CENTER FOR DEVELOPMENT OF
ADVANCED COMPUTING

Customer Segmentation Using RFM Analysis

PG-DBDA September 2022

Submitted by:

Project Team 5

Tushar Shirsath

Somesh Rewadkar

Ayush Singh

Srujack Gedam

Shubham Mane

1. INTRODUCTION

In today's highly competitive business landscape, understanding and catering to customers' needs and preferences are crucial for a company's success. Customer segmentation is a powerful marketing technique that helps companies gain insights into their customers and create tailored marketing strategies to meet their specific needs. By dividing customers into smaller groups based on shared characteristics such as demographics, psychographics, or behavior, businesses can identify unique patterns and behaviors within their customer base.

Customer segmentation has become increasingly important in recent years, as advances in technology and data analytics have made it easier for companies to collect and analyze customer data. With the rise of e-commerce and social media, businesses can track customers' browsing and purchasing behavior, as well as their likes, interests, and social connections, to gain a better understanding of their preferences and needs.

To implement customer segmentation effectively, businesses need to collect and analyze large volumes of data from various sources. One way to do this is by using big data technologies such as Apache Spark and Amazon EMR (Elastic MapReduce). Amazon EMR is a fully-managed cloud service that enables businesses to process large amounts of data using popular big data frameworks such as Spark, Hive, and Hadoop.

In addition to data processing and segmentation, data visualization is another crucial aspect of customer analytics. By visualizing data, businesses can gain a better understanding of their customers' behavior and preferences and identify trends and patterns that may not be immediately apparent from raw data. Power BI is a powerful data visualization tool that allows businesses to create interactive reports and dashboards that help them make data-driven decisions.

In this project, behavioral segmentation was used to group customers based on their purchasing behavior. This approach is particularly effective because it provides insights into how customers interact with a company's products or services.

RFM (Recency, Frequency, Monetary) analysis is a popular technique for customer segmentation that helps businesses identify their most valuable customers based on their purchasing behavior. It involves analyzing three key metrics: how recently a customer has made a purchase (recency), how frequently they make purchases (frequency), and how much money they spend (monetary). By analyzing these metrics, businesses can identify their high-value customers and tailor their marketing strategies to meet their specific needs.

To perform RFM analysis, the following steps were taken:

Data Collection: Data on customer purchases and interactions with the company's products or services were collected.

Data Preparation: The data was cleaned and normalized to ensure accuracy and completeness.

Analysis: RFM scores were calculated for each customer based on their recency, frequency, and monetary value.

Segmentation: Customers were segmented into distinct groups based on their RFM scores.

2. PROBLEM STATEMENT

A company wants to increase customer loyalty and retention by developing more targeted and effective marketing strategies. To achieve this, the company needs to segment its customer base and identify the most valuable customers using RFM analysis.

The company is facing the challenge of not being able to effectively reach and engage its customers. It has a large customer base, but lacks the necessary understanding of their behaviors, preferences, and needs. As a result, the company's marketing campaigns are not as effective as they could be, resulting in lower sales and revenue.

To overcome this challenge, the company needs to conduct customer segmentation and RFM analysis to gain deeper insights into its customers' behavior and preferences. By segmenting customers based on their purchasing behavior, the company can create more targeted marketing campaigns and promotions that resonate with each customer segment. RFM analysis will help the company identify its most valuable customers and develop strategies to retain them.

The goal of this project is to conduct customer segmentation and RFM analysis to help the company develop more effective marketing strategies and increase customer loyalty and retention. By understanding its customers better, the company can create more personalized and relevant experiences that will drive customer satisfaction and increase revenue.

LITERATURE SURVEY

3.1 Introduction

Customer segmentation is the process of dividing a customer base into groups of individuals who share similar characteristics, needs, and behaviors. This technique is widely used by businesses to better understand their customers and create targeted marketing campaigns, product offerings, and customer experiences. Customer segmentation projects typically involve collecting and analyzing customer data, such as demographic information, purchase history, and behavior patterns, in order to identify distinct customer groups. The insights gained from customer segmentation can help businesses improve customer retention, increase sales, and enhance overall customer satisfaction.

1. Customer Segmentation:

Customer segmentation is the process of dividing customers into distinct groups based on their behaviors, preferences, and needs. According to Yim et al. (2004), customer segmentation helps businesses to better understand their customers and develop more effective marketing strategies.

The study suggests that customer segmentation can improve customer satisfaction, loyalty, and retention.In another study, Verhoef et al. (2010) highlight the importance of customer segmentation for customer relationship management. The study suggests that businesses can use customer segmentation to develop personalized marketing strategies that are more likely to resonate with customers.

2. RFM Analysis:

RFM analysis is a method used to identify a company's most valuable customers based on their purchasing behavior. RFM stands for Recency, Frequency, and Monetary Value. According to Fader and Hardie (2010), RFM analysis is a valuable tool for businesses to identify their most profitable customers and develop strategies to retain them.

In a study by Gupta and Lehmann (2006), the authors suggest that RFM analysis can be used to predict future customer behavior and help businesses to develop targeted marketing campaigns. The study also suggests that RFM analysis can be combined with other data analysis techniques such as predictive modeling to improve the accuracy of customer segmentation.

3. Customer Loyalty and Retention:

Customer loyalty and retention are critical for businesses to maintain a sustainable customer base and increase revenue. According to Reichheld (1996), customer loyalty is essential for businesses to achieve long-term success. The study suggests that businesses can increase customer loyalty by providing high-quality products and services and developing strong relationships with their customers.

In a study by Kim et al. (2010), the authors suggest that businesses can use customer segmentation and RFM analysis to identify customers who are most likely to defect and develop strategies to retain them. The study highlights the importance of customer retention for businesses to maintain a loyal customer base and increase revenue.

4. Amazon EMR and Apache Spark

Amazon Elastic MapReduce (EMR) is a web service that allows businesses to process large amounts of data using a distributed computing framework. One of the most popular distributed computing frameworks supported by Amazon EMR is Apache Spark. Spark is an open-source, in-memory distributed computing framework that provides high performance and scalability for big data processing. Amazon EMR makes it easy to set up and manage Spark clusters on the cloud, allowing businesses to focus on data processing rather than infrastructure management.

Spark is well-suited for data processing tasks such as ETL (extract, transform, load), data mining, machine learning, and graph processing. Spark's ability to process data in-memory provides faster performance compared to traditional disk-based processing frameworks. Spark's API also supports multiple programming languages such as Python, Java, and Scala, making it a versatile choice for data processing tasks.

Amazon EMR provides pre-configured Spark clusters that can be easily customized based on specific business needs. EMR also offers integration with other AWS services such as Amazon S3 (Simple Storage Service), Amazon Redshift (data warehouse service), and Amazon Kinesis (real-time data streaming service), making it easy to ingest and process data from various sources.

In conclusion, Amazon EMR and Spark provide a powerful combination for processing large amounts of data in a scalable and cost-effective manner. Spark's in-memory processing capabilities and multi-language API make it a versatile choice for a wide range of data processing tasks, while Amazon EMR simplifies cluster management and integrates with other AWS services for seamless data processing workflows.

5. KNN Algorithm

The K-nearest neighbors (KNN) algorithm is a type of supervised learning algorithm used for classification and regression tasks. The KNN algorithm works by finding the K closest training examples in the feature space to a given test example, and using the class labels of these neighbors to predict the label of the test example.

The KNN algorithm is a simple but effective algorithm that can be used for both binary and multi-class classification problems. One of the advantages of the KNN algorithm is that it does not require any assumptions about the underlying distribution of the data, making it a non-parametric algorithm. Additionally, KNN can handle non-linear decision boundaries, which makes it useful for a wide range of problems.

However, the main disadvantage of the KNN algorithm is its computational complexity. As the size of the dataset grows, the cost of finding the K-nearest neighbors for each test example can become prohibitively expensive. Additionally, KNN can be sensitive to the choice of K, and choosing the optimal value of K can be a challenging task.

In conclusion, the KNN algorithm is a powerful tool for solving classification and regression problems, particularly when the underlying distribution of the data is unknown or non-linear. However, its high computational cost and sensitivity to the choice of K should be taken into account when deciding whether to use it for a particular task.

3. LIBRARIES USED

1. Pandas

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

2. Numpy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It is open-source software. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

3. Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib

4. Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Important features of scikit-learn:

- Simple and efficient tools for data mining and data analysis. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, etc.
- Accessible to everybody and reusable in various contexts.
- Built on the top of NumPy, SciPy, and matplotlib.
- Open source, commercially usable – BSD license

Online Sales & Customer Segmentation

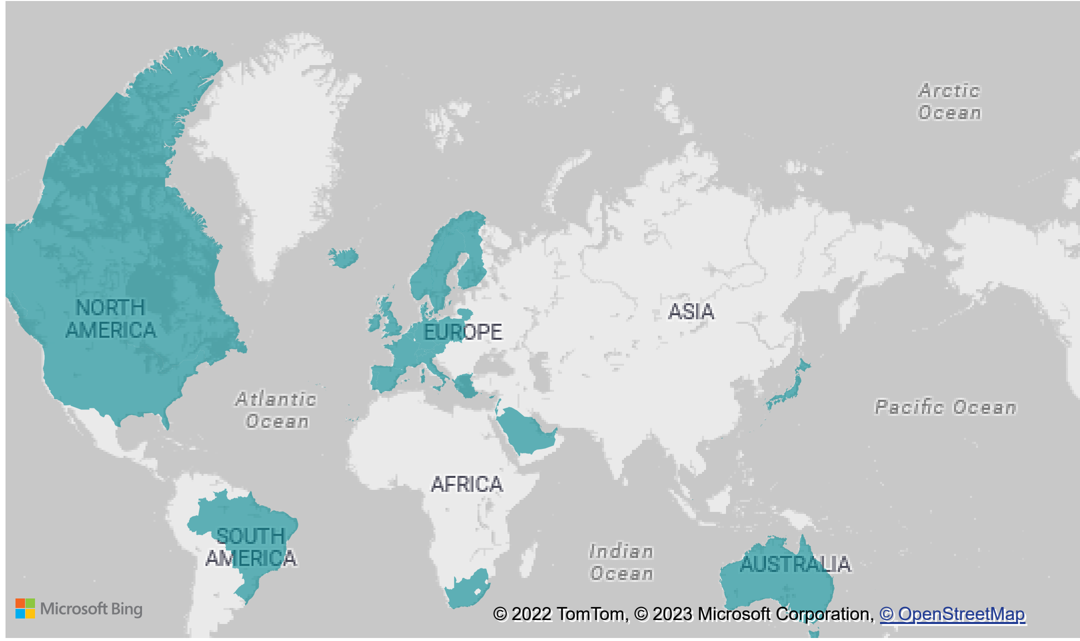
36
COUNTRIES

4371
CUSTOMERS

9,748K
TOTAL REVENUE

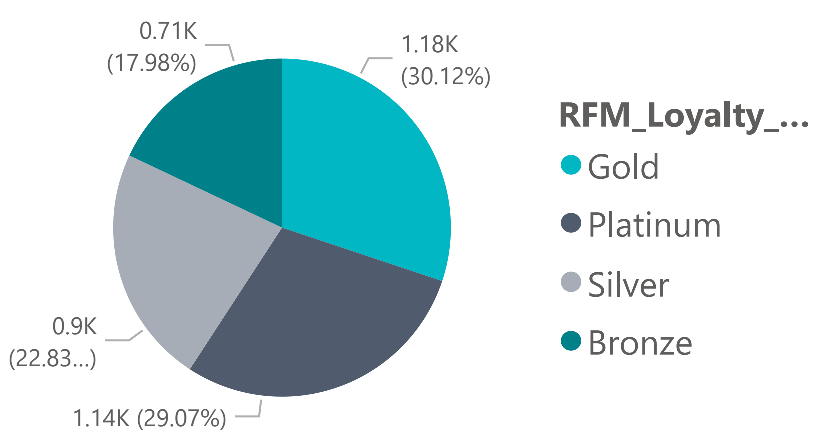
5310804
PRODUCTS

Total Revenue	Country	Year	Quantity
75,11,063.74	United Kingdom	2011	4100278
6,76,742.62	United Kingdom	2010	299081
2,75,877.06	Netherlands	2011	192741
2,54,246.87	EIRE	2011	136992
2,07,135.37	Germany	2011	110724
1,87,828.54	France	2011	105501
1,36,072.17	Australia	2011	82891
55,080.43	Switzerland	2011	29610
52,930.85	Spain	2011	25946
39,101.05	Belgium	2011	21397
33,949.61	Sweden	2011	31923
31,376.34	Norway	2011	15663
27,635.55	Japan	2011	21125



© 2022 TomTom, © 2023 Microsoft Corporation, © OpenStreetMap

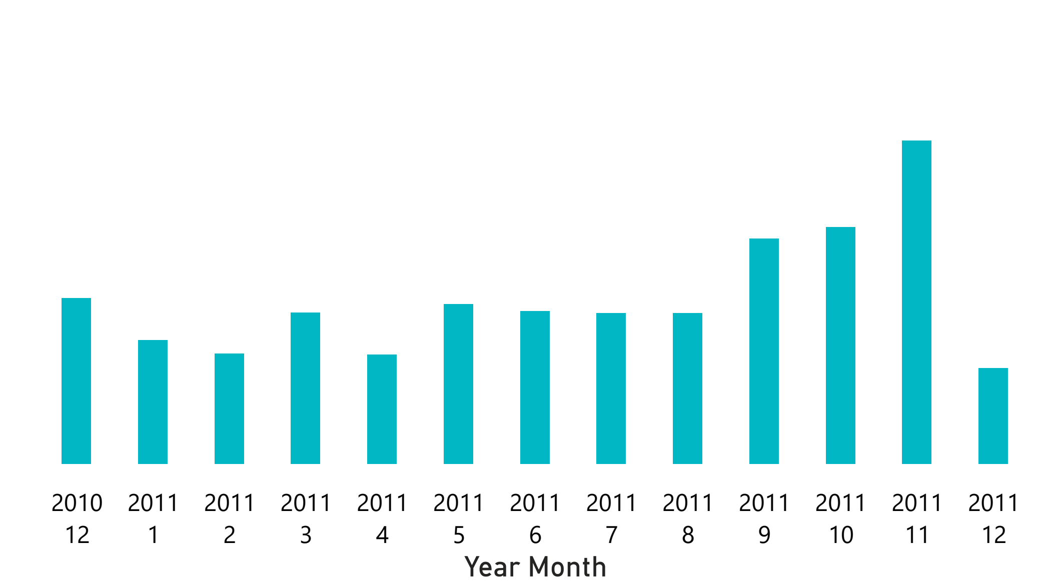
Count of RFM_Loyalty_Level by RFM_Loyalty_Level



Products Quantity and Total Revenue

Product	Quantity	Total Revenue
WORLD WAR 2 GLIDERS ASSTD DESIGNS	53847	13,587.93
JUMBO BAG RED RETROSPOT	47359	92,356.03
ASSORTED COLOUR BIRD ORNAMENT	36381	58,959.73
POPCORN HOLDER	36334	33,969.46
PACK OF 72 RETROSPOT CAKE CASES	36039	21,059.72
WHITE HANGING HEART T-LIGHT HOLDER	35313	99,668.47
RABBIT NIGHT LIGHT	30680	66,756.59
Total	**5310804**	**97,47,747.93**

Total Revenue by Year and Month



5. CONCLUSION & FUTURE SCOPE

Conclusion:

In conclusion, customer segmentation is a powerful technique that allows businesses to gain a deeper understanding of their customers and create tailored marketing strategies that cater to each customer group's specific needs. By using big data technologies such as Spark and Amazon EMR, businesses can analyze large volumes of data from various sources and identify their most valuable customers using RFM analysis. Power BI provides a powerful data visualization tool that helps businesses gain insights into customer behavior and preferences and make data-driven decisions.

Future Scope:

1. Integration with AI and Machine Learning: As AI and machine learning technologies continue to advance, businesses will be able to analyze customer data more effectively and make more accurate predictions about customer behavior. Integrating these technologies into the customer segmentation process could lead to even more precise and personalized marketing strategies.

2. Real-time segmentation: Currently, customer segmentation is often done using historical data, which may not reflect current customer behavior. Real-time segmentation would allow businesses to analyze customer behavior as it happens, enabling them to respond to changes in customer preferences and needs more quickly

3. Expansion to other channels: This project focuses on analyzing customer behavior in the context of e-commerce. However, businesses can also benefit from analyzing customer behavior across other channels such as social media, mobile apps, and offline interactions. Future work could explore how to integrate data from these channels into the customer segmentation process.

4. Personalization at scale: Personalized marketing is becoming increasingly important as customers expect more personalized experiences from businesses. However, personalizing marketing at scale can be challenging. Future work could explore how to use customer segmentation to create personalized marketing strategies for a large customer base.

5. Integration with customer feedback: Customer feedback can provide valuable insights into customer preferences and needs. Integrating customer feedback into the customer segmentation process could lead to even more accurate and effective segmentation.

6. REFERENCES

1. Yim, C. K., Tse, D. K., & Chan, K. W. (2004). Strengthening customer loyalty through intimacy and passion: Roles of customer-firm affection and customer-staff relationships in services. Journal of Marketing Research, 41(3), 281-292.

2. Verhoef, P. C., Reinartz, W. J., & Krafft, M. (2010). Customer engagement as a new perspective in customer management. Journal of Service Research, 13(3), 247-252.

3. Fader, P., & Hardie, B. (2010). Customer-base analysis in a discrete-time non-contractual setting. Marketing Science, 29(6), 1086-1108.

4. Gupta, S., & Lehmann, D. R. (2006). Customer metrics and their impact on financial performance. Marketing Science, 25(6), 718-739.

5. Reichheld, F. F. (1996). The loyalty effect: The hidden force behind growth, profits, and lasting value. Harvard Business Press.

6. Kim, Y. J., Lee, J. H., & Kim, W. Y. (2010). The role of customer classification in customer relationship management: An application to the banking industry. Expert Systems with Applications, 37(9), 6143-6150.