Predicting customer preferences using RFM analytics

Faculty of Science
Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam

Research Paper Business Analytics

Veronika Zhezhela

Supervisor: Prof.dr. Sandjai Bhulai

December 2017

Abstract

Online retailers seek approaches to build valuable relationships with their existing customers. In order to develop a pleasant and long-term relationship, increase customer satisfaction and loyalty retailers have to be able to hit the right customers with the right offers at the right times. To do so, retailers can leverage the insights gained from transactional data to provide consumers with a personalized shopping experience. This research aims to segment customers of an online retailer based on behavioral variables and predict distinct segment preferences. On the basis of the k-means algorithm and RFM model-based clustering, customers of the store have been segmented into five meaningful groups. For each target group a number of products were predicted by means of logistic regression. The proposed approach enables online retailers to predict relevant products by value-based customer segmentation. The insights can be used for the customizing marketing strategies and fulfilling different customers' needs by allocating resources effectively and efficiently.

Contents

1 Introduction 2

2 Literature review 4

2.1 Customer segmentation 4

2.1.1 k-means clustering 4

2.1.2 RFM analytics 6

2.2 Consumer preference prediction 7

2.2.1 Logistic regression 7

3 Understand your data 9

3.1 Data exploration 9

3.2 Feature engineering 14

4 Model implementation 15

4.1 k-means clustering 15

4.2 RFM analytics 16

4.3 Logistic regression 16

5 Conclusions 19

1. Introduction

According to the latest industry trends, challenges that online retail is facing are:

1. customers are no longer limited to location or channel, therefore retailers have to be flexible and adapt fast to new trends, customer preferences in order to be relevant and useful;
2. consumers' demand and expectations increased, thus retailers have to provide personalized shopping experience, which entails offering to the right customer at the right time for the right price the right product or content;
3. retailers have to protect customers data to build loyal and trustworthy relationships, fulfill all their privacy needs.

It is a very rapidly growing and competitive industry, small and midsized chains, who wish to continue their growth, face an even greater set of challenges than single stores or large retailers. There are many ways to address these challenges, enabling retailers to establish a long-lasting relationship with consumers.

The advantage of online retail nowadays is the possibility of tracking each step of the customer, record it and convert it into knowledge. Retailers collect a large number of transactions daily. Knowledge extracted from this data can help the organization to get to know their customers, optimize inventory levels at different locations, improve the store layout and sales promotions, optimize logistics by predicting seasonal effects, minimize losses. Moreover, they can use the data for competitive advantage.

Companies can get meaning of the data utilizing data mining, often referred to as "analytical intelligence". Data mining discovers patterns and relationships hidden in data, and is actually part of a larger process called "knowledge discovery", which describes the steps that must be taken to ensure meaningful results [1]. In addition, data mining tools predict future trends and behaviors, help organizations to make proactive knowledge-driven decisions [2].

In this research a case study of using data mining techniques for segmentation and prediction of customer preferences for the online retailer is presented. The goal was to learn about the customers as much as possible and use obtained knowledge to derive profit drivers. The proposed method can help retailers to provide a more personalized customer experience in the future.

This paper structured as follows. In the second section techniques are described with a literature

review on their applications. The section "Understanding your data" discusses in detail the data analysis. The subsequent section presents the implementation of the segmentation and prediction algorithms and shows the obtained results. The section after provides conclusions and final remarks.

2. Literature review

2.1 Customer segmentation

When the transaction size of a retailer increases, it is necessary to classify customers based on their similarities into specific and homogenous segments which are non-homogenous by pairs. In this way, the value of different classes can be calculated [3]. The classification can be approached in different ways. One of them is clustering. The clustering algorithm aims to identify groups of customers that minimizes differences among people in the same cluster while maximizing the difference between clusters. The segmentation of customers is a standard application of cluster analysis [4].

There are different clustering algorithms each having strengths and weaknesses. In this research we will first use the unsupervised machine learning algorithm k-means. The assumption about the optimal number of clusters for the dataset will be based on the performance of the algorithm and the silhouette measure. Subsequently, for the optimal number of clusters we will conduct RFM segmentation, which allows us to interpret and understand each cluster.

The goal is to segment customers in order to provide distinct products based on the customers' value. The output of the segmentation is discrete customer groups that differ based on both needs and behavior.

2.1.1 k-means clustering

k-means [5] is a non-hierarchal clustering method that assigns objects to a pre-specified number of clusters.

The clustering process starts by randomly assigning objects to a determined number of clusters with k centers. Each cluster center is updated to be the mean of its constituent instances. The objects are then successively reassigned to other clusters such that the within-cluster variation is minimized, which is the distance metric from each observation to the cluster centroid. The distance metric used in this research is the Manhattan distance, which is the sum of the absolute values of the differences of the coordinates. The distance is calculated as $\sum_{i=1}^{n} |x_i - y_i|$, where $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ are two vectors in an n-dimensional real vector space.

The algorithm converges when there is no further change in the reassignment of instances to

clusters. Depending on the input data distribution and initialized number of clusters the required number of iterations might differ from few to thousands.

The algorithm is easy to implement and performs faster than hierarchal clustering techniques if k is small. However, it requires some prior knowledge of the number of clusters (k-value) and initial seeds have a strong impact on the final result. Besides, one of the assumptions is balanced dataset, which in practice is highly unlikely. Lastly, we cannot assign meaning to the cluster, which is the focus of this research.

We will use the algorithm as an easy and fast way to decide if the clustering is appropriate and use the silhouette measure to determine the initial number of clusters for our dataset.

Silhouette measure

The silhouette measure captures information about cluster goodness and an approach to defining the optimum value for the number of clusters during k-means algorithm. It was first described by Peter Rousseeuw in 1986 [6]. Intuitively, good clusters have the property that instances within the cluster are close to each other and far from instances of other clusters.
Before calculating the silhouette measure:

1. for each cluster member calculate dissimilarity within the cluster: average distance from all other instances within the cluster;
2. average dissimilarity for a cluster, which measures compactness of the cluster;
3. average distance for instance to members of the neighboring cluster.

Low dissimilarity within the cluster means good fit for the instance to the assigned cluster. It is worth mentioning that two members of the same cluster may have different neighboring clusters. For points that are close to the boundary between two clusters, the two dissimilarity scores may be nearly equal.

The silhouette of the cluster is the ratio of an instance's dissimilarity to its own cluster to its dissimilarity with its nearest neighboring cluster:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.1}$$

where $a(i)$ = average dissimilarity of i to all other objects of cluster A,
$b(i) = min_{C \neq A} d(i, C)$,
$d(i, C)$ = average dissimilarity of i to all objects of cluster C.

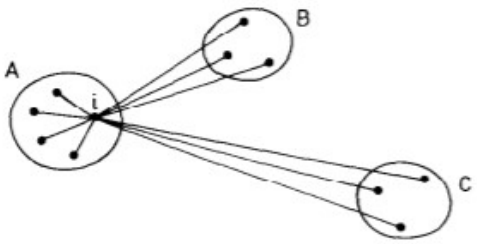


Figure 2.1: An illustration of the elements involved in computation of *s(i)*

Silhouette values lie in the range of $[-1, 1]$, where a higher value indicates better cluster configuration. 1 indicates that the sample is far away from its neighboring cluster and very close to the cluster to which it is assigned. Likewise, a value of -1 indicates that the instance is closer to its

neighboring cluster than to the assigned cluster. A silhouette value of 0 means that object is at the boundary of the distance between the two clusters.

By looking at the average silhouette over all instances with respect to the minimum needed number of clusters the optimal number of clusters can be defined.

2.1.2 RFM analytics

The RFM (recency, frequency and monetary) model is a behavior-based model used to analyze the behavior of a customer and then make predictions based on the behavior in the database [7]. RFM is a widely-used model for measuring the strength of the relationship established with customers as it can efficiently identify valuable customers [3]. The method is based on calculating three behavioral variables for each customer: recency, frequency and monetary value. Recency represents the length of a time period since the last purchase, frequency denotes the number of purchases within a specified time period and monetary means the amount of money spent in this specified time period [8].

These variables give us insights on when people buy, how often and how much they buy. In this way, retailers can analyze customers' behavior, segment the market and identify the most and least promising customers. The value of each customer can be calculated based on the assigned RFM class.

First, the RFM values for each customer has to be determined:

1. recency: most recent purchase date;
2. frequency: number of transactions within the period;
3. monetary value: total sales for the customer.

Each value has to be assigned into one of the RFM categories. There are different approaches to decide on the number of categories. The cut-offs imposed on each of the three parameters can be fixed ranges or quantiles. By using quantiles, we will divide data into 4 equal parts. Thus, each value will fall into one of the quantiles. The RFM score is the number between 1 and 4 based on the quantile. Four equal groups across three variables create 64 different customer segments. The RFM Class is a combined RFM score, which is simply the three individual scores concatenated into a single value.

For example, the customer who is within the group who purchased most recently (R score = 1), who purchased most often (F score = 1), who spent the most (M score = 1) will have the combined RFM score = 111. These customers are more likely to respond to the offer. In Table 2.1 a few of the groups which the retailer can access using this technique are presented with the description and the appropriate marketing strategy. From the table, it is very easy to filter out different types of customers based on their class.

However, the RFM values are inclined to be firm-specific and are based on the nature of the products [8]. The RFM-model was applied in different areas of business and should be modified according to the business needs. Chan [9] combined the customer segmentation and the targeting approach. He showed that the RFM approach provides better results for targeting valuable customers compared to a random selection approach. Cheng and Chen presented a method to combine the RFM factors and k-means algorithm to enhance segmentation accuracy and develop classification rules enabling organizations to achieve a successful customer relationship model [10]. Hsieh applied RFM to the banking customers' segmentation, defining the recency measure as the average time distance between the day of making a charge and the day of paying the bill. The frequency measure as the average number of credit card purchases made, and monetary measures as the amount of consumption spent during a yearly time period [11]. Also, the model has found applications in nonprofits and

Segment	*RFM Class*	*Description*	*Marketing*
Best Customers	111	Bought most recently and most often, and spend the most	No price incentives, new products, and loyalty programs
Loyal Customers	X1X	Buy most frequently	Use R and M to further segment
Big Spenders	XX1	Spend the most	Market your most expensive products
Almost Lost	311	Have not purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Customers	411	Have not purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Cheap Customers	444	Last purchase long ago, purchased few, and spent little	Do not spend too much trying to re-acquire

Table 2.1: Key RFM segments

financial organizations [12], government agencies [13], on-line industries [14], telecommunication industries, travel industries [15].

There are obvious advantages of implementing RFM. For instance, the model is very effective as the purchase behavior can be summarized by using a very small number of variables. Although, there are different thoughts on the interpretation of the segments. Miglautsch [16] referred to the customers with the highest RFM score as not "best customers" but the biggest customer segment that may have the greatest untapped potential. The simplicity of the RFM model has been overemphasized and its ability to differentiate has little to be considered. RFM focuses on a company's current customers and cannot be applied to the prospecting for new customers as a marketer does not have transactions for prospects [17].

2.2 Consumer preference prediction

2.2.1 Logistic regression

Multinomial logit regression models are routinely applied as a means of classifying individuals or instances in various domains including marketing and market modelling [18]. Accurate prediction of shopping preferences has become an important issue for retailers seeking to maximize customer loyalty.

The logistic regression model is a discriminative classifier that builds a linear model based on a transformed target variable. The objective is to establish a classification system based on the model for predicting class membership.

Unlike linear regression, the prediction for the output is transformed using a non-linear function called the logistic function. The model predicts the probability whether the customer belongs to the class or not by fitting data to a logit function. For the binary output variable Y there is a set of explanatory variables $X = (X_1, X_2, ..., X_k)$ which can be discrete, continuous, or a combination. x_i is the observed value of the explanatory variables for observation i. We want to model the conditional

probability $Pr(Y = 1|X = x)$ as a function of x. Any unknown parameters in the function are to be estimated by maximum likelihood. The logistic regression model:

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta x$$

Solving for π_i gives

$$\pi_i = Pr(Y_i = 1|X_i = x_i) = \frac{e^{(\beta_0+\beta x)}}{1+e^{(\beta_0+\beta x)}} = \frac{1}{1+e^{-(\beta_0+\beta x)}}.$$

.

3. Understand your data

Part of any data mining project is learning about and understanding the nature of your data. It is no secret that 80% of the time and effort it spent on preprocessing data, because it is crucial to reshape and refine the dataset, which can be leveraged for models. In order to investigate the data you have to know if there are discrepancies in the names or codes, outliers or errors, what are the attributes of interest for analysis.

We will start with data exploration. Then bridging the gap between data and model to achieve better performance, we will transform the data and create features for better representation of the data. After getting the data into the right format and extracting meaningful features, we will apply a supervised learning algorithm and evaluate the performance.

3.1 Data exploration

Our retail data is a transnational dataset of the online retailer with the 541 910 instances in the period between 01/12/2010 and 09/12/2011.

The original data contains the following variables (columns):

1. InvoiceNo: invoice number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation, ‘d’ denotes discount.
2. StockCode: product (item) code uniquely assigned to each distinct product.
3. Description: product (item) name.
4. Quantity: the quantities of each product (item) per transaction.
5. InvoiceDate: invoice date and time, the day and time when each transaction was generated.
6. UnitPrice: unit price. Product price per unit in sterling.
7. CustomerID: customer number uniquely assigned to each customer.
8. Country: country name. The name of the country where each customer resides.

First exploration analysis showed that there were 25 900 transactions of 4 070 products with 4 372 unique customers among 38 countries. "CustomerID" is the only variable with missing values, precisely 135 080 records for 9 countries. There were 9 288 cancellations made and 77 purchases with discounts.

While exploring each variable to identify inconsistencies, we found 2 records with a negative Unit Price, 2 512 with Unit Price equal to zero and 10 624 records with a negative quantity (through which 9 288 cancellations and 1 336 with unknown customers, zero unit price and all coming from the United Kingdom). For further analysis these instances will be excluded.

For 530 104 records with positive quantity and unit price sales are 1 066 6684.5 currency units. We can see in Figure 3.1 that most purchases are below 20 units. Lost sales based on canceled transactions are 896 812.4 units. Mean sales per transaction is 20.12 units. The Netherlands has the highest mean sales per transaction, which is 121 units. The 22nd country is Germany 25.31, France with 24.95. The United Kingdom is thirty fifth with 18.60 units.

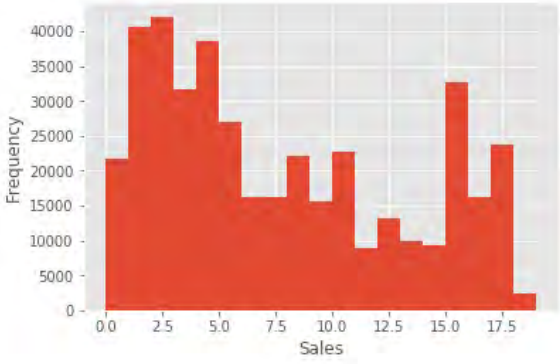


Figure 3.1: Histogram of the Sales

We decided to take a closer look on a distribution of transactions and sales among countries (see Figure 3.2). The United Kingdom has the highest number of transactions 18 021 and sales are 9 025 222.08 and also the number of customers 3920. The Netherlands' revenue is 285 446.3 with only 94 number of transactions and 9 customers. Intuitively, we can understand that these customers are important.

To see the distribution of customers among the countries we calculated the number of customers per country (Figure 3.3a). To get the number of placed orders per customer we computed the average number of transactions per customer (Figure 3.3b) and identified top customers in each country by percentage of the sales they made in the country.

The trend is that there is one or two big customers, who generated revenue in the country, with a difference only in their purchases. From that point of view, the Netherlands is the most promising country, because sales and number of products per transaction are the highest (Figure 3.4b, Figure 3.4d). The potential profit can be achieved by keeping customers from the Netherlands and finding a way to stimulate purchases for customers from France and Germany. France and Germany stood out as countries with relatively low sales compared to the countries with smaller number of transactions, such as EIRE or the Netherlands (Figure 3.2). However, France and Germany are countries with high number of customers (Figure 3.3a) and overall quantity of purchased items (Figure 3.4a).

Campaign for the Netherlands can be developed taking into account the fact that average sales per item are among lowest, so customers are more likely to buy cheaper products (Figure 3.4c). We took a step further and created time series to see if there is any seasonal trend, relationship with number of transactions for the countries of interest (Figure 3.5). We can see that there is a drop in April and July for all of them, and the rate in October for France is low. The captured information allows us to focus and target customers with innovative business strategy and marketing campaign more actively before low-sales period to raise the income.

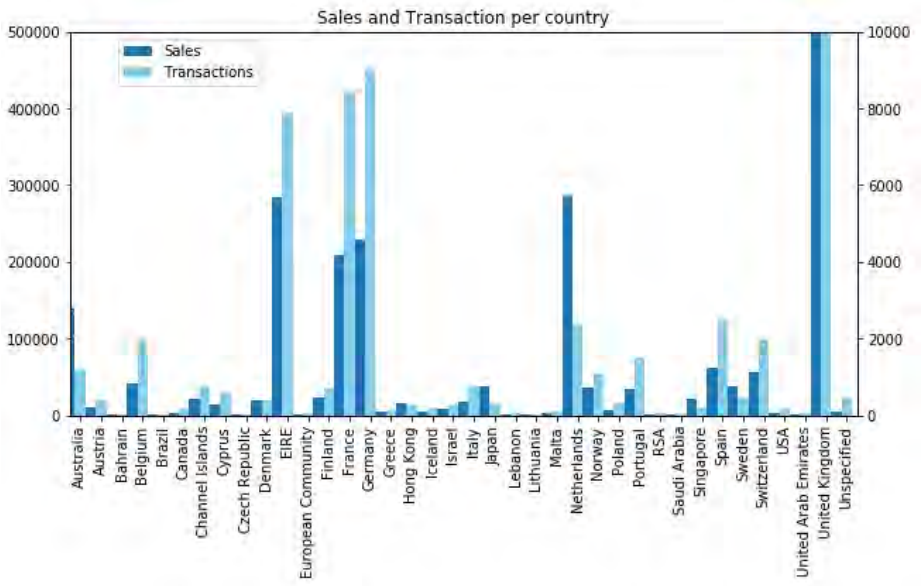
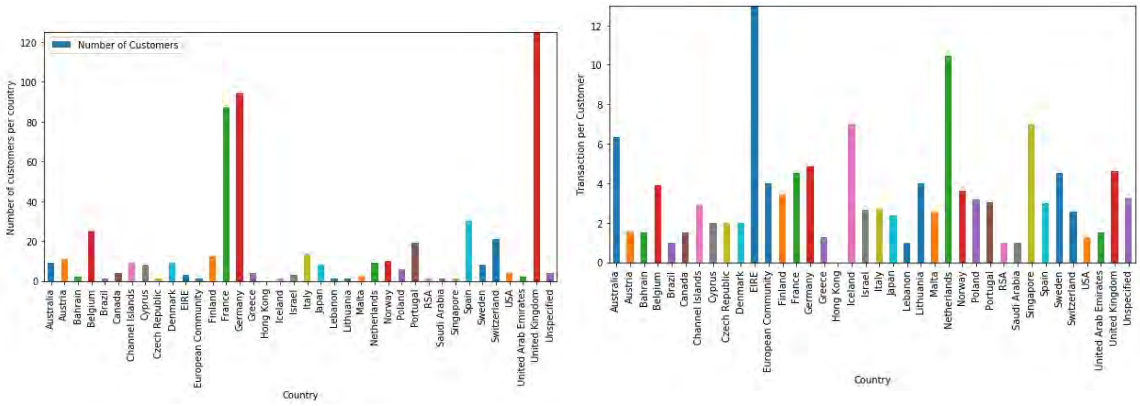


Figure 3.2: Sales and Number of transactions per country

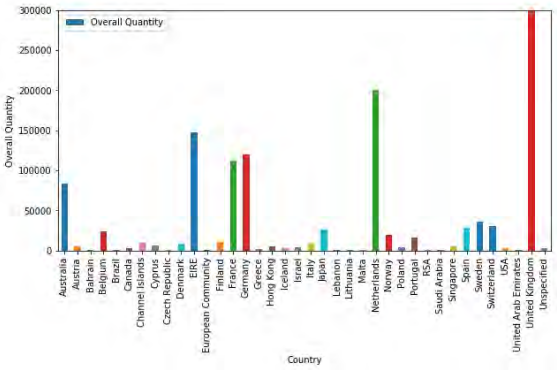


(a) Number of customers

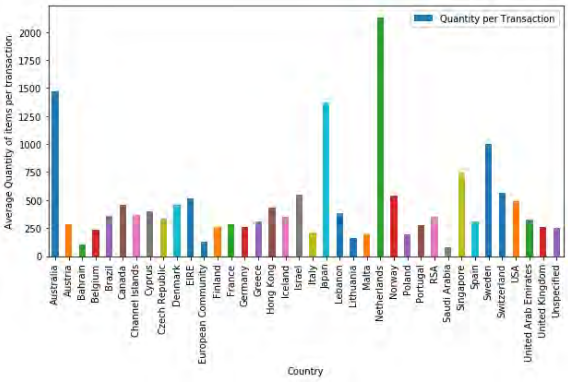
(b) Average number of transactions per customer (EIRE has 96)

Figure 3.3: Customer analysis

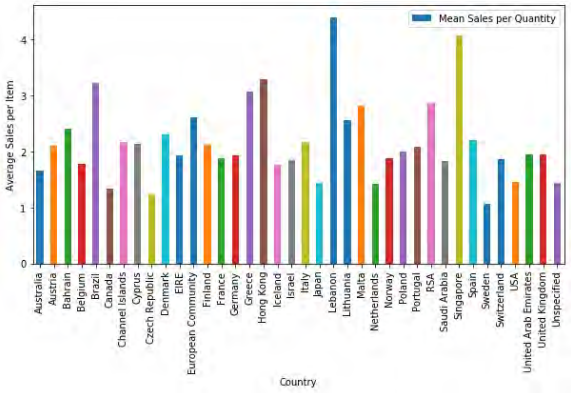
To get to know better the customers, we investigated the variable "CustomerID". We wanted to know who are the top 10 and where do they come from, results are in Table 3.1. The most important customer is from the Netherlands with 2 076 number of transactions and mean check of 134.97 currency units. It can be seen that 6 customers out of the top 10 are from the United Kingdom, second one in the top with the highest number of sales and only 60 transactions, third one is following with 46 number of transactions and fourth out of 10 conducted only two transactions. The highest number of transactions through the whole dataset is 7 983 for the customers from the UK.



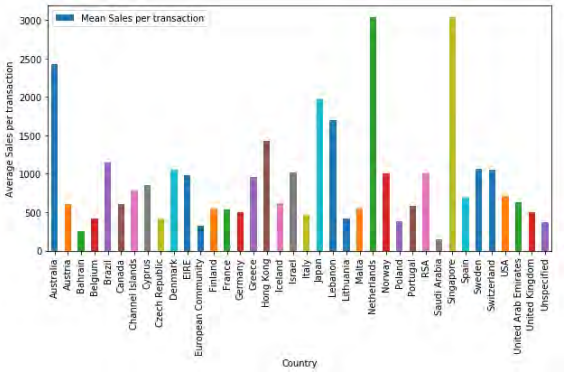
(a) Total quantity of items per country



(b) Average quantity of items per transaction per country



(c) Average sales per item per country



(d) Average sales per transaction per country

Figure 3.4: Country analysis

(a) Time series for the number of transactions for the Netherlands

(b) Time series for the number of transactions for Germany and France

Figure 3.5: Time series

		Sales	*Sales%*	*Country*
	CustomerID			
1	14646	280 206	3.14	Netherlands
2	18102	259 657	2.91	United Kingdom
3	17450	194 550	2.18	United Kingdom
4	16446	168 472	1.89	United Kingdom
5	14911	143 825	1.61	EIRE
6	12415	124 914	1.4	Australia
7	14156	117 379	1.31	EIRE
8	17511	91 062	1.02	United Kingdom
9	16029	81 024	0.91	United Kingdom
10	12346	77 183	0.87	United Kingdom

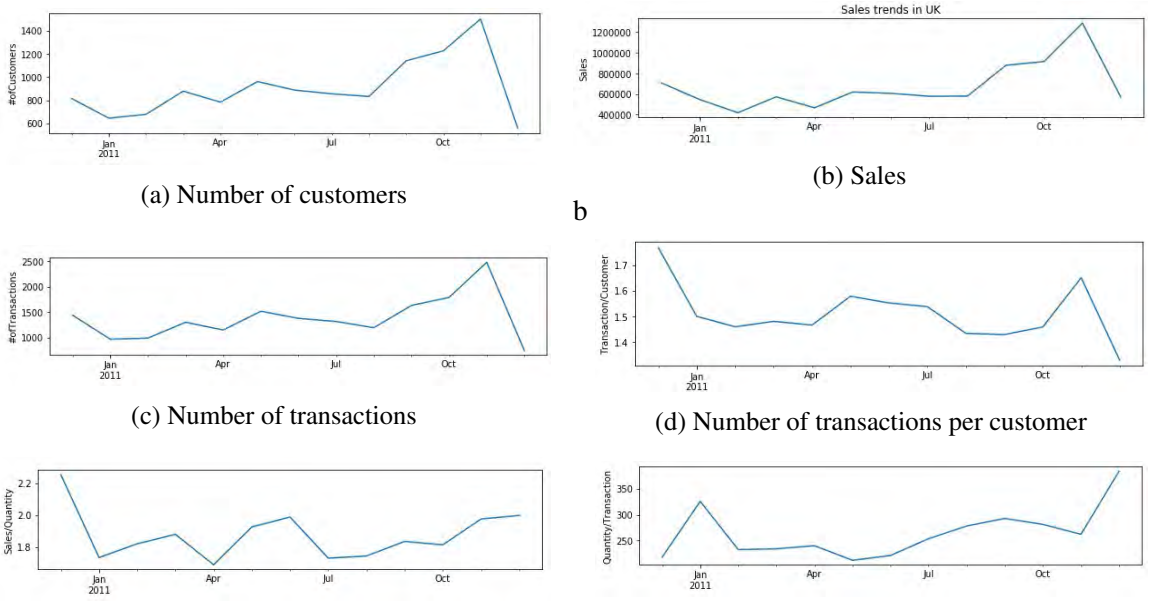
Table 3.1: Top 10 customers with the highest sales

In order to analyze retailer's assortment, we explored purchasing trends by pinpointing the best sellers. We used "UnitPrice" to sort from the most expensive to the cheapest products, mean is 4.45 units. The most expensive one is "Amazon Fee" with 2 records for 13 541.33 currency units, which is obviously a tax not an item. Products with "Postage", "Dotcom postage", "Manual" description were among products with the highest price as well. We looked into each, for example "Postage" had 1 832 records. Considering that it is not a product from the catalog either as "Manual", we excluded them one by one from the dataset. There were more cheaper products excluded, namely "Carriage", "Bank charges". The dataset of items contains 527 806 records, from which products with the highest demand were obtained, results can be observed in Table 3.2. While exploring we compared the number of items with unique records for "StockCode" and "Description" and 220 items with multiple descriptions were found, all of them no more than 2. When we extracted these products, the difference was in punctuations marks or a switch in the order of the words.

	StockCode	*Quantity*	*Description*
1	23843	80995	'PAPER CRAFT, LITTLE BIRDIE'
2	23166	78033	MEDIUM CERAMIC TOP STORAGE JAR
3	22197	56921	SMALL POPCORN HOLDER', 'POPCORN HOLDER'
4	84077	55047	WORLD WAR 2 GLIDERS ASSTD DESIGNS
5	85099B	48474	JUMBO BAG RED RETROSPOT'

Table 3.2: Top 5 products

Based on the literature and data analysis, we decided to choose one market to have better understanding and interpretation of the results. This comes from the concept that for each country, population and mentality approaches and insights should differ, what can work for one country possibly can be a bad idea for another, so generalization cannot take place in the research. Due to the biggest market share, we will focus on the United Kingdom. In addition to previous results, we were interested in the behavior of the market through the time. In Figure 3.6 the time series for the UK are shown for variables: number of customers, sales, number of transactions, number of transactions per customer, sales per quantity in the transaction, quantity per transaction. The time series provide past experience and can help improve predictions, generate models that identify the



(a) Number of customers

(b) Sales

b

(c) Number of transactions

(d) Number of transactions per customer

(e) Sales per quantity

(f) Quantity per transaction

Figure 3.6: UK time series analysis

significant factors. Moreover, this information can be used to develop actions that are crucial for retailer managers, for example knowing when demand is the highest will help to maintain sufficient stock level of the products. We can observe that there are high values for all features in November, so basically before holidays and Christmas break, which make sense that people are buying more in this time frame. From the retailer's perspective, this intention should be saved and stimulated with suitable advertisement campaign, promotion and discount deals. There is a positive trend of growing number of customers, sales and number of transactions.

3.2 Feature engineering

Any type of tasks starts with creating - engineering and selection of the variables. It is worth mentioning that this step of the process is often the most challenging in practice, because of the direct relation to the quality of the algorithms.

The description of created variables is presented in Table 3.3:

1. we define recency as the number of days between the last transaction for each customer and the last transaction date for the dataset 09/12/2011;
2. frequency is calculated as the number of transactions per customer;
3. monetary value is calculated as the sum of the sales per customer.

	Monetary	***Frequency***	***Recency***
mean	1854.99	4	91
min	3.75	1	0
25%	299.34	1	17
50%	650.43	2	50
75%	1571.03	5	143
90%	3407.25	9	261
max	259657.30	206	373

Table 3.3: Description of RFM variables

4. Model implementation

4.1 k-means clustering

During the pre-processing we created the customer-product matrix 4.1. The matrix has 3918 customers-rows and 3641 products-columns. The entries of the matrix are the quantity of the product that the customer has purchased.
The silhouette measure for different number of clusters is presented in Table 4.2. We can see that 2 and 3 clusters have the highest silhouette measure. However, this number of segments is not enough to provide a more personalized experience for the customers. Meaning assigned to 2 or 3 groups is too general and a lot of personal information can be lost. The segmentation by five, six or eight clusters seems to have a clearer interpretation of the dataset than the ones by 4 clusters.

StockCode / *CustomerID*	**10002**	**10080**	**10120**	**10123C**	**10133**	**10135**	**11001**	**15030**
12346	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12747	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12748	1.0	0.0	6.0	0.0	28.0	36.0	32.0	5.0
				...				
12821	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12872	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 4.1: Customer-product matrix

	Number of clusters	*Silhouette score*
1	2	0.9754
2	3	0.9734
3	4	0.9333
4	5	0.9391
5	6	0.9396
6	7	0.9398
7	8	0.9400
8	9	0.9345
9	10	0.8131

Table 4.2: Silhouette score for different number of clusters

4.2 RFM analytics

The number of custom segments depends on the business needs and opportunities. Creating a targeting campaign for a big number of segments can be challenging and costly, the segments should be easy to grasp and distinguish. That is the reason why many times the segments are combined. In our dataset, "Almost Lost" segment represents also "Lost Customers" according to the segmentation in Table 2.1. This decision is based on the fact that total number of "Lost Customers" is 12 and both segments have the same business value and win-back target campaign.

We summarize our customer database in 5 distinct segments from Table 2.1, which is a relatively small number for segmentation. Firstly, 5 clusters is a good fit for our dataset based on the silhouette measure for the k-means algorithm. Secondly, it significantly reduces costs and allows the retailer to do more experimentation. Lastly, while exploring different RFM classes, we did not identify another significant target group. The final division of the customer database is shown in Figure 4.1 below. Each segment is completely different from the other.

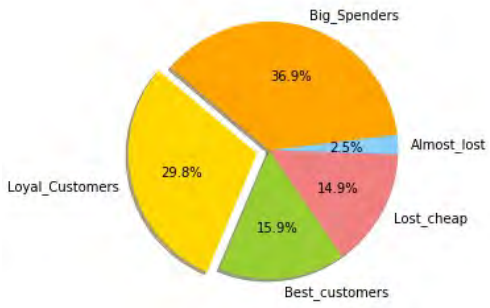


Figure 4.1: Customer segments

Segment	*Number of customers*
Loyal_Customers	789
Best_Customers	422
Almost_Lost	65
Lost_Cheap	396
Big_Spenders	979

Table 4.3: Number of customers in each of the segments

4.3 Logistic regression

The explanatory variables for classification are values from the matrix 4.1. The target variable is categorical variable for each of the RFM segments. In case that the customer belongs to the segment the outcome equals to 1, otherwise 0.

We divided the dataset into 2 sets for training and testing purposes. The training test consists of 75% of the data, which is 2937 instances and testing on unknown data is the remaining 25%

(979 customers). The performance will be evaluated using the accuracy of the model and 10-fold validation. Accuracy is the percentage of correctly classified customers. 10-fold validation assesses how well the model performs in predicting the target variable on different subsets of the data. We partition the dataset into 10 equally sized parts, so-called folds. One fold is held out for validation while the other 9 folds are used to train the model and then used to predict the target variable in our testing subset. The same procedure is repeated 10 times and accuracy for each fold is recorded. The accuracy and mean for all folds for each segment are shown in Table 4.4. We can see that 10-fold shows higher or same accuracy for all segments.

	Best customers	***Loyal Customers***	***Big Spenders***	***Almost lost***	***Lost Cheap***
Accuracy	0.88	0.86	0.91	0.96	0.89
mean for 10-k folds	0.91	0.88	0.92	0.97	0.89

Table 4.4: Evaluation metrics for the Logistic regression

Getting back to the main goal, we want to identify drivers for each class. After obtaining the vector with mean probabilities of all 10 folds for the products "StockCode", we select top 5 values and extract the description from the original dataset. The profit drivers for meaningful segments are presented in Table 4.5. These are the products that customers within the segment are most likely to buy. It is noticeable that products are different for each target group.

StockID	*Top 5 for "Best Customers"*
22265	EASTER DECORATION NATURAL CHICK
22795	SWEETHEART RECIPE BOOK STAND
23284	DOORMAT KEEP CALM AND COME IN
47559B	TEA TIME OVEN GLOVE
85014B	RED RETROSPOT UMBRELLA
	Top 5 for "Loyal Customers"
23091	ZINC HERB GARDEN CONTAINER
22423	REGENCY CAKESTAND 3 TIER
21524	DOORMAT SPOTTY HOME SWEET HOME
21591	COSY HOUR CIGAR BOX MATCHES
23284	DOORMAT KEEP CALM AND COME IN
	Top 5 for "Big Spenders"
48138	DOORMAT UNION FLAG
82484	WOOD BLACK BOARD ANT WHITE FINISH
21158	MOODY GIRL DOOR HANGER
22624	IVORY KITCHEN SCALES
22726	ALARM CLOCK BAKELIKE GREEN
	Top 5 for "Almost Lost"
22926	IVORY GIANT GARDEN THERMOMETER
22598	CHRISTMAS MUSICAL ZINC TREE
22464	HANGING METAL HEART LANTERN
21090	SET/6 COLLAGE PAPER PLATES
79160	HEART SHAPE WIRELESS DOORBELL
	Top 5 for "Lost Cheap"
22335	HEART DECORATION PAINTED ZINC
22419	LIPSTICK PEN RED
21043	APRON MODERN VINTAGE COTTON
21984	PACK OF 12 PINK PAISLEY TISSUES
21198	WHITE HEART CONFETTI IN TUBE

Table 4.5: Top 5 products for each segment

5. Conclusions

The goal of this research was to derive meaningful and actionable customer segments for an online retailer and forecast their future preferences. Ultimately, the objective was to cluster customers in order to understand their perceived value and predict profitable products using regression analysis for each segment. The solution for the segmentation task was provided by the RFM model taking into account the average silhouette measure to determine that five clusters were appropriate for our dataset. The meaning assigned to the optimal number of clusters is the key to customer-centric business intelligence. Based on past purchase behavior and by means of logistic regression we identified top 5 products for each segment. The results can be used for creating a targeting campaign and allow the retailer to provide a more personalized experience, addressing one of the biggest challenges in the market nowadays. The approach can be adapted to increase customer satisfaction and build a loyal relationship.

However, the solution works only for existing customers. This can be changed by extending the model and creating a recommendation system with user-collaborative filtering to predict preferences for the new user. Future work can consist of meeting the aim to discover new patterns in the given dataset. One of the possible research questions can be finding association rules for the products that will help the retailer optimize product assortment, validate promotions and use the data to better understand the spending patterns, communication preferences, and merchandising preferences of customers.

Bibliography

[1] S. B. Patil, "Data mining techniques for customer relationship management in organized retail industry", (cit. on p. 2).

[2] B. M. Ramageri and B. Desai, "Role of data mining in retail sector", *International Journal on Computer Science and Engineering*, vol. 5, no. 1, p. 47, 2013 (cit. on p. 2).

[3] B. H. H. Maskan, "Proposing a model for customer segmentation using wrfm analysis (case study: An isp company)", *Int. J. Econ. Manag. Soc. Sci*, vol. 3, no. 12, pp. 77–80, 2014 (cit. on pp. 4, 6).

[4] J. S. Larson, E. T. Bradlow, and P. S. Fader, "An exploratory look at supermarket shopping paths", *International Journal of research in Marketing*, vol. 22, no. 4, pp. 395–414, 2005 (cit. on p. 4).

[5] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations", in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA., vol. 1, 1967, pp. 281–297 (cit. on p. 4).

[6] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987 (cit. on p. 5).

[7] I.-C. Yeh, K.-J. Yang, and T.-M. Ting, "Knowledge discovery on rfm model using Bernoulli sequence", *Expert Systems with Applications*, vol. 36, no. 3, pp. 5866–5871, 2009 (cit. on p. 6).

[8] J.-T. Wei, S.-Y. Lin, and H.-H. Wu, "A review of the application of rfm model", *African Journal of Business Management*, vol. 4, no. 19, p. 4199, 2010 (cit. on p. 6).

[9] C. C. H. Chan, "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer", *Expert systems with applications*, vol. 34, no. 4, pp. 2754–2762, 2008 (cit. on p. 6).

[10] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via rfm model and rs theory", *Expert systems with applications*, vol. 36, no. 3, pp. 4176–4184, 2009 (cit. on p. 6).

[11] N.-C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers", *Expert systems with applications*, vol. 27, no. 4, pp. 623–633, 2004 (cit. on p. 6).

[12] B. Sohrabi and A. Khanlari, "Customer lifetime value (clv) measurement based on rfm model", *Iranian Accounting & Auditing Review*, vol. 14, no. 47, pp. 7–20, 2007 (cit. on p. 7).

[13] S. F. King, "Citizens as customers: Exploring the future of crm in uk local government", *Government Information Quarterly*, vol. 24, no. 1, pp. 47–63, 2007 (cit. on p. 7).

[14] Y.-M. Li, C.-H. Lin, and C.-Y. Lai, "Identifying influential reviewers for word-of-mouth marketing", *Electronic Commerce Research and Applications*, vol. 9, no. 4, pp. 294–304, 2010 (cit. on p. 7).

[15] S.-T. Li, L.-Y. Shue, and S.-F. Lee, "Business intelligence approach to supporting strategy-making of isp service management", *Expert Systems with Applications*, vol. 35, no. 3, pp. 739–754, 2008 (cit. on p. 7).

[16] J. Miglautsch, "Application of rfm principles: What to do with 1–1–1 customers?", *Journal of Database Marketing & Customer Strategy Management*, vol. 9, no. 4, pp. 319–324, 2002 (cit. on p. 7).

[17] J. A. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression", *Journal of business research*, vol. 60, no. 6, pp. 656–662, 2007 (cit. on p. 7).

[18] H. Hruschka, W. Fettes, and M. Probst, "An empirical comparison of the validity of a neural net based multinomial logit choice model to alternative model specifications", *European Journal of Operational Research*, vol. 159, no. 1, pp. 166–180, 2004 (cit. on p. 7).