Universitat de Girona
Escola Politècnica Superior

Master´s Thesis

Study: Master in Data Science

Title: Comparative Evaluation of Imputation Strategies for Missing Data and Their Influence on Machine Learning Models in Marketing Analytics. An Application to the Pharmaceutical Sector

Document: Memory

Student: Oksana Klymenko

Tutor: Maria Beatriz Lopez Ibañez
Department: Electrical, Electronic and Automation Engineering
Area: Systems and Automatization Engineering
Tutor: Albert Pla Planas
Sanofi
Area: Date for New Technology (D4NT)

Call: June 2025

Universitat de Girona
Escola Politècnica Superior

MASTER´S THESIS

Comparative Evaluation of Imputation Strategies for Missing Data and Their Influence on Machine Learning Models in Marketing Analytics.

An Application to the Pharmaceutical Sector

Author:
Oksana KLYMENKO

June 2025

Master in Data Science

Tutors:
Maria Beatriz LOPEZ IBAÑEZ
Albert PLA PLANAS

Summary

This master's thesis explores the critical role of data imputation in the context of marketing analytics, with a particular focus on how different imputation strategies affect the performance of machine learning (ML) models. Incomplete data is a common issue in real-world datasets, and marketing is no exception - missing values can significantly hinder insights, predictions, and ultimately, decision-making processes. Therefore, selecting appropriate imputation methods becomes a key step in preparing data for analysis.

The main objectives of this research are to analyze the impact of various imputation methods on ML model performance, to compare imputation approaches in terms of their advantages and limitations, to investigate the relationship between missing data and model accuracy, to determine the most suitable imputation strategy for marketing goals, and to explore how the choice of imputation method can influence data-driven marketing decisions.

The study is based on a real-world dataset obtained from a pharmacy network, which includes information about members and their purchasing behavior. The predictive task involves identifying changes in customer clusters, making the cluster change variable the main target of ML models. Two classification algorithms are applied: Logistic Regression (LR) and Random Forest (RF), chosen for their interpretability and robustness in dealing with structured marketing data.

The methodology is structured in several stages: initial dataset preparation and partitioning, application of different imputation techniques, and evaluation of both imputation quality and downstream model performance. Four distinct scenarios are tested: using only features with no missing values, adding features with missing values but removing all rows with nulls, imputing missing values with statistical measures such as mode, and using K-Nearest Neighbors (KNN) for imputation.

The results of each scenario are compared to assess how the quality of the imputation affects the predictive power of the models. The findings contribute valuable information on the practical implications of imputation choices in marketing analytics and provide recommendations for optimizing data preprocessing pipelines to support more accurate, reliable, and actionable business decisions.

Acknowledgments

I would like to express my deepest gratitude to all those who have supported and guided me throughout the journey of completing this Master's thesis and the entire Master's in Data Science program at the *Escola Politècnica Superior*, Universitat de Girona.

First of all, I would like to sincerely thank **Dra. Beatriz López**, my thesis supervisor, for her invaluable support, encouragement, and insightful guidance throughout this project. Her expertise, clarity, and patience have been fundamental in helping me transform initial ideas into meaningful results. It has been a privilege to work under her supervision and I am truly grateful for the learning experience.

I am also grateful to **Albert Pla**, whose contribution played an important role in shaping this thesis. He shared a link to an article that included the dataset [11] used in this research. This resource served as the foundation for my work and enabled me to explore real-world data with academic rigor. In addition, his valuable feedback helped me refine and improve the quality of this thesis.

Special thanks go to **Mateu Villaret**, who supported my application and gave me the opportunity to join this master program. His confidence in my potential made it possible for me to embark on this academic journey.

Finally, I would like to thank the Universitat de Girona for giving me the opportunity to enroll in this master's program. The education and resources provided by the university have equipped me with the knowledge and skills necessary to successfully complete this thesis. I am deeply grateful for this opportunity and the foundation it has provided me with to advance in the field of data science.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In today's era of advanced analytics, organizations constantly face the challenge of handling large volumes of information to drive strategic marketing decisions. However, a persistent obstacle in this process is the inevitable presence of missing or incomplete data, which can significantly compromise the integrity and reliability of ML models used in marketing analytics.

The approach to addressing these information gaps can profoundly impact the accuracy of predictions, customer segmentation, offer personalization, and ultimately, the return on investment of marketing campaigns. This Master's Thesis emerges from the critical need to systematically evaluate the various imputation strategies available for treating missing data in the specific context of marketing analytics.

Using a comprehensive database of pharmacy chain customers as the primary dataset, this research investigates how different imputation techniques affect model performance. The pharmaceutical retail environment presents unique analytical challenges due to the sensitive and highly variable nature of customer data. As pharmacy chains increasingly compete through personalized services and targeted marketing initiatives, the accuracy of customer analytics has become a crucial competitive advantage.

1.1 Motivation

Handling missing data is one of the most critical and pervasive challenges in data-driven domains, especially in marketing analytics, where incomplete datasets can severely hinder accurate customer segmentation, behavior analysis, and strategic decision-making. Data imputation plays a fundamental role in maintaining the integrity and usability of such datasets. The goal is to recover the original data distribution as closely as possible while minimizing information loss.

Recent literature has emphasized the importance and complexity of data imputation in applied contexts [19]. The findings indicate that the choice of the imputation method significantly influences the outcomes of both supervised and unsupervised learning, especially when dealing with heterogeneous attributes [2]. Furthermore, these studies reinforce the idea that accurate imputation not

only preserves statistical properties but also enhances the performance of downstream models, particularly in predictive analytics applications [18].

The study by [11] highlights that reliable data imputation is essential to improve prediction quality and designing effective reengagement strategies in customer management systems. In this context, missing values may obscure meaningful patterns unless properly handled. The authors propose an advanced pharmacy membership management system that leverages artificial neural networks (ANNs) to activate "sleeping" customers using behavioral data.

This research is motivated by the need to better understand how different imputation strategies - ranging from simple single imputation methods such as mean or regression, to more sophisticated techniques like KNN and multiple imputation - affect the performance of ML models in real-world marketing settings. Gaining insights into these dynamics can inform best practices for data preprocessing and improve the quality of data-driven decision-making in marketing and customer analytics.

1.2 Objectives

This study aims to systematically evaluate the influence of various data imputation techniques on ML model effectiveness within pharmaceutical marketing analytics. The research addresses the critical gap between theoretical imputation approaches and their practical application in real-world marketing scenarios, where data completeness significantly impacts strategic decision-making.

The investigation focuses on understanding how different missingness patterns and imputation strategies interact with the unique characteristics of pharmaceutical customer datasets. By examining this relationship, the study seeks to provide actionable insights for marketing professionals who must balance analytical rigor with operational constraints when handling incomplete data. The specific objectives include:

1. Implement and systematically evaluate a spectrum of imputation techniques, from traditional statistical methods to advanced ML-based approaches, applied to representative pharmacy customer datasets.

2. Quantify the impact of each imputation strategy on key performance metrics for various ML algorithms used in pharmacy marketing, such as customer classification and conversion prediction (segment change).

3. Analyze the interaction between the specific characteristics of the pharmacy customer data and the effectiveness of different imputation techniques.

4. Develop a decision framework that guides marketing professionals and data scientists in selecting the most appropriate imputation strategy based on the type of marketing problem, the ML algorithm employed, and the characteristics of the pharmacy customer dataset.

1.3 Justification and Relevance

The justification for this research lies in several converging factors that underscore its importance from both academic and practical perspectives.

First, modern marketing departments operate in an environment increasingly dependent on data-driven decision-making. According to recent studies, more than 60% of companies consider data a critical business asset, but less than 40% fully trust the quality of their marketing data [15]. This gap represents a significant opportunity to improve the effectiveness of analytical models through more sophisticated imputation techniques.

Second, the retail pharmacy sector specifically stands at the intersection of healthcare and retail, creating a unique data ecosystem where missing values can have significant consequences for customer understanding and service personalization. As pharmacies evolve from traditional dispensing models to comprehensive health service providers, their ability to leverage complete and accurate customer data becomes increasingly vital for strategic positioning and competitive differentiation.

Finally, the advancement of increasingly complex ML algorithms has intensified the need for complete and accurate data. The sensitivity of these models to the quality of input data makes choosing the right imputation strategy a critical decision with direct ramifications for model performance and, by extension, the effectiveness of marketing strategies based on these insights.

1.4 Thesis structure

This master's thesis is structured as follows:

- State of the art: This includes both traditional and ML–based approaches, with a focus on their use in marketing contexts. Relevant studies such as [2], [18], and [11] are discussed to highlight recent developments and gaps in the field.

- Preliminaries: This section introduces the foundational terminology and concepts relevant to both marketing analytics and ML. It aims to provide

the reader with a basic understanding of the domain, the nature of marketing datasets, and the significance of missing data in real-world applications.

- Planning and Methodology: This section outlines the methodological framework used in the study, drawing general vision from the structured approach proposed by [13]. It details each stage of the research - from dataset preparation and missing data simulation to imputation, model training, and performance evaluation. The four imputation scenarios considered are described here, as well as the use of LR and RF models for comparative analysis.

- Methodological Contribution: This part discusses the specific contribution of this thesis in terms of evaluating imputation methods within a marketing context. It reflects on how the approach can support better preprocessing practices and guide data scientists working with incomplete marketing data.

- Results: A summary and interpretation of the findings from the experiments, focusing on how the different imputation strategies affect model accuracy and reliability. The comparative performance of single versus multiple imputation methods is analyzed in detail.

- Conclusions and future work: The thesis concludes with reflections on the main findings, practical implications for marketing analytics, and potential directions for future research, such as exploring deep learning–based imputation techniques or extending the study to other business domains.

CHAPTER 2

State of the Art

In recent years, the management of missing data has become a critical issue in data science and marketing analytics. The performance of ML models is highly dependent on the quality and completeness of the underlying datasets, particularly in marketing applications, where decisions are frequently informed by behavioral, transactional, and segmentation data.

2.1 Impact of Imputation on Machine Learning Models in Marketing

In marketing analytics, ML models are employed for various tasks such as customer segmentation, churn prediction, and response modeling. These models require complete datasets for effective training. The presence of missing values can lead to biased training, increased error variance, and loss of statistical power [24].

Empirical evidence consistently demonstrates that imputation enhances the predictive accuracy of algorithms including decision trees, support vector machines, and neural networks, especially when missingness is systematically modeled [8]. For instance, [11] implemented imputation within an artificial neural network framework to detect and re-engage inactive users in a pharmacy membership system. Their findings underscore the importance of preprocessing in maximizing marketing return on investment and customer lifetime value.

Moreover, [20] caution against the widespread use of listwise deletion, a method that remains prevalent despite its detrimental effects on sample size and inferential power. They promote advanced imputation techniques as a means to safeguard data quality, particularly in contexts where observations are expensive or hard to obtain, such as panel or customer-level data.

Recent research by [19] further emphasizes the critical role of imputation quality in the performance of ML classifiers. Their study reveals that the percentage of missingness in the test data significantly affects classifier performance, with higher missingness rates leading to considerable declines in accuracy. Additionally, they introduce a novel class of discrepancy scores based on the sliced Wasserstein distance to better assess imputation quality. Their findings indicate

that commonly used measures for assessing imputation quality may lead to imputed data that poorly matches the underlying data distribution, thereby compromising the interpretability of models built on such data.

These insights underscore the necessity of carefully selecting and evaluating imputation methods in marketing analytics to ensure the reliability and interpretability of ML models, ultimately informing more effective marketing strategies.

2.2 Imputation Strategies for Missing Data in Marketing Analysis

Imputation techniques in marketing analytics range from basic to highly sophisticated approaches. Simple strategies such as mean, median, or mode imputation are computationally efficient but can distort data distributions, especially under Missing Not at Random (MNAR) conditions [1]. This distortion can lead to biased parameter estimates and misleading inferences that impact marketing decision-making.

Multiple Imputation by Chained Equations (MICE) represents a more advanced method suitable for Missing at Random (MAR) scenarios. It iteratively models each incomplete variable as a function of others, yielding valid inferences while accounting for uncertainty in the imputation process [22], [21]. As [17] emphasize in their statistical framework, MICE offers robust performance across diverse missing data patterns while providing valid standard errors that reflect imputation uncertainty - a critical consideration for marketing analyses requiring precise confidence intervals.

KNN imputation leverages similarity between observations to estimate missing values. Its nonparametric nature makes it especially valuable in marketing datasets where complex, nonlinear relationships may govern consumer behavior [12]. [17] demonstrate that distance-based methods like KNN can preserve variable relationships when missingness mechanisms are complex, though they note its performance degrades with high-dimensional data - an important consideration given the expansive feature sets common in modern marketing analytics.

Recent advances in deep learning have led to the development of autoencoder-based imputation models and generative adversarial networks (GANs), which show promise in capturing high-dimensional, nonlinear dependencies inherent in big marketing datasets [9]. At the same time [17] highlight that these methods excel particularly when dealing with mixed data types (continuous, categorical, and ordinal) frequently encountered in marketing contexts, though they

caution that computational complexity and interpretability remain challenges.

The selection of imputation strategy carries significant implications for downstream marketing analytics. In clustering tasks, for example, poor imputation can obscure latent group structures, leading to inefficient targeting and misallocated resources [16].

[17] contribute valuable insights regarding sensitivity analysis, recommending that marketing analysts routinely compare results across multiple imputation strategies to identify potential areas of model fragility. Their framework introduces the concept of "imputation diagnostics" that quantify the impact of missing data assumptions on final marketing insights - a methodological advance that supports more transparent and reliable analytics.

In summary, the most promising research avenues include combining advanced imputation methods (e.g., MICE, KNN, deep learning) with automated ML workflows and integrating missingness modeling into the prediction pipeline itself. These advances, supported by the statistical foundations established in [17], are expected to further reduce the impact of missing data and support data-driven, personalized marketing strategies in an era of increasing data complexity and customer expectations. The findings serve as a valuable foundation for the development of master's thesis research in marketing analytics, particularly when addressing the ubiquitous challenge of incomplete consumer data.

CHAPTER 3

Preliminaries

3.1 Domain

Marketing analytics has evolved into a complex and data-intensive discipline, essential for understanding consumer behavior, optimizing customer experiences, and enhancing business performance. The effectiveness of analytical methods depends heavily on the quality and completeness of the underlying data.

These challenges become even more pronounced in specialized and highly regulated industries, such as pharmaceuticals. Factors such as privacy regulations, voluntary data reporting, and system integration limitations frequently result in missing or inconsistent data entries, complicating comprehensive analyses. Data governance and completeness are therefore critical. As emphasized in [14], pharmaceutical companies are increasingly turning to advanced analytics to improve medical insights, commercial effectiveness. However, the effectiveness of such approaches is directly tied to data quality and harmonization.

[23] highlights the pivotal role of data completeness in pharma marketing, noting that timely and accurate insights into healthcare professional engagement can significantly influence campaign performance and agility in a competitive market. Similarly, [6] underscores that poor data quality can undermine the deployment of ML in pharma marketing, reducing the value of segmentation, forecasting, and targeting efforts.

Thus, marketing analytics in the pharmaceutical sector not only requires advanced technologies but also demands stringent attention to data completeness, integration, and governance.

3.2 Terminology

This section introduces essential marketing strategy terminology and fundamental ML concepts to establish a common understanding of the project framework.

3.2.1 Terminology on Marketing Strategy

The marketing strategy development in this research is grounded in the following established frameworks that emphasize understanding and serving individ-

ual customer needs:

- **Customer segmentation** refers to the process of dividing customers into distinct groups based on shared characteristics, behaviors, or needs. In retail pharmacy contexts, effective segmentation enables tailored marketing approaches that address specific customer requirements.
- **Customer Lifetime Value (CLV)** represents the projected total value a customer will generate throughout their entire relationship with a business. Marketing resources are optimally allocated when directed toward customer segments with high CLV potential.
- **Customer Relationship Management (CRM)** is a data-driven strategy used in marketing to manage interactions with current and potential customers, aiming to improve customer satisfaction, retention, and business outcomes. CRM systems collect, store, and analyze customer data, which is often subject to missing values, making effective imputation strategies critical for accurate marketing analytics and model performance.
- **Recency, Frequency, Monetary (RFM) analysis** is a customer segmentation technique, designed to evaluate consumer behavior based on three dimensions: recency (the time elapsed since the last purchase), frequency (the number of transactions within a given period), and monetary value (the total expenditure by the customer).

3.2.2 Terminology on Machine Learning

Some essential ML terminology to facilitate understanding of the project are the following:

- **Missing values** refer to the absence of data in one or more variables for a given observation within a dataset. Missing data can significantly affect the validity of statistical analyses and ML models, as it may introduce bias, reduce statistical power, and complicate data interpretation.
- **Imputation method** stands for statistical or ML techniques used to estimate and replace missing values in a dataset. Imputation techniques range from simple strategies such as mean, median, or mode substitution, to more sophisticated approaches including multiple imputation, KNN, regression imputation, and ML-based models such as RF or deep learning. The choice of imputation method depends on the pattern and mechanism of the missing data, as well as the analytical goals and properties of the dataset.

- **Clustering**: An unsupervised learning technique used to group similar data points without predefined labels. It is often applied in customer segmentation or data imputation, where natural structures in the data are identified.
- **Model**: In ML, a model is a mathematical representation of a real-world process built from training data. It is used to make predictions or decisions based on new input data.
- **Logistic Regression (LR)**: A statistical model used for binary classification tasks. It estimates the probability of an outcome based on one or more independent variables and is commonly used in churn prediction and campaign response modeling.
- **Random Forest (RF)**: An ensemble learning algorithm that constructs multiple decision trees and aggregates their outputs to enhance prediction accuracy and control overfitting. It is robust against noise and useful in complex classification tasks.
- **Shapley Values (SHAP)**: A method from game theory used to explain the contribution of each feature to the prediction made by a model. SHAP provides interpretability, helping stakeholders understand and trust the results of ML models.

3.3 Background

3.3.1 Classification of Missingness Models

Missing data mechanisms are typically categorized using the framework introduced by Rubin [3], [4]: Missing Completely at Random (MCAR), MAR, and MNAR. Each mechanism carries implications for data analysis and imputation strategies.

- MCAR occurs when the probability of a data point being missing is independent of both observed and unobserved data. In a marketing context, this may happen if data is lost due to system crashes or random technical failures. Although rare in practice, MCAR allows for unbiased analysis under simple deletion methods (e.g., listwise deletion), albeit at the cost of reduced sample size [1].
- MAR is more common in marketing datasets. It assumes that the probability of missingness is related only to the observed data. For instance, in customer surveys, higher-income individuals might be more likely to skip

questions about monthly spending, but their income is known from other variables. Under MAR, valid inference is possible through techniques like multiple imputation or maximum likelihood estimation [1], [22].

- MNAR occurs when missingness depends on the unobserved data itself. For example, churn-prone customers might intentionally avoid providing satisfaction feedback. This mechanism is particularly challenging, as it violates the assumptions of most standard imputation methods and requires model-based or sensitivity analyses to appropriately account for the missingness structure [16].

Despite the widespread use of this tripartite classification, recent critiques have emerged regarding its practical applicability. Notably, [10] argues that the MCAR/MAR/MNAR framework may oversimplify the complex nature of missingness in multivariable settings typical of real-world marketing datasets. She advocates for a model-driven paradigm that integrates missingness modeling directly into the analytical process.

This includes the pre-specification of missing data assumptions, incorporation of missing data indicators into predictive models, and systematic sensitivity analyses. This approach encourages researchers and practitioners to move beyond rigid classification, favoring pragmatic and context-sensitive strategies that better reflect the underlying data-generating processes [10].

In this thesis, particular attention was given to the MAR and MNAR mechanisms, as these are most representative of the missingness patterns encountered in the pharmaceutical sector. Specifically, the analysis will be conducted within the framework of a network of pharmacy retail stores, where customer and transactional data often exhibit incomplete entries due to selective nonresponse, purchase anonymity, or system integration limitations. Understanding and appropriately modeling these forms of missingness is essential for developing robust and interpretable marketing analytics pipelines tailored to the operational realities of the pharmaceutical domain.

3.3.2 Imputation Approaches

Imputation methodologies can be categorised into two primary approaches: *single imputation and multiple imputation techniques*. Single imputation methods generate a single value for each missing entry, thereby creating a complete dataset for subsequent analysis. In contrast, multiple imputation approaches generate multiple values for each missing observation, providing explicit acknowledgement of imputation uncertainty [4].

Single Imputation Methods

Single imputation replaces each missing value with a single estimate. Common techniques in this category include:

- *Mean/Median/Mode Imputation* method involves substituting missing values with the mean, median, or mode of the observed data. While simple and computationally efficient, it may underestimate variability and distort distributional properties.
- *KNN Imputation* identifies the closest observations and imputes missing values using the average or most frequent values among the neighbors. It preserves local data structure but may be sensitive to outliers and the choice of distance metric.
- *Regression Imputation* involves fitting a regression model to predict missing values using other observed variables. It can yield reasonable estimates when the predictors are strongly correlated with the missing variable but may lead to overfitting or biased estimates in some cases.

Multiple Imputation Methods

Multiple imputation addresses the limitations of single imputation by generating several complete datasets, each with different imputed values, to reflect the uncertainty inherent in the imputation process. The most widely used approaches include:

- *MICE* constructs regression models for each variable with missing values, cycling through them until convergence. It is flexible and can handle different data types and missingness patterns [5].
- *Bootstrap Imputation* involves resampling the observed data to generate multiple complete datasets. It is often used to estimate variance and confidence intervals for statistical inference.

The choice of an appropriate imputation strategy is context-dependent and influenced by various factors, including the type of missingness, the nature of the variables, and the goals of the analysis.

3.4 Data

The primary dataset utilized in this study originates from the supplementary materials provided in the open-access article published in Heliyon [11]. It is important to note that while the main article does not directly employ the data described in Appendix A, this research is based specifically on that attached archive, which contains two separate datasets:

- "Data set of Member Info.xlsx" – a dataset comprising anonymized customer records, including some demographic and behavioral attributes.
- "Data set of Sale Info.xlsx" – a limited dataset capturing transactional activity for a single day, which was deemed non-representative for longitudinal modeling and therefore excluded from this analysis.

The "members" dataset contains 10,000 unique customer profiles and 123 variables, including customer demographics, loyalty program details, engagement metrics, and inferred behavioral patterns. Initial inspection revealed that only 37% of these variables are fully populated, while approximately 20% are considered analytically valuable based on completeness, relevance to marketing objectives, and predictive potential.

The dataset covers the period from January 1, 2016 to December 12, 2021 (based on the “last consume time” field), allowing it to be classified as longitudinal. The data appear to have been recorded automatically, likely capturing user transactions or interactions within an application or web site on a daily basis.

CHAPTER 4

Planning and Methodology

4.1 Planning

The planning encompassed data preprocessing and imputation, implementation of ML models, comparative evaluation of imputation strategies, and the preparation of conclusions and documentation. The following table 4.1 outlines the main activities scheduled for each week of the project, providing a clear overview of the research workflow and its temporal distribution.

4.2 Methodology

This study was structured to ensure the rigorous evaluation of imputation strategies and their influence on ML model performance in the context of marketing analytics for pharmacy retail.

1. **Data Preparation** includes an exploratory data analysis, data cleaning procedure and feature engineering. To maintain data integrity and ensure analytic reliability, preprocessing steps included the removal of duplicate entries, normalization of categorical encoding, and the standardization of data types. Feature engineering was required to deal with the high dimensionality of the dataset. Selection was guided by domain knowledge, correlation analysis, and variance thresholds, ensuring the inclusion of only those features that held statistical or business significance.

2. **Filtering data.** Different filtering strategies were analysed to induce missingness into the data.

3. **Imputation of Missing Data.**To account for the complex nature of missingness - as the dataset exhibits characteristics of both MAR and MNAR - missing data imputation was performed using multiple approaches. In one scenario, basic techniques such as mode substitution were applied, while in another, the KNN imputation algorithm was employed. These strategies were systematically compared to evaluate their impact on subsequent learning models.

Activity	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
Data preparation	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-
Remaining steps with complete data (8,481 samples x 10 features)	-	-	-	X	X	X	-	-	-	-	-	-	-	-	-	-
Remaining steps with complete data (2,687 samples x 10 features)	-	-	-	-	-	-	X	X	X	-	-	-	-	-	-	-
Remaining steps with complete data (10,000 samples x 15 features)	-	-	-	-	-	-	-	-	-	X	X	X	-	-	-	-
Remaining steps with complete data (10,000 samples x 15 features)	-	-	-	-	-	-	-	-	-	-	-	-	X	X	X	-
Writing and conclusions	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	X

Table 4.1: Project Activity Timeline.

4. **Clustering Analysis.** To uncover natural customer groupings, several clustering algorithms were employed: K-Means, Mean Shift, Agglomerative clustering, DBSCAN, and Gaussian Mixture Models (GMM). These methods were validated using internal metrics including the Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, and ARI. The clustering outputs provided a structural basis for further segment-based model training and marketing analysis.

5. **Machine Learning Model.** Two widely accepted supervised learning algorithms were selected to assess the downstream influence of imputation: LR and RF. These models were chosen for their distinct characteristics - LR offering linear transparency and probabilistic outputs, while RF provide non-linear modeling capacity and robustness to noise.

6. **Model Evaluation.** Model performance was evaluated through metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). These measures were calculated using cross-validation across multiple imputation scenarios to quantify the effect of data completeness on prediction accuracy and robustness.

7. **Model Validation.** Each imputation strategy was evaluated under identical validation schemes to ensure fair comparisons.

8. **Interpretability and Shapley Value Analysis.** Finally, Shapley values were used to interpret the contribution of each variable to the predictive outcomes within each modeling scenario. This layer of analysis provided actionable insights into which features - imputed or not - were most influential in predicting customer behaviors, enabling the translation of technical findings into strategic marketing decisions.

CHAPTER 5

Methodological Contribution

This chapter details the specific methodological approaches used throughout the investigation. As can be observed in Fig.5.1, different missingness approaches

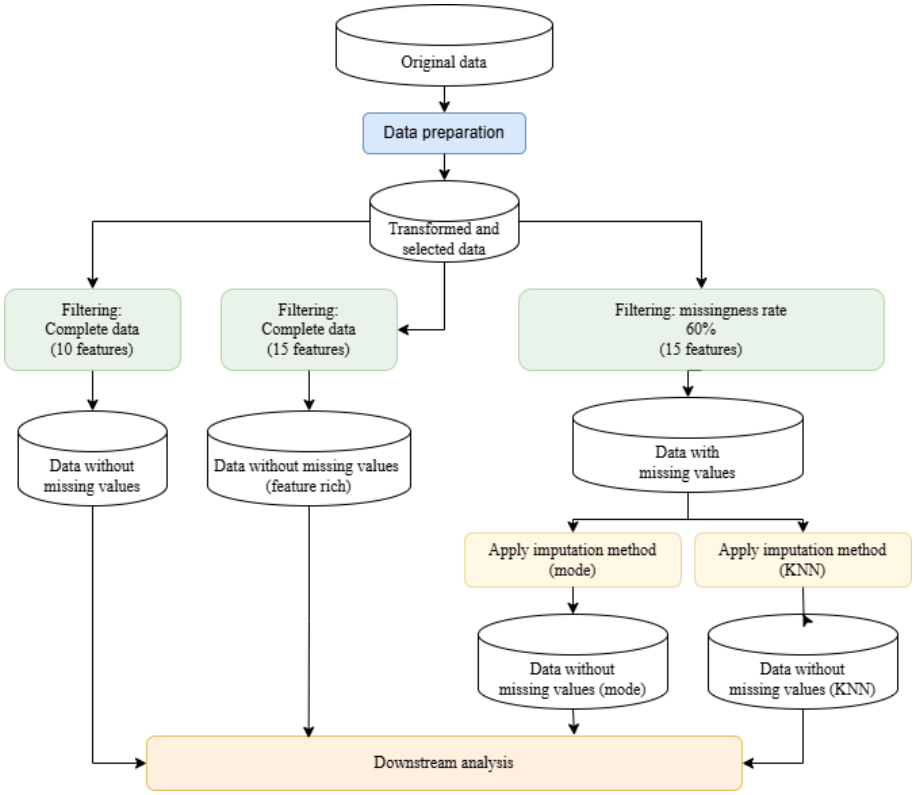

Figure 5.1: Methodological Framework

are categorized based on data completeness and missingness rate. In the completeness-based approach, the number of samples is reduced to include only those that have complete values for a given set of features. In contrast, under the missingness-rate-based approach - specifically when the rate of missing data is less than 60% - the full dataset of 10,000 records is retained, assuming it includes a sufficient proportion of observed values for meaningful analysis. The downstream analysis is illustrated in Fig.5.2, and the individual methodological steps are explained in detail throughout this chapter.

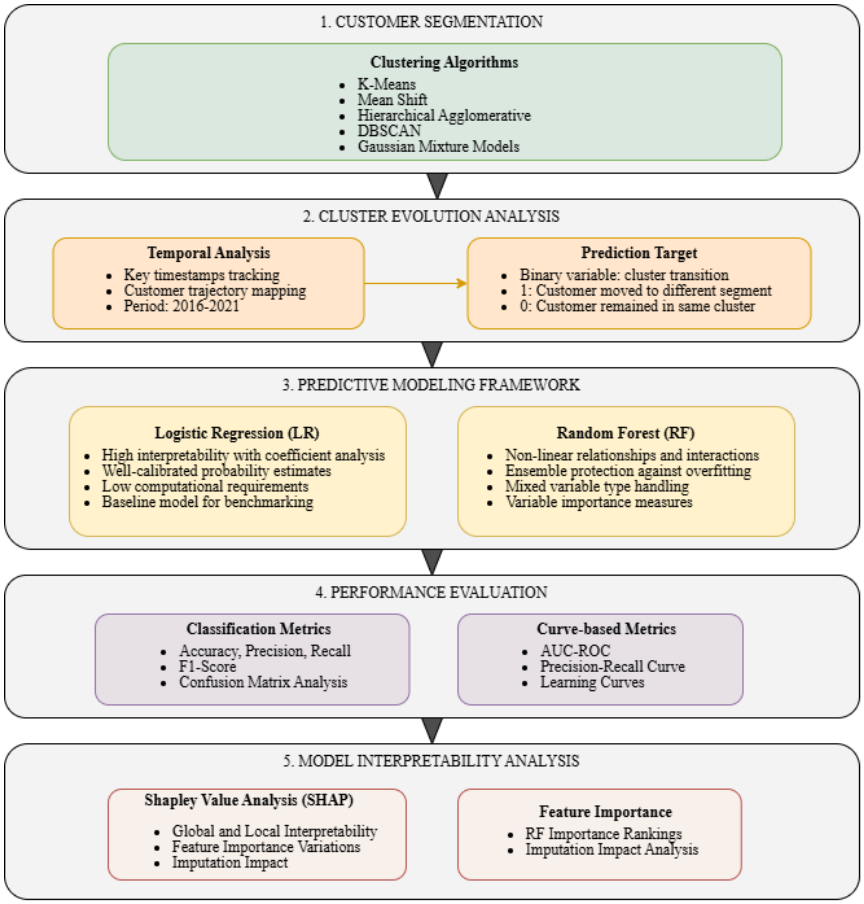

Figure 5.2: Downstream Analysis

5.1 Data Acquisition and Preparation

Fig.5.3 shows the different steps followed to prepare the data for their analysis.

5.1.1 Exploratory Data Analysis.

A thorough exploratory data analysis revealed numerous data quality issues within the dataset:

- Language inconsistencies: the dataset contained a mixture of English and Chinese variable names and values, particularly in demographic fields: The "sex" variable contained Chinese characters representing male female and none "card type" categories were labeled in Chinese, requiring translation before analysis. Several descriptive fields contained untranslated Chinese text, necessitating additional preprocessing.

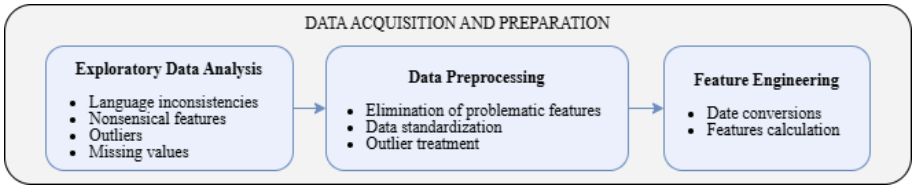

Figure 5.3: Data Acquisition and Preparation Workflow

- Nonsensical features: several variables contained values that lacked interpretable meaning in a marketing context. For example: "fate" contained seemingly random numeric codes [nan, 102918, 113467, 102833, 113735, 102853, 102792] without any documented significance; "value level" showed no variance, with all records containing the value "1"; "daoqi flag" contained "yes" for all entries, rendering it analytically useless "amount" column uniformly contained zeros, providing no discriminatory value, etc.
- Distributional analysis: numerical features exhibited significant skewness and outliers, for instance, "point" value fields contained outliers exceeding standard deviations from the mean, "age" distribution revealed unrealistic values (e.g., ages > 100 and < 10), suggesting data entry errors.

5.1.2 Data Preprocessing

Data preprocessing include elimination of problematic features: variables with zero variance were removed. Features with unintelligible values or codes (e.g., "fate") were excluded after confirming their lack of analytical utility.

Outlier treatment: outlier management was conducted through a systematic approach to identify and address extreme values that could potentially skew analytical results. Standard statistical methods were employed to detect observations falling outside expected ranges. For continuous variables, outliers were identified using established threshold techniques based on distributional properties.

Data standardization: categorical encodings were normalized to ensure consistency (e.g. "member class", "define class", "card type"). Numerical ranges were examined and standardized where inconsistent measurement scales were detected.

5.1.3 Feature engineering

Date fields were converted to datetime format and validated for chronological integrity. Derived features were created, including: "age" and "recency" (time elapsed since the customer's last transaction) (Chapter 3) of calculated as the

difference between the most recent data update timestamp ("updatetime") and the "last consume time".

Through this comprehensive data preparation process, the original 123 variables were reduced to 15 analytically viable features. The resulting dataset provided a clean and structured foundation for subsequent stages of feature selection, imputation, and modeling.

5.2 Filtering

Datasets were formed based on varying levels of completeness and missingness:

- Complete-Case Analysis: Initial clustering was performed on a subset containing only records with complete data across 10 core features, eliminating the need for imputation but reducing the dataset to 8,481 customers.
- Feature-Rich Complete Case: A more feature-rich approach was employed by expanding to 15 variables, but requiring complete cases, which further reduced the dataset to 2,687 customers using listwise deletion (dropna).

5.3 Data Imputation

Among the wide range of imputation techniques (Chapter 3), this study adopts a pragmatic approach by selecting two representative methods: mode imputation and KNN imputation. These methods were chosen based on their relevance to categorical data, computational feasibility, and practical interpretability within the context of pharmaceutical retail analytics. While mode imputation provides a simple and transparent solution suitable for variables with low to moderate levels of missingness, KNN imputation offers a more nuanced alternative capable of capturing local patterns and underlying relationships in more heterogeneous datasets.

5.3.1 Mode Imputation

Mode imputation was selected as a baseline single imputation strategy to address missing values in categorical variables. In the context of pharmaceutical retail, several key features - such as "member class", "sex", "card type", "define class", and "last consume time" - are categorical in nature and directly related to customer segmentation and purchase behavior. While this method may reduce variability and underrepresent minority groups, its simplicity, transparency, and computational efficiency make it particularly well-suited for retail environments

where timely insights and interpretability are essential. In this study, it served as a practical benchmark to evaluate the relative performance of more advanced imputation techniques.

5.3.2 KNN Imputation

KNN imputation was applied to the same set of categorical features as mode imputation under a more sophisticated framework that accounts for the structure and heterogeneity of the data. Unlike mode imputation, which imputes based solely on frequency, KNN identifies the k most similar customer profiles using a defined distance metric and imputes missing values based on the most representative patterns among those neighbors. This makes KNN particularly suitable when the data suggest non-linear relationships or subgroup-specific behaviors, as is common in pharmacy chains where customer preferences and health-related consumption vary considerably. By preserving local structures and incorporating relational nuances, KNN imputation offers a more robust alternative for enhancing data completeness in marketing analytics tasks.

5.4 Customer Segmentation

Customer segmentation was performed using five clustering algorithms:

- K-Means clustering with optimal cluster counts determined through the elbow method, which identified the point of diminishing returns in explained variance.
- Mean Shift clustering with adaptive bandwidth selection based on the distribution of pairwise distances in the feature space.
- Hierarchical Agglomerative clustering using Ward's linkage criterion to minimize within cluster variance while maintaining interpretable segment hierarchies.
- DBSCAN with epsilon neighborhood parameters optimized through distance distribution analysis, particularly valuable for identifying non-convex clusters.
- Gaussian Mixture Models with component selection guided by Bayesian Information Criterion, allowing for probabilistic cluster assignments.

Clustering was applied in each period to identify evolving customer segments over time. Given that the dataset spans from 2016 to 2021 ("last consumption date") (Chapter 3), it qualifies as longitudinal, capturing dynamic of

customer behavior across multiple years. To investigate these temporal dynamics, the data were partitioned into five consecutive yearly intervals as follows:

- 2016–2017: customers active between 2016 and 2017.
- 2017–2018: customers active between 2017 and 2018.
- 2018–2019: customers active between 2018 and 2019.
- 2019–2020: customers active between 2019 and 2020.
- 2020–2021: customers active between 2020 and 2021.

Each subset was created by filtering the dataset using the customer's activity year. This segmentation strategy enables a comparative analysis of clustering results across time, revealing structural changes in customer profiles and consumption behavior.

Cluster solutions were validated through multiple complementary approaches: Internal validation metrics including Silhouette Score (to assess cluster cohesion and separation), Calinski-Harabasz Index (to evaluate the ratio of between-cluster to within-cluster variance), Davies-Bouldin Index (to measure the average similarity between clusters), and Adjusted Rand Index (ARI) (to quantify the agreement between the clustering assignments).

5.5 Cluster Evolution Analysis

Building upon the static customer segmentation, a dynamic analysis of cluster evolution was conducted to capture customer migration across segments over time. For example, to enable this longitudinal tracking, the unique identifier "card id" was used to match customers across periods and evaluate their movement between clusters.

5.5.1 Temporal Analysis

Temporal analysis focused on detecting customer migration trajectories and the evolution of segment profiles over time.

The most substantial behavioral shifts began in 2020, a trend potentially linked to the introduction of new strategic marketing initiatives during that period. The observed transition dynamics intensified in the final observation window leading up to December 2021, which exhibited the highest segment volatility across the entire study period, as shown in Fig. 5.4 and Fig. 5.5 for the values "costumer level" and "point" of the data without miising values.

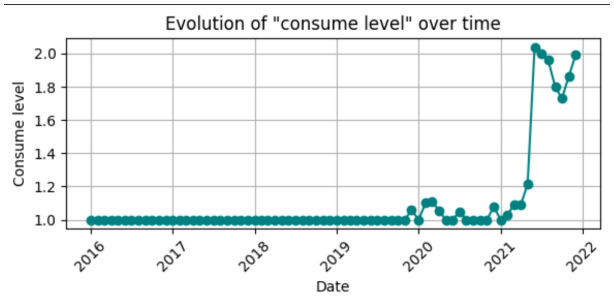

Figure 5.4: Evolution of "consume level" over Time

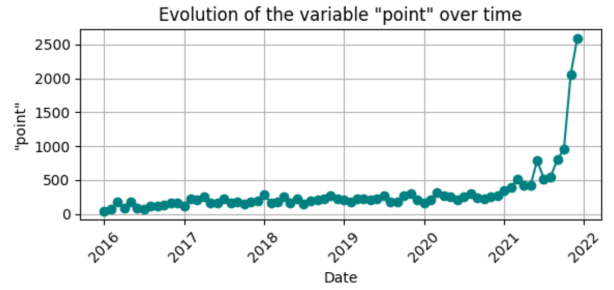

Figure 5.5: Evolution of the Variable "point" over Time

To deeper investigate how customer behavior evolved over time, transitions between clusters were analyzed across consecutive year periods (Fig. 5.6). This transition analysis was enabled by leveraging the unique customer identifier, "card id", which allowed individual customer trajectories to be tracked longitudinally. By matching records across yearly intervals, it was possible to identify whether customers remained in the same cluster or migrated to a different one in subsequent periods.

The analysis revealed varying degrees of cluster stability over time, with the most pronounced changes occurring during the 2020–2021 interval (Table 5.1). Although the increase in retained customers indicates that more customers continued their engagement year over year, the sharp drop in cluster consistency suggests a significant shift in customer behavior or segmentation patterns.

	2018 → 2019	2019 → 2020	2020 → 2021
Retained customers	47.1%	52.3%	55.9%
Same cluster	61.4%	64.8%	20.4%

Table 5.1: Between-period Retention Analysis.

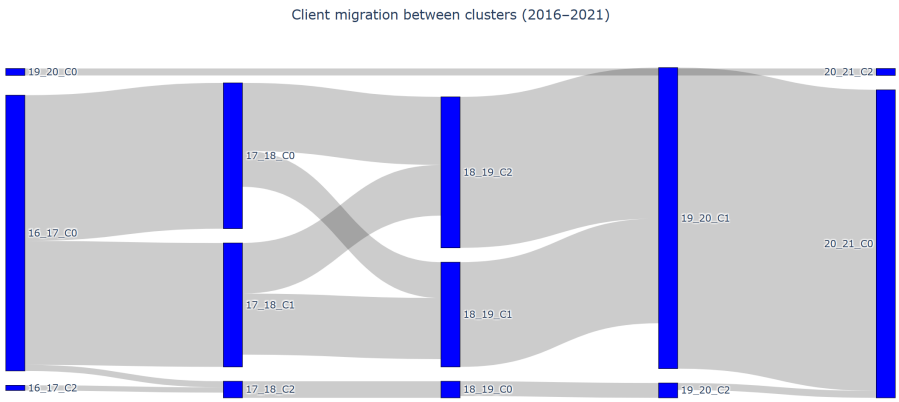

Figure 5.6: Complete Customer Jorney over Time

5.5.2 Prediction target

As a result of this temporal analysis, the principal prediction target for the modeling phase was defined as the probability of a cluster transition within a subsequent time window. This was operationalized as a binary variable indicating whether a customer moved to a different segment (1) or remained in the same cluster (0) during the next observed period.

This approach allows to describe changes in customer behavior. By predicting segment transitions, the models offer the potential to identify customers at pivotal decision points in their engagement lifecycle, thereby enabling more proactive and strategically aligned marketing interventions.

5.6 Predictive Modeling Framework

The selection of appropriate ML algorithms was guided by both theoretical considerations and practical requirements specific to marketing analytics applications in the pharmacy retail context. After careful evaluation of various modeling approaches, two complementary algorithms were used as the core predictive framework: LR and FR.

5.6.1 Logistic Regression

LR is selected primarily for its high degree of interpretability, offering straightforward coefficient interpretation. This transparency is critical in marketing contexts where stakeholders need to understand which factors drive customer behavior. The model produces well-calibrated probability estimates rather than

just binary classifications, enabling nuanced customer scoring and tiered marketing approaches based on predicted likelihood of response.

Given the need to train multiple models across different imputation scenarios, LR's low computational requirements allowed for extensive experimentation and validation. As a parametric linear model, LR served as an effective baseline against which more complex models could be benchmarked, helping isolate the impact of imputation strategies versus model complexity.

5.6.2 Random Forest

RF is chosen for its ability to capture complex non-linear relationships and interactions between variables without explicit specification, addressing the potential limitations of LR's linearity assumption. The ensemble nature of RF, combining multiple decision trees trained on bootstrapped samples, offered inherent protection against overfitting - a particular concern when working with imputed data that might introduce artificial patterns.

The algorithm provides robust variable importance measures that complement and contrast with LR coefficients, offering deeper insights into predictive drivers across different imputation scenarios. RF's ability to effectively process both categorical and continuous variables without extensive preprocessing made it well-suited for marketing datasets with mixed variable types.

This dual-model approach created a comprehensive framework for evaluating imputation strategies through both linear and non-linear modeling lenses, providing complementary perspectives on how different approaches to missing data influence predictive performance in marketing analytics applications. For each algorithm, separate models were trained using datasets produced by each imputation method, creating a comprehensive matrix of model-imputation combinations for comparative analysis.

5.7 Performance Evaluation

The performance of the predictive models was comprehensively assessed using a range of quantitative evaluation techniques to ensure both robustness and interpretability. Standard classification metrics were applied, including accuracy, precision, recall, and the F1-score, to capture different aspects of predictive capability and to address potential trade-offs between false positives and false negatives.

To further evaluate model discriminative power, particularly in the presence of class imbalance, two key curve-based metrics were employed: the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the Precision-

Recall Curve. These metrics provided a more nuanced understanding of model performance, especially when the target variable (cluster transition) exhibited skewed distribution.

In addition to point-wise evaluations, a cross-validation strategy was employed to ensure the generalizability and reliability of model performance across different subsets of the data. Specifically, 5-fold cross-validation was used to reduce the variance associated with random train-test splits, offering a more robust estimate of model performance.

To further examine model behavior and convergence, learning curves were generated by plotting training and validation performance as a function of the training dataset size. These curves allowed for the evaluation of underfitting or overfitting tendencies across different imputation strategies, providing insight into whether certain techniques introduced noise or instability that impaired model learning. For instance, learning curves that displayed high variance between training and validation accuracy could signal overfitting caused by biased imputation, whereas consistently low performance might indicate underfitting or information loss during preprocessing.

The confusion matrix was also examined for each model to understand the distribution of prediction errors across the binary classification outcomes and to identify any systematic misclassification patterns. In the context of this research - focusing on the comparative evaluation of imputation strategies for missing data in marketing analytics - this analysis played a critical role in assessing how different imputation techniques influence model behavior and predictive reliability.

By analyzing confusion matrices across various imputation scenarios, it was possible to detect consistent tendencies toward false positives (i.e., incorrectly predicting a customer transition when none occurred) or false negatives (i.e., failing to detect an actual transition), which might be linked to the noise or bias introduced during data imputation.

Furthermore, feature importance rankings, particularly from the RF classifier, served as a foundational analysis for subsequent model interpretability. These importance values guided the application of Shapley value decomposition, enabling a detailed inspection of feature contributions at the individual prediction level. This interpretability phase provided critical insight into the underlying drivers of customer behavior and transitions, thereby informing the development of targeted marketing strategies aligned with specific customer dynamics.

5.8 Model Interpretability Analysis

This section presents a critical assessment of model interpretability across imputation strategies, focusing particularly on Shapley values in comparison with traditional RF importance metrics. The ability to interpret ML models is paramount in marketing contexts, where stakeholders must understand the drivers behind model predictions to deploy them effectively in business applications.

Shapley values, derived from cooperative game theory, [7] provide a mathematically robust approach to assessing feature contributions to individual predictions. Unlike traditional feature importance metrics that offer global perspectives, Shapley values enable both global and local interpretability, making them especially valuable for marketing applications where understanding individual customer behaviors is essential. The Shapley value for feature j is calculated as:

$$\phi_j(f,x) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!} [f_x(S \cup \{j\}) - f_x(S)]$$

where N is the set of all features, S represents a subset of features not containing j, and $f_x(S)$ denotes the prediction for instance x using only the features in subset S.

In this thesis, SHAP (SHapley Additive exPlanations) was implemented to compute Shapley values across models trained on datasets processed with different imputation strategies. This approach allows to:

- Quantify how feature importance shifts across imputation methods.
- Assess whether certain imputation techniques systematically alter the relative contribution of features.
- Identify potential distortions in the relationships between predictors and target variables.

The interpretability analysis reveals several important considerations for marketing applications. First, the substantial variation in feature importance depending on the imputation method highlights the potential risk of relying on a single imputation strategy to derive marketing insights. This variability can lead to inconsistent or even misleading conclusions.

Moreover, SHAP values prove especially effective in uncovering complex feature interactions that are often overlooked by traditional importance metrics like those from RF. This capacity enhances the understanding of nuanced customer behavior patterns. It is also evident that simpler imputation techniques, such as mean imputation, tend to introduce the most distortion in the assessment of

feature importance, thereby increasing the likelihood of misinterpretation and suboptimal strategic decisions.

Finally, the ability of SHAP values to provide local, instance-level explanations adds significant value to personalized marketing efforts, as it enables a clearer understanding of the rationale behind specific model recommendations for individual customers.

This comprehensive methodological framework provided a rigorous foundation for evaluating how different imputation strategies influence the performance and reliability of ML models in pharmacy retail marketing analytics, with particular attention to the downstream effects on decision-making processes.

CHAPTER 6

Results

The implementation of the system has been done with Python from the Google Colab environment. The code can be found at GitHub.

6.1 Complete-Case Results

Starting with fully observed variables ensures that the clustering results are not influenced by imputation artifacts or assumptions regarding the missing data mechanism. This scenario serves as a baseline to which subsequent, more complex imputation-based approaches can be compared.

The analysis begins with a complete-case analysis based on a subset of the dataset containing only those variables that are fully observed for all customers. Although this approach limits the analysis to a reduced number of features (10 variables) and a smaller sample size (8,481 records out of 10,000 due to the removal of outliers), it offers several methodological advantages that justify its initial use.

6.1.1 Segmentation Results

To determine the optimal number of clusters in the complete-case scenario, the Elbow Method was first applied to the K-Means clustering algorithm (Fig. 6.1). The resulting graph exhibited a steep decline in inertia at lower values of K, followed by a visible "elbow" at K=2. However, despite K=2 being statistically optimal according to the Elbow Method, a minimum of three clusters was selected based on the RFM analysis (Chapter 3). This prior segmentation identified three meaningful behavioral groups - sleep, active, and intermediate - which align more closely with marketing objectives and customer lifecycle stages.

To reinforce the analysis, multiple internal validation metrics were computed for several clustering configurations (Table 6.1). The Mean-Shift algorithm achieved the highest Silhouette Scores, however, produces 32 clusters, a granularity that is impractical for the current business context. The Gaussian Mixture Model (GMM) and K-Means with 4 clusters both showed competitive internal metrics. Nonetheless, increasing the number of clusters beyond three risks over-segmentation. Considering these factors, K-Means with 3 clusters presents

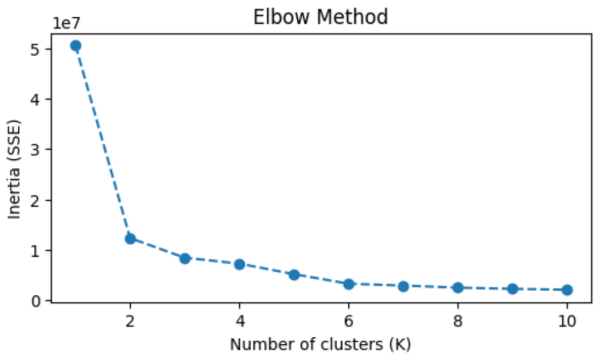

Figure 6.1: Complete-Case Analysis: Elbow Method.

Method	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means (3 clusters)	0.617801	3191.869251	1.274137
K-Means (4 clusters)	0.660912	3168.424106	1.066105
Mean-Shift	0.707130	1627.974304	0.761820
Agglomerative clustering	0.554817	3161.797355	1.2137093
DBSCAN	0.494605	317.3564784	1.423389
GMM	0.598063	3824.766953	0.766933

Table 6.1: Complete-Case Analysis: Clustering Algorithms.

a balanced solution. Additionally, Agglomerative Clustering with 3 clusters was considered due to its hierarchical nature and comparable internal metrics.

Following the K-Means analysis, Agglomerative clustering was applied using the same number of clusters (K=3) for consistency and comparison. To assess the consistency and agreement between clustering algorithms, a comparative analysis was conducted between K-Means and Agglomerative clustering, both applied using three clusters on the same complete-case dataset (Table 6.2). The

cluster	K-Means Count	Agglomerative Count
0	5,589	5,585
1	2,880	2,884
2	12	12

Table 6.2: Complete-Case Analysis: Cluster Distributions.

numerical distributions of clusters between the two methods are remarkably similar, suggesting a high degree of structural consistency in the way both algorithms partition the data.

To formally quantify the similarity between the cluster assignments, *the ARI was computed, yielding a score of 0.967*. This ARI score indicates excellent agreement, validating that both methods arrive at highly comparable segmentations

despite their distinct computational principles (centroid-based vs. linkage-based).

Furthermore, the 3D scatter plots produced for both clustering methods - based on the "point", "consume level", and "define class" features - exhibited near-identical cluster structures. This visual similarity reinforces the quantitative findings, suggesting that the underlying segmentation patterns are robust and algorithm-independent (Fig. 6.2).

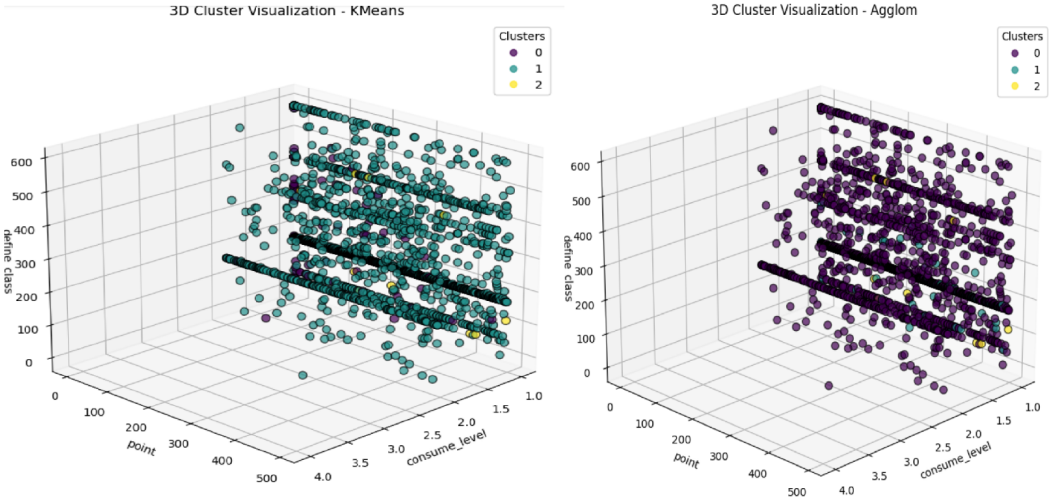

Figure 6.2: Complete-Case Analysis: 3D Cluster Visualization.

Strategic Justification for Feature Selection

The variables “point”, “consume level”, and “define class” were selected as core dimensions for three-dimensional cluster visualization and customer segmentation, due to their direct strategic relevance for marketing operations and customer lifecycle management within the pharmacy retail network.

“Point” represents an accumulated customer score - typically linked to loyalty programs or purchase activity. It reflects both frequency and monetary value of transactions, aligning closely with CLV and engagement. High point values usually correspond to the most loyal and valuable customers.

“Consume level” captures the intensity of purchasing behavior, distinguishing high-activity customers from occasional or passive users. From a marketing perspective, this variable is essential for prioritizing marketing spend and identifying segments that respond better to promotions, cross-selling, or upselling strategies.

“Define class” is a categorical segmentation variable typically assigned by the CRM system based on pre-defined business rules or thresholds (e.g., regular, low-tier). It is often used in campaign targeting and service personalization. Monitoring how this class evolves over time provides insight into customer migration dynamics and the effectiveness of loyalty interventions.

Together, these three features form a multidimensional view of customer value, behavioral intensity, and strategic classification, which are key to designing and evaluating tailored marketing actions.

By incorporating these features into clustering and visualizations, the analysis enables the tracking of customer transitions between clusters over time. This is especially valuable in understanding:

- Which customers upgrade to more valuable segments following marketing interventions.
- Which customers show signs of disengagement or attrition, as reflected in dropping consume levels or points.
- How well the internal CRM classes match the results of data-driven segmentation.

This insight allows pharmacy retailers to proactively adjust their marketing strategies, such as reward programs, personalized offers, or retention efforts, based on measurable changes in customer segmentation profiles.

In summary, the selection of “point,” “consume level,” and “define class” bridges the gap between algorithmic segmentation and actionable business strategy, ensuring that clustering results can inform real-world decisions and drive tangible outcomes.

6.1.2 Cluster Evolution Results

Following the establishment of a three-cluster segmentation structure, the subsequent step focused on analyzing the temporal dynamics of customer behavior. Specifically, cluster memberships were tracked across two consecutive time periods, enabling the quantification of customer transitions between clusters.

The analysis of customer migration patterns (Fig. 6.3) revealed that the majority of customers initially assigned to clusters 0 and 1 remained within their respective clusters, with only a small proportion transitioning to others. In contrast, customers from cluster 2 exhibited greater mobility, with approximately 5% changing their cluster membership in the following period.

Based on this observation, a binary target variable "change cluster" was defined to support predictive modeling tasks. This variable takes the value of 1 if the customer changed clusters in the most recent period, and 0 if the customer remained in the same cluster. This binary outcome was subsequently used as the dependent variable in classification models aimed at predicting customer migration behavior.

Figure 6.3: Complete-Case Analysis: Cluster Migration Analysis.

6.1.3 Predictive Results

To predict whether a customer is likely to change cluster status, two classification models were developed and compared: LR and RF classifier. Both models were trained on a comprehensive feature set representing historical behavior, transactional patterns, and categorical classifications derived from the CRM system. Model performance was evaluated using standard metrics such as accuracy, precision, recall, and AUC-ROC, ensuring both predictive power and business relevance.

LR Results

The LR model was trained and validated. This model served as a baseline due to its interpretability and effectiveness in binary classification tasks. The model demonstrated predictive performance, as evidenced by the metrics shown in the Table 6.3. These metrics reflect the model's effectiveness in distinguishing

Metric, %	Class 0	Class 1	Macro Avg	Weighted Avg
Precision	98	18	58	94
Recall	82	69	76	82
F1-Score	90	28	59	86

Table 6.3: Complete-Case Analysis: LR Model Performance

positive and negative cases while maintaining stable predictive accuracy. However, class-wise evaluation reveals a significant imbalance in performance. The majority class 0 achieved high precision (0.98) and recall (0.82), leading to a strong F1-score of 0.90. In contrast, the minority class 1 showed low precision (0.18) but moderate recall (0.69), resulting in a poor F1-score of 0.28. This indicates the model tends to under-predict the minority class, potentially missing many true positive cases. The macro-averaged F1-score of 0.59 highlights this imbalance, despite a weighted average F1-score of 0.86 driven by the dominant class. These findings suggest further model optimization is needed to improve minority class detection depending on the application's critical requirements.

With an AUC of 0.86 (ROC Curve Fig. 6.4), the model exhibits moderate discriminatory power between classes.

The AUC of 0.26 in Precision-Recall Curve (Fig. 6.4) is low, meaning precision drops as recall increases, indicating class imbalance.

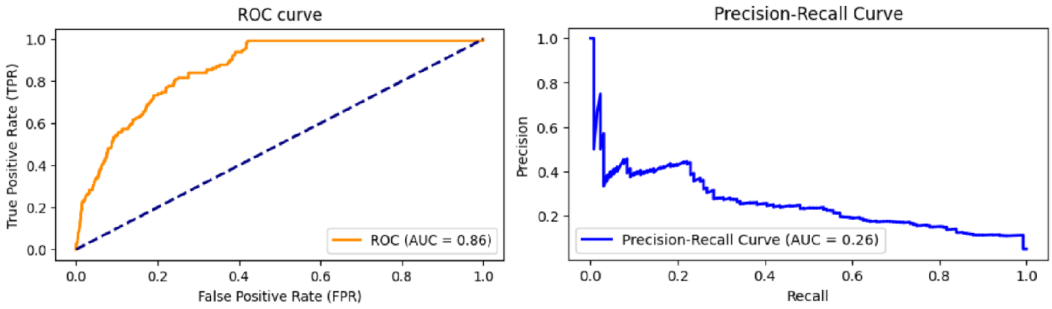

Figure 6.4: Complete-Case Analysis: LR Performance Metrics.

A breakdown of predictions in the confusion matrix further confirms the model's precision (Fig. 6.5).

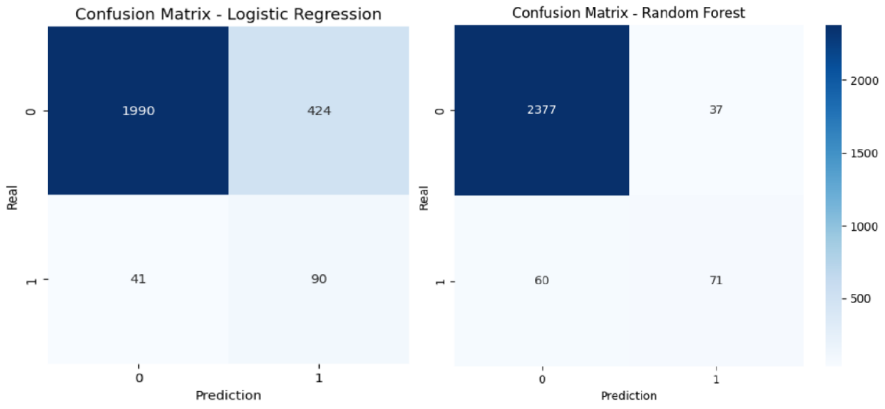

Figure 6.5: Complete-Case Analysis: Confusion Matrix.

The learning curve (Fig. 6.6) provides valuable insights into model behavior as a function of training set size.

The train accuracy starts high (0.95) and gradually decreases as the dataset size increases. This suggests that with more data, the model is generalizing better but losing some of its initial overfitting benefits. The test accuracy starts lower (0.943), then improves as the dataset size increases but later slightly drops, indicating potential saturation in learning. As more data is added, the gap between train and test accuracy decreases, it means the model is learning better and becoming more generalized. However, the drop in test accuracy at

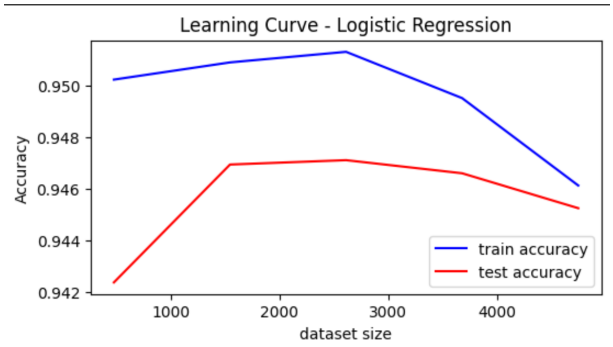

Figure 6.6: Complete-Case Analysis: LR Learning Curve.

larger dataset sizes could imply that the model has reached its limit in improvement.

RF Results

To complement the LR baseline, a RF classifier was trained. The RF model surpasses the LR results in some metrics (Table 6.5)

Metric, %	Class 0	Class 1	Macro Avg	Weighted Avg
Precision	98	66	82	96
Recall	98	54	76	96
F1-Score	98	79	98	96

Table 6.4: Complete-Case Analysis: RF Model Performance

The model demonstrates strong discriminative performance, as reflected by a high average ROC AUC score during cross-validation (0.9464) and a consistent ROC AUC on the test set (0.9498). The overall test accuracy reaches 96%, indicating that the model correctly classified the vast majority of cases. However, a closer examination of class-specific performance reveals limitations in detecting the minority class (class 1). The confusion matrix confirms these findings (Fig. 6.5). Despite a small number of false negatives and false positives, the model achieves high sensitivity and specificity, performing well in both class 0 and class 1 predictions. This is noteworthy given the considerable class imbalance - the dataset contains significantly more class 0 samples than class 1 samples.

The learning curve of RF offers key insights (Fig. 6.7). The model achieves high accuracy on the training data very quickly. The validation accuracy stabilizes around 0.96, though lower than training, this performance is still good. However, small fluctuations indicate diminishing returns from increasing the number of trees.

The comparison of LR and RF shown in the table 6.5 supports the selection

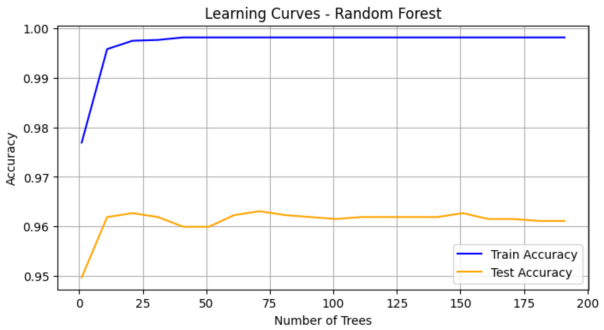

Figure 6.7: Complete-Case Analysis: RF Learning Curve.

of RF as the most suitable algorithm for the use case.

Metric, %	LR	RF
Accuracy	82	96
ROC AUC (cv)	87	94
ROC AUC	85	94
Precision	94	96
Recall	82	96

Table 6.5: Complete-Case Analysis: Logistic Regression vs Random Forest

6.1.4 Model Interpretability Results

To enhance understanding of the predictive mechanisms behind the cluster change models, we conducted a feature importance analysis using RF built-in feature importance and SHAP values. This dual approach enables a more nuanced interpretation by comparing algorithmic importance with model-agnostic explanations.

The results (Fig. 6.8) show that the discrepancies between applied methods. SHAP assigns greater importance to "level", suggesting its strong influence on individual predictions, whereas Random Forest considers point more impactful in overall model accuracy. Features like "age" and "consume level" receive higher importance from Random Forest, indicating their relevance across the dataset, though SHAP suggests their contributions may be more context-dependent. Meanwhile, "is suff" has minimal significance in both approaches, reinforcing its limited role in predictive outcomes. The SHAP waterfall (Fig. 6.8) highlights how "level" contributes +0.17 towards increasing the predicted probability of cluster change for this instance - making it the dominant driver in the decision. Other features like "point", "consume level", and "age" contribute marginally, while "is suff" have no impacts.

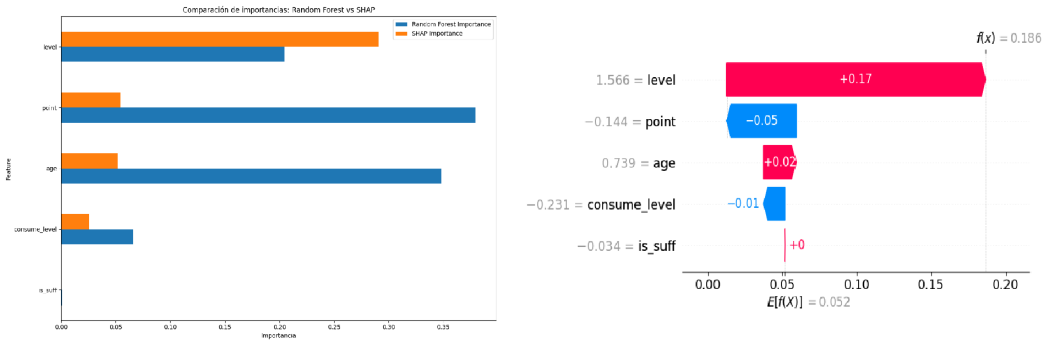

Figure 6.8: Complete-Case Analysis: Shap Importance Comparison.

Such granular insights from SHAP complement the aggregate view offered by RF, providing a more trustworthy and explainable model. In sensitive applications such as customer segmentation and behavioral prediction, this transparency can guide business interventions with greater confidence.

6.2 Feature-Rich Complete Case Results

A more comprehensive approach was implemented by expanding the dataset to include 15 features. These features include: "*member class*", "*last consume time*", "*sex*", "*card type*", *and* "*define class*". To ensure the reliability of the analysis, listwise deletion (dropna) was applied, resulting in a final dataset of 2,687 customers with complete cases.

To better understand the nature of the added features, it is important to describe several of them:

- "Member class" is a categorical variable with 3,407 unique values, potentially reflecting different levels of customer segmentation or purchase behavior.

- "Card type" is also categorical and indicates the type of loyalty card held by the customer, such as "Membership Card", "Peace Pharmacy Silver Card", and others.

- "Define class" is another categorical feature with 586 unique categories, possibly linked to customer profiling or service classification.

Importantly, the inclusion of the last "consume time" variable enabled the computation of "recency", a key metric in customer behavior analysis. Recency refers to the number of days since the customer's most recent transaction. It is

a critical component in RFM analysis and is widely used in customer segmentation, especially in clustering algorithms. High recency (i.e., recent activity) typically indicates an engaged customer, whereas low recency may signal attrition or reduced interest. Incorporating recency thus enhances the model's ability to distinguish between active and inactive customers, making it invaluable for both strategic marketing and ML applications.

6.2.1 Segmentation Results

To determine the optimal number of clusters, the Elbow Method was first applied to the K-Means clustering algorithm. The resulting plot (Fig. 6.9) illustrates a steep decline in inertia - initially very high at K = 1 - which drops sharply until K = 3. After this point, the rate of decrease diminishes, forming the characteristic "elbow" shape. This suggests that the optimal number of clusters lies at K = 3 or K = 4, beyond which additional clusters contribute minimally to improving model performance. Therefore, K = 3 or K = 4 is considered the most appropriate choice for this dataset. To reinforce these findings, several internal

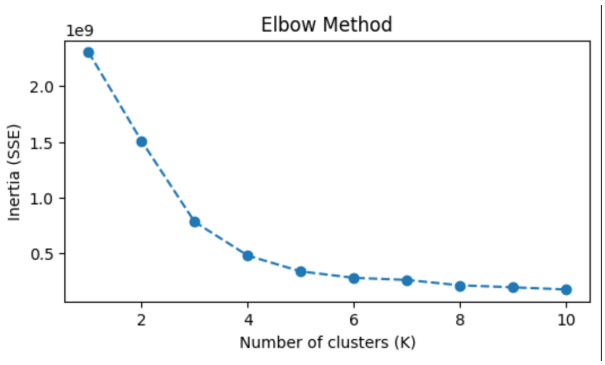

Figure 6.9: Feature-Rich Complete Case: Elbow Method.

validation metrics were computed across different clustering configurations (Table 6.6). Below is an interpretation based on three widely accepted evaluation

Method	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means (3 clusters)	0.232807	707.665318	1.851334
K-Means (4 clusters)	0.162489	557.327747	2.057643
Mean-Shift	0.629916	74.595265	0.594951
Agglomerative clustering	0.238390	828.100843	1.627926
DBSCAN	-0.321963	29.665030	1.314028
GMM	0.230904	699.893147	1.435733

Table 6.6: Feature-Rich Complete Case: Clustering Algorithms.

criteria:

- Silhouette Score: Mean-Shift achieves the highest silhouette score (0.6299), indicating strong intra-cluster cohesion and inter-cluster separation; DBSCAN produces a negative score (-0.322), reflecting poor cluster definition and possibly excessive noise; K-Means and Agglomerative clustering both yield moderate scores (0.23), suggesting fair clustering performance.

- Calinski-Harabasz Index: Agglomerative clustering has the highest score (828.10), indicating well-separated and compact clusters; K-Means (3 clusters) follows with a strong value of 707.67; Mean-Shift, with a much lower score (74.60), indicates weakly defined clusters.

- Davies-Bouldin Index (lower is better): Mean-Shift again performs best (0.595), suggesting well-separated and compact clusters; K-Means (3 clusters) and Agglomerative clustering show moderate performance; K-Means (4 clusters) (2.06) and DBSCAN (1.31) indicate poor separation and compactness.

Considering both cluster compactness and density, Agglomerative clustering emerges as the most effective method. However, for a balanced approach in terms of interpretability and robustness, K-Means with 3 clusters is also a strong candidate. To assess the consistency between clustering methods (Table 6.7), the ARI was computed between K-Means (3 clusters) and Agglomerative clustering. The ARI score of 0.484 indicates a moderate degree of similarity between the two cluster assignments. This reinforces the conclusion that both algorithms capture similar structural patterns in the data, thus enhancing confidence in the robustness of the segmentation. Further support for the clustering results comes

cluster	K-Means Count	Agglomerative Count
0	1,272	1,403
1	759	1,270
2	656	14

Table 6.7: Feature-Rich Complete Case: Cluster Distributions.

from 3D scatter plots generated using the features "point", "member class", and "recency" (Fig. 6.10). These visualizations exhibit highly similar cluster structures for both K-Means and Agglomerative clustering, suggesting that the underlying customer segmentation patterns are robust and largely independent of the algorithm used. The rationale behind selecting these three features is as follows:

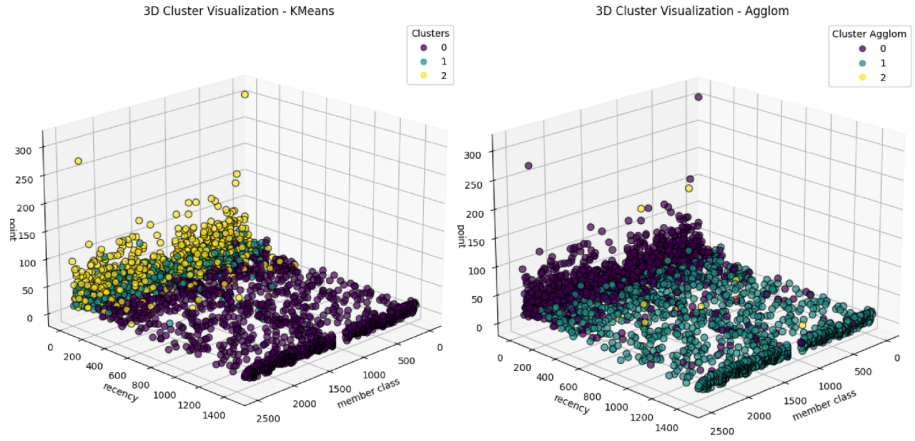

Figure 6.10: Feature-Rich Complete Case: 3D Cluster Visualization.

- "Point": Represents accumulated customer rewards or activity level, often indicative of engagement or purchasing power.
- "Member class": A high-cardinality categorical feature reflecting detailed segmentation, potentially linked to behavioral or demographic traits.
- "Recency": Measures how recently a customer made a purchase.

These features capture a balanced combination of customer behavior ("recency"), engagement ("point"), and segmentation ("member class"), making them particularly suitable for clustering.

6.2.2 Cluster Evolution Results

In this dataset, the analysis showed that *35.5% of customers changed their cluster membership* between the last two observed periods ("last consume date": 2019 - 2020 and 2020 - 2021). Such moderate degree of movement correspond to the volume of the previous case. The Fig. 6.11 provides more details about client migration between clusters. While cluster 1 has remained mostly stable, cluster 0 has seen significant redistribution, with 61.4% of its elements migrating to cluster 2. On the other hand, cluster 2 has been the most fragmented, with elements distributed across all three groups, indicating high variability in its characteristics, which may be caused by a minoritarian presence.

6.2.3 Predictive Results

The next step involved developing predictive models to anticipate whether a customer is likely to change cluster membership in future periods (0 = no

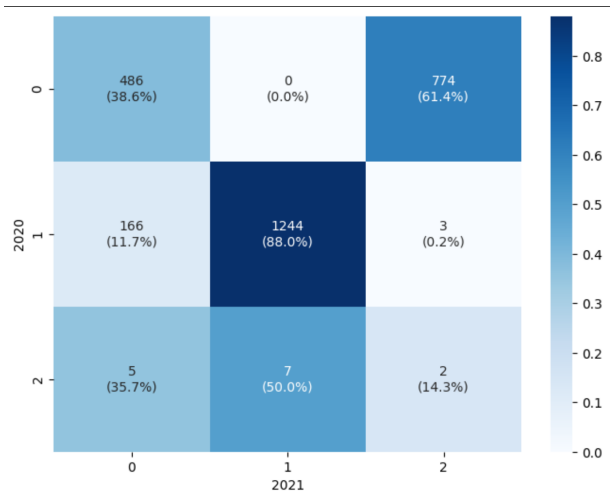

Figure 6.11: Reature-Rich Complete Case: Cluster Migration Analysis.

change, 1 = change). Two ML algorithms, LR and RF, were employed to model this classification task to identify the key drivers of customer. Comparison of the performance metrics listed in the Table 6.8

Metric, %	LR	RF
Accuracy	97	98
ROC AUC (cv)	99	99
ROC AUC	99	99
Precision	97	98
Recall	97	98

Table 6.8: Feature-Rich Complete Case: Logistic Regression vs Random Forest

The visual analysis of LR performance metrics (Fig. 6.12) reveals strong classification capabilities. The ROC Curve with an AUC of 0.9, indicate excellent discriminative performance. Similarly, the Precision-Recall Curve shows an AUC of 1.00, reflecting a perfect balance between precision and recall. This suggests that the model rarely produces false positives while maintaining high precision when identifying positive cases.

These observations are further supported by the confusion matrix (Fig. 6.13), which confirms the model's high accuracy and minimal misclassification. The learning curve for LR (Fig. 6.14) reveals important insights into the model's behavior as the dataset size increases. Initially, a sharp decline in training accuracy is observed, as the training set grows, training accuracy gradually decreases, which paradoxically suggests that the model is learning to generalize rather than memorizing. Meanwhile, test accuracy remains relatively stable, reflecting

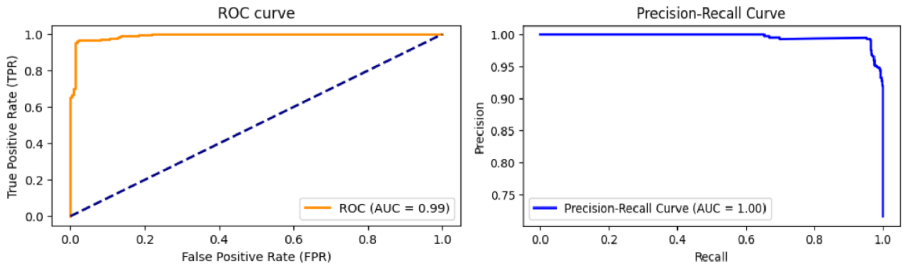

Figure 6.12: Feature-Rich Complete Case: Performance Metrics.

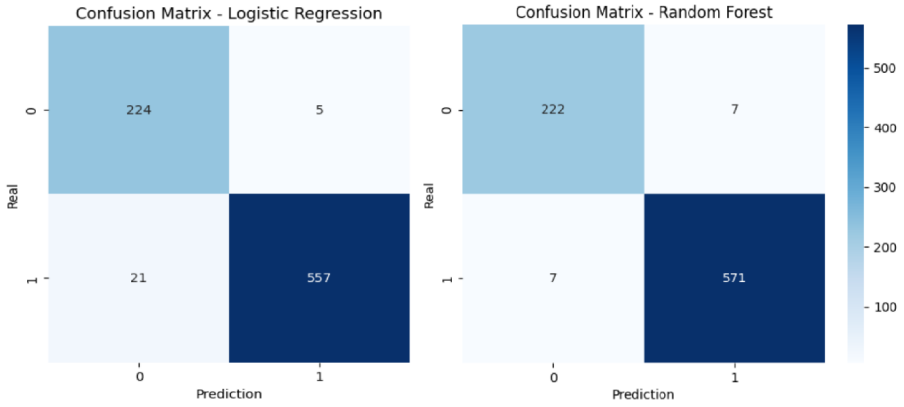

Figure 6.13: Feature-Rich Complete Case: Confusion Matrix

a consistent ability to generalize to unseen data. At larger dataset sizes, a modest upward trend in both training and test accuracy becomes apparent, implying that additional data enhances model robustness and further reduces overfitting. Overall, the learning curve suggests that the LR model benefits from increased data volume, resulting in improved stability and predictive performance.

The learning curve for the RF classifier (Fig. 6.15) highlights key aspects of its performance and generalization capabilities. The training accuracy rapidly reaches 1.000 and remains flat, indicating that the model perfectly fits the training data. In contrast, the test accuracy fluctuates slightly around 0.980, demonstrating that the model maintains strong generalization to unseen data despite its perfect training performance. Notably, increasing the number of trees beyond 100 does not significantly improve test accuracy, implying that the model has reached a performance plateau and that additional complexity does not yield further generalization gains. Overall, the curve reflects a robust and high-performing model.

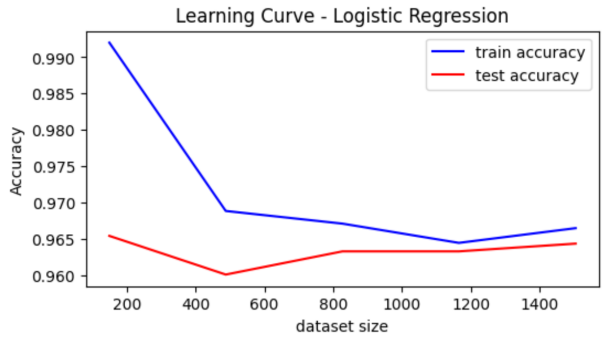

Figure 6.14: Feature-Rich Complete Case: LR Learning Curve

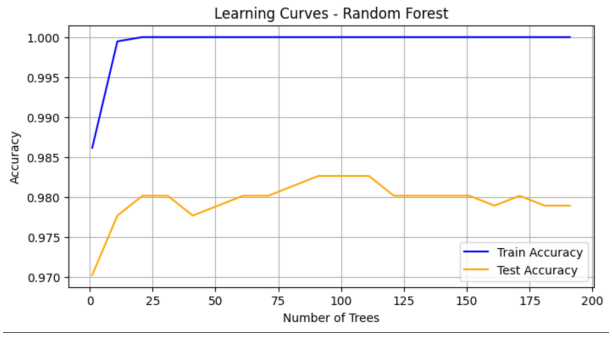

Figure 6.15: Feature-Rich Complete Case: RF Learning Curve

6.2.4 Model Interpretability Results

To gain a deeper understanding of the model's decision-making process, SHAP value analysis was employed. This is particularly valuable for ensuring transparency, accountability, and trust in predictive modeling, especially in domains where decision interpretability is critical. In the following part, SHAP is applied to RF models to explore how different features influence prediction outcomes and to identify the most impactful drivers of model behavior. The analysis reveals strong overall concordance between FR and SHAP importance rankings, with 7 out of 10 features showing identical ranks across both methods (Fig. 6.16). This alignment suggests robustness in the identification of feature relevance. The observed discrepancies, particularly for "card type", highlight methodological differences in how these techniques evaluate feature contributions. RF importance might be more sensitive to interaction effects or might overestimate the importance of categorical variables with multiple levels, which could explain the higher ranking of "card type" in the RF assessment.

Both RF importance and SHAP values identify "recency," "level," and "define class" as the three most influential features, providing strong evidence for their significance in the model. The consistent ranking of lower-importance features

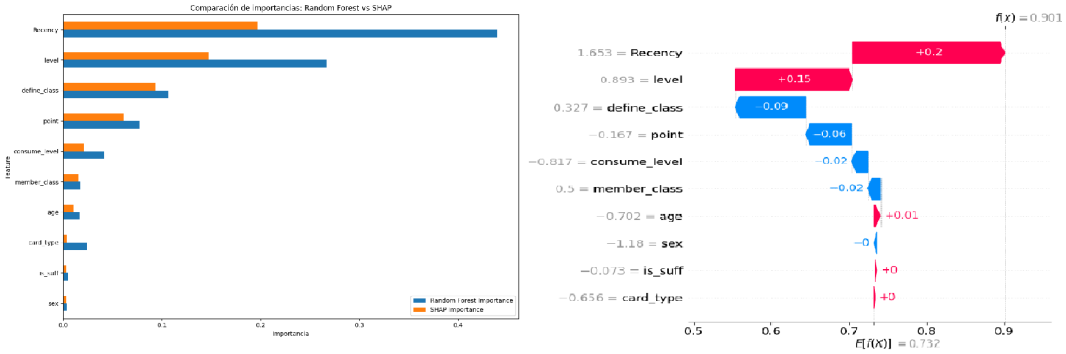

Figure 6.16: Feature-Rich Complete Case: Shap vs RF Features Importance.

like "is suff" and "sex" further reinforces the reliability of both methodologies. The moderate discrepancies observed for "member class," "age," and particularly "card type" suggest that complementary use of both techniques provides a more comprehensive understanding of feature importance. For critical applications requiring robust feature selection, features with consistent rankings across both methodologies should be prioritized, while features with discrepant rankings warrant further investigation to understand the nature of their contribution to the model.

6.3 Mode Imputation Results

The next phase of our study involves the application of mode imputation to address missing data. Unlike the listwise deletion approach, which significantly reduced the sample size, mode imputation allows for the retention of all records by filling in missing values with the mode of the corresponding feature.

As a result, the dimensionality of the dataset increased substantially, yielding a final shape of 10,000 observations across 15 features. This method assumes that data are MAR and that the mode provides a reasonable estimate for missing values without introducing major distortions. By maintaining the full dataset, mode imputation enhances statistical power and allows for more comprehensive modeling while preserving the structure of the original variables. While mode imputation does not capture the uncertainty associated with missing data as multiple imputation techniques do, it offers a pragmatic and computationally efficient solution that supports model training on larger and more representative samples. Its application here aims to strike a balance between data completeness and simplicity of implementation.

6.3.1 Segmentation Results

As in the previous cases, the optimal number of clusters was determined using the Elbow Method. The shape of the inertia curve closely resembles that of the earlier Feature-Rich Complete Case analysis (Fig. 6.17). Initially, as the number of clusters (K) increases, the inertia decreases sharply, indicating that the clustering improves as within-cluster variance is minimized. However, beyond a certain point the rate of decrease in inertia slows significantly. This inflection point suggests that adding more clusters results in diminishing returns in terms of clustering performance. In this case, the elbow appears around $K = 2$ or $K = 3$, implying that these values strike an optimal balance between model complexity and explanatory power. Therefore, 3 clusters were considered as appropriate candidates for further analysis.

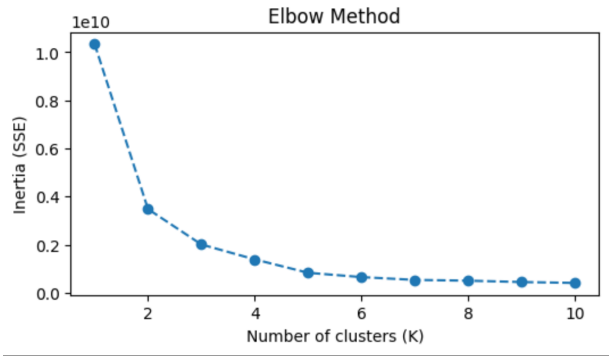

Figure 6.17: Mode Imputation: Elbow Method.

To validate the cluster quality and compare the effectiveness of different algorithms, three internal validation metrics were computed (Table 6.9). The

Method	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means (3 clusters)	0.487808	3794.915501	1.347339
K-Means (4 clusters)	0.486470	2882.656371	1.691801
Mean-Shift	0.542897	582.658198	1.050127
Agglomerative clustering	0.436540	3664.496863	1.318242
DBSCAN	0.410044	223.121669	1.3069278
GMM	0.482593	2870.447045	1.396751

Table 6.9: Mode Imputation: Clustering Algorithms.

K-Means with 3 clusters and Agglomerative clustering emerge as the most effective models, depending on whether the priority is interpretability or statistical robustness. The ARI score of 0.854 reflects a strong agreement between the two methods, indicating that both algorithms identified highly similar underlying

structures in the data (Table 6.10). This high level of concordance strengthens the reliability of the segmentation and supports the robustness of the identified clusters. Additional evidence supporting the clustering outcomes is provided by

cluster	K-Means Count	Agglomerative Count
0	1,440	1,453
1	5,843	5,616
2	2,717	2,931

Table 6.10: Mode Imputation: Cluster Distributions.

the 3D scatter plots constructed using the features "point", "member class", and "recency" (Fig. 6.18). These visual representations reveal a high degree of structural similarity between the clusters formed by both K-Means and Agglomerative clustering. This visual consistency implies that the segmentation patterns observed are stable and not heavily influenced by the specific clustering algorithm employed.

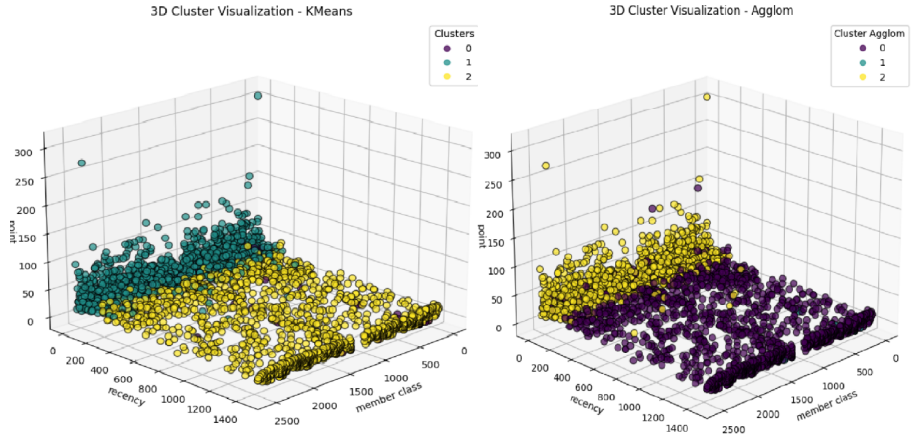

Figure 6.18: Mode Imputation: 3D Cluster Visualization.

6.3.2 Cluster Evolution Results

In this data set, analysis revealed that *99.78% of the customers shifted their cluster membership* between the two most recent time periods - an even higher transition rate than that observed in the fully feature rich data set (Fig. 6.19).

This notable degree of change reflects the curve´s behavior in time period 2020-2021 in Fig. 5.5, underscoring the importance of ongoing monitoring and analysis. In response, the subsequent phase of the study involved the development of predictive models to predict whether a customer is likely to switch clusters in future periods.

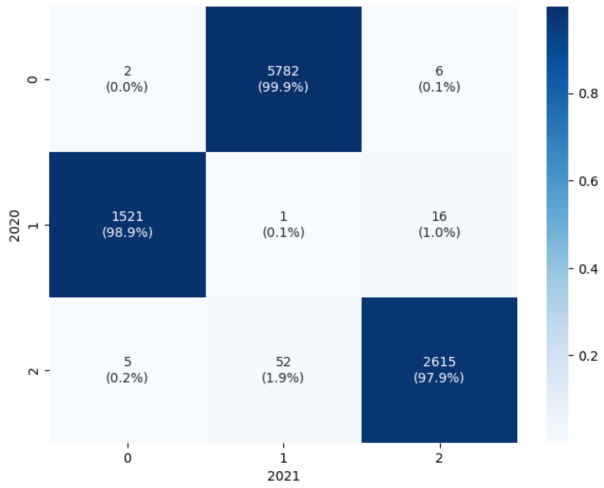

Figure 6.19: Mode imputation: Migration Clustering Analysis.

6.3.3 Predictive Results

For this classification task, LR and RF algorithms were applied, and their performance metrics are compared in Table 6.11. The graphical evaluation of LR

Metric, %	LR	RF
Accuracy	97	99
ROC AUC (cv)	98	99
ROC AUC	98	99
Precision	97	99
Recall	97	99

Table 6.11: Mode Imputation: Logistic Regression vs Random Forest

performance metrics (Fig. 6.20) highlights the model's robust classification ability. This indicates that the classifier consistently identifies positive and negative cases with high accuracy. Additionally, the Precision-Recall Curve demonstrates an ideal balance between precision and recall, implying that the model maintains a low false positive rate while effectively recognizing relevant instances. Together, these visualizations underscore the strength and reliability of the LR model in distinguishing class membership.

These findings are further validated by the confusion matrix (Fig. 6.21), which reinforces the model's strong predictive accuracy and demonstrates a low rate of classification errors. The behavior of the learning curves (Fig. 6.22) illustrates the model's performance evolution as the dataset size increases. Initially, the training accuracy is high, around 0.982, when only a small subset of data is used. This high accuracy stems from the model's ability to perfectly memo-

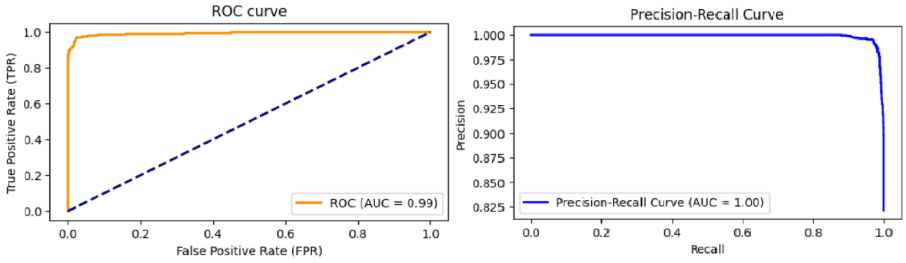

Figure 6.20: Mode Imputation: Performance Metrics.

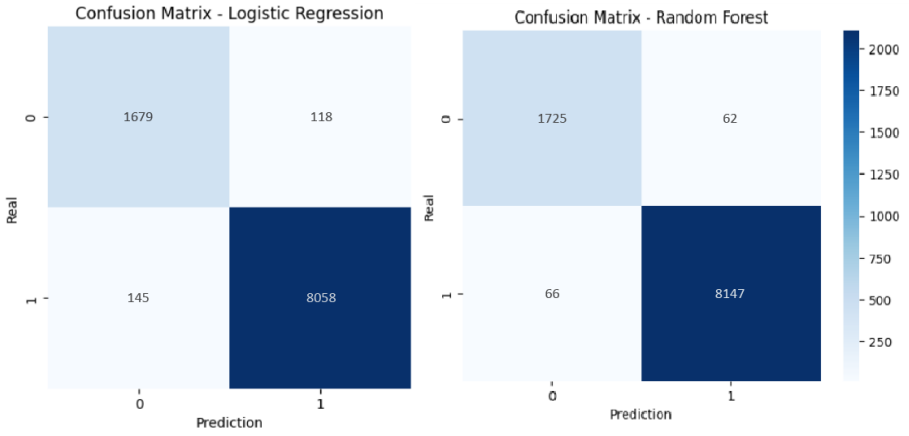

Figure 6.21: Mode imputation: Confusion Matrix.

rize the limited data, which typically results in overfitting. However, as more data is introduced, the training accuracy gradually declines, reflecting a shift toward more generalizable learning rather than memorization. In contrast, the test accuracy begins lower, approximately 0.970, indicating that the model initially struggles to generalize well from the small dataset. As the dataset size increases, the test accuracy improves steadily, suggesting that the model starts identifying more reliable and consistent patterns, enhancing its performance on unseen data.

Notably, at around 5,000 samples, both training and test accuracy curves converge toward 0.974. This convergence demonstrates that the model maintains robust generalization capabilities while minimizing overfitting, making it well-suited for predictive tasks with this dataset.

The analysis of the learning curves for the RF model (Fig. 6.23) demonstrates clear trends in model behavior as complexity increases. Training accuracy rapidly reaches 1 with only a small number of trees and remains constant thereafter. This indicates that the model perfectly fits the training data, a characteristic inherent to ensemble methods like RF, which aggregate multiple decision trees and are capable of fully capturing complex patterns within the training set.

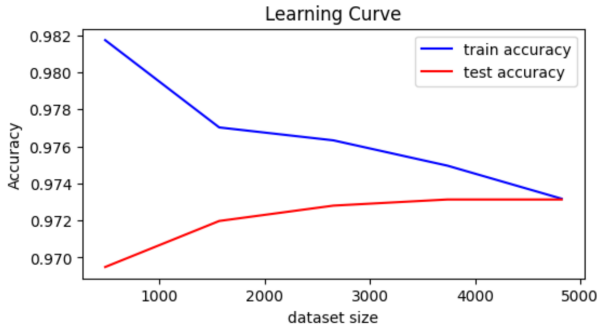

Figure 6.22: Mode Imputation: LR Learning Curve.

In contrast, the validation accuracy begins at a lower value but quickly rises,

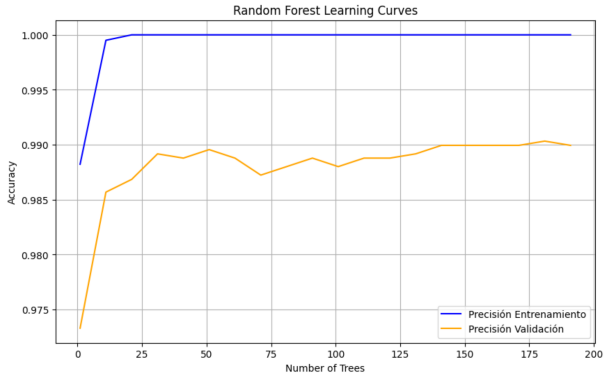

Figure 6.23: Mode Imputation: RF Learning Curve.

stabilizing around 0.990 as the number of trees increases. This improvement reflects the model's enhanced generalization capacity with growing complexity. However, once a certain number of trees is reached, the validation accuracy plateaus, suggesting that adding more trees yields diminishing returns in terms of predictive performance on unseen data.

6.3.4 Model Interpretability Results

The comparison between SHAP and RF feature importances offers complementary insights (Fig. 6.24). SHAP enhances our understanding by quantifying individual-level contributions, capturing subtleties missed by aggregate measures in RF. "Consume level" and "level" emerge as core features, consistently impacting model outcomes.

SHAP reorders the perceived relevance of several features, such as elevating the significance of "age" and "sex", while downgrading features like "define class"

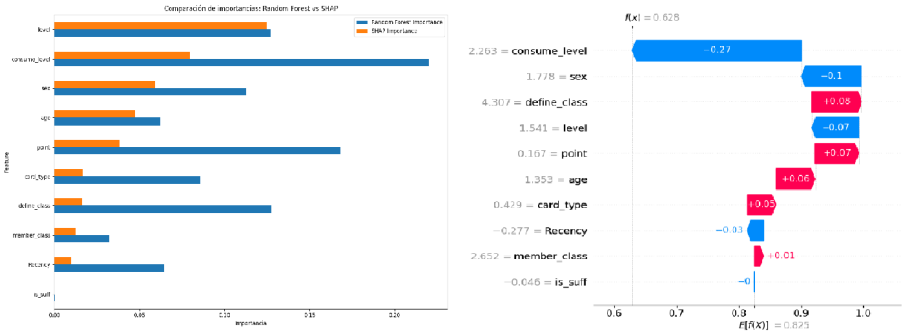

Figure 6.24: Mode Imputation: Shap vs RF Features Importance.

and "point", previously overestimated by RF importance. The alignment on low-importance features like "is suff" adds confidence to the model's interpretability.

This dual-analysis approach strengthens confidence in model transparency and suggests a refined feature prioritization strategy for both explanation and optimization purposes. It also confirms that model interpretability tools like SHAP are crucial in real-world applications, where understanding how and why predictions are made is as important as the predictions themselves.

6.4 KNN Imputation Results

In this subsection, we outline the process of applying KNN imputation, which follows a similar methodology to that of the previous subsections, with the key difference being the imputation technique employed. The dataset used includes 15 features and 10,000 records, ensuring a sufficient sample size for imputation.

To optimize the performance of KNN imputation, GridSearch was utilized to fine-tune the hyperparameters of the KNN algorithm. The hyperparameters adjusted during this step include the number of neighbors (k), distance metric, and weight function. The Table 6.12 presents a list of the hyperparameter metrics, along with the corresponding accuracy for each feature after applying KNN imputation. This table serves as a valuable reference to assess the effectiveness of KNN imputation across the different features considered in this analysis.

Feature	N neighbors	Mistance metric	Weight
"sex"	3	manhattan	distance
"card type"	11	euclidean	distance
"member class"	3	nan euclidean	uniform
"last consume time"	5	manhattan	distance

Table 6.12: KNN Imputation: Hyperparameter Metrics

Overall, the hyperparameters were selected based on the specific nature of each feature. For continuous or ordinal features, Euclidean or Manhattan distance with distance-based weighting tends to work well, while categorical features benefit from a non-Euclidean approach and uniform weighting. The combination of these choices ensures that the KNN imputation process is tailored to the inherent structure of the data, leading to more accurate and reliable imputations for missing values.

6.4.1 Segmentation Results

After imputing the missing data, we proceed with clustering analysis, starting with the application of the Elbow Method to determine the optimal number of clusters. The generated graph (Fig. 6.25) displayed a significant drop in inertia for lower values of K, followed by a clear "elbow" at K=2. This inflection point suggests that increasing the number of clusters beyond K=2 yields only marginal improvements in variance reduction, indicating that K=2 is a reasonable choice for the initial clustering configuration. It is worth noting that the shape of this curve closely resembles the pattern observed in the Complete-Case Analysis (Fig. 6.1), almost replicating it.

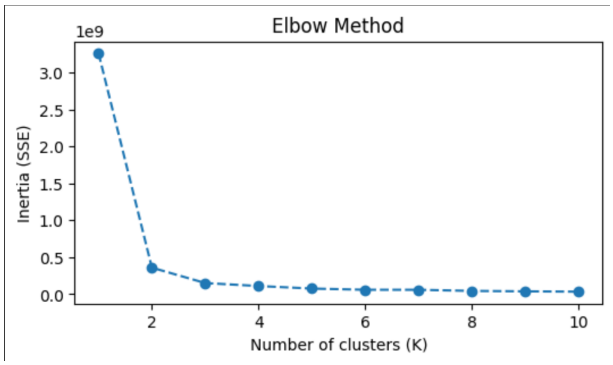

Figure 6.25: KNN Imputation: Elbow Method.

To strengthen the robustness of the analysis, internal validation metrics were calculated for different clustering configurations, as presented in the Table 6.13.

Among the evaluated models, Mean-Shift achieved the highest Silhouette Score (0.623) and the lowest Davies-Bouldin index (0.967), indicating strong cohesion and separation, though its Calinski-Harabasz score was relatively low (1031.39). In contrast, K-Means with 3 clusters showed the best balance across all metrics, with the highest Calinski-Harabasz score (4947.11), a solid Silhouette Score (0.549), and acceptable separation (1.171). Agglomerative clustering and GMM performed moderately, while DBSCAN showed weak results on

Method	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means (3 clusters)	0.512040	4548.305145	1.200486
K-Means (4 clusters)	0.549195	4947.114158	1.171252
Mean-Shift	0.549195	4947.114158	1.171252
Agglomerative clustering	0.478274	4363.383313	0.954902
DBSCAN	0.519133	237.459019	1.325992
GMM	0.576075	4014.685293	1.350657

Table 6.13: KNN Imputation: Clustering Algorithms.

all metrics. Overall, K-Means (3 clusters) appears to offer the most consistent and interpretable clustering structure for the imputed dataset.

To evaluate the consistency and alignment between clustering algorithms, a comparative analysis was performed using K-Means and Agglomerative clustering, each configured with three clusters and applied to the same complete-case dataset Table 6.14. The resulting cluster distributions from both methods were notably similar, indicating a strong structural alignment in how the data was segmented.

This high level of concordance strengthens the reliability of the segmentation and supports the robustness of the identified clusters. To formally measure

cluster	K-Means Count	Agglomerative Count
0	4,918	4,832
1	3,504	3,715
2	1,578	1,453

Table 6.14: KNN Imputation: Cluster Distributions.

the similarity of the cluster assignments, *the ARI was calculated, producing a high score of 0.924*, nearly identical to the result obtained in the Complete-Case Analysis. This high ARI value confirms a strong agreement between the two methods.

The visual resemblance (Fig. 6.26) supports the quantitative results, indicating that the underlying segmentation patterns are stable and not dependent on the specific clustering algorithm used.

6.4.2 Cluster Evolution Results

In this scenario, the analysis showed that *99.93% of customers changed clusters* between the two most recent time periods (Fig. 6.27), reflecting the dynamic presented in Fig. 5.4, 5.5. This shift suggests a significant transformation in classification or client behavior from one year to the next. The reasons behind

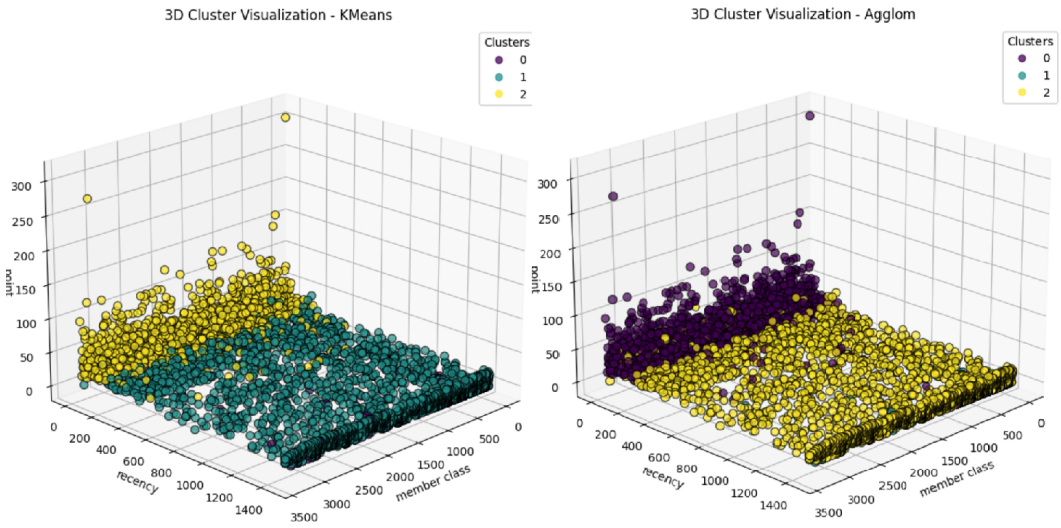

Figure 6.26: KNN Imputation: 3D Cluster Visualization.

this change could be related to evolving trends or changes in the data categorization process.

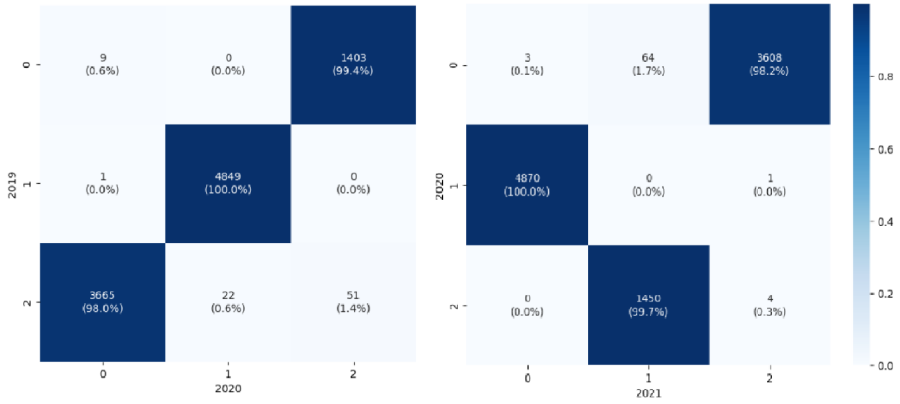

Figure 6.27: KNN Imputation: Migration Clustering Analysis.

6.4.3 Predictive Results

To further explore this, LR and RF models were developed, with their performance compared in Table 6.15. The graphical representation of LR performance (Fig. 6.28) demonstrates the model's strong classification capabilities. The smooth contours in the plots suggest gradual transitions in decision boundaries, while the model effectively differentiates between classes. High AUC scores of 0.95 and 0.94 confirm that LR is well-suited for precise classification tasks.

Metric, %	LR	RF
Accuracy	89	93
ROC AUC (cv)	96	98
ROC AUC	95	98
Precision	90	93
Recall	89	93

Table 6.15: KNN Imputation: Logistic Regression vs Random Forest

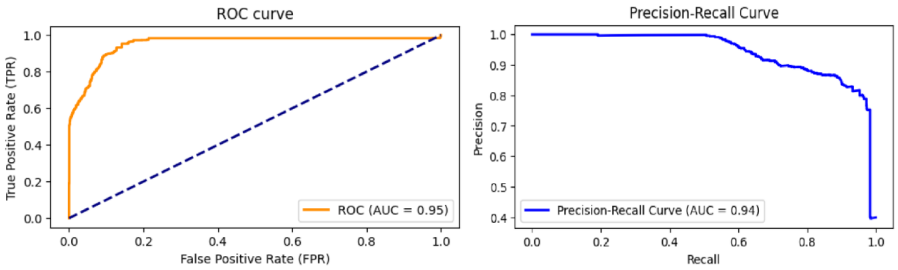

Figure 6.28: KNN Imputation: Performance Metrics.

These results are visually supported by the confusion matrix (Fig. 6.29), which shows high predictive accuracy and a low misclassification rate, reinforcing the LR model's reliability in identifying distinct customer segments.

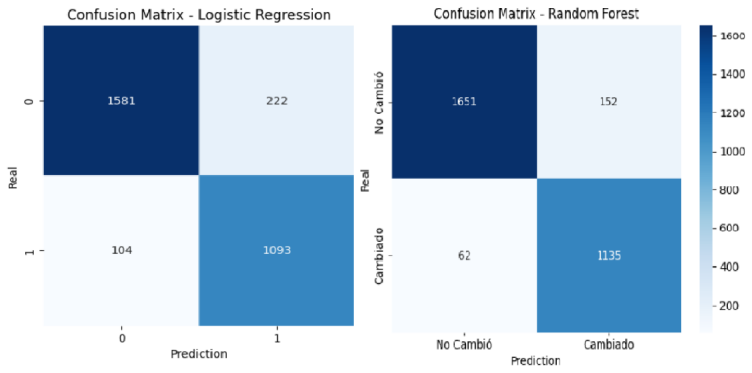

Figure 6.29: KNN Imputation: Confusion Matrix.

The learning curve of LR (Fig. 6.30) indicates that the model experiences notable performance gains with increasing data up to approximately 3,000 samples, beyond which the marginal benefits begin to decline. The convergence of training and testing accuracies between 3,000 and 5,000 observations suggests that this range represents an optimal dataset size for the model. Furthermore, the narrow accuracy interval implies consistent model performance across different configurations, with a maximum variation of only 2%. These findings

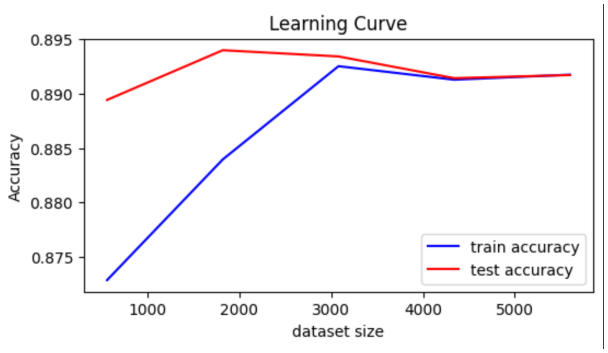

Figure 6.30: KNN Imputation: LR Learning Curve.

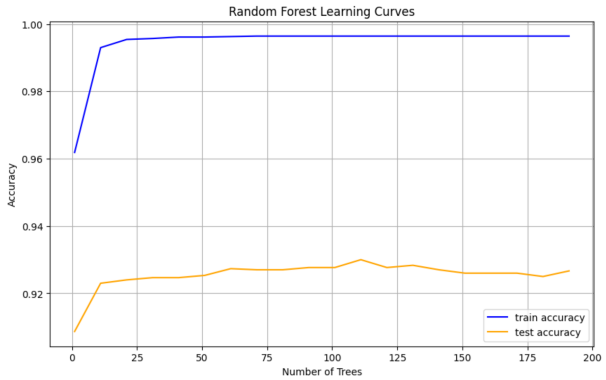

Figure 6.31: KNN Imputation: RF Learning Curve.

suggest that the model has likely reached its learning capacity, as additional data beyond this threshold does not result in substantial improvements.

The training accuracy of RF model (Fig. 6.31) remains consistently high, whereas the test accuracy is lower and exhibits instability. The widening gap between training and test performance indicates that increasing the number of trees does not enhance the model's ability to generalize. The observed fluctuations in test accuracy suggest that the model may be overfitting, capturing noise or specific patterns in the training data rather than learning generalized, robust structures.

6.4.4 Model Interpretability Results

The comparative analysis of feature importance using both RF and SHAP values (Fig. 6.32) reveals that both methods agree on the general relevance of features like "point", "card type", and "recency", notable rank differences - such as "recency" being top-ranked by SHAP but only fifth by RF - suggest that each interpretability approach captures distinct aspects of customer behavior.

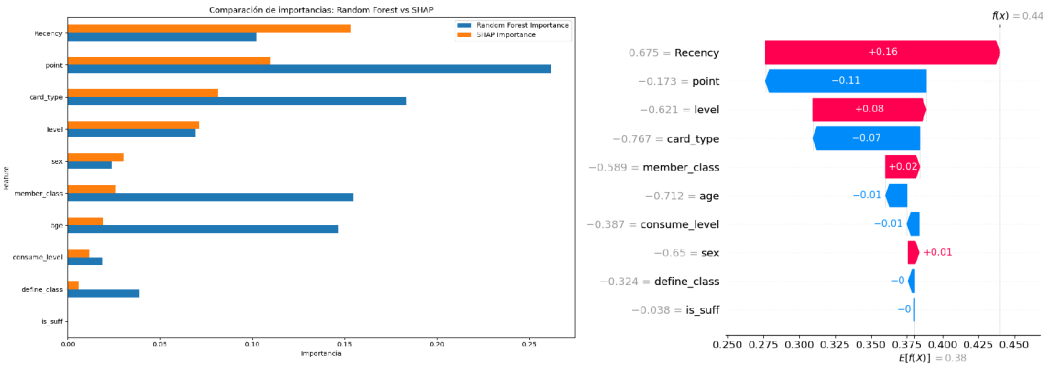

Figure 6.32: KNN Imputation: Shap vs RF Features Importance.

These variations underscore how the choice of imputation and interpretation methods can shape strategic focus. For example, if "recency" is highly influential according to SHAP, retention-focused strategies such as personalized re-engagement campaigns might be prioritized. Conversely, RF emphasizes "point" and "card type", which could guide strategies toward loyalty program design or segmentation based on card types.

Therefore, in marketing analytics, relying on a single feature ranking may lead to biased targeting or misaligned initiatives. Cross-validating the importance of the features through multiple methods ensures that marketing strategies are grounded in a more holistic and accurate understanding of the imputed data.

6.5 Discussion

The findings conclusively demonstrate that imputation decisions are not merely technical preprocessing steps but strategic choices with significant implications for marketing analytics and decision-making. Each imputation approach revealed different patterns of customer segment transitions, with KNN imputation particularly reflecting the evolution of fully completed features such as "point" and "consume level" (Table 6.16).

Scenario	ARI,%	LR, % Accuracy	RF, % Accuracy
Complete-Case Analysis	96	82	96
Feature-Rich Complete Case	48	97	98
Mode Imputation	89	97	99
KNN Imputation	92	89	93

Table 6.16: Comparative Analysis of Imputation Strategies

The lower accuracies observed in the case of KNN imputation of both ML models (89% and 93%) highlight the advantages of this method, as it better preserve the dynamic nature of customer segment transitions, which is crucial for marketing analytics applications.

Thus, despite the differences in observed transition rates, both LR and RF models maintained impressive predictive performance across all scenarios. This indicates robust generalization capabilities regardless of the imputation method employed, though the underlying patterns captured by these models varied significantly.

Furthermore, the analysis reveals a clear dependency between the number of features, the number of retained records, and the performance of both clustering and classification models. The Complete Case approach, which had the smallest dimensionality but higher sample size, demonstrated the lowest accuracy.

The comparative analysis of feature importance using both RF metrics and SHAP values (Fig. 6.33) revealed that while certain features - notably "*recency*", "*level*" *and* "*point*" - consistently demonstrated high importance across methods, their relative significance fluctuated depending on the imputation approach. This highlights the need for complementary interpretability methods to gain a comprehensive understanding of feature influence.

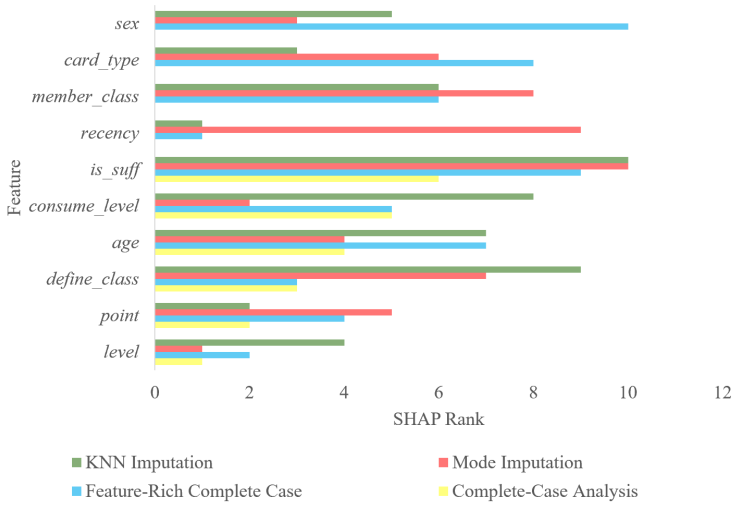

Figure 6.33: Comparative Analysis of SHAP Feature Importance Across Methods

Furthermore, we observed that the identification of customer segments remained largely consistent between clustering algorithms (K-Means and Agglomerative clustering) for each imputation scenario, as evidenced by high ARI scores (Table 6.16). This supports the robustness of the segmentation framework, suggesting that the identified customer groups represent genuine market structures rather than artifacts of the analytical process.

For marketing analytics practitioners, this research demonstrates that the choice of imputation strategy should be aligned with business objectives. If identifying stable customer segments is the goal, KNN imputation might be preferred. If highlighting potential changes in customer behavior is the priority, mode imputation could be more suitable.

6.6 Limitations

While this research provides valuable insights, several limitations should be acknowledged:

- Feature Selection Boundaries: The research restricted analysis to features with less than 60% missing values, potentially excluding variables that might have significant predictive value despite their incomplete nature.
- Limited Imputation Methods: While distinct scenarios were evaluated, other sophisticated techniques such as multiple imputation, regression imputation, or deep learning-based approaches were not included in the comparison.
- Binary Classification Focus: The predictive modeling targeted a binary outcome (change in cluster vs. no change) rather than predicting specific transitions between identified segments, which might offer more nuanced insights.

CHAPTER 7

Conclusions

7.1 Overview and Conclusions

This study examined the impact of different imputation strategies on clustering stability and classification performance within the context of marketing analytics in the pharmaceutical sector, using customer data with varying degrees of missingness and feature richness.

The results highlight the clear trade-offs between dataset richness, feature richness, and model performance. The RF classifier consistently outperformed LR in terms of classification accuracy and interpretability, confirming its suitability for high-dimensional real-world marketing datasets.

Importantly, to ensure ethical compliance and reproducibility, the dataset used in this study is publicly available and was obtained from the open-access article published in [11], which provides anonymized pharmaceutical marketing data suitable for academic research, guaranteeing transparency, reproducibility, and compliance with data use regulations.

Based on the research findings, here are *key recommendations for developing effective marketing strategies*:

- Customer Migration Management: implement early warning systems using the predictive models to identify customers likely to change segments, develop proactive retention strategies for high-value customers showing signs of downward migration, create "uplift" campaigns to encourage positive segment transitions

- Feature-Driven Personalization: focus personalization efforts on the consistently important features identified across imputation methods: "recency" (trigger time-sensitive offers for customers approaching inactivity thresholds), "level" and "point" (design graduated reward systems that incentivize progress between tiers), "define class" (align marketing communications with customer classification).

- Imputation Strategy Selection: choose imputation methods aligned with strategic goals. For conservative strategies (avoiding false positives): KNN

imputation provides more stable segmentation. For aggressive acquisition/retention (minimizing false negatives): mode imputation highlights more potential segment changes.

By implementing these recommendations, pharmacy retailers can transform data preprocessing decisions into strategic marketing advantages, ultimately enhancing customer retention, increasing customer lifetime value, and optimizing marketing return on investment.

7.2 Future Work

Several promising avenues for **future research** emerge from this study:

- Extended Imputation Methodology: Future work should expand the comparison to include more advanced imputation techniques, such as MICE (captures the uncertainty associated with missing values, preserves relationships between variables (such as the relationship between recency and consume level), generates confidence intervals for imputations, allowing sensitivity analysis).
- Multi-class Transition Prediction: Developing models that predict not just whether a customer will change segments but specifically which segment they will transition to would provide more actionable insights for targeted marketing interventions.
- Hybrid Imputation Strategies: Investigating the effectiveness of hybrid approaches that combine multiple imputation methods based on feature characteristics or missing data mechanisms could yield more robust preprocessing frameworks.
- Cross-Industry Validation: Applying the same comparative methodology across different industries would help establish the generalizability of the findings and identify sector-specific considerations for imputation strategy selection.

By addressing these future research directions, scholars and practitioners can further refine approaches to missing data in marketing analytics, ultimately enhancing the accuracy, reliability, and actionability of customer insights derived from real-world datasets.

Bibliography

[1] Little R. J. A. and Rubin D. B. *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc., 2002. (Cited on pages 6, 11 and 12.)

[2] Rabea Aschenbruck, Gero Szepannek, and Adalbert F. X. Wilhelm. Imputation strategies for clustering mixed type data with missing values. *Journal of Classification*, 40(2):2–24, 2023. (Cited on pages 1 and 3.)

[3] Rubin D. B. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. (Cited on page 11.)

[4] Rubin D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Inc., 1987. (Cited on pages 11 and 12.)

[5] Rubin D. B. An overview of multiple imputation. in proceedings of the survey research methods section of the american statistical association. *Princeton, NJ, USA: Citeseer*, pages 79–84, 1988. (Cited on page 13.)

[6] Tom Carter. Data Quality: A Pillar for AI and Machine Learning in Pharma Marketing. https://solli.global/solli-hub/data-quality-a-pillar-for-ai-and-machine-learning-in-pharma-marketing, 2024. Accessed: April 2025. (Cited on page 9.)

[7] Chen Hugh, Covert Ian C., and Lundberg Scott M. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5:590–601, 2023. (Cited on page 29.)

[8] Wang J. and Wu W. Missing data and model performance in marke- ting analytics. *Marketing Science*, 38(6), 2019. (Cited on page 5.)

[9] Yoon J. and Shin K. Data imputation using deep learning models. *Journal of Artificial Intelligence Research*, 69:223–249, 2020. (Cited on page 6.)

[10] Katherine J. Lee, John B. Carlin, and Julie A. Simpson. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the mcar/mar/mnar classification. *International Journal of Epidemiology*, 52(4):1268–1275, 2023. (Cited on page 12.)

[11] Jing Liang, Xin Zhou, Chong Yuan, and Yong Chen. Optimization of pharmacy membership management system based on big data: Sleeping member activation and awakening methods using ann modeling. *Heliyon*, 10, 2024. (Cited on pages iii, 2, 3, 5, 13 and 61.)

[12] Templ M., Kowarik A., and Hothorn T. Imputation of missing values in r. journal of statistical software. *Journal of Statistical Software*, 56(3):1–19, 2015. (Cited on page 6.)

[13] V. Anand Varsha Mamidi. Multiple imputation of missing data in marketing. *International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy*, 2020. (Cited on page 4.)

[14] Medical Affairs Professional Society (MAPS). 2024 Benchmark Report: Digital, Advanced Analytics and AI in Medical Affairs. https://medicalaffairs.org/wp-content/uploads/2024/10/MAPS-Digital-Advanced-Analytics-Artificial-Intelligence-Report-2024.pdf, 2024. Accessed: April 2025. (Cited on page 9.)

[15] McKinsey. Meeting changing consumer needs: The US retail pharmacy of the future. https://www.mckinsey.com/industries/healthcare/our-insights/meeting-changing-consumer-needs-the-us-retail-pharmacy-of-the-future, 2023. Accessed: April 2025. (Cited on page 3.)

[16] Ainsworth R. and Reinders P. Handling missing data in clustering algorithms for marketing segmentation. *Journal of Data Science in Marketing*, 25(4):200–215, 2020. (Cited on pages 7 and 12.)

[17] Carpenter James R. and Smuk Melanie. Missing data: A statistical framework for practice. *Biometrical Journal*, 63:915–947, 2021. (Cited on pages 6 and 7.)

[18] Syed Tahir Hussain Rizvi, Muhammad Yasir Latif, Muhammad Saad Amin, Achraf Jabeur Telmoudi, and Nasir Ali Shah. Analysis of machine learning based imputation of missing data. *Journal of Statistical Computation and Simulation*, 2023. (Cited on pages 2 and 3.)

[19] Tolou Shadbahr, Michael Roberts, and Jan Stanczuk. The impact of imputation quality on machine learning classifiers for datasets with missing values. *COMMUNICATIONS MEDICINE*, 3, 2023. (Cited on pages 1 and 5.)

[20] Dirk Temme and Sarah Jensen. Missing data – better “not to have them”, but what if you do? *MARKETING*, 41(4), 2019. (Cited on page 5.)

[21] van Buuren S. *Flexible Imputation of Missing Data. 2nd ed*. Chapman and Hall/CRC, 2018. (Cited on page 6.)

[22] van Buuren S. and Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. (Cited on pages 6 and 12.)

[23] Christopher Wooden. The Crucial Role of Real-Time Data Use in Pharma Marketing. https://www.iqvia.com/blogs/2024/11/the-crucial-role-of-real-time-data-use-in-pharma-marketing, 2024. Accessed: April 2025. (Cited on page 9.)

[24] Zhang X. and Liu F. The impact of missing data on the performan- ce of predictive analytics models in marketing. *Journal of Business Analytics*, 8(2):105–118, 2021. (Cited on page 5.)

Appendix

N°	Variable Name	Description	Data Type	NaN	N of Unique Values	Unique values
1	card_depart_id	ID of the card department	int64	100%	0	[nan]
2	district_name	Name of the district	float64	100%	0	[nan]
3	fate14	Fate status 14	float64	100%	0	[nan]
4	disease	Disease indicator	float64	100%	0	[nan]
5	fate8	Fate status 8	float64	100%	0	[nan]
6	status_code	Status code of the record	float64	100%	0	[nan]
7	last_consume_dp_id	Last consume department ID	float64	100%	0	[nan]
8	fate13	Fate status 13	float64	100%	0	[nan]
9	fate12	Fate status 12	float64	100%	0	[nan]
10	fate11	Fate status 11	float64	100%	0	[nan]
11	corp_tag	Corporate tag	float64	100%	0	[nan]
12	last_workwx_time	Last work time for the WX department	float64	100%	0	[nan]
13	last_sale_man	Last sales representative	float64	100%	0	[nan]
14	head_pic	Profile picture	float64	100%	0	[nan]
15	parent_card_id	ID of the parent card	float64	100%	0	[nan]
16	city_name	Name of the city	float64	100%	0	[nan]
17	delete_user_id	User ID of the person who deleted	float64	100%	0	[nan]
18	province_name	Name of the province	float64	100%	0	[nan]
19	fate10	Fate status 10	float64	100%	0	[nan]
20	fate9	Fate status 9	float64	100%	0	[nan]
21	medication_record_typ	Type of medication record	float64	100%	0	[nan]
22	account_modifytime	Time of account modification	float64	100%	0	[nan]
23	jifen_modifytime	Time of Jifen modification	float64	100%	0	[nan]
24	operator	Operator handling the record	float64	100%	0	[nan]
25	fate1	Fate status 1	float64	100%	0	[nan]
26	fate2	Fate status 2	float64	100%	0	[nan]
27	fate3	Fate status 3	float64	100%	0	[nan]
28	fate4	Fate status 4	float64	100%	0	[nan]
29	fate5	Fate status 5	float64	100%	0	[nan]
30	fate6	Fate status 6	float64	100%	0	[nan]
31	fate15	Fate status 15	float64	100%	0	[nan]
32	fate7	Fate status 7	float64	100%	0	[nan]
33	apply_card_date	Date of card application	float64	100%	0	[nan]
34	latitude	Latitude coordinate	float64	100%	0	[nan]
35	county_code	Code of the county	float64	100%	0	[nan]
36	city_code	Code of the city	float64	100%	0	[nan]
37	member_gms	Member GMS value	float64	100%	0	[nan]
38	province_code	Code of the province	float64	100%	0	[nan]
39	fate22	Fate status 22	float64	100%	0	[nan]
40	fate21	Fate status 21	float64	100%	0	[nan]
41	street_code	Code of the street	float64	100%	0	[nan]
42	longitude	Longitude coordinate	float64	100%	0	[nan]
43	fate20	Fate status 20	float64	100%	0	[nan]
44	fate19	Fate status 19	float64	100%	0	[nan]
45	fate18	Fate status 18	float64	100%	0	[nan]
46	member_date	Date of membership	float64	100%	0	[nan]
47	state	State of the record	float64	100%	1	[nan 1.]
48	last_time	Last modification time	object	100%	1	[nan '2020-03-03 10:02:41']
49	union_id	Union ID for the record	object	100%	1	[nan 'oKeA_6sUBmTA8--japNbh1UwKCeA']
50	filing_time	Filing time of the record	object	100%	14	[nan '2019-12-31 15:01:05']
51	file_update_time	Last update time of the file	object	100%	15	[nan '2019-12-31 15:04:03' '2019-10-08 14:26:58']
52	ncd_update_time	Last update time for NCD	object	100%	35	[nan '2019-12-31 15:04:36' '2021-05-19 20:49:54']
53	fate23	Fate status 23	float64	98%	132	[nan 102918, 113467, 102833, ...]

Figure 1: Dataset Description: "Data set of Member Info.xlsx"

N°	Variable Name	Description	Data Type	NaN	N of Unique Values	Unique values
54	user_id	User ID of the record	object	98%	171	[nan 'wm0Q75CgAAHnS9FLADr3AtL9k_QNJLYw' 'wm0Q75CgAAfTuQw0IrsXyYz9UIs-pdQ'...]
55	fate24	Fate status 24	float64	98%	85	[nan 270045, 270000, 270183, ...]
56	fate25	Fate status 25	object	98%	180	[nan '2021-01-29 12:18:41' '2021-06-05 19:48:44' ...]
57	nick_name	Nickname of the user	object	98%	1	[nan 'old']
58	user_card_code	User card code	float64	98%	183	nan 8,68878418e+11 7,91915847e+11 5,67953956e+11...]
59	is_workwx_reply	Indicates if there is a WX reply	float64	98%	2	[nan 1, 0]
60	mini_open_id	Mini program ID	object	98%	184	[nan 'oL4iC4s6bJCn7_W_PnjDoik4JPY0' 'oL4iC4lU2AUXTeNfqICexavTOlpg']
61	gukeid	ID for Guke	object	98%	184	[nan '00193697' '00074083' '00096289']
62	pharmacist_id	ID of the pharmacist	float64	97%	175	[nan 4,00022532e+08 4,00022269e+08]
63	tel	Telephone number	object	97%	244	[nan '183rbon5745' '134gnqb1969' '187obcr4198']
64	open_id	Open ID for the record	object	94%	588	[nan 'oKqxBs2S3ywooyFGc5YWsP8Ur4Ks' ...]
65	fate17	Fate status 17	object	90%	105	[nan '2020-02-18' '2021-11-10' '2020-06-09']
66	last_visit_time	Last visit time of the member	object	90%	1013	[nan '2020-02-22 11:43:15' '2021-11-11 22:19:47']
67	maintain_task_ids	IDs of maintenance tasks	object	87%	580	[nan ',2848,' ',2935,' ',2957,2962,3191,3396,']
68	fate16	Fate status 16	object	78%	123	[nan '2020-06-15 11:36:44' '2020-06-23 09:55:42' '2020-06-23 09:05:00']
N°	**Variable Name**	**Description**	**Data Type**	**NaN**	**N of Unique Values**	**Unique values**
69	sfz	SFZ code	object	75%	2472	['510224fardcdpq4379' nan '510223oarqqroh1425' ... ' '510213parcdqpc4427']
70	address	Address of the user	object	67%	6	[nan '同心家园' '重庆市南岸区福利社' '巴黎左岸' '荣昌安富' '江南山庄' '无']
71	balance	Account balance	float64	67%	4	[0, 5, nan 70, 15,]
72	member_class	Class of the member	object	59%	3407	[nan ',null,fsl,pfwyl,xdl,qrjdl,']
73	last_consume_time	Last consume time of the user	datetime64[ns]	58%	4214	NaT', '2020-01-22 09:18:15'
74	phone	Phone number	object	54%	4476	['136opaq7233' '135omrc0269' '157ebob6164' ...]
75	shop_id	Shop ID	float64	43%	3	[9897. nan 29933. 30917.]
76	card_type	Type of card	object	24%	8	会员卡, 和平药房铜卡, 和平药房银卡, 和平药房金卡, 和平内部员工卡, 和平药房钻石卡, 和平药房集团VIP卡,
77	define_class	Defined class	object	1%	586	['jz88stkhy294,' 'jz88stkhy294,yqdhyx216639,'...]
78	value_level	Value level	int64	0%	1	[0]
79	active_level	Active level	int64	0%	1	[1]
80	is_suff	Indicator for sufficiency	int64	0%	2	[0 1]
81	child_num	Number of children	int64	0%	2	[0 3]
82	id	Register unique identifier	int64	0%	10000	[1, 2, 3 ... 9998, 9999, 10000]
83	has_filed	Indicator for filed status	int64	0%	2	[0 1]
84	consume_level	Level of consumption	int64	0%	5	[1 2 3 5 4]
85	level_monlog	Monitoring level	object	0%	2832	[' ... 2018-04:23,2018-02:24,2018-03:24,2018-05:24..]
86	updatetime	Last update time	datetime64[ns]	0%	794	['2021-09-13 06:18:05', '2021-11-28 05:31:55'...]
87	is_check	Indicator for checking	int64	0%	2	[0 1]
88	is_ncd	Indicator for NCD	int64	0%	2	[0 1]

Figure 2: Dataset Description: "Data set of Member Info.xlsx" (continued)

N°	Variable Name	Description	Data Type	NaN	N of Unique Values	Unique values
89	level	Level of the record	int64	0%	9	[2 3 24 0 22 23 21 5 4]
90	is_sms	Indicator for SMS notifications	int64	0%	1	['yes']
91	stat	Status indicator	int64	0%	1	[1]
92	isEffective	Indicator for effectiveness	int64	0%	2	[0 1]
93	daoqi_flag	Flag for expiration	int64	0%	1	['yes']
94	yongyao_flag	Flag for medication usage	int64	0%	1	['yes']
95	jiance_flag	Flag for monitoring	object	0%	1	['yes']
96	createdate	Date of record creation	object	0%	3287	['2016-08-15 09:25:32' '2016-09-02 09:03:17'...]
97	depart_id	Department ID	int64	0%	274	[270297 270155 270117 270114 ...]
98	point	Points accrued	float64	0%	1409	['510224fardcdpq4379' nan '510223oarqqroh1425' ... '510122parcceoe3215']
99	birthday	Birthday of the user	object	0%	3808	'510202narqdgod2622' '510213parcdqpc4427']
100	sex	Gender of the user	object	0%	3	[男, 女, 无]
101	password	User password	int64	0%	1	[123456]
102	assign	Assignment status	int64	0%	1	[0]
103	amount	Amount of transaction	int64	0%	1	[0]
104	invalid_type	Type of invalid record	int64	0%	2	[0 2]
105	consume_times	Number of times consumed	int64	0%	1	[0]
106	a7	Variable a7	float64	0%	1649	[0, -279,58 24, ... -1835,8 142,4 112,5]
107	card_id	ID of the card	object	0%	10000	['；001936560？' '；003208730？' '；004053650？' ... '91006011' '00763764']
108	a5	Variable a5	float64	0%	1534	[0, 9,5 -59,2 ... 333,55 154,4 47,5]
109	a4	Variable a4	float64	0%	1232	[0, 9,5 48, ... 110,12 88,4 92,7]
110	a3	Variable a3	float64	0%	985	[0,00000e+00 3,65000e+01 1,35000e+01 ...]
111	a2	Variable a2	float64	0%	607	[0,00000e+00 3,65000e+01 1,35000e+01 7, ...]
112	a1	Variable a1	float64	0%	146	[0,00000e+00 3,65000e+01 1,35000e+01 7,...]
113	t7	Variable t7	int64	0%	43	[0 -2 1 3 -3 5 -1 4 -5 ...]
114	t6	Variable t6	int64	0%	44	[0 -6 3 1 -2 5 18 ...]
115	sum_enddate	End date of sum	object	0%	738	['2021-09-13' '2021-11-28 05:31:55'
116	t5	Variable t5	int64	0%	39	[0 1 -2 4 3 2 -1 5 -3 16 13 12 -4 7 9 15 -6 6 8 -5 20 10 21 11 24 19 31 14 -7 23 18 -8 36 -9 39 29 17 26 27]
117	t4	Variable t4	int64	0%	36	[0 1 2 4 3 5 6 13 14 8 16 12 9 7 10 29 26 15 19 18 11 22 35 20 17 24 21 27 25 56 37 45 33 31 23 50]
118	t3	Variable t3	int64	0%	26	[0 2 3 1 6 7 8 4 13 5 10 21 17 9 15 11 14 20 18 19 16 31 26 12 22 27]
119	t2	Variable t2	int64	0%	13	[0 2 1 6 5 3 4 10 9 8 7 14 11]
120	t1	Variable t1	int64	0%	6	[0 1 2 3 4 5]
121	last_sendsms_time	Last SMS send time	datetime64[ns]	0%	1823	['1900-01-02 00:00:00', '2020-04-16 09:40:09']
122	is_get_card		int64	0%	1	[0]
123	a6	Variable a6	float64	0%	1550	[0, -15,61 -368,28 ... 333,55 -6,5 -38,8]

Figure 3: Dataset Description: "Data set of Member Info.xlsx" (continued)