Project Report

On

Customer Segmentation Through Purchase History Data

Submitted in partial fulfillment for the award of
Post Graduate Diploma in Big Data Analytics (PGDBDA)
From KnowIT (Pune)

Guided by:

Miss Trupti Joshi & Prasad Deshmukh Sir

Submitted By:

Tejas Shinde(230943025050)
Shruti Kharsade(230943025046)
Pranita Wani(230943025038)
Jay Chaudhari(230943025025)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Tejas Shinde(230943025050)
Shruti Kharsade(230943025046)
Pranita Wani(230943025038)
Jay Chaudhari(230943025025)

Have successfully completed their project on

Customer Segmentation through Purchase History Data

Under the guidance of Miss Trupti Joshi and Prasad Deshmukh sir

ACKNOWLEDGEMENT

This project Customer Segmentation Through Purchase History Datawas a great learning experience for us and we are submitting this work to CDAC KnowIT (Pune). We all are very glad to mention the name Miss TruptiJoshi for her valuable guidance to work on this project. Her guidance and support helped us to overcome various obstacles and intricacies during the course of project work. We are highly grateful to Mr. Vaibhav Inamdar Manager (KnowIT), CDAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PGDBDA) through CDAC ACTS, Pune. Our most heartfelt thanks goes to Mr Prasad Deshmukh sir (Course Coordinator, PGDBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in CDAC KnowIT, Pune.

TABLE OF CONTENTS

Abstract

This project utilizes the capabilities of Spark for efficient data processing and employs advanced ML techniques such as RFM (Recency, Frequency, Monetary Value) analysis and K-means clustering to segment customers based on their purchase history. By exploring the patterns in customer behavior, it aims to extract valuable insights crucial for targeted marketing strategies. Utilizing Tableau for visualizations enhances the interpretation of results, empowering stakeholders to make informed decisions and drive business growth effectively.

By integrating these tools, our project aims to provide businesses with insights into their customer base. This understanding enables targeted marketing campaigns, personalized recommendations, and optimized strategies. Ultimately, the goal is to enhance customer satisfaction, drive growth, and maintain competitiveness. Through effective segmentation and visualization of purchase history data, businesses can make data-driven decisions leading to tangible improvements in engagement and performance.

1. INTRODUCTION

In the dynamic landscape of modern business, understanding and catering to the unique needs of customers has become a crucial element for success. One powerful tool for achieving this is customer segmentation, a strategy that involves dividing a customer base into distinct groups based on shared characteristics or behaviors. By doing so, businesses can tailor their marketing efforts, optimize product offerings, and enhance overall customer satisfaction.

Businesses struggle to extract insights from vast customer purchase data. Inefficient analysis and lack of integration hinder effective segmentation and targeted strategies, impacting growth and competitiveness. Customer segmentation is pivotal in modern marketing, dividing a diverse customer base into smaller, homogeneous groups. By understanding distinct customer segments, businesses can tailor strategies, enhance marketing effectiveness, and drive growth through personalized experiences, divides customers into groups with similar characteristics or behaviors , Used to tailor marketing strategies effectively based on specific customer needs and behaviors .It helps businesses improve customer satisfaction, increase sales, and optimize marketing ROI.

Through the implementation of advanced data analysis techniques and machine learning algorithms, this project seeks to uncover hidden patterns within the purchase history data that may not be immediately apparent. The ultimate goal is to create meaningful and actionable customer segments that allow businesses to make informed decisions and allocate resources more effectively.

Dataset Collection and Features

Data Sources

This file contains purchase data from April 2020 to November 2020 from a large home appliances and electronics online store.Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users.Data collected by Open CDP project.This dataset was acquired via Kaggle, a well-known website that hosts datasets and competitions for data research.

Data Structure

Contains purchase data, enabling analysis of customer behavior and preferences for segmentation and targeted strategies. 2.6M rows, 8 columns: event_time, order_id, product_id, category_id, category_code, brand, price, user_id.

Dataset Size

The dataset comprises 2.6M records containing valuable information on customer Purchase history

Features/Attributes

here's an overview of the key features within your purchase data:

Event Time (Timestamp): This column represents the timestamp when the purchase event occurred. It provides valuable information about the timing and frequency of purchases, which can be analyzed to identify patterns and trends in customer behavior over time.

Order ID (Bigint): Each purchase is associated with a unique order ID, which serves as a reference to identify and track individual transactions. Analyzing order IDs can help in understanding the volume of orders and the distribution of purchases across different orders.

Product ID (Bigint): This column contains the unique identifier for each product purchased. It enables tracking of product sales and identification of popular or frequently purchased items.

Category ID (Bigint): Indicates the category to which the purchased product belongs. Analyzing category IDs allows for the segmentation of products into different product categories, facilitating market basket analysis and understanding customer preferences across product types.

Category Code (String): Provides a textual representation of the product category. This column may contain hierarchical information about the product category, such as the main category and subcategories, which can be useful for organizing and analyzing products based on their classification.

Brand (String): Specifies the brand associated with the purchased product. Brand information is valuable for brand analysis, brand affinity studies, and understanding brand loyalty among customers.

Price (Double): Represents the price of the purchased product. Analyzing price data allows for understanding the distribution of product prices, identifying price segments, and detecting anomalies or outliers in pricing.

User ID (Bigint): Indicates the unique identifier of the user who made the purchase. Analyzing user IDs enables customer segmentation, personalized marketing, and understanding customer lifetime value (CLV) based on purchasing behavior.

These key features provide valuable insights into customer purchase behavior, product preferences, and market dynamics, which can be leveraged for various analytical tasks such as customer segmentation, product recommendation, and sales forecasting.

Here is an overview of the key features (attributes) within our dataset:

2. SYSTEM REQUIREMENTS

Hardware Requirements

Computer: A computer with sufficient processing power and memory to run data processing and analysis tasks. A modern multicore processor and at least 8 GB of RAM are recommended.

1. Storage: Adequate storage space to store the generated dataset and any additional datasets if required. An SSD (Solid State Drive) is recommended for faster data access.
2. Internet Connection: A stable internet connection for downloading and installing software packages and libraries, as well as for any online resources needed during the project.

Software Requirements

1. Operating System: Windows 10 or higher
2. Apache Spark: If your project involves big data processing, consider installing Apache Spark. You can use PySpark to interact with Spark using Python.
3. Integrated Development Environment (IDE): Choose a Pythonfriendly IDE, such as PyCharm, Jupyter Notebook, Visual Studio Code, or your preferred text editor.

Visualization Software

1. Tableau: If you plan to visualize and analyze data with Tableau, install Tableau Desktop.

3. **FUNCTIONAL REQUIREMENTS**

(1) Apache Spark:

What is Spark: Apache Spark is an opensource distributed computing system designed for processing large volumes of data.

Key Features: Spark provides a number of key features that make it wellsuited for processing big data, including inmemory processing, support for various data sources and formats, faulttolerance, and scalability.

Spark also provides a range of APIs, including SQL, streaming, machine learning, and graph processing, making it a versatile platform for a wide range of use cases.

(2) Tableau:

Data visualization is the graphical representation of information and data.

It helps create interactive elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Tableau is widely used for Business Intelligence but is not limited to it.

It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.

All of this is made possible with gestures as simple as drag and drop.

ARCHITECTURE

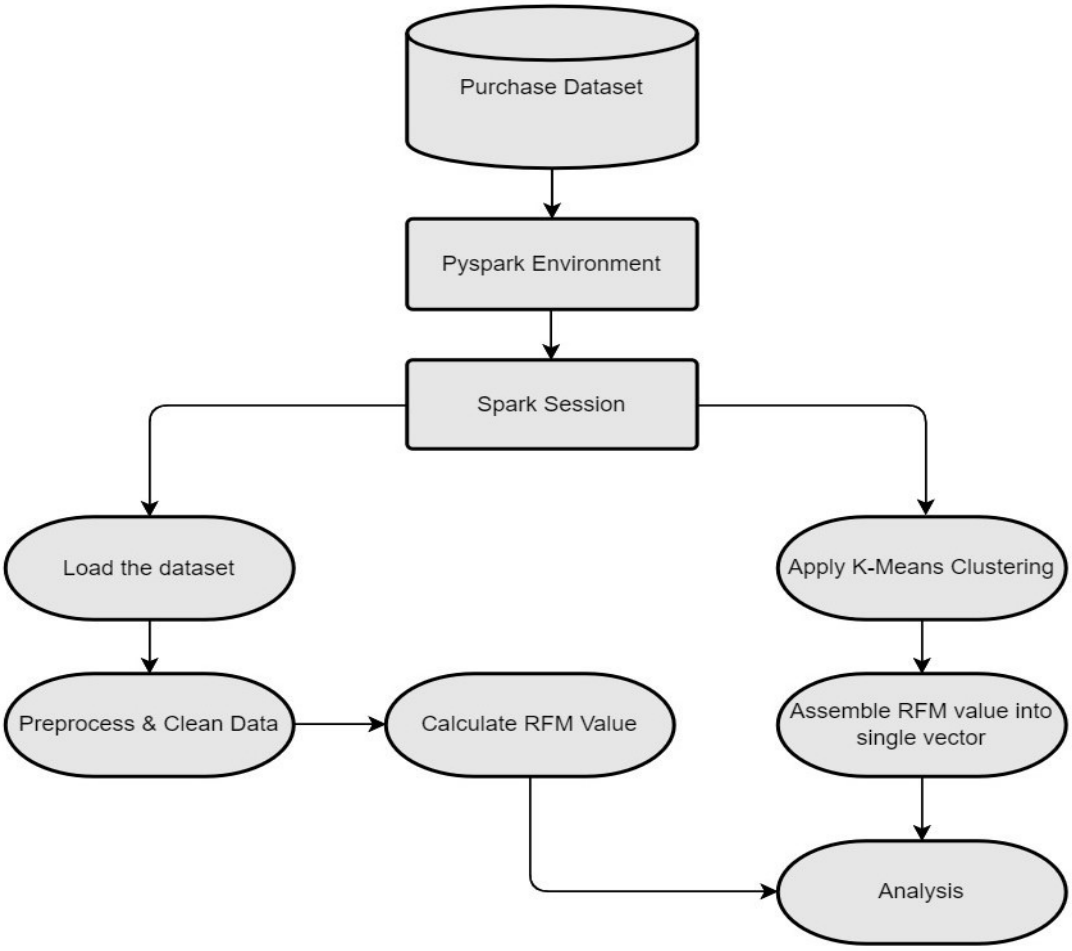


Fig: System Architecture Of customer segmentation through perches history data

MACHINE LEARNING ALGORITHMS

RMF MODEL

From the dataset above, we need to create multiple customer segments based on each user's purchase behavior. The variables in this dataset are in a format that cannot be easily ingested into the customer segmentation model. These features individually do not tell us much about customer purchase behavior.

Due to this, we will use the existing variables to derive three new informative features - recency, frequency, and monetary value (RFM) is commonly used in marketing to evaluate a client's value based on their:

Recency: How recently has each customer made a purchase?

Frequency: How often have they bought something?

Monetary Value: How much money do they spend on average when making purchases?

We will now preprocess the dataframe to create the above variables.

KMeans Clustering

Algorithm Overview:

Customer segmentation is a vital task in marketing and ecommerce analytics. It involves dividing customers into distinct groups based on shared characteristics, enabling businesses to tailor marketing strategies and services more effectively. In our project, we employ the KMeans Clustering algorithm for customer segmentation.

Explanation:

KMeans Clustering is an unsupervised machine learning algorithm used for partitioning data into 'K' distinct, nonoverlapping clusters or groups. Each cluster represents a set of data points that are similar to each other based on selected features. Here's how KMeans Clustering works in our project:

1. **Feature Selection:**

We select relevant features from our customer dataset, such as purchase history, demographics, and behavioral data. These features help define the similarity between customers.

2. **Normalization:**

Before applying KMeans, we normalize the data to ensure that features with different scales do not dominate the clustering process.

3. **KMeans Algorithm:**

We initialize 'K' centroids (cluster centers) randomly within the feature space. The algorithm iteratively assigns each customer to the nearest centroid based on a chosen distance metric (usually Euclidean distance). After all customers have been assigned, the centroids are recalculated as the mean of all data points within their respective clusters.Steps 2 and 3 are repeated until convergence, i.e., until the centroids no longer change significantly.

4. **Cluster Interpretation:**

Once clustering is complete, customers are grouped into 'K' clusters. Each cluster represents a segment of customers who exhibit similar behaviors or characteristics.

5. **Business Insights:**

These customer segments can provide valuable business insights. For example, one cluster might contain highvalue, frequent customers, while another might represent occasional shoppers. Marketing strategies, product recommendations, and promotions can then be customized for each group.

Benefits:

1. Targeted Marketing:

Personalized Communication: Segmentation helps in crafting messages and offers that resonate with specific segments, leading to higher engagement.

2. Market Expansion:

Identifying New Markets: Segmentation can reveal new, untapped markets or segments that align with a company's strengths.

3. Enhanced Profitability:

Optimized Pricing: Segments may have varying price sensitivities, allowing for pricing strategies that maximize revenue.

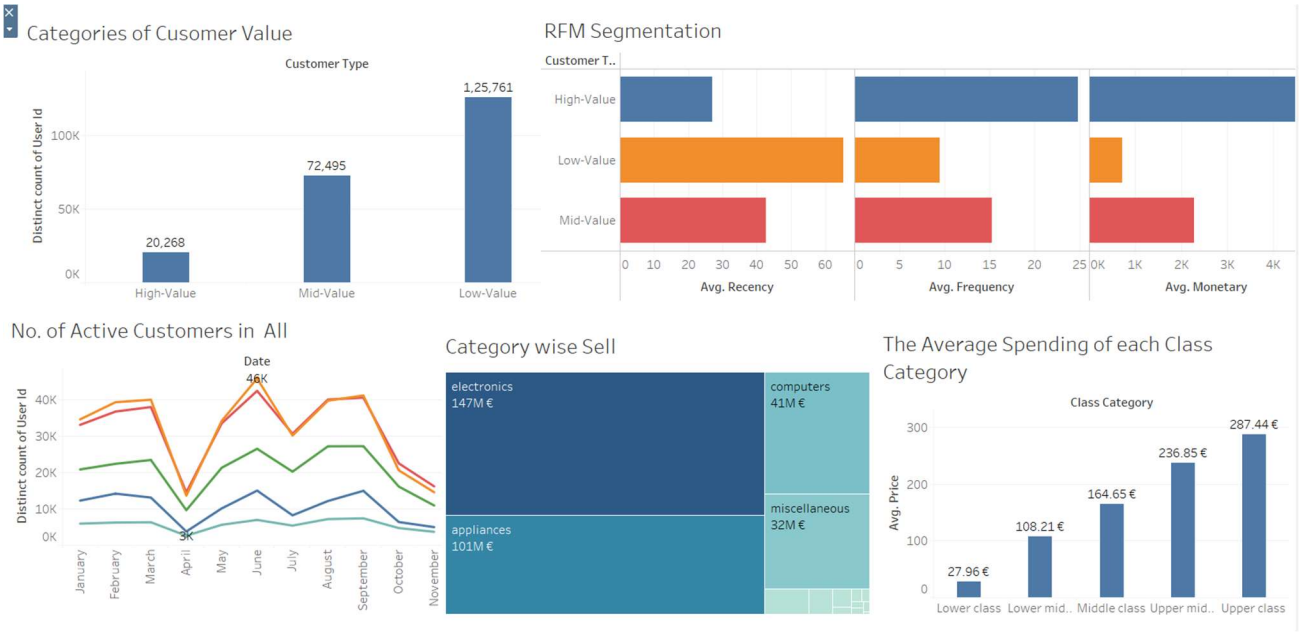
4. Risk Mitigation:

Diversification: If one segment is impacted (e.g., economic downturn), diversified segments can help stabilize revenue.

Conclusion:

K-Means Clustering is a powerful tool for customer segmentation in our project. It helps businesses gain insights into customer behavior, preferences, and needs, ultimately driving more effective marketing and sales strategies.

DATA VISUALIZATION AND REPRESENTATION



CONCLUSION AND FUTURE SCOPE

Conclusion:

- In the rapidly evolving landscape of customer segmentation through purchase history data offers valuable insights for e-commerce stakeholders.
- Utilizing RFM analysis and K-means clustering, we segment customers effectively based on their purchasing patterns, enabling targeted marketing strategies.
- This approach reveals variations in profitability among customer segments and provides a structured framework for optimizing business strategies and driving growth.

Our project has revolved around several key aspects:

1. Data Cleaning and preprocessing:

Clean and preprocess the data to ensure accuracy and consistency. This may involve removing duplicates, handling missing values, and standardizing data formats.

4. Data Analysis with Apache Spark:

- Apache Spark, a powerful data processing framework, facilitated data analysis. We demonstrated how to extract, transform, into Spark DataFrames, allowing for indepth analysis.

5. Machine Learning and Customer Segmentation:

- We utilized the K-Means Clustering algorithm to perform customer segmentation. This ML technique partitioned customers into distinct groups, enabling businesses to personalize marketing strategies.

6. Data Visualization and Insights:

- Visualizations created using Tableau showcased real-time data trends, providing businesses with valuable insights for decision-making.

7. Future Scope and Enhancements:

- Our project has vast potential for expansion. Future enhancements could include incorporating real data sources, implementing more advanced machine learning models, and deploying predictive analytics for sales forecasting.

8. Business Relevance:

- The project underscored the immense value of data analysis in purchase history dataset. , the insights derived from this analysis can significantly impact business success.

Future Scope:

- The project, of customer segmentation has laid a solid foundation for exploring various avenues of enhancement and expansion. While the current implementation has demonstrated the , approach that reveals variations in profitability among customer segments and provides a structured framework for optimizing business strategies and driving growth. there are several exciting possibilities for future development and improvement:

1. Advanced Machine Learning Models:

- Enhance the machine learning component by incorporating more advanced algorithms such as Random Forests, Gradient Boosting, or Neural Networks. These models can provide deeper insights into customer behavior, product recommendations, and sales forecasting.

2. Predictive Analytics:

- Implement predictive analytics to forecast future sales trends and customer behavior. Predictive models can help businesses anticipate demand, optimize inventory management, and plan marketing campaigns effectively.

3. A/B Testing and Personalization:

- Incorporate A/B testing frameworks to evaluate the effectiveness of different marketing strategies, website designs, or product placements. Additionally, invest in personalized recommendation systems to enhance the customer shopping experience.

4. Cloud-Based Scalability:

- Migrate the data processing and storage components to cloud platforms like AWS, Azure, or Google Cloud. This ensures scalability, high availability, and cost-efficiency as data volumes grow.

The future scope of this project is not limited to these suggestions; it extends as far as your imagination and business objectives. Continual adaptation and innovation are key to thriving in the dynamic customer history purchase landscape, and leveraging data insights is central to this endeavor.

By exploring these avenues, businesses can remain competitive, adapt to changing customer behaviors, and make data-driven decisions that lead to sustainable growth and success.

References

1. https://towardsdatascience.com/top-3-python-packages-to-generate-synthetic-data33a351a5de0c

2. Apache Spark. [https://spark.apache.org/]

3. scikit-learn. [https://scikit-learn.org/]

4. Confluent Python Client. [https://docs.confluent.io/platform/current/clients/confluentkafka-python/html/index.html]