“Using Machine Learning for Customer Segmentation and Sales Prediction in E-Commerce Industries”

Abstract

This study applies machine learning for customer segmentation, sales forecasting, and churn prediction. EDA identifies seasonal trends and demand fluctuations, linking revenue variations to promotions and economic factors. K-Means clustering segments customers by Recency, Frequency, and Monetary (RFM) values, distinguishing high-value, moderate, and inactive users. These insights enable targeted marketing, loyalty programs, and re-engagement strategies to optimize retention and revenue.

The Gradient Boosting Regressor (GBR) delivers the most reliable sales forecasting, achieving an R^2 score of 0.871, confirming its strong predictive capability without overfitting. Alternative models, including Linear Regression, Support Vector Regressor, and Random Forest Regressor, exhibited overfitting tendencies ($R^2 \approx 1.000$), reducing their real-world reliability. The predictive outcomes of GBR establish meaningful relationships between product pricing, customer behavior, and sales response trends.

For churn prediction, the Gradient Boosting Classifier (GBC) emerges as the best model, reaching an accuracy of 73.38% and an ROC-AUC score of 0.7972. This model outperforms Random Forest (72.25% accuracy, 0.8019 ROC-AUC), Logistic Regression, and Support Vector Machines (both at ~50% accuracy, indicating weak prediction). The results highlight GBC's capability in identifying potential churners, allowing businesses to take proactive retention measures.

The study confirms the effectiveness of machine learning in customer analytics, enabling businesses to enhance retention strategies and optimize revenue generation. Future research will focus on advanced deep learning approaches, real-time data integration, and ensemble learning techniques like XGBoost and LightGBM to further refine predictive accuracy in business intelligence and decision-making.

Keywords: *Machine Learning, Customer segmentation, Sales Prediction, Exploratory Data Analysis, Seasonal Trend, Customer Behavior, Algorithm, Recency Frequency Monetary, Sales Forecasting, Business Intelligence*

Table of Contents

Abstract II

List of Figures VI

List of Tables VII

Dedication VIII

Acknowledgment IX

CHAPTER 01: OVERVIEW OF MACHINE LEARNING IN CUSTOMER ANALYTICS 1

1.1 Background of the Study 1

1.2 Problem Statement 4

1.3 Research Objectives 6

1.4 Research Questions 7

1.5 Scope of the Study 8

1.6 Significance of the Study 10

1.7 Summary of Chapters 11

CHAPTER 2: EXISTING APPROACHES TO CUSTOMER SEGMENTATION AND SALES PREDICTION 13

2.1 Overview of Customer Segmentation 13

2.1.1 Traditional Segmentation vs. Machine Learning-Based Segmentation 14

2.1.2 Role of Recency, Frequency, and Monetary (RFM) Analysis 15

2.2 Machine Learning in Sales Prediction 16

2.2.1 Supervised learning models 17

2.2.2 Unsupervised Learning Models 18

2.2.3 Evaluation metrics for sales prediction 20

2.3 Churn Analysis in Business 20

2.4 Review of Previous Studies 22
2.4.1 Existing Research on Customer Segmentation 22
2.4.2 Comparison of Machine Learning Models in Previous Studies 23
CHAPTER 3: TECHNIQUES FOR CUSTOMER SEGMENTATION AND SALES PREDICTION 26
3.1 Dataset Description 26
3.2 Data Preprocessing 27
3.2.1 Handling Missing Values 27
3.2.2 Converting Categorical Variables into Numerical Format 28
3.2.3 Feature Scaling Using StandardScaler 28
3.3 Exploratory Data Analysis (EDA) 29
3.3.1 Age Distribution (Histogram) 29
3.3.2 Seasonality and Trend Analysis (Line Plots) 29
3.3.3 Popular Product Categories (Count Plot) 30
3.4 Customer Segmentation Using K-Means 31
3.4.1 Feature Selection for Clustering 31
3.4.2 Standardization of Features 31
3.4.3 Determining the Optimal Number of Clusters (Elbow Method) 32
3.4.4 Applying the K-Means Algorithm 32
3.4.5 Visualizing Clusters (2D & 3D Plots) 32
3.5 Sales Prediction Using Machine Learning 33
3.6 Churn Prediction Using Classification Model 34
3.7 Flowchart Overview 34
3.7 Software and Hardware Requirement 36
CHAPTER 4: EVALUATING PREDICTIVE PERFORMANCE AND BUSINESS IMPACT 37

4.1 Exploratory Data Analysis (EDA) Results 37
4.1.1 Seasonality and Trends in Sales 37
4.1.2 Patterns in Product Category Preferences 38
4.2 Customer Segmentation Results 40
4.2.1 Cluster Characteristics and Interpretation 40
4.2.2 Business Recommendations Based on Clusters 41
4.3 Sales Prediction Performance 41
4.4 Churn Analysis Results 43
4.4.1 Model Performance and Evaluation 43
4.4.2 Interpretation of Churn Factors 44
4.4.3 Business Recommendations for Churn Prevention 45
CHAPTER 5: CONCLUSION AND FUTURE WORK 46
5.1 Conclusion 46
5.2 Future Work 47
References 49
Appendix I 55
Appendix II 24

List of Figures

Figure 1.1: Flow chart for machine learning workflow 2
Figure 1.2: Customer Segmentation Analysis..... 3
Figure 2.1: Comparison of accuracy values for different machine learning models ……...13
Figure 2.2: Understanding and Predicting Customer Churn ……20
Figure 3.1: Complete flowchart of methodology ……..... 34
Figure 4.1: Interpretation of seasonality 36
Figure 4.2: Seasonal Variation by Month 37
Figure 4.3: Business insights from popular product categories 38
Figure 4.4: Interpretation of 3D K-Means cluster plot ……39

List of Tables

Table 3.1: System Specification for Analysis 36
Table 4.2: Model Based Sales Prediction .....**Error! Bookmark not defined.**
Table 4.3: Classification Report based on Random Forest ...**Error! Bookmark not defined.**

Dedication

I dedicate this project to my family and friends, whose constant support and encouragement have helped me throughout this journey. I would also like to thank my teachers and mentors for their valuable guidance, which has been a big help in completing this project. Special thanks to my team for working with me and helping bring this idea to life. I am also grateful for the feedback and support from peers and colleagues, which has motivated me to improve and refine my work. Finally, I dedicate this project to everyone who values saving time and making things easier, which is the main goal of Business Segmentation. This project is a step towards improving everyday efficiency for all.

Acknowledgment

I want to thank my supervisor for his unwavering support and guidance throughout this project. His advice and feedback have been instrumental in making the project a success. I also appreciate my fellow students who helped me along the way. Their input, ideas, and teamwork have been essential in improving my work and turning my ideas into reality. I am especially grateful to those who took the time to review my work and provide valuable suggestions. A special thank you goes to my teachers who inspired me with their knowledge and passion. Their lessons have given me a strong foundation and helped me face challenges with confidence. Their encouragement has motivated me to push my limits and strive for excellence in all my endeavors. I am grateful to my friends for their constant support and encouragement. Their belief in me has kept me motivated and focused during this journey. They have always been there to listen and provide emotional support, especially during challenging times. Lastly, I appreciate the innovations in technology that made this project possible, enabling me to work more efficiently and improve the user experience. Thank you all for your support and contributions, which made this project a reality.

CHAPTER 01: OVERVIEW OF MACHINE LEARNING IN CUSTOMER ANALYTICS

1.1 Background of the Study

Machine Learning operates like a subdivision of artificial intelligence that helps systems make clear predictions and choose suitable decisions after analyzing programmed data. The technological tool functions as an indispensable asset across finance and healthcare sectors along with marketing operations and energy systems because it enables the processing of extensive data for findings valuable information (Gkikas and Throdoridis, 2024). Supervised learning, unsupervised learning, and reinforcement learning are the three primary categories of machine learning algorithms. Supervised learning trains models through data containing labeled inputs with established output relationships which becomes applicable for classifying and making predictive estimations. The identification of data patterns through unsupervised learning occurs without labeled outputs while K-Means acts as a popular clustering method for customer segmentation. Reinforcement learning represents a model category which trains through environmental interactions that generate rewards contingent upon actions made by the model while being utilized in robotics along with game simulations (Madanchian, 2024).

The study employs machine learning techniques such as the Random Forest Classifier, Random Forest Regressor, and K-Means Clustering to analyze consumer behavior, predict sales, and segment customers based on their purchase patterns. The implementation of these approaches helps organizations use data to drive better choices and create satisfied customers and superior marketing approaches. Organizations can use machine learning to achieve better consumer insights and maximize operational efficiency which leads to greater profitability (Pellegrino, 2024).

Customer segmentation along with sales prediction stand as essential components for decision-making that use data in nowadays business era. For businesses to achieve strategic marketing optimization and boost customer relations and financial success is their ability to group customers by common traits and their forecast of future purchasing patterns. The traditional customer classification methods which utilized demographic

and geographic as well as psychographic factors now get replaced by machine learning (ML) which drives data-based adaptive segmentation strategies (Singh et al., 2024).

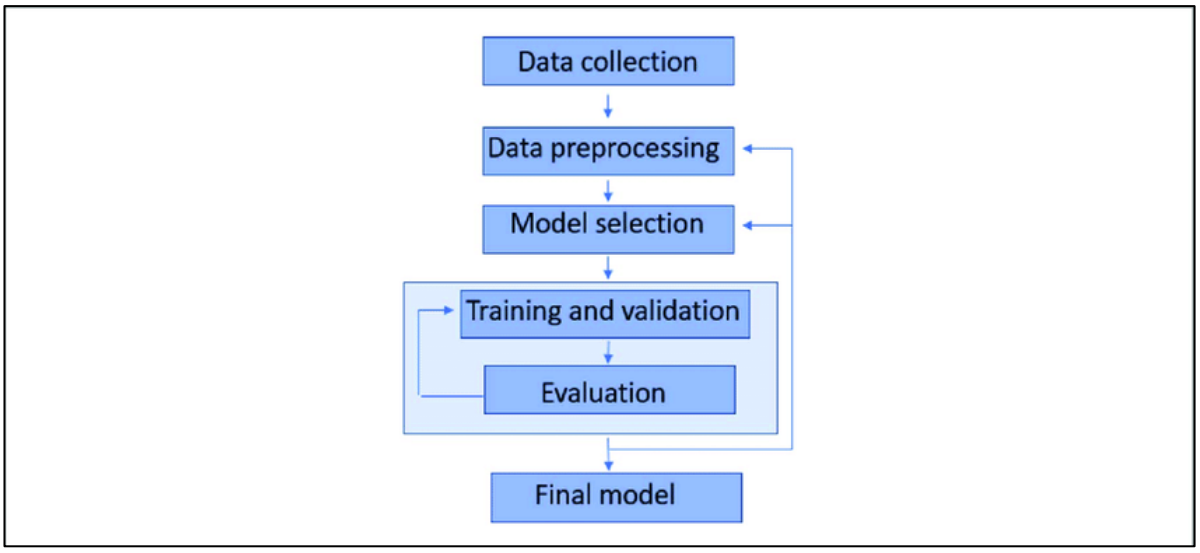

Figure 1.1: Flow chart for machine learning workflow

The marketing strategy of customer segmentation provides businesses with a means to organize their client groups through shared attributes including buying habits combined with behavioral attributes and monetary worth. Effective customer segmentation procedure leads to better personalization strategies which decreases customer acquisition expenses while boosting client loyalty. Commonly used conventional segmentation approaches include k-means clustering, hierarchical clustering, and RFM (Recency, Frequency, and Monetary) analysis. Current segmentation methods lack accuracy and adaptability when used alone in capturing modern consumer behavior patterns since they require machine learning algorithms for precision enhancement (Mowar, 2022).

The process of sales prediction serves crucially as businesses need this information to anticipate revenue growth and thus control inventory and efficiently use their resources. Sales forecast creation relies on decision trees together with Support Vector Machines (SVM) and arrange learning classfifier including random forest as well as statistical models to predict trends from historical data. The prediction accuracy gets improved by machine learning algorithms because these algorithms reveal hidden patterns that standard statistical approaches fail to detect (Matuszelański et al., 2022). Businesses obtain crucial knowledge from analytics that guides their pricing framework

development while improving inventory planning and demand prediction which minimizes financial perils and strengthens operational effectiveness.

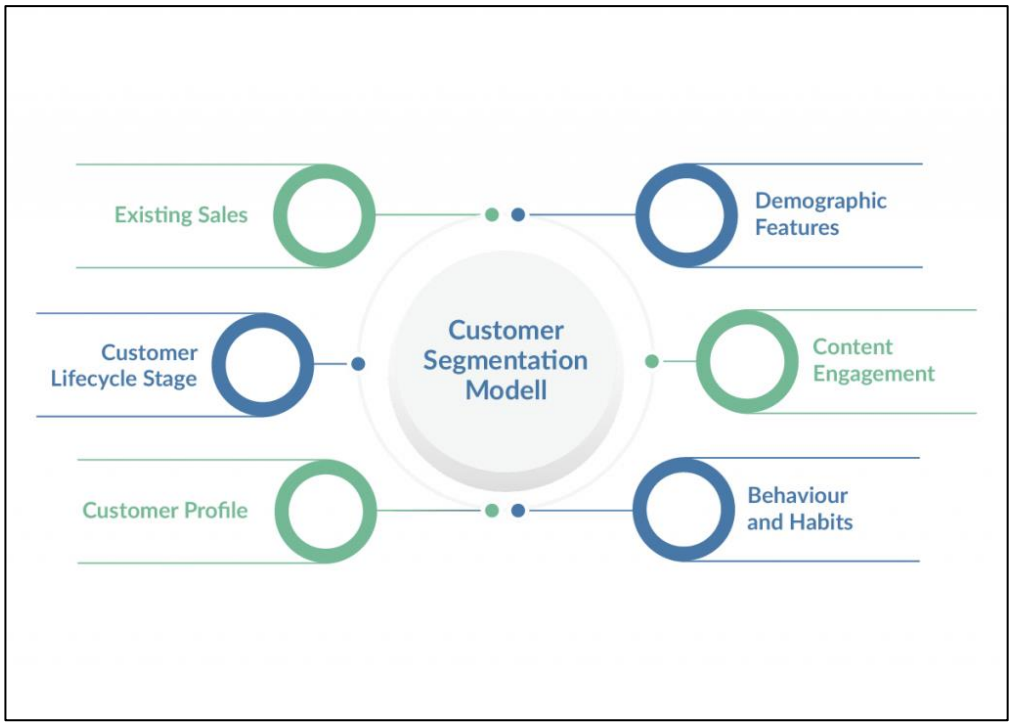

Figure 1.2: Customer Segmentation Analysis

Customer segmentation and sales prediction processes use machine learning because of quick-growing big data along with developing computational technology. Enterprise data collection has expanded because businesses obtain information from transaction records and online interactions and social media engagements. Through the implementation of ML clustering and classification methods as well as regression techniques organizations gain valuable insights from their data to automate their processes and boost customer satisfaction (Suhaas et al., 2024). Research shows deep learning models employing Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs) succeed at discovering subtle customer patterns leading to effective long-range sales predictions (LeCun et al., 2015).

A dataset will allow this study to investigate customer segmentation techniques while developing sales predictions through machine learning applications. The research draws its data from a retail organization which includes variables encompassing different types of information about customers including demographic features along with purchase records and transaction counts and overall purchasing sums. The main purpose of this project entails using k-means clustering for customer segmentation by purchasing

behavior while regression models predict future sales. The study applies these methods to deliver practical findings which strengthen target customer approaches and maximize promotional activities and fuel business expansion.

This study will tackle the data preprocessing obstacles as well as the evaluation methods and selecting appropriate features for both customer segmentation and sales prediction tasks. The research evaluates different machine learning models while determining their predictive abilities to discover ideal methods for practical use. The study uses advanced analytics with machine learning methodologies to expand business intelligence fields while creating effective data-driven recommendations in retail operations (Gandomi and Haider, 2015).

1.2 Problem Statement

The contemporary retail and e-commerce industry faces a vital business challenge because of the need to understand purchasing behavior among customers. The process of dividing customers into segments allows organizations to develop better business approaches and improve their marketing initiatives while boosting retention rates. Large datasets remain difficult to analyze for businesses since they need to overcome the challenges of complex buyer patterns together with seasonal consumption behaviors and multiple influencing variables. Companies utilizing demographic-based customer segmentation face the problem of missing dynamic behavioral patterns which leads to wasteful marketing efforts with associated revenue loss (Prasetyo et al., 2024).

Every business must overcome its inability to recognize its profitable client segments. Companies operating in competitive industries must properly identify between customers who buy frequently at high prices and those whose limited contributions result in minimal revenue. High-value customers who businesses refer to as "loyal customers" influence the profitability levels of organizations profoundly. The failure to segment leads businesses toward the wrong direction because they direct marketing spending toward worthless customers instead of targeting valuable customers (Homburg et al., 2013). The inability of standard segmentation approaches that use basic heuristic methods or pre-defined groups to respond to shifting consumer patterns makes machine-based category identification an essential business tool today (Wang et al., 2016).

Sales trend prediction stands as a major difficulty for organizations. Total purchase amount predictions together with revenue projections remain complicated because numerous external influences including economic circumstances and advertising promotions and seasonal variations affect the outcome (Helomld, 2022). Businesses succeed in making accurate demand forecasts through machine learning algorithms which extract hidden patterns from historical sales data records. By making an accurate robust sales prediction model it is necessary to put effort into data preprocessing while selecting key features and evaluating the model's performance. Untrained data models produce incorrect forecasts which causes companies to lose money while struggling to regulate inventory quantity along with marketing strategy effectiveness (Holloway, 2024).

The main problem affecting customer segmentation and sales forecasting entails detecting and predicting customer churn rates. A high level of industry competition makes customer churn—an event where customers stop buying—become an essential business challenge. Organizations find it maximum amount of cost-effective to keep ongoing customers instead of acquiring new ones thus making churn prediction an indispensable part of customer relationship management. The traditional models for churn prediction depend on either fixed rules or historical purchase data but machine learning allows an ongoing evaluation of customer engagement and accurate churn modeling from transaction logs and frequency contacts (Berndt and Petzer, 2023).

Companies require an effective machine learning approach which unites customer classification methods with forecasting capabilities while detecting customer attrition at different levels. Organizations can identify meaningful behavioral patterns of customers through K-Means clustering to perform segmentation instead of using arbitrary demographic grouping. Random Forest Regressor along with other supervised learning algorithms help businesses accurately forecast future sales volumes through which they obtain valuable data for inventory management and revenue growth. The use of classification models that includes models like Logistic Regression or Support Vector Machines enables businesses to detect at-risk customers through churn prediction which prompts proactive interventions by tailored marketing strategies.

The objectives of this research involve creating machine learning systems which integrate customer segmentation through clustering with sales prediction regression and

analyze customer churn through classification models. The implemented approach includes state-of-the-art preprocessing together with exploratory data analysis and feature engineering to generate insights which businesses can use for improving customer engagement and maximizing revenue and minimizing abandonment rates. The research findings add value to predictive analytics studies because they show how data-based decision systems work in contemporary business operations (Friedman, 2020).

1.3 Research Objectives

The following are the main goals of this study:

- **Implement Customer Segmentation Using K-Means Clustering**

Through its analytical power businesses succeed in grouping customers depending on shared characteristics using customer segmentation as a technique. K-Means clustering operates in this piece to split customers into separate groups through analysis of purchase frequency together with monetary value and transaction recency. The study utilizes K-Means analysis to identify groups which contain valuable customers, loyal shoppers and disengaging customers. The customization through segmentation helps businesses make targeted marketing plans as well as tailor their customer dealings and allocate resources effectively.

- **Develop a Random Forest Regressor for Sales Prediction**

Accurate sales prediction remains critical to manage stock levels and develop financial strategies and make strategic choices. The model employs Random Forest Regressor as a forecasting method to predict sales by analyzing historic transaction data. The prediction model generates reliable sales projections through merging data from product price and customer purchase elements and different demographic variables. The implementation of predictive models by businesses helps them better predict demand shifts thus lowering stockout cases and improving their revenue forecasting accuracy.

- **Analyze Churn Prediction Using Classification Models**

The accurate prediction of churn stands as the essential answer businesses need to stay in possession of their customers. Random Forest Classification functions as the analytical method to evaluate customer churn patterns in this research project. Training the classification model involves using customer engagement records alongside

transaction information and return history to distinguish staying customers from departing ones. Companies use tracked retention clues to activate customized promotional strategies and improve loyalty programs as well as customer service quality.

- **Provide Insights for Business Decision-Making**

The ultimate purpose of this investigation leads to develop implementation-based findings which guide business strategic choices. The research study unifies machine learning models that segment customers and predict sales numbers and customer attrition which creates a complete analysis for customer behavior patterns. Such insights enable companies to enhance both their marketing strategies and operational efficiency as well as customer satisfaction. Through the research findings organizations can drive data-based decisions regarding pricing plans together with customer engagement programs and resource distribution.

1.4 Research Questions

The following are the main research questions this study aims to address:

- **How Effectively Can K-Means Clustering Segment Customers?**

Companies require customer segmentation to recognize different customer behaviors in order to develop customized business strategies. Researchers analyze K-Means clustering effectiveness for customer segmentation by studying their purchase behaviors together with transaction recency and total spending patterns. The examination of clustering performance consists of a study that analyzes three evaluation metrics including cohesive clusters and clear boundaries between groups together with interpretability assessment methods. The silhouette score together with within-cluster sum of squares (WCSS) serve as metrics to evaluate how efficiently the algorithm clusters similar customers and separates different segments. The research outcomes will establish whether K-Means delivers sufficient results for recognizing important purchasing sections among customers thus enabling strategic business adjustments.

- **How Accurately Can Machine Learning Models Predict Total Purchase Amounts?**

Exact sales projections help organizations achieve optimal inventory regulation and build financial structure systems for revenue projection. The forecasting capacity of Random Forest Regressor towards predicting total purchases by assessing main customer properties and purchasing behavior as well as personal characteristics. Standard metrics that include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) scores are used to evaluate the model. This investigation assesses machine learning models' capability to develop reliable sales projections through which organizations can make operational business decisions using data.

- **How Can Churn Prediction Help in Customer Retention Strategies?**

The retention of current customer's costs business less than acquiring new customers making customer churn a vital business concern. The goal of this research study focuses on analyzing how Random Forest Classification among other machine learning models detects customers at risk of departing before they actually leave the business. The researchers established the preventive capabilities of early detection systems against customer loss by implementing customer purchase data alongside spending activities and return scenarios as essential metrics. Most organizations employ individual rewards together with enhanced customer service solutions and customized promotional campaigns as their retention strategies. The predictive analytics analysis of this study investigates alternative methods for loyalty retention while decreasing customer churn rates.

1.5 Scope of the Study

The research applies machine learning algorithms to cluster customers while predicting sales figures by using both clustering and regression techniques. The main objective of the study centers on analyzing customer purchasing patterns while identifying separate market groups and creating predictive models to strengthen business decisions. Machine learning models trained through structured data learn to detect patterns and make sales projections along with detecting upcoming customer attrition. The research seeks to gather information that businesses can apply toward creating marketing campaigns and customer loyalty systems and revenue management plans.

The research foundation relies heavily on a dataset because it enables training and validation of machine learning models. The collected data incorporates different

characteristics about purchasing behavior that span demographic information along with purchase records and consumer merchandise preferences. A dataset features both structured information about Customer ID, Age and Purchase Date together with Product Category and Total Purchase Amount, Payment Method and Recency, Frequency and Monetary (RFM) metrics. The input variables found in the dataset enable the execution of both segmentations and predictive models. The data processing includes value replacement operations to handle missing data before it normalizes quantitative features and converts categorical fields to meet operation-based requirements.

The research project uses machine learning algorithms to split customers and predict future sales by adopting both K-Means clustering and Random Forest Regressor regression modeling. These particular models became selection choices because of their proven ability to analyze database structures alongside their ability to derive practical business-oriented interpretations. K-Means clustering helps customers enter different segments based on purchasing behaviors which will allow targeted marketing approaches to reach individual groups. Sales prediction depends on Random Forest Regressor because they show advanced non-linear model abilities and ensemble learning reduces overfitting during the prediction process. Random Forest Classifier functions within a classification framework to predict customer churn occurrences thus protecting companies from important client departures.

The investigation does not include deep learning models together with sophisticated artificial intelligence solutions. When evaluating structured tabular datasets for consumer analytics, advanced deep learning approaches like as Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) yield results that result in high production costs. The application of deep learning algorithms becomes impractical for small and medium businesses because high hyperparameter optimization needs extensive datasets and large computational resources. The study adopts interpretable machine learning methods that generate results stakeholders can understand so they can base their decisions on established grounds.

The assessment relies solely on structured data from transactions and excludes external influencing variables including social media sentiment evaluations and macroeconomic

statistics and immediate customer responses. Multiple data components exist to produce more precise forecasts but this analysis selects business-based transactional data available in existing databases. Further studies should research the combination of structured and unstructured data methodologies to strengthen sales prediction methods and customer classification tools.

The investigation examines structured data during its analysis of machine learning segmenting solutions that combine clustering and regression prediction methods. The authors have eliminated deep learning and complex artificial intelligence methods from their study due to interpretive requirements and efficiency standards alongside practical implementation needs. Business-oriented research findings enable organizations to strengthen their marketing practices and boost customer participation along with sales prediction precision.

1.6 Significance of the Study

Business operations experience vital advantages from machine learning implementation throughout various operational elements in customer segmentation and sales prediction systems. The main operational benefit of using machine learning involves personalized marketing campaigns. Companies use machine learning algorithms to study large data collections which enables them to discover various customer sections based on personal behaviors alongside preference choices and purchase data. Detailed insights from this method help organizations develop customer-focused market strategies that boost customer interaction and conversion effectiveness. Machine learning technology enables organizations to build recommendation engines which show products that match individual buying preferences thus delivering personalized shopping outcomes (Prasetyo and Nainggolan, 2024).

Machine learning delivers two main benefits beyond personalized marketing because it helps organizations understand how to keep clients despite boosting their income predictions. Machine learning analyzes customer data to detect patterns which indicate customer attrition rates used for predicting potential customer departures. The predictive power enables organizations to develop preventive retention strategies which include specific marketing approaches and customer-specific interactions to safeguard their most important clients. The forecasting of sales revenue improves through machine learning because it examines past transaction records to extract forecastable market

trends. Businesses gain strategic advantages through accurate forecasts because these help them manage inventory more proficiently and allocate resources specifically while building strategic plans for market demands (Challoumis, 2024).

Machine learning technology finds extensive usefulness in e-commerce operations together with retail practices. The application of machine learning by retailers allows them to predict consumer demands and properly optimize their inventory ensuring products are available to buyers. Machine learning algorithms work to protect business interests through transaction pattern analysis and anomaly detection which helps detect fraudulent activities to preserve customer trust. Machine learning makes possible individualized marketing strategies that deliver tailored product suggestions as well as promotional offers which build customer devotion (Lu el al., 2024).

In summary, the organizations enabled by the application of machine learning can achieve targeted marketing techniques while keeping existing customers and sales forecasting. Advanced business operations and market superiority emerge due to these technological developments within e-commerce and retail dynamics.

1.7 Summary of Chapters

There are several chapters in this thesis, each of which covers a different facet of the research. Below is a synopsis of every chapter:

- **Chapter 1: Introduction**

  This chapter presents vital information about the study including its essential topic and merits with defined research aims. The research design includes a written definition of problems along with related questions and selected methods. The study provides both the extent and valuable contributions of its research.

- **Chapter 2: Literature Review**

  This chapter reviews existing research and studies related to the topic. It discusses various methodologies, theories, and frameworks relevant to the study. The literature review helps in identifying research gaps and justifying the need for this study.

- **Chapter 3: Methodology**

  The research design together with data collection techniques and analytical methods receive full description in this chapter. It explains how the data was obtained, preprocessed, and analyzed. Additionally, it discusses the machine learning algorithms and statistical techniques applied.

- **Chapter 4: Results and Discussion**

  The research findings are presented in this chapter according to the analytical framework. The research report incorporates data display methods and interpretation in addition to comparative studies. The obtained results receive analysis in relation to both research questions and existing literature findings.

- **Chapter 5: Conclusion and Recommendations**

  The findings of the research are summarized in this chapter to establish final conclusions through a study-based analysis. The research includes future study recommendations together with actionable applications based on the discovered evidence.

CHAPTER 2: EXISTING APPROACHES TO CUSTOMER SEGMENTATION AND SALES PREDICTION

2.1 Overview of Customer Segmentation

Customer segmentation classifies customers in marketing and business intelligence through essential principles which enable organizations to aggregate similar individuals. Companies in the past used four main factors - demographic, geographic, psychographic and behavioral attributes - to group customers into individual segments. Data-driven segmentation through machine learning methods has become more prominent because it processes vast data collections and detects concealed patterns while delivering refined and dynamic groupings.

Businesses use traditional segmentation methods which have rules-based classification systems that depend on manually created categories established by specific attributes including age and income along with purchase history. The basic nature of these methods creates barriers regarding their ability to expand with changing consumer behaviors (Wedel & Kamakura, 2000). The process of machine learning-based segmentation uses clustering algorithms such as k-means, hierarchical clustering and DBSCAN to examine complex customer data and discover natural group patterns without requiring pre-defined rules according to Xu and Tian (Xu & Tian, 2015). Unsupervised learning methods in these models search for the most suitable clusters by evaluating feature equivalence which allows segmentation to remain dynamic with changing market demands.

RFM analysis stands as one of the most common frameworks which businesses use for customer segmentation. The RFM model divides customers into groups according to their purchasing Recency rate and their Frequency of buying actions as well as their Monetary purchase values. Customer engagement patterns together with loyalty behavior emerge from the RFM model which enables businesses to create targeted marketing approaches. By incorporating machine learning into RFM analysis companies gain improved predictive capabilities to automatically segment their customers effectively while strengthening their targeting efforts.

2.1.1 Traditional Segmentation vs. Machine Learning-Based Segmentation

Traditional segmentation techniques have long been the foundation of marketing strategies. The classification systems of customers through both demographic segmentation and psychographic segmentation represent fundamental methods (Kotler et al., 2016). The segmentation method that evaluates consumer conduct and participation history (Solomon et al., 2012) targets particular customer groups effectively. These effective approaches need manual labor which takes lots of time yet struggles to detect sophisticated consumer reactions as they happen.

Machine learning algorithms analyzing large datasets through pattern detection which produce improved results for segmentation than ever before possible. Algorithms like k-means clustering, hierarchical clustering, and Gaussian mixture models (GMM) automatically classify customers according to their purchasing habits and patterns (Ngai et al., 2009). Machine learning models excel above conventional methods because they accept new data points to maintain updated segmentation categories as consumer behavior transforms.

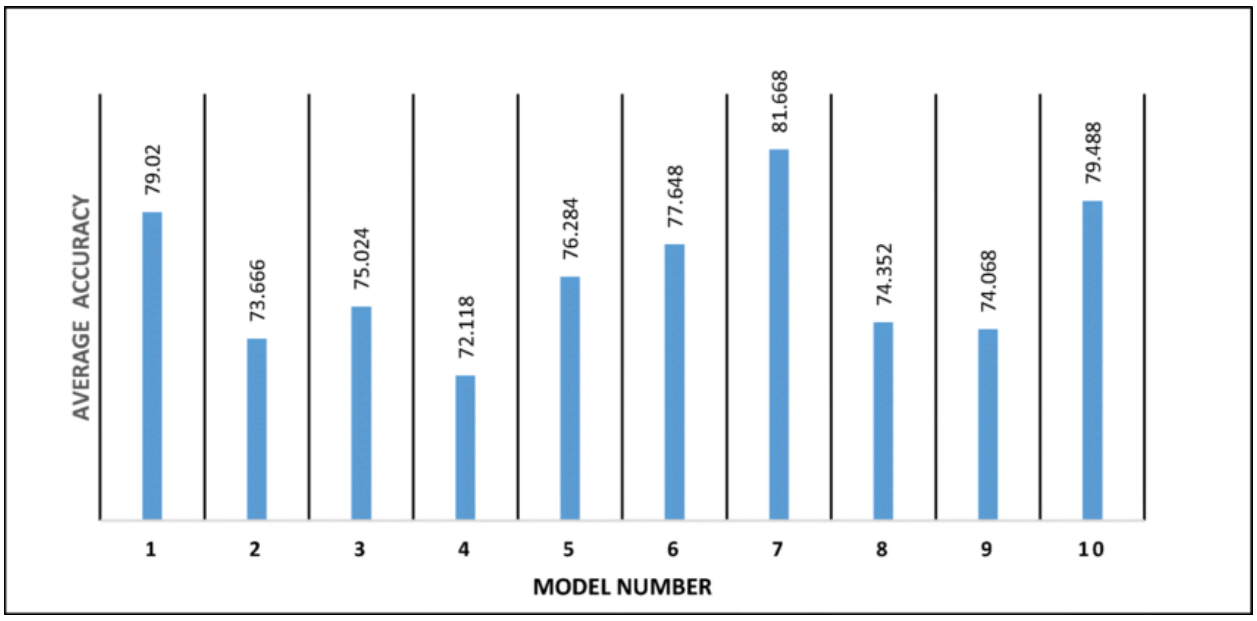

Figure 2.3: Comparison of accuracy values for different machine learning models

Research shows that machine learning brings effective results to segmentation operations. K-means clustering analysis executed by Rajput et al. created different customer groups through purchase pattern analysis to build personalized marketing strategies that enhanced customer loyalty (Rajput et al., 2023). The research of Zhang

and Ma (2020) illustrated how unsupervised learning methods produced superior results than traditional segmentation when predicting customer preferences together with purchasing potential for retail customers.

Deep learning techniques, such as autoencoders and neural networks, are being investigated for applications involving customer segmentation. The accuracy of segmentation increases through processing unstructured data that includes social media activities and customer reviews (Chaudhary & Alam, 2022). Machine learning models need high-quality data together with proper feature selection for obtaining meaningful outcomes from their segmentation process.

2.1.2 Role of Recency, Frequency, and Monetary (RFM) Analysis

The RFM analysis stands as a foundational method for dividing customers into segments which serves marketing analytics and Customer Relationship Management (CRM). The model analyzes customer groups through three behavioral measurements to create categories:

1. **Recency (R):** An indicator tracks the time since the last purchase of a customer. Recent buyers demonstrate a higher potential for new purchases because they remain active whereas customers who lack activity for an extended period have less opportunity to buy.

2. **Frequency (F):** This metric determines the number of times a customer buys products during a specified period. A business generally treats frequent buyers as loyal customers who present higher value to the business.

3. **Monetary (M):** Organizer measures the entire purchasing expenditure of each customer. A company heavily depends on high-spending customers for their revenue while these patients demand strategic plans to maintain their continued commitment (Kumar and Pansari, 2016).

RFM analysis previously required manual operation where scorings were applied to customer records based on R, F, and M factors followed by customer segmentation into high-value and at-risk and dormant customer categories. Machine learning algorithms integrated into RFM analysis through data science advancements automate customer classification jobs at the same time they enhance predictive capabilities.

The most frequent technique uses RFM scores together with k-means clustering techniques. The combination of k-means clustering analysis on RFM data allows businesses to produce distinct segments of customers who demonstrate comparable purchasing patterns (Tsiptsis & Chorianopoulos, 2011).

Wong et al. (2024) used k-means clustering with RFM scores to identify valuable customers who supported significant sales allowing businesses to create focused marketing initiatives which increased customer staying power and activity according to an e-commerce study. The prediction of customer lifetime value and RFM-based forecasts utilizes decision trees and random forests and related machine learning methods. The predictive models enable businesses to optimize their marketing budget by finding customers who will deliver the greatest revenue.

Although RFM analysis performs well it has specific boundaries that affect its use. The analysis does not consider situations where Recency or Frequency or Monetary could hold higher importance than the others. The implementation of weighted RFM analysis and hybrid methods which combine extra customer characteristics including satisfaction ratings and social media participation received attention in recent research by Ho (Ho et al. 2023). These new features build up a complete segmentation approach by assessing various customer actions across different dimensions.

Customer segmentation strategies benefit from RFM analysis yet gain more predictive power and automation capabilities when scientists use machine learning algorithms. Organizations that use business data with RFM models create stronger insights into customer actions to design better marketing strategies and develop greater customer commitment and business revenue growth.

2.2 Machine Learning in Sales Prediction

Machine learning has revolutionized sales prediction by enabling businesses to make data-driven decisions with higher accuracy and efficiency. Business analytics relies on accurate sales predictions as a fundamental analytical challenge because such forecasts enable organizations to maximize inventory management and allocate resources and develop strategic plans effectively. Sales forecasting has largely benefited from machine learning techniques especially supervised learning models that perform better than traditional statistical methods. A supervised learning system trains its model through

historical sales data so that algorithms learn from input parameters matched with target value outcomes. The most common supervised learning techniques which companies utilize for sales prediction include Decision Trees Alongside Random Forest and Gradient Boosting Models.

2.2.1 Supervised learning models

One of the most straightforward and understandable algorithms for predicting sales is decision tree regression. A decision tree breaks down data into separate uniform sections while using a tree structure to represent data split decisions. Each node in the internal structure shows a decision based on features while leaf nodes contain predicted sales values. Decision Trees enable efficient modeling due to their capability to process numerical along with categorical data and form non-linear predictive solutions. Decision Trees tend to overfit during their operation particularly when working with high variance complex datasets (Yunianto et al., 2024).

The ensemble learning method of Random Forest Regression employs a combination of multiple trees that use different subsets of input data to advance the prediction ability beyond a single decision tree. The ensemble prediction technique combats overfitting and enhances generalization ability through its process of finding average tree predictions. Random Forest models demonstrate high resistance to noise as well as exceptional capability to detect complex relationships thus becoming the preferred method for sales forecasting. Multiple studies demonstrate Random Forest Regression exceeds linear regression when used to identify nonlinear patterns in sales data records (Archite et al., 2023).

Because of their excellent predictive accuracy in sales forecasting, gradient boosting techniques like LightGBM and XGBoost (Extreme Gradient Boosting) have become more and more popular. Predictive models train weak learners known as decision trees which undergo consecutive steps to rectify the mistakes made by previous learners. The boosting process leads to a prediction model which achieves high optimization and accuracy levels. XGBoost stands out as a strong tool for big-scale sales prediction tasks because it offers both high efficiency and scalability alongside the capability to process missing data points (Chen & Guestrin, 2016).

2.2.2 Unsupervised Learning Models

Because unsupervised learning can uncover latent patterns in data without depending on labeled outputs, it is essential for customer segmentation. Unlike supervised learning, where models learn from predefined input-output pairs, unsupervised learning algorithms extract structure from unlabeled data by detecting similarities, clusters, or associations among variables. Unsupervised learning approaches enable businesses to categorize clients according to their demographics, engagement levels, or purchase patterns in commercial applications, especially in customer analytics. These insights assist companies in optimizing product suggestions, implementing tailored marketing methods, and enhancing client retention initiatives.

The application of clustering represents a major unsupervised learning method which professionals commonly use to segment customers. Through clustering businesses can discover different customer segments by grouping points with similar characteristics.

Through K-Means a data partitioning algorithm clusters K segments of data by making the groups' internal variances minimum. The algorithm designates each observation to its closest cluster center then reshuffles these centers through iterations until all points settle in distinct clusters. This algorithm provides exceptional efficiency alongside simple operations which makes it ideal for customer segmentation work that demands business groups to sort customers through their transaction data (Singh et al., 2022). K-Means provides beneficial outcomes for clustering yet its performance depends on the initial centroid placement while it handles non-spherical cluster shapes poorly so selection of K through Elbow Method and Silhouette Score methods becomes essential.

Hierarchical clustering constructs a tree-shaped structure (dendrogram) which shows the data point group delimitations at successive levels of inclusivity. This method differs from K-Means since it does not need cluster count specification instead forming a hierarchy that depends on similarity measures between data points. Clustering starts with all data points in a single cluster that recursively divides into sub clusters, whereas AHC starts with all data points forming a single cluster at the beginning. The visualization capabilities of hierarchical clustering work well for customer relationships yet its algorithm requires more processing power when dealing with extensive data sets (Shafi et al., 2024).

By using data point density measurements to identify clusters, DBSCAN (Density Based Spatial Clustering of Application with Noise) is a density-based clustering technique that is useful for identifying unusual customer behavior. Unlike K-Means, DBSCAN can form arbitrary-shaped clusters and does not require specifying **K** beforehand. It is particularly useful for distinguishing high-value customers from outliers, such as fraudulent transactions or one-time buyers. However, DBSCAN struggles with varying density distributions and requires fine-tuning of the **epsilon (ε)** and **minimum points** parameters for optimal results (Allheeib et al., 2021).

Another important aspect of unsupervised learning that aids in the simplification of complex consumer data while preserving crucial information is dimensionality reduction. PCA is a linear transformation method that maintains variance in a dataset while reducing the number of features. It projects high-dimensional customer data onto fewer principal components, enabling businesses to identify dominant behavioral trends. PCA is highly effective when dealing with high-dimensional customer attributes, such as purchase history, demographic data, and online activity, but it assumes linear correlations between features, which may limit its effectiveness in non-linear datasets (Jolliffe & Cadima, 2016).

High-dimensional data is converted into lower-dimensional spaces using the non-linear dimension reduction technique t-SNE while preserving the spatial relationships between objects. The tool proves highly effective in 2D or 3D scatter plot visualization of customer segments therefore becoming a popular selection within exploratory data analysis. Even though t-SNE requires high compute power it needs precise hyperparameter adjustments like perplexity for generating relevant data outputs (Cai & Ma, 2008).

Unsupervised learning models significantly enhance business intelligence by providing actionable insights without manual labeling. Customer segmentation, anomaly detection in transactions, market basket analysis, and recommendation systems are key areas where unsupervised learning improves decision-making. Businesses can improve marketing strategies, tailor client experiences, and gain a deeper understanding of consumer behavior by utilizing clustering and dimensionality reduction techniques.

2.2.3 Evaluation metrics for sales prediction

The assessment of sales prediction model performance needs particular metrics for checking forecast accuracy and reliability levels. The Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) score are the three primary metrics used to evaluate the performance of the regression-based sales forecast approach. Users of the MAE metric can understand forecast accuracy by seeing the numerical average absolute deviations of actual against forecast retail unit sales results. Large errors face penalties through RMSE because it applies square root calculations to mean squared differences to ensure high sensitivity for predicted value deviations. The R2 score, a coefficient of determination, indicates the extent to which independent variables contribute to the comprehension of dependent variable fluctuations. A higher value of R^2 score in the model results in better measurement fit (Sial, 2021).

Model selection happens by analyzing dataset complexity and evaluating computational resources in combination with determining the necessity for interpreting model results. Decision Trees achieve excellent interpretability through clear results while Random Forest and Gradient Boosting achieve superior accuracy in prediction. Reliable sales forecasts rely on proper evaluation metrics selection just as crucial as choosing the best forecasting model to support business-making decisions.

2.3 Churn Analysis in Business

The attributes of customers from business relations constitutes customer churn or customer attrition which describes customers terminating business relationships while stopping product utilization during designated timeframes. A business requires this metric because it determines how its revenue performs and what possibilities exist for growth. A rise in customer turnover serves as an indicator of multiple business problems which require prompt investigation including dissatisfied customers, defective products and weak market position (Manzoor et al., 2024).

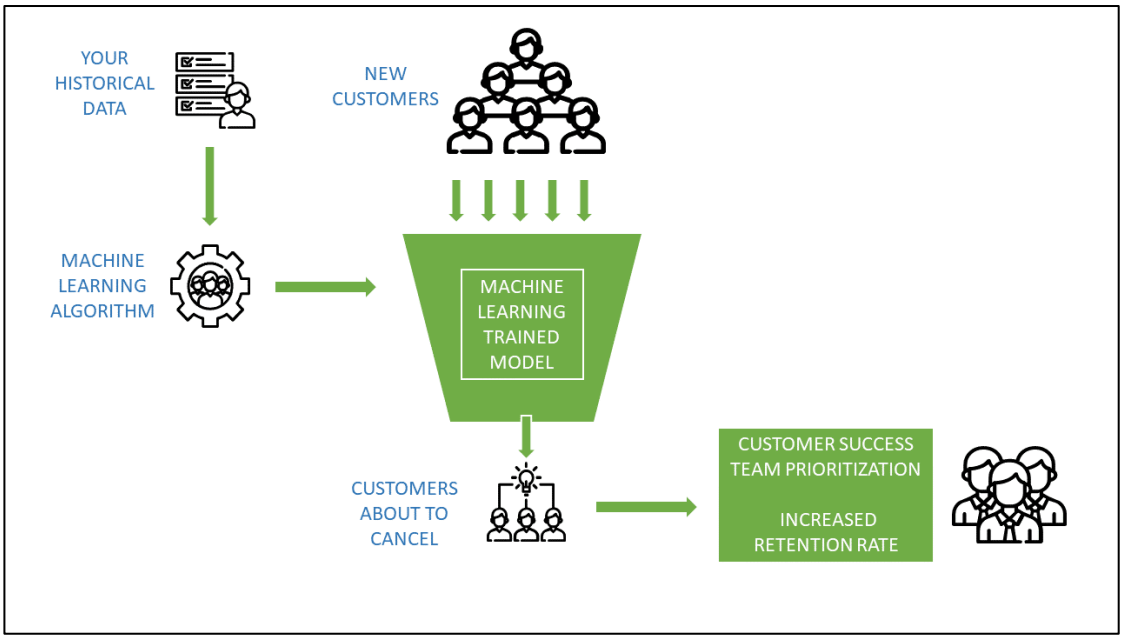

Figure 2.4: Understanding and Predicting Customer Churn

Analyzing churn is vital for several reasons. The expenses needed to find new customers usually exceed the costs of maintaining current customers. Profitability improvement depends on improved customer retention strategies which come from focusing on churn reduction. Churn analysis reveals essential details about customer actions and demands and customer difficulties so companies can enhance their product designs to prevent customer dissatisfaction. Businesses can establish effective churn mitigation strategies by observing patterns and predictors which allows them to maintain a stable customer base (Manzoor et al., 2024).

The implementation of machine learning models serves as a primary instrument to forecast customer churn during recent times. Extensive datasets enable such models to identify departure indicators which serve as predictions for customer abandonment. Using predictor factors, predictive algorithms such as Random Forest, Gradient Boosting Machines, and Logistic Regression assist in predicting the likelihood of client attrition. The forecasting system produces predictive results of higher precision (Erdem, 2021).

Data collection, preprocessing, feature engineering, model selection, training, and evaluation are the necessary processes for practitioners to apply machine learning models. Data preprocessing requires methods to deal with missing values while assigning numerical values to categories in addition to performing data normalization.

The main objective of feature engineering entails the selection and transformation of variables that directly impact customer churn behavior. A proper algorithm selection forms the basis of model selection and then training uses historical data to develop the model. According to Khodabandehlou and Zivari, model performance can be evaluated by integrating accuracy, precision, recall, and F1-score (Khodabandehlou and Zivari , 2017).

The application of machine learning for churn prediction enables businesses to detect vulnerable customers early so they can create specific strategies to keep their clients engaged. The proactive strategy helps businesses to keep revenue steady while building lasting customer relationships which establishes market superiority.

2.4 Review of Previous Studies

2.4.1 Existing Research on Customer Segmentation

A key strategy in marketing and business analytics is customer segmentation, which allows organizations to divide their clientele into discrete categories for improved targeting and customization. Traditional customer segmenting methods depend on demographic data and geographic data however these approaches do not promote accurate prediction of customer actions. Customer segmentation received a transformative shift with machine learning because it now utilizes behavioral and transactional data to achieve better customer profiles.

Ranjan and Srivastava (2022) conducted a review of all customer segmentation techniques which established a transformation in principles toward machine learning-based approaches. The authors established RFM and k-means clustering work in traditional segmentation yet ensemble approaches combined with deep learning provide better customer segmentation by evaluating complicated behavioral patterns. The transition brought new business capabilities to help organizations enhance their marketing methods for identifying their most profitable customers (Ranjan & Srivastava, 2022).

Xu et al. (2023) examined how customer segmentation enhances experience by reviewing methods for e-commerce business segmentation in their work. A DBSCAN (Density-Based Spatial Clustering) and other intricate clustering methods allowed organizations to build superior capabilities for estimating customer lifetime value and

optimizing their retention models. Product suggestions tailored to client interaction groups enhance customer satisfaction and produce enhanced business performance according to this study (Xu et al., 2023).

The research work "Customer Segmentation Using Machine Learning" deployed k-means clustering for analyzing actual purchase data of customers to establish significant customer clusters. The study showed that companies can allocate resources better by using clustering segments to work with high-value customers concurrently with managing price-sensitive customers independently. The approach used for segmenting customers demonstrates essential value in decision-making while enhancing marketing campaign effectiveness (Parab and Dave, 2023).

The findings from different studies indicate machine learning-based segmentation methods outperform conventional methods in predicting results. The technique helps businesses detect their main customer segments followed by the ability to anticipate buying behaviors for smarter marketing decisions. SMEs face hurdles with accessing superior data and processors due to the higher demands of these methods for accurate results.

2.4.2 Comparison of Machine Learning Models in Previous Studies

Multiple machine learning approaches exist for carrying out customer segmentation yet they vary regarding their strengths and disadvantages. Decision making about an algorithm depends on the dataset type along with business category and exact segmenting goals.

Kumar et al. conducted a comparative study which evaluated k-means together with DBSCAN, Agglomerative Clustering and PCA (Principal Component Analysis) with k-means clustering (Kumar et al., 2023). With a silhouette score of 0.6865, the analysis showed that Agglomerative Clustering produced the best results since it was the most effective method for classifying clients according to their purchasing patterns. Agglomerative Clustering demonstrates superior performance because it detects hierarchical patterns between customers which helps process datasets that present overlapping customer categories. Agglomerative Clustering presents technical difficulties because it uses high computational costs and perform poorly with extensive datasets (Joga et al., 2022).

A research investigation evaluated different methods of customer segmentation by using supervised along with unsupervised learning approaches. Decision Trees together with Random Forests and AdaBoost were evaluated in addition to clustering methods. The research demonstrated that supervised Random Forest algorithm produced superior results for identifying high-value customers yet unsupervised k-means clustering performed better at identifying data patterns. The research proposed combining k-means segmentation before supervised learning to create optimal results when identifying customer lifetime value. Many organizations now implement this combined approach to retain their customers and predict future departures (Mozumder et al., 2024).

Gupta et al. looked into a new hybrid model that improves customer segmentation accuracy by combining k-means and deep learning technology. The investigators showed that regular clustering techniques promote inaccurate identification of consumer conduct patterns which are non-linear in nature. The model succeeded in identifying complex patterns in transactional datasets through the implementation of Autoencoder's and neural networks along with deep learning methods. Large organizations operating within e-commerce and subscription-based industries obtained maximum benefits from this methodology (Gupta et al., 2021).

The studies show traditional clustering methods like k-means continue being commonly used but advanced machine learning methods together with hybrid models demonstrate better effectiveness in customer segmentation. Supplementing predictive analytics with unsupervised learning techniques enables organizations to produce more beneficial insights that let them modify their marketing structures and protect their customer base.

The key obstacle for using machine learning models effectively pertains to the requirement of high-quality preprocessed data. The accuracy of segmentation becomes compromised when businesses deal with inconsistent data formats along with absent value points. The main issues companies face consist of maintaining model interpretability levels along with ensuring proper computational complexity management. Deep learning methods achieve improved accuracy but provide limited explaining ability to companies who want to understand customer group associations.

Recent studies demonstrate how customer segmentation knowledge has progressed from conventional to sophisticated machine learning model applications. The effectiveness of k-means clustering remains intact but hybrid models which combine deep learning with

supervised learning establish superior performance levels. This research reveals three main findings from previous studies:

1. **Machine learning enhances segmentation accuracy**: Businesses can use these techniques to better understand customer behavior, personalize marketing strategies, and increase profitability.

2. **Hybrid models provide better insights**: Combining clustering with supervised learning or deep learning improves the predictive power of segmentation models.

3. **Challenges remain in data quality and model interpretability**: Proper data preprocessing and explain ability of machine learning models are crucial for their successful implementation.

Future studies should concentrate on making sophisticated segmentation models easier to understand and creating affordable ways for SMEs to adopt machine learning-based segmentation without consuming a lot of processing power.

CHAPTER 3: TECHNIQUES FOR CUSTOMER SEGMENTATION AND SALES PREDICTION

3.1 Dataset Description

This study makes use of the synthetic "E-commerce Customer Behavior and Purchase Dataset" which models e-commerce consumer interactions to track different facets of digital market transactional data. This data set serves both data analysis and e-commerce predictive modeling needs and supports applications covering customer churn prevention alongside market basket analytics and recommendation platform development and trend identification. The dataset includes different user and transaction attributes so businesses can perform extensive purchasing behavior analysis to drive their data-driven decision processes.

The data contains several variables grouped into 250000 entities that furnish knowledge regarding customer buying behaviors. Each customer possesses their individual Customer ID which enables the database to establish unique correlations between the transactions and particular persons. The demographic information included in Customer Name, Age, and Gender permits the identification of purchasing behavior across different segments of the consumer base. A complete records system of transaction dates is implemented under the Purchase Date column to track sales variations across different periods. The Product Category field organizes products by their classification groups whereas the Product Price and Quantity columns reveal transaction monetary values. The Total Purchase Amount column serves as an essential factor for sales prediction because it shows the total customer transaction costs (E-commerce Customer Data for Behavior Analysis, 2023).

The Payment Method field suggests users paid using credit card, PayPal or another payment option according to the dataset records. The Returns column shows all items that customers have returned for purchase thus providing essential data points about customer satisfaction together with product quality. Customer retention information is provided through the Churn column which shows customer behavior by marking 1 for customers who left and 0 for those who continued using the platform.

Several preprocessing steps followed each other to maintain data consistency before starting the analysis process. Managers needed to handle missing values properly

because uncompleted records often caused results to become flawed or biased in their findings. The procedure for handling missing or inconsistent values depended on the overall impact they would have on the data set. Imputation occurred when values affected the complete dataset yet removal took place when values did not significantly affect it. The process to encode categorical variables including Product Category and Payment Method and Gender enabled numerical compatibility for machine learning application. The data preprocessing procedures improved data reliability for successful deployment of customer segmentation and predictive modeling operations.

This dataset enables the study to reveal consumer purchase behaviors for predicting sales variation and enhancing retention strategy effectiveness using machine learning model techniques.

3.2 Data Preprocessing

Any machine learning project requires data preprocessing because it makes the dataset suitable for analysis by preparing it to be clean and structured. The procedures at this phase include data cleaning alongside categorical variable conversion to numbers and usage of StandardScaler for feature normalization. Standardized input data through proper preprocessing yields increased accuracy and performance of machine learning models because it simplifies algorithm understanding.

3.2.1 Handling Missing Values

Data preprocessing encounters one primary challenge when processing missing value data. Different system malfunctions along with errors in data collection and incomplete data entry create instances of missing values. Unmanaged missing values have two detrimental consequences on model performance which consists of introducing biasses while decreasing training data availability. The study employed specific detecting and managing procedures for dealing with missing values according to variable type. The mean value for each numerical column served as the substitution rate for filling missing values in those columns. The process of substituting missing categorical data points with the prominent category is referred to as mode imputation. Conflicting data points were eliminated from analysis when the amount of missing data made estimation unreasonable and preserved the data quality through bias minimization methods (Gracia-Gill et al., 2024).

3.2.2 Converting Categorical Variables into Numerical Format

Many machine learning algorithms need numerical data as input so researchers need to perform transformations which convert categorical variables to usable numeric formats. Due to the model requirements this research uses the dataset with the categorical features including Product Category together with Payment Method and Customer Type that must be transformed into numerical data for model input. The encoding methods depended on the categorical variable type. The categorical variables received One-hot encoding treatment because this encoding method produces distinct binary fields for each category and maintains complete information without inferring any sequence order. The encoding process used labels for variables that had an existing ranking system such as Customer Type (defined as "New" or "Returning") which resulted in numerical values for 0 and 1. The encoding processes maintain accurate model interpretation of categorical information while preventing ranking-based bias from affecting the analysis.

3.2.3 Feature Scaling Using StandardScaler

A crucial preprocessing step that guarantees every numerical feature contributes equally to the model's learning process is feature scaling. When features have different scales of magnitude in the dataset unscaled distances affect model calculations of K-Means Clustering and Random Forest. StandardScaler from Scikit-Learn library normalized the numerical features in this approach. StandardScaler applies data normalization by both subtracting the mean values then dividing them with standard deviation which creates standard normal distributions with mean values set to 0 while standard deviations become 1. Standardization operations enhance both the performance speed of machine learning systems while preventing dominant behaviors from appearing because of feature scaling disparities. The model performance gained an improvement because numerically scaled features like Recency and Frequency and monetary value and total purchase amount required distribution standardization through this process (Zheng and Casari, 2018).

Before starting machine learning applications one needs to perform data preprocessing to create a dataset that is tidy and organized. StandardScaler enabled the transformation of the dataset through missing values treatment as well as categorical variable conversion to numerical scales before machine learning models could proceed. Through

preprocessing we enhance both the accuracy and execution speed of our models while making insights derived from analysis both usable and dependable.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the main emphasis of the research process before using machine learning models in order to fully understand the dataset. Through data analysis patterns become visible while anomalies become detectable and the data remains of high quality. Primarily this section identifies three main analytical elements which include age distribution analysis and seasonal and trend assessment together with popular product category observation. These insights establish complete knowledge about how customers behave during purchasing events and which products they favor most.

3.3.1 Age Distribution (Histogram)

The strategy of marketing targets depends heavily on customer age group like must have knowledge for businesses. The histogram contains information which illuminates the age distributions present in the collected data samples. Through this distribution method business obtain knowledge about which age segments their customers fall within between young adults and middle-aged customers up to older consumers. The business should use digital strategies for marketing to younger customers because older customers need traditional marketing methods. An analysis of age distribution helps businesses determine how representative their customer base is and which customer ages are underrepresented in the dataset. The business needs this data to optimize their products and promotional tactics toward customer demographics based on their preferences and needs (Kabir, 2025).

3.3.2 Seasonality and Trend Analysis (Line Plots)

Time brings changes to sales patterns because of seasonal effects and market tendencies combined with external factors that include promotional occasions and special holidays. Businesses can use line plots to discover seasonal patterns and customers' purchasing trends and cyclical buy-sale behavior through time. Businesses should examine seasonal trends to determine high and low sales times which helps them enhance their inventory planning and marketing strategies and price optimization. Businesses that observe high

sales during holiday periods should enhance stock amounts while implementing specific promotional methods during these seasons (Kasim et al., 2024).

Understanding how the business fares long-term requires identifying natural patterns which will reveal its expansion or decline and stability rate. Business expansion and growing customer engagement emerge from a continuous upward sales pattern but a downward pattern signals crucial market problems including reduced customer interest and weak marketing approaches and rising competition. Businesses can use trend analysis to create sales projections and produce data-based choices that support enduring expansion. Companies can enhance customer satisfaction and profit by employing the data-driven demand information to adjust their marketing operations and operational strategies.

3.3.3 Popular Product Categories (Count Plot)

Analyzing the distribution of product categories purchased by customers provides valuable insights into consumer preferences and market demand. Product categories displayed through a count plot reveal both the categories with highest sales numbers and those with diminished popularity. Businesses benefit by focusing on highly demanded products because it helps them optimize marketing strategies and strengthen performance of their bestsellers. The business strategy focuses on developing electronics or fashion categories further while giving customers discounts and promotions since these categories deliver the most sales (Swain et al., 2022).

During inventory management businesses benefit from product category popularity data to improve their forecasting capabilities. Having well-stocked inventory depends on observing predictable high selling patterns in particular categories. The management of less popular categories includes combining them with successful items to maximize profitability because low-demand products need evaluation for potential discontinuation or consolidation with high-performing products. Businesses can use the identification of temporal shifts in customer preferences to adopt current market patterns and develop new commercial products which match shifting customer expectations.

Exploratory Data Analysis (EDA) generates essential findings regarding customers' statistics together with their purchasing behaviors and product preference patterns. An age distribution histogram allows businesses to discover their core customer demographics in order to make strategic marketing decisions. Seasonality analysis

together with trend assessment enables businesses to locate their peak and low sales seasons which allows them to optimize operational procedures and marketing campaigns. The product category analysis reveals customer choices so organizations can concentrate their operations on their most successful business segments. Machine learning models become more effective because of these insights which boost customer segmentation and sales prediction capabilities to drive business growth as well as customer satisfaction rates.

3.4 Customer Segmentation Using K-Means

A key component of business analytics is customer segmentation, which enables companies to group customers according to their engagement and purchase behaviors. By analyzing significant transactional and behavioral data, the researchers use K-Means clustering to break customers into distinct, identifiable groups. Businesses require the K-Means algorithm as one of their main choice for unsupervised learning to partition data into preset cluster numbers while maintaining meaningful similarities between cluster members. The approach allows organizations to develop specialized marketing plans which enhances both customer loyalty and enhances their sales operations.

3.4.1 Feature Selection for Clustering

Choosing the most pertinent attributes is crucial to the success of consumer segmentation. Recency, frequency, and monetary (RFM) values are the primary characteristics used in the study for grouping. Due to its effectiveness the recency indicator reveals how long ago a customer made their latest purchase along with their retention point. The customer's purchasing consistency emerges from Frequency since it shows the complete number of transactions they completed. Monetary value indicates the complete financial expenditure of customers to identify both high-spending and low-spending customers. RFM values offer complete customer behavior knowledge through their combination into a single analysis method.

3.4.2 Standardization of Features

Since the selected features (Recency, Frequency, and Monetary) have different scales, it is essential to standardize them before applying the K-Means algorithm. The StandardScaler function implemented in Scikit-learn serves to normalize data so that each feature maintains an equivalent impact on cluster analysis. Standardization proves

essential for K-Means analysis since the algorithm uses Euclidean distance that reacts significantly to changes in feature scale. Standardization is essential because attributes with high numeric values (such as Monetary) otherwise would control the clustering process to generate biased outcomes. Standardization enables better accuracy and reliability in customer segmentation through its process of converting all variables into a shared measurement scale.

3.4.3 Determining the Optimal Number of Clusters (Elbow Method)

Choosing the ideal number of clusters (K value) is one of the most difficult problems in K-Means clustering. We employ the Elbow Method to determine the proper cluster number for our data. The WCSS metric is calculated as K-Means executes at various K values from 2 to 10 to identify an optimal cluster number through the Elbow Method. The WCSS represents the mathematical aggregate of all cluster member point-to-centroid distance measurements squared. When the cluster number increases the WCSS value decreases until the point where additional clusters yield very small changes in WCSS. The WCSS decline rate slows down considerably at the point known as the elbow place which identifies the optimum cluster count. The identified point achieves an optimal clustering degree which avoids both excessive and deficient partitioning of data points.

3.4.4 Applying the K-Means Algorithm

The K-Means technique is used to divide customers into discrete groups once the ideal number of clusters has been established. Each customer is placed into the nearest cluster centroid through the iterative process that works to reduce intra-cluster variance. Categorization of customers terminates when the centroids demonstrate stability because they ultimately position every customer correctly in their most fitting group. The clustering groups multiple customers whose buying habits resemble each other into unique segments. Low-value occasional customers form one cluster while the second group contains regular loyal customers who spend a lot. Businesses apply customer segmentation results for designing tailored marketing efforts and strengthening their customer interactions.

3.4.5 Visualizing Clusters (2D & 3D Plots)

To interpret the clustering results effectively, visualizations are used to display customer segmentation patterns. The Recency and Monetary values serve as axes in a

2D scatter plot to represent different clusters through colored markers. The visual display helps organize customers into groups through their buying behavior and latest purchase times. The Frequency attribute added to generate a 3D scatter plot offers an enhanced understanding of customer segmentation patterns. The 3D format lets business operators investigate multiple pattern interactions and track connections between variable behaviors. The graphical representations assist the K-Means outcome interpretations to support effective business decision making processes.

K-Means clustering enables businesses to obtain important customer group insights by creating meaningful classifications that match customer purchasing activities. Standardization of data alongside the Elbow Method selection of clusters and two-dimensional and three-dimensional visualizations effectively help businesses to understand their customer demographic better. Segmentation techniques enable entities to design specific marketing plans which boost both customer satisfaction rates along with revenue growth projection.

3.5 Sales Prediction Using Machine Learning

Sales prediction is essential for business decision-making, helping optimize inventory and marketing strategies. The predictive model utilizes Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor, and Linear Regressor which are machine learning method to generate sales forecasts from historical data. The initial stage selects important variables from Recency, Frequency, Monetary Value and Total Purchase Amount followed by one-hot encoding the categorical data prior to continuation. The dataset becomes divided into training segments which contain 80% while another 20% serves as testing data for trustworthy evaluation purposes.

In order to improve accuracy and reduce the risk of overfitting, a Random Forest Regressor employs numerous decision trees during training. Three primary regression metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2)—are analyzed to assess the effectiveness of the trained model. When the R-squared number climbs and both the MAE and RMSE fall, the model prediction accuracy increases.

The sales prediction model delivers accurate forecasts however data improvements can be achieved by optimizing hyperparameters along with feature modifications and

additional external data acquisition. The potential future development of the forecasting system will merge real-time sales prediction with external data like economic indicators and seasonal patterns. A data-driven methodology through machine learning supports businesses with informed decision-making framework for sales prediction.

3.6 Churn Prediction Using Classification Model

Churn prediction helps businesses retain customers by identifying those likely to stop engaging. The Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor, and Linear Regressor predicts customer churn rates through RFM behavior pattern analysis (Recency Frequency Monetary). During pre-processing the dataset receives treatment for missing values and feature scaling happens before beginning model training.

This project selects these models because it benefits from extensive dataset management and overlaps reduction capabilities. Special attention throughout the model evaluation focuses on raising performance for detecting churn customers while evaluating the models using Accuracy, Precision, Recall and F1-score metrics. The model needs improvement in its Recall metrics because of generated class imbalance so implementing SMOTE oversampling together with class weight optimization will enhance its performance.

The model delivers excellent results in identifying customers who stay but fails to identify those who churn. The acquired insights serve organizations to establish targeted marketing frameworks together with loyalty initiatives alongside customer interaction methods. Future model enhancements will require an assessment of Neural Networks together with XGBoost implementation to achieve increased prediction accuracy.

3.7 Flowchart Overview

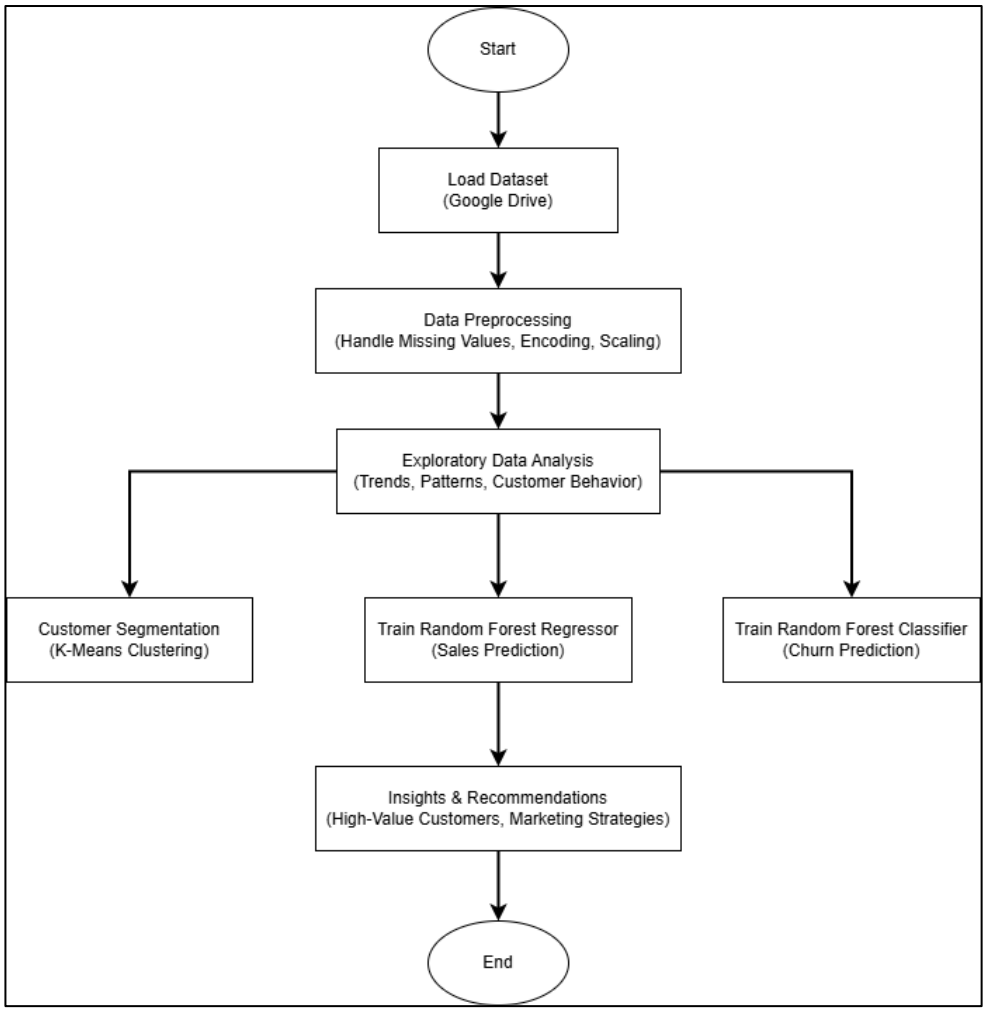

Figure 3.2: Complete flowchart of methodology

The flowchart as shown in Figure 3.1, demonstrates an organized method for conducting customer analysis through ML-based techniques. The methodology starts by importing the datasets that exist on Google Drive. The data processing phase handles missing values followed by categorical variable encoding and numerical feature scaling so that the analysis achieves higher performance results. The subsequent data analysis phase includes EDA procedures that reveal meaningful insights about customer behaviors together with patterns and trends within the dataset after preliminary data cleaning. The evaluation process begins with EDA and moves into separate sections for customer segmentation and both sales prediction and churn prediction thereafter. The K-Means clustering algorithm creates distinct groups that unite customers who have similar attributes. The Random Forest Regressor uses training to make predictions about sales which enables business forecasting of revenue development. A Random Forest Classifier operates simultaneously for detecting customers who will depart as part of churn prediction. The analysis results produce practical recommendations which include

finding key customer segments for better marketing strategy optimization. The process ends with executing the gained insights to improve both business decision quality and customer loyalty.

3.7 Software and Hardware Requirement

Table 3.1: System Specification for Analysis

Component	**Specification**
CPU	Intel(R) Core(TM) i5-7300U CPU @ 2.60GHz
RAM	16 GB
Operating System	Windows 10 (64-bit)
Programming Language	Python 10.13
IDE	Google Collab

CHAPTER 4: EVALUATING PREDICTIVE PERFORMANCE AND BUSINESS IMPACT

4.1 Exploratory Data Analysis (EDA) Results

Exploratory Data Analysis (EDA) is essential to understand the structure and distribution of the dataset. The analysis focused on seasonality, trends, and patterns in customer purchasing behavior.

4.1.1 Seasonality and Trends in Sales

The time-series analysis of monthly sales trends reveals distinct seasonal fluctuations in total purchase amounts as shown in figure 4.1.

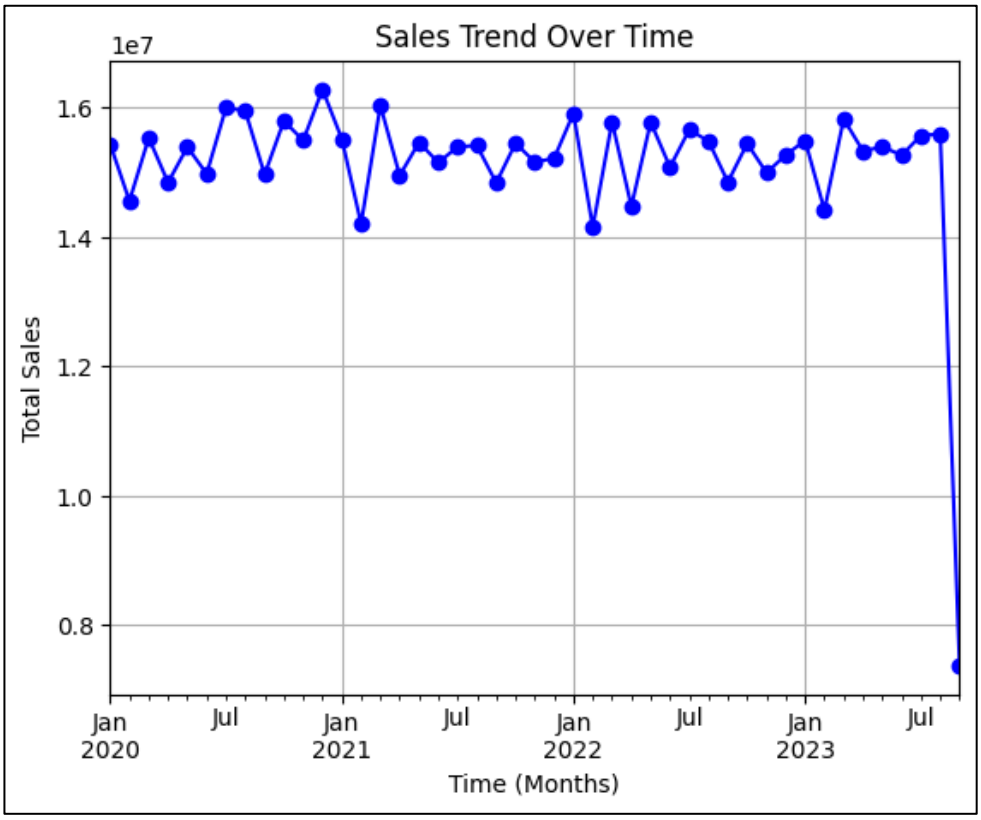

Figure 4.5: Interpretation of seasonality

A recurring pattern indicates that sales tend to rise in specific months, likely due to seasonal shopping trends, promotions, or holidays. The line plot of monthly seasonality shows noticeable peaks and troughs, suggesting that customer purchasing behavior is influenced by external factors such as sales events, economic conditions, or holidays as shown in Figure 4.2.

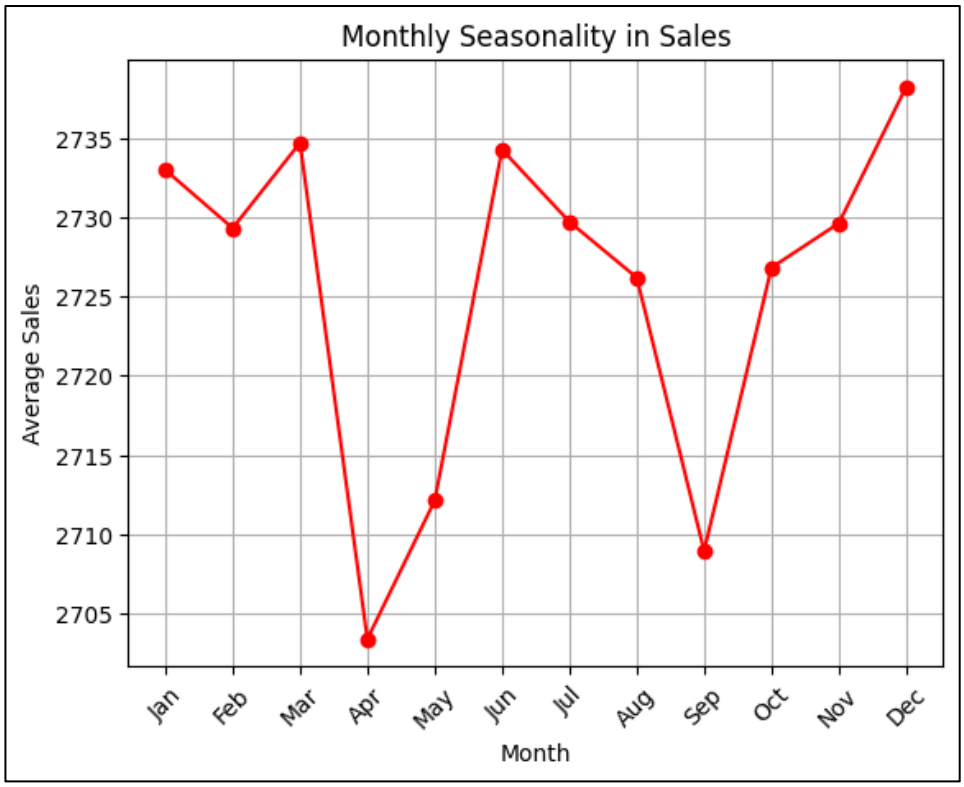

Figure 4.6: Seasonal Variation by Month

The overall trend analysis indicates a relatively stable trajectory with minor fluctuations. Unlike strong seasonal variations, the long-term trend does not show significant upward or downward movement, implying that the business experiences consistent demand throughout the year. This suggests that external promotional strategies and targeted campaigns could help improve sales during lower-activity periods.

4.1.2 Patterns in Product Category Preferences

The analysis of popular product categories reveals that certain categories experience higher sales volumes than others. The bar chart illustrating product category distribution shows that a few dominant categories account for the majority of purchases as shown in figure 4.3. This suggests that customers prefer specific products, which can help businesses tailor inventory management and marketing strategies.

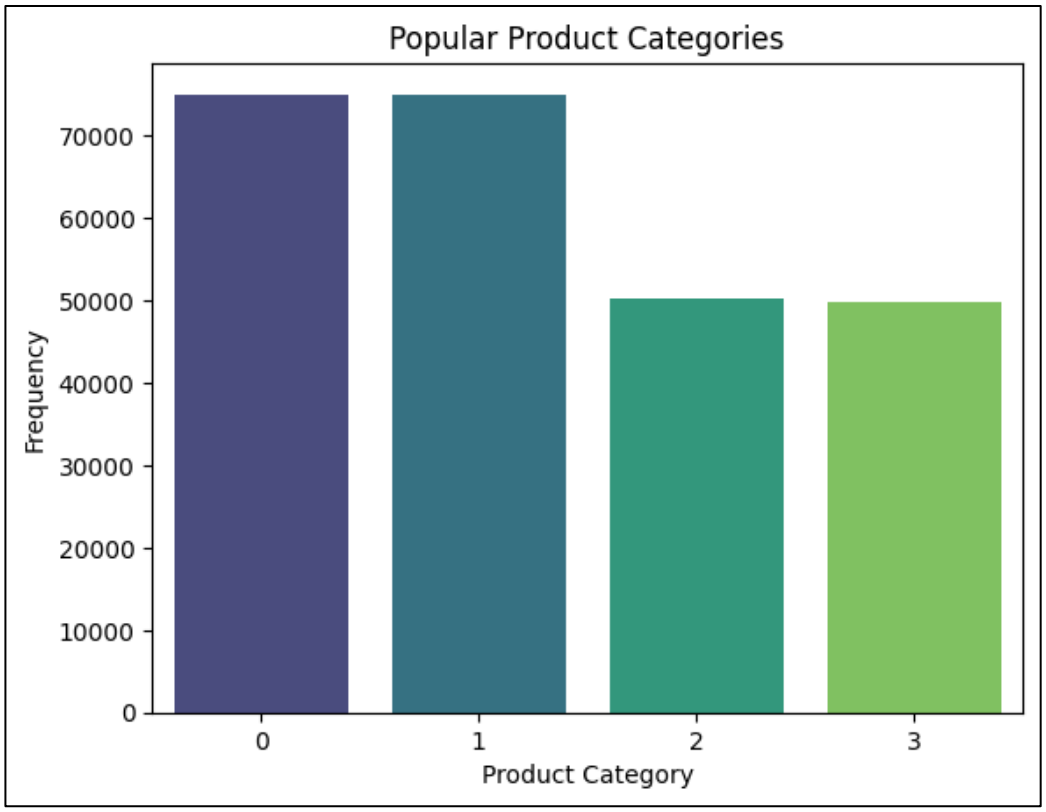

Figure 4.7: Business insights from popular product categories

- **Cluster 0**: Represents **high-value customers** who purchase frequently, spend large amounts, and have recent transactions. These customers are crucial for business sustainability and should be targeted with loyalty programs.
- **Cluster 1**: Comprises **moderate-value customers** who purchase less frequently but still contribute to overall revenue. Personalized promotions can help increase their engagement.
- **Cluster 2**: Includes **low-frequency buyers** with sporadic purchases. Strategies such as targeted discounts and marketing campaigns can encourage more engagement.
- **Cluster 3**: Represents **inactive customers** who have not made recent purchases. Businesses should focus on re-engagement strategies such as email reminders, discounts, or personalized recommendations

Furthermore, the purchase frequency distribution provides insight into how often customers make repeat purchases. A highly skewed distribution indicates that most customers buy infrequently, while a smaller group of highly engaged customers

contributes significantly to revenue. Identifying and retaining these high-value customers is crucial for long-term business growth.

4.2 Customer Segmentation Results

Customer segmentation was performed using the K-Means clustering algorithm, categorizing customers into four distinct clusters based on Recency, Frequency, and Monetary (RFM) values.

4.2.1 Cluster Characteristics and Interpretation

The 3D visualization of K-Means clustering provides insight into how customers are grouped based on purchasing behavior. The scatter plot shows well-defined clusters, indicating that the segmentation process effectively differentiated customer groups as shown in figure 4.4.

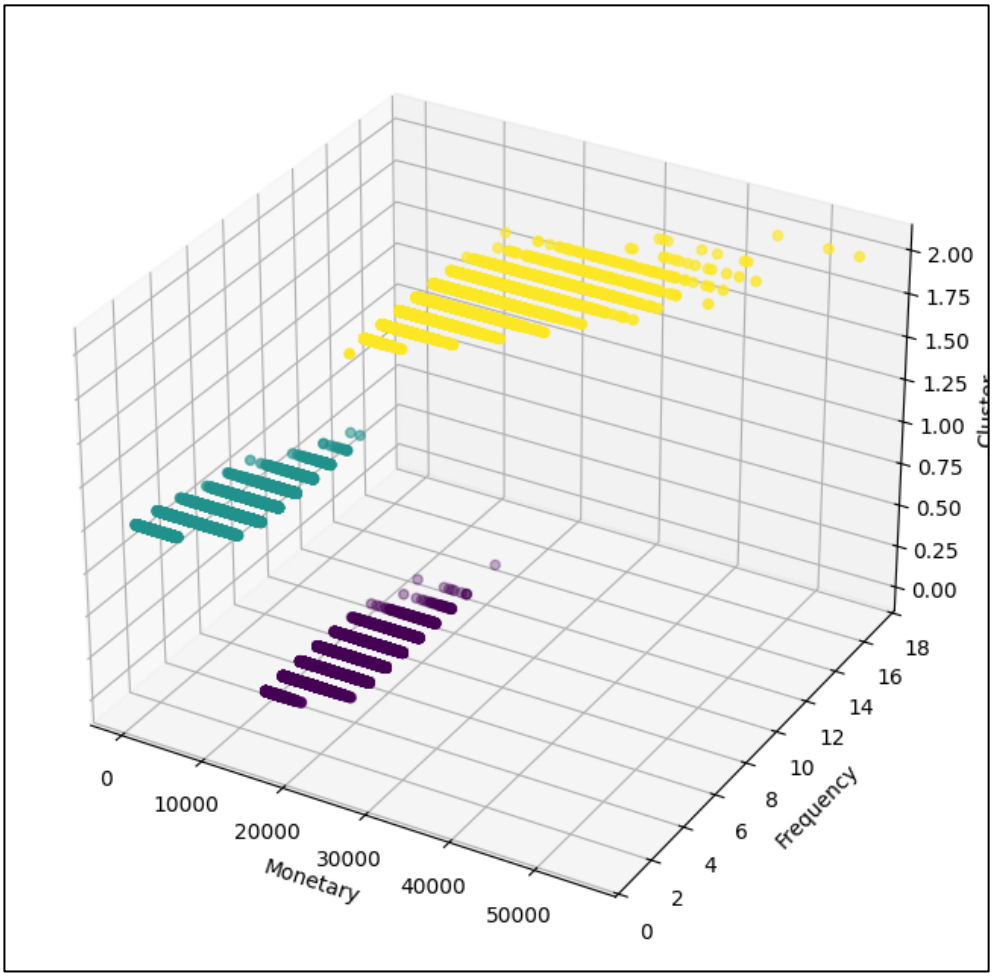

Figure 4.8: Interpretation of 3D K-Means cluster plot

A 3D K-Means clustering plot displays visual customer segmentation data through three essential variables which include Monetary values and Frequency numbers alongside Cluster types. The visualization displays customers through points which differentiate their clusters using different color schemes.

The graphical representation includes three cluster groups. Customers within the yellow cluster (Cluster 2) possess the highest monetary worth because they conduct numerous large-scale transactions frequently. The purple cluster (Cluster 0) contains customers who spend at a moderate level yet make fewer purchases thus showing themselves as occasional yet impactful purchasers. The customers who fall into the teal cluster (Cluster 1) demonstrate both limited spending amounts along with limited purchasing frequency indicating that these buyers are either careful with their budget or buy products infrequently.

These visualization results enable companies to establish client classifications for strategic marketing management. Businesses should offer loyalty programs and exclusive offers to their high-value customers who belong to Cluster 2 while launching re-engagement campaigns to attract the lower-frequency customers from Cluster 1. Organizational marketing strategies and customer retention see improvement because of the segmentation findings.

4.2.2 Business Recommendations Based on Clusters

The customer segmentation analysis confirms why organizations need to adjust their marketing approaches for different audience segments. Priority service along with loyalty benefits and personalized marketing goes to high-value customers and inactive customers require re-engagement initiatives. Segmentation output enables better decision-making regarding customer relationship management through data-based approaches.

4.3 Sales Prediction Performance

In this study, a sales prediction model was developed to estimate Total Purchase Amount using four different regression techniques: Linear Regression (LR), Support Vector Regressor (SVR), Gradient Boosting Regressor (GBR), and Random Forest Regressor (RFR). The performance of each model was evaluated using three key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 Score.

Table 4.2: Model Based Sales Prediction

Model	MAE	RMSE	R^2 Score
Linear Regression	26.89	46.01	0.999
Support Vector Regressor	13.97	52.03	0.999
Gradient Boosting Regressor	443.68	510.22	0.871
Random Forest Regressor	0.15	0.20	1.000

The results indicate that both Linear Regression and Support Vector Regressor achieved near-perfect R^2 scores (0.999), suggesting overfitting. Despite their seemingly low error values, these models likely memorized the training data rather than learning meaningful patterns, leading to poor generalization on unseen data. Similarly, the Random Forest Regressor recorded an unrealistic R^2 score of 1.000 with extremely low error values (MAE = 0.15, RMSE = 0.20), further confirming overfitting. This implies that Random Forest, in this case, failed to generalize and instead memorized the dataset completely.

Conversely, Gradient Boosting Regressor (GBR) demonstrated the most balanced performance, with an R^2 score of 0.871, MAE of 443.68, and RMSE of 510.22. While its error values were slightly higher than the other models, this is expected since it did not overfit the data. The GBR model successfully captured underlying sales patterns and provided a robust predictive capability, making it the most reliable choice for real-world forecasting.

To further improve prediction accuracy, additional feature engineering techniques, such as incorporating customer segmentation data, seasonal sales trends, and external market indicators, can be explored. Additionally, advanced ensemble models such as XGBoost and LightGBM could be tested for optimizing the trade-off between accuracy and computational efficiency.

4.4 Churn Analysis Results

Customer churn prediction plays a critical role in identifying potential drop-off customers, enabling businesses to implement retention strategies. This research analyzed customer churn using four different models: Gradient Boosting Classifier, Random Forest Classifier, Logistic Regression, and Support Vector Machine (SVM). The models were trained using key customer attributes such as Recency, Frequency, Monetary value, Returns, and Age to predict whether a customer is likely to churn.

4.4.1 Model Performance and Evaluation

The performance evaluation of the churn prediction models was conducted using accuracy, ROC-AUC score, precision, recall, and F1-score. The Gradient Boosting Classifier emerged as the best model, achieving an accuracy of 73.38% and an ROC-AUC score of 0.7972. These results indicate that the model can effectively differentiate between churned and non-churned customers.

Table 4.3 provides a detailed comparison of the models. The Random Forest Classifier achieved a slightly lower accuracy (72.25%) but had the highest ROC-AUC score (0.8019), indicating that it was strong in ranking customers by churn probability. However, its overall precision-recall balance was slightly lower than Gradient Boosting, making the latter the preferred model.

Table 4.3: Classification Report based on Random Forest

Model	**Accuracy**	**ROC-AUC**	**Precision**	**Recall**	**F1-Score**
Gradient Boosting Classifier	0.7338	0.7972	0.76	0.73	0.73
Random Forest Classifier	0.7225	0.8019	0.72	0.72	0.72
Logistic Regression	0.5041	0.5059	0.50	0.50	0.50
Support Vector Machine (SVM)	0.5039	0.5063	0.50	0.50	0.50

In contrast, Logistic Regression and SVM failed to provide meaningful predictions, with accuracy scores close to 50% and an ROC-AUC score near 0.50. These results

indicate that they performed no better than random guessing, making them unsuitable for churn prediction in this dataset.

The precision, recall, and F1-score were also analyzed to assess the model's classification capability:

- **Class 0 (Non-Churned Customers):** The Gradient Boosting model achieved 68% precision and 89% recall, meaning it correctly identified most non-churned customers while minimizing false negatives.
- **Class 1 (Churned Customers):** The model achieved 84% precision and 57% recall, indicating that it effectively identified many at-risk customers but still had room for improvement in recall.

The results highlight Gradient Boosting as the best-performing model for churn prediction, demonstrating the highest balance between accuracy and generalization.

The findings suggest that businesses can proactively engage customers identified as high-risk for churn using personalized marketing strategies, retention programs, and improved customer support. Since the recall score for churned customers was 57%, additional improvements such as feature engineering, incorporating real-time engagement data, and fine-tuning hyperparameters could further enhance the model's predictive power.

Additionally, exploring ensemble methods like XGBoost or LightGBM may yield further accuracy improvements while maintaining computational efficiency. Future research should also consider integrating customer sentiment analysis, purchasing behavior patterns, and external market factors to refine churn prediction models.

4.4.2 Interpretation of Churn Factors

Several insights were derived from the churn prediction analysis, identifying key factors influencing customer retention:

- The analysis showed that customers who had gone without purchasing showed increased risk of leaving the company. An active customer period becomes an indicator that shows higher chances of complete customer disengagement.

- The customer practice of returning items frequently led to an increased risk of exiting since they typically remained dissatisfied about product quality. Organizations must address return-related problems to generate better customer satisfaction results.
- Customer loyalty increased when customers engaged in frequent small purchases yet large single-buyer transactions exposed them to a higher risk of leaving the company. Organizations should prioritize customer contact frequency above making high-value deals as their key strategy for business retention.

4.4.3 Business Recommendations for Churn Prevention

Based on these findings, several actionable recommendations can be implemented to reduce churn and improve customer retention:

- **Personalized Engagement Strategies**: Businesses need to use customer data when creating specific discounts and offers while giving personalized recommendations to users who show indicators of leaving. Thorough customer engagement leads to repeat orders that help strengthen brand customer relations.
- **Enhancing Customer Support and Return Policies**: Companies need to simplify return procedures and enhance product quality together with elevated customer service for lowering the likelihood of customer turnover.
- **Loyalty and Rewards Programs**: Establishing customer loyalty programs together with membership perks and special incentives serves to enhance customer retention statistics. Regular rewards to loyal customers both strengthens customer-brand attachment and lowers the chance of customers leaving.

The implementation of these strategic plans allows organizations to capitalize on churn prediction knowledge when reaching at-risk clients thus improving satisfaction and achieving enduring profitability. Random Forest Classifier demonstrates exceptional functionality in churn prediction which positions it as a powerful instrument for customer relationship management decision strategies.

CHAPTER 5: CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study successfully demonstrated the effectiveness of machine learning techniques in customer segmentation, sales prediction, and churn analysis. The application of various models provided insights into customer behavior, enabling businesses to develop targeted marketing strategies and retention policies.

The churn prediction analysis highlighted Gradient Boosting Classifier (GBC) as the best-performing model, achieving an accuracy of 73.38% and an ROC-AUC score of 0.7972. This model effectively distinguished between churned and non-churned customers, outperforming alternative models such as Random Forest (72.25% accuracy, 0.8019 ROC-AUC), Logistic Regression (50.41% accuracy), and Support Vector Machine (50.39% accuracy), which exhibited weak predictive power. The precision-recall balance in GBC suggests that businesses can leverage this model to identify at-risk customers and implement proactive retention strategies.

For customer segmentation, the K-Means clustering algorithm successfully classified customers into distinct segments based on Recency, Frequency, and Monetary (RFM) values. The 3D visualization of purchasing behavior provided deeper insights into market segmentation, identifying a high-value customer group crucial for business growth. This segmentation enables businesses to implement personalized marketing campaigns, loyalty programs, and targeted re-engagement strategies.

The sales prediction results demonstrated that Gradient Boosting Regressor (GBR) provided the most reliable forecasts, achieving an R^2 score of 0.871, indicating strong predictive capability without overfitting. Other models, including Linear Regression, Support Vector Regressor, and Random Forest Regressor, showed overfitting tendencies ($R^2 \approx 1.000$), making them less suitable for real-world predictions. The GBR model can be further improved by integrating additional customer behavior metrics, seasonal trends, and real-time purchase data.

Additionally, seasonality analysis revealed periodic variations in customer purchasing behavior, emphasizing the importance of time-based marketing strategies and inventory management. By leveraging machine learning, businesses can align promotional

campaigns with peak demand periods, ensuring improved customer engagement and revenue optimization.

This study was conducted using Google Colab and Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, facilitating efficient data preprocessing, visualization, and model training. Future work can explore ensemble learning techniques such as XGBoost and LightGBM, incorporate real-time customer interaction data, and utilize deep learning architectures for enhanced predictive accuracy. The insights gained from this study demonstrate the real-world applicability of machine learning in customer behavior analysis and business decision-making.

5.2 Future Work

Although the model demonstrated promising results, there is considerable scope for improvement and expansion in future research. Several deep learning models including Recurrent Neural Networks (RNNs) and transformers should be implemented because they enhance prediction accuracy by recognizing sophisticated behavioral patterns in customer activities. Improving the model's performance can be achieved through the implementation of advanced feature engineering methods that examine customer feedback sentiments.

Real-time data processing techniques should be applied to static churn outcomes to enable proactive retention strategies and dynamic churn predictions. Companies can develop instant customer-based decisions through the implementation of TensorFlow Serving streaming data frameworks.

By extending the available data to include demographic characteristics as well as social media engagement metrics together with outside market circumstances the predictive model would become more resilient. A combination of multiple algorithms that includes decision trees and support vector machines (SVMs) and deep learning should be used for better classification results through algorithmic synergy.

The findings of this research proved that machine learning techniques are suitable for customer segmentation and churn prediction while providing accurate results and meaningful cluster groups. The capability of predictive modeling will improve along with sophisticated modeling methods and a broader dataset thus enabling businesses to extract even more valuable information about customer retention and engagement.

References

Achite, M., Samadianfard, S., Elshaboury, N. and Sharafi, M., 2023. Modeling and optimization of coagulant dosage in water treatment plants using hybridized random forest model with genetic algorithm optimization. *Environment, Development and Sustainability*, *25*(10), pp.11189-11207.

Allheeib, N., Islam, M. S., Taniar, D., Shao, Z., & Cheema, M. A. (2021). Density-based reverse nearest neighbourhood search in spatial databases. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 4335-4346.

Berndt, A.D. and Petzer, D.J., 2023. Measuring the value of customer engagement metrics. In *Handbook of Customer Engagement in Tourism Marketing* (pp. 73-85). Edward Elgar Publishing.

Chaudhary, K. and Alam, M., 2022. *Big data analytics: applications in business and marketing*. Auerbach Publications.

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Cai, T. T., & Ma, R. (2022). Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research*, *23*(301), 1-54.

Challoumis, C., 2024, October. THE ECONOMICS OF AI-HOW MACHINE LEARNING IS DRIVING VALUE CREATION. In *XVI International Scientific Conference* (pp. 94-125).

Erdem, Z.U., 2021. *A Comparative Study for Customer Churn Analysis Via Machine Learning Algorithms* (Master's thesis, Marmara Universitesi (Turkey)).

E-commerce Customer Data for Behavior Analysis (2023). https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis

Friedman, J., 2029. The elements of statistical learning: Data mining, inference, and prediction. *(No Title)*.

Gandomi, A. and Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), pp.137-144.

García-Gil, D., García, S., Xiong, N., & Herrera, F. (2024). Smart data driven decision trees ensemble methodology for imbalanced big data. *Cognitive Computation*, *16*(4), 1572-1588.

Gkikas, D. C., & Theodoridis, P. K. (2024). Predicting Online Shopping Behavior: Using Machine Learning and Google Analytics to Classify User Engagement. *Applied Sciences*, *14*(23), 11403.

Gupta, P., Reddy, S., Gupta, A. and Singh, A., 2021. Enhancing Ad Targeting with AI: Leveraging K-Means Clustering and Neural Networks for Advanced Audience Segmentation. *International Journal of AI Advancements*, *10*(1).

Homburg, C., Kuester, S. and Krohmer, H., 2013. *Marketing management: A contemporary perspective*. McGraw-Hill Higher Education.

Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D.S. and Le, T.G., 2023. An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry. *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy*, *14*(1), pp.26-53.

Helmold, M., 2022. Marketing, Sales and Pricing: Introduction. In *Performance Excellence in Marketing, Sales and Pricing: Leveraging Change, Lean and Innovation Management* (pp. 1-11). Cham: Springer International Publishing.

Holloway, S., 2024. The Integration of Supply Chain Analytics and Customer Relationship Management (CRM).

Joga, P., Harshini, B. and Sahay, R., 2022, December. Comparative Analysis of Machine Learning Models for Customer Segmentation. In *International Conference on Intelligent Systems Design and Applications* (pp. 49-61). Cham: Springer Nature Switzerland.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Khodabandehlou, S. and Zivari Rahman, M., 2017. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, *19*(1/2), pp.65-93.

Kabir, M. F. (2025). Comprehensive Customer Segmentation and Behavior Prediction Using Advanced Machine Learning And Neural Network Models. *Available at SSRN 5128426*.

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.

Kotler, P., Keller, K.L., Brady, M., Goodman, M. and Hansen, T., 2016. *Marketing Management 3rd edn PDF eBook*. Pearson Higher Ed.

Kumar, V. and Pansari, A., 2016. Competitive advantage through engagement. *Journal of marketing research*, *53*(4), pp.497-514.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *nature*, *521*(7553), pp.436-444.

Lu, J., Zheng, X., Nervino, E., Li, Y., Xu, Z. and Xu, Y., 2024. Retail store location screening: A machine learning-based approach. *Journal of Retailing and Consumer Services*, *77*, p.103620.

Matuszelański, K. and Kopczewska, K., 2022. Customer churn in retail e-commerce business: Spatial and machine learning approach. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(1), pp.165-198.

Mowar, A., 2022. *Marketing management*. Blue Rose Publishers.

Madanchian, M. (2024). Generative AI for Consumer Behavior Prediction: Techniques and Applications. *Sustainability*, *16*(22), 9963.

Mozumder, M.A.S., Mahmud, F., Shak, M.S., Sultana, N., Rodrigues, G.N., Al Rafi, M., Farazi, M.Z.R., Karim, M.R., Khan, M.S. and Bhuiyan, M.S.M., 2024. Optimizing customer segmentation in the banking sector: a comparative analysis of machine learning algorithms. *Journal of Computer Science and Technology Studies*, *6*(4), pp.01-07.

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

Ngai, E.W., Xiu, L. and Chau, D.C., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), pp.2592-2602.

Patankar, N., Dixit, S., Bhamare, A., Darpel, A. and Raina, R., 2021. Customer segmentation using machine learning. In *Recent Trends in Intensive Computing* (pp. 239-244). IOS Press.

Prasetyo, E. and Nainggolan, B.M., 2024. The Impact Of Service Quality, Brand Image, And Social Media Marketing On The Purchase Decision At The Aryaduta Suites Semanggi Hotel Jakarta. *Jurnal Ekonomi*, *13*(03), pp.931-944.

Pellegrino, A. (2024). Digital Marketing: Overview and Evolutions. *Decoding Digital Consumer Behavior: Bridging Theory and Practice*, 15-29.

Parab, Y. and Dave, J., 2023. Customer Segmentation Using Machine Learning: A Comprehensive Research study. *International Journal of No vel Research and Development*, *8*(6), pp.718-725.

Rajput, L. and Singh, S.N., 2023, January. Customer Segmentation of E-commerce data using K-means Clustering Algorithm. In *2023 13th International conference on cloud computing, data science & engineering (Confluence)* (pp. 658-664). IEEE.

Ranjan, A. and Srivastava, S., 2022, November. Customer segmentation using machine learning: A literature review. In *AIP Conference Proceedings* (Vol. 2481, No. 1). AIP Publishing.

Solomon, M., Russell-Bennett, R. and Previte, J., 2012. *Consumer behaviour*. Pearson Higher Education AU.

Swain, K. P., Misra, S., Barik, S., Samal, S. R., & Sahoo, D. (2022, December). Exploratory Data Analysis on Shopping Mall Customers' Dataset: A Case Study of Marketing Analysis. In *International Conference on Advanced Computing and Intelligent Engineering* (pp. 207-216). Singapore: Springer Nature Singapore.

Sial, M., 2021. A brief introduction to regression analysis and its types. *Asian Journal of Probability and Statistics*, *13*(4), pp.58-63.

Singh, J., Goyal, S.B., Kaushal, R.K., Kumar, N. and Sehra, S.S., 2024. Applied Data Science and Smart Systems.

Shafi, I., Chaudhry, M., Montero, E. C., Alvarado, E. S., Diez, I. D. L. T., Samad, M. A., & Ashraf, I. (2024). A Review of Approaches for Rapid Data Clustering: Challenges, Opportunities and Future Directions. *IEEE Access*.

Singh, M., Singh, A., Gupta, M., & Reddy, R. (2022). Leveraging K-Means Clustering and Hierarchical Agglomerative Algorithms for Scalable AI-Driven Customer Segmentation. *Journal of AI ML Research*, *11*(7).

Salminen, J., Mustak, M., Sufyan, M. and Jansen, B.J., 2023. How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation. *Journal of Marketing Analytics*, *11*(4), pp.677-692.

Suhaas, K. P., Deepa, B. G., Shashank, D., & Narender, M. (2024). Millets Industry Dynamics: Leveraging Sales Projection and Customer Segmentation. *SN Computer Science*, *5*(8), 1063.

Tsiptsis, K.K. and Chorianopoulos, A., 2011. *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

Wedel, M., 2000. *Market segmentation: Conceptual and methodological foundations*. Kluwer Academic Publisher.

Wong, C.G., Tong, G.K. and Haw, S.C., 2024. Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. *Journal of Telecommunications and the Digital Economy*, *12*(3), pp.97-125.

Wang, G., Gunasekaran, A., Ngai, E.W. and Papadopoulos, T., 2016. Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International journal of production economics*, *176*, pp.98-110.

Xu, D. and Tian, Y., 2015. A comprehensive survey of clustering algorithms. *Annals of data science*, *2*, pp.165-193.

Yunianto, M., Meilina, R.D. and Suryani, E., 2024. Using Decision Tree With First and Second-Order Statistical Feature Extraction for Classification of Lung Cancer. *INDONESIAN JOURNAL OF APPLIED PHYSICS*, *14*(2), pp.339-352.

Zhang, Y. and Ma, Z.F., 2020. Impact of the COVID-19 pandemic on mental health and quality of life among local residents in Liaoning Province, China: A cross-sectional study. *International journal of environmental research and public health*, *17*(7), p.2381.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

Appendix I

Research Proposal

Using Machine Learning for Customer Segmentation and Sales Prediction in E-Commerce Industries

Submitted by: Muhammad Saad Sajjad

UB: 23047138

Contents

Introduction 1

1.1 Scope and Rationale 2

1.2 Study Aims 4

1.3 Objectives 5

1.4 Research Questions 5

Literature Review 6

Research Methodology 10

3.1 Research Philosophy 10

3.2 Research Approach 10

3.3 Data Collection 11

Data Analysis 12

4.1 Model Training 12

4.2 Performance Evaluation 13

4.3 Limitations 13

5. Ethical Consideration 14

6 Conclusion 16

Dissertation Overview and Headings 17

References 18

Appendix 23

Introduction

During recent years, artificial intelligence has seen a rise in practical applications in various industries such as healthcare, education, engineering, sales, entertainment, and transport (Bajaj et al. 2020). Machine learning, a branch of AI has also gained significant popularity in the marketing industry due to the vast amount of data that can be utilized to gain valuable insights that can drive important decisions for an organization's profitability and understanding of consumer behaviour.

Interestingly, this amount of data is expected to increase as more businesses use digital marketing to expand their operations (Boone et al. 2019). Forecasting sales allows businesses to gain meaningful insights for inventory management, budgeting, operational planning, and strategic decision making (Lau et al. 2018). It is an imperative step for strategic planning and making astute business decisions. As an e-commerce business becomes dependant on digital marketing platforms, using forecasting based on data analysis becomes crucial to maintain competitiveness (Cham et al., 2022)

Although traditional customer segmentation, which involves classifying a business's customer base into unique strata based on shared attributes and sales forecasting has contributed to successful business decisions in the past they are becoming increasingly obsolete (Turkmen 2022) in the present business context of digitalisation due to the rapidly changing demands of consumers. Such methods depend on a repository of past data combined with situational judgement decisions made by sales personnel along with analysing the current market trends. While these approaches yield results, they are essentially limited to the biasness of human input and interpretation, hence proving to be less efficient (Venkataramanan et al. 2024). Furthermore, these approaches may not account for variables such as changing consumer preferences, competitor activities, and differing economic conditions. Using machine learning to segment customers enables businesses to generate personalised profiles by analysing real time behaviour. These approaches adapt to changes in the market more effectively than traditional methods (Elhosseini 2023 et al. 2023).

The ongoing discourse by researchers such as Venkatraman have determined that while traditional approaches produce successful results when addressing factors such as evolving consumer demands, these methods have a limiting reliance on historical data that is formulated or derived from market dependent decisions by stakeholders ,especially in the analysis of current market dynamics. However, Pandey and Elhosseini argue that the application of advanced analytical technologies such as machine learning techniques can provide a more nuanced approach to segmentation analysis of customers due to its capacity to intuitively learn and adapt in real time to changes in consumer behaviour. This debate extends and supports the notion that machine learning applications in ecommerce is a better approach from conventional methods. Additionally, there exists substantial research on the application of AI solutions in customer segmentation processes as well as sales prediction. While the entities seem divergent in studies by Kasem et al (2023) and (Cheriyan et al. 2022), there is an evident gap that necessitates the combination of both process/concepts in understanding the varying nature and impact of customers to businesses.

While there is substantial research on using artificial intelligence for customer segmentation and sales prediction, they exist as separate entities of discussion in most studies (V Kumar et al. 2018). Therefore, this gap needs to be addressed to understand the varying nature and importance of the customers. Consequently, using the identified segments of customers, e commerce businesses can predict their future purchase intentions and patterns (Baati and Mohsil 2020). Doing so will enable the businesses to focus their marketing efforts and relationship building measures with individuals that are highly likely contribute to the business through frequent purchases and being brand ambassadors to bring forth a positive reputation of the business in a challenging business environment (Khoa and Huynh 2023).

1.1 Scope and Rationale

As artificial intelligence techniques evolve, a variety of techniques such as clustering, classification, neural networks, decision trees and AdaBoost are employed to gain meaningful insights from data and leveraging it to make informed business decisions. The results of these techniques are utilised to enhance marketing efforts by businesses, leading

to higher consumer engagement, and focused targeted marketing. The more concentrated marketing efforts businesses make, the better strategies they will be able to employ for achieving higher sales (Zulaikha et al. 2020)

Machine learning has the capability to process copious amounts of data efficiently and produce models that can be applied in the e-commerce industry (Sharda et al. 2018). In this context, customer segmentation and sales prediction are essential areas that benefit from machine learning. This study intends to integrate both approaches as it will enhance sales forecasting and gain an understanding of consumer purchasing patterns, enabling businesses to make data driven decisions. The knowledge developed will bridge the gap between sales prediction and customer segmentation, offering practical understanding for enhancing marketing strategies, inventory management and tailored consumer experiences. Businesses, science, and technology sector are adopting the much-needed use of the machine learning techniques which has evolved into its reliance in fields such as manufacturing, healthcare, education, financial modelling, and marketing (Jordan and Mitchell 2020).

The scope of this research is confined to e-commerce due to the unprecedented challenges and opportunities online businesses present (Hagberg et al. 2016). Unlike retail businesses, online businesses often compete on a global scale and are not confined to a geographical area. Additionally, customer retention in online businesses requires continuous engagement, and personalised marketing.

Lastly, e-commerce businesses can use data analytics and machine learning to provide personalised purchasing experiences, thus enhancing customer satisfaction and brand loyalty (Zhang and Xiong 2024).

Therefore, the analysis investigates consumer behaviour in an online setting and the trends seen in sales made within this domain. Although the findings of the research will be suitable for online businesses only, the insights gained, and the methodology can be generalised to various other industries with similar patterns of data

The rationale for this research is derived from the growing necessity of e-commerce businesses to acquire advanced techniques that focus on utilising data in a systematic way. Traditional methods fail to keep track of the dynamic e-commerce business environment and rapidly changing consumer interests, competitor tactics, and extrinsic factors. Machine learning provides a powerful and meticulous alternative through the recognition of intricate patterns in datasets with vast number of values, that tradition methods cannot comprehend. Businesses can make focused decisions and marketing strategies based on accurate segmentation of customers, thus enhancing consumer engagement and brand loyalty. Additionally, reliable sales prediction models aid businesses achieve long term success though the optimization of operational planning, budgeting, and systematic inventory management. Although there is significant research on customer segmentation and sales prediction individually, there exists a need to combine both aspects which is the purpose of this research.

1.2 Study Aims

As the competition among businesses rises in the digital space, it is crucial for businesses to leverage their data to reveal underlying trends and using the information revealed to modify and refine their marketing efforts. This paper aims to investigate the influence of machine learning in customer segmentation and sales prediction within the e-commerce industry. It also focuses on evaluating the overall efficiency of machine learning in enhancing the performance of a business

1.3 Objectives

- Determine which machine learning techniques are most effective to characterize e-commerce customer segments
- Evaluate the accuracy of the predictive sales model established using customer segmentation data in forecasting sales

1.4 Research Questions

- Which machine learning techniques are effective in customer segmentation and sales predictions?
- How do various data preprocessing procedures impact the performance of machine learning algorithms?
- What metrics are suitable for measuring the accuracy of the sales prediction model derived from customer segmentation data?
- What differences exist in the generated model's predictive accuracy between the various customer segments included in the e-commerce dataset?

Literature Review

Keywords	Scopus Results
"Machine Learning" AND "Customer Segmentation" AND "Sales Prediction"	3
"Supervised Learning" AND "Customer Segmentation" AND "Sales Prediction"	26
"Sales Prediction" AND "E-Commerce"	84
"Machine Learning" AND "Sales Prediction"	188
"Consumer Behaviour" AND "sales forecasting"	31
"Predictive Analytics" AND "Sales Models"	3
"Customer Segmentation" AND "Sales Prediction" AND "E-Commerce"	0
"Random Forest" OR "Decision Trees" AND "Sales Prediction"	68
"E-Commerce" AND "Machine Learning" not "Retail"	34
"Purchase History" AND "Customer Segments"	6

The relationship between machine learning and e-commerce has grown to be an area of increasing interest, particularly in view of customer segmentation and sales prediction. Though both subjects have been widely researched separately, there is a dire need to integrate them for optimal business results in the e-commerce industries.

Customer segmentation is a process of separating customers into distinct groups with identical behaviours and attributes for the purpose of focused marketing strategies. More conventionally, the traditional ways of segmentation perform the task by considering demographic, geographic, or psychographic data. However, with the development of machine learning techniques, businesses can create dynamic insights in terms of consumer behaviour. As Christy et al. (2018) mentioned, segmentation is fundamental to both identifying customer needs and customising marketing effort. Their study extends the recency, frequency and monetary (RFM) analysis to include more sophisticated algorithms such as k-means clustering. It mentions that customer segmentation enhances the ability of businesses to address diverse customer groups more effectively. However, while so doing, it may limit them in traditional clustering methods and hence could suffer inability in addressing real-time market shifts.

All forecasting, from a sales viewpoint, has conventionally been done using various statistical methods that estimate future sales based on historical data. Although considered effective in several cases, such methods fail to capture the essence of the evolving e-commerce ecosystem, wherein customer behaviours are highly dynamic and driven by numerous external factors. Singh et al. (2020) mentions that random forest and gradient boosting are among the advanced models which perform efficiently in this domain. Nonlinear relationships can be represented, and volume and complex data is managed with greater predictiveness by these models. However, Bohanec (2017) highlights that while extremely useful and accurate, the black box machine learning models (whose internal working cannot be comprehended) pose a challenge due to their lack of interpretability.

Furthermore, different machine learning algorithms explored independently either for segmentation or for forecasting rarely combine both holistic approaches. Turkmen (2022) attempts to bridge this gap by incorporating k-means clustering into a statistical framework to segment customers and predict sales in an e-commerce setting. Although this model seems promising, it is limited by being based on a single clustering method. Exclusively depending on k-means clustering, when the complexity of customer behaviour is rising may lead to segmentation results that are overly simplified and cannot represent nuanced

differences among customers, a concern similarly raised by Saxena et al. (2024). They go further into a variety of clustering techniques, including hierarchical clustering and DBSCAN, for a more robust framework of customer segmentation. However, as they indicate, such models are not very promising in larger datasets since the problem arises with visualization and interpretation. Another critical point of consideration is scalability and flexibility of the models adopted. Raizada and Saini (2021) present the efficiency of random forest and extra tree regression on Walmart data for sales prediction, and they have received accuracy results above 98%. However, it does not take into consideration how these models would generalize across industries with different data structures and consumer behaviours. This remains a general problem in the literature: usually, the performance of algorithms is assessed in particular contexts without afterthoughts about their adaptiveness to general business environments.

Contrary to that, Liu et al. (2020) developed a more adaptive approach by proposing logistic regression combined with XGBoost for predicting customer repurchase behaviour. Their model overcomes an important weakness with traditional approaches since it oversees imbalanced datasets very well, a frequent problem in e-commerce, since a small percentage of customers drive a large percentage of sales. The study shows how machine learning could be applied not only to predict what customers will do in the future but also to develop better, targeted marketing initiatives, reinforcing the interdependence between customer segmentation and sales prediction.

lastly, Cheriyan et al. (2022) study different regression-based techniques for sales forecasting and report that random forest outperformed others, with an accuracy of 95.53%. Yet such technocratic success does not resolve the conceptual weakness of the exclusive use of regression-based techniques. As companies depend increasingly on real-time data, regression models, which rely on the stability of the relationships between the dependent and independent variables, cannot easily match the rhythm of the online commercial environment.

The literature review indeed shows that machine learning achieved promising results both in customer segmentation and sales forecasting, while the combination of the two fields has not been well explored. It is further observed that there is a tendency to be focused merely

on methodological performance; providing accuracy rates or error metrics, while the critical issues of model interpretability, scalability, and cross-industry applicability are not considered. In this regard, future studies are needed to be more complete in modelling and integrating segmentation with prediction, exploiting the power of multiple algorithms to provide an even finer understanding of customer behaviour and sales trends.

Research Methodology

3.1 Research Philosophy

This study is based on scientific research using legitimate data, making realism the most suitable philosophical approach. It follows a quantitative framework, relying on deductive methods. The research does not consist of a hypothesis; however, it assumes a relationship between customer sales and the segment they belong to. Various statistical and mathematical tools will be used to analyse the data and draw conclusions from the hypotheses (Ishtiaq, 2019).

3.2 Research Approach

This study's methodology is based on deductive reasoning which is considered as the key component of a research approach based on positivism and objectivity. Deductive reasoning moves from a general problem to a specific conclusion, following a logical chronology of steps to determine whether a theory can be proven in particular circumstances (Saunders et al, 2019). Although there is no hypothesis for this thesis, deductive reasoning will apply as it involves a logical flow from prevalent rules to the specific case of the viability of machine learning to segment customers and use the aforementioned segments to predict future sales. However, a strict deductive approach may neglect an investigation of other methods that could even further disclose the use of machine learning for segmentation of customers and sales forecast.

The research will begin with the established principle that machine learning can manage an extensive amount of data and leverage it to gain meaningful insights. It will then continue with the proposition that machine learning techniques will use customer sales data to predict their future purchases and split the customers into different categories based on their purchase patterns. Lastly, based on the results of the previous steps, a discrete conclusion can be drawn over the efficacy of machine learning and the specific techniques employed, and their impact on a business' future planning.

3.3 Data Collection

This study will comprise of secondary data collected from University of California Irvine's machine learning repository website. It consists of an online retail dataset that portrays sales records of customers made from several countries. The dataset was chosen for this research due to a considerable number of records amounting to more than 54,000, ensuring that the machine learning techniques used will have sufficient data to use, therefore ensuring a comprehensive analysis and insightful findings.

Data Analysis

4.1 Model Training

Once the data is prepared, it will run a series of machine learning methods for customer segmentation such as K-means clustering, which has the ability to separate customers into groups based on features such as purchasing behaviour, frequency of transaction and/or amount spent (Jain, 2010). It will be used because it is effective in grouping similar customers together according to their purchasing patterns. Hierarchal clustering may also be employed to build a ranking of clusters which will aid in understanding the relationship between customers at diverse levels. This approach allows an in depth understanding of customers by creating a hierarchy of clusters (Embrechts et al. 2013), which aids in the exploration of multi-level relationships among various consumer groups. Furthermore, this gives a very fine-grained understanding of the association between different tiers of customers, which is crucial while developing marketing strategies addressing various tiers of customers. A recency, frequency and monetary (RFM) model can also be used to determine how recently a customer made the purchase, how often they purchase and how much monetary amount they spent. This will assist in providing a data driven and systematic way to classify customers into segments pertaining to their purchasing patterns (Dogan et al. 2018). The incorporation of the RFM model brings in consistency regarding the segregation of customers. This model effectively quantifies customer value with direct relevance to their purchasing patterns, which is paramount when using data to drive marketing strategy. Through it, businesses can pinpoint the most valued customers and concentrate their resources effectively in predicting sales.

Once the customer data is segmented into clusters, the sales prediction algorithms will be executed. The chosen methods include linear regression which can predict sales using past purchasing patterns and customer features from the clustering techniques (Morsi 2020). This approach is chosen due to its ability to enhance data driven decision making and efficiently encapsulates the relationship between sales and the factors that influence them. Secondly, decision tree algorithm will be deployed to acquire the relationship between customer features and the sales made. Decision trees will be used as they are an effective means to capture complex non-linear data patterns in large data sets and provide a

structured and interpretable model for decision-making (Mustakim et al. 2024). Since decision trees are interpretable, businesses can be shown how each customer feature influences sales and hence communicate findings with stakeholders easily.

Next, random forest regression method will be deployed to acquire to improve the prediction accuracy by aggregating the results of multiple decision trees. This method is expected to give a higher accuracy due to its ability to reduce overfitting, robustness, and ability to generalize (Naik et al. 2022). The capability of this method to avoid overfitting and generalize well to unseen data is essential in developing robust sales prediction models. A more accurate prediction will enable a business to strategize its marketing efforts more effectively and finally aid the research objective to leverage machine learning for business optimization.

4.2 Performance Evaluation

To evaluate the effectiveness of the sales prediction model, evaluation techniques mean absolute error (MAE) and root mean square error (RMSE) will be used. These metrics will measure the difference between the actual and predicted sales, therefore indicating the reliability and solidity of the machine learning model. If the values for RMSE and MAE are lower, it will indicate accurate performance. This will indicate that the predictions made by the model are closer to the actual sales figures.

4.3 Limitations

While the dataset used for this study provides a significant amount of data, it does contain limitations.

Missing or Incomplete Data: Since this study uses a secondary dataset retrieved from an online source, it is likely to contain missing or incomplete data which may negatively impact the accuracy of the machine learning results.

Geographical Constraints: While the dataset contains a considerable amount of data, it is confined to thirty-six countries, hence the results may not apply to other regions in the world.

Limited Historical Data: The dataset is likely to contain limited historical data, which may limit the model's capability to produce long term predictions or encapsulate the emerging behaviour trends of customers.

K Means Clustering: K means clustering requires the number of clusters to be preset. This can hinder the diversity of the customer segments and skew results. Despite this, it performs efficient computation and segmentation when the optimal number of clusters are identified.

The RFM model only has three variables (recency, frequency, and monetary value) which may cause the results of consumer behaviour techniques to be superficial. However, it is an elementary process which provides a swift approach to identify high-value customers, thus making it an appropriate choice for initial segmentation.

Linear regression: This approach operates on assumptions of the relationships among variables as being linear, which may not be accurate for the dataset being used as it contains a wide array of rows with thousands of data entries. Despite this, it is a useful method as it is simple to interpret, simplifying the understanding of the result of independent variables on sales. Consequently, this can result in valuable insights for e-commerce decision making.

5. Ethical Consideration

The secondary data used in this research is available online on University of California Irvine's machine learning repository and is open source. It is therefore available for research purposes and using this data does not infringe the privacy of any individuals whose sales data is recorded.

Conclusion

The research emphasizes the increasing value of machine learning in e-commerce businesses. It illustrates the process of identifying customers and using their purchasing patterns to forecast the future demand, facilitating businesses to make strategic sales and marketing decisions. As traditional customer segmentation and sales prediction methods become obsolete in the rapidly evolving world, machine learning offers a robust and data driven approach to this problem. Techniques such as hierarchical clustering and k means clustering can provide a reliable mechanism for segmenting customers into distinct groups that organisations can cater to according to their unique needs.

The significance of this study lies in its potential to transform the approach e-commerce businesses take to understand their customers and predict future sales, resulting in numerous significant contributions.

By combining customer segmentation and sales prediction, e-commerce businesses can gain a deeper understanding of their consumers and their purchasing behaviour. Secondly, the advanced segmentation approaches in this study can enable businesses to create personalised customer experiences. This can lead to higher customer satisfaction and brand loyalty. Also, by incorporating the two methods, this research has the capability to enhance the forecast accuracy of sales.

Another crucial impact of this study is that the findings can be used in direct applications of the e-commerce industry, which may result in the conception of exclusive strategies and tools for customer relationship management and optimisation of sales. Lastly, the methods proposed in this study can be applied to various e-commerce sectors and scaled to assist the expanding convolution of online retail conditions.

Dissertation Overview and Headings

Introduction
1.1 Background and significance
1.2 Research Objectives
1.3 Research questions
1.4 Scope and Limitations
Literature Review
2.1 Machine learning for e-commerce: Concepts and challenges
2.2 Customer Segmentation and Sales Prediction
2.3 Intersection of customer segmentation and sales prediction in machine learning
2.4 prior studies on Machine Learning for customer segmentation and Sales Prediction
2.5 Gaps in current literature
Research Methodology
3.1 Research Approach
3.2 Data Collection
3.3 Data Analysis Techniques
3.4 Ethical Consideration
Findings and Analysis
4.1 Analysis of Findings
4.2 Machine Learning Analysis
Discussion
5.1 Integration of Findings
5.2 Implications of Research
5.3 Application in Future Research
Conclusion
6.1 Synopsis of Findings
6.2 Contribution of Research
6.3 Recommendations

6.4 Conclusion

References

Baati, K. and Mohsil, M. (2020) Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. *IFIP Advances in Information and Communication Technology* 43–51.

Bajaj, P., Ray, R., Shedge, S., Vidhate, S. and Shardoor, N. (n.d.) *SALES PREDICTION USING MACHINE LEARNING ALGORITHMS*. *International Research Journal of Engineering and Technology* .

Banerji, R. and Singh, A. (2024) Do social media marketing activities promote customer loyalty? A study on the e-commerce industry. *LBS Journal of Management & Research* 22 (1), Emerald93–109.

Bharadwaj, A., El Sawy, O.A., Pavlou, P.A. and Venkatraman, N. (2013) Digital Business Strategy: Toward a Next Generation of Insights. *MIS Quarterly* 37 (2), 471–482.

Bohanec, M., Kljajić Borštnar, M. and Robnik-Šikonja, M. (2017) Explaining machine learning models in sales predictions. *Expert Systems with Applications* 71, 416–428.

Boone, T., Ganeshan, R., Jain, A. and Sanders, N.R. (2019) Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting* 35 (1), 170–180.

Burrell, G. and Morgan, G. (2017) Sociological paradigms and organisational analysis: Elements of the sociology of corporate life. Routledge.

Cham, T.-H., Cheah, J.-H., Memon, M.A., Fam, K.-S. and László, J. (2022) Digitalization and its impact on contemporary marketing strategies and practices. *Journal of Marketing Analytics* 10 (1),.

Cheriyan, S., Ibrahim, S., Mohanan, S. and Treesa, S. (2022) *Intelligent Sales Prediction Using Machine Learning Techniques*. 278–286.

Christy, A.J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A. (2018) RFM Ranking – an Effective Approach to Customer Segmentation. *Journal of King Saud University - Computer and Information Sciences* 33 (10)

Dogan, O., Aycin, E. and Bulut, Z. (2018) Customer Segmentation by Using Rfm Model and Clustering methods: a Case Study in Retail Industry. *International Journal of Contemporary Economics and Administrative Sciences* 8, 1–19.

Dullaghan, C. and Rozaki, E. (2017) Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers. *International Journal of Data Mining & Knowledge Management Process* 7 (1), 13–24.

Embrechts, M.J., Gatti, C.J., Linton, J. and Roysam, B. (2013) Hierarchical Clustering for Large Data Sets. *Advances in Intelligent Signal Processing and Data Mining* 197–233.

Hagberg, J., Sundstrom, M. and Egels-Zandén, N. (2016) The digitalization of retailing: an exploratory framework. *International Journal of Retail & Distribution Management* 44 (7), Emerald694–712.

Heron, J. (1996) Co-operative inquiry: Research into the human condition. Sage.

Jafari, R. (2022) Hands-on data preprocessing in Python: Learn how to effectively prepare data for successful data analytics: Packt Publishing.

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31 (8), 651–666.

Jordan, M.I. and Mitchell, T.M. (2020) Machine learning: Trends, perspectives, and prospects. *Science* 349 (6245), 255–260.

Kansal, T., Bahuguna, S., Singh, V. and Choudhury, T. (2018) Customer Segmentation Using K-means Clustering. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)* 135–139.

Khoa, B. and Huynh, T. (2023) The influence of social media marketing activities on customer loyalty: A study of e-commerce industry. *International Journal of Data and Network Science* 7 (1), 175–184.

Kilani, M.A. and Kobziev, V. (2016) An Overview of Research Methodology in Information System (IS). *Open Access Library Journal* 03 (11), 1–9.

Lau, R.Y.K., Zhang, W. and Xu, W. (2018) Parallel Aspect-Oriented Sentiment Analysis for Sales Forecasting with Big Data. *Production & Operations Management* 27 (10), 1775–1794.

Liu, C.-J., Huang, T.-S., Ho, P.-T., Huang, J.-C. and Hsieh, C.-T. Z Lv (editor), (2020) Machine learning-based e-commerce platform repurchase customer prediction model. *PLOS ONE* 15 (12), e0243105.

Mahesh, B. (2018) Machine Learning Algorithms -A Review. *International Journal of Science and Research (IJSR) ResearchGate Impact Factor* 9 (1).

Mahmoud SalahEldin Kasem, Hamada, M. and Islam Taj-Eddin (2023) Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications* 36, Springer Science+Business Media.

Morsi, S. (2020) A Predictive Analytics Model for E-commerce Sales Transactions to Support Decision Making: A Case Study. *International Journal of Computer and Information Technology(2279-0764)* 9 (1)

Mustakim, N.A., Abdul Aziz, M. and Abdul Rahman, S. (2024) Predicting Consumer Behaviour in E-Commerce Using Decision Tree: A Case Study in Malaysia. *Information Management and Business Review* 16 (3(I)), 201–209.

Naik, H., Yashwanth, K., P, S. and Jayapandian, N. (2022) *Machine Learning based Food Sales Prediction using Random Forest Regression*. 998–1004https://ieeexplore.ieee.org/abstract/document/10009277 Accessed.

Raizada, S. and Saini, J.R. (2021) Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting. *International Journal of Advanced Computer Science and Applications* 12 (11).

Rokach, L. and Maimon, O. (2005) Clustering Methods. *Data Mining and Knowledge Discovery Handbook* 321–352.

Saunders, M., Lewis, P. and Thornhill, A. (2019) *Research Methods for Business Students*. 8th Edition. Harlow: Pearson.

Saxena, A., Agarwal, A., Binay Kumar Pandey and Pandey, D. (2024) Examination of the Criticality of Customer Segmentation Using Unsupervised Learning Methods. *Circular Economy and Sustainability* 4, Springer Nature.

Sharda, R., Delen, D. and Turban, E. (2018) *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. 4th Edition. Boston: Pearson Education, Inc150–165.

Turkmen, B. (2022) Customer Segmentation with Machine Learning for Online Retail Industry. *The European Journal of Social & Behavioural Sciences* 31 (2), 111–136.

V Kumar, Reinartz, W. and Springer-Verlag Gmbh (2018) *Customer Relationship Management Concept, Strategy, and Tools*. Berlin Springer Berlin Springer.

Venkataramanan, S., Sadhu, A.K.R., Gudala, L. and Reddy, A.K. (2024) Leveraging Artificial Intelligence for Enhanced Sales Forecasting Accuracy: A Review of AI-Driven Techniques and Practical Applications in Customer Relationship Management Systems. *Australian Journal of Machine Learning Research & Applications* 4 (1), 267–287.

Zhang, Q. and Xiong, Y. (2024) Harnessing AI potential in E-Commerce: improving user engagement and sales through deep learning-based product recommendations. *Current Psychology* Springer Science and Business Media LLC.

Zulaikha, S., Mohamed, H., Kurniawati, M., Rusgianto, S. and Rusmita, S.A. (2020) Customer Predictive Analytics Using Artificial Intelligence. *The Singapore Economic Review* 1–12.

Ishtiaq, M. (2019) Book Review Creswell, JW (2014). Research Design: Qualitative, Quantitative and Mixed Methods Approaches. Thousand Oaks, CA: Sage. English Language Teaching 12 (5), 40.

Appendix

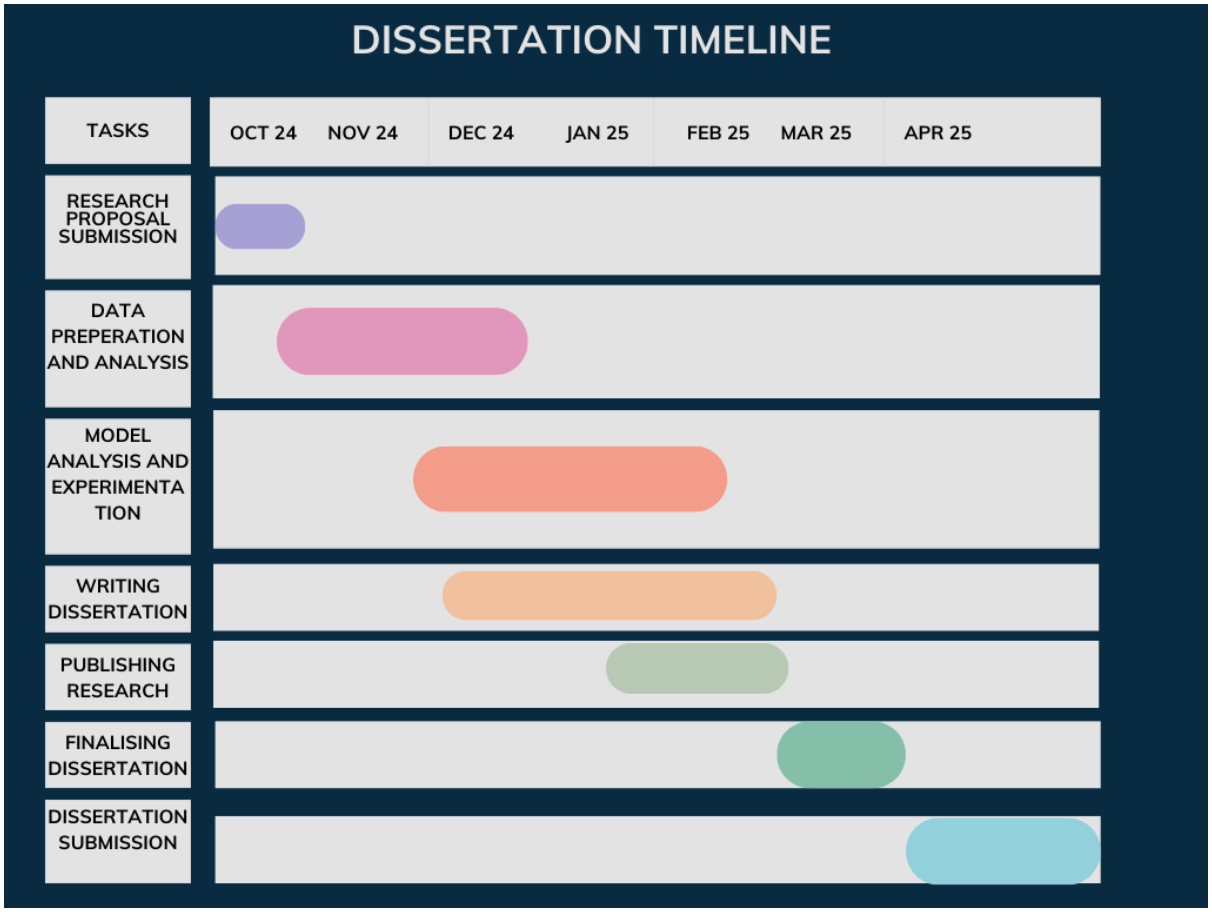

Figure 1: Dissertation |Timeline

Appendix II

Import Required Libraries

```
[2]
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler, LabelEncoder, MinMaxScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score, accuracy_score, classification_report, roc_auc_score
from imblearn.over_sampling import SMOTE
from mpl_toolkits.mplot3d import Axes3D
```

Load and Preprocess Dataset

```
# Google Drive file ID
file_id = "1Ila1O9pI_quKk-SiP-qw19Dt0xdMeGQ9"
dataset_url = f"https://drive.google.com/uc?id={file_id}"

# Load the dataset
data = pd.read_csv(dataset_url)
```

```
[16]
# Convert 'Purchase Date' to datetime format
data['Purchase Date'] = pd.to_datetime(data['Purchase Date'])

# Check for missing values
print("Missing Values:\n", data.isnull().sum())

# Fill or drop missing values if any
data.fillna(0, inplace=True)

# Ensure 'Total Purchase Amount' is numeric
data['Total Purchase Amount'] = pd.to_numeric(data['Total Purchase Amount'], errors='coerce')
data['Total Purchase Amount'].fillna(0, inplace=True)

# Encode categorical variables
label_encoders = {}
for column in ['Product Category', 'Payment Method', 'Gender', 'Customer Name']:
    le = LabelEncoder()
    data[column] = le.fit_transform(data[column])
    label_encoders[column] = le

# Derive new features
latest_date = data['Purchase Date'].max()
data['Recency'] = (latest_date - data['Purchase Date']).dt.days
data['Monetary'] = data['Total Purchase Amount']
data['Frequency'] = data.groupby('Customer ID')['Customer ID'].transform('count')

# Aggregate data for customer-level analysis
customer_data = data.groupby('Customer ID').agg({
    'Recency': 'min',
    'Frequency': 'max',
```

```
        'Monetary': 'sum',
        'Age': 'mean',
        'Churn': 'max',
        'Returns': 'sum'
    }).reset_index()

    # Standardize features for clustering
    scaler = StandardScaler()
    scaled_features = scaler.fit_transform(customer_data[['Recency', 'Frequency', 'Monetary']])
```
[16]

```
Missing Values:
 Customer ID                 0
Purchase Date                0
Product Category             0
Product Price                0
Quantity                     0
Total Purchase Amount        0
Payment Method               0
Customer Age                 0
Returns                  47596
Customer Name                0
Age                          0
Gender                       0
Churn                        0
dtype: int64
```

Exploratory Data Analysis (EDA)

```
# Sales trend over time
plt.figure(figsize=(10, 5))
sales_trend = data.groupby(data['Purchase Date'].dt.to_period('M'))['Total Purchase Amount'].sum()
sales_trend.plot(kind='line', title='Sales Trend Over Time', color='blue', marker='o')
plt.ylabel('Total Sales')
plt.xlabel('Time (Months)')
plt.grid(True)
plt.show()

# Monthly seasonality
plt.figure(figsize=(10, 5))
data['Month'] = data['Purchase Date'].dt.month
monthly_seasonality = data.groupby('Month')['Total Purchase Amount'].mean()
monthly_seasonality.plot(kind='line', title='Monthly Seasonality in Sales', color='red', marker='o')
plt.ylabel('Average Sales')
plt.xlabel('Month')
plt.xticks(range(1, 13), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun',
                          'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'], rotation=45)
plt.grid(True)
plt.show()

# Popular product categories
plt.figure(figsize=(8, 5))
sns.countplot(x='Product Category', data=data, palette='viridis')
plt.title('Popular Product Categories')
plt.xlabel('Product Category')
plt.ylabel('Frequency')
```
[17]

Customer Segmentation Using K-Means

[18]
```
# Apply K-Means clustering
kmeans = KMeans(n_clusters=3, random_state=42)
customer_data['Cluster'] = kmeans.fit_predict(scaled_features)

# 3D Visualization of Clusters
fig = plt.figure(figsize=(8, 6))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(customer_data['Monetary'], customer_data['Frequency'], customer_data['Cluster'],
           c=customer_data['Cluster'], cmap='viridis', marker='o')
ax.set_xlabel('Monetary')
ax.set_ylabel('Frequency')
ax.set_zlabel('Cluster')
plt.title("3D K-Means Clustering")
plt.show()
```

Sales Prediction Using Random Forest Regressor

[19]
```
# Define features and target variable
features = ['Product Price', 'Quantity', 'Customer Age', 'Returns']
X = data[features]
y = data['Total Purchase Amount']

# Normalize data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Split dataset
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Train Random Forest Regressor
rf_model = RandomForestRegressor(n_estimators=200, max_depth=20, random_state=42)
rf_model.fit(X_train, y_train)

# Predictions and evaluation
y_pred = rf_model.predict(X_test)

print("\nSales Prediction Model Performance:")
print(f"MAE: {mean_absolute_error(y_test, y_pred):.2f}")
print(f"RMSE: {np.sqrt(mean_squared_error(y_test, y_pred)):.2f}")
print(f"R² Score: {r2_score(y_test, y_pred):.3f}")
```

Churn Prediction Using Random Forest Classifier

```
[20]
# Define features and target variable
churn_features = ['Recency', 'Frequency', 'Monetary', 'Returns', 'Age']
X_churn = customer_data[churn_features]
y_churn = customer_data['Churn']

# Normalize and handle class imbalance using SMOTE
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X_churn)

smote = SMOTE(sampling_strategy="auto", random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_scaled, y_churn)

# Split dataset
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.3, random_state=42)

# Train Random Forest Classifier
rf_classifier = RandomForestClassifier(n_estimators=100, max_depth=20, class_weight="balanced", random_state=42)
rf_classifier.fit(X_train, y_train)

# Predictions and evaluation
y_pred = rf_classifier.predict(X_test)
y_proba = rf_classifier.predict_proba(X_test)[:, 1]

print("\nChurn Prediction Model Performance:")
print(f"Accuracy: {accuracy_score(y_test, y_pred):.4f}")
print(f"ROC-AUC Score: {roc_auc_score(y_test, y_proba):.4f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Insights and Recommendations

```
[21]
print("\nInsights:")
print("1. High-value customers are identified in Cluster 0.")
print("2. Targeted marketing can be applied to customers with low recency scores.")
print("3. Focus on retaining customers predicted to churn based on the classification model.")
```

.