National College of Ireland

Optimising Direct Marketing Through Data-Driven Analytics and Predictive Models

MSc Research Project
Artificial Intelligence for Business

Karla Priscila Vale de Sousa
Student ID: 23152516

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Karla Priscila Vale de Sousa
Student ID:	X2315254
Programme:	MSc Artificial Intelligence for Business **Year:** 2023
Module:	MSc Research Project
Supervisor:	Victor Del Rosal
Submission Due Date:	16/09/2024
Project Title:	Optimising Direct Marketing through Data-Driven Analytics and Predictive Models
Word Count:	6.282 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Karla Priscila Vale de Sousa

Date: 16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimising Direct Marketing through Data-Driven Analytics and Predictive Models

Karla Priscila Vale de Sousa
23152516

Abstract

Direct marketing campaigns are essential for preserving and growing market leadership in today's demanding environment. This research focuses on enhancing the effectiveness of these campaigns through data-driven analytics, employing advanced methodologies such as exploratory data analysis (EDA), customer segmentation via K-means clustering, and predictive classification models including logistic regression, decision trees, and K-Nearest Neighbours (KNN). The study aims to optimise marketing campaigns by identifying profitable customer segments and accurately predicting customer responses. The research employs the CRISP-DM methodology to methodically address business objectives, prepare data, build models, and evaluate their performance using a dataset from iFood, Brazil's leading food delivery service. The findings offer practical insights that help direct marketing campaigns, improve client interaction, reduce costs, and reverse profit declines. Beyond resolving business challenges, this study contributes to the theoretical knowledge of data-driven marketing strategies and provides insightful information for both academic research and real-world implementations in the industry.

Keywords: Direct Marketing, Data-Driven Analytics, Exploratory Data Analysis, Customer Segmentation, Predictive Classification Models.

1 Introduction

In modern industries, direct marketing initiatives are essential for maintaining and expanding market leadership. According to Sagala and Permai (2021), the objective of these campaigns is to generate specific responses from targeted customer groups, and success mostly depends on the quality of prospect data. By leveraging data mining techniques, companies can better identify and segment customers, which enables more specialised and successful marketing campaigns.

Technological advancements have driven a surge in data volume, which has led marketers to optimise their strategies by utilising machine learning and predictive analytics in conjunction with substantial customer behaviour data. These rich data sources and developments in distributed data processing allow for accurate customer behaviour predictions, including expected actions and responses to marketing campaigns. These kinds of data-driven analytics greatly improve the ability to predict specific consumer reactions to marketing initiatives.

Efficient tools and analyses are essential for managing big volumes of data and ensuring business success in today's data-rich environment. Marketing teams frequently create several campaigns with different goals, therefore allocating resources carefully is necessary to minimise conflicts and maximise value (Lu and Boutilier, 2014). As stressed by Olbrich and Lindenbeck (2016), concentrating on profitable clients and prospects is essential for reducing costs and avoiding unnecessary losses. Focusing on the appropriate target can also reduce the

possibility of upsetting non-candidates with irrelevant advertising, leading to better customer satisfaction and building stronger relationships.

The purpose of this study is to improve campaign success using data-driven analytics to address challenges in the direct marketing industry. The research focuses on using predictive models to increase campaign profitability and identify customer characteristics likely to lead to purchases. The central research question is: "How can integrating data-driven analytics improve the efficacy of direct marketing campaigns by leveraging advanced techniques such as customer segmentation and predictive classification models?" Given the significant role of marketing in a company's growth, investigating the efficiency of data-driven analytics in optimising direct marketing campaigns is essential.

The research employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which provides a structured and flexible approach to aligning data mining with business objectives (Schroer *et al.*, 2021). This methodology is chosen for its emphasis on business and data understanding, which is essential for optimising direct marketing campaigns. Through exploratory data analysis (EDA), consumer segmentation using K-means clustering, and predictive classification models including logistic regression, decision trees, and K-Nearest Neighbours (KNN), the study aims to improve campaign efficacy.

This project is organised as follows: introduction, literature review, research methodology, design specification, implementation, evaluation and results analysis, conclusion and discussion. The introduction establishes the context for the study problem and questions by outlining the significance of data analytics and direct marketing. The literature review section reviews existing literature on data-driven marketing, customer segmentation, and predictive modelling. The research methodology details the CRISP-DM approach, explaining its relevance to the project. The design specification outlines the project's framework, including selected algorithms and evaluation criteria. The implementation section describes the process of data preparation, model building, and testing. The evaluation and results analysis section presents and discusses the model results and their effectiveness in improving marketing campaigns. Finally, the conclusion and discussion summarise key findings, discuss implications, and suggest areas for future research, contributing to the field of data-driven marketing.

2 Literature Review

Data is essential in predicting customer behaviour and responses to marketing campaigns. Lu and Boutilier (2014) emphasise the importance of predictive analytics, machine learning, and distributed data processing in achieving accurate predictions. Efficient marketing optimisation is important to maximise insights by strategically allocating resources to targeted marketing activities. Johnson *et al.* (2019) highlight the importance of implementing a data-driven approach within an organisation, asserting that data-driven decision-making is essential for enhancing marketing strategies. However, the practical challenges of incorporating new technologies into existing workflows are not adequately covered by the authors.

In data-driven marketing, decisions are based on the analysis of customer data to determine needs, preferences, and behaviour. According to Ali (2023), data insights are essential for enhancing performance through message customisation in marketing. Hossain *et al.* (2023)

further support this by emphasising the significance of sustained customer analytics in maintaining a competitive edge. These studies, however, frequently ignore the challenges associated with sustaining and growing analytics skills in large organisations.

Furthermore, Pour and Emami (2023) emphasise the necessity of an integrated methodology for strategic planning in data-driven marketing, proposing a framework that includes strategic positioning, contextualising strategies, and performance management. This approach aligns with Srikasem *et al.* (2022), who noted that data-driven technologies improve interactive engagement with prospects and facilitate customer-centric business operations. Despite their thorough methodologies, both studies could benefit from more empirical evidence demonstrating the long-term impacts of their proposed frameworks.

2.1 Customer Segmentation

Customer segmentation is a critical strategy in various industries, involving the use of algorithms like K-means clustering to analyse and categorise customer groups based on different characteristics and behaviours. Bose and Chen (2009) emphasise the importance of customer profiling in direct marketing, identifying methods such as RFM (Recency, Frequency and Monetary) profiling, latent class analysis, and K-means clustering as essential for predicting future purchases. Anitha and Patil (2022) applied the RFM model using the K-means algorithm to identify potential customers, finding that segmentation based on purchasing patterns can result in effective strategies. Kasem *et al.* (2024) highlighted the importance of RFM analysis and boosting algorithms in customer profiling and sales prediction, identifying high-value customers and customising marketing initiatives to their preferences.

In order to achieve more precise consumer segmentation, Christy *et al.* (2021) expanded the traditional RFM analysis paradigm by incorporating algorithms like K-means and RM K-means. Qu (2022) explored unsupervised learning techniques like Self-Organizing Map (SOM) and K-means, concentrating on how clustering interpretation improves behavioural analysis. Venkatesan *et al.* (2021) discussed the importance of segmentation in marketing analytics based on key characteristics, which helps business to tailor their marketing strategies and maximise returns on marketing investments, optimising resource allocation by targeting the right customers.

Overall, customer segmentation and profiling, encompassing customer clustering and pattern recognition, are fundamental elements of contemporary marketing approaches. They empower businesses to develop customised and individualised campaigns for distinct customer segments, allowing them to create targeted and personalised marketing strategies that align with the unique needs and preferences of each customer cluster. By utilising advanced techniques such as K-means clustering and RFM analysis, marketers can effectively segment customers, identify high-value prospects, and design customised marketing initiatives that align with specific customer preferences and behaviours.

2.2 Predictive Classification Models

The dataset characteristics have a significant impact on how well data mining methods perform. In classification tasks, it is standard practice to test several algorithms, including

decision trees, neural networks, and logistic regression, to determine which one delivers the best outcomes (Olson and Chae, 2012). Kara *et al.* (2011) investigate predictive modelling, suggesting that classification models, such as artificial neural networks and Support Vector Machines (SVM), perform better than level estimation models, particularly in forecasting trends, which emphasises the potential of predictive models in guiding marketing campaigns.

Apampa (2016) analysed the effectiveness of various classification algorithms in predicting customer responses, finding that a balanced dataset improves performance metrics and key features like past campaign responses influencing campaign success. Similarly, Sagala and Permai (2021) utilised the SVM to enhance the effectiveness of direct marketing by predicting customer responses, employing CRISP-DM methodology, for data quality and model optimisation, through rigorous cleaning and transformation processes and advanced techniques like random oversampling, one-hot encoding and PCA.

In another study, Chate (2022) explored RFM modelling and machine learning techniques like Random Forest and AdaBoost to predict customer reviews, highlighting the importance of integrating machine learning for predictive analytics. Additionally, Zaki *et al.* (2024) examined how predictive analytics and machine learning can improve bank marketing strategies, emphasising the importance of EDA in model development. They used models like Logistic Regression, Random Forest and SVM. The effectiveness of these models are evaluated through accuracy, precision, and F1 score, finding that data-driven insights boost conversion rates.

In conclusion, using predictive models like logistic regression, decision trees, and KNN, as well as techniques like EDA and K-means clustering, to combine data analytics with direct marketing greatly improves campaign effectiveness. Even though significant progress has been made, there are still challenges with scaling these models for large datasets and improving the interpretability of complex models without sacrificing performance. Exploring hybrid models that combine various techniques could address these challenges. The findings of this research aim to address challenges of customer engagement, market leadership, and profitability, establishing best practices and strategic frameworks for more effective direct marketing campaigns.

3 Research Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology provides a comprehensive strategy for the proposed solution to optimise direct marketing effectiveness through data-driven analytics, advanced techniques, and predictive models. Its emphasis on aligning the data mining process with business objectives and its iterative nature, which allows for continuous refinement based on feedback and evolving business needs, makes CRISP-DM particularly effective (Schroer *et al.*, 2021). This business-oriented approach ensures that the final models and insights are directly applicable and beneficial for marketing strategies.

This methodology comprises six phases (Figure 1): business understanding, data understanding, data preparation, modelling, evaluation, and deployment. By following the CRISP-DM framework, the goal is to systematically navigate through each stage to build robust predictive models that identify potential customers for targeted marketing campaigns, maximising profitability and improving customer engagement.

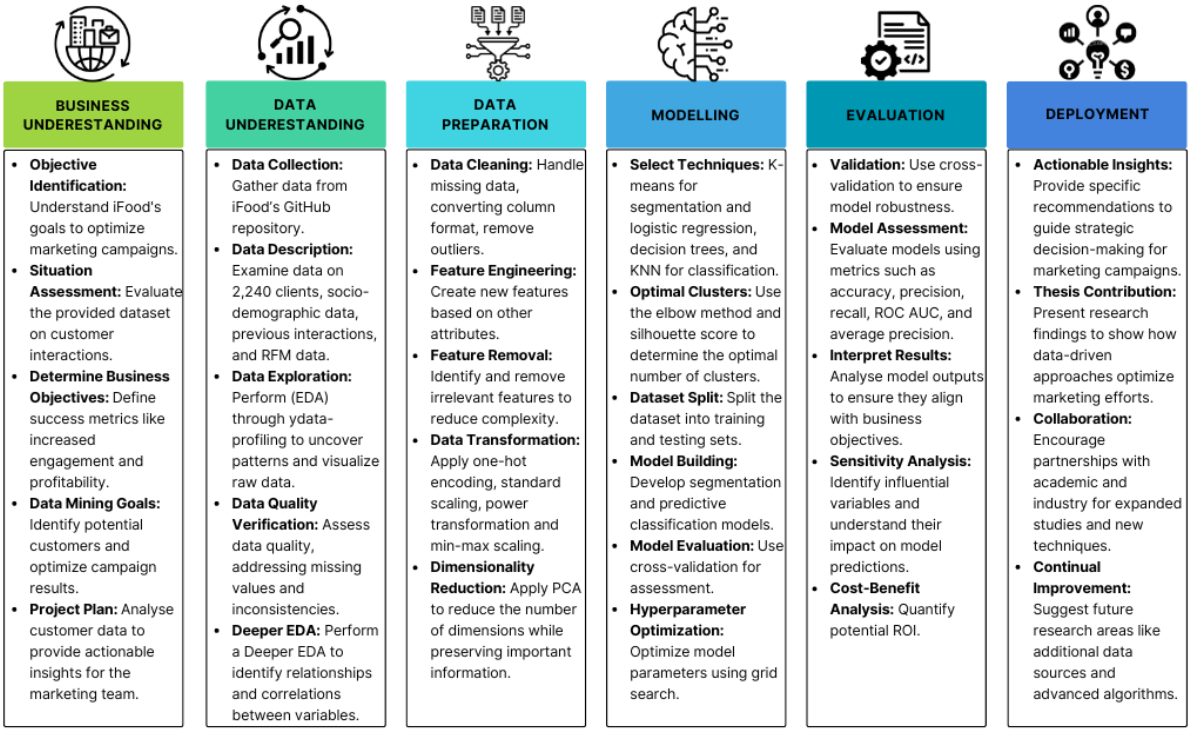

Figure 1: CRISP-DM Methodology Adapted for the Research

3.1 Business Understanding

The Business Understanding phase is crucial as it involves comprehending iFood's objectives, opportunities, and goals through a data-centric lens. For this project, iFood, the leading food delivery app in Brazil, has provided a sample dataset from a real selection process for hiring Data Analysts for the iFood Brain team, made available via their GitHub repository (Oliveira, 2020). The iFood Brain team focuses on AI and Data, managing billions of data points generated monthly within the app. The primary objective is to leverage this data to optimise marketing campaigns, increasing customer engagement and profitability. The dataset includes metadata of customer interactions with iFood campaigns, to understand customer responses and evaluate campaign effectiveness.

The project's success will be measured by how well it can increase customer retention rates, conversion rates, and overall sales growth, as only 15% of the customers responded positively in the last campaign (target). The objective of data mining is to identify potential customers likely to respond positively to marketing campaigns and optimise campaign results by predicting customer responses. The structured approach includes data collection, preprocessing, exploratory data analysis, model building, and evaluation. The insights gained can help the marketing team tailor their strategies to different customer profiles, boosting iFood's profitability.

3.2 Data Understanding

After the establishment of business objectives, the next step is to gather and become familiar with the initial data. This involves collecting data from iFood's GitHub repository (Oliveira, 2020), which includes metadata on 2,240 clients covering socio-demographic information, previous interactions with the platform and campaigns, and RFM data in 29 columns. This pilot campaign dataset is essential for understanding customer purchasing patterns, estimating product purchase probabilities, and improving accuracy in customer segmentation and campaign targeting. Performing exploratory data analysis using tools such as ydata-profiling[1] helps uncover patterns, visualise raw data, and provide an initial assessment of the dataset. Appendix 1 shows an overview of the dataset variables, detailing the attributes and their description.

Assessing data quality is essential to identify issues such as missing values, outliers, and inconsistencies to ensure reliability. This phase includes a deeper EDA to identify relationships and correlations between variables, offering a comprehensive overview of the data structure and characteristics. Visualising the data through charts, graphs, and other visual aids can uncover trends and anomalies not immediately apparent from raw data. Understanding the distribution of key variables and identifying potential correlations between different attributes provide valuable insights into the underlying patterns, setting the foundation for subsequent modelling and analysis.

3.3 Data Preparation

The Data Preparation phase is crucial for transforming and constructing variables to facilitate optimal model creation. Some tasks include data cleaning to handle missing data through imputation, converting column formats, and removing outliers, ensuring consistency and reliability. Feature engineering creates new features from existing attributes to enhance model performance, while irrelevant features are removed to reduce complexity.

Data transformation techniques such as one-hot encoding, standard scaling, power transformation, and min-max scaling are applied to standardise the data. For clustering, dimensionality reduction through Principal Component Analysis (PCA) simplifies the dataset by reducing the number of dimensions while preserving essential information. These steps ensure a well-prepared dataset, balancing comprehensive data representation with practical considerations of model performance and robustness, setting a solid foundation for building accurate, generalisable models that effectively predict customer behaviour and optimise marketing strategies.

3.4 Modelling

The Modelling phase involves creating models to predict or analyse data based on project objectives. Initially, a customer segmentation model is developed using the K-means algorithm

[1] YData Profiling: https://docs.profiling.ydata.ai/latest/

to categorise customers into distinct segments based on behaviour, preferences, and other relevant attributes. The elbow method and silhouette score are utilised to determine the optimal number of clusters, ensuring effective segmentation. This process helps identify target customer groups and refines the clustering model.

For predictive classification, models are developed to meet specific business objectives using techniques such as logistic regression, decision trees, and KNN, tailored to the classification tasks and data characteristics. The dataset is split into training and testing sets for model development and validation. A pipeline is created, incorporating Random Under-Sampling to address class imbalance. Cross-validation techniques are applied to evaluate model performance, returning scores for assessment. After selecting the best-performing model, hyperparameter optimisation is conducted using grid search to fine-tune the model parameters, ensuring optimal performance and robustness.

3.5 Evaluation

During the Evaluation phase, the primary focus is on assessing the performance, reliability, and effectiveness of the models to ensure they meet project objectives. Both the clustering segmentation and predictive classification models undergo a comprehensive evaluation. The quality of clusters generated by the segmentation model is assessed using the elbow method and silhouette score. For predictive classification models, metrics such as accuracy, precision, recall, ROC AUC, and average precision are employed. Cross-validation techniques are utilised to validate the models, ensuring they are robust and can generalise well to new, unseen data.

A thorough analysis of the model outputs is conducted to ensure alignment with business objectives. This includes interpreting the results and understanding their implications. Sensitivity analysis is performed to identify the most influential variables and understand their impact on model predictions. Additionally, a cost-benefit analysis helps quantify the potential return on investment (ROI) of applying the model. By meticulously evaluating the models, the project ensures that the developed solutions are effective and actionable, enhancing iFood's marketing strategies.

3.6 Deployment

Actionable insights show how data-driven approaches optimise marketing campaigns by offering specific recommendations based on research findings to drive strategic decision-making in marketing initiatives. The thesis contribution highlights the significant impact of these approaches, which is supported by data and analysis. Collaboration between academic and industry partners can lead to the expansion of research studies and develop innovative techniques, resulting in a more comprehensive understanding of marketing dynamics.

Following the CRISP-DM methodology, this research aims to establish a structured and iterative process that aligns with business goals and adapts to feedback and evolving business needs. The objective is to enhance the effectiveness of direct marketing campaigns through

advanced data-driven techniques, addressing current business challenges and providing a robust framework for continuous improvement and innovation in marketing strategies.

4 Design Specification

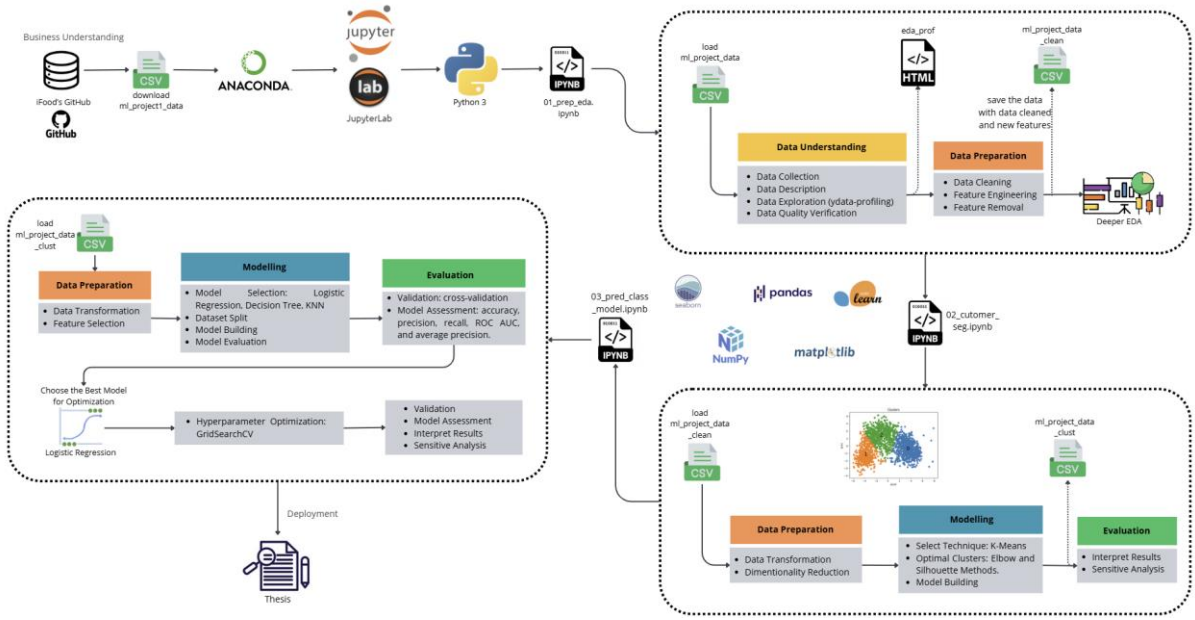

Figure 2: Design Specification of Implementation

The research project provides an integrated platform for data analysis, visualisation, and model development by utilising Jupyter Notebook and Anaconda. Python is the primary programming language due to its versatility and extensive support for data science applications. Important libraries used in the project include Seaborn, Matplotlib, NumPy, Scikit-learn, Pandas, and NumPy.

Pandas is used for data manipulation and analysis, providing the data structures and operations needed to work with numerical data frames and time series and scientific numerical computations are performed with NumPy. Clustering, classification, and regression models construction are built using Scikit-learn, which also facilitates data mining and machine learning algorithms. Matplotlib is used for data visualisation, creating static, animated, and interactive visualisations and Seaborn improves statistical visualisations, providing a sophisticated interface for drawing attractive and informative statistical graphics.

Throughout the entire study process, ethical considerations are crucial. Anonymised data is used to maintain confidentiality and adhere to ethical standards. This study addresses privacy concerns and promotes appropriate data management methods by using an open-source database and focussing on aggregate patterns and behaviours rather than individual characteristics Additionally, in order to ensure fairness and equity, the research also attempts to detect and mitigate biases in predictive models. This approach guarantees that the decision-making processes derived from the models are fair and do not disproportionately affect any specific group.

5 Implementation

The implementation phase produced key outputs, including transformed data, developed code, models, and actionable insights. Raw data underwent extensive preprocessing, resulting in a clean, structured dataset ready for analysis. Python scripts were developed in three different notebooks for exploratory data analysis, clustering, and predictive modelling, with the code modularised for reusability and ease of maintenance.

The developed models include a customer segmentation model and predictive classification models. A K-means clustering model segmented customers based on behaviour, preferences, and socio-demographic attributes, effectively grouping them into distinct segments. Classification models, including logistic regression, decision tree, and K-nearest neighbours (KNN), were developed to predict customer behaviour and evaluated using some metrics to ensure their effectiveness.

By utilising these tools and methodologies, the implementation phase successfully transformed raw data into actionable insights, developed robust predictive models, and provided a framework for optimising direct marketing campaigns. The outputs produced during this phase are integral to the overall success of the research and demonstrate the practical application of data-driven analytics in enhancing marketing strategies.

5.1 EDA

Exploratory Data Analysis (EDA) is the process of examining and visualising datasets to summarise their main characteristics, often using statistical graphics and other data visualisation methods. EDA can help analysts to understand patterns, detect anomalies, test hypotheses, and check assumptions. This initial analysis is crucial as it provides a foundation for further data modelling, ensuring that any subsequent analyses or models are built on an accurate and insightful understanding of the data, helping in making informed decisions, identifying potential areas for further investigation, and ensuring the integrity and quality of the data being analysed.

To ensure good organisation and understanding of the data, the case files were structured, and a comprehensive data dictionary with detailed column descriptions was created. Initial analyses of the dataset highlighted key patterns and potential issues, and ydata_profiling was used for an initial exploratory data analysis. This allowed for the generation of a detailed report, from which insights for data cleaning and transformation were taken. Null values were minimal and removed. Key columns were transformed, some data in columns were simplified, others were combined, and several new features were introduced for comprehensive EDA. Outliers in key columns were identified and addressed. Unnecessary columns were removed, and visualisations like pairplots and boxplots were created to analyse relationships between various customer attributes. The cleaned dataset was then saved.

5.2 Customer Segmentation: K-Means Clustering

Clustering is a machine learning technique used to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other

groups. In a different notebook, the first part of clustering focused on preprocessing the data according to the guidelines in the Scikit-learn documentation (Scikit-learn, 2024). Preprocessing included applying scaling techniques such as standard scaler for normally distributed columns, power transformer for columns with different distributions, and min-max scaler for other columns. Categorical variables were encoded using one-hot encoding. These preprocessing steps were essential to standardise the dataset and improve the performance of clustering algorithms. Once preprocessing was complete, the dataset expanded to 64 columns. The new features were visualised to ensure the transformations were correctly applied.

The optimal number of clusters was determined using the elbow and silhouette methods, and 3 clusters were chosen to an ideal balance between both methods. A pipeline was created to automate the preprocessing and K-means clustering steps, maintaining a consistent workflow and ensuring reproducibility. PCA with 2 dimensions was performed to further refine the clustering process, reducing dimensionality and enhancing interpretability and efficiency. After assigning clusters to each customer, key columns were selected for further analysis, generating pairplots and boxplots. A detailed comparison of the clusters was conducted to understand their characteristics, and a new file was saved with the cluster column added.

5.3 Predictive Classification Models

During the classification phase, a new notebook was created, and the updated dataset with 35 columns was loaded. The columns used for data preprocessing were selected based on the outcomes of the clustering phase. The dataset was divided into train and test sets (80-20). Three predictive classification models were selected: Decision Tree, Logistic Regression, and K-Nearest Neighbours (KNN) chosen for their simplicity and explainability. The objective was to compare these models and identify the one with the best performance. A Dummy Classifier with the strategy="stratified" was employed as a baseline, making predictions that ignore input features, and serving as a simple reference point to compare the performance of more complex classifiers.

A function was built to perform preprocessing, pipeline and cross-validation, returning score metrics. Given the dataset imbalance, StratifiedKFold was used to maintain the target variable's distribution in each fold. To handle the numerous columns, the SelectKBest feature selection method was adopted, selecting the top 10 features for classification tasks by default. The Random Under Sampler strategy was also used to address the imbalance issue. The models were evaluated using various metrics such as accuracy, precision, recall, ROC AUC, and average precision, providing a comprehensive performance comparison.

5.4 Logistic Regression Hyperparameter Optimisation

After comparing the models, Logistic Regression was selected for further development. The primary focus was on optimising the model's parameters. A comprehensive pipeline was generated to ensure the model passed through all necessary stages, including Column Transformers, SelectKBest, and RandomUnderSampler, before reaching the Logistic Regression phase. Interestingly, the top 10 columns selected by SelectKBest were all related to previous marketing campaigns, suggesting a significant influence on the model's

performance. Given the initial selection was limited to 10 out of 65 columns, optimising this parameter for better accuracy was considered. Additionally, the ROC and precision-recall curves were examined to establish a baseline before hyperparameter optimisation.

GridSearchCV, a robust machine-learning technique for finding the best hyperparameters, was employed to optimise the model's parameters. By evaluating multiple parameter combinations through cross-validation, GridSearchCV identified the optimal settings for the Logistic Regression model. The chosen metric for refitting was average precision, due to the dataset imbalance. The best parameters identified were: 'feature_selection__k': 25, 'model__C': 1000.0, 'model__penalty': 'l1', and 'model__solver': 'liblinear'. The results from the grid search were compiled into a data frame for detailed analysis.

6 Evaluation and Results Analysis

The evaluation presents a comprehensive analysis of the study's results, providing deep insights into the effectiveness of data-driven approaches in optimising direct marketing strategies. By employing a variety of statistical tools, the analysis critically assesses the reliability and significance of the experimental research outputs. The use of visualisations, such as charts and graphs, further enhances understanding by illustrating complex data patterns, making the findings more accessible. This comprehensive approach underscores the importance of leveraging customer data for more targeted and effective marketing campaigns, highlighting the transformative potential of data analytics in direct marketing.

6.1 EDA

The EDA phase, utilising ydata_profiling, was instrumental in identifying patterns and potential issues within the dataset. The examination of the data provided important insights into its structure, relationships, and key characteristics. ydata_profiling generated detailed summaries for each variable, highlighting statistics, distributions, missing values, outliers, and correlations. These insights facilitated the selection of key columns for deeper analysis. Figure 3 represents some pivotal columns that were further analysed, showcasing their distributions and relationships.

Further analysis revealed that customers with basic education tend to have lower incomes and spend less, whereas higher education levels correlate with greater financial resources and increased spending. This pattern supports the general understanding that higher educational attainment leads to improved financial stability. Additionally, the analysis indicated that younger customers typically possess lower educational levels. Interestingly, there is no significant correlation between marital status and income levels.

The analysis of the acceptance of previous campaigns and response rates indicated that most customers did not accept any marketing campaigns. However, those with higher incomes were more likely to respond positively to such efforts. Customers who accepted campaigns also tended to have higher spending and income levels, further highlighting the link between financial capacity and marketing responsiveness.

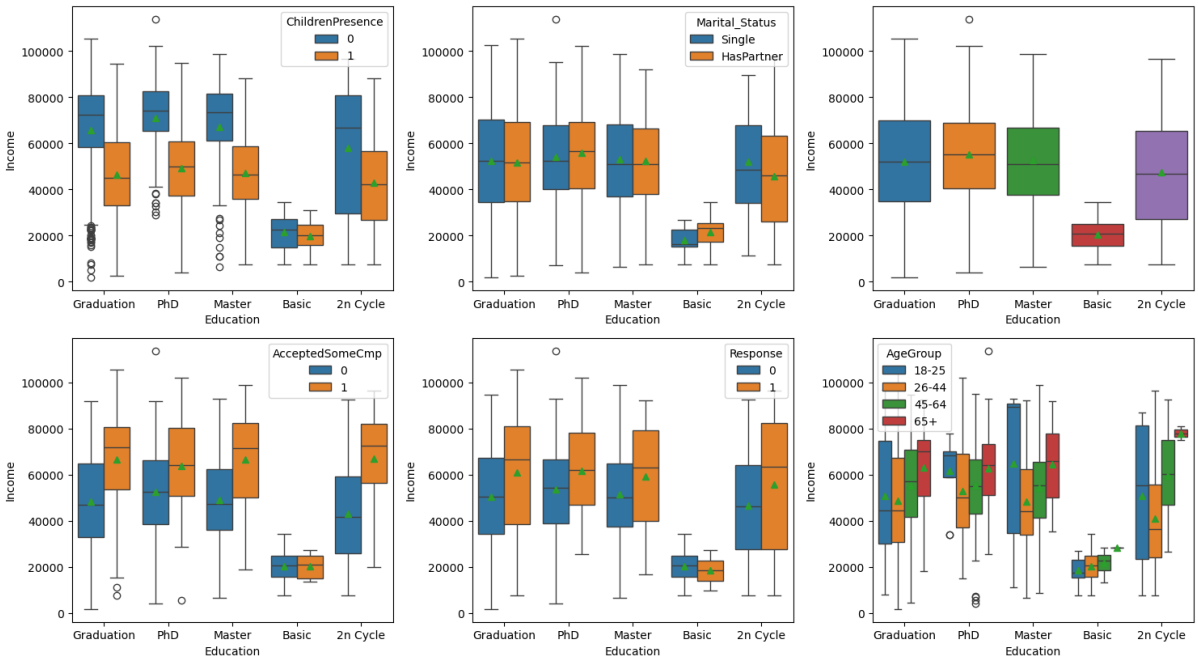

Figure 3: Income Distribution by Education Level and Demographic Characteristic

Analysing the correlation matrix and the relationship between features and customer response to the last campaign (Figure 4), several key insights emerge. Higher income is linked to fewer web visits, suggesting that wealthier customers may require fewer visits to make a purchase decision. Additionally, higher income tends to correlate with fewer children at home, confirming earlier conclusions. Customers with basic education tend to have lower incomes, while higher income is associated with more purchases and higher spending, especially on regular products. Moreover, higher income correlates with greater acceptance of marketing campaigns, indicating that financially stable customers are more receptive to marketing efforts.

Regarding response to campaigns, higher income levels slightly increase the likelihood of a positive response, suggesting wealthier customers are more responsive to marketing efforts. More recent purchases are linked to higher conversion rates, indicating recent engagement boosts responsiveness. Longer enrolment correlates with higher conversion rates, emphasising customer loyalty's importance. Conversely, having children or a partner reduces conversion likelihood. Acceptance of multiple campaigns significantly increases the likelihood of a positive response, showing that customers who previously engaged with campaigns have a higher probability of future engagement. Higher spending also correlates with higher conversion rates, reinforcing the link between financial capacity and marketing responsiveness.

These insights provide a comprehensive understanding of the dataset, highlighting key customer segments and behaviours crucial for targeted marketing strategies and customer relationship management. The data cleaning and exploratory data analysis laid a strong foundation for identifying significant trends and patterns within the customer base. Targeting higher-income segments for marketing campaigns can enhance engagement and effectiveness. Potential actions include focusing on high-income, high-spending segments for personalised campaigns, developing tailored products and services for different educational backgrounds, creating age-specific offers targeting older age groups who tend to spend more, and refining campaigns to improve acceptance rates and boost spending among respondents.

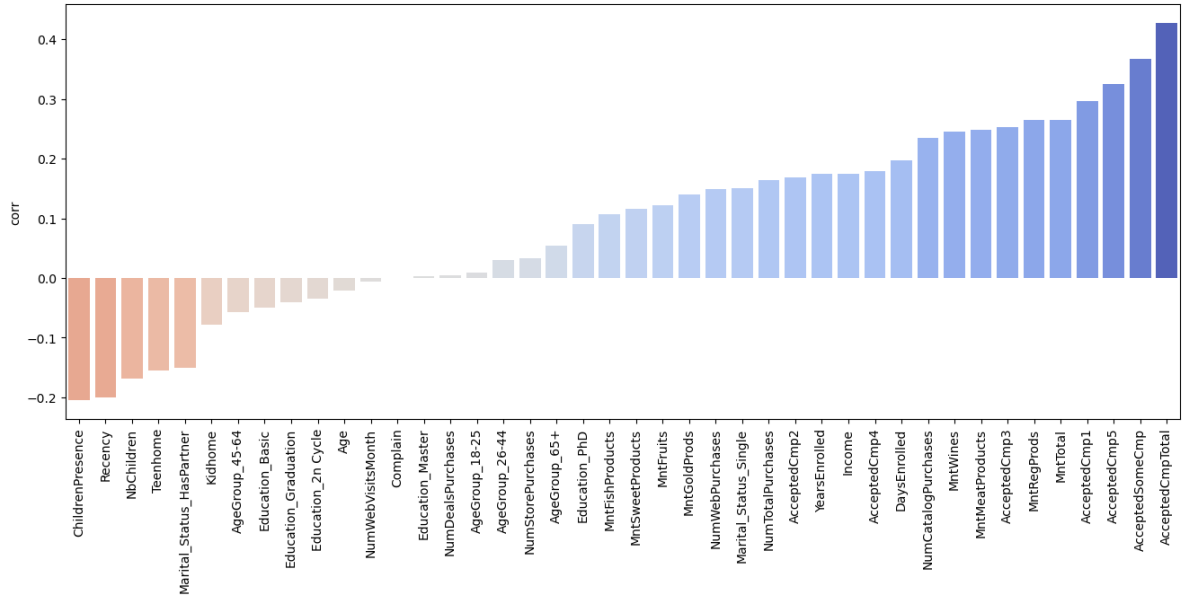

Figure 4: Feature Correlation with Customer Response

6.2 Customer Segmentation: K-Means Clustering

Preprocessing steps were important to standardise the dataset and improve the performance of clustering algorithms. These steps expanded the dataset to 64 columns by transforming the features. To ensure that the transformations were correctly applied, these features were visualised and examined in detail. The optimal number of clusters was determined using the elbow and silhouette methods, as illustrated in Figure 5. These methods indicated that three clusters were the ideal number, providing a balanced and effective segmentation of the data.

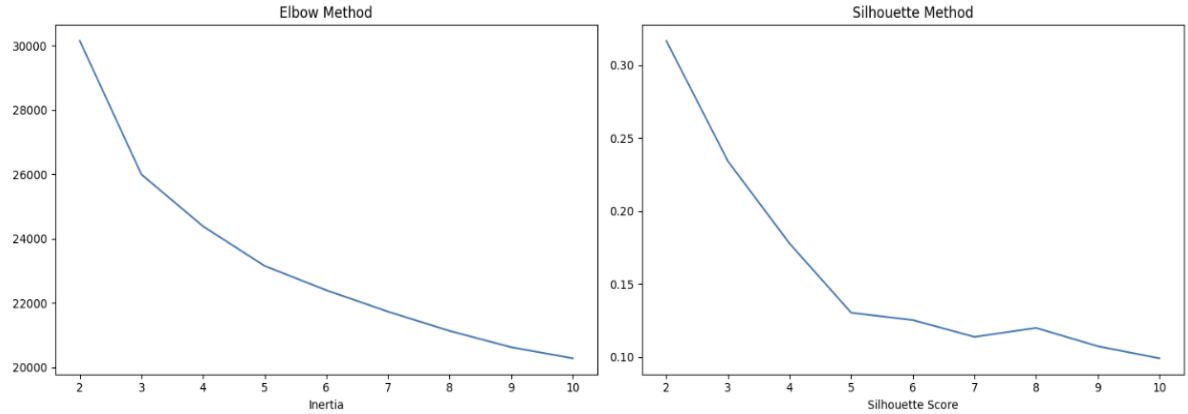

Figure 5: Methods for Optimal Numbers of Clusters

After determining the optimal number of clusters, a comprehensive pipeline was created to automate the process, ensuring a consistent workflow and reproducibility. This pipeline included preprocessing steps, PCA for dimensionality reduction, and K-means clustering. The results and analysis of the clusters showcased through diverse visualisations (including Figures 6 and 7) revealed significant insights.

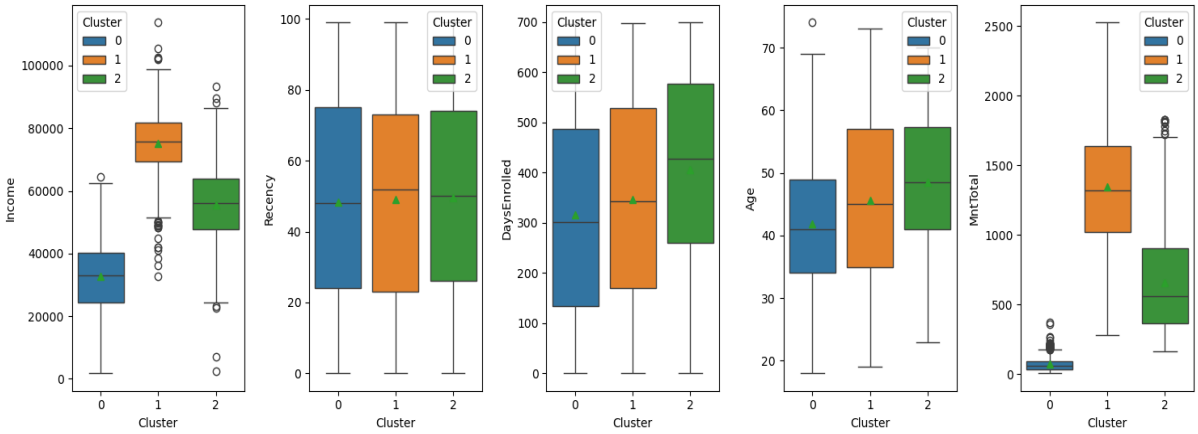

Figure 6: Boxplots of Key Metrics Across Customer Clusters

Figure 7: Distribution of Key Attributes and Marketing Engagement by Cluster

Based on the analysis, Cluster 0 is characterised by low income (median around 30,000), younger customers, low education levels, high child presence (88.5%), and low spending (most under 500). They have a very low acceptance of campaigns (7.9%) and a moderate-low response rate (23.4%). Cluster 1 has the highest income (median around 70,000), spans all age groups, high spending (500-2500), and most do not have children (80.7%). They show a high propensity to accept campaigns (52.6%) and a high response rate (49.2%). Cluster 2 features mid-range income (median around 50,000), middle-aged customers, high child presence (93.9%), and intermediate spending (500-1500). Their campaign acceptance and response rates are moderately high (27.3%).

Marketing strategies should be carefully tailored to the unique characteristics of each cluster to maximise effectiveness and customer engagement. Cluster 0, characterised by low income, high child presence, and low spending, requires budget-friendly products and promotions. Given their moderate-low response rate and very low acceptance of campaigns, marketing efforts should focus on affordability and value. Tailored campaigns that highlight discounts, special offers, and cost-effective solutions can help improve engagement and attract this price-sensitive segment.

Cluster 1, with the highest income and spending, spans all age groups but has a significant proportion of older adults. This cluster shows a high propensity to accept campaigns and a high response rate, making them ideal candidates for high-value products and regular engagement. Marketing strategies for Cluster 1 should emphasise premium offerings, loyalty programmes, and personalised experiences. Regular engagement through exclusive events, personalised recommendations, and high-end products can leverage their higher income and willingness to engage with campaigns, fostering long-term loyalty and increasing revenue.

Cluster 2, which has middle-aged customers with mid-range income and spending, shows moderate-high campaign acceptance and response rates. This cluster is suited for premium products and personalised marketing strategies that cater to their balanced spending habits and educational backgrounds. Strategies should focus on premium and high-quality products, coupled with personalised marketing messages that resonate with their lifestyle and preferences. Offering tailored experiences, such as customised product bundles and targeted promotions, can enhance their engagement and drive sales.

6.3 Predictive Classification Models

A comparison of the three predictive classification models (Decision Tree, Logistic Regression, and KNN) and Dummy Classifier were evaluated to identify the one with the best performance (Figure 8). After running the models, the evaluation showed that KNN had the best performance in accuracy and precision but performed poorly in recall. Logistic Regression offered the best overall balance, particularly in metrics critical for imbalanced datasets, such as recall, ROC AUC, and average precision, with a good time efficiency. Given its consistent performance across key metrics, Logistic Regression was chosen for further improvement and hyperparameter optimisation.

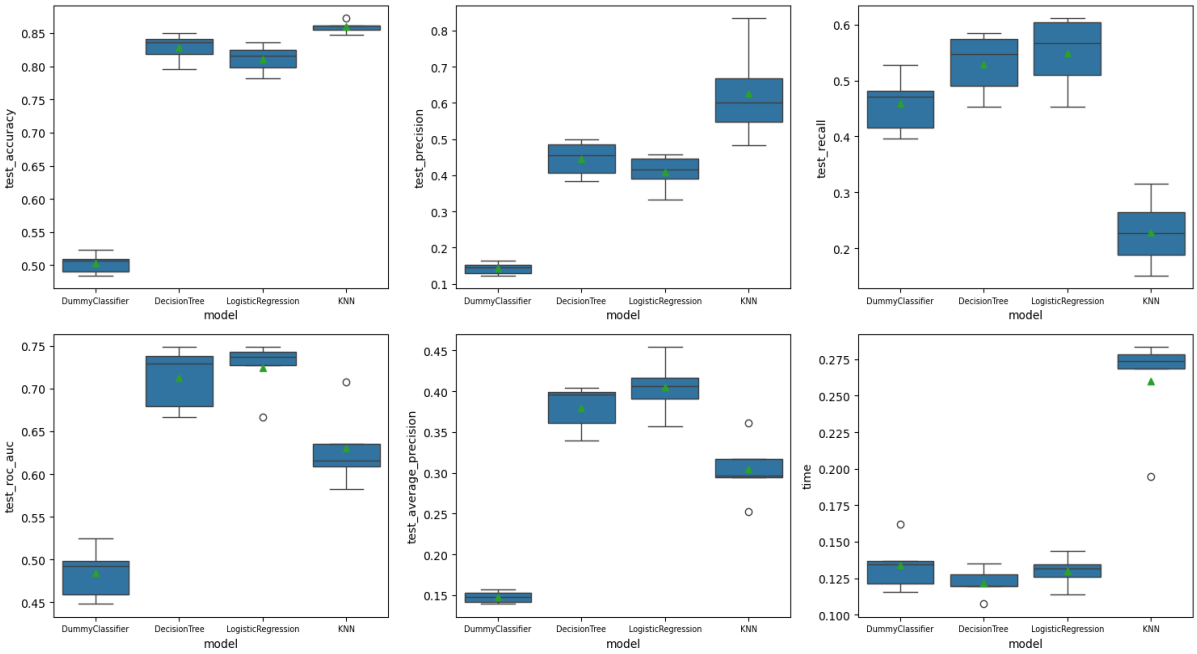

Figure 8: Performance Comparison of Classification Models Across Metrics

6.4 Logistic Regression Hyperparameter Optimisation

GridSearchCV was employed to systematically identify the optimal parameters for the Logistic Regression model, significantly enhancing its performance. As shown in Figure 9, hyperparameter optimisation led to substantial improvements in recall, ROC AUC, and average precision, boosting the model's ability to accurately identify true positives and differentiate between classes. While accuracy and precision remained stable, reflecting consistent performance before and after optimisation, the slight increase in training time suggests that the process was efficient and did not add significant computational complexity.

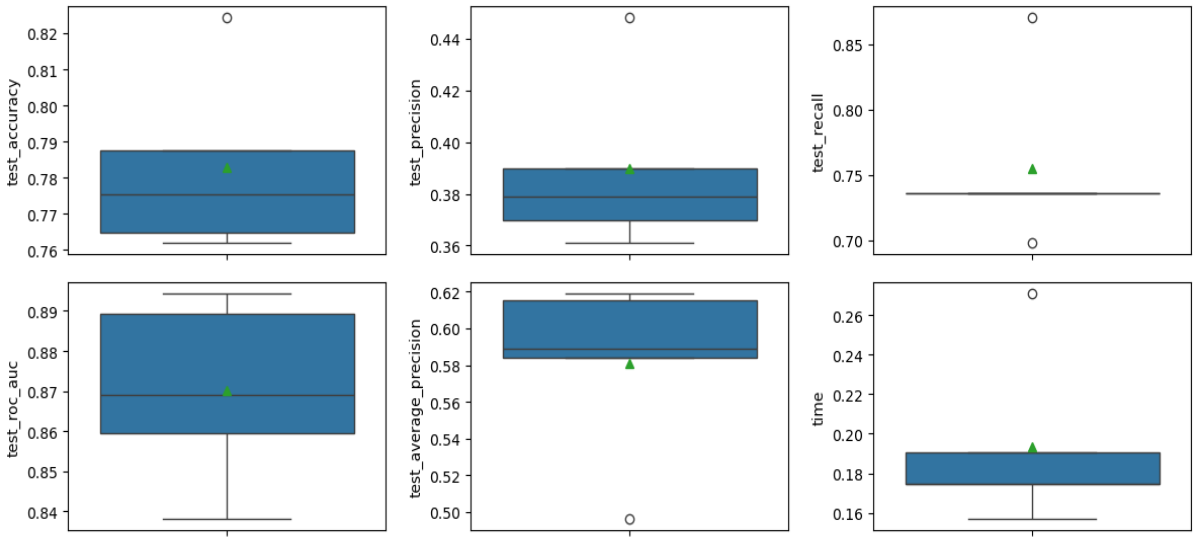

Figure 9: Performance Metrics of Logistic Regression Model after Hyperparameter Optimisation

Also, the most important features identified during model training were examined across the entire dataset to understand their impact on customer behaviour. The feature importance values reveal that campaign acceptance, the duration of customer registration, and spending habits are key predictors of the target behaviour. However, the negative importance of some campaign acceptance features suggests complex dynamics that may need further investigation. Understanding these patterns can help refine marketing strategies and improve customer engagement practices, ensuring more targeted and effective campaigns.

When applying the optimised model to the test set, as shown in Figures 10 and 11, there was a significant improvement in performance metrics. The ROC AUC increased from 0.69 to 0.78, indicating a better ability to distinguish between positive and negative classes. Additionally, the Average Precision (AP) rose from 0.43 to 0.57, reflecting an improved balance between precision and recall. These enhancements make the model more reliable and robust for practical applications. The optimised model is now better equipped to handle actual data, providing more reliable support for decision-making.

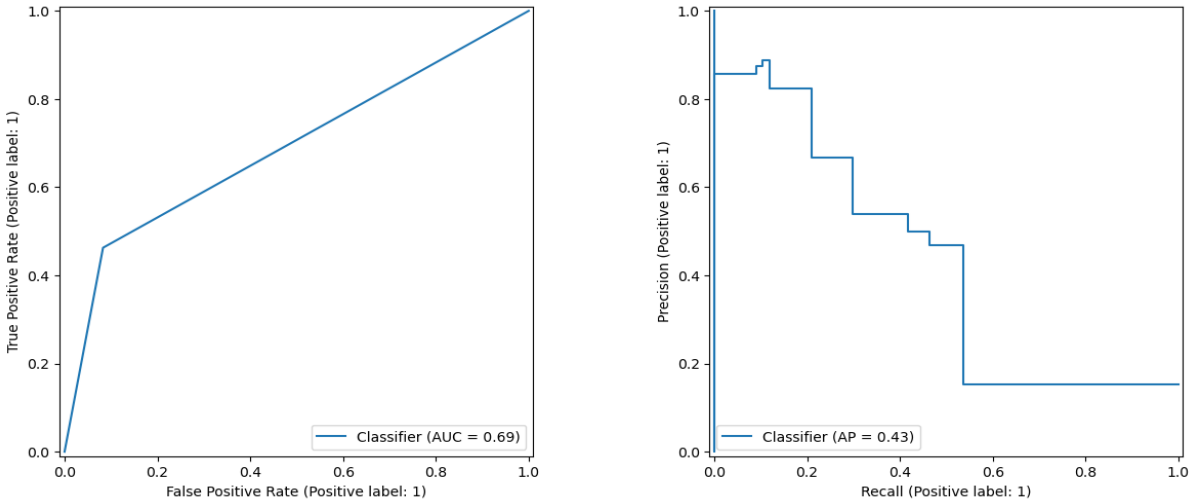

Figure 10: ROC and Precision-Recall Curves before Hyperparameter Optimisation

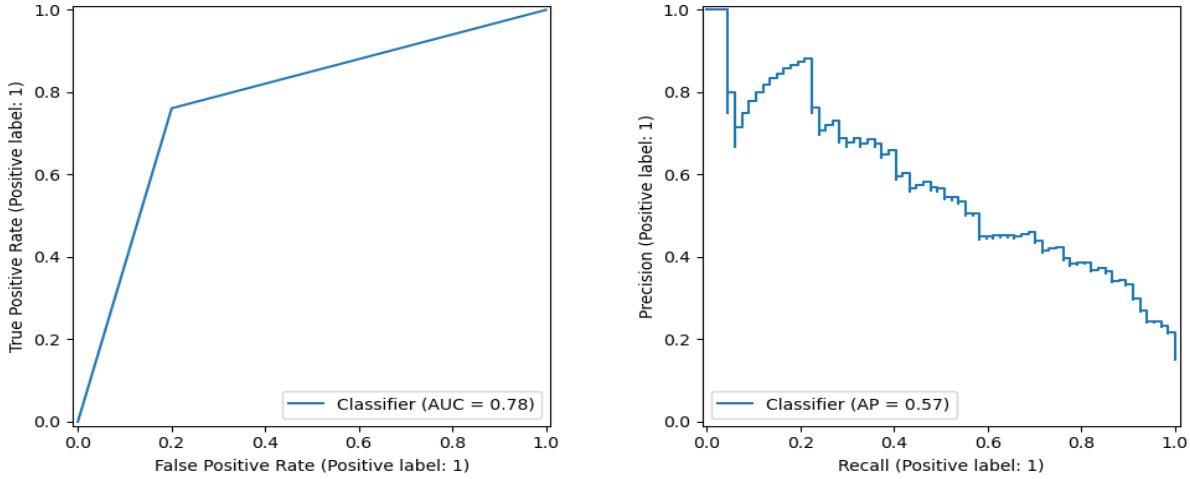

Figure 11: ROC and Precision-Recall Curves after Hyperparameter Optimisation

The confusion matrix for the test data (Figure 12) shows that post-optimisation, the Logistic Regression model achieved strong recall, effectively identifying a high proportion of positive cases, but with lower precision, resulting in a significant number of false positives. While the model's overall accuracy of 79.3% is robust, the trade-off between recall and precision suggests that there is potential for improvement. Enhancing precision would reduce false positives, making the model more reliable, especially in scenarios where the cost of incorrect positive predictions is high. This balance would further refine the model's effectiveness in practical applications.

By using Logistic Regression, the likelihood of each customer responding positively to a marketing campaign can be accurately predicted. Customers with a high predicted probability of responding can be prioritised for contact, while those with a low probability can be excluded. This targeted approach ensures that marketing resources are efficiently allocated, focusing on customers most likely to generate revenue, thereby maximising return on investment (ROI).

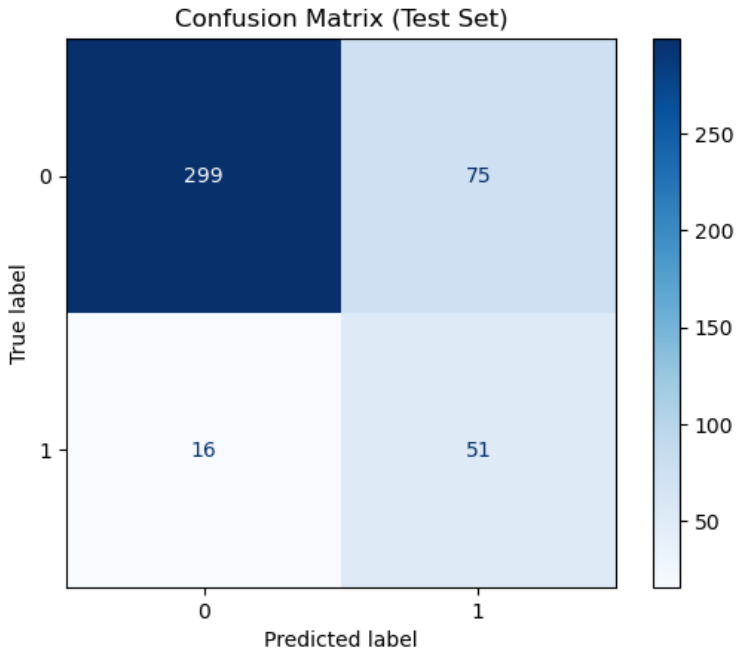

Figure 12: Confusion Matrix of Logistic Regression Model on Test Set

As highlighted in the case description, and in the context of the Z columns removed from the dataset during data cleaning, the estimated cost of contacting each customer is 3MU, while the potential revenue per positive response is 11MU. In a baseline scenario where all 441 customers in the test set are contacted, the total cost amounts to 1,323MU (441 customers × 3MU per contact). However, only 67 customers responded, generating revenue of 737MU (67 responders × 11MU per response), resulting in a net loss of 586MU.

After applying Logistic Regression, where only the 126 customers predicted to respond positively are contacted, the total cost drops significantly to 378MU. With 51 true positive responses, the generated revenue is 561MU, leading to a net profit of 183MU. This demonstrates how the model effectively reduces unnecessary contact costs while enhancing revenue by targeting the most promising customers, transforming a potentially unprofitable campaign into a profitable one.

In summary, the evaluation phase underscored the effectiveness of the developed models, illustrating the practical application of data-driven analytics in optimising marketing strategies. The insights gained are crucial to the overall success of the research, offering valuable contributions to both academic and industry perspectives on leveraging advanced analytics for direct marketing optimisation.

7 Conclusion and Discussion

This research aimed to improve the effectiveness of direct marketing campaigns by applying data-driven analytics, focusing on customer segmentation and predictive classification models. Through the CRISP-DM methodology, the study systematically analysed customer behaviour, leading to actionable insights that can enhance marketing strategies. The application of Logistic Regression, Decision Trees, and K-Nearest Neighbours (KNN) allowed for a detailed comparison of predictive models, and Logistic Regression was

identified as the most balanced and effective model, particularly after hyperparameter optimisation, which significantly improved its ability to predict customer responses accurately.

The use of clustering to segment customers provided additional insights, allowing the development of customised marketing strategies that target different customer groups. For example, Cluster 0, characterised by low income and high child presence, requires budget-friendly products and value-driven marketing efforts. In contrast, Cluster 1, with higher income and spending levels, is better suited for premium offerings and regular engagement, while Cluster 2, which is in the middle, benefits from personalised and premium marketing strategies.

The use of historical data, which might not accurately reflect the dynamic nature of customer behaviour and preferences, is one of the research's limitations. Furthermore, even though the generated models performed well on the test data, overfitting is always a possibility, especially with highly tuned models like the optimised Logistic Regression. Future research could address these limitations by incorporating real-time data streams and exploring more advanced machine learning techniques, such as ensemble methods or deep learning models, which may offer better generalisation capabilities.

The implications of this research are substantial for various stakeholders, including marketing professionals, data scientists, and business strategists. For companies, the insights gained from this study can guide the development of more effective marketing campaigns, resulting in better customer engagement and increased profitability. For academics, this research contributes to the growing body of knowledge on data-driven marketing strategies, offering a framework that can be further refined and tested in different contexts.

In conclusion, this study demonstrates the practical benefits of integrating data-driven analytics into direct marketing strategies. By applying advanced segmentation and predictive models the research demonstrated how data analytics can be applied to improve customer engagement and increase revenue. The research also provides a foundation for further exploration and potential commercialisation of these findings, offering valuable contributions to both academic and industry practices.

References

Ali, N. (2023) ‘Influence of data-driven digital marketing strategies on organizational marketing performance: Mediating role of IT infrastructure’, in *Conference on Sustainability and Cutting-Edge Business Technologies*, Springer, pp. 337–347.

Anitha, P. and Patil, M. M. (2022) ‘RFM model for customer purchase behavior using K-Means algorithm’, *Journal of King Saud University-Computer and Information Sciences*, 34(5), pp. 1785–1792.

Apampa, O. (2016) ‘Evaluation of classification and ensemble algorithms for bank customer marketing response prediction’, *Journal of International Technology and Information Management*, 25(4), p. 6.

Bose, I. and Chen, X. (2009) ‘Quantitative models for direct marketing: A review from systems perspective’, *European Journal of Operational Research*, 195(1), pp. 1–16.

Chate, P. A. (2022) *Behavioral Modelling of Customer Marketing Patterns and Review Prediction Using Machine Learning Techniques*. PhD thesis. Dublin: National College of Ireland.

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

Hossain, M. A., Akter, S., Yanamandram, V. and Wamba, S. F. (2023) 'Data-driven market effectiveness: The role of a sustained customer analytics capability in business operations', *Technological Forecasting and Social Change*, 194, p. 122745.

Johnson, D. S., Muzellec, L., Sihi, D. and Zahay, D. (2019) 'The marketing organization's journey to become data-driven', *Journal of Research in Interactive Marketing*, 13(2), pp. 162–178.

Kara, Y., Boyacioglu, M. A. and Baykan, Ö. K. (2011) 'Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange', *Expert Systems with Applications*, 38(5), pp. 5311–5319.

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

Lu, T. and Boutilier, C. (2014) 'Dynamic segmentation for large-scale marketing optimization', in *ICML-2014 workshop on customer life-time value optimization in digital marketing*, 31st international conference on machine learning (ICML 2014), Beijing, June 21–26.

Olbrich, R. and Lindenbeck, B. (2016) 'Targeting direct marketing campaigns by a more differentiated view on generated sales', in *Proceedings of the 15th International Marketing Trends Conference*, Venedig, Italien, vol. 314.

Oliveira, V. C. (2020) 'ifood-data-business-analyst-test', Available at: https://github.com/ifood/ifood-data-business-analyst-test/ [Accessed 03 March 2024].

Olson, D. L. and Chae, B. K. (2012) 'Direct marketing decision support through predictive customer response modeling', *Decision Support Systems*, 54(1), pp. 443–451.

Pour, M. J. and Emami, S. A. (2023) 'Designing an integrated methodology for data-driven marketing strategic planning', in *2023 9th International Conference on Web Research (ICWR)*, IEEE, pp. 289–293.

Qu, Y. (2022) 'Using data mining techniques to discover customer behavioural patterns for direct marketing', in *2022 7th International Conference on Big Data Analytics (ICBDA)*, IEEE, pp. 361–365.

Sagala, N. T. M. and Permai, S. D. (2021) 'Predictive model using SVM to improve the effectiveness of direct marketing', in *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IEEE, pp. 1–6.

Schröer, C., Kruse, F. and Gómez, J. M. (2021) ‘A systematic literature review on applying CRISP-DM process model’, *Procedia Computer Science*, 181, pp. 526–534.

Scikit-learn (2024) ‘sklearn.preprocessing’, Scikit-learn: Machine Learning in Python. Available at: https://scikit-learn.org/stable/api/sklearn.preprocessing.html#module-sklearn.preprocessing/ [Accessed 10 July 2024].

Srikasem, C., Sureepong, P., Dawod, A. Y. and Chakpitak, N. (2022) ‘Data-driven approach to raise the marketing and trade strategy based on Halal food product transactions of China market’, in *2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, IEEE, pp. 95–101.

Venkatesan, R., Farris, P. W. and Wilcox, R. T. (2021) *Marketing Analytics: Essential Tools for Data-Driven Decisions.* University of Virginia Press.

Zaki, A. M., Khodadadi, N., Lim, W. H. and Towfek, S. (2024) ‘Predictive analytics and machine learning in direct marketing for anticipating bank term deposit subscriptions’, *American Journal of Business and Operations Research*, 11(1), pp. 79–88.

APPENDIX 1 – Overview of Dataset Variables

Category	Variable	Description
Personal data	ID	Unique customer identifier
	Year_Birth	Year of birth of the customer
	Education	Customer education level
	Marital_Status	Marital status of the client
	Income	Annual income of the client's family
	Kidhome	Number of children in the customer's home
	Teenhome	Number of teenagers in the client's home
	Dt_Customer	Date of registration of the customer with the company
	Recency	Number of days since the customer's last purchase
	Complain	If the customer complained in the last 2 years
Product data	MntWines	Amount spent on wine in the last 2 years
	MntFruits	Amount spent on fruits in the last 2 years
	MntMeatProducts	Amount spent on meat in the last 2 years
	MntFishProducts	Amount spent on fish in the last 2 years
	MntSweetProducts	Amount spent on sweets in the last 2 years
	MntGoldProds	Amount spent on gold products in the last 2 years
Purchase location data	NumWebPurchases	Number of purchases made through the company website
	NumCatalogPurchases	Number of purchases made using a catalog
	NumStorePurchases	Number of purchases made directly in stores
	NumWebVisitsMonth	Number of visits to the company website in the last month
Campaign data	NumDealsPurchases	Number of purchases made with a discount
	AcceptedCmp1	If the customer accepted the offer in the 1st campaign
	AcceptedCmp2	If the customer accepted the offer in the 2nd campaign
	AcceptedCmp3	If the customer accepted the offer in the 3rd campaign
	AcceptedCmp4	If the customer accepted the offer in the 4th campaign
	AcceptedCmp5	If the customer accepted the offer in the 5th campaign
	Response (target)	If the customer accepted the offer in the last campaign (pilot)
	Z_CostContact	Cost of contacting customers for the pilot
	Z_Revenue	Revenue from the new gadget (customer contact)