Customer Segmentation

**ABSTRACT**

Since the increased importance is placed on customer equity in today’s business environment, many firms are focusing on the notion of customer loyalty and profitability to increase market share. The main goal of this research is to identify “golden customers' ' who have the most purchases and the least on the verge of churning out. Two approaches are used: in the first approach, a variant of RFM (Recency, Frequency, and Monetary) model is used in order to segment customers and in the second approach, we use data-driven method and more specific clustering method. The input dataset we used is the UK's E-commerce dataset from the UCII repository between 01/12/2009 and 09/12/2011 for machine learning which is based on customers’ purchasing behavior. With the time window being one year, we have two “golden customers'' sets corresponding to two approaches given the dataset in 2010. After all, we used two business metrics in line with desired business outcomes to find out which approach had more accuracy with golden customers given the dataset in 2011.

## **INTRODUCTION**

In the light of data segmentation, customers are divided into a set of individuals with distinct similarities. The major industries wherein customer segmentation and for data mining can be applied are the Retail Industry (Han et al., 2011), because it requires a vast amount of data on sales, transportation, consumption ratio, redelivery service and many others. Also, Retail data mining helps in identifying and effectively mapping customer behavior and related patterns during the entire life-cycle of business transactions. This ultimately leads to improved customer service, effective sales and distribution strategies and many more (Han et al., 2011).

Therefore, this paper focuses on the transaction data of a customer of UK online retail e-commerce. This work mainly focuses on finding the golden customers which have the high revenue, and less ability to churn out based on results of business metrics. By providing the right treatment for golden customers, a company can form an effective and targeted strategy for the future.

The rest of this study is organized as follows. Section 2 reviews related work on customer segmentation methods, traditional RFM model and variant of RFM model which pairing the Recency, Frequency, Monetary scores with the Customer Lifetime Value model to segment customers. Section 3 outlines the new approach of customer segmentation including their benefit and algorithm idea behind. Section 4 describes our research methodology, case study and two approaches which are used to define golden customers. Finally, Section 5 draws conclusions and discussion.

## **LITERATURE**

1. **Customer Segmentation methods**

According to Havard's study, what industry you’re in, acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing one. It makes sense: you don’t have to spend time and resources going out and finding a new client — you just have to keep the one you have happy. If you’re not convinced that retaining customers is so valuable, consider research done by Frederick Reichheld of Bain & Company (the inventor of the net promoter score) that shows increasing customer retention rates by 5% increases profits by 25% to 95%.“Think about the customers you want to serve up front and focus on acquiring the right customers. The goal is to bring in and keep customers who you can provide value to and who are valuable to you”.Therefore, E-commerce organizations may enhance their profits and business by retaining existing customers, especially the golden customers. But identifying these customers, and satisfying their demands of them is a very complex task. This is because customers may be different according to their demands, desires, preferences, and so on. Instead of a “one-size-fits-all” approach, customer segmentation clusters the customers into groups sharing the same properties or behavioral characteristics. Customer segmentation is a strategy of dividing the market into homogeneous groups, from that, identifying the group of golden customers. This allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies and identify customer retention. There are many methods for clustering customers but we consider 2 main methods: The traditional customer segmentation methods( The traditional RFM model, and the variant of the RFM model which pairs the Recency, Frequency, and Monetary scores with the Customer Lifetime Value model to segment customers.); The unsupervised learning method (Machine Learning model algorithm for classification such as PCA, Kmeans, GMM, Hierarchical Clustering,...). We don’t compare the two methods because each one has disadvantages and advantages, so it depends on the purpose of the company and choosing which method is more suitable.

1. **RFM model and its variant (RFM+CLV)**

Recency, Frequency and Monetary (RFM) analysis model is a famous and applied model that uses transaction history to define behavior-based customer segmentation.

*The recency of the last purchase (R)*: This variable is pointing at recency, the interval between the last purchase by the customer to the end of a special period. The shorter interval shows the higher value of this variable in the model.

*Frequency of the purchases (F)*: This variable shows exchanging several transactions. It is the number of transactions that a customer is going through at a special period. A higher number shows a higher variance in this model.

*The monetary value of the purchases (M)*: This variable shows purchase monetary value. It is the amount of money that a customer spent during a special period for purchase.

CLV is one method that organizations use during the customer lifecycle. According to Safari, Safari, and Montazer, (2016, CLV can be used to determine the present value of customers. CLV should be an important construct in designing and budgeting a number of marketing decisions such as customer acquisition programs (Berger and Nasr, 1998). There are two common methods to calculate CLV:

*Historical Approach*

* Aggregate Model: calculating the CLV by using the average revenue per customer based on past transactions. This method gives us a single value for the CLV.

CLV = ((Average Sales x Purchase Frequency) / Churn) x Profit Margin

Average Sales = Total Sales / Total no. of orders

Purchase Frequency = Total no. of orders / Total unique customers

Retention rate = Total no. of orders greater than 1 / Total unique customers

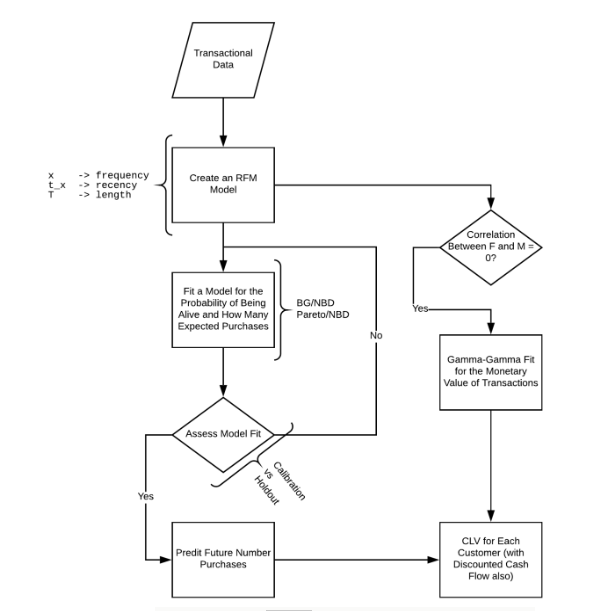
Churn rate = 1 – Retention rate

Profit Margin = Based on business context

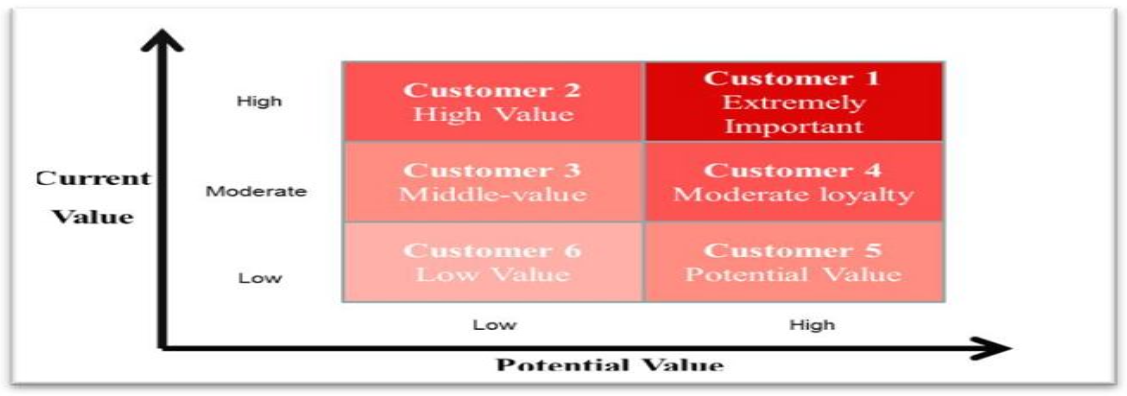
* Cohort Model: Grouping the customers into different cohorts based on the transaction date and calculating the average revenue per cohort. This method gives CLV value for each cohort.

*Predictive Approach*

* Machine Learning Model: using regression techniques to fit past data to predict the CLV . Because we do not typically know the life span of customers, we often try to estimate CLV over the course of a certain period. It can be done by estimating a customer’s 12-month CLV, or 24-month CLV. We can build machine learning models that predict customers’ CLV over the course of a certain period. We are going to learn how to build a regression model that predicts customers’ 3-month CLV.
* Probabilistic Model: it tries to fit a probability distribution to the data and estimates the future count of transactions and monetary value for each transaction. Using the Lifetimes package which has workflow like the figure below.



RFM segmentation focuses on just three variables while there could be others that are critical for a business. Another downside of RFM customer segmentation is it only takes historical data points into consideration, so there is an advanced technique, corporate RFM with CLV to improve performance. In our project, we calculate the RFM score and CLV using the ‘lifetimes’ package, after that, we will corporate RFM scores and CLV value prediction; we call that the variant of RFM. RFM methods segment the dataset into 3 clusterings (High, Mid, Low) and CLV segment into 2 clusterings (High and Low). Then we have 6 clusterings like the figure below. We named that 6 clusterings: 'low value', 'potential value', 'middle value', 'moderate loyalty', 'high value', 'extremely important’ (those are golden customer)



## **III. NEW APPROACH**

1. **What is the benefit of this new approach**
2. **Clustering method (GMM + K-Means)**

**2.1. Gaussian mixture models**

Gaussian mixture models are a kind of probabilistic model that helps to represent normally distributed subpopulations with a complete population. Mixture models learn subpopulation in an automatic manner without knowing the data point. As the assignment of a subpopulation is unknown, these compose unsupervised learning. Gaussian mixture models are extensively utilized in mining data, recognition of patterns, machine learning, and statistical analysis. In several applications, their parameters are detected using maximal likelihood and EM algorithm and are modeled as latent variables.

* *Learning the model*

If the count of components k is known, the EM is the method used frequently for evaluating the mixture model parameters. In frequentist probability theory the models are learned considering maximum likelihood estimation method which seems to increase the likelihood or probability of the pragmatic data specified the model attributes. However, determining the maximal likelihood solution for mixture models by distinguishing the log-likelihood and addressing for 0 is analytically not possible.

EM is a method for estimating the maximum likelihood and is employed to evaluate closed-form expressions to update the model parameters. EM is an iterative technique that possesses suitable property in which the maximal likelihood of the data maximization is certified to come near a local maximum.

* *EM for Gaussian mixture models*

EM for mixture models contains two steps:

* The first step is termed as expectation step or E-step which contains computation of the expectation of the component assignments for each data point provided the model parameters.
* The second step is termed as maximization step or M-step which contains maximization of expectations computed in E-step based on model parameters.

Each step contains the updation of parameter values. The complete iterative process continues until the algorithm converges providing a maximal likelihood estimate. Instinctively, the algorithm works by knowing the component assigned for each data point and makes parameter solving in an easier manner. The expectation step is based on the latter case, while the maximization step is linked to the former case. Thus, by considering consecutive values, the non fixed values are computed in an effective manner.

* *Gaussian mixture model clustering*

A GMM assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with certain parameters. Expectation - maximization (EM) algorithm is used to fit the GMM to the dataset. GMM learns the representation of a multimodal data distribution as a combination of unimodal distributions. GMM assumes the data in a specific cluster are generated by a specific Gaussian distribution/component. GMM fits *K* Gaussian components to the dataset by parameterizing the weight, mean, and covariance of each cluster, where *i* is the cluster number. If there are *K* clusters in the dataset, Gaussian mixture model fits the dataset by optimizing the following sum of Gaussian distributions/components:

A picture containing text

Description automatically generated

*Where:*

* is the data point vector
* is Gaussian distribution
* is the number of clusters
* is the mean of a cluster
* is the covariance matrix
* is the weight/parameter to be learnt by the GMM algorithm. The sum of the weights of all distributions equals to 1.

After fitting the data with multiple Gaussian distributions, the results can be used to cluster any new data point into one of the identified clusters.

**2.2. K-Means**

K-Means clustering is a type of unsupervised learning, which is used when you have unlabeled data. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-Means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically.

* *Algorithm*

The *Κ*-Means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters *Κ* and the data set. The data set is a collection of features for each data point. The algorithms starts with initial estimates for the *Κ*centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

*Step 1*: *Data assignment step*

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if *ci* is the collection of centroids in set *C*, then each data point *x* is assigned to a cluster based on

Text

Description automatically generated with medium confidence

where*dist*( *·*) is the standard (*L*2) Euclidean distance. Let the set of data point assignments for each *ith* cluster centroid be *Si*.

*Step2:Centroid update step*

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

Text

Description automatically generated

The algorithm iterates between steps one and two until a stopping criteria is met. This algorithm is guaranteed to converge to a result. The result may be a local optimum, meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

**3. Number of cluster**

*The elbow method*

The elbow method is used to determine the optimal number of clusters in clustering. The elbow method plots the value of the cost function produced by different values of k (number of clusters). As you know, if k increases, the average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement declines the most is called the elbow, at which we should stop dividing the data into further clusters.

**4. Metrics to evaluate clustering method**

**4.1.Inertia**

The Inertia or within cluster of sum of squares value gives an indication of how coherent the different clusters are. Equation 1 shows the formula for computing the Inertia value.

Text

Description automatically generated

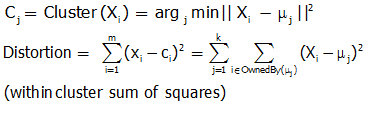
*Where*:

* N is the number of samples within the data set
* C is the center of a cluster.

So the Inertia simply computes the squared distance of each sample in a cluster to its cluster center and sums them up. This process is done for each cluster and all samples within that data set. The smaller the Inertia value, the more coherent are the different clusters. When as many clusters are added as there are samples in the data set, then the Inertia value would be zero

**4.2. Distortion**

It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used



**4.3. Silhouette score**

Letter

Description automatically generated with medium confidence

*Where:*

* is the average distance between the data point i and all the other data points in the cluster to which i belongs
* is the minimum average distance from i to all clusters to which i does not belong

The value of the silhouette coefﬁcient is between [-1, 1]. A score of 1 denotes the best meaning that the data point **o**is very compact within the cluster to which it belongs and far away from the other clusters**.**The worst value is -1. Values near 0 denote overlapping clusters.

**4.4. AIC and BIC**

The definition of AIC (and thus BIC) might differ in the literature. In this section, we give more information regarding the criterion computed in scikit-learn.

The AIC criterion is defined as:

Logo

Description automatically generated

Where:

* is the maximum likelihood of the model
* d is the number of parameters (as well referred to as degrees of freedom in the previous section).

The definition of BIC replace the constant 2 by log(N):

Text

Description automatically generated with medium confidence

Where:

* N is the number of samples.

**5. Feature important**

In Scikit-learn, Gini importance is used to calculate the node impurity and feature importance is basically a reduction in the impurity of a node weighted by the number of samples that are reaching that node form the total number of samples. This is known as node probability. Let us suppose we have a tree with two child nodes, the equation we have is:

Text

Description automatically generated

Where:

* : node j importance
* : weighted number of samples reaching node j
* : the impurity value of node j
* left(j): child node on left of node j
* right(j): child node on right of node j

This equation gives us the importance of a node j which is used to calculate the feature importance for every decision tree. A single feature can be used in the different braches of the tree. Thus, we calculate the feature importance as follow:

Text

Description automatically generated

The features are normalized against the sum of all feature values present in the tree and after dividing it with the total number of trees in our random forest, we get the overall feature importance.

## **IV. CASE STUDY: Customer Segmentation for Online Retail dataset**

**Phase 1: Business Understanding**

Stages of business focus on understanding the purpose of needs based on business valuation. After understanding the business initial data mining plan is designed to reach the goal. And our purpose is finding out what our golden customers are to make righted business strategies based on that thereby avoiding future churn of less responsive customers and giving different attention to potential customers. The study of this paper is of an online retail E-commerce website of a UK retail which consists of sales transactions from December 2009 to December 2011.

**Phase 2: Data Understanding**

This dataset contains 1,067,371 records, and 8 variables with a total of 3 data types: float64 (2), int64 (2), object (5), almost all variables have no null values with the exception of Description and CustomerID, which corresponds to:

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Data Type |
| InvoiceNo | A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation. | Nominal |
| StockCode | Product code. A 5-digit integral number uniquely assigned to each distinct product. | Nominal |
| Description | Product name. | Nominal |
| Quantity | The quantities of each product per transaction | Numeric |
| InvoiceDate | Invoice date and time. The day and time when a transaction was generated. | Numeric |
| Price | Product price per unit in sterling. | Numeric |
| CustomerID | Customer number. A 5-digit integral number uniquely assigned to each customer. | Nominal |
| Country | Country name. The name of the country where a customer resides. | Nominal |

**Phase 3: Data preparing**

*Data Reduction*

This dataset has many bad records that we will discover each variable as we proceed.

* Missing Values

Many machine learning algorithms cannot process missing values. Hence we need to decide whether to drop these observations or to impute them with values. If we drop them, it may reduce the accuracy of the prediction’s output, whereas imputing these observations may increase the bias in our model. Both approaches will give the predicting model adverse impact if not tread carefully. In our dataset, from the first look, two variables have null values (i.e., Description and CustomerID). This problem might happen due to the customer not wanting to reveal their personal information.

There are 247,389 null observations in our dataset, given that the null values from both of features that have null values are independent from each other, which will translate to about 23% of our total dataset.

We have discovered that this dataset may share a fair amount of unusual entries and it can be difficult to detect. For example, one of the strange occurrences we have encountered was that not less than 4382 observations had unit price 0.0. We do not know if it is because of transaction error or customer gifting. So because of it, we decided to handle missing values by dropping the majority of them.

* Outliers

Outliers refer to data points that are significantly different from other datapoints in a dataset. If we left outliers untreated, problems might occur such as the model output will not be accurate; or prolong the training process. In our dataset, variable Quantity has extreme outliers (meaning the values are greater than 3 times its interquartile range), these extreme outliers will be very likely to leave an adverse impact on our model predict we leave it as is, so we use log- transformed Quantity to find and drop these extreme values without reducing dataset information too much.

*Feature Engineering*

* Feature Engineering for RFM + CLV method

|  |  |
| --- | --- |
| Variable Name | Description |
| recency | Time since last order or last engaged with the product |
| frequency | Total number of transactions or average time between transactions/engaged visits |
| monetary | Average transactions value |

* Feature Engineering for clustering method

|  |  |
| --- | --- |
| Variable Name | Description |
| total\_invoice | total number of unique invoices per customer |
| avg\_invoice | total number of unique invoices per customer divided by the total number of unique months in a year that the customer made a transaction. |
| min\_invoice | invoice from the month that have the smallest amount of transaction |
| max\_invoice | invoice from the month that have the largest amount of transaction |
| std\_invoice | standard deviation of invoice from a customer |
| avg\_money | average amount of money spent by a customer per month |
| min\_money | smallest amount of money spent in a month by a customer |
| max\_money | largest amount of money spent in a month by a customer |
| std\_money | standard deviation of money spent by a customer |
| total\_product | total number of products collected from all the invoice generated by the customer |
| avg\_product | total product divided by the number of invoices recency & frequency: explained from above |
| monetary\_value | total amount of money spent |
| mode\_hour | which hour that the customer is most active (meaning made lots of transaction) |
| avg\_product\_month | average product purchase per month |
| avg\_sale | average amount of money spent per purchase |

*Data transformation*

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. One of the purposes of scaling is to equalize the weights of the values ​​of features. These are effective for models based on the distance between features such as K-Means clustering. Numerical values ​​are basically used for features when dealing with machine learning. Numerical values ​​do not have the unit information that we usually handle. Therefore, the model may capture numerical values ​​with different units on the same scale during training. Scaling exists to prevent this.

In this essay we will use the Min-Max scaling method (also known as Normalization). Min-Max scaler is performed by subtracting each value in the feature from the minimum value and dividing by the value obtained by subtracting the minimum value from the maximum value in the feature. One thing to keep in mind when using it is that normalization is sensitive to outliers. If very large or small values ​​are present as outliers, the other values ​​will be pulled to near 0 or 1. Therefore, we have dealt with outliers before using normalization.

**Phase 4: Modeling**

1. ***RFM and CLV***

*First*, we calculated the RFM scores for the 2010 dataset.

* Dividing Recency, Frequency, and Monetary into 5 labels (using ‘qcut’).
* RFM scores = Recency + Frequency + Monetary
* Converted to categorical value : 3 segments (High, Middle, Low)

*Second*, we calculate CLV using the ‘lifetimes’ package.

* Summary data using summary\_data\_from\_transaction\_data function.
* Fitting the BG/NBD model using the BetaGeoFitter function.
* Calculate the customer alive probability
* Predict future transactions for the next 30 days based on historical data using
* Modeling the monetary value using the Gama-Gama model
* Calculate the conditional expected average profit for each customer per transaction
* Checking the expected average value and the actual average value to make sure the values are good.
* Predicting Customer Lifetime Value for the next 30 days.
* Segmenting into 2 clustering (High and Low) by using the result of CLV value prediction above.

*Third*, we corporate RFM score and CLV value prediction.

The RFM score has 3 clusterings (High, Middle, Low) and CLV value prediction has 2 clusterings (High and Low). Then we have 6 clustering and named them by the following and extremely important are golden customers which our customers need to care about

|  |  |  |
| --- | --- | --- |
| RFM segment | CLV segment | Overall |
| Low | Low | Low value |
| Low | High | Potential value |
| Middle | Low | Middle value |
| Middle | High | Moderate loyalty |
| High | Low | High value |
| High | High | Extremely important |

1. ***Clustering***

*At the beginning*, we built a baseline model with the K-Means algorithm using all the features that we have created. We used 3 metrics (distortion, inertia, silhouette score) to evaluate the model. We noticed that the model achieves the best result with 3 clusters, with the elbow point of distortion and inertia at 3 and highest silhouette score of 0.4 (which is still pretty low)

|  |
| --- |
| Chart, line chart  Description automatically generated |
| *K-Means* |
| Chart, line chart  Description automatically generated |
| *GMM* |

We continued to try out Gaussian Mixture Model and used 2 different metrics (BIC and AIC). The result of this algorithm is pretty similar to K-Means, with the elbow point and best silhouette score at 3 clusters. However, when we look at this scatter plot of the dataset after dimensionality reduction (T-SNE) with each color representing a cluster of customers, the data points are separated better than using K-Means. Gaussian Mixture Model seems to be more robust than K-Means because it finds the distribution of the data, furthermore, by default, Gaussian Mixture Model in sklearn is initialized with K-Means, so it still has some benefits of K-Means. For these reasons, we decided to use a Gaussian Mixture Model.

|  |  |
| --- | --- |
| *K-Means* | *GMM* |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |

*Then***,** we used the labels of the clustering model as the classes for a classification problem of the same dataset and fitted it with a Random Forest model. After that, we derived the feature importance of the model to find out the variables that have the most contribution to the model.

Chart, waterfall chart

Description automatically generated

*After that*, we removed the features that have low importance so our data set has 8 columns left (total\_invoice, avg\_invoice, min\_invoice, max\_invoice, std\_invoice, std\_money, frequency, monetary\_value). We continued to fit new the dataset with Gaussian Mixture Model and achieved the following results:

Chart, line chart

Description automatically generated

This time the model has a significantly higher silhouette score (around 0.65) at 3 and 5 clusters. We did silhouette analysis to choose between 3 and 5 clusters. For both 3 and 5 clusters, all the points are above the average silhouette score. The silhouette scores and thickness of the silhouette plots when n\_clusters=3 are very consistent but when n\_clusters=5, the silhouette plots for cluster 0 and 1 are so much thinner compared to others so this is probably not a good choice. As a result, 3 can be considered as the best n\_clusters.

Chart

Description automatically generated

*Analysis to identify Golden customers*

Chart, radar chart

Description automatically generated

After fitting the data into 3 clusters with Gaussian Mixture Model, we drew these radar charts. It can clearly be seen that the customers in cluster number 0 are the most important because on average, they have the highest number of invoices, money spent and purchase frequency.

**Phase 5: Business metric to evaluate**

Our purpose is to find out the golden customers who are potential customers and less possible responsive churn out. And in this part, we will define which method is better. Then we choose two metrics that are the most consistent for our purpose:

*Metric 1:* In the top 10% of customers with the highest monetary value in 2011, how many customers are golden of the two methods?

*Metric 2:* Among customers who buy for 6 continuously consecutive months in 2011, how many percent of the golden customers of the two methods?

We defined the golden customers of GMM methods as cluster 0; the golden customers of RFM and CLV methods are an ‘extremely important ’ segment.

RESULT

|  |  |  |
| --- | --- | --- |
|  | GMM | RFM + CLV |
| Metric 1 | 351/413 (84%) | 317/413 (74%) |
| Metric 2 | 383/369 (91%) | 282/369 (76%) |

With metric 1, in 413 customers with the highest monetary value in 2011, cluster 0 of GMM method accounts 351 people corresponding to 84% while the extremely important group of RFM and CLV just have 317 customers corresponding to 74%. In 369 customers who bought for 6 continuous months in 2011 of metric 2, the golden customer of GMM accounts for up to 91% (383 customers) while the percentage of traditional method is 76% Looking at the result above, one can easily say that Unsupervised Learning (GMM) is a much better option over RFM and CLV for clustering customers on the given dataset and desired business outcomes.