REPORT

ON

“CUSTOMER SEGMENTATION ON E-COMMERCE DATA”

Submitted by

Rajashri Chitti

ABSTRACT

Customer segmentation is the process of separating a company's customers into groups based on their shared characteristics. Customer segmentation has the ability to help marketers reach out to each customer in the most efficient way possible. A customer segmentation study uses the huge amount of data available on customers (and future customers) to identify distinct groups of consumers with a high degree of accuracy based on demographic, behavioral, and other characteristics. The purpose of customer segmentation is to determine how to relate to customers in each category in order to optimize each customer's value to the company. This paper aims to at analyzing the content of an E-commerce database that lists purchases made by ∼∼4000 customers over a period of one year (from 2010/12/01 to 2011/12/09). Based on this analysis, we develop a model that allows to anticipate the purchases that will be made by a new customer, during the following year and this, from its first purchase.

1. Introduction

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

When creating a focused marketing strategy, it's also important to separate clients into these groups based on how they want to communicate. Businesses must demonstrate that they understand their customers by providing only relevant, tailored information. Customers want to be treated as individuals and cherished, yet this level of customer understanding is impossible to accomplish for all but the smallest of organizations.

Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioural and other indicators. Since the marketer's goal is usually to maximize the value (revenue and/or profit) from each customer, it is critical to know in advance how any particular marketing action will influence the customer. Sometimes referred to as market segmentation, customer segmentation is a method of analysing a client base and grouping customers into categories or segments which share particular attributes. The selected criteria are used to create customer segments with similar values, needs and wants.

2. Objectives:

- ✓ To analyse the content of Ecommerce dataset that lists purchases made by ~4000 customers over a period of one year (from 2010/12/01 to 2011/12/09).
- ✓ Based on this analysis, we develop a model that allows to anticipate the purchases that will be made by a new customer, during the following year and this, from its first purchase.

3. Dataset Details:

This dataset is based on E-commerce that lists purchases made by ~ 4000 customers over a period of one year (from 2010/12/01 to 2011/12/09).

Dataset which we have used is available on Kaggle.

Dataset consists of 541909 rows and 8 columns InvoiceNo., Stockcode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID and Country.

Data frame Dimensions: (406829,8)

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.

StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Description: Product (item) name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.

4. Implementation

Implementation steps are as given below:

- Data Preparation
  In this step, first we load the dataset. Then we check for missing values if any and we are finding duplicate values if any. Then we removed duplicate values.

- Data Exploration

- In data exploration, we plotted number of orders per country and found that the UK is largely dominating amongst the order values as shown in the Fig. 4.1.

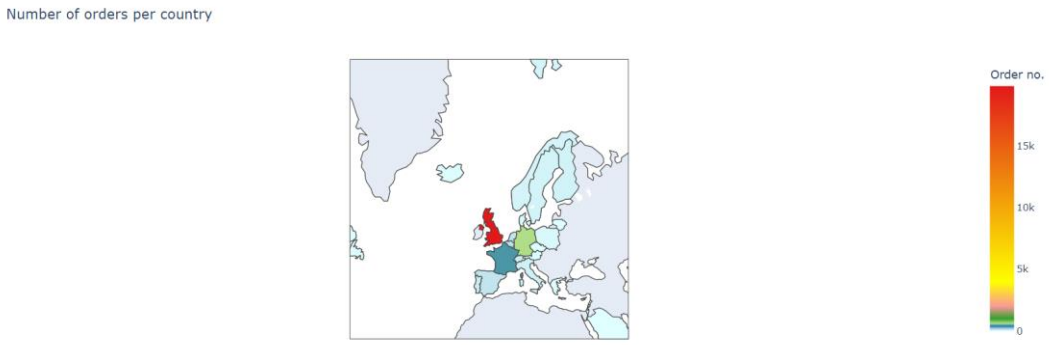

Figure 4.1: No of orders per country

- Now, we have 3684 products, 22190 transactions and 4372 customers in our dataset.
- We note that the number of cancellations is quite large (∼∼16% of the total number of transactions).
- We see that there are several types of peculiar transactions, connected e.g. to port charges or bank charges.
- From the following Fig. 4.2, It can be seen that the vast majority of orders concern relatively large purchases given that ∼∼65% of purchases give prizes in excess of £ 200.

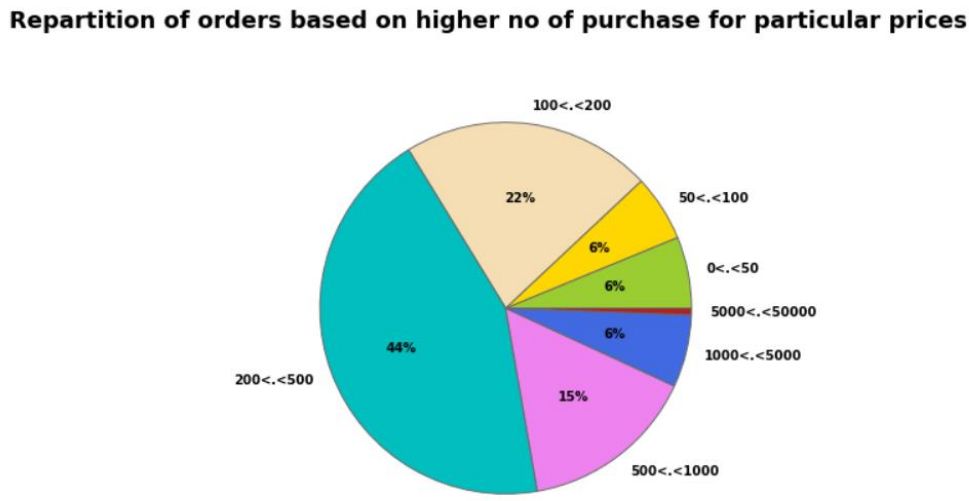

Figure 4.2: Pie chart for distribution of orders.

➢ Clustering based on products using Kmeans

- In order to define (approximately) the number of clusters that best represents the data, we use the silhouette score.
- We found that with best average silhouette score is number of clusters is equal to 5. So, we divided into 5 clusters based on products as shown in Fig. 4.3.

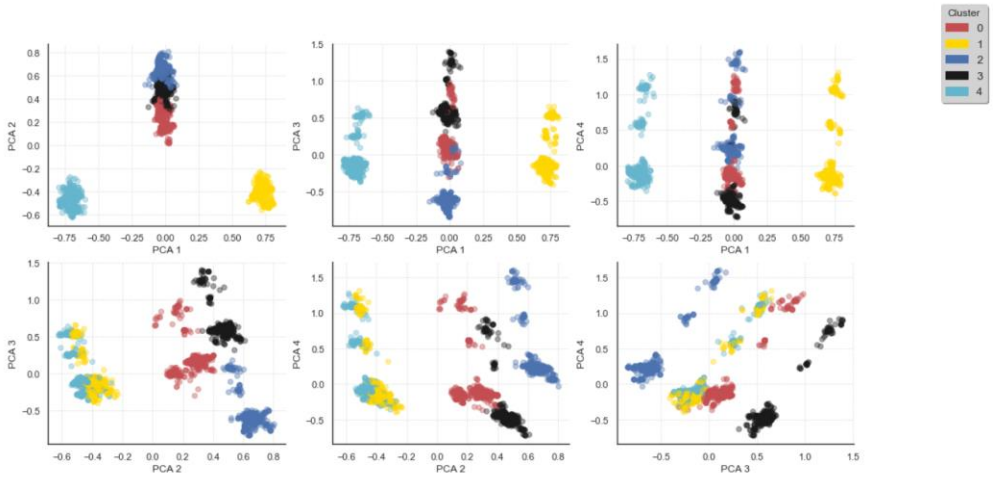

Figure 4.3: PCA Analysis

- Categorizing the customer's clusters
  - We have created the categorical variable categ_product where I indicate the cluster of each product. Next, we create the categ_N variables (with N∈[0:4]) that contains the amount spent in each product category. Collected all the information of a particular order, and put into a single entry.
  - We then create a new data frame that contains, for each order, the amount of the basket, as well as the way it is distributed over the 5 categories of products.
  - We have separated data over time.
  - The data frame basket_price contains information for a period of 12 months.
  - In order to be able to test the model in a realistic way, we split the data set by retaining the first 10 months to develop the model and the following two months to test it.
  - In a next step, we group together the different entries that correspond to the same user. we thus determine the number of purchases made by the user, as well as the minimum, maximum, average amounts and the total amount spent during all the visits.
  - We again use the Kmeans method for clustering the customers. For defining the number of clusters we used silhouette score and found that the best case is with 11 clusters.
  - Number of customers/clients in each clusters is observed.
  - Finally, we re-organize the content of the data frame by ordering the different clusters: first, in relation to the amount spent in each product category and then, according to the total amount spent.

- Classifying the customers

  The dataset was split into train and test with 80:20 ratios. Then we apply machine learning algorithm. Since the goal is to define the class to which a client belongs and this, as soon as its first visit, we only keep the variables that describe the content of the basket, and do not take into account the variables related to the frequency of visits or variations of the basket price over time. We used different machine learning algorithms such as SVM, Random

Forest classifier, Logistic Regression, K nearest Neighbor, Adaboost Classifier, Gradient Boosting Classifier. Precision is calculated for each model.

- Testing the predictions
  We are testing the model the last two months of the dataset, that has been stored in the set_test data frame.

5. Result

5.1 Models used

- SVM (Support Vector Machine)

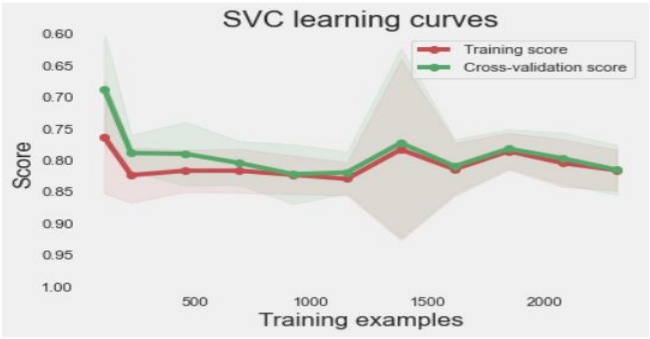

Figure 5.1.1: SVC learning curve

Precision 82.13%

- Logistic Regression

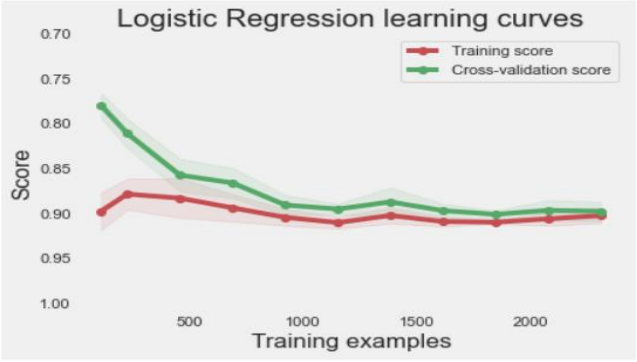

Figure 5.1.2: Logistic Regression learning curve

Precision 91.14%

- K-Nearest Neighbours

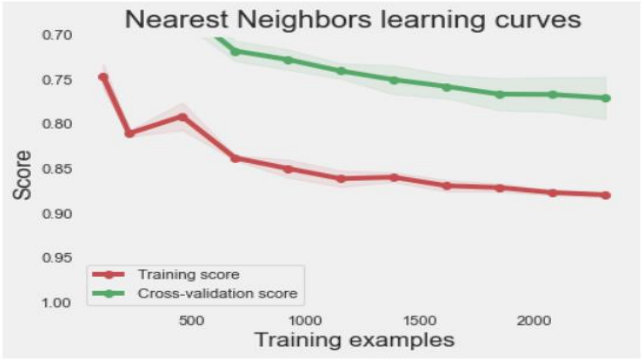

Figure 5.1.3: K- Nearest Neighbour learning curve

Precision 80.19%

- Decision Tree

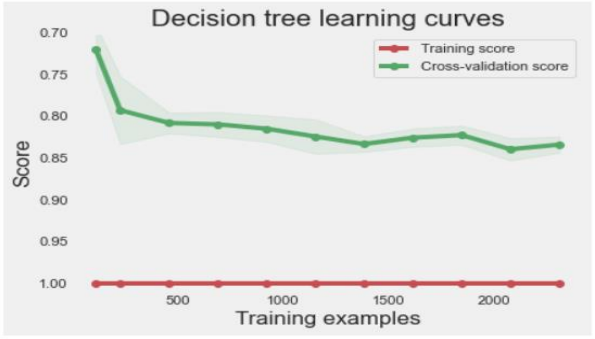

Figure 5.1.4: Decision Tree learning curve

Precision 86.29%

- Random Forest

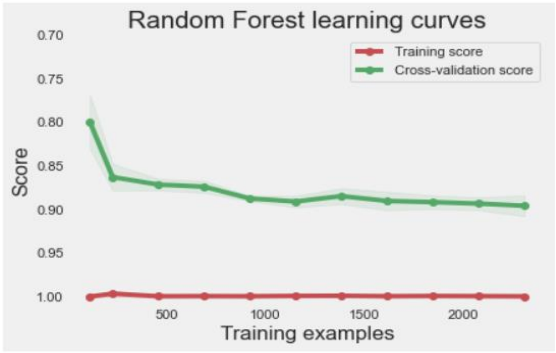

Figure 5.1.5: Random Forest learning curve

Precision 91.27%

- Ada Boost Classifier

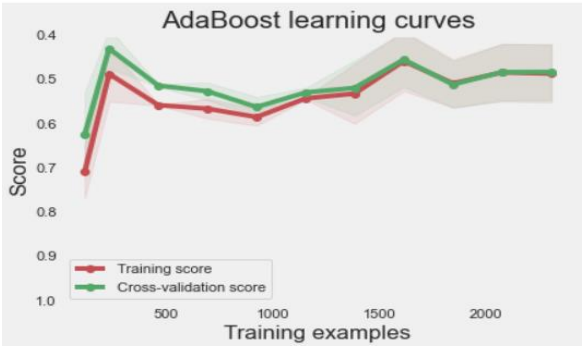

Figure 5.1.6: AdaBoost learning curve

Precision 52.91%

- Gradient Boosting Classifier

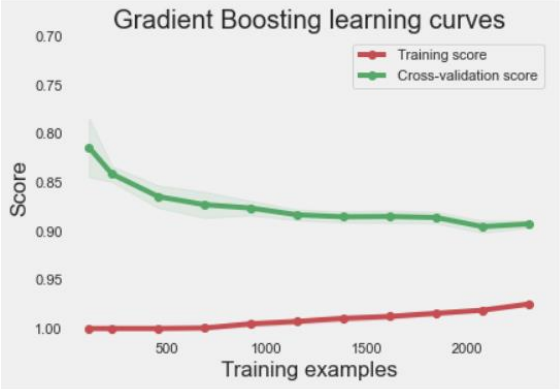

Figure 5.1.7: Gradient Boosting learning curve

Precision 89.47%.
After testing the data, following results was observed for prediction on different models, as given below:

Support Vector Machine
Precision: 67.31 %

Logistic Regression
Precision: 75.42 %

---

k-Nearest Neighbors
Precision: 66.88 %

Decision Tree
Precision: 72.09 %

Random Forest
Precision: 75.30 %

Gradient Boosting
Precision: 75.46 %

6. Conclusion and applicability

For this, the classifier is based on 5 variables which are: mean: amount of the basket of the current purchase categ_N with N∈[0:4] : percentage spent in product category with index N.

Finally, the quality of the predictions of the different classifiers was tested over the last two months of the dataset. The data were then processed in two steps: first, all the data was considered (ober the 2 months) to define the category to which each client belongs, and then, the classifier predictions were compared with this category assignment. We then found that 75% of clients are awarded the right classes. The performance of the classifier therefore seems correct given the potential shortcomings of the current model. In particular, a bias that has not been dealt with concerns the seasonality of purchases and the fact that purchasing habits will potentially depend on the time of year (for example, Christmas).

In practice, this seasonal effect may cause the categories defined over a 10-month period to be quite different from those extrapolated from the last two months. In order to correct such bias, it would be beneficial to have data that would cover a longer period of time. Through customer segmentation and personalized marketing campaigns, you can reduce the risk of showing ads to uninterested consumers. This clearly increases the efficiency of the campaign and hence produces ROI-producing efforts.