Cluster Analysis for Customer Segmentation with Open Banking Data

Catja Bartels (Initials: C. Bartels)
University of Edinburgh, United Kingdom
catja.bartels@outlook.com

ABSTRACT

Segmenting customers into different groups using their characteristics and behaviors has always been an important topic. Customer segmentation can lead to better customer understanding and targeting, which in turn leads to more effective product tailoring and marketing strategies. Data mining methods are powerful techniques that can be used in customer segmentation to find customers with similar characteristics. Past research that evaluated different data mining techniques has often had drawbacks, such as using too time-consuming methods or conducting studies with smaller data sets. Density-based clustering algorithms for customer segmentation, such as the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) has only been examined by a few research papers. This study, which summarized the main findings of the unpublished dissertation of Bartels [2021], aimed to classify the segmentation of customers using a Recency, Frequency and Monetary Value (RFM) Model and the clustering techniques, K-Means and DBSCAN, to find groups of similarities and differences and to discover potential valuable and vulnerable customers. The data used was from Open Banking data sets, including anonymized transactions from various bank customers in the UK for three months in 2017 and mainly focused on different types of expenses. K-Means found three clusters each month that represent the most, medium, and least valuable customers. While the most valuable customers have the highest average values per attribute, the least valuable customers have the lowest average values. The found clusters were analyzed and evaluated to find potential vulnerable and valuable groups, which can help with future product tailoring and marketing, especially for unforeseen emergency circumstances such as a pandemic. K-Means outperformed DBSCAN, as the latter showed negative silhouette coefficients.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Learning paradigms**; • **Unsupervised learning**; • **Cluster analysis**;

KEYWORDS

Cluster Analysis, K-Means, DBSCAN, RFM Model, Open Banking

ASSE' 22, February 24–26, 2022, Macau, Macao

ACM ISBN 978-1-4503-8745-3/22/02…$15.00
https://doi.org/10.1145/3523181.3523194

ACM Reference Format:
Catja Bartels (Initials: C. Bartels). 2022. Cluster Analysis for Customer Segmentation with Open Banking Data. In *2022 3rd Asia Service Sciences and Software Engineering Conference (ASSE) (ASSE' 22), February 24–26, 2022, Macau, Macao.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3523181.3523194

1 INTRODUCTION

Segmenting customers into various groups depending on their characteristics and behaviors is an essential marketing strategy for any industry. In the financial sector in particular, customer segmentation can be used to target different customer groups with specific marketing strategies [30] and to find potential valuable or vulnerable groups. Some customer groups are more vulnerable to sudden unforeseen emergency circumstances, such as a financial crisis, sudden unemployment, or a pandemic. Research has found that the COVID-19 pandemic, for example, has led to a severe decrease in the spending of customers, and that there has been a significant effect on the labor market in the UK, with suggestions of a potential rise in the unemployment rate after the pandemic [3, 7].

The purpose of the study conducted by Bartels [2021] was to classify the segmentation of customers based on their value and other characteristics, in order to identify similarities and differences in and between segments. Finding common patterns and behaviors (for example, in spending and saving), identifying potential risk or vulnerable groups and value groups can lead to enhanced customer understanding and product tailoring and marketing. The data used is accessed through the Open Banking model in the UK. This model enables bank customers to share their transaction data with registered third-party companies, which can be used to analyze customers for a better market understanding and product tailoring [23].

In the past, extensive research has been undertaken on various data mining techniques for customer segmentation. However, many of these techniques have experienced drawbacks. Neural Networks or Artificial Neural Networks represent a group of connected points that function as input and output, with each connection having a related weight [13]. One of its disadvantages is its long training times, leading to inconsistencies in the output [5, 13]. A Decision Tree is a flowchart-like tree structure consisting of nodes, branches, and tree leaves, with the nodes indicating tests on attributes, branches indicating the test results, and the leaves indicating classes or distributions of classes [13, 31]. But possible disadvantages of this model might be the risk of having too large decision trees due to too many data points, leading to an imprecise creation of relationships and a reduction in the classification accuracy rate [5, 16]. Hierarchical clustering techniques break down a given data set hierarchically

into clusters, with approaches being either agglomerative or divisive [13, 21]. One of the most significant disadvantages of these techniques is their possible high computational cost when dealing with large data sets [21].

Past research that used the RFM Model and K-Means often analyzed smaller data sets when exploring customer segmentation. Similarly, studies using other methods, such as Khalili-Damghani, Abdi and Abolmakarem [2018], who used a hybrid soft computing approach, or Sun, Zuo, Liang, Ming, Chen and Qiu [2021], who used Gaussian Peak Heuristic-based Clustering, did not examine high-dimensional data. There were only a few studies that used a density-based clustering approach such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for evaluating segmentation. Moreover, much of the research in the finance sector did not consider the categorization of transactions. As customer transaction data, such as from the Open Banking model, is not accessible to anyone, the research on the topic is limited. Thus, this study could contribute to the field by using clustering algorithms on large data sets, using the categorization of transactions [4].

The study shows how Bartels [2021] first used the RFM Model to analyze different types of expenses and income in detail. Based on the results, the clustering algorithms K-Means, a popular partitioning clustering algorithm, and DBSCAN, a density-based clustering algorithm, were then used for the customer segmentation. In Section 2 of this paper an overview then describes related work, in Section 3, the proposed methods are presented. Section 4 describes the used data, and Section 5 discusses the results of the procedures. Finally, in Section 6, a conclusion is drawn.

2 RELATED WORKS

The following section reviews past related studies of customer segmentation, the RFM Model, and different customer segmentation methodologies.

2.1 Customer Segmentation

"Customer Segmentation" was introduced as a term to the marketing community in 1956 by Wendell Smith (as cited in Weinstein [2004]). This concept, which is crucial for customer targeting, is achieved by grouping all customers into smaller segments which share similar traits or characteristics, as explained by Weinstein [2004]. The customers in the segments, therefore, tend to share similar purchasing behaviors [17, 30]. The precise usage of customer segmentation in company marketing strategies is essential for successful marketing activities, according to Smith [1956], as this can improve not only a company's position with the competition but also increase company sales, market share, and recognition [30]. Furthermore, other advantages of using customer segmentation are listed as being able to use target groups for product planning and designing and adjusting marketing campaigns accordingly [18, 30]. Customer segmentation not only helps to understand existing marketing strategies but can also help to discover potential new business sectors [18, 30]. There are a number of approaches to segmenting customers that have been used in the past.

2.2 Customer Lifetime Value

"Customer Lifetime Value" can be defined as a way to measure the existing relationship with a customer and is calculated as the current value of the future cash flows assigned to the customer relationship [11]. It should be the objective of companies to improve this value as much as possible [11]. The calculation of the customer lifetime value can be challenging as customers might buy at different prices and at different times [26]. The Customer Value Analysis is a technique to learn the characteristics of a customer [5]. One possible technique that is often used to estimate customer lifetime value and customer loyalty is the RFM Model [5, 6, 10].

2.3 RFM Model

The Recency, Frequency, and Monetary (RFM) Model is a technique to analyze customer value [15]. This technique can help predict the likelihood of a customer's value for companies and to find the customers who purchased most recently, most frequently, and who spent the most, and it can be used to determine the relationship strength of a company and its customers [15, 24]. The attributes of the RFM Model are as follows [15, 26]:

- "Recency" defines how recently a customer has purchased from a company.
- "Frequency" is the number of times the customer purchased from a company within a specific time period.
- "Monetary" refers to monetary value and shows the total spending within a specific time interval.

The most recent and frequent purchasers were suggested to be the most loyal customers who respond best to marketing [15]. However, a low monetary value can indicate a new customer, and as it takes time for customers to reach the highest quantile, this attribute should therefore be regarded with caution [15]. The RFM Model is also a very effective method for customer segmentation that can be used for a database [15, 22, 32]. Furthermore, there are different opinions on whether the three variables should receive different weights and therefore different levels of importance, but it is said that it depends on the industry [15, 19, 26, 29]. There were no weights used in the study of Bartels [2021].

2.4 Clustering Algorithms for Customer Segmentation

2.4.1 K-Means. K-Means is a partitioning cluster method that was introduced by Forgy [1965] (as cited in Cheng and Chen [2009] and Han, Pei and Kamber [2011]). This technique allocates data points to the clusters with the nearest centroid, which stands for the mean value of all data points within the cluster [13, 20]. The advantages of K-Means are that the technique is relatively measurable and can efficiently operate with big data sets [13]. Problems of the method can be its sensitivity to outliers and noise as they can significantly influence the mean of the clusters and that the algorithm can usually only work with numerical values [13]. This clustering algorithm has often been used in past research, as for example, together with the RFM Model and rough set theory to classify customer value segmentation on company data of the electronic industry in Taiwan [5].

2.4.2 DBSCAN. DBSCAN is a density-based cluster algorithm based on connected regions with high density and was first introduced by Ester et al. in 1996 (as cited in Maimon and Rokach [2010]). This technique works by measuring the density (number of

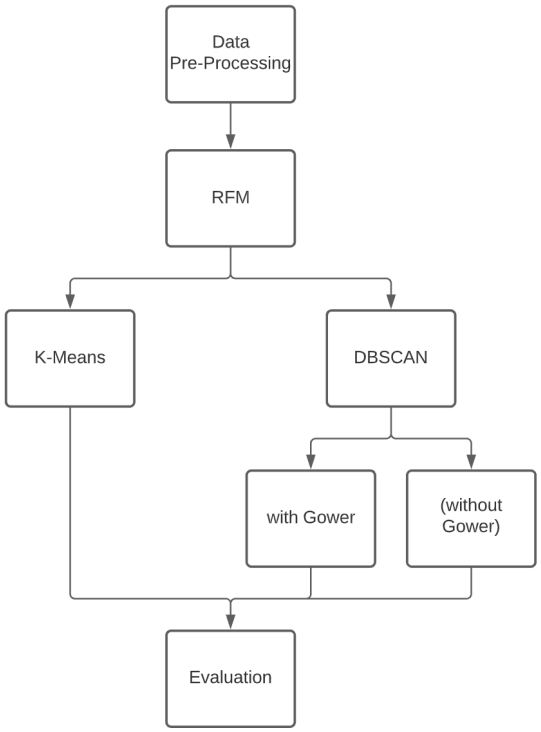

Figure 1: Process of Methodology (simplified, based on Bartels [2021])

data points) of a neighborhood (the area around the cluster) of the cluster and combining data points and their neighborhoods to form dense clusters [13]. Advantages of this cluster algorithm are the ability to find clusters of arbitrary shape and help remove outliers and noise [13, 21]. Past works, even though conducted with rather small data sets, have already shown successful usage of DBSCAN in customer segmentation with examples being Hossain [2017] and Wang, Zhou, Yang, Yang, Ji, Wang, Chen and Zheng [2020]

3 METHODOLOGY

The overview of the most important conducted methodology is shown in Figure 1 [4]. The first step (data pre-processing) consisted of removing missing values and choosing the attributes for the RFM Model and cluster analysis. In the second step, the attributes for the RFM Model were calculated and labeled. The third step consisted of conducting the two cluster algorithms K-Means and DBSCAN. Originally, DBSCAN was conducted using the Gower distance and the Euclidean distance. In the final step, the results of cluster algorithms were evaluated and compared. Python was used to implement all processes [4].

3.1 RFM Model

In the study by Bartels [2021], the RFM Model was used to analyze customer attributes leading to better customer differentiation and to obtain the first ideas for valuable or vulnerable customer groups and their spending patterns and behaviors. The calculation of the chosen attributes for the RFM Model per customer is as follows [4]:

- Recency represents the difference between the last transaction date and the set date.
- Frequency stands for the number of transactions within the examined time period.
- Monetary value refers to the sum of all transactions within the examined time period.

For each examined month, the set date was the first day of the following month. Therefore, the highest number of recency possible was 30 days. For each of the RFM attributes, the data was sorted in descending order and then divided into equal-sized quintiles, as suggested by Hughes [1994]. Each quintile represents around 20% of the data. The labels were given according to the following table (Table 1) with the quintiles 1, 2, 3, 4, and 5 representing a very low, low, medium, high, and very high value, respectively, for recency, frequency, and monetary value [4].

3.2 K-Means

The K-Means algorithm works as follows [2, 13]:

- Knumber of initial centroids or centers of each cluster (μ_k), which represents the mean of all data points of the cluster, is chosen. The calculation is as follows, with N_k representing the number of data points allocated to cluster k:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

- Iteratively, the following steps are conducted:
- Each data point is assigned to its nearest center and a new centroid based on all assigned points is created.
- The difference between old and new centers $i \cdot (x_{i1}, x_{i2}, \ldots x_{ip})$ and $j \cdot (x_{j1}, x_{j2}, \ldots x_{jp})$ is then calculated with the Euclidean distance, with p representing the numeric attributes of the centers:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

- The steps a.and b. are repeated until the difference is less than a set threshold, and from this point the centers of the clusters do not change significantly.

In general, the K-Means clustering algorithm aims to minimize the within-cluster-sum-of-square criterion or inertia.

3.3 DBSCAN

The DBSCAN algorithm is explained in depth as follows [9, 13]: Firstly, DBSCAN selects a data point p randomly and examines the number of *minPts* (parameter stating the minimum density threshold of dense regions) objects in the neighborhood (within the radius ϵ, which defines the neighborhood). If p has a greater number of *minPts* objects than the set threshold, p is considered a *core object*, and a cluster is generated for p. All neighbors of p within the radius ϵ are added to the same cluster and are called *direct density reachable*. If they are, however, also *core objects*, then they are identified as *density reachable*. If they are not, then they are called *border points*, which are called *density connected*. Points are considered *noise* when they are not *density reachable* from any surrounding core point and are not part of the cluster. DBSCAN checks any object in the neighborhood once to decide if it can be added to the cluster or not.

Table 1: Labeling for RFM Model [4]

Quintile	Recency	Frequency	Monetary Value
1	Very Low (least recent)	Very Low (least frequent)	Very Low (least amount spent)
2	Low	Low	Low
3	Medium	Medium	Medium
4	High	High	High
5	Very High (most recent)	Very High (most frequent)	Very High (highest amount spent)

When every data point has been controlled, the cluster building is done. The algorithm then takes a new random data point as a *core object* from the data points that have not been evaluated until then. In the study by Bartels [2021], the number of minimum *minPts* objects and the radius ϵ was chosen as 100 and 0.3 (as 0.3 is commonly used, and 100 was manually chosen). For the evaluation of the performance, the silhouette coefficient was used. In the study by Bartels [2021], DBSCAN was additionally conducted with the Gower Distance; but, as there were no valid or significant results, it will not be further discussed.

4 DATA

The data sets used by Bartels [2021] were provided through Open Banking, a model that was introduced to customers in the UK in 2018 and enabled registered third-party companies to access account and transaction information of small and medium-sized enterprises and customers based on their consent [23]. The system started to provide a fairer competition and to promote innovation in the market of Personal Current Accounts (PCA) in the UK, and the data is accessed and shared through secure Application Program Interfaces (APIs) [8]. Before, new fintech companies and financial institutions had more difficulty accessing and growing in the retail banking market in the UK than older, larger banks [23]. With the first introduction in 2018, the nine largest banks and building societies were all instructed to enable their customers to share their account information; later on, around 40 other companies joined [8]. The advantages of Open Banking are the enabling of more competition, which encourages companies to better tailor products to customer needs and behaviors, more availability of products that combine bank accounts (for example, personal finance apps) [1, 23], and finally, the study of Bartels [2021] was able to use the customer transaction data.

From the available data, two data sets were used. The first data set (table "customer transactions", example shown in Appendix A.1) included the monthly aggregated transaction amounts per category. The attributes in this data set were the user ID, the start date of the month, saving capacity (which is the difference between monthly income and monthly expenses, a negative saving capacity can be seen as a debt), basic expenses, discretionary expenses, luxury expenses, recurrent income, and total income. The second data set (table "transactions", example shown in Appendix A.2) included all unprocessed individual transactions for each customer. The initial 15 attributes of the data set were: category, subcategory, transaction type, account id, account provider, account type, amount, company id, "credit-debit" (transaction direction, income is set as a credit transaction and expense is set as a debit transaction), merchant business line, description (general description of transactions), provider category (category of the provider who provides the data entry), transaction date, transaction id and user id.

The two transaction types (Table 2) "income", and "expense", stand for the following: "Income" is any cash that is paid into the account, categorized into recurring or not recurring, and divided into six subcategories. The subcategories are, for example, salary for recurrent income and expense refund for nonrecurrent income. The Transaction Type "expense" refers to any money that leaves the account, categorized into basic, discretionary, and luxury, and into 22 subcategories. The subcategories include, for example, for basic expenses: groceries and housing; discretionary expenses can be food, drinks, and entertainment; and luxury expenses can be holidays and luxury products.

From the raw data sets, only a few attributes were used for the RFM Model and the clustering algorithms. The data extracted focused on the months January, July, and December 2017. Per customers, RFM attributes were calculated for basic, discretionary and luxury expenses, as well as income and total expenses (basic, discretionary, and luxury combined), as shown in Table 3 [4].

The first data set (table "customer transactions") included 3291143 rows. The table "transactions" for January, July, and December consisted of 7337408, 9003008, and 10288405 rows, respectively. Example extracts are included in Appendix A.1 and A.2. For the cluster algorithms K-Means and DBSCAN, only the numeric values of the RFM attributes of basic, discretionary, and luxury expenses were used [4]. Invalid data entries (for example, customers with missing values for income or expenses) were excluded from further analysis.

5 FINDINGS

The following section discusses the findings of the study. Through descriptive statistics, the study by Bartels [2021] found that around 70% of customers had no savings in January, and was similar in July and December with around 68% and 62%, respectively.

5.1 Findings of RFM Model

The average frequency, recency, and monetary value per month seem quite similar, except for December, where the average recencies of income and expenses are slightly higher than July and January. This suggests that fewer people spend money during the last days of December than in July and January. To consider the thresholds of the labels for the RFM attributes, Table 4 shows the thresholds for the expenses of January. To belong to the label "Very Low", the frequency and monetary value of basic, discretionary,

Table 2: Simplified Categorization of Transaction Types Based on Bartels [2021]

Transaction Type	Category	Example Subcategory
Income	Recurrent	Salary, benefits, interest income
	Nonrecurrent	Expense refund
Expenses	Basic	Groceries, housing, utilities
	Discretionary	Food & drink, entertainment, products, services, cash
	Luxury	Holidays, luxury products, luxury services

Table 3: Used Customer Attributes for RFM Model, Based on [4]

Customer Attributes Used for First Analysis	
Recency	Basic Expenses
Frequency	Discretionary Expenses
Monetary Value	Luxury Expenses
	Income
	Expenses (basic, discretionary, and luxury)

Table 4: Thresholds for RFM-Labels for Detailed Expenses for January 2017 [4]

Label	Recency	Frequency Basic Expenses	Monetary Value Basic Expenses	Frequency Discretionary Expenses	Monetary Value Discretionary Expenses	Frequency Luxury Expenses	Monetary Value Luxury Expenses
Very Low	24–30	0	0	0	0	0	0
Low	18–23	9	304.78	14	892.35	1	29
Medium	12–17	17	694.16	26	1623.73	539	27104.61
High	6–11	26	1279.17	38	2641.87	/	/
Very High	0– 5	39	2456.76	54	4932.16	/	/

and luxury expenses must be under 9, 304.78, 14, 892.35, 1, and 29, respectively. In contrast, the label "Very High" for frequency and monetary value of basic and discretionary are set at a threshold of 39, 2456.76, 54, and 4932.16, respectively. Regarding luxury expenses, the thresholds for the highest group are 539, and 27104.61 for frequency and monetary value, respectively [4]. The thresholds for all recencies were chosen manually, as there were only 30 days available. As there were not many data points containing luxury expenses, only three groups were created (as seen in Table 4). The thresholds for the labeling of December and July do, in the majority, not differ significantly from those of January [4].

5.2 Findings of K-Means

For all three months, three was chosen as the number of clusters, based on the results of the silhouette coefficients. The smallest cluster is cluster 0, with 9714 data points. This cluster can also be called "least valuable group", as it had the lowest means for recency, frequency, and monetary value for basic and discretionary expenses with 5.834877, 10.365864, 1381.259174, 3.056825, 16.714639, and 3068.152179, respectively. Regarding luxury expenses, the averages for recency, frequency, and monetary value were slightly higher (lower for recency) than of cluster 2 (with around 1, 0.07, and 2.90 more, respectively) [4].

Cluster 2 is the largest cluster, with 42263 data points, and can be called the "medium valuable group". It includes the lowest means for the recency of basic and discretionary expenses with 0.542437 and 0.306911, respectively. Regarding luxury expenses, cluster 2 presents the lowest mean in frequency and monetary value of 0.112249, and 2.676574 respectively, and the highest average in recency with 29.446608. For the frequency and monetary value of basic and discretionary expenses, cluster 2 is in between clusters 0 and 1 (with around 3.37, 453.16, 6.25, and 709.84 less than cluster 1, respectively) [4].

The last cluster is cluster 1 and consists of 12285 data points. This cluster can be called the "most valuable group" as it has the highest means for frequency and monetary value for basic, discretionary, and luxury expenses (28.719577, 2292.986075, 42.150753, 4615.480471, 2.108751, 94.785005, respectively). For the recency of basic and discretionary expenses, the averages of cluster 1 are slightly higher than the averages of cluster 2 (with around 0.2 and 0.7 more, respectively). Furthermore, regarding the recency of luxury expenses, cluster 1 shows the lowest average with around 12 days [4].

The averages of every RFM attribute per expense type for January are listed in Table 5 [4].
July and December showed mainly similar results [4].

Table 5: Averages of RFM attributes of Expenses in January 2017 [4]

Cluster	Basic Expenses			Discretionary Expenses			Luxury Expenses		
/Mean	R	F	M	R	F	M	R	F	M
0	5.834877	10.365864	1381.259174	3.056825	16.714639	3068.152179	28.407762	0.185300	5.576335
1	0.744648	28.719577	2292.986075	0.374847	42.150753	4615.480471	12.039967	2.108751	94.785005
2	0.542437	25.351679	1839.823437	0.306911	35.898422	3905.645178	29.446608	0.112249	2.676574

Table 6: Comparison of K-Means and DBSCAN Performance [4]

Month /Method	K-Means			DBSCAN		
	Cluster Number	Outliers	Silhouette Coefficient	Cluster Number	Outliers	Silhouette Coefficient
January	3	17909	0.3105621087950614	13	32252	-0.154
July	3	20859	0.2531113832038402	13	45791	-0.188
December	3	22353	0.18548982217668522	11	56470	-0.087

5.3 Findings of DBSCAN

There were a few problems conducting DBSCAN using the Gower Distance as a distance metric. The calculation of the distance matrix was time-consuming, and the algorithm only found one cluster. Therefore, only the results of DBSCAN using the Euclidean distance were valid. In January, the number of clusters found was 13, and 32252 data points were considered noise points (meaning outliers). However, for all months, the silhouette coefficient was negative (in January: -0.154), suggesting wrongly assigned clusters and were therefore not significant [4]. Thus, no descriptive statistics of the clusters will be further discussed.

5.4 Discussion

The RFM Model was able to give some first indications on the identification of potential risk or vulnerable groups. The conducted methods were able to segment customers due to differences and similarities. The results of K-Means show a similar pattern for the three clusters in all three months. One cluster represents the biggest spenders, one the least spenders, and one with customers in between. The first cluster, which includes customers who spent the most recently, frequently and the highest amounts, was consequently classified as the "most valuable" [4]. As customers in this cluster are the most responsive to marketing (as discussed in Section 2), tailored advertisements to directly target this group can be developed. Customers in this cluster can still be vulnerable, as their capacity to spend these amounts of money is unknown. However, customers in the group of the least or lowest spenders can also be vulnerable customers, or they are simply very careful with their expenses [4]. It was difficult to classify any exact group as vulnerable because only the spending behavior was analyzed in depth, not the income [4]. However, these findings are still helpful for better understanding bank customers and to help with product and marketing development.

Comparing K-Means and DBSCAN, it is clear that K-Means performed better as DBSCAN only showed negative silhouette coefficients for the analyzed data set. Table 6 shows an overview of the performance of K-Means and DBSCAN [4]. DBSCAN determined the number of clusters itself, resulting in eleven to thirteen clusters, whereas the number of clusters for K-Means was manually set to three. Moreover, DBSCAN classified more outliers with 33352, 45791, and 56470 noise points for January, July, and December than K-Means with 17909, 20859, and 22353, respectively. While all silhouette coefficients for K-Means with 0.311, 0.253, and 0.185 were positive, the coefficients for DBSCAN were all negative with -0.154, -0.188, and -0.87 for January, July, and December, respectively [4]. Therefore, it can be said that the combination of RFM Model and K-Means worked on large data sets, while DBSCAN did not give valid clustering results, hence ways must be found to improve the algorithm, for example, with different starting conditions.

6 CONCLUSION, LIMITATIONS, FUTURE WORK

The study by Bartels [2021] proposed building the RFM Model and then using K-Means and DBSCAN for the customer segmentation. Larger data sets of anonymized transactions from Open Banking in the UK from 2017, focusing on the detailed expenses, were used. Not only could different thresholds for the RFM labels be identified, but additional customers with the highest and lowest RFM attributes could be evaluated. Furthermore, the study was able to find three customer segments based on expense behavior and patterns, which helps to better understand bank customers in the UK. The three clusters that K-Means found can indicate which customers can be seen as vulnerable or valuable. The group with the "most valuable" customers should be considered most responsive to marketing and targeted with marketing advertisements by financial institutions, banks, or companies. Regarding the customers that belong to the group of "least valuable" customers, strategies should be found to encourage more spending. It was not possible to determine the vulnerability or risk of potential customers, as attributes as income and savings were not included in the clustering algorithm but should be evaluated in future work [4]. In general, financial companies or

banks can and should use these findings in their advertising and product development planning.

The results show and support previous research that the RFM Model and K-Means are both effective methods for customer segmentation. DBSCAN did not perform as well as K-Means with negative silhouette coefficients, and in usage with the Gower Distance, there were no valid results. However, DBSCAN should be tested with different data or bank data in future work [4], as well as different conditions and distance measures.

There were some drawbacks of the study that can and should be addressed in future research. Future work can include the evaluation of different weights of the RFM Model on Open Banking data. Moreover, implementing a customer segmentation strategy in different countries might lead to different results. The vulnerability of customers should be addressed in future research, for example, using data of income and savings. Finally, especially considering the COVID-19 pandemic, future analyses of the transaction data of different years should be evaluated to identify possible changes and trends and to help with product tailoring and marketing.

REFERENCES

[1] (n.d.). The Global Open Finance Centre of Excellence. Retrieved 24.08. 2021 from https://ddi.ac.uk/case-studies/the-global-open-finance-centre-of-excellence/

[2] (n.d.). 2.3. Clustering. Retrieved 11.08. 2021 from https://scikit-learn.org/stable/modules/clustering.html#

[3] Abel, William , Key, Tomas and Shaw, Catherine. 2020. How persistent will the impact of Covid-19 on unemployment be? Retrieved 03.08. 2021 from https://www.bankofengland.co.uk/bank-overground/2020/how-persistent-will-the--impact-of-covid-19-on-unemployment-be

[4] Bartels, Catja. 2021. Cluster Analysis for Customer Segmentation with Open Banking Data. Master's thesis, Business School, University of Edinburgh.

[5] Cheng, Ching-Hsue and Chen, You-Shyang. 2009. Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications, 36, 3, Part 1 (2009/04/01/ 2009), 4176-4184. https://doi.org/10.1016/j.eswa.2008.04.003

[6] Dhandayudam, Prabha and Krishnamurthi, Ilango. 2014. A Rough Set Approach for Customer Segmentation. Data Science Journal, 13 (2014), 1-11. https://doi.org/10.2481/dsj.13-019

[7] Economist, The. 2021. Will the economic recovery survive the end of emergency stimulus? Retrieved 03.08. 2021 from https://www.economist.com/finance-and-economics/2021/07/18/will-the-economic-recovery-survive-the-end-of-emergency-stimulus

[8] Edmonds, Timothy. 2018. Open Banking: banking but not as we know it? Retrieved from https://commonslibrary.parliament.uk/research-briefings/cbp-8215/

[9] Ester, Martin, Kriegel, Hans Peter , Schubert, Erich, Sander, Jörg and Xu, Xiaowei 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Trans. Database Syst., 42, 3 (2017), 1–21. https://doi.org/10.1145/3068335

[10] Fader, Peter S., Hardie, Bruce G.S. and Lee, Ka Lok. 2005. RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. Journal of Marketing Research, 42, 4 (2005), 415-430. https://doi.org/10.1509/jmkr.2005.42.4.415

[11] Farris, P.W., Bendle, N., Pfeifer, P.E. and Reibstein, D. 2010. Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. Pearson Education.

[12] Forgy, E. 1965. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. Biometrics, 21 (1965), 768-769

[13] Han, J., Pei, J. and Kamber, M. 2011. Data Mining: Concepts and Techniques. Elsevier Science.

[14] Hossain, A. S. M. S. 2017. Customer segmentation using centroid based and density based clustering algorithms. Khulna, Bangladesh. https://doi.org/10.1109/EICT.2017.8275249

[15] Hughes, Arthur Middleton. 1994. Strategic database marketing : the masterplan for starting and managing a profitable, customer-based marketing program. Probus Pub. Co.

[16] Khalili-Damghani, Kaveh, Abdi, Farshid and Abolmakarem, Shaghayegh. 2018. Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. Applied Soft Computing, 73 (2018/12/01/ 2018), 816-828. https://doi.org/10.1016/j.asoc.2018.09.001

[17] Liao, Shu-hsien, Chen, Yin-ju and Lin, Yi-tsun. 2011. Mining customer knowledge to implement online shopping and home delivery for hypermarkets. Expert Systems with Applications, 38, 4 (2011/04/01/ 2011), 3982-3991. https://doi.org/10.1016/j.eswa.2010.09.059

[18] Ling, Raymond and Yen, David C. 2001. Customer Relationship Management: An Analysis Framework and Implementation Strategies. Journal of Computer Information Systems, 41, 3 (2001/03/01 2001), 82-97. https://doi.org/10.1080/08874417.2001.11647013

[19] Liu, Duen-Ren and Shih, Ya-Yueh. 2005. Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management, 42, 3 (2005/03/01/ 2005), 387-400. https://doi.org/10.1016/j.im.2004.01.008

[20] MacQueen, James. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA.

[21] Maimon, O. and Rokach, L. 2010. Data Mining and Knowledge Discovery Handbook. Springer US.

[22] Newell, Frederick. 1997. The New Rules of Marketing: How to Use One-To-One Relationship Marketing to Be the Leader in Your Industry. McGraw-Hill.

[23] Open Banking Limited. 2021. ABOUT THE OBIE. Retrieved 24.08. 2021 from https://www.openbanking.org.uk/about-us/

[24] Schijns, Jos M. C. and Schröder, Gaby J. 1996. Segment selection by relationship strength. Journal of Direct Marketing, 10, 3 (1996), 69-79. {https://doi.org/10.1002/(SICI)1522-7138(199622)10:3}<69::AID-DIR6>3.0.CO;2-W

[25] Smith, Wendell R. 1956. Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing, 21, 1 (1956), 3-8. 10.1177/002224295602100102

[26] Stone, B. 1984. Successful Direct Marketing Methods: The Bob Stone Direct Marketing Book. Crain Books.

[27] Sun, Zhao-Hui, Zuo, Tian-Yu, Liang, Di, Ming, Xinguo, Chen, Zhihua and Qiu, Siqi. 2021. GPHC: A heuristic clustering method to customer segmentation. Applied Soft Computing, 111 (2021/11/01/ 2021), 107677. https://doi.org/10.1016/j.asoc.2021.107677

[28] Wang, X. , Zhou, C., Yang, Y. , Yang, Y. , Ji, T. , Wang, J. , Chen, J. and Zheng, Y. . 2020. Electricity Market Customer Segmentation Based on DBSCAN and k-Means: A Case on Yunnan Electricity Market. Conference Location, https://doi.org/10.1109/AEEES48850.2020.9121413

[29] Wei-jiang, Liu, Shu-Yong, Duan, Xue, Yang and Xiaofeng, Wang. 2011. Determination of customer value measurement model RFM index weights. African Journal of Business Management, 5 (2011), 5567-5572. 10.5897/AJBM11.290

[30] Weinstein, Art. 2004. Handbook of Market Segmentation : Strategic Targeting for Business and Technology Firms, Third Edition. Taylor & Francis Group.

[31] Witten, I.H. and Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Elsevier Science.

[32] Wu, Jing and Lin, Zheng. 2005. Research on customer segmentation model by clustering. In Proceedings of the Proceedings of the 7th international conference on Electronic commerce. Association for Computing Machinery, Xi'an, China, 316–318. https://doi.org/10.1145/1089551.1089610

A APPENDICES

A.1 First Three Entries From Table "Customer Transactions" [4]

index	userid	startdate	savingcapacity	basicexp	discretionaryexp	luxuryexp	recurrentincome	totalincome	transactions number
0	5	2017-01-06 00:00:00.000000	-12381	10322.2	6474.04	128.79	0	4543.93	139
1	5	2017-02-01 00:00:00.000000	1189.42	5336.16	4329.84	48	0	10903.4	104
2	5	2017-03-01 00:00:00.000000	24640.3	4056.56	5327.67	173	0	34197.6	139

A.2 First Three Entries From Table "Transactions" in January [4] (Divided Into Two Parts for Clearer Visualization)

index	category	subcategory	transactiontype	accountid	accountprovider	accounttype	amount
0	nonRecurrent	expenseRefund	income	309153	Lloyds Bank	Current	0.47
1	nonRecurrent	other	transfers	309152	American Express	Credit Card	4.58
2	nonRecurrent	other	transfers	410936	Saga Credit Card	Credit Card	7.49

index	companyid	creditdebit	merchantbusinessline	description	providercategory	transactiondate	transactionid	userid
0	No Merchant	Credit	Account Provider	interest (gross)	Interest charges	2017-01-03 00:00:00.000000	1.8E+08	1
1	No Merchant	Credit	Account Provider	payment received - thank you	Credit Card	2017-01-23 00:00:00.000000	1.83E+08	1
2	No Merchant	Credit	Account Provider	direct debit - thank you	Credit Card	2017-01-12 00:00:00.000000	1.81E+08	1