PREDICTIVE MODELING OF RETURN OCCURRENCE IN E-COMMERCE APPAREL MARKET: A COMPARATIVE STUDY OF LOGISTIC REGRESSION, LASSO, XGBOOST AND RANDOM FOREST TECHNIQUES

A Master's Thesis

by
ASİYE ASLI KUTLU

Department of Management
İhsan Doğramacı Bilkent University
Ankara
May 2024

ASIYE ASLI KUTLU
PREDICTIVE MODELING OF RETURN OCCURRENCE IN E-COMMERCE APPAREL MARKET: A COMPARATIVE STUDY OF LOGISTIC REGRESSION, LASSO, XGBOOST AND RANDOM FOREST TECHNIQUES
Bilkent University 2024

To Yiğit

PREDICTIVE MODELING OF RETURN OCCURRENCE IN E-COMMERCE APPAREL MARKET: A COMPARATIVE STUDY OF LOGISTIC REGRESSION, LASSO, XGBOOST AND RANDOM FOREST TECHNIQUES

Graduate School of Economics and Social Sciences
of
İhsan Doğramacı Bilkent University

by

ASİYE ASLI KUTLU

In Partial Fulfilment of the Requirements for the Degree of
MASTER OF SCIENCE

THE DEPARTMENT OF MANAGEMENT
İHSAN DOĞRAMACI BİLKENT UNIVERSITY
ANKARA

May 2024

PREDICTIVE MODELING OF RETURN OCCURRENCE IN E-COMMERCE APPAREL MARKET: A COMPARATIVE STUDY OF LOGISTIC REGRESSION, LASSO, XGBOOST AND RANDOM FOREST TECHNIQUES
by Asiye Aslı Kutlu

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science in Business Administration.

____

Yasemin Limon Kahyaoğlu
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science in Business Administration.

____

Banu Yüksel Özkaya
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science in Business Administration.

____

Canberk Üçel
Examining Committee Member

Approval of the Graduate School of Science in Business Administration.

____

Refet S. Gürkaynak
Director

ABSTRACT

PREDICTIVE MODELING OF RETURN OCCURRENCE
IN E-COMMERCE APPAREL MARKET: A COMPARATIVE STUDY OF
LOGISTIC REGRESSION, LASSO, XGBOOST AND RANDOM FOREST
TECHNIQUES

Kutlu, Asiye Aslı
M.S., Department of Management
Supervisor: Assist. Prof. Dr. Yasemin Limon Kahyaoğlu
Co-Advisor: Assoc. Prof. Dr. Fehmi Tanrısever

May 2024

This study focuses on the development of a predictive model for return occurrence in the apparel segment of an e-commerce company based in Turkey. Leveraging data provided by the company, the research employs various machine learning techniques to explore the impact of various factors on return. Models are developed, incorporating predictor variables related to product, supplier, customer and shopping information with the final model also including interaction of these variables. LASSO is applied to simplify the final model and select the most relevant variables. Performance metrics; AUC score, accuracy, precision, and recall are evaluated for the models, with comparisons made between logistic regression, LASSO, XGBoost, and Random Forest. Findings indicate that logistic regression models outperform XGBoost and Random Forest in terms of AUC score.

Keywords: return, online shopping, logistic regression, prediction models.

ÖZET

E-TİCARET GİYİM PAZARINDA İADE TAHMİNİ MODELLEME: LOJİSTİK REGRESYON, LASSO, XGBOOST VE RASTGELE ORMAN TEKNİKLERİNİN KARŞILAŞTIRMALI BİR ÇALIŞMASI

Kutlu, Asiye Aslı

Yüksek Lisans, İşletme

Tez Danışmanı: Yrd. Doç. Dr. Yasemin Limon Kahyaoğlu

2. Tez Danışmanı: Doç. Dr. Fehmi Tanrısever

Mayıs 2024

Bu çalışma, Türkiye'de faaliyet gösteren bir e-ticaret şirketinin giyim segmentinde iade oluşumu için tahmin modelinin geliştirilmesine odaklanmaktadır. Şirket tarafından sağlanan verileri kullanarak, araştırma çeşitli makine öğrenimi tekniklerini kullanarak iade oluşumları üzerinde çeşitli faktörlerin etkisini keşfetmektedir. Modeller, ürün, tedarikçi, müşteri ve alışveriş bilgileri ile ilgili tahmin edici değişkenlerin yanı sıra bu değişkenlerin etkileşimini de içerecek şekilde geliştirilmiştir. Nihai modeli basitleştirmek ve en etkili değişkenleri seçmek için LASSO uygulanmıştır. Modeller, AUC skoru, doğruluk, kesinlik ve duyarlılık gibi performans metrikleri ile karşılaştırılmıştır. Lojistik regresyon modelinin AUC skoru açısından XGBoost ve Rastgele Orman'dan daha iyi performans sergilediği görülmektedir.

Anahtar Kelimeler: iade, cevrimiçi alışveriş, e-ticaret, lojistik regresyon, tahmin modelleri.

ACKNOWLEDGMENTS

This master's program has been a great experience, and I am deeply appreciative of the support provided by my professors and family. Their support has been invaluable throughout this journey. My sincere gratitude to Prof. Yasemin Limon for her academic support and patience; she was helpful and accessible whenever I needed her assistance. I am also grateful to her for overseeing my master's process from the beginning. I am thankful to Prof. Fehmi Tanrısever for being my research mentor and guiding my studies. Additionally, I want to express my appreciation to Prof. Canberk Üçel and Prof. Banu Yüksel Özkaya for their valuable contributions as members of my thesis committee. To my mum and dad, I always had the chance to follow my own way thanks to your trust and encouragement. Finally, my deepest thanks go to my husband, who supported me throughout this entire master's process. Thank you for always being there when I needed, and thank you for learning to teach me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Online shopping has shown an increasing trend for a decade, especially making a peak in 2020 (8% growth from 2019 to 2020) as a result of Covid-19 ("Impact of COVID Pandemic on eCommerce", 2024). During the COVID-19 pandemic, as customers turned to online shopping, businesses also took their place in the online market. However, conducting online sales also brought about the necessity for businesses to manage the issue of returns. As stated in "The Right Fit: How AI is Changing eCommerce Apparel Returns" (2024), online retailers can encounter return rates exceeding 20% in the apparel sector, whereas this rate is 9% for brick-and-mortar retailers. The National Retail Federation (NRF) reported that online returns in the US increased from 18.1% in 2020 to 20.8% in 2021. Apparel is the second most returned category with an average return rate of 12.2% according to the NRF survey conducted in 2022 (Inman, 2022). Besides the pandemic, apparel shopping has clearly shifted to online mediums. According to a report from Freedman (2023), online shoppers purchased more apparel online from June 2022 to June 2023. Besides, Fokina (2024) shows that fashion, clothing, and accessories are the most bought categories by both female and male internet users. Figure 1 is taken from "Impact of COVID Pandemic on

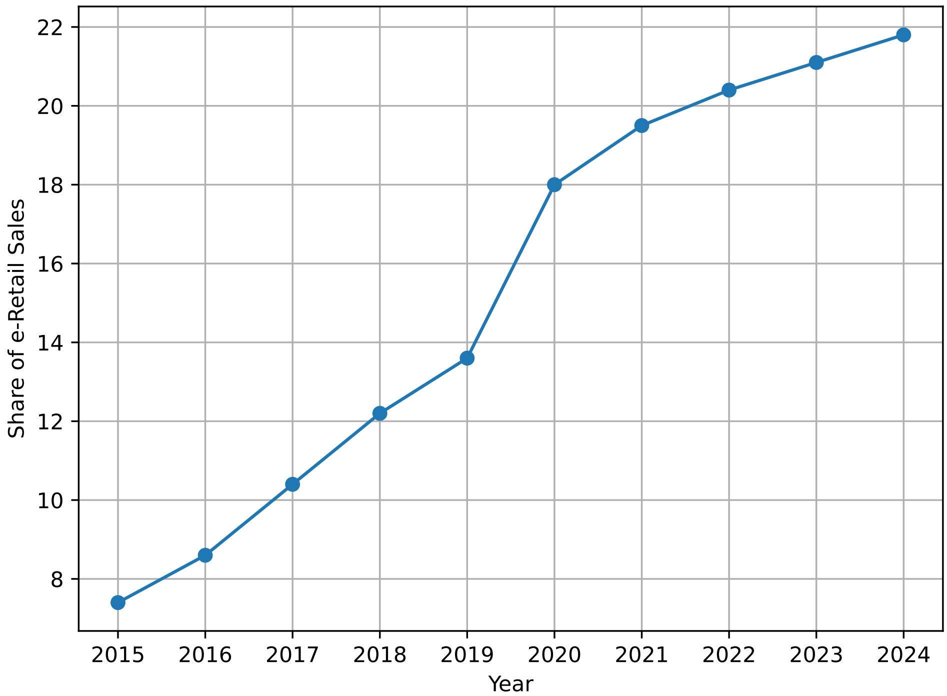

Figure 1: E-Retail Share of Total Global Retail Sales 2015-2024

eCommerce" (2024) in order to demonstrate the increasing trend of e-commerce share in total global retail sales. As Inman (2024) projections indicate that non-store and online sales will experience growth ranging between 7% and 9% in 2024, surpassing the sales value of the previous year, which amounted to $1.38 trillion in 2023.

Return is a vital issue for e-commerce in terms of operational and environmental aspects. According to Kim et al. (2022), online shopping produces 4.8 times more waste than brick-and-mortar shopping when the same amount is spent and also shipping of returned items contributes to transport emissions. Online retailers have to provide an easy return process for their customers in the competitive e-commerce market but also need to keep returned product amounts as low as possible in order to protect their profit rates. Today, with the rise of social media and developed product picturing, many consumers benefit from the ease of online shopping. However, it is still not possible to enable consumers to feel the texture and try the items online. When the product is a garment, sizing is often arbitrary, as different producers use different sizing molds. Many countries enforce long return acceptance periods or no-reason-to-return rights with their laws on Consumer Protection to protect citizens from fraud or disappointment. This situation creates a big disadvantage for online retailers. Returns are an inevitable part of the electronic commerce industry, impacting customer satisfaction, brand reputation, competitive advantage, and operational efficiency. Hence, e-commerce companies have adopted easier return processes year by year in order to build steady customer relationships. While returns are inevitable in the e-commerce industry, they are still predictable. Predicting return rates for online retailers offers opportunities to improve operations, enhance customer satisfaction, reduce costs, and support business growth in today's competitive online marketplace. The primary objective of this study is to build an evidence-based model to predict the return status of future transactions. The study was conducted in collaboration with a leading e-commerce company in Turkey. The study focuses on predicting future returns based on several predictor variables, which can be categorized as price, supplier, product, customer, and time effects. Through analyzing the influence of various predictor variables and investigating machine learning techniques—specifically, Logistic Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting (XGBoost), and Random Forest—this study endeavors to develop effective analytic models for predicting returns for the e-commerce sector. The remainder of the thesis is organized as follows: The review of the related literature is provided in Chapter 2. Chapter 3 describes the data structure and explains the steps involved in cleaning and preprocessing. Chapter 4 is dedicated to model selection, where the predictors important for return occurrence are discussed, and different classification methods are applied. Lastly, in Chapter 5, the conclusion of the study and suggestions for future work are presented.

1.1 Societal Impact

As stated earlier in this chapter, return of online purchase has huge impact on environment specifically because of packaging and transportation. In 2022, over £9.5 billion of returned products were disposed of in landfills due to resell operations and costs (George, 2024). Besides, e-commerce returns contribute up to 24 million metric tons of carbon dioxide emissions annually (Optoro, 2022). Another fact is that the carbon emissions from returns contribute an additional 30% from the initial delivery (Davison, 2024). According to a survey conducted by Narvar (2020), almost two-thirds of online shoppers bought multiple sizes or colors of the same item in 2020, with the intention of returning some of the items. This practice, referred to as bracketing in the e-commerce industry, is especially common among luxury goods buyers and shoppers under 30 (Ivanova, 2020). Returns of online purchased items are also crucial in terms of their economic results. Stambor (2022) reports that approximately 91% of businesses are experiencing a faster increase in the rate of returns compared to their revenue growth. NRF & Retail (2022) shows that out of \$1.29 trillion in total online sales in 2022, \$212 billion of purchases were returned, and among these returns, \$22.8 billion were detected as fraudulent. Additionally, US retailers reported \$816 billion of lost due to returns processes in 2022 (Inman, 2022). It is quite often that returned goods end up in landfills. Reselling a returned item requires operations such as inspection, logistics, packaging, etc., and in some cases, reselling is not possible due to legal reasons (Ivanova, 2020). Hence, companies opt for disposal instead of reselling. The disposal of returned goods is detrimental not only to the environment but also economically. The societal impact of online returns extends beyond environmental and economic consequences therefore, addressing this issue requires a collective effort from businesses, consumers, and policymakers to develop more sustainable practices and reduce the negative impacts of online returns.

CHAPTER 2

LITERATURE REVIEW

This research draws upon existing operations management literature on optimization, prediction modeling, and marketing literature on e-commerce and product return. Since our aim is to build a data driven prediction model with logistic regression. Theoretical backgrounds of logistic regression have been touched in this chapter. In addition to logistic regression, model performance has been investigated with application of the Least Absolute Shrinkage and Selection Operator (LASSO), XGBoost and Random Forest algorithms as a part of feature selection. Therefore, this section also discusses the widespread usage and backgrounds of LASSO, XGBoost, and Random Forest.

2.1 Prediction in Online Shopping

Ambilkar et al. (2022) and Duong et al. (2022) provide comprehensive reviews of the literature on product returns management and future research directions. As stated in these studies, the literature on the domain of product returns has extensively developed in recent years. In this section, we focus on product returns in online channels. Unal & Park (2023) investigates the impact of adopting one-click buying on customer behavior in the context of a specific retailer and show that the adoption of one-click buying led to a significant increase in sales and customer engagement. Specifically, Unal & Park (2023) found that customers who used one-click buying had a higher purchase frequency, spent more per transaction, and were more likely to return to the retailer's website. This study proves the relationship between ease of use and customer loyalty. We can assume a similar relationship between ease of return processes and customer loyalty as well. Rodrigues & Alves (2018) is another study that focuses on return issues in the e-commerce industry, particularly in the luxury garment segment. This study focuses on cost reduction in return management by using machine learning to redirect customer returns to new clients. Rodrigues & Alves (2018) show that big data can be used to improve supply chain efficiency in returns by providing insights and solutions that can improve return management. Cui et al. (2020) is one of the latest studies that develops a return prediction model for an automotive accessories manufacturer in the US. The authors in this study explore several factors that contribute to predicting return volume, such as sales, time, product features, retailers, and production processes. They build on previous research on consumer product returns and extend it by using advanced machine learning methods including high-dimensional variable/model selection methods and tree-based statistical machine learning methods to predict return volume. While research on predicting product returns remains limited, there is a significant body of literature focusing on predicting customer behavior in retail through historical data analysis. For instance, Moe & Fader (2004) offer a model that estimates the probability of a purchase given a customer's visit history. This model allows for the inclusion of customer characteristics and mentions two components that affect an individual's conversion behavior: visit effects and purchasing threshold effects. Moe & Fader (2004) compare the proposed stochastic model with the logistic regression model along several dimensions, including in-sample log-likelihood, holdout log-likelihood, and predictive accuracy for an individual's next visit. Another predictive study proposed by T. Huang et al. (2019) claims that predictive analytics involve the use of statistical and machine learning techniques to analyze historical data and make predictions about future events or behaviors whereas prescriptive analytics, involve the use of optimization and simulation techniques to identify the best course of action to take in a given situation. Kadiyala et al. (2021) aims to maximize the promotion's profitability and e-mail click-through rate for retailers. They argue that LASSO can help to improve the model's predictive accuracy and reduce overfitting by shrinking the coefficients of less important variables to zero (Kadiyala et al., 2021).

2.2 Use of Prediction Models in other Businesses

Prediction is a vital issue not only within the retail industry but also across diverse domains, including healthcare and hospital operations. As demonstrated by Somanchi et al. (2022), timely prediction is critical in emergency service scenarios. They propose a two-stage classification methodology to achieve swift predictions while maintaining high accuracy levels. The first stage of the framework uses only information that is available at the time of patient arrival to make a preliminary prediction of whether the patient will be admitted to an interior hospital unit or discharged from the emergency department whereas the second stage of the framework uses additional information that becomes available later in the patient's ED visit to refine the prediction made in the first stage. Many other studies leverage data-driven and machine-learning techniques. Apart from predicting returns, other studies focusing on data-driven predictions have been explored, such as the research conducted by Chang et al. (2021). The study uses deep learning techniques such as Long Short-Term Memory and classical supervised learning approaches to analyze data from Continuous Emission Monitoring

Systems and identify anomalies in reporting data that may indicate environmental violations. Ferreira et al. (2016) offers a machine learning algorithm (regression trees in specific) to overcome two challenges faced by the online retailers which are estimating lost sales due to stock-outs and predicting demand for items that have no historical sales data. Chou et al. (2022) use probabilistic modeling and machine learning techniques to predict customer repurchase in an online retail setting. They incorporate the estimates of BG/BB (BetaGeometric/Beta-Bernoulli) models into LASSO regression and show that the interdisciplinary integration of two modeling approaches improves prediction accuracy and interoperability. Falkenberg & Spinler (2022) build a predictive maintenance model to forecast equipment breakdowns and find that K-Nearest-Neighbors and Random Forest Classifier are cost-optimal. Next, we discuss the relevant literature on logistic regression, LASSO, XGBoost, and random forests in more detail.

2.3 Logistic Regression

Hence our primary focus is to use logistic regression models. Logistic regression has been used for the analysis and prediction of a dichotomous outcome since it was first proposed as an alternative to OLS regression and discriminant analysis in the late 1960s (Peng et al., 2010). It is a statistical method used to model the relationship between a binary dependent variable and one or more independent variables, which makes it an appropriate technique for this study while modeling returned or not returned (binary dependent variable). Before establishing a return prediction model, eight articles using logistic regression were reviewed. These articles include an empirical analysis to derive a logistic regression as well as regression coefficients, which is similar to what is aimed in our study. Studies of Alexander et al. (1996), Diamond et al. (2000), Machamer & Gruber (1998), McNeal (1998), Meisels & Liaw (1993), Rush & Vitale (1994), Smith

(1997) and Trusty (2000) guided us to understand application and evaluation of the logistic regression. As summarized in Peng et al. (2010), a researcher can evaluate their logistic regression model by an overall evaluation of the model, statistical tests of individual predictors, goodness-of-fit statistics, and validations of predicted probabilities. Pearce & Ferrier (2000) is another study that evaluates the performance of a logistic regression model by two components: reliability and discrimination capacity. For reliability, Pearce & Ferrier (2000) describes techniques for evaluating calibration bias and spread error. For discrimination capacity, Pearce & Ferrier (2000) describes the use of a 2x2 classification table (confusion matrix) and measuring discrimination capacity in terms of the area under a relative operating characteristic (ROC) curve.

2.4 LASSO, XGBoost, Random Forest

The Least Absolute Shrinkage and Selection Operator (LASSO) was first introduced by Tibshirani (1996) with the aim of reducing the complexity of the model by penalizing large coefficients. This leads to simpler and more interpretable models. It selects important variables by setting some coefficients to exactly zero, thus improving feature selection and model simplification. Finally, LASSO enhances prediction accuracy by finding a balance between bias and variance. Logistic regression and LASSO may be incorporated in several domains of research. For instance, T. T. Wu et al. (2009) demonstrates the utility of LASSO penalized logistic regression in improving the accuracy, efficiency, and interpretability of genome-wide association analyses, particularly in the context of complex diseases and large-scale genetic datasets. Ma et al. (2016), employs the LASSO technique in the variable selection stage where they aim to contribute to the field of marketing analytics by proposing a systematic approach that combines advanced statistical techniques with domain-specific knowledge to optimize demand forecasting for retail SKUs. By using LASSO regression, Ma et al. (2016) is able to identify the most influential variables that impact the sales forecasts of SKUs, while also addressing the high dimensionality of the data. The LASSO method helps in automatically selecting important variables and shrinking the coefficients of less relevant variables toward zero, thus improving the model's interpretability and predictive performance. Similarly, Cui et al. (2020) employs LASSO as part of the methodology to select a small number of interaction terms that are useful in predicting return volume from a large pool of potential predictor variables and finds that the LASSO model is effective in selecting important variables for prediction and achieves high accuracy in predicting future return volume. Another machine learning algorithm applied in the model selection part of our study is the Extreme Gradient Boosting (XGBoost) algorithm, which was first proposed by Chen & Guestrin (2016). This algorithm is introduced as an efficient and scalable implementation of gradient-boosting decision trees. Chen & Guestrin (2016) introduced the advantages of XGBoost over existing tree-based algorithms in terms of speed, performance, and scalability, making it a popular choice in various machine-learning competitions and applications. The Random Forest algorithm was initially introduced by Breiman (2001), demonstrating its efficacy in achieving accurate predictions by combining multiple decision trees. This algorithm utilizes random features and employs bagging techniques to mitigate overfitting and enhance reliability (Breiman, 2001). J.-C. Huang et al. (2020) compares different modeling techniques including XGBoost and Random Forest and demonstrates that RF performs better with lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values on the testing dataset compared to XGBoost in the context of blood pressure prediction during hemodialysis. In addition to the aforementioned literature, our study is based on the product returns in an online retail medium and we aim to propose a model with an acceptable predictive performance with elemental features which can easily be obtainable for retailers' databases. Furthermore, we develop the main effect model with advanced machine learning techniques; specifically LASSO, Random Forest, and XGBoost.

CHAPTER 3

EMPIRICAL SETTING

In this study, data is provided by one of the leading online retailer companies in Turkey. The aim is to develop a model and utilize it to predict whether a product will be returned or not. In this chapter, we describe the data and outline the steps involved in cleaning and preprocessing it. In Chapter 4, subsequent to dataset preparation, we employ logistic regression, LASSO, XGBoost, and Random Forest to predict product return likelihood.

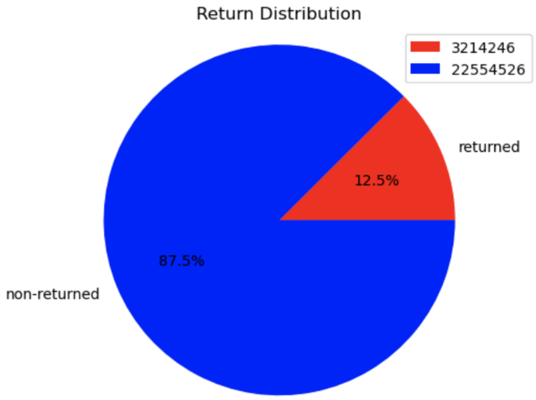

Figure 2: Return Distribution

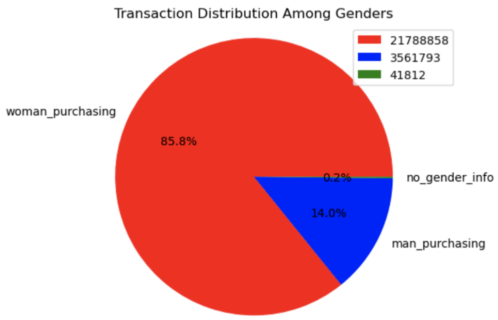

Figure 3: Customer Gender Distribution

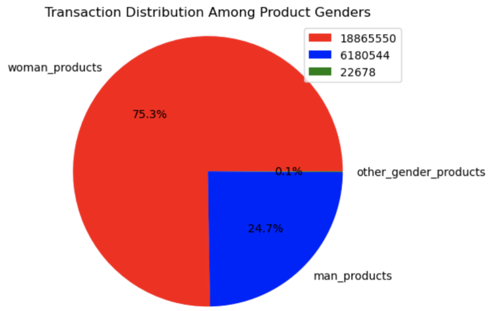

Figure 4: Product Gender Distribution

3.1 Data Description

Eight transaction data files, obtained from the Turkish retailer, provide a data set containing 25.768.772 transactions that occurred between May 2021 and July 2021. The descriptions for transaction data variables are given in Table 1. As shown in Figure 2, 3.214.246 number of transactions ended with a return while 22.554.526 were non-returned which gives us the return ratio as 12.47%. When we study shopping distribution among genders, Figure 3 shows 21.788.858 transactions made by women (84.5%), 3.561.793 made by men (13.8%) and 418.12 (1.6%) transactions has no gender information. 7.957.608 number of transactions made by elite users (30.9%) and 17.811.164 made by non-elite users (69.1%). The e-tailer provides a wallet feature in their app and website so customers can transact money from their credit cards to their wallets and use it for future orders. In the given data set 10.424.445 number transactions were made with a wallet (40.5%) whereas 15.344.327 were made without a wallet (59.5%). Another payment option for customers is to save their credit card in the app or the website so they do not need to type information each time. Data analysis shows that 9.443.875 transactions were made with a saved card (36.65%) and 16.324.897 made without a saved card (63.35%). Order time also gives us an insight as 12.854.931 orders were given daytime (49.89%) whereas 12.913.841 orders were given after 6 pm (50.11%). Finally, when ordered products were analyzed it was seen that

18.865.550 products were bought for women (75.26%), 6.180.544 products were bought for men (24.65%), and 22.678 products were bought for another gender category (0.09%) as seen in Figure 4. Also, the age distribution of customers can be found in Figure 5.

Table 1 demonstrates variables given in the transaction dataset, some critical variables in this dataset are order_date, user_id_hashed, is_elite_user, original_price and is_wallet_trx. User demographics dataset is shown in Table 3, birth_date, membership_date and gender is used to derive some other useful variables. Product dataset contains color_name, gender_name and other variables seen in Table 2. Supplier related information such as supplier_return_rate and supplier_return_rate_defect are given in Table 4, Table 5, Table 6.

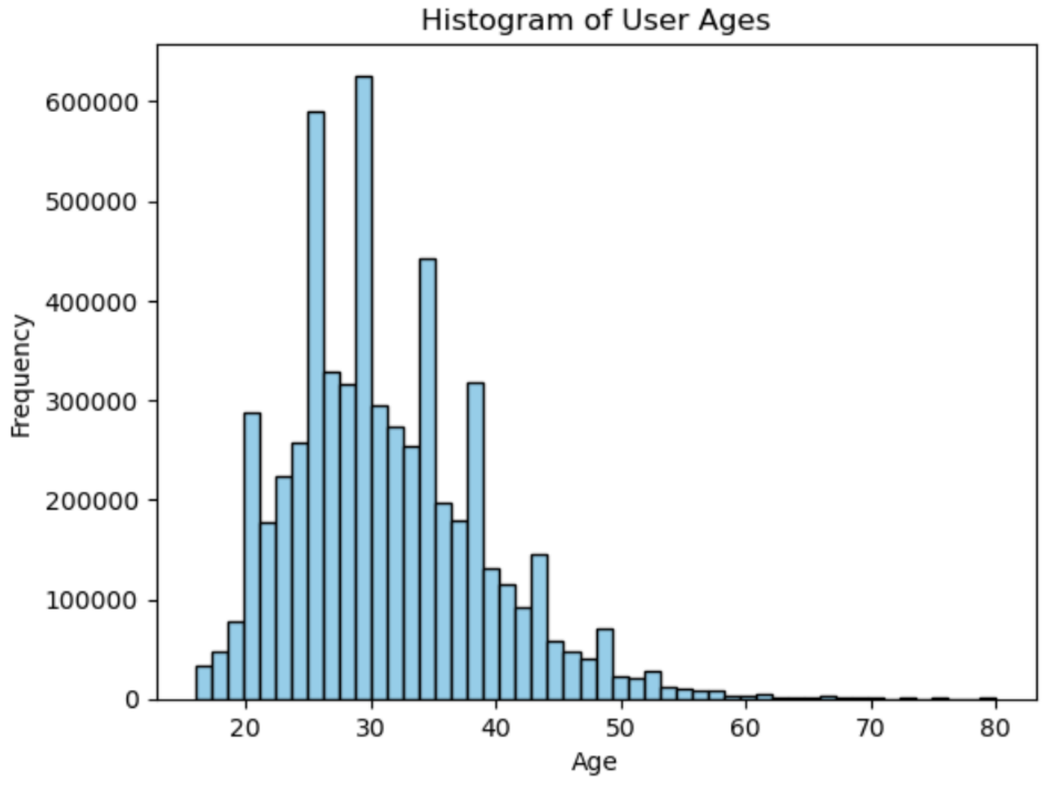

Figure 5: Age Distribution of Users

Data files explained in this section were imported as data frames by using the Pandas library in Python.

Table 1: Transaction Dataset Variables

Variable Name	Definition
order_date	The date when the transaction occurred
user_id_hashed	Unique ID of the user
is_elite_user	Whether the user has a user status at the time of the order
supplier_id_hashed	Unique ID of the seller
order_line_item_id_hashed	Unique ID for every sold item within the basket
order_parent_id_hashed	Unique ID for the placed order (same for all items within the basket)
product_content_id	ID for the product independent of its size. Different sizes of the same colored product have the same product_content_id
product_variant_id	ID of the product. Different sizes of the same colored product have different product_variant_id
original_price	List price of the product
discounted_price	The price at which the product is bought after the promotion is applied. Coupon discount is not included in this price
ship_cost	Shipping cost incurred
coupon_id	ID of the discount coupon used for the transaction if any
coupon_discount	The TL discount applied by the coupon if any
promotion_name	Name of the promotion applied if any
promotion_award_value	How much discount the promotion applies if any. It can be a percentage discount or a flat discount
is_wallet_trx	Whether the transaction occurred using app wallet or not
is_saved_card_trx	Whether the transaction occurred using the customer's saved credit card or not
is_returned	Whether the product is returned by the customer or not

Table 2: Products Dataset Variables

Variable Name	Definition
product_id	ID for the product independent of its color and size
product_content_id	ID for the product independent of its size.
	Different sizes of the same colored products have the same product_content_id
product_variant_id	ID of the product.
	Different sizes of the same colored products have different product_variant_id
product_name	Name of the product
brand_id	Brand ID of the product
brand_name	Brand name of the product
gender_id	Designated gender ID of the product if any
gender_name	Designated gender name of the product if any
category_id	Category ID of the product
category_name	Category name for the product
color_id	Color id of the product if any
color_name	Color of the product if any
supplier_color_name	Color given by the supplier
attribute_name	Name of the attribute, always equals to size in the dataset
attribute_value	Size of the product

3.2 Data Cleaning And Preprocessing

This section involves data sanitization and basic feature engineering steps wherein additional columns derived from existing features were introduced as well as

Table 3: **User Demographics Dataset Variables**

Variable Name	**Definition**
user_id_hashed	Unique ID of the user
birth_date	Birth date of the user if given
membership_date	When the user became a member on the app or the website
gender	Gender of the user if given

Table 4: Supplier Return Dataset Variables

Variable Name	**Definition**
supplier_id_hashed	Unique ID for the seller
supplier_return_rate	Ratio of returned products to the total number of sold products by the same seller

Table 5: Supplier Defective Return Dataset Variables

Variable	**Definition**
supplier_id_hashed	Unique ID for the seller
supplier_return_rate_defect	Ratio of returned defective products to the total number of sold products by the same seller

Table 6: Supplier Disputed Return Dataset Variables

Variable	**Definition**
supplier_id_hashed	Unique ID for the seller
total_claim	Total number of returns for the seller
unresolved_claim	Total number of disputed claims for the seller
unresolved_accepted_claim	Total number of disputed claims that are decided on behalf of the buyer
unresolvedclaim_percentage	Ratio of unresolved_accepted_claim to unresolved_claim

uninformative rows and columns were removed. Some columns are found to be uninformative since the information given by them can be captured by other features or they are basically not practical to analyze. Excluded rows are those that mainly contain faulty entries or missing values.

The data excluded from the dataset is shown in Table 7 and Table 8.

Table 7: Rows Excluded from the Dataset

Uninformative Rows
Transaction dataset rows lacking return information
Product dataset rows with an unknown attribute of gender_name
Transaction dataset rows if original_price value is zero or negative
Transaction dataset rows if coupon_discount value is positive
User demographics rows with deleted_data_protection_request attribute

Table 8: Features Excluded from the Dataset

Uninformative Columns
coupon_id
promotion_name
attribute_name
color_id
supplier_color_name
product_name

In addition to excluding the uninformative data, we also derived additional variables aiming to better explain return occurrences. New features can be seen in Table 9.

In order to run logistic regression it was required to convert categorical data to integers. One hot encoding technique applied to convert each categorical value into a new column and assigns a binary value of 1 or 0 to those columns. To be more specific, gender_name is changed with is_cloth_for_female, is_cloth_for_male and is_cloth_for_unisex in order to determine genders of the products sold. gender is transformed to is_male and is_female for customer genders. Similar transformation is made for transaction_hour thus, order time is shown with is_time_00_06, is_time_06_12, is_time_12_18, is_time_18_24. color_name

Table 9: **Derived Data from Existing Features**

Derived Data	Description
abused	Orders containing products with the same colors but different sizes by using order_id and product_content_id
promotion_percentage	original_price - discounted_price / original_price
discount_percentage	coupon_discount / original_price
membership_len_in_days	order_date - membership_date
paid_amount	discounted_price + coupon_discount + shipping_cost
ship_cost_percentage	ship_cost / paid_amount
user_age	order_date - birth_date
original_log_price	Natural logarithm of original_price
discounted_log_price	Natural logarithm of discounted_price
paid_log_amount	Natural logarithm of paid_amount

is analyzed and colors are categorized in three group as is_black, is_white, is_others for the sake of simplicity.

After applying one-hot encoding, the dataset was suitable for logistic regression analysis. Termed the ”merged data frame”, this final dataset contains clean transaction, product, user demographic, and supplier return data into a unified framework.

Before running logistic regression, correlations among all predictive variables were checked with correlation matrix and observed quite high correlations among original_price and discounted_price as seen in the Figure 6.

Since original_price and discounted_price are highly correlated and we created promotion_percentage as a derivation of these two variables, we decided to eliminate discounted_price from the logistic regression models to not to increase multicollinearity.

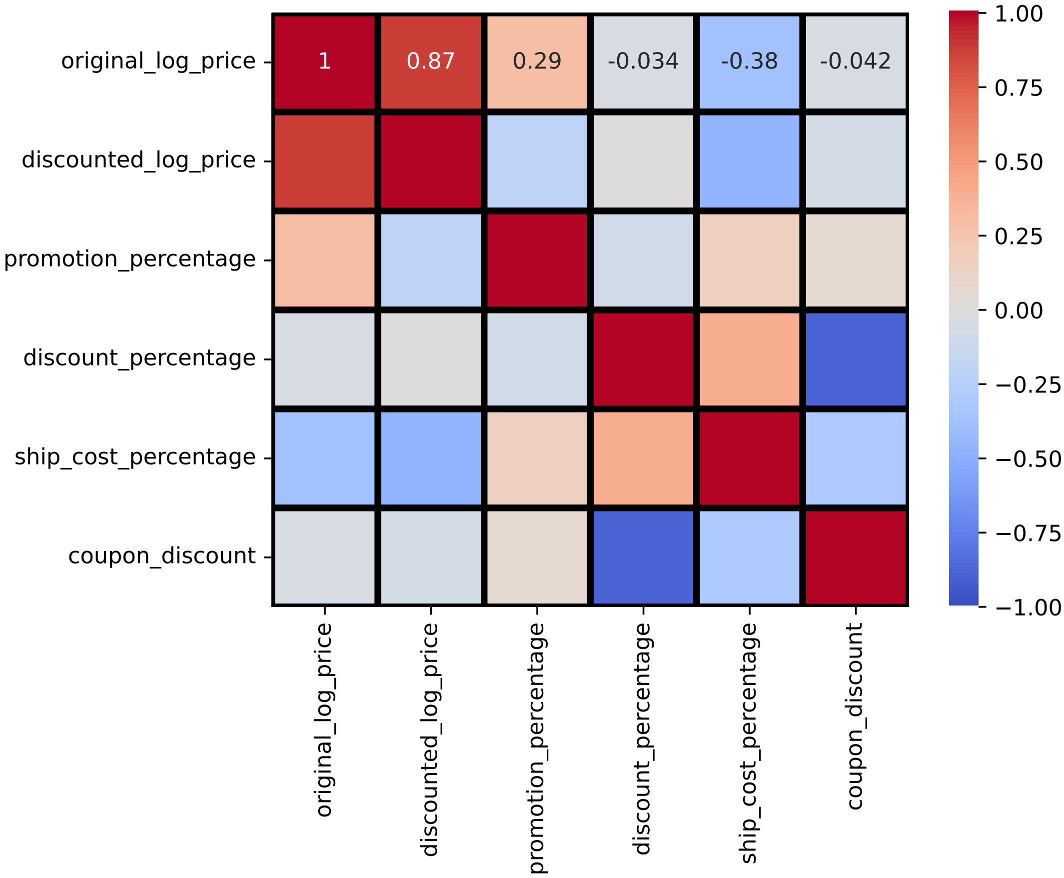

Figure 6: Correlation Matrix of Price Effect Variables

CHAPTER 4

MODEL SELECTION

In the first part of model selection, it is aimed to develop main effect models with the variables provided in the given data without further feature engineering. By doing so, we desire to reach some models with acceptable predictive performance with the basic data that can be easily derived from the databases of e-commerce companies. Main effects models were established by the addition of different variable categories consecutively. We created interaction terms from the variables given and added these terms to the final model. Then we presented the logistic regression results of each model. We utilized LASSO with the most extensive model, to select informative variables and interaction terms proposed. Finally, LASSO, XGBoost, and Random Forest techniques were applied and compared with logistic regression results at the end of this section.

4.1 Predictive Modelling Methods

4.1.1 Logistic Regression

Logistic regression is a widely used statistical technique for modeling the relationship between a binary outcome variable (such as "success" or "failure", "yes" or "no", "1" or "0") and one or more predictor variables. Unlike linear regression, which is used for continuous outcome variables, logistic regression is specifically designed for binary outcomes. In logistic regression, the relationship between the predictor variables and the probability of the outcome occurring is modeled using the logistic function, also known as the sigmoid function (Saeed, 2021). This function maps the linear combination of predictor variables to a probability value between 0 and 1, representing the likelihood of the event happening as described in Hosmer et al. (2013). In logistic regression, the relationship between the predictor variables and the probability of the outcome occurring is modeled using the logistic function

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}} \ ,$$

where $P(Y = 1 \mid X)$ is the probability of the outcome being 1 given the values of the predictor variables X, e is the base of the natural logarithm, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the logistic regression model and $X_1, X_2, \ldots, X_k$ are the coefficients of the logistic regression model.

The coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are estimated from the data using techniques like maximum likelihood estimation. These coefficients represent the change in the log-odds of the outcome variable associated with a one-unit change in the corresponding predictor variable, assuming all other variables are held constant. Once the model is fitted, it can be used to predict the probability of the outcome variable being 1 for new observations based on their predictor variable values. A common threshold (often 0.5) is used to classify observations into one of the two categories based on their predicted probabilities. In this study, we selected the threshold value that maximizes the F1 score.

As stated in Agresti (2012) model evaluation is an essential step in logistic regression analysis. Various metrics can be used to assess the performance of the model, including accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (ROC) curve. Since logistic regression is a statistical method used for modeling the relationship between a categorical dependent variable and one or more independent variables it is generally applied in research questions involving binary outcomes or classification tasks.

4.1.2 The Least Absolute Shrinkage and Selection Operator (LASSO)

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regression analysis method applied for variable selection and regularization. As introduced in Friedman et al. (2009), the objective function of LASSO includes a penalty term which is the sum of the absolute values of the coefficients multiplied by a tuning parameter called lambda or alpha. The penalty term leads shrinkage of coefficient estimates towards zero enabling variable selection and regularization. The degree of regularization can be controlled by adjusting the tuning parameter. To be more clear; larger values of lambda lead to more aggressive shrinkage and stronger variable selection, while smaller values allow for less regularization (Friedman et al., 2009). LASSO regression produces a simpler and more interpretable model by setting some coefficients to zero. Interpretation of LASSO results involves examining the coefficients of the selected variables, with non-zero coefficients indicating significant predictors and zero coefficients implying non-significance and exclusion from the model.

LASSO aims to estimate the coefficients β_j while promoting sparsity in the model (Hastie et al., 2015). This is achieved by adding a penalty term to the ordinary least squares (OLS) objective function. The LASSO objective function can be expressed as

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} ,$$

where β is the vector of regression coefficients to be estimated, X is the design matrix with dimensions $n \times p$, containing the predictor variables, λ is the regularization parameter (also known as the tuning parameter), controlling the strength of the penalty, $\| \cdot \|_2$ denotes the L2 norm (Euclidean norm) of a vector, representing the sum of squared values, $\| \cdot \|_1$ denotes the L1 norm (absolute value norm) of a vector, representing the sum of absolute values. LASSO is used in regression analysis in order to investigate prediction, variable selection, and model interpretation particularly in the case of high-dimensional data where the number of predictors is large relative to the number of observations.

4.1.3 XGBoost

XGBoost is a powerful machine learning algorithm mostly utilized in various prediction tasks, including classification and regression. Data preparation includes handling missing values, encoding categorical variables, and splitting the data into training and testing sets for XGBoost modeling. When the data is ready, parameters of XGBoost model (learning rate, maximum tree depth and regularization parameters) should be defined Raschka & Mirjalili (2019). When parameters are defined training phase begins with building an ensemble of decision trees in a chronological sequence. Each tree aims to correct the errors made by the previous trees. The model parameters are optimized in order to minimize the loss function during the training phase. Hyperparameter tuning is a crucial step for the performance of the XGBoost model Putatunda & Rama

(2018). Grid search or random search are some of the techniques that used for hyperparameter tuning. Trained and tuned model is evaluated using the testing data to interpret its performance and generalization ability Choudhary (2023). XGBoost classification tasks are generally evaluated by metrics as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC), while XGBoost regression tasks are evaluated by metrics such as mean squared error (MSE) or mean absolute error (MAE).

XGBoost builds an ensemble of decision trees to make predictions. The final prediction is obtained by summing the predictions from each individual tree. As Safhi et al. (2023) showed in their paper that, the objective function for XGBoost can be expressed as

$$\text{Obj} = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \ ,$$

where n is the number of training samples, $\mathcal{L}(y_i, \hat{y}_i)$ is the loss function, which measures the difference between the true target y_i and the predicted target $\hat{y}_i$, K is the number of trees in the ensemble, f_k represents the k-th decision tree, $\Omega(f_k)$ is the regularization term applied to each tree to prevent overfitting.

XGBoost performs well in regression and classification tasks by efficiently handling high-dimensional and imbalanced datasets. It is also robust to overfitting and scalable to handle large datasets.

4.1.4 Random Forest

Random forest is a robust machine learning technique for both regression and classification tasks. Breiman (2001) introduces the Random Forest algorithm, explaining its principles, construction, and applications. Similar to aforementioned techniques, first step is data cleaning and preprocessing followed by splitting the data into training and testing sets. Random Forest consists of a collection of decision trees, each trained on a subset of the data using a process called bootstrapping Pramoditha (2020). The hyperparameters of the model such as, number of trees in the forest, the maximum depth of the trees and the number of features considered at each split are specified. After that, decision trees in the forest are trained independently on a randomly selected subset of the training data. Random selection of subset enables to reduce overfitting. During the training phase, each decision tree is constructed by repeatedly dividing the data into smaller subsets with the aim of maximizing the purity of the created subsets in terms of the target variable. Training process lasts until a stopping criterion for example it can end when the maximum tree depth is met Er (2024). When training is completed for all decision trees predictions can be made using the Random Forest model. In the case of classification, the class labels are aggregated using a majority voting scheme to find the final predicted class for a given input sample while, in the case of regression the output of each tree is averaged to obtain the final predicted value. Model evaluation metrics are differentiate according to task types as well. Accuracy, precision, recall, F1-score, and area under the curve (AUC) are valid performance metrics for classification tasks while, mean squared error (MSE) or mean absolute error (MAE) are valid for regression tasks. Selecting the optimal values for hyperparameters (the number of trees in the forest, the maximum depth of the trees, the number of features considered at each split) is critical for Random Forest efficiency. The best combinations of hyperparameter values that maximizes model performance can be determined by grid search or random search Priya (2023). Mathematically, the prediction of a Random Forest model can be expressed as follows

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} T_i(X) \,,$$

where $\hat{y}$ is the predicted target value, $T_i(X)$ is the prediction of the i-th decision tree, and n is the number of trees in the ensemble.

Each individual decision tree T_i is trained using a subset of the training data and features, typically determined by hyperparameters such as the number of trees in the forest, the maximum depth of each tree, and the number of features considered at each split. Overall, Random Forest combines the predictions of multiple decision trees to improve predictive performance and reduce overfitting. Random forest is applicable for both regression and classification tasks, accommodating non-linear relationships and handling categorical, numerical, and mixed data types effectively. Random Forest is also robust to overfitting and capable of handling imbalanced datasets.

4.2 Models Proposed

In this section, we introduce six models that have been run in logistic regression with the purpose of return prediction. The merged data file comprises various columns that could serve as predictor variables. However, in line with the advantages mentioned in Caignya et al. (2018), including better predictive performance, prevention of overfitting on noisy data, and improving the interpretability of the model, we opted to include only the variables that we find helpful in explaining returns. is_returned aimed to be predicted by binary classification. When the train data set optimizes the coefficients of the model, the test data compares the results with the real labels. The predictor variables categorized in price effect, supplier effect, product effect, customer effect and time effect are given in detail in Table 10. All six models explained in this section can be found in Table 11. After conducting logistic regression on all six models, the highest-performing model is chosen for subsequent application of LASSO, XGBoost, and Random Forest techniques.

Table 10: Explations of Effects

Price Effect	Supplier Effect	Product Effect	Customer Effect	Time Effect
original_log_price discount_percentage promotion_percentage ship_cost_percentage	supplier_return_rate supplier_return_rate_defect	is_black is_white is_cloth_for_female is_cloth_for_male	is_female membership_len_in_days user_age is_elite_user abused is_wallet_trx	is_time_6_12 is_time_12_18 is_time_18_24 is_bank_holiday is_weekend is_May is_July

Table 11: Predictor Variables Across The Models

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
original_log_price promotion_percentage discount_percentage ship_cost_percentage	original_log_price promotion_percentage discount_percentage ship_cost_percentage supplier_return_rate supplier_return_rate_defect	original_log_price promotion_percentage discount_percentage ship_cost_percentage supplier_return_rate supplier_return_rate_defect is_black is_white is_cloth_for_female is_cloth_for_male	original_log_price promotion_percentage discount_percentage ship_cost_percentage supplier_return_rate supplier_return_rate_defect is_black is_white is_cloth_for_female is_cloth_for_male is_female membership_len_in_days user_age is_elite_user abused is_wallet_trx	original_log_price promotion_percentage discount_percentage ship_cost_percentage supplier_return_rate supplier_return_rate_defect is_black is_white is_cloth_for_female is_cloth_for_male is_female membership_len_in_days user_age is_elite_user abused is_wallet_trx is_time_6_12 is_time_12_18 is_time_18_24 is_bank_holiday is_weekend is_May is_July	original_price promotion_percentage discount_percentage ship_cost_percentage supplier_return_rate supplier_return_rate_defect is_black is_white is_cloth_for_female is_cloth_for_male is_female membership_len_in_days user_age is_elite_user abused is_wallet_trx is_time_6_12 is_time_12_18 is_time_18_24 is_bank_holiday is_weekend is_May is_July two-way interactions

Addition to Table 11, construction of the models can be seen in the equations given as following

Model 1:

$$P(\text{is\_returned} = 1 | \text{price\_effect})$$

Model 2:

$$P(\text{is\_returned} = 1 | \text{price\_effect, supplier\_effect})$$

Model 3:

$$P(\text{is\_returned} = 1 | \text{price\_effect, supplier\_effect, product\_effect})$$

Model 4:

$$P(\text{is\_returned} = 1 | \text{price\_effect, supplier\_effect, product\_effect} \\ \text{, customer\_effect})$$

Model 5:

$$P(\text{is\_returned} = 1 | \text{price\_effect, supplier\_effect, product\_effect} \\ \text{,customer\_effect, time\_effect})$$

Model 6:

$$P(\text{is\_returned} = 1 | \text{price\_effect, supplier\_effect, product\_effect} \\ \text{,customer\_effect, time\_effect, interaction\_terms})$$

Two-way interactions were created mostly among price, customer demographics, time and customer behaviour related features. To be more specific, discount_percentage x is_female is created to examine how female customers return behaviours change with discount percentage, is_cloth_for_female x is_female to study changes when female customers' buy female products and user_age x abused investigates how abusive buying behaviour change as age of the customer increases. Interaction terms created with is_elite_user, is_female and abused can be found in Table 12, logistic regression results of Model 6 (Table 21) demonstrates all interaction terms in detail.

Table 12: Interaction Terms

x is_elite_user	x is_female	x abused
is_time_6_12	is_time_6_12	ship_cost_percentage
is_time_12_18	is_time_12_18	promotion_percentage
is_time_18_24	is_time_18_24	discount_percentage
is_bank_holiday	is_bank_holiday	user_age
is_weekend	is_weekend	supplier_return_rate
is_wallet_trx	is_white	is_time_6_12
is_female	is_black	is_time_12_18
user_age	is_cloth_for_female	is_time_18_24
	is_cloth_for_male	is_bank_holiday
	discount_percentage	is_weekend
	promotion_percentage	
	abused	
	user_age	

4.3 Evaluation Criteria

4.3.1 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a measure used for model selection, particularly in the context of statistical modeling and regression analysis. It balances the goodness of fit of the model with its complexity, penalizing models with more parameters to prevent overfitting (Stoica & Selen, 2004). Lower AIC values indicate better-fitting models with fewer parameters, suggesting a better balance between goodness of fit and model complexity. In our study, we used AIC as an indicator to compare logistic regression results of the models since AIC allows us to compare models with different numbers of predictors and select the one that provides the best trade-off between goodness of fit and complexity. Besides, the comparison of Model 6 and LASSO also contains AIC, LASSO gives a negative AIC score at -9905220 which is lower than the AIC of Model 6 3444117. AIC is calculated as follows (Akaike, 1974)

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \, .$$

4.3.2 Area Under Curve

The Area Under the Curve (AUC), also known as the Receiver Operating Characteristic (ROC) curve, is a measure of the discriminative ability of a classification model, particularly for binary outcomes (Hilden, 1991). Hence, AUC is a legitimate measure for comparing logistic regression models. AUC scores range from 0 to 1, where higher values indicate better model performance. Specifically, AUC values lower than 0.5 refers to a classifier performing worse than random guessing. If AUC equals to 0.5 indicates that the model has no discriminative ability, it is equivalent to random guessing. Finally, AUC = 1 indicates a perfect classifier that perfectly separates the positive and negative classes ("Interpreting Logistic ROC Curves", n.d.). Since we employ logistic regression, LASSO, XGBoost and Random Forest algorithms for a classification task in this study, AUC is applicable in model comparison. As seen in Table 23 logistic regression has the highest AUC score, closely followed by LASSO, XGBoost and Random Forest have lower scores yet all four models scored higher than 0.5.

4.3.3 Accuracy

Accuracy is another metric for the performance evaluation of classification models. Accuracy measures the proportion of correctly predicted instances out of the total number of instances and provides an overall measure of predictive performance. It is calculated as follows

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Prediction}} .$$

An example of a confusion matrix given in Table 13 provides insights for accuracy.

Table 13: Example of A Confusion Matrix for Binary Classification

Actual	Predicted Positive	Predicted Negative
Positive	True Positives (TP)	False Negatives (FN)
Negative	False Positives (FP)	True Negatives (TN)

In our case, XGBoost and Random Forest resulted in higher accuracy values compared to logistic regression and LASSO, as seen in Table 23.

4.3.4 Recall

Recall, also known as sensitivity, is another performance metric employed in our model comparison since it is effective to evaluate the ability of a classification model to correctly identify positive instances from all actual positive instances in the dataset (Powers, 2017). Recall is calculated as follows

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} .$$

In our model evaluation, logistic regression and LASSO resulted in quite similar recall ratios. Recall ratio of XGBoost is extremely low compared to other models, given in Table 23.

4.3.5 Precision

Precision is a performance metric used to evaluate the quality of positive predictions made by a classification model. Hence a higher precision value indicates that the model is making fewer false positive predictions, when the cost of false positives is high is it critical to have a high precision value. In our case false positive is predicting not returned transactions as returned. Calculation of precision is as follows

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives } + \text{False Positives}}\,.$$

XGBoost performs best in terms of precision value, it is followed by Random Forest. Logistic regression and LASSO precision scores are lower, details can be found in Table 23.

4.3.6 F-1 Score

The final measure employed in model evaluation is F1 score, which is the harmonic mean of precision and recall, which provides a single metric to evaluate the balance between precision and recall in a classification model (Logunova, 2023). The F1 score ranges from 0 to 1, where a higher value indicates better model performance. It is especially valuable when the cost of false positives and false negatives are both significant, and the goal is to achieve a balance between minimizing both types of errors (Logunova, 2023). The F1 score is calculated as follows

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\,.$$

Among the four models, logistic regression gives the highest F1 score as seen in Table 23.

4.4 Results of Logistic Regression

In Model 1, our analysis focuses solely on examining the impact of price on return occurrence, incorporating original_log_price, promotion_percentage, discount _percentage and ship_cost_percentage as predictor variables. The logistic regression results of the first model are presented in Table 14. All proposed variables are significant and discount_percentage has a significant coefficient with a negative sign which corresponds to when the discount percentage of a product increases, the likelihood of this product to be returned decreases. ship_cost_percentage and promotion_percentage have a similar negative effect on return. Adversely, original_log_price has a positive coefficient which implies that price increase leads to increase in return incidence.

Table 14: Logit Regression Results of Model 1 – Price Effect

Dep. Variable:	is_returned	**No. Observations:**	4600827
Model:	Logit	**Df Residuals:**	4600822
Method:	MLE	**Df Model:**	4
Date:	Tue, 09 Apr 2024	**Pseudo R-squ.:**	0.05824
Time:	10:11:51	**Log-Likelihood:**	-1.7968e+06
converged:	True	**LL-Null:**	-1.9080e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-3.6163	0.012	-313.927	0.000	-3.639	-3.594
original_log_price	0.5739	0.002	236.118	0.000	0.569	0.579
promotion_percentage	-1.1865	0.007	-160.682	0.000	-1.201	-1.172
discount_percentage	-20.8581	0.163	-127.748	0.000	-21.178	-20.538
ship_cost_percentage	-3.4677	0.027	-128.944	0.000	-3.520	-3.415

In Model 2, supplier effect is added to price effect with supplier_return_rate and supplier_return_rate_defect predictor variables. Table 15 shows the findings of the logistic regression of Model 2. Both of the new variables are significant and positive which means if a supplier has a high return rate or high return rate due to defective products in the past, probability of return of their products increases.

Table 15: Logit Regression Results of Model 2 – Price & Supplier Effect

Dep. Variable:	is_returned	**No. Observations:**	4600827
Model:	Logit	**Df Residuals:**	4600820
Method:	MLE	**Df Model:**	6
Date:	Tue, 09 Apr 2024	**Pseudo R-squ.:**	0.07331
Time:	10:16:46	**Log-Likelihood:**	-1.7681e+06
converged:	True	**LL-Null:**	-1.9080e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-3.9514	0.012	-334.138	0.000	-3.975	-3.928
original_log_price	0.4281	0.003	167.044	0.000	0.423	0.433
promotion_percentage	-0.8383	0.008	-110.442	0.000	-0.853	-0.823
discount_percentage	-20.7184	0.163	-126.814	0.000	-21.039	-20.398
ship_cost_percentage	-2.9132	0.027	-108.075	0.000	-2.966	-2.860
supplier_return_rate	6.5708	0.040	163.488	0.000	6.492	6.650
supplier_return_rate_defect	71.8881	1.264	56.896	0.000	69.412	74.365

Model 3 contains product effects in addition to price and supplier effects. Additional variables are is_black, is_white to indicate colour of the product wherein, is_cloth_for_female and is_cloth_for_male to specify product gender. In the empirical setting, colours are classified into three groups as black, white and other colours, and before to run logistic regression other colours are chosen as the reference, therefore is_others excluded from Model 3. Similarly, product genders were grouped in three groups which are is_cloth_for_female, is_cloth_for_male, and is_cloth_for_unisex and unisex product category is chosen as reference, so it is also excluded from the model prior to logistic regression. Results seen in Table 16 shows that black and white products are returned with a lower probability compared to other colours. We can also interpret the odds ratios for black and white products, as when we exponentiate the coefficient of is_black (-0.0608) we get approximately 0.9408. Hence for every one-unit increase in the product being black the odds of return occurrence decrease by about 5.92%, as 1 - 0.9408 = 0.0592. The exponential of the coefficient of is_white (-0.0492) is approximately 0.9517 therefore, for every one-unit increase in the is_white variable (the product being white), the odds of return occurrence decrease by about 4.83% (1 - 0.9517 = 0.0483). We can say that black products are more likely to be kept by

Table 16: Logit Regression Results of Model 3
Price & Supplier & Product Effect

Dep. Variable:	is_returned	**No. Observations:**	4600827
Model:	Logit	**Df Residuals:**	4600816
Method:	MLE	**Df Model:**	10
Date:	Tue, 09 Apr 2024	**Pseudo R-squ.:**	0.07850
Time:	10:20:02	**Log-Likelihood:**	-1.7582e+06
converged:	True	**LL-Null:**	-1.9080e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-4.3241	0.018	-247.086	0.000	-4.358	-4.290
original_log_price	0.4339	0.003	167.664	0.000	0.429	0.439
promotion_percentage	-0.6995	0.008	-90.899	0.000	-0.715	-0.684
discount_percentage	-20.6658	0.164	-126.378	0.000	-20.986	-20.345
ship_cost_percentage	-3.0768	0.027	-113.349	0.000	-3.130	-3.024
supplier_return_rate	5.6717	0.041	138.816	0.000	5.592	5.752
supplier_return_rate_defect	64.2410	1.264	50.841	0.000	61.764	66.718
is_black	-0.0608	0.003	-18.451	0.000	-0.067	-0.054
is_white	-0.0492	0.004	-11.989	0.000	-0.057	-0.041
is_cloth_for_female	0.5612	0.013	43.416	0.000	0.536	0.587
is_cloth_for_male	0.0575	0.013	4.352	0.000	0.032	0.083

the users compared to other colours. The gender specificity of the product has a statistically significant impact on the likelihood of return occurrence. However, the magnitude of the effect differs between the genders, with the coefficient for female products (0.5612) being larger than that for male products (0.0575). When we take exponential for is_cloth_for_female it is about 1.7537 and 1.7537 - 1 = 0.7537 means that for every one-unit increase in the product being for females (compared to unisex products) the odds of return occurrence increase by about 75.37%. Exponentiating the coefficient of is_cloth_for_male yields approximately 1.0594, so for every one-unit increase in the product being specifically for males (compared to unisex products), the odds of return occurrence increase by about 5.94% as 1.0594 - 1 = 0.0594.

In Model 4, the customer effect is added to previous variables. Customer effect brings; is_female, membership_len_in_days, frequent_user, user_age, customer_return_percentage, is_elite_user, abused and is_wallet_trx to the model. As customer gender is categorized in two groups, is_female and is_male, the male gender is chosen as reference and excluded from the model. As seen in logistic regression results in Table 17, is_female is significant with a coefficient of 0.5071. In other words, return occurrence for female customers are higher compared to male customers, holding other variables constant. Another significant variable is abused which refers to orders containing the same product with different sizes. This variable is created with the aim of questioning the customers' intention to return when the order is occurred. The coefficient of abused is 0.4032, the odds ratio is 1.497 which results in 1.497 - 1 = 0.497. Hence, if a transaction is labeled as abused, it is approximately 49.7% more likely to be returned compared to non-abused transactions, all else being equal.

Table 17: Logit Regression Results of Model 4
Price & Supplier & Product & Customer Effect

Dep. Variable:	is_returned	**No. Observations:**	4600827
Model:	Logit	**Df Residuals:**	4600810
Method:	MLE	**Df Model:**	16
Date:	Sun, 14 Apr 2024	**Pseudo R-squ.:**	0.09218
Time:	12:37:32	**Log-Likelihood:**	-1.7321e+06
converged:	True	**LL-Null:**	-1.9080e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-5.2603	0.019	-272.220	0.000	-5.298	-5.222
original_log_price	0.4795	0.003	180.798	0.000	0.474	0.485
promotion_percentage	-0.7324	0.008	-94.109	0.000	-0.748	-0.717
discount_percentage	-21.1935	0.165	-128.409	0.000	-21.517	-20.870
ship_cost_percentage	-2.4682	0.028	-89.060	0.000	-2.522	-2.414
supplier_return_rate	5.8226	0.041	140.554	0.000	5.741	5.904
supplier_return_rate_defect	64.0319	1.278	50.088	0.000	61.526	66.538
is_black	-0.0605	0.003	-18.223	0.000	-0.067	-0.054
is_white	-0.0612	0.004	-14.807	0.000	-0.069	-0.053
is_cloth_for_female	0.5269	0.013	40.553	0.000	0.501	0.552
is_cloth_for_male	0.1612	0.013	12.134	0.000	0.135	0.187
is_female	0.5071	0.005	99.902	0.000	0.497	0.517
membership_len_in_days	0.0002	1.21e-06	163.048	0.000	0.000	0.000
user_age	-0.0075	0.000	-38.172	0.000	-0.008	-0.007
is_elite_user	-0.0496	0.003	-17.386	0.000	-0.055	-0.044
abused	0.4032	0.006	66.779	0.000	0.391	0.415
is_wallet_trx	0.0682	0.003	24.384	0.000	0.063	0.074

In the next model, Model 5, the time effect is investigated. A day-time is divided by four time intervals as is_time_00_06, is_time_06_12, is_time_12_18 and

Table 18: Logit Regression Results of Model 5
Price & Supplier & Product & Customer & Time Effect

Dep. Variable:	is_returned	**No. Observations:**	4600827
Model:	Logit	**Df Residuals:**	4600803
Method:	MLE	**Df Model:**	23
Date:	Sun, 14 Apr 2024	**Pseudo R-squ.:**	0.09352
Time:	12:41:57	**Log-Likelihood:**	-1.7295e+06
converged:	True	**LL-Null:**	-1.9080e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-5.0280	0.020	-254.605	0.000	-5.067	-4.989
original_log_price	0.4776	0.003	179.845	0.000	0.472	0.483
promotion_percentage	-0.7473	0.008	-95.559	0.000	-0.763	-0.732
discount_percentage	-21.4453	0.166	-129.340	0.000	-21.770	-21.120
ship_cost_percentage	-2.4374	0.028	-87.909	0.000	-2.492	-2.383
supplier_return_rate	5.7388	0.042	138.269	0.000	5.657	5.820
supplier_return_rate_defect	63.5888	1.279	49.719	0.000	61.082	66.096
is_black	-0.0627	0.003	-18.862	0.000	-0.069	-0.056
is_white	-0.0612	0.004	-14.813	0.000	-0.069	-0.053
is_cloth_for_female	0.5205	0.013	40.051	0.000	0.495	0.546
is_cloth_for_male	0.1554	0.013	11.695	0.000	0.129	0.181
is_female	0.5071	0.005	99.822	0.000	0.497	0.517
membership_len_in_days	0.0002	1.21e-06	161.941	0.000	0.000	0.000
user_age	-0.0073	0.000	-37.096	0.000	-0.008	-0.007
is_elite_user	-0.0573	0.003	-19.997	0.000	-0.063	-0.052
abused	0.4032	0.006	66.717	0.000	0.391	0.415
is_wallet_trx	0.0621	0.003	22.178	0.000	0.057	0.068
is_time_6_12	-0.0909	0.005	-17.536	0.000	-0.101	-0.081
is_time_12_18	-0.1280	0.005	-28.150	0.000	-0.137	-0.119
is_time_18_24	-0.0965	0.005	-21.329	0.000	-0.105	-0.088
is_bank_holiday	-0.0565	0.005	-12.266	0.000	-0.066	-0.047
is_weekend	-0.0035	0.003	-1.126	0.260	-0.010	0.003
is_May	-0.1997	0.003	-58.546	0.000	-0.206	-0.193
is_July	-0.1037	0.004	-29.256	0.000	-0.111	-0.097

is_time_18_24, before to run logistic regression, is_time_00_06 is selected as reference. Model 5 results presented in Table 18 indicate that transactions have been made from 00.00 to 06.00 are more likely to be returned than transactions occurred at other times of the day. Also, if the transaction has been made during a bank holiday, it has a lower likelihood of return. Another result is that is_weekend did not show significance as a variable in predicting returns.

Finally, Model 6 incorporates the effect of interaction terms on return. Some significant resulted interactions are discount_percentage x abused with -11.7049, discount_percentage x is_female with -2.1091, and supplier_return_rate x abused with 6.2278. When we analyze discount_percentage x is_female, is_female has a positive coefficient in previous models and also in Model 6, this implies that females have a higher probability of return compared to males. However as discount percentage increases, return probability decreases for female customers. In other words, female customers might prefer to keep the products as the discount percentage increases. The interactions found to be insignificant are, is_time_6_12 x is_elite_user, is_bank_holiday x is_female, is_time_6_12 x abused and is_weekend x abused, although most of them are significant as single predictors. The results of Model 6 are presented in Table 20 and Table 21.

Table 19 presents the performance metrics of each model after logistic regression has been applied.

Table 19: Test Data Performance Across The Logistic Regression Models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AIC Score	3593692	3536161	3516386	3464194	3459085	3444117
AUC Score	0.6234	0.6437	0.6461	0.6566	0.6573	0.6593
Accuracy	0.6386	0.6886	0.7126	0.7184	0.7188	0.7193
Recall	0.6018	0.5803	0.5521	0.5694	0.5704	0.5747
Precision	0.2243	0.2526	0.2660	0.2749	0.2756	0.2769
F-1 Score	0.3268	0.3520	0.3590	0.3708	0.3716	0.3737

There are no substantial differences among the models in terms of other performance metrics as seen in Table 19 yet, Model 6 is the best performing model in terms of all measures. After Model 6 was determined as the baseline main effects model LASSO regression was applied in order to enhance the predictive performance and interpretability.

Table 20: Logit Regression Results of Model 6
Price & Supplier & Product & Customer & Time Effect & Interaction Terms

Dep. Variable:	is_returned	**No. Observations:**	4600827
Model:	Logit	**Df Residuals:**	4600768
Method:	MLE	**Df Model:**	58
Date:	Sun, 14 Apr 2024	**Pseudo R-squ.:**	0.09746
Time:	12:46:03	**Log-Likelihood:**	-1.7220e+06
converged:	True	**LL-Null:**	-1.9080e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

Table 21: Table 20 (cont'd)

	coef	std err	z	P> |z|	[0.025	0.975]
const	-4.8731	0.043	-112.744	0.000	-4.958	-4.788
original_log_price	0.4754	0.003	177.762	0.000	0.470	0.481
promotion_percentage	-0.8240	0.022	-37.989	0.000	-0.866	-0.781
discount_percentage	-19.0746	0.562	-33.931	0.000	-20.176	-17.973
ship_cost_percentage	-2.4274	0.028	-86.603	0.000	-2.482	-2.373
supplier_return_rate	5.8839	0.148	39.639	0.000	5.593	6.175
supplier_return_rate_defect	61.3040	1.282	47.823	0.000	58.792	63.817
is_black	-0.0147	0.011	-1.328	0.184	-0.036	0.007
is_white	-0.0308	0.013	-2.285	0.022	-0.057	-0.004
is_cloth_for_female	-0.0421	0.032	-1.329	0.184	-0.104	0.020
is_cloth_for_male	0.1003	0.031	3.190	0.001	0.039	0.162
is_female	0.0742	0.041	1.790	0.073	-0.007	0.155
membership_len_in_days	0.0002	1.21e-06	153.045	0.000	0.000	0.000
user_age	0.0076	0.001	10.654	0.000	0.006	0.009
is_elite_user	-0.1511	0.019	-8.120	0.000	-0.188	-0.115
abused	0.0223	0.043	0.525	0.600	-0.061	0.106
is_wallet_trx	0.1620	0.004	43.441	0.000	0.155	0.169
is_time_6_12	-0.3260	0.018	-18.626	0.000	-0.360	-0.292
is_time_12_18	-0.2655	0.015	-17.856	0.000	-0.295	-0.236
is_time_18_24	-0.1639	0.015	-11.212	0.000	-0.193	-0.135
is_bank_holiday	-0.0284	0.015	-1.842	0.066	-0.059	0.002
is_weekend	0.0957	0.010	9.255	0.000	0.075	0.116
is_May	-0.2564	0.009	-28.159	0.000	-0.274	-0.239
is_July	-0.0373	0.010	-3.764	0.000	-0.057	-0.018
is_time_6_12 x is_elite_user	-0.0087	0.011	-0.827	0.408	-0.029	0.012
is_time_12_18 x is_elite_user	-0.0208	0.009	-2.233	0.026	-0.039	-0.003
is_time_18_24 x is_elite_user	-0.0288	0.009	-3.102	0.002	-0.047	-0.011
is_bank_holiday x is_elite_user	-0.0208	0.009	-2.300	0.021	-0.038	-0.003
is_weekend x is_elite_user	-0.0130	0.006	-2.078	0.038	-0.025	-0.001
is_wallet_trx x is_elite_user	-0.2278	0.006	-40.340	0.000	-0.239	-0.217
is_female x is_elite_user	0.6008	0.011	53.134	0.000	0.579	0.623
user_age x is_elite_user	-0.0101	0.000	-26.263	0.000	-0.011	-0.009
is_time_6_12 x is_female	0.2653	0.018	14.574	0.000	0.230	0.301
is_time_12_18 x is_female	0.1689	0.016	10.884	0.000	0.139	0.199
is_time_18_24 x is_female	0.0965	0.015	6.315	0.000	0.067	0.126
is_bank_holiday x is_female	-0.0080	0.016	-0.511	0.609	-0.039	0.023
is_weekend x is_female	-0.0982	0.011	-9.192	0.000	-0.119	-0.077
is_white x is_female	-0.0333	0.014	-2.352	0.019	-0.061	-0.006
is_black x is_female	-0.0576	0.012	-4.975	0.000	-0.080	-0.035
is_cloth_for_female x is_female	0.6187	0.035	17.891	0.000	0.551	0.687
is_cloth_for_male x is_female	-0.0216	0.035	-0.625	0.532	-0.090	0.046
discount_percentage x is_female	-2.1091	0.587	-3.591	0.000	-3.260	-0.958
promotion_percentage x is_female	0.1201	0.022	5.368	0.000	0.076	0.164
abused x is_female	0.4597	0.020	23.570	0.000	0.421	0.498
user_age x is_female	-0.0081	0.001	-14.041	0.000	-0.009	-0.007
ship_cost_percentage x abused	-2.3176	0.144	-16.132	0.000	-2.599	-2.036
promotion_percentage x abused	-0.1866	0.029	-6.377	0.000	-0.244	-0.129
discount_percentage x abused	-11.7049	1.292	-9.060	0.000	-14.237	-9.173
user_age x abused	-0.0109	0.001	-13.469	0.000	-0.012	-0.009
supplier_return_rate x abused	6.2278	0.154	40.557	0.000	5.927	6.529
is_time_6_12 x abused	-0.0213	0.023	-0.919	0.358	-0.067	0.024
is_time_12_18 x abused	-0.0686	0.021	-3.329	0.001	-0.109	-0.028
is_time_18_24 x abused	-0.0764	0.021	-3.692	0.000	-0.117	-0.036
is_bank_holiday x abused	-0.0935	0.020	-4.669	0.000	-0.133	-0.054
is_weekend x abused	-0.0061	0.014	-0.442	0.658	-0.033	0.021
user_age x supplier_return_rate	-0.0164	0.004	-3.923	0.000	-0.025	-0.008
supplier_return_rate x May	0.5641	0.077	7.330	0.000	0.413	0.715
supplier_return_rate x July	-0.5994	0.082	-7.304	0.000	-0.760	-0.439
is_weekend x is_bank_holiday	-0.0339	0.011	-3.175	0.001	-0.055	-0.013

Table 22: Feature Selection of LASSO

Non-Zero Indices	Zero Indices
original_log_price, promotion_percentage, discount_percentage, ship_cost_percentage, is_black, is_white, is_cloth_for_female, is_cloth_for_male, is_female, membership_len_in_days, user_age, abused, is_wallet_trx, is_time_6_12, is_time_12_18, is_time_18_24, is_weekend, is_May, is_July, is_time_18_24 x is_elite_user, is_wallet_trx x is_elite_user, user_age x is_elite_user, is_time_6_12 x is_female, is_time_18_24 x is_female, is_weekend x is_female, is_white x is_female, is_black x is_female, is_cloth_for_female x is_female, is_cloth_for_male x is_female, discount_percentage x is_female, promotion_percentage x is_female, abused x is_female, user_age x is_female, promotion_percentage x abused, user_age x abused, supplier_return_rate x abused, user_age x supplier_return_rate	supplier_return_rate, supplier_return_rate_defect, is_elite_user, is_bank_holiday, is_time_6_12 x is_elite_user, is_time_12_18 x is_elite_user, is_bank_holiday x is_elite_user, is_weekend x is_elite_user, is_time_12_18 x is_female, ship_cost_percentage x abused, discount_percentage x abused, is_time_6_12 x abused, is_time_12_18 x abused, is_time_18_24 x abused, is_bank_holiday x abused, is_weekend x abused, supplier_return_rate x May, supplier_return_rate x July, is_weekend x is_bank_holiday

4.5 Results of LASSO, XGBoost and Random Forest

Feature selection of LASSO shows that a considerable number of the independent variables included in Model 6 were eliminated. Table 22 represents which features resulted in non-zero indices and which of them have zero indices. Price, product, customer, and time related variables are found relevant according to LASSO.

As the final step of model selection in addition to logistic regression and LASSO; XGBoost and Random Forest have been applied for Model 6. Comparison of results in given in Table 23.

Table 23: Test Data Performance Across The Models

	Model 4	Model 6	LASSO	XGBoost	Random Forest
AIC Score	3464194	3444117	-9905220	N/A	N/A
AUC Score	0.66	0.66	0.65	0.52	0.58
Accuracy	0.72	0.72	0.71	0.86	0.86
Recall	0.57	0.57	0.56	0.04	0.17
Precision	0.27	0.28	0.27	0.62	0.57
F-1 Score	0.37	0.37	0.36	0.07	0.27

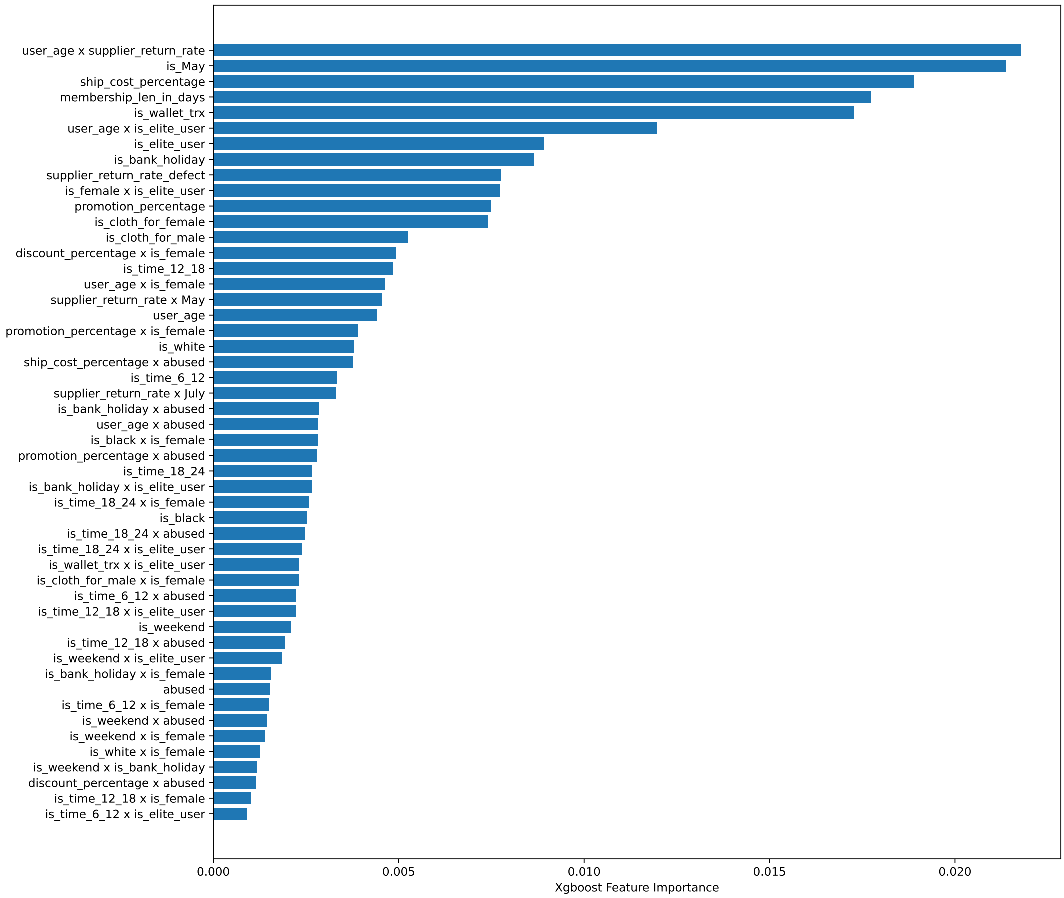

Figure 7: XGBoost Feature Importance

The most important fifty variables selected by the XGBoost algorithm are shown in Figure 7. Feature importance allows us to understand which features are most influential in making return predictions. According to the XGBoost Model, the interaction of user_age and supplier_return_rate is the most relevant feature in return prediction. This is followed by is_May, and other important variables are ship_cost_percentage, membership_len_in_days, is_wallet_trx, user_age x is_elite_user. Customer-related variables and some time relevant variables are selected as the most important predictors of return occurrence according to XGBoost.

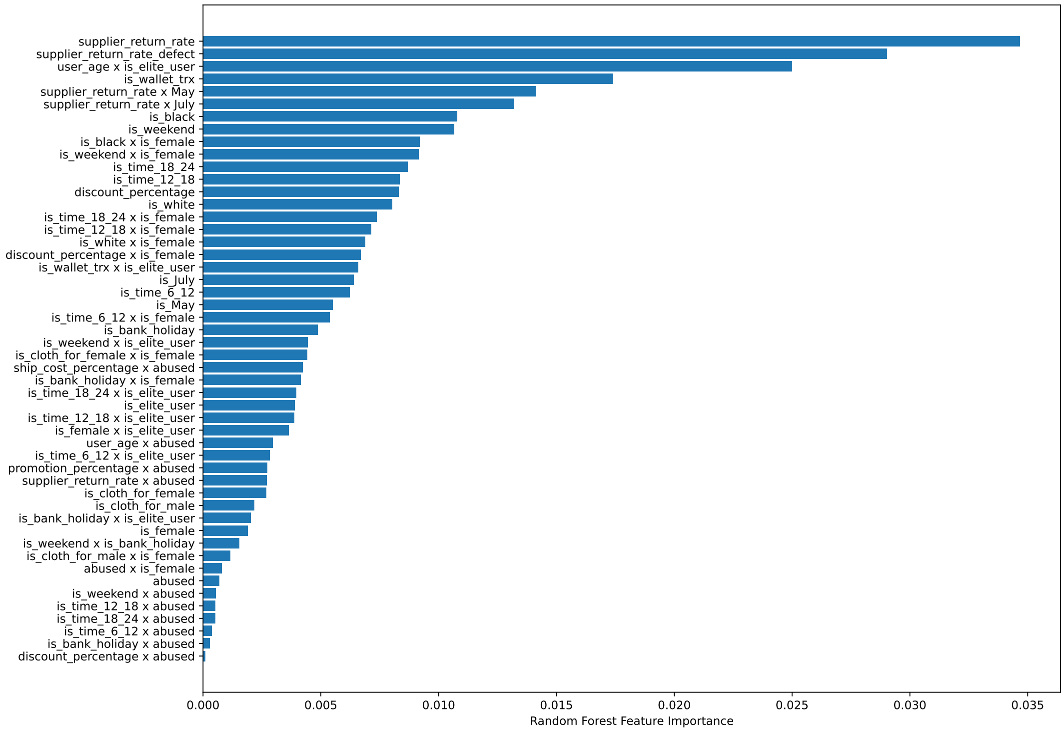

Figure 8: Random Forest Feature Importance

The importance gradation between predictive variables differs in the Random Forest Model. The most important fifty variables are given in Figure 8, the most important variable in return prediction is supplier_return_rate, it is followed by supplier_return_rate_defect, user_age x is_elite_user and is_wallet_trx respectively. Unlike XGBoost, Random Forest gives more importance to supplier-related predictors. Some other remarkable variables for Random Forest are user_age x is_elite_user, is_black, is_weekend, is_black x is_female and is_weekend x is_female.

CHAPTER 5

CONCLUSION

The study aimed to develop an evidence-based prediction model using data provided by one of the leading e-commerce companies in Turkey. Our primary focus was on apparel transactions and their associated returns. We were provided with several data files containing a wide range of information about products, customers, prices, transactions, and suppliers. After data cleaning and feature engineering, the data was prepared for logistic regression analysis. Six models were developed incrementally. Specifically, the first model investigated the effect of price related variables on returns, while the second model examined both price and supplier effects. Model 3 incorporated product related variables in addition to price and supplier variables. In Model 4, customer effects were included as predictor variables, and Model 5 incorporated time effects alongside previously mentioned variables. The final model included two-way interaction terms in addition to other variables. Logistic regression results indicated no significant improvement from Model 4 to Model 6. Due to the excessive number of predictor variables in Model 6, LASSO regularization was applied to obtain a simplified model with the most relevant variables. LASSO effectively eliminated nearly half of the variables which is seen in Table 22. As demonstrated in Table 23, there was minimal difference in performance metrics between Model 6(logistic regression) and the LASSO model. XGBoost and Random Forest models exhibited higher accuracy and precision, yet lower recall scores compared to the LASSO and logistic regression. Logistic regression resulted with the highest AUC score. This preliminary study aims to demonstrate the importance of return issue for e-commerce companies, particularly in apparel segment, while investigating which factors could be effective in return occurrence. It is evident that price-related factors, product features and customer demographics as well as customer behavior (such as buying same product with different sizes or buying for opposite sex) affect returns. Our findings indicate that female customers have a higher probability of returning items compared to male customers. Transactions made between 00:00 and 06:00 are more likely to be returned than transactions occurring at other times of the day. The data we used in the study can be extracted from the databases of any corporate company, and similar models can be developed by the companies themselves to predict returns. Besides industrial application, we hope to inspire future research on return prediction in e-commerce due to the financial and ecological effects of a huge number of returned goods. Considering the human resources allocated for return operations and the disposal processes of non-resalable products, it is important to predict or prevent returns if possible. The improvement of return prediction is expected to result in better operational functionality for the companies and environmental sustainability.

5.1 Societal Impact

The return of online purchases has a significant environmental impact due to packaging and transportation. Economically, the high rate of returns imposes substantial costs on businesses. Addressing the societal impact of online returns requires a collective effort from businesses, consumers, and policymakers to develop more sustainable practices and reduce the negative impacts associated with online returns. By improving return prediction models and promoting responsible consumer behavior, it is possible to mitigate these adverse effects, leading to better environmental sustainability, economic efficiency, and social responsibility.

BIBLIOGRAPHY

Ader, J., Adhi, P., Chai, J., Singer, M., Touse, S., & Yankelevich, H. 2021. Returning to order: Improving returns management for apparel companies. Retrieved from https://www.mckinsey.com/industries/retail/our-insights/returning-to-order-improving-returns-management-for-apparel-companies

Agresti, A. 2012. *Categorical data analysis.* John Wiley & Sons.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. doi: 10.1109/TAC.1974.1100705

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. 1996. Children in motion: School transfers and elementary school performance. *The Journal of Educational Research*.

Ambilkar, P., Dohale, V., Gunasekaran, A., & Bilolikar, V. 2022. Product returns management: a comprehensive review and future research agenda. *International Journal of Production Research*, *60*(12), 3920–3944.

Analyzing the returns landscape. 2023. Retrieved from https://www.uspsdelivers.com/analyzing-the-returns-landscape/

Bernstein, F., Modaresi, S., & Sauré, D. 2019. A dynamic clustering approach to data-driven assortment personalization. *Management Scince*.

Breiman, L. 2001. Random forests. *Machine Learning*. doi: 10.1023/A:1010933404324

Brownlee, J. 2019. Classification accuracy is not enough: More performance measures you can use. Retrieved from https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

Caignya, A. D., Coussementa, K., & Bock, K. W. D. 2018. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*.

Chang, X., Huang, Y., Li, M., Bo, X., & Kumar, S. 2021. Efficient detection of environmental violators: A big data approach. *Production and Operatitions*

Management, *30*(5), 246–1270. doi: 10.1111/poms.13272

Chen, T., & Guestrin, C. 2016. *Xgboost: A scalable tree boosting system.* Association for Computing Machinery.

Chou, P., Chuang, H. H. C., Chou, Y. C., & Liang, T. P. 2022. Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. *European Journal of Operational Research*, *296*(2), 635–651.

Choudhary, J. 2023. Mastering xgboost: A technical guide for intermediate machine learning practitioners. Retrieved from https://medium.com/@jyotsna.a.choudhary/mastering-xgboost-a-technical-guide-for-intermediate-machine-learning-practitioners-f7ad167c6865

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

Davison, T. 2024. The environmental impact of returning online products. Retrieved from https://blog.cleanhub.com/ecommerce-returns-environmental-impact

Dekking, Kraaikamp, Lopuhaa, & Meester. 2005. *A modern introduction to probability and statistics.* Springer.

Diamond, K. E., Reagan, A. J., & Bandyk, J. E. 2000. Parents' conceptions of kindergarten readiness: Relationships with race, ethnicity, and development. *The Journal of Educational Research*.

Duong, Q. H., Zhou, L., Meng, M., Nguyen, T. V., Ieromonachou, P., & Nguyen, D. T. 2022. Understanding product returns: A systematic literature review using machine learning and bibliometric analysis. *International Journal of Production Economics*, *243*, 1–19.

Er, S. 2024. Understand random forest algorithms with examples. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Falkenberg, S. F., & Spinler, S. 2022. The role of novel data in maintenance planning: Breakdown predictions for material handling equipment. *Computers & Industrial Engineering*, *169*, 1–12.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. 2016. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*.

Fokina, M. 2024. Online shopping statistics: Ecommerce trends for 2024. Retrieved from https://www.tidio.com/blog/online-shopping-statistics/

Freedman, L. 2023. Apparel and accessories shoppers accelerate online buying. Retrieved from https://www.digitalcommerce360.com/2023/07/24/apparel-accessories-shoppers-accelerate-online-buying/

Friedman, J., Hastie, T., & Tibshirani, R. 2009. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.*.

George, S. 2024, February. *The hidden cost: Understanding the environmental impact of online purchase returns* (Tech. Rep. No. 2). Partners Universal Innovative Research Publication.

Gross, D., & Juttijudata, M. 1997. Sensitivity of output performance measures to input distributions in queueing simulation modeling. In D. H. Withers, B. L. Nelson, S. Andradottir, & K. J. Healy (Eds.), *Proceedings of the 1997 winter simulation conference* (p. 296-302). Institute of Electrical and Electronics Engineers, Piscataway, NJ.

Hastie, T., Tibshirani, R., & Wainwright, M. 2015. *Statistical learning with sparsity the lasso and generalizations.* CRC Press.

Hilden, J. 1991. The area under the roc curve and its competitors. *Medical Decision Making*. doi: 10.1177/0272989X9101100204

Hosmer, D., Lemeshow, S., & Sturdivant, R. 2013. *Applied logistic regression.* Addison-Wesley.

Huang, J.-C., Tsai, Y.-C., Wu, P.-Y., Lien, Y.-H., Chien, C.-Y., Kuo, C.-F., ... Kuo, C.-H. 2020. Predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector regression, xgboost, lasso regression and ensemble method. *Computer Methods and Programs in Biomedicine*. doi: 10.1016/j.cmpb.2020.105536

Huang, P. W., & Lee, C. H. 2009. Automatic classification for pathological prostate images based on fractal analysis. *IEEE Transactions on Medical Imaging*, *28*(7), 1037 - 1050.

Huang, T., Bergman, D., & Gopal, R. 2019. Predictive and prescriptive analytics for location selection of add-on retail products. *Production and Operatitions Management*, *28*(7), 1858–1877. doi: 10.1111/poms.13018

Impact of covid pandemic on ecommerce. 2024. Retrieved from https://www.trade.gov/impact-covid-pandemic-ecommerce

Inman, D. 2022. 2022 retail returns rate remains flat at $816 billion. Retrieved from https://nrf.com/media-center/press-releases/2022-retail-returns-rate-remains-flat-816-billion

Inman, D. 2024. Nrf forecasts retail sales to reach at least $5.23 trillion in 2024. Retrieved from https://nrf.com/media-center/press-releases/nrf-forecasts-retail-sales-reach-least-523-trillion-2024

Interpreting logistic roc curves. n.d.. Retrieved from https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_logistic_roc_curves.htm

Ivanova, I. 2020. Those holiday returns come with a massive environmental footprint. Retrieved from https://www.cbsnews.com/news/holiday-returns-evironmental-costs-shipping/

Jagabathula, S., Subramanian, L., & Venkataraman, A. 2018. A model-based embedding technique for segmenting customers. *Operations Research*.

Kadiyala, B., Özalp Özer, & Şimşek, A. S. 2021. Data-driven approaches to targeting promotion e-mails: The case of delayed incentives. *Production and Operatitions Management*, *30*(3), 766–782. doi: 10.1111/poms.13316

Katircioglu, K. 1996. *Essays in inventory control* (PhD Thesis). University of British Columbia, Vancouver, BC, Canada.

Kim, Y., Kang, J., & Chun, H. 2022. Is online shopping packaging waste a threat to the environment? *Economics Letters*.

Knuth, D. E. 1973. *Fundamental algorithms*. Addison-Wesley.

Lindner, J. 2023. Must-know ecommerce return statistics. Retrieved from https://gitnux.org/ecommerce-return-statistics/#:~:text=41%25%20of%20online%20shoppers%20buy,%25%2C%20depending%20on%20the%20industry.

Logunova, I. 2023. A guide to f1 score. Retrieved from https://serokell.io/blog/a-guide-to-f1-score

Ma, S., Fildes, R., & Huang, T. 2016. Demand forecasting with high dimensional data: The case of sku retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*. doi: 10.1016/j.ejor.2015.08.029

Machamer, A. M., & Gruber, E. 1998. Secondary school, family, and educational risk: Comparing american indian adolescents and their peers. *The Journal of Educational Research*.

McNeal, R. B. 1998. High school extracurricular activities: Closed structures and stratifying patterns of participation. *The Journal of Educational Research*.

Meisels, S. J., & Liaw, F.-R. 1993. Failure in grade: Do retained students catch up? *The Journal of Educational Research*.

Moe, W. W., & Fader, P. S. 2004. Dynamic conversion behavior at e-commerce sites. *Management Scince*.

Nagtegaal, I., & van Krieken, J. 2002. The role of pathologists in the quality control of diagnosis and treatment of rectal cancer. *European Journal of Can-*

cer, *38*(7), 964 - 972. Retrieved from http://www.sciencedirect.com/science/article/pii/S0959804902000564 doi: 10.1016/S0959-8049(02)00056-4

Narvar. 2020. 2022 state of returns. Retrieved from https://corp.narvar.com/resources/2022-the-state-of-returns

NRF, & Retail, A. 2022, October. *2022 consumer returns in the retail industry* (Tech. Rep.). National Retail Federation.

Optoro. 2022. 2022 state of returns. Retrieved from https://info.optoro.com/2022stateofreturns

Pearce, J., & Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. 2010. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*.

Powers, D. M. W. 2017, December. *Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation* (Tech. Rep. No. Technical Report SIE-07-001). School of Informatics and Engineering Flinders University of South Australia.

Pramoditha, R. 2020. Random forests — an ensemble of decision trees. Retrieved from https://towardsdatascience.com/random-forests-an-ensemble-of-decision-trees-37a003084c6c#:~:text=In%20a%20random%20forest%2C%20each,with%20replacement%20is%20called%20bootstrapping

Priya, C. B. 2023. Hyperparameter tuning: Gridsearchcv and randomizedsearchcv, explained. Retrieved from https://www.kdnuggets.com/hyperparameter-tuning-gridsearchcv-and-randomizedsearchcv-explained

Putatunda, S., & Rama, K. 2018. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of xgboost. *SPML '18: Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*. doi: 10.1145/3297067.3297080

Raschka, S., & Mirjalili, V. 2019. *Python machine learning: Machine learning and deep learning with python, scikit-learn and tensorflow.* Packt Publishing.

The right fit: How ai is changing ecommerce apparel returns. 2024. Retrieved from https://www.retaildive.com/library/rakuten-whitepaper-the-right-fit/

Rodrigues, A. M., & Alves, O. 2018. *Improving supply chain efficiency in e-tail by redirecting returns: a machine learning approach* (Master Thesis). Universidade Do Porto.

Rogers, D. S., & Tibben-Lembke, R. S. 1999. *Going backwards:reverse logistics trends and practices.* (Reverse Logistics Executive Council)

Rush, S., & Vitale, P. A. 1994. Analysis for determining factors that place elementary students at risk. *The Journal of Educational Research*.

Saarela, U., Leiviska, K., & Juuso, E. 2003, June. *Modelling of a fed-batch fermentation process* (Tech. Rep. No. Report A No. 21). University of Oulu, Department of Process and Environmental Engineering.

Saeed, M. 2021. A gentle introduction to sigmoid function. Retrieved from https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function/

Safhi, A. M., Dabiri, H., Soliman, A., & Khayat, K. H. 2023. Prediction of self-consolidating concrete properties using xgboost machine learning algorithm: Part 1–workability. *Construction and Building Materials*.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., . . . Lin, C.-T. 2017. A review of clustering techniques and developments. *Neurocomputing*. doi: https://doi.org/10.1016/j.neucom.2017.06.053

Sethi, A. 2019. One hot encoding vs. label encoding using scikit-learn. Retrieved from https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/

Sinaga, K. P., & Yang, M.-S. 2020. Unsupervised k-means clustering algorithm. *IEEE Access*. doi: 10.1109/ACCESS.2020.2988796

Smith, J. B. 1997. Effects of eighth-grade transition programs on high school retention and experiences. *The Journal of Educational Research*.

Somanchi, S., Adjerid, I., & Gross, R. 2022. To predict or not to predict: The case of the emergency department. *Production and Operatitions Management*, *31*(2), 799-818. doi: 10.1111/poms.13580

Stambor, Z. 2022. The value of us online returns will reach 279.03*billionthisyear*, *morethandoublin*

Stoica, P., & Selen, Y. 2004. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*. doi: 10.1109/MSP.2004.1311138

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. doi: 10.1111/j.2517-6161.1996.tb02080.x

Trusty, J. 2000. High educational expectations and low achievement: Stability of educational goals across adolescence. *The Journal of Educational Research*.

Unal, M., & Park, Y.-H. 2023. Fewer clicks, more purchases. *Management Science*. doi: https://doi.org/10.1287/mnsc.2023.4716

Wu, J., Zhang, Z., & Zhou, S. X. 2022. Credit rating prediction through supply chains: A machine learning approach. *Production and Operatitions Management*, *31*(4), 1613–1629. doi: 10.1111/poms.13634

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. 2009. Genome-wide association analysis by lasso penalized logistic regression. *BIOINFORMATICS*. doi: 10.1093/bioinformatics/btp041