An automatic machine learning based customer segmentation model with RFM analysis

Tran Thi Xuan *, Nguyen Thai Hoc

Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam
ttxuan@ictu.edu.vn,dtc205480201clc0021@ictu.edu.vn

Abstract. Big Data analysis has played a great role in extracting valuable insights from a vast amount of data, which can help companies make better business decisions and gain significant benefits. This paper investigates a Big Data based solution which applies the K-means clustering algorithm integrated with RFM analysis in a Spark processing pipeline for customer segmentation. Spark plays as a primary tool to deploy the entire experimentation process into a pipeline for automation, reusability, scalability, and enhanced computational power and speed for large datasets. The experimental results indicate the algorithm's effectiveness in consumer segmentation with the proposed Big Data based model.

Keywords: Customer segmentation · K-Means clustering · RFM model · Spark framework

1 Introduction

The focus of many companies is to provide the best products and services to attract attention in the market. Each customer has different preferences due to variations in age, gender, and other personal factors. Purchasing behavior is a significant indicator that helps determine customer's preferences. To achieve this, they must find the way to classify customers with similarities into segments. Customer segmentation based on their direct or indirect interaction behavior with the company can be challenging due to the difficulty in selecting key features that highlight the interactions [1], [3], [9].

RFM model that refers to the three key features of Recency, Frequency, and Monetary value has been considered as an effective technique to expose valuable insights of customers' behaviors [2]. Some studies have addressed that applying the K-means algorithm combined with the RFM model can be a promising solution for customer segmentation [1],[4],[10].

With the continuous growth of generated data, it is crucial to deploy a machine learning based segmenting model in a Big data system. Hadoop [8] and Spark [11] are among best Big data storage and processing technologies. In this

* Corresponding author

study, we propose an automatic, engaged machine learning based customer segmentation solution developed by Spark application framework while costumer data are stored in the HDFS storage.

The rest of the paper is organized as follows. Section 2 presents the background and the methodology of our proposal. Section 3 addresses the experiments and output results with discussion. Finally, Section 4 makes the conclusion for the study.

2 Methodology and Proposal

2.1 RFM analysis

The RFM analysis is based on the three factors of recency, frequency, and monetary value. The frequency (the times a costumer purchases or visits) and monetary value (the intention of a customer to spend) affects a customer's lifetime value while recency (the time since the last visit or purchase) indicates a measure of the engagement.

The RFM model plays an important role to get perceptive insights into customers and rank them into specific segments, which help marketers target customers with messages and offers that best match their relationship with a brand. Using RRM scores, a brand can specify such customer RFM segments as champion, best spender, loyalty, or at-risk, and so forth.

2.2 K-Mean clustering algorithm

K-Means is an widely used unsupervised learning algorithms to discover relationships among data features and group them into clusters that exhibit similarity and meaning. The main concept of K-Means is to partition the data into K distinct clusters, where K is predetermined. The K-Means algorithm utilizes the Euclidean distance metric to calculate distances between vectors in the feature space. The steps of K-mean clustering are illustrated in Fig. 1.

Input:
D= {t1, t2, …. Tn } // Set of elements
K // Number of desired clusters
Output:
K // Set of clusters
K-Means algorithm:
Assign initial values for *m1, m2,…. mk*
repeat
assign each item t_i to the clusters which has the closest mean;
calculate new mean for each cluster;
until convergence criteria is met;

Fig. 1: K-Means clustering

2.3 Spark framework

Spark framework [11] plays the role as a high-performance, unified, memory-based processing engine for big data. Spark provides numerous processing services in parallel for big data, ranging from preprocessing task, analytics, to machine learning modeling. Spark uses three data API so-called RDD (Resilient Distributed Dataset), DataFrame, and DataSet for various types of input data. In addition, it is straightforward to convert between two types of Spark data. That makes it suitable to develop a complete and automatic application that carries out different stages of data processing and yields the expected outcomes of customer segmentation.

2.4 HDFS

HDFS (Hadoop Distributed File System) [8] is a highly available distributed data lake that can store various types of data. Inside HDFS, data are split into blocks and stored across the whole computer cluster with replication. The replication mechanism in HDFS with the default factor of 3 guarantees the high availability for data. As Spark runs on top of Hadoop framework, it is convenient to store our data inside HDFS.

2.5 Proposed model

The unified machine learning model developed by Spark and integrated in a processing pipeline is illustrated in Fig. 2 . In what follows, all process steps will be described.

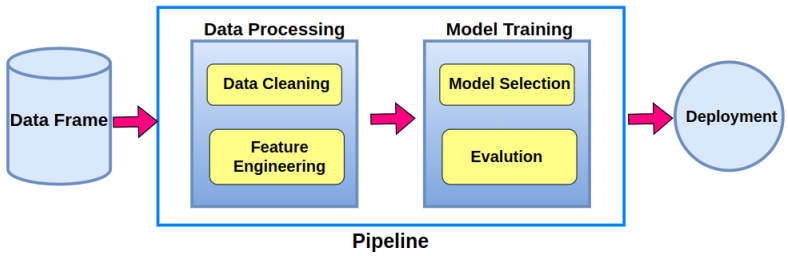


Fig. 2: The proposed analysis model

Data Preprocessing: After being fetched from the HDFS storage, data are pre-processed to ensure clean data that meets the input requirements of the model, such as removing noise, errors, and missing values. The three features of Recency, Frequency and Monetary value are derived in this step.

Feature Engineering: After pre-processed, some remaining issues with data discrepancies and uneven distribution have been observed. These need to

be addressed and adjusted to reduce model bias during training on this dataset. We have applied Box-Cox [6] transformation method. It is noted that to use this technique, the numerical values must be positive.

Training and validating model: In our Spark based model, K-Means clustering algorithm is applied to segment customers using the three features of RFM. The most critical task in deploying the K-Means algorithm is selecting the initial number of clusters K, to optimally partition the data into groups.

To address this, two techniques called the Elbow method [5] and Silhouette score [7] have been applied to achieve the appropriate value of K. Elbow method aims at computing the cost function by calculating the sum of squared distances from each data point to its nearest cluster centroid. The optimal number of clusters is identified at the elbow point on the curve plot, where the cost function begins to decrease more slowly. On the other hand, Silhouette score uses a graphical representation to succinctly present the degree of classification, measuring the optimality when an observation or data point is assigned to any cluster. The Silhouette score ranges from -1 to 1, where values reaching 1 indicate relatively well-defined clusters, and conversely for values closer to -1. Silhouette score however has limitations when data lacks clear structure or clusters have uneven sizes, which drives us to use it in a combination with the Elbow method for more reliable evaluation.

Model Deployment: After determining the appropriate cluster number K, the model is deployed to segment customers and provide the report of outcomes. The results then can be used by marketers to generate better offers and services for their customers.

3 Experimental Results and Discussion

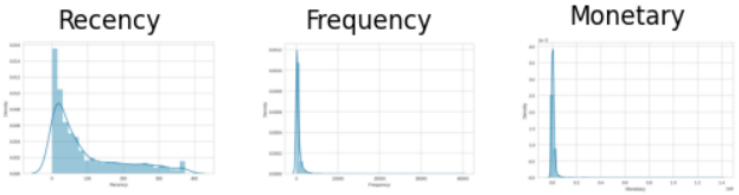
3.1 Data Preparation

The dataset used in this study was collected from an online retail e-commerce website in the UK in the period between 2010 and 2011. The dataset has a size of approximately 2 GBytes and includes information such as purchase quantities, purchase dates, etc. The first five samples of the dataset are shown in Fig. 3.

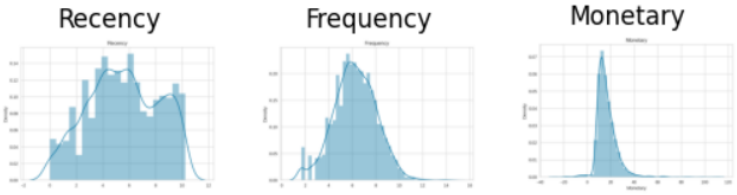
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Fig. 3: Customer datasets

The data is passed through the Data cleaning stage to fix and remove incorrect or missing data. The three features of RFM are also derived in this stage.



(a) Distribution of RFM features



(b) Normalized RFM features

Fig. 4: Data normalization using Box-Cox

Given in Fig. 4a, all three features do not follow normal distribution, which may cause serve problems for learning such as biased performance, imbalanced classes, inaccurate statistical measures and so forth. We apply the Box-Cox method for feature engineering, which successfully reshapes likely normal distributions as shown in Fig. 4b.

3.2 Training model and validation

The Elbow method and Silhouette methods analysis have been used to determine an appropriate number of cluster K. The Elbow method (Fig. 5) suggests that data should be grouped in 5 clusters.

Applying Silhouette scoring method verifies the outcome of Elbow method. Fig. 6 shows that all five clusters pass the threshold score (the red straight line), which indicates that all clusters are well-defined. As a result, $K = 5$ is the most appropriate cluster number for the considered dataset.

3.3 Model deployment

With the optimally chosen $K = 5$, the proposed model is deployed to classify customers in segments. Fig. 7 plots the customer segments from the input dataset.

It addresses that the pink cluster show the best customer segment with a high lifetime value (high frequency and monetary) and engaged relationship (low value of recency). The loyal and faithful customers is grouped in the blue

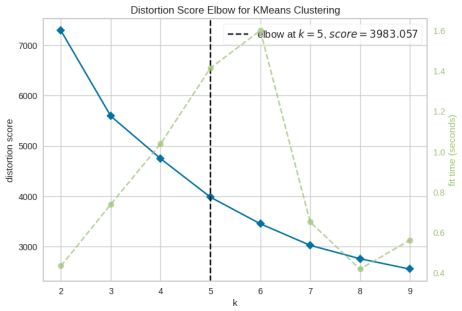


Fig. 5: Elbow method

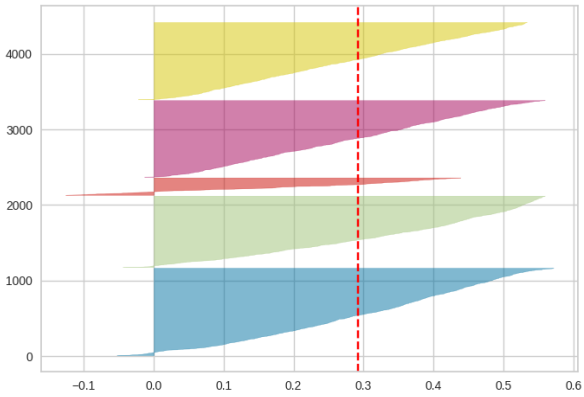


Fig. 6: Clusters vs. Silhouette score

and yellow segments where the monetary value and frequency are moderate yet the recency is relatively low.

Orange group shows a segment of customers who had high number of times interacting with the market but neither spend much for the products nor engage with the market lately. Therefore they might be at risk of churn. Purple cluster indicates a group of customer with a far interaction (high recency) and low purchase frequency. Likely not having engaged with the business recently and spent less money, these customers have the most loose relationship with the market.

In addition, we plot the clusters achieved from data without feature engineering in Fig. 8 for comparison. It is evident that the algorithm struggles significantly with data that does not adhere to a normal distribution, greatly affecting the model's effectiveness.

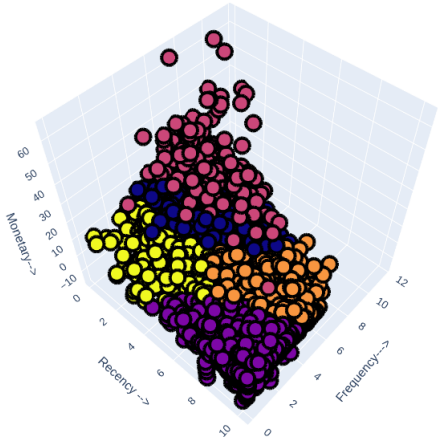


Fig. 7: Customer segments with $K = 5$ clusters

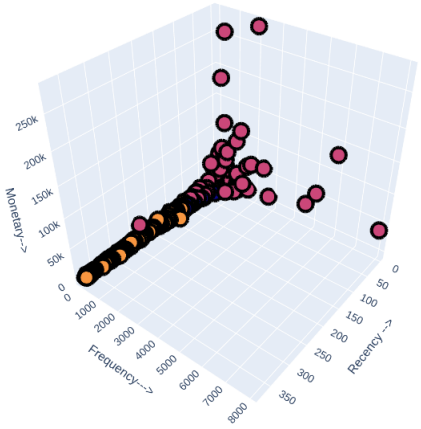


Fig. 8: Customer segments without feature engineering techniques

4 Conclusion

This study utilizes the Spark framework to construct a complete pipeline for a machine learning model that performs the customer segmentation. In the proposal, we adopt K-means clustering algorithm combined with the RFM model for customer segmentation tasks. Experimental results demonstrated that our model effectively classified and identified meaningful customer groups from the dataset. The outcomes also emphasized the importance of feature engineering on numerical data for optimizing model performance. It should be noted that the RFM model may not capture all aspects of customer behavior, such as personal preferences or browsing behavior, which could provide a more comprehensive view of customer segments. In future, more data aspects traced from production

industries should be taken into account of study. Moreover, K-means clustering may struggle with non-spherical clusters, which affects on the segmentation accuracy. Though beyond the scope of this research, other clustering algorithms as DBSCAN, Hierarchical Clustering, or deep learning models should be considered for comparison.

References

1. P. Anitha and Malini M. Patil. Rfm model for customer purchase behavior using k-means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5):1785–1792, 2022.
2. Richard Colombo and Weina Jiang. A stochastic rfm model. *Journal of Interactive Marketing*, 13(3):2–12, 1999.
3. ASM Shahadat Hossain. Customer segmentation using centroid based and density based clustering algorithms. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–6. IEEE, 2017.
4. Ina Maryani, Dwiza Riana, Rachmawati Darma Astuti, Ahmad Ishaq, Sutrisno, and Eva Argarini Pratama. Customer segmentation based on rfm model and clustering techniques with k-means algorithm. In *2018 Third International Conference on Informatics and Computing (ICIC)*, pages 1–6, 2018.
5. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
6. Remi M Sakia. The box-cox transformation technique: a review. *Journal of the Royal Statistical Society Series D: The Statistician*, 41(2):169–178, 1992.
7. Erich Schubert. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter*, 25(1):36–42, 2023.
8. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10, 2010.
9. Shreya Tripathi, Aditya Bhardwaj, and E Poovammal. Approaches to clustering in customer segmentation. *International Journal of Engineering & Technology*, 7(3.12):802–807, 2018.
10. Jo-Ting Wei, Shih-Yen Lin, and Hsin-Hung Wu. A review of the application of rfm model. *African journal of business management*, 4(19):4199, 2010.
11. Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache spark: a unified engine for big data processing, oct 2016.