

**Master Studies**

**Project II. Segmentation model.**

Field of Study: Advanced Analytics-Big Data Advanced Business analytics

|  |  |
| --- | --- |
| **Student names:**  Oleksandr Bachynsky | **Album nr:**  83457 |
| Oleksandr Romanchenko | 83459 |
| Kelija Vecens | 85949 |

----------------------------------------

Instructor: Dr. Adam Korczyński

----------------------------------------

Warsaw, 2020

1. Introduction

Every single business goal is to be as profitable as possible. If only companies could have a magic power to attract the customers into their stores so that a client by walking in the shop would purchase the products and services customized for their preferences, they would have already done that. However, it doesn't really work that way. In order to have a full picture of the customer, there must be certain systems in place that are capable of precisely gathering and analysing data. Moreover, companies must establish the right approach by applying data accordingly to their needs and be ready to compromise to achieve the best result.

In this digital era most of the companies have the automated systems that provide the deep insight into data by the ability of gathering, storing, reporting and analysing the large amounts of data. Companies are interested in understanding the customer and provide useful services and products. The problem comes with business ability to finding compromise. Ideally, the firm would want to target every single customer individually to meet its preferences and needs, however it is not possible. Therefore, it is crucial to perform customer segmentation. In other words, the company needs to divide customers in certain groups or clusters based on the similarities between customer’s behaviour or the characteristics and then aim to meet the needs of certain cluster. Moreover, once the customers are divided into the clusters, company still can offer customized products or service within the cluster. Segmentation helps the company to get to know their customers and understand their needs. Moreover, it is essential for developing targeted marketing and delivering the right message to its customers. Additionally, it plays a crucial role in developing a differentiated product or service, based on customer needs.

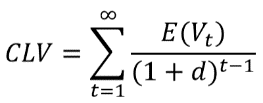
The aim of this project is to perform segmentation analysis and compare the results by using multiple models: such as: Quantile, K-means, Birch, Gaussian Mixture model. In order to achieve it, first we will perform data preparation by variable normalisation, then we will group the variables and do the model analysis. As our dataset, we chose "Online Retail Data Set", a dataset representing the transactional behaviour of online retail store that mainly sells unique all-occasion gifts. We believe that the customers in monetary group (with high amount of spending) will be classified in the best customer group in our model.

1. Theoretical overview

Customer Lifetime Value (CLV)

Customer lifetime value (CLV) is mostly defined as a total net income that company may receive from the customer over the duration of the relationship with the business. Essentially, CLV is the total worth of a customer to a business over the whole period of their relationship. It’s an important metric as it costs less to keep an existing customer than it does to acquire new ones, so increasing the value of your existing customers is a great way to drive growth.

From the practical point of view the CLV is just the discounted sum of future cash flows contributed by the customer and calculated in the following way:



Where:

* 𝑉𝑡 - the customer’s net contribution in period t
* d – the discount rate

The reason why CLV is so essential is because it gives the business information about the customer acquiring costs. These costs come from company trying to acquire or retain the customer by means of promotions, discounts, targeted marketing etc. In general, company’s goal should be to maximise the CLV, therefore the acquiring costs should be lower than customer lifetime value for the business strategy to make sense.

The CLV can be classified in two type of models:

* Gone for good model – assumes that once customer has stopped the relationship with business, he/she will not return
* Always a share model – opposite to the first model, assumes that customer may return in some point in time

Companies invest in customer in order to generate more revenue. Therefore, some customers become more valuable in terms of CLV, however there are always customers that pulls down business profitability. Being said that, the company's task is to identify customer behaviour, segment customer and take appropriate actions.

Segmentation models

All customers have different-different kind of needs. With the increase in customer base and transaction, it is not easy to understand the requirement of each customer. Identifying potential customers can improve the marketing campaign, which ultimately increases the sales.

Customer segmentation is defined as a method of dividing customers into groups or clusters on the basis of common characteristics. By using segmentation the company can target specific clusters of customers with similar expectations by offering products or services customized to cover wants of different individuals within the cluster.

This helps the business to determine appropriate product pricing, develop customized marketing campaigns, choose specific product features for deployment and prioritize new product development efforts. All these actions help the company retain customers and hence, increase the CLV.

There are numerous methods to perform segmentation, varying in rigor, data requirements, and purpose. One of the most widespread is the cluster analysis which is used in this project. Other approaches include:

* **Chi-square Automatic Interaction Detector (CHAID)** is a decision tree classification method that creates nodes or groupings of consumers enabling smaller group analysis (McCarty & Hastak, 2006).
* **Logistic Regression** is a modeling method used on a dichotomous or binary dependent variable (McCarty & Hastak, 2006).
* **Association rule mining (Market basket analysis)** is especially helpful in purchasing behavior segmentation for retail businesses interested in finding items commonly purchased together, and how that may coincide with more typical demographic, psychographic, geographic, or behavioral data (Griva, Bardaki, Pramatari, & Papakiriakopoulos, 2018).

RFM analysis

As mentioned above we will use the cluster analysis approach. Cluster analysis is a method of grouping, or clustering, consumers based on their similarities. There are 5 primary types of cluster analysis leveraged in market segmentation: partitioning methods, hierarchical clustering, fuzzy clustering, density-based clustering and model-based clustering (Miller, 2015). We will use the RFM analysis, a subdivision of partitioning methods, for its simplicity and straightforwardness.

RFM stands for Recency, Frequency, and Monetary value and is a behaviour-based approach to grouping customers into certain segments. The goal of RFM segmentation is to identify new customers and provide a better service for already existing customers.

The customers are grouped based on their past transactions and considering the following questions: how recently, how often, and how much did a customer buy. The RFM parameters are defined as:

* **Recency (R)** answers the question “Who has purchased recently?” and is measured as number of days since last purchase (smallest recency).
* **Frequency (F)** answers the question “Who has purchased frequently?” and is measured as the total number of purchases (highest frequency).
* **Monetary Value (M)** answers the question “Who has a high purchase amount?” and is measured as the total money the customer spent (highest monetary value).

Once the RFM values are calculated from the purchase history, a score is assigned to each of the 3 parameters individually for each customer. A final RFM score is calculated simply by combining individual RFM score numbers.

The key issue is to define a proper approach to assigning a score for each of the parameters. Some of the examples are:

* **Simple fixed ranges**: for this scoring method the reviewer decides what range is ideal for recency, frequency and monetary values. The downturn of such approach is that the score ranges may need frequent adjustments. If you have a recurring payment business, but with different payment terms – monthly, annual etc – the calculations go wrong.
* **Quantiles** approach allows to divide the data into several equal parts. It solves a lot of problems in fixed range method, but does not always account for specifics of data. For example, a value which is closer to group 1 can be assigned to group 2 simply because it falls into that quantile.
* **Cluster analysis** utilises unsupervised machine learning algorithms to divide the data into clusters. It is often used for discovering interesting patterns in data by finding natural groups or clusters in the feature space. There are many clustering algorithms to choose from which means that this approach is the most universal one.

1. Data preparation

As mentioned in the theoretical overview, there are several steps in the RFM analysis that have to be carried out:

1. Calculate the Recency, Frequency, Monetary values for each customer.
2. Segment each parameter using a specific approach .
3. Calculate the overall RFM score.

This section focuses on the first step.

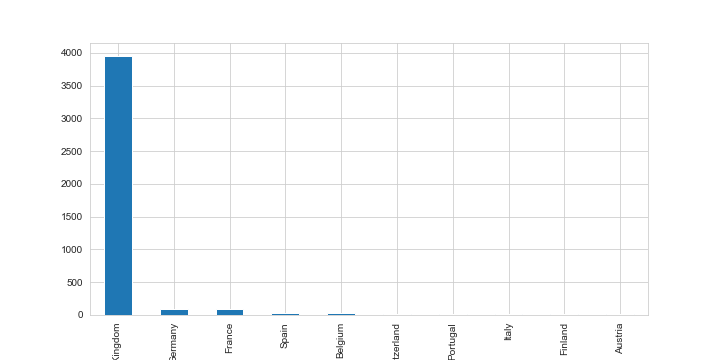
Data overview

The data used for this project is taken from a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

There are 541909 observations in the dataset with the following attributes:

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Data type |
| 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 | Invoice 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. |

Since the data mostly covers UK firms (as shown in the plot below), it is reasonable to filter out the dataset.

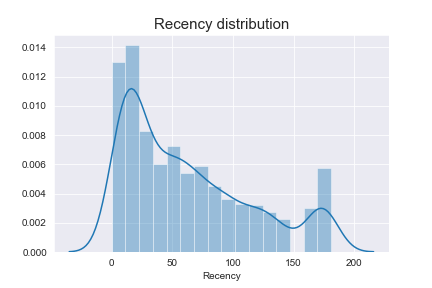


*Figure 3.1. Counties of origin*

Based on the current form of the data, it is hard to determine which RFM values each client has. In order to get this information we need to do the following:

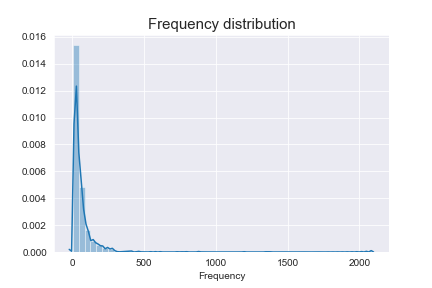
1. Calculate the total purchase as quantity\*unit price
2. Split the dataset into two periods – before and after 1.06.2011 – one for calculating the RFM, the other for model testing.
3. Calculate the Recency, Frequency and Monetary Value as:
   1. Recency = the most recent purchase date for client i – the most recent purchase date available in the dataset
   2. Frequency = count of all invoices for client i
   3. Monetary = Sum of total purchases for client i

The calculation results are shown in the plots below.



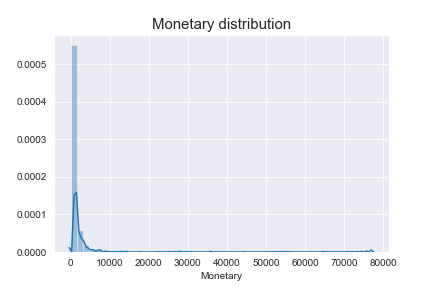
*Figure 3.2. Recency distribution*

The recency distribution shows that the majority of clients made purchased in the last 30 days. However, there is a spike of clients ~180 days before. This spike can be explained by the dates – our last available date is 1.06.2011 and 6 months before we had Christmas which is a time of excessive spending.



*Figure 3.3. Frequency distribution*

Both the frequency and monetary distribution show a similar trend – the majority of clients spend a little and not often but there are those who can afford to spend large sums of money.



*Figure 3.4. Monetary distribution*

1. Segmentation model

After preparing the RFM data, the next step is to segment it into several groups. We will use two approaches mentioned in the section 2 – the quantiles method and the clustering approach.

Normalisation of Variables

The data obtained in section 3 shows that the RFM values have different scales. This can affect the proper division into groups, so it is necessary to normalise the data before conducting cluster analysis.

There are three benefits:

* Normalizing the data to have them on the same scale means each attribute is weighed properly (has the same scale and variance).
* Having non-normalized data would simply disregard the attribute with the smaller range.
* The K-means approach used here is "isotropic" in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. In this situation leaving variances unequal is equivalent to putting more weight on variables with smaller variance.

There are three approaches that can be used:

* Normalisation based on the z-score formula (standardization)
* Normalisation based on min-max values that scales the features values to [0, 1].
* Normalisation based on transforming the values of the attributes in a rank. Such transformation is more robust to outliers compared to the other two normalization methods.

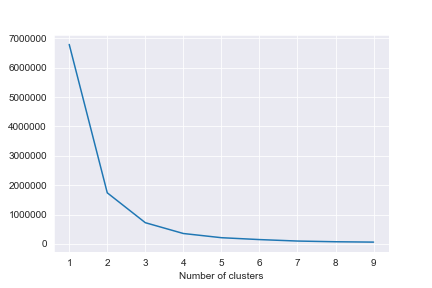
In our case we will use the last approach.

Grouping variables

Once the data is prepared, it is necessary to determine how many groups we will have in each parameter. In general, there are 2 approaches to determining the optimal number of clusters:

* **Direct methods** which consist of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named elbow and silhouette methods.
* **Statistical testing methods** which consists of comparing evidence against the null hypothesis. An example is the gap statistic.

In our case we used the elbow method. Its result is shown below. The graph suggests that 4 clusters would the most optimal. Three tiers can also be used (resulting in 27 segments); however using more than four is not recommended (the difficulty in use outweighs the small benefit gain from the extra granularity).



*Figure 4.1. Elbow method*

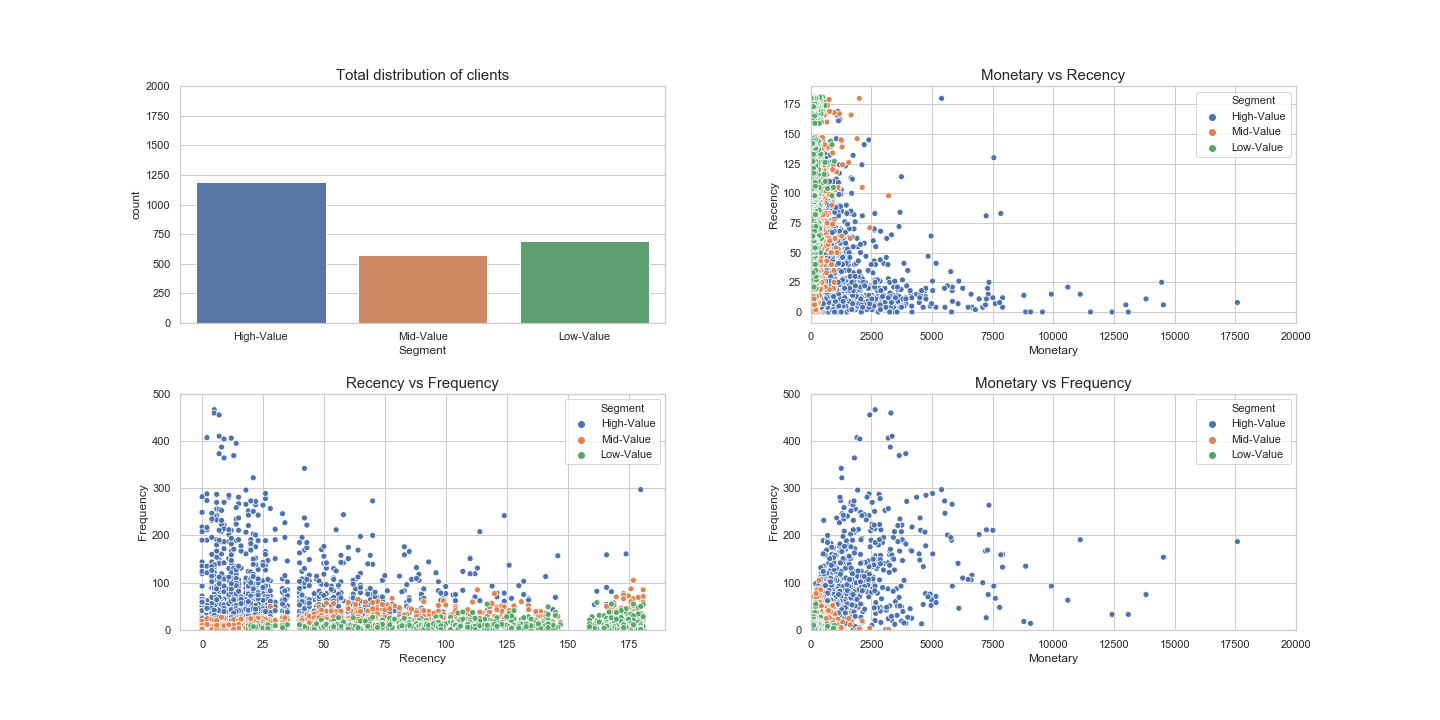
We will assign each cluster a value from 0 to 3, where 3 indicates a client who purchased recently, often or spends large amounts. In total, we will have 64 (4x4x4) different customer segments with overall score ranging from 0 (worst client) to 9(best client).

Based on the overall score, we will distinguish the following groups of clients:

* **Low Value** [0;3): Customers who bought a long time ago, rarely make purchases and generate very low, sometimes negative revenue.
* **Mid Value** [3;5): Customers that purchased recently, are more active, and spend large amounts of money.
* **High Value** [5;9]: Customers that made a transaction recently, do so often and spend more than other customers.
  + 1. Quantile

The first approach that is used is the quantile approach, which we will use as our baseline. As mentioned before, the quantile approach basically divides each parameter into 4 equal groups.

The result of such division is shown in the dashboard below. Overall, we can see that we have a large group of high-value customers. The scatterplots do not show any discrepancies regarding the data – as expected, high-value customers have low recency, high frequency and high monetary values.



*Figure 4.2. Quantile dashboard*

* + 1. Kmeans

K-Means Clustering may be the most widely known clustering algorithm and involves assigning examples to clusters in an effort to minimize the variance within each cluster. The K-means algorithm aims to choose centroids that minimise the **inertia**, or **within-cluster sum-of-squares criterion**:



Inertia can be recognized as a measure of how internally coherent clusters are. It suffers from various drawbacks:

* Inertia makes the assumption that clusters are convex and isotropic, which is not always the case. It responds poorly to elongated clusters, or manifolds with irregular shapes.
* Inertia is not a normalized metric: we just know that lower values are better and zero is optimal. But in very high-dimensional spaces, Euclidean distances tend to become inflated (this is an instance of the so-called “curse of dimensionality”).

It is implemented via the KMeans class in python and the main configuration to tune is the “n\_clusters” hyperparameter set to the estimated number of clusters in the data.



*Figure 4.3. Kmeans dashboard*

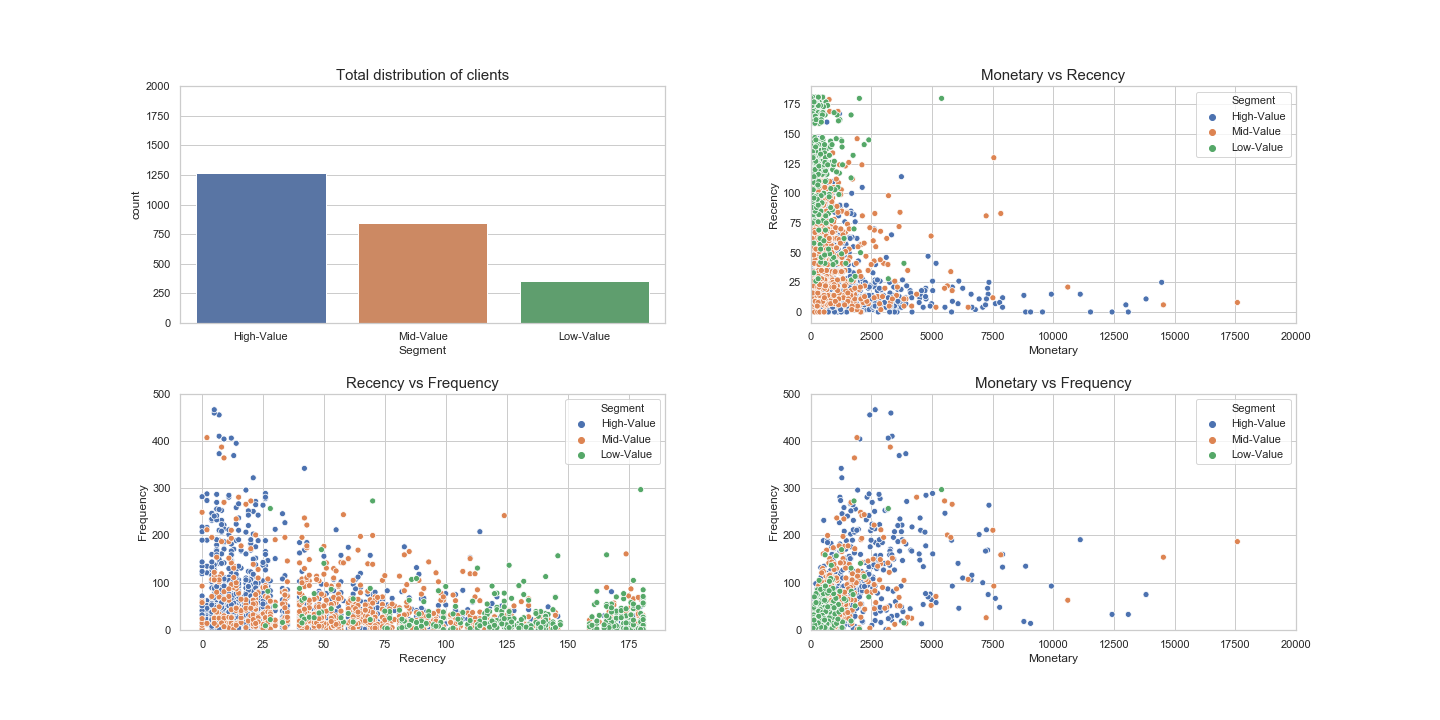
In the kmeans case, we can see that the division between high-value/low-value groups is not as clear as in the quartile case sometimes high-value clients appear among mid-value and even low-value segments. Such discrepancy can be explained by the fact that the scatterplot looks only at 2 parameters at a time, while we have 3. So if a client has low recency and frequency, but a high monetary value, he can still end up in a high-value group. Nevertheless, the pattern still holds.

In general, the number of people with low value has reduced compared to the quantile case.

* + 1. Birch

BIRCH Clustering (BIRCH is short for Balanced Iterative Reducing and Clustering using Hierarchies) involves constructing a CFT tree structure from which cluster centroids are extracted. This algorithm can be viewed as an instance or data reduction method, since it reduces the input data to a set of subclusters which are obtained directly from the leaves of the CFT.

It is implemented via the Birch class and the main configuration to tune is the “threshold” and “n\_clusters” hyperparameters, the latter of which provides an estimate of the number of clusters.



*Figure 4.4. Birch dashboard*

The Birch algorithm generates a similar pattern as the k-means approach. However, the division of groups is more clear – for example, in the “Monetary vs Recency” plot we have few low value groups for recency < 50. In the kmeans case there was a large cluster there.

Overall, in this algorithm we have the smallest low-value group compared to the rest.

* + 1. Gaussian Mixture

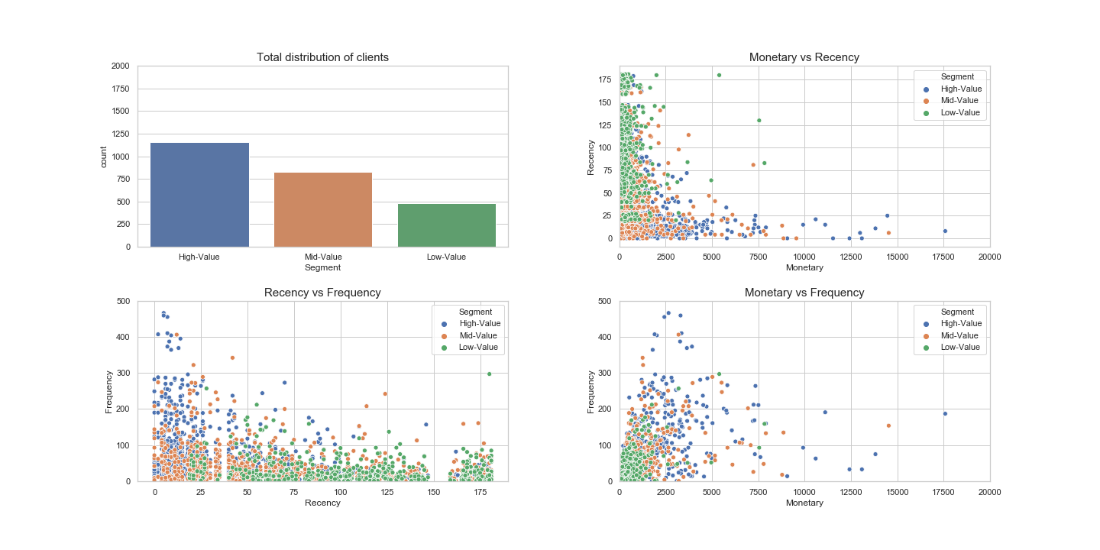
A Gaussian mixture model (GMM) attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset.

Under the hood, a Gaussian mixture model is very similar to k-means: it uses an expectation–maximization approach which qualitatively does the following:

* Choose starting guesses for the location and shape
* Repeat until converged:
  + E-step: for each point, find weights encoding the probability of membership in each cluster
  + M-step: for each cluster, update its location, normalization, and shape based on all data points, making use of the weights

The result of this is that each cluster is associated not with a hard-edged sphere, but with a smooth Gaussian model.

It is implemented via the GaussianMixture class and the main configuration to tune is the “n\_clusters” hyperparameter used to specify the estimated number of clusters in the data.



*Figure 4.5. Quantile dashboard*

The gaussian mixture model demonstrates similar patterns as the kmeans approach. However, the number of high-value clients is lower by 100 compared to the other approaches. This could affect the model effectiveness.

Final results

The algorithms used in the previous section have demonstrated different approaches to clustering the clients. However, it is hard to determine which algorithm is the most effective.

We will use the data from the second half of the year in order to calculate the revenue that each client will bring us. Based on the revenue, we will classify the users into 3 groups and use the XGBoost model to predict the revenue group based on the rfm groups. The accuracy ratio results are presented below.

Overall, we can see that all models are better than a random guess. However, their effectiveness differs – the most effective was the Birch model, while the least effective was the quantile approach.

|  |  |  |
| --- | --- | --- |
| Model | Train dataset | Test dataset |
| Quantile | 0.57 | 0.57 |
| Kmeans | 0.81 | 0.81 |
| Birch | 0.92 | 0.92 |
| Gaussian mixture | 0.58 | 0.60 |

*Figure 4.5. Accuracy ratios for models*

That is why for our model it is advisable to use to use the Birch model.

1. Business interpretation

Marketing approach

In our model we used a simple division of clients into 3 groups. However, for a more effective business strategy it is necessary to take into account the position of each client in the recency, frequency and monetary rating separately. Such granular approach can allow the business to approach each client individually. For example we can divide the clients into the following groups:

* **Best Customers (rfm score: 333)** have the highest score on all 3 parameters. They bought most recently, most often, and are heavy spenders. These customers likely generate a disproportionately high percentage of overall revenues and thus focusing on keeping them happy should be a top priority. Further analysing their individual preferences and affinities will provide additional opportunities as they can become early adopters for new products and will help promote the company.
* **Can’t Lose Them** **(rfm score: ?33)** clients are those, who have a high frequency and monetary score, but low recency. These customers used to visit and purchase quite often, but not anymore. One way to approach them would be to offer them promotions and renewals to encourage another purchase.
* **Potential Loyalists (rfm score: 3?3)** are customers with high recency and monetary value, but low frequency. It is advised to build relationships with these customers by making them feel valued and appreciated. One approach is to offer incentives like membership or loyalty programs to increase their purchases.
* **Lowest-Spending Active Loyal Customers** **(rfm score: 33?)** are customers who purchased recently and with high frequency, but the amounts they spend are low. Marketers should create campaigns for this group that make them feel valued, and incentivize them to increase their spend levels.

Such segmentation can be made even further depending on the clients` importance to the business:

* **Loyalists (?3?)** - Customers who buy the most often.
* **“Whales” (??3)** - Customers who have generated the most revenue.
* **Faithful customers (?31, ?30)** - Customers who return often, but do not spend a lot.
* **Rookies (30?)** - First time buyers.
* **Slipping (00?)** - Great past customers who haven't bought in awhile.

Different business models

In our approach we treated each aspect of the RFM model equally. However, depending on the business model, each aspect can have a different importance on CLV.

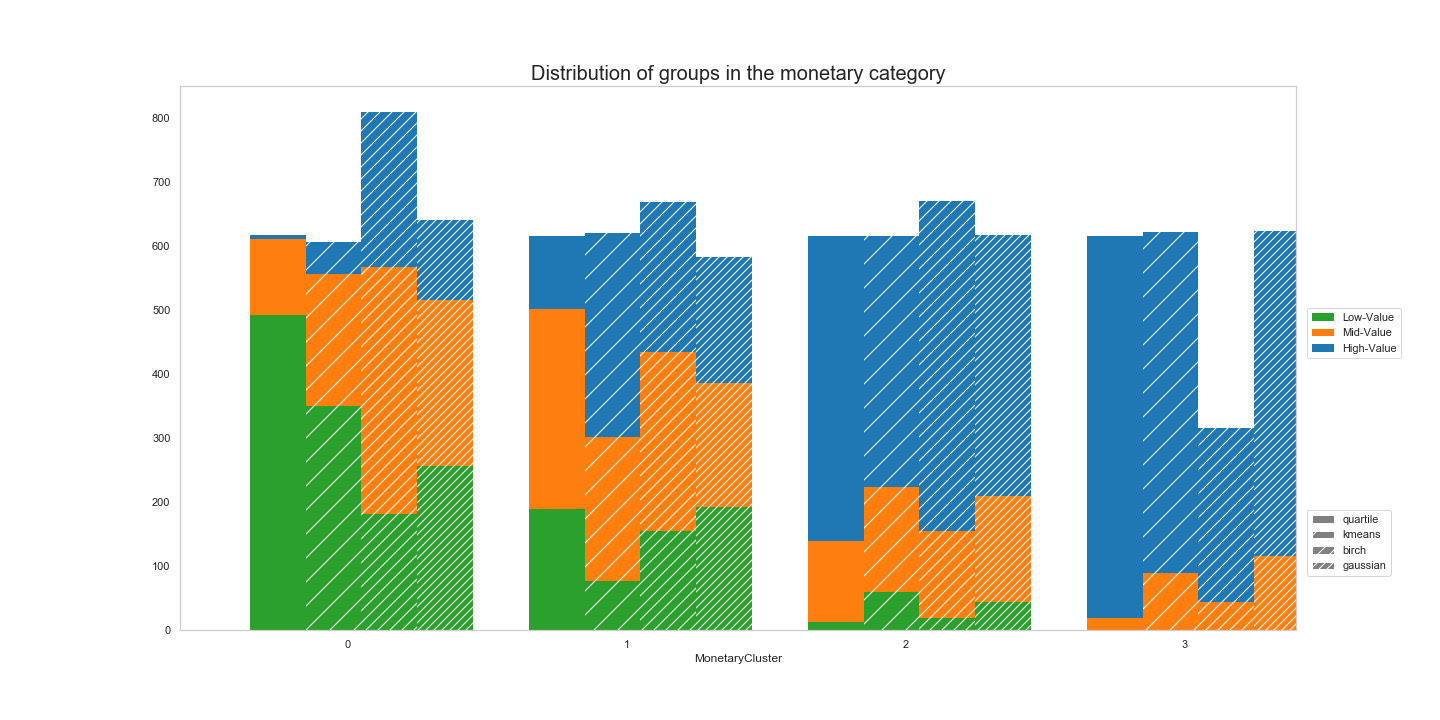
For example, in retail business, consumers typically purchase low-value goods, e.g. clothes. In this case recency and frequency will be more important than the monetary score and the business has to adjust marketing strategies accordingly. On the other hand, when purchasing consumer durables (products that are expected to last a long time, e.g. cars), the monetary value per transaction is normally high but frequency and recency is low. In this case, the marketer should give more weight to monetary and recency aspects rather than the frequency aspect.

1. Conclusions

Looking at the results of all our modules we can sum it up saying that in most cases High-Value customers are those clients who have low recency, high frequency and high monetary. In other words, these are those customers whose total sum of purchases is quite big, they’ve recently made a purchase and they actually purchase a lot.

However, it’s worth to note here that even though high monetary value of customer is almost always a sign that this customer is related to High-Value group (which confirms our initial hypothesis that customers with high amount of spending will be classified as the best customers in our model), it is not a “requirement” to be there, since most of customers from this group have a pretty average monetary value but high frequency and low recency scores. Likewise, a High-Value customer doesn’t necessarily have a huge total sum of his/her purchases.

The best prediction score has been achieved with the BIRCH model showed the result of 92% accuracy which is significantly higher than the second-best result achieved by k-means (81%). Another thing we would like to emphasize here is that the BIRCH model is the only one that classified a decent number of customers with low monetary value as High-Value clients. Therefore, it’s fair to conclude that monetary score shouldn’t be a key factor for decision-making and it’s better to focus on 2 other factors (frequency and recency) that have less variation in the way they affect the value of a customer.



1. References

* <https://www.optimove.com/resources/learning-center/rfm-segmentation>
* <https://towardsdatascience.com/divide-and-conquer-segment-your-customers-using-rfm-analysis-68aee749adf6>
* <https://www.datacamp.com/community/tutorials/introduction-customer-segmentation-python?fbclid=IwAR2lQXiYxMEcrr-Qfd6bua6N7PNMmyff6OP-XlL8du6tqh-k4P1cwdC181U>

1. Code

data = pd.read\_excel("Online Retail.xlsx")

uk\_data=data[data.Country=='United Kingdom']

uk\_data = uk\_data[(uk\_data['Quantity']>0)]

#create tx\_user for assigning clustering

rfm = pd.DataFrame(uk\_data\_3m['CustomerID'].unique())

rfm.columns = ['CustomerID']

#calculate recency score

max\_purchase = uk\_data\_3m.groupby('CustomerID').InvoiceDate.max().reset\_index()

max\_purchase.columns = ['CustomerID','MaxPurchaseDate']

max\_purchase['Recency'] = (max\_purchase['MaxPurchaseDate'].max() - max\_purchase['MaxPurchaseDate']).dt.days

rfm = pd.merge(rfm, max\_purchase[['CustomerID','Recency']], on='CustomerID')

#calcuate frequency score

frequency = uk\_data\_3m.groupby('CustomerID').InvoiceDate.count().reset\_index()

frequency.columns = ['CustomerID','Frequency']

rfm = pd.merge(rfm, frequency, on='CustomerID')

#calcuate revenue score

uk\_data\_3m['Monetary'] = uk\_data\_3m['UnitPrice'] \* uk\_data\_3m['Quantity']

revenue = uk\_data\_3m.groupby('CustomerID').Monetary.sum().reset\_index()

rfm = pd.merge(rfm, revenue, on='CustomerID')

# dataset

quartile = pd.DataFrame(rfm).copy(deep=True)

quartile['RecencyCluster'] = pd.qcut(normalized\_rfm['Recency'], 4, ['3','2','1','0']).astype(int)

quartile['FrequencyCluster'] = pd.qcut(normalized\_rfm['Frequency'], 4, ['0','1','2','3']).astype(int)

quartile['MonetaryCluster'] = pd.qcut(normalized\_rfm['Monetary'], 4, ['0','1','2','3']).astype(int)

#overall scoring

quartile['OverallScore'] = quartile['RecencyCluster'].astype(int) + quartile['FrequencyCluster'].astype(int) + quartile['MonetaryCluster'].astype(int)

quartile['Segment'] = 'Low-Value'

quartile.loc[quartile['OverallScore']>2,'Segment'] = 'Mid-Value'

quartile.loc[quartile['OverallScore']>4,'Segment'] = 'High-Value'