UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS

FINAL PROJECT REPORT

DATA ANALYTICS IN BUSINESS

TOPIC: CUSTOMER SEGMENTATION AND RESPONSE PREDICTION IN IFOOD'S MARKETING CAMPAIGNS USING MACHINE LEARNING

Group: 05

Instructor: Assoc. Prof. Ho Trung Thanh, Ph.D.

Ho Chi Minh City, October, 2025

Members of Group 5

No.	**Full name**	**Student ID**	**Point / 10 (Individual Contribution)**	**Signature**
1	Phạm Tuyết Nhung	K224111460	10	
2	Vũ Quỳnh Như	K224111461	10	
3	Lê Nguyễn Minh Thảo	K224111462	10	
4	Nguyễn Ngọc Diễm Thúy	K224111466	10	
5	Phan Thị Thùy Trang	K224111470	10	

Acknowledgments

First of all, Group 5 would like to express our sincere gratitude to the University of Economics and Law, Vietnam National University Ho Chi Minh City and the Faculty of Information Systems for providing us with an excellent academic environment for learning, research and practical application. The knowledge and skills acquired from the university and faculty have served as a solid foundation for the completion of this project.

We would like to express our deep appreciation to Assoc. Prof. Ho Trung Thanh, Ph.D., instructor of the course Data Analysis in Business for his dedicated guidance, valuable insights and continuous support throughout this project. His feedback and encouragement have been an invaluable source of motivation, helping our group complete the project in a more scientific and practical manner.

Finally, we sincerely thank all group members for their cooperation, responsibility and continuous effort during the project. Although there are still some limitations due to time and knowledge constraints, we have done our best to complete the project with a spirit of diligence and continuous learning. We sincerely hope to receive valuable feedback from our instructor and faculty to further improve this work in the future.

Commitment

Group 5 hereby declares that the project titled “Customer Segmentation and Response Prediction in Ifood’s Marketing Campaigns using Machine Learning” is the result of the group’s own research and work, conducted under the supervision of Assoc. Prof. Ho Trung Thanh, Ph.D., instructor of the course Data Analysis in Business. All contents of this project are developed based on the knowledge gained from the course, together with references to academic materials, research papers and publicly available datasets, all of which are properly cited. The group takes full responsibility for the integrity, accuracy and transparency of the project’s content. In the event of any plagiarism or violation of academic integrity, the group will take full responsibility before the university and the supervising instructor.

Table of Contents

List of Tables

List of Figures

GANTT Chart

The Gantt chart is an essential tool for managing and monitoring project progress, providing an overview of the timeline for key tasks and milestones. This enables the team to easily track progress, allocate resources efficiently and make timely adjustments to ensure the project achieves its objectives within the specified timeframe. Below is the link to access the team's Gantt chart: Link

STT	Task	Start Date	End Date	Duration	Completion Progress	Priority (1 = Highest)	Person in Charge	Phase 1		Phase 2	
								31/08/2025	07/09/2025	14/09/2025	21/09/2025
	Overview of the Project						**All**				
1	Draw a Gantt chart	27/08/2025	05/09/2025	54 days	100%	3	Như				
2	Define the direction of the topic	27/08/2025	30/08/2025	4 days	100%	1	All				
3	Find the suitable dataset	30/08/2025	31/08/2025	2 days	100%	1	All				
4	Design a complete outline	31/08/2025	31/08/2025	1 day	100%	1	Như				
5	Acknowledgement + Commitment	31/08/2025	31/08/2025	1 day	100%	4	Nhung				
6	Write Abstract	17/10/2025	18/10/2025	2 days	100%	2	Như				
	Project Overview						**All**				
7	Business problems/Challenges	01/09/2025	03/09/2025	3 days	100%	2	Nhung, Thảo				
8	Business Objectives, Business Questions, Objects and scopes	01/09/2025	02/09/2025	2 days	100%	1	Thuý				
9	Experimental method/process	01/09/2025	02/09/2025	2 days	100%	1	Như				
10	Tools and Programming languages	01/09/2025	01/09/2025	1 day	100%	2	Thuý				
11	Structure of project	01/09/2025	01/09/2025	1 day	100%	1	Trang				
	Chapter 1. Theoretical Background						**All**				
12	Chapter overview	03/09/2025	07/09/2025	5 days	100%	3	Trang				
13	Food Order and Delivery sector	03/09/2025	06/09/2025	4 days	100%	2	Thảo				
14	Introduction to Machine Learning	03/09/2025	06/09/2025	4 days	100%	2	Thảo				
15	Customer Response Prediction	03/09/2025	06/09/2025	4 days	100%	2	Như				
16	Customer Segmentation	03/09/2025	06/09/2025	4 days	100%	2	Như				
17	Predictive Classification Models	03/09/2025	06/09/2025	4 days	100%	1	Nhung,Trang				
18	Clustering Models	03/09/2025	06/09/2025	4 days	100%	1	Thuý				
19	Format word Project Overview + Chapter 1	07/09/2025	07/09/2025	1 day	100%	4	Nhung				

Abstract

Tóm tắt: Trong bối cảnh cạnh tranh ngày càng gay gắt và ngân sách marketing bị giới hạn, việc hiểu rõ hành vi khách hàng và dự đoán khả năng phản hồi đóng vai trò quan trọng trong tối ưu chi phí và nâng cao hiệu quả chiến dịch. Nghiên cứu này tập trung vào phân khúc khách hàng và dự đoán phản hồi trong các chiến dịch marketing đa kênh của iFood - nền tảng giao đồ ăn trực tuyến lớn nhất tại Brazil. Dựa trên khung phương pháp CRISP-DM, nhóm nghiên cứu tiến hành toàn bộ quy trình từ xử lý và chọn lọc dữ liệu đến mô hình hóa và trực quan hóa kết quả. Dữ liệu gồm hơn 2,000 khách hàng của iFood, phản ánh hành vi mua sắm trên nhiều kênh. Sau bước xử lý dữ liệu, nhóm áp dụng K-Means để chia khách hàng thành ba cụm có đặc điểm nhân khẩu học và hành vi tiêu dùng khác nhau. Tiếp đó, nhóm xây dựng các mô hình dự đoán phản hồi riêng cho từng cụm, sử dụng các thuật toán Logistic Regression, Random Forest và XGBoost để tìm ra mô hình hiệu quả nhất. Kết quả thực nghiệm cho thấy cả ba mô hình đều đạt hiệu quả dự báo tốt, với ROC-AUC dao động từ 0.785 đến 0.866, độ chính xác từ 0.75 đến 0.90, và F1-score trong khoảng 0.40-0.62. Mô hình Random Forest thể hiện hiệu suất ổn định và vượt trội hơn cả, đặc biệt ở cụm 2, đạt ROC-AUC = 0.843 và Accuracy = 0.898. Các kết quả này chứng minh mô hình có khả năng phân loại khách hàng phản hồi với độ tin cậy cao, qua đó hỗ trợ doanh nghiệp nhận diện các nhóm khách hàng tiềm năng, phân bổ ngân sách hợp lý và thiết kế chiến lược tiếp thị cá nhân hóa theo từng phân khúc.

Abstract: In an increasingly competitive market with limited marketing budgets, understanding customer behavior and predicting response likelihood are crucial for optimizing costs and improving campaign effectiveness. This study focuses on customer segmentation and response prediction in iFood's multichannel marketing campaigns, conducted by the largest online food delivery platform in Brazil. Following the CRISP-DM framework, the research covers the entire process from data preprocessing and feature selection to modeling and visualization. The dataset consists of more than 2,000 iFood customers, reflecting their purchasing behavior across multiple channels. After data preparation, the K-Means algorithm was applied to segment customers into three clusters

with distinct demographic and behavioral characteristics. Subsequently, separate response prediction models were developed for each cluster using Logistic Regression, Random Forest and XGBoost to determine the most effective approach. Experimental results show that all three models achieved good predictive performance, with ROC-AUC ranging from 0.785 to 0.866, Accuracy between 0.75 and 0.90 and F1-scores between 0.40 and 0.62. The Random Forest model demonstrated the most consistent and superior performance, especially in Cluster 2, achieving ROC-AUC = 0.843 and Accuracy = 0.898. These findings confirm the models' ability to accurately classify customer responses, thereby supporting businesses in identifying high-potential customer groups, optimizing marketing budgets and designing personalized marketing strategies for each segment.

Project Overview

Business problems

The food delivery market in Brazil is expanding rapidly, with a 15% compound annual growth rate (CAGR) expected to reach USD 1.29 billion in 2024, according to a report by IMARC Group, a top market research firm. Among the companies in this market, iFood was established in 2011 and is presently Brazil's biggest online food delivery platform, linking over 300,000 restaurants with millions of consumers via its website and mobile app. With a strong technological ecosystem and an extensive partner network, iFood holds an 87% market share, which necessitates optimizing marketing strategies to both maintain its leading position and ensure efficient budget utilization. As noted by GhorbanTanhaei et al. (2024), understanding customer behavior is a critical factor in enhancing marketing effectiveness. By capturing customer needs and purchasing behaviors, businesses can tailor their marketing campaigns more effectively, thereby significantly improving business performance. However, in practice, iFood faces several core challenges related to budget allocation, customer data analysis and regulatory compliance.

First, a common problem is the waste of marketing funds as a result of poorly targeted audiences. More than 50% of marketers waste between 0% and 20% of their budgets on ineffective channels, and over 15% waste over 40%, according to a global survey conducted by Rakuten Marketing and eMarketer (2018). Similarly, according to Forrester Research, inaccurate targeting or out-of-date customer data wastes about 37% of advertising budgets. More importantly, Next&Co (2024) found that brands squandered about USD 97.1 million in Q1 2024 alone, which is equal to 42% of the total audited digital media spend. This amount rose to USD 6.149 billion in 2023, making up 43% of Australia's total spending on digital advertising. Notably, waste increased further in Q2 2024, reaching USD 123.1 million, the highest amount ever recorded (44% of total audited digital ad spend) (Branding in Asia, 2024). These figures highlight a key challenge: the

lack of effective targeting and budget optimization strategies in marketing campaigns can lead to inefficient spending, directly affecting business profitability.

Within the constraints of available marketing budgets, online platforms frequently work toward particular business goals, like raising user engagement or retention rates. However, allocating promotional incentives effectively remains a significant challenge, as each user responds differently to various types of offers. Ineffective campaigning and budget waste can arise from poor allocation. Therefore, developing strategies to optimize the distribution of incentives across diverse user segments is essential to maximize marketing efficiency and return on investment (Ziang Yan et al., 2023).

Second, imbalanced customer response data poses difficulties in building accurate predictive models. In iFood's dataset, only about 15% of customers respond to campaigns, creating a severe imbalance that causes machine learning models to be biased toward the "non-response" group. This leads to high accuracy but low recall, resulting in missed potential customers. In practice, a survey by HubSpot Blog Research revealed that although the average email marketing open rate is 46-50%, the click-through rate is only 2.6-3%, clearly indicating that most customers do not take action after receiving messages.

Third, the lack of multi-channel data integration and suboptimal customer segmentation is a major barrier to personalized experiences. Gao et al. (2020) claim that fragmented customer experiences and decreased multichannel campaign effectiveness result from a lack of channel integration. Omnichannel customers spent 10% more online and 4% more per in-store purchase, according to a different Harvard Business Review study with 46,000 participants. More significantly, they increased repeat business by 23% over a six-month period. Therefore, customer segmentation will be imprecise if iFood does not efficiently use data from channels like the website, catalogs, stores, or monthly visits. This will result in unreliable response prediction models and decreased marketing effectiveness.

Finally, privacy and regulatory compliance risks are also critical challenges. Saura (2024) warns that exploiting personal data such as income, age and behavioral history in digital marketing carries potential privacy violations if transparency is lacking. With

stringent regulations such as GDPR in Europe and LGPD in Brazil, iFood may face financial penalties and loss of customer trust if transparency in data collection and processing is not ensured.

In summary, iFood confronts four main groups of challenges: marketing budget wastage, imbalanced customer response data, lack of multi-channel data integration and privacy risks. These are not only technical issues but also critical determinants of business efficiency and competitive position. Analyzing customer segmentation and response prediction using machine learning in marketing campaigns is key to helping iFood optimize resources, increase ROI and create seamless customer experiences in an increasingly competitive landscape. Consequently, this can generate positive outcomes for the company in terms of both revenue and brand value.

Related Researches

In recent years, numerous studies have focused on predicting customer responses to marketing campaigns and customer segmentation in multichannel retail. These works not only provide important theoretical and methodological foundations but also highlight the potential applications of machine learning models in optimizing the effectiveness of marketing campaigns.

In the context of modern marketing, predicting customer responses to campaigns is a key task, especially when marketing resources need to be optimized. El-Hajj & Pavlova (2024) focused on developing and evaluating classification models to predict whether a customer would respond in their study on customer responses to marketing campaigns. The models tested included Decision Tree, Random Forest, Gradient Boosting and Logistic Regression, while also assessing the impact of class imbalance and data imbalance handling strategies such as undersampling. Additionally, the study explored another important aspect: extracting key features and decision rules to support interpretability for marketing professionals. The results indicated that, before resampling, the Decision Tree model achieved high accuracy (~87.3%) but low recall (~44%), demonstrating that accuracy alone does not reflect true performance on the minority class. When

undersampling was applied, recall improved, but the total number of samples decreased, posing a risk to generalizability. Ensemble and boosting methods (XGBoost, LightGBM) achieved higher overall performance but reduced interpretability compared to Decision Tree. This highlights two phenomena: metrics sensitive to imbalance, such as Recall, F1-score and AUC-PR, should be prioritized and strong models should be combined with explanation tools like SHAP to achieve both predictive effectiveness and interpretability for marketers.

In the banking sector, predicting customer responses to marketing campaigns plays an important role in optimizing cost and profit. Apampa, Olatunji (2016) evaluated the effectiveness of multiple classification models, including Logistic Regression, Decision Tree, Naïve Bayes and Random Forest, while also examining the impact of data preprocessing, feature selection and class imbalance on model performance. The study applied the CRISP-DM framework, ensuring logical consistency from business understanding to model deployment. The results showed that, when data were balanced, models achieved better performance, with Logistic Regression and Decision Tree providing easily interpretable results through coefficients and rules, whereas Random Forest and other tree-based ensembles achieved better overall performance but were harder to interpret. The paper noted that overfitting occurs easily with CART when data are small or features are numerous and improper application of sampling techniques can lead to data leakage. The study emphasized the trade-off between interpretability and performance and recommended using explanation tools such as SHAP when applying ensembles for marketing decision-making.

Both studies by El-Hajj & Pavlova (2024) and Apampa, Olatunji (2016) highlight the challenge of class imbalance and the necessity of using imbalance-sensitive metrics instead of relying solely on accuracy. Decision Trees offer good interpretability but lower performance, while ensemble and boosting methods provide high performance but are harder to explain. Sampling strategies, imbalance handling and the application of CRISP-DM are suggested to optimize both technical and business outcomes. Therefore, a

reasonable pipeline for a project should combine strong models for scoring with explanation tools to optimize lift/precision while also providing insights for marketers.

In parallel, Guido et al. (2011) investigated the applicability of Neural Networks in customer targeting for direct marketing. The results showed that ANNs can outperform Logistic Regression and Decision Trees in complex and nonlinear data scenarios, thereby confirming the potential of nonlinear models in predicting customer responses. However, ANNs face challenges in interpretability, which limits their ability to provide direct insights for marketers. From this, future research needs to find ways to balance high performance and model transparency, particularly in a multichannel context.

Meanwhile, in the direction of combining segmentation and prediction, Olson et al. (2012) emphasized that using RFM together with predictive models such as Logistic Regression, Decision Trees, or Neural Networks enhances accuracy and marketing effectiveness. This is especially meaningful for iFood, where different customer groups exhibit distinct behaviors across channels that need to be modeled separately. However, a limitation of this study is that it did not take multichannel data into account and did not evaluate business impacts. This opens a direction for our group to link segmentation-prediction models with real multichannel contexts and assess campaign-level effectiveness.

In addition, Chian et al. (2024) evaluated models on banking data and concluded that Random Forest achieved the best performance, particularly after selecting important features. The results highlight the role of feature selection in improving model performance. This provides important insights for retail businesses when applying machine learning in multichannel marketing. However, the limitation lies in the relatively modest accuracy for real-world applications and the lack of attention to interpretability. This suggests a future research direction for our group to enhance both performance and transparency of models in multichannel contexts.

Moreover, Lin (2025) implemented multiple machine learning models to predict consumer behavior and support precision marketing. The findings revealed that boosting algorithms such as XGBoost and CatBoost achieved superior performance with high F1-

scores and ROC-AUC values. This study confirms the strength of boosting models in handling complex data, thereby optimizing campaign effectiveness. Nonetheless, the limitation is that the research primarily focused on accuracy, with little discussion of practical applications and implementation costs. This opens a necessary direction to consider overall efficiency in multichannel contexts.

Other studies expanded input data; notably, Dai et al. (2021) used behavioral data from social networks to successfully predict customer engagement, showing that clickstream data and content features play a crucial role in forecasting responses. This provides evidence that online behavioral data can be integrated to improve prediction in multichannel marketing. However, the limitation is that the study focused mainly on engagement metrics without directly linking them to business conversions, which opens a research direction for our group to connect behavioral responses with actual performance indicators in multichannel settings.

From the segmentation perspective, Gautam (2022) affirmed that K-Means can cluster customers into homogeneous groups to support sustainable strategies. This forms an important basis for application in multichannel retail environments, where diverse customer behaviors require clear segmentation. However, the method depends heavily on the number of clusters chosen and does not evaluate changes in segmentation over time. This suggests a research direction for our group to examine the stability of segments and extend them across multiple channels.

Also related to campaign optimization, Mandapaka et al. (2014) emphasized the role of response models in call centers to identify potential customers and optimize resources in direct marketing. The results indicated that predicting customer behavior in advance helps improve outreach effectiveness and reduce costs. This provides valuable experience for application in multichannel retail, where customer interactions occur simultaneously across various channels. However, the limitation is that the study focused only on call centers without considering other channels. This opens a direction for our group to expand the scope across multiple channels to build a more comprehensive picture.

In addition to the primary studies, several other works provide supporting insights and contribute supplementary value to the methodological foundation. Chioma Ikeh (2025) focuses on real-time campaign optimization based on multimedia interaction data, thereby suggesting approaches for handling complex data and enhancing behavior prediction capabilities. Yaiprasert and Hidayanto (2023) emphasize the role of ensemble and boosting techniques in improving model performance, particularly useful when implementing stacking methods. Meanwhile, Ghorban Tanhaei et al. (2024) synthesize and compare multiple forecasting models, reinforcing the theoretical basis for analyzing customer trends and behaviors. Although these studies do not directly shape the iFood-style segmentation or response framework, they clarify technical aspects and data processing methodologies within the research context.

Additionally, another group of studies contributes academic value through strategic, ethical and customer experience perspectives. Saura et al. (2024) raise critical ethical concerns and privacy paradoxes in AI-based marketing, highlighting limitations in practical applications. In parallel, Balbín Buckley (2024) investigates the impact of channel integration on customer experience, reflecting strategic trends in multichannel marketing. Finally, Gerea (2021) provides a systematic overview of customer experience management in omnichannel environments, offering a foundational theoretical perspective. Overall, this body of work does not directly support methodological development but plays a crucial role in broadening the discussion, supplementing the theoretical framework and suggesting potential directions for strategic application in the study.

In summary, the body of research reviewed highlights significant progress in the fields of customer response prediction and segmentation within multi-channel retail and marketing. Previous studies have confirmed the role of segmentation and prediction in enhancing marketing effectiveness, yet they still lack integration with the multi-channel context and pay limited attention to model interpretability. This study will build upon prominent achievements such as boosting, RFM combined with K-Means and imbalance handling via SMOTE/class-weight, while additionally incorporating uplift analysis and

model interpretability (SHAP) to support decision-making. Accordingly, the research aims to contribute an optimized solution for iFood's multichannel marketing campaigns, helping the company allocate resources efficiently, improve ROI and enhance customer experience in an increasingly competitive environment.

Business Objectives

The project focuses on building a tool to predict customer behavior in marketing campaigns at multi-channel retail stores. We apply modern machine learning algorithms to improve the effectiveness of each campaign.

In the context of increasingly fierce markets and constantly escalating customer acquisition costs, the ability to accurately predict how customers respond has become the key to successful marketing activities. The research aims at five specific objectives:

First, evaluate performance across sales channels: Understand the relationship between customers' favorite shopping channels (directly in stores, online, or through catalogs) and their level of engagement in marketing programs. Based on this analysis, we can identify the channels that deliver the strongest results.

The second objective involves pinpointing key drivers of customer behavior. We'll examine whether elements like income levels, age groups, purchasing patterns, and product preferences shape how customers engage with marketing efforts.

For the third goal, we're taking a data-driven approach to customer segmentation. By applying K-Means clustering techniques, we can organize customers into distinct groups based on who they are and how they shop. This segmentation allows us to craft targeted marketing approaches that resonate with each specific group.

Fourth, predict response: Develop machine learning models that can predict the probability that a customer will participate in a marketing campaign.

Fifth, optimize marketing resources: Apply these models to identify customer groups with high conversion potential, thereby allocating budgets more reasonably and cutting costs in ineffective channels.

Through this research, the team hopes to create a scientific analytical framework that businesses can apply in practice to: refine marketing strategies, increase sales across multiple channels, provide better customer experiences, and promote long-term business growth.

Business Questions

We're looking to improve how marketing campaigns perform across different sales channels. Here's what we want to find out:

Q1: How do customers' preferred purchasing channels (store, web, catalog) relate to their likelihood of responding to marketing campaigns?

Q2: Which customer attributes contribute most to predicting the probability of campaign response?

Q3: What are the main customer segments identified through clustering?

Q4: How do these customer segments differ in their campaign response patterns, and what strategic implications can be drawn for targeted marketing?

Objects and scopes

Objects

The study used a public pilot dataset of more than 2,000 customers, shared by iFood on GitHub for recruitment on consumer behavior analysis and customer response to marketing campaigns in an omnichannel retail environment.

Scopes

Time Scope: The dataset reflects iFood customer information from 2018 to 2020, including shopping data and results from five marketing campaigns.

Space Scope: The dataset focuses on iFood customers in Brazil, capturing their demographic characteristics and consumer behavior, as well as their interactions with marketing campaigns and various purchase channels.

Experimental method

This study employs the CRISP-DM model (Cross-Industry Standard Process for Data Mining), a prevalent methodology for enhancing the performance of data mining initiatives (Chapman et al., 2000). The model offers a step-by-step but flexible process that helps connect data analysis with business goals (Schroer et al., 2021). As Martínez-Plumed et al. (2019) mention, CRISP-DM is still relevant today in data science, especially for projects that have clear objectives and workflows, while exploratory ones might need a looser approach. Many recent studies also show that CRISP-DM is still applied a lot in data mining, especially when it comes to predicting customer reactions to marketing campaigns (Apampa, 2016; Lin, 2025). Following this model, our study goes through six main steps, which are illustrated in Figure 0.1.

Figure 0.1 CRISP-DM framework applied in this study. (Source: Research team)

The research steps are specified as follows:

- **Business Understanding:** Clearly outline the project goals and ensure they support business objectives. In this study, the main aim is to predict customer responses to marketing campaigns and segment customers. This helps businesses increase response rates, allocate resources more effectively and run campaigns more efficiently.
- **Data Understanding:** Examine the data to understand its structure and contents. This includes checking distributions, identifying missing values and outliers and performing exploratory data analysis (EDA) to uncover patterns and relationships.
- **Data Preparation:** Prepare the data by handling missing values, removing outliers and standardizing formats. Generate new features from customer behavior, eliminate irrelevant variables to simplify the dataset and encode and normalize data for model training.
- **Modeling:** Apply classification models such as Logistic Regression, Random Forest and XGBoost, along with K-Means clustering.
- **Evaluation:** Assess the performance of classification models using Accuracy, Precision, Recall, F1-score and ROC-AUC. For clustering, evaluate using the Silhouette score and analyze the characteristics of each customer segment.
- **Deployment:** Present the findings through dashboards and turn them into practical recommendations, including improving marketing campaigns, developing personalized engagement strategies and setting up ongoing monitoring and model retraining to keep up with changing customer behavior.

Tools and Programming languages

In this project, the team will use Python as the primary programming language due to its flexibility and rich libraries designed for data science and machine learning. Python supports the entire analytical workflow, including data cleaning, exploratory analysis,

visualization, customer segmentation and predictive modeling, key steps for evaluating customer responses to marketing campaigns.

Although the dataset used in this study is not very large, it contains numerous customer attributes and diverse campaign information, requiring efficient preprocessing, integration and model development. Python is particularly suitable here as it enables seamless handling of structured demographic and behavioral data, while also offering scalable methods for building both classification and clustering models to predict customer responses.

The main development environment chosen by the team is Google Colab, a cloud-based platform with pre-installed libraries, GPU support and seamless integration with Google Drive. This environment not only accelerates model training but also facilitates collaboration and reproducibility, ensuring efficiency and accessibility.

Additionally, to support business reporting and decision-making, Power BI is used as a supplementary tool in this study. While Python libraries meet the technical and in-depth analytical requirements, Power BI allows the transformation of complex model results into interactive dashboards and actionable insights. This combination ensures that the research findings can be effectively communicated to stakeholders, bridging the gap between technical accuracy and managerial needs.

Structure of project

This project consists of 4 main chapters along with supporting sections such as Abstract, References and Appendices. The structure is organized as follows:

- **Project Overview** - Introduce the research context, state the problem, objectives, scope and contributions of the project.
- **Chapter 1: Theoretical Background** - Present fundamental theories and related works, summarize key concepts and models relevant to the project.
- **Chapter 2: Data Preparation** - Describe the dataset, preprocessing steps, feature engineering and methods used to clean and transform the data for analysis.

- **Chapter 3: Experimental Results and Evaluation** - Present the experimental setup, results obtained, model performance and provide evaluation using appropriate metrics.
- **Chapter 4: Visualization and Discussion** - Provide visualization of results, analyze findings, discuss implications, limitations and propose directions for future work.
- **Conclusion and Future Works** - Summarize main findings, contributions of the project and suggest possible future research directions.

Chapter 1. Theoretical Background

Chapter 1 presents the theoretical foundation of the study by first providing an overview of the food order and delivery sector and introduces machine learning concepts applied in marketing, focusing on customer response prediction and segmentation. The chapter highlights the use of predictive classification models such as Random Forest and XGBoost, as well as clustering approaches, with a special emphasis on the K-Means algorithm for customer segmentation. These methods are employed to analyze customer behavior patterns and identify distinct customer groups. This framework establishes the basis for applying data-driven techniques in later chapters to address key business challenges in the online food delivery industry.

1.1 Online Food Delivery (OFD) Industry

Online Food Delivery (OFD) refers to the segment of online food retail in which consumers order prepared or ready-to-eat meals via digital platforms (mobile apps or websites) and receive them at a chosen location through platform-mediated payment and last-mile logistics; OFD is characterized by platform ordering, menu/merchant aggregation and service attributes such as convenience, variety and delivery reliability (Bennett et al., 2025; Meena et al., 2022)

The global Online Food Delivery (OFD) market has been experiencing significant growth in recent years. According to Grand View Research (2024), the OFD market is estimated to reach approximately USD 288.84 billion in 2024 and is projected to grow to USD 505.50 billion by 2030, with a compound annual growth rate (CAGR) of around 9.4% during the period from 2025 to 2030.

The rise of online food delivery (OFD) services can be attributed to the evolving lifestyles of urban consumers. People turn to these services for many reasons, but the most common one is unsurprising: the need for quick and convenient meals throughout the workday, whether during office hours or after a long day at work. With multiple platforms readily available, consumers no longer have to worry about meal planning, whether it

involves cooking at home, dining in at a restaurant, or picking up food to take back to the office or home (Lau et al., 2019).

OFD services have reshaped consumer habits, particularly in urban areas, to the point where ordering food online has become a normal, routine part of everyday life. In recent years, more people have adopted food delivery not only because of the fast-paced nature of modern living but also because it offers opportunities to explore a wider variety of restaurants. For many busy city dwellers, food delivery provides a practical option that fits seamlessly into their daily routines. It allows them to enjoy fresh and healthy meals at home or at the office without interrupting their workflow or spending extra time traveling and waiting at restaurants. In essence, OFD services save time and effort by eliminating the need to leave home or the office to get food. Beyond convenience, these services are also gradually transforming the food and beverage industry. They open new avenues for business growth, improve employee productivity and ensure greater accessibility to diverse dining experiences (Lau et al., 2019).

1.2 Introduction to Machine Learning

Machine Learning is the science of developing algorithms and statistical models that computer systems use to perform tasks based on patterns and inference without specific instructions. Machine learning has the main purpose of training computers to automatically "learn" without human intervention or assistance to perform and adjust actions. Computer systems use machine learning algorithms to process large volumes of historical data and identify data patterns. This allows them to predict results more accurately from the same set of input data (Dr.Chitra et al., 2013).

Within the field of machine learning, two of the most fundamental paradigms are supervised learning and unsupervised learning. These approaches differ primarily in how data is structured and how algorithms learn from it.

Supervised learning refers to a method in which algorithms are trained on labeled datasets. Each data instance consists of an input variable and a corresponding output variable, allowing the model to learn the mapping from inputs to outputs. During the

training phase, the algorithm identifies patterns and relationships between the two and once the process is complete, it can predict the appropriate label or value for unseen data based on prior knowledge. Supervised learning tasks are commonly categorized into two types: regression, where the objective is to predict continuous values and classification, where the goal is to assign discrete labels to inputs (Hiran et al., 2022).

In contrast, unsupervised learning operates on datasets without labels, meaning only input variables are available and no corresponding outputs are provided. In this case, there is no "supervisor" guiding the process and no predefined target to learn from. Instead, the algorithm autonomously explores the underlying structure of the data by identifying patterns, correlations, or similarities. The primary objective is to group similar data points or extract meaningful features for more efficient representation. Common approaches include clustering and dimensionality reduction (Hiran et al., 2022).

1.3 Customer Response Marketing Campaign Prediction

In the context of intense competition and rising marketing costs, predicting customer responses to marketing campaigns has become a crucial factor for businesses to optimize their resources. The application of customer classification models allows businesses to reduce the number of customers they need to approach while still improving response rates, thereby saving costs for marketing campaigns (Moro et al., 2014).

Various machine learning algorithms have been applied to tackle the problem of customer response prediction, each with its own advantages and drawbacks. In their study of a Portuguese bank's telemarketing campaign, Moro et al. (2014) found that SVM gave the most accurate predictions, while Decision Tree was simpler for managers to understand. Apampa (2016) also compared Random Forest, Logistic Regression, Decision Tree and Naïve Bayes and demonstrated that when handling imbalanced data, Decision Tree and Random Forest significantly improved accuracy and recall, making it easier to identify potential customers. Chian et al. (2024) showed that XGBoost and Random Forest outperformed traditional models in terms of Accuracy and AUC, highlighting the power of ensemble algorithms in optimizing marketing campaigns. Similarly, Sousa (2024) applied

Decision Tree, Logistic Regression and KNN on the IFood dataset, confirmed that these models not only provided strong predictive capability but also supported managers in making more efficient marketing resource allocation decisions.

1.4 Customer Segmentation

Customer segmentation is the process of dividing the market into distinct, homogeneous and meaningful customer groups based on different characteristics and attributes. This approach is a key tool in differentiated marketing, enabling businesses to better understand customers and to develop strategies tailored to each target segment (Kotler & Keller, 2011). When combined with business knowledge, clustering algorithms can facilitate more effective behavior-based segmentation, thereby improving customer relationship management and enhancing competitive advantage (Ziafat & Shakeri, 2014). According to Shirole et al. (2021), integrating the RFM (Recency, Frequency, Monetary) model with the K-Means algorithm can generate clearer customer groups, helping businesses identify loyal customers, potential customers and those at risk of churn. In addition, Sousa (2024) applied K-Means clustering to segment customers, laying the foundation for more effective customer response prediction models. While traditional methods often rely on RFM models to measure customer value, modern algorithms such as K-Means allow businesses to leverage larger and more complex datasets.

1.5 Predictive Classification Models

1.5.1 Logistic Regression

Logistic Regression is a statistical model designed to describe the probability of the occurrence of a binary dependent variable based on a linear combination of independent variables. In contrast to standard linear regression, logistic regression maps the output values into the interval [0,1] using the logistic (or sigmoid) function, which makes it easy to interpret them as probabilities. The algorithm has developed into a key technique in statistics and machine learning since David Cox first presented it in 1958.

According to Cox (1958), this relationship can be expressed through the logit function:

$$log\frac{p}{1-p} = \alpha + \beta x$$

Where $p = Pr(Y = 1|X)$ denotes the probability that the dependent variable takes the value 1 given the independent variable X. Here, α is the intercept term and β is the regression coefficient, which reflects the change in the log-odds of the outcome for a one-unit increase in X. From this formulation, the probability p can also be written using the logistic function:

$$p = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

Numerous scientific and practical domains have made extensive use of logistic regression. This model is frequently used in medicine to forecast post-operative survival rates, disease risk, or the efficacy of treatments (Hosmer et al., 2013). In epidemiology and biomedical research, logistic regression is essential for building clinical data-based predictive models that aid in risk assessment and treatment selection (Huang et al., 2025). According to Balyemah et al. (2024), e-commerce uses logistic regression to forecast consumer buying patterns based on attributes like age, gender, browsing history, and product reviews. This helps to improve operational efficiency and optimize marketing strategies. In CRM and marketing, the model is also applied to analyze and forecast conversion rates in loyalty programs, helping businesses adjust marketing expenditures appropriately and monitor campaign effectiveness (Singh & Rao, 2025).

Thanks to its simplicity, probability-based interpretability and high efficiency. Logistic Regression remains a trusted foundational model despite the emergence of more complex techniques.

1.5.2 Random Forest

Random Forest is an ensemble learning method that belongs to the group of supervised learning methods, used for both classification and regression, proposed by Breiman in 2001. The algorithm constructs multiple decision trees on bootstrap samples and combines their predictions by majority voting (for classification) or averaging (for regression) (Breiman,

2001). The random selection of feature subsets at each split helps reduce overfitting and improve the accuracy of the model. (Hastie et al., 2009).

Basic Formula

With B decision trees, the average prediction for an observation x′ is calculated as follows:

$$\hat{f}(x') = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Where, $f_b(x')$ denotes the prediction of the b^{th} tree.

The feature importance is determined based on the decrease in impurity:

$$Imp(x) = \frac{1}{N_T}\sum_{trees}\sum_{nodes\ split\ on\ x} p(j)\Delta i(j)$$

Where $p(j)$ is the proportion of samples at node j and $\Delta i(j)$ is the decrease in impurity caused by the split on feature x.

Random Forest has been widely applied across various research domains, particularly in data science and business. In the field of consumer behavior, Joshi et al. (2018) employed Random Forest to predict online shopping behavior of customers in the Indian market. The findings indicated that the model was able to accurately classify potential customer groups with higher response rates, thereby supporting retailers in planning more effective multichannel marketing strategies. In addition, Random Forest has also been leveraged in big data analysis to evaluate feature importance. For example, Chen et al. (2020) applied the algorithm to compare the effectiveness of feature selection with traditional methods such as SVM and KNN. Their results demonstrated that Random Forest not only achieved high predictive accuracy but also provided valuable insights into which features had the strongest impact on prediction outcomes, thereby simplifying the model without compromising performance.

In the healthcare domain, Random Forest has proven effective in predicting the likelihood of diabetes (Sisodia et al, 2018; Kavakiotis et al., 2017) and in medical image classification, particularly in diagnosing diabetic retinopathy and analyzing MRI scans (Islam et al., 2017). These applications highlight that Random Forest is not only a powerful tool for classification but also a valuable decision-support method in both academic research and practical contexts.

1.5.3 XGBoost

The XGBoost (Extreme Gradient Boosting) algorithm, developed by Chen & Guestrin (2016), is an optimized version of Gradient Boosting. Similar to other boosting methods, XGBoost sequentially builds multiple decision trees, where each new tree is trained to reduce the residual errors left by the previous model. The strengths of XGBoost lie in its use of a second-order approximation of the loss function, a regularization mechanism to mitigate overfitting and optimization for large-scale data processing. As a result, XGBoost has become one of the most popular and effective algorithms in supervised machine learning.

Basic Formula

At iteration t, the loss function is approximated by the second-order Taylor expansion:

$$bj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Where

- $g_i = \partial_{f_{t-1}} L(y_i, f_{t-1}(x_i))$ is the gradient,
- $h_i = \partial^2_{f_{t-1}} L(y_i, f_{t-1}(x_i))$ is the Hessian,
- $\Omega(f_t)$ is the regularization term.

In the banking and finance sector, numerous studies have employed XGBoost to predict customer behavior. For example, Yang et al. (2025) applied XGBoost to forecast customers' likelihood of accepting term deposit products. The model demonstrated superior accuracy

compared to traditional classification methods, enabling banks to enhance conversion rates and optimize marketing costs.

Another study by (Zheng, 2025) applied XGBoost in combination with imbalanced data handling techniques to predict the effectiveness of direct marketing campaigns. The model was able to segment customers based on demographic characteristics such as age, occupation, income and education level, thereby supporting banks in designing more personalized and effective marketing strategies than traditional direct marketing approaches.

In the context of e-commerce and digital marketing, XGBoost has also been widely applied. Wang et al. (2023) developed a model to predict user purchasing behavior based on behavioral data (such as login time and multi-channel interaction frequency). By integrating XGBoost, the model analyzed feature importance and accurately predicted purchasing behavior, achieving an accuracy of 0.9761, an F1-score of 0.9763 and an ROC of 0.9768. These results outperformed many widely used algorithms such as KNN, SVM, Random Forest and traditional neural networks.

These findings demonstrate that XGBoost is not only a theoretically powerful algorithm but also delivers substantial practical value across multiple domains, particularly in marketing, finance and e-commerce.

1.6 Clustering Models: K-Means

The K-Means clustering algorithm is a method used to group data based on a partitioning system (Jain & Dubes, 1988). K-Means works by dividing a dataset into one or more clusters so that objects with similar characteristics are grouped together in the same cluster, while objects with different characteristics are assigned to separate clusters. The advantage of K-Means is its simplicity, ease of implementation and wide application in areas such as customer segmentation. However, the main limitation of this algorithm is that it is highly sensitive to the initialization of the initial clusters (Rahman et al., 2017).

In the K-Means algorithm, certain basic principles must be followed during its application. First, the number of clusters k needs to be determined at the outset. The data attributes used in the algorithm must be numeric to allow accurate distance calculations. The algorithm also has limitations regarding the types of attributes it can process. Additionally, the complexity of K-Means is linear with respect to the dataset size, enabling it to efficiently handle large datasets (Rahman et al., 2017).

The basic process of K-Means clustering is as follows (Johnson & Wichern, 2002):

(1) Determine k, the number of clusters to be created.

(2) Randomly initialize k centroids (cluster centers).

(3) Calculate the distance of each data point to each centroid. One common method for distance calculation is the Euclidean Distance (Lubis, 2017).

$$\text{Euclid}(x,y) = \sqrt{\sum_{i}^{n} (xi - yi)^2} \quad , i$$

In which:

- x: the first object
- y: the second object
- n: the number of attributes of the first and second objects

(4) Each data point is assigned to the nearest centroid.

(5) The new centroid positions are determined by calculating the average value of all data points belonging to the same centroid.

(6) Return to step 3 if the new centroid positions differ from the previous ones.

Customer segmentation is a crucial step in marketing strategy, helping businesses understand groups of customers with similar behaviors and needs. Many studies have applied the K-Means algorithm to address this issue. Perapu (2025) used K-Means to segment customers based on shopping behavior and demographic characteristics at a retail store, helping identify similar customer groups and design more effective personalized marketing strategies.

In the banking sector, Barkhordar et al. (2021) combined K-Means with deep learning techniques to classify customers based on transaction behaviors, highlighting K-Means' ability to cluster accurately and flexibly when handling different types of data.

In Kenya, Omol et al. (2024) also applied K-Means to segment customers in grocery stores, demonstrating the algorithm's effectiveness in handling real-world data and supporting local marketing strategies.

Chapter 2. Data Preparation

This chapter focuses on the data preparation process to support customer segmentation and response prediction. The data cleaning stage removes missing values and duplicates, while feature engineering is applied to create derived variables that better capture customer characteristics. Next, descriptive statistics and exploratory data analysis (EDA) are conducted to understand the data distribution, detect outliers and uncover important patterns in customer behavior. Finally, the pre-processing stage includes outlier handling, encoding categorical variables and scaling numerical features, ensuring that the dataset is standardized and ready for clustering and predictive modeling in Chapter 3.

Link Code Chapter 2

2.1 Data Cleaning

After collecting the iFood Marketing Campaigns dataset from Kaggle, the team conducted an initial exploration to determine the shape and size of the dataset. The original dataset contained 2,205 observations and 30 variables, including numerical variables such as Income, Age, Recency, MntWines, MntMeatProducts and categorical variables such as marital_status and education. The inspection results indicated that the dataset contained no missing values and no duplicate records. Some variables that did not contribute to the analysis, such as the Index, Complain column, were removed to avoid redundancy. After the cleaning steps, the final dataset consisted of 2,205 observations and 28 variables, ensuring integrity and readiness for further analysis.

2.2 EDA

2.2.1 Demographics

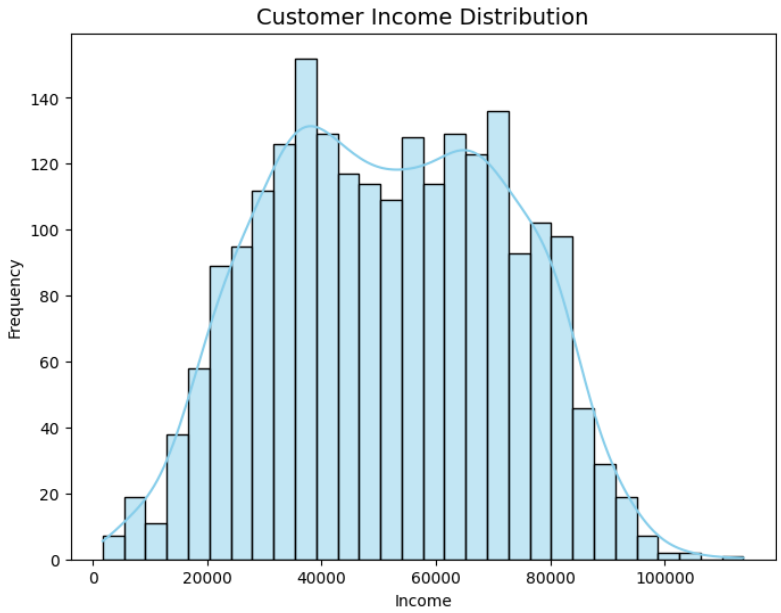

Figure 2.1 Histogram of Customer Income Distribution

The chart reveals income spread across customers in a bell-curve pattern, with most people earning between 30,000 and 70,000. The bulk fall into middle-income territory. A handful earn over 90,000; these are the high-end shoppers. There's a slight tilt to the right, meaning a few customers pull in notably higher incomes. The sweet spot sits around 40,000-50,000, where we see the highest concentration. The data looks clean without major

outliers, which makes it practical for dividing customers into groups and studying how they spend.

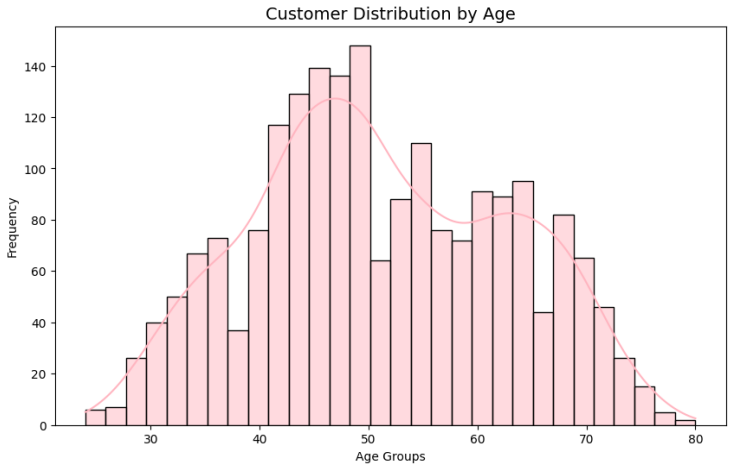

Figure 2.2 Histogram of Customer Distribution by Age

Most customers are between the ages of 40 and 60, which is the dominant age group in the data set. The distribution is roughly bell-shaped but slightly skewed towards older age groups, peaking around 45-50. We see fewer customers under 30 or over 70. This suggests that the core customer base is middle-aged, often with a stable income and is the main target for marketing activities.

2.2.2 Purchasing Behavior

Looking at how people buy reveals a lot about their relationship with products and channels. Things like how often they purchase, when they last shopped, and what they spend on different categories show their engagement level, buying patterns, and overall worth to the business.

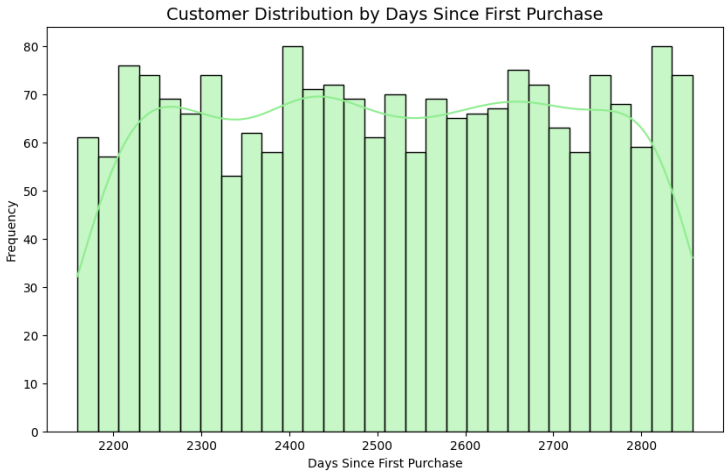

Figure 2.3 Histogram of Customer Distribution by Days Since First Purchase

The graph shows that customer retention is spread out across the system, with no particular timeframe standing out. This means that there are some long-term customers and some new customers coming in, and there is no spike in growth at any one stage. Instead, it appears that the business has been attracting new customers at a steady rate over time, rather than seeing all the growth happen all at once.

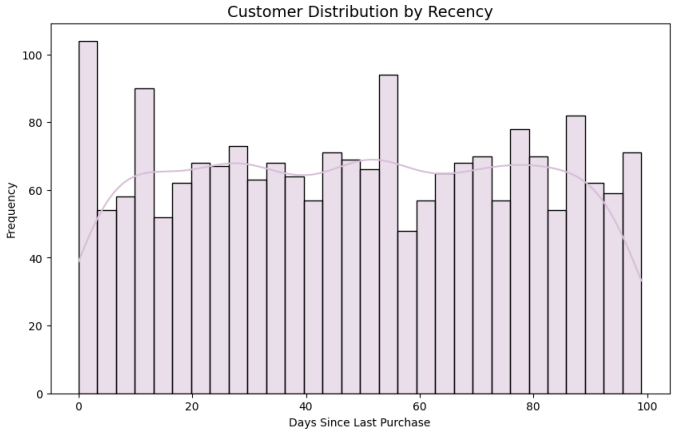

Figure 2.4 Histogram of Customer Distribution by Recency

The chart illustrates the distribution of Recency, which represents the number of days since the last purchase. Days since last purchase is fairly evenly distributed and spread out, with no particular customer group dominating. A notable segment of customers has a very low Recency (less than 10 days), indicating they have made a recent purchase, while another segment shows a high Recency (over 80 days) who have not made a purchase in a long time. Overall, the chart shows that purchasing activity is steady over time.

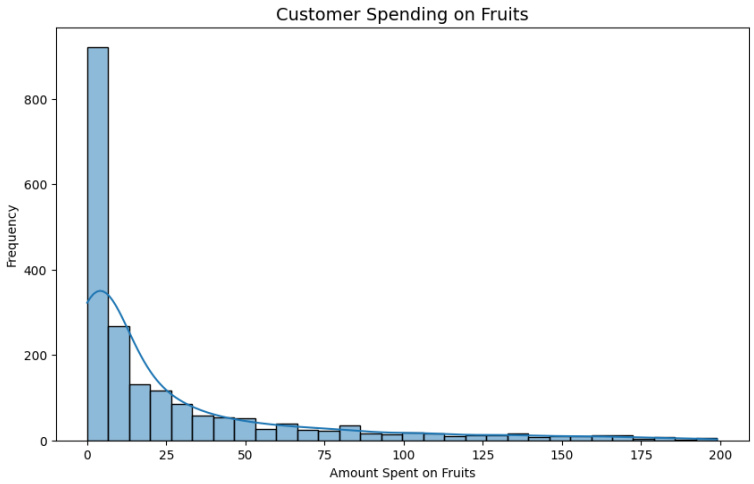

Figure 2.5 Histogram of Customer Spending on Fruits

The graph shows the distribution of customer spending on fruits. As shown in the graph, most customers spend very little, mostly under 20 units, while only a small group spends much more than the average. This distribution is strongly skewed to the right, indicating a large segment of customers spending at the low end and a small number of high spenders. This suggests that fruits are not a major product group for most customers, but there is a small, loyal or high-spending segment that can be targeted through specialized marketing campaigns.

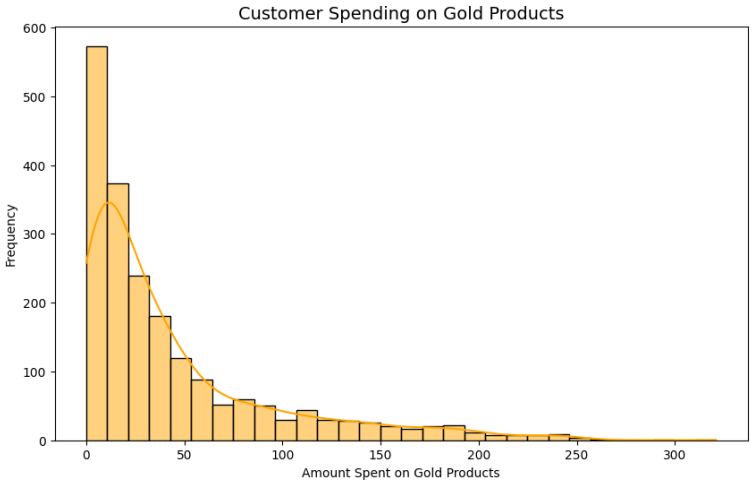

Figure 2.6 Histogram of Customer Spending on Gold Products

The chart shows that most customers spend very little on gold products and most of them spend less than 50 units. There is a small group that spends significantly more. The distribution is heavily skewed to the right, meaning that most spend little, but there are a few who spend very much. They may be VIP customers or loyal fans of this product. This suggests that gold products are niche products, not everyone has a need to buy them, but still bring in good revenue thanks to the high-value customer group.

Figure 2.7 Histogram of Customer Distribution by Number of Web Purchases

The chart shows the number of times customers buy online. Most only buy 1-5 times, and buying more than 10 times is quite rare. The distribution is heavily skewed to the right, meaning that online shopping is not the main channel for most people. However, the small group of frequent buyers can be a notable target for developing an e-commerce strategy, building a customer loyalty program or launching special offers for the online channel.

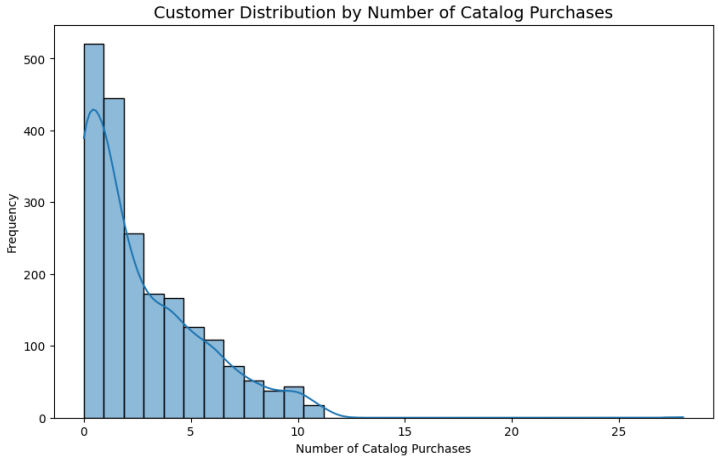

Figure 2.8 Histogram of Customer Distribution by Number of Catalog Purchases

The chart shows that most customers buy from catalogs 0-3 times, while more than 5 times is rare. The distribution is heavily skewed to the right, meaning that catalogs are not a popular channel. However, there is a small group of loyal customers who buy regularly through this channel. This group is worth retaining and developing through personalized campaigns or special offers to encourage them to buy more catalogs.

Figure 2.9 Histogram of Customer Distribution by Number of Store Purchases

The chart shows that most customers make 2-5 in-store purchases, with 3 being the most common. The distribution is slightly skewed to the right, meaning there is a small group that makes purchases more frequently than average, but the majority are still in the low to medium range. This confirms that physical stores are still an important channel. Companies can encourage more frequent purchases with loyalty programs or in-store promotions.

2.2.3 Outlier Analysis

Analyzing outliers is an important step in assessing the stability and quality of the data, while also helping to identify customers with behaviors or characteristics that differ significantly from the majority. Outliers may sometimes indicate data errors, but in many cases, they provide valuable insights into special customer groups, such as those with exceptionally high spending levels or incomes outside the common range.

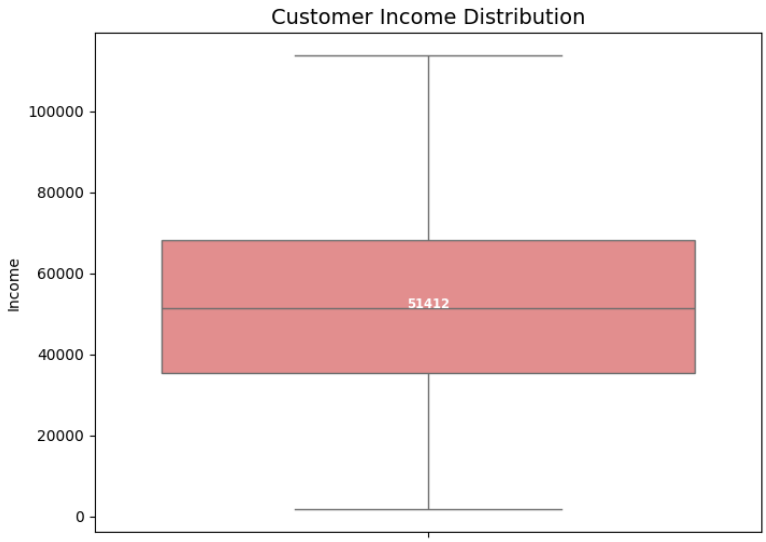

Figure 2.10 Box Plot of Customer Income Distribution

The box plot illustrates the distribution of customer income, with most values falling within the range of 30,000 to 70,000 and a median of around 50,000. The relatively wide range reflects diversity in income levels among customers. Notably, no outliers are observed, indicating that the income data is stable and not influenced by extreme values, making it suitable for further analyses such as customer segmentation or spending prediction.

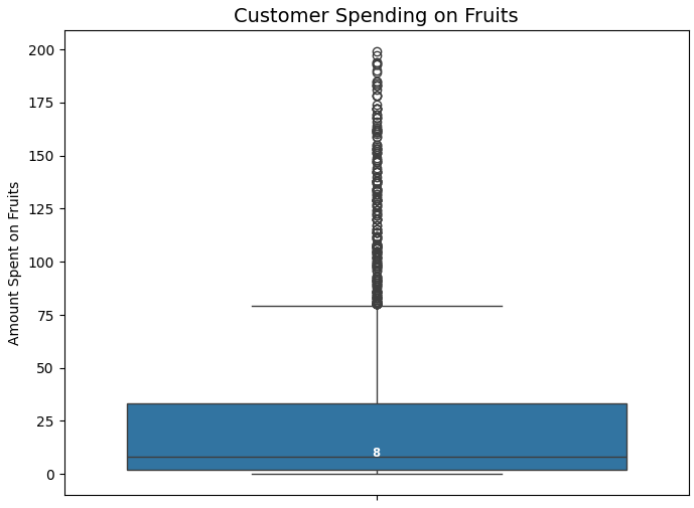

Figure 2.11 Box Plot of Customer Spending on Fruits

The box plot illustrates the distribution of customer spending on fruits. Most spending values are below 25, with a median of around 8 units, indicating that the majority of customers spend very little on this product category. However, the plot shows many outliers appearing above the main range, representing customers who spend significantly more on fruits than the average. This reflects a large disparity in spending behavior, where most customers purchase minimally while a small group spends heavily. These high-spending customers may represent a loyal or niche segment with a particular preference for fruit products.

Figure 2.12 Box Plot of Customer Spending on Gold Products

The box plot illustrates the distribution of customer spending on gold products. Most customers spend below 50 units, with a median of around 25 units, indicating that overall spending on this product category is relatively low. However, the plot shows numerous outliers above the main range, representing a small group of customers with very high spending levels, possibly exceeding 200 to 300 units. This highlights a significant disparity in spending behavior, where most customers make minimal purchases while a small segment spends substantially more, likely representing premium or loyal customers of gold products.

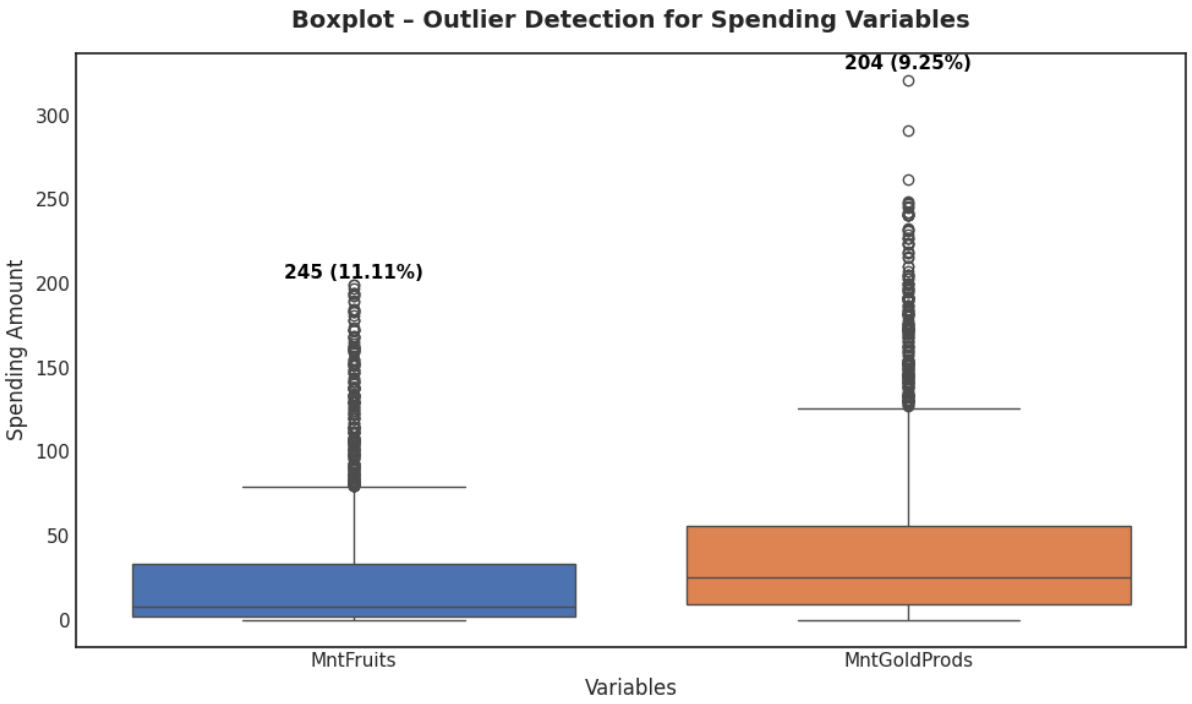

Figure 2.13 Boxplot - Outlier Detection for Spending Variables (MntFruits, MntGoldProds)

The box plot illustrates the spending distribution of customers across two product categories: MntFruits and MntGoldProds. Both variables show a notable presence of outliers, accounting for 11.11% and 9.25% of the observations, respectively, indicating significant variability in spending behavior. Most customers spend relatively little, while a small portion exhibit unusually high spending, likely representing premium or high-value customers. Identifying and managing these outliers is crucial for improving the accuracy of machine learning models in response prediction, ensuring that extreme behaviors do not distort overall insights.

2.2.4 Customer Behavior Insights

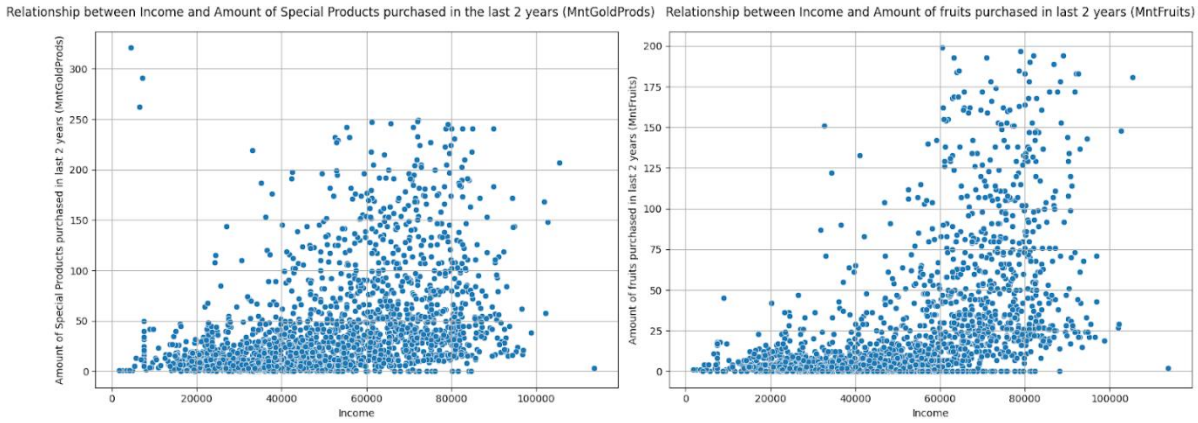

Figure 2.14 Relationship between Income and Amount Purchased in Last 2 Years (Special Products & Fruits)

The plots illustrate the relationship between income and product purchases over the past two years, covering special (MntGoldProds) and fruit (MntFruits) products. A slight positive trend appears; higher-income customers tend to spend more, especially on special products. However, wide data dispersion indicates that income alone doesn't determine purchasing behavior. The high-income but low-spending group emerges as a potential target for premium campaigns, while medium-income customers show stable fruit purchasing behavior driven by daily needs. Overall, variables like Income and MntGoldProds are valuable for customer segmentation and response prediction in iFood's marketing strategy.

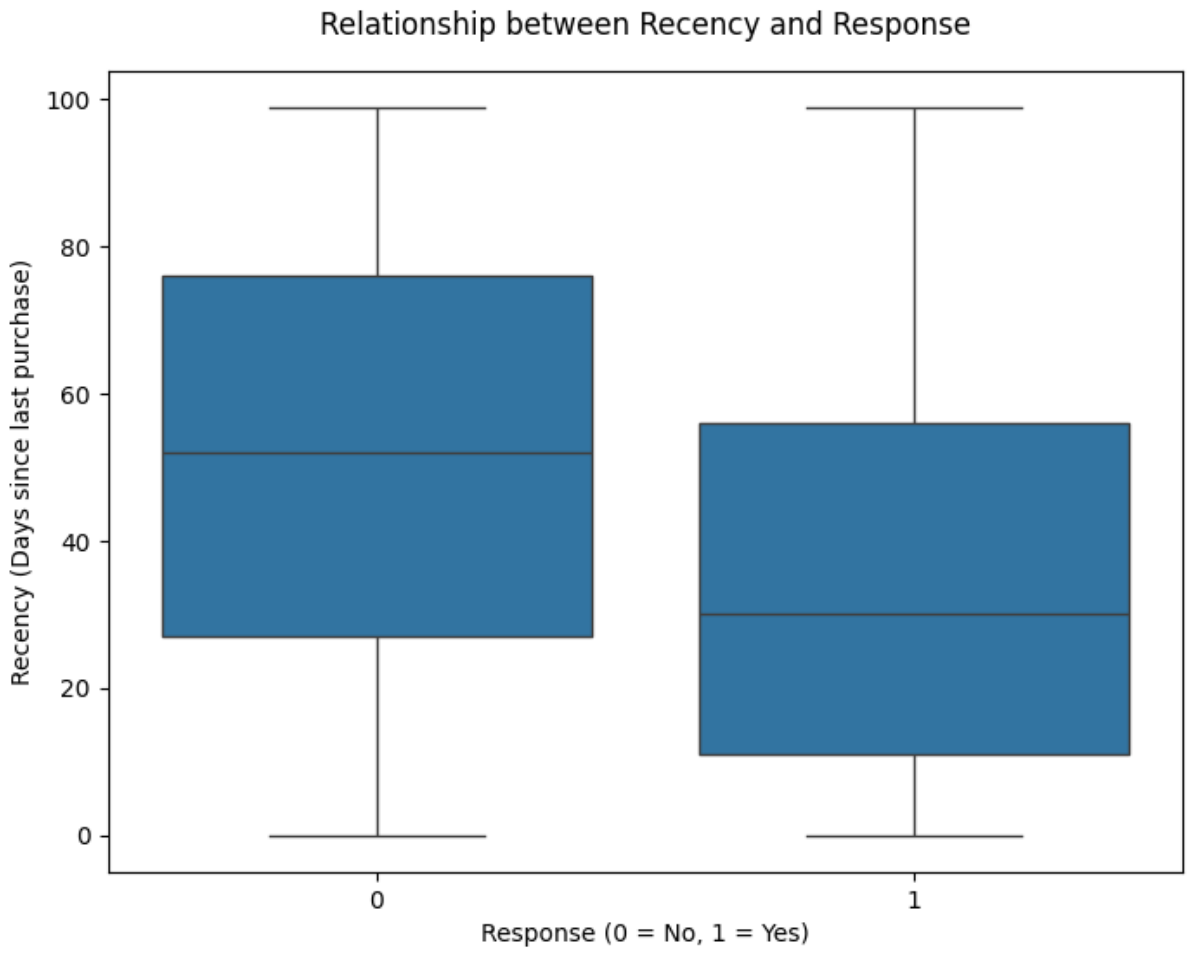

Figure 2.15 Relationship between Recency and Response

The boxplot illustrates the relationship between Recency (days since last purchase) and Response to marketing campaigns. Results indicate that customers who responded positively (Response = 1) tend to have lower Recency values, meaning they purchased more recently. In contrast, non-responders (Response = 0) show higher Recency, suggesting longer inactivity. This highlights that recent engagement strongly influences campaign responsiveness. Therefore, Recency serves as a key predictive variable in response prediction models and should be prioritized for identifying potential customers in remarketing strategies.

2.2.5 Correlation Heatmap Analysis

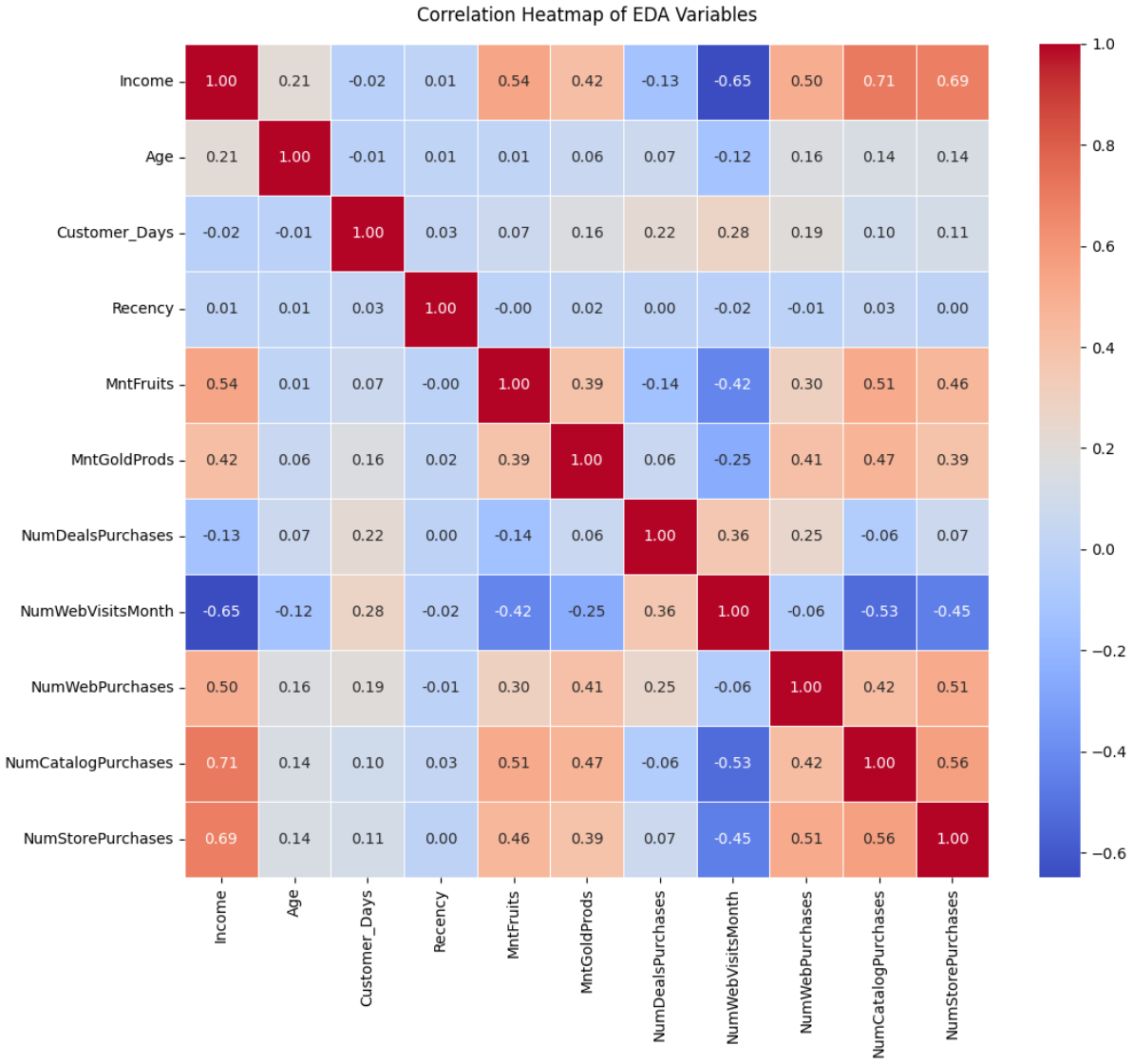

Figure 2.16 Correlation Heatmap of EDA Variables

The heatmap illustrates the correlation among the variables analyzed in the EDA process. Results show that Income is strongly positively correlated with store purchases (0.69), catalog purchases (0.71) and spending on both fruits (0.54) and special products (0.42), indicating that higher-income customers tend to spend more and engage across multiple purchasing channels. Conversely, Recency has a negative correlation with purchase frequency, meaning customers who purchased recently tend to be more active buyers.

Additionally, the high correlations among web, catalog and store purchases suggest a prevalent multi-channel shopping behavior. These insights help identify potential customer segments based on income and purchasing patterns, providing a strong foundation for segmentation and response prediction models in iFood's marketing campaigns.

2.2.6 Household Composition on Response Rate

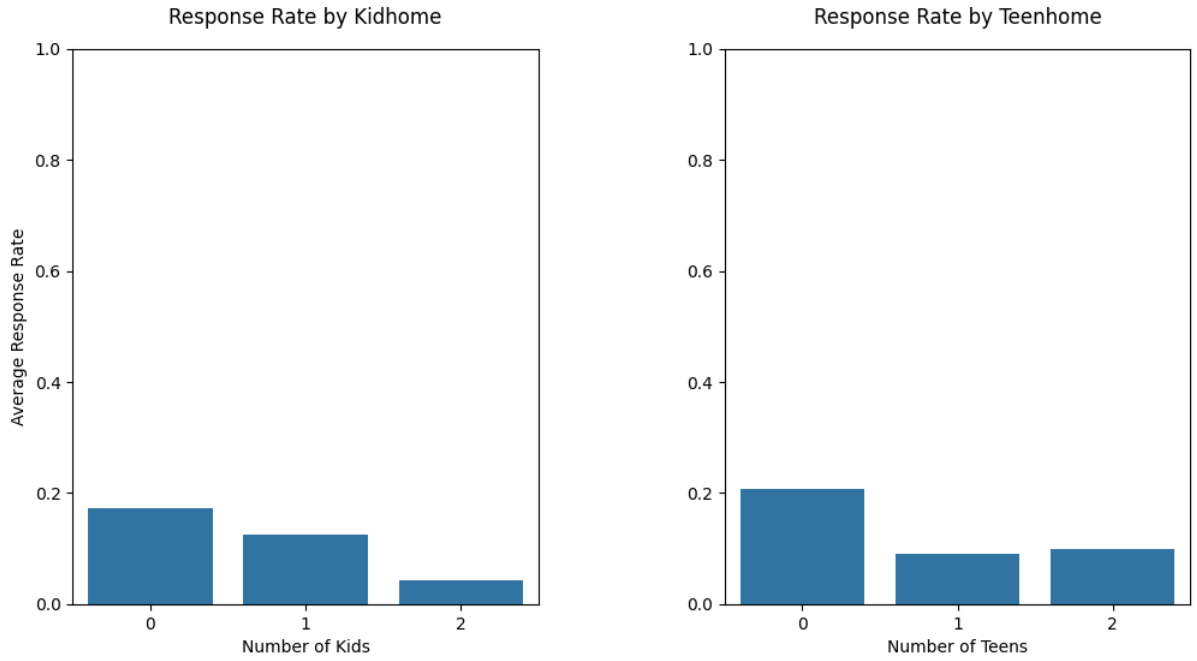

Figure 2.17 Response Rate by Kidhome and Teenhome

The charts illustrate the relationship between average response rate and the number of young children (Kidhome) and teenagers (Teenhome) in each household. The results indicate that customers without children or teenagers tend to respond more positively to iFood's marketing campaigns. As the number of children increases, the response rate declines, suggesting that households with more dependents are less likely to engage with promotions. This highlights family composition as a key predictor in response modeling, supporting machine learning models in identifying high-potential customer segments.

2.3 Feature Engineering

In addition to data cleaning steps, the research team also constructed new features from existing variables to better describe customers' demographic characteristics and consumption behaviors. Creating these new features enriched the dataset and improved its usefulness for subsequent analyses, including customer segmentation and campaign response prediction.

First, the variable Living_With was created based on marital_status. The categories Married and Together were grouped into Partner, while Single, Divorced, Widow, YOLO and Absurd were combined into Alone. This recoding simplified the data and highlighted behavioral differences between customers living with a partner and those living alone.

Next, the team created the variable Children by summing the number of children in the household (Kidhome + Teenhome). From this, the variable Is_Parent was derived to distinguish between customers with and without children. Additionally, the Family_Size variable was developed to represent household size, calculated as the number of adults (1 if Alone, 2 if Partner) plus the number of children. These variables enabled the model to capture differences in family structure and consumer behavior linked to demographic traits.

For the education variable, grouping was applied to reduce data fragmentation. Specifically, Basic and 2n Cycle were merged into Undergraduate, Graduation remained as Graduate and Master and PhD were combined into Postgraduate. This recoding preserved the educational hierarchy while ensuring sufficient observations within each group for meaningful comparisons.

Beyond demographic attributes, the team also developed indicators reflecting customers' preferred purchasing channels. Based on the number of purchases made through the website, catalog and store (NumWebPurchases, NumCatalogPurchases, NumStorePurchases), the ratios of purchases per channel were calculated - Ratio_NumWebPurchases, Ratio_NumCatalogPurchases and Ratio_NumStorePurchases.

Representing these as ratios allowed the identification of customers' preferred channels within their overall purchasing behavior, directly supporting the research question about the relationship between preferred purchasing channels and campaign response likelihood.

As a result, the Feature Engineering process introduced a set of important new variables, including Living_With, Children, Family_Size, Is_Parent, the recoded version of education and the purchasing channel ratio variables. These enhancements made the dataset more comprehensive in representing customers' family structures, educational backgrounds and consumption behaviors, thereby providing a solid foundation for both clustering and prediction tasks in subsequent chapters.

2.4 Data Pre-processing

To prepare the dataset for modeling, preprocessing steps included handling outliers, encoding categorical variables and standardizing numerical features. All transformations were fitted on the training set and applied to the test set to prevent data leakage, while the target variable Response was excluded from these operations.

2.4.1 Handling Outliers

During the exploratory data analysis, our team observed that some variables contained outliers. In particular, spending-related variables such as MntWines, MntMeatProducts and MntRegularProds had customers with exceptionally high spending, reaching up to nearly 2,500 units. Other spending variables like MntFishProducts, MntSweetProducts and MntFruits, although having lower average values, also included unusually high spending cases.

If left unchanged, these extreme values could distort the clustering process, as the K-Means algorithm relies on the Euclidean distance, which is highly sensitive to large values. However, these are also "high-value customers" that the business is particularly interested in, so they cannot simply be removed. Therefore, our team applied Winsorizing using the

IQR (Interquartile Range) method to the spending variables. This approach preserves the information about high-spending customers while reducing the excessive influence of outliers on clustering and prediction outcomes.

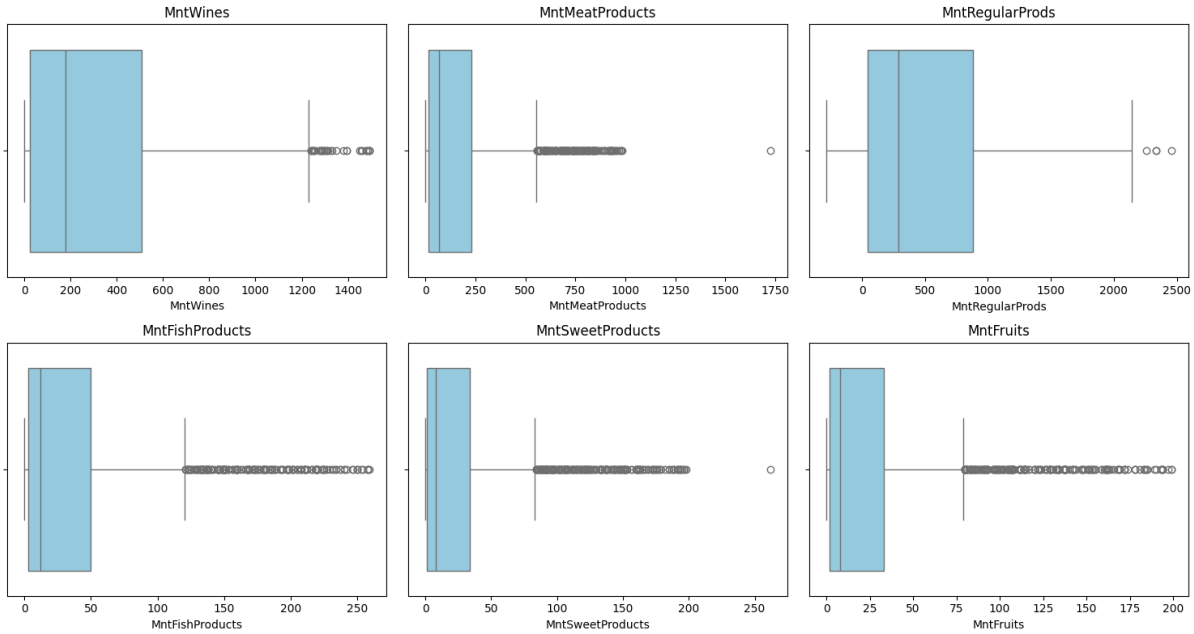

Figure 2.18 Box plots of spending variables before outlier handling

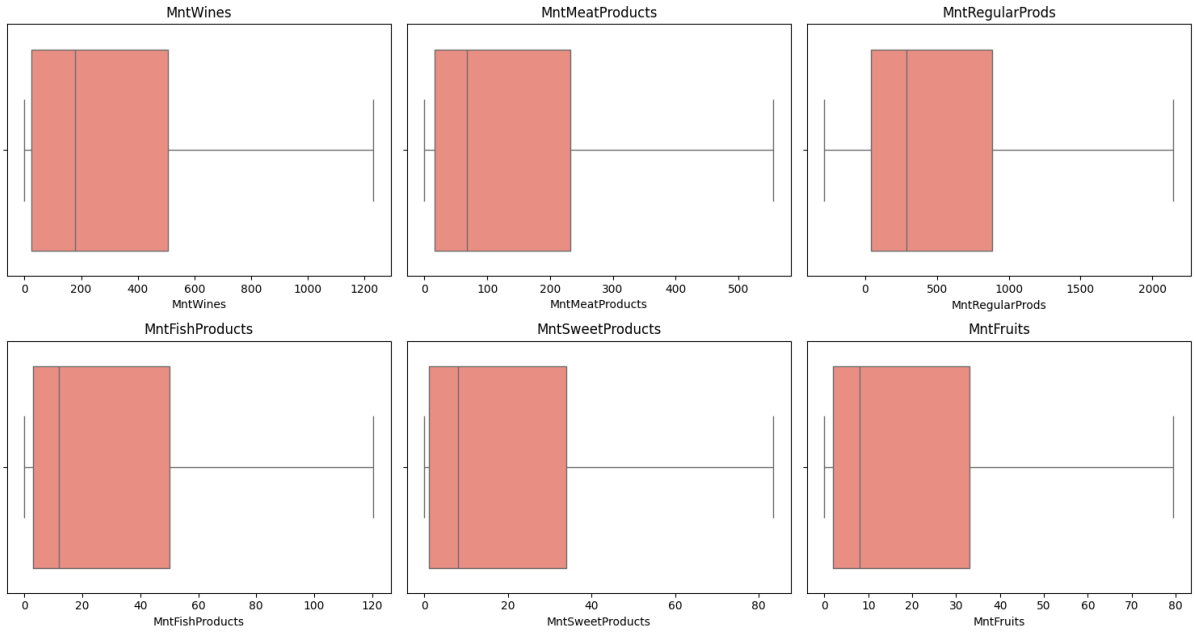

Figure 2.19 Box plots of spending variables after outlier handling

The results show that the whiskers became more compact and extreme values were brought closer to the boundaries without being removed. This helps balance the data distribution while still retaining the characteristics of high-spending customer groups.

In addition, for the Income variable, although there are some high values exceeding 100,000, our team decided to keep them unchanged, as these observations reflect the high-income customer segment, an important characteristic to preserve in the analysis. Other variables such as Age, Recency and the Num* group did not exhibit significant outliers, so they were retained as-is.

2.4.2 Encoding Categorical Variables

In this dataset, marital_status, education and the engineered variable Living_With are categorical strings (marital_status includes Married, Single, Together, Divorced, Widow; education includes Undergraduate, Graduate, Postgraduate). Because these labels have no natural order, mapping them directly to integers (label encoding) would impose spurious ordinality and can bias models, especially distance-based and linear methods.

Therefore, we apply One-Hot Encoding (OHE): each label is represented by an independent 0/1 indicator column. This preserves the nominal nature of the features and allows models such as Logistic Regression, SVM, k-NN, KMeans and tree-based methods (Decision Tree, Random Forest, XGBoost) to consume the information without introducing false order. In implementation, OneHotEncoder is configured with handle_unknown='ignore' so inference remains stable if unseen labels appear.

After encoding, columns of the form marital_status_*, education_* and Living_With_* take values in {0, 1}, correctly reflecting the state of each label.

2.4.3 Standardizing Quantitative Features

Quantitative features vary widely in scale (Income can reach tens or hundreds of thousands, while Recency ranges from 0-99; spending measures Mnt*, behavioral measures Num*, and tenure Customer_Days also differ substantially). Without standardization, large-magnitude features can dominate distance-based algorithms and slow optimization for scale-sensitive models.

We therefore use scikit-learn's StandardScaler to transform each numeric feature to a z-score:

$$z = \frac{x - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation estimated on the training set. This balanced feature contributions to objectives/distances and improves stability and convergence for Logistic Regression, SVM, k-NN, KMeans, and PCA. Note that AcceptedCmpOverall is a count (0–5), so it is treated as numeric and also standardized (rather than treated as binary), allowing the model to exploit interaction intensity.

After standardization, numeric features have means ≈ 0 and standard deviations ≈ 1, whereas OHE columns remain 0/1 (not standardized) - ensuring a consistent, interpretable feature matrix.

2.5 Feature Selection

In machine learning models, the number and quality of features directly affect analytical performance. If the dataset contains too many redundant or irrelevant variables, the model can become noisy, leading to unstable analytical results and limited interpretability from a business perspective. Conversely, a concise and information-rich set of variables that both reflects actual customer behavior and aligns with research objectives enables the model to perform efficiently and yield meaningful insights.

In this study, the feature selection process was divided into two distinct directions: one for customer clustering (unsupervised learning) and another for campaign response prediction (supervised learning). This separation was necessary due to the fundamental differences between unsupervised and supervised learning problems.

2.5.1 Feature Selection for Clustering

The goal of clustering is to group customers into distinct segments based on behavioral and demographic similarities, without using the target variable Response. Therefore, feature selection for clustering focuses on representativeness, eliminating information leakage and reducing redundancy caused by high correlations among variables.

The feature selection process for clustering consists of the following steps:

- Remove information-leakage and non-informative variables: This includes the target variable Response, previous campaign outcome variables (AcceptedCmp1..5, AcceptedCmpOverall), technical flags generated during data processing (e.g., outlier flags *_was_capped) and identifier variables such as Index.

- Remove zero-variance variables: Variables with no variation across customers do not contribute to differentiation during analysis.
- Filter by Spearman correlation: Pairs of variables with very high correlations ($|\rho| \geq 0.7$) were examined and only one representative variable was retained. A "whitelist" was used to ensure that key business variables such as Recency, Income, Age, Children and the purchase channel ratios (Ratio_NumWebPurchases, Ratio_NumCatalogPurchases, Ratio_NumStorePurchases) were preserved.

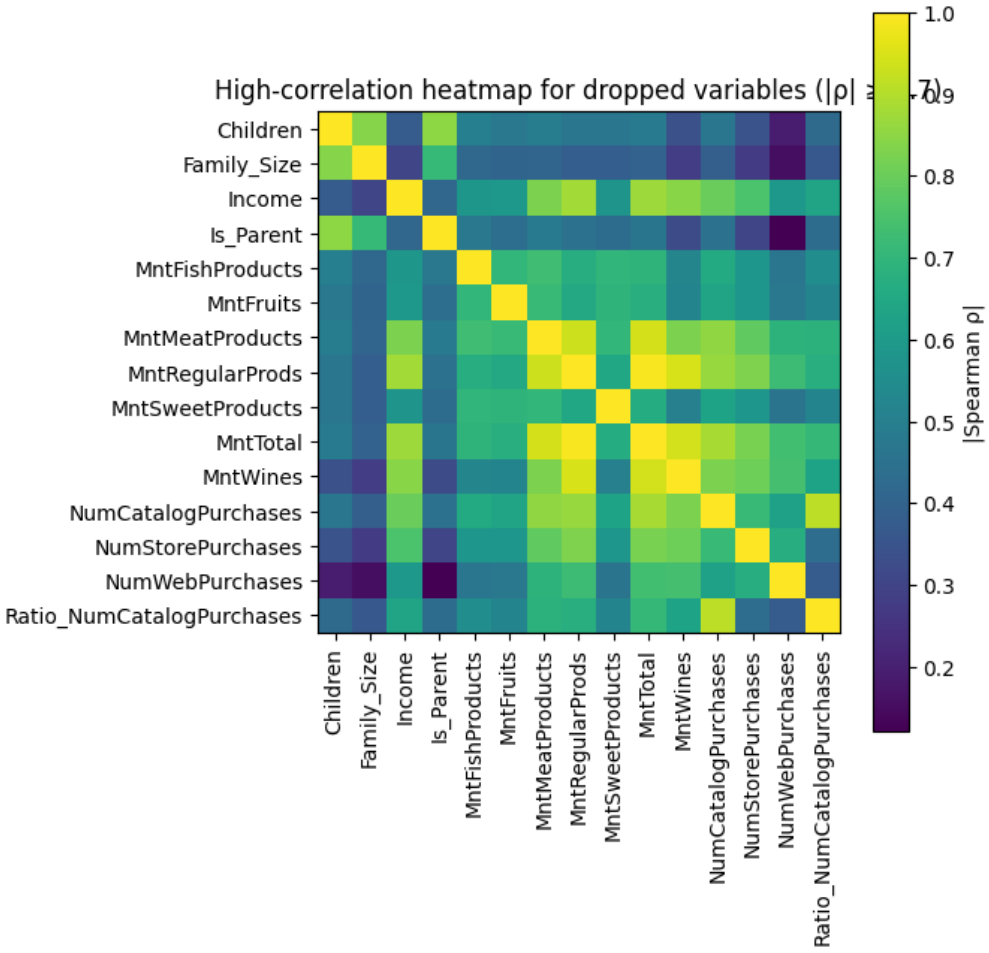

Figure 2.20 Spearman correlation heatmap of highly correlated features ($|\rho| \geq 0.7$)

After applying business logic filtering, variance screening and Spearman correlation filtering ($|\rho| \geq 0.7$), the dataset retained 14 features for clustering, while 11 variables were

removed due to high correlation or lack of relevance. The retained variables include demographic factors such as Income, Age, Kidhome, Teenhome, Children, Customer_Days; behavioral indicators such as Recency, NumDealsPurchases, NumWebVisitsMonth; spending variables like MntFruits, MntGoldProds; and channel preference ratios (Ratio_NumWebPurchases, Ratio_NumCatalogPurchases, Ratio_NumStorePurchases). This final feature set ensures a balanced representation of both demographic and behavioral perspectives, effectively supporting meaningful customer segmentation.

2.5.2 Feature Selection for Prediction

Unlike clustering, the Response prediction problem is a supervised learning task, allowing direct exploration of relationships between features and the target variable. Hence, feature selection here follows a supervised feature selection approach.

The process was implemented in three consecutive steps:

- Pre-filtering using Spearman correlation: At this stage, the Spearman correlation between each independent variable and the target variable Response was computed. Variables with an absolute correlation lower than $|\rho| < 0.02$ were removed, as they showed almost no statistical relationship with Response and were unlikely to add value to the model.

- Ranking using Mutual Information (MI): For the remaining variables, the Mutual Information metric was used to measure their dependency on the target variable. Unlike correlation, which captures only linear relationships, MI can detect both linear and nonlinear relationships, which are common in customer behavior data. This allowed the identification and retention of features with complex but significant effects on Response (online purchase behavior, campaign participation history).

- Selecting Top-K features: After ranking by MI, a subset of top-scoring features was selected. With around 35 processed variables, selecting the Top-15 was determined to be a reasonable cutoff, covering nearly half of all features. This number effectively removes low-information variables while retaining enough features for the model to capture key aspects of customer behavior.

Based on the Spearman filtering and MI ranking results, the final dataset retained 15 most important features for predicting Response. The prominent variables include AcceptedCmpOverall, Ratio_NumStorePurchases, MntTotal, MntMeatProducts, MntRegularProds, Income, AcceptedCmp5, MntGoldProds, MntWines, Family_Size, AcceptedCmp1, Ratio_NumCatalogPurchases, Is_Parent, NumCatalogPurchases and Recency.

2.6 Description of Variables

2.6.1 Feature selection for clustering

After applying business-rule filtering, variance filtering, and Spearman correlation filtering ($|\rho| \geq 0.7$) in the feature selection stage, the dataset retained 14 features for clustering. The detailed description of the variables is presented in the following table:

Table 2.1 Description of the Variable in Feature selection for clustering

No	Variable	Description	Data type	Total of values
1	Income	The yearly income for each customer.	int	2205 values: 1730 - 113734

2	Kidhome	Number of small children in customer's household.	int	3 values: 0 - 2
3	Teenhome	Number of teenagers in customer's household.	int	3 values: 0 - 2
4	Recency	Number of days since last purchase.	int	100 values: 0 - 99
5	MntFruits	Amount of fruits purchased in last 2 years.	int	158 unique values: 0 - 199
6	MntGoldProds	Amount of Special Products purchased in last 2 years.	int	212 unique values: 0 - 321
7	NumDeals Purchases	Number of purchases made with discount.	int	15 unique values: 0 - 15
8	Ratio_NumWebPurchases	Ratio of purchases made through company's website.	float	
9	Ratio_NumCatalog Purchases	Ratio of purchases made using catalog.	float	

10	Ratio_NumStorePurchases.	Ratio of purchases made directly in the store.	float	
11	NumWebVisitsMonth	Number of visits to company's website in the last month.	int	16 unique values: 0 - 20
12	Age	The customer's age.	int	24 - 80
13	Customer_Days	Days since registration.	int	2159 - 2858

2.6.2 Feature selection for Prediction

After applying Spearman correlation filtering ($|\rho| \geq 0.02$) and Mutual Information (MI) ranking, a total of 15 features were retained for the supervised prediction task. The detailed description of the selected variables is presented in the following table:

Table 2.2 Description of the Variable in Feature selection for Prediction

No	Variable	Description	Data type	Total of values	Variable type
1	Response	Response to the latest campaign	int (0/1)	0: 1872 1: 333	dependent
2	Income	The yearly income for each customer.	int	2205 values: 1730 - 113734	independent

3	Recency	Number of days since last purchase.	int	100 values: 0 - 99	independent
4	MntMeatProducts	Amount of meat purchased in last 2 years. In	int	551 unique values: 0 - 1725	independent
5	MntWines	Amount of wine purchased in last 2 years.	int	775 unique values: 0 - 1493	independent
6	MntRegularProds	Amount of Regular Products purchased in last 2 years.	int	974 unique values: -283 - 2458	independent
7	MntGoldProds	Amount of Special Products purchased in last 2 years.	int	212 unique values: 0 - 321	independent
8	MntTotal	Total amount of everything purchased in last 2 years.	int	897 unique values: 4 - 2491	independent
9	NumCatalog Purchases	Number of purchases made using catalog.	int	13 unique values: 0 - 28	independent
10	NumStore Purchases	Number of purchases made directly in the store.	int	14 unique values: 0 - 13	independent

11	AcceptedCmp1	1 if customer accepts the offer in first campaign, 0 for otherwise.	int (0/1)	0: 2063 1: 142	independent
12	AcceptedCmp5	1 if customer accepts the offer in fifth campaign, 0 for otherwise.	int (0/1)	0: 2044 1: 161	independent
13	AcceptedCmpOverall	Total number of marketing campaigns that customer accepted.	int	0: 1747 1: 322 2: 81 3: 44 4: 11	independent
14	Family_size	Total number of people in customer's household (1 + Kidhome + Teenhome)	int	5 unique values: 1, 2, 3, 4, 5	independent
15	Ratio_NumCatalogPurchases	Ratio of purchases through catalogs.	float		independent
16	Is_parent	1 if customer has at least one child or teen at home, 0 otherwise.	int (0/1)	0: 1248 1: 957	independent

Chapter 3. Experimental results and evaluation

This chapter details the experimental steps, from clustering customer data to building and evaluating separate predictive classification models for each segment. In the Clustering Models section, the objective is to partition the customer base into groups (clusters) using the K-Means Algorithm, based on similar characteristics regarding behavior, demographics and spending. The Predictive Classification Models section will focus on constructing distinct predictive classification models for each customer cluster identified in Section I, aiming to optimize prediction performance.

Link Code Chapter 3

3.1 Clustering Models

3.1.1 Clustering

The data clustering process was conducted using the K-Means algorithm to group customers with similar characteristics based on numerical data. This approach helps identify the natural structure within the dataset and supports a deeper understanding of potential customer segments.

The variables used in the model were carefully selected based on their importance and ability to reflect customer behavior, including spending patterns, promotional response levels, and demographic characteristics.

To determine the optimal number of clusters (k), two common methods, the Elbow Method and the Silhouette Score, were applied to evaluate the quality and the level of separation between clusters within the model.

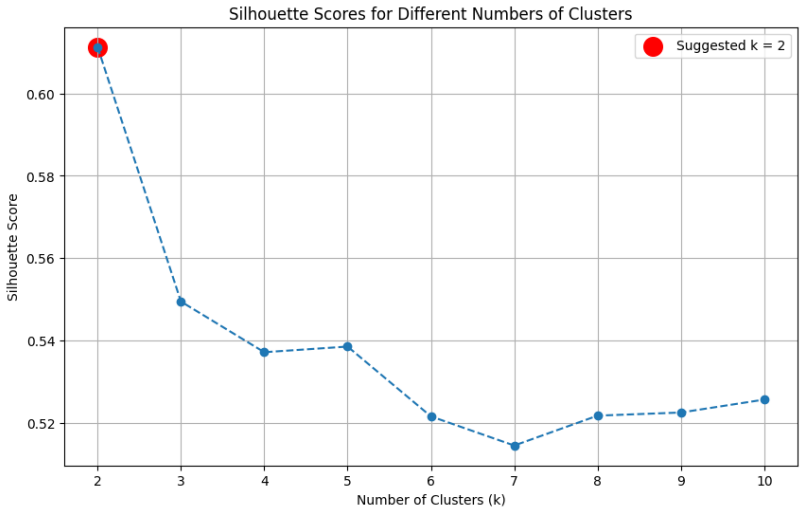

Figure 3.1 Silhouette scores for determining the optimal number of clusters (k)

Table 3.1 Determination of the Optimal Number of Clusters (k) Based on Silhouette Score

Number of Clusters (k)	Silhouette Score Value
2	0.6113
3	0.5495
4	0.5372
5	0.5386
6	0.5216
7	0.5145
8	0.5218
9	0.5225

10	0.5257

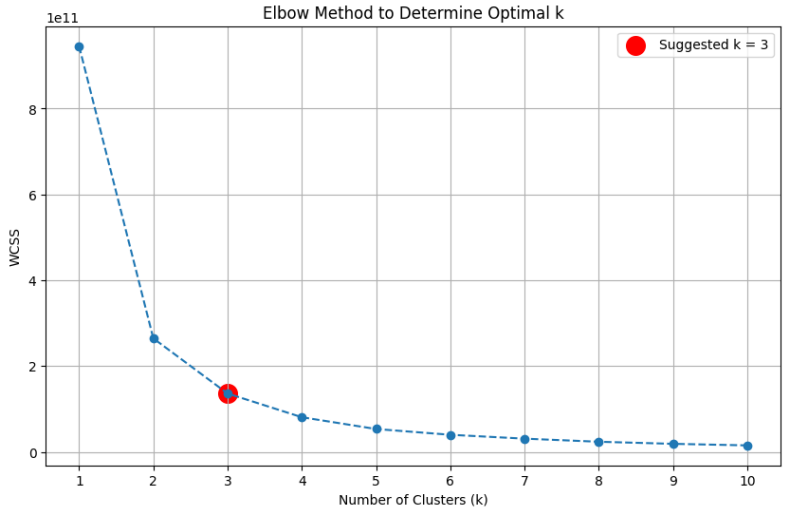

Figure 3.2 Elbow method for determining the optimal number of clusters (k)

Table 3.2 Determination of the Optimal Number of Clusters (k) Based on WCSS Value

Number of Clusters (k)	WCSS Value
1	945,684,458,243.81
2	264,574,470,199.80
3	136,132,428,399.52
4	80,864,605,520.87
5	53,379,187,178.94
6	39,943,223,307.90
7	30,915,739,114.00

8	23,691,422,166.00
9	18,850,871,314.44
10	15,204,163,229.62

The Silhouette score is highest when k = 2 (0.6113). It gets lower as k increases. This means splitting the data into two groups gives the best separation. Still, the difference between k = 3 and k = 5 is small, so the data structure stays pretty stable in that range.

The WCSS drops fast from k = 1 (945 billion) to k = 2 (264 billion), then goes down again to 136 billion at k = 3. After k = 4 (80 billion), the drop slows down a lot, and the change between values becomes small.

Looking at this pattern, the clear “elbow point” is at k = 3. That means this is likely the best number of clusters. At this point, the model keeps a good balance - less variation inside each group, but not too many clusters overall. It helps make the customer groups clear and useful.

Even though k = 2 gives a higher Silhouette score, choosing k = 3 gives better, more detailed results. It shows customer differences more clearly and helps understand how each group behaves. So, k = 3 is chosen as the best number of clusters for this segmentation.

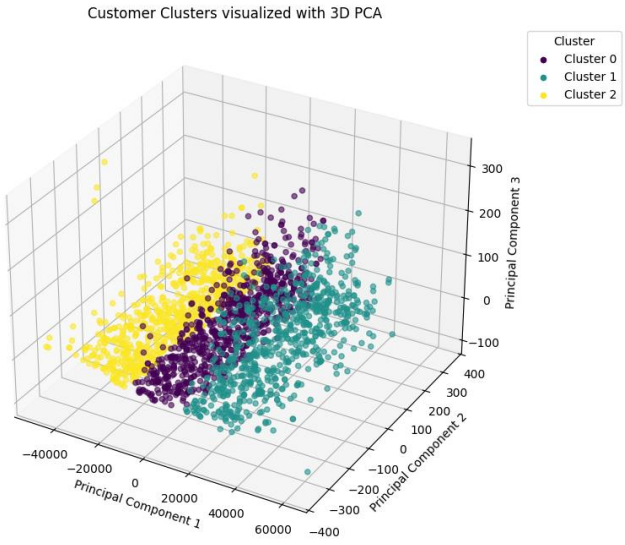

Figure 3.3 Visualization of Customer Clusters

The graph shows the customer segmentation results in 3D space, collapsed by PCA for better visibility. Each point is a customer, different colors correspond to the three clusters created by K-Means: Cluster 0, Cluster 1 and Cluster 2. From the image, it is clear that the clusters are quite clearly separated, the boundaries between the groups are easily recognizable in the PCA space.

Overall, K-Means with k = 3 effectively segments customers, creating three groups with clear differences in behavior and consumption. The PCA image reinforces the reasonable choice of k = 3, as it both clearly separates the clusters and helps to gain a deeper understanding of the customer data structure.

3.1.2 Description of Cluster Characteristics

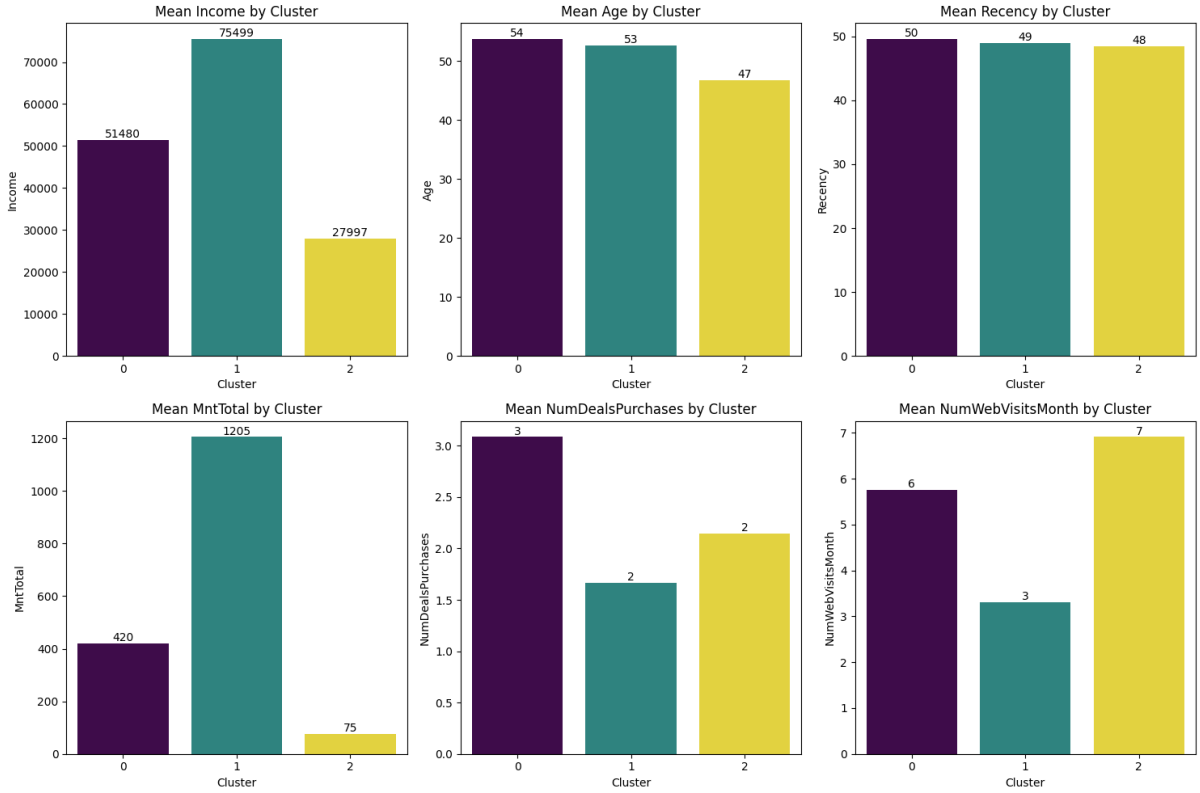

Figure 3.4 Customer Profile by Cluster

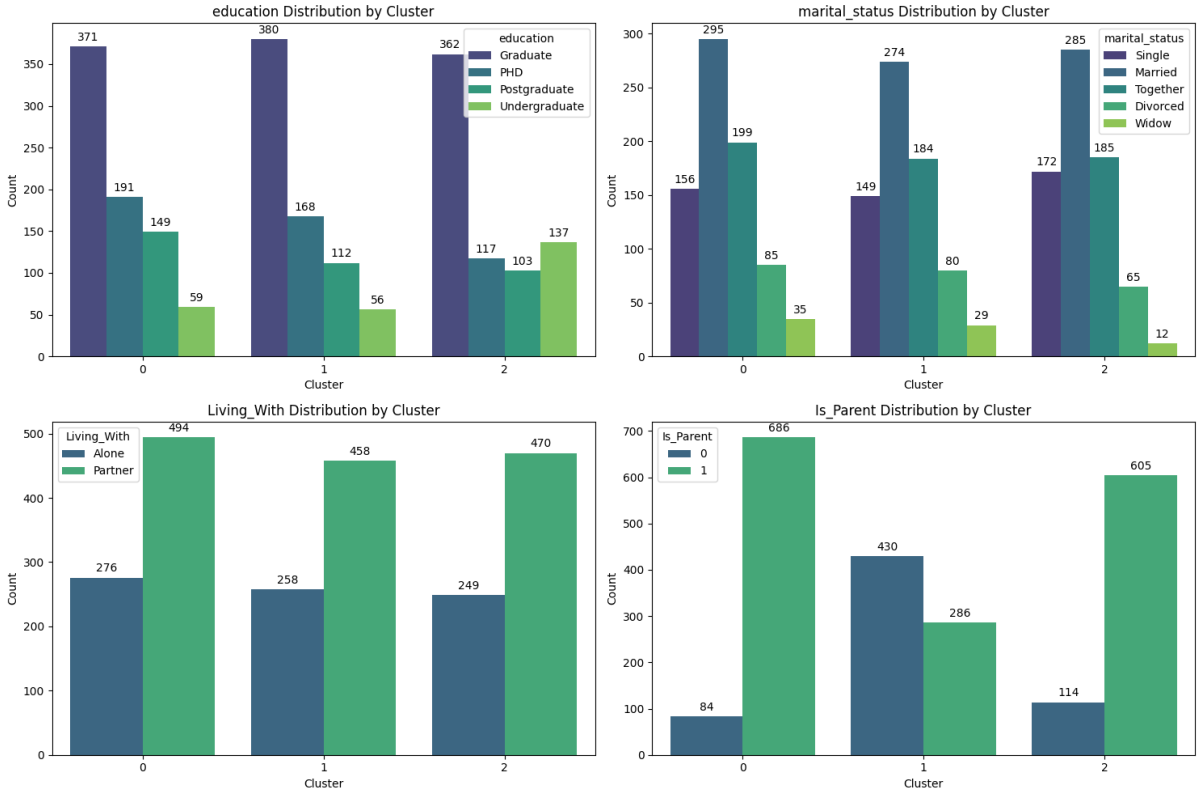

Figure 3.5 Demographic Distribution by Cluster

The clustering results show that the customer data is divided into three clearly distinct groups based on demographics and purchasing behavior. These groups differ significantly in terms of income, total spending and shopping behavior, while age and recency show only minor variations.

In terms of behavior, Income and MntTotal are the two variables that best reflect customer value:

- One group has high income and high spending - representing high-value customers who shop less frequently online but spend heavily.
- Another group has low income and low spending but frequent website visits - the deal-seekers.
- The third group has moderate income, stable spending and good responsiveness to promotions - the potential customers.

From a demographic perspective, most customers hold at least a university degree, with higher-income groups often having postgraduate education. The majority are married or living with a partner, suggesting greater stability and higher spending potential. The Is_Parent variable also differentiates the groups: those with children tend to spend more, while younger, single customers without families are more cautious in their spending.

Overall, the model highlights a clear stratification in both customer value and life stage: educated, family-oriented customers are the main spenders, whereas younger, low-income singles show limited spending despite frequent engagement.

3.1.2.1 Cluster 0 - Traditional Family Spenders

Cluster 0 consists of the oldest customer group, who mostly have families and children, with an average income level. They spend relatively little - especially on meat, fish, fruits and sweets - resulting in the lowest total expenditure. In terms of shopping behavior, this group has been less active recently, prefers in-store purchases over catalog or online channels, but visits the website fairly often. Regarding marketing, they show the lowest response and acceptance rates among all clusters. In summary, this is a group of older, family-oriented customers with average income, low spending and limited engagement with marketing activities, primarily shopping directly in stores.

3.1.2.2 Cluster 1 - Affluent Mature Buyers

Cluster 1 consists of relatively younger customers with the highest income among the three clusters, who often live alone. They have smaller family sizes with fewer children and spend heavily across most product categories, especially on wine, meat, fish, sweets and fruits, resulting in the highest total expenditure. This group prefers shopping via web and catalog channels rather than in-store, although they visit the website less frequently. They also have the highest response and acceptance rates to marketing campaigns and possess above-average education levels. Overall, Cluster 1 represents young, wealthy,

independent and high-value customers, making them a highly promising segment for marketing activities.

3.1.2.3 Cluster 2 - Budget-Conscious Young Families

Cluster 2 has the following key characteristics: the lowest average income and age; more young children, resulting in larger family sizes; and the most recent purchasing activity (lowest Recency). However, they have the lowest spending across most product categories, especially on wine, meat and regular products. Their total number of purchases is low, with a preference for in-store shopping; they visit the website frequently but make few online purchases. They also have lower education levels and the lowest response rate to marketing campaigns. In summary, Cluster 2 represents young, low-income customers with small children, low spending and recent activity who are loyal to in-store shopping channels, a segment with growth potential if targeted with affordable products and in-store promotions.

3.2 Predictive Classification Models

3.2.1 Cluster 0 - Traditional Family Spenders

After identifying the behavioral characteristics of Customer Cluster 0, the dataset was divided into 70% for training and 30% for testing. The preliminary analysis showed that the proportion of responding customers (class 1) was relatively low, around 15-18%, indicating a class imbalance issue. Therefore, the team applied the SMOTE (Synthetic Minority Oversampling Technique) method on the training set to increase the number of minority class samples. This helped the model learn better decision boundaries while maintaining objectivity during evaluation on the test set.

Next, three machine learning models - Logistic Regression, Random Forest and XGBoost - were implemented to predict customer response within the cluster. The models were evaluated based on ROC-AUC, F1-score, Precision, Recall and Accuracy. The results are summarized in the table below.

Table 3.3 Performance of response prediction models (Cluster 0)

Model	ROC-AUC	F1	Precision	Recall	Accuracy
Logistic Regression	**0.813**	0.400	0.268	**0.792**	0.753
Random Forest	0.796	**0.423**	**0.393**	0.458	**0.870**
XGBoost	0.785	0.305	0.257	0.375	0.823

From the results, Logistic Regression achieved ROC-AUC = 0.8126 and F1-score = 0.40, indicating a fairly good ability to distinguish between responding and non-responding customers. Although nonlinear models such as XGBoost and Random Forest recorded higher Accuracy (0.87 and 0.82 respectively), Logistic Regression showed a notably higher Recall (0.79), meaning it successfully identified the majority of customers likely to respond. This characteristic is valuable in marketing contexts, where the objective is to capture as many potential responders as possible.

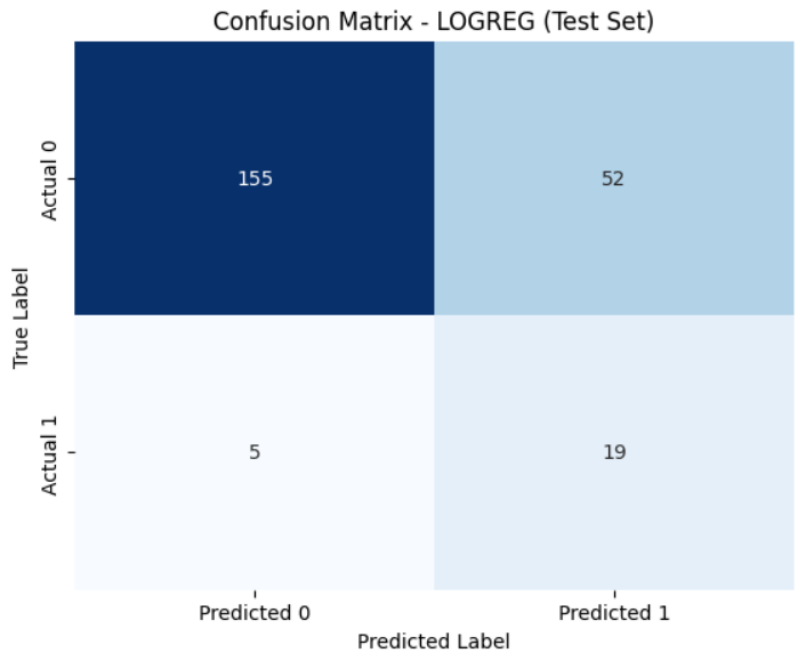

Figure 3.6 Confusion Matrix of the Logistic Regression model on the test set (Cluster 0)

Figure 3.6 illustrates the Confusion Matrix of the Logistic Regression model. The model correctly predicted 155 non-responding customers (True Negatives) and 19 responding customers (True Positives), while it misclassified 52 customers as responders (False Positives) and missed 5 actual responders (False Negatives). The calculated performance metrics are Precision = 0.27, Recall = 0.79, Accuracy = 0.75 and F1-score = 0.40. These results suggest that Logistic Regression tends to prioritize coverage (high recall) over strict precision (low precision). Although it produces a relatively high number of false positives, the model ensures that only a small number of potential responders are missed - a desirable property for proactive marketing strategies that aim to maximize customer outreach.

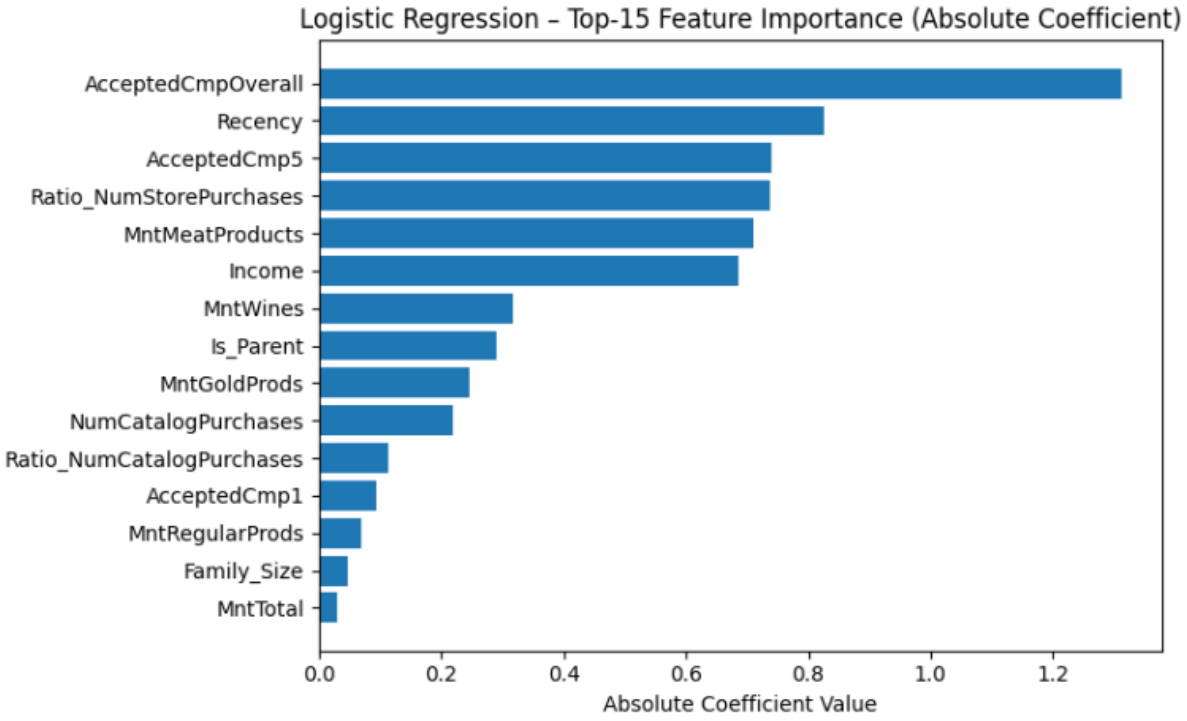

Figure 3.7 Top 15 Feature Importance in the Logistic Regression model (Cluster 0)

The Feature Importance chart (based on absolute coefficient values) indicates that variables related to past response history and recent customer interactions play the most significant roles in predicting behavior within Cluster 0. The variable AcceptedCmpOverall holds the highest importance, suggesting that customers who

previously responded to marketing campaigns are much more likely to respond again in the future. The variable Recency has a negative coefficient, indicating that customers with shorter time gaps since their last purchase tend to have higher response probabilities. Other influential features include AcceptedCmp5 and Ratio_NumStorePurchases, which reflect recent campaign participation and the frequency of in-store purchases, respectively - both contributing positively to response likelihood. Conversely, variables representing regular spending patterns such as MntWines, MntMeatProducts and Income show moderate importance, implying that consistent spending behavior has less influence on campaign responsiveness compared to direct marketing engagement.

These findings suggest that businesses should prioritize customers with prior response history and recent purchasing activity when planning re-engagement campaigns, as this approach can improve conversion rates while minimizing unnecessary marketing costs.

3.2.2 Cluster 1 - Affluent Mature Buyers

After identifying the behavioral characteristics of customer cluster 1, the dataset was split into 75% for training and 25% for testing. The initial analysis showed that the proportion of responding customers (class 1) accounted for only 23.6%, indicating a data imbalance problem. To address this, the team applied the SMOTE (Synthetic Minority Oversampling Technique) method on the training set to increase the number of minority class samples. This approach allowed the model to better learn the decision boundary while maintaining objectivity in performance evaluation using the untouched test set.

Next, three machine learning models were implemented to predict customers' campaign response: Logistic Regression, Random Forest and XGBoost. The models were evaluated using the following metrics: ROC-AUC, F1-score, Precision, Recall and Accuracy. The results are summarized in the table below.

Table 3.4 Performance of response prediction models (Cluster 1)

Model	ROC-AUC	F1	Precision	Recall	Accuracy
Logistic Regression	0.857	0.6	0.517	**0.714**	0.777
Random Forest	**0.866**	**0.617**	**0.641**	0.595	**0.827**
XGBoost	0.862	0.575	0.605	0.547	0.81

The comparison among the three models - Logistic Regression, Random Forest and XGBoost - shows that all models achieved good predictive performance, with ROC-AUC values exceeding 0.85. Among them, Random Forest achieved the best overall result with ROC-AUC = 0.866, slightly outperforming XGBoost (0.862) and Logistic Regression (0.857). Based on the primary metric ROC-AUC - which measures the model's ability to distinguish between responders and non-responders - Random Forest demonstrated the best generalization capability.

When considering secondary metrics, Random Forest continued to show advantages with F1 = 0.617, precision = 0.641 and recall = 0.595, reflecting a good balance between identifying actual responders and minimizing false positives. XGBoost ranked second with an F1-score of 0.575, while Logistic Regression achieved the highest recall (0.714) but lower precision (0.517). This suggests that the linear model tends to overpredict positive responses, resulting in more false alarms among non-responding customers. Overall, Random Forest emerged as the most stable and effective model for this cluster, achieving both a high ROC-AUC and solid overall accuracy.

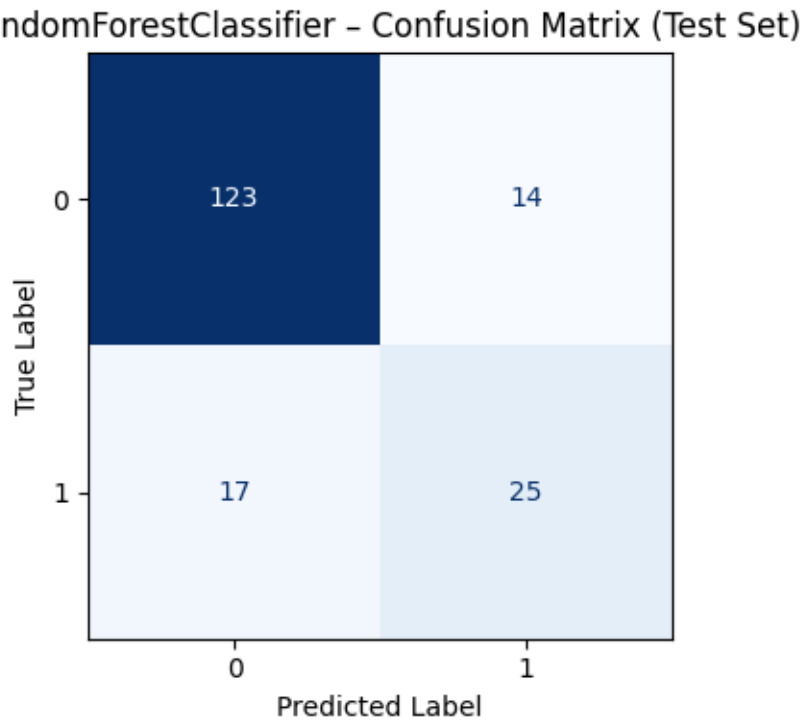

Figure 3.8 Confusion Matrix of the Random Forest model on the test set (Cluster 1)

The confusion matrix of the Random Forest model provides detailed insights into its classification performance on the test set. The model correctly predicted 25 responding customers and 123 non-responding customers, while misclassifying 17 actual responders as non-responders and 14 non-responders as responders. These results indicate that the model performs well in identifying non-responders - the majority group in the dataset - while still capturing a significant portion of true responders.

With an overall accuracy of approximately 82.7%, Random Forest shows strong stability and effective separation between the two customer groups. Although a small portion of potential responders were missed (actual responders incorrectly predicted as non-responders), this error margin is acceptable given the natural class imbalance in marketing data. In practice, the model can be effectively used to identify high-potential customers who are more likely to respond to future campaigns, supporting businesses in optimizing marketing resource allocation.

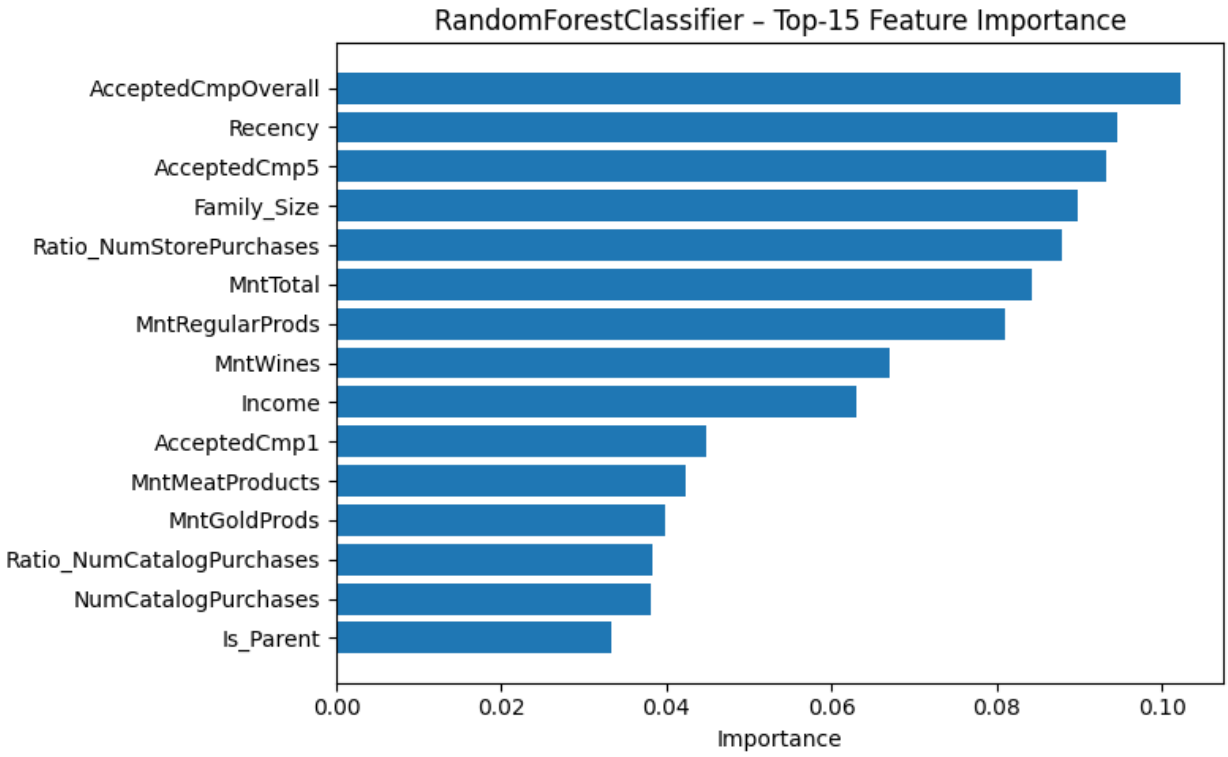

Figure 3.9 Top 15 Feature Importance in the Random Forest model (Cluster 1)

The feature importance chart of the Random Forest model highlights the variables that most influence customer response predictions. The variable "AcceptedCmpOverall", representing the total number of successful past marketing campaigns a customer has joined, stands out as the most significant - confirming that past response behavior is the strongest predictor of future response. The next most important variable, "Recency", reflects the time since the last interaction, indicating that recently active customers are more likely to respond. Other key features include "AcceptedCmp5", "Family_Size", "MntTotal", "MntRegularProds" and "MntWines", all of which represent overall spending levels and purchasing habits. Additionally, "Ratio_NumStorePurchases" contributes notably, suggesting that customers who frequently purchase in-store tend to have a higher likelihood of responding to campaigns.

These insights not only enhance model accuracy but also provide valuable business perspectives, reaffirming that past engagement and real purchasing behavior are core factors driving marketing campaign effectiveness.

3.2.3 Cluster 2 - Budget-Conscious Young Families

After identifying the behavioral characteristics of Cluster 2, the dataset was divided into 70% for training and 30% for testing. Initial analysis revealed that the proportion of responding customers (class 1) was less than 20%, indicating a class imbalance problem. To address this, the SMOTE (Synthetic Minority Oversampling Technique) method was applied to the training set to increase the number of minority-class samples. This helped the model learn better decision boundaries while maintaining the objectivity of performance evaluation on the test set.

Subsequently, three machine learning models were implemented to predict customer response propensity: Logistic Regression, Random Forest and XGBoost. The models were evaluated using standard metrics: ROC-AUC, F1-score, Precision, Recall and Accuracy. The results are summarized in Table 3.5.

Table 3.5 Performance of response prediction models (Cluster 2)

Model	ROC-AUC	F1	Precision	Recall	Accuracy
Logistic Regression	0.821	0.450	0.327	**0.720**	0.796
Random Forest	**0.843**	**0.560**	**0.560**	0.560	**0.898**
XGBoost	0.820	0.519	0.483	0.560	0.880

The results indicate that all three models achieved fairly high performance, with ROC-AUC values ranging from 0.82 to 0.84, demonstrating good discriminative capability between the two target classes. Among them, the Random Forest (RF) model achieved the

best overall performance (ROC-AUC = 0.843, Accuracy = 0.898, F1 = 0.56), reflecting a good balance between Precision and Recall. The Logistic Regression model obtained the highest Recall (0.72), which is advantageous when the goal is to maximize coverage of potential customers; however, its low Precision (0.33) led to more false positives (FP), resulting in unnecessary marketing costs. Meanwhile, XGBoost achieved Precision = 0.483 and Recall = 0.56, reducing false positives but still underperforming compared to RF. Therefore, Random Forest was selected as the optimal model for Cluster 2 due to its ability to capture non-linear relationships, robustness to noise and balanced performance between accuracy and coverage.

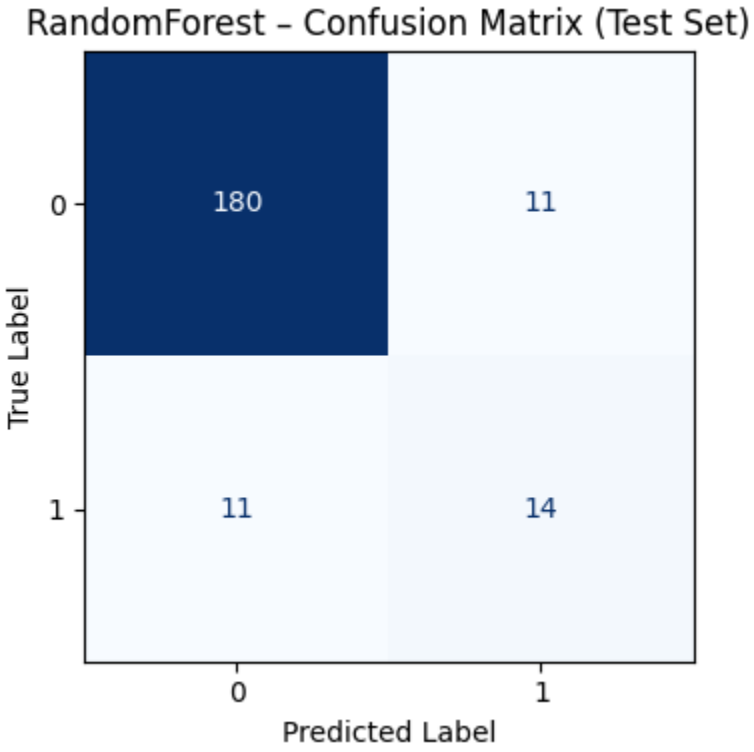

Figure 3.10 Confusion Matrix of the Random Forest model on the test set (Cluster 2)

Figure 3.10 presents the Confusion Matrix of the Random Forest model on the test set. The model correctly predicted 180 non-responders (True Negative) and 14 responders (True Positive), while 11 cases were misclassified as responders (False Positive) and 11 true responders were missed (False Negative). Based on these results, the model achieved

Precision ≈ 0.56 and Recall ≈ 0.56, indicating a balanced performance in identifying potential customers while minimizing wasted marketing costs. This demonstrates that the Random Forest model is well-suited for real-world marketing operations, where firms must balance contact costs (outreach and promotions) with conversion potential.

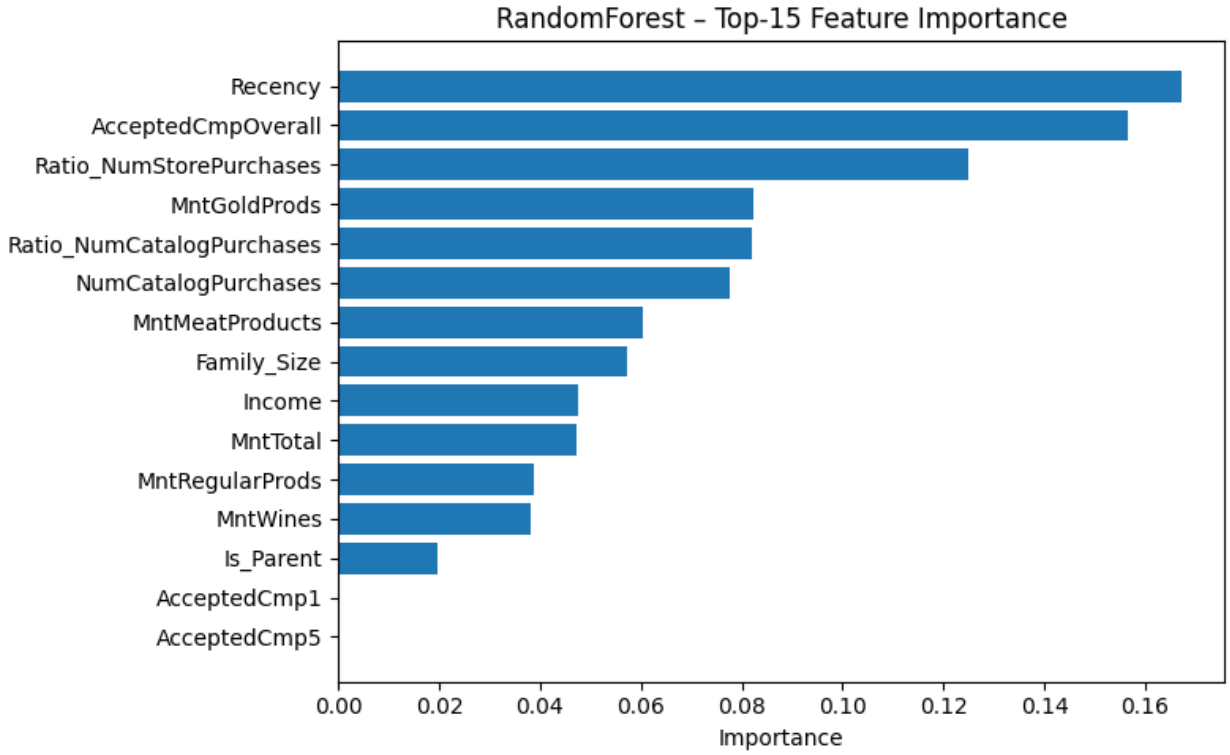

Figure 3.11 Top 15 Feature Importance in the Random Forest model (Cluster 2)

The Feature Importance analysis of the Random Forest model reveals that behavioral variables such as Recency and AcceptedCmpOverall have the greatest influence on customer response likelihood. In addition, factors related to purchase frequency and sales channel (in-store or catalog purchases), together with total spending and basic demographic attributes, also contribute significantly to the model's predictive performance.

Chapter 4 plays a pivotal role in transforming the complex experimental results from Chapter 3 into visual, easily understandable and actionable information for the business. This chapter focuses on designing three main Dashboards: the Customer Overview Dashboard, the Segment Analysis Dashboards and the Response Analysis Dashboard Crucially, the discussion section will delve into the interpretation of insights derived from the models and charts, emphasizing the management and business implications. The goal is to fully leverage the analytical results to build marketing strategies/campaigns, enhance customer retention, or optimize resources (such as more precise targeting based on Feature Importance per cluster), thereby supporting managerial decision-making based on data.

4.1 Customer Overview Dashboard

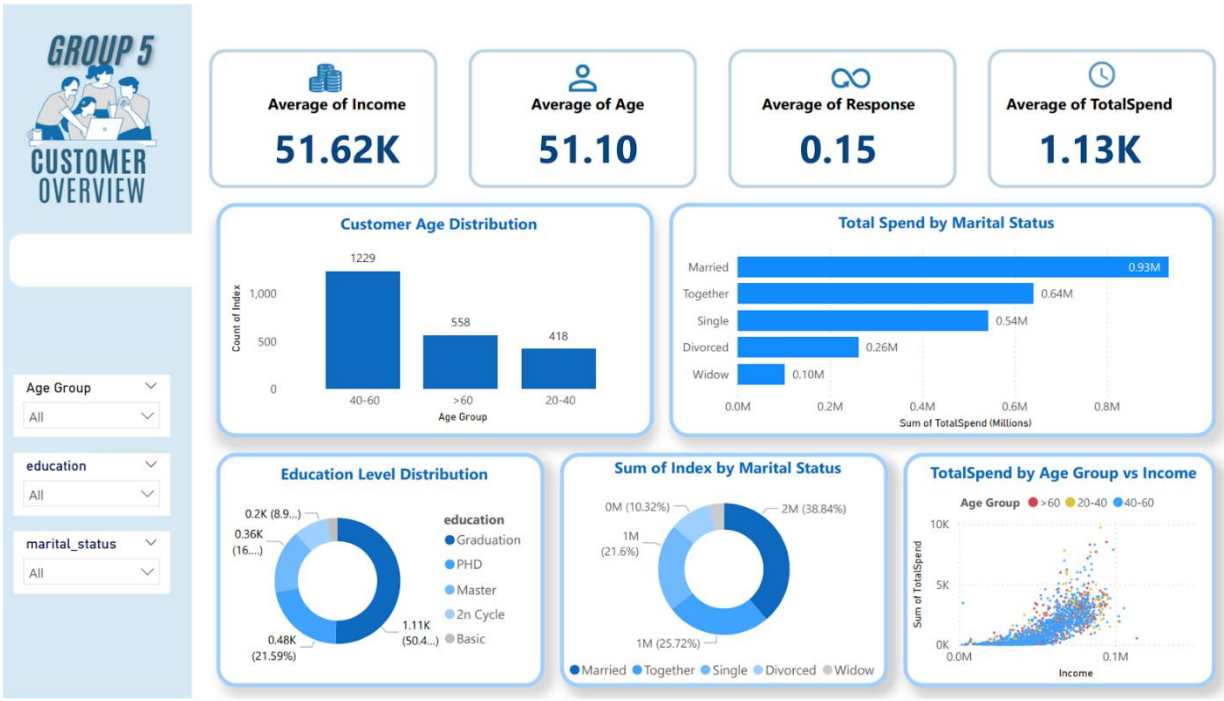

Figure 4.1 Customer Overview Dashboard

The "Customer Overview" dashboard provides a comprehensive view of customer characteristics through both demographic and behavioral indicators. Its layout is clean and

logically structured, balancing key performance indicators (KPIs) with visual charts. The color palette of blue and white enhances clarity and professionalism, allowing users to quickly grasp the overall picture of the customer base and make data-driven decisions for marketing and retention strategies.

The four main KPIs displayed at the top - Average Income (51.62K), Average Age (51.10), Average Response (15%) and Average TotalSpend (1.13K) - reveal that most customers fall within the 40-60 age group, with medium-to-high income and consistent spending levels. This indicates a strong potential for marketing strategies targeting mature, financially stable consumers with sustainable purchasing behavior.

An analysis by marital status shows that customers who are Married or Living Together have the highest total spending values, 0.93M and 0.64M, respectively. This implies that family-based customers tend to have higher purchasing demand, especially for household-related products and services. Therefore, the company should focus on developing long-term loyalty programs, family-oriented promotions and bundled offers tailored to this segment.

In terms of education level, the Graduation group accounts for 50.48%, followed by the PhD group at 21.59%. This suggests that most customers are well-educated, financially capable and likely to value premium products and high-quality service. Marketing messages should therefore emphasize brand reliability, product excellence and personalized experiences to resonate with this customer group and increase loyalty and response rates.

The scatter plot TotalSpend by Age Group vs. Income highlights a clear positive correlation between income and spending - customers with higher incomes tend to spend more, particularly those in the 40-60 age range, who represent the most profitable segment. Based on this insight, the business can design premium product packages or special retention programs for middle-aged customers with stable income levels.

In conclusion, to enhance the effectiveness of marketing campaigns, businesses should focus on high-income, middle-aged customer groups, such as married and together segments, through family-oriented offers and loyalty programs. Campaigns should be personalized based on the behavioral characteristics of each customer group to optimize response rates and improve marketing return on investment.

4.2 Segment Analysis

4.2.1 Overview

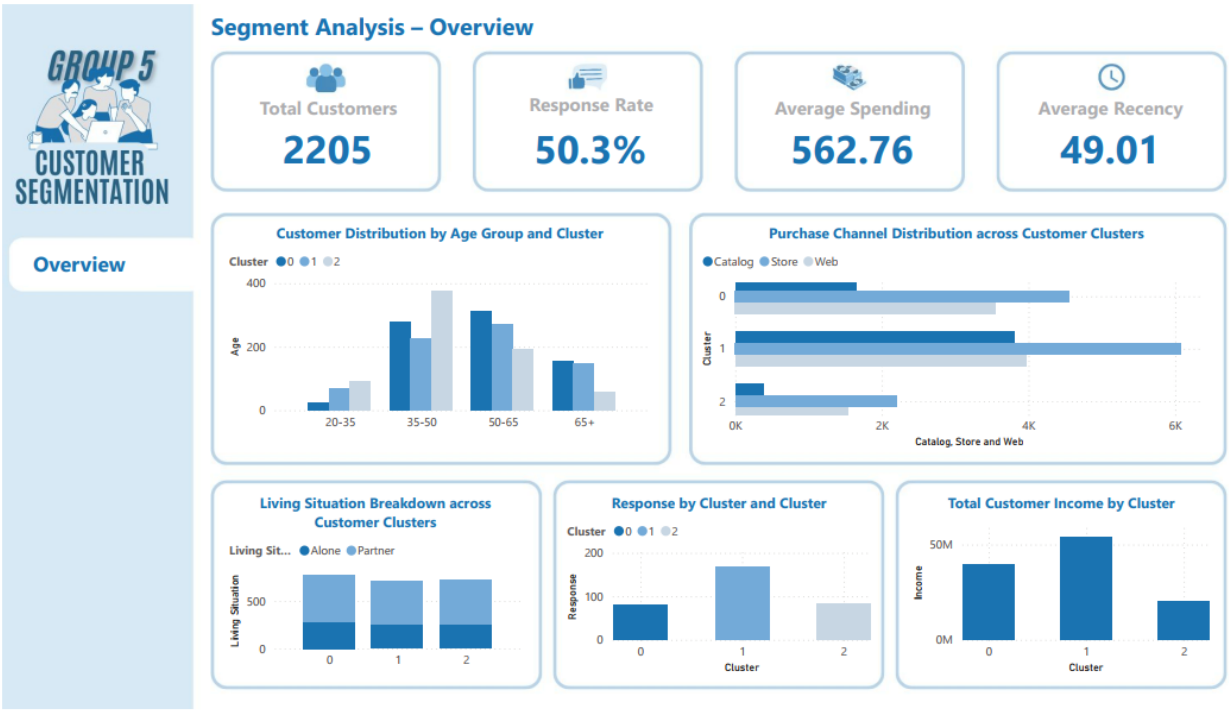

Figure 4.2 Customer Segmentation Overview Dashboard (1)

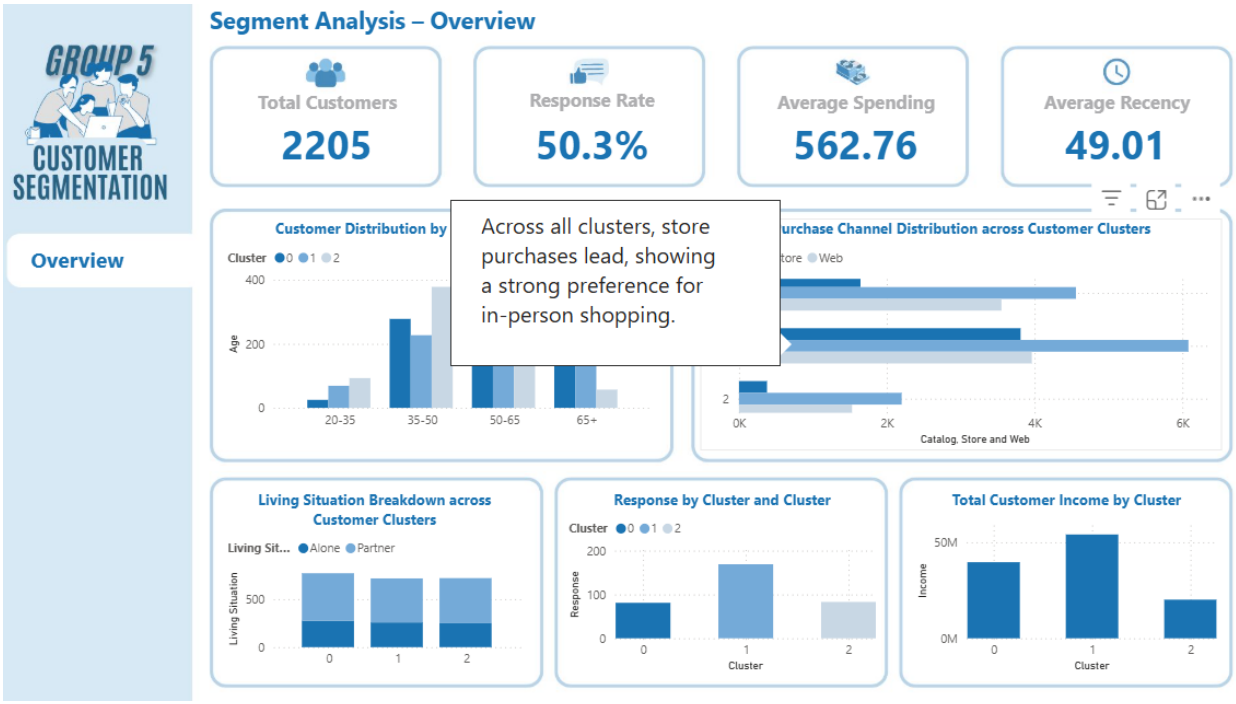

Figure 4.3 Customer Segmentation Overview Dashboard (2)

iFood's customer segmentation dashboard aggregates the results from demographic and behavioral data. From 2,205 customers, K-Means has been divided into three distinct groups with an average response rate of 50.3%, an average spend of 562.76, and an average recency of 49.01 days. These three groups differ in income, marital status, preferred shopping channels, and interaction styles, helping iFood better understand its customers. The three groups are: Group 0 - Traditional families with moderate spending, Group 1 - Middle-aged wealthy professionals and Group 2 - Young families who care about price.

Group 0 is made up of middle-aged people aged 50-65, with a stable income, mostly living with their spouses. Their response rate is the lowest of the three groups, indicating a medium level of engagement. They buy most in-store, next online, and least through catalogs. This shows that they still prefer to buy in person but are getting used to online. Overall, this is a stable, moderately loyal group with long-term potential if approached correctly.

Group 1 is also aged 50-65, but has a high income and is the most responsive. Most live with a partner, showing financial stability and strong brand attachment. They buy mainly in-store, then online and catalog, combining both traditional and modern habits. This is the most valuable group with high spending power, loyalty and consistent engagement.

Group 2 consists of customers aged 35-50, with lower income and moderate response. They often live alone or in small families, buy mainly in-store, then online, least catalog. Low spending and price sensitivity indicate low loyalty. This group is young, dynamic with flexible habits, suitable for short-term expansion strategies focused on engagement.

4.2.2 Cluster 0 - Traditional Family Spenders

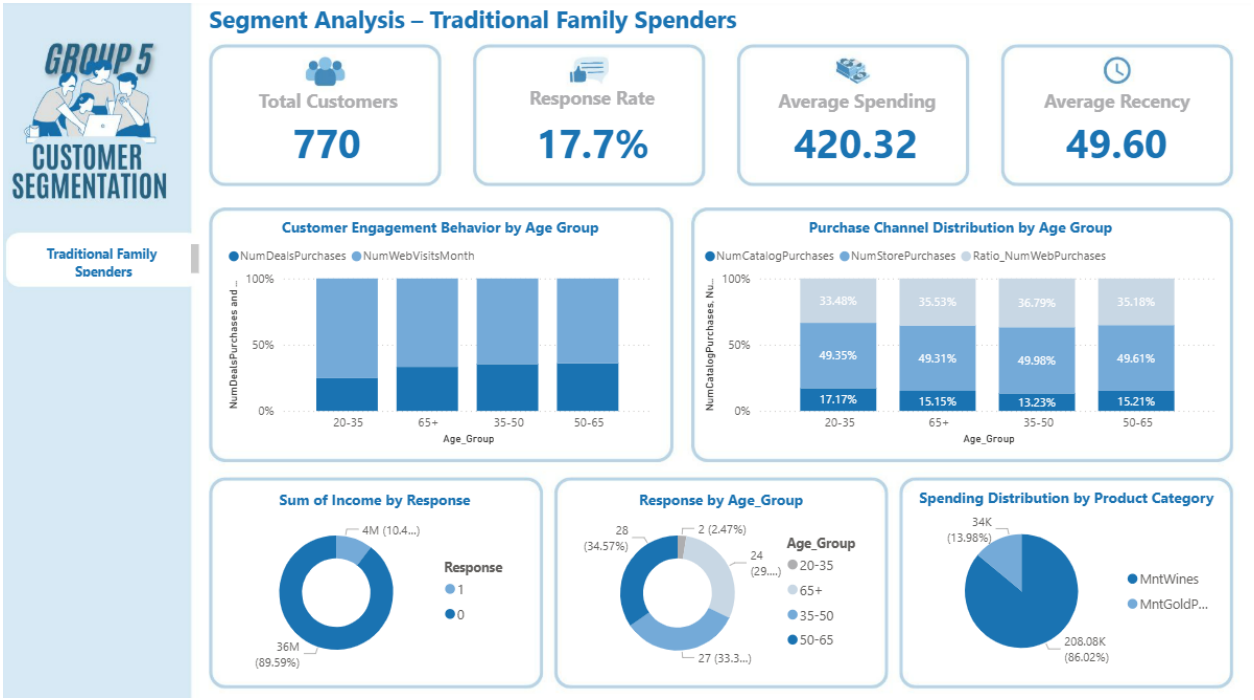

Figure 4.4 Customer Segment Overview of Cluster 0 - Traditional Family Spenders

Group 0 - Stable spending families have about 770 customers. They shop regularly, not a lot but quite regularly, and stick with the brand for a long time. Although the response rate is only about 17.7%, they still maintain an average spending level of about 420, quite

high compared to the general level. In other words, this group does not react strongly to marketing campaigns, but still buys regularly, so it is still a stable source of revenue.

Most of them are cautious in spending. They like familiarity, do not often try new things. These customers often buy because they trust and are familiar with the brand, not because of advertising or promotions. For them, quality and reliability are more important than price or trends. Many people in this group only buy when they feel confident based on previous experience.

Data analysis shows that they spend the most on wine, far exceeding other product groups. This shows that they prefer things that bring a sense of relaxation or enjoyment, rather than buying for actual needs. This is an opportunity for businesses to build emotional campaigns, focusing on experiences and spiritual values, instead of talking too much about price or product features.

In terms of behavior, they still maintain interactions on many channels, especially the 50-65 age group. They participate in promotions, sometimes buy online, but not too often. Although the response is low, they are loyal out of habit - as long as the products and services do not change, they will continue to buy.

With this group, the best way is to maintain a stable relationship, without having to push short-term sales. Businesses should take care of after-sales, have gratitude programs or small personalized incentives to maintain trust. In addition, they can encourage them to try some new forms of shopping such as ordering online and picking up at the store, helping them gradually get used to digital channels.

In short, group 0 is a stable, trustworthy group of customers that does not create rapid growth but helps businesses maintain long-term revenue and maintain a strong brand image.

4.2.3 Cluster 1 - Affluent Mature Buyers

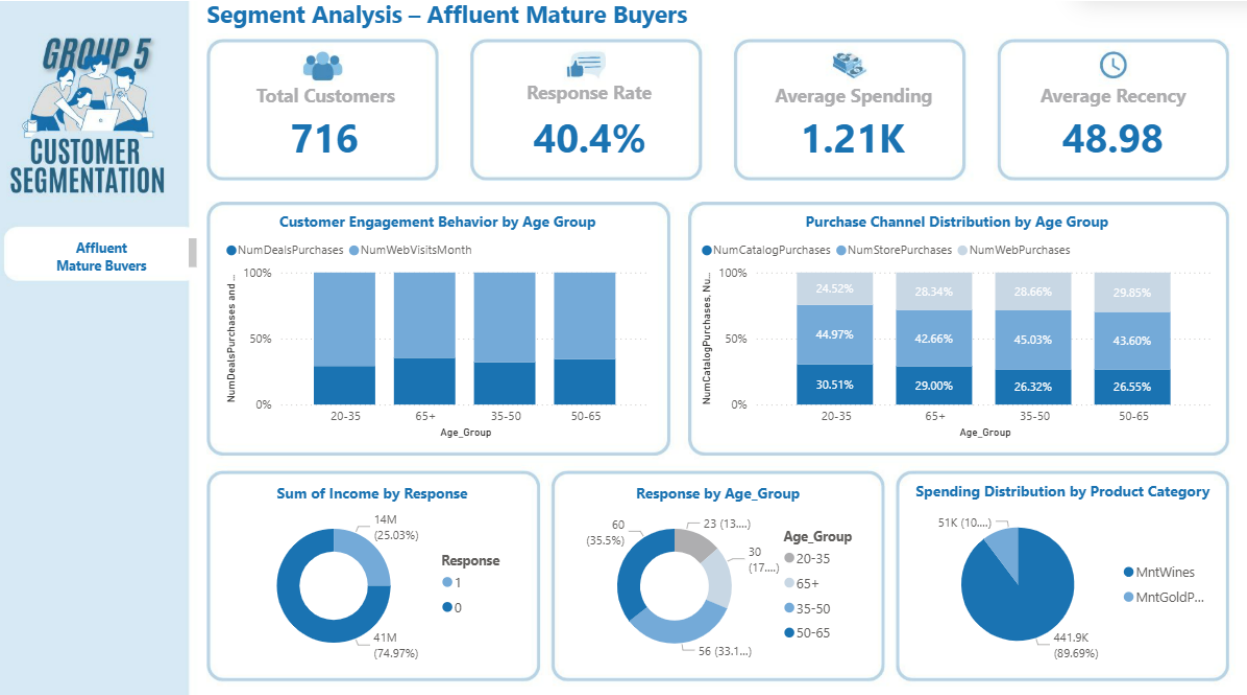

Figure 4.5 Customer Segment Overview of Cluster 1 - Affluent Mature Buyers

Customer Cluster 1 - Affluent Mature Buyers includes 716 customers. They have the highest response rate, about 40.4%. Their average spending is around 1,210, which is almost three times higher than Group 0. The average recent purchase gap is about 49 days, meaning they still buy quite often.

These customers spend a lot, but they still prefer traditional ways of shopping. Most of them buy in stores, some use catalog orders, and few shop online. Even though everything is digital now, they still like to see and touch the product first before buying. They care more about the real experience than convenience.

Most of them are between 35 and 65 years old. They usually respond well to marketing, showing that they still pay attention to the brand. But since they're busy with work and family, not everyone reacts often. They are traditional but open-minded. They don't mind

spending more money if the product feels worth it is something high quality, elegant, or satisfying.

For this group, companies should focus on building long-term relationships instead of quick campaigns. The best approach is to keep good quality, offer nice service, and give real experiences; for example, in-store consulting or loyalty programs. It's also smart to stay active in traditional channels like stores and catalogs, while adding personal touches to make them feel noticed.

In short, Group 1 is wealthy, stable, and reliable. They respond better when approached the right way. In the long run, they bring steady revenue and trust, helping the brand stay strong and respected.

4.2.4 Cluster 2 - Budget-Conscious Young Families

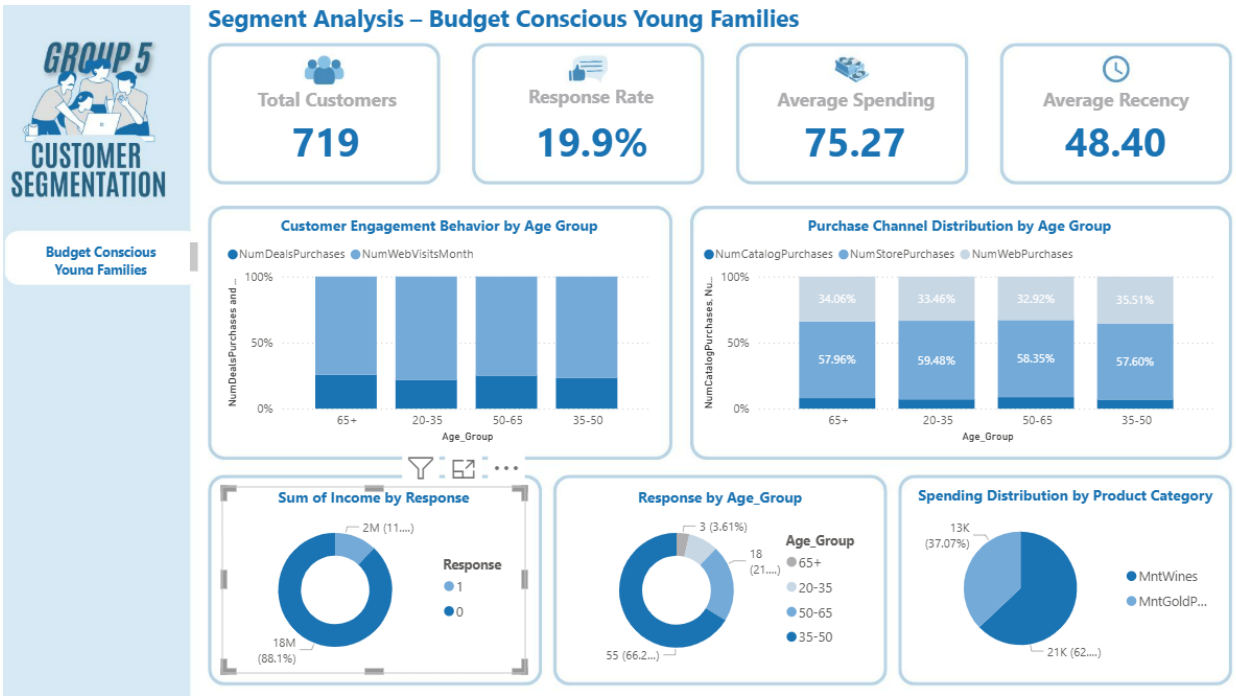

Figure 4.6 Customer Segment Overview of Cluster 2 - Budget-Conscious Young Families

Customer Cluster 2 - Budget-Conscious Young Families has about 719 customers, accounting for the same proportion as the other two groups. However, this group has a response rate of 19.9%, the lowest average spending level of about 75.27, and an average time of last purchase of 48 days. These numbers show low engagement and limited spending ability, so this group has less potential than the other groups.

Most of them still shop in stores. After that comes online, and catalogs are used the least. They visit the website a lot, more than other groups, just to look around or check prices. But they don't buy online that much, maybe because they don't fully trust online payment yet or just prefer seeing things in person. When we look at what they buy, most of the money goes to wine. It's the same trend as the other groups. Even though they try to save money, they still spend on small treats that make them feel relaxed or happy. So, they are careful but still want some enjoyment.

People aged 35-50 react the most to marketing. They click on promos, read emails, and visit the website. The younger and older ones join less. It looks like the 35-50 group cares about the brand when the message fits their needs or makes sense to them.

For business, this group fits discounts or reward programs. They often check online first but buy in the store, so companies should improve digital ads, make honest online info, and connect online to in-store offers to help them decide faster.

In short, this is a younger middle-age group that's practical and open to change. They don't spend much, but if reached the right way with clear deals and easy shopping. They can still bring short-term sales growth.

Conclusion: For iFood's marketing plan, Cluster 2 - Young, frugal families was chosen as the main target group. This group is younger and more tech-savvy, which fits iFood's business model. They are online, look for deals, and care about price and speed. Even if they don't spend much per order, they still shop more frequently, which helps iFood grow rapidly in the short term. This segment is also more responsive to digital marketing efforts,

such as app-based promotions, social media ads, and personalized online offers, making them highly suitable for iFood's customer acquisition and engagement campaigns. By focusing on this group, iFood can expand its reach, increase order frequency, and attract new users through promotions and app campaigns. Meanwhile, Cluster 1 customers already have strong spending habits and brand trust, but they mainly shop offline. iFood can still retain them through loyalty programs, but they are not the main focus of growth. On the other hand, Cluster 2 represents the most scalable and conversion-ready segment for iFood's digital-first strategy, offering both short-term sales uplift and long-term digital engagement potential.

4.3 Response Analysis

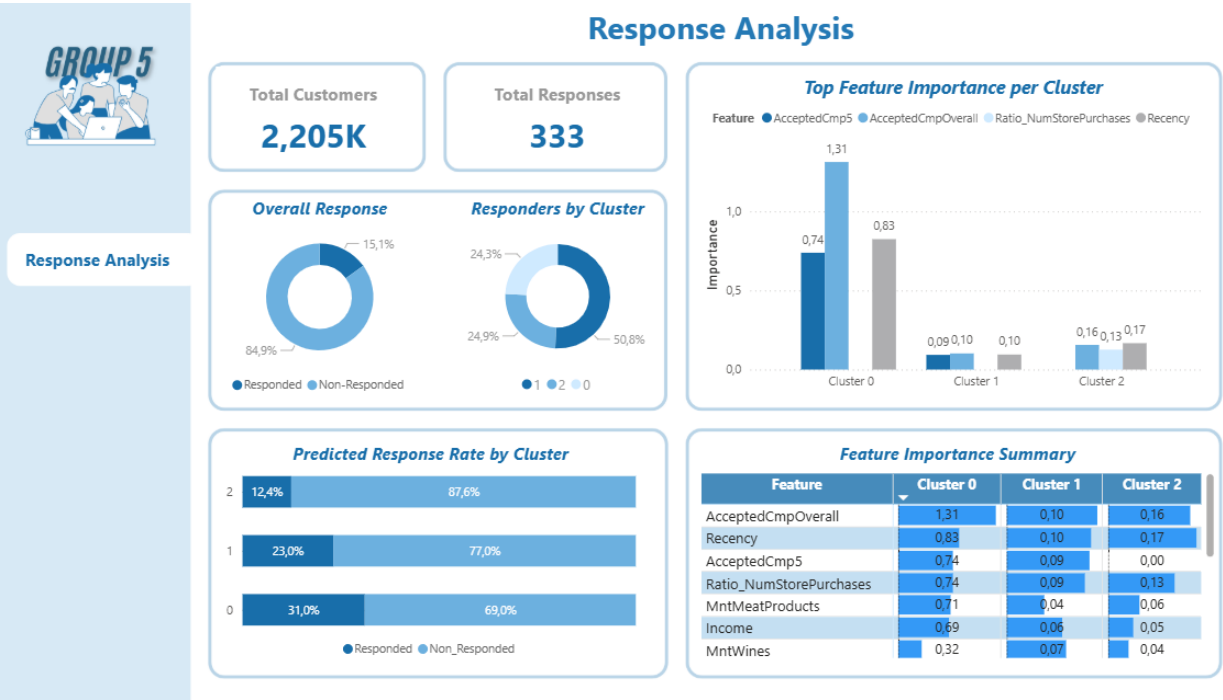

Figure 4.7 Response Analysis Dashboard

The Response Analysis Dashboard gives an overview of how customers respond to iFood's marketing campaigns. Out of 2,205 customers, only 333 responded. This represents 15.1% of the total, while the remaining 84.9% did not respond. This low

response rate shows the challenge of keeping customer interest in a competitive market with rising marketing costs. Therefore, effective customer grouping and predicting response behavior are crucial for businesses to use resources well and tailor campaigns.

When looking at each customer group, Cluster 1 has the highest response rate at 50.8%. This is followed by Cluster 2 at 24.9% and Cluster 0 at 24.3%. However, the predictive model reveals marked differences in potential response levels for the next campaign among the clusters. This shows the gap between actual response behavior and predicted trends, indicating different stages in the customer-brand interaction cycle. Cluster 1 has the highest actual response rate at 50.8%, but the average predicted response likelihood is only 23.0%. This shows message saturation and a drop in engagement. In comparison, Cluster 0 makes up only 24.3% of actual responses but has the highest predicted likelihood at 31.0%. This means that this group is in a reactivation phase, with recent high activity and a strong willingness to buy. This group is the most promising for future campaigns. Meanwhile, Cluster 2 makes up 24.9% of actual responses but has the lowest predicted response probability at 12.4%. This indicates low brand attachment and limited interaction with digital campaigns, likely because of their preference for traditional shopping channels. Thus, the three clusters represent three behavioral states: Cluster 0 is increasing its engagement, Cluster 1 is stable but saturated, and Cluster 2 is declining in interaction. This shows that customer response behavior is dynamic, changing based on engagement level, purchase frequency, and the most recent brand experience.

The Feature Importance analysis shows which attributes most influence customers' likelihood to respond in each cluster. In Cluster 0, the variables AcceptedCmpOverall (1.31) and Recency (0.83) are the most significant. This indicates that customers who have participated in campaigns and made recent purchases are more likely to respond again. This group represents loyal customers who are very engaged with the brand and can be easily brought back through repeat or follow-up campaigns. In Cluster 1, the key factors are Recency (0.10) and AcceptedCmpOverall (0.10), but their lower values suggest this group is less responsive. This might be due to lower interest or weaker attachment to the brand.

In contrast, Cluster 2 is primarily driven by Recency (0.17), AcceptedCmpOverall (0.16), and Ratio_NumStorePurchases (0.13). This shows that recent purchases and how often customers shop in-store are the strongest influences. This group tends to respond better to traditional channels and point-of-sale promotions. Overall, Recency and AcceptedCmpOverall consistently appear as the two most important predictors across all clusters. This confirms that recent interaction history and prior campaign participation are key in predicting how customers will respond. Therefore, keeping customers engaged and retaining historical response data is essential for improving the model's predictive accuracy and overall campaign performance in the future.

Based on the results, we can draw several management implications. First, businesses should focus on keeping highly responsive customers (Cluster 1) by using personalized strategies and offering exclusive benefits, like loyalty programs or appreciation campaigns. Second, for Cluster 0, companies should work to reactivate potential but less responsive customers with email reminders, brand notifications, or incentives for return purchases. Third, since Cluster 2 is strongly influenced by in-store channels, companies should enhance on-site promotions, provide discount vouchers, or create point-accumulation programs to encourage responses. Finally, by using a predictive response model, businesses can improve their marketing budgets. They can focus on customer groups that are likely to respond. This boosts investment efficiency and lowers customer acquisition costs.

Conclusion and Future Works

Conclusion

In an increasingly competitive market with limited marketing budgets, understanding customer behavior and predicting response probability have become essential for optimizing marketing performance. This study analyzed customer segmentation and response prediction in iFood's multichannel marketing campaigns, applying the CRISP-DM framework to ensure a systematic and business-oriented analytical process.

Using a dataset of more than 2,000 iFood customers reflecting multi-channel purchasing behaviors, the research team conducted a complete analytical workflow, including data preprocessing, feature engineering, clustering, predictive modeling and visualization. The K-Means algorithm successfully segmented customers into three clusters, each exhibiting distinct demographic and behavioral profiles. These clusters provided the foundation for developing tailored predictive models and marketing strategies for different customer groups.

For response prediction, three supervised learning models such as Logistic Regression, Random Forest and XGBoost, were trained and evaluated across clusters. Experimental results showed consistently strong predictive performance, with ROC-AUC ranging from 0.785 to 0.866, Accuracy between 0.75 and 0.90 and F1-scores from 0.40 to 0.62. Among them, the Random Forest model achieved the most stable and superior results, especially in Cluster 2, with ROC-AUC = 0.843 and Accuracy = 0.898. These findings confirm the model's reliability in identifying customers with high response potential, allowing iFood to optimize budget allocation and enhance personalization in its marketing campaigns.

In addition, the visualization component using Power BI transformed complex analytical outputs into interactive dashboards, providing actionable insights for management and strategic decision-making. The integration of clustering, prediction and visualization

effectively bridges technical analytics and business practice, supporting data-driven marketing management.

Overall, this project demonstrates the practical value of machine learning in improving marketing efficiency, particularly in customer targeting and resource optimization. Future work may expand this approach by integrating real-time behavioral data, employing uplift modeling to measure campaign impact and automating model retraining to adapt to evolving customer behaviors and market conditions.

Limitation

While the analytical results offer useful insights for optimizing customer outreach, this study still has several limitations that should be considered with caution. These limitations are primarily related to the scope and structure of the data, the class imbalance that affects model training and evaluation, the relatively strong dependence on historical interaction variables and the gap between technical metrics and business effectiveness in real-world deployment. Clarifying these limitations not only enables a more balanced interpretation of the results but also guides methodological improvements and empirical validation in future research.

To begin with, the data scope is narrow, which constrains generalizability. The dataset is a pilot sample of iFood customers in Brazil from 2018-2020, containing many behavioral variables and responses to marketing campaigns. However, the findings may not fully capture other markets, more complex omnichannel operating models, or consumer behaviors outside this time period and geographic context.

The data are also affected by class imbalance, which influences how the model should be evaluated. The customer response rate is relatively low (around 15% across the dataset), necessitating techniques such as SMOTE during training. That said, oversampling can distort the distribution and increase overfitting risk if not accompanied by out-of-sample, time-based validation or real-world A/B testing to confirm effectiveness.

The evaluation metrics currently emphasize model accuracy but do not sufficiently account for measures suitable for imbalanced data or for real cost-benefit considerations. As a result, the reported performance may not fully reflect effectiveness for marketing decision-making.

The feature set relies heavily on historical interactions, reflecting the rule of thumb that "those who responded before are more likely to respond again." This approach helps the model perform well on the current data, but it may be less effective for new customers or those with limited history and it is more vulnerable when market behavior changes rapidly (concept drift).

Finally, there remains a gap between model metrics and business outcomes. Although ROC-AUC by segment appears strong, the report does not measure incremental lift/uplift or conduct field A/B tests to quantify actual impact on budget and revenue. Consequently, there is not yet sufficiently strong empirical evidence of ROI for deployment at operational scale.

Future Work

Building upon the identified limitations, several directions can be pursued to enhance the robustness, interpretability and practical value of this study.

First, future research should expand and diversify the dataset across both temporal and geographical dimensions. Extending data collection to more recent years and integrating customer information from multiple channels - such as websites, mobile apps, social media and offline stores - would allow the model to better capture the omnichannel nature of consumer behavior and improve its generalizability to other markets.

Second, to address the issue of class imbalance and improve model reliability, more advanced resampling and cost-sensitive learning techniques - such as ADASYN, ensemble-based sampling, or hybrid oversampling-undersampling methods - should be explored. In addition, time-based validation and out-of-sample testing should be

implemented to ensure the model remains stable and realistic under dynamic marketing conditions.

Third, subsequent research should incorporate uplift analysis and real-world experimentation to evaluate the true business impact of predictive models. Conducting A/B tests or field experiments will help quantify the incremental lift generated by campaigns and establish a clearer connection between predictive outcomes and return on investment (ROI).

Fourth, future studies should aim to reduce dependence on historical interaction data and develop more adaptive models capable of handling concept drift - the gradual change in customer behavior over time. Introducing real-time behavioral features, such as browsing frequency or social media engagement, can further enhance prediction accuracy and responsiveness.

Fifth, to bridge the gap between technical performance and managerial understanding, future work should emphasize model interpretability by integrating Explainable AI tools such as SHAP or LIME. This will enable marketing managers to better understand why a prediction was made and translate model insights into actionable business strategies.

Finally, a long-term development goal is to build a dynamic decision-support system that integrates predictive models into interactive dashboards. Such a system would allow real-time monitoring of campaign outcomes, automate model updates as new data are collected and support data-driven decision-making in marketing management.

References

[1] Mandapaka, A. K., Kushwah, A. S., & Chakraborty, G. (2014), 'Role of Customer response models in Customer Solicitation Center's Direct Marketing campaign', *Conference: SAS Global Forum*, Washington D.C, 1713-2014.

[2] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013), *Applied Logistic Regression*, Wiley series in probability and statistics.

[3] Huang, J., Lyu, S., Pan, B., Wang, H., Ma, Y., Jiang, T., He, Q., & Lang, R. (2025), 'A logistic regression model to predict long-term survival for borderline resectable pancreatic cancer patients with upfront surgery', *Cancer Imaging*, 25(1), 10.

[4] Balyemah, A. J., Weamie, S. J. Y., Bin, J., Jarnda, K. V., & Joshua, F. J. (2024), 'Predicting Purchasing Behavior on E-Commerce Platforms: A Regression Model Approach for Understanding User Features that Lead to Purchasing', *International Journal of Communications Network and System Sciences*, 17(06), 81-103.

[5] Singh, S., & Rao, A. (2023), 'Application of logistic regression models in marketing', *Proceedings of the Twenty Second AIMS International Conference on Management*, 1657-1661.

[6] Liu, Y., & Yang, S. (2022), 'Application of Decision Tree-Based Classification Algorithm on Content Marketing', *Journal of Mathematics*, 2022, 1-10.

[7] Matzavela, V., & Alepis, E. (2021), 'Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments', *Computers and Education: Artificial Intelligence*, 2, 100035.

[8] Mustakim, N. A., Aziz, M. A., & Rahman, S. A. (2024), 'Predicting consumer behavior in E-Commerce using Decision Tree: a case study in Malaysia', *Information Management and Business Review*, 16(3(I)), 201-209.

[9] Zhou, X. (2024), 'Analysis and mining of user behavior on e-commerce platforms based on decision trees and RFM models', *Highlights in Business Economics and Management,* 33, 243-250.

[10] Guido, G., Prete, M. I., Miraglia, S., & De Mare, I. (2011), 'Targeting direct marketing campaigns by neural networks', *Journal of Marketing Management*, 27(9-10), 992-1006.

[11] Olson, D. L., & Chae. (2012), 'Direct marketing decision support through predictive customer response modeling', *Decision Support Systems*, 54(1), 443-451.

[12] Dai, Y., & Wang, T. (2021), 'Prediction of customer engagement behaviour response to marketing posts based on machine learning', *Connection Science,* 33(4), 891-910.

[13] Gautam, N., & Kumar, N. (2022), 'Customer segmentation using k-means clustering for developing sustainable marketing strategies', *Business Informatics*, 16(1), 72-82.

[14] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.; Wirth, R. (2000), *CRISP-DM 1.0. Step-by-step data mining guide*, CRISP-DM Consortium, Belgium.

[15] Schröer, C., Kruse, F., & Gómez, J. M. (2021), 'A Systematic Literature Review on Applying CRISP-DM Process model', *Procedia Computer Science*, 181, 526-534.

[16] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N. & Flach, P. (2019), 'CRISP-DM twenty years later: From data mining processes to data science trajectories', *IEEE Journals & Magazine*, 33(8), 3048-3061.

[17] Apampa, O. (2016), 'Evaluation of classification and ensemble algorithms for bank customer marketing response prediction', *Journal of International Technology and Information Management*, 25(4), Article 6.

[18] Moro, S., Cortez, P. & Rita, P. (2014), 'A data-driven approach to predict the success of bank telemarketing', *Decision Support Systems*, 62, 22-31.

[19] Lee, X. C., Khaw, K. W., Chew, X., Alnoor, A. & Ng, W. C. (2024), 'Bank direct marketing campaign success prediction', *Applied Mathematics and Computational Intelligence*, 13(3), 95-114.

[20] Vale de Sousa, K. P. (2024), 'Optimising direct marketing through data-driven analytics and predictive models', MSc thesis, National College of Ireland, Dublin, Ireland.

[21] Kotler, P. & Keller, K. L. (2011), *Marketing management*, 14th ed., Prentice Hall, USA.

[22] Ziafat, H. & Shakeri, M. (2014), 'Using data mining techniques in customer segmentation', *International Journal of Engineering Research and Applications*, 4(9), 70-79.

[23] Shirole, R., Salokhe, L. & Jadhav, S. (2021), 'Customer segmentation using RFM model and K-means clustering', *International Journal of Scientific Research in Science and Technology*, 8(3), 591-597.

[24] Wei, J. T., Lin, S. Y. & Wu, H. H. (2010), 'A review of the application of RFM model', *African Journal of Business Management*, 4(19), 4199-4206.

[25] IMARC Group (2025), *Brazil Food Delivery Market Size, Share, Trends and Forecast by Business Model, Order Type, Payment Method, Platform Type, and Region, 2025-2033*, India.

[26] Gao, W., Fan, H., Li, W., & Wang, H. (2020), 'Crafting the customer experience in omnichannel contexts: The role of channel integration', *Journal of Business Research*, 126, 12-22.

[27] Yaiprasert, C., & Hidayanto, A. N. (2023), 'AI-driven ensemble three machine learning to enhance digital marketing strategies in the food delivery business', *Intelligent Systems With Applications*, 18, 200235.

[28] El-Hajj, M., & Pavlova, M. (2024), 'Predictive modeling of customer response to marketing campaigns', *Electronics*, 13(19), 3953.

[29] Lin, J. (2025), 'Application of machine learning in predicting consumer behavior and precision marketing', *PLOS ONE*, 20(5), e0321854.

[30] Bennett, R., Gomez-Donoso, C., Zorbas, C., Sacks, G., White, C. M., Hammond, D., Gupta, A., Cameron, A. J., Vanderlee, L., Contreras-Manzano, A., & Backholer, K. (2025), 'Prevalence of online food delivery platforms, meal kit delivery, and online grocery use in five countries: an analysis of survey data from the 2022 International Food Policy Study', *International Journal of Obesity,* 49, 1307-1316.

[31] Meena, P., & Kumar, G. (2022), 'Online food delivery companies' performance and consumers expectations during Covid-19: An investigation using machine learning approach', *Journal of Retailing and Consumer Services*, 68, 103052.

[32] Grand View Research (2024), *Online Food Delivery Market Size, Share & Trends Analysis Report By Type (Restaurant to Consumer, Platform to Consumer), By Product (Grocery Delivery, Meal Delivery), By Payment Method (Online, CoD), By Region, And Segment Forecasts, 2025 - 2030,* United States.

[33] Lau, T., & Ng, D. (2019), 'Online food delivery services: Making food delivery the new normal', *Journal of Marketing Advances and Practices*, 1(1), 62-77.

[34] Al-Homery, H.A., Ashari, H., & Ahmad, A. (2023), 'Customer Relationship Management: A Literature Review Approach', *International Journal of Global Optimization and Its Application*, 2(1), 20-38.

[35] Payne, A., & Frow, P. (2004), 'The role of multichannel integration in customer relationship management', *Industrial Marketing Management*, 33, 527-538.

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

[37] Ma, Z. (2025), 'Strategies for Enhancing Customer Lifetime Value through Data Modeling', *European Journal of Business, Economics & Management*, 1(1).

[38] Dr. K. Chitra, B. Subashini (2013), 'Data Mining Techniques and its Applications in Banking Sector', *International Journal of Emerging Technology and Advanced Engineering*, 3(8).

[39] Hiran, K. K., Jain, R. K., Lakhwani, K., & Doshi, R. (2022), *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples*, BPB Publications, India.

[40] Aviliani, Sumarwan, U., Sugema, I. & Saefuddin, A. (2011), 'Segmentasi nasabah tabungan mikro berdasarkan recency, frequency, dan monetary: Kasus Bank BRI', *Finance and Banking Journal*, 13(1), 95-109.

[41] Maryani, I., Riana, D., Astuti, R. D., Ishaq, A., Sutrisno, & Pratama, E. A. (2018), 'Customer segmentation based on RFM model and clustering techniques with K-Means algorithm', *Proceedings of the Third International Conference on Informatics and Computing (ICIC)*, IEEE, Bandung, Indonesia, 1-6.

[42] Jain, A. K. & Dubes, R. C. (1988), *Algorithms for clustering data*, Prentice-Hall, New Jersey.

[43] Rahman, A. T., Wiranto & Anggrainingsih, R. (2017), 'Coal trade data clustering using K-Means (case study PT. Global Bangkit Utama)', *ITSMART: Jurnal Teknologi dan Informasi*, 6(1), 24-31, retrieved on September 5th 2025, from ITSMART database.

[44] Johnson, R. A. & Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, 6th edition, Prentice Hall, New Jersey.

[45] Lubis, A. H. (2016), 'Model segmentasi pelanggan dengan kernel K-Means clustering berbasis customer relationship management', *Jurnal Penelitian Teknik Informatika*, 1(1), 36-41.

[46] Perapu, P. (2025), 'Customer segmentation using K-Means clustering for personalized marketing campaigns', *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(3), 810-815, retrieved on September 5th 2025, from IJSRCSEIT database.

[47] Barkhordar, E., Shirali-Shahreza, M. H. & Sadeghi, H. R. (2021), 'Clustering of bank customers using LSTM-based encoder-decoder and dynamic time warping', *arXiv* , retrieved on September 5th 2025, from arXiv database.

[48] Omol, E., Onyangor, D., Mburu, L. & Abuonji, P. (2024), 'Application of K-Means clustering for customer segmentation in grocery stores in Kenya', *International Journal of Science, Technology & Management*, 192-200.

[49] Ester, M., Kriegel, H-P., Sander, J. & Xu, X. (1996), 'A density-based algorithm for discovering clusters in large spatial databases with noise', *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96),* AAAI Press, Portland, Oregon, 226-231.

[50] Çelik, M., Dadaşer-Çelik, F. & Dokuz, A. Ş. (2011), 'Anomaly detection in temperature data using DBSCAN algorithm', *2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, IEEE, Çeşme, Turkey, 91-95.

[51] Pulabaigari, V. & Rajwala, P. (2006), 'l-DBSCAN: A fast hybrid density based clustering method', *18th International Conference on Pattern Recognition (ICPR 2006)*, IEEE, Hong Kong, 1-6.

[52] Sang, Y. & Yi, Z. (2008), 'Motion Determination Using Non-uniform Sampling Based Density Clustering', *Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, Portland, Oregon, 113-120.

[53] Govind, A. & Syam, R. (2024), 'Using DBSCAN to Identify Customer Segments with High Churn Risk on Amazon Consumer Behavior Data', *TechRxiv,* retrieved on September 8th, 2025, from TechRxiv database.

[54] Yan, X., Li, Y., Nie, F. & Li, R. (2025), 'Bank Customer Segmentation and Marketing Strategies Based on Improved DBSCAN Algorithm', *Applied Sciences*, 15(6), 3138, retrieved on September 5th 2025, from MDPI database.

[55] Saragih, H. & Manurung, J. (2024), 'Customer segmentation analysis using DBSCAN method in marketing research of retail company', *Jurnal Teknik Informatika C.I.T Medicom*, 16(5), 321-328, retrieved on September 6th 2025, from Medicom database.

[56] Breiman, L. (2001), 'Random Forests', *Springer Nature*, 45(1), 5-32.

[57] Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020), 'Selecting critical features for data classification based on machine learning methods', *Journal of Big Data*, 52(1), 1-26.

[58] Chen, T., & Guestrin, C. (2016), 'XGBoost: A Scalable Tree Boosting System', *KDD '16*, (2016), 785-794.

[59] Hastie, T., Tibshirani, R., & Friedman, J. (2017), *The Elements of Statistical Learning Data Mining, Inference, and Prediction,* 2nd Edition, Spinger, Stanford, California.

[60] Joshi, R., Gupte, R., & Saravanan, P. (2018), 'A random forest approach for predicting online buying behavior of indian customers', *Theoretical Economics Letters*, 8(3), 448-475.

[61] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, Nicos, Vlahavas, I., & Chouvarda, I. (2016), 'Machine Learning and Data Mining Methods in Diabetes Research', *Computational and Structural Biotechnology Journal*, 15(2017), 104-116.

[62] Rahman, A. M., Mamun, A. A., & Islam, A. (2017), 'Programming challenges of chatbot: Current and future prospective', *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dhaka, Bangladesh, 75-78.

[63] Sisodia, D., & Singh Sisodia, Dilip. (2018), 'Prediction of Diabetes using Classification Algorithms', *Procedia Computer Science*, 132(2018), 1578-1585.

[64] Wang, W., Xiong, W., Wang, J., Tao, L., Li, S., Yi, Y., … Li, C. (2023), 'A user purchase behavior prediction method based on XGBoost', *Electronics*, 12(9).

[65] Yang, Y. (2025), 'A Study on Bank Marketing Prediction Based on XGBoost', *Advances in Economics, Management and Political Sciences*, 185(1), 127-134.

[66] Zheng, Y. (2025), 'Prediction of the effectiveness of bank marketing strategies using the XGBoost model', *Advances in Economics, Management and Political Sciences*, 170(1), 17-28.

[67] Cox, D. R. (1958), 'The regression analysis of binary sequences', *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 20(2), 215-232.

[68] Quinlan, J. R. (1986), 'Induction of decision trees', *Machine Learning*, 1(1), 81-106.