**ALY 6020 Data Predictive Analytics**

Northeastern University

Professor Marcos

A picture containing black, photo, red, sitting

Description automatically generated

College of Professional Studies, Northeastern University, Boston, MA 02115

This report was created as part of the Final Project for this course. The following report uses raw data UCI Data repository.

.

Submitted By:

Vinoothna Peruri

Contact : [peruri.v@husky.neu.edu](mailto:peruri.v@husky.neu.edu)

**SUMMARY**

This report aims at analyzing the content of a Uk based marketing database containing the list of purchases made by 4000 customers over one year to predict the purchases that will be made by a new customer, during the following year and this, from its first purchase.

 Each row in the dataset describes the purchase of a product, by a customer and at a given date. In total, approximately 4000 customers appeared in the database. Given the available information, I decided to develop a classifier that allows for predicting the type of purchase that a customer will make, and this from their first visit to the E-commerce website.

Stage I

The first stage of this work consisted of explaining the different products sold by the site, which was the subject of the first classification. Then, I grouped the different products into 5 main classes of items using K-means clustering.

Stage II

In a second step, I performed a classification of the buyers by analyzing their consumption habits. I have classified clients into 11 main categories based on the variety of products they usually buy, the number of visits they make and the amount they spent during the initial 10 months. For this, the classifier is based on 5 variables which are:

* **mean:** the amount of the cart/basket of the current purchase
* **categ\_N :** with N∈[0:4]: percentage spent in a product category with index N.

Once these categories established, I finally trained several classifiers like kNN, Decision tree, Random Forest and Logistic Regression whose purpose is to be able to classify consumers in one of these 11 categories and this from their first purchase. The random forest model obtained the highest accuracy of 92.63%.

**INTRODUCTION**

This report focuses on analyzing the content of an E-commerce database, mainly focusing on customer segmentation. It is the practice of dividing a customer base into groups of individuals that are similar in distinct ways. To provide diverse value propositions to different customer groups. Customer segments are usually determined on similarities, such as personal traits, preferences or practices that should correlate with the same behaviours that drive customer profitability. (Rahula Raj, Sept 2013).

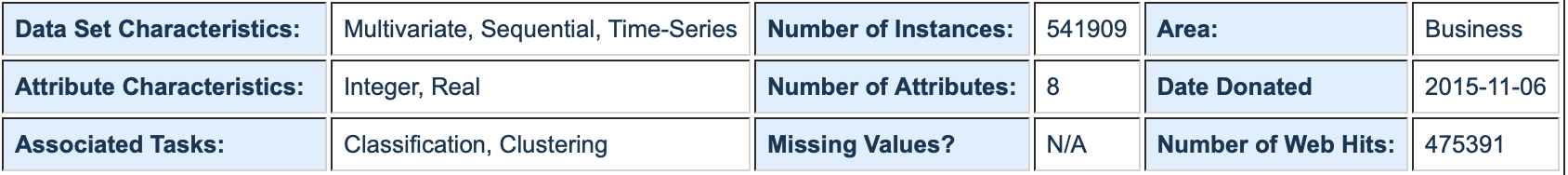
**ANALYSIS**

**Where did the data come from and why did we choose this data?**

This is a transnational data set from UCI machine Learning Repository which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

*Link to the DATASET* : https://archive.ics.uci.edu/ml/datasets/online+retail#

Figure 1: Dataset Characteristics of Online Retail data



It’s natural for companies to use more costs on certain customer segments than others. Not all visitors to a company's website will want all the things the company is willing to offer. **B**enefits of customer segmentation include staying a step ahead of the competition and identifying new products that existing or potential customers could be interested in. This would help me understand how the Online Retail businesses operate and the crucial relationship between the customer and the business.

**What is the business/social problem we are trying to solve?**

## Static Segments Fail to Provide Enough Actionable Detail -

**There are two main reasons for this failing of detail: The machinery and transparency. A segmentation system using basic statistical modelling simply will not be able to capture the thousands of fields and rows that originate from**[**multiple data sources**](https://simmachines.com/machine-learning-prediction-methodology/demos/query-multiple-databases/) **(Simachines,** Emily Webber)

* Traditional segmentation analysis Aren’t Updated Based on Changing Customer Behaviors and cannot have timely, relevant information for decision-making.

**IMPLEMENTATION**

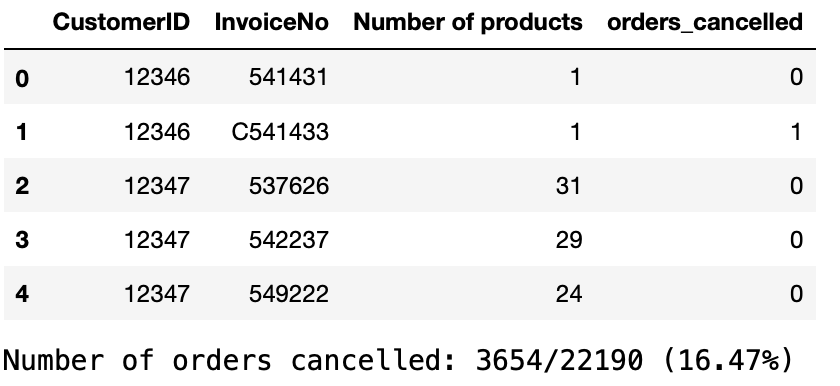
1. **Cleaning Data**

* Data Preparation

Firstly, we import the retail dataset from UCI repository. This data frame contains 8 variables. U*sing ‘dtypes’* we check it’s summary, types of variables, the number of null values and their percentage.

* Null Values

Figure 1: Identifying existence of entries with the prefix C as Cancellation orders.



We observed the number of null values in the data frame ‘CustomerID’ and noted that about 16.7% of the entries are not assigned to a particular customer. With the data available that is available to us, it is undesirable for us to impute values for the user and these entries are thus not useful for the current assignment. So I deleted them from the data frame. After the deletion of those entries, I checked for duplicate entries and found that they were associated with the cancelled orders. The existence of entries with the prefix C in the InvoiceNo variable, have found to be indicating transactions that have been 'cancelled'.

I also observed that when an order is cancelled, we have another transaction in the data frame, mostly identical except for the Quantity and InvoiceDate variables. To check if the cancellations are always with multiple entry counterparts, I counted the number of transactions corresponding to cancelled orders and noted that the number of cancellations is ∼16% of the total number of transactions which is quite substantial. On these few lines, we see that when an order is cancelled, we have another transaction in the data frame, mostly identical except for the Quantity and InvoiceDate variables. I decided to check if this is true for all the entries. To do this, I decide to locate the entries that indicate a negative transaction and check if there is systematically an order indicating the same quantity but positive, with the same description (CustomerID, Description and UnitPrice).

I stored these entries in the 'doubtfull\_entry' list corresponding to the entries indicating a cancellation but for which there is no command beforehand. I chose to delete all of these entries, which account for ∼1.4% and 0.2% of the data frame entries.

1. **Exploratory Analysis**

* The highest number of orders are coming from the United kingdom. TOP 5 countries (including United Kingdom) that place the highest number of orders are as below:

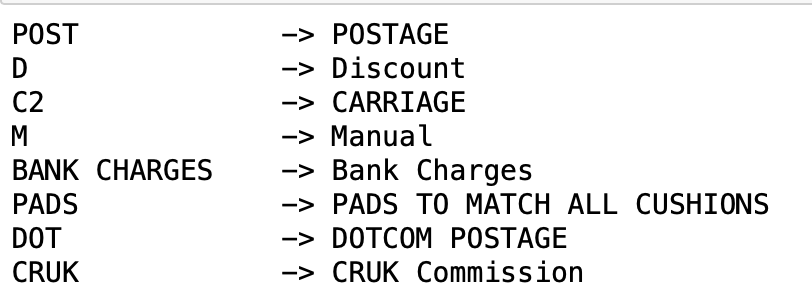
1. United Kingdom
2. Germany
3. France
4. Ireland
5. Spain

Figure 2: Number of orders per country

* **Netherlands has the highest money spent customer.**
* The company also tends to **give out FREE items for purchases occasionally.**
* The total number of transactions carried are 22190.
* The entries with a prefix C in the column **“InvoiceNo”** represents transactions that have been cancelled.
* There is an existence of users who had only come once and only purchased one product and also the existence of frequent users that place a large order of numerous items at once.
* Some values of the column “Stock Code” indicate particular transactions like Discounts, Bank charges, commisions etc.

Figure.3 : Peculiar transactions

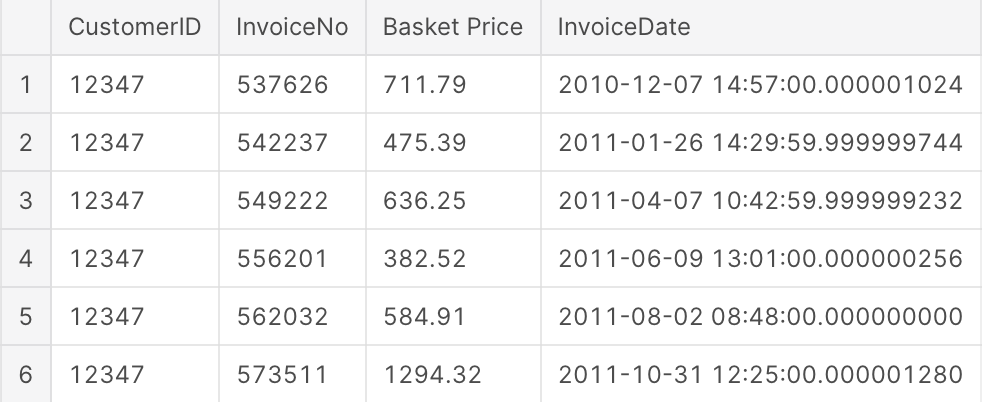


* It can be seen that the vast majority of orders concern relatively large purchases given that ∼65% of purchases give prizes in excess of £ 200.

1. **Feature Engineering**

Next, we created a new variable called the "basket\_price" that indicates the total price of every purchase. Each entry of this data frame indicates prices for a single kind of product. Since we have orders that are split into several lines, we accumulate all the purchases made during a single order to recover the total order price.

Figure 4: Basket price obtained along with Customer ID



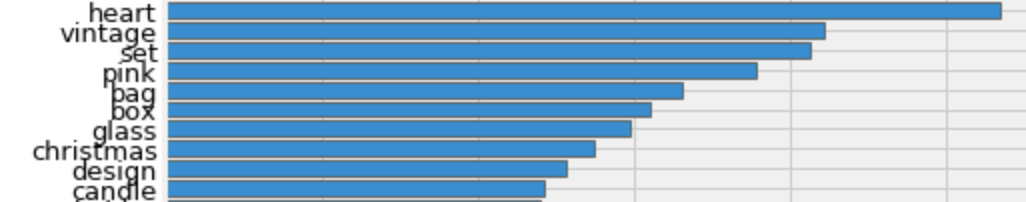
1. **Defining Product categories**

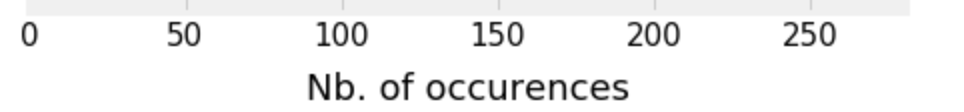
This function analyzes the content of the **Description** column by performing the following operations:

* extract the names both proper, common appearing in the products description, then extracts the root of the word for each of them and aggregates the set of names associated with this particular root.
* count the number of times each root appears in the data frame

In this process, we created a 'keywords\_roots' ; a dictionary where the values are the lists of words associated with those roots. Which then systematically selects the singular variants called 'count\_keywords'. The number of keywords returned is 1483. Further, we analyze the description of these various products by converting the count\_keywords dictionary into a list and sort the keywords according to their occurrences.

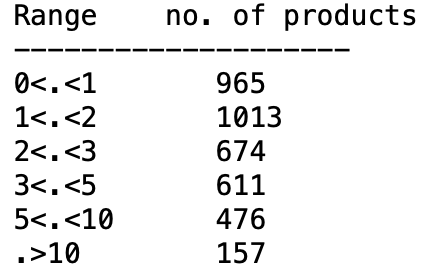
Figure 5: No of occurences of the keywords



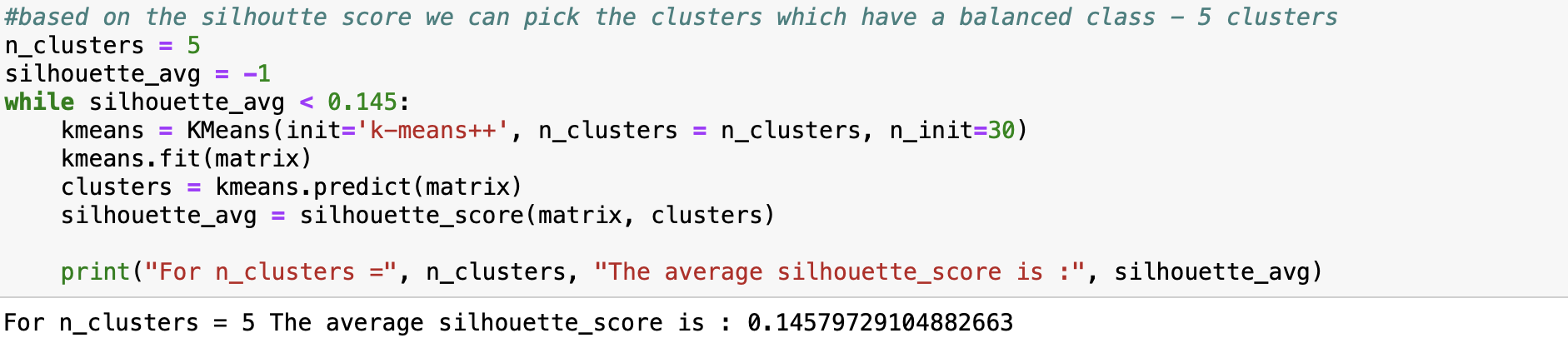
We check the number of products in these different product categories. Next we define product categories with the obtained '1483' keywords obtained. We kept the keywords that have appeared more than 13 times and discarded the keywords like colours that don't carry much information about a customer. We one hot encoded using a matrix X based on if the product description contains a keyword or otherwise.  We have found that introducing the price range would result in a much more balanced groups in terms of element numbers. Hence, we added 6 extra columns to this matrix, where I indicate the price range of the products.

Figure 6: Price range of products



In order to approximately define the number of clusters that best represents the data, we use the silhouette score, which is the measure of how similar an object is to its own cluster.

Figure 7: Below is the code that is used to check the silhouette scores.



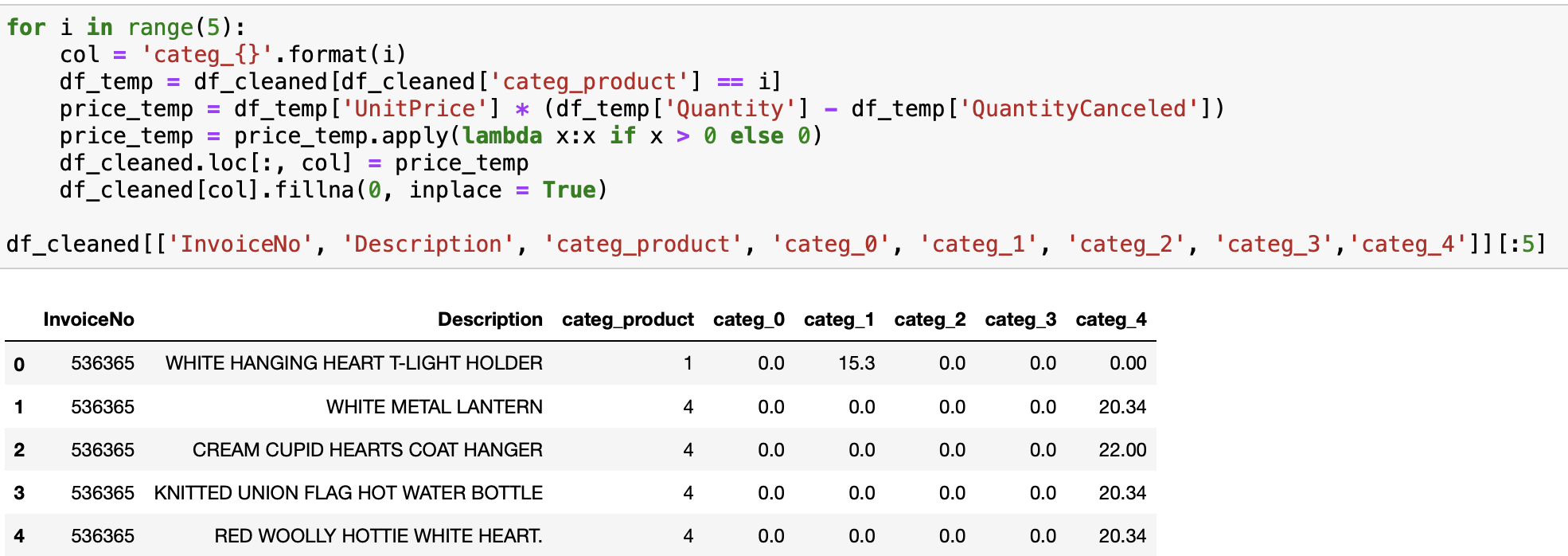
In our critique, we found that beyond 5 clusters, there contained very few items. So we choose to separate into only 5 clusters also ensuring best Silhouette score.

1. **Defining Customer categories**

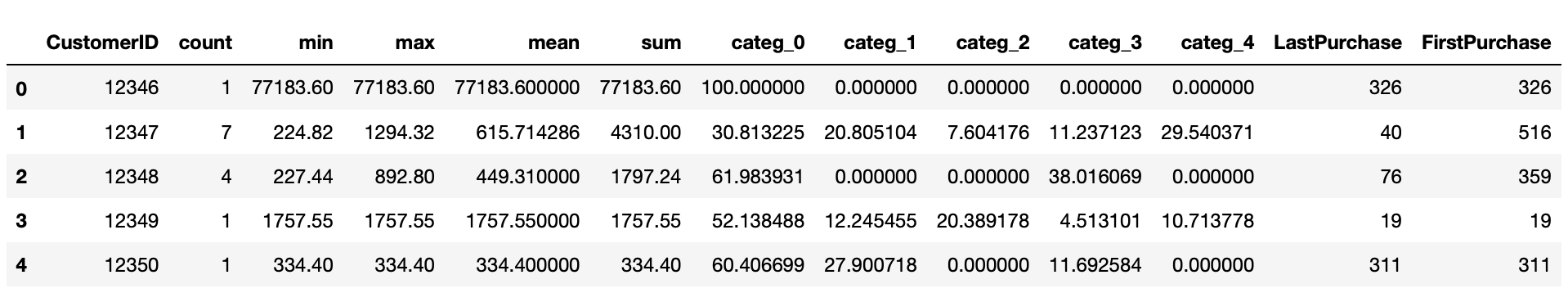
Following, we create Customer categories, we create the categorical variable '**categ\_product'** and '**categ\_N**' where we indicate the cluster of each product and the respective amount spent in each category.  I therefore created a new dataframe that contains, for each order, the amount of the basket, as well as the way it is distributed over the 5 categories of products.

Figure 8: Below is the code used to indicate the cluster of each product and it’s output containing the categories/classifications of each product.

.

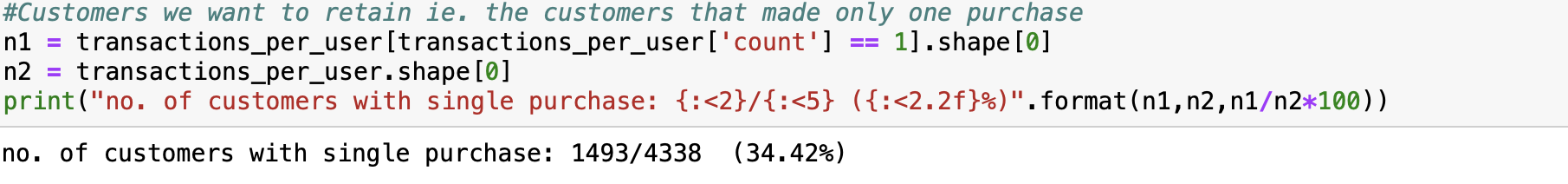


Then I finally grouped together with the number of purchases made by the user, as well as the minimum, maximum, average amounts and the total amount spent during all the visits and two additional variables that give the number of days elapsed since the first purchase and the number of days since the last purchase.

Figure 9: Table representing min-max, count and sum of purchases, product categories and their 1st and last purchases.  


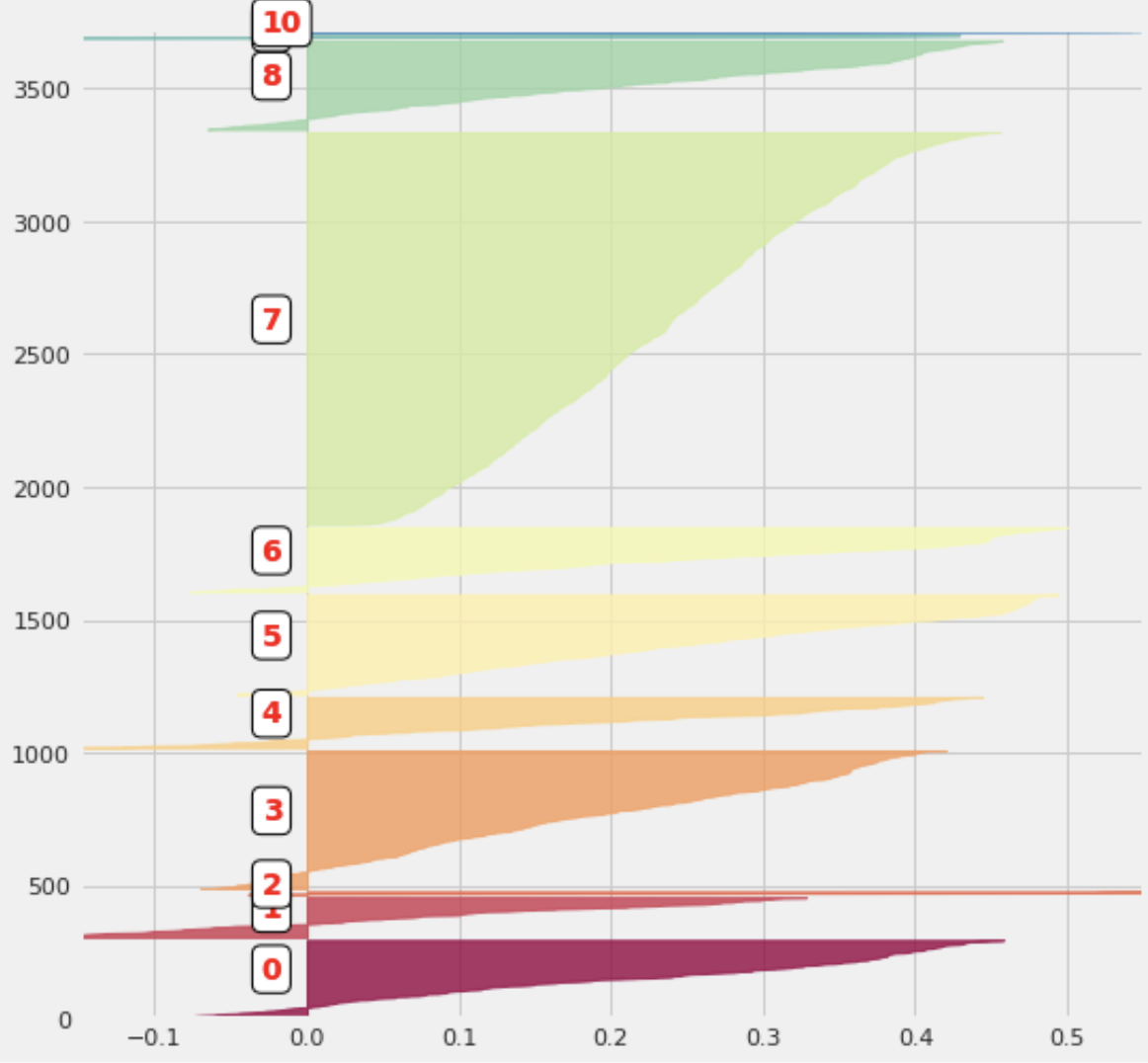
Since the goal is to define the class to which a client belongs and this, as soon as its first visit, to target these customers in order to retain them. In part, I find that this type of customer represents 1/3 of the customers listed in the dataset which I believe is good for our project.

Figure 10: Customers that made only one purchase.



Next, I created a matrix (Data encoding) where these data are standardized. Then I defined clusters of clients from the standardized matrix using the k-means algorithm from sci-kit-learn. The way to look at the quality of the separation of the clusters is through the silhouette score, so we plotted quality of seperation using Intra-cluster Silhoutte scores.

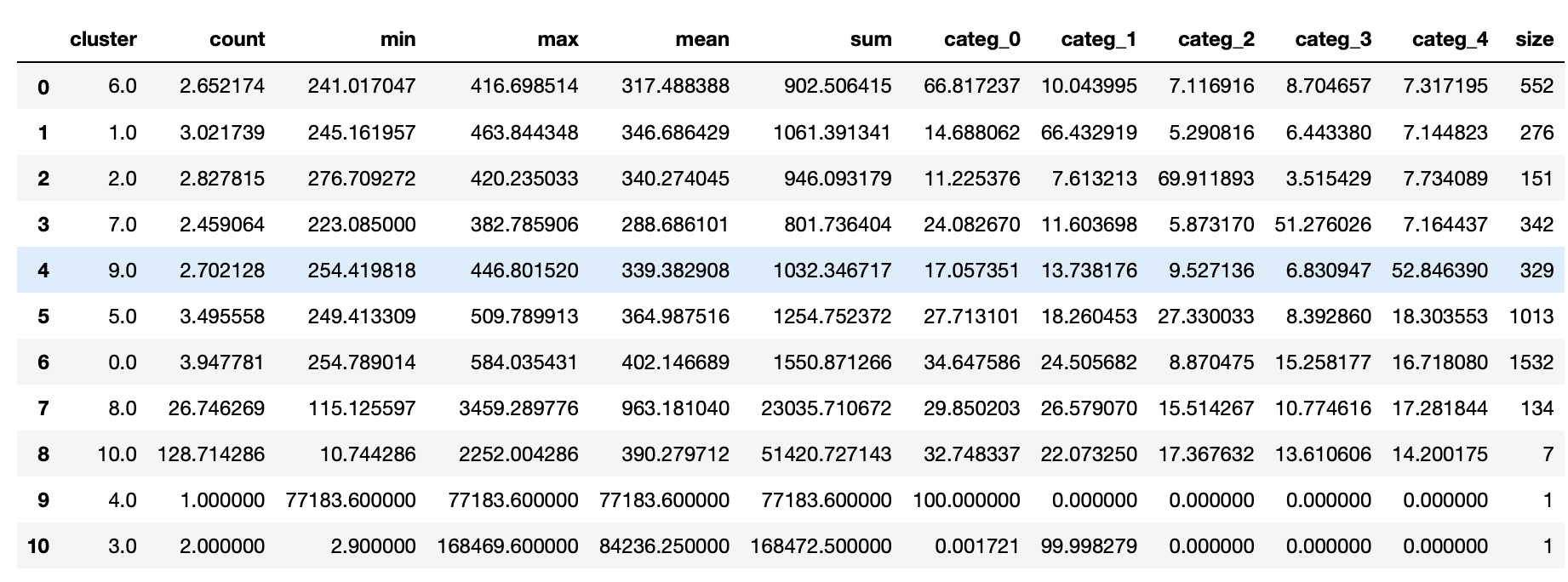
Figure 11: Intra-cluster Silhoutte score plot



Validated the ‘Customer Clusters’ using ‘Silhouette scores’ and verified that the different clusters are indeed disjoint.

Then, we check the average the contents of this data frame by first selecting the different groups of clients. This gives us a better understanding of the average baskets price, the number of visits, the total amount spent by the clients of the different clusters. I also determine the number of clients in each group under 'variable **size'.**

Figure 11: Averages of contents in the different client clusters.



Finally we split the dataset into train and test datasets.

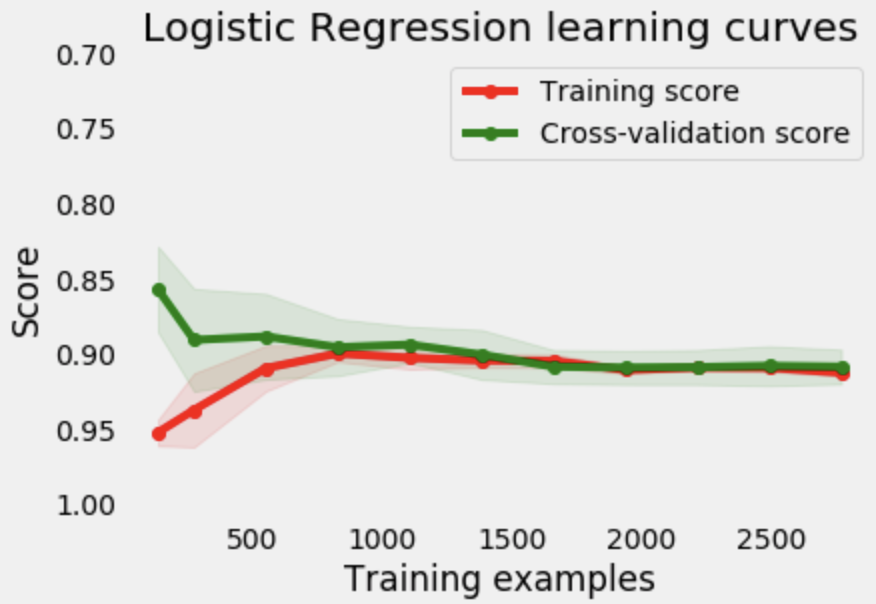
1. **Data Modelling**

We tested several classifiers implemented in scikit-learn like logistic regression, KNN, Decision tree and random forest to classify consumers in one of the 11 client categories**.**

LOGISTIC REGRESSION

We created an instance of the Class\_Fit class to adjust the model on the training data and see how the predictions compare to the real values.

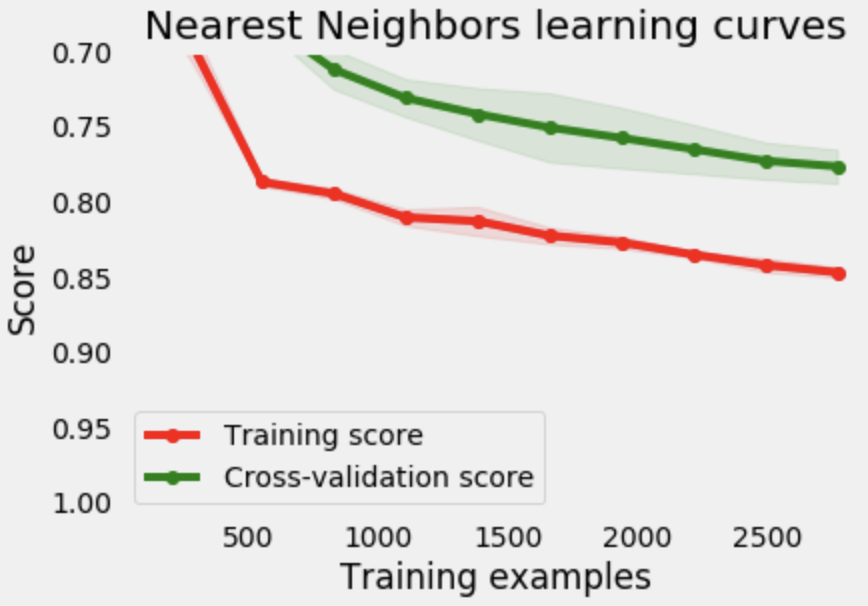
Figure 12: Logistic regression Plot



**We have a prediction accuracy of 89.17%.**

K NEAREST NEIGHBOURS (kNN)

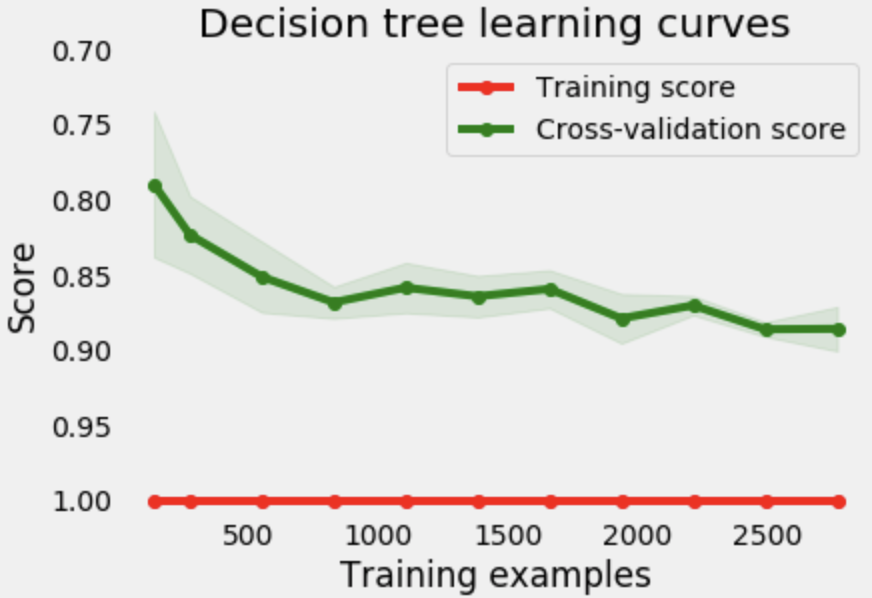
Figure 12: KNN Plot



**We have had a precision of 80.76%.**

DECISION TREE

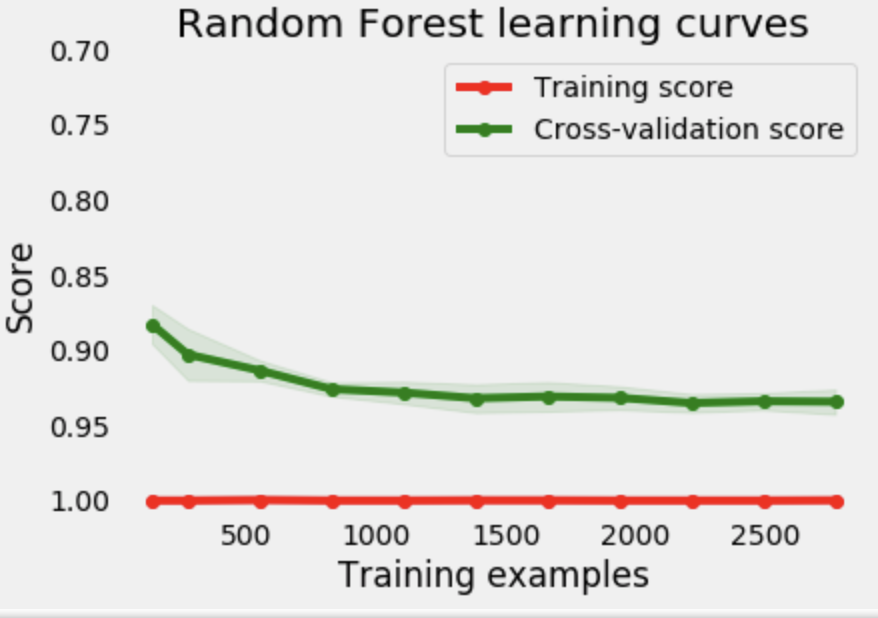
Figure 13: Decision tree Plot



**The accuracy obtained is 87.33%**

RANDOM FOREST

Figure 14: Random forest plot



**The accuracy obtained is 92.63%.**

**DATA ANALYSIS**

Comparing the accuracies of applied 4 models.

|  |  |
| --- | --- |
| * Logistic regression | * 89.17% |
| * k-NN | * 80.76% |
| * Decision Tree | * 87.33% |
| * Random Forest | * 92.63% |

Among these optimization techniques, the Random forest gives the best model with highest

prediction accuracy.

We found that around 90% of clients are awarded in 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 of the seasonality of purchases because purchasing habits will potentially depend on the time of year, for example, Christmas or Black Friday. This seasonal effect might cause the categories defined over a 10-month period to bias. In order to correct such bias, it would be beneficial to have data that would cover a longer period of time, maybe a few years.

We prove our novel proposal on how implicit features obtained through clickstream/weblogs, time spent on the website, items bought and amount spent, etc act as significant features in establishing customer behaviour, experience and hence can be used as features to find much narrower customer segments and customer churn. In our case, it was useful to target the first buyers in order to retain them.

**DISCUSSION**

The results of the different classifiers presented in the the abovesegement can be combined to improve the classification model. This can be achieved by selecting the customer category as the one indicated by the majority of classifiers. Another popular optimization technique, Gradient Boosting, can also be used to further optimize the model.

Once the model is settled by tuning the hyper-parameters, the area under the ROC Curve can be used to tweak the threshold probability such that we can tune the classifier to deliver the best predictions for a given quadrant in the confusion matrix as desired by us. Also identify and analyse the cost of mis-classifying non-churning customers.

Another important avenue worth exploring for addressing customer segementation is the application of Time series analysis. As the businesses grow and get more complex there are more additional data channels that continuously open up holding valuable information. This needs to be captured and harnessed for such or similar applications.

An optimisation technique would be The popular RFM method; Recency, Frequency, Monetary analysis method and can be used for data reduction. It decreases negative reactions from customers due to controlled targeting and is implemented with email marketing .

**References:**

[1] Emily Webber

https://simmachines.com/5-market-segmentation-challenges-solved/

[2] Rahula Raj, Sept 2013- <https://www.howhiz.com/accepting-failure-and-stop/>

[3] <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html>

[4] Sklearn documentation- <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html>

[5]https://lib.dr.iastate.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=7030&context=etd