E-commerce Customer Segmentation via Unsupervised Machine Learning

Boyu Shen, University of Wisconsin-Madison, bshen27@wisc.edu (mailto:bshen27@wisc.edu)

DOI: https://doi.org/10.1145/3448734.3450775 (https://doi.org/10.1145/3448734.3450775)
CONFCDS 2021: The 2nd International Conference on Computing and Data Science (https://doi.org/10.1145/3448734), Stanford, USA, January 2021

Customer segmentation through data mining could help companies conduct customer-oriented marketing and build differentiated strategies targeted at diverse customers. However, there has not been a guideline for systematic implementation of customer segmentation given the raw transaction data. This study focuses on a real-world database from an online transaction platform with the purpose to develop a guideline for customer segmentation for the business. Since the raw data are unlabeled, unsupervised machine learning methods are utilized. This study firstly employs the RFM model to create behavioral features; next, the TF-IDF method is applied to the product descriptions to generate product categories; then, K-means clustering algorithm is used to group customers. After customers are grouped, association rules mining by Apriori Algorithm is used to analyze purchased products. Principle Component Analysis (PCA) and T-Distributed Stochastic Neighbor Embedding (T-sne) methods are utilized to reduce the dimension of data in order to create visualizations. Finally, some concrete recommendations for the business based on the results are provided accordingly.

CCS Concepts: • **Computing methodologies** → **Modeling methodologies**;

KEYWORDS: Customer Segmentation, RFM model, TF-IDF, Unsupervised Machine Learning, K-means, PCA, T-sne, Association Rules, Apriori Algorithm

ACM Reference Format:
Boyu Shen. 2021. E-commerce Customer Segmentation via Unsupervised Machine Learning. In *The 2nd International Conference on Computing and Data Science (CONFCDS 2021), January 28-30, 2021, Stanford, CA, USA.* ACM, New York, NY, USA, 9 Pages. https://doi.org/10.1145/3448734.3450775 (https://doi.org/10.1145/3448734.3450775)

1 Introduction

E-commerce sales have been increasing rapidly in the past few years. According to Statista, retail e-commerce sales amounted to 3.53 trillion US dollars in 2019 and e-retail revenues are projected to grow to 6.54 trillion US dollars in 2022 [1]. This significant growth indicates that the shopping ways of customers have dramatically changed.

Compare with traditional sales, E-commerce has a unique characteristic that all the transaction information including the shopping time, items, and prices can be tracked and stored accurately. Based on detailed transaction data, the business is allowed to gain more insights into customers' behaviors and preferences by classifying them into meaningful groups to satisfy their needs more efficiently. For customer segmentation, there have been many studies focusing on RFM (Recency, Frequency, Monetary Value) model and clustering algorithms, but associations between customer groups and the products they have purchased have not been substantially studied. Furthermore, there has not been a guideline for customer segmentation given the raw transaction data.

Consequently, the main purpose of this study is to develop a systematic implementation of customer segmentation for the business. To distinguish diverse customers, customers' behavioral characteristics are obtained from the RFM model (Recency, Frequency, Monetary Value). Besides, purchased products are classified into different categories by Term Frequency - Inverse Document Frequency (TF-IDF) and clustering methods to reflect customers' product preferences. For the modeling part, customers have been segmented into several groups using the k-means clustering algorithm, and the clusters are visualized after reducing the data dimension by Principle Component Analysis (PCA) and T-Distributed Stochastic Neighbor Embedding (T-sne) methods. The main characteristics of the consumers in each segment have been explored and identified. Finally, association rules mining by Apriori algorithm is used to find associations between products for a specific group of customers. Some concrete recommendations for the business based on the results are provided accordingly. In appliance with that, real-world online transaction data was used to group customers based on the proposed methodologies.

2 Related works

This section explores the theoretical parts of different algorithms and methods that are employed in this study, and some related applications of the methods are also included. This study reviews related studies of RFM model, K-means Clustering Algorithm, Association Rules and Dimension Reduction Algorithms (PCA& T-sne).

2.1 RFM model

Hughes (1994) put forward the RFM analytic model which is usually utilized to characterize important customers from a large dataset by three variables: Recency, Frequency, and Monetary value [2]. The detailed definitions of the RFM model are presented as below:

- Recency of a customer (R)

The time interval between the most recent date in the data and the date of the most recent transaction of the customer. The shorter the interval is, the bigger the R is.

- Frequency of a customer (F)

The number of transactions of the customer in a specific period, in this case, a year. The bigger the frequency is, the bigger the F is.

- Monetary value of a customer (M)

The sum of the amount spent in each transaction of the customer. The bigger the sum is, the bigger M is.

Razieh et al. (2012) utilized the RFM model to measure customer loyalty in their article "Developing a model for measuring customer loyalty and value with RFM technique and clustering algorithms" [3]. According to this paper, understanding their customers' behaviors and priorities based on the data allows the companies to provide personalized offers and increase customers' usage time of existing products. The authors discussed the RFM technique and how to use RFM scores as inputs of the clustering algorithm.

2.2 TF-IDF method

TF-IDF serves as a statistical measure that evaluates how relevant a word is to a document in a collection of documents. Shi et al. (2009) summarized that TF-IDF was implemented by multiplying two metrics: how many times a word appears in a document and the inverse document frequency of the word across all the documents [4].

2.3 K-means Algorithm

Clustering is the process of grouping a set of observations into groups of similar observations. A cluster is a collection of observations that are similar to one another within the same cluster and are dissimilar to the observations in other clusters. The K-means clustering is a simple and popular approach for partitioning a data set into K distinct clusters. To perform K- means clustering, the desired number of clusters K has to be specified, then assign each observation to exactly one of the K clusters. The detailed calculating process for K-means is presented as below:

Step 1: Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

Step 2: Iterate until the cluster assignments stop changing.

a) For each of the K clusters, computing the cluster centroid.

b) Assign each observation to the cluster whose centroid is closest (closest is defined by Euclidean distance).

Razieh et al. (2012) also gave an example of how to implement the k-means algorithm to group customers [3]. They provided detailed calculations of the algorithm and explained how to determine the loyalty of customers based on which the customers are classified after clustering.

2.4 Association Rules

An association rule is usually in the form of X→ Y, in which X is the rule's antecedent and Y is the consequent of the rule [5]. There are three common ways to measure the reliability of the association:

Measure 1: Support. Support value is measured by the proportion of transactions in which both X and Y appear. The support threshold is a minimum proportion of the itemset, and itemsets with support values above this threshold are regarded as significant itemsets.

Measure 2: Confidence. The confidence is measured by the proportion of transactions with item X, in which item Y also appears.

Measure 3: Lift. The lift value is measured by the conditional probability of Y on the occurrence of X divided by the probability of occurrence of Y.

There are different algorithms mining the association rules such as Apriori Algorithm and FP-Growth Algorithm. Mythili (2013) gave a comparison of these two algorithms and provided the benefits and drawbacks of them [5]. According to this article, FP-Growth Algorithm generally performs better than Apriori Algorithm, but Apriori Algorithm is extremely convenient to be implemented and easy to be understood.

2.5 Dimension Reduction Algorithms

It is impossible for humans to visualize space with more than three dimensions so feature dimensions need to be reduced in order to create visualizations of the results. Two algorithms are used to reduce dimension in order to create the visualization: PCA and T-sne. The detailed mathematical calculations of these two methods can be found in this article [6].

T-Distributed Stochastic Neighbor Embedding (T-SNE) is a technique for dimensionality reduction and is particularly well suited for the visualization of high-dimensional datasets. "The SNE transformation is a bijection from the original multidimensional feature space to the mapping space (2-dimensional or 3-dimensional)" [6]. T-SNE minimizes the divergence between two distributions: a distribution that describes pairwise similarities of the original input observations and a distribution that describes pairwise similarities of the corresponding low-dimensional points in the embedding.

Principal component analysis (PCA) is a technique to compute the principal components of the data, and the first few principal components are usually used and the rest are ignored. It is commonly used for dimensionality reduction by projecting each data point onto the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

3 Research methodology

This section introduces the research methods of this study and the proposed procedure for grouping customers. This study is divided into several stages, and Fig.1 shows an overview of the steps taken in this project.

With the business objective of building differentiated strategies according to customers' characteristics, customer segmentation is used to help the business better understand its customers. The customer transaction data used in this project included 8 variables as shown in Table 1, and it contains all the transactions occurring in 2019. As the given data was real-world data, the quality was not good enough to work on it. Further data processing and feature engineering are needed before implementing clustering algorithms.

Table 1 Variables in the transaction dataset

Variable Name	Description
Transaction ID	A 6-digit number uniquely assigned to each transaction
Item ID	A 5-digit number uniquely assigned to each distinct product
Description	Item Name
Quantity	The quantity of each item per transaction
Unit Price	Product Price Per Unit
Customer ID	A 5-digit number uniquely assigned to each distinct customer
Time	The day and time when each transaction was generated
Country	Deliver Address Country

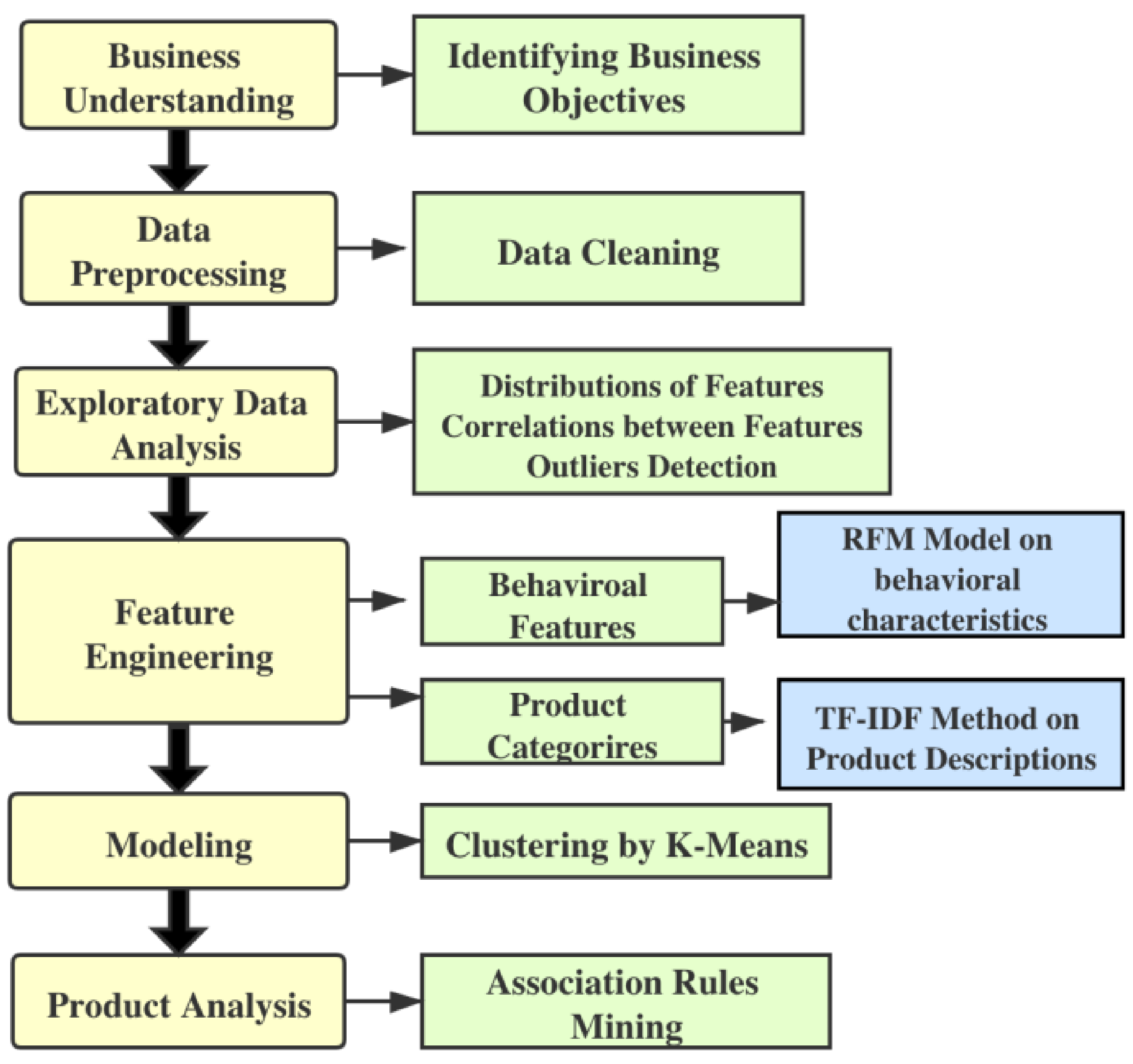

Fig. 1. Overview of the Research Methodology

3.1 Data Preprocessing

Two columns contain missing data which are Customer ID and Product Description. Customer ID is a nominal variable with no useful hints for filling the missing values, and it is essential in our project since all the analyses were centered around customers. Thus, the missing values of Customer ID were dropped. Product Description is also a nominal variable and contains various categories. Given the small percentage of missing values and difficulty to fill the data, missing Descriptions were also dropped. After comparing all the columns, some data duplicates were detected and dropped afterward.

3.2 Exploratory Data Analysis

Distributions and correlations of variables are explored and outliers detection is also conducted in this section. The distribution of variable country shows the data is largely dominated by orders made from China with a small part from other countries. One noticeable finding is that the Transaction

IDs are not only numerical since there is a "C" before the other numbers for every negative value in the quantity column, which could mean that the order was canceled. After deeper exploration, it's been detected that some cancellation orders canceled only a part of a previous transaction. An additional column named "Quantity Canceled" is added to record the quantity canceled for each transaction with a cancellation counterpart, and this variable is important for building the following RFM model. After the additional column was added, all the cancellation transactions were deleted.

Another interesting finding is that the item IDs variable included some special codes that do not characterize the customers, but they indicated particular transactions or have other meanings instead. The discovered special codes and their contents are listed in Table 2. All of these special codes were dropped since they would not provide any useful information on customers' products preference.

Table 2 Special codes in item id

Special Codes	Contents
BANK CHARGES	Bank Charges
C2	Carrriage
M	Manual
PADS	Pads to Match All Cushions
POST	Postage

3.3 Feature Engineering

The raw data only contains variables describing transaction histories which are not appropriate to serve as inputs for clustering algorithms, so further transformations are needed to gain more meaningful features. In this study, two kinds of features were created: behavioral characteristics features and products categories.

RFM model was employed in this section to calculate customer values and then obtain behavioral characteristics of customers. Six variables from the original data were chosen to create the features, the chosen variables are UnitePrice, Quantity, Quantity canceled, Time, Customer ID and Transaction ID. One variable was added to record the total price of each row which is calculated by the following equation.

$$Total\Pr ice = Unit\Pr ice * (Quantity - QuantityCanceled)$$

Then using groupby operation on Customer ID, these six existing features were transformed to RFM features which are Recency, Frequency and Monetary value.

After the value of R, F and M were obtained, a score was assigned to each of the three measures of each customer. The scores were acquired by grouping each measure in quantiles. For each measure, scores of 1, 2, 3, 4 were created based on the quantile of each value in the measure.To avoid the influence of skewness, standardized scaler from sklearn package in Python was employed to scale R, F and M to center the mean and standard deviations. Finally, the standardized value of R, F and M served as features depicting customer behavioral characteristics. Then a RFM score was assigned to each customer by simply putting the scores for R, F and M in sequence.

By classifying products into different categories and recording purchase history of products in each category, information on customers' products preference could be obtained. Unique product descriptions were extracted from the description column with word stemming and stop-word removal performed on words in the descriptions. Then, Term Frequency-Inverse Document Frequency (TD-IDF) method was applied to extract keywords of each description. After TF-IDF matrix was obtained, k-means clustering algorithm was performed on the matrix to group unique products into different categories. In this step, all the products in the raw data were classified into 85 categories by the k-means algorithm.

For each category, a column was created to record the proportion of purchased products in this category in all the purchased products of this customer. This manipulation served as a way to balance the difference in numbers of purchased products of each customer and thus focus the classification on difference between categories.

At the end of feature engineering stage, the original variables were transformed into customer behavioral features and 85 product categories which record the proportions of products in this category in all the products purchased by this customer. Table 3 shows the final features for customer segmentation.

Table 3 Input features for clustering

Feature Name	Description
Recency	Quantile value of the time interval between the customer's most recent transaction and the most recent date in the dataset
Frequency	Quantile value of the number of transactions of the customer in a year
Monetary Value	Quantile value of the sum of the amount spent in each transaction of the customer
Products Category 1	The proportion of products in this category in all the products purchased by this customer
...	...
Products Category 85	The proportion of products in this category in all the products purchased by this customer

3.4 Clustering

The above features are used as inputs with equal weights for clustering algorithm. K-means clustering algorithms is used in this section. A good clustering is one for which the within cluster variation is as small as possible. "Elbow Method" was used to determine the number of clusters and thus choose the best clustering. To obtain the optimal K, model should be built several times by setting different numbers of centroids and for each of the model within cluster sum of errors should be calculated. Fig.2 displays within cluster sum of errors for each K. Based on the slope of the figure, the decision about the number of clusters is made.

According to Fig. 2,value of 6, 8, 9 could be chosen. In this study, 8 was chosen for the clustering algorithm, since the value of the slope after 8 is relatively small. Thus, all the customers in the dataset are grouped into 8 clusters.

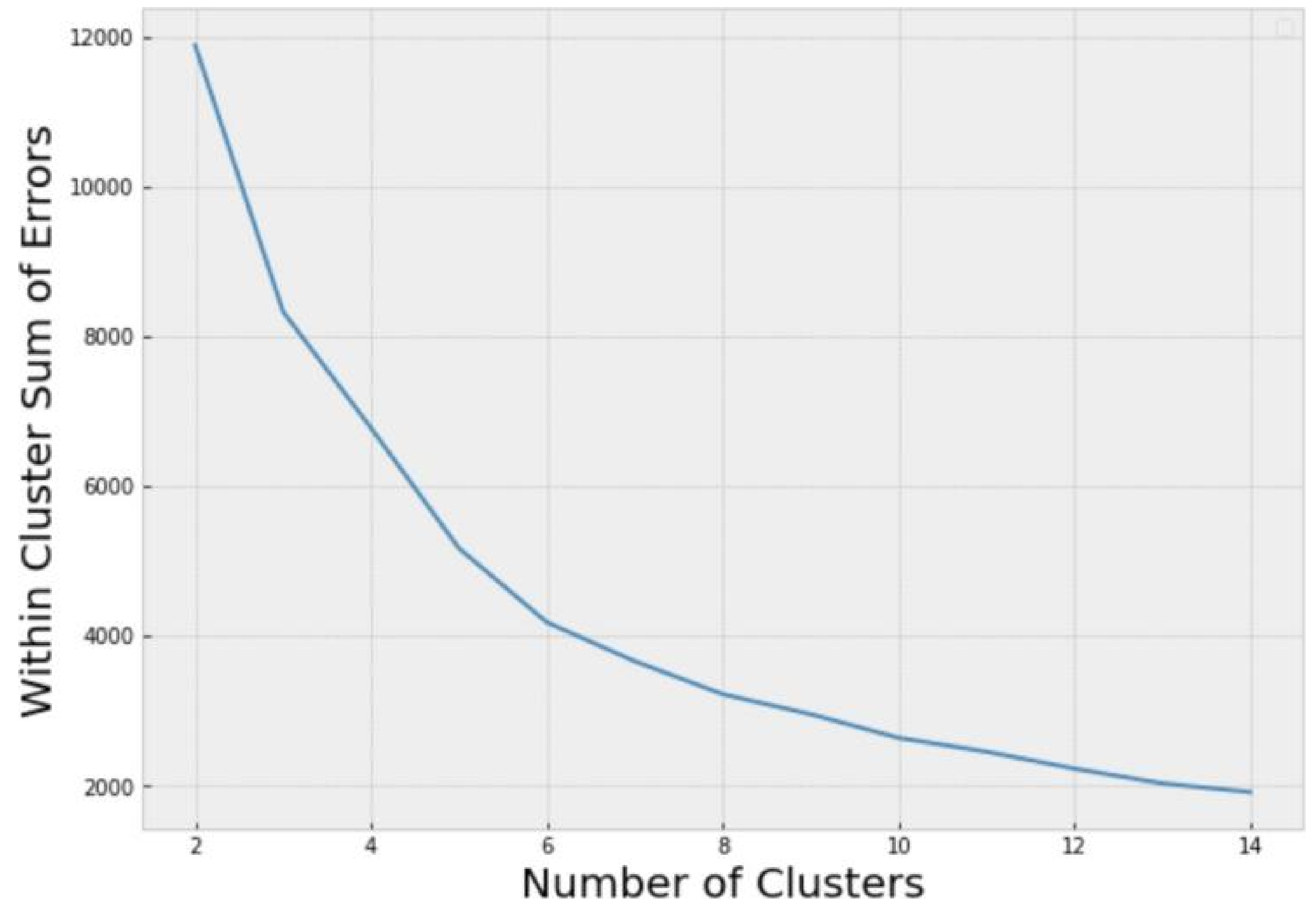

Fig. 2. Within Cluster Sum of Errors for each K

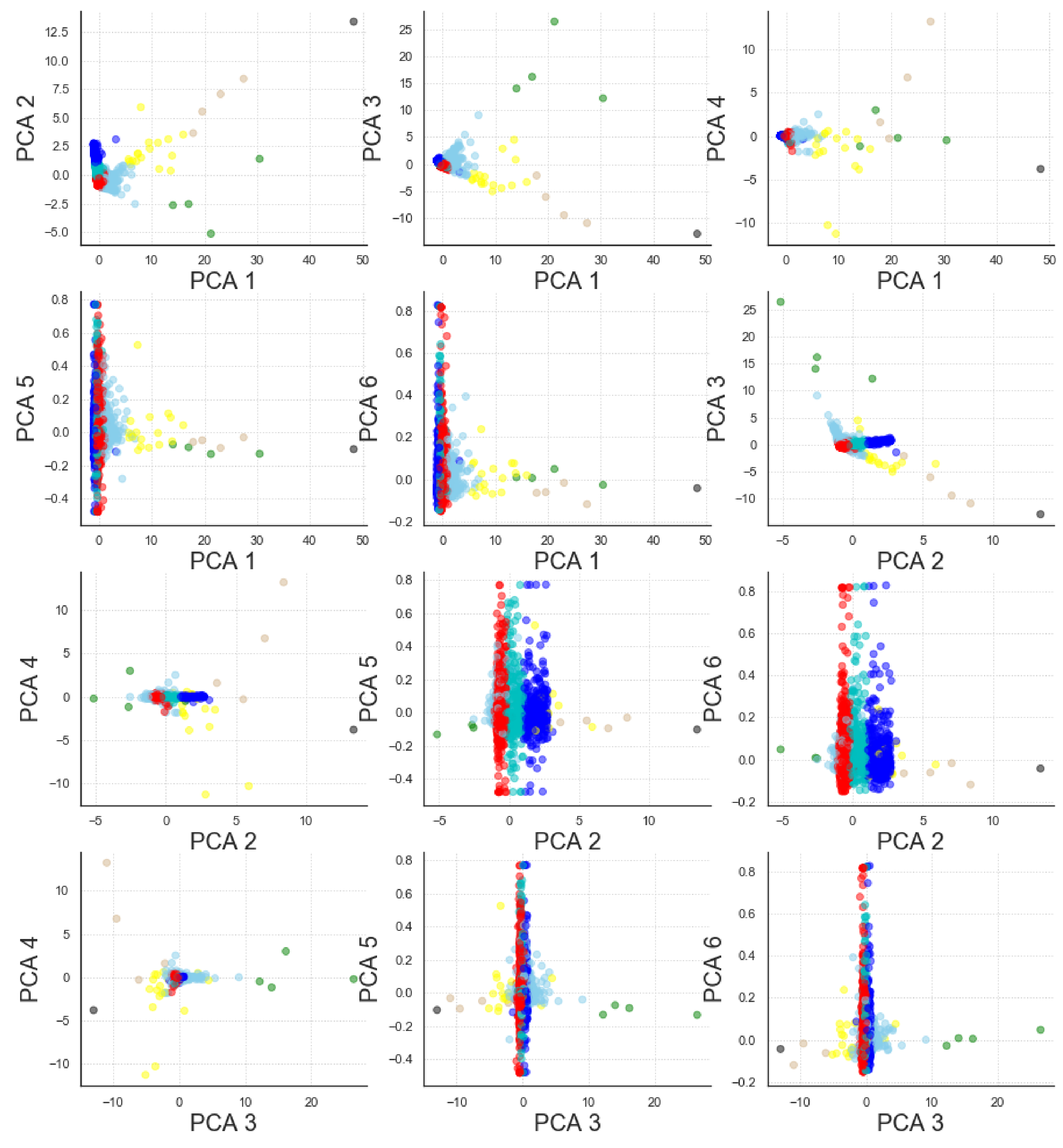

Fig. 3. Visualization of Clusters with PCA

3.5 Product Analysis

For each cluster, the purchased products can be further analyzed to develop product promotion strategies for customers in this cluster. In this section, cluster 3 is used as an example. Item-level product analysis is conducted to find out which products the customers in this cluster have purchased, which products have been purchased together most frequently and in which sequence the products have been purchased.

A technique to uncover the underlying associations between items is association rule. In this study, the association rule mining is split up into two separate steps:

1) Find all frequent itemset: An itemset with support value above the support threshold.
2) Generate strong association rules from the frequent itemsets: The rules should satisfy minimum Confidence and minimum Lift.

Due to the convenience and simplicity of implementing Apriori Algorithm, this algorithm is chosen in this study to mine the association rules. Support threshold was set as a very small value of 0.03. Since there are 2622 unique purchased items in cluster 3, analysis of such large inventories involve a large number of item configurations, which leads to a matrix full of 0s and causes the support of basket occurrences to drop drastically. Thus, the support threshold has to be lowered to detect certain associations. This situation is common in E-commerce business where a large number of item configurations are possible in a single basket among an even larger number of baskets. Furthermore, the minimum confidence is set as 0.6 and the minimum of lift is set as 4.

4 Results and discussions

4.1 Cluster Visualization

To check whether the clusters are truly distinct, visualization of the clusters was created in this study. Since there are 98 dimensions in the data set, it is hard to display the clusters directly without dimensionality reduction. Thus, dimensionality reduction was conducted to visualize the clusters. Two dimension reduction algorithms are applied to created visualizations: PCA and T-sne methods. By applying PCA to the feature space, the feature space is reduced to the first 6 principal components while preserving 97% of data's variation. The visualization of the various clusters is shown in Fig .3.By applying T-SNE to the feature space and reducing the dimensionality to 2, 2D graph depicting the clusters, each with distinct color is plotted and shown in Fig. 4.

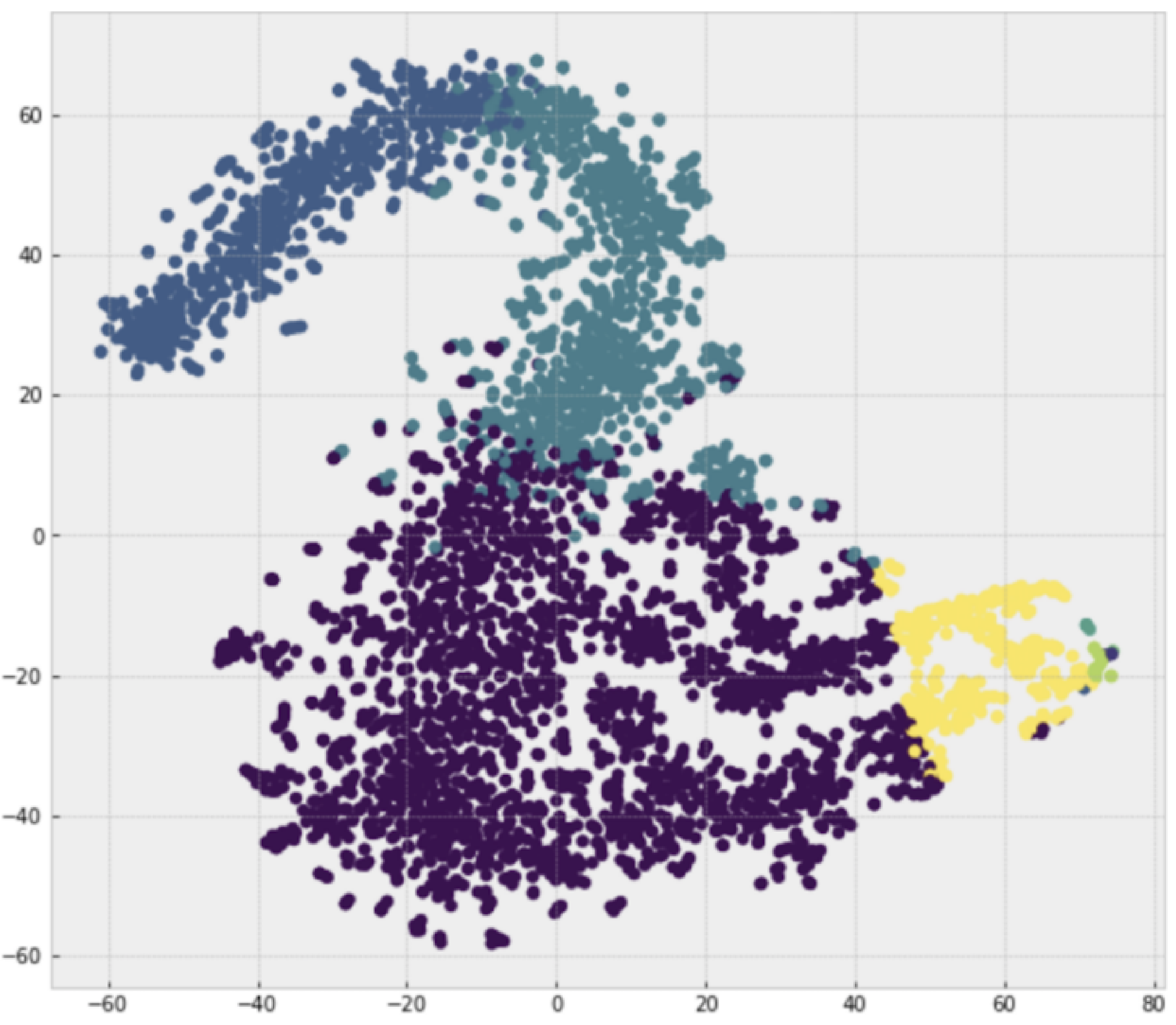

Fig. 4. Visualization of Clusters with T-sne

4.2 Cluster Interpretation

Frequency and Percentage of common RFM scores in Cluster 1,3,7,8

Among the total 8 clusters, some clusters only contain a very small number of customers, thus interpreting those clusters would not generate informative results. The four clusters with more than 300 customers (clusters 1,3,7 and 8) were interpreted in order to better understand the customers. RFM scores of customers in each cluster are analyzed to gain information on customers' behaviors. For each of the four clusters, the data are grouped by customers, then the frequency and percentage of each RFM score in the cluster was calculated. The most common RFM scores and their percentages in each cluster are labeled in Fig. 5. The number of customers and the number of possible RFM scores in each cluster are listed in Table 4.Cluster 1 seems to be quite heterogeneous, while the four most represented categories are 444, 333, 433, 344, the scores of whom are lost cheap customers and not worth investments, the scores of other customers who need aggressive new products promotion such as 311, 411, 322,422 also account for a significant percentage of all scores. Furthermore, scores of normal customers who need price incentives and new product promotion such as 233, 222, 211 account for a small proportion. Cluster 3 seems to consist mostly of best buyers with high recency, frequency and monetary value, who must be taken care of and worth further investments. Cluster 7 contains normal customers who buy recently, but some of them spend little and do not make purchases frequently. These customers need careful promotion strategies since some of them might be potentially highly profitable or unprofitable at all in the long term. Cluster 8 is clearly full of lost cheap customers who are not worth further investments. Understanding the behaviors of customers can help the business better target their strategies at specific customers and generate larger profits.

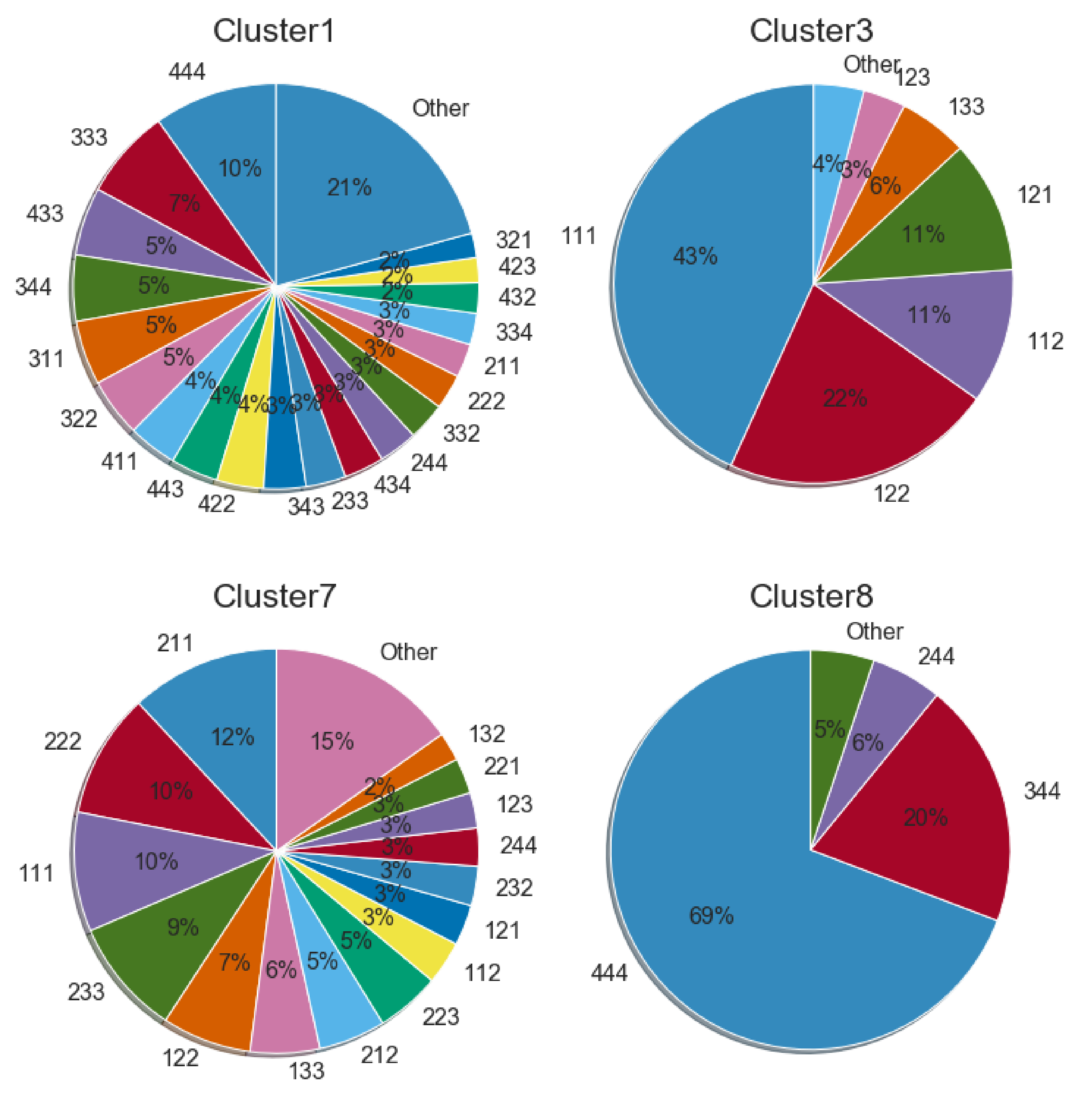

Fig. 5. Percentage of Most Common RFM scores in Cluster 1,3,7,8

Table 4 Frequency and percentage of common RFM scores

Cluster	Number of Customers	Number of Possible RFM scores
Cluster1	2345	48
Cluster3	690	15
Cluster7	956	30
Cluster8	318	10

After counting the frequency of each product, the 10 most popular items are Regency Cake stand 3 Tier, White Hanging Heart T-Light Holder, Party Bunting, Assorted Colour Bird Ornament, Jam Making Set With Jars, Set of 3 Cake Tins Pantry Design, Natural Slate Heart Chalkboard, Pack of 72 Retrospot Cake Cases, Rex Cash+Carry Jumbo Shopper and Jam Making Set Printed. The most frequently purchased items seem to be party supplies including decorations and eating utensils. Customers in cluster 3 tend to hold parties frequently and may have a high demand for party supplies. After obtaining the buying patterns, the following results focuses on the associations between products. Based on the given thresholds, the strong association rules for customers in cluster 3 are generated and listed in table 5.

Table 5 Strong association rules

Antecedent	Consequent	Support	Confidence	Lift
Green Regency Teacup and Saucer	Pink Regency Teacup and Saucer	0.035	0.711	17.624
Pink Regency Teacup and Saucer	Green Regency Teacup and Saucer	0.035	0.865	17.624
Roses Regency Teacup and Saucer	Green Regency Teacup and Saucer	0.039	0.720	14.672
Green Regency Teacup and Saucer	Roses Regency Teacup and Saucer	0.039	0.800	14.672
Pink Regency Teacup and Saucer	Roses Regency Teacup and Saucer	0.031	0.757	13.879
Green Regency Teacup and Saucer	Regency Cake stand 3 Tier	0.033	0.667	5.608

Product promotion strategies can be formulated according to both the buying patterns and the strong associations between products. For example, since customers in cluster 3 are likely to buy more party supplies, the business may reinforce their advertisement before special festivals and include more pro- motions of decorations or eating utensils. Advertisement on the consequent product of the rule could be targeted at customers who bought the antecedent product of the rule. Discounts can be applied to one of the two items in a frequent itemset since a lower price of one item may prompt the customers to buy both of them. Other promotion strategies based on the obtained patterns and rules can also be formulated.

5 Conclusion

An example of customer segmentation has been provided in this article to elaborate how differentiated strategies targeted at diverse online customers can be formulated by unsupervised machine learning. The identified customer groups in this study can help the business better understand its customers' behaviors or product preferences and adopt corresponding marketing strategies.

In this study, two steps are very important and the most time-consuming: feature engineering and cluster interpretation. Different ways of creating the features can greatly influence the experiment results and different interpretations of the clusters can lead to different promotion strategies. In the feature engineering part, RFM model is used to quantify the purchasing behaviors. Further research can incorporate other customers' characteristics such as demographic or product preference into the feature engineering part. The cluster interpretations in this study focus on the RFM scores which display the purchasing behaviors of customers, and they are based on the author's personal understanding of the model. Further research may focus its interpretation on customers' products preference or other possible characteristics of the customers.

REFERENCES

[1] Statista. (2020b, October 26). E-commerce worldwide - statistics & facts. **https://www.statista.com/topics/871/online-shopping/ (https://www.statista.com/topics/871/online-shopping/)** Navigate to

[2] Hughes, Arthur M. ”Strategic database marketing: the masterplan for starting and managing a profitable.” Customer-based Marketing Pro- gram, Irwin Professional (1994). Navigate to

[3] Qiasi, Razieh, *et al.* ”Developing a model for measuring customer's loyalty and value with RFM technique and clustering algorithms.” The Journal of Mathematics and Computer Science 4.2 (2012): 172-181. Navigate to

[4] Shi, Congying, Chaojun Xu, and Xiaojiang Yang. ”Study of TFIDF algorithm." Journal of Computer Applications 29.6 (2009): 167-170. Navigate to

[5] Mythili, M. S., and AR Mohamed Shanavas. ”Performance evaluation of apriori and fp-growth algorithms.” International Journal of Computer Applications 79.10 (2013). Navigate to

[6] Razmochaeva, Natalya V., Dmitry M. Klionskiy, and Vladimir V. Cher- nokulsky. ”The Investigation of Machine Learning Methods in the Prob- lem of Automation of the Sales Management Business-process.” 2018 IEEE International Conference” Quality Management, Transport and Information Security, Information Technologies”(IT&QM&IS). IEEE, 2018. Navigate to

CONFCDS 2021, January 28–30, 2021, Stanford, USA

ACM ISBN 978-1-4503-8957-0/21/01…$15.00.
DOI: **https://doi.org/10.1145/3448734.3450775 (https://doi.org/10.1145/3448734.3450775)**