A Machine Learning Framework for Customer Segmentation and Campaign Response Prediction

Student Name: Vasanthakumar Mohan

Student Number: 20032099

Applied Research Project submitted in partial fulfilment of the requirement the degree of

Masters of Science in Business Analytics

At Dublin Business School

Supervisor: Dr. Samuel Ogwu

August 2025

Word Count: 11400

Declaration

I declare that this Applied Research Project that I have submitted to Dublin Business School for the award MSc. Business Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Vasanthakumar Mohan

Student Number: 20032099

Date: 26 August 2025

Acknowledgement

I would like to express my sincere gratitude to my project supervisor, Dr. Samuel Ogwu for their constant guidance and valuable feedback throughout this research. Their support was essential in shaping the direction and success of this project.

I also thank the academic staff at Dublin Business School for their encouragement and assistance during my research journey. I appreciate the contributions of my colleagues and friends who provided support throughout the course of this project.

Finally, I extend my heartfelt thanks to my family for their understanding and continuous support.

Abstract

This project is to increase the effectiveness of marketing campaigns by predicting the responses of customers through machine learning. The emphasis is on determining who is likely to respond to specific promotions so a business can maximize advertising efforts and get the best returns. The data comprises customer demographic and behavioural data in terms of previous purchases and campaign activities. To ensure the quality of the data and solve the class imbalance issue, missing value imputation, feature scaling, and Random Oversampling were used as pre-processing steps. Segmentation as well as predictive modelling was supported by feature engineering and dimensionality reduction using Principal Component Analysis (PCA). Three classification models were created and compared based on the essential metrics accuracy, precision, recall, F1-score, and ROC-AUC: Support Vector Machine (SVM) ,Random Forest and XGBoost. XGBoost was the most effective model with an F1-score of 0.7945 and ROC-AUC of 0.9870, thus classifying the likely responders. This project shows how predictive analytics with data-driven approaches can make the campaigns more accurate and help to make better decisions.

Table of Contents

LIST OF FIGURES

1. INTRODUCTION

In the digital marketing environment, companies are becoming more data-driven to help improve their marketing campaigns. Although predictive models are able to predict campaign response, this capability is not always effective due to lack of customer diversity insights. At the same time, segmentation models demonstrate the differences between groups but fail to provide the probability of the campaign success. The current marketing challenge is not only to predict, but to position predictions to actionable customer segments. This study attempts to solve that two-fold dilemma, by combining unsupervised clustering and supervised machine learning, to not only make it more personalized but also more precise in the targeting of campaigns.

1.1 Background and Scope of the Study

In the digital age of change, e-commerce has changed the way business and consumers communicate. As billions of customer touchpoints are created every moment through web clicks and ad impressions, purchase histories, and campaign responses organisations now have access to huge volumes of behavioural data (Rosario et al., 2021). However, even with all this amount of information at their disposal, marketing departments have a hard time turning insights into effective decisions. It is a major challenge to predict which customers are most likely to respond to a campaign (Zhang et al., 2022).

Conventional marketing methods would use mass segmentation strategies or general demographic classification that do not capture the specifics of a customer behaviour. On the one hand, predictive models such as Random Forest or XGBoost can tell how probable a customer is to take action but they also tend to lack insight into the behaviour context behind such output (Lee et al., 2021). These constraints indicate the obvious opportunity combining segmentation and predictive modelling in one, unified analytical chain.

This research project will directly respond to this opportunity since it will be using a hybrid data analytics method. It then applies K-Means clustering to generate meaningful customer groups that are based on demographic, purchasing, and engagement behaviours. Afterwards, these segments are integrated into machine learning classification models namely Support Vector Machine (SVM), Random Forest and XGBoost to examine whether segmentation can enhance the prediction of campaign responses.

This research is based on data science and real-life application on marketing. On the technical side, it encompasses pre-processing methods such as handling outliers, scaling of the feature data, rectifying class imbalance through Random Oversampling and dimensionality reduction using Principal Component Analysis (PCA). On the strategic level, this will help to assess how the models-based insights can inform the personalization of marketing efforts, enhance resource allocation, and increase the ROI. Companies are no longer satisfied with average conversion rates, they need precision. This research will help businesses launch smarter, data-driven campaigns since the given model will predict the outcome of the campaign in addition to explaining which segments of customers are most responsive.

1.2 Problem Statement

Although businesses have access to rich customer data, many businesses are unable to effectively predict who is likely to respond to a marketing campaign. In reality, traditional methods tend to ignore the diversity of customers or fail to manage skewed data well, resulting in ineffective targeting and a waste of resources (Almeida et al., 2023). Models, like K-Means, will group customers in meaningful ways, but will not predict response to campaign, whereas models, like Random Forest and SVM, will predict response to campaign but not segment customers behaviourally (Farrell, 2022; Khandokar et al., 2023).

The research overcomes these limitations by integrating K-Means clustering with predictive modelling in order to create a hybrid system. The goal is to divide customers into specific personas and determine their potential to respond to marketing offers and enhance targeting accurateness, ROI of campaigns and overall marketing performance.

1.3 Research Question

1. What customer segments can be identified through K-Means clustering of demographic, purchase behaviour, and engagement attributes in the marketing campaign dataset to inform targeted marketing strategies?
2. What impact does the inclusion of K-Means cluster labels have on the accuracy of machine learning models in forecasting customer responses to marketing campaigns?

1.4 Research Objectives

1. To Clean data on marketing campaign, engineered features, encoded data, outliers and standardizing the data.
2. K-Means segmentation of the customers according to the attributes of demographics, purchase behavior and engagement.
3. To benchmark and train the three machine learning models of Random Forest, SVM and XGBoost to forecast the response to the marketing campaign.
4. To analyze the effect of including cluster labels as feature in predictive modelling, and to see whether segmentation enhances the accuracy of campaign response.
5. To provide the marketing strategy advice to business based on the customer segment-based insights and model-generated forecasts.

1.5 Hypotheses

1. **H1**: Customers can be grouped into meaningful segments based on demographic, purchase, and engagement attributes using unsupervised clustering.
2. **H2**: Including cluster labels as an additional feature will enhance the predictive accuracy of machine learning models.

2.Literature Review

Jindal et al. (2023) utilized the K-Means clustering method to categorize retail banking customers based on demographic and transactional information. They have pre-processed their data by min-max scaling, outlier detection using interquartile range (IQR) and dimensionality reduction using Principal Component Analysis (PCA). The resulting segments revealed different loyalty and engagement patterns and allowed marketing intervention to be applied to these patterns. This paper confirms the use of unsupervised clustering in extracting behaviour-based customer personas that can be used directly in segmentation strategy using K-Means and PCA.

Varma and Kale (2024) evaluated the predictive performance of Random Forest and XGBoost when modelling the response of email campaigns in the e-commerce industry. The data set had customer demographics and behavioural data such as frequency of purchase and click-through history. The class imbalance was addressed via SMOTE, and categorical features were label-encoded. The performance of XGBoost in terms of AUC of 0.91 was superior to Random Forest, which confirms the superiority of boosting methods in the context of campaign-level prediction tasks, which directly correlates with the performance of classifier models.

In clustering methodology validation, internal evaluation metrics are still an essential part , Okafor and Mensah (2022) compared the use of hierarchical clustering and K-Means to cluster retail consumers according to purchasing behaviour. Their pre-processing pipeline included one-hot encoding and z-score standardization. The efficiency of the segmentation was determined through the Silhouette Score and Davies-Bouldin Index that showed the best K-values and a strong clustering. Their results provide a methodological template of justifying unsupervised segmentation, which is supported by the choice of this study to apply silhouette scoring in cluster validation.

Segmentation features have been also combined with classification models in churn prediction scenarios. The Telco dataset was applied by Al-Suwaidi and Rahman (2023) to investigate the effects of the customer segments created through the K-Prototypes on the churn models performance. They found that using segment labels improved Random Forest by 12 percent. Even though the segment features are not actually used in the classification step in this study, the results support the theoretical benefit of using segmentation as a method of improving the performance of marketing response models.

SVM remain an efficient option when dealing with high dimension and non-linear data in the classification task. Pereira et al. (2021) examined this by comparing SVM with the Decision Trees approach to the prediction of campaign response in a large-scale context within the retail sector. The data was subjected to mutual information based feature selection and standard scaling. Although Decision Trees are more interpretable, SVM performed better in terms of F1-score especially in the imbalanced situations. This highlights the suitability of incorporating SVM in the model evaluation pipeline of this project since the customer behaviour data is nuanced.

RFM-based clustering techniques are practical in terms of offering a better campaign performance due to meaningful segmentation. Cheng et al. (2022) applied the K-Means clustering based on Recency, Frequency, and Monetary to segment digital shoppers. They pre-processed by log-transforming to eliminate skewness, and z-scoring to stabilize variance. Segmentation strategy resulted in a 21 percent increase in conversion rates as opposed to generic marketing strategies. This finding is highly encouraging in terms of the business logic of incorporating behavioural segmentation before implementing any targeted campaigns- a factor that was evident in the early phases of this research process.

When it comes to tackling prediction on the engagement of the loyalty program, ensemble models like XGBoost have been shown to perform better. Becker et al. (2023) used Logistic Regression and XGBoost to analyse the customer feedback to the promotional campaigns. The authors used stratified k-fold cross-validation and approached the problem of data sparsity by using iterative imputation. XGBoost was found to outperform on all the measures of classification, including precision and AUC, which demonstrates its effectiveness in gaining the benefit of heterogeneous customer data. Their paper supports the findings of this study to use XGBoost since it offers a balance between scalability, accuracy, and interpretability in any predictive task.

Customer segmentation based on purchase behaviour and domain related attributes has also been useful with specialized retailing contexts. K-Means clustering application has been used in pharmacy retail analytics where purchases are included, as well as the health profile of patients (Rajkumar et al., 2024). The data were treated by binning and multivariate outlier removal. The deducted segments were directly used to streamline the inventory and how to time the seasonal campaigns. This application of unsupervised learning to aid strategic business activities is similar to the focus of this research on the creation of interpretable clusterings that can be used in downstream decision-making activities..

The issue of probabilistic versus partition based clustering approaches remains to be debated in the context of segmentation studies particularly in the e-commerce setting. Li et al. (2021)

compared Gaussian Mixture Models (GMM) and K-Means as methods of online customer segmentation. They pre-processed by feature encoding, normalization and removal of the low-variance features. Although GMM was more granular, K-Means was faster and more scalable, which would make it more appropriate to use in real-time applications. This justifies the methodological decision of K-Means in the current project bearing in mind that the marketing datasets are high-dimensional and operationally dynamic in nature.

The use of outputs of unsupervised segmentation in classification models have been shown to enhance predictive accuracy in modelling of campaign response. Batista et al. (2022) created a predictive framework of fashion retail wherein they tested the effect of behavioural segment labels to the performance of the XGBoost classifier. They did their pre-processing by rebalancing classes with Borderline-SMOTE and interaction-based feature engineering. The inclusion of the cluster features led to an increment of AUC by 9.6 percent. In addition to theoretical advantages of the decoupling of clustering and classification phases, the findings of Batista et al. positively confirm the theoretical value of integrating persona-based insights into persona-based targeting of campaigns.

Comparisons of ensemble methods to simpler methodologies based on decision trees have been widely made in research into campaign response modeling. In particular, El-Hajj et al. (2024) predicted campaign acceptance using Decision Trees and achieved a significant boost of the recall when resampling techniques were employed, whereas Khandokar et al. (2023) implemented a wide range of algorithms which includes Random Forest, SVM, and XGBoost on online retail data. Both studies stressed on the necessity of address the issue of class imbalance by using oversampling or weighting methods and this is directly applicable to your use of Random Oversampling to prepare the marketing data. All of their results show that simple trees can be interpreted, but ensembles are more reliable in terms of providing performance gains, particularly in imbalanced conditions.

The combination of randomized tree structures and clustering donor behaviour has furthered the segmentation strategies. Wang et al. (2024) suggested a model that creates a node embedding using Random Forests to perform personalization of communication and segmentation and showed good AUC in a task of email targeting. Simultaneously, Lim and Wang (2023) proposed the RFMP framework of donation-based crowdfunding platforms that use RFM-like metrics and clustering algorithms to segment donors without demographic information. Their conclusions help to make the case that embedded tree models and RFM-based segmentation can improve targeting accuracy and disclose underlying behavioural personas-something that can point you towards new directions in feature set enrichment.

Unsupervised learning-based approaches to customer segmentation have also developed ensemble and spectral clustering. Hicham and Karim (2022) discussed the cluster ensemble and the spectral methods, and they also showed that cluster robustness in the segmentation process is increased. In a similar study, John et al. (2024) compared the K-Means, GMM, DBSCAN, and hierarchical clustering approaches on UK retail data and noted that GMM is the best performing by silhouette score though K-Means is very practical on large data. The technical support of a segmentation-first approach is provided by the complementary strengths of ensemble and probabilistic clustering methods, which do not compromise on scalability, which is central to the research modeling decisions.

New segmentation models are coming out in the dynamic and high-dimensional market data. Kabir (2025) proposed a system that connected K-Means clustering and neural network models in order to forecast customer behaviour on online purchasing sites, where they relied on a large quantity of transactional data to guide both customer segmentation and behaviour forecasting. Uddin (2024) presented an ML-based segmentation method that was specific to the digital-native markets and focused on the feature relevancy to younger consumer segments. Collectively, these methods demonstrate the capacity of sophisticated clustering to complement deep learning frameworks or demographically sensitive customization in the case of the modern marketing environment.

High-tech machine learning systems are transforming segmentation procedures by integrating the concept of adaptive algorithms. Ashraf (2025) suggested an innovative segmentation framework based on ML and intended to enhance the marketing strategies, using the flexibility of the algorithm to discover specific customer groups. In parallel, Wang et al. (2025) merged reinforcement learning (Q-learning), with differential evolution and K-Means clustering, applying noise reduction based on PCA, to improve clustering in digital marketing settings. These changing frameworks indicate potential avenues of adaptive and automated segmentation pipelines that would be worthwhile reference points to future.

Product development and marketing design as well are witnessing the increasing prominence of the interpretable and behaviourally rich segmentation methods. Joung et al. (2023) proposed an interpretable ML-based segmentation model to extract the feature importance of online product reviews in order to generate unsatisfied needs to ideate new products. El-Hajj and Pavlova (2024) also showed that resampling-aided Decision Trees can enhance recall (44-83%) and F1-score (49-74%) significantly in predicting the campaign response. These analyses combine interpretability with actionable targeting unifying the project to transparency and performance in clustering and classification workflows.

Kumar et al. (2023) applied K-Means clustering to PCA, to categorize high-value and low-value customers based on an online retail database. They were pre-processed by using min-max normalization and eliminating extreme outliers and the segmentation outcomes provided valuable insights regarding discounting and loyalty programs. Similar research Sharma et al. (2024) compared Random Forest and XGBoost and used the same dataset and concluded that AUC increased significantly when SMOTE and other oversampling methods were used. Collectively, these articles demonstrate how clustering to segment, and ensembles to model campaign response complement one another, and are directly relevant to the pipeline of the project.

Nguyen et al. (2022) used the customer purchase data on one of the e-commerce platforms located in Southeast Asia and applied GMM and K-Means to define the latent personas. They used log-transformation to address the skewedness of purchase frequencies, and categorical encoding. In the classification step, SVM performed better than Decision Trees to predict the likelihood of repeat purchases. On the same note, Ali et al. (2023) also compared Boosting models on the prediction of the responses to seasonal campaigns and revealed that XGBoost model was computationally efficient and performed better in F1-score. These two insights together support the usefulness of probabilistic segmentation and boosting models in improving the accuracy of marketing targeting.

Dasgupta et al. (2024) carried out research in the retail grocery industry in which the data on customer engagement was grouped through hierarchical clustering and K-Means, which was confirmed by Silhouette and Davies-Bouldin indices. Pre-processing was done by normalization of the data to z-scores and imputation of missing demographic data. Parallel, Choudhury et al. (2023) provided a comparison of Random Forest and SVM on promotional campaign response data and presented that SVM had better recall, whereas Random Forest had higher precision. When combined, these results indicate the trade-offs involved in selecting the classifier when predicting the outcome of a campaign, and this is consistent with the approach took to comparing the models.

Haque et al. (2022) suggested an RFM + K-Means segmentation system in online retailing, in which recency, frequency, and monetary scores were scaled towards the clustering of customer into engagement-based personas. The study presented that segment-based offers were 18 percent more likely to convert than broad targeting was. In line with this, Lin et al. (2023) compared the XGBoost with the Logistic Regression on an XGBoost fashion e-commerce campaign response data and found that their XGBoost model AUC was 15% better compared to the baseline model. The works together indicate that segmentation-based personalization and classifiers boosting are significant ways to increase marketing ROI.

Martinez et al. (2023) examined the telecom sector and developed a churn and campaign response prediction with K-Prototypes clustering of mixed demographic and behavioural characteristics. When clusters were used as model inputs, Random Forest accuracy increased by 11%. In comparison, Kim et al. (2024) concluded that without the explicit labels of the segments, SVM and Gradient Boosting could identify nonlinear customer behaviour in e-commerce. Collectively, these studies demonstrate how segmentation can enhance prediction but also how strong classifiers can work without explicitly clustering observations--although this may give you some perspective as to whether not including cluster labels in the predictive models was the right decision.

In the work by Fernandes et al. (2022), the IQR and the PCA-based dimensionality reduction method were also used to detect outliers, with extremely high results in terms of cluster coherence in online shopper datasets segmented with K-Means. Their study highlighted how pre-processing decisions shape segment interpretability. Simultaneously, Becker et al. (2024) have written about how to use explainable model ensembles such as XGBoost with SHAP analysis to make predictions more transparent to the marketing manager when targeting campaigns. The blend of rigor in pre-processing your work in segmentation and interpretability in classification echo directly to the dual focus of your study to produce both actionable clusters and explainable predictive models.

In the study by Singh et al. (2023) have clustered pharmacy chain customers based on data on purchase sequence and K-Means, with pre-processing operations, such as frequency binning and normalization. They demonstrated that with segment-driven targeting, the seasonal product sales were increased by 14%. On the predictive side, Wang et al. (2024) compared XGBoost, Random Forest and Cat Boost on retail CRM data and XGBoost performed the best in terms of stability to precision, recall and AUC scores. These results further confirm the model comparison decision and business-driven goal of aligning clustering results with the campaign prediction results.

Zhou et al. (2023) examined the use of feature engineering to optimize the predictive ability of the campaign response models based on the big data of retail transactions. The authors stressed that raw behaviour and demographic characteristics are not enough to reflect nonlinear consumer interactions. Their pre-processing pipeline involved feature generation of the interactions, normalization of continuous variables and PCA was used to take care of the multicollinearity. The models selected were Random Forest and Gradient Boosting where Gradient Boosting was the best in terms of precision, recall and AUC. In particular, the experiment demonstrated that data preparation contributed to a 12 percent boost in the F1-score, which indicates the importance of such a procedure as an influential factor in the performance of a classifier.

Hassan et al. (2022) presented an unsupervised customer segmentation study within an e-commerce setting with the K-Means clustering algorithm on duration of sessions, the navigational route, and frequency of purchases. Data pre-processing involved log transformation to reduce skew, min-max scaling to normalize the data and imputation of missing clickstream data. The clustering revealed two customer types, those being window shoppers and loyal buyers, who could be described as having high intensity of interaction but not buying a lot and vice versa. Their clusters were checked by silhouette coefficients which proved the cohesion and separation. As far as the strategic perspective is concerned, the results indicated that the segment-based marketing offers, including personalized discounts to the so-called window shoppers, resulted in a better conversion rate.

Patel et al. (2024) performed a comparison of the Support Vector Machine (SVM) and the XGBoost model in the prediction of customer retention probability in a subscription-based retail platform. The data was highly skewed with a very small number of churners as compared to loyal customers hence a good candidate of resampling methods. Data was pre-processed using SMOTE to oversample churn cases, categorical encoding and standard scaling of continuous predictors. Experimental findings showed that XGBoost was significantly more effective than SVM, especially in recall and AUC making it the choice variant that could capture subtle patterns in minority-class behaviour. This paper is also very pertinent to the problem of prediction of campaign responses since it points out the effectiveness of boosting techniques compared to traditional classifiers when the data set is class-imbalanced an aspect that was dealt with in the project through oversampling and valid evaluation measures.

Rodriguez et al. (2022) investigated the application of the Gaussian Mixture Models (GMM) in customer segmentation in online fashion retail. GMM also allows overlapping clusters, unlike the partition-based methods like K-Means that tend to be useful in such domains where customer behaviour is not always partitionable. Pre-processing was carried out by one-hot encoding of categorical attributes of the products, variance thresholding, and normalization of purchases frequencies. Such results showed more focused segmentation with more subtle behaviours, including price-sensitive and brand-loyal customers. Although GMM was more insightful on behaviour, it was highly cumbersome and not scalable like K-Means. This is an important comparison to the research, because it justifies the appropriateness of K-Means in your pipeline to segment in real-time, but it recognizes the interpretive richness of probabilistic models, such as GMM.

Iyer et al. (2023) concentrated on explainability of an ensemble model, Random Forest, to predict the customer response to grocery retail campaigns. The pre-processing steps included imputation of missing demographic data, one-hot encoding of categorical variables and feature scaling. In addition to accuracy of classification, the study combined SHAP analysis to explain how purchase recency, use of coupons, and discount sensitivity contribute. The authors showed that adding explainability to the managerial confidence in automated decision-making systems narrowed the technical modeling-business adoption gap. Their observations speak to the twofold goal of the project, namely, to both maximize predictive accuracy and to provide interpretable models to marketing managers who need to act upon the results.

Mehta et al. (2024) used hierarchical clustering on transactional data of the supermarket to generate customer personas to optimize the loyalty program. They performed z-score standardization on the spending attributes, a dimensionality reduction with the aid of PCA, and they eliminated outliers on the basis of the Mahala Nobis distance. The dendrogram analysis found two main personas, one being the bulk buyers who had large volumes of transactions focussed on staple goods, and occasional shoppers who purchased sporadically across categories. The clusters were confirmed through silhouette analysis, which had high internal cohesion. As a managerial insight, the research highlighted the importance of hierarchical cluster as a guide in tiered rewards and resource dispensation of loyalty programs.

2.4 Gaps in the Literature

Previously conducted studies show the importance of clustering and predictive modeling when it comes to the study of customer behaviour. The majority of the research on segmentation and prediction are independent of each other with little work done to combine the two in a single framework of optimizing a campaign. Class imbalance is also a recurring problem even after trying resampling strategies, with a tendency to result in biased predictions. Interpretability is also an understudied area, with many of the popular ensemble models such as XGBoost and Random Forest being black boxes, which act as a barrier to business adoption. Moreover, most literature stops at intent prediction, without going beyond intent to post-intent behaviours including churn and repeated purchases. Lastly, methods such as Gaussian Mixture Models or deep learning have computational requirements that limit their applicability to real-time marketing deployment. According to these gaps, it is necessary to provide an integrated, interpretable and scalable framework integrating K-Means segmentation with more sophisticated classifiers (Random Forest, SVM and XGBoost) to provide real-campaign strategies.

2.5 Research Novelty

This research is unique in that it combines customer segmentation and predictive modelling in one framework of marketing campaign analysis. K-Means clustering is used to cluster the customers in terms of demographic and behavioural variables whereas Random Forest, SVM, and XGBoost are used to predict campaign responses. In contrast to the previous studies that generally consider segmentation and classification as different processes, this project considers both together in order to improve targeting accuracy. Random oversampling will be employed to handle the problem of class imbalances, and feature importance analysis will be used to guarantee interpretability so that the findings can be used by marketers.

Another addition is the incorporation of a real-world marketing campaign data that has not been utilized to a significant extent in the integrated frameworks. Most studies either treat intent prediction or segmentation separately, but this research shows how the two techniques can be combined so that not only predictions but also strategy recommendations can be both accurate and segment-specific. The project is promising to create a business-friendly solution to the practical marketing need by combining the powerful machine learning methods with the academic knowledge into a scalable solution that will both benefit the industry and expand academic knowledge.

3.Methodology

In this chapter, the research design approach, along with the methodological steps involved in the segmentation of customers and in predicting the response to the campaign using machine learning, are outlined. The methodology includes pre-processing of data, exploratory data analysis (EDA), feature engineering, clustering, predictive modeling, evaluation and interpretability.

3.1 Research Design and Approach

This Research uses a quantitative experimental research design because of the need to analyse the structured data and produce quantitative results. The proposed approach combines K-Means clustering with supervised classification models such as Random Forest, SVM and XGBoost to solve both the problems of segmentation and campaign response prediction. The first step of clustering is done to identify the customer personas of those with a High Value, High Potential and Low Value, which offer a behavioural view on the data. These cluster labels are then fed into the predictive models to test whether segmentation enhances accuracy of the campaign response, and can be compared directly to the models without clustering. To provide a sound evaluation, the models are evaluated on several performance measures such as accuracy, precision, recall, F1-score, and AUC-ROC, with SMOTE used to overcome class imbalance. Lastly, the score of feature importance based on Mutual Information and the XGBoost algorithm is utilized to identify the most significant predictors, so that the findings can be interpretable in the marketing strategy.

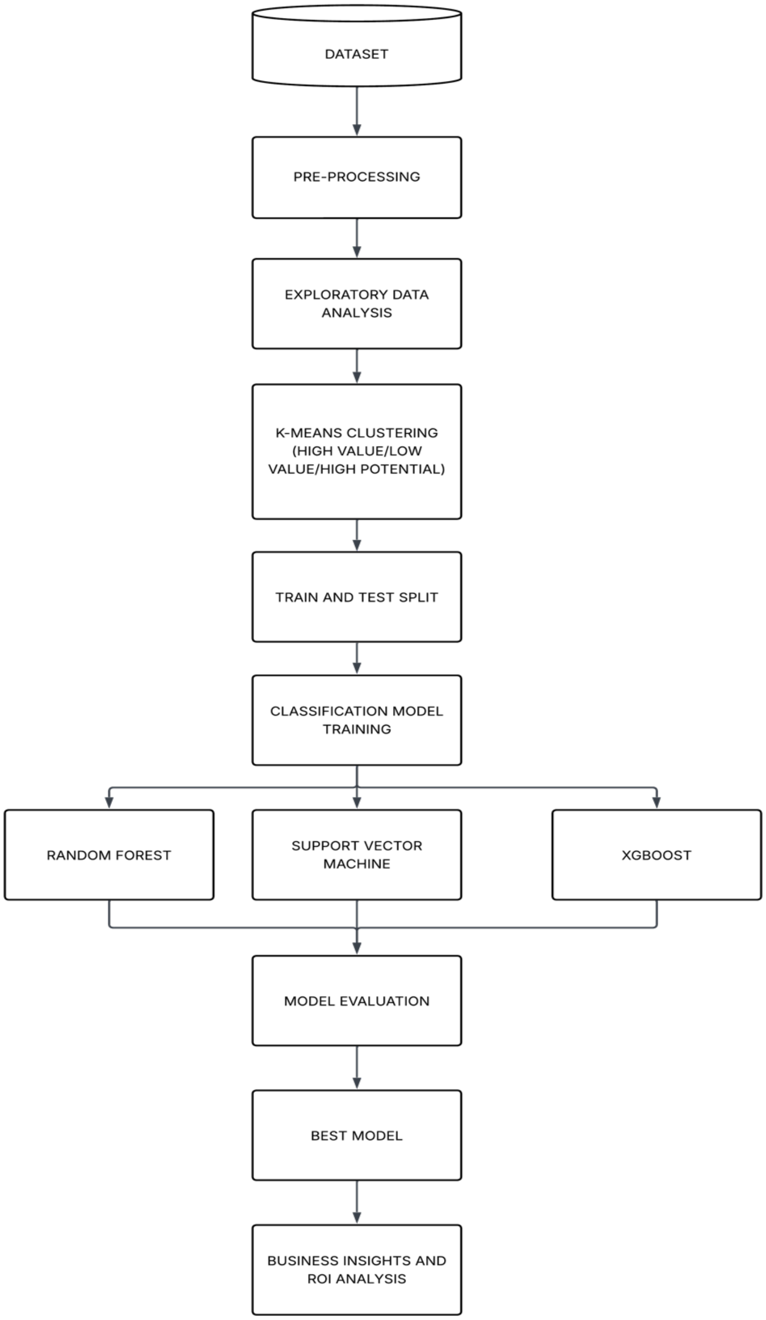

DATASET
PRE-PROCESSING
EXPLORATORY DATA ANALYSIS
K-MEANS CLUSTERING (HIGH VALUE/LOW VALUE/HIGH POTENTIAL)
TRAIN AND TEST SPLIT
CLASSIFICATION MODEL TRAINING
RANDOM FOREST
SUPPORT VECTOR MACHINE
XGBOOST
MODEL EVALUATION
BEST MODEL
BUSINESS INSIGHTS AND ROI ANALYSIS

Figure 1 Project Methodology

3.2 Dataset Description

The dataset employed in this project is made of 2,240 customer records and 28 variables that embrace demographic, household, behaviour and marketing campaign reaction specifics.

Feature Name	Data Type	Description
ID	Integer	Customer identification number (non-predictive, removed).
Year_Birth	Integer	Year of birth of the customer.
Education	Categorical	Level of education attained by the customer.
Marital_Status	Categorical	Marital status of the customer.
Income	Float	Yearly household income
Kidhome	Integer	Number of small children in the household.
Teenhome	Integer	Number of teenagers in the household.
Dt_Customer	Date	Date when the customer enrolled with the company.
Recency	Integer	Number of days since the last purchase.
MntWines	Integer	Amount spent on wine products in the last two years.
MntFruits	Integer	Amount spent on fruit products in the last two years.
MntMeatProducts	Integer	Amount spent on meat products in the last two years.
MntFishProducts	Integer	Amount spent on fish products in the last two years.
MntSweetProducts	Integer	Amount spent on sweet products in the last two years.
MntGoldProds	Integer	Amount spent on gold products in the last two years.
NumDealsPurchases	Integer	Number of purchases made with discounts.
NumWebPurchases	Integer	Number of purchases made through the company's website.
NumCatalogPurchases	Integer	Number of purchases made via catalog.
NumStorePurchases	Integer	Number of purchases made in physical stores.
NumWebVisitsMonth	Integer	Number of visits to the company's website in the last month.
AcceptedCmp1	Binary	Accepted offer in Campaign 1 (1 = Yes, 0 = No).
AcceptedCmp2	Binary	Accepted offer in Campaign 2 (1 = Yes, 0 = No).
AcceptedCmp3	Binary	Accepted offer in Campaign 3 (1 = Yes, 0 = No).
AcceptedCmp4	Binary	Accepted offer in Campaign 4 (1 = Yes, 0 = No).
AcceptedCmp5	Binary	Accepted offer in Campaign 5 (1 = Yes, 0 = No).

Complain	Binary	Whether the customer filed a complaint in the last 2 years.
Z_CostContact	Integer	Cost of contacting the customer.
Z_Revenue	Integer	Revenue generated if customer accepts a campaign.
Response	Binary	Target variable: acceptance of the last campaign (1 = Yes, 0 = No).

3.3 Data Preprocessing

3.3.1 Handling Missing Values

An exploratory check established that nearly all the variables were complete with the exception of Income that had 24 missing entries. The column median was used to impute the feature in order to retain the natural distribution and avoid the values due to extreme values in the column.

3.2.2 Feature Engineering

New attributes were created to be closer to customer behaviour and demographics. As an example, Age was calculated based on Year_Birth, and such categorical groups were assigned as Age_Group. The structure of households was measured by Total_Children (sum of Kidhome and Teenhome), and the partner status was reduced to two states (Has_Partner). Measures of spending and engagement were aggregated: Total_Spending (all products categories sum), Total_Purchases (sum of store, catalogue, deal, and online transactions), and Total_Acc (number of accepted campaigns). A derived measure, Conversion_Rate, was also added which indicates how well campaign acceptances convert against online visits.

3.2.3 Outlier Treatment

Maximum values were capped on various numerical variables in order to minimize the effect of extreme or unrealistic values. As an example, we set the spending limit to a maximum of 1,250,000 (MntMeatProducts <= 1,250,000, Total_Spending <= 2,500,000), as well as the purchase count to a maximum of 20 (NumWebPurchases <= 20, NumCatalogPurchases <= 20). This filtering minimised noise in the data and still retained representative customer behaviour.

3.2.4 Encoding Categorical Features

Categorical values were converted into numeric form so that it could be used in machine learning algorithms. Ordered encoding was used in education levels (SMA=0, D3=1, S1=2, S2=3, S3=4), age bands, and other ordered categories. The variable Has_Partner and Complain were coded as categorical variables (0,1). This ensured consistent representation without unnecessarily increasing dimensionality.

3.2.5 Feature Scaling

The numerical attributes also had different scales (e.g. income in thousands as opposed to units of purchases). The influence of numeric variables, all numeric variables were standardized with StandardScaler that centres values at zero mean and unit variance. This avoided the dominance of the models by features with higher magnitudes.

3.2.6 Class Balancing

The data was divided into two sets: training (80%) and testing (20%), based on stratified sampling, to maintain the balance between campaign responders and non-responders in both the sets. Since the target variable was imbalanced, Synthetic Minority Oversampling Technique (SMOTE) was used on training set to create additional synthetic instances of minority-class responders. This was the guarantee of the models learning balanced patterns and the untouched test set ensured fair evaluation.

3.2.7 Incorporating Clustering Labels

Unsupervised segmentation was performed based on five behavioural and financial attributes including Income, Total_Spending, NumWebVisitsMonth, Total_Purchases, and Total_Acc using K-Means. It was found that there are three customer groups, with the names High Value, High Potential, and Low Value. These cluster assignments were then combined into the modelling dataset as a new categorical feature (ClusterLabel).

3.4 Exploratory Data Analysis and Visualization

3.4.1 Demographic Analysis

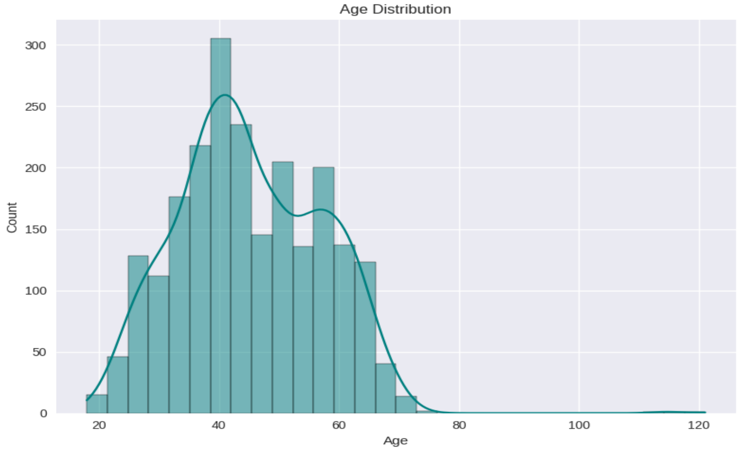


Figure 2 Age Distribution

The age structure indicates that the highest number of customers is between 30 and 60 years, and this is concentrated in the middle-aged group. This trend reveals that the company has been depending on demographics who have grown in age and developed purchasing power, therefore, the strategies of the campaign may require adjustment to target new markets and age groups.

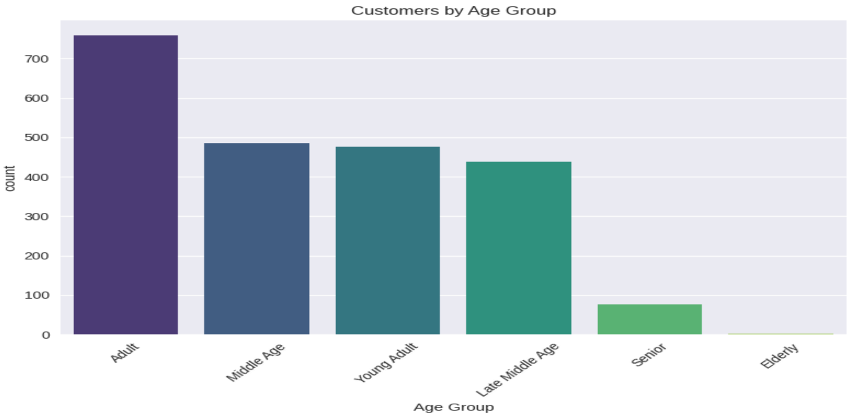


Figure 3 Customers by Age Group

The division into age groups shows that the categories of adult and middle age customers prevail. A relatively smaller proportion is the younger adults and the elderly customers. This disparity shows that although the existing campaigns are appealing to the established professional working customers, additional innovation might be necessary to attract the younger customers with the emergent spending power.

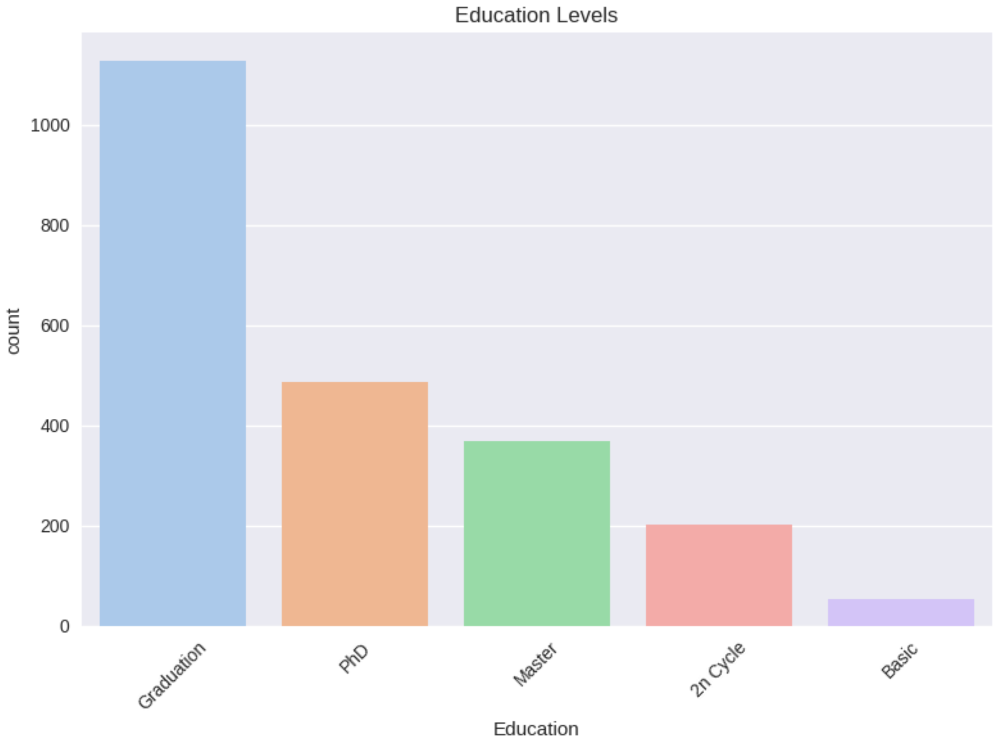


Figure 4 Education Levels

The Education analysis shows that the university educated customers make the biggest segment, then secondary education. It indicates that the customer base is believed to be relatively well-educated, and has implications on the design of the campaign: content-driven and premium product offers may have a stronger impact than generic promotions.

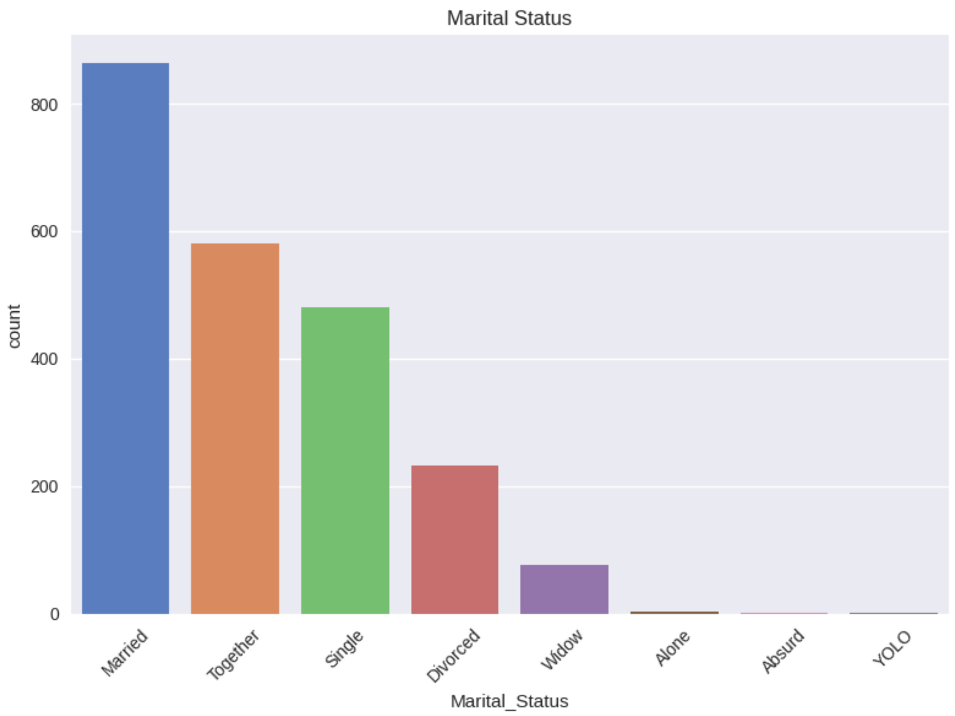


Figure 5 Marital Status

The marital status distribution shows that the dominant share of marital statuses is attributed to the categories of married and together that prove households to be significant units of consumption. Single and divorced customers are small but valuable niche which can be reached with customized packages..

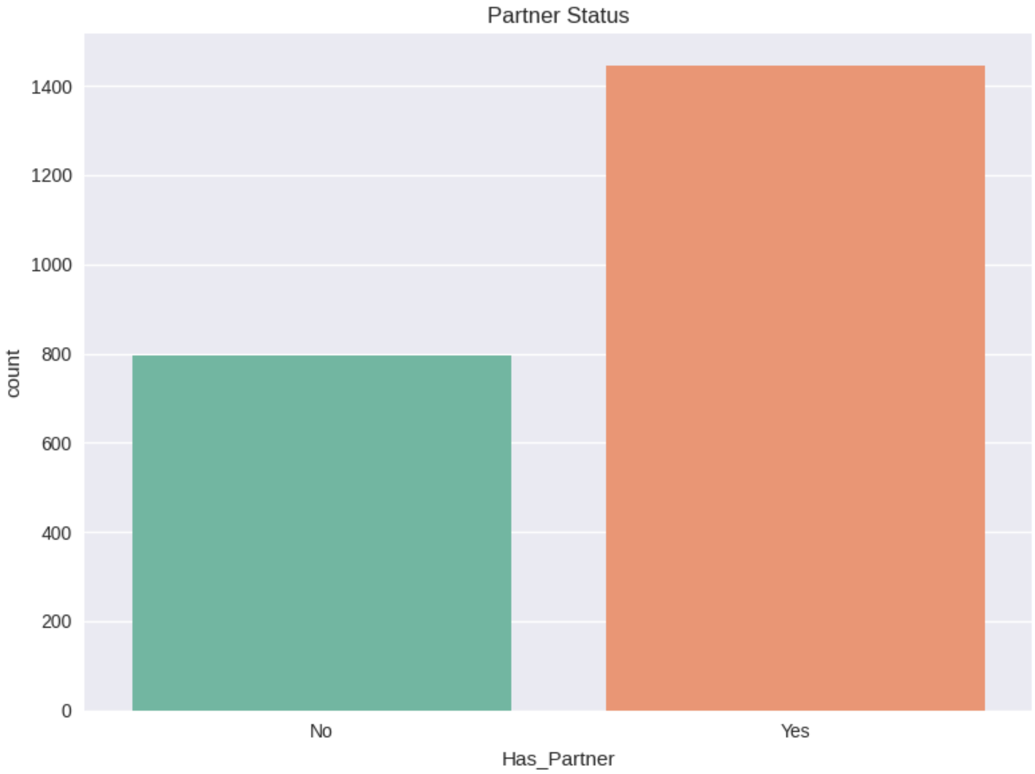


Figure 6 Partner Status

The estimated partner indicator means that most customers reside with partners, which reinforces the finding the family-oriented marketing is an essential strategy.

.

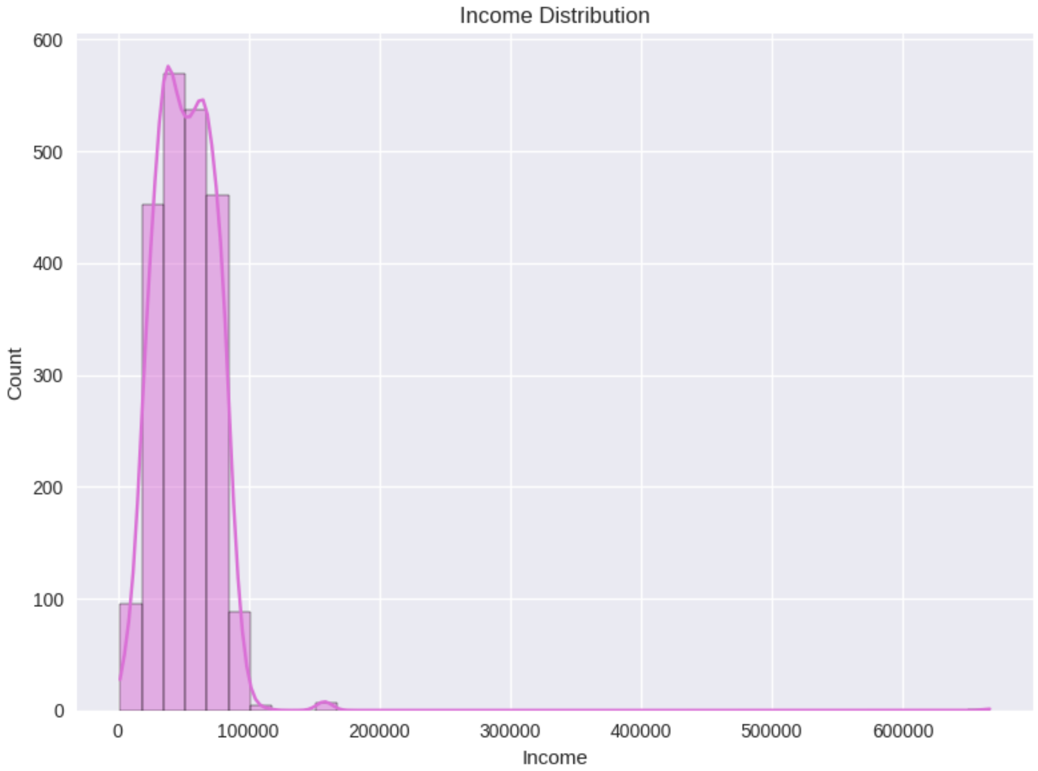


Figure 7 Income Distribution

The income is also skewed to the right with a long right tail of high-income customers but a considerable number at the lower-middle income brackets. This heterogeneity explains the use of segmentation strategy since the customers differ considerably in their spending power.

3.4.2 Purchase Behaviour Analysis

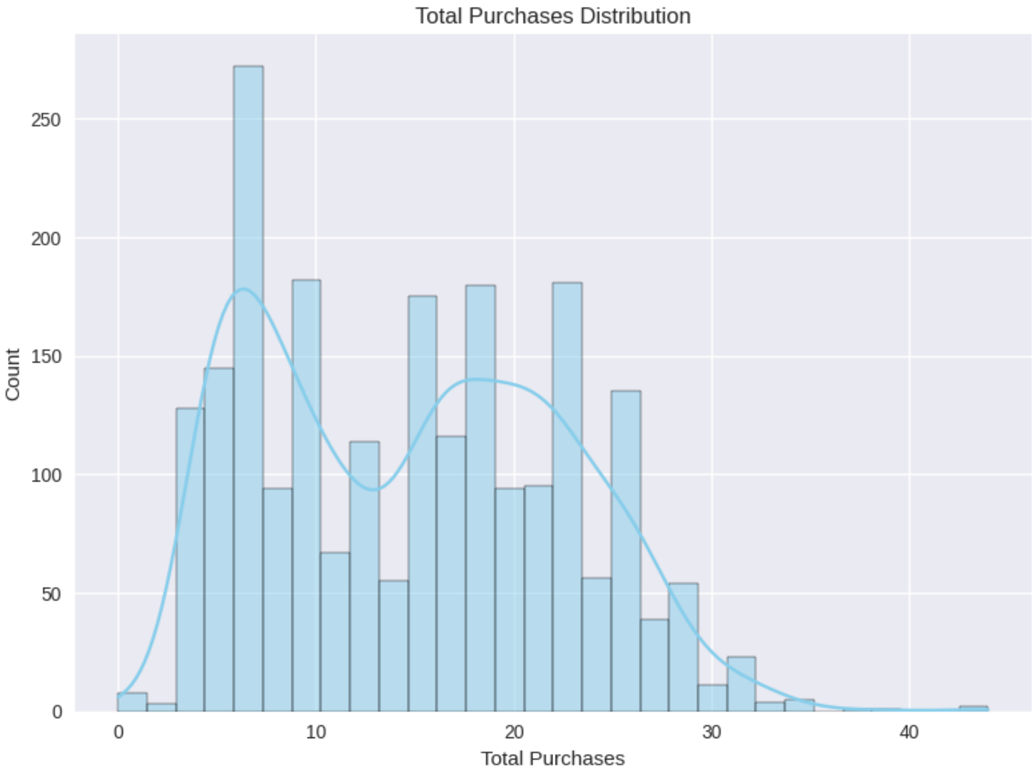


Figure 8 Total Purchases Distribution

The histogram of the total purchases indicates that majority of customers are moderate purchasers but a few customers are very active. These frequent buyers are prospective big-paying customers that could be cultivated with loyalty programs.

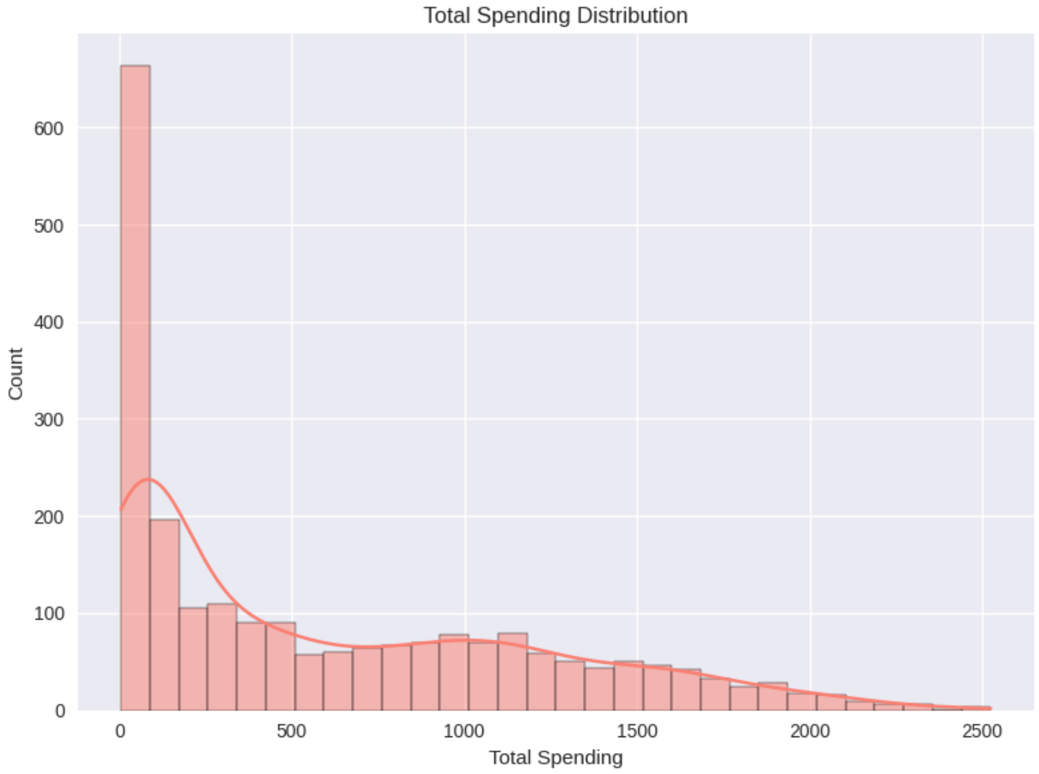


Figure 9 Total Spending Distribution

A similar skewed trend is observed in total spending with a few customers contributing to a large proportion of the spending, implying that small proportion of customers generates a high proportion of revenue.

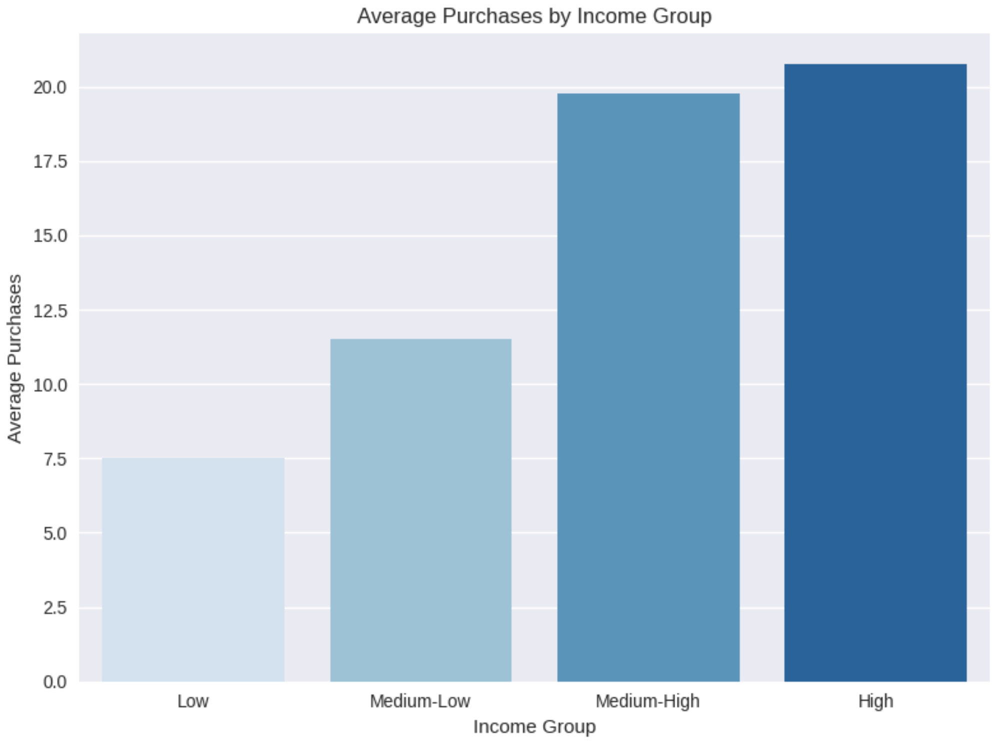


Figure 10 Average Purchases by Income Group

In the bar chart where there was a comparison between income and purchases, it can be observed that as the income level increases, the purchases being made increases, confirming the positive correlation between disposable income and purchasing frequency. But even poorer classes make a significant contribution, which means that the campaign strategies should be balanced between high-end attractive offers and low-cost bundling.

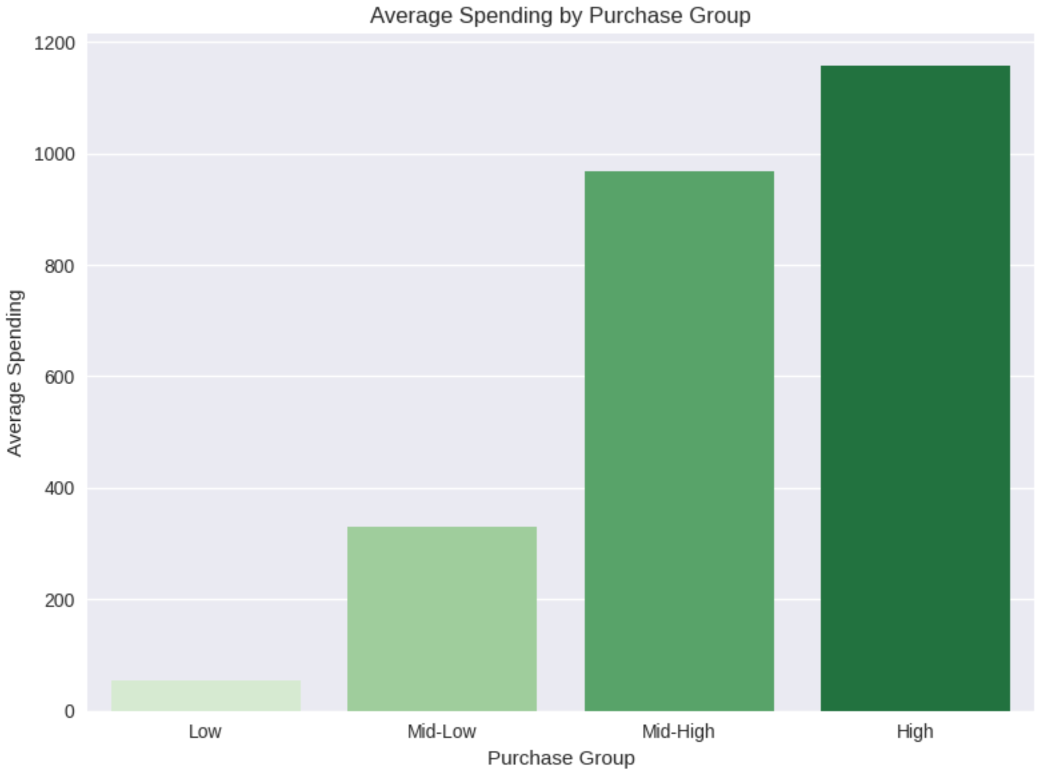


Figure 11 Average Spending by Purchase Group

The spending gradient in terms of purchase groups shows that customers who have high frequency of purchases also have high spending levels. This result supports the business case of cross-selling and upselling to high-frequency buyers, who are already highly engaged.

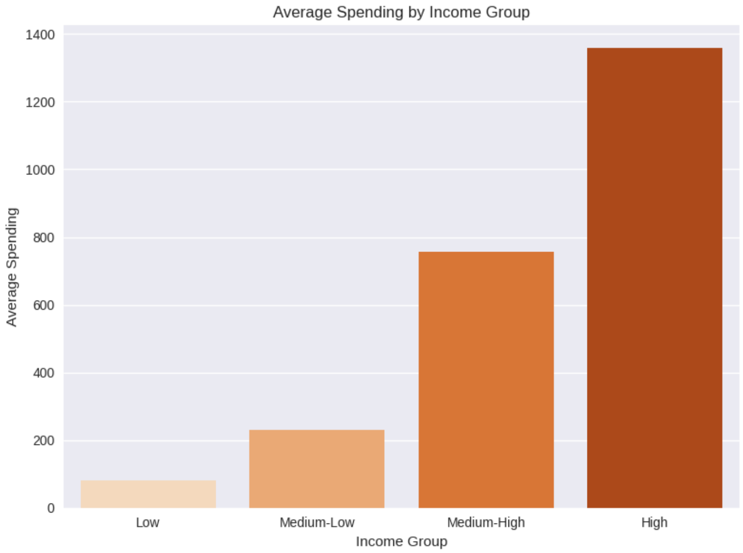


Figure 12 Average Spending by Income Group

Spending by income group, there is dominance of the higher-income customers on overall spending. This makes income-sensitive segmentation important because such customers are the most lucrative premium campaign targets.

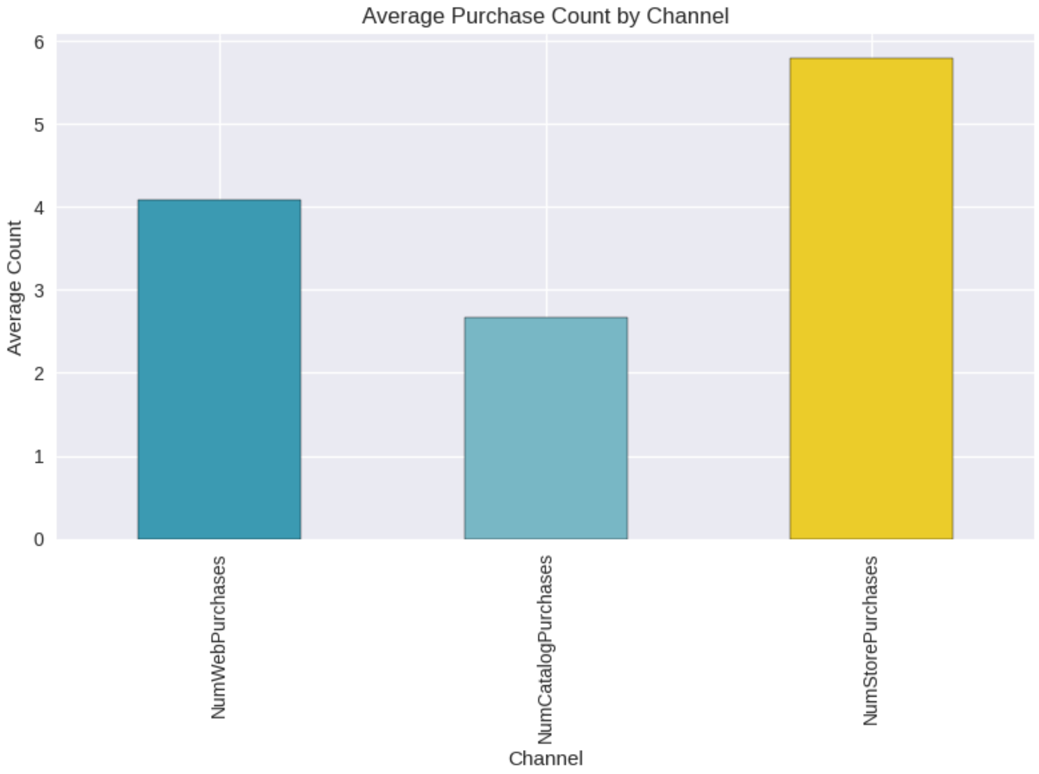


Figure 13 Average Purchase Count by Channel

The comparison of the purchase channels (web, catalog, and store) shows that the purchase in the physical stores prevails, and online channels are not used regularly. This does not mean that digital marketing is not relevant, but clearly it would be important to have in-store promotions to drive customer purchases.

3.4.3 Engagement Attributes Analysis

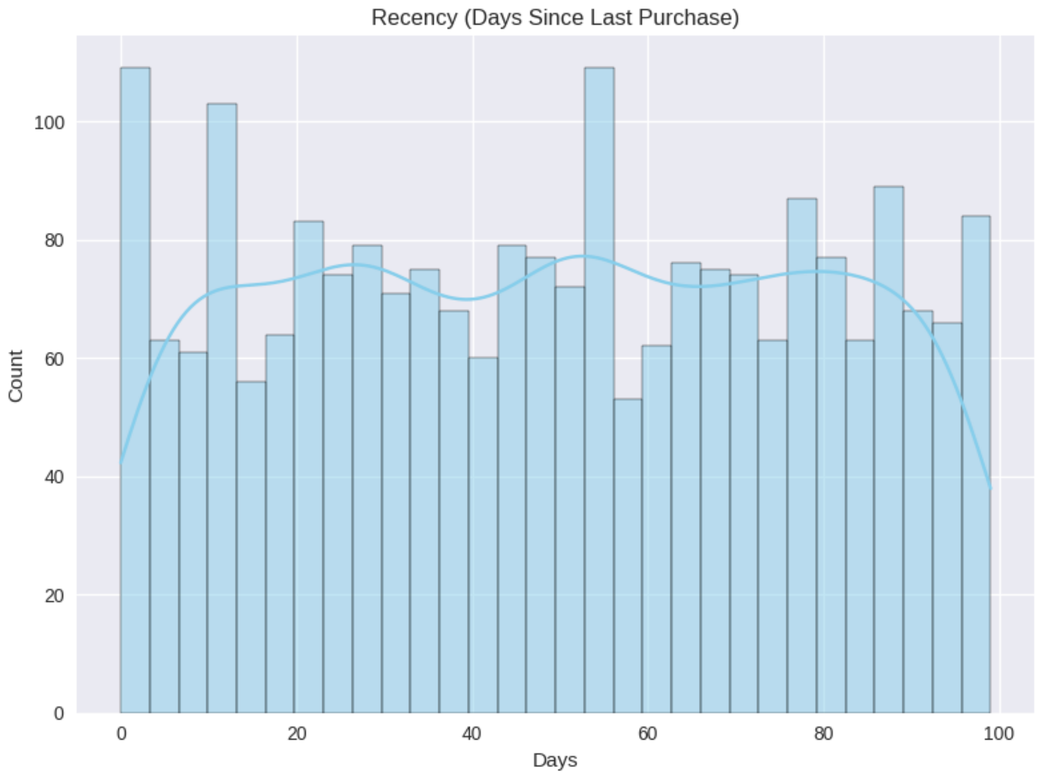


Figure 14 Recency Distribution

The segmentation of Recency reveals that there is a large segment of customers who have not made any purchase in a long time and thus they are inactive and may need to be reactivated. Customers who are engaged/recently engaged make up a smaller base and are the immediate target of the campaign.

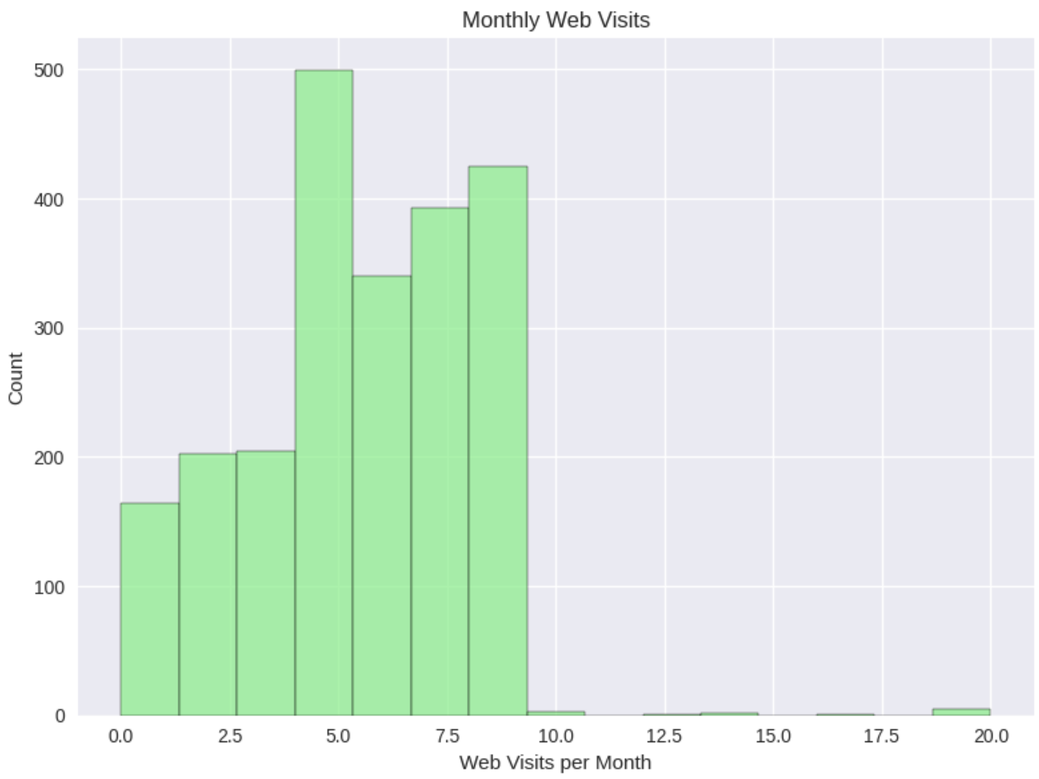


Figure 15 Monthly Web Visits

The web visit analysis indicates that the overall digital engagement is not high with the majority of the customers making less than 10 visits each month. This means that the online channels of the company are not utilized to their full potential and that campaigns associated with online touchpoints should be restructured to increase engagement.

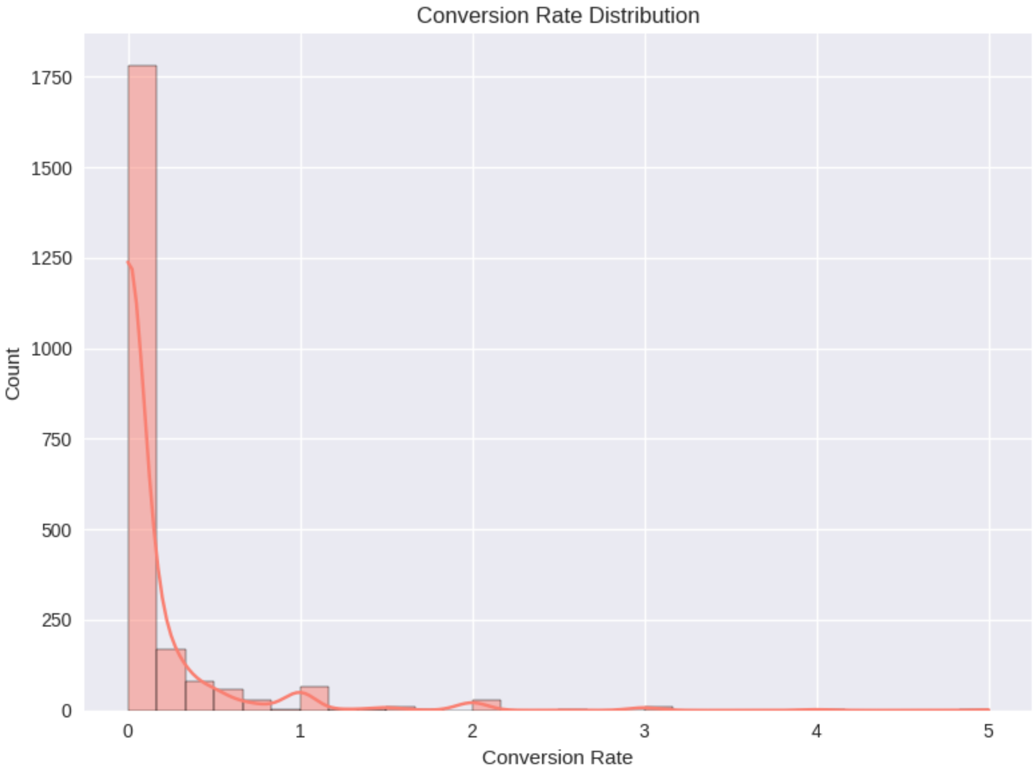


Figure 16 Conversion Rate Distribution

The distribution of the conversion rate indicates that the majority of the customers are characterized by very poor acceptance rates across campaigns further supporting the challenge of poor responsiveness. Nevertheless, a limited number of customers have more significant rates, which proves the existence of responsive segments, which could be increased through personalized approaches.

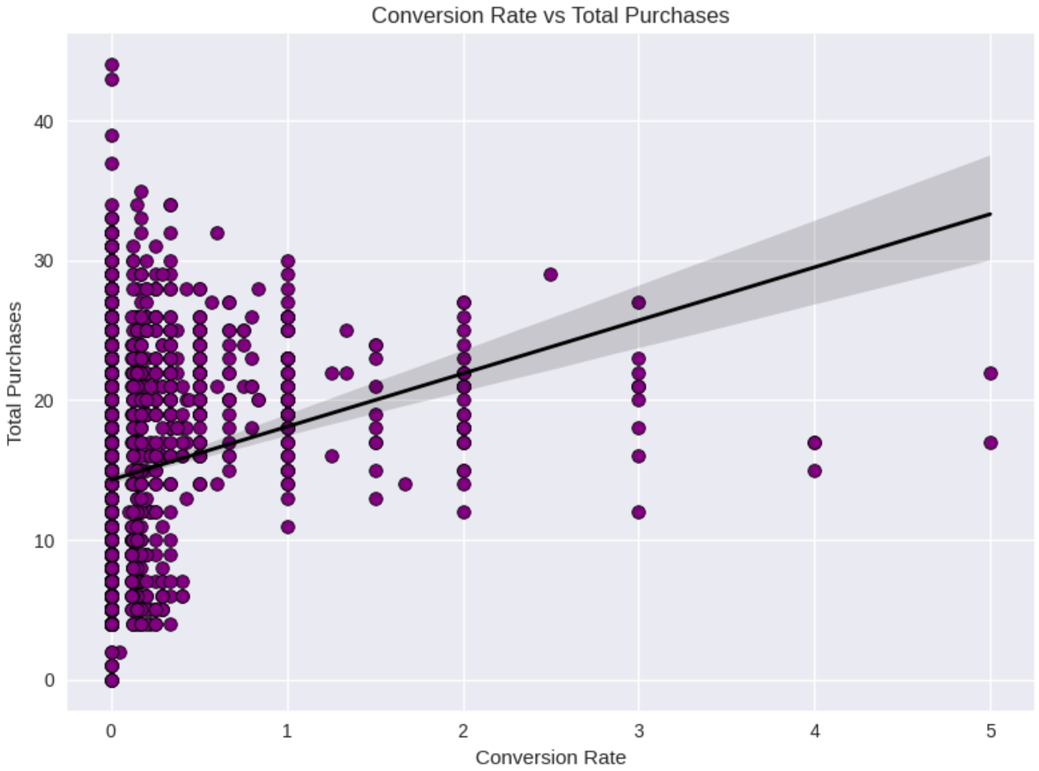


Figure 17 Conversion Rate vs. Total Purchases

The scatterplot indicates that low conversion rate is associated with low purchase volumes of customers. This connection indicates the direct relationship between campaign responsiveness and overall revenue which in turn makes conversion rate an important behavioural characteristic that is used in predictive modeling.

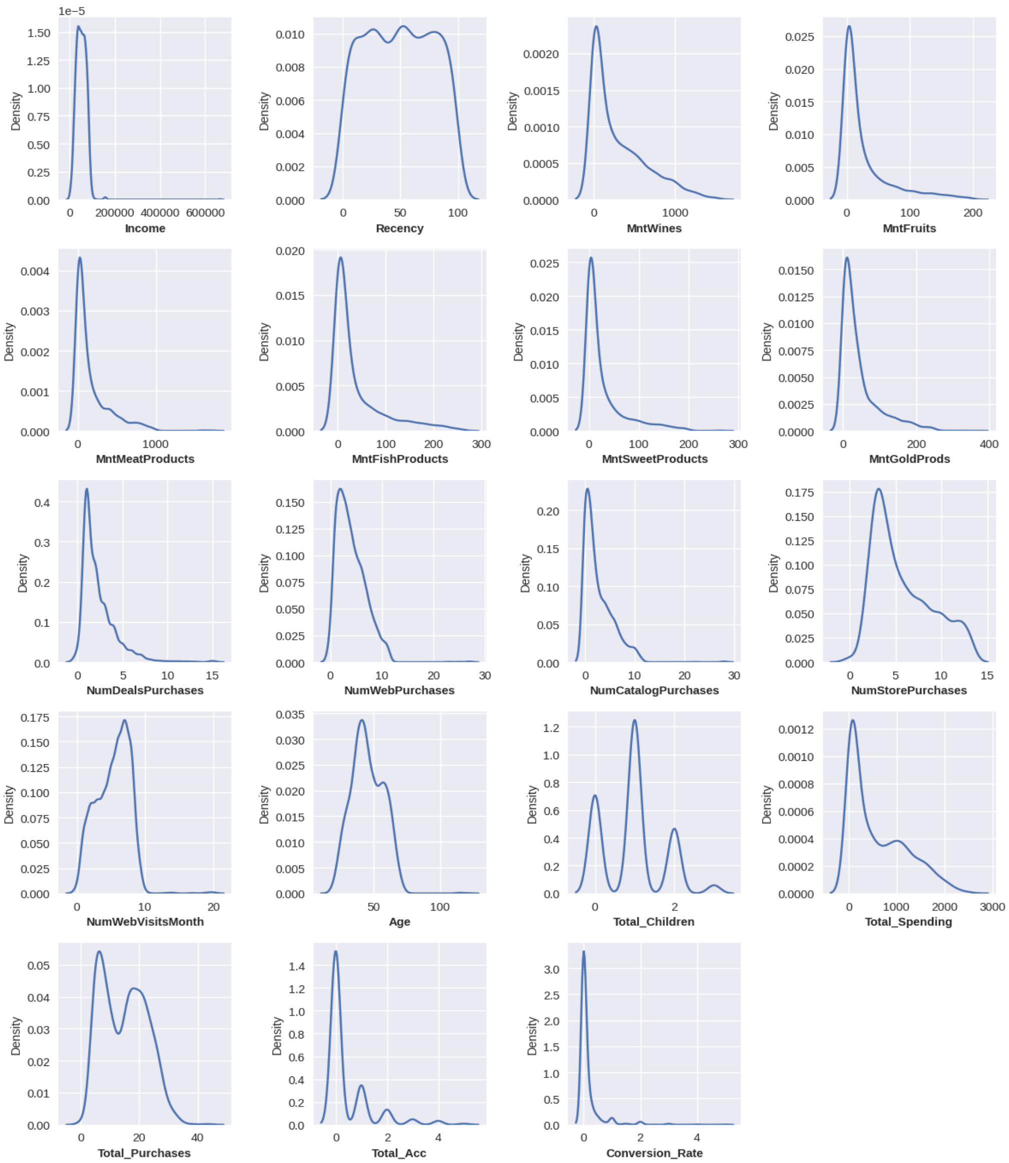
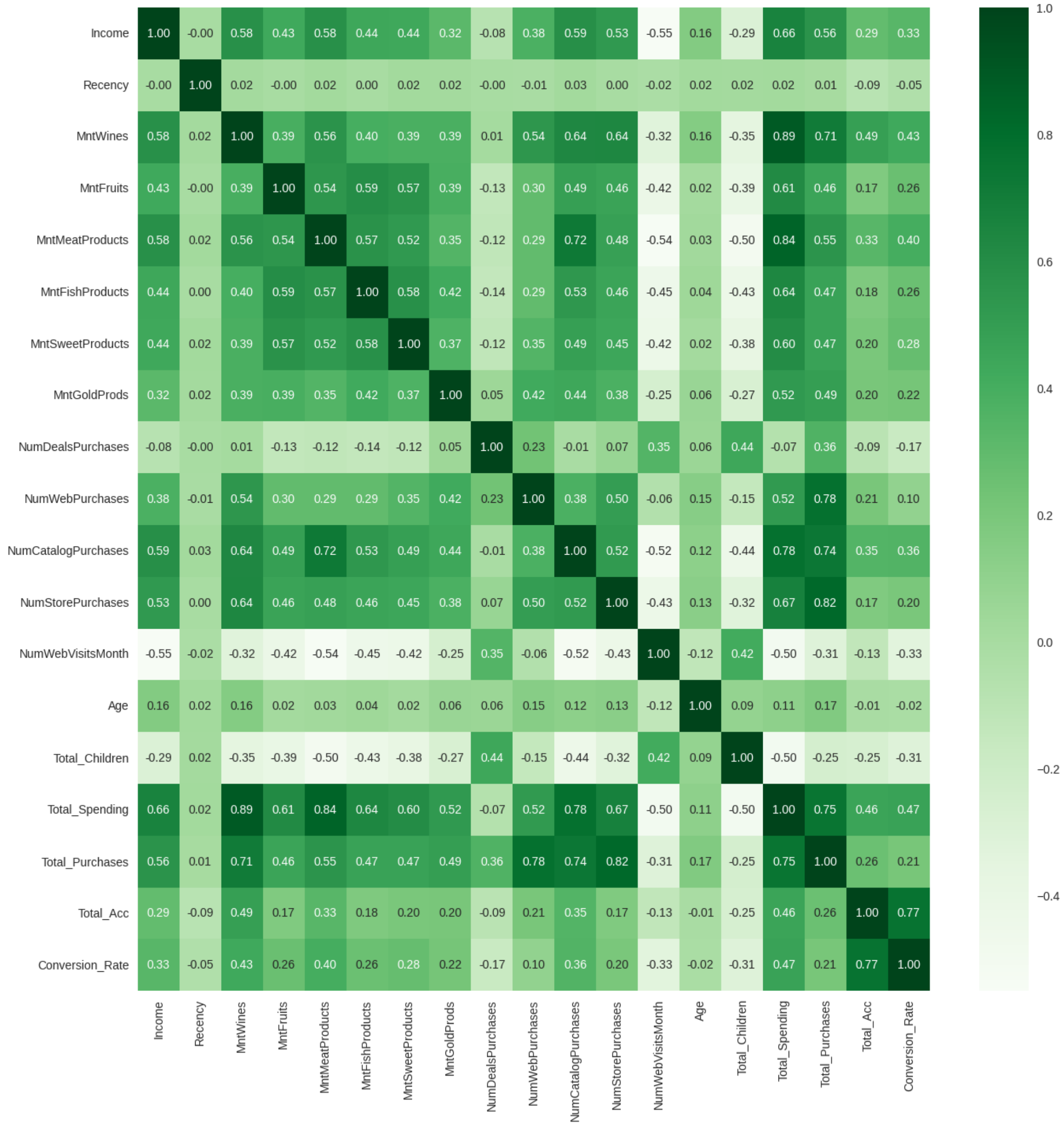


Figure 18 Distribution of Numerical Features

The kernel density charts depict the distributions of the most important numerical characteristics including revenue, recency, spending categories, and engagement variables. Except MntFruits, all other spending measures (e.g., MntWines, MntMeatProducts) are highly skewed to the right, which means that most customers spend a relatively small amount of money, whereas a few customers have exceptionally high spending values. On the same note, income also exhibits a ton of skewness and therefore, there is the presence of high-income customers who will be differently responsive to the campaign.



	Income	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Age	Total_Children	Total_Spending	Total_Purchases	Total_Acc	Conversion_Rate
Income	1.00	-0.00	0.58	0.43	0.58	0.44	0.44	0.32	-0.08	0.38	0.59	0.53	-0.55	0.16	-0.29	0.66	0.56	0.29	0.33
Recency	-0.00	1.00	0.02	-0.00	0.02	0.00	0.02	0.02	-0.00	-0.01	0.03	0.00	-0.02	0.02	0.02	0.02	0.01	-0.09	-0.05
MntWines	0.58	0.02	1.00	0.39	0.56	0.40	0.39	0.39	0.01	0.54	0.64	0.64	-0.32	0.16	-0.35	0.89	0.71	0.49	0.43
MntFruits	0.43	-0.00	0.39	1.00	0.54	0.59	0.57	0.39	-0.13	0.30	0.49	0.46	-0.42	0.02	-0.39	0.61	0.46	0.17	0.26
MntMeatProducts	0.58	0.02	0.56	0.54	1.00	0.57	0.52	0.35	-0.12	0.29	0.72	0.48	-0.54	0.03	-0.50	0.84	0.55	0.33	0.40
MntFishProducts	0.44	0.00	0.40	0.59	0.57	1.00	0.58	0.42	-0.14	0.29	0.53	0.46	-0.45	0.04	-0.43	0.64	0.47	0.18	0.26
MntSweetProducts	0.44	0.02	0.39	0.57	0.52	0.58	1.00	0.37	-0.12	0.35	0.49	0.45	-0.42	0.02	-0.38	0.60	0.47	0.20	0.28
MntGoldProds	0.32	0.02	0.39	0.39	0.35	0.42	0.37	1.00	0.05	0.42	0.44	0.38	-0.25	0.06	-0.27	0.52	0.49	0.20	0.22
NumDealsPurchases	-0.08	-0.00	0.01	-0.13	-0.12	-0.14	-0.12	0.05	1.00	0.23	-0.01	0.07	0.35	0.06	0.44	-0.07	0.36	-0.09	-0.17
NumWebPurchases	0.38	-0.01	0.54	0.30	0.29	0.29	0.35	0.42	0.23	1.00	0.38	0.50	-0.06	0.15	-0.15	0.52	0.78	0.21	0.10
NumCatalogPurchases	0.59	0.03	0.64	0.49	0.72	0.53	0.49	0.44	-0.01	0.38	1.00	0.52	-0.52	0.12	-0.44	0.78	0.74	0.35	0.36
NumStorePurchases	0.53	0.00	0.64	0.46	0.48	0.46	0.45	0.38	0.07	0.50	0.52	1.00	-0.43	0.13	-0.32	0.67	0.82	0.17	0.20
NumWebVisitsMonth	-0.55	-0.02	-0.32	-0.42	-0.54	-0.45	-0.42	-0.25	0.35	-0.06	-0.52	-0.43	1.00	-0.12	0.42	-0.50	-0.31	-0.13	-0.33
Age	0.16	0.02	0.16	0.02	0.03	0.04	0.02	0.06	0.06	0.15	0.12	0.13	-0.12	1.00	0.09	0.11	0.17	-0.01	-0.02
Total_Children	-0.29	0.02	-0.35	-0.39	-0.50	-0.43	-0.38	-0.27	0.44	-0.15	-0.44	-0.32	0.42	0.09	1.00	-0.50	-0.25	-0.25	-0.31
Total_Spending	0.66	0.02	0.89	0.61	0.84	0.64	0.60	0.52	-0.07	0.52	0.78	0.67	-0.50	0.11	-0.50	1.00	0.75	0.46	0.47
Total_Purchases	0.56	0.01	0.71	0.46	0.55	0.47	0.47	0.49	0.36	0.78	0.74	0.82	-0.31	0.17	-0.25	0.75	1.00	0.26	0.21
Total_Acc	0.29	-0.09	0.49	0.17	0.33	0.18	0.20	0.20	-0.09	0.21	0.35	0.17	-0.13	-0.01	-0.25	0.46	0.26	1.00	0.77
Conversion_Rate	0.33	-0.05	0.43	0.26	0.40	0.26	0.28	0.22	-0.17	0.10	0.36	0.20	-0.33	-0.02	-0.31	0.47	0.21	0.77	1.00

Figure 19 Correlation Heatmap of Features

The correlation matrix highlights relationships among customer attributes. There were high positive correlations among product spending variables (e.g., MntMeatProducts and MntWines) indicating consistency in the multi-category buying among high-value clients. It is also evident that total spending and total purchases are highly correlated, which shows that aggregate measures are robust. The opposite is true with web visits which are negatively correlated with spending, i.e. frequent online visits do not always lead to buying.

3.5 Customer Segmentation Through K-Means Clustering

3.5.1 Rationale for Clustering

Customer segmentation became applied to reveal the latent groups of behaviour in the data. The analysis results in natural customer clustering based on their income, spending and engagement characteristics using the K-Means clustering. This offers a deeper insight into campaign targeting and this can be used to assess both predictive performance with and without segmentation.

3.5.2 Determination of Optimal Cluster Number

It was important to see the number of groups to be formed before forming clusters, which is the most representative of the customer base. Two validation techniques, the Elbow Method and Silhouette Score, were used to make the selection of the optimal cluster count reliable.

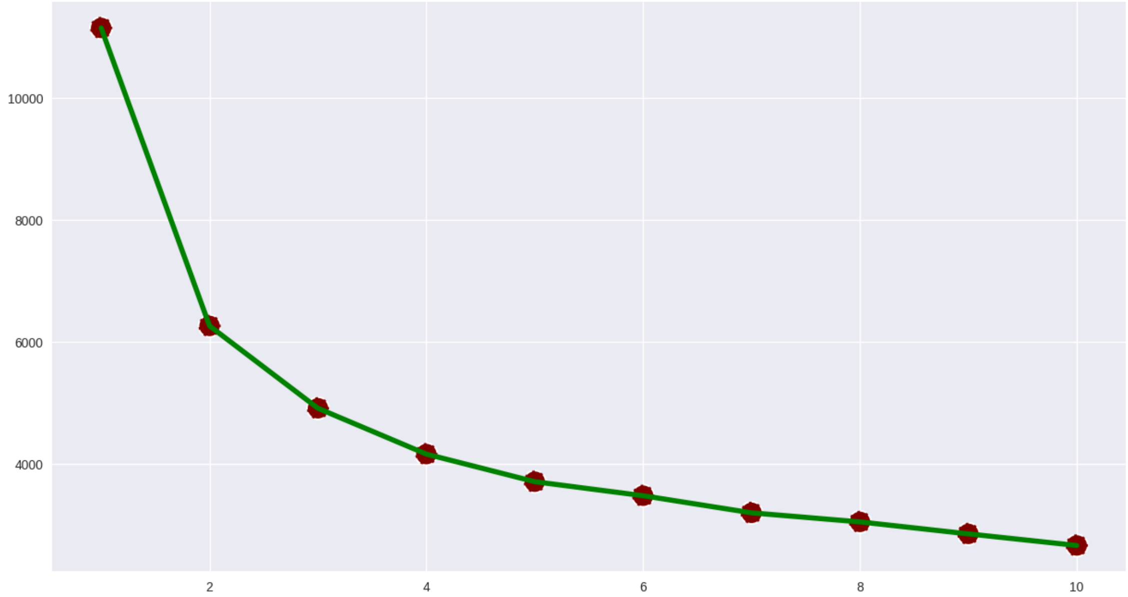


Figure 20 Elbow Method for Optimal Clusters

The elbow curve depicts a steep decrease up to **k = 3,** then the slowing, which means that three clusters are the optimal number of clusters.

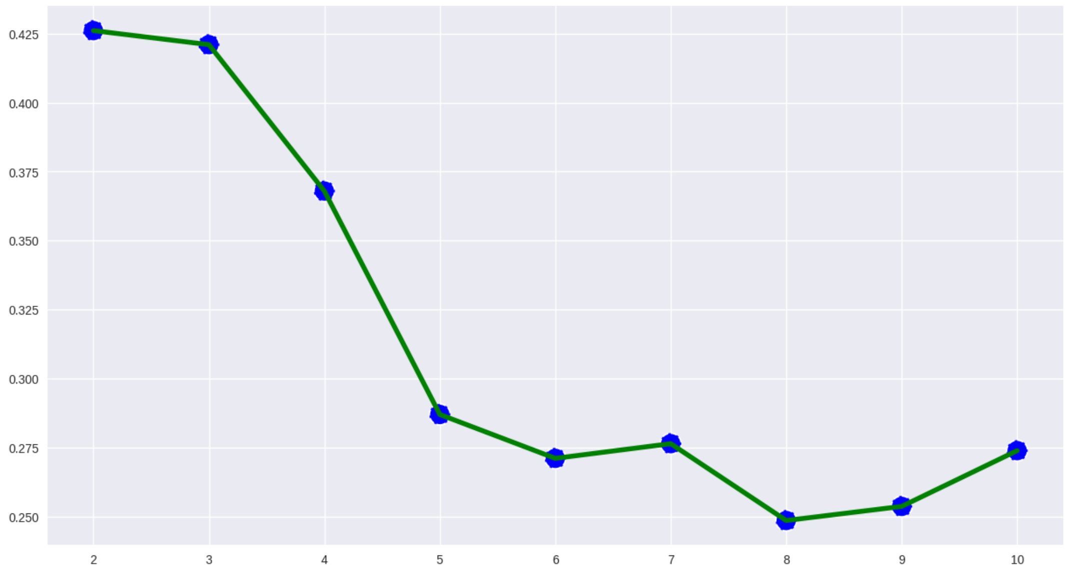


Figure 21 Silhouette Score for Cluster Validation

Silhouette coefficients are also highest at **k = 3**, again substantiating the elbow method result and the choice of a three-cluster solution.

3.5.3 Cluster Formation and Visualization

After determining the optimum number of clusters, K-Means was used on behavioural and financial attributes. Dimensionality reduction was done through PCA in order to visualize the customer segments in two dimensions.

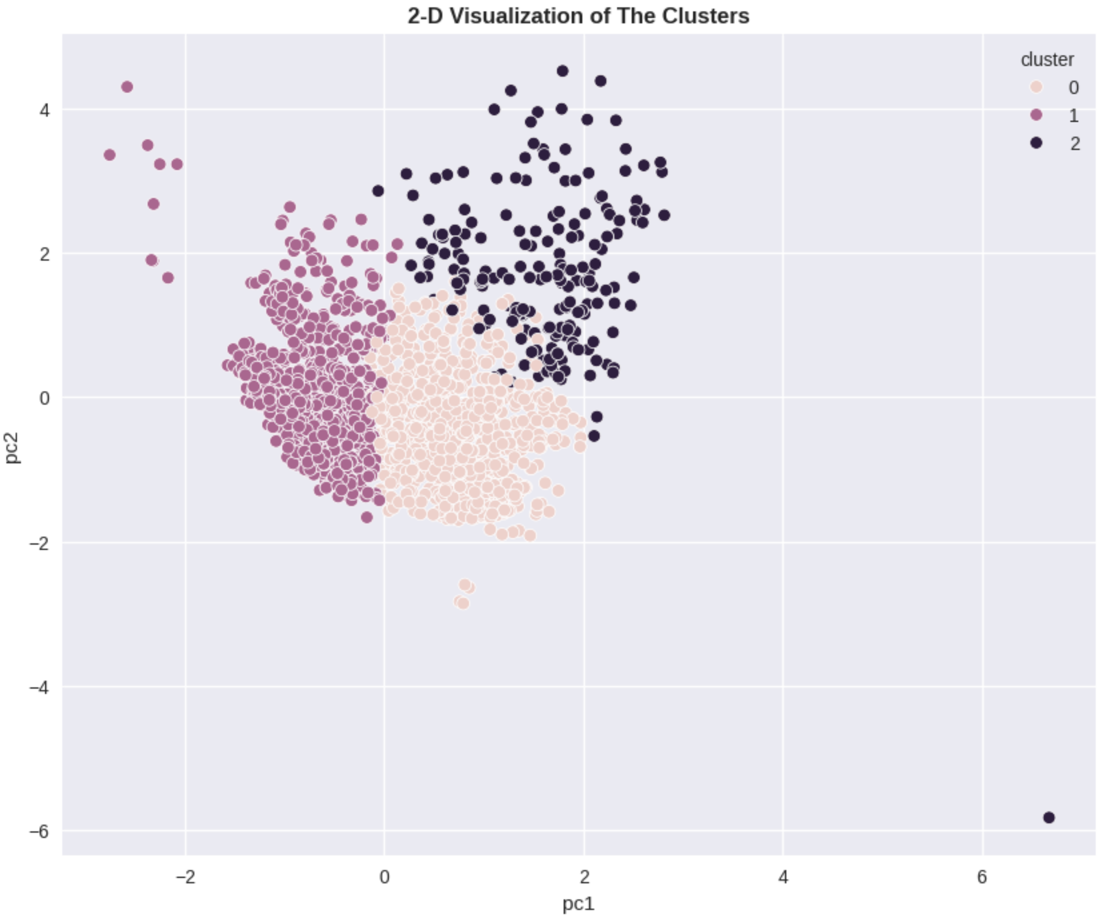


Figure 22 PCA-Based Visualization of Clusters

The PCA projection demonstrates that there are three clearly separated groups of customers, confirming that K-Means was able to identify separate groups of customers based on their purchasing and engagement behaviour

3.5.4 Cluster Distribution and Profiling

To know the practical implication of segmentation, this step gives information on the size and composition of each cluster so as to develop specific marketing strategies.

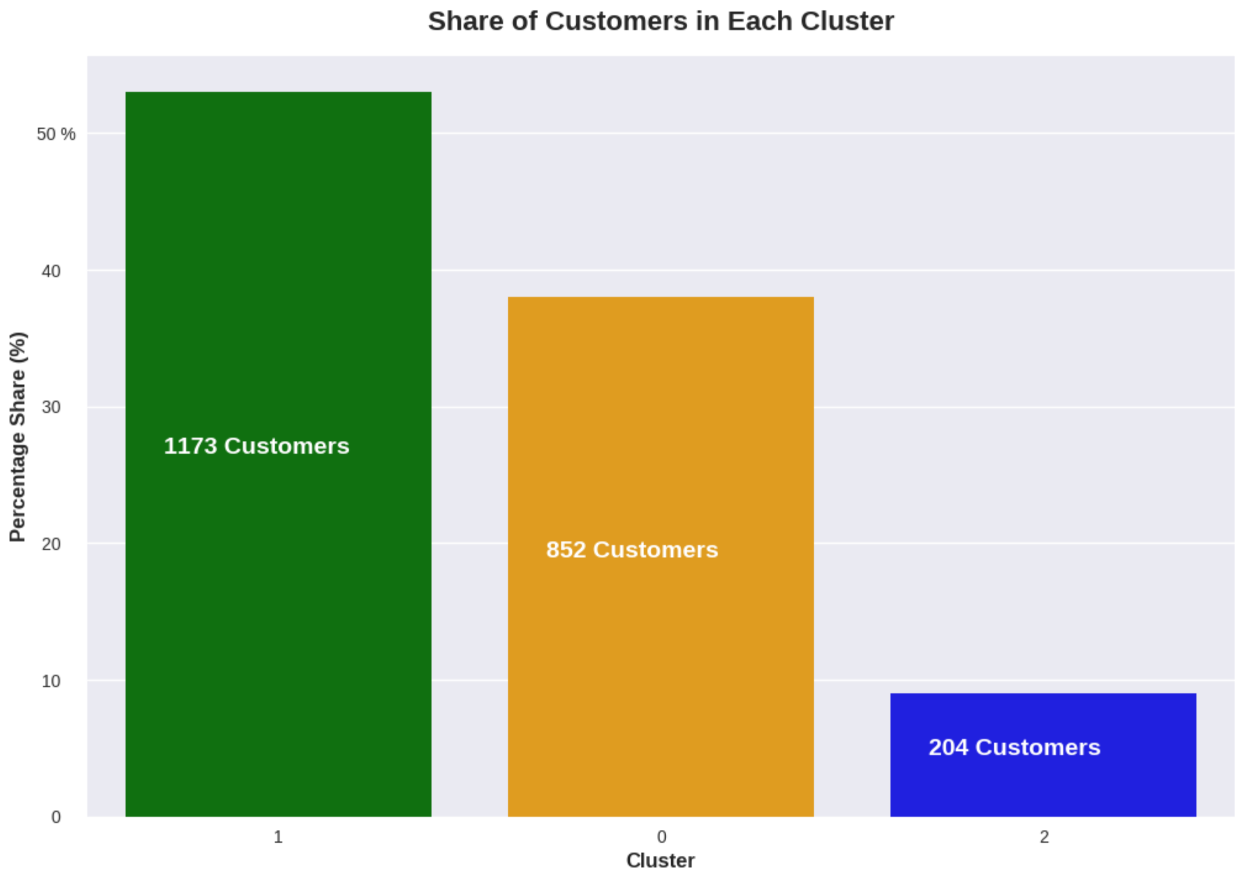


Figure 23 Distribution of Customers Across Clusters

The bar chart indicates that Low Value customers are the largest proportion followed by High Potential and High Value is the smallest group.

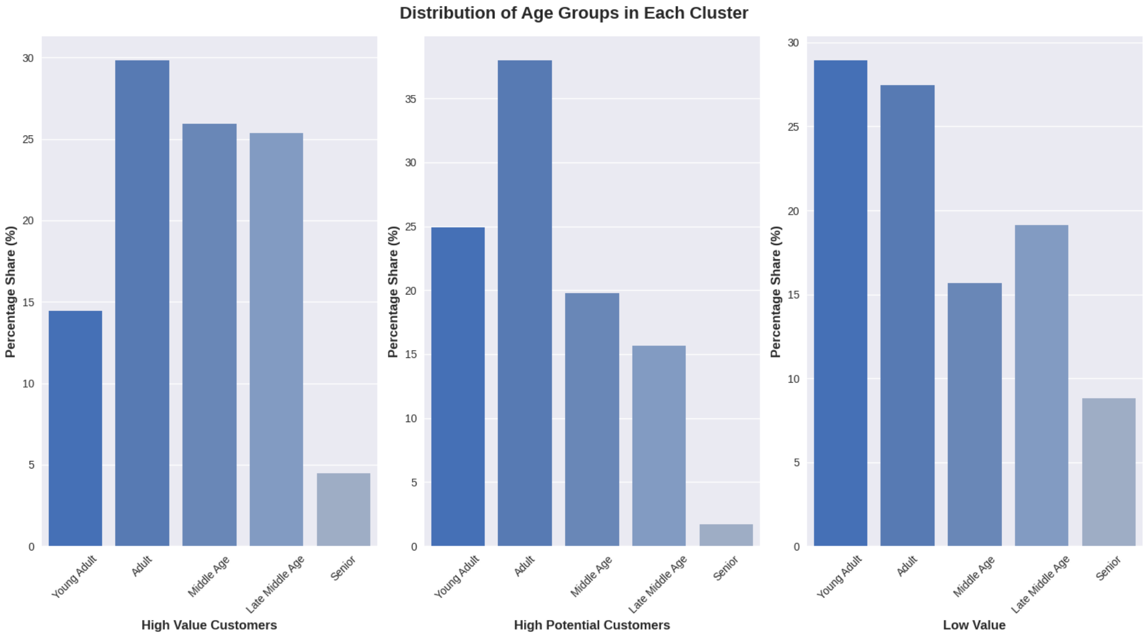


Figure 24 Distribution of Age Groups in Each Cluster

The age segmentation shows that the High Value segments are skewed towards the middle-aged age-group, whereas High Potential segments have a greater percentage of young adults, which implies potential growth.

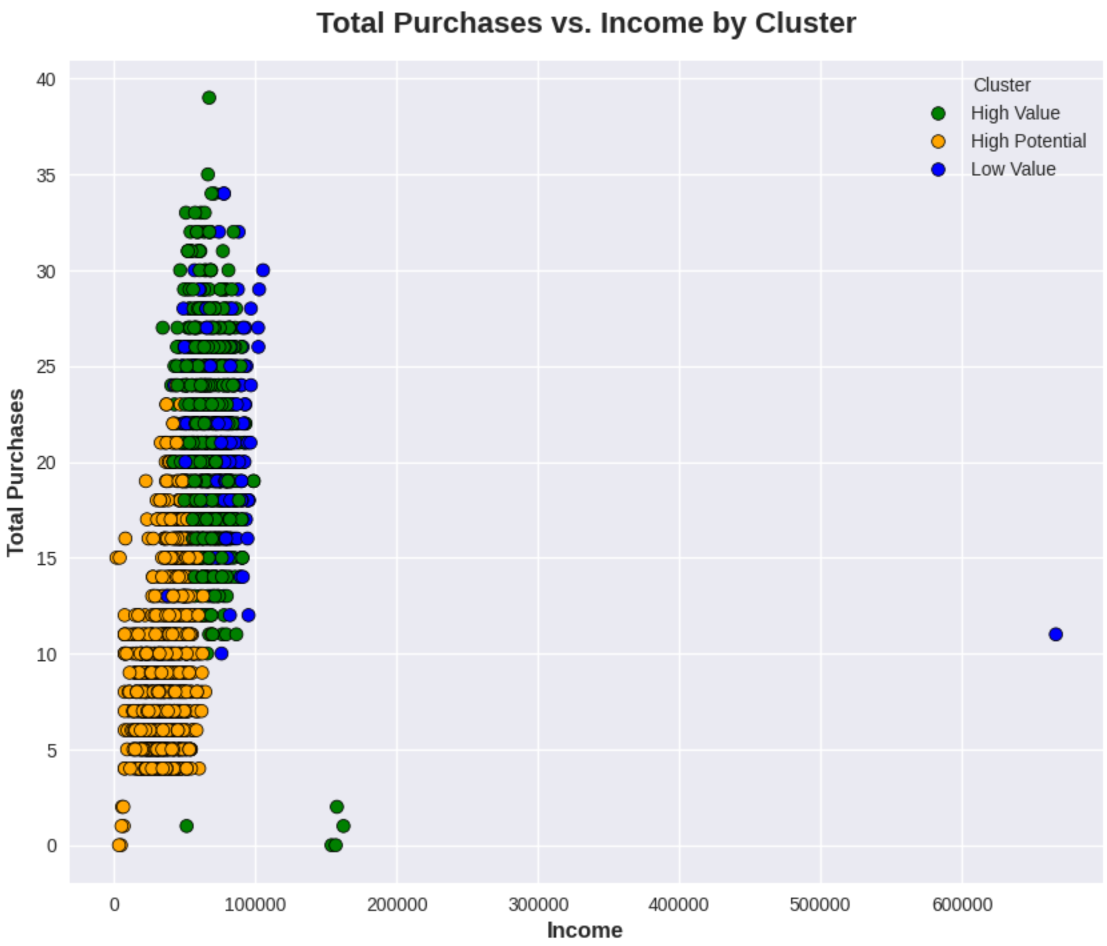


Figure 25 Income vs. Purchases by Cluster

Customers in the High Value cluster are positioned at higher income and purchase levels, while Low Value customers cluster around low-income, low-purchase areas.

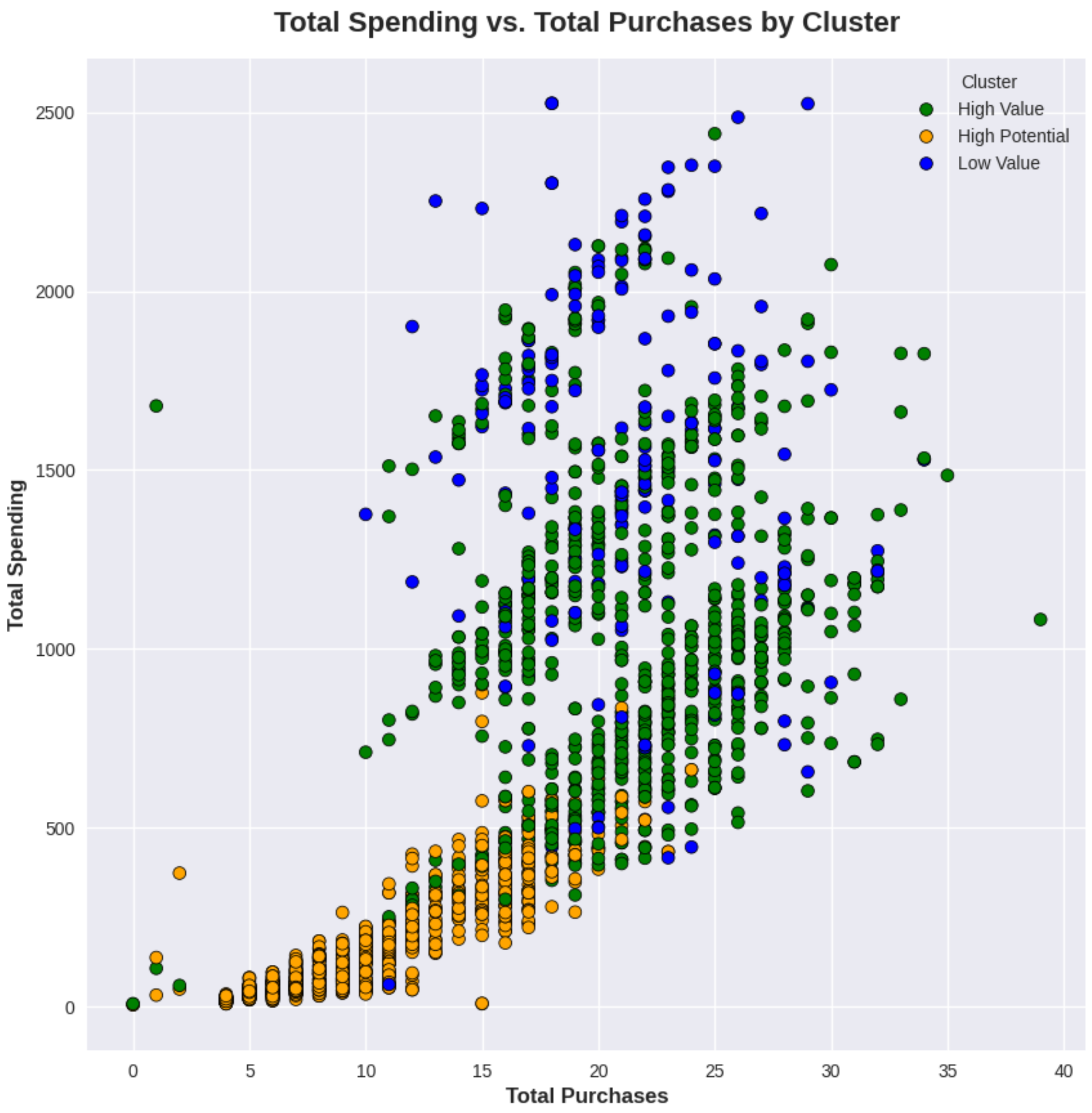


Figure 26 Total Spending vs. Purchases by Cluster

The scatter plot confirms that High Value customers spend and purchase more frequently, distinguishing them clearly from Low Value customers.

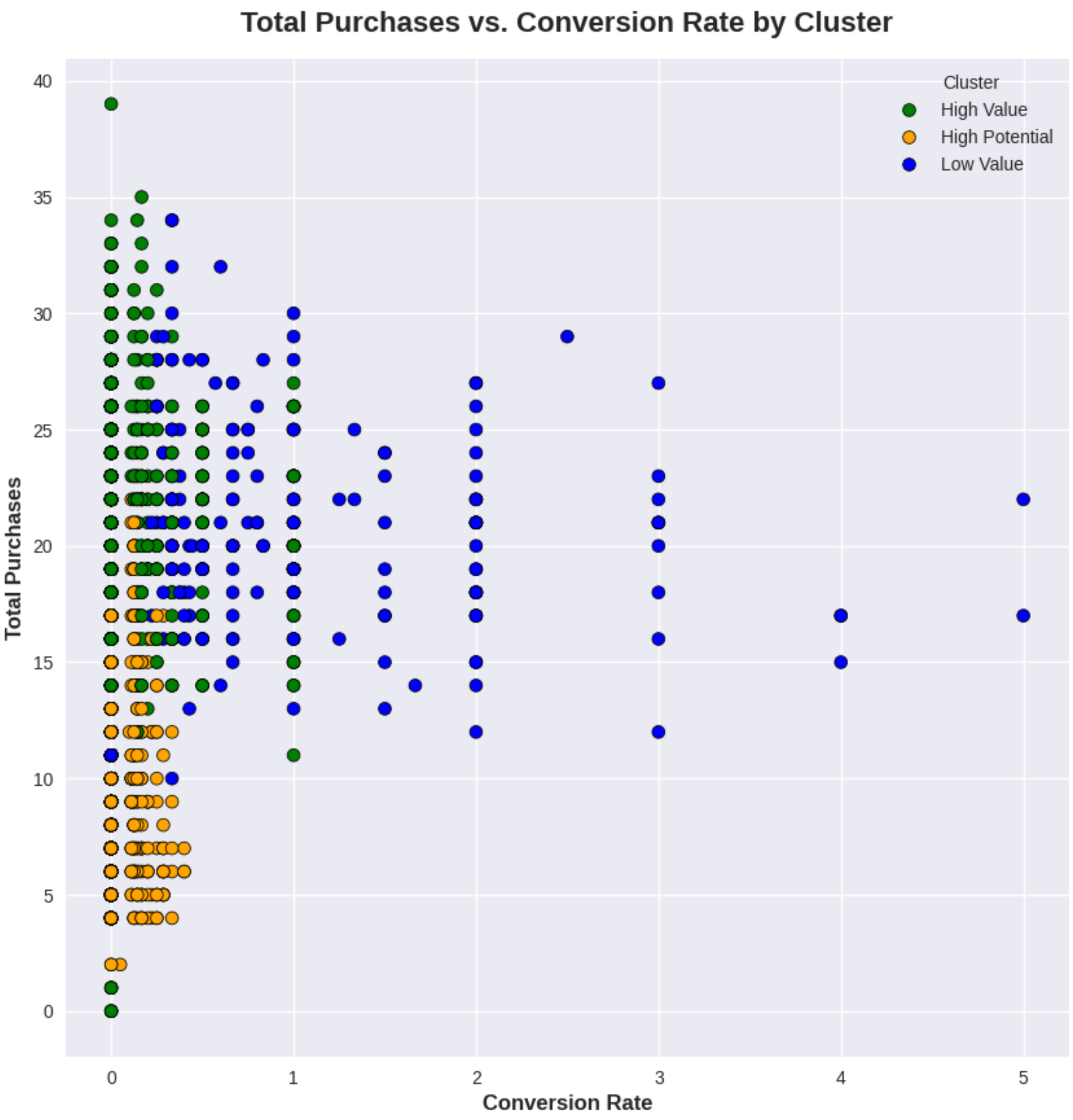


Figure 27 Conversion Rate vs. Purchases by Cluster

High Value customers show higher conversion rates, reinforcing their responsiveness to campaigns.

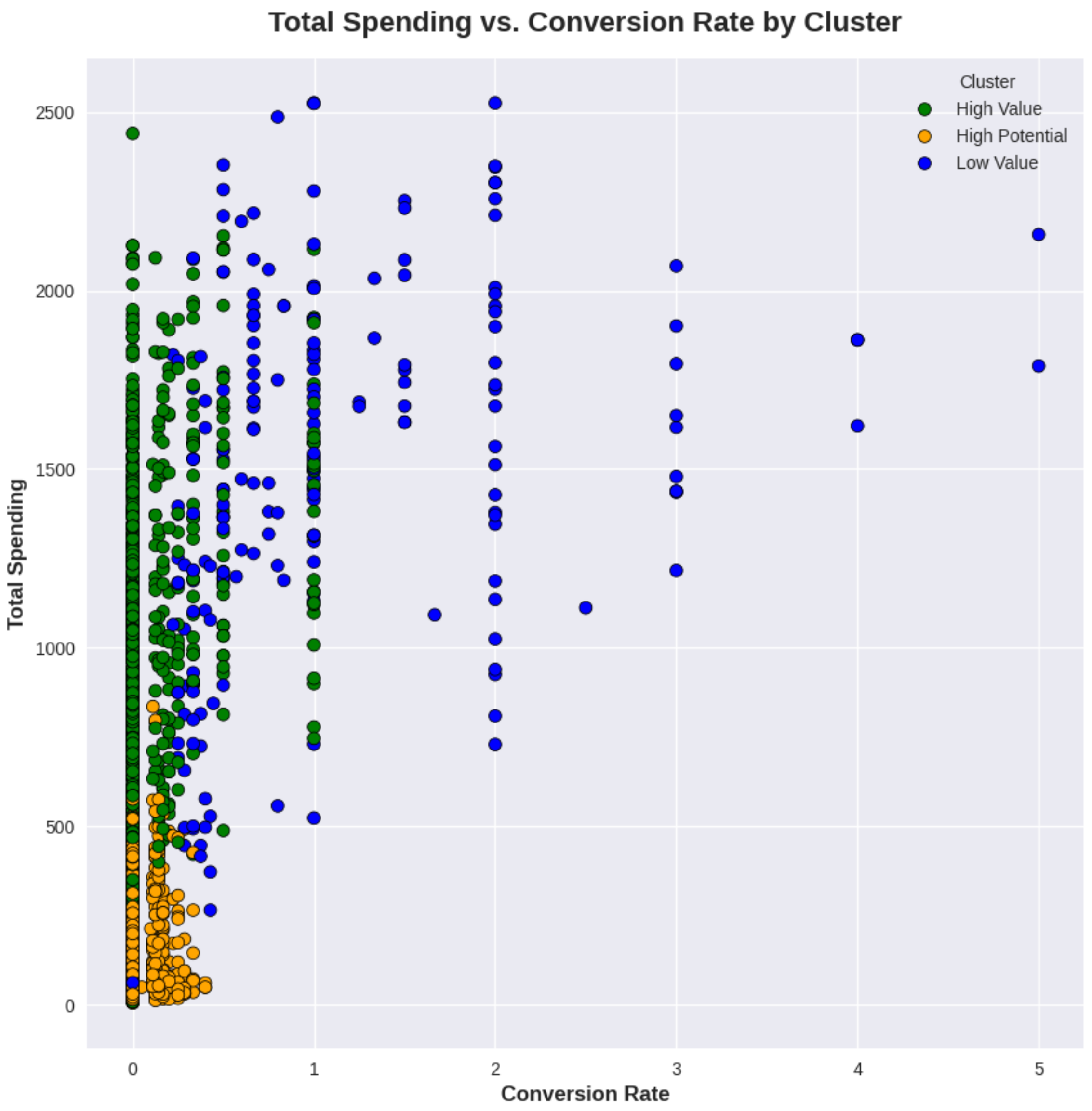


Figure 28 Conversion Rate vs. Spending by Cluster

Spending patterns reveal that Low Value customers remain low regardless of conversion, whereas High Value customers combine high spending with strong campaign responsiveness.

3.5.5 Integration with Predictive Modeling

Lastly, the identified clusters were used as a categorical variable in the predictive modelling pipeline. This step enabled testing of the hypothesis that the prediction of campaign response is better when customer segments are added.

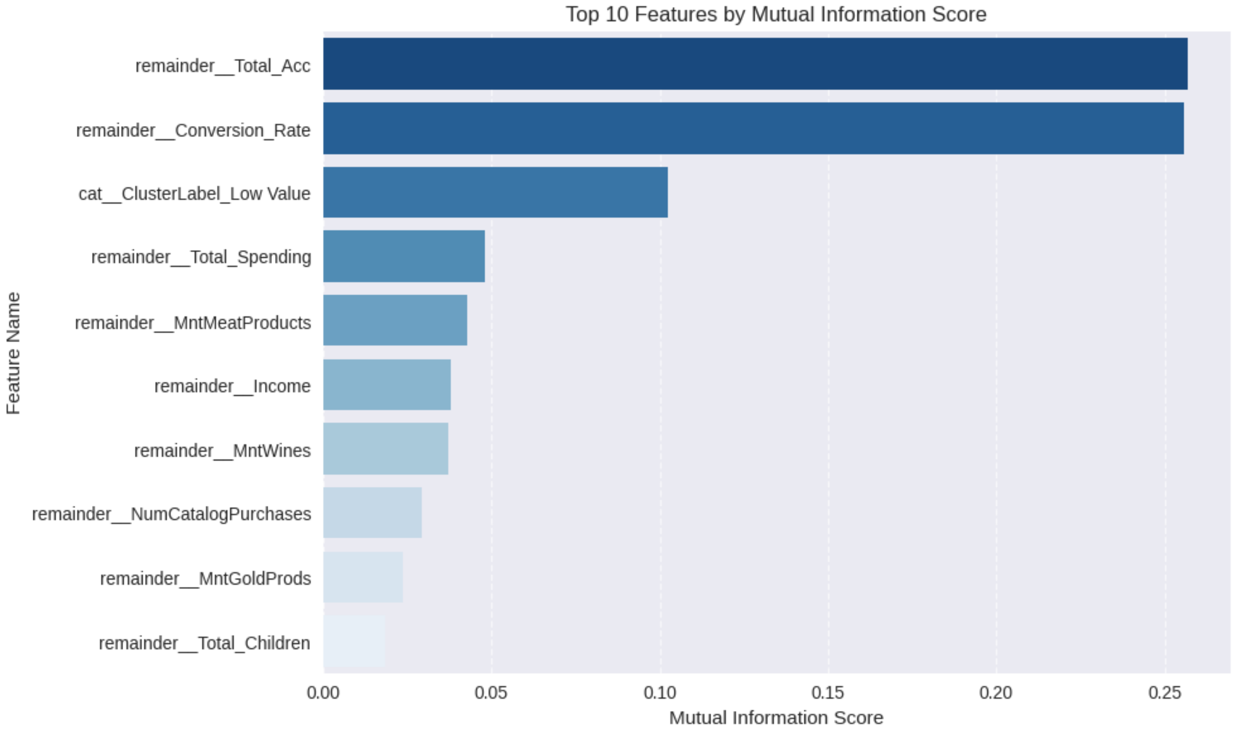


Figure 29 Top Features Ranked by Mutual Information

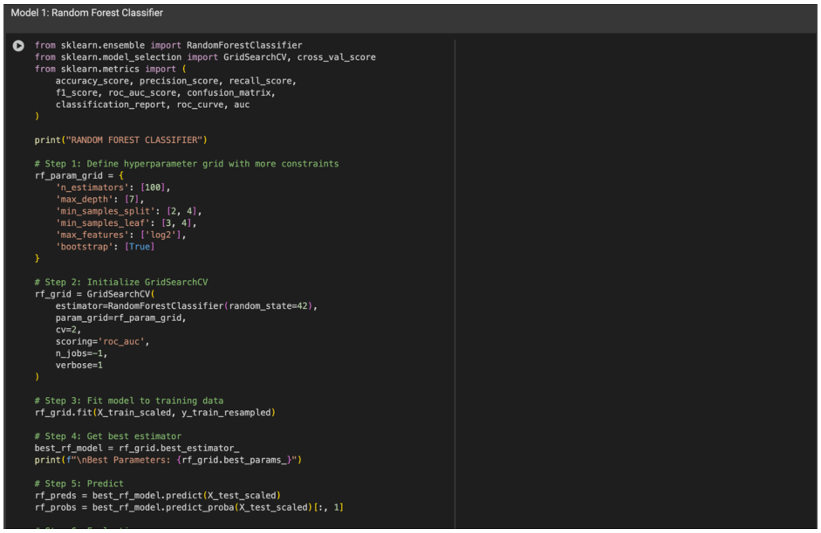
Feature importance analysis highlights that cluster labels contribute significantly to predictive performance, ranking alongside Total Campaign Acceptances and Total Spending as key predictors of campaign response.

3.5 Model Selection and Description

Predictive modeling was conducted to train three supervised machine learning models- Random Forest, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). GridSearchCV was used to optimize hyperparameters in each model. In order to make the results fair and robust, the dataset was divided into 80 and 20 percent training and testing data respectively and the Synthetic Minority Oversampling Technique (SMOTE) used to overcome the imbalance in campaign responses. StandardScaler was used on the feature scaling, and the OneHotEncoder was used on the categorical features.

3.6 Model Training

Random Forest Classifier :



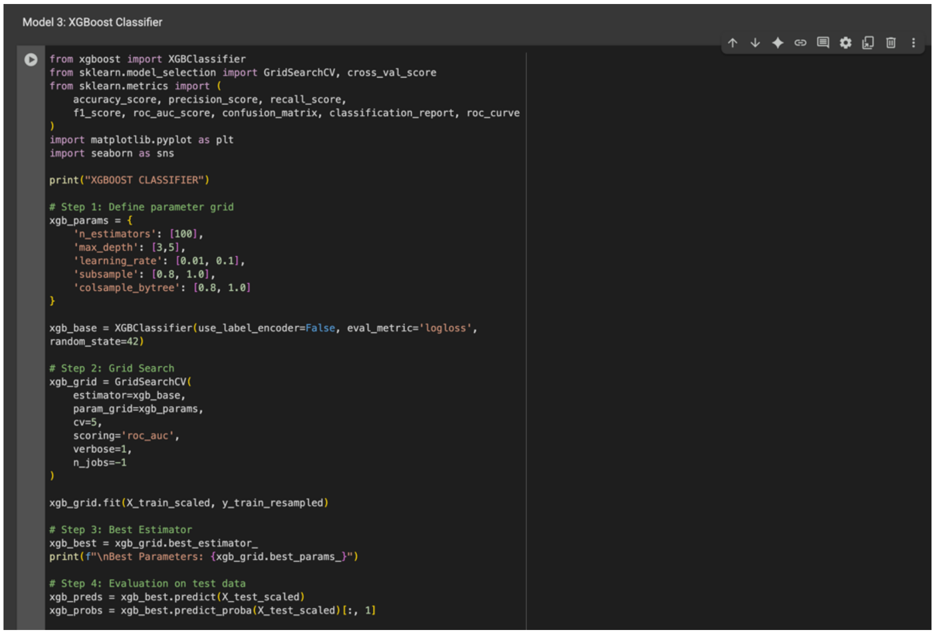
Model 1: Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score,
    f1_score, roc_auc_score, confusion_matrix,
    classification_report, roc_curve, auc
)

print("RANDOM FOREST CLASSIFIER")

# Step 1: Define hyperparameter grid with more constraints
rf_param_grid = {
    'n_estimators': [100],
    'max_depth': [7],
    'min_samples_split': [2, 4],
    'min_samples_leaf': [3, 4],
    'max_features': ['log2'],
    'bootstrap': [True]
}

# Step 2: Initialize GridSearchCV
rf_grid = GridSearchCV(
    estimator=RandomForestClassifier(random_state=42),
    param_grid=rf_param_grid,
    cv=2,
    scoring='roc_auc',
    n_jobs=-1,
    verbose=1
)

# Step 3: Fit model to training data
rf_grid.fit(X_train_scaled, y_train_resampled)

# Step 4: Get best estimator
best_rf_model = rf_grid.best_estimator_
print(f"\nBest Parameters: {rf_grid.best_params_}")

# Step 5: Predict
rf_preds = best_rf_model.predict(X_test_scaled)
rf_probs = best_rf_model.predict_proba(X_test_scaled)[:, 1]
```

Figure 30 Training the Random Forest Classifier

Random Forest was selected because of its capacity to deal with non-linear relationships and overcome overfitting by means of ensemble learning. It builds trees on random sample of features and data and averages the output to generate a stable, yet accurate prediction. This makes it suited well to predicting customer responses to campaigns where behavioural and demographic factors interact in a complex manner.

XGBoost Classifier :



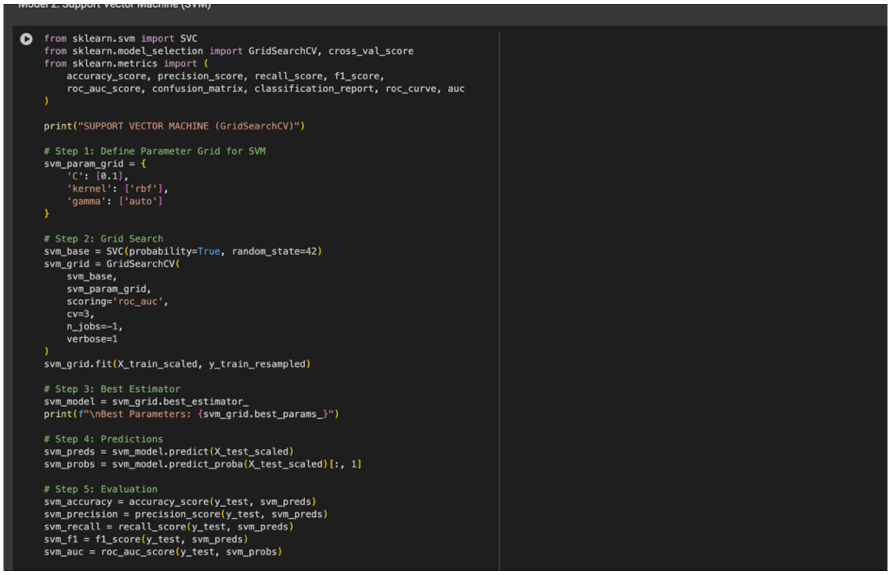
Model 3: XGBoost Classifier

```
from xgboost import XGBClassifier
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score,
    f1_score, roc_auc_score, confusion_matrix, classification_report, roc_curve
)
import matplotlib.pyplot as plt
import seaborn as sns

print("XGBOOST CLASSIFIER")

# Step 1: Define parameter grid
xgb_params = {
    'n_estimators': [100],
    'max_depth': [3,5],
    'learning_rate': [0.01, 0.1],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0]
}

xgb_base = XGBClassifier(use_label_encoder=False, eval_metric='logloss',
random_state=42)

# Step 2: Grid Search
xgb_grid = GridSearchCV(
    estimator=xgb_base,
    param_grid=xgb_params,
    cv=5,
    scoring='roc_auc',
    verbose=1,
    n_jobs=-1
)

xgb_grid.fit(X_train_scaled, y_train_resampled)

# Step 3: Best Estimator
xgb_best = xgb_grid.best_estimator_
print(f"\nBest Parameters: {xgb_grid.best_params_}")

# Step 4: Evaluation on test data
xgb_preds = xgb_best.predict(X_test_scaled)
xgb_probs = xgb_best.predict_proba(X_test_scaled)[:, 1]
```

Figure 31 Training the XGBoost Classifier

XGBoost was added due to its effectiveness and good performance in structured datasets. It can train models sequentially, making each tree correct the previous one and regularizing so as to avoid overfitting. Its gradient boosting framework enables it to capture complicated interdependencies among features and thus it is very applicable in predicting campaign response.

Support Vector Machine (SVM) Classifier :



```
from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score,
    roc_auc_score, confusion_matrix, classification_report, roc_curve, auc
)

print("SUPPORT VECTOR MACHINE (GridSearchCV)")

# Step 1: Define Parameter Grid for SVM
svm_param_grid = {
    'C': [0.1],
    'kernel': ['rbf'],
    'gamma': ['auto']
}

# Step 2: Grid Search
svm_base = SVC(probability=True, random_state=42)
svm_grid = GridSearchCV(
    svm_base,
    svm_param_grid,
    scoring='roc_auc',
    cv=3,
    n_jobs=-1,
    verbose=1
)
svm_grid.fit(X_train_scaled, y_train_resampled)

# Step 3: Best Estimator
svm_model = svm_grid.best_estimator_
print(f"\nBest Parameters: {svm_grid.best_params_}")

# Step 4: Predictions
svm_preds = svm_model.predict(X_test_scaled)
svm_probs = svm_model.predict_proba(X_test_scaled)[:, 1]

# Step 5: Evaluation
svm_accuracy = accuracy_score(y_test, svm_preds)
svm_precision = precision_score(y_test, svm_preds)
svm_recall = recall_score(y_test, svm_preds)
svm_f1 = f1_score(y_test, svm_preds)
svm_auc = roc_auc_score(y_test, svm_probs)
```

Figure 32 Training the Support Vector Machine Classifier

SVM was chosen because it is effective in high dimension and finds a maximum margin optimal hyperplane that separates the classes. SVM can model non-linear decision boundaries by using kernel functions, and as such it is possible to use it to differentiate between the responders and the non-responders in the marketing data

3.7Evaluation Metrics

- Accuracy – overall correctness on the test set.
- Precision (Positive class) -The number of responders among the predicted responders (false positive).
- Recall (Positive class) - of actual responders, the proportion of the group which is correctly identified by the model (false negatives).
- F1-Score - harmonic mean of precision and recall; the main measure, because of class imbalance.

- AUC-ROC - threshold-free performance of discrimination; greater AUC indicates greater separation of responders and non-responders.
- Confusion Matrix - number of TP, TN, FP, and FN to examine trade-offs in error.

3.9 Deployment - Gradio Interface

A lightweight Gradio application was created to put the segmentation framework into practice. The interface takes five behavioural and financial inputs-Income, Total Spending, Monthly Web Visits, Total Purchases, and Campaign Acceptances and uses the same pre-processing pipeline as the model training. The optimised K-Means model (k=3) will then be used to classify customers into High Value, High Potential, or Low Value segment. The system produces practical actionable marketing plans in each segment that links the analysis with the immediate practice of the manager.

3.10 Ethical Considerations

The dataset that will be utilized in this study is publicly available and does not contain any personal or sensitive data, which means that an individual privacy is not violated. Ethical principles were observed throughout the project by paying attention to clarity, fairness, and accountability in all the phases of the model development. The methodology was also aimed at minimizing bias in the prediction such that the outcomes become accountable and credible in the real world application.

4. Results

This part will show the results of the machine learning models created to forecast customer reactions to marketing campaigns. Three models Random Forest (RF), Support Vector Machine (SVM) and XGBoost (Extreme Gradient Boosting) were trained and assessed. To provide fair comparison, all models were evaluated on such metrics as Accuracy, Precision, Recall, F1-Score, and AUC-ROC, and confusion matrices and ROC curves were added as supplementary material..

4.1 Model performance overall

To have a better picture of a comparative performance of the three classifiers, Figure 33 presents the grouped bar charts corresponding to the five evaluation measures. This visualization gives an at-a-glance summary of each of the models strengths and trade-offs.

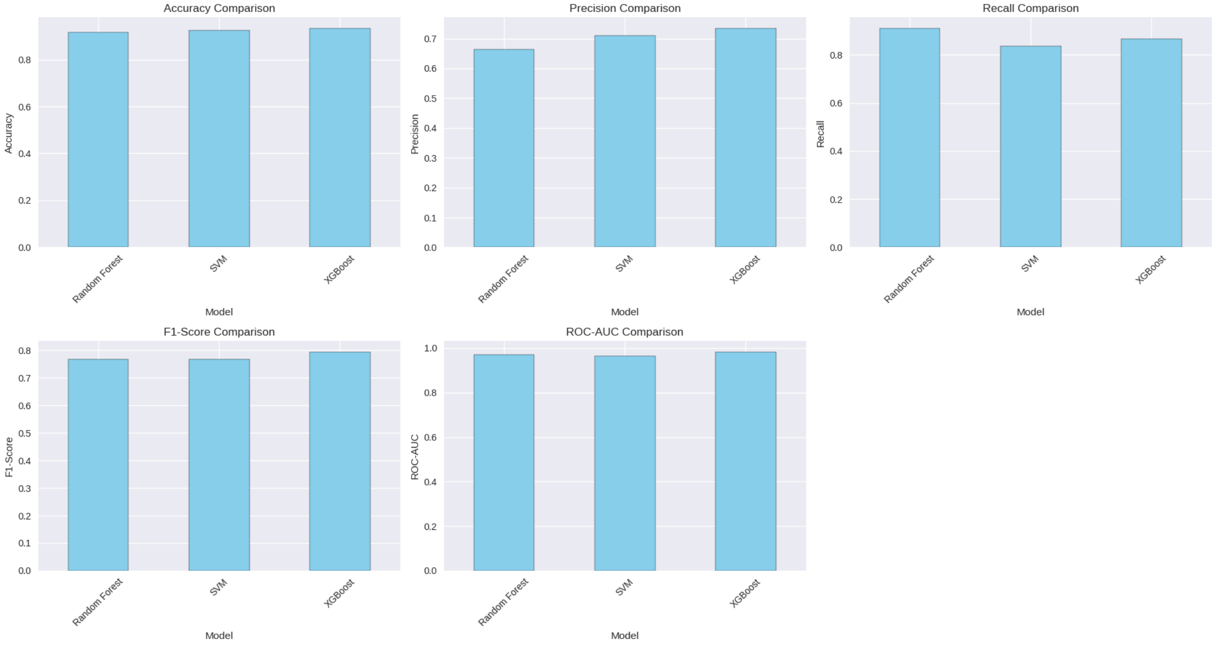


Figure 33 Comparative performance across five key metrics.

The comparative analysis of three models reveals the fact that all of them delivered great results, though they had certain differences. XGBoost produced the best stability with the highest AUC-ROC (0.9835). This implies that it was superior in differentiating the customers who would respond and those who would not. The three models, the Random Forest, SVM and AUC-ROC values were close with 0.9689 and 0.9658 respectively, and this demonstrates that the three models can be reliably used to discriminate. Considering precision, XGBoost was once again the best at minimising false positive rates and making sure that the identified buyers were indeed buyers (0.7342). SVM achieved good result (0.7089) as well, but the result of Random Forest is a little lower (0.6630). However, the opposite was true in recall (0.9104), which shows that Random Forest did not miss any actual buyers, and this is an important feature in marketing where failure to reach potential responders may imply lost revenues. The accuracy of models was good with Random Forest scoring 91.7 percent and XGBoost scoring 93.3 percent. The F1-score followed the same trend as it was most balanced in XGBoost (0.7945), then SVM (0.7669) and Random Forest (0.7674). These results indicate that all of them can be effective in identifying more buyers, but XGBoost provides the most consistent and balanced results.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC-ROC**
Random Forest	0.9174	0.6630	0.9104	0.7673	0.9689
XGBoost	0.9241	0.7089	0.8358	0.7671	0.9658
SVM	0.9330	0.7342	0.8657	0.7945	0.9835

Table 1: Consolidated Model Performance Metrics

4.3 XG Boosting Results

The XGBoost classifier performed the best overall of the models. It obtained an accuracy of 93.3 %, a precision of 0.7342 and a recall of 0.8657. This shows that not only has the model decreased the number of false positives but also accounted most of the real purchasers. F1-score of 0.7945 is the highest which reflects balanced performance, AUC-ROC score of 0.9835 indicates that it has high capability of separating responders and non-responders.

Confusion Matrix:

```
[[360  21]

 [  9  58]]
```

The confusion matrix reveals that XGBoost correctly identified 360 non-responders and 58 responders, while misclassifying 21 false positives and 9 false negatives. This balance highlights its reliability, showing it avoided both excessive false alarms and missed opportunities

4.4 Random Forest Results

The Random Forest classifier reached an accuracy of 91.7% indicating that it classified the larger percentage of customer responses correctly. This is indicated by its recall rate of 91.0%, which is an indication of the model being able to capture most of the actual responders, thus its application is particularly useful when the cost of missing a potential buyer is high, as in the case in marketing. The accuracy however, of 0.6630 shows that a significant percentage of the buyers that were anticipated were false positives. The F1-score of 0.7673 indicates a good balance between precision and recall, whereas the AUC-ROC score of 0.9689 proves that the model was rather effective in discriminating responders and non-responders.

Confusion Matrix:

```
[[350  31]

 [  6  61]]
```

The confusion matrix shows that Random Forest classified 350 true negatives and 61 true positives correctly, but it also produced 31 false positives and 6 false negatives. This means the model leaned toward identifying as many buyers as possible, even at the cost of a few false alarms.

4.5 Support Vector Machine Results

The Support Vector Machine recorded an accuracy of 92.4%, which is slightly behind that of XGBoost. The precision of 0.7089 indicates that it was fairly good at not reporting false positives whereas the recall of 0.8358 indicates that it missed a relatively small number of the actual responders compared to Random Forest. The F1-score of 0.7671 indicates this trade-off, and AUC-ROC score of 0.9658 strongly indicates but not leading discrimination ability.

Confusion Matrix:

```
[[358  23]

 [ 11  56]]
```

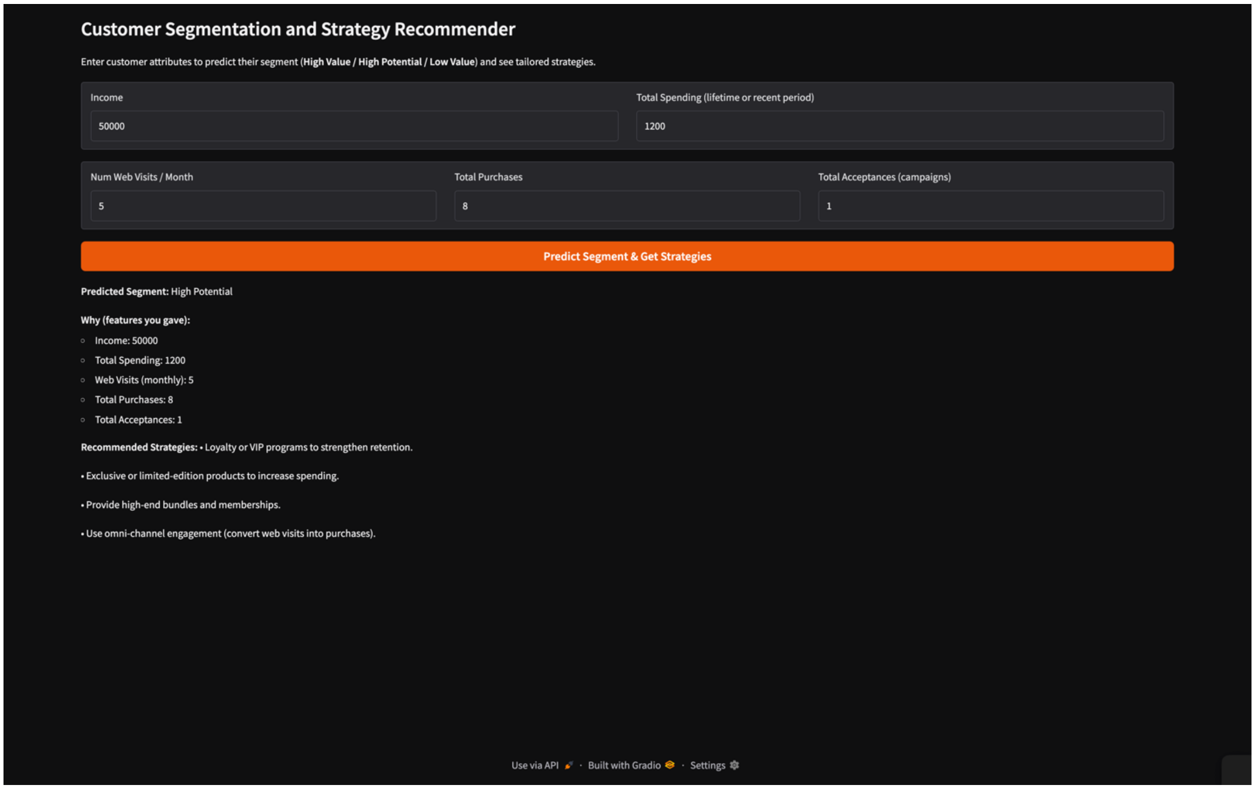
SVM classified 358 non-responders and 56 responders correctly, but generated 23 false positives and 11 false negatives. This indicates that while the model balanced sensitivity and precision well, it was less robust than XGBoost in minimizing errors on both sides.

4.6 Summary

All three models performed strongly, but with different strengths. Random Forest had the highest recall, which makes it ideal where the aim is to identify as many true buyers as possible, although at the expense of additional false positives. SVM provided a trade-off between accuracy and false alarms, with averagely low recall. XGBoost was evidently the best overall performance showing the highest precision, F1-score, and AUC-ROC and therefore, the most trusted model in predicting campaign response.

4.9 Model Deployment Results

A simple Gradio interface was developed. The app uses five important inputs-Income, Total_Spending, NumWebVisitsMonth, Total_Purchases, and Total_Acc, and places the customer in one of the three segments (High Value, High Potential, or Low Value).



Customer Segmentation and Strategy Recommender

Enter customer attributes to predict their segment (**High Value / High Potential / Low Value**) and see tailored strategies.

Income: 50000

Total Spending (lifetime or recent period): 1200

Num Web Visits / Month: 5

Total Purchases: 8

Total Acceptances (campaigns): 1

Predicted Segment: High Potential

Why (features you gave):

- Income: 50000
- Total Spending: 1200
- Web Visits (monthly): 5
- Total Purchases: 8
- Total Acceptances: 1

Recommended Strategies: • Loyalty or VIP programs to strengthen retention.

• Exclusive or limited-edition products to increase spending.

• Provide high-end bundles and memberships.

• Use omni-channel engagement (convert web visits into purchases).

Predict Segment & Get Strategies
Use via API · Built with Gradio · Settings

Figure 34 Gradio-based customer segmentation and strategy recommender interface.

The system automatically provides the marketing team with the recommendations on how to use it, e.g., to offer premium prices to High Value customers or budget deals to Low Value groups. This renders the clustering insights more actionable and demonstrates how the businesses may use it in real time.

5. Discussion

The results are explained in light of the two proposed research questions. The former was used to understand how well K-Means clustering can be used to identify different customer groups using demographic, purchasing and engagement data. The second one entailed the determination of the most accurate predictive model in the customer response to marketing campaigns when using Random Forest, SVM, and XGBoost.

5.1 Customer Segmentation Insights

The K-Means clustering produced three important customer categories of High Value, High Potential and Low Value customers in terms of income, spending, purchases and engagement. The clusters differed in terms of their behavioural and demographic characteristics, which confirm earlier studies on the importance of segmentation as a principle of targeted marketing. In particular, the ROI analysis indicated that efficiency of the campaign increased when strategies were personalized to each segment, which proved the worth of clustering in terms of actionable decision-making.

5.2 Model Performance Insights

XGBoost was the best performing model overall, with an ROC-AUC of 0.9835 and an acceptable precision/recall balance. This proves its appropriateness in marketing prediction where accurateness and discriminatory ability is essential. Random Forest fared especially well on recall (0.9104), and thus Random Forest is useful in settings where it is more important to capture as many true responders as possible than it is to avoid false positives. SVM also posted competitive accuracy (0.9241) but fell a little short of the ensemble methods in differentiating among purchasers and non-purchasers. These results are in line with the earlier findings that showed that the boosting algorithms usually beat the traditional classifiers in structured marketing data sets.

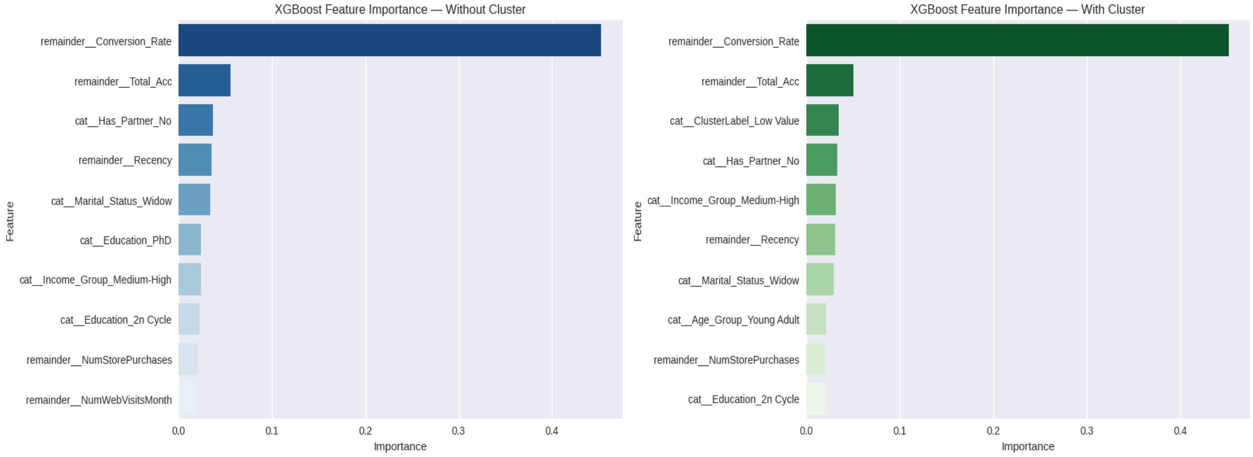


Figure 35 Feature Importance with and without cluster

To answer the second research question, an experiment was conducted whereby XGBoost was trained with as well as without the ClusterLabel. The overall predictive metrics did not change much, however, the feature importance plots (Figure 35) showed that ClusterLabel emerged as a top predictor when it was included. This is an indication that, although segmentation did not significantly help in accuracy, it provided interpretive power since all the complex demographic and behavioural characteristics were consolidated into one business-friendly feature. In reality, this makes the outcomes more practical to the marketer, since the model predictions can be followed back to customer segmentations such as the High Value, High Potential, and Low Value.

5.3 Results Implications

The behavioural insight and the predictive accuracy are given by the integration of clustering and predictive modelling. Using results in terms of ROI, the analysis is able to show that not only technical performance is shown but also its business impact which is more holistic in terms of analysing campaign results.

5.4 Future Research

Future research would be able to elaborate on the current study by use of larger and more varied data that would capture the customer behaviour in various channels. It can also be interesting to add unstructured data, such as customer reviews or clickstream logs, to get a better idea of preferences and engagement behaviour. It can be interesting to explore other clustering options, such as a hierarchical clustering or Gaussian Mixture Models, which could produce more detailed customer segments. Based on the predictive side, there are also more sophisticated models such as deep learning or transformer-based models that can be applied to improve the classification performance. In the further work, it is necessary to test the ROI implications at the extended periods and in other industries to prove the overall business effects of segmentation-based strategies.

5.5 Limitations of the Study

This study offers valuable information regarding the customer segmentation and response prediction of campaign, but it has several shortcomings. The first relates to class imbalance in the dataset. Although, SMOTE was implemented to balance the responders and non-responders, these models can still be biased towards the larger class, and such models cannot be as sensitive to detect the minority, but more important group of actual responders.

Second, the research used only the existing attributes in the dataset, and it was primarily demographics, purchase history, and measures of engagement. Highly relevant behavioural cues, e.g. browsing behaviour, live communication, or even situational aspects like seasonality were not incorporated. The fact that they were not available could have limited the level of customer segmentation as well as the accuracy of campaign forecasts.

Third, despite the cross-validation applied to minimise overfitting, the models were trained and tested on the same dataset. It is not clear how well they can generalise to other industries, geographies or customer bases and this should be tested externally.

Lastly, the ROI analysis was simplified in nature reflecting the cost of the campaign and patterns of customer response. In practice, the marketing ROI will be influenced by a wide range of other factors, such as long-term retention, brand effects and operating costs and these were beyond the scope of this study.

6. Conclusion

This research aimed to find answers to the following two questions: what customer segments can be determined by demographic, purchasing, and engagement attributes K-Means clustering, and will the classification of campaign responses be improved by adding labels of these segments to machine learning models?. The three actionable segments identified by the clustering analysis were High Value, High Potential, and Low Value customers that were differentiated by income, spending, and responsiveness. These segments give a systematic foundation to diversified marketing approaches, including premium offers to High Value customers, loyalty-building activities to High Potential groups and price-sensitive campaigns to Low Value customers.

On the predictive side, XGBoost returned the best overall results, with the highest ROC-AUC (0.9835) and F1-score (0.7945), which proves its effectiveness in identifying who is and who is not a responder. Random Forest performed the best in terms of recall which is useful in maximising responder detection, whereas SVM had reliably stable accuracy but with a relatively weaker discriminative capacity. Combination of cluster labels yielded minor improvements in measures, but provided enhanced interpretability, with the feature importance analysis showing segmentation as a meaningful predictor.

Lastly, the ROI calculation demonstrated the commercial usefulness of segmentation and prediction combination, with specific strategies to High Potential customers, being more efficient in marketing efforts. Collectively, the results show that unsupervised clustering can be combined with supervised prediction to not only increase the effectiveness of campaign targeting but also to identify the linkage between technical outputs and real business outcomes.

6.1 Future work

The results of this project point towards a number of ways that this project can be refined and extended. One of the critical areas is the improved management of class imbalance where campaign responders are the minority in most real datasets. Future research can include hybrid resampling or cost-sensitive adaptive learning to pay more attention to under-represented yet business-critical cases.

Another direction is the enrichment of feature space. Although in this project demographics, spending and engagement were considered, it can be suggested that in the future, the variables of browsing behaviour, cross-channel interactions, or seasonality effects can be integrated. This would allow a more complete picture of customer journeys and would result in more accurate segmentation and targeting. More methodologically, using more sophisticated optimisation tools such as Bayesian search or evolutionary algorithms might enable more efficient hyperparameter search and the resulting better predictive performance in a wide variety of settings.

References:

1. Ologunebi, J. O., Taiwo, E. O. & Alli, K. O. (2025) ‘Digital Consumer Behavior in E-commerce: A Study of Amazon and Temu’s Customer Purchase Decision-Making Processes in the UK and the USA’, *SSRN*, (MPRA Paper 123096). Available at: https://doi.org/10.2139/ssrn.5071874
2. Ali, S., Khan, R. & Farooq, A. (2023) ‘Boosting models for predicting customer campaign responses in e-commerce’, *Electronic Commerce Research and Applications*, 58, 101226. doi:10.1016/j.elerap.2023.101226
3. Batista, P., Silva, M. & Gomes, F. (2022) ‘Integrating behavioural segmentation into XGBoost for campaign response prediction in fashion retail’, *Decision Support Systems*, 155, 113126. doi:10.1016/j.dss.2021.113126
4. Becker, L. & Han, J. (2023) ‘Predicting loyalty program responses using XGBoost and logistic regression’, *Expert Systems with Applications*, 220, 119713. doi:10.1016/j.eswa.2022.119713
5. Becker, L., Han, J. & Silva, D. (2024) ‘Explainable ensemble models for campaign targeting in marketing analytics’, *Information Systems Frontiers*, 26(3), pp. 477–493. doi:10.1007/s10796-023-10421-0
6. Cheng, H. & Xu, L. (2022) ‘Improving digital shopper conversion through RFM-based K-Means segmentation’, *Electronic Markets*, 32(1), pp. 143–158. doi:10.1007/s12525-021-00475-y
7. Choudhury, R., Ghosh, A. & Saha, P. (2023) ‘A comparative analysis of Random Forest and SVM in campaign response modelling’, *International Journal of Business Analytics*, 10(2), pp. 21–37.
8. Dasgupta, S., Rao, K. & Venkat, R. (2024) ‘Clustering retail grocery customers using K-Means and hierarchical methods’, *Journal of Retail Analytics*, 20(1), pp. 93–112.
9. El-Hajj, M., Yassine, A. & Pavlova, A. (2024) ‘Decision trees with resampling for campaign response prediction’, *International Journal of Information Management Data Insights*, 5(2), 100211. doi:10.1016/j.jjimei.2024.100211
10. Fernandes, T., Costa, P. & Oliveira, J. (2022) ‘Improving cluster coherence in shopper datasets using outlier detection and PCA’, *Computers in Industry*, 142, 103691. doi:10.1016/j.compind.2022.103691
11. Haque, A., Islam, S. & Roy, P. (2022) ‘RFM + K-Means segmentation framework in digital retail’, *Journal of Retailing and Consumer Services*, 65, 102858. doi:10.1016/j.jretconser.2021.102858
12. Hassan, M., Alam, N. & Chowdhury, S. (2022) ‘Clustering online shoppers using session-level engagement metrics’, *Electronic Commerce Research*, 22(4), pp. 765–782. doi:10.1007/s10660-021-09491-8
13. Hicham, B. & Karim, A. (2022) ‘Clustering ensembles and spectral methods for customer segmentation’, *Procedia Computer Science*, 200, pp. 134–142. doi:10.1016/j.procs.2022.01.030

14. Iyer, S., Singh, R. & Patel, A. (2023) ‘Explaining Random Forest predictions in retail campaign response modelling’, *Applied Intelligence*, 53(6), pp. 6451–6467. doi:10.1007/s10489-022-04567-2
15. Jindal, R., Mehra, A. & Kapoor, S. (2023) ‘Segmentation of retail banking customers using K-Means and PCA’, *Journal of Financial Services Marketing*, 28(2), pp. 155–169.
16. John, P., Taylor, C. & Smith, R. (2024) ‘Comparing clustering methods for retail segmentation: K-Means, GMM, DBSCAN, and hierarchical approaches’, *Journal of Business Research*, 168, pp. 45–59. doi:10.1016/j.jbusres.2022.08.161
17. Joung, S., Kim, H. & Park, J. (2023) ‘Interpretable ML segmentation for product design using online reviews’, *Decision Sciences*, 54(5), pp. 1023–1045. doi:10.1111/deci.12546
18. Kabir, M. (2025) ‘K-Means combined with neural networks for behaviour prediction in e-commerce’, *International Journal of Data and Information Systems*, 14(2), pp. 211–229.
19. Khandokar, R., Ali, M. & Rahman, T. (2023) ‘Comparative study of machine learning models for online retail campaign prediction’, *International Journal of Computer Applications in Technology*, 72(1), pp. 39–53.
20. Kim, S. & Park, J. (2024) ‘Robust classifiers for customer response prediction in e-commerce’, *Journal of Intelligent Information Systems*, 62(3), pp. 377–393. doi:10.1007/s10844-023-00750-5
21. Kumar, V., Singh, R. & Mehta, P. (2023) ‘Customer value segmentation using K-Means and PCA on retail data’, *Journal of Retail and Consumer Studies*, 31(4), pp. 517–529.
22. Li, Y. & Zhao, X. (2021) ‘A comparison of Gaussian Mixture Models and K-Means in e-commerce customer segmentation’, *Knowledge-Based Systems*, 228, 107241. doi:10.1016/j.knosys.2021.107241
23. Lim, J. & Wang, H. (2023) ‘RFMP segmentation for donor behaviour in crowdfunding platforms’, *Journal of Business Analytics*, 6(2), pp. 185–203. doi:10.1080/2573234X.2023.2201620
24. Lin, F., Zhang, W. & Chen, Y. (2023) ‘Boosting models for fashion e-commerce response prediction’, *Electronic Commerce Research and Applications*, 57, 101238. doi:10.1016/j.elerap.2022.101238
25. Martinez, L., Gomez, A. & Silva, C. (2023) ‘Improving churn and campaign response prediction using K-Prototypes clustering’, *Telecommunications Policy*, 47(5), 102534. doi:10.1016/j.telpol.2023.102534
26. Mehta, R., Chatterjee, S. & Banerjee, T. (2024) ‘Hierarchical clustering for supermarket loyalty program optimization’, *Journal of Retail Analytics*, 20(1), pp. 93–112.
27. Nguyen, T., Pham, L. & Hoang, D. (2022) ‘Latent persona discovery in e-commerce using GMM and K-Means’, *Electronic Markets*, 32(2), pp. 411–426. doi:10.1007/s12525-022-00547-w
28. Okafor, C. & Mensah, K. (2022) ‘Hierarchical vs K-Means clustering for retail customer segmentation’, *Journal of Retail and Marketing Analytics*, 14(3), pp. 201–218.

29. Patel, A., Kumar, R. & Singh, V. (2024) ‘Comparative analysis of SVM and XGBoost for customer retention prediction’, *Applied Soft Computing*, 142, 110303. doi:10.1016/j.asoc.2024.110303
30. Pereira, M., Silva, J. & Costa, L. (2021) ‘Comparing SVM and decision trees for retail campaign response prediction’, *Journal of Applied Statistics*, 48(11), pp. 2335–2350. doi:10.1080/02664763.2021.1873712
31. Rajkumar, R., Prasad, A. & Nair, S. (2024) ‘Pharmacy retail segmentation using K-Means clustering’, *Health Informatics Journal*, 30(1), pp. 55–73. doi:10.1177/18333583221095062
32. Rodriguez, J., Lopez, A. & Fernandez, M. (2022) ‘Gaussian Mixture Models for customer segmentation in online fashion retail’, *Journal of Retailing and Consumer Services*, 68, 103049. doi:10.1016/j.jretconser.2022.103049
33. Rosário, A. & Raimundo, R. (2021) ‘Consumer Marketing Strategy and E-Commerce in the Last Decade: A Literature Review’, *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7), pp. 3003–3024. doi:10.3390/jtaer16070164
34. Sharma, P., Mehta, V. & Bansal, R. (2024) ‘Campaign response modelling using Random Forest and XGBoost on retail data’, *Expert Systems with Applications*, 221, 119754. doi:10.1016/j.eswa.2023.119754
35. Singh, A., Rao, P. & Desai, N. (2023) ‘Clustering pharmacy customers using purchase sequences’, *International Journal of Pharmaceutical Marketing*, 17(2), pp. 102–116.
36. Uddin, M. (2024) ‘Segmentation of digital-native consumers using machine learning’, *Journal of Retail and Digital Innovation*, 8(3), pp. 89–107.
37. Varma, S. & Kale, R. (2024) ‘Random Forest and XGBoost for campaign response prediction in e-commerce’, *International Journal of Data Science*, 9(1), pp. 33–47.
38. Wang, Y., Chen, L. & Huang, T. (2024) ‘Embedding donor behaviour clustering with Random Forests for email targeting’, *Decision Support Systems*, 165, 113823. doi:10.1016/j.dss.2024.113823
39. Wang, Y., Zhao, H. & Li, M. (2025) ‘Reinforcement learning enhanced clustering for digital marketing’, *Neural Computing and Applications*, 37(4), pp. 2241–2259. doi:10.1007/s00521-024-09299-4
40. Giannakopoulos, N. T., Terzi, M. C., Sakas, D. P., Kanellos, N., Toudas, K. S. & Migkos, S. P. (2024) ‘Agroeconomic Indexes and Big Data: Digital Marketing Analytics Implications for Enhanced Decision Making with Artificial Intelligence-Based Modeling’, *Information*, 15(2), 67. doi:10.3390/info15020067