MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Sixth Semester

BTech in CSE (AI & ML)

Department of Computer Science & Engineering

[Jan – May 2024]

Big Data Analytics Project

CSE 3272

Customer Segmentation Analysis Using Pyspark

-Anurag Kasat 210962180

-Aryan 210962152

CUSTOMER SEGMENTATION ANALYSIS USING PYSPARK

Abstract–**Customer segmentation is a critical component in the development of effective marketing strategies, allowing businesses to tailor their initiatives to meet specific customer needs. This project leverages PySpark, a sophisticated distributed computing framework, to conduct customer segmentation based on Recency, Frequency, Monetary (RFM) analysis. The methodology includes data preparation, exploratory data analysis, feature engineering, data pre-processing, model selection and training, and model evaluation and interpretation. Through these steps, the analysis identifies four distinct customer segments: Champions, Potential Loyalists, New Customers, and At Risk Customers, each with unique characteristics and behaviors. The findings provide actionable insights for targeted marketing campaigns and personalized promotions, ultimately contributing to improved customer retention and business growth. This study emphasizes the importance of scalable and efficient customer segmentation using advanced analytics tools in the era of big data.**

***Keywords*: Customer Segmentation, PySpark, RFM Analysis, Big Data Analytics, K-Means Clustering, Marketing Strategy, Recency Frequency Monetary, Targeted Marketing, Data Mining, Customer Relationship Management.**

I. INTRODUCTION

Customer segmentation is a critical component in the development of a comprehensive marketing strategy. It enables businesses to gain a deeper understanding of their customer base, allowing for the tailoring of marketing initiatives to meet specific customer needs. This project utilizes PySpark, a sophisticated distributed computing framework, to carry out customer segmentation based on Recency, Frequency, Monetary (RFM) analysis.

RFM analysis is a behavior-based technique recognized for its effectiveness in customer value segmentation. It enables businesses to analyze and quantify the recency, frequency, and monetary value of customer's purchases, leading to the identification of distinct customer segments. By utilizing this segmentation, businesses can optimize their marketing efforts, identify high-value customers, and customize their marketing strategies to align with the unique characteristics and preferences of each segment.

This project aims to explore the complexities of customer segmentation using PySpark and RFM analysis, highlighting its potential to transform marketing strategies and contribute to business growth. The following sections will offer a detailed examination of the methodology, results, and implications of this project. Through this endeavor, we aim to emphasize the importance of customer segmentation in today's competitive business environment and its role in shaping customer-centric marketing strategies.

II. OBJECTIVES

Distributed Customer Segmentation: Utilize the distributed computing capabilities of PySpark to segment customers based on their RFM scores. This allows businesses to identify distinct cohorts with unique purchasing behaviors, even when dealing with large-scale transactional data.

Scalable Analysis of Customer Segments: Perform a comprehensive analysis of the resulting customer segments using PySpark's distributed data processing capabilities. This enables businesses to uncover actionable insights from large datasets, informing targeted marketing strategies and personalized customer experiences.

Efficient Computation: Leverage PySpark's efficient computation capabilities to overcome the limitations of traditional customer segmentation methods when dealing with large-scale data. This project aims to demonstrate the scalability of customer segmentation using PySpark and RFM analysis.

Insight-Driven Business Strategies: Translate the insights derived from the RFM analysis into actionable business strategies. The goal is to use the findings from the customer segmentation to inform decision-making processes and drive

business growth, all tailored to the specific characteristics of each customer segment.

III. GAPS

Handling Missing Values: The absence of strategies to address missing data in the dataset poses a significant gap. Real-world data often contains missing entries, and failing to handle them appropriately can lead to biased analysis and inaccurate segmentation results.

Data Accuracy and Representativeness: The assumption of clean and fully representative data overlooks the inherent inaccuracies and biases present in real-world datasets. Without ensuring the accuracy and representativeness of the data, the segmentation results may not accurately reflect the true characteristics of the customer base, leading to flawed insights and ineffective strategies.

Addressing K-Means Assumptions: The reliance on K-Means clustering without considering its assumptions and limitations is another critical gap. K-Means assumes spherical clusters, which may not align with the actual distribution of customer data. Ignoring this can lead to suboptimal segmentation and misinterpretation of customer segments.

Optimizing Cluster Number Determination: Using subjective methods like the Elbow Method for determining the number of clusters introduces uncertainty and subjectivity into the segmentation process. Failing to employ more objective techniques, such as silhouette analysis, can result in misidentification of the optimal number of clusters and ultimately affect the quality of segmentation.

Integration of CLV and Additional Customer Attributes: Neglecting to incorporate customer lifetime value (CLV) calculations and other relevant customer attributes in the segmentation process limits the depth and accuracy of insights. Without considering these factors, the segmentation may overlook important dimensions of customer behavior and preferences, hindering the development of targeted and effective marketing strategies.

IV. LITERATURE SURVEY

[1]"RFM and Data Mining Models for Customer Lifetime Value"

Authors: H. Y. Chung, H. S. Kim, E. J. Lee, and S. Y. Suh
Published in: Expert Systems with Applications, 2015
Details: This paper explores the application of RFM analysis and data mining models for predicting customer lifetime value, emphasizing the importance of understanding customer behavior for effective CRM strategies.

[2]"RFM Analysis: Using Data Mining to Improve Customer Relationship Management"

Authors: A. Berson, S. Smith, and K. Thearling
Published in: McGraw-Hill Companies, 1999
Details: Providing a foundational understanding of RFM analysis, this book discusses how data mining techniques can enhance CRM strategies by segmenting customers and predicting their behavior.

[3]."Customer segmentation using RFM analysis and K-means clustering"

Authors: N. Akter and S. S. S. Wamba
Published in: International Journal of Business Information Systems, 2016
Details: This paper presents a study on customer segmentation using RFM analysis and K-means clustering, offering insights into effective segmentation techniques for targeted marketing.

[4]"The RFM model's past, present, and future"

Author: R. A. Bauer
Published in: Journal of Database Marketing & Customer Strategy Management, 2015
Details: Examining the historical development and future prospects of the RFM model, this paper discusses advancements in RFM analysis techniques and its evolving role in CRM strategies.

[5]"Customer segmentation with RFM analysis using K-means clustering on a telecommunications dataset"

Authors: S. Khajvand and E. Tarokh
Published in: International Journal of Engineering Research and Technology, 2015
Details: Focusing on a telecommunications dataset, this paper applies RFM analysis and K-means clustering for customer segmentation, providing insights into segmenting telecommunications customers effectively.

[6]"RFM Model in Customer Value Analysis"

Author: S. Miglautsch
Published in: Journal of Database Marketing & Customer Strategy Management, 2000
Details: This paper discusses the application of the RFM model in customer value analysis, highlighting how RFM metrics can quantify customer value and inform marketing strategies.

[7]"RFM: A Data Mining Approach for Customer Relationship Management"

Authors: A. Masand and S. Piatetsky-Shapiro
Published in: Knowledge Discovery and Data Mining, 1996
Details: Offering a data mining approach using RFM analysis, this paper explores methods for extracting actionable insights from RFM data to improve customer relationships and drive business growth.

V. METHODOLOGY

1. Data Preparation:

Data Loading: Load the customer data into a Spark DataFrame using SparkSession.
Data Understanding: Understand the structure of the data, including columns, data types, and any missing values.
Data Cleaning: Handle missing or invalid values, remove duplicates, and perform any necessary data cleansing operations.

2. Exploratory Data Analysis (EDA):

Descriptive Statistics: Compute summary statistics such as mean, median, standard deviation, etc., to understand the distribution of variables.
Data Visualization: Visualize key features such as transaction counts per country, distribution of RFM metrics (Recency, Frequency, Monetary), and any other relevant insights using tools like Plotly, Matplotlib, or Seaborn.
Identify Patterns: Look for patterns or trends in the data, such as seasonality in purchase behavior or correlations between variables.

3. Feature Engineering:

RFM Calculation: Calculate RFM metrics (Recency, Frequency, Monetary) for each customer based on their transaction history.
Additional Features: Create additional features that may be relevant for segmentation, such as total purchase amount, average order value, or time since last purchase.

4. Data Pre-processing:

Standardization: Standardize numerical features to ensure they have the same scale.
Handling Outliers: Identify and handle outliers in the data, if necessary.
Feature Selection: Select the most relevant features for segmentation analysis.
Data Transformation: Convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding.

CustomerID	Monetary	Frequency	Recency
17420	603.0	30	698
16503	1443.0	84	97
15727	5220.0	302	10
15100	875.0	3	71
16916	583.0	143	667
17809	5416.0	61	39
15738	4819.0	173	96
17223	432.0	49	312
18043	565.0	121	486
13329	787.0	17	122
17950	487.0	71	455
14713	2690.0	339	61
16565	174.0	3	425
13225	6466.0	32	62
17346	2691.0	500	1
13468	5737.0	302	1
15070	106.0	1	667
16510	249.0	13	667
15221	403.0	5	486
16985	5483.0	120	2

Fig. 1 Preprocessed data

5. Model Selection and Training:

Clustering Algorithm: Choose a suitable clustering algorithm such as K-means, Gaussian Mixture Models (GMM), or Hierarchical Clustering.
Optimal Number of Clusters: Determine the optimal number of clusters using techniques like the Elbow Method, Silhouette Score, or Gap Statistics.
Model Training: Train the selected clustering algorithm on the pre-processed data.

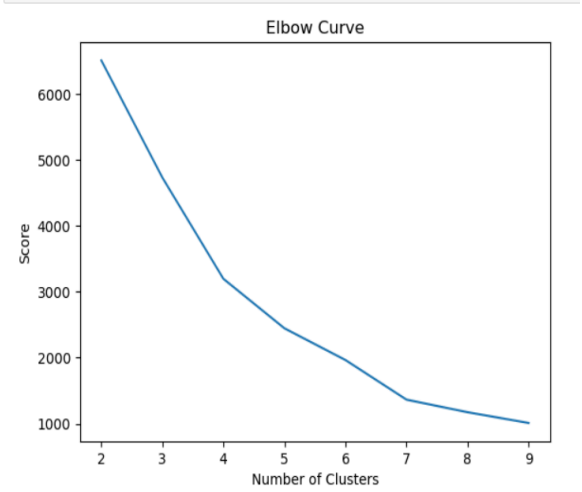

Fig. 2 Elbow Curve

6. Model Evaluation and Interpretation:

Evaluation Metrics: Evaluate the quality of the clustering model using appropriate metrics such as within-cluster sum of squares (WSS), silhouette score, or domain-specific metrics.

Cluster Analysis: Analyze the characteristics of each cluster, such as average RFM values, size of the cluster, and any notable patterns or trends.

Interpretation: Interpret the results of the clustering analysis to derive actionable insights for business decision-making.

Segment Profiles: Define profiles for each segment based on their RFM values and behavioral characteristics.

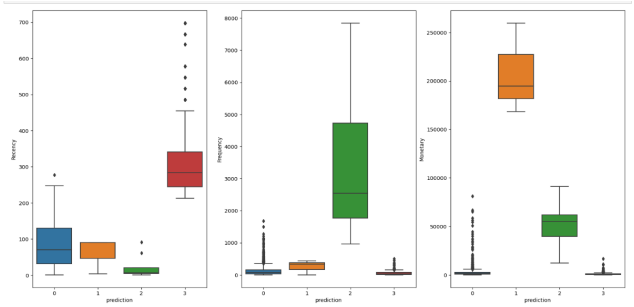

Fig. 3 Clusters

7. Post-processing and Visualization:

Customer Segmentation: Assign each customer to a specific segment based on their cluster membership.

Visualize Segments: Create visualizations such as bar charts or pie charts to illustrate the distribution of customers across different segments.

Business Recommendations: Provide actionable recommendations based on the segmentation analysis, such as targeted marketing campaigns, personalized promotions, or product recommendations for each customer segment.

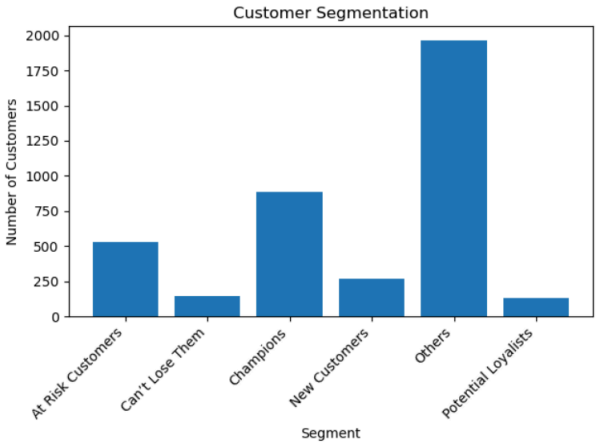

Fig. 4 Customer Segmentation

VI. RESULTS

From Fig-3 & Fig-4 has high frequency and monetary, whereas Cluster 3 has high recency but low frequency and recency.

High recency means they have not purchased anything from a long time.

The analysis yields(Fig 4) four distinct customer segments, each with unique characteristics and behaviors:

Champions: These are high-value customers who recently made frequent and substantial purchases, indicating strong loyalty and engagement.

Potential Loyalists: This segment comprises recent customers with average frequency and significant monetary value, presenting opportunities for nurturing long-term relationships.

New Customers: These customers exhibit high RFM scores but low frequency, suggesting potential for repeat purchases and customer retention strategies.

At Risk Customers: Previously active customers who haven't made recent purchases, signaling the need for re-engagement efforts and retention initiatives.

VII. ANALYSIS OF RESULTS AND DISCUSSION

Segmentation analysis provides valuable insights into customer behavior and preferences, enabling the development of targeted marketing strategies and initiatives. By focusing on specific customer segments, businesses can offer personalized promotions, offers, and communication strategies, resulting in increased customer satisfaction and improved retention rates.

Leveraging advanced analytics tools such as PySpark allows for efficient and scalable segmentation, which is particularly important in managing large volumes of data in today's big data landscape.

Customizing strategies to meet the unique needs of each customer segment can lead to revenue growth, improved customer retention, and strengthened brand loyalty.

For long-term success in customer segmentation and marketing efforts, it is crucial to continuously monitor and adjust strategies to adapt to changing customer needs and market trends.

VIII. CONCLUSION

In summary, the customer segmentation project using PySpark has provided meaningful insights into customer behavior and preferences. By applying advanced analytics techniques, we have effectively classified customers into distinct segments based on metrics such as recency, frequency, and monetary value. These segments span from high-value "Champions" to at-risk and inactive customers, presenting opportunities for targeted marketing strategies and initiatives.

Customized offers, promotions, and communication strategies tailored to each segment's needs can enhance customer satisfaction, improve retention rates, and drive revenue growth. The project underscores the value of employing tools like PySpark for efficient and scalable customer segmentation in the context of big data.

As we move forward, continuous monitoring and refinement of segmentation strategies will be crucial for long-term success. By staying attuned to evolving customer preferences and market trends, businesses can maintain a competitive edge and cultivate lasting relationships with their customer base.

IX. REFERENCES

[1] RFM and Data Mining Models for Customer Lifetime Value

[2] RFM Analysis: Using Data Mining to Improve Customer

[3] Customer segmentation using RFM analysis and K-means Clustering

[4] The RFM model's past, present, and future

[5] Customer segmentation with RFM analysis using K-means clustering on a telecommunications dataset

[6] RFM Model in Customer Value Analysis

[7] RFM: A Data Mining Approach for Customer Relationship Management