/10.1002/mar.20986

Li, J., He, J. and Zhu, Y.(2018, July) 'E-tail product return prediction via hypergraph-based local graph cut'. *In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 519-527. doi:10.1145/3219819.3219829

Li, Y., Xu, L. and Li, D.(2013) 'Examining relationships between the return policy, product quality, and pricing strategy in online direct selling'. *International Journal of Production Economics,* 144(2), pp.451-460. doi:https://doi.org/10.1016/j.ijpe.2013.03.013

Ma, J. and Kim, H.M.(2016) 'Predictive model selection for forecasting product returns'. *Journal of Mechanical Design*, 138(5), p.054501.doi: 10.1109/ITMC.2018.8691285

Makkonen, M., Frank, L. and Kemppainen, T.(2021) *The effects of consumer demographics and payment method preference on product return frequency and reasons in online shopping*. Available at: https://jyx.jyu.fi/handle/123456789/77000 (Accessed: 16 July 2022).

Meng, J. and Li, H.(2017) 'An efficient stochastic approach for flow in porous media via sparse polynomial chaos expansion constructed by feature selection'. *Advances in Water Resources,* 105, pp.13-28. doi: https://doi.org/10.1016/j.advwatres.2017.04.019

Minnema, A., Bijmolt, T.H., Gensler, S. and Wiesel, T.(2016). 'To keep or not to keep: Effects of online customer reviews on product returns'. *Journal of retailing,* 92(3), pp.253-267. doi:10.1016/j.jretai.2016.03.001

Mukhopadhyay, S.K. and Setoputro, R.(2005) 'Optimal return policy and modular design for build-to-order products'. *Journal of Operations Management,* 23(5), pp.496-506. doi:https://doi.org/10.1108/09600030410515691

Pei, Z. and Paswan, A.(2018) 'CONSUMERS'LEGITIMATE AND OPPORTUNISTIC PRODUCT RETURN BEHAVIORS IN ONLINE SHOPPING'. *Journal of Electronic Commerce Research*, 19(4), pp.301-319. Available at:http://www.jecr.org/sites/default/files/201819041.pdf (Accessed: 16 July 2022).

Petersen, J.A. and Kumar, V.(2010) 'Can product returns make you money?'. *MIT Sloan Management Review*, 51(3), p.85. Available at:https://apprissretail.com/wp-content/uploads/sites/4/2017/02/Can-Returns-Make-You-Money_White-Paper.pdf (Accessed: 20 July 2022).

Piron, F. and Young, M.(2000) 'Retail borrowing: insights and implications on returning used merchandise'. *International Journal of Retail & Distribution Management.* doi:10.1108/09590550010306755

Potdar, A. and Rogers, J.(2012) 'Reason-code based model to forecast product returns'. *Foresight*. doi:10.1108/14636681211222393

Rabuzin, K., Varazdin, C., Karthika, S., Tamil Nadu, I., Bose, S., Kannan, A. and Keyvanpour, M.R.(2014) *International Journal of Data Warehousing and Mining*. https://www.igi-global.com/pdf.aspx?tid%3D106857%26ptid%3D91422%26ctid%3D15%26t%3Dtable+of+contents (Accessed: 20 July 2022).

Ram, A. (2016) 'UK retailers count the cost of returns'. *Financial Times.* Available at: https://www.ft.com/content/52d26de8-c0e6-11e5-846f-79b0e3d20eaf. (Accessed 20July 2022)

Sahoo, N., Dellarocas, C. and Srinivasan, S. (2018) 'The impact of online product reviews on product returns'. *Information Systems Research*, 29(3), pp.723-738. doi:10.1287/isre.2017.0736

Sala, S., Benini, L., Beylot, A., Castellani, V., Cerutti, A., Corrado, S., Crenna, E., Diaconu, E., Sanyé-Mengual, E., Secchi, M. and Sinkko, T.(2019) 'Consumption and Consumer Footprint: methodology and results'. I*ndicators and Assessment of the Environmental Impact of European Consumption*. Luxembourg. Available at:https://core.ac.uk/download/pdf/268886158.pdf (Accessed 20July 2022)

Salehzadeh, R., Tabaeeian, R.A. and Esteki, F.(2020) 'Exploring the consequences of judgmental and quantitative forecasting on firms' competitive performance in supply chains'. *Benchmarking: An International Journal*, 27(5), pp.1717-1737. doi:https://doi.org/10.1108/BIJ-08-2019-0382

Saranya, C. and Manikandan, G. (2013) 'A study on normalization techniques for privacy preserving data mining'. *International Journal of Engineering and Technology (IJET)*, 5(3), pp.2701-2704. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.411.1996&rep=rep1&type=pdf (Accessed 20July 2022)

Sayagavi, V. and Charhate, S. (2017) 'Applications of soft tools to solve hydrological problems for an integrated Indian catchment'. *International Journal of Water Resources and Environmental Engineering,* 9(7), pp.150-161. Available at:https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&as_vis=1&q=Sayagavi+Charhate%2C+2017&btnG= (Accessed 28 July 2022)

Schmidt, R.A., Sturrock, F., Ward, P. and Lea-Greenwood, G.(1999) 'Deshopping–the art of illicit consumption'. *International Journal of Retail & Distribution Management.* doi: https://doi.org/10.1108/09590559910288569

Sharghi, E., Nourani, V., Soleimani, S. and Sadikoglu, F.(2018) 'Application of different clustering approaches to hydroclimatological catchment regionalization in mountainous

regions, a case study in Utah State'. *Journal of Mountain Science,* 15(3), pp.461-484. doi:10.1016/j.seta.2014.09.001

Speights, D. and Rittman, T.(2015) *Fighting return fraud during the holiday season. White Paper, The Retail Equation.* Available at: https://risnews.com/fighting-return-fraud-holiday-season (accessed 25 November 2015).

Stock, J., Speh, T. and Shear, H.(2002) 'Many happy (product) returns'. *Harvard business review,*80(7), pp.16-17. Available at:https://www.elibrary.ru/item.asp?id=4535555 (accessed 25 July 2022).

Su, X.(2009) 'Consumer returns policies and supply chain performance'. *Manufacturing & Service Operations Management*, 11(4), pp.595-612. doi:10.1287/msom.1080.0240

Teo, T.S. and Yeong, Y.D. (2003) 'Assessing the consumer decision process in the digital marketplace'. *Omega*, 31(5), pp.349-363. doi:10.1016/S0305-0483(03)00055-0

Tüylü, A.N.A. and Eroglu, E. (2022) 'The prediction of product return rates with ensemble machine learning algorithms'. *Journal of Engineering Research.* doi: https://doi.org/10.36909/jer.13725

Urbanke, P., Kranz, J. and Kolbe, L.(2015) *Predicting product returns in e-commerce: the contribution of mahalanobis feature extraction.* Available at:https://www.researchgate.net/profile/Johann-Kranz/publication/283270887_Predicting_Product_Returns_in_E-Commerce_The_Contribution_of_Mahalanobis_Feature_Extraction/links/5720d84c08aead26e721322b/Predicting-Product-Returns-in-E-Commerce-The-Contribution-of-Mahalanobis-Feature-Extraction.pdf

Walsh, G., Albrecht, A.K., Kunz, W. and Hofacker, C.F.(2016) 'Relationship between online retailers' reputation and product returns'. *British Journal of Management*, 27(1), pp.3-20. doi :https://doi.org/10.1111/1467-8551.12120_

Wang, W., Lesner, C., Ran, A., Rukonic, M., Xue, J. and Shiu, E.(2020, April) 'Using small business banking data for explainable credit risk scoring'. *In Proceedings of the AAAI Conference on Artificial Intelligence,*Vol. 34, No. 08, pp. 13396-13401. doi:: https://doi.org/10.1609/aaai.v34i08.7055

Wang, Y., Anderson, J., Joo, S.J. and Huscroft, J.R.(2019) 'The leniency of return policy and consumers' repurchase intention in online retailing'. *Industrial Management & Data Systems.* doi:https://doi.org/10.1108/IMDS-01-2019-0016

Witteveen, A., Nane, G.F., Vliegen, I.M., Siesling, S. and IJzerman, M.J.(2018) 'Comparison of logistic regression and Bayesian networks for risk prediction of breast cancer recurrence'. *Medical decision making,* 38(7), pp.822-833. doi:https://doi.org/10.1177/0272989X18790963

Wood, S.L.(2001) 'Remote purchase environments: The influence of return policy leniency on two-stage decision processes'. *Journal of Marketing Research*, 38(2), pp.157-169. doi:10.1509/jmkr.38.2.157.18847

Yan, R. and Cao, Z.(2017) 'Product returns, asymmetric information, and firm performance'. *International Journal of Production Economics*, 185, pp.211-222. doi :https://doi.org/10.1016/j.ijpe.2017.01.001

Yoo, S.H.(2014) 'Product quality and return policy in a supply chain under risk aversion of a supplier'. *International Journal of Production Economics,* 154, pp.146-155. doi:https://doi.org/10.1016/j.ijpe.2014.04.012

Zhou, L., Dai, L., & Zhang, D. (2007) 'Online shopping acceptance model – A critical survey of consumer factors in online shopping'. *Journal of Electronic Commerce Research*, 8(1), 41–62. Available at:https://asset-pdf.scinapse.io/prod/2150312165/2150312165.pdf (accessed 25 July 2022).

Zhou, L., Xie, J., Gu, X., Lin, Y., Ieromonachou, P. and Zhang, X.(2016) 'Forecasting return of used products for remanufacturing using Graphical Evaluation and Review Technique (GERT)'. *International Journal of Production Economics*, 181, pp.315-324. doi:https://doi.org/10.1016/j.ijpe.2016.04.016

Zhou, W., Hinz, O. and Benlian, A.(2018) 'The impact of the package opening process on product returns'. *Business Research,* 11(2), pp.279-308.doi:https://doi.org/10.1007/s40685-017-0055-x

Zhu, Y., Li, J., He, J., Quanz, B.L. and Deshpande, A.A.(2018) 'July. A Local Algorithm for Product Return Prediction in E-Commerce'. *In IJCAI* ,pp. 3718-3724. Available at:https://www.researchgate.net/profile/Yada_Zhu/publication/326203879_A_Local_Algorithm_for_Product_Return_Prediction_in_E-Commerce/links/5cf504ce299bf1fb18538ae3/A-Local-Algorithm-for-Product-Return-Prediction-in-E-Commerce.pdf (accessed 25 July 2022).