UnSupervised Learning Competition

Rules:

- Teams of maximum 3 people [recommended]. Doesn't have to be your project team.
- You can discuss general ideas with other teams but not repeat their exact analytics process.
- One submission per team.
- Don't use any external data sources.
- Your submission MUST include:
    i. A 1-4 pages report stating: (1) Team members' names and student IDs, (2) the used software packages, (3) instructions on how to download and install them, (4) a detailed description of the analytics process (Exploration, preparation and modeling) used with justification for each step, (5) a brief description of the other approaches you tried but didn't work out and (6) the accuracy you achieved (Must show the relevant clustering assessment metrics).
    ii. All code developed to produce the predictions.
    iii. Clusters' distribution, statistics and description as mentioned below.
- Create a folder for your submission and compress the folder into an archive that you upload to OnQ.
- Late submissions will be penalized 1 point for each late day.

Customer Segmentation for Online Retail

For the past 10 years, we have witnessed a steady and strong increase of online retail sales. According to the Interactive Media in Retail Group (IMRG), online shoppers in the United Kingdom spent an estimated £50 billion in year 2011, a more than 5000 per cent increase compared with year 2000[1]. This remarkable increase of online sales indicates that the way consumers shop for and use financial services has fundamentally changed.

Compared with traditional shopping in retail stores, online shopping has some unique characteristics: each customer's shopping process and activities can be tracked instantaneously and accurately, each customer's order is usually associated with a delivery address and a billing address, and each customer has an online store account with essential contact and payment information. These desirable, special online shopping characteristics have enabled online retailers to treat each customer as an individual with personalized understanding of each customer and to build upon customer-centric business intelligence. In relation to customer-centric business intelligence, online retailers are usually concerned with the following common business concerns:

- Who are the most/least valuable customers to the business? What are the distinct characteristics of them?
- Who are the most/least loyal customers, and how are they characterized?
- What are customers' purchase behavior patterns? Which products/items have customers purchased together often? In which sequence the products have been purchased?
- Which types of customers are more likely to respond to a certain promotion mailing?

RFM (Recency, Frequency, Monetary) analysis[2] is a marketing technique used to determine quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary). RFM analysis is based on the marketing axiom that "80% of your business comes from 20% of your customers." With the advent of e-mail marketing campaigns and customer relationship management software, RFM ratings have become an important tool. Using RFM analysis, customers are assigned a ranking number of 1,2,3,4, or 5 (with 5 being highest) for each RFM parameter. The three scores together are referred to as an RFM "cell". The database is sorted to determine which customers were "the best customers" in the past, with a cell ranking of "555" being ideal.

Although RFM analysis is a useful tool, it does have its limitations. A company must be careful not to over-solicit customers with the highest rankings. Experts also caution marketers to remember that customers with low cell rankings should not be neglected, but instead should be cultivated to become better customers.

[1] Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197-208, 2012

[2] http://searchdatamanagement.techtarget.com/definition/RFM-analysis

Given a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Attribute Information (Online Retail.xlsx):

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

Objective:

Segment the customer into various **meaningful groups** using clustering algorithms on the RFM metrics, and **clearly identify the main characteristics** of the consumers in each segment. Based on your segmentation, a customer-centric marketing strategy will be devised to reach out to different customer types.

In your report:

- Show the distribution of the customers within each cluster (Ratio & Percentage of customers per cluster divided by the total number of customers)
- Show the statistics of each cluster (Min, Max, Median of clustering RFM attributes)
- Describe the distinct features of each cluster. For example, a cluster identifies the least or most profitable customer group, the loyal (most frequent) group, old customers with no recent purchases (low recency, high frequency and medium monetary), or potential highly profitable customers (recent and medium monetary).

Hints:

- Group same customer transactions (multiple records for same customer) into a single record (one record per customer).
- Create the three aggregated variables Recency, Frequency and Monetary (Sum, Min, Median, Max) for each customer.
    - **Recency**: Number of months between the last order and the last date of the study (09/12/2011). Zero is most recent.
    - **Frequency**: Number of orders made in the study period (Count transactions per customer).
    - **Monetary:** Money value for the purchases per customer.
- Clustering algorithm can be sensitive to outliers or variables that are of incomparable scales or magnitudes.
    - Outliers should be isolated from the majority and treated separately.
    - Incomparable scales variables should be normalized before the clustering analysis.