CUSTOMER SEGMENTATION & MARKET BASKET ANALYSIS

A PROJECT REPORT

Submitted by

SHUBHRADEEP CHATTERJEE

ABSTRACT:

In the realm of successful E-commerce, two key pillars are innovation and understanding customer desires. Today's user-friendly E-commerce platforms encourage purchases, driven by innovation's ability to captivate customers with products. However, the extensive product range can leave customers bewildered. To tackle this, businesses often categorize customers into segments based on their behaviour: High (frequent, high-spending, recent visitors), Medium (less spending, moderate visits), and Low (potential churn). Machine Learning (ML) steps in with algorithms that uncover hidden data patterns, aiding informed decision-making. Our proposed customer segmentation concept utilizes ML algorithms like Support Vector Machine Classifier, Random Forest, Logistic Regression, and XGBOOST. This approach empowers B2C companies to excel by tailoring products and services to specific segments, effectively reaching potential customers. For retail, sales are paramount. Instead of assuming customers know all offerings, businesses should showcase options to boost engagement and sales. ML, with its ability to discern hidden patterns, plays a pivotal role here. Algorithms unveil insights that drive sales growth. Furthermore, ML is essential for Retail Industry, as it harnesses vast customer purchasing data. Our paper utilizes Market Basket Analysis employing concepts like Apriori algorithms. This aids retailers in filtering and making personalized recommendations based on customer purchasing history. This technique enhances the shopping experience and boosts sales.

ABOUT THE DATA:

The dataset contains records of retail transactions, offering a glimpse into the world of buying and selling. Each row represents a unique transaction, revealing the interplay between customers, products, and time. Let's break down the columns to understand what's inside:

- **InvoiceNo:** This special code marks the identity of each transaction. It's like a fingerprint that distinguishes one purchase from another.
- **StockCode:** Each product or item in the store has its own StockCode, a sort of label that helps us recognize what's being bought.
- **Description:** This field provides a brief description of the product or item being sold. It's like the title of a book that gives you an idea of what's inside.
- **Quantity:** How many of each product were bought in a transaction? This column gives us that vital piece of information.
- **InvoiceDate:** Every transaction has a date and time stamp, showing exactly when it took place. It's like a timestamp in your photo album that tells you when each picture was taken.
- **UnitPrice:** This is the cost of one unit of the product. Just like the price tag on a product in a store, this tells us how much each individual item costs.
- **CustomerID:** Each customer is given a unique ID, which helps us track who made each purchase. It's similar to a membership card at a store that identifies you as a valued shopper.
- **Country:** The dataset spans different countries, each represented in this column. It's like an international flavour, showing where these transactions took place around the world.

With this dataset, we can uncover fascinating insights. We can understand which products are popular, when shopping spikes, and even uncover patterns where certain products are frequently bought together.

SYSTEM REQUIREMENTS:

A python program is developed and it imports essential libraries for data analysis and visualization, sets up visualization configurations, and ensures smooth plotting. This setup is designed to facilitate data exploration, analysis, and visualization, making it suitable for in-depth analysis and visual representation of data.

- **pandas:** Data manipulation, analysis.
- **numpy:** Numeric computations.
- **matplotlib:** Visualization library.
- **seaborn:** Statistical graphs.
- **nltk:** Text analysis toolkit.
- **Orange:** Data mining, ML.
- **sklearn:** Machine learning toolbox.
- **mlxtend:** ML extensions.

METHODOLOGY:

The methodology employed in this study revolves around **two key areas**: **Customer Segmentation** and **Market Basket Analysis**. To begin, the data is **prepared** by loading it and addressing any **missing information**. A comprehensive understanding of the dataset's **attributes and structure** is established, followed by the implementation of strategies to handle missing data effectively. Exploring variable **distributions and relationships** provides initial insights into the dataset. Moving forward, a detailed **market analysis** is conducted, shedding light on **customer-product dynamics**. This includes the examination of **stock codes**, **basket prices**, and gaining insights from **product descriptions and categories**.

The core of customer segmentation involves the **RFM analysis** approach, encompassing **Recency, Frequency, and Monetary Value** factors. Based on this analysis, **distinct customer categories** are created. Classification algorithms, including **Support Vector Machine**, **Logistic Regression**, **K Nearest Neighbour**, **Random Forest**, and **XGBoost** are then employed to classify customers further. Enhancing predictive accuracy, a **Voting Classifier** is introduced and tested to evaluate predictions.

Additionally, the study delves into **cross-selling opportunities** through **Market Basket Analysis**, facilitated by **association rule mining**. A transaction dataset is carefully defined and pre-processed, leading to the extraction of **frequent item sets**. These insights are gleaned through the lens of **Association Rule Mining**. The methodology concludes by summarizing findings from the **Market Basket Analysis**, presenting a comprehensive picture of the potential for **cross-selling strategies**.

DATA READING AND PREPARATION

Commencing with the dataset loading, the **"Customer_Segmentation.csv"** file was ingested into a pandas DataFrame. Subsequently, a thorough investigation of **missing values** was conducted, highlighting columns with notable absence. This accentuated their potential impact on forthcoming analyses. It is imperative to address these gaps through measures such as **imputation** or **column removal** to preserve data credibility and result accuracy.

ANALYZING MARKET TRENDS AND CUSTOMER – PRODUCTS DYNAMICS

This report presents a **comprehensive analysis** that encompasses multiple facets of the dataset. Starting with a **thorough examination** of market trends, customer-product interactions, stock codes, and basket prices, the analysis provides **crucial insights** for strategic decision-making, particularly in **customer engagement** and **product portfolio optimization**. The **geographic distribution** of orders is vividly depicted through **bar and pie charts**, with the UK emerging as a **dominant market**. **High-value transactions** are scrutinized, and **cancellations** are appropriately addressed to refine the dataset. A focused **DataFrame** is crafted, detailing unique alphabetical **"StockCode"** values and their corresponding **"Description"**. **Purchase patterns** are explored, revealing that over **65% of orders** surpass £200, with approximately **44% ranging from £200 to £500**. Additionally, a **bump plot** effectively highlights **rank shifts over time**. Furthermore, a function is devised to **extract keywords** from product descriptions, culminating in a **binary matrix** used for **k-means clustering**, despite the potential limitations of categorical data fitting. Overall, this analysis culminates in **strategic insights** to guide decision-making, underpinned by a range of **analytical techniques**.

SEGMENTING CUSTOMERS: RFM AND ADVANCED ML TECHNIQUES

Upon implementing the **RFM (Recency, Frequency, Monetary Value)** methodology, the analysis adopts a thorough **data preparation approach**, employing **data pre-processing techniques** to ensure the accuracy and dependability of subsequent stages. With this groundwork established, the focus shifts towards **establishing distinct customer categories based on the perceptive attributes derived from RFM analysis**. This pivotal step forms the bedrock of **the segmentation process**, facilitating a more personalized customer targeting strategy. Central to the analysis is **the classification of customers**. Initially, the model is trained using data from the first ten months, **partitioned into training and test sets**. A diverse array of **machine learning algorithms**, including **Support Vector Machine, Logistic Regression, K Nearest Neighbour (KNN), Random Forest, and XGBoost (Extreme Gradient Boosting)** are harnessed to effectively group customers into meaningful segments. To enhance the performance of these algorithms, **a strategic approach employing a Voting Classifier is introduced**. This approach yields a heightened level of predictive accuracy and robustness in **the segmentation outcomes**. Crucially, the analysis goes beyond this by subjecting its conclusions to **rigorous testing**. This **testing phase** critically evaluates the effectiveness of **the segmentation**, ensuring that the insights generated not only hold theoretical merit but also demonstrate practical resilience. This thorough **validation process** bolsters the credibility of **the customer segmentation model** and instils confidence for businesses to incorporate these insights into their strategic endeavours.

The algorithm works as follows:

1. Apply RFM methodology (Recency, Frequency, Monetary Value).
2. Calculate recency, frequency, and monetary metrics.
3. Categorize products based on attributes.
4. Group products into 5 main categories.

5. Analyse customer habits over 10 months.
6. Classify customers into 11 categories.
7. Split data into train/test sets.
8. Use SVM, Logistic Regression, KNN, RF, XGBoost.
9. Apply Voting Classifier for improved accuracy.
10. Test segmentation accuracy.
11. Compare predictions with real categories.

In conclusion, the study **culminates** in **customer segmentation**, wherein products and clients are grouped based on **purchasing patterns**. The resulting **classifier**, utilizing **basket averages** and **category spending percentages**, achieves a **75% accuracy rate** during evaluation on the **final two months** of data. The study **recognizes** the impact of **unaccounted seasonal influences** and recommends a **more extensive data collection period** to **enhance predictive outcomes**.

EXPLORING MARKET BASKET ANALYSIS WITH ASSOCIATION RULE-MINING ALGORITHMS

Market Basket Analysis is a data mining technique that uncovers item relationships within transactions, aiding businesses in understanding customer buying behavior and enhancing strategies. This report explores the application of Association Rule Mining in Market Basket Analysis, focusing on key steps and metrics.

Terminologies:

- **Support:** Support measures the relative frequency of a specific itemset in a dataset, indicating how often the items appear together in transactions.
- **Confidence:** Confidence quantifies the likelihood that a consequent item is purchased when the antecedent item is purchased, indicating the strength of the association between the items.
- **Lift:** Lift measures the ratio of the observed support of an itemset to the expected support if the items were independent, revealing the strength of the association between items beyond chance.
- **Antecedent and Consequent:** In association rules, the antecedent refers to the item or items that are present in the transaction, while the consequent refers to the item that is likely to be present as a result of the antecedent's occurrence.

Methodology: The methodology of this study involved a systematic approach. First, a transaction dataset was utilized for analysis, providing data on customer purchases. Following this, the dataset was pruned to identify frequent itemsets by applying a specified support threshold. Lastly, Association Rule Mining techniques were employed to uncover relationships between items in transactions, leveraging algorithms like Apriori.

Transaction Dataset: In this step, data was collected encompassing customer purchases, including pertinent information like products, customers, and purchase dates. This dataset formed the basis for subsequent analysis.

Pruning for Frequent Itemsets: The dataset underwent a process of pruning where infrequent items were removed, resulting in a streamlined dataset enriched with frequent itemsets. This step involved the setting of a minimum support threshold to ensure the relevance of retained item combinations.

Association Rule Mining:

Bread + Milk + Egg ➔ Basket 1

Bread + Milk + Oats ➔ Basket 2

Bread + Milk + Wheat ➔ Basket 3 …

"Bread + Milk" as one rule

Association rule learning, a rule-based approach, reveals significant connections in vast databases by generating both established and novel rules. Utilizing techniques like Apriori, it identifies item associations within pruned datasets, unveiling valuable relationships spanning transactions. The resultant dataframe highlights the leading 15 association rules based on confidence, providing insights into frequently co-purchased items and their association strength. These rules offer guidance for optimizing product placement, promotions, and cross-selling tactics, ultimately enhancing customer experiences and driving sales growth.

Conclusion of Market Basket Analysis: The Market Basket Analysis process yielded insightful outcomes. Item co-occurrences were revealed, shedding light on patterns and relationships within customer purchases. This conclusion was supported by the application of metrics like Support, Lift, Confidence, and Conviction, which quantified these associations.

Metrics: Support, quantifying itemset frequency in transactions; Lift, gauging association strength between items; Confidence, measuring the reliability of association rules; and Conviction, indicating the effect of item independence, were key metrics used to analyze and interpret the associations.

Benefits: The insights garnered from Market Basket Analysis offered several benefits. Businesses could optimize product placement strategies and enhance cross-selling techniques based on identified item relationships. Moreover, promotional efforts could be refined to align with the associations discovered through this analysis.

Conclusion: Market Basket Analysis, facilitated by Association Rule Mining, emerged as a powerful tool to unveil customer behavior trends. The utilization of metrics like Support, Lift, Confidence, and Conviction provided a robust foundation for identifying significant item relationships. This comprehensive approach underscored the significance of data-driven decision-making, elucidating hidden patterns and enriching customer experiences.