THE CHINESE UNIVERSITY of HONG KONG, SHENZHEN

MKT 4220

BIG DATA MARKETING

Project Report

Author:	*Student Number:*
Dharma Setiawan	120040017
Joseph Ariel Christopher	120040002
Darren Boesono	120040022
Fan Wu	120090812
Kaiyou Wu	120090839
Xi Huang	120040071

December 24, 2023

Contents

1 Introduction

1.1 Background

In order to maximize profits, decision makers need to focus on customer relationship management (CRM) to promote customer spending. According to the Pareto principle, 20% of customers contribute more to a company's revenue than the rest (Alsayat 2023). In order to increase the willingness of customers to spend, customer segmentation strategies need to be developed and the nature of their future behavior needs to be assessed (Glanz et al. 2020). Machine learning methods such as RFM, clustering can be used to categorize the customer into different levels based on similarities in behavior and characteristics (Zeybek H. 2018). This technique can help decision makers to recognize the needs of potential customers and match the corresponding marketing strategies.

1.2 Literature review

Jiang and Tuzhilin (2009) proposed the K-Classifiers Segmentation algorithm, which combines customer segmentation and buyer targeting to allocate most of the resources to high-level customers. To reduce the cluster error criterion, Shah and Singh (2012) proposed an execution method similar to K-means algorithm and K-medoids algorithm. However, this method still has limitations when applied to all conditional cases. Cho and Moon (2013) used a weighted frequent pattern mining technique. The weights were reconfigured to improve the accuracy of target customer categorization through RFM modeling. In addition, Lu et al. (2014) used logistic regression to separate transaction data to create a model to analyze customer churn. In this model, customers with the largest churn rate are identified, which makes it easier to analyze the causes of customer churn.

1.3 Research questions

In this research, we aim to segment customer information through different machine learning models. The following are the questions we need to address in our research：

First and foremost, which is the best approach between RFM analysis, Clustering combined with classification, and a standalone classification model based on customer response to previous campaigns, in terms of their abilities to segment customers and provide actionable business insights for optimizing marketing strategies?

Secondly, given our clusters, what segments do we get? How much is the potential Gross Merchandise Value (GMV) and optimization cost? What is the personalized approach marketing strategy for each segment?

Finally, how might a company implement a customer segmentation service in their internal stack?

2. Exploratory Data Analysis

2.1 Understanding the Data & Descriptive Statistic

We collected data from Marketing Campaign (kaggle.com). This data has 2240 rows, 29

features, 0 duplicates and 24 missing values. The description of data is shown in Appendix Table 1 after dropping the missing values.

2.2 Univariate Analysis

As shown in Appendix Figure 1, most of the dataset has highly positively skewed distribution. And as shown in Appendix Figure 2, there are some outliers in each feature and most features are not symmetric.

2.3 Multivariate Analysis

Through Appendix Figure 3 we can find the following 3 points. First, customers are more likely to buy wine and meat products. Second, customers tend to buy products directly in stores. Third, the number of visits to the company's website and number of children in customers' households contribute little to the response.

2.4 Feature Engineering

To further analyze the data in detail, we created some new features based on the original features. And we applied the method of normalization and standardization on processing the data of numerical features and label encoding on processing the data of categorical features to make them more suitable to be used in our model.

3. Methodology

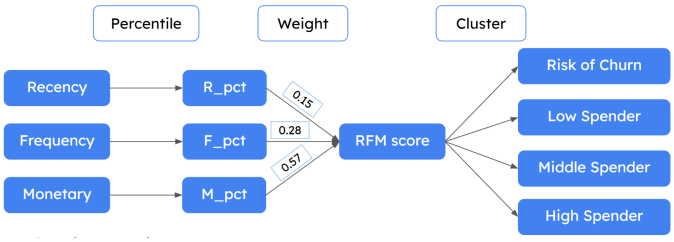
Silhouette score: a metric used to calculate the goodness of a clustering technique. It can be calculated using the following steps. First, calculate the average distance. For each data point i calculate the average distance of i to all other data points in the same cluster (intra-cluster distance) as a_i and the average distance of i to all data points in the nearest cluster (inter-cluster distance) as b_i. Second, calculate silhouette score for each point using the formula Silhouette score$(i) = \frac{max(a_i, b_i)}{b_i - a_i}$. Third, calculate the overall silhouette score by averaging the individual silhouette score of all the points. For the following methods used, we use silhouette score as the evaluation metric.

3.1 RFM

Motivation: RFM stands for recency, frequency, and monetary, respectively. Specifically, recency refers to how recently has the customer made a transaction. Frequency refers to how frequent the customer is in buying products. Monetary refers to how much does the customer spend on purchasing products.

Methodology: In business practice, RFM analysis is often used to divide customers into different segments based on their RFM values. In our case, recency is directly given in the dataset. Frequency is calculated as the column sums of NumDealsPurchases, NumWebPurchasses, NumCatalogPurchases, and NumStorePurchases. Monetary is calculated as the column sums of MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, and MntGoldProducts. Then the percentile ranks of the R, F, M variables are calculated. Finally, RFM score is calculated as a weighted sum of

the three percentile ranks. The segment criteria is as follows, RFM score ≤ 1, risk of churn; 1 < RFM score ≤ 2, low spender; 2 < RFM score ≤ 3, middle spender; RFM score > 3, high spender.



The reason why we chose 0.15, 0.28, and 0.57 as the weights for recency, frequency, and monetary is that the weights are commonly used in a general business context. We used it as a prior knowledge and we didn't focus too much on it although we can treat it as a hyperparameter and perform hyperparameter tuning to find the "best" weights that returns the highest silhouette score. In general, the weights should be determined based on a thorough understanding of the business context. Therefore, there is no one-size-fits-all answer to the choice of the weight. It needs to be tailored to the specific needs and objectives of each business.

Evaluation: We use silhouette score to evaluate the performance of our models.

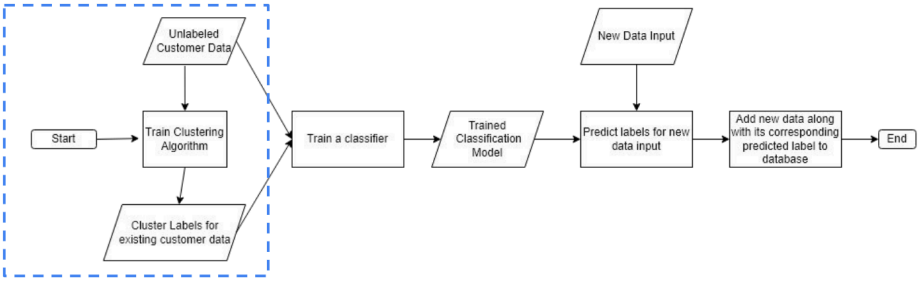
Result: The silhouette score for RFM analysis is 0.528. Here are some interesting findings of the RFM analysis (see Appendix Figure 4). 1. After segmenting the customers, we find that high spenders typically have higher income, while low spenders usually have lower income. However, the type of customers is educational independent, meaning the customers' educational level is evenly distributed among the four segments (Fig. 4a). 2. To avoid Simpson's Paradox, we stratify the data with customer segments. We find that high spenders typically have higher frequency, but recency can be varied (Fig. 4c and 4d). 3. Frequency and recency; monetary and recency are significantly positively correlated within customer segments. Although monetary and frequency seem to be positively correlated, they are significantly negatively correlated within the high spender segment (Fig. 4e and 4b). 4. Customers with higher income may not necessarily spend a lot on purchases, but low monetary customers generally have lower incomes (Fig. 4f). 5. Generally, customers with higher income purchase products more frequently (Fig. 4g). 6. The correlation between recency and income is not significant (Fig. 4h).

Cluster Interpretation:

3.2 Clustering + Classification

3.2.1 Clustering

Motivation: Although RFM clustering method can be done to segment existing customers, RFM method is not robust enough. We will use advanced machine learning methods to identify the data of the customers.



Methodology: To find the most suitable model for our specific use case, we tested out many different clustering models, as shown in Table 2 in the appendix. First, we load the data. After that, we handle missing and duplicated data. In feature engineering, we create several features to enhance the model. Then, we scaled the data using standardization. Lastly, we fit the data into several algorithms and evaluate the silhouette score.
From table 2, Ordering Points To Identify Cluster Structure (OPTICS) perform the best with the highest silhouette score. OPTICS is a density-based clustering technique that allows partitioning data into groups with similar characteristics (clusters). It addresses one of the DBSCAN's major weaknesses. The problem of detecting meaningful clusters in data of varying density. In a density based clustering, clusters are defined as dense regions of data points separated by low-density regions.
Evaluation: OPTICS produces a hierarchical clustering result, providing a more comprehensive view of the data structure. This hierarchical information can be valuable for understanding varying densities and shapes of clusters within the dataset. Moreover, OPTICS is also robust to noise and outliers in the data. It can identify clusters of varying shapes and sizes while gracefully handling data points that do not belong to any cluster. This model is also Well-suited for datasets with clusters of varying densities. It can adapt to clusters that have different levels of density, making it effective in scenarios where traditional methods like K-Means may struggle
Result: As we can observe, advanced models, especially those who can handle nonlinear data have better silhouette score.The code for this section is available in Code 3 and 4 of the attached zip file.
Cluster Interpretation: As depicted in Figure A, high spender customers exhibit distinct characteristics. Notably, they tend to possess the highest income, aligning with expectations. Surprisingly, despite their high spending habits, these customers infrequently visit the website and seldom take advantage of discounted purchases. Additionally, high spender customers demonstrate a notable high conversion rate. Their age distribution is diverse, but a noteworthy pattern emerges— the younger demographic, specifically Generation Y, is more prone to churn, indicating a tendency to discontinue engagement with the product when preferences change.Interestingly, a common trait among high spender customers is the absence of children in their households. In contrast, low spender customers typically have 2-3 children.

3.2.2 Classification using generated clusters

Motivation: Although clustering can segment existing customers, not all clustering algorithms can efficiently cluster new data points (future customer data). To solve this problem, we train a classifier, using the generated cluster labels as ground truth labels, to segment new customers efficiently.

Methodology: To find the most suitable model for our specific use case, we tested out many different classification models, as shown in Table 3 below. We split out existing data into three parts: training, validation, and testing. We employ 5-fold cross-validation for all models. Our testing data consists of 20% of our overall data, our training data consists of 80%, and our validation data for each fold consists of 20% of our training data (16% of our overall data).
Evaluation: We use accuracy to evaluate the performance of our models.
Result: As we can observe, most models could accurately predict the customer segment given customer data. K-Nearest Neighbours (KNN) and Decision Tree Classifier performed the best in our specific situation here, so we can pick either to infer the segments for new customers. The code for this section is available in Code 6 of the attached zip file.

Model Names	**KNN**	SVM	Logistic Regression	Naive Bayes	Random Forest Classifier	**Decision Tree Classifier (DTC)**	XGBoost Classifier	Fully-Connected Neural Network
Testing Accuracy (%)	**98.91**	98.23	98.07	97.92	98.43	**98.91**	98.89	98.77

Table 3

3.3 Classification using customer data with response to past campaigns
Motivation: We also need to predict whether the customers will join in the next campaign or not. We will use supervised learning binary classification methods to segment the customer.
Methodology: To find the most suitable model for our specific use case, we tested out many different classification models, as shown in Table 4 below. We will handle duplicate rows, removing invalid values, and split the dataset into training (75%) and testing (25%) sets. We will test out 6 different supervised machine learning models shown in Table 4.
Evaluation: We use accuracy, recall, and F1 score to evaluate the performance of our models.
Result: As we can observe, most models could accurately predict the customer response given customer campaign data. Logistic regression and support vector machines performed the best in this situation here, so we can pick either to predict whether a customer will join the next campaign or not.

Model	Accuracy	Precision	Recall	F1 Score	Cross Val F1 (k=5)	ROC AUC	Cross Val ROC AUC (k=5)
LogisticRegression	0.861000	0.523000	0.675000	0.589000	0.510000	0.900000	0.885000
LinearSVC	0.859000	0.520000	0.614000	0.564000	0.448000	0.895000	0.884000
MLPClassifier	0.846000	0.483000	0.518000	0.500000	0.558000	0.830000	0.887000
GradientBoostingClassifier	0.846000	0.482000	0.494000	0.488000	0.487000	0.725000	0.752000
KNeighborsClassifier	0.845000	0.482000	0.639000	0.549000	0.400000	0.842000	0.810000
DecisionTreeClassifier	0.838000	0.456000	0.494000	0.474000	0.488000	0.699000	0.718000

Table 4

4. Implications

4.1 Theoretical Implications

In this section, we compare the methods explored in this paper across four key categories: performance, results interpretability, model explainability, and computational cost.

Performance: We assess performance using silhouette scores for clustering algorithms and accuracy for classification algorithms. RFM analysis achieved a silhouette score of 0.528, while the optimal clustering algorithm (OPTICS) reached a score of 0.72. The clustering + classification approach, particularly with KNN and DTC models, outperformed pure classification using customer data with response to past campaigns, achieving an accuracy score of 98.91%, compared to 86.1%. Overall, our clustering + classification approach demonstrated superior performance.

Results Interpretability: RFM analysis and the clustering + classification approach provide easily interpretable results, allowing for actionable insights by analyzing the generated clusters and customer features. In contrast, pure classification offers limited interpretability, as it only indicates whether a customer is predicted to engage with the new campaign without providing insights into the decision rationale.

Model Explainability: RFM analysis and pure classification, especially with a logistic regression model, exhibit high explainability. However, the clustering + classification approach, due to prior dimensionality reduction during cluster generation, faces challenges in determining feature importance, impacting model explainability.

Computational Cost: RFM analysis boasts the lowest computational cost. Pure classification has a moderate cost, while the clustering + classification approach has the highest computational cost among the explored methods due to dimensionality reduction being applied and subsequent model complexity.

4.2 Marketing Implications

Our approach holds theoretical promise and offers significant commercial potential. Companies can tailor campaigns to specific segments by effectively segmenting customers and optimizing profitability. Our analysis categorized customers into four segments: high spenders, middle spenders, low spenders, and those at risk of churn. Examples of targeted campaigns for each segment are as follows:

High Spenders: Offer opulent and exclusive services like personalized concierge service, valet parking, and exclusive preview sales.

Middle Spenders: Limited-time offers, upgrade and save campaigns, referral rewards programs.

Low Spenders: Frequent shopper discounts, free shipping vouchers, and first-purchase discounts.

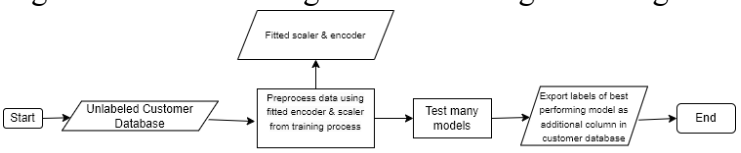
Risk of Churn: Reactivate customers through email campaigns with exclusive offers.

These targeted marketing strategies, facilitated by automated customer segmentation, can significantly boost profits. For instance, projecting a 15% conversion of middle spenders into high spenders could result in an additional $35,154 in revenue. This estimate is derived by calculating the difference in average spending between high and middle spenders and multiplying it by the target conversion rate of 15% among middle spenders (approximately 102 individuals).

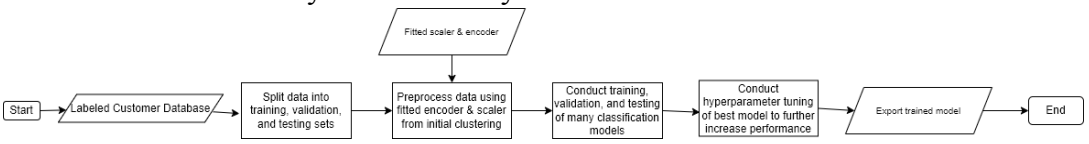
4.3 Managerial Implications

Given the substantial theoretical and marketing benefits that a customer segmentation service provides to a company. It is natural for a company to seek an automated, scalable, efficient, and accurate customer segmentation service. Our paper proposes a 4-step solution, which is as follows:

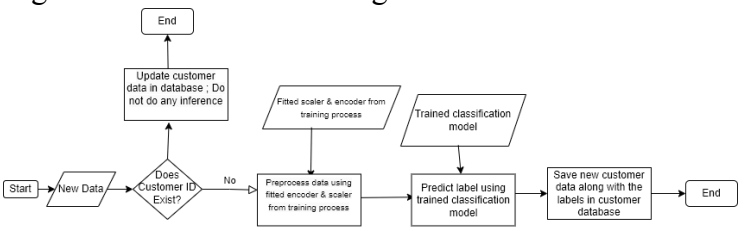
1. Segmentation of existing customers using clustering methods.



2. Train a classification model using the cluster labels from the previous step as ground truth labels for efficiency and scalability of future inference



3. Segment new customers using the trained classification model



4. Conduct model monitoring & maintenance to ensure the reliability of results. Examples are cluster stability, data drift, and classification model performance monitoring.

Our 4-step solution provides a simple framework for implementing the Clustering + Classification approach we explored in this paper, which offers a general guide towards implementing this approach under different use cases.

Appendix

Figure 1: Distplot Analysis

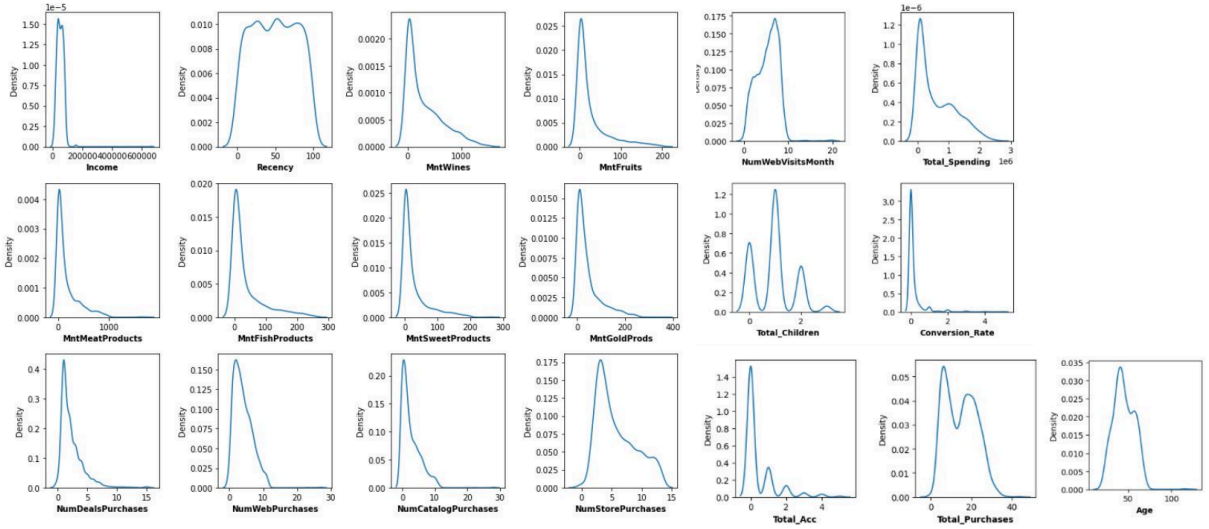


Figure 2: Box Plot Analysis

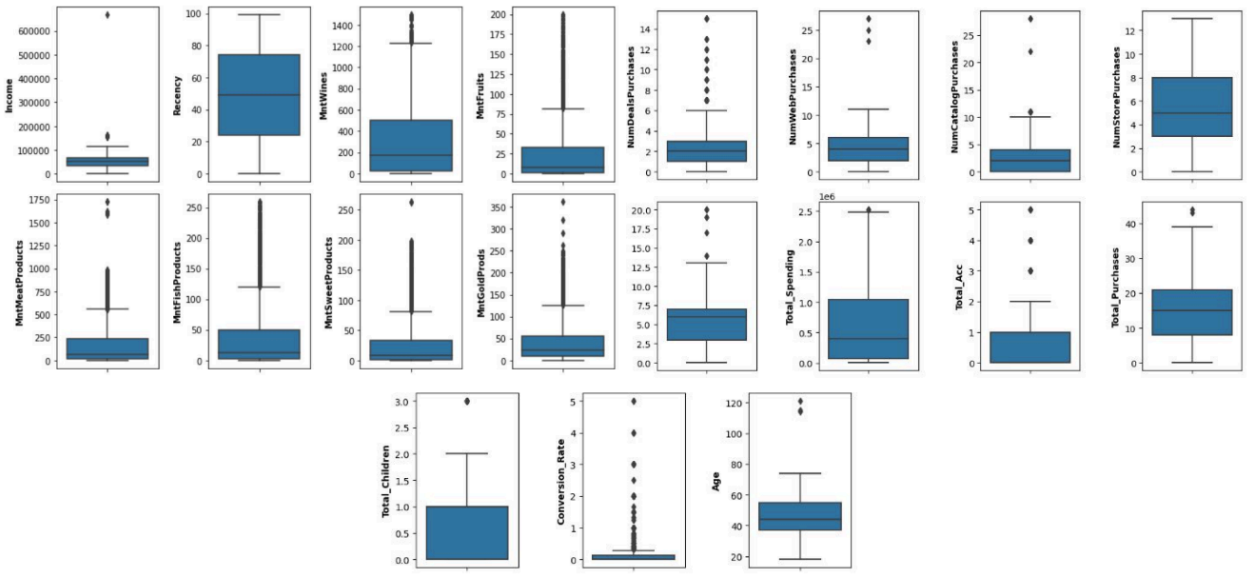


Figure 3: Correlation Heatmap

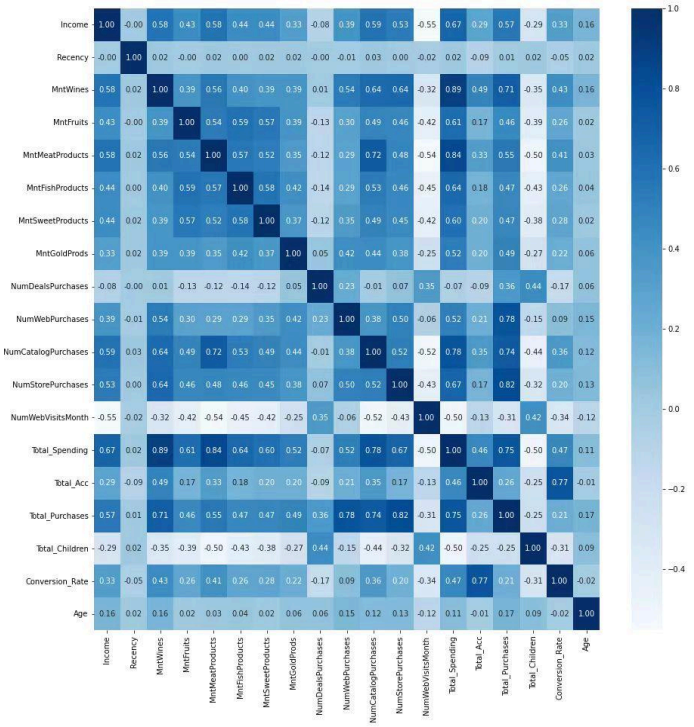
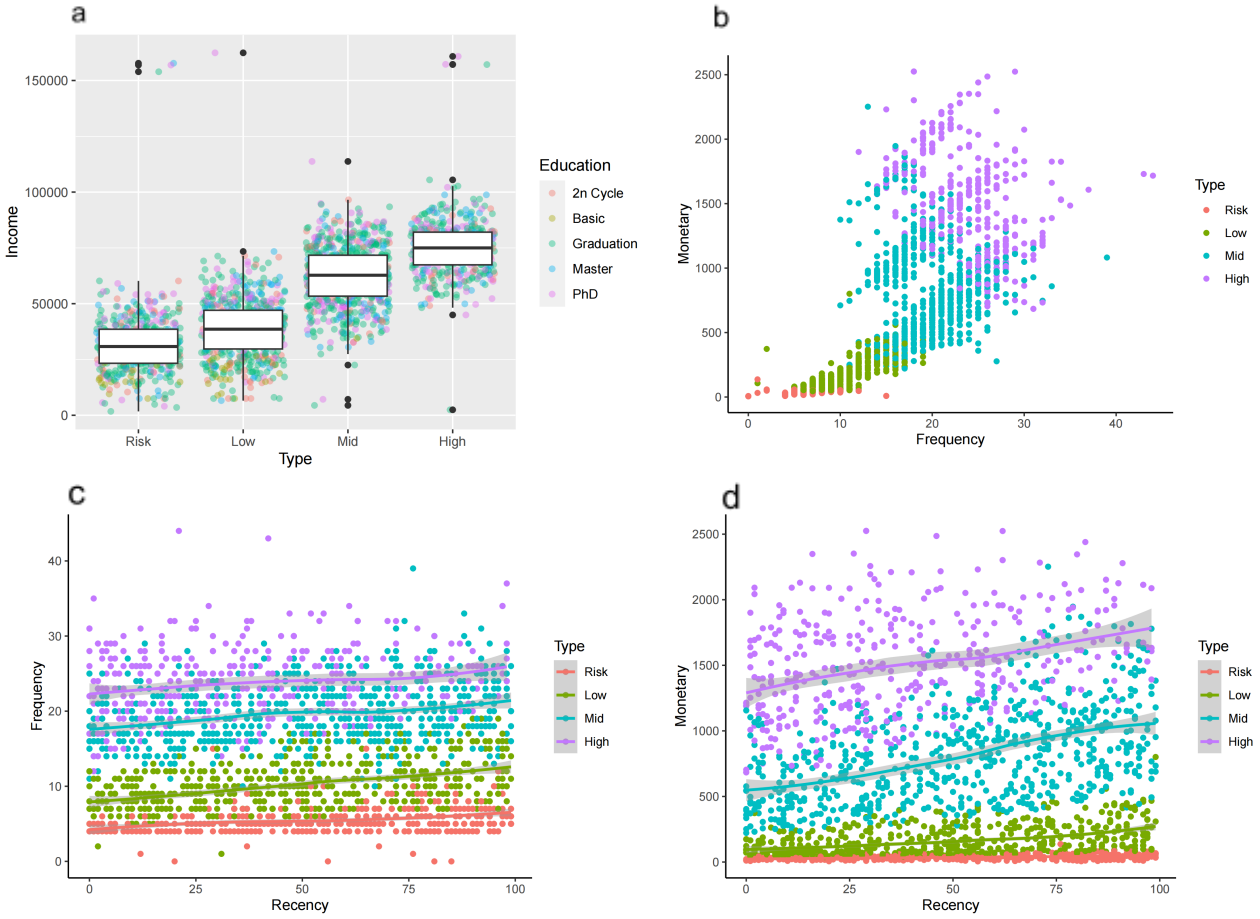


Figure 4: RFM analysis



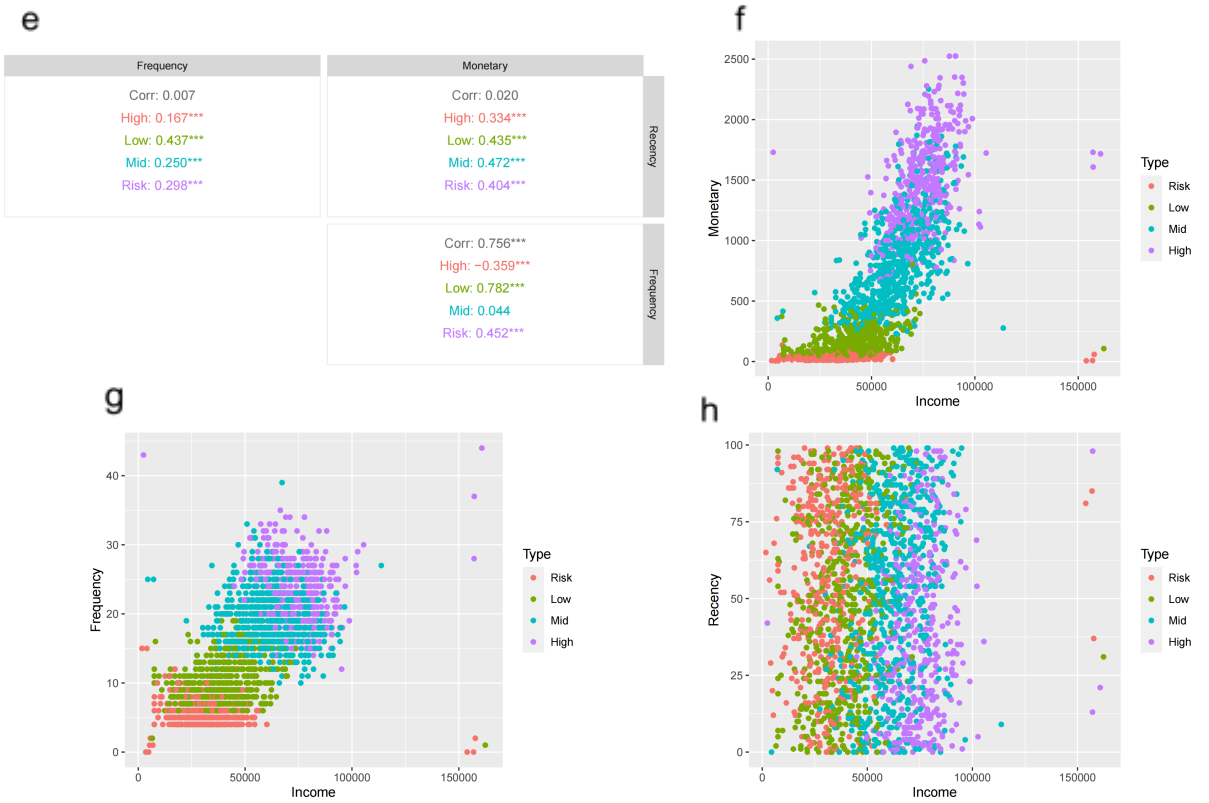


Table 1-1 Description of features

	count	mean	std	min	25%	50%	75%	max
ID	2216.0	5588.353339	3249.376275	0.0	2814.75	5458.5	8421.75	11191.0
Year_Birth	2216.0	1968.820397	11.985554	1893.0	1959.00	1970.0	1977.00	1996.0
Income	2216.0	52247.251354	25173.076661	1730.0	35303.00	51381.5	68522.00	666666.0
Kidhome	2216.0	0.441787	0.536896	0.0	0.00	0.0	1.00	2.0
Teenhome	2216.0	0.505415	0.544181	0.0	0.00	0.0	1.00	2.0
Recency	2216.0	49.012635	28.948352	0.0	24.00	49.0	74.00	99.0
MntWines	2216.0	305.091606	337.327920	0.0	24.00	174.5	505.00	1493.0
MntFruits	2216.0	26.356074	39.793917	0.0	2.00	8.0	33.00	199.0
MntMeatProducts	2216.0	166.995939	224.283273	0.0	16.00	68.0	232.25	1725.0
MntFishProducts	2216.0	37.637635	54.752082	0.0	3.00	12.0	50.00	259.0
MntSweetProducts	2216.0	27.028881	41.072046	0.0	1.00	8.0	33.00	262.0
MntGoldProducts	2216.0	43.965253	51.815414	0.0	9.00	24.5	56.00	321.0
NumDealsPurchases	2216.0	2.323556	1.923716	0.0	1.00	2.0	3.00	15.0
NumWebPurchases	2216.0	4.085289	2.740951	0.0	2.00	4.0	6.00	27.0
NumCatalogPurchases	2216.0	2.671029	2.926734	0.0	0.00	2.0	4.00	28.0
NumStorePurchases	2216.0	5.800993	3.250785	0.0	3.00	5.0	8.00	13.0
NumWebVisitsMonth	2216.0	5.319043	2.425359	0.0	3.00	6.0	7.00	20.0
AcceptedCmp1	2216.0	0.064079	0.244950	0.0	0.00	0.0	0.00	1.0
AcceptedCmp2	2216.0	0.013538	0.115588	0.0	0.00	0.0	0.00	1.0
AcceptedCmp3	2216.0	0.073556	0.261106	0.0	0.00	0.0	0.00	1.0
AcceptedCmp4	2216.0	0.074007	0.261842	0.0	0.00	0.0	0.00	1.0
AcceptedCmp5	2216.0	0.073105	0.260367	0.0	0.00	0.0	0.00	1.0
Complain	2216.0	0.009477	0.096907	0.0	0.00	0.0	0.00	1.0
Z_CostContact	2216.0	3.000000	3.000000	3.0	3.00	3.0	3.00	3.0
Z_Revenue	2216.0	11.000000	11.000000	11.0	11.00	11.0	11.00	11.0
Response	2216.0	0.150271	0.357417	0.0	0.00	0.0	0.00	1.0

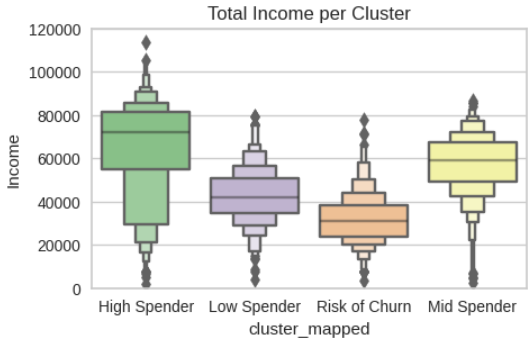
Table 1-2 Content of some features

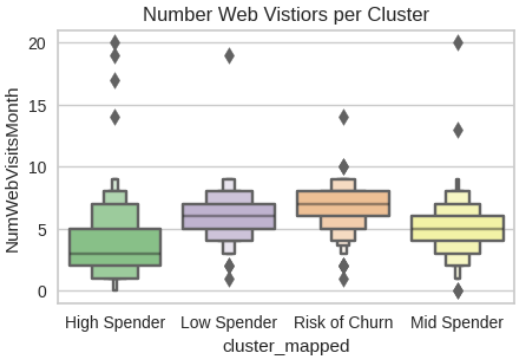
Education	customer's level of education
Marital	customer's marital status
Income	customer's yearly household income
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Recency	number of days since the last purchase
MntWines	amount spent on wine products in the last 2 years
MntFruits	amount spent on fruits products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFishProducts	amount spent on fish products in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntGoldProducts	amount spent on gold products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumWebPurchases	number of purchases made through company's web site
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebVisitsMonth	number of visits to company's web site in the last month
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Complain	1 if customer complained in the last 2 years
Response(target)	1 if customer accepted the offer in the last campaign, 0 otherwise

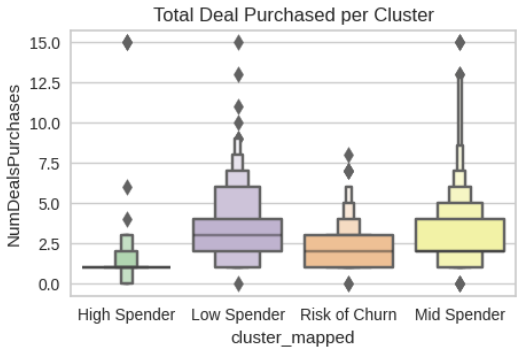
Table 2 Model Comparison

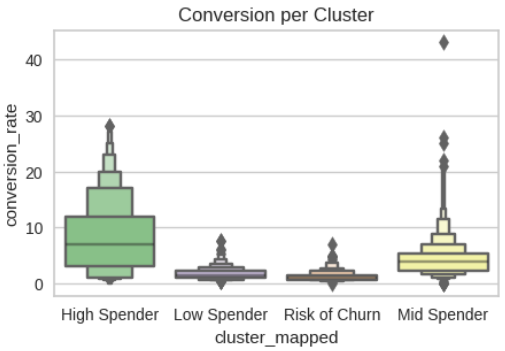
Model / Performance	Description	Silhouette Score
K-Means	K-Means is a centroid-based clustering algorithm that partitions data into K clusters, assigning each data point to the cluster with the nearest centroid.	0.1666
K-Means PCA	K-Means with PCA helps to handle high-dimensional data and improve clustering performance.	0.5786
K-Means tSNE	K-Means with t-SNE incorporates t-SNE dimensionality reduction to visualize and cluster data points in a lower-dimensional space, preserving local structures.	0.6723
Mini Batch K-Means	Mini Batch K-Means is a variant of K-Means that processes random subsets (mini-batches) of the data at each iteration, making it computationally more efficient for large datasets	0.6702
Agglomerative	Agglomerative Clustering is a hierarchical clustering algorithm that	0.6681

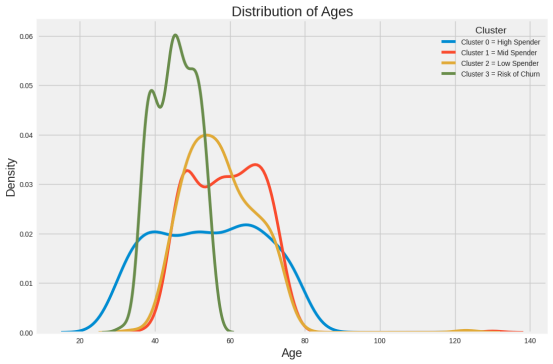
Clustering	starts with individual data points and progressively merges them into clusters based on similarity, forming a tree-like structure	
DBSCAN	DBSCAN is a density-based clustering algorithm that identifies clusters based on the density of data points, discovering clusters of varying shapes and handling noise effectively	0.6651
OPTICS	OPTICS (Ordering Points To Identify the Clustering Structure) is a density-based clustering algorithm that extends DBSCAN by providing a hierarchical view of clusters and identifying varying densities within the data.	0.7199
BIRCH	BIRCH is a hierarchical clustering algorithm designed for large datasets, using a tree structure to efficiently group data points and reduce memory requirements	0.6714
SOM	SOM is an unsupervised neural network that is trained using unsupervised learning techniques to produce a low dimensional, discretized representation from the input space of the training samples, known as a map and preserves the topological properties of the input space	0.16

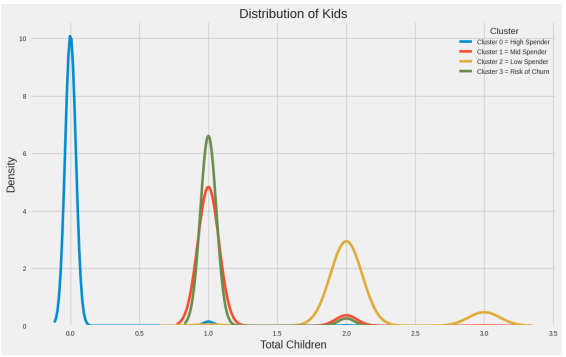


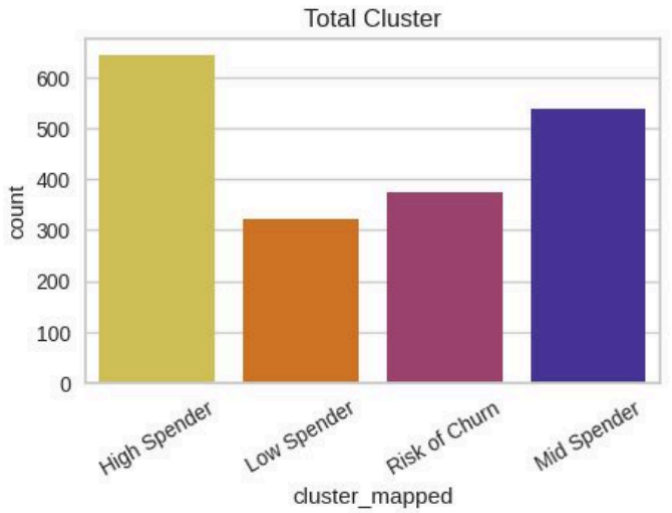







Distribution of Ages
Cluster
Cluster 0 = High Spender
Cluster 1 = Mid Spender
Cluster 2 = Low Spender
Cluster 3 = Risk of Churn
Density
Age


Distribution of Kids
Cluster
Cluster 0 = High Spender
Cluster 1 = Mid Spender
Cluster 2 = Low Spender
Cluster 3 = Risk of Churn
Density
Total Children


Total Cluster
count
600
500
400
300
200
100
0
High Spender
Low Spender
Risk of Churn
Mid Spender
cluster_mapped

Reference

Alsayat, A. (2023). Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. *Neural Computing and Applications*, 35(6), 4701-4722.

Glanz, K., Chung, A., Morales, K. H., Kwong, P. L., Wiebe, D., Giordano, D. P., et al. (2020). A. The healthy food marketing strategies study: Design, baseline characteristics, and supermarket compliance. *Translational Behavioral Medicine*, 10, 1266–1276.

Gui biebei. (2020). Research on customer churn prediction based on customer segmentation. Yunnan University of Finance and Economics, pp. 14-15.

H. Lu, J.Lu. Lin, G. Zhang. (2014). A customer churn prediction model in the telecom industry using boosting. *IEEE Trans. Ind. Inf.*, 10 (2), pp. 1659-1665.

Shah, S., Singh, M., 2012. Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm. *In: 2012 International Conference on Communication Systems and Network Technologies, Rajkot,* pp. 435–437.

T.Jiang, A.Tuzhilin. (2009). Improving personalization solutions through optimal segmentation of customer bases. *IEEE Trans. Knowledge Data Eng.*, 21 (3) (March 2009), pp. 305-320.

Young Cho, S.C. Moon. (2013). Weighted mining frequent pattern-based customer's RFM score for personalized u-commerce recommendation system. *J. Converg.*, 4(2013), pp. 36-40.

Zeybek H. (2018) Customer segmentation strategy for rail freight market: The case of Turkish State Rail ways, *Research in Transportation Business & Management,* 28 (2018), pp. 45-53.