****

**By**

**Sunil Raj Thota**

**Group Alpha  
ALY6110 - Data Management and Big Data**

**Class Number: 202135**

**Class Name: Spring 2021 CPS Quarter – Second Half**

**CRN: 80524**

**Customer Segmentation with Online Retail Data**

**Week 3 Assignment – Final Project Proposal**

**thota.su@northeastern.edu**

**Summary:**

In this assignment, I am going to propose an idea and how I am going to implement it in my final group project. We took a real data set with a real problem and to solve this real-world problem, we plan to employ big data platforms and tools such as Apache Spark, Cloudera, Hive, Hadoop, Hue, Impala, and others to process, store, query, and analyze a vast amount of data. We will also use additional analytical tools such as Python/ R in addition to these. For this, we try to use as many ways as possible to analyze and predict/ recommend the best strategies for the online retail company. More customers equal more revenue. To be honest, it was our first time dealing with such a large amount of data, but we thought it would be great to experiment and learn new approaches. Our goal is to perform Market Analysis on this data set and determine a new study on how to use Spark to derive business-valued insights for an online retailer. Intuitively, it's reasonable to infer that a company's ability to attract a growing number of clients is proportional to its sales growth. As is typical of big data projects, a major portion of the study consists of data cleansing and basic exploratory analysis. Increasing the number of customers, on the other hand, might be a vanity statistic that fails to deliver on its promise of driving revenue. Following those fundamental processes, we can deduce more complicated customer behavior patterns, such as which products are usually purchased together.

**Content:**

**Problem:**

Analysts all around the world are coming up with new ways to leverage accessible data connected to online retail to create models, analyses, and graphs that can help understand the size of client purchasing behavior in various locations and, to some extent, make forecasts. Data allows us to have a deeper understanding of the aspects that influence the internet market. An e-commerce company wants to segment its customers and determine marketing strategies according to these segments. To this end, we will define the behavior of customers and create groups according to clusters in these behaviors. In other words, we will take those who exhibit common behaviors into the same groups and try to develop special sales and marketing techniques for these groups.

**Objective:**

For customers, online shopping is becoming a prerequisite for a safe, sustainable, and convenient shopping experience. A Big Data system will be developed as part of the project to assess and offer data-driven strategies and solutions for discovering diverse patterns.

**Methodology:**

There are numerous datasets of various sizes and dimensions that are utilized in retail analytics research and analysis. In this, we will use UCI Machine Learnings open-source data repository on Online Retail II Data Set to find out things. Within the Cloudera environment, this data will be used in Spark using the PySpark library. Spark will also be utilized to create machine learning python code that can train a model using classification methods to forecast whether or not an area will become a retail hotspot. Also, Spark SQL can be used to create relational tables. Integrate Tableau/ Power BI to create a visualization to determine the patterns.

**Dataset Description:**

I'd want to work on a project using a massive data set of 1067371 rows! This Online Retail II data collection includes all transactions made by a UK-based and registered online retailer between December 1, 2009, and December 9, 2011. The company primarily sells one-of-a-kind all-occasion gifts. Most of their customers are wholesalers. Wholesalers make up a large portion of the company's clientele. Think of it as promotional items. The description implies that the sales transactions are a mix of wholesale and retail transactions; nonetheless, this distinction is critical because wholesale and retail clients' purchasing behavior differs greatly.

We saved the data set as a .csv file and went over all of the rows quickly. Because of the dimensionality issue, we recognized that we might not be able to see the most crucial patterns and features. However, we found that several cells are blank, indicating that values are missing. For the sake of convenience, they were recoded into NA. We also discovered where and how many missing values there are.

**Attribute Details:**

* **InvoiceNo:** Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
* **StockCode:** Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
* **Description:** Product (item) name. Nominal.
* **Quantity:** The quantities of each product (item) per transaction. Numeric.
* **InvoiceDate:** Invoice date and time. Numeric. The day and time when a transaction was generated.
* **UnitPrice:** Unit price. Numeric. Product price per unit in sterling (Â£).
* **CustomerID:** Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
* **Country:** Country name. Nominal. The name of the country where a customer resides.

**Key Steps:**

Recency, frequency, and monetary (RFM) analysis is a well-known and effective database marketing strategy. It's a popular way to rank clients based on their previous purchases. RFM analysis is used in a variety of applications that involve a large number of customers, such as online shopping, retailing, and so on. This method divides customers into three groups based on three dimensions.

**Expected Outcome:**

* Targeting New Users
* Creating an atmosphere of trust
* Increasing the Frequency of Purchasing
* Lifetime segmentation

Certain assumptions must be kept in mind before acting on actionable recommendations, based on our analysis and the limitations of our dataset. Before making a judgment, data trends should usually be confirmed with complementary tests.

**Conclusion:**

To remedy this challenge, it is recommended that the present dataset be combined with information about each customer's demographics and shopping behaviors for future research. This will provide useful information to the clustering method, which can be used to further segment customers, and the regression model, which can be used to spot relevant associations among the variables in the data.

Customer segmentation is an important marketing approach that businesses should employ to gain a better understanding of the market and make more informed decisions to increase sales. This is a simple but effective approach that businesses can use. Finally, to optimize our marketing success, we must keep the RFM client segmentation up to date.

**Dataset Source:**

The dataset was retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

**References:**

[1] Online Retail II Data Set. (Sept 2019). *UCI Machine Learning Repository*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Retail+II