EXPLORING CUSTOMER SEGMENTATION IN FASHION E-COMMERCE THROUGH MACHINE LEARNING TECHNIQUES

A THESIS

SUBMITTED TO ABDULLAH GÜL UNIVERSITY

SOCIAL SCIENCES INSTITUTE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

By

Nazlınur Madenoğlu

April, 2025

Kayseri, Türkiye

Nazlınur Madenoğlu

EXPLORING CUSTOMER SEGMENTATION IN FASHION E-COMMERCE THROUGH MACHINE LEARNING TECHNIQUES

AGU 2025

EXPLORING CUSTOMER SEGMENTATION IN FASHION E-COMMERCE THROUGH MACHINE LEARNING TECHNIQUES

A THESIS

SUBMITTED TO ABDULLAH GÜL UNIVERSITY

SOCIAL SCIENCES INSTITUTE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

By

Nazlınur Madenoğlu

April, 2025

Kayseri, Türkiye

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: Nazlınur Madenoğlu

Signature :

REGULATORY COMPLIANCE

M.S. thesis titled Exploring Customer Segmentation In Fashion E-Commerce Through Machine Learning Techniques has been prepared in accordance with the Graduate Thesis Preparation Guidelines of the Abdullah Gül University, Social Sciences Institute.

Prepared By	Advisor
Nazlınur Madenoğlu	Associate Professor Faruk Güven
Signature	Signature

Head of the Data Science Program

Umut Türk

Signature

ACCEPTANCE AND APPROVAL

M.S. thesis titled Exploring Customer Segmentation In Fashion E-Commerce Using Machine Learning Techniques and prepared by Nazlınur Madenoğlu has been accepted by the jury in the …………………………………….. Graduate Program at Abdullah Gül University, Social Sciences Institute.

……….. /……….. / ………..

(Thesis Defense Exam Date)

JURY:

Advisor : (Name Surname)……………………………............................... ….

Member : (Name Surname)……………………………… ....

Member : (Name Surname)……………………………… ....

Member : (Name Surname)……………………………… ....

Member : (Name Surname)……………………………… ....

APPROVAL:

The acceptance of this M.S. thesis has been approved by the decision of the Abdullah Gül University, Social Sciences Institute, Management Board dated ….. /….. / ……….. and numbered ………….……. .

….. /….. / …

(Date)

Director of Social Sciences Institute

Name Surname : Nazlınur Madenoğlu

Program : Data Science (Master of Science)

Advisor : Associate Professor Faruk Güven

Thesis Title : Exploring Customer Segmentation In Fashion E-Commerce Through Machine Learning Techniques

Date and Place : April, 2025 – Kayseri, Turkey

ABSTRACT

In today's world where technology is developing very rapidly, internet usage is also increasing proportionally. This change has revealed that brands attach importance to the sector. The significance of e-commerce is to the advantage of brands because there have been decreases in some fixed expenses of companies. With the increase in online shopping, personal analyses of customers can also be made by customer relationship management (CRM). It is necessary to divide customers into segments for customer-oriented marketing. Customer segmentation is a widely used form of analysis. There is an increasing demand to develop a deep awareness of individual customer needs and desires. Segmentation, a commonly utilized method for achieving this understanding, has undergone continuous refinement in recent years. This study targets to present a detailed analysis of various segmentation approaches and their evolution. In this study, RFM (Recency, Frequency, Monetary) analysis was used for segmenting the customers. Customers were divided into segments by scoring them on the last shopping time, shopping frequency and total spending. Four customer groups were created with K-means and the values of each segment were analyzed. Churn rate analysis determined customers who did not shop for 90 days as lost. Churn estimation was performed with the LightGBM model using the machine learning technique. In addition, the Predictive CLV (customer lifetime value) model was developed using the machine learning technique Ridge Regression. The accuracy rate was increased and low, medium and high CLV segments were created. As a result; RFM , K-means and CLV estimation were used to optimize customer relationships and increase revenues. E-commerce data of a private brand was analyzed using machine learning techniques. Nowadays, there is an increase in computing power and rapid developments in machine learning/artificial intelligence algorithms. This has recently enabled the application of more advanced techniques.

Keywords: Customer Segmentation, K-Means, RFM Analysis, Churn Prediction, Customer Lifetime Value

Ad Soyad : Nazlınur Madenoğlu

Anabilim Dalı, Program : Veri Bilimi Yüksek Lisans

Tez Danışmanı : Associate Professor Faruk Güven

Tez Başlığı : Exploring Customer Segmentation In Fashion E-Commerce Through Machine Learning Techniques

Tarih ve Yer : April, 2025 – Kayseri, Türkiye

ÖZET

Teknolojinin çok hızlı geliştiği günümüzde internet kullanımı da orantılı olarak artmaktadır. Bu değişim markaların e-ticaret sektörüne önem verdiğini ortaya koymuştur. E-ticaretin önemi markaların lehinedir çünkü şirketlerin bazı sabit giderlerinde azalmalar olmuştur. Online alışverişin artmasıyla birlikte müşterilerin kişisel analizleri de yapılabilmektedir. Müşteri ilişkileri yönetimi (CRM) önem kazanmıştır. Müşteri odaklı pazarlama için müşterileri segmentlere ayırmak gerekmektedir. Müşteri segmentasyonu yaygın olarak kullanılan bir analiz biçimidir. Her bir müşterinin ilgi ve motivasyonlarını derinlemesine anlamak için artan bir talep vardır. Bu anlayışı elde etmek için yaygın olarak kullanılan bir yöntem olan segmentasyon son yıllarda sürekli olarak iyileştirilmektedir. Bu makale çeşitli segmentasyon yöntemlerinin ve bunların mevcut gelişim durumlarının iyi yapılandırılmış bir genel görünümünü sunmayı amaçlamaktadır. Bu çalışmada müşteri segmentasyonu için RFM (Recency, Frequency, Monetary) analizi kullanılmıştır. Müşteriler son alışveriş zamanı, alışveriş sıklığı ve toplam harcamalarına göre puanlanarak segmentlere ayrılmıştır. K-Means ile dört müşteri grubu oluşturulup her bir segmentin değerleri analiz edilmiştir. Churn oranı analizi ile 90 gün boyunca alışveriş yapmayan müşteriler kayıp olarak belirlenmiştir. Churn tahmini, makine

öğrenmesi tekniği kullanılarak LightGBM modeli ile yapılmıştır. Ayrıca, Ridge Regresyonu makine öğrenmesi tekniği kullanılarak Tahmini CLV modeli geliştirilmiştir. Doğruluk oranı artırılarak düşük, orta ve yüksek CLV segmentleri oluşturulmuştur. Sonuç olarak, müşteri ilişkilerini optimize etmek ve gelirleri artırmak için RFM analizi, K-Means ve CLV tahmini kullanılmıştır. Özel bir markanın e-ticaret verileri makine öğrenmesi teknikleri kullanılarak analiz edilmiştir. Günümüzde, hesaplama gücünde artış ve makine öğrenmesi/yapay zeka algoritmalarında hızlı gelişmeler yaşanmaktadır. Bu durum son zamanlarda daha gelişmiş tekniklerin uygulanmasına olanak sağlamıştır.

Anahtar kelimeler: Müşteri Segmentasyoıu, K-Means, RFM Analizi, Kayıp Müşteri Tahmini, Müşteri Yaşam Boyu Değeri

ACKNOWLEDGEMENT

I want to thank Assoc. Prof. Dr. Faruk Güven, who has supported and given me knowledge in this study,

My mother Dr. Nermin Turan and my father Dr. Harun Turan, who have supported me throughout my life,

My husband Orhan Madenoğlu, who has supported my work both materially and spiritually, and my sons Halit Madenoğlu and Turan Madenoğlu, who have given me strength.

TABLE OF CONTENT

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AOV	Average Order Value
CLV	Customer Lifetime Value
CRM	Customer Relationship Management
F	Frequency
GBDT	Gradient Boosting Decision Tree
KDE	Kernel Density Estimate
LightGBM	Light Gradient Boosting Machine
M	Monetary
MAE	Mean Absolute Error
ML	Machine Learning
R	Recency
RFM	Recency, Frequency, Monetary
RMSE	Root Mean Square Error
ROI	Return of Investment

LIST OF TABLES

LIST OF GRAPHS

1. INTRODUCTION

Digital changes have accelerated after the pandemic. The e-commerce sector has become the most important platform for shopping. Consumer habits have changed with the pandemic. Online shopping has come to the fore. It has become easily accessible everywhere on the internet. This has largely shifted the traditional shopping method to online shopping. This change has revealed the need for companies to develop customer-oriented strategies. Companies have understood that the e-commerce sector will become important and have shifted their strategies to customers in this direction.

In particular, in this products sector places great emphasis on ensuring repeat purchases, building brand loyalty, and effectively managing customer relationships. These factors are critical for gaining a competitive edge in this highly competitive environment.

In this study, customer segmentation was performed using machine learning techniques for a company operating in the e-commerce fashion sector. Consumer behavior has also changed in the digital world. For this reason, traditional segmentation methods may also be inadequate. Consumer behavior, shopping habits, purchase history and demographic data form the basis for analysis and are very important. Recency, Frequency, Monetary (RFM) analysis was used instead of classical segmentation methods. In this way, unique customer groups are created. The research methodology includes pre-processing and evaluation of customer data sets. The findings resulting from these analyses aim to show how machine learning methods can increase the effectiveness of customer segmentation for companies selling on e-commerce platforms.

In sensitive and highly competitive markets, such as the fashion e-commerce sector, employing these advanced methods can give companies a significant edge. As a result of this study, marketing strategies adapted to company data will be developed and valuable information will be created in the sector. The e-commerce company selling fashion products will better understand customer demands and needs. Customer satisfaction will increase and customer loyalty will also increase. Innovative approaches will help these businesses remain competitive and maintain strength in an intensely challenging market.

Customer segmentation can be performed in fashion e-commerce using machine learning techniques. This segmentation helps to better understand customer behavior. It can enable to increase customer value and develop personalized marketing strategies.

How to perform customer segmentation and increase customer value in fashion e-commerce using RFM analysis, K-Means and LightGBM?

Machine learning techniques aim to analyze customer behavior more accurately by extracting meaningful information from large data sets. This enables more effective segmentation. While traditional segmentation methods are usually based on demographic data in the modern e-commerce environment, customers' past purchasing behavior, spending habits and interaction levels are much more decisive. At this point, RFM analysis has offered a powerful approach for customer segmentation.

RFM analysis evaluates the relationship between customers and the business through the variables Recency (recency), Frequency (frequency) and Monetary (spending) and creates customer groups. Recency means that the time the customer last made a purchase and shows how close the customer is to the business. Frequency measures the total buying made by the customer over a given time. Monetary shows the customer's total spending. The combination of these three variables better describes the value of customers and enables the development of personalized marketing strategies.

In this research, RFM analysis was performed for customer segmentation and then customer groups were created using the K-means. The K-means is a commonly known ML method to determine customer groups. During the clustering process, the Elbow method was utilized for calculating the most appropriate total clusters.

Churn (customer churn) estimation was also performed to understand how the obtained segmentation results can be used by businesses. Customer churn is a big problem for companies, and given the expense of bringing in new customers, keeping current customers is a main strategic objective. In order to keep customers, it's necessary to know the central drivers of loyalty and predict when they will leave so that effective retention efforts can be planned (Manzoor et al., 2024).

Maintaining existing customers is less expensive than bringing in new ones, so online retailers should focus more on retaining their current customer base. (Gadgil et al., 2023). Therefore, customer churn estimation was performed using the LightGBM algorithm and customers who could potentially leave the business were identified.

However, Ridge regression was used to forecast CLV. CLV identifies the overall revenue to be gained from a customer by a company in the long term and is thus a vital element of customer management practices. The success of mass application of machine production techniques in practice has increased the value of such techniques in CLV estimation and loss estimation (Bauer et al., 2021). Thanks to correctly estimated CLV values, businesses use their marketing budgets more efficiently.

The literature reveals a lack of research applying CRM, RFM, CLV, Churn Prediction, and K-Means analysis together in this sector, making this research unique and a potential model for other businesses.

The primary point of the research is to show how ML-supported consumer segmentation can deliver value to businesses. The findings will make customer segmentation more efficient, along with aiding businesses in devising strategies to prevent customer churn and attain optimum customer lifetime value. Especially in the business of fashion e-commerce, understanding customers' requirements better and providing them with personalized offers will be highly valuable in a highly competitive

Therefore, the present study considered how machine learning techniques can be utilized to enable e-commerce firms to become more effective in managing their customers. Techniques such as K-means, RFM analysis, churn estimation, and CLV calculation have put firms in a position to create customer-centric strategies. According to the data obtained, it was concluded that companies operating in the fashion e-commerce sector can increase customer loyalty and revenues by managing customer segmentation more consciously.

2. LITERATURE REVIEW

Retailers collect and analyze massive amounts of data daily, offering opportunities for growth, innovation, and improved customer relationships. However, many companies struggle to leverage this data effectively for marketing, customer acquisition, and personalized recommendations in e-commerce. Customer segmentation lays the groundwork for product development and service decisions, making it essential for effective personalization strategies (Griva et al., 2024). Customer segmentation is important to increase marketing strategies and increase online personal shopping rates. Leveraging segmentation helps retain customers while enabling the company's marketers to gradually become more persuasive than their competitors (Jethva et al., 2024). It offers the firm a platform for making sound judgments on market activities (e.g., running campaigns, creating production plans, and choosing suppliers) while managing limited resources. Performing inappropriate customer segmentation fails to meet customer expectations, and this may result in poorly constructed market simplification (Sun et al., 2021).

An accomplished personalized marketing strategy in e-commerce depends on advanced data analytics solutions to gather and process information regarding customers, their tastes, and needs, create customized content, segment audiences, and continuously analyze and optimize strategies for maximum impact (Mykhalchenko et al., 2023). For marketing strategies and improving service personalization, segmenting customers based on their behavioral traits is becoming more important (Nugroho et al., 2024).

There are many varied algorithms that a person can use while performing customer segmentation. They also involve other approaches such as association algorithms, clustering algorithms, classification algorithms, and regression algorithms (Rungruang et al., 2024). These algorithms are usually based on the customer purchase's history. Most of these research have centered on analyzing buying patterns and behaviors in order to establish distinct customer segments for companies (Kasem et al., 2024).

2.1 Customer Relationship Management

When the focus is customer segmentation, it becomes important to recognize the importance of Customer Relationship Management (CRM). CRM means that set of technology used for the management of the customer relationship, customer experience development, and extended customer loyalty. Machine learning technique can evaluate the behavior of a customer based on data from the CRM system so that they could deliver improved customer satisfaction and shape long-term associations.

The management style that aims to upset the business processes and technologies a company builds up to manage customer behavior is called customer relationship management (Li et al., 2021).

It is a method of understanding the behavior of customers based on the best communication for customer retention, loyalty, addition of new customers and profitability. The goals of CRM are to spread customers and keep this situation to the highest degree for the company throughout its lifecycle (Khajvand et al., 2011). Marketers who use customer segmentation have now understood the importance of customer relationship management even better. Through the use of customer segmentation, marketers can focus more on customer relationship management that is not possible with existing mass marketing methods.

CRM aims to have a better understanding of particular customers, such as variations in loyalty behavior, and are typically measured through segmentation. The concept aims for segmentation of a marketplace into distinct groups of behavior, demographics using varied indicators such as frequency of purchases, average time elapsed between purchases, duration of activity, and basket size (Lloyd et al., 2018). For being a successful business, retaining new customers are not only the vital objective, but also ensuring profitability from each individual customer is critical (Jethva et al., 2024).

A successful web-based CRM can assist in enhancing the web-based customer interactions of a company along with offering opportunities to enhance revenue by creating an e-commerce strategy and rebuilding its website (Wu et al., 2011). CRM encompasses sales, marketing, service, and operations functions, and the cross-functional nature of CRM has led to fragmented AI-CRM studies across business disciplines. In light of this context, it seems appropriate for business and academic communities alike to distill the existing literature on AI in CRM into a unified corpus of knowledge that can guide managers and inspire subsequent research for academics (Ledro et al., 2022).

2.2 K-Means

K-Means technique has been one of the most frequent segmentation techniques among the researchers since it accommodates clustering consumers in data-rich and diversified settings, allowing for better perception of their preferences (Feng et al., 2024). For making effectively marketing performance, it is very important to analyze customer segmentation alongside the buyer's advantages. These two aspects are combined in a step-by-step approach, but achieving optimal integration remains a challenge. To tackle this, the authors introduced the K-classifiers segmentation algorithm (Kasem et al., 2024).

K-classifiers are techniques of clustering. The use of ML used to cluster data points with similar data in a dataset is called k-means analysis. The algorithm will group the data into k clusters. It is a metric-based approach and groups the data numerically into different clusters (Syakur et al., 2018). The primary goal of clustering is to ensure meaning that data points in the same cluster possess high similarity and minimum similarity with other clusters (Sarkar et al., 2024). The second goal of K-means clustering is to make points within a cluster as similar as possible and as distinct as possible from each other. This results in well-defined clusters that have similar points.

K-Classifiers prioritize allocating more resources to customers who generate higher returns for the company (Christy et al., 2021). Among the many data mining techniques, clustering is highly significant in grouping similar data points together. K-Means Clustering is highly popular because of it is very simple and effective (Sulianta et al., 2024).

The clustering method has been useful in grouping data based on similar characteristics, offering deeper insights into purchasing behavior. This helps companies develop more targeted and impactful marketing strategies (Nugroho et al., 2024). With the help of K - Means Clustering on e-commerce data based on the internet, businesses can identify clusters of customers with similar attributes, so they can more effectively customize marketing tactics and product portfolios accordingly. Based on literature review, The k-means approach is one of the best techniques used for segmentation of customers. Best cluster counts are achieved through the Elbow Method. This technique graphs WCSS and optimal clusters can be calculated where this point starts to decrease (Sulianta et al., 2024).

K-means analysis is a straightforward and fast dataset clustering algorithm. To cluster based on RFM scores, K-means analysis ought to be used after conducting RFM analysis. Using K-means analysis in conjunction with RFM, the customer segmentation result is more accurate, allowing for the development of Better-segmented branding approaches.

2.3 RFM

The Recency, Frequency, Monetary (RFM) is a quantitative approach to calculate customer value by determining key indicators. According to this model, customer value can be quantified, segmented, and ranked effectively (Feng et al., 2024). RFM analysis is a marketing technique to determine customer value, and it relies on three primary criteria.

Customer segmentation is one of the primary approaches utilized in the stage of acquiring customers to identify and select customers for each segment. The analysis of RFM is then employed to highlight and focus on characteristics based on three key variables: Recency (R), Monetary (M), and Frequency (F) (Sarkar et al., 2024). Recency is the timeframe since the previous purchase by the consumer. Frequency is a count of the number of times the customer purchased something within a given period. Monetary is a count of all amounts spent by the customer during the period (Alves Gomes et al., 2023). If the R value decreases, it highlights that the customer is active and thus an active customer. A rise in the F value indicates that the customer repeats purchase. As the M value increases, it indicates that the customer is spending more and thus is more valuable. Decreases in these values would place the customer in lower-value segments. It employs variables R, F, M to divide customers and calculate scores by splitting the customers. Scores within segmentation are employed together with clustering methods such as k-means to define customer sets (Barrera et al., 2024).

RFM model is extensively used in segmenting consumers in targeted marketing due to a variety of reasons. To start with, it is cheap to acquire and quantify basic customer behavior analysis. Because customer details and buying history can be stored in an easily accessible web-based database. Secondly, The model can summarize buying behavior using a few simple factors. Third, the model can make a firm more profitable in the short term by accurately predicting customer responses and it can be empowering the firm to manage customer inquiries and activities better (Chouaten et al., 2024). Using the RFM models that segmentate customers into three important measures, e-commerce businesses have a clearer picture of their segments of customers. They can implement specifically tailored strategies for every group on response to the information, resulting in customers being satisfied and loyal.

The RFM is an essential technique for evaluating consumer’s potential. Customer segmentation is used in a lot of fields, including sales, marketing, and operations management. (Sivaguru et al., 2024). The RFM model is typically used in consumer evaluation, researchers have adapted this model to various contexts. (Wu et al., 2021).

Customer segmentation, churn forecasting, and marketing strategies can be formulated using RFM analysis. Special campaigns can be formulated for customers based on scores from the analysis. As long as the data is fresh and up to date, RFM analysis is simple and data-driven. RFM is particularly a strong tool in the e-commerce sector, promoting marketing efficiency and customer relationship management.

2.4 Churn Prediction Analysis

After segmenting customers, churn analysis is conducted to define customers at threatened of leaving. Churn prediction analysis is a tool businesses use to estimate the risk of losing their existing customers. With the increasing importance of predicting and preventing churn across different industries, companies are using various machine learning techniques. After creating an accurate model to predict churn, it plays a vital role in businesses to focus on their customers to prevent potential risks (Suh Y, 2023). Customer churn prediction models have proven to be highly valuable beyond the telecommunications industry, especially in sectors like digital marketing, e-commerce, and banking, where understanding and reducing churn is just as important (Poudel et al., 2024). It is a crucial analysis in highly competitive sectors, such as fashion, where customer loyalty tends to be low. Churn analysis is the type of analysis that investigates the reasons behind customer losses for a business over a specific period.

Customer churn prediction analysis use individual data to forecast assessing the likelihood of customers terminating their engagement with a company within a certain time frame (Mena et al., 2024). Machine learning techniques are used in churn rate analysis.

To study the core cause of customer churn and retain existing customers, various researchers have deployed integrating machine learning algorithm usage to create efficient predictive models. Among these, methods are typically utilized for prediction and classification in this regard (Peng et al., 2023). LightGBM is an efficient open-source machine learning technique utilized in classification, regression, and ranking. Unlike common level-wise methodologies, LightGBM takes the leaf-wise growing strategy, minimizing the tree's decision nodes while enhancing its power of generalization (Yamini et al., 2024). The main role of machine learning to customer retention is to observe and predict customer churn based on behavioral changes. Machine learning allows companies to define consumers whose can be most presumably to go away (Singh et al., 2024). This assessment is intended to develop a customer loyalty. It is important in increasing revenue since developing strategy in focused form for at-risk customers is more desirable than acquiring new customers.

In conclusion, churn prediction analysis helps businesses predict the risk of losing their existing customers. Churn prediction analysis will typically identify early enough whether the customers go back to the business within a predetermined time limit. It will often factor in customer behavior, frequency of purchase, and spending trends. Based on this data, it determines which customers will churn. With it, businesses are able to design targeted marketing campaigns for at-risk customers, increase customer loyalty, and stem revenue loss. It is also related to customer lifetime value since it keeps those customers who will generate revenue in the long run.

2.5 Customer Lifetime Value (CLV)

CLV allows one to understand the total revenue a consumer adds to a company. CLV is a tool that allows for an estimation of a consumer’s worth relationship with a business. It is especially significant when designing marketing schemes and making campaign designs in an attempt to promote customer loyalty with the progress of data mining techniques and computing power improves. Customer lifetime model researchers improve the functionality, accuracy and applicability of numerous models with machine learning techniques (Sun et al., 2023).

Estimated CLV is a technique for computing the future value a consumer will bring to a firm. Companies use Ridge Regression, an advanced machine learning technique, to compute this. This is a regularized linear regression. Ridge regression addresses common problems such as multicollinearity (the features in the data are strongly correlated) and overfitting (the model is strongly reliant on the training data). With Ridge Regression, businesses are able to elevate the precision and reliability of their CLV predictions. Retail data sets usually have strong relationships between variables such as marketing expenditures and promotions. Ridge Regression reduces the regression coefficients and it provides more stable estimates and addresses the problems by reducing the effects of multicollinearity (Zhang, 2024).

CLV is the earning businesses expect to receive from a single customer and is a value-based measurement that captures value on each purchase on the first purchase (Ali et al., 2024). CLV measurement creates the data based on customer purchasing behavior and patterns, which in turn creates data for analysis for business development and long-term decision making (Elveny et al., 2024).

If you want to determine strategic resources, CLV calculations provides useful information for businesses like conducting marketing activities, and seeing the long-term value of customers (Firmansyah et al., 2024).

CLV approach is of very high importance in long-term evaluation processes as it helps determine the segments where the customers have higher value by using the information obtained. In this way, it allows companies to set the appropriate strategy for activities in the framework of CRM (Jasek et al., 2018). CLV which is executed as per customer segmentation allows a holistic approach to be followed by segmenting customers based on common attributes. Each of these segments is then assigned a specific CLV based on its individual buying patterns and revenue prospects (Ali et al., 2024). The CLV model must be developed using either traditional methods or ML methods based on the company's data. These techniques offer different advantages in terms of accuracy and usability. The appropriate method can be selected depending on the available data.

This literature review provides information on how businesses operating in the fashion online retail sector can apply RFM, CLV, Churn Prediction and K-Means analysis to achieve deeper insights into their customers and make better decisions. At its core, CRM is simply the build-up of ongoing relationship with consumers, fueling loyalty, and reducing churn. RFM analysis makes this possible by dividing customers into multiple segments depending on how they buy. Firms can create personalized marketing strategies with such information that address one segment at a time, thus guaranteeing.

In addition, K-Means clustering analysis segments customers according to shared characteristics and, upon incorporation with the utilization of RFM and CLV analyses, becomes more precise in segmenting customers, thereby providing the business with strategic insights. In churn prediction, the risk of customer loss is identified, and it is observed whether keeping a current customer is cheaper than a new consumer.

In conclusion, these approaches enable a fashion e-commerce firm to manage customer relationships more effectively, gaining a competitive edge. This study aims at fostering long-term customer loyalty and may serve as an example for other e-commerce companies. The methodology of this study will be explained in the next section.

3. MATERIAL AND METHOD

3.1 Problem Definition

In today's e-commerce environment, customer relationship management and customer satisfaction are the key for companies to achieve sustainable success. For an e-business company operating in the fashion industry, a customer behavior study, proper customer segmentation, and strategic decision-making based on these studies are imperative.

With the information of current customers and company sales history, there is a possibility of creating customer loyalty, preventing customer churn, and converting current customers into more loyal ones. Nevertheless, extracting insights from the information and determining proper strategies require use of advanced techniques of data analysis.

The level of competition and consumer behavior complexity within the fashion e-commerce sector are very high. Effective customer segmentation has made it necessary to establish a marketing plan. The critical issues in the sector are named as follows:

- The need to develop marketing campaigns based on customers due to the heterogeneous structure of customers

- Low ROIs (Return on Investments) due to untargeted marketing campaigns

- The need to reduce customer churn rate

- Trying to gain new customers instead of customers who will leave is more costly.

- The need to make a strategic marketing plan by estimating long-term customer value (CLV).

- Identifying profitable customer groups and making more investments in this area

The purpose of the research is to categorize the customers in terms of behaviors through RFM analysis initially, classify the customers through the K-Means clustering method secondly, minimize the customer churn risk based on customer churn prediction thirdly, and identify their lifetime values to the business by calculating the CLV.

This problem enables the firm to understand its customers better and make more effective customer relationship management. Lastly, in a competitive sector, the objective is to enhance customer satisfaction, prevent customers from leaving, and retain loyal customers in the long term.

3.2 Data Information and Preparation

3.2.1 Data Information

Customer movement data of a company that specializes in the online fashion retail industry was used for this study. Customer buying habits are part of the information in this data. These data were organized to be used in customer segmentation, churn prediction and CLV calculation processes by analyzing the shopping behavior of the customers. The labels of the variables were described as follows:

MüşteriID: A unique identifier for each customer, anonymized to protect customer identity. It is used for customer-level analyses.

Sipariş Tarihi: The date when the order was placed. This data is used to analyze the frequency of customer purchases and the date of their most recent purchase. It is utilized to calculate **Recency**.

Ürün Kodu: A unique code assigned to each product. It can be used for product-based analyses.

Ürün Adı: The name assigned to the sold product. It can be used to evaluate product-based sales performance.

Beden: The size of product.

Renk: The color of product.

Ürün Cinsiyeti: The target gender for the sold product. It is used in demographic analyses.

Şehir: The city where the order will be delivered. Regional customer distribution can be identified through basic statistical analyses, and city-based sales performance can be evaluated.

SatisYeri: Indicates where the product was sold, such as the two online platforms where the sale took place.

Sipariş Adedi: The number of orders placed by the customer. It is important for observing the customer's purchasing behavior and can be used to analyze sales volume.

Sipariş No: It is the number that uniquely identifies orders. Shopping frequency is determined using this data.

Toplam Fiyat: The total amount of the order. It can be used for customer revenue analysis (Monetary) and is also used in calculating Customer Lifetime Value (CLV).

Table 3.2.1.2 Descriptive Statistic of The Data

	CustomerID	SalesPlace	Order Quantity	Total Price
Count	241	241	241	241
Mean	269,9	1,81	1,0	2149,02
Std	371,79	0,39	0	1055,12
Min	1	1	1	979
25%	55	2	1	1449
50%	104	2	1	1849
75%	150	2	1	2699
Max	1041	2	1	6749

The data set gives a wide range of variability in both the customer IDs and order numbers, indicating there is immense variation in data, and that since the order quantity is always constant for all records at 1, that variable holds very little analytical value. The sales channel variable contains two different options, and the average value indicates there's a majority of the second channel. The overall cost varies between 979 TL and 6749 TL with the average expenditure of approximately 2149 TL, and there are large variations in customer buying behavior. The overall dataset is generally suitable for advanced analysis such as segmentation and spending behavior measurement for different customer and order profiles.

Table 3.2.1.2 Correlation Matrix

	MüşteriID	SatisYeri	Sipariş No	Toplam Fiyat
Müşteri ID	1,00	-0,98	0.99	0,11
SatısYeri	-0,98	1,00	-1,00	-0,11
Sipariş No	0,99	-1,00	1,00	0,10
Toplam Fiyat	0,11	-0,11	0,10	1,00

There is a strong positive correlation (0.99) between CustomerID and Order Number, which reflects that both variables are rising simultaneously with the passage of time and likely reflect sequential data entry or chronological order. There is a strong negative correlation (-1.00) between Order Number and Sales Channel, which reflects that different sales channels were utilized more frequently in different time periods. Similarly, the extremely negative correlation (-0.98) between Sales Channel and CustomerID shows that certain groups of customers used specific channels with high relative specificity. Conversely, Total Price has low correlations (about 0.1) with all of the other variables, which reflects that spending amounts are not highly regulated by sales channel, customer ID, or position of order.

This dataset contains customer records and is anonymized. It has been made in accordance with privacy principles. The dataset is structured to be used both to understand customer behavior and to improve decision support systems.

3.2.2 Data Cleaning

Missing Values: Filling in missing fields such as birth year or spending amount with median or mean. Removing missing values in critical fields on a row basis.

3.2.3 Data Standardization

Recency, Frequency and Monetary variables have been brought to the same scale for normalization. Labeling has been done to classify Churn and active customers correctly. The date section has been made suitable for use in coding processes.

3.3 Method

3.3.1 RFM Analysis

RFM is useful in learning about customer behavior and segmentation of customers. RFM analysis consists of Recency, Frequency, and Monetary values. Each one of these metrics provides us with different information about customer purchases.

Recency: Gives the time elapsed since the customer's previous purchase. Low value customers for recency are generally considered as active and loyal customers. This score is of special importance in the calculation of how recent the consumer's usage of the firm's product.

● Calculation: Date of the last transaction minus the analysis date in days.

● Usage: Identifying customers according to their segments and developing marketing strategies according to their segments.

Frequency: Demonstrates the total purchases done by the customer over a specified amount of time. Frequency helps us know how frequently the customer does business with the company.

- Calculation: Determined by summing up the number of times the customer has been making purchases in a certain timeframe.

- Usage: Used to provide customer loyalty feedback. Marketing initiatives can also be designed for regularly shopping customers.

Monetary (Amount Spent): It is the total cost of money consumers spend. It is a significant factor in quantifying the consumer's monetary contribution to the company.

- Calculation: Determined by adding up all the purchases of the customer.

- Usage: Special offers can be given to high-value customers.

RFM Scoring: Each customer is rated using a scale of 1 to 5 for Recency, Frequency, and Monetary metrics. Customers with the lowest Recency (recent buyers) are given 5 points, while those with the highest Recency (long-term no buyers) are given 1 point. Customers with the highest Frequency are given 5 points, while those with the lowest Frequency are given 1 point. Customers with the highest Monetary are given 5 points, while those with the lowest Monetary are given 1 point. These scores are combined to create segments. There are 10 segments in this study.

These segments are as follows:

can't lose them: A customer class that should not be lost despite frequent shopping.

at risk: A customer class that makes relatively frequent shopping but has not made a purchase for a long time.

hibernating: Sleeping customers, who have made few purchases and haven't buy anything in time.

about to sleep: A consumer class that has not made a purchase frequently and has not made a purchase for a certain period of time.

need attention: Customers who will move into the risky group if not focused on.

new customers: A customer class that has not made a purchase frequently and has not made a purchase recently.

promising: A customer class that has not made a purchase frequently and has not made a purchase recently.

potential loyalists: A customer group that makes medium-frequent purchases and has not made a long purchase since its last purchase.

loyal customers: A customer class that makes very frequent purchases and has not made a long purchase since its last purchase.

champions: A customer class that makes very frequent purchases and has made their last purchase in a very short time. They are also the highest-yielding customers.

3.3.2 K-Means Clustering

How the K-Means method works is explained as follows (Paramita et al., 2024):

K centers are selected as the starting point in the data set. Every data is put in the closest to the center. Cluster centers are recomputed by calculating the average of the data placed.

This repetition goes on until the cluster centers remain unchanged or the number of iterations is completed.

Strengths of this method are that it is easy and fast. Weakness is that pre-knowledge of value for number of clusters (k) needs to be provided and there are performance problems in imbalanced clusters.

Determination of K-Value with the Elbow Method:

In order to determine the best value of K-Means clusters, Elbow Method is utilized. For different sizes of clusters, total inertia is calculated and a chart is drawn. Elbow method works best when the value of k is small. It identifies the squared difference for dissimilar values of k. As the value of k goes up, the average irregularity goes down, the number of samples per cluster becomes smaller, and the samples move closer to the centroid. The optimal k value is identified at the point where the improvement in distortion begins to slow down significantly, which corresponds to the "elbow" point (Cui, 2020).

Total Intrinsic Variance:

It calculates the sum of squared differences between data points and cluster centers. It is a measure of homogeneity of the clusters.

Elbow Point:

On the graph of total intrinsic variance vs. the value of K, it is the point where the decreasing curve of the variance steeply goes down. It is the point which shows the optimal number of clusters.

Application Steps:

Different k values are tested using the K-Means algorithm (for example, from 1 to 10). Total intrinsic variance is calculated for each k value. Total intrinsic variance graph corresponding to the K value is created. The "elbow" point is determined on the graph and this k value is selected.

Interpretation of the Graph:

After the Elbow point, the change in the decrease of total intrinsic variance decreases rapidly. Therefore, the k value up to the Elbow point shows that the clusters are both homogeneous and significant.

3.3.3 Churn Prediction

Churn Prediction is an application that will forecast the likelihood of customers of a company churning (departing from the company) within a specific time interval. This forecast allows companies to act ahead of time in order to prevent losing customers. By doing so, companies can plan ahead of time the customers that they will lose and plan their marketing strategies.

LightGBM (Light Gradient Boosting Machine) is a fast and strong machine learning model that is largely used for churn prediction since it effectively manages speed and accuracy in classification tasks. LightGBM was launched in 2017 and is an algorithm of Gradient Boosting Decision Tree (GBDT) which was designed to solve the computational complexities that often pose a huge issue in traditional machine learning methods (Ke et al., 2017).

For the customer segmentation in this research, LightGBM (Light Gradient Boosting Machine) was used. Model selection was made in view of the amount of data, working mechanisms of the algorithms, and prediction accuracy considerations.

The consideration that the dataset contains small numbers of observations is a determining factor in choosing models. Random Forest is a predictor that makes predictions by averaging many decision trees. However, it is known that tree-based algorithms tend to overfit in moderately small datasets. Random Forest trains for each tree by dividing the dataset into different subsample groups. This method generalizes better in large datasets. It may not provide enough model diversity in small datasets. In contrast, LightGBM is a method that uses the gradient boosting algorithm and gives more successful results especially in small and medium-sized data sets.

Its first advantage is that it has regularization mechanisms that prevent over-learning even in small data sets. Another advantage is that it processes features more efficiently thanks to the histogram-based discretization method. In addition, it can make good predictions even in data sets with a small number of observations with advanced division algorithms. For these reasons, LightGBM was used instead of Random Forest when working with the data set containing 241 observations used in this study.

Based on past customer data, customers who haven’t bought anything within the specified time are tagged. According to RFM scores, customers who will churn are revealed. LightGBM is trained using key customer behavior features such as purchase frequency, recency, and monetary value, enabling precise churn predictions.

Discounts, promotions, or personalized campaigns are created for customers with a high churn risk. Some measures are taken to increase customer satisfaction for risky customer groups. LightGBM feature importance analysis helps companies determine which factors contribute the most to churn. Its performance is continually monitored and retrained if necessary. LightGBM's ability to support incremental learning allows companies to react quickly to changing behavior among customers and continually enhance their churn prevention strategies.

3.3.4 CLV

CLV can be used in calculating the summation lifetime revenue a consumer takes into a firm. It helps in marketing optimization by measuring customer value.

Approaches to Calculate CLV:

Historical CLV: CLV value calculated based on past purchase behavior. It is calculated on the total of the customer's previous spending.

Estimated CLV: Calculated by estimating future customer behavior. It is estimated based on RFM analysis and machine learning algorithms.

Predictive CLV: Predictive CLV aims to estimate future revenue potential by analyzing customer behavior and shopping trends. Since traditional CLV methods are based on historical data, Predictive CLV is used in this study. Predictive CLV is a method used to predict the total revenue that customers will bring to the company in the future and is calculated by machine learning models using RFM analysis data. Ridge Regression was used in the course of this estimation to prevent over-learning and reduce multicollinearity between variables. It is programmed to develop marketing and customer segmentation strategies based on Predictive CLV and Ridge Regression combined.

CLV Calculation Formula:

AOV (Average Order Value): Average order value.

Purchase Frequency: Represent the number of buying during the time.

Churn Rate: Shows the customer's churn rate within a certain time period.

CLV Usage:

Segmentation: Different strategies are determined by dividing customers into segments.

Personalized Marketing: Special marketing strategies are offered to customers with high CLV.

Budget Optimization: The marketing budget is made more efficient based on customer value.

Thanks to the CLV calculation, customer relationship management becomes more strategic and long-term profitability is increased, providing an advantage to the company.

4. FINDINGS

4.1 RFM Analysis

RFM analysis is used extensively in order to segment the customers and analyze customer behavior. RFM analysis was implemented using Python, and libraries NumPy, Matplotlib, Pandas, Seaborn and Scikit-learn were used for implementation. Analysis and visualization were performed.

When performing RFM analysis, first the R, F and M values are assigned.

Recency — Recency (R) — Time elapsed from the last sale to the present

Frequency — Frequency (F) — Total number of purchases

Monetary — Monetary (M) — Values are calculated as the monetary total of all purchases and assignments are made.

Each customer is assigned a separate recency (R), frequency (F) and monetary (M) score. This score is scored from 1 to 5.

In this study, customers were scored from 1 to 4. 1 was defined as the best and 4 as the worst. In other words, since we have 4 possibilities for each score, 4x4x4 = 64 different combinations were created. Thus, sufficient combinations were obtained to obtain detailed information.

To calculate the recency value, the difference between the day the analysis was made and the day the last purchase was made was taken. According to these values, there are customers who shopped 583 days ago and 513 days ago. You can see the 5 recency values in Table 4.1.1.

Table 4.1.1 Recency Values

CustomerID	**Recency**
1	518 days
2	518 days
3	514 days
4	504 days
5	504 days

Frequency values were calculated by looking at how often the customer makes purchases. Since it is a newly established brand, the shopping frequency of customers is generally one. It is also shown in the table. You can see the values in the next table.

Table 4.1.2 Frequency Values

CustomerID	Frequency
1	1
2	1
3	1
4	1
5	1

Monetary value represents the total spending value of customers. You can see the values in the next table.

Table 4.1.3 Monetary Values

CustomerID	Monetary
1	1449
2	2648
3	1999
4	1399
5	1199

RFM values are scored from 1 to 5 and the RFM score is created according to these scores. Score-based segments are created. You can see the values of RFM scores in Table 4.1.4 and segment based on RFM scores is in the Table 4.1.5.

Table 4.1.4 RFM Scores

CustomerID	Recency	Frequency	Monetary	RecencyScore	FrequencyScore	MonetaryScore	RFM_SCORE
1	518	1	1449	1	1	2	112
2	518	1	2648	1	1	4	114
3	514	1	1999	1	1	3	113
4	504	1	1399	1	1	1	111
5	504	1	1199	1	1	1	111

Table 4.1.5 Segments Based on RFM Scores

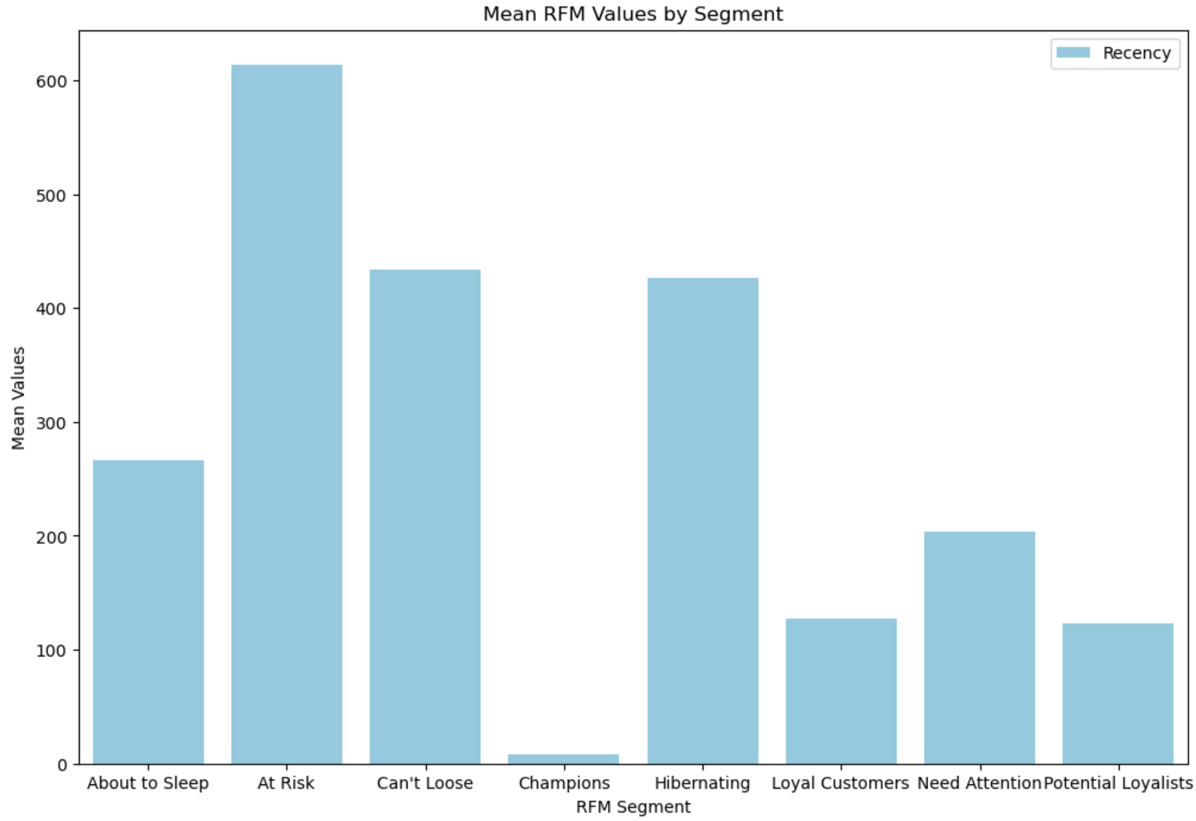
MüşteriID	Recency	Frequency	Monetary	RecencyScore	FrequencyScore	MonetaryScore	RFM_SCORE	Segment
1	518	1	1449	1	1	2	112	Hibernating
2	518	1	2648	1	1	4	114	Hibernating
3	514	1	1999	1	1	3	113	Hibernating
4	504	1	1399	1	1	1	111	Hibernating
5	504	1	1199	1	1	1	111	Hibernating

The segments and the number of customers in these segments are shown in Table 4.1.6. According to this table, special marketing plans can be designed for each segment. In this way, the champions group is the best segment. The average recency value is 8.59 and this is the lowest value. This means that the shortest time between shopping times belongs to customers in this segment. If we look at the frequency average, it is 1.703 and they shop frequently. If we look at the total value they spend, the highest value is 3864.75 according to the Monetary average. In this way, all groups can be evaluated. In Graph 4.1.1, you can see the bar chart of the segments according to their average recency values.

Table 4.1.6 RFM Analysis

Segment	Avg Recency	Avg Frequency	Avg Monetary	Count
About to Sleep	266.357	1.000	1965.429	28
At Risk	613.545	1.000	1998.909	22
Can't Lose	433.500	2.000	4219.667	6
Champions	8.595	1.703	3864.757	37
Hibernating	426.056	1.000	2101.944	54
Loyal Customers	127.529	1.176	3136.353	17
Need Attention	203.500	1.000	2121.500	8
Potential Loyalists	122.970	1.000	2024.152	33

Graph 4.1.1 Average Recency Bar Chart



Graph 4.1.2 analyzes the relationship between customers' Recency and Monetary. According to this analysis, it shows which segment is the most valuable, which customers are at risk of churn, and which customers are new or loyal.

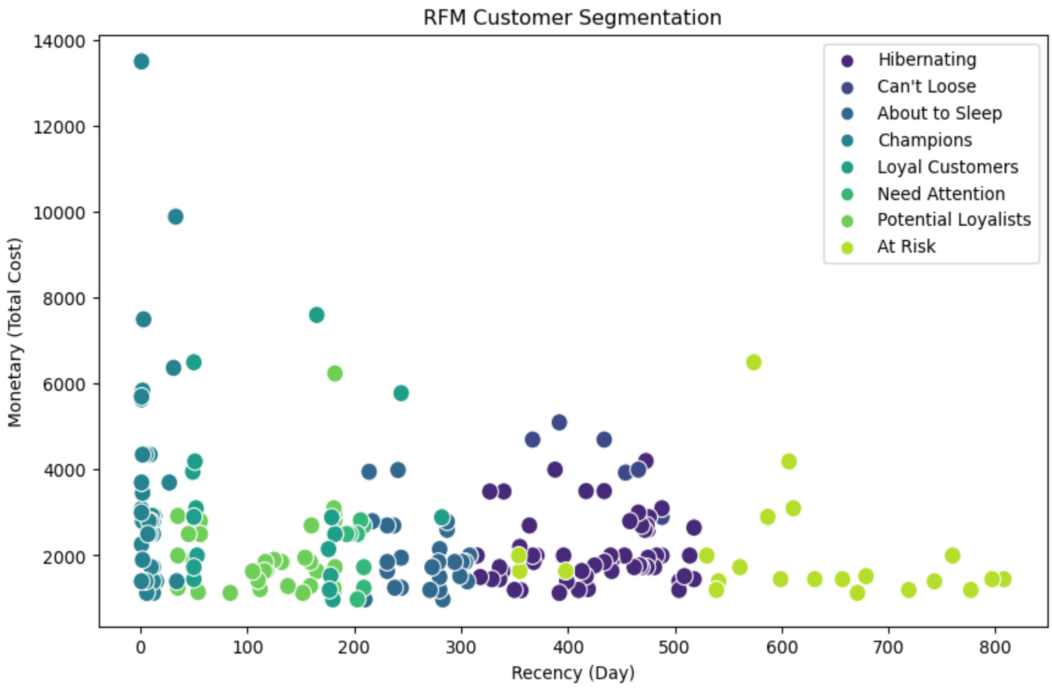
Top left corner (Low Recency, High Monetary): It shows customers who have recently shopped and spent a lot. These customers are the most valuable customers.

Top right corner (High Recency, High Monetary): They are customers who have not purchased for a long time but allocated a lot of money in the past. They are valuable customers who are at risk of losing.

Bottom left corner (Low Recency, Low Monetary): They are customers who have recently shopped but have dedicated less budget. They are new customers.

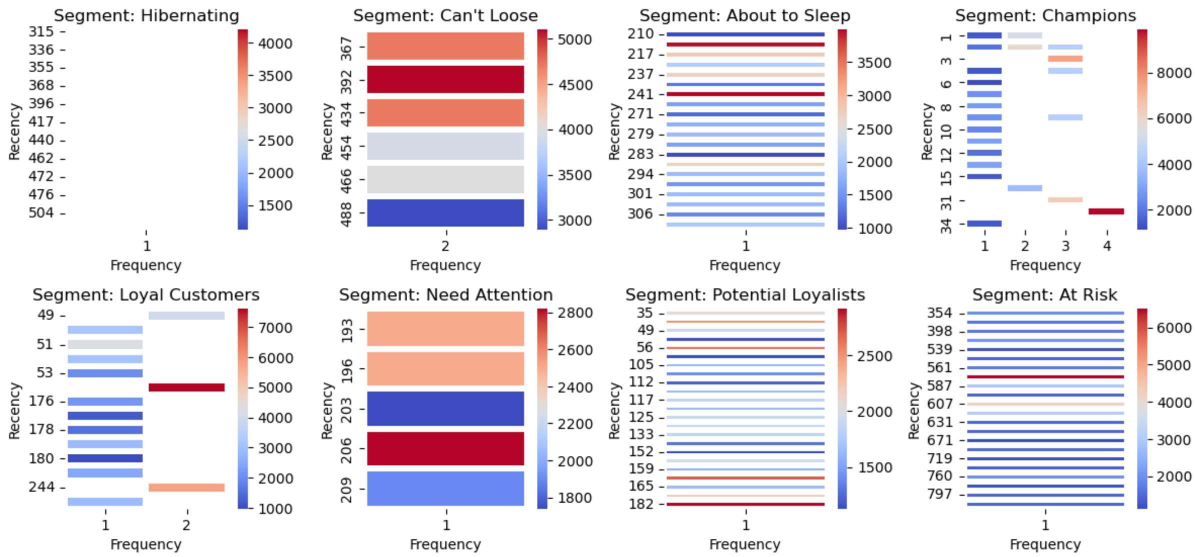
Bottom right corner (High Recency, Low Monetary): They are customers who have not purchased for a long time and have spent less of amount money. These customers are churned.

Graph 4.1.2 Scatter Plot of Recency and Monetary



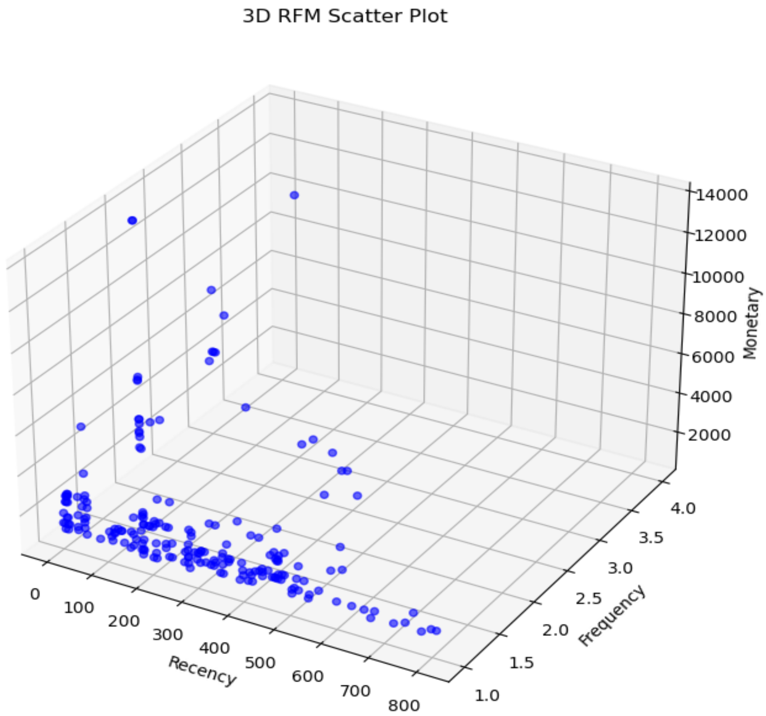
Heat map was used to evaluate the relationship between Recency and Frequency within each customer segment. According to this map, it was observed how frequency and recency changed on a segment basis. Recency-Frequency matrix was created separately for each segment and colored according to Monetary (Total Expenditure) values. Consumers who purchased a lot in the recent past are at the bottom right, those who purchased not long ago are at the top left, those who purchased a lot in the recent past but rarely are at the bottom left, and those who lost out completely are those with the highest cold color. You can see the heat map in Graphic 4.1.3.

Graph 4.1.3 Heat Map



In the graph, the points clustered in the upper part show customers who spend a lot. If the frequency value on the Y-axis is small, most customers have shopped less frequently. Customers with a high recency value show consumers who have not shopped for a long time. You can see the 3D scatter plot graph in Graphic 4.1.4.

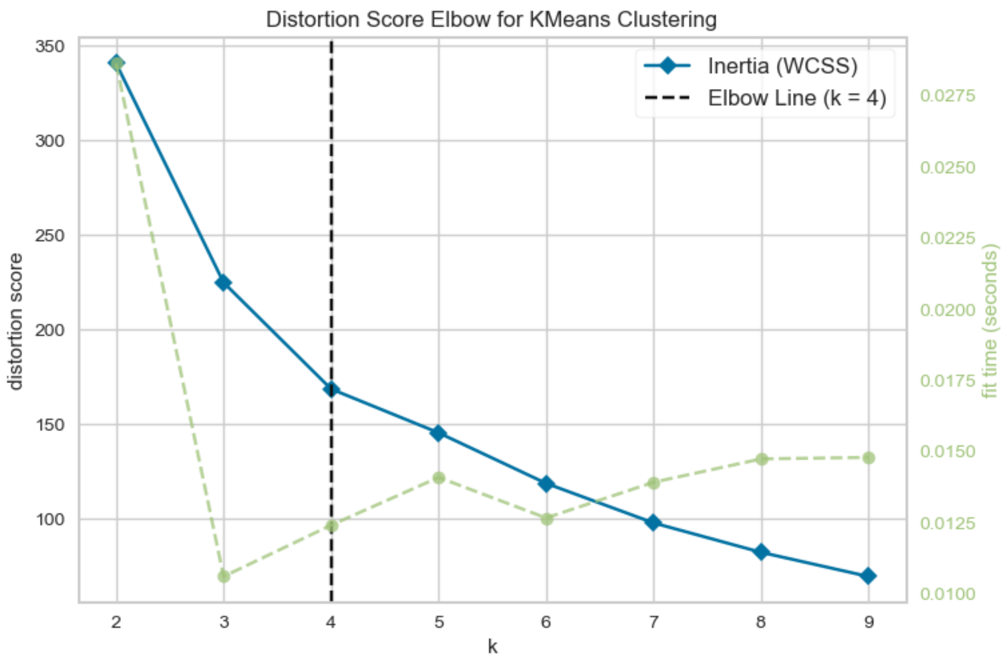
Graph 4.1.4 3D RFM Scatter Plot



4.2 K-Means Clustering

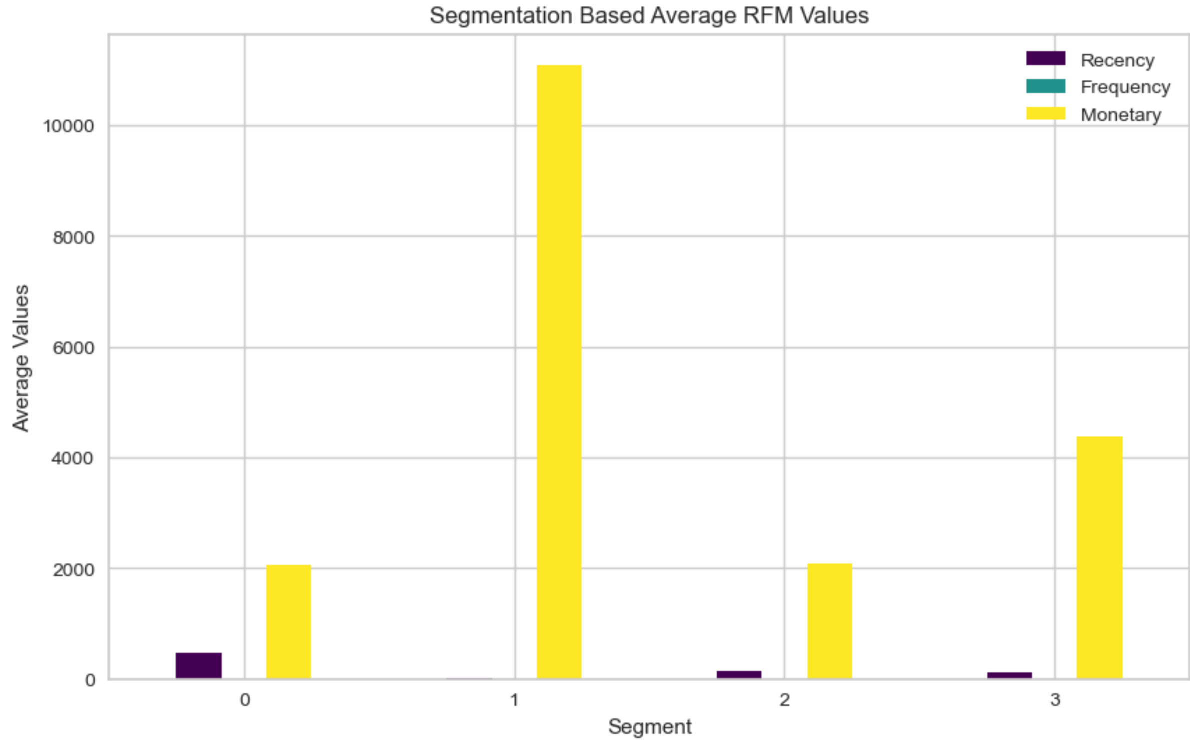
K-means analysis divides customers through RFM data. Customer segments are found and shown with this analysis. K-means algorithm is sensitive to scale. This is the reason why RFM data is scaled. Elbow technique calculates the cluster numbers in K-means. Graph 4.2.1 has Elbow Method graph. As seen on this graph, the biggest break occurs at point 4. It is the best number of clusters. Segmentation is done by splitting customers into 4 groups.

Graph 4.2.1 Elbow Chart



To understand which segment is more valuable, the average R, F and M values of each cluster were analyzed. Segments with high F and M values are a valuable customer group. In Graph 4.2.2, this customer group is the best because the Monetary value is the highest in the 1st group.

Graph 4.2.2 Segmentation Based RFM Values



In the K-Means Customer Segmentation Chart, each point represents a customer. Customers who have recently made purchases and are frequent shoppers are on the top left. Customers who have previously made purchases frequently but are no longer active are on the top right. Customers who have recently made purchases but are not yet frequent customers are on the bottom left. Customers who have not made purchases for a long time and have made very few purchases are on the bottom right. These patterns are shown in Graphic 4.2.3.

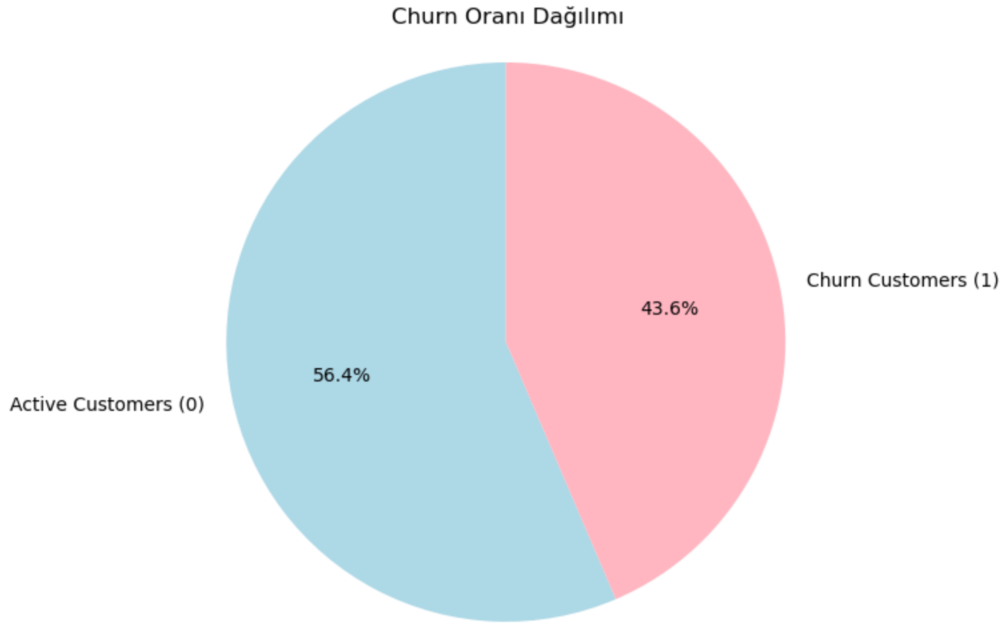
Graph 4.2.3 K-Means Customer Segmentation

K-Means is a ML technique regularly used for application to unsupervised learning problems such as segmentation of customers. RFM analysis is utilized so that marketing endeavors can be reduced to a systematized method through customer grouping according to what they buy.

4.3 Churn Rate Analysis

To perform churn rate analysis, a code was written in Python that analyzes customer churn. Accordingly, churn analysis examined customer behavior and helped predict customers who left the business in a certain period. The company determined that customers become churn customers when they did not shop within 90 days and the analysis was done in this way. The analysis results made according to this constraint are shown in Graphic 4.3.1.

Graph 4.3.1 Churn Distribution



According to this graph, customers shown in orange are active customers, and customers shown in blue are churn customers. The churn rate has increased to 56.43%. This means that half of the customers have started to take the risk of churn and special campaigns can be prepared for these customers.

Churn analysis was also performed using machine learning techniques. LightGBM (Light Gradient Boosting Machine) classifier model was created. This model works with gradient boosting technique. The model was trained with training data. Then, prediction was made with the test set. Test accuracy rate: 0.694

LightGBM model showed good performance with 69% accuracy rate. It is seen that the model is successful in recognizing churn customers; precision is 72%, recall is 75% and F1 score is 74% for churn class. This shows that the model has good performance in detecting churn customers. Outcomes are shown in following the table.

Table 4.3.1 LightGBM Results

	precision	recall	f1-score	support
0	0.65	0.62	0.63	21
1	0.72	0.75	0.74	28
accuracy			0.69	49
macro avg	0.69	0.68	0.69	49
weighted avg	0.69	0.68	0.69	49

Random forest algorithm can also be used in this type of data, but LightGBM gives better results in small data sets. Random Forest Results are:

Accuracy: 0.5918 (~59.2%)

Precision: 0.52 for Class 0, 0.64 for Class 1

Recall: 0.52 for Class 0, 0.64 for Class 1

F1-Score: 0.52 for Class 0, 0.64 for Class 1

LightGBM Result:

Accuracy: 0.6939 (~69.4%)

Precision: 0.65 for Class 0, 0.72 for Class 1

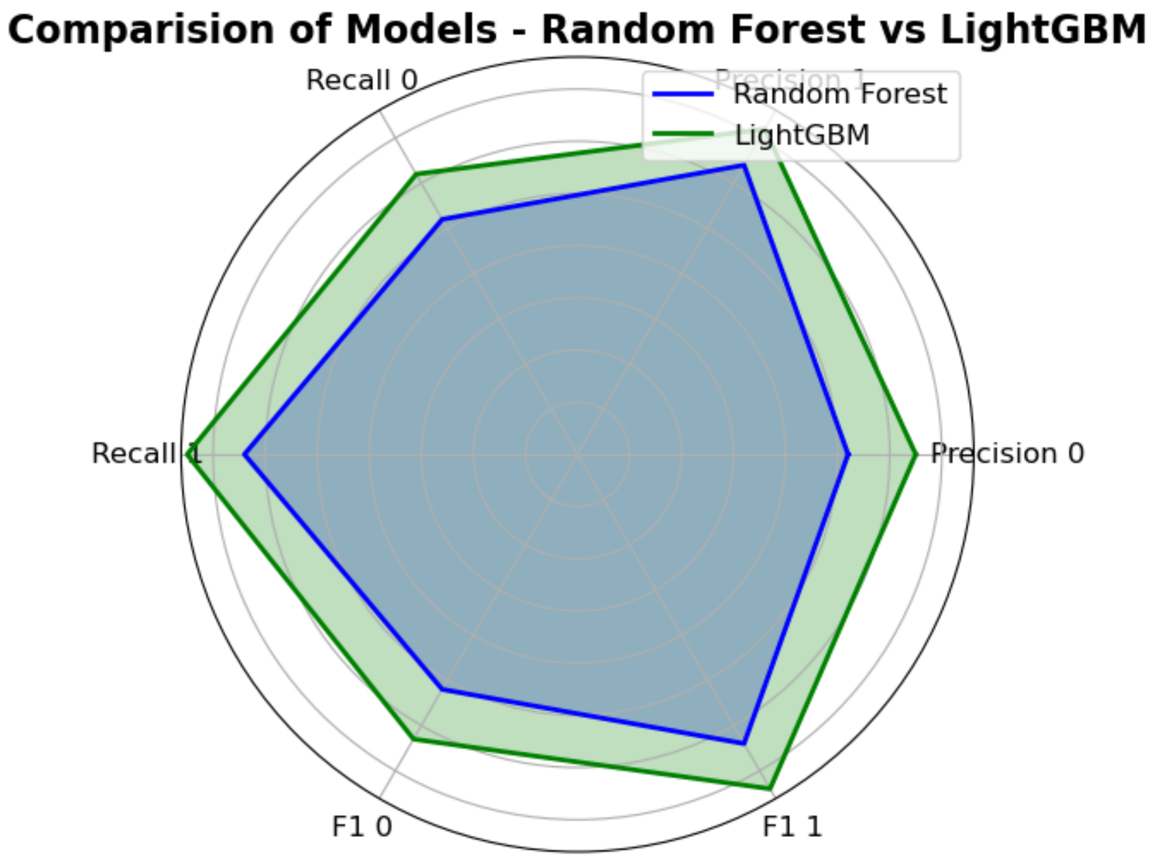
Recall: 0.62 for Class 0, 0.75 for Class 1

F1-Score: 0.63 for Class 0, 0.74 for Class 1

The results are as follows. Accordingly, LightGBM has a higher accuracy rate than the Random Forest model (69.4% vs. 59.2%). LightGBM has higher precision and recall values for both classes. It performs much better especially in Class 1. LightGBM provides 75% sensitivity (recall) in correctly predicting Class 1, which is a better result. As for F1-Score, LightGBM's F1-score is higher for both classes. Especially for Class 1, a high value of 0.74 was obtained.

LightGBM was selected as the better model. LightGBM model performed better overall was found. It has higher values of accuracy, precision, recall, and F1-score. You can see comparisons in Graphic 4.3.2.

Graph 4.3.2 Model Performance Comparison



4.4 Customer Lifetime Value

CLV is the revenue that a company will receive from a consumer over their lifetime. CLV analysis helps to determine the most valuable customers and allows strategies to be developed accordingly. Purchase history-related metrics were used in CLV calculations. Customers with high CLV values are those who demonstrate long-term loyalty and generate the most revenue for the business. Customers with low CLV values are those who will churn.

Predictive CLV model was used while estimating CLV and this model was trained with Ridge Regression. CLV estimation was made with machine learning. CLV estimation was calculated for a new customer and the accuracy of the Model Mean Absolute Error (MAE) was checked.

In this study, the Ridge Regression algorithm has been utilized to predict Customer Lifetime Value (CLV), and model performance was comparatively analyzed before and after feature engineering techniques were applied. In the baseline model, simple RFM (Recency, Frequency, Monetary) values and some categorical attributes (such as city and product gender) have been taken into consideration. The baseline model's performance was bad with Mean Absolute Error (MAE) = 881.53, Root Mean Squared Error (RMSE) = 1129.34, and R^2 = 0.19 only. These performance indicators inform us that the model could not effectively capture customer behavior, leading to very large prediction deviations.

To improve model performance, some new features were introduced by adding new variables through feature engineering. These included Tenure, which shows the duration between the customer's first and last purchase; FrequencyPerDay, a normalization of frequency of purchases over time; MonetaryPerOrder, for average spending per order; and SalesChannelsDiversity, showing diversity of sales channels used by the customer. Also, the existing categorical variables were translated into numerical form through one-hot encoding and introduced into the model.

With this expanded feature set, the performance of the new model improved significantly. The MAE decreased to 314.55, the RMSE dropped to 524.33, and the R^2 score improved to 0.83. These indicate that the enhanced model produces more accurate and consistent customer-level revenue forecasts. The feature engineering applied assisted in gaining a dramatic improvement in predictive ability by adding variables that capture customer behavior more accurately.

The rise in MAE and RMSE appears to be negative because the errors have decreased — which is a good thing. The R^2 score, however, has increased significantly, demonstrating that the model now correctly predicts much more.

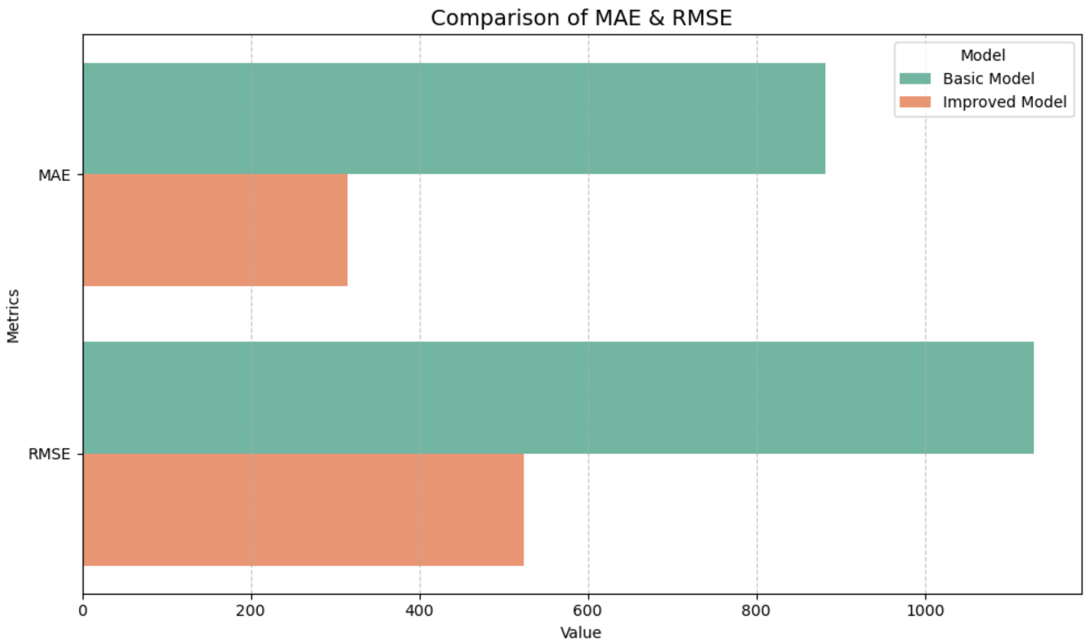
Table 4.4.1 New Model Performance Evaluation

Metric	**Old Model**	**New Model**	**Development**
MAE	881.53	314.55	🔽 -64.32% (Better)
RMSE	1129.34	524.33	🔽 -53.58% (Better)
R^2 Score	0.19	0.83	🔼 Better

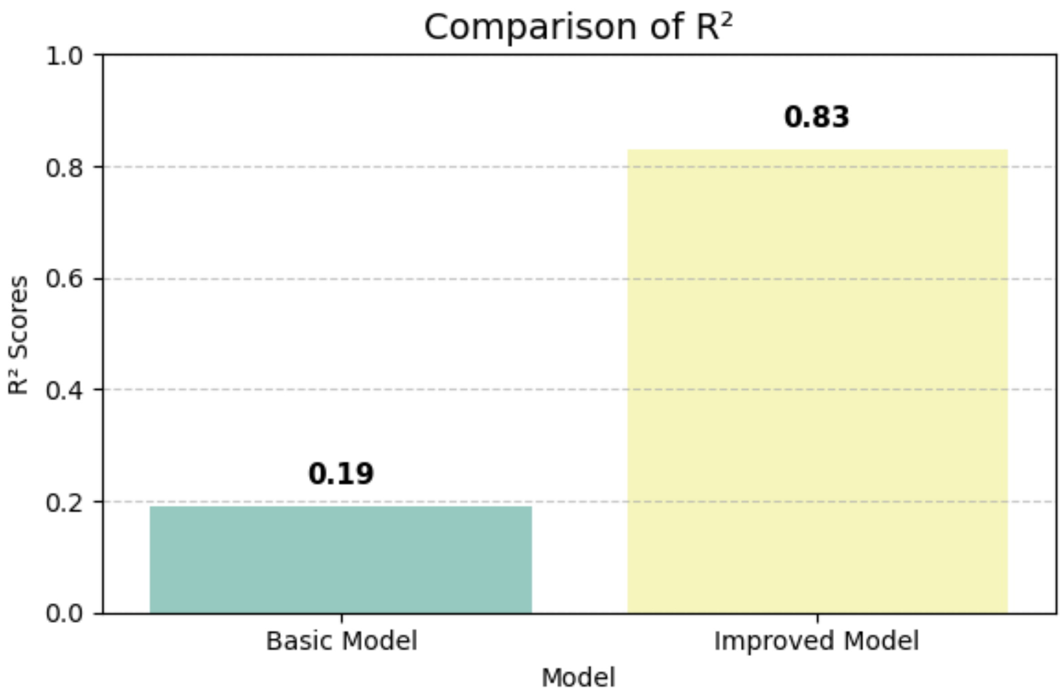
Notes: Ridge regression model was applied with $\lambda = 1.0$. The new model showed better performance in all metrics. MAE: Mean Absolute Error, RMSE: Root Mean Square Error, R^2: Coefficient of Determination.

Data: Fashion E-Commerce Data

Graph 4.4.1 Comparison of MAE & RMSE



Graph 4.4.2 Comparison of R^2



The complex model with better feature engineering was significantly superior to the basic model. The error measures (MAE and RMSE) fell significantly, but the R^2 score improved by leaps and bounds. It suggests that the complex model works much better as a predictor of customer value. Graph 4.4.1 and Graph 4.4.2. shows this.

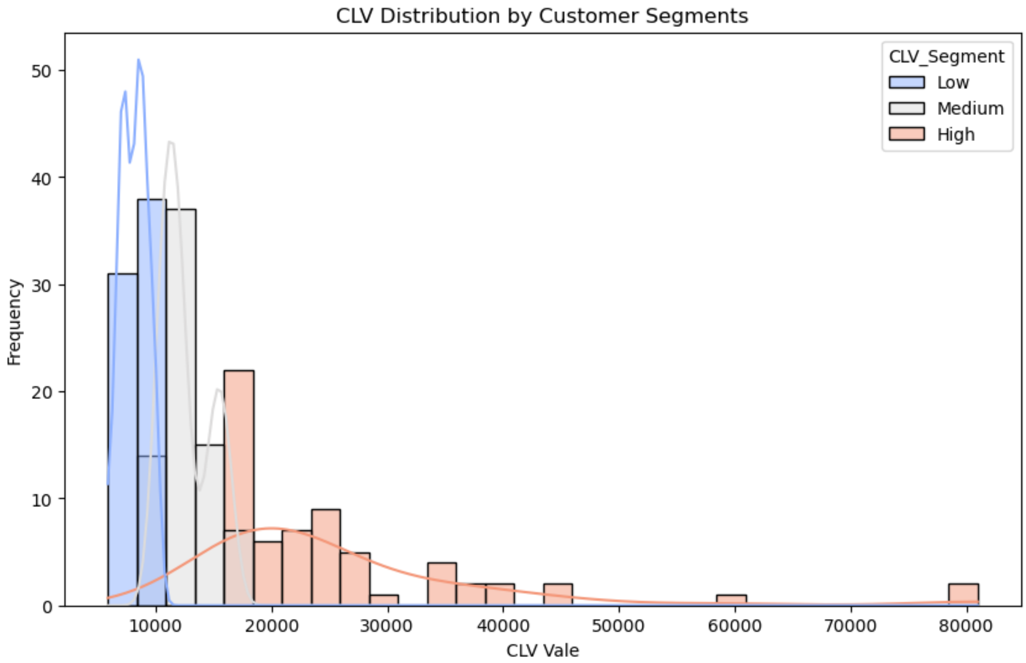
Then, as a result of the CLV (Customer Lifetime Value) analysis performed on the dataset, three different customer segments were created: Low, Medium and High CLV segments. First, the mean CLV values for each segment were examined. The mean value for the low CLV segment was 8107.57, for the medium segment 12582.33 and for the high segment 25865.90. This data showed that the high CLV segment consisted of more valuable customers and that this segment made high expenditures with longer-term customer relationships. The low CLV segment showed customers who spent less and had shorter-term relationships with the company. The medium CLV segment, on the other hand, was in between the two segments and showed that there was a target audience that needed to be developed.

In addition, when the order quantities of the segments are examined, the order quantity in each segment is approximately 1. This is due to the company entering a new sector. Tenure (customer relationship duration) is significantly longer in the high CLV segment. This shows that customers are more valuable because they have established longer-term relationships. As a result, a specific marketing strategy was developed for each of the High, Medium and Low segment customers. These results are shown in Table 4.4.2.

Table 4.4.2 Summary of the Segments

CLV Segments	CLV_Target Mean	CLV_Target Median
Low	8107.565217	8394.0
Medium	12582.328767	11994.0
High	25865.904762	20994.0

Graph 4.4.3 CLV Distribution by Customer Segments Histogram



The histogram is used to see the distribution of CLV values of each segment in more detail. This is especially useful for understanding the distribution of segments. You can also see the smoothed line of the CLV distribution of each segment using KDE (Kernel Density Estimate) in Graph 4.4.3.

As a result, RFM analysis, churn prediction using k-means clustering, and CLV prediction have helped companies optimize their customer relationships and increase their revenues.

5. DISCUSSION

This research applied machine learning techniques for segmenting the customers. However, it has some limitations.

First of all, this study used data from one e-commerce platform in a specific sector. This may make it hard to directly extend the findings to other e-commerce platforms or to different sectors. It should be noted that customer behavior may vary across platforms.

In addition, the inherent limitations of machine learning models should be taken into account. The algorithms used may be biased in certain data structures. Model outputs may also be biased, especially if they do not adequately represent certain customer groups. Therefore, care has been taken in data processing and model evaluation processes to ensure that model performance is fair and balanced, but it cannot be guaranteed that all possible biases have been completely eliminated.

The effects of these limitations can be reduced by conducting studies on different data sets and tests on different platforms. In addition, advanced model balancing techniques and comparative analyses with different machine learning algorithms can be performed to reduce model biases.

6. CONCLUSION

This thesis aimed to improve marketing strategies, increase customer loyalty and increase profitability by using various machine learning techniques while performing segmentation in fashion E-Commerce industry. This research includes basic analysis such as RFM analysis, K-means, Elbow method, churn estimation, LightGBM, Ridge regression and CLV estimation to understand customer behavior in more depth and perform segmentation effectively.

First, the RFM model provided a solid foundation to understand customer interaction and allowed analysis based on recency, frequency, and monetary metrics to identify profitable customers. Such data was extremely important in segmenting the customer segment and developing unique marketing strategies for each group. Significant customer segments were effectively created.

Meaningful customer segments were successfully created based on purchasing behavior using the K-means technique. The Elbow method is effective in determining the best cluster count to ensure that each segment is meaningful and applicable. RFM metrics were also used here while performing K-means. This segmentation allowed for personalized experiences to be provided that would increase customer happiness and retention.

Churn prediction is a critical issue for loyal customers. It has been addressed using LightGBM. Marketing strategies have been developed to reduce churn rates by accurately predicting the customers at risk of churn.

Additionally, Customer Lifetime Value (CLV) was estimated using Ridge regression and feature engineering, providing a quantitative metric that will help prioritize high-value customers. Data obtained from CLV analysis performs a crucial function in personalizing marketing strategies and allocating resources more efficiently.

The combination of these machine learning techniques has not only provided a clearer understanding of customer segments but also produced actionable e-commerce strategies for businesses. By using data-driven intelligence, fashion e-commerce businesses can make more informed choices, increase customer loyalty, and improve profitability.

To conclude, this thesis presents the potential of ML in transforming e-commerce strategies, specifically how it is able to create value in the fashion industry. Techniques such as K-means, churn prediction, and CLV prediction can be beneficial in making decisions for future marketing and operations.

Future research should consider ethical rules and concerns when conducting customer segmentation studies in fashion e-commerce. With the fast development of computing power and ML techniques, actionable insights are being gained. Sensitivity has also increased by applying deep learning models.

From a business perspective, companies that utilize real-time data processing and Artificial Intelligence-driven personalization systems contribute to enhancing customer experience and retention. Secondly, explainable and understandable ML and deep learning models are able to increase transparency, trust and reliability in customer segmentation models.

From the point of view of academic research, future work can be done on the interpretability of segmentation models, including multimodal data (i.e., image, text, and transaction data). In addition, work may be conducted on the use of reinforcement learning methods for dynamic customer targeting. The effect of ethical AI applications on consumer trust and engagement can also be a promising research direction.

7. REFERENCES

[1] Ali, N., & Shabn, O. S. (2024). Customer lifetime value (CLV) insights for strategic marketing success and its impact on organizational financial performance. *Cogent Business & Management*, *11*(1), 2361321.

[2] Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, *21*(3), 527-570.

[3] Bauer, J., & Jannach, D. (2021). Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(5), 1-37.

[4] Barrera, F., Segura, M., & Maroto, C. (2024). Multiple criteria decision support system for customer segmentation using a sorting outranking method. *Expert Systems with Applications*, *238*, 122310.

[5] Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, *33*(10), 1251-1257.

[6] Chouaten, K., Rodriguez Rivero, C., Nack, F., & Reckers, M. (2024). Unlocking high-value football fans: unsupervised machine learning for customer segmentation and lifetime value. *Frontiers in Sports and Active Living*, *6*, 1362489.

[7] Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5-8.

[8] Elveny, M., Syah, R. B., & Nasution, M. K. (2024). An boosting business intelligent to customer lifetime value with robust M-estimation. *Int J Artif Intell ISSN*, *2252*(8938), 1633.

[9] Feng, Y., Wu, X., Mu, W., Fan, M., & Feng, J. (2024). Segmentation of Chinese consumer preference for wine extrinsic attributes based on stratification and weighted clustering algorithm. *Food Research International*, *196*, 115097.

[10] Firmansyah, E. B., Machado, M. R., & Moreira, J. L. R. (2024). How can Artificial Intelligence (AI) be used to manage Customer Lifetime Value (CLV)—A systematic literature review. *International Journal of Information Management Data Insights*, *4*(2), 100279.

[11] Gadgil, K., Gill, S. S., & Abdelmoniem, A. M. (2023). A meta-learning based stacked regression approach for customer lifetime value prediction. *Journal of Economy and Technology*, *1*, 197-207.

[12] Griva, A., Zampou, E., Stavrou, V., Papakiriakopoulos, D., & Doukidis, G. (2024). A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data. *Journal of Decision Systems*, *33*(1), 1-29.

[13] Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018, January). Modeling and application of customer lifetime value in online retail. In *Informatics* (Vol. 5, No. 1, p. 2). MDPI.

[14] Jethva, D. A., Khamkar, S. M., Pachchhapur, A. A., & Kulkarni, S. (2023, December). Customer Segmentation Analysis using K-means Algorithm with Elbow method and Dendrogram. In *2023 6th International Conference on Advances in Science and Technology (ICAST)* (pp. 296-301). IEEE.

[15] Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, *36*(9), 4995-5005

[16] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

[17] Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia computer science*, *3*, 57-63.

[18] Ledro, C., Nosella, A., & Vinelli, A. (2022). Artificial intelligence in customer relationship management: literature review and future research directions. *Journal of Business & Industrial Marketing*, *37*(13), 48-63.

[19] Li, M., Wang, Q., Shen, Y., & Zhu, T. (2021). Customer relationship management analysis of outpatients in a Chinese infectious disease hospital using drug-proportion recency-frequency-monetary model. *International Journal of Medical Informatics*, *147*, 104373.

[20] Lloyd, A., Cheshire, J., Squires, M., Longley, P., & Singleton, A. (2018). The Provenance of Customer Loyalty Card Data. *Consumer Data Research*, 28-39.

[21] Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. *IEEE Access*.

[22] Mena, G., Coussement, K., De Bock, K. W., De Caigny, A., & Lessmann, S. (2024). Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. *Annals of Operations Research*, *339*(1), 765-787.

[23] Mykhalchenko, H., & Tytarenko, M. (2023). Data analytics and personalized marketing strategies in E-commerce platforms. *Futurity Economics&Law*, *3*(3), 114-138.

[24] Nugroho, B. I., Rafhina, A., Ananda, P. S., & Gunawan, G. (2024). Customer segmentation in sales transaction data using k-means clustering algorithm. *Journal of Intelligent Decision Support System (IDSS)*, *7*(2), 130-136.

[25] Paramita, A. S., & Hariguna, T. (2024). Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce. *Journal of Digital Market and Digital Currency*, *1*(1), 43-62.

[26] Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *Plos one*, *18*(12), e0289724.

[27] Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. *Machine Learning with Applications*, *17*, 100567.

[28] Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, *237*, 121449.

[29] Sarkar, M., Puja, A. R., & Chowdhury, F. R. (2024). Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis. *Journal of Business and Management Studies*, *6*(2), 54-60.

[30] Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). *Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. Data Science and Management, 7 (1), 7–16.*

[31] Sivaguru, M. Dynamic customer segmentation: A case study using the modified dynamic fuzzy c-means clustering algorithm. *Granul. Comput.* 8, 345–360. https://doi.org/10.1007/s41066-022-00335-0(2023).

[32] Sulianta, F., Ulfah, K., & Amalia, E. (2024). Revealing Consumer Preferences in the Fashion Industry Using K-Means Clustering. *International Journal of Engineering Continuity*, *3*(2), 34-53.

[33] Suh, Y. Machine learning based customer churn prediction in home appliance rental business. *J Big Data* 10, 41 (2023). https://doi.org/10.1186/s40537-023-00721-8

[34] Sun, Y., Liu, H., & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, *9*(2).

[35] Sun, Z. H., Zuo, T. Y., Liang, D., Ming, X., Chen, Z., & Qiu, S. (2021). GPHC: A heuristic clustering method to customer segmentation. *Applied Soft Computing*, *111*, 107677.

[36] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering* (Vol. 336, p. 012017). IOP Publishing.

[37] Yamini, B., Ramana, K. V., Nalini, M., Devi, D. C., Maheswari, B., & Siva, S. R. (2024). Customer churn prediction model in enterprises using machine learning. *International Journal of Advanced Technology and Engineering Exploration*, *11*(110), 94.

[38] Wu, J., Shi, L., Yang, L., Niu, X., Li, Y., Cui, X., ... & Zhang, Y. (2021). User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Analysis. *Wireless Communications and Mobile Computing*, *2021*(1), 9982484.

[39] Wu, R. S., & Chou, P. H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, *10*(3), 331-341.

[40] Zhang, C. (2024). Exploring Key Factors Influencing Retail Enterprise Performance: A Comprehensive Analysis Using Multivariate Regression Models. In *SHS Web of Conferences*(Vol. 208, p. 04019). EDP Sciences.