Data Mining Project:

Customer Segmentation for Tailored Marketing Strategies

Md Mobusshar Islam
M.Sc. in CSE (Applied Computing)
University of Oulu
mislam23@student.oulu.fi
2305578

Taufiq Ahmed
M.Sc in CSE (Artificial Intelligence)
University of Oulu
tahamed23@student.oulu.fi
2307644

Arash Nedaei Janbesaraei
M.Sc in Business Analytics (CSE)
University of Oulu
arash.nedaeijanbesaraei@student.oulu.f
i
2308916

Kavinda Kulasinghe
M.Sc. in Business Analytics (IPS)
University of Oulu
kkulasin23@student.oulu.fi
2304823

Zohreh Yousefi Dahka
M.Sc in Business Analytics
University of Oulu
zohreh.yousefidahka@student.oulu.fi
2310628

This study aims to optimize marketing strategies by effectively targeting and engaging distinct customer segments. Utilizing the Customer Personality Insight dataset, we employ clustering methods to detect unique customer segments based on relevant demographic, income, and behavioral factors. Our objectives include accurately categorizing customers into segments and predicting the segment to which new customers belong, which will enable the development of tailored marketing strategies.

Keywords— marketing strategies, customer segmentation, predictive modeling, customer behavior analysis, demographic factors, income segmentation, machine learning algorithms, targeted advertising, customer profiling, customer retention strategies

I. INTRODUCTION

In today's increasingly competitive market, knowing consumer behavior and preferences is critical for organizations looking to establish effective marketing strategies and increase customer engagement using modern data mining techniques and machine learning algorithms, marketers may obtain important insights into consumer segmentation and predictive modeling, enabling for more specialized and focused approaches to customer outreach and retention. Consumer segmentation and predictive modeling are important topics for firms looking to better their marketing tactics and consumer engagement. Understanding consumer behavior and preferences helps organizations to better adapt their marketing efforts, resulting in higher customer satisfaction and loyalty. We are interested in studying this topic because of its practical implications for businesses across various industries. Our project intends to evaluate customer data to create different consumer groups based on demographic and financial criteria, as well as develop predictive algorithms to categorize new customers into these categories. By doing so, we want to deliver actionable information for businesses to learn targeted marketing tactics that resonate with their target demographic, resulting in business development and success.

This article delves into the realms of customer segmentation and predictive modeling, presenting a comprehensive analysis aimed at unraveling the complexities of consumer preferences and behavior. The article aims to provide practical insights for firms looking to employ data-driven tactics for increased marketing success by meticulously analyzing a dataset that includes demographic, economic, and purchasing behavior factors. Furthermore, predictive modeling approaches, like as neural networks and decision trees, are applied to forecast consumer behavior, particularly about campaign acceptability. The findings underscore the significance of factors such as income per household and total promotions accepted in predicting customer responses to marketing campaigns.

In addition to giving insights into client segmentation and predictive analytics, the study has practical implications for firms trying to properly adjust their marketing tactics. While acknowledging limitations such as dataset representativeness and potential biases, the findings contribute to advancing understanding in the field, paving the way for further research and exploration. Ultimately, the insights gleaned from this study enable businesses to drive improved customer engagement, satisfaction, and overall marketing performance.

This paper is organized as follows: Section 2 gazes into related works, while Section 3 summarizes the study's aims. In Section 4, we provide information regarding the dataset used in our research. Following that, Section 5 elaborates on the data analysis techniques used. Chapter 6 outlines the findings, and Chapter 7 provides a thorough explanation of them. Section 8 delves into the reflection on group work, while Section 9 includes the sources cited throughout the study.

II. Related Work

Previous investigations into consumer psychological categorization and prediction have predominantly employed machine learning, neural networking, and deep learning methodologies, each presenting unique advantages and drawbacks. While certain studies have opted for conventional machine learning algorithms such as logistic regression, decision trees, and support vector machines (SVM) to forecast consumer behavior, these endeavors frequently lack thorough comparative assessments of algorithm performance and are commonly confined to particular industry sectors or geographical locales [1].

Transitioning to customer profiling in digital start-ups, Kasem et al. (2023) showcase the efficacy of AI techniques, particularly RFM analysis coupled with K-means clustering, in delineating distinct customer clusters. Their identification of new, best, and intermittent customer segments underscores the need for tailored engagement strategies to maximize marketing efforts and sales performance in the digital realm. [2]

Christy (Year) echoes the importance of segmentation in satisfying customer needs, leveraging RFM analysis and K-means algorithms to refine customer segmentation processes. This iterative approach underscores the evolution of segmentation methodologies to accommodate the diverse needs and behaviors of consumers. [3].

Chen et al. (2012) provide a comprehensive case study illustrating the leverage of data mining techniques, including RFM model integration with k-means clustering and decision tree induction. Their findings unveil unique consumer groups characterized by diverse shopping behaviors, laying the groundwork for targeted marketing strategies tailored to specific customer segments. [4]

In contrast, Olson & Chae (2012) shed light on the comparative effectiveness of RFM-based predictive models and classical data mining techniques in direct marketing decision support. While RFM methods offer simplicity, classical data mining algorithms demonstrate superior prediction accuracy, signaling the potential for enhanced decision-making through their integration. [5]

Kashwan's (2013) model for continuous analysis and online prediction of sales in e-commerce organizations introduces a hybrid approach, blending K-means clustering with statistical tools. This integrated system empowers managers with intelligent insights for rapid decision-making, aligning with the demand for real-time adaptability in retail strategies. [6]

X He and C Li provide a three-dimensional technique for increasing customer lifetime value (CLV), satisfaction, and behavior. Their study emphasizes the need of segmentation in understanding consumers' various needs and expectations, thereby allowing the delivery of improved service. [7]

In the meantime, using data mining techniques like clustering, decision trees, and SVM methods Hegde et al. (2019) investigated persona classification. However, the research was dependent on a narrow dataset may impact the overall system of its outcomes [8].

Wong & Wei's (Year) comprehensive model amalgamates data mining, customer segmentation, and predictive analysis to decipher online shopping tendencies within an online travel agency. By segmenting customers based on RFM and CLV models, they uncover high-value customers and anticipate their travel behaviors, enabling targeted marketing tactics and tailored promotions [9]

Another example, Vo et al. (2021) used unorganized call log data to produce estimation, using text mining and machine learning approaches. But this approach has drawbacks, such as a constrained dataset and a lack of vast algorithm comparison [10]. Likewise, Sun et al. (2021) used virtual individuals for identifying personality traits with CNN for feature extraction and for classification used RNN. Still, the study's tiny dataset and narrow focus on personality factors may influence the applicability of its conclusions. [11].

McCrary's (2009) multi-stage targeting methodology represents a paradigm shift in retail marketing, incorporating customer spending and profit alongside traditional response models. This holistic approach leads to significant profit increases compared to single-stage models, underscoring the importance of multifaceted targeting strategies in driving revenue growth. [12]

Chaudhuri et al. (2021) utilized a Multi-Layer Perceptron (MLP) neural network to study client purchase behaviors. Despite its achievements, the tiny dataset of the study lacked a complete assessment of deep learning techniques. [13]. In addition, Hassanein et al. (2018) using topic modeling and SVM classifiers identified personality characteristics from social media material, however they focused only on a limited range of characteristics [14].

Beginning with the seminal work of Jiang and Tuzhilin (2008), the integration of customer segmentation and buyer targeting emerges as a crucial avenue for optimizing marketing performance. Their emphasis on direct clustering methods using transactional data underscores the evolving nature of segmentation techniques, moving beyond computed statistics to capture nuanced customer behaviors. [15]

Lastly, Bradlow et al.'s discourse on big data's impact on retailing emphasizes the role of predictive analytics and statistical tools in harnessing vast datasets for informed decision-making. Despite the advent of big data, the article underscores the enduring significance of theoretical frameworks in guiding retail analysis and deriving meaningful insight. [16]

In conclusion, while previous research has made significant advances in understanding consumer personality analysis and turnover prediction, there are still significant gaps and limitations. Our study attempts to close these gaps by developing an alternative collection model.

III. Objectives

The article explores two primary objectives: firstly, to identify distinct customer groups based on demographic, income, and purchasing behavior attributes with aims to provide insights into customer behavior, aiding businesses in tailoring marketing strategies. Secondly, the article investigates predictive modeling approaches to anticipate customer behavior, particularly focusing on campaign acceptance. The goal is to predict the customer segment to which a new customer belongs, enabling personalized marketing strategies. Ultimately, the article aims to offer practical insights for businesses seeking to utilize data-driven approaches in customer segmentation and predictive analytics to enhance marketing effectiveness.

The objectives of our project revolve around understanding customer behavior and preferences through data analysis and modeling techniques, with the ultimate goal of informing tailored marketing strategies. The research questions guiding our investigation are as follows:

1. What clustering techniques can be employed to accurately identify distinct customer segments based on relevant demographic, income, ... factors?
2. Predicting the customer segment to which a new customer belongs to devise tailored marketing strategies.

The expected results include insights into distinct customer segments and the development of predictive models to customize marketing strategies for improved customer engagement and retention.

IV. Data

The dataset used here in this study consisted of 2240 data points and 29 attributes. It was collected by Dr. Omar Romero-Hernandez, The dataset includes various attributes falling into four main categories:

Customer Information: Attributes that include birthdate, level of education, relationship status, and annual household income provide demographic information on customers.

Products: Includes attributes indicating the amount spent on various products, including wine, fruits, meats, fish, sweets, and gold. These attributes reflect customers' purchasing behaviors and preferences.

Promotion: Attributes in this category describe customers' responses to promotion campaigns, including whether they accepted the promotion and the number of purchases made with discounts. This information helps evaluate the effectiveness of marketing promotions.

Place: Attributes related to customers' interaction and engagement with the company's online platforms and catalogs, such as the number of website visits and purchases made online or directly through catalogs. These variables offer insights into customers' preferred shopping channels and behavior.

The dataset is available on Kaggle as a CSV file. You can access the dataset through the following link: Kaggle Customer Personality Analysis Dataset. Each row represents a unique customer, and each column represents a different attribute. The variables in the dataset include categorical variables (e.g., marital status), numerical variables (e.g., yearly household income), and binary variables (e.g., whether a customer accepted a promotion).

A. Pre-processing

During this process, it was learned that the dataset contained missing values in the income variable and outliers in age and income. These issues were addressed by removing instances with missing income values and eliminating outliers, respectively.

Initially, 24 instances with missing income values were encountered. Two options were considered: removing the data points with missing values or replacing them with the median income. Since both approaches did not significantly affect the distribution, it was decided to remove these instances to maintain data integrity. Outliers in the age and income variables were identified and subsequently removed. This step ensured that the data used for analysis was more representative and less affected by extreme values.

Several new features were created to enrich the dataset and provide more insights into customer behavior. These features included Age (derived from the latest year in the dataset), Spent (total spending on products), Children (merging kids and teenagers), Adults, Family Size, Days_a_customer (based on the latest date in the dataset), Total Purchases, and Income per person (income divided by family size). Categorical variables were encoded using one-hot encoding to create a numerical format appropriate for analysis.

V. Methods

Descriptive Analysis:

For the first research question, we used the K-means clustering technique to categorize clients based on demographic, economic, and purchase habit characteristics. Prior to clustering, we applied scaling to ensure that all features contributed equally to the clustering process. The optimal number of clusters was determined using the Elbow method, which identified the point of maximum curvature in the within-cluster sum of squares (WCSS) plot. Subsequently, K-means was employed to partition the dataset into distinct clusters. Through this approach, we gained valuable insights into customer segmentation, allowing us to understand the distinct characteristics and preferences of each cluster. Furthermore, we refined our findings using the RFM (Recency, Frequency, Monetary) analysis technique, resulting in a thorough understanding of client behavior and targeted marketing tactics tailored to each segment's specific requirements and tastes.

Predictive Analysis:
As for the second research question for prediction on the dataset has been done using two different approaches.

In the first approach classification models are used to predict if a customer accepts a campaign or not using the "Total_Promos" attribute which was produced during data preprocessing.

In the second approach, clustering is applied on the attributes that stores are able to gather about customers, based on their purchase behaviour and then predict the cluster using more personal data in order to see if stores can find insights about their customers using their purchase behaviour.

Moreover, in the prediction phase, two new attributes, Unnecessary_Purchases (total spending on gold and sweet products, and wine) and Necessary_Purchases (total spending on meat, fish and fruit) are added to the dataset.

1. Classification:
The target attribute is as below which demonstrates an uneven dispersion of the amounts.

TABLE I. UNEVEN DISPERSION OF THE AMOUNTS

Attributes	*Amounts*
0	1606
1	367
2	139
3	50
4	36
5	10

As the distribution of the data is uneven especially in the higher numbers, the data is divided into two categories, including 0 which stands for the customers that did not accept any of the campaigns and 1 for the ones accepting at least 1 campaign. In order to balance the data distribution, SMOTE technique was used to increase the data samples for class 1 to 1606 samples.

2. Cluster prediction:
Here the elbow method is used which resulted in two clusters as it is demonstrated in the chart Figure 1 below to find the optimal number of clusters. For the clustering, the Agglomerative Clustering model is used which resulted in a more different clusters than K-means. This method employs a bottom-up strategy in which each data point begins in its own cluster and pairs of clusters where they are combined altogether based on some similarity metric until all data points fall to a single cluster or a stopping requirement is satisfied.

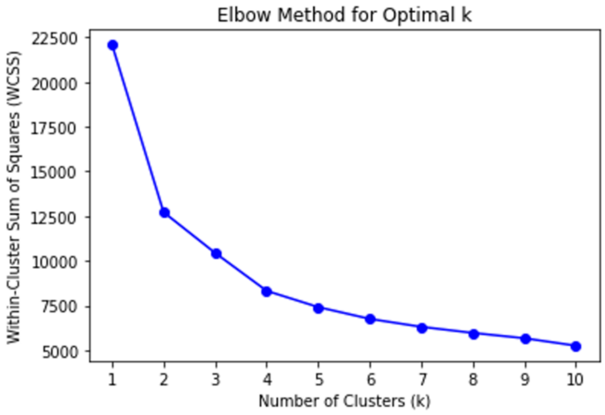

Figure 1: Elbow method for Optimal K

In the prediction part, all the attributes used in clustering are removed from dataset and the cluster number is considered as target data.

For this part also, neural network is chosen as the predictive model, with 3 layers, SGD as optimizer, learning rate of 0.1 and 45 epochs.

VI. RESULTS

Descriptive Analysis:

K- Means Clustering:
Based on their socioeconomic and behavioral characteristics, the customer's base was divided into two main clusters on Figure 2 by applying the k-means clustering technique. A clear bend in the inside-cluster sum of squares (WCSS) plot, which indicates the greatest reduction in variance within the clusters, indicated that two was the ideal number of clusters on below, as established by the elbow technique.

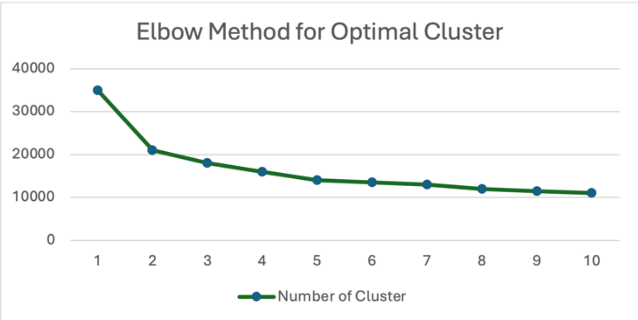

Figure 2: Optimal numbers of clusters

TABLE II. CHARACTERISITICS OF CUSTOMER CLUSTER

Attribute Name	**Cluster 0**	**Cluster 1**
Average income	$42,024	$73,679
Average spending	$275	$1,329
Family size	2.9	1.9
Number of Children	1.2	0.3

Larger families, a higher probability of having children, poorer income and expenditure levels are the characteristics of Cluster 0. Customers in Cluster 1 on the other hand had smaller families, fewer children, and higher incomes and spending power. Because of this segmentation, marketing tactics may be specifically designed to meet the demands of each group.

RFM Analysis:

Based on the customer's purchase habits, the RFM research divided the client base into four unique groups. Three variables of customer analysis were used to create this segmentation: monetary value, frequency, and recency. Every dimension was given a score, and each consumer was then categorized into one of four groups based on the total RFM score: Top, High, Middle, or Low.

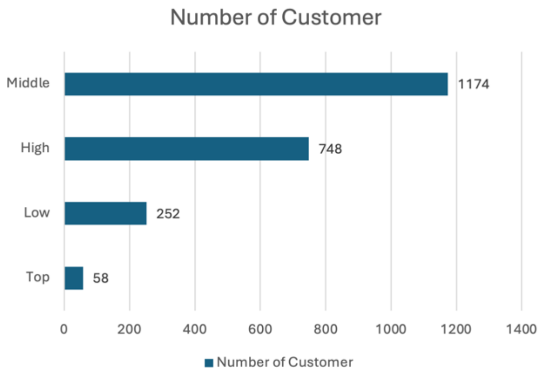

Fig. 9: RFM segmentation of customers

Top Customers, with an RFM score of 12, consist of 58 extremely valuable clients who spend a lot, interact often, and have recently made purchases, indicating the need for retention and individualized service tactics. With 748 active users and a score between 9 and 11, these customers are considered High-Value and can be targeted for marketing campaigns and upselling. 1174 customers, categorized as Middle-Value Customers, have ratings between 5 and 8, indicating room for growth in the form of promotions meant to boost spending and transaction volume. Finally, 252 less active consumers are included in the Low-Value consumers category, with a score of 4 or lower, necessitating re-engagement tactics like campaigns and special offers. Because of this segmentation, marketing techniques may be tailored to improve client engagement and allocate resources as profitably as possible.

Predictive Analysis:

1. Classification result:

For classification, Decision Tree, Neural Network, Logistic Regression and SVC are used which led to an average accuracy of 82.75% and the best accuracy was for neural network to be about 85 percent. The details on accuracy for each model are shown in the chart below.

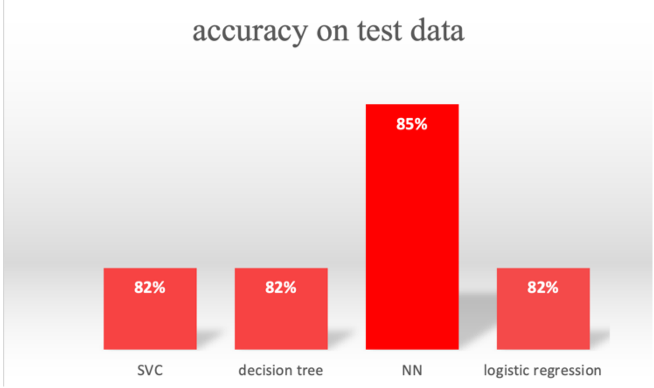

Fig. 1: Accuracy on test data

In this Fig.1 by visualizing the highest accuracy, the main model is chosen to be Neural Network. The Neural Network model in this project consists of 3 Dense layers, including 128 neurons in the first two layers and one neuron for the last one. SGD is the optimizer in the model, with learning rate of 0.05 and momentum to be 0.9 and finally the number of epochs is 45 which led to the best accuracy. To find the influential factors on the model, Permutation Feature Importance is used. In this technique, the change in model error (like MAE, r-squared, or accuracy) is measured after a single model feature's values have been permuted. The result of this technique is demonstrated in the chart Fig. 2 below which shows income per household, total purchases, necessary purchases (meat, fish and fruit) and unnecessary purchases (wine, sweet and gold products) to be the most influential factors on the model.

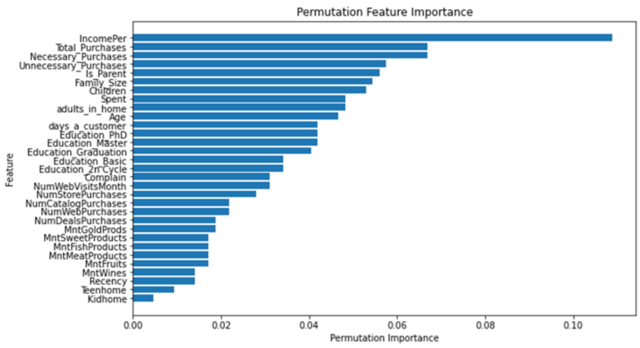

Fig. 2: Permutation Feature Importance

In order to have better understanding of the data samples in each class and their differences based on the influential factors, the below charts Fig. 3 seem to be helpful. The charts demonstrate that the number of children and family size are bigger in the customers of class 0 and on the other side, the customers of class 1 have done more purchases both in necessary and unnecessary categories and they also have more income per household.

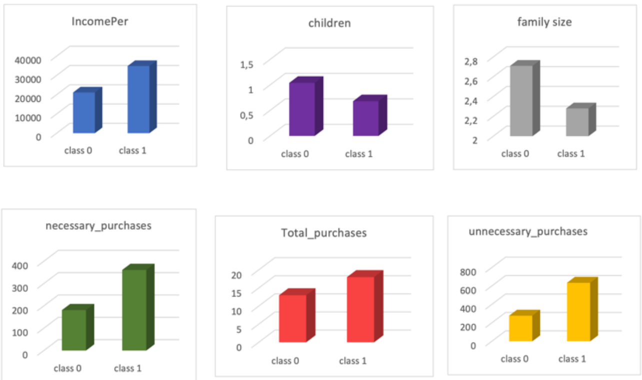

Fig. 3: Data samples in each class and their differences

2. Cluster prediction result:

The clustering model resulted in two different cluster of customers based on their purchase behavior. The chart below fig. 5 is demonstrative of the distribution of data in the clusters:

Fig. 5: Distribution of the clusters

Below is the table demonstrating the difference between clusters based on the average of the attributes used in clustering. Based on the below table, the cluster 1 belongs to the customers that pay less on both necessary and especially on unnecessary products and they have fewer number of purchases. Moreover, in cluster 1 the customers have more web purchases and in cluster 0, customers have more catalog and store purchases. Furthermore, it is demonstrated that the customers in cluster 1, have more purchases using discounts.

TABLE III. DIFFERENCES BETWEEN CLUSTERS

Cluster number	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Complain	Spent	Unnecessary_Purchases	Necessary_Purchases	Total_Purchases
0	1.99	5.76	5.16	8.53	3.80	0.00	1088	666	421	21
1	2.56	2.85	0.83	3.80	6.43	0.01	136	92	44	8

Furthermore, in order to find the difference between clusters based on other attributes, cluster number was added to the primary dataset as a new column. The attributes in which the clusters are significantly different, are shown in charts below which show that the cluster 0, have fewer children and they are less likely to be parents and also, they accept more campaigns than the customers in cluster 1.

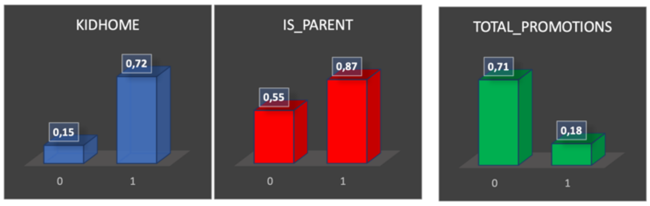

Fig. 5: Difference between clusters based on other attributes

In the chart fig. 6 below, the data distribution is demonstrated based on income per household and the total spending, differentiated based on clusters as colors. In this chart also, the cluster 0 seems to have higher income per household and more spending than cluster 1.

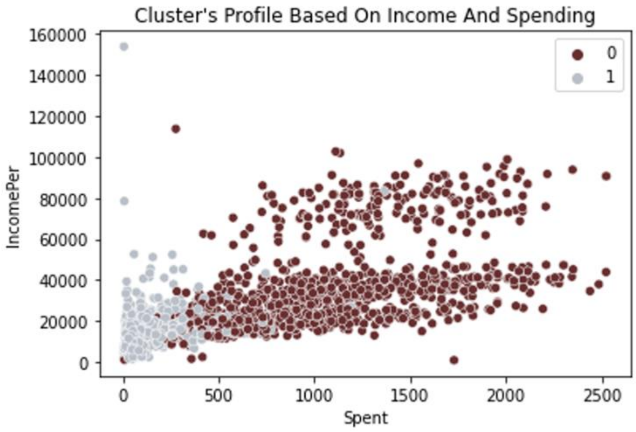

Fig. 6: Cluster's profile based on income and spending

For this part also, neural network is chosen as the predictive model, with 3 layers, SGD as optimizer, learning rate of 0.1 and 45 epochs which resulted in 89% accuracy on the test data. In order to find the influential factors, again the permutation importance model is applied which resulted in the chart below showing income per household to be the most influential factor and after that, total promotions accepted, family size and parental status are effective factors on the model. The results in this model shows that, finding the cluster in which a customer stands based on purchase behaviour can give insights regarding their demographics.

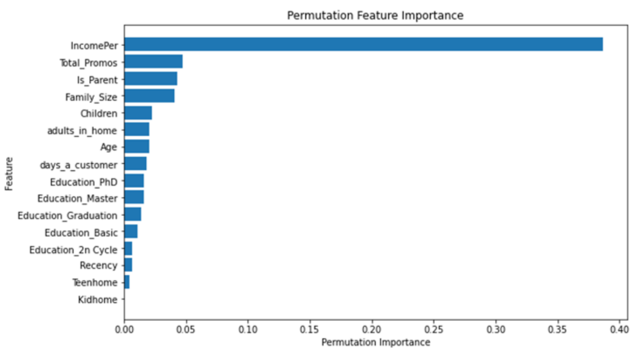

Fig. 7: Permutation feature importance

VII. Discussion

The results of this study's RFM analysis, predictive modeling, and k-means clustering offer a detailed picture of consumer behavior and segmentation. Targeted marketing methods and enhanced customer relationship management greatly benefit from these techniques, which successfully separate consumer groups based on purchasing habits and receptivity to marketing campaigns.

Using purchasing behavior, two separate client groups are identified in clustering, which provides useful information for focused marketing. For instance, Cluster 1 may be more open to premium offers and loyalty programs based on their higher spending and campaign responsiveness, whereas Cluster 0 may react better to cost-effective bargains and discount promotions.

The classification approach's high predictive accuracy of 85%, attained by Neural Network models, indicates a great capacity to forecast client responses based on a wide range of features, such as spending patterns and demographic information. This implies that comprehensive consumer profiles can greatly improve marketing apps' forecasting accuracy.

Our results align with previous research that highlights the value of precise segmentation of customers and tailored marketing tactics to increase sales and engagement. However, by combining two approaches to improve the prediction of campaign responses, our use of Agglomerative Clustering to divide customers based on purchasing patterns and then using Neural Network models for prediction expands on earlier work.

The marketing strategy can be improved by immediately implementing the outcome of our work. Businesses can target customers who are most likely to convert by allocating resources more effectively by knowing which customers are most likely to respond to campaigns. Furthermore, by using clustering findings, marketing communications can be more specifically tailored to the needs and tastes of various client segments.

This research has certain limitations, including its dependence on past purchase data that might not take into consideration changes in customer trends over time or other factors that influence the choice to buy. Because of the high dimensionality of input features in comparison to the size of the dataset, the study's prediction models may also be overfit.

Potential areas for further study could include the dynamic updating of consumer profiles and predictions through the integration of real-time data. It would also be easier to comprehend how consumer behavior evolves over time and in response to various marketing tactics if longitudinal research were conducted. Additionally, exploring other machine learning models and feature selection methods might improve the robustness and generalizability of the predictions.

VIII. Reflection on Group Work

Reflecting on our experiences conducting this data mining project as a group, several aspects stand out. The most interesting part of the project was undoubtedly the process of uncovering insights from the dataset and seeing how clustering techniques like K-means could reveal distinct customer segments. It was fascinating to observe how different demographic and behavioral factors could lead to unique customer profiles. However, the most difficult and time-consuming component of the project was data preprocessing, which included handling values that were missing, abnormalities, and feature engineering. It required careful attention to detail and thorough understanding of the dataset. Throughout the project, we learned valuable lessons about the importance of data quality and the significance of feature selection and engineering in enhancing model performance.

Deciding on the research questions to tackle was a collaborative process, involving brainstorming sessions where we discussed the objectives of the project and the potential avenues of exploration. We aimed to address questions that would provide actionable insights for marketing strategies, leading us to focus on customer segmentation and predictive modeling.

Task division within the group was based on each member's strengths and expertise. Some members took the lead in data preprocessing and feature engineering, while others focused on implementing clustering algorithms and predictive models. Regular communication and coordination were essential to ensure that everyone stayed on track and that the project progressed smoothly.

While the overall workflow followed our initial plan, there were some deviations and adjustments along the way based on the insights gained from the data. For example, we decided to include RFM analysis to further refine our segmentation and provide additional insights into customer behavior, which wasn't initially part of our plan but proved to be valuable.

In terms of contributions, each group member played a vital role in different aspects of the project. Some members were responsible for data cleaning and preprocessing, others focused on model implementation and evaluation, and some contributed to the interpretation of results and the writing of the final report. Overall, the project's success was the result of effective collaboration, communication, and the collective effort of all group members.

IX. References

[1] L. Rives, S. Ruiz de Maya and A. Rubio Bañón, "The Role of Identity Salience in the Effects of Corporate Social Responsibility on Consumer Behavior.," *Journal of Business Ethics,* pp. 65-78, 2009.

[2] M. Kasem, M. Hamada and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," *Neural Computing and Applications*, vol. 36, pp. 4995-5005, 2023.

[3] A. Christy, A. Umamakeswar, L. Priyatharsini and A. Neyaa, "RFM ranking–an effective approach to customer segmentation," *Journal of King Saud*

University-Computer and Information Sciences, vol. 33, no. 10, pp. 1251-1257, 2021.

[4] D. Chen, S. Sain and K. Gao, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *Journal of Database Marketing &Customer Strategy Management*, pp. 197-208, 2012.

[5] D. Olson and B. Chae, "Direct marketing decision support through predictive customer response modeling," *Database Marketing & Customer Strategy Management*, vol. 19, no. 3, pp. 197-208, 2012.

[6] K. Kashwan and C. Velu, "ustomer segmentation using clustering anddata mining techniques," *International Journal of Computer Theory and Engineering*, vol. 5, no. 6, pp. 856-872, 2013.

[7] X. He and C. Li, "The research and application of customer segmentation on e-commerce website," in *6th International Conference on Digital Home*, 2016.

[8] R. Hegde, "Personality classification using data mining approach," vol. 354, no. 1, p. 354–359, 2019.

[9] E. Wong and Y. Wei, " Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model," *International Journal of Retail & Distribution Management*, vol. 46, no. 4, pp. 406-420, 2018.

[10] N. Vo, S. Liu, X. Li and G. Xu, "Leveraging unstructured call log data for customer churn prediction," *Knowledge-Based Systems*, no. 106586, p. 212, 2021.

[11] J. Sun, Z. Tian, Y. Fu, J. Geng and C. Liu, "Digital twins in human understanding: a deep learning-based method to recognize personality traits.," *International Journal of Computer Integrated Manufacturing*, vol. 34, pp. 1-14, 2020.

[12] M. McCrary, "Enhanced customer targeting with multi-stage models: Predicting customer sales and profit in the retail industry," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 17, no. 4, pp. 273-295, 2009.

[13] N. Chaudhuri, G. Gupta, V. Vamsi and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," *Decision Support Systems*, vol. 149, no. 113622, 2021.

[14] M. Hassanein, W. Hussein, S. Rady and T. Gharib, "Predicting Personality Traits from Social Media using Text Semantics," *Int. Conf. Comput. Eng. Syst. (ICCES)*, pp. 184-189, 2018.

[15] T. Jiang and Tuzhlin, A., "Improving personalization solutions through op-timal segmentation of customer bases," *IEEE transactions on knowledgeand data engineering*, vol. 21, no. 3, pp. 305-320, 2008.

[16] E. Bradlow, M. Gangwar, P. Kopalle and S. Voleti, "The Role of Big Data and Predictive Analytics in Retailing," *Journal of Retailing*, vol. 93, no. 1, pp. 79-95, 2017.