DEFINING THE DECISIVE FACTORS INFLUENCING PURCHASE DECISION BY USING FEATURE IMPORTANCE METHODS IN E-COMMERCE AND COMPARING THE METHODS' PERFORMANCES

(E-TİCARETTE SATIN ALMA KARARLARINI ETKİLEYEN FAKTÖRLERİN ÖZELLİK ÖNEMİ METODLARI İLE TESPİTİ VE METODLARIN KIYASLANMASI)

by

Erman Demir, B.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

SMART SYSTEMS ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

Supervisor: Assist. Prof. Ahmet Teoman NASKALİ

Feb 2024

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Öğr. Üyesi Ahmet Teoman NASKALİ, my thesis advisor, for his invaluable guidance, unwavering support, and mentorship throughout the entire research process. His expertise and commitment to academic excellence have been instrumental in shaping this thesis.

I am also grateful to Dr. Öğr. Üyesi Serhan Daniş for his insightful feedback and constructive suggestions have significantly enriched the quality of this work. I am thankful for his willingness to share his knowledge and expertise.

I extend my appreciation to the entire team at Hepsiburada ARGE for their support and collaboration. Their resources and assistance were pivotal in conducting this research, and I am thankful for the opportunities they provided.

I would like to acknowledge the support of my family and friends who stood by me during the challenging moments of this academic journey. Your encouragement and belief in my abilities have been a constant source of motivation.

Lastly, I wish to express my gratitude to all those who, directly or indirectly, contributed to the completion of this thesis. Your support has been invaluable, and I am sincerely thankful for your involvement.

.

Feb 2024

Erman Demir

TABLE OF CONTENTS

LIST OF SYMBOLS

SQL : Structured Query Language
E-commerce : Electronically buying or selling products on online services
E-platform : Electronic trading platform
CRM : Customer Relation Management
ML : Machine Learning
ADASYN : Adaptive Synthetic Sampling
MLP : Multi Layer Perceptron
LSTM : Long Short-Term Memory
SMOTE : Synthetic Minority Oversampling Technique
SME : Small and medium-sized enterprises
OLS : Ordinaly Least Square
QCA : Qualitative Comparative Analysis
RNN : Recurrent Neural Network
AUC : Area Under the Curve
SmartPLSÇ : Partial least square-based software
Ad: Advertisement
Id: Session Identification Number
CART: Classification and Regression Tree
PCA: Principal Component Analysis
ASM: Attribute Selection Measures

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Online retail enterprises engage in two primary activities to foster revenue generation and thrive in the market. The initial focus involves augmenting user traffic to the online shopping platform, followed by the subsequent task of converting this traffic into tangible revenue. Marketing endeavors are dedicated to attracting customers to online shopping platforms, often incurring substantial costs. Given the financial implications of obtaining traffic, it becomes imperative to not only draw users but also incentivize them to make purchases on the platform. Consequently, comprehending the pivotal factors influencing customer purchase decisions is crucial for online shopping platforms.

This study aims to scrutinize the factors shaping consumer purchase decisions by utilizing an illustrative case from an e-commerce platform. The objective is to ascertain the most influential factor among various possibilities, such as traffic source types (Google, campaign, direct, etc.), customer persona or segment, types of pages or components visible on the platform, product placement on the page, customer participation in campaigns or discounts, product review scores and counts, and whether the product was recommended by the platform's recommendation system.

Throughout the data collection phase, performance issues will be addressed through SQL optimization and other techniques. Additionally, data quality concerns will be rectified to ensure uniform and reliable results. Subsequently, statistical methods, supervised machine learning, and deep learning approaches will be employed on the data to derive feature importance values. These values will illuminate which factors play the most decisive role in customer purchase decisions. Finally, the study will compare and evaluate the performances of seven methods, presenting a comprehensive examination unprecedented in the existing literature in terms of methodological diversity.

ÖZET

E-ticaret şirketleri gelir elde etmek ve pazarda varlıklarını sürdürmek için iki faaliyete odaklanmaktadır. İlk faaliyet, online alışveriş platformunun kullanıcı trafiğini artırmak olup, ikinci faaliyet ise bu trafiği gelire dönüştürmektir. Trafiği elde etmenin maliyetleri nedeniyle e-ticaret platformuna gelen kullanıcılara sipariş verdirmek son derece önemlidir. Bu nedenle, e-ticaret platformları kullanıcıların sipariş verme oranlarını arttırmak için, müşterilerin satın alma kararını etkileyen faktörlerin hangilerinin daha önemli olduğunu anlamalıdırlar. Bu çalışmada, bir e-ticaret platformu verileri üzerinden tüketicilerin satın alma kararlarını etkileyebilecek faktörler incelenerek hangilerinin satın alma kararı üzerinde en belirleyici olduğu tespit edilmeye çalışılacaktır. E-ticaret terimlerini ve çalışmaya konu olan e-ticaret platformunun tanımlarını içeren bu faktörlere örnek olarak şunlar verilebilir: e-ticaret platformuna gelen kullanıcı trafiği tipi (Google ile, bir kampanya aracılığı ile veya doğrudan), müşteri segmenti, e-platformda hangi ekranları gördüğü, ürünün ekrandaki konumu, müşterinin kampanya veya indirimden yararlanıp yararlanmadığı, ürün yorum adetleri, kullanıcı puanlamaları, ürünün öneri sistemi tarafından önerip/önerilmediği. Bu çalışmada veri toplama aşamasında karşılaşılan performans sorunları SQL optimizasyonu ve diğer yöntemlerle çözülecektir. Veri kalite sorunları ise tutarlı sonuçlar elde etmek için çeşitli yöntemler ile düzeltilecektir. Ardından özellik önemi değerlerini tespiti için istatistiksel yöntemler, denetimli makine öğrenimi ve derin öğrenme yöntemleri kullanılacaktır. Uygulanan yöntemler ile elde edilen özellik önemi değerlerine göre kullanıcını satın alma kararında en çok belirleyici faktörler belirlenecektir. Kullanılan yöntemlerin sonuçları kıyaslanacaktır. Ayrıca kullanılan yöntemlerin performansları da kıyaslanarak hangi yöntemin verimli çalıştığı, yöntemlerin başarılı olduğu ve olmadığı noktalar tespit edilecektir. Bu çalışma, özellik önemi değeri tespiti konusunda literatürde en çok sayıda yöntem kullanan ve kıyaslama yapan çalışmadır.

1. INTRODUCTION

Following the global outbreak of the pandemic in 2020, physical stores were shuttered across most countries. Even consumers who were previously hesitant about online shopping, due to various reasons, found themselves turning to digital alternatives. This shift to online shopping resulted in increased profits for the sector. With growing consumer interest in online platforms, the industry invested significantly in enhancing functionality and convenience. Companies began offering quality guarantees, free returns, the ability to choose precise delivery and return times, and more. Additionally, new entrants intensified competition in the sector.

Responding to heightened competition, online retailers increased their investments in technology, particularly in expanding their technology teams well before the pandemic. The importance of data in decision-making became paramount, with every customer interaction on online platforms being analyzed to glean valuable consumer insights.

Technologies for tracking, storing, and analyzing customer data assumed a critical role. Consequently, companies made substantial investments in understanding user behaviors through data analysis, enabling them to provide personalized and tailored experiences.

While many customers initially access retailers' apps and web platforms, it's not enough for a business that customers merely land on their platforms; actual purchases are essential for revenue generation and the survival and growth of the business.

During the pandemic, companies allocated significant budgets to marketing due to the surge in customer traffic, resulting in substantial revenue generation. However, as the pandemic waned, customer traffic decreased, and regulations were implemented to restrict marketing activities. Consequently, the focus shifted to generating revenue from the existing customer traffic on the platforms, necessitating the creation of more user-friendly and effective platforms.

Once customers reach the platform, they must navigate through the conversion funnel, encompassing landing, discovery, checkout, and the ultimate purchase, to complete their transaction.

- Landing: Customers arrive on online platforms through diverse channels, including the Google search engine, advertisements on various platforms, emails, social media platforms, push notifications, and links shared by friends or influencers. Additionally, customers can directly access the platforms by opening them on their devices.
- Discovery: During this phase, customers explore products, seeking to find the items they require. Following their landing on the platform, customers peruse the homepage, navigate through category and brand pages, or conduct searches for specific product names. Once they locate the desired products, they proceed to add them to their shopping basket.
- Checkout: During this phase, customers review the items in their basket, verify purchase details, confirm the transaction, select payment options, and provide delivery address information.
- Purchase: This marks the final step of the conversion funnel. The customer confirms the payment and is directed to the purchase success page.

Customers may encounter challenges at any point in the conversion funnel, leading them to abandon the journey toward completing a purchase. Therefore, it is essential for the conversion funnel steps to be seamless, straightforward, practical, and user-friendly. Customers should be able to easily navigate the platform, browse products, and successfully complete the purchase process. The success metric for this process is the conversion rate, representing the proportion of customers on the platform who successfully make a purchase. Given the high cost associated with attracting customers to the platform, it becomes crucial to convert customer visits into actual orders.

Identifying obstacles or incentives that influence a purchase is vital. In this thesis, the factors predominantly affecting purchase decisions will be delineated.

Which factors can play a decisive role in a customer's purchase decision? The factors outlined below, encompassing terms within the e-commerce domain and specific definitions related to an example e-commerce platform, are generally influential in customer purchase decisions based on industry knowledge in the e-commerce domain.

- Traffic Source Type: This refers to the channel through which customers access the platform, such as the Google search engine, price comparison websites, push notification campaigns, influencer links, etc.

- Customer Persona: This entails representing customers' shopping habits, with categories including "Moms - Women," "Digital Parents," "Gamers," "Fashionists," "Students," and other persona groups.

- Customer Segment: This indicates the frequency of a customer's engagement with the platform, with segments including new, inactive, retained, reactivated, or churned.

- Customer Tag: These are loyalty program classifications to which customers belong, such as new member, premium, efso, new member & premium, efso & new member & premium, efso & new member, efso & premium.

- Page Type: Denotes the type of page customers browse, including homepage, search results pages, product detail pages, merchant and brand pages, etc.

- Page Component: Signifies the type of page component with which customers interact, such as buybox, widgets, and variants.

- Campaign Label and Label Type: Refers to labels on product images that display offers related to payment and delivery conveniences, benefits.

- Discount: Discounts can significantly influence purchase decisions.

- Product Position on Page: Indicates whether the product is easily accessible on the page or not.

- Notification Permission Status: Represents whether the customer permits notifications, indicating their openness to campaign communications.

- Product Review Counts and Scores: Customers often rely on other customers' reviews about products when making purchase decisions.

- Recommendation: Reflects whether the product is recommended to the customer by a recommendation system or not.

This paper will delve into the examination of factors influencing customers' purchase decisions using a case study from an e-commerce retail platform. The data associated with the factors outlined earlier will be collected and compiled into a dataset.

Data gathering and pre-processing pose notable challenges in this study. The substantial volume of data gives rise to performance issues during data processing. Optimizations are required for SQL queries, and pipelines must be established for efficient data feeding to the model. Additionally, addressing quality issues within the data constitutes another significant challenge that demands careful attention.

In this paper, these steps will be implemented.

- Data needed for the study will be collected from e-commerce platform transactions stored in Google's BigQuery environment. Performance issues will be mitigated through SQL optimization and other methods.

- Data quality issues will be resolved by addressing missing values and replacing unstandardized data with appropriate information suitable for the model.

- Unbalanced data problems will be managed to ensure consistent and reliable results.
- Supervised machine learning (ML) and deep learning models will be employed on the data for feature selection and feature importance procedures.
- Results from different models will be compared.
- The performance of various models will be evaluated and compared.

2. LITERATURE REVIEW

Customer purchasing decisions can be influenced by a multitude of factors, a topic extensively explored in the literature. Teye & Missah (2020) delved into the relationship between platform usability and customer purchase decisions. Their model considered parameters such as the number of pages visited, duration of page visits, bounce and exit rates, alongside traffic type, visitor type, special days, and weekends. Utilizing adaptive synthetic (ADASYN) sampling to address dataset imbalance, they applied one-hot encoding to convert categorical data into numerical data. Feature selection using Sklearn's SelectKBest method resulted in 15 impactful features, and outliers were removed for enhanced accuracy. Employing a random forest classifier, they achieved an accuracy of 95.12%, emphasizing the significance of factors like navigation time, special days, and pricing and design components in customer decisions.

Sakar et al. (2019) conducted real-time analysis of online shopping behavior, simulating a physical store salesperson to understand customer purchasing intentions. Their two-module system predicted shopping intent and abandonment likelihood. For the first module, they employed random forest, support vector machines, and multilayer perceptron classifiers, achieving the best performance with a multilayer perceptron accuracy of 87.24%. The second module, using the LSTM-RNN model, identified customer website abandonment with a 74.3% accuracy after a single action.

Baati & Mohsil (2020) replicated Sakar et al.'s study, utilizing the same dataset. Their system, interpreting customer purchasing intention, involved two parts offering marketing content and more generous offers. Employing naive Bayes, decision tree, and random forest classifiers, the random forest method yielded the highest accuracy (86.78%) and an F1 score of 0.60, minimizing the risk of customer website abandonment.

Topal (2019) applied different methods on the same dataset as Teye & Missah (2020), utilizing the Fisher score method for feature selection and the decision tree classifier for classification. Addressing imbalanced data with K Fold Cross-validation, the study identified "Page Value" as the most powerful feature, achieving an accuracy of 88.39%.

Fatta et al. (2018) proposed a hypothesis categorizing conversion rate influencers into promotional and quality factors. Factors like free shipping, free return services, discount policies, and sale seasons positively affected the conversion rate. OLS regression analysis and qualitative comparative analysis (QCA) were employed on data from six SMEs in e-commerce, revealing that discounts and free shipping together positively influence the conversion rate. Notably, discounts during sale seasons had a positive impact on conversion rates, with page speed consistently contributing positively regardless of other factors.

It is certain that there is relationship between conversion rate and website design features. In this scope, McDowel et al. (2016) made a study to test these hypotheses in 2016.

- Hypothesis 1: E-commerce website features that enhance purchase intention within the visitor greeting stage of the website are associated with conversion.
- Hypothesis 2: E-commerce website features that enhance purchase intention within the catalog pages of the website are associated with conversion.
- Hypothesis 3: E-commerce website features that enhance purchase intention within the shopping cart pages of the website are associated with conversion.
- Hypothesis 4: E-commerce website features that enhance purchase intention within the checkout pages of the website are associated with conversion.

In this study, the dependent variable is the conversion rate, while independent variables include various website design features. The study investigates the impact of recommended and featured products on greeting and catalog pages, the presence of a shopping cart icon in the navigation bar, the display of shipping charges in the shopping cart, one-time credit card registration on the checkout page, and the provision of links to site pages. Analysis is conducted using Pearson correlation and p-values. Results highlight the critical role of early customer interaction, with recommended and featured products on the greeting page and the shopping cart icon positively influencing the conversion rate. Conversely, links to other pages have a negative impact. On catalog pages, special offers and subject tabs show a strong positive correlation, while recommended products in shopping carts show no significant association. In shopping carts, features like instantaneous pricing, displaying shipping charges, and offering related products positively impact the conversion rate. However, features like "Return to main catalog" and "Order tracking" have a negative impact. On the checkout page, the requirement to provide an email address negatively affects the conversion rate, while human contact information and instantaneous pricing have a positive impact. The study suggests a relationship between effective web design features and the conversion rate, emphasizing the positive impact of flow-enhancing features.

Cezar and Öğüt (2016) explored the role of customer reviews, recommendations, and rank order in search listings on the conversion rate in hotel bookings. Data from Booking.com on Barcelona and Paris hotels were analyzed using fractional regression models, quasi-maximum likelihood estimation, and a regression model with a beta distribution. High customer ratings, a high number of recommendations, and high ranks in search listings were found to positively impact the conversion rate. Higher ranks in search listings had a positive effect when combined with high numbers of recommendations and high customer ratings, while service rating and hotel star level did not significantly influence the conversion rate. Additionally, a higher price had a negative impact on the conversion rate.

Diamantaras et al. (2021) employed the LSTM-RNN method to understand customer purchase intention using web server logs categorized into different customer actions. The study achieved 98.15% accuracy with a window size of 20.

Sheil et al. (2018) used the RNN method to predict purchase intent, achieving 98.4% accuracy using datasets from The RecSys 2015 Challenge and Retail Rocket Kaggle.

Nazir et al. (2023) investigated the influence of artificial intelligence on social media engagement and purchase intentions. The study distributed questionnaires to customers who booked hotels or flights via Booking.com in three regions of Oman, finding positive relationships between AI technology and social media engagement, customer satisfaction, and repurchase intention. The study concludes that AI in social media campaigns effectively converts users into customers.

Hu et al. (2023) applied feature extraction to understand shopping behavior. Features are divided into five categories.

- User characteristics: The count of various behaviors of the users before the prediction day.
- Product characteristics: the number of users visiting the product before the prediction day.
- Product category characteristics: Which category a product belongs to.
- User product characteristics: The count of various actions of users on a product before the prediction day.
- User - product category characteristics: The user preference on certain product categories.

The utilization of these feature categories provides essential insights for the analysis and prediction of shopping behavior, enabling a thorough exploration of various facets related to users, products, and their interactions.

The study employed a deep forest model, a construct derived from random forest structures in a manner analogous to deep learning models. Different types of random forests were applied to each layer of the model, enhancing overall fault tolerance. The study concluded with an impressive F1 score of 9.51 achieved within a training time of forty-one seconds. Additionally, various other models, including support vector machine, random forest, XGBoost, and deep neural network models, were experimented with. The deep forest model outperformed the others, achieving the highest F1 score. This study underscores the superior performance of the deep forest model in specific cases compared to alternative models.

Table 2.1 provides a comprehensive overview of all methods in the literature related to the topic, listing them by author name.

Table 2-1 Methods Comparison in Literature

Authors	Methods
Teye and Missah	Random Forest
Sakar et al.	MLP, Random Forest, LSTM-RNN, SVM
Baati and Mohsil	Naive Bayes, Decision Tree, Random Forest
Topal	Decision Tree
Fatta et al.	OLS Regression Analysis, QCA
McDowel et al.	Pearson Correlation
Cezar and Öğüt	Fractional Regression
Diamantaras et al.	LSTM-RNN
Sheil et al.	RNN
Nazir et al	Partial Least Square
Hu et al	Deep Forest, SVM, Random Forest, XGBoost, Deep Neural Network

As indicated in Table 1.1, authors employ diverse methodologies such as correlation-based, decision tree-based, and deep learning methods in their studies. Notably, permutation-based methods are conspicuously absent across all studies, and there is a noticeable gap in research that systematically applies all types of methods to a single dataset for comparative purposes.

However, a significant portion of the studies featured in the literature relies on small dataset sizes, often providing limited details about the data collection process. This reliance on small datasets raises concerns about the potential for misleading results, emphasizing the necessity of comprehensive information for accurate result interpretation.

Furthermore, a considerable number of studies share the same dataset, limiting the diversity and scale of the data used in analyses. Addressing these concerns, we propose a mixed-methods study that investigates the factors influencing customer purchase decisions. This study will leverage a large dataset collected from transactions on a major e-commerce platform, aiming to provide more robust and realistic results.

The overarching goal of this study is to contribute to both academic knowledge and practical interventions by unraveling the complex relationships between potential factors in e-commerce and customer purchase decisions, grounded in business knowledge. Acknowledging potential limitations, such as challenges associated with processing diverse-sized data with limited computational resources and collecting various types of customer interactions, is crucial.

In conclusion, this proposed study seeks to enhance existing literature by delving into the factors influencing customers' purchase decisions. By utilizing a large dataset and employing a variety of methods related to the topic, the study aims to offer valuable insights that can inform both research and practical interventions in this critical area.

3. DATA PREPARATION

The data utilized in this study is sourced from customers' "add to cart" transactions, where products are added to the online platform's shopping cart. The selection of "add to cart" transactions as the primary dataset is based on the understanding that this action represents a distinct interaction event initiated by the customer within the platform. When customers engage in "add to cart" transactions, it signals an intention to make a purchase, reflecting genuine interest in the online platform and its offerings. It's noteworthy that "add to cart" transactions can originate from any page with various features.

The primary objective of this study is to assess the factors within "add to cart" data and their correlation with whether a customer ultimately completes a purchase during the session. The methodology employed will strive to pinpoint the factors most strongly associated with the purchase decision.

In the pursuit of identifying factors related to purchases, the study will scrutinize the following types of factors within "add to cart" transactions:

- **Related to screen design characteristics:**
  Page Type: The dataset encompasses information on the page type when a product is added to the cart. Customers have the flexibility to initiate the 'add to cart' action from various pages such as product detail pages, listing pages, and the homepage, each characterized by unique design elements. The study aims to investigate whether the page type significantly impacts the purchase decision, recognizing the potential influence of different page contexts

Product Position on the Page: Examining the position of products on the page, particularly on listing pages in the mobile app, where customers initially view only two products without scrolling. This factor addresses the potential impact of product visibility and placement on the purchase decision, acknowledging the limited initial product exposure.

Campaign Label on Product Image: Assessing whether the presence of a campaign label on the product image influences the purchase decision. This factor delves into the potential effect of promotional labels on customer choices and their likelihood to proceed with a purchase.

These screen design characteristics collectively aim to explore the visual and contextual aspects of the online platform that may contribute to customers' decisions to add products to their cart and ultimately make a purchase. The study recognizes the importance of understanding how design elements and layout considerations influence user behavior and purchasing patterns.

- **Related to customer characteristics:**

  Customer Persona: This factor encompasses information about the customer's persona, providing insights into their shopping habits and preferences. For example, a customer with a "gamer" persona might be inclined to shop frequently in the electronics category, while a customer with a "mother-children" persona might show preferences for baby food and daily household needs. Understanding customer personas helps identify targeted marketing strategies and personalized recommendations.

  Customer Segment: Examining customer segments based on their shopping frequency on the online platform. For instance, a customer who shops regularly is categorized as a retained customer, whereas a customer who recently made their first purchase falls into the new customer segment. Identifying customer segments aids in tailoring engagement strategies for different customer groups.

Customer Tags: Analyzing customer tags, which indicate whether a customer is part of the company's loyalty program, such as Premium or Efso. Loyalty programs influence customer retention and engagement, and understanding the impact of these programs on purchase decisions is crucial for effective customer management.

Notification Permission Status: Evaluating whether a customer receives marketing campaign messages from the company, indicating their sensitivity to campaigns. This data type provides insights into customers' preferences regarding communication and helps optimize marketing strategies. It is noteworthy that this type of data is not commonly explored in existing literature, highlighting the novel aspect of this study.

These customer-related characteristics delve into the individual preferences, behaviors, and interactions of customers on the platform. By examining these factors, the study aims to uncover the nuanced aspects of customer decision-making processes during the 'add to cart' transactions.

- **Related to How Customers Start to Interact with the Online Platform:**
  Traffic Source Type: This factor focuses on the type of traffic source through which customers reach the platform. Customers may directly access the platform or arrive via search engines such as Google, Instagram ads, and other channels. The study aims to explore the diverse points of entry that customers use to access the platform, recognizing the potential influence of different traffic sources on their subsequent actions.

  This aspect addresses the initial touchpoints that customers have with the platform and acknowledges the various online channels that drive traffic. By incorporating this data into the study, the goal is to understand how the source of customer traffic may impact their decision to add products to the cart and make a purchase. This type of data is considered novel in the literature, highlighting the study's contribution to exploring previously uncharted aspects of customer behavior.

- **Related to product characteristics:**

  Presence of a Price Discount: This factor examines whether a product has a price discount. The presence of discounts is a known influencing factor in customer purchase decisions, as highlighted by Fatta et al. Discounts can serve as incentives for customers to proceed with the purchase, and their impact on decision-making will be assessed in the study.

  Product Review Counts and Ratings: Analyzing the number of product review counts and the associated ratings. High product ratings are known to positively influence customer purchase decisions, as noted by Cezar and Öğüt. Customer reviews and ratings contribute to building trust and confidence in the product, potentially affecting the likelihood of a purchase.

  Recommendation by the Data Science Engine: Investigating whether the product is recommended by the data science engine. Product recommendations play a significant role in guiding customer choices, and their influence on the purchase decision will be explored. Cezar and Öğüt also emphasize the positive impact of recommendations on customer decisions.

  These product-related characteristics delve into the features and attributes associated with individual products on the platform. By examining these factors, the study aims to understand how product-related elements contribute to customers' decisions during the 'add to cart' transactions and subsequent purchases. The insights gained from these characteristics can inform strategies for promoting products and enhancing the overall customer experience.

3.1 Raw Data Collection

The data for this study was sourced from customer transactions on the e-commerce company's online platform, specifically from both IOS and Android native apps, with the web platform excluded from consideration.

Initially, data on "add to cart" transactions was retrieved from the database, encompassing details such as page type, location, campaign label specifics, product position on the page, discounted price status, customer notification permission, product review count, and score. The "add to cart" click data included information on screen design, product details, device characteristics, and platform specifications.

Following this, customer persona, tag, and segment information were gathered from the system and integrated into the "add to cart" data. Customer segment values included new, churn, inactive, retained, and reactivated, while customer tag values consisted of categories such as new member, premium, efso, and various combinations thereof. Customer persona values encompassed categories like gamer, student, no_persona, anne_kadın, fashionista, and digital_ebeveyn.

Subsequently, information on all customers' traffic types was acquired from the system, categorizing sessions as SEO, paid, or direct traffic. SEO indicated customers arriving from search engine searches, paid denoted customers coming from paid advertisements across platforms, and direct signified customers directly opening the online platform.

Finally, purchase information was incorporated into the "add to cart" data using session id details. The "is_order" column in the dataset was set to true if users added items to the cart and made a purchase in the same session, and false if they added items to the cart but did not make a purchase.

In conclusion, the raw dataset, containing 14 features and one target variable ("is_order"), has been finalized. This dataset reflects factors influencing customers' purchase actions in a session and consists of 991,615 rows.

Table 3-1 Raw Data Structure

Data Columns	Description
user id	id of the user in the system
session id	id of the session in the system
page type	page type can be home, product detail, category
location	location of the page, it can be buybox, compare
label type	label is an image shown on the product image to attract customers. Label types can be campaign, value added service.
position	position rank of product on the page
is price discount	whether price of the product is shown as discounted
is notifications on	whether customer give permission to notifications
customer review count	count of customer reviews related the product
customer review score	score of customer reviews related the product
segment	customer segment is defined according to shopping frequencies of the customer
tag	whether customer is member of the loyalty program, Premium etc.
persona	persona info describes shopping habits of the customer like gamer, mother etc.
is_direct	whether the customer visit the online platform directly without any interaction point.
is_paid	whether the customer visit the online platform through marketing activities
is_seo	whether customer visit the online platform through search engine like google
is_order	whether customer purchased in the session

3.2 Data Preprocessing

Upon reviewing the dataset, it was noted that certain feature values were null, an occurrence deemed impractical and attributed to technical issues. Specifically, these null values originated from the "page_type" and "location" features, totaling 23 instances, which were subsequently dropped. Following this, an examination of the data types revealed that the "is_notification_on" feature was initially in object format and was subsequently converted to boolean type.

Moreover, some features were categorical, and to facilitate model development using the one-hot coding method, these categorical values were converted to numerical format. The affected columns in our dataset include 'page_type', 'location', 'label_type', 'segment', 'tag', and 'persona'. The one-hot coding method was chosen for its superior performance, as it encodes feature values into a binary vector array in separate columns. Given the substantial size of our dataset, this method is well-suited to our case. After the application of this method, the column count in our dataset increased to 68.

Furthermore, as numerical features in the dataset had different ranges, normalization was applied to enhance model performance and ensure accurate results. The MinMaxScaler method was employed for data normalization, which rescales variables into the range [0,1].

4. METHODOLOGY AND EXPERIMENTS

4.1. Methodology

To pinpoint the critical factors influencing customer purchase decisions, the study will analyze, apply, and compare feature importance methods. These methods assign scores to the features based on their effectiveness in predicting the target value. In this context, the target value is binary, indicating whether the customer made a purchase or not.

Previously, potential features that may impact purchase decisions were identified using domain knowledge. Feature importance methods will further refine this selection, highlighting the most significant features through statistical analysis, thereby offering valuable insights into the dataset.

There are three primary types of methods for determining feature importance.

- The first type involves calculating coefficients. Coefficients illustrate the relationship between features and the target variable, indicating how much the target variable changes, positively or negatively, when a feature changes. Both logistic regression and linear regression methods use coefficients when predicting target values. While linear regression predicts continuous target variables, logistic regression predicts categorical variables.

  In our study, where the target variable is the decision to purchase or not, characterized by categorical binary classes, we employ the logistic regression method.

- The second type is tree-based methods. Tree-based methods are powerful in terms of explainability. In this method, the importance of each feature is determined by how much it reduces impurity in a tree. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. In our study, we will use CART, Random Forest, and XGBoost methods as tree-based methods.

  While CART employs a single decision tree, random forest and XGBoost methods use an ensemble of many decision trees.

  Random forest combines many independently trained trees. On the other hand, XGBoost trains decision trees sequentially. Each tree is trained based on the learnings from previous trees, and adjustments are made according to the errors of the previous tree.

- The third type involves permutation-based methods. In permutation-based feature importance methods, some feature values are shuffled in the model. If the model error increases significantly, the shuffled feature is assumed to be more important. This implies that the feature has a more substantial impact on predictions.

  Permutation-based methods cannot be used alone. Before applying permutation-based methods, a model must be trained using one of the classification methods. In this study, we will employ the permutation method with Random Forests, XGBoost, and Multilayer Perceptron classifiers. Once a model has been trained with any of these classifiers, the shuffling process is applied to the model to identify features that primarily contribute to an increase in model errors.

4.2. Experiments

We conducted experiments following the procedures mainly:

1. Preprocessed the data, then it was normalized.
2. Subsequently, the dataset was divided into training and testing sets.
3. The training data was then utilized to train the relevant method.
4. Following the training phase, accuracy, F1 were computed using the test data.

The details of our experiments are elaborated below.

4.2.1 Logistic Regression

"Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables." (Saini, 2021)

In our dataset, the dependent variable represents the customer's purchase decision in a session, categorized as either a purchase or no purchase, constituting a binary class. These values are discrete, not continuous. Logistic regression is employed when the dependent variable is not continuous. To determine whether customers are likely to purchase or not, a threshold is set between the purchase and no purchase classes.

Linear regression, being unbounded, is unsuitable for detecting classes of customers. Consequently, we opted for logistic regression instead of linear regression.

The logistic regression method predicts customers' purchase decisions in a session using the sigmoid function. This function incorporates all independent features and assigns customer sessions to binary classes: 1 (purchase) or 0 (no purchase).
In more detail, the sigmoid function transforms the output of linear regression to logistic. It is a mathematical function that maps the continuous output of linear regression to values

between 0 and 1. If the linear regression output tends towards negative infinity, it maps the value to 0, representing no purchase in our example. Conversely, if the linear regression output tends towards positive infinity, it maps the value to 1, indicating a purchase.

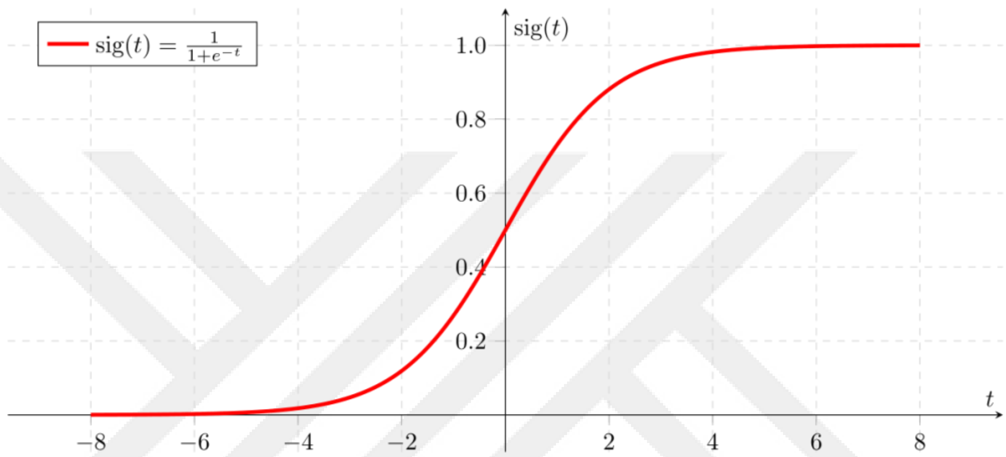

Figure 4.1 Sigmoid Function

In the initial attempt, an unbalanced dataset was used to train the model, resulting in a test accuracy of 0.83. Since the data was unbalanced, the f1 score was examined, which measures a model's accuracy. The purchase f1 score was low (0.34), while the no purchase f1 score was high (0.9). Due to the high count of no purchase sessions, the no purchase f1 score was disproportionately high, indicating that the results for the purchase class were not reliable. To address this issue, the model was retested with balanced data.

To balance the data, the undersampling method from the imblearn library was applied. After this procedure, the model's accuracy dropped to 0.71. However, the f1 scores for no purchase and purchase became closer, with a no purchase f1 score of 0.64 and a purchase f1 score of 0.76. The same procedure was applied using the oversampling method, another data balancing technique. However, the accuracy and f1 scores remained unchanged in the results.

Normalization of the data was performed using the StandardScaler method, which was then replaced with the MinMaxScaler method. However, neither the accuracy nor the f1 score showed significant changes.

Subsequently, PCA, a feature selection procedure, was applied, reducing the feature count to 10. Logistic regression was then employed, but the accuracy score did not improve.

Finally, the accuracy score was 0.65, with a no purchase f1 score of 0.68 and a purchase f1 score of 0.62. The method identified the ten most important features and their importance values, as shown in Table 4.1.

Table 4-1 Logistic Regression Results

Feature_name	**Importance**
segment_Reactivated Customer	2.16
segment_Retained Customer	2.00
segment_New Customer	1.97
location_buy_ticket	1.69
page_type_add_basket	1.69
persona_Gamer	1.12
segment_Inactive Customer	1.09
persona_Student	0.93
segment_Churn Customer	0.86
tag_New Member & Premium	0.84

4.2.2 Decision Tree

“A decision tree is a hierarchical model used in decision support that depicts decisions and their potential outcomes, incorporating chance events, resource expenses, and utility. This algorithmic model utilizes conditional control statements and is non-parametric, supervised learning, useful for both classification and regression tasks. The tree structure is comprised of a root node, branches, internal nodes, and leaf nodes, forming a hierarchical, tree-like structure.” (Saini, 2021)

To comprehend the decision tree method, it's essential to introduce some terminology. Nodes essentially represent features in the dataset. Root nodes act as the initial node that divide the dataset. Each split generates sub-trees, as root nodes are further divided into decision nodes. The process of splitting continues until leaf nodes are identified.

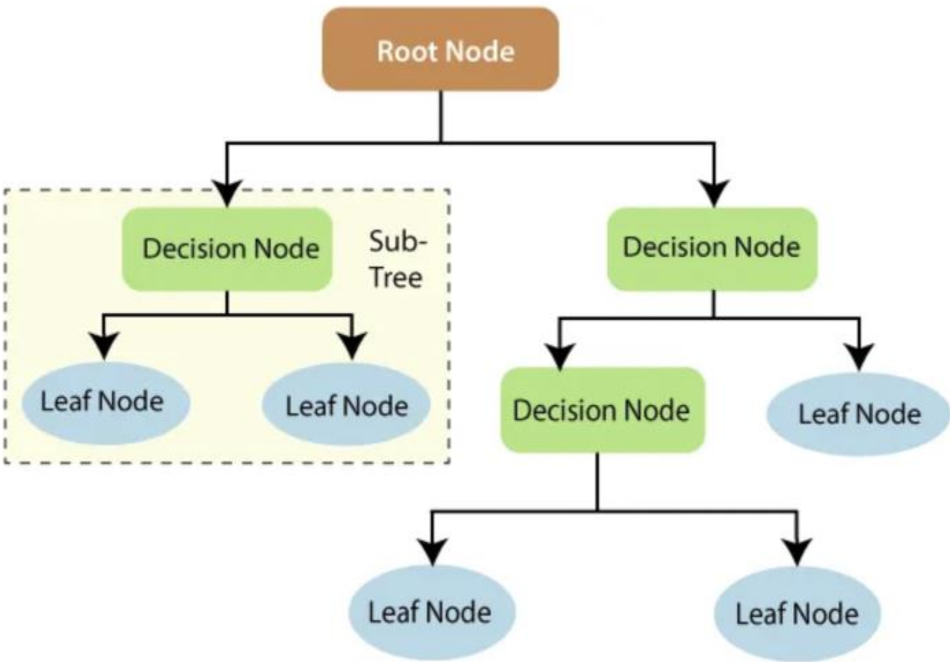

Figure 4.2 Decision Tree

This algorithm initially selects the best feature in the dataset as the root node using Attribute Selection Measure (ASM). ASM identifies the feature that will create the most homogeneous subsets of data. Thus, ASM splits the root node into subsets, which can be decision nodes or leaf nodes. If the subset is a decision node, the split process continues; if it is a leaf node, the split procedure stops. By repeating this process, many decision trees are created.

ASM uses Information Gain and Gini Indexes. Information gain measures the changes in entropy after splitting the dataset, identifying features that provide significant information about the class. The Gini Index looks at the purity measure in the dataset when splitting, with a preference for a low Gini index. The splitting process stops when there is a high information gain and purity in the datasets, but it is crucial to avoid much higher information gain or impurity to prevent overfitting.

Using unbalanced data to train the model results in a test accuracy of 0.82. However, the f1 score for the no purchase class is much higher than the purchase class, with a no purchase f1 score of 0.90 and a purchase f1 score of 0.37. To address this imbalance, the model is tested with balanced data using the oversampling method. While the f1 scores get closed to each other (0.69 for no purchase, 0.78 for purchase), the accuracy of the model decreases to 0.74.

Normalization of the data is achieved by applying MinMaxScaler instead of StandardScaler, but this adjustment does not impact accuracy or f1 scores.

In an effort to enhance model accuracy, the parameters of the decision tree model are tuned. The best parameters are determined using the RandomizedSearchCV method, which randomly tests a set of parameters and calculates the model score. However, despite tuning parameters such as setting max_depth to 10, max_features to 0.8, and trying both "best" and "random" for the splitter parameter, these adjustments do not lead to an improvement in accuracy.

The method identifies the ten most important features and their importance values, as shown in the table below.

Table 4-2 Decision Tree Results

Feature_name	Importance
location_buy_ticket	0.27
segment_Inactive Customer	0.19
segment_Churn Customer	0.10
position	0.10
tag_No-Tag	0.06
segment_Reactivated Customer	0.05
tag_Premium	0.04
persona_No-Persona	0.03
segment_Retained Customer	0.03
customerreviewcount	0.02

4.2.3 Random Forest

"Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result." (Donges, 2023)

Random Forest is an ensemble of decision trees that utilizes the bagging method. Bagging involves obtaining random subsets from the dataset and independently training each subset, thereby creating multiple models. Each training process results in separate outputs, and each output selects a class after training. These outputs are then combined using the majority voting method. The final output of the process is determined by the class selected by the majority of models.

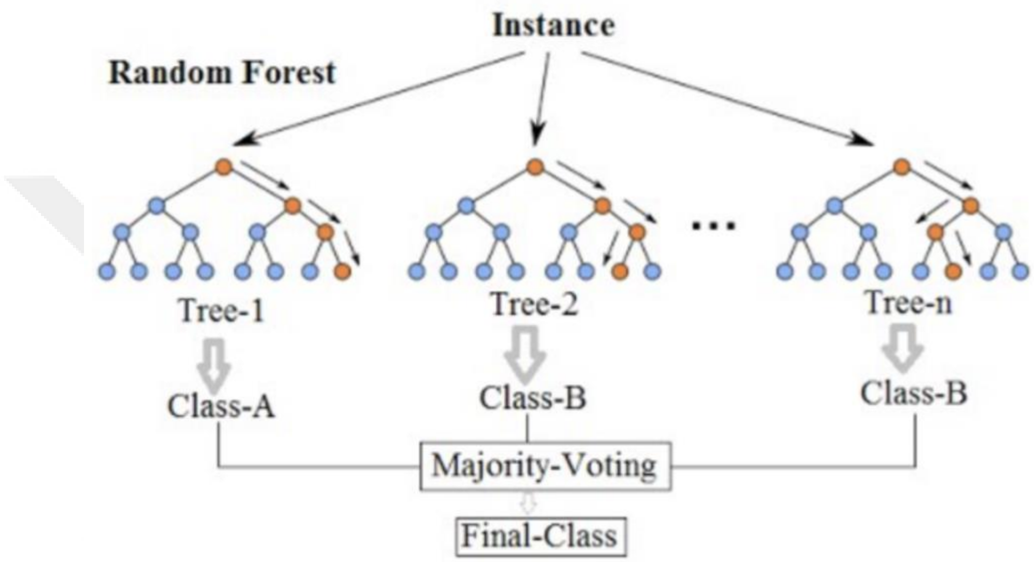

Figure 4.3 Random Forest

When the model was trained with unbalanced data, the accuracy of the model on the test data was 0.83. However, the f1 score for the no purchase class was significantly higher than the purchase class, indicating that the results for the purchase class were not reliable. The no purchase f1 score was 0.9, while the purchase f1 score was 0.35. To address this issue, the model was retested with balanced data using the oversampling method. After balancing the data, the f1 scores became closer, with a no purchase f1 score of 0.67 and a purchase f1 score of 0.77. However, the accuracy of the model decreased to 0.73.

To optimize the learning process of the model, Random Forest parameters were tuned. Tuning involves trying different hyperparameters in the model learning process to find the ones that yield the best results. Two main methods for hyperparameter tuning are

Grid Search and Randomized Search, and in this case, the Randomized Search method was employed.

The best parameters were identified for the model:

n_estimators = 115
min_samples_split = 2
min_samples_leaf = 1
max_features = 'log2'
max_depth = 19
criterion = 'gini'

Even after applying the best parameters, the model accuracy remained at 0.73. The no purchase f1 score was 0.67, and the purchase f1 score was 0.78.

The method defined the ten most important features and their importance values, as shown in the table below.

Table 4-3 Random Forest Results

Feature_name	**Importance**
position	0.14
is_price_discount	0.13
isnotoficationon	0.11
customerreviewcount	0.10
customerreviewscore	0.09
is_direct	0.09
is_paid	0.07

is_seo	0.06
page_type_	0.02
page_type_add_basket	0.02

4.2.4 XGBoost

Gradient Boosting is a popular boosting algorithm where each predictor corrects its predecessor's error. XGBoost, short for Extreme Gradient Boosting, is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created sequentially. Weights play a crucial role in XGBoost, as they are assigned to all independent variables and fed into the decision tree to predict results. The weight of variables predicted incorrectly by the tree is increased, and these variables are then fed into the second decision tree. These individual classifiers or predictors then ensemble to create a strong and more precise model. XGBoost is versatile and can work on regression, classification, ranking, and user-defined prediction problems.

XGBoost shares similarities with the Random Forest algorithm. While Random Forest trains each model simultaneously, XGBoost trains sequentially, correcting errors from predecessors at each training step.

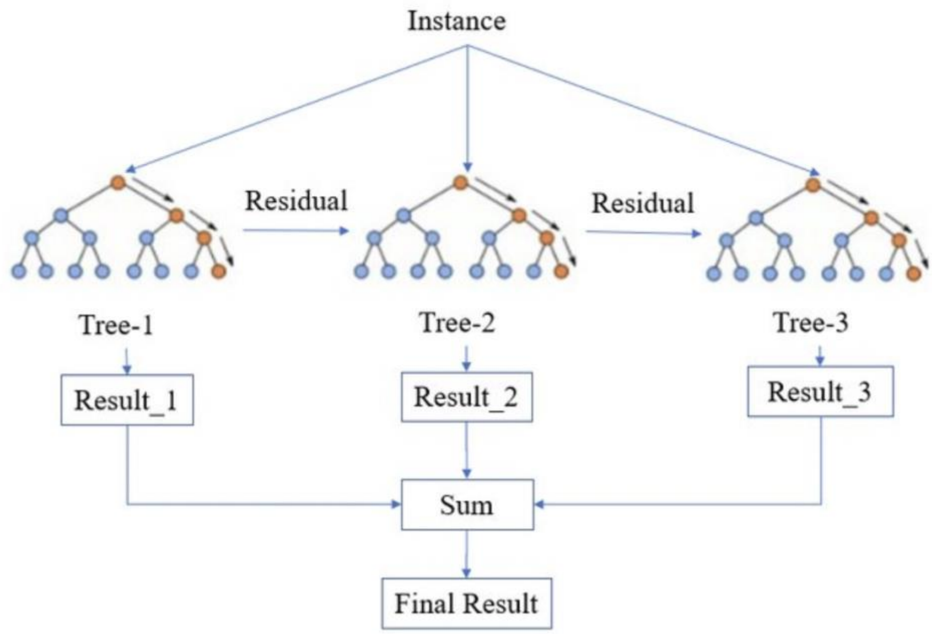

Figure 4.4 XGBoost

When the model was trained with unbalanced data, the accuracy of the model on the test data was 0.83. However, the f1 score for the no purchase class was significantly higher than the purchase class, indicating that the results for the purchase class were not reliable. The no purchase f1 score was 0.9, while the purchase f1 score was 0.35. To address this issue, the model was retested with balanced data using the oversampling method. After balancing the data, the f1 scores became closer, with a no purchase f1 score of 0.66 and a purchase f1 score of 0.76. However, the accuracy of the model decreased to 0.72.

To improve model accuracy, XGBoost parameters were tuned using the Randomized Search method to find the best parameters.

The best parameters were identified for the model:

- subsample = 0.7
- n_estimators = 250
- learning_rate = 0.1
- max_depth = 15

- colsample_bytree = 0.9
- colsample_bylevel = 0.9

After applying the best parameters, the model accuracy slightly increased to 0.74. Additionally, f1 scores showed slight improvement, with a no purchase f1 score of 0.68 and a purchase f1 score of 0.77.

The method defined the ten most important features and their importance values, as shown in the table below.

Table 4-4 XGBoost Results

Feature_name	**Importance**
position	0.721
is_price_discount	0.164
isnotificationson	0.042
customerreviewcount	0.024
customerreviewscore	0.005
is_direct	0.003
is_paid	0.002
is_seo	0.002
page_type_	0.002
page_type_add_basket	0.002

4.2.5 Permutation with Random Forest

“The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature.” (scikit-learn.org)

We trained a model using Random Forest, as described in section 4.2.3. Subsequently, we applied the permutation method to this model. The permutation method operates based on the predictions of the model: it shuffles the values in selected features and monitors the model's error increase. If the model error increase is substantial for a selected feature, that feature is considered important.

The results of this method have defined the ten most important features and their corresponding importance values, as outlined in the table below."

Table 4-5 Permutation with Random Forest Results

Feature_name	**Importance**
segment_Retained Customer	0.0557 ± 0.0006
segment_Reactivated Customer	0.0514 ± 0.0005
tag_Premium	0.0321 ± 0.0009
position	0.0286 ± 0.0005
is_direct	0.0178 ± 0.0006
segment_Inactive Customer	0.0165 ± 0.0006
tag_No-Tag	0.0150 ± 0.0003
page_type_search	0.0103 ± 0.0003

segment_New Customer	0.0090 ± 0.0002
is_paid	0.0089 ± 0.0002

4.2.6 Permutation with XGBoost

"We trained a model using XGBoost, as described in section 4.2.4. Following that, we applied the permutation method.

The results of this method have defined the ten most important features and their corresponding importance values, as outlined in the table below.

Table 4-6 Permutation with XGBoost Results

Feature_name	Importance
segment_Retained Customer	0.0598 ± 0.0009
segment_Reactivated Customer	0.0390 ± 0.0004
position	0.0271 ± 0.0006
tag_Premium	0.0213 ± 0.0006
segment_Inactive Customer	0.0185 ± 0.0004
is_direct	0.0120 ± 0.0005
segment_New Customer	0.0104 ± 0.0001
page_type_add_basket	0.0095 ± 0.0004
segment_Churn Customer	0.0079 ± 0.0003
tag_No-Tag	0.0073 ± 0.0004

4.2.7 Permutation with Multilayer Perceptron

We did not create a multilayer perceptron model previously. Therefore, it was necessary to perform a multilayer perceptron model before applying the permutation method.

“The perceptron is very useful for classifying data sets that are linearly separable. The MultiLayer Perceptron (MLPs) breaks this restriction and classifies datasets which are not linearly separable. They do this by using a more robust and complex architecture to learn regression and classification models for difficult datasets.” (deepai.org)

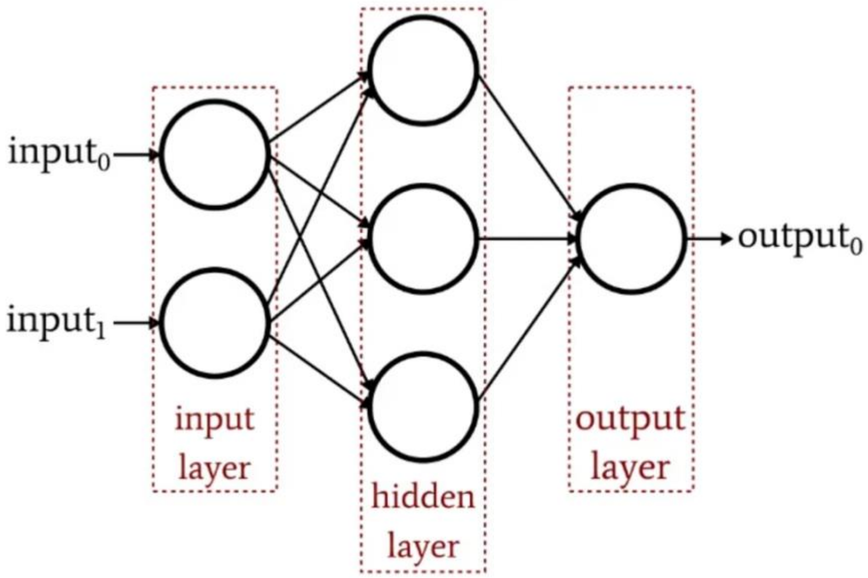

Figure 4.5 Multilayer Perceptron

When the model was trained with unbalanced data, the accuracy of the model on the test data was 0.83. However, the f1 score for the no purchase class was significantly higher than the purchase class, indicating that the results for the purchase class were not reliable. The no purchase f1 score was 0.9, while the purchase f1 score was 0.35. To address this issue, the model was retested with balanced data using the oversampling method. After balancing the data, the f1 scores became closer, with a no purchase f1 score of 0.67 and a purchase f1 score of 0.75. However, the accuracy of the model decreased to 0.72.

In an attempt to improve model accuracy, multilayer perceptron parameters were tuned using the Grid Search method, and the best parameters were identified:

- Activation: 'relu'
- Alpha: 0.0001
- Hidden Layer Sizes: (50, 100, 50)
- Learning Rate: 'adaptive'
- Solver: 'adam'

Despite applying the best parameters, the model's accuracy and f1 scores did not improve or change.

The method defined the ten most important features and their importance values, as shown in the table below.

Table 4-7 Permutation with MultiLayer Perceptron Results

Feature Name	Importance
segment_Retained Customer	0.0415 ± 0.0006
segment_Reactivated Customer	0.0387 ± 0.0003
tag_Premium	0.0210 ± 0.0005
segment_Inactive Customer	0.0166 ± 0.0005
is_direct	0.0098 ± 0.0002
segment_Churn Customer	0.0081 ± 0.0004
persona_No-Persona	0.0079 ± 0.0002
position	0.0073 ± 0.0002
is_paid	0.0065 ± 0.0003
tag_No-Tag	0.0049 ± 0.0004

5. EXPERIMENT RESULTS

In this segment, the outcomes of the experiments and the performances of the models will be analyzed and contrasted.

5.1 Logistic Regression

The logistic regression method suggests that all features exert an influence on purchase behavior. Among them, three features associated with customer segment information stand out as the most impactful factors in influencing customers' purchase decisions. Specifically, the reactivated, retained, and new customer segments are identified as being highly likely to result in a purchase. Examining the segment descriptions provided, it is noted that customers in these segments have placed orders within the last 30 days. This observation leads to the conclusion that customers who have made recent orders are more inclined to make a purchase.

The segment descriptions:

- New Member: user that do not have order yet.
- New Customer: the user in a 30 days period of after first order.
- Retained: the user that have order in last 31-60 days, also have order in last 30 days.
- Reactivated: the user that have not order last 31-60 days, but have order in last 30 days.
- Churn: the user that have order in last 31-60 days, but do not have order in last 30 days.
- Inactive: the user that do not have order in last 31-60 days, also do not have order in last 30 days.

The features 'location_buy_ticket' and 'page_type_add_basket' are associated with the 'Scratch and Win' game within our app. In this context, customers engage in purchasing tickets with small amounts to participate in draws within the game. The logistic regression results highlight that customers who buy 'Scratch and Win' tickets are significantly more likely to make a purchase.

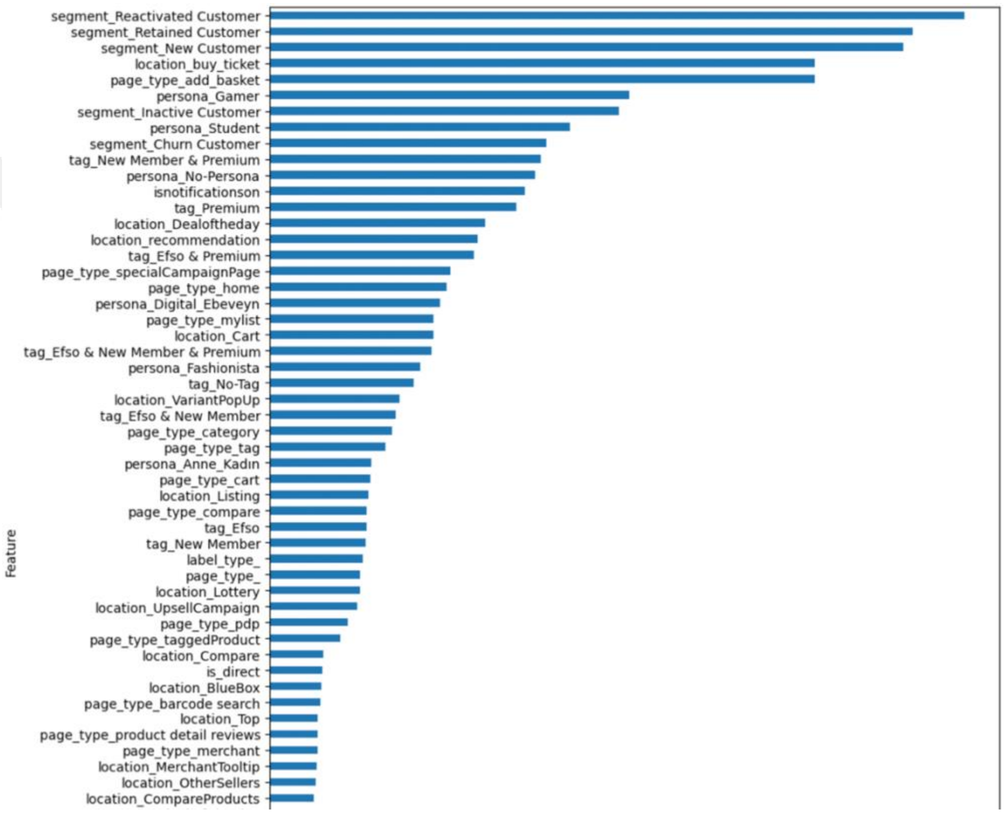

Figure 5.0.1 Logistic Regression Results

5.2 Decision Tree

In contrast to logistic regression, the decision tree method does not indicate that all features impact the purchase decision. Notably, the features “location_buy_ticket,” “segment_inactive_customer,” “segment_churn_customer,” and “position” are identified as impactful on the purchase decision, according to the results.

Interestingly, the “location_buy_ticket” feature, which logistic regression also found impactful, stands out from other features in the decision tree method. It suggests that customers purchasing tickets for the “Scratch and Win” game are significantly more likely to make a purchase.

Unlike logistic regression, the decision tree method reveals distinct segments related to purchases. It suggests that customers who haven't made a purchase in the last 30 days are more likely to make a purchase, contrary to the logistic regression finding.

Another notable insight from this method is the significance of the "position" feature in influencing purchase decisions compared to other features. Logistic regression did not identify the position as relevant to purchases.

While the logistic regression method considers the retained customer segment as the second most important feature, it ranks as the ninth important feature in the decision tree method.

Considering business domain knowledge, customer review counts and scores are deemed crucial in purchase decisions. Logistic regression did not identify these features as relevant to purchases, but the decision tree method recognizes them as the tenth most important feature.

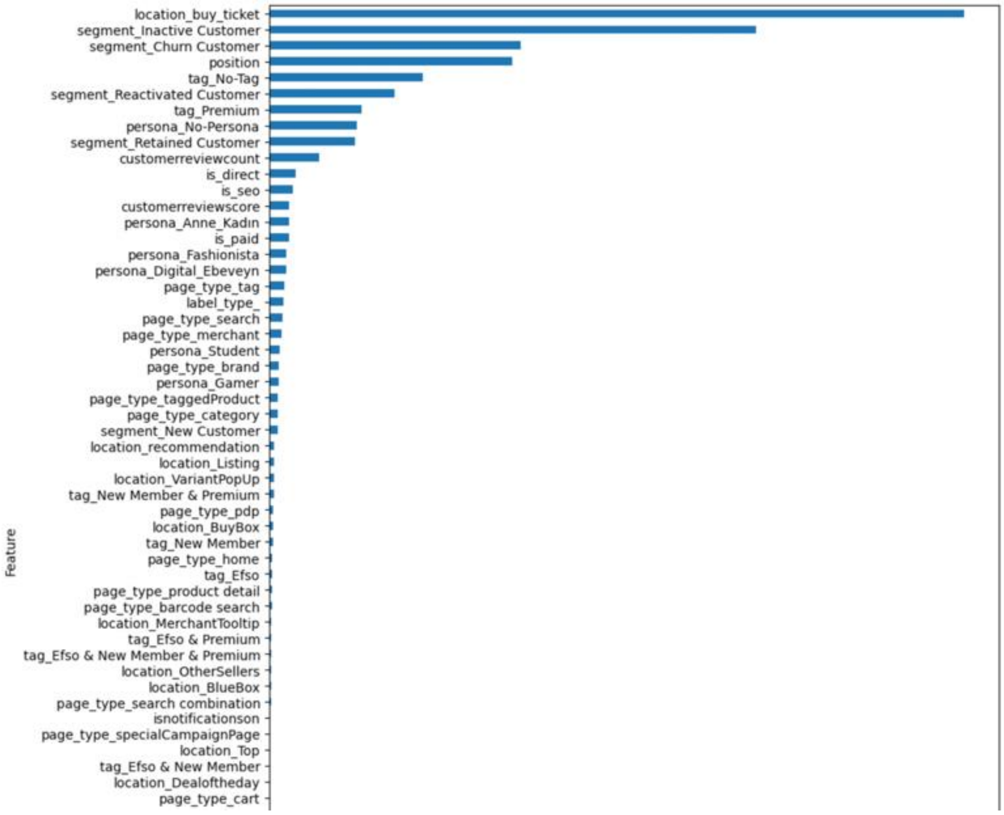

Figure 5.0.2 Decision Tree Results

5.3 Random Forest

The Random Forest method has identified several features that are strongly associated with the purchase decision: “position,” “is_price_discount,” “isnotificationson,” “customerreviewcount,” “customerreviewsscore”. These findings align well with business domain knowledge.

Notably, the product's position is highlighted as a critical factor. A higher product position enhances visibility to customers, increasing the likelihood of a purchase.

The results also underscore the significance of customer reviews in the purchase decision. Higher review counts and scores are correlated with a higher probability of customers making a purchase.

Another noteworthy finding, in line with business domain knowledge, is the impact of price discounts. In the e-commerce sector, discounts are widely recognized for their effectiveness in influencing purchase decisions.

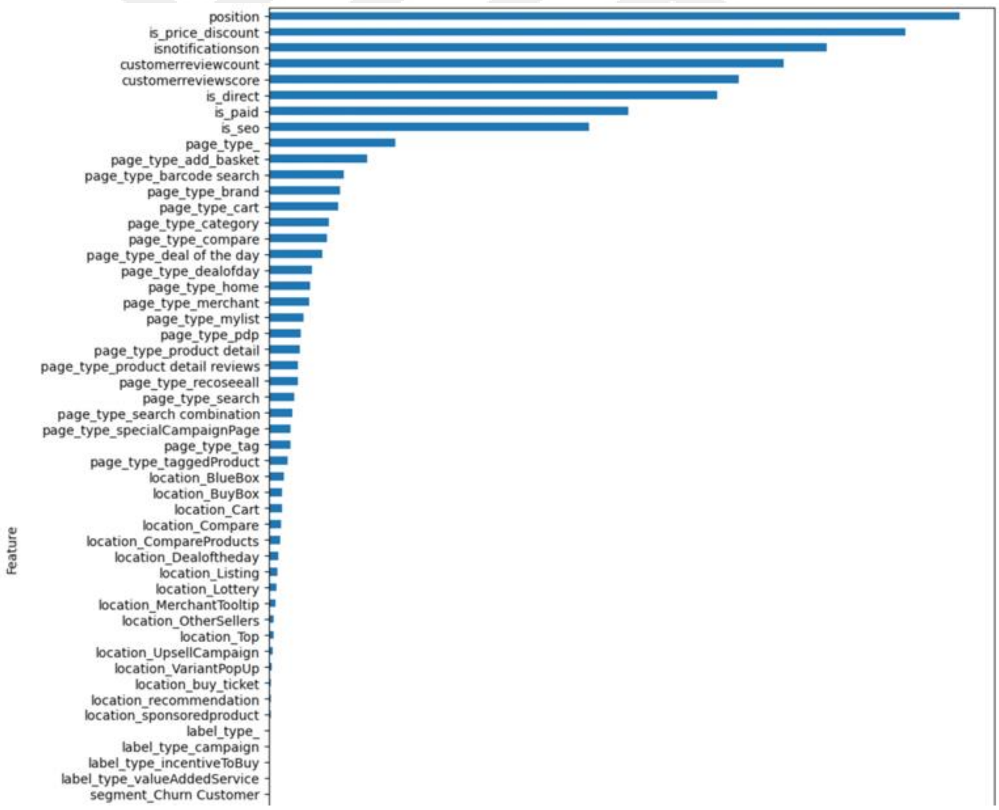

Figure 5.0.3 Random Forest Results

5.4 XGBoost

The XGBoost method produces results that closely mirror those of the Random Forest. Both methods identify the same features in the same order of importance. However, XGBoost provides clearer outputs. While Random Forest identifies additional features related to the order but with small importance ratios, XGBoost assigns most feature importances as zero.

Random Forest is a cluster of decision trees where each tree trains samples of the dataset simultaneously. The final decision is based on the prediction of the majority of decision trees. In contrast, XGBoost trains samples of datasets sequentially. Each model learns from its predecessor's mistakes and attempts to correct them. Our findings suggest that XGBoost presents much clearer results than Random Forest.

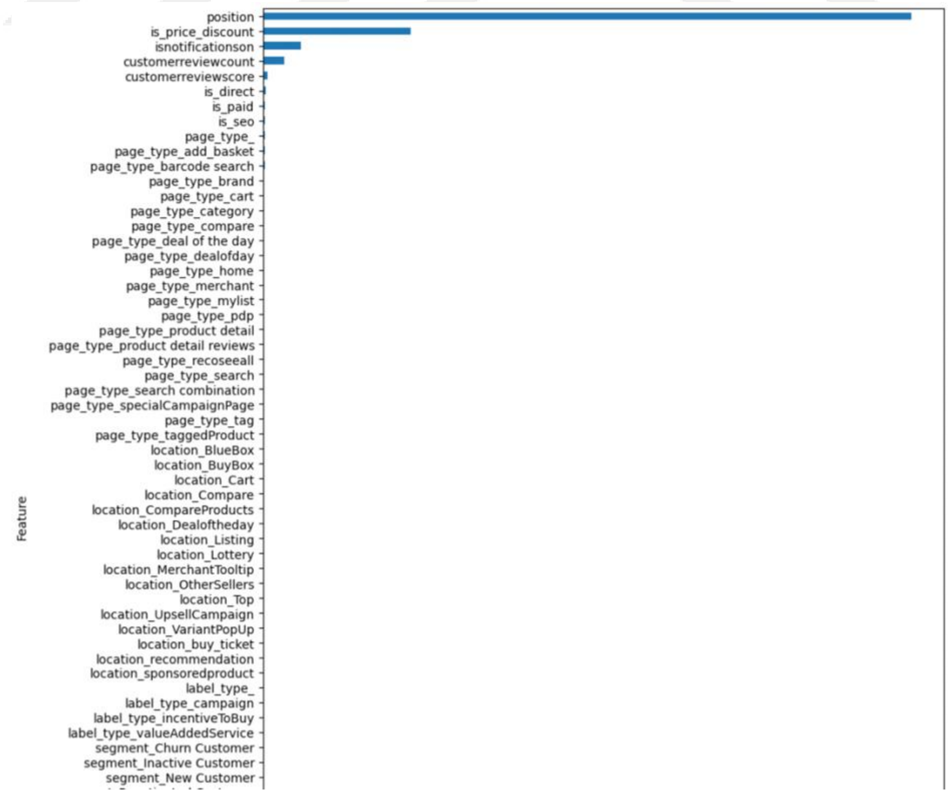

Figure 5.0.4 XGBoost Results

5.5 Permutation

The permutation method was applied to three different models: Random Forest, XGBoost, and Multilayer Perceptrons. Remarkably, permutation generated similar results across all these models.

Retained and reactivated customers emerged as the most important features, aligning with the findings of the logistic regression method. Interestingly, premium customers were identified as much more likely to purchase, a feature only defined as important by the permutation method. Furthermore, consistent with Random Forests and XGBoost methods, the position feature was also deemed important by the permutation method.

6. CONCLUSION

In this thesis, we embarked on an in-depth exploration of the influential factors, includes the terms of e-commerce domain or example e-commerce platform definitions, drive purchase decisions in the realm of e-commerce applications. We had to two objectives: first, to detect and rank these decisive features, and second, to compare various methods for feature importance assessment.

Our research journey took us through three distinct categories of feature importance methods, each offering a unique perspective on the significance of different features:

Coefficient-Based Methods: The first category of methods focused on quantifying the relationship between features and the binary target variable—whether a purchase was made or not. In this regard, we employed logistic regression, a well-suited tool for categorical binary target variables like ours. Coefficients obtained from logistic regression shed light on how a change in a feature impacts the likelihood of a purchase occurring. This method helped us uncover the initial insights into feature importance.

Tree-Based Methods: The second category delved into the power of tree-based methods, known for their interpretability. These methods assessed feature importance by examining how each feature contributes to reducing the impurity within decision trees. We employed a range of tree-based methods, including CART, Random Forest, and XGBoost. While CART provided insights through a single decision tree, Random Forest and XGBoost leveraged ensembles of decision trees. Interestingly, Random Forest and XGBoost yielded remarkably similar results, indicating the robustness of these methods in feature importance assessment.

Permutation-Based Methods: The third category introduced permutation-based feature importance methods, a unique approach where feature values were shuffled within the model. The increase in model error resulting from such shuffling indicated the importance of specific features. We applied permutation methods in conjunction with Random Forests, XGBoost, and Multilayer Perceptron classifiers. While permutation-based methods don't directly yield accuracy and F1 scores, they provided valuable insights into feature relevance.

The choice of using "add to cart" transactions as the focal point of our study was rooted in its significance as a clear indication of customer intent to make a purchase. These actions showcase a genuine interest in both the products and the online platform itself. "Add to cart" transactions can occur from any page with varying features, making them a rich source of data for analysis. Our research aimed to discern which factors within these transactions were most closely linked to purchase decisions.

The culmination of our study brought forth several noteworthy findings:

- Notably, Random Forest and XGBoost methods consistently generated analogous results, underlining their robustness in identifying feature importance.

- When we assessed the results of these methods according to business knowledge, it was evident that Random Forest and XGBoost most closely mirrored the expectations grounded in practical business experience.

- Although the permutation method doesn't provide its own accuracy and F1 scores, its insights were invaluable. In terms of model performance, tree-based methods consistently outperformed logistic regression.

- All methods generated similar accuracy and F1 scores.

- Although the Multilayer Perceptron is considered the most complex method, the decision-tree based methods achieved slightly higher accuracy.

- Tuning procedures were employed for decision-tree based methods as well as the multilayer perceptron method in an effort to optimize model parameters. However, these tuning efforts did not result in significant improvements in model accuracy and F1 scores.

- When the methods were applied to unbalanced data, the models yielded high accuracy results. However, the 'no purchase' scores were low, and the 'purchase' F1 score was high, indicating a significant imbalance in the scores. This imbalance suggested a bias towards the 'purchase' class. To obtain more accurate and unbiased results, we employed oversampling and undersampling methods to balance the data.

- Comparing the accuracy results with those found in the existing literature, our results appear to be relatively lower. This discrepancy can be attributed to the utilization of more realistic and complex data, which encompasses a multitude of features.

Table 6-1 Model Performances Comparison

Methods	Accuracy	No Purchase F1 Score	Purchase F1 Score
Logistic Regression	0.71	0.64	0.76
Decision Tree	0.74	0.69	0.78
Random Forest	0.73	0.67	0.78
XGBoost	0.74	0.68	0.77
MultiLayer Perceptron	0.72	0.67	0.75

In summary, our research endeavors have contributed valuable insights into the complex world of e-commerce purchase decisions. By employing diverse methodologies and rigorously comparing their results, we've provided a roadmap for e-commerce businesses seeking to understand and harness the most influential factors driving customer purchases.

These findings can empower businesses to fine-tune their strategies and optimize their platforms, ultimately enhancing customer satisfaction and competitiveness in the ever-evolving e-commerce landscape.

In closing, this thesis serves as a valuable resource for academics, researchers, and industry professionals alike, offering a comprehensive understanding of feature importance assessment and its application to e-commerce purchase decision analysis.

6.1. Position of the Study in the Literature

This section contextualizes our research within the broader landscape of existing studies in the field of feature importance analysis, particularly in the context of e-commerce platforms. Our investigation draws upon several key observations and gaps in the existing literature:

6.1.1. Methodological Diversity

A survey of the literature reveals a wide range of methodologies employed to assess feature importance in the context of e-commerce platforms. Some researchers have favored correlation-based methods, while others have turned to decision tree-based techniques or delved into the intricacies of deep learning methods. However, a notable absence in the literature is the utilization of permutation-based methods, which offer unique insights into feature relevance. Furthermore, no study to date has sought to comprehensively apply and compare all these methodological approaches within a single dataset.

6.1.2. Dataset Limitations

A recurring limitation in many prior studies is the relatively small size of the datasets used for analysis. Furthermore, a significant portion of these studies lacks comprehensive information regarding the data collection process. Such limitations can lead to the generation of potentially misleading results. In contrast, our study addresses this concern by utilizing a large and meticulously collected dataset, ensuring the robustness and reliability of our findings.

6.1.3. Dataset Diversity

Another prevalent trend in the literature is the repeated use of the same datasets across multiple studies. This practice, while common, may hinder the generalizability of findings and restrict the exploration of factors affecting customer purchase decisions. Our study sets itself apart by leveraging a diverse and extensive dataset obtained from a significant-scale e-commerce platform, promoting more concrete and realistic results.

6.1.4. Proposed Mixed-Methods Approach

In response to these identified gaps, we propose a mixed-methods study that aims to comprehensively investigate the factors influencing customer purchase decisions. Our research employs a large dataset acquired from transactions on a substantial e-commerce platform, offering a holistic view of customer interactions and behaviors.

6.1.5. Contribution to Knowledge and Practice

The primary objective of this study is to contribute to both academic knowledge and practical interventions in the realm of e-commerce. By shedding light on the intricate relationships between various factors, as informed by established business knowledge, and customer purchase decisions, our research seeks to advance the understanding of this vital area.

6.1.6. Conclusion and Expected Contributions

In conclusion, our proposed study endeavors to enrich the existing literature by comprehensively exploring the factors, includes terms of e-commerce domain or example e-commerce platform definitions, influencing customers' purchase decisions. By employing a large dataset and encompassing all relevant methodological approaches, our aim is to provide valuable insights that can inform both research endeavors and practical interventions in this pivotal domain. Through this study, we aspire to bridge the gaps in prior research and offer a nuanced understanding of the intricate dynamics governing customer purchase behavior in e-commerce.

6.2. Limitations and Future Work

Our study encountered several limitations, which are important to acknowledge:

6.2.1. Limitations

Our study encountered several limitations, which are important to acknowledge:

6.2.1.1. Limited Data Processing Scope

Due to constraints in individual computational infrastructure, this study primarily relied on customer "add to cart" actions as the basis for detecting decisive factors in purchase decisions. Future research should aim to expand the scope to encompass a wider range of customer actions and interactions within e-commerce platforms.

6.2.1.2. *Data Source Restrictions*

In this study, the data source was restricted to "add to cart" actions due to practical constraints. Future work should explore the possibility of accessing and analyzing a more comprehensive dataset, encompassing various customer interactions, to offer a more nuanced understanding of e-commerce behaviors.

6.2.2. Future Work

Building upon the findings and limitations of this study, several promising directions for future research emerge:

6.2.2.1. Comprehensive Customer Behavior Analysis

Future studies can extend beyond the confines of "add to cart" actions to encompass a holistic analysis of various customer behaviors and interactions within online e-commerce platforms. This broader scope would provide a more comprehensive view of the factors influencing purchase decisions.

6.2.2.2. Expanding to Online E-commerce Platforms

While our study focused on e-commerce features derived from business knowledge, future research could delve into more granular aspects of online e-commerce platforms. Investigating specific features unique to online platforms would further enrich our understanding and contribute to a more specialized body of knowledge.

6.2.2.3. Guiding Future Research

Our study serves as a pioneering endeavor in feature importance analysis for online e-commerce platforms. As prior literature primarily dealt with different domains, this study's approach offers a foundation for future studies in the e-commerce context. We anticipate that our findings will guide and inspire subsequent research endeavors in this domain.

6.2.2.4. Application of Study Results

The insights garnered from this study have implications beyond the boundaries of our research. In future studies, we intend to leverage the findings to inform and refine our analytical approaches. Additionally, these results can be applied to enhance our understanding of e-commerce behavior and influence decision-making processes in subsequent investigations.

In conclusion, while our study has provided valuable insights into feature importance analysis within the e-commerce domain, it is important to acknowledge its limitations. These limitations open the door to exciting opportunities for future research that can broaden our knowledge, encompass a wider array of customer interactions, and contribute to a deeper understanding of e-commerce behavior and decision-making.

7. DISCUSSION

In this section, we contextualize the results within the existing literature and research landscape, while also assessing their significance in both academic and business domains. Specifically, we examine how these findings can be leveraged to benefit business operations.

Our research questions pertaining to the academic domain focused on evaluating whether feature importance methods yield results that align with the dynamics of the business domain and effectively identify features influencing customer purchase decisions. In that section, we will discuss model performances according to their results.

Also, we will discuss performances of models in terms accuracy rate and F1 scores.

7.1. Performance of the Methods in Generating Results Aligned with Domain Knowledge

Our analysis reveals that Random Forest and XGBoost methods outperform logistic regression and MultiLayer Perceptron in identifying features that significantly impact purchase decisions. This discrepancy can be attributed to the inherent suitability of tree-based methods to our research context. Unlike regression-based techniques, such as logistic regression and multilayer perceptron, tree-based methods possess a distinct advantage due to their ability to discern decisive features amidst dispersed feature values across sessions. By employing Attribute Selection Measure (ASM) and segregating data into two classes based on the most influential features, tree-based methods streamline the identification process, thereby enhancing clarity in determining feature effectiveness.

To further evaluate the efficacy of regression methods, we conducted an experiment wherein we reduced the feature set to five variables: position, is_price_discount, is_notifications_on, customer_review_count, and customer_review_score. These features, identified as influential by Random Forest and XGBoost results alongside domain expertise, were assessed under logistic regression. Interestingly, while the results underscored the significance of customer review scores in purchase decisions, the effectiveness of customer review counts was not corroborated. This discrepancy contradicts domain knowledge, as the interplay between review scores and counts is deemed integral in influencing purchase decisions. Consequently, this experiment underscores the limitations of logistic regression in accurately capturing influential factors in purchase decisions.

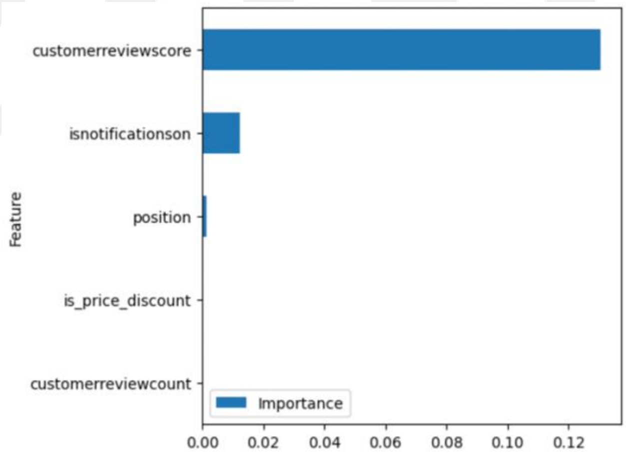

Figure 7.1 Logistic Regression Results with five feature dataset

Conversely, when we replicated the same experiment using the XGBoost method, the results mirrored those obtained using the entire feature set. Notably, the hierarchical importance assigned to review scores and counts aligns with domain expertise, indicating the efficacy of XGBoost in capturing nuanced relationships between features. Thus, it can be concluded that XGBoost demonstrates robust performance in this experiment, effectively aligning with domain knowledge.

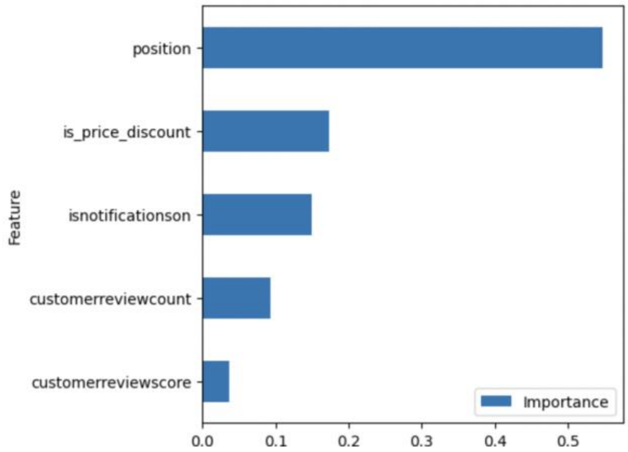

Figure 7.2 XGBoost Results with five feature dataset

7.2. Practical Suggestions for Business Domain

The findings underscore the paramount importance of product positioning in shaping customer purchase decisions within e-commerce settings. Positioning directly correlates with product visibility on the webpage, significantly influencing users' purchasing behavior. Products positioned within the view port, obviating the need for scrolling, exhibit higher purchase propensity. In light of this revelation, e-commerce managers are advised to strategically increase the number of products within the view port, prioritizing product cards over ad banners to optimize revenue streams. Such adjustments necessitate meticulous validation through A/B testing. Furthermore, personalized product listings based on users' shopping histories within the view port could further enhance order rates.

Equally pivotal is the prominence of price discounts in influencing customer purchase decisions. The visibility of discounts on product cards from the outset of user interaction is imperative. Data science methodologies validate the profound impact of discounts on customer behavior. Leveraging visual cues such as badges to highlight discounted products can effectively draw customer attention and bolster their likelihood of purchasing such items.

Moreover, the study underscores the significance of marketing notifications in driving customer orders, particularly among users who have granted permission to receive them. E-commerce enterprises are encouraged to invest in initiatives aimed at boosting notification permission rates. Incentivizing users with special offers in exchange for opening notification permissions holds promise for augmenting commission rates.

Additionally, the influence of product reviews on purchase decisions emerges as a pivotal finding. Customers seek assurance regarding product authenticity, quality, and delivery processes through reviews. The abundance of reviews instills confidence in prospective buyers and serves as a crucial determinant in their product selection process. E-commerce entities are urged to prioritize efforts to increase review counts for products, offering incentives and benefits to incentivize customers to submit reviews.

In summary, the insights gleaned from this study offer actionable recommendations for e-commerce managers seeking to optimize their platforms and enhance customer engagement and conversion rates. Strategic adjustments informed by these findings stand to significantly impact revenue generation and customer satisfaction levels.

7.3. Performance of the Methods in terms of Accuracy Rate

In the literature, accuracy rates typically range from 80% to 90%. We attribute this variability to the inherent complexity of our e-commerce case. The e-commerce sector is highly competitive, characterized by numerous factors influencing purchase decisions. Dynamic elements such as promotional campaigns and special discounts continuously shape customer behavior. Existing studies predominantly focus on sectors other than e-commerce, making our investigation the pioneering endeavor in this field.

Furthermore, beyond the scope of our dataset, various external factors impact customer purchase decisions. For instance, competitors' pricing strategies during similar time frames and evolving shopping trends contribute to fluctuations. Our data is collected from Turkey, where economic conditions are subject to constant change, thereby influencing customer purchase intent within this economic context. Consequently, there exist numerous determinants influencing order decisions beyond design components or customer segmentations.

Despite the complexity of our problem, we opted to streamline our dataset to assess whether simplicity would yield improved accuracy rates. Employing the XGBoost method, we restricted our dataset to five key features: position, is_price_discount, isnotificationson, customerreviewcount, and customersreviewscore. These features were identified as the top five correlates with orders based on our comprehensive dataset. We hypothesized that utilizing only these highly correlated features might enhance our accuracy rate. However, contrary to our expectations, our accuracy rate decreased to 60%, compared to the 73% achieved when utilizing the entire dataset. This decline suggests that reducing the number of features may result in the loss of crucial information pertinent to influencing order decisions.

REFERENCES

Baati K., Mohsil M. (2020). Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest, Artif. Intell. Appl. Innov. AIAI 2020. IFIP Adv. Inf. Commun. Technol.
URL: https://link.springer.com/book/10.1007/978-3-030-49161-1.

Banerjee A. (2022). Hyperparameter Tuning Using Randomized Search,
URL: https://www.analyticsvidhya.com/blog/2022/11/hyperparameter-tuning-using-randomized-search/

Cezar A. and Öğüt H. (2016). Analyzing conversion rates in online hotel booking, International Journal of Contemporary Hospitality Management (2016).
URL: https://www.emerald.com/insight/content/doi/10.1108/IJCHM-05-2014-0249/full/html

Diamantaras K., Salampasis M., Katsalis A., Christantonis K. (2021). Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks, 10thInternational Conference on Data Science, Technology and Applications.
URL:https://pdfs.semanticscholar.org/99ea/7a6a1c7aa63d752a8f33efa0128b7c36b330.pdf

Donges N. (2023). Random Forest: A Complete Guide for Machine Learning,
URL: https://builtin.com/data-science/random-forest-algorithm

Fatta D., Patton D. (2018). The determinants of conversion rates in SME e-commerce websites,Journal of Consumer and Retailing Services .

Hu X., Yang Y., Chen L., Zhu S. (2023). Research on a Prediction Model of Online Shopping Behavior Based on Deep Forest Algorithm, International Conference on Artificial Intelligence and Big Data .
URL: https://ieeexplore.ieee.org/abstract/document/9137436

McDowell W., Wilson R. (2016). An examination of retail website design and conversion rate,Journal of Business Research .
URL: https://www.sciencedirect.com/science/article/pii/S014829631630203X

Nazir S., Khadim S., Asadullah M., Syed N. (2023). Exploring the influence of artificial intelligence technology on consumer repurchase intention: The mediation and moderation approach, Technology in Society .
URL: https://www.sciencedirect.com/science/article/pii/S0160791X22003311

Saini A. (2021). Conceptual Understanding of Logistic Regression for Data Science Beginners,
URL: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

Saini A. (2021). Decision Tree Algorithm – A Complete Guide,
URL: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

Sakar C.O., Polat S.O., Katircioglu M., Kastro Y. (2019). Real- time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks.
URL: https://link.springer.com/article/10.1007/s00521-018-3523-0

Sheil H., Rana O.,Reilly R. (2018). Predicting purchasing intent: Automatic Feature Learning using Recurrent Neural Networks, SIGIR eCom.
URL: https://arxiv.org/abs/1807.08207.

Teye M., Missah Y.M. (2020). Investigating Factors that Affect Purchase Intention of Visitors of E-commerce Websites Using a High Scoring Random Forest Algorithm, International Journal of Engineering and Technical Research (2020). **URL:**https://www.researchgate.net/publication/351840990_Investigating_Factors_that_Affect_Purchase_Intention_of_Visitors_of_E-commerce_Websites_Using_a_High_Scoring_Random_Forest_Algorithm

Topal I. (2019). Estimation of Online Purchasing Intention Using Decision Tree, journal of management and geonomics research.
URL: https://dergipark.org.tr/en/pub/yead/issue/50655/542249.

PUBLICATIONS

- Title: "Defining the Decisive Factors on Purchase and Comparing Feature Importance Methods"
  Authors: Erman Demir, F. Serhan Danis
  Publication Source: International Conference of Advanced Technologies
  Publication Date: August 17-19,2023
  DOI: https://doi.org/10.58190/icat.2023.18

- Title: "Financial Applications of Graph Neural Networks"
  Authors: Erman Demir, Gülfem Işıklar Alptekin, Özlem Durmaz İncel
  Publication Source: Technology Analysis, Fintech and Financial Services
  Publication Date: March 11-13,2022