Customer Segmentation for Ecommerce

Harshil Bharatkumar Darji
Computer Engineering
San Jose State University
San Jose, United States
harshilbharatkumar.darji@sjsu.edu

Jainam Hareshbhai Patel
Software Engineering
San Jose State University
San Jose, United States
jainamhareshbhai.patel@sjsu.edu

Abstract—**Analysing customer behaviour over time helps to predict what kind of products a customer will purchase. Companies use this idea of customer segmentation extensively to boost sales. The goal of this paper is to analyse customer behaviour using various techniques and derive inferences about customers with a motive to segment customers into best possible groups. Such a classification is performed using various available techniques and algorithms and the quality of these predictions is tested and the precision of each algorithm is compared.**

Index Terms—**segmentation, classification, prediction**

I. Introduction

Ecommerce has found its way in all types of industries, helping countless businesses grow with the help of latest technologies. Ecommerce marketplaces have been on the rise around the world since the mid-1990s with the launch of market giants like Amazon, Alibaba and others. The success of these companies is hugely reliant on their customers. For the same reason, these Ecommerce giants tend to provide a broad spectrum of selection of products to their customers, causing an overload of information for the customers, confusing them on what to buy. So, the idea is to show the customers a personalized view of products, which will boost the sales of the company. Analysing a customer's behaviour over a period of time will hint into anticipating what kind of product the customer will tend to purchase.

II. Literature Review

A. Customer Segmentation

The key to boosting sales for a company is knowing exactly what the customer wants. The interactions of the customer with the Ecommerce platform helps in building up the generic characteristics of the customer. For long term associations with the customer, such interactions play an important role in identifying healthy Customer Relationship Management(CRM) systems. However, trying to personalise each customer individually is very difficult, making it necessary to divide each customer into groups having similar characteristics. This is called Customer Segmentation. Baer states that customer segmentation is the activity to categorise or classify an item or subject to a group that has been identified to have in common [1]. The goal of such a classification is to increase profitability, reduce operational cost and enhance customer service. The following section will discuss existing research on Customer Segmentation techniques.

B. Obtaining Data

Customer Segmentation data can be obtained from various sources and analysed differently. Magento categorises data into internal and external data. The former includes data about customer profile, customer registration and purchase history and the latter includes census data, media browsing, surverys and market research data, web and social analysis [2]. Colica uses customer database and purchase history on customer segmentation methods [3].

C. Methods for Customer Segmentation

There are numerous methods for Customer Segmentation. Schneider divides customer segmentation methods into geographic, demographic, psychographic and usage-based market segmentation. Demographic segmentation is based on age, gender, family size, income, education, religion or ethnicity. Psychographic segmentation is based on social class, personality or approach to living [4]. Magento divides customer segmentaion into several variables described by:

- *Profit Potential:* Using variable transaction frequency, date of last purchase, average order value, customer lifetime value.
- *Past Purchases:* using the variable of product type/attribute, product price, payment/shipping method used, product benefit sought (price, quality, prestige), product satisfaction.
- *Demographic:* using the variable of geographic location (city state, country, region), age, gender, household size, income, occupation, education, ethnicity, browsing device (laptop, PC, tablet, smartphone) and type (vendor and model), traffic source (organic search, banner link, referral site).
- *Psychographic:* using the variable of hobbies and interest, leisure and recreational activity, affiliations (religious, professional, cultural, political, institutional), personal traits (social vs. private; modern vs. traditional; spontaneous vs. cautious).
- *Behaviour:* using the variable of pages viewed, responses to offers and promotions, participation in reward programs, channel management.

Migros employs variety of methods to effectively segmented its customer base. These include value, behavorial, lifestyle, lifecycle and activity-based segmentation schemes [5]. Usually, a conjunction of multiple techniques yields a better result for segmentation. Baer segments customer using business rules method, quantile membership method, supervised clustering with decision tree method and unsupervised clustering method using k-means algorithm. Other segmentation methods used by Lieberman include combination Business Rule, Customer Profiling to analyse monthly customer spending and tally monthly customer visits [7]. Based on above techniques, Customer Segmentation can be broadly classified into categories shown in Fig. 1. The following sections describe the implementation steps performed for Customer Segmentation.

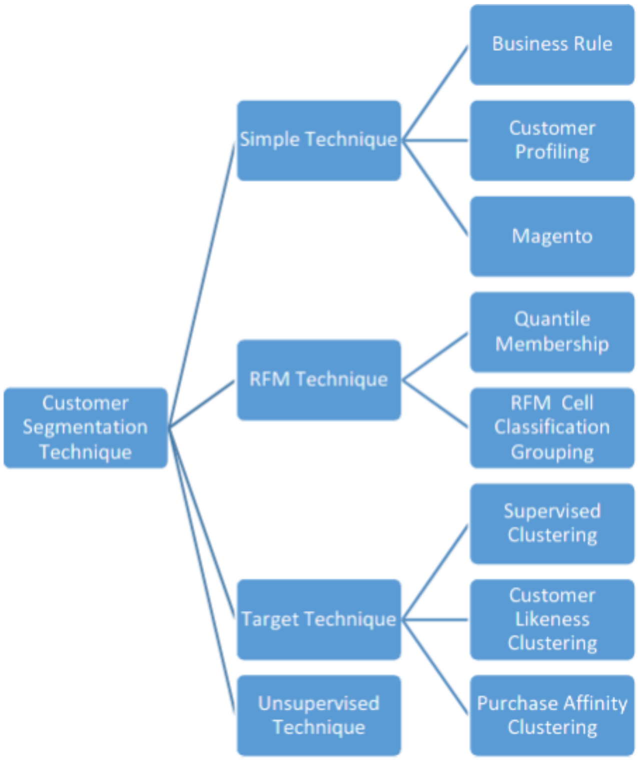


Fig. 1: Customer Segmentation Techniques

III. Data Preparation

For Customer Segmentation, we are using a transnational data set which contains all transactions for a UK-based online retail store for a period of over one year [8]. The data set contains 541909 rows and 8 columns as shown in Table I. The data is imported as a pandas dataframe and later processing for customer segmentation is performed. The following sections will discuss the steps involved.

IV. Data Preprocessing

As the first step, the null values are counted and dealt with. While looking at the dataframe, it was observed that approximately 25% of the entries are not assigned to a particular customer. With available data, it is not possible to impute these values for customer and hence turn out to be meaningless for further processing. Hence, these entries are removed from the dataset. Later, duplicate entries are found and removed. Total number of rows now remaining in the dataset are 401604. The next important step involves gaining essential insights into the dataset.

V. Exploratory Data Analysis

This section discusses some key inferences dervied from different variables of the dataset. The variables go as follows:

1) Country: It is found that the orders were made from 37 different countries. The maximum number of orders were made from the United Kingdom. The dataframe consists of nearly 400,000 entries.

2) Customer, Product and Invoices: On processing, it turned out that a total 4372 customers ordered 3684 unique products with a sum total of approximately 22,000 invoices generated. On sorting the number of products per transaction it is found that it varies upto a count of 542. There exist customers who visited once and just brought one product and there also exist customers who were frequent and brought large number of items per transaction.

3) Cancelled Orders: It is important to find out how many orders were cancelled. On counting orders with a prefix of 'C' in the quantity it is found that a total of 3654 orders were cancelled, which amounts to 16.47% of the total orders, which is quite large. On observing these entries, it is found that the cancelled order entries are mostly identical to the previous transaction except for Quantity and InvoiceDate variables. For the cancelled orders it is important to check for two important hypotheses:

- *Hypothesis-1 Check if all such cancelled entries are identical to their counterparts:*
  To check this, for all entries it is checked whether there is a a negative quantity and if there is corresponding same quantity (i.e. positive), with the same description as that of the former one. On processing, it is found that the hypothesis fails. This happened because there exists Discount entries in the dataset which have negative quantities.
- *Hypothesis-2 Check cancelled entries discarding Discount entries:*
  To check this, for all entries it is checked whether there is a a negative quantity and if there is corresponding same quantity (i.e. positive), with the same description (excluding Discount) as that of the former one. On processing, it is found that the hypothesis again fails.

From above analysis, it can be verified that the cancellations do not necessarily correspond to orders that would have been made beforehand. This could be due to the reason that these orders were made before the dataset was created. Hence, it is important to have some variable indicating a cancelled order. A variable is introduced called *QuantityCanceled* which is indicated when a cancelled order is found to exist along with a counterpart. Cancelled orders found without its counterpart are also found. This sums up to approximately 2.2% entries of the dataset. All such entries are removed from the dataset.

4) StockCode: The contents of the StockCode variable are searched where it contains only letters. It is inferred that there

TABLE I: Variables in dataset

Variable Name	Data type	Description
InvoiceNo	Nominal	Invoice number; Six-digit number to uniquely assigned to each transaction. Prefix 'C' stands for cancellation
StockCode	Nominal	Product (item) code; Five-digit number uniquely assigned to each product
Description	Nominal	Product Name and short Description
Quantity	Numeric	Quantities of each product (item) per transaction
UnitPrice	Numeric	Product price per unit in sterling pounds per unit
InvoiceDate	Numeric	The date and time when each transaction was generated
CustomerID	Nominal	Five-digit integral number assigned to each customer
Country	Nominal	Country name of customer residence

are several peculiar types of transactions like bank charges, post charges etc.

5) Total Price: A new variable is introduced called *TotalPrice* which represents total price associated with each entry. It is observed from the analysis that approximately 65% of the orders are above 200 pounds. To get a global view of how purchases are divided, the following visualisation is created:

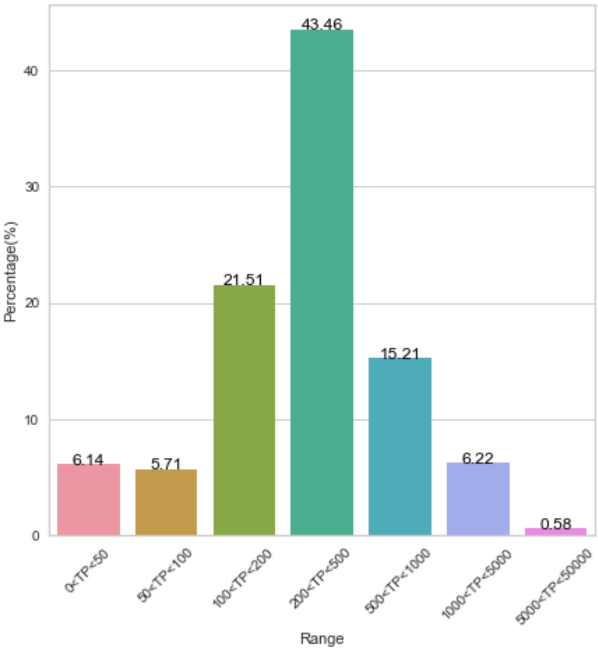


Fig. 2: Percentage of total price lying between range

VI. Product Categories Insights

The products are uniquely identified through *StockCode* variable. A short description of products is given by the *Description* variable. The idea is to use to contents of the description variable to group products into different categories. To perform this, we make use of *Natural Language Toolkit (NLTK)*, which is a suite of libraries and programs for symbolic and statistical natural language processing for English written in Python. The NLTK Snowball stemmer library is used, which removes morphological affixes from words and leave only the root words.

The logic of this function involves extracting names appearing in the products description. For each of these extracted names, the root word if found using NLTK library and a set of names associated with this root are aggregated. When several words are listed for the same root, the keyword associated with this root is the shortest name. This systematically selects the singular variant when there are singular/plural variants. This analysis yields the number of keywords in Description attribute to be equal to 1483. Figure 3 displays the top 50 keywords with their frequency extracted using above analysis.

On further examination, it is found that 94 keywords contain special characters. Keywords including colors like pink, blue, tag, green, orange don't carry any information. Also, few keywords with length less than two are not relevant and have a total count of 803. These keywords are need to be filtered and removed. On removing these keywords, a final count of keywords leads to a total number of 193 keywords.

A. Data Encoding - One hot encoding

The idea of data encoding is to use the previously found keywords to create groups of products. The matrix of such one-hot encoded vectors can be briefly shown as follows:

$$\begin{array}{cl}
\begin{array}{ccccc} w1 & w2 & . & . & wN \end{array} & \\
\begin{pmatrix}
a_{11} & a_{12} & . & . & a_{1N} \\
a_{21} & a_{22} & . & . & a_{2N} \\
. & . & . & . & . \\
. & . & . & a_{ij} & . \\
. & . & . & . & . \\
a_{M1} & a_{M2} & . & . & a_{MN}
\end{pmatrix} &
\begin{array}{c} Product1 \\ Product2 \\ . \\ Producti \\ . \\ ProductM \end{array}
\end{array}$$

In the matrix, the columns represent the description keywords for each product. The coefficient a_{ij} is 1 if the description of product i contains the keyword j, and 0 otherwise. To get a better division and balanced groups of products, products are divided into price ranges. To choose the appropriate price ranges, the number of products in each group is checked. This is shown in Table II.

TABLE II: Price ranges and products

Unit Price range	Number of products
$0 < UPR < 1$	964
$1 < UPR < 2$	1009
$2 < UPR < 3$	673
$3 < UPR < 5$	606
$5 < UPR < 10$	470
$UPR > 10$	156

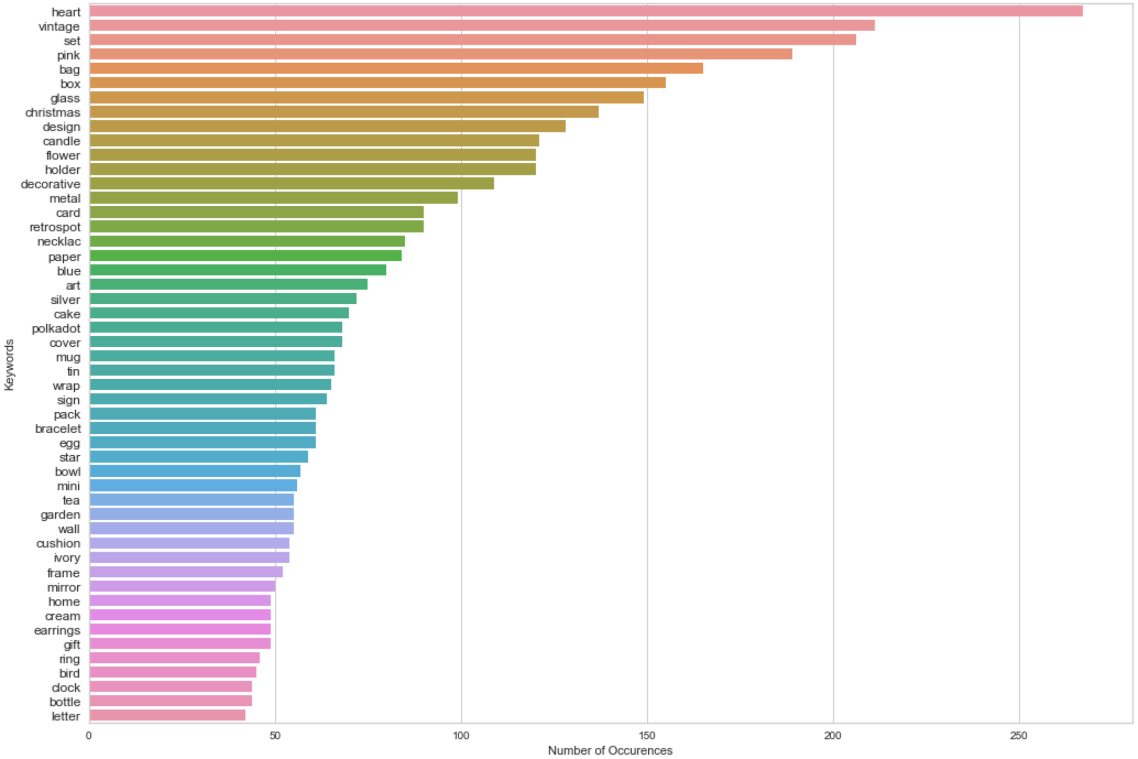


Fig. 3: Top 50 keyword occurrences

B. KModes Clustering

After performing above analysis, the products are grouped into different classes. The best metric for binary matrices is the Hamming metric which is used in conjunction with KModes clustering. In order to find out the optimal number of clusters, Silhouette scoring is used along with the KModes clustering method. The Silhouette score is a metric used to calculate the goodness of a clustering technique and is considered the best metric among others as it deals with higher dimensions of data pretty well. After trying KModes clustering with different number of clusters, the Silhouette scores for different number of clusters are shown in the following figure:

```
The Average silhouette score for 2 is: 0.1247260125946426
The Average silhouette score for 3 is: 0.1675404012581198
The Average silhouette score for 4 is: 0.17934477919317687
The Average silhouette score for 5 is: 0.2570111837759065
The Average silhouette score for 6 is: 0.14333178896562104
The Average silhouette score for 7 is: 0.21026576564401667
The Average silhouette score for 8 is: 0.16935566918500014
```

Fig. 4: Silhouette scores for products for varied no. of clusters

It is found that the highest score is for 5 clusters. To compare different metrics for scoring, the Silhouette score is compared with Calinski Harabasz score. The optimal cluster number visualization turned out as shown in Figure 6.

Performing KModes clustering with number of clusters as 5 leads to even frequency in all clusters. This can be visualized in the following figure:

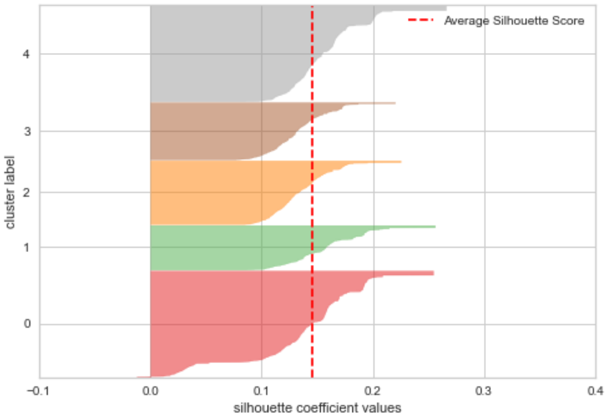


Fig. 5: Silhouette plot for Products with 5 clusters

C. Word clouds

To get more idea about the clusters, we can have a look at the type of objects that each cluster represents. To do this, we find which keywords are most frequent in each item. The result is output in the form of word clouds. The word cloud visualisation is shown in Figure 7.

From these different cluster of words, it can be inferred that cluster n_0 could be associated with gifts, as it includes words like christmas, art, card etc. Hence, different products were split into five clusters. The next step is to categorise customers based on the product categories derived.

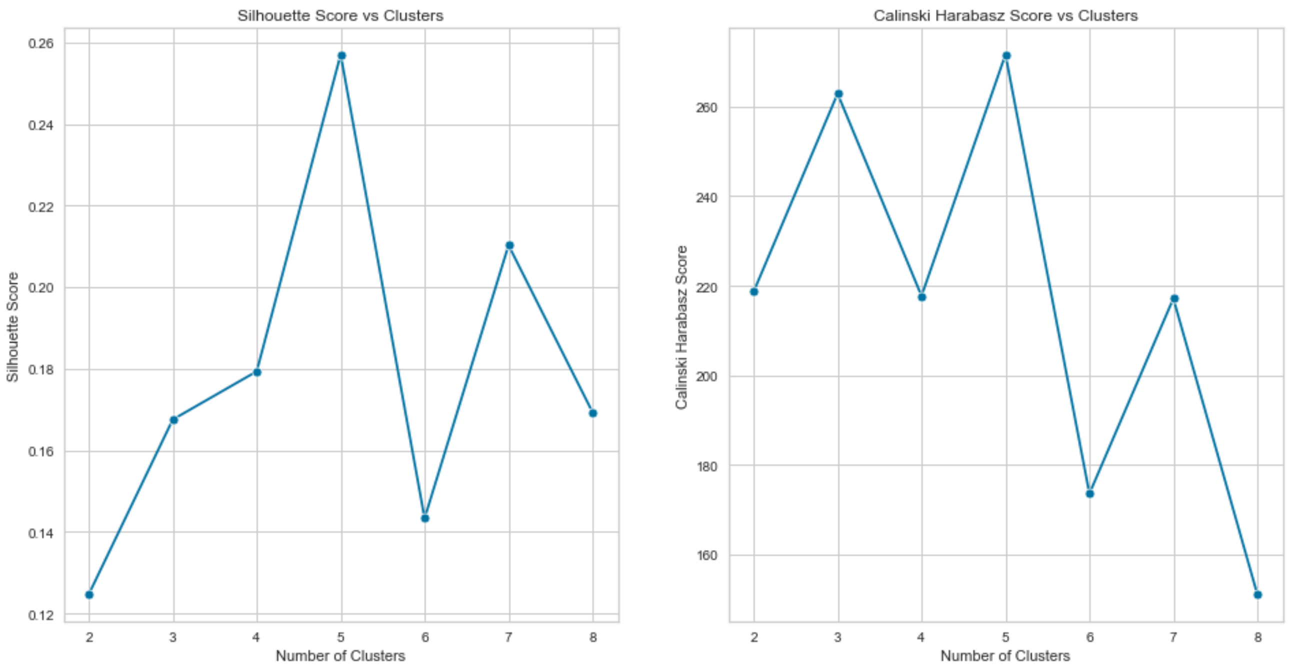


Fig. 6: Product Silhouette scores

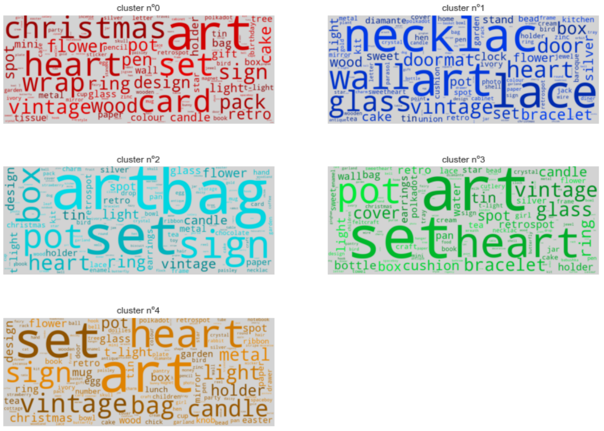
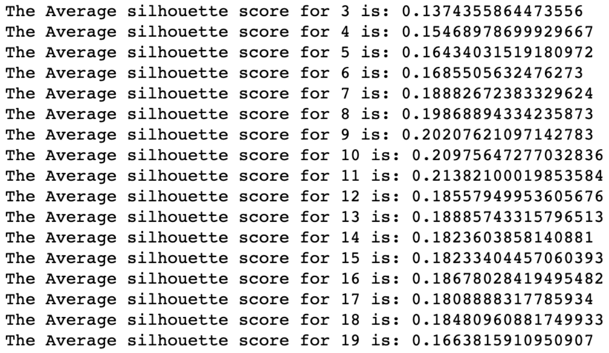


Fig. 7: Word Cloud Visualisation

VII. Customer Categories

For splitting the data, since the dataset was consisted 12 months of data, 10 months of data was used for training and 2 months of data was used as test data. For customer categories, we merged multiple entries of one order into one row. The data was joined using the CustomerID and InvoiceNo attributes. For these customer order combinations, the number of purchases, minimum amount, maximum amount, mean amount and the sum of amounts was calculated for each customer. Two attributes were added called FirstPurchase and LastPurchase which indicated the number of days elapsed since the first purchase and last purchase respectively. From this analysis, it turned out that 40% of customers (1445) were only one-time customers. One of the objectives of any company could be to target these users and try to retain them.

For these customers, KMeans clustering method was used. KMeans was preferred over KMode method only because matrix for customers is not binary matrix like it was for the products. Here too, just like product categories, Silhouette metric was used to find out the optimum number of clusters. The objective then is to make a classification possible for the new customers coming at the first visit.



```
The Average silhouette score for 3 is: 0.1374355864473556
The Average silhouette score for 4 is: 0.15468978699929667
The Average silhouette score for 5 is: 0.16434031519180972
The Average silhouette score for 6 is: 0.1685505632476273
The Average silhouette score for 7 is: 0.18882672383329624
The Average silhouette score for 8 is: 0.19868894334235873
The Average silhouette score for 9 is: 0.20207621097142783
The Average silhouette score for 10 is: 0.20975647277032836
The Average silhouette score for 11 is: 0.21382100019853584
The Average silhouette score for 12 is: 0.18557949953605676
The Average silhouette score for 13 is: 0.18885743315796513
The Average silhouette score for 14 is: 0.1823603858140881
The Average silhouette score for 15 is: 0.18233404457060393
The Average silhouette score for 16 is: 0.18678028419495482
The Average silhouette score for 17 is: 0.1808888317785934
The Average silhouette score for 18 is: 0.18480960881749933
The Average silhouette score for 19 is: 0.1663815910950907
```

Fig. 8: Silhouette scores for customers for varied no. of clusters

The maximum Silhouette score for 11 customer clusters turned out to be maximum. So, we split the customers into 11 groups as it is most optimal Silhouette score. The Silhouette scores plot for different clusters can be visualised in Figure 9.

To determine how even the distribution of data is among these 11 clusters, the Silhouette plot can be visualised. The plot turns out to be as shown in Figure 10. It can be inferred from the plot that the 11 clusters have an even distribution

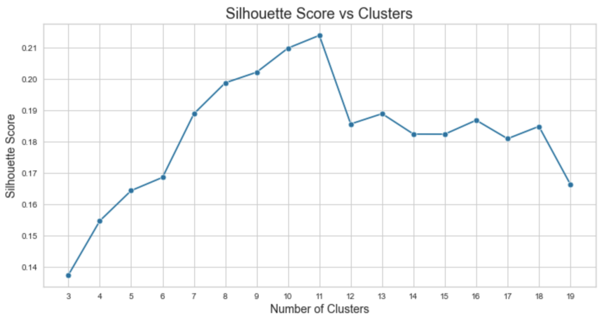


Fig. 9: Customer Silhouette scores

and also the peaks of each cluster are greater than the average Silhouette score value.

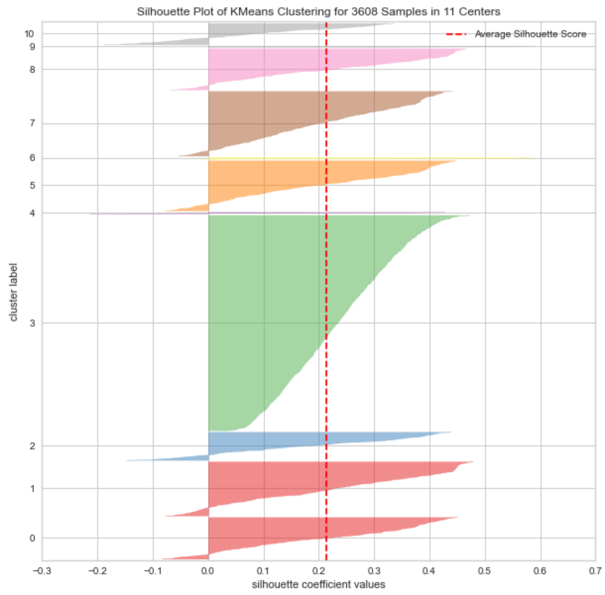


Fig. 10: Silhouette plot for Customers with 11 clusters

To get a global view of each of these formed clusters, radar charts can be used. Radar charts were plotted for each of these clusters to try and understand how strong preponderance is represented towards the 5 categories of products. The radar charts can be seen in Figure 10. From the radar charts it can be observed that the 5 clusters $n_0, n_1, n_2, n_3, n_4, n_5$ have strong inclination towards each of the 5 categories. For clusters n_6, n_7, it can be observed that either the customer visits the Ecommerce site a lot or he made a lot of purchases with low amounts.

VIII. Customer Classification

Customer classification included two essential steps:

A. Hyperparameter Tuning

Hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. For our case, we have used GridSearchCV instead of RandomizedSearchCV as it looks through all possible combinations. The KFold value taken for cross-validation is equal to five. For each hyper parameter, we plot the learning curves to decide whether underfitting or overfitting occurs or not. The plots can be seen in Figure 12. For various curves, it was either underfitting or overfitting, however for SVM and Logistic Regression, the plots generalised well. The different parameters that gave optimal values along with their precision counts can be found in Table III.

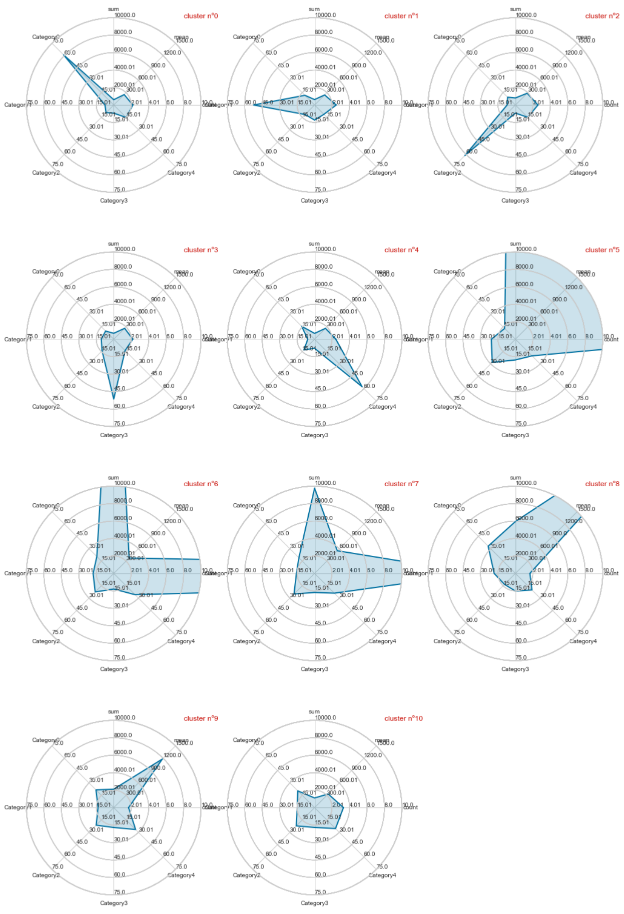


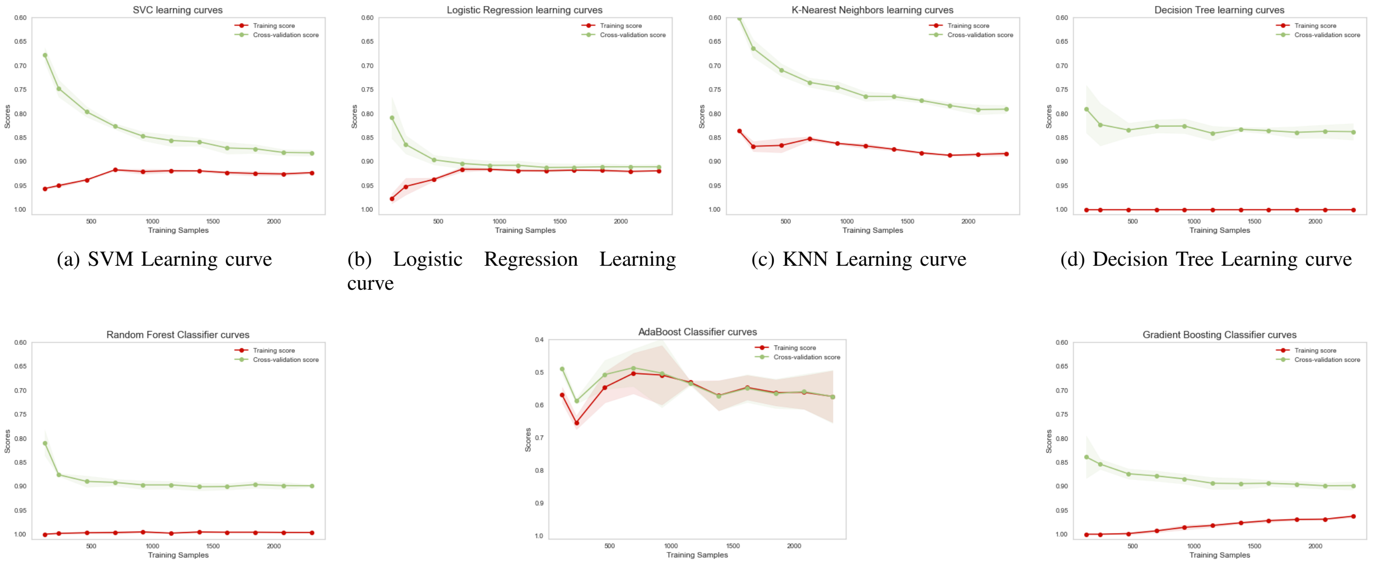
Fig. 11: Radar chart for Customers

B. Ensemble Voting Classifier

The next step consisted of building the ensemble voting classifier. This classifier was built using the best combination of various classifiers from the previous section. Random Forest, Gradient Boosting, KNN and Logistic Regression were used to build the ensemble. Soft voting was used for the ensemble since it allows classification based on probabilities and weights associated with each classifer and hence building a better classifier. The precision of the ensemble turned out to be 91.97%.

C. Prediction on Test Data

Since the data split for testing only included data for 2 months, it is really important to make this data equivalent with



(a) SVM Learning curve

(b) Logistic Regression Learning curve

(c) KNN Learning curve

(d) Decision Tree Learning curve

(e) Random Forest Learning curve

(f) Adaboost Learning curve

(g) Gradient Boost Learning curve

Fig. 12: Customer classification on train data

TABLE III: Learning curve insights

Classifier	Best parameters	Precision	Insights
SVM RBF	c=10, gamma=0.0001, kernel=rbf	90.03%	Generalises well
Logistic Regression	c=0.01623, max_iter=10000	91.14%	Generalises well
KNN	n_neighbours=5, p=1	80.75%	Underfitting
Decision Trees	criterion=entropy, max_features=log2	85.73%	Underfitting
Random Forest	criterion=entropy, max_features=sqrt, n_estimators=100	91.00%	Underfitting
Adaboost	n_estimators = 20	59.97%	Overfitting
Gradient Boosting	n_estimators = 50	90.44%	Underfitting

that of the test data. This is done by updating the sum and count values to account for the time difference between train and test sets. The final matrix will be similar to the matrix we built for the training set. The KMeans model built previously will classify the test data customers into 11 clusters which will act as benchmark. These cluster classification values will be used as labelled data for the main classifier. Finally, we compare the Voting classifier with the benchmark KMeans Model. The final precision value turned out to be 76.64% for the test data.

IX. Conclusion

The products were split into 5 different categories. The classification of customers was done by analysing their consumption habits over a period of 10 months. The customers were split into 11 different segments. A classifier was trained to be able to classify a new customer into one of these eleven categories based on their first purchase. The classifier produced the correct output 76% of the time. Since, the test dataset was only available for 2 months, it would have been beneficial and a better classifier could have been trained if the available data covered a longer period of time.

References

[1] Baer, D., 2012. Customer Segmentation Intelligence for Increasing Profits. [online] Support.sas.com. Available at: https://support.sas.com/resources/papers/proceedings12/103-2012.pdf.

[2] magento.com. 2019. [online] Available at: https://magento.com/sites/default/files8/2019-01/introduction-to-customer-segmentation-v2.pdf.

[3] Collica, R., 2017. Customer Segmentation and Clustering Using SAS Enterprise Miner, Second Edition, 2nd ed.

[4] Schneider, G., 2010. Electronic Commerce. 9th ed.

[5] Cooil, B., Aksoy, L. and Keiningham, T., 2006. (PDF) Approaches to Customer Segmentation. [online] ResearchGate. Available at: https://www.researchgate.net/publication/230557972_Approaches_to_Customer_Segmentation.

[6] Sari, J., Nurgroho, L., Ferdiana, R. and Santosa, P., 2011. (PDF) Review on Customer Segmentation Technique on Ecommerce. [online] ResearchGate. Available at: https://www.researchgate.net/publication/313737530_Review_on_Customer_Segmentation_Technique_on_Ecommerce.

[7] Lieberman, M., 2009. Target 'golden egg' consumer to achieve maximum ROI. [online] mvsolution.com. Available at: https://www.mvsolution.com/wp-content/uploads/Enhanced-Customer-Segmentation-Targeting-the-Golden-Egg.pdf.

[8] Chen, D., Sain, S., Guo, K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. J Database Mark Cust Strategy Manag 19, 197–208 (2012). https://doi.org/10.1057/dbm.2012.17