Data Mining Approach to Target Customers for Marketing Campaigns Based on Customer Segmentation

Venkat Vijay Gudapati, Mahidhar Reddy Narala, Mauricio Espinoza

1. Introduction

With rapid growth in technology and increase in the data stored by the enterprises, it has changed how they manage their marketing strategies. In the recent times we have observed how highly competitive Ecommerce industry has become and how difficult it is to retain an existing customer as well as to acquire a new customer. To design marketing strategies and design advertisements company should be able to understand characteristics of their customers. Enterprises are investing in CRM enabled with Big Data to store customer related data which can be used in identifying needs, preferences and behavior of their customers. Customer segmentation is a process of identifying customers with similar characteristics and club them in the same group. These characteristics can be obtained from customer's behavior such as transaction frequency and monetary value. Clustering is the most common method in customer segmentation literatures. They are developed with similarity and distance so that the data with similar characteristics is grouped near to each other.

The main goal of this project is to develop a comprehensive algorithm to group the customers into well defined groups and build a real time predictive machine learning model to predict which group a customer belongs to. Customer segmentation is done through behavioral characteristics such as recency, frequency and monetary value.

2. Literature Survey

Initially proposed by Arthur Hughes, the RFM (Recency Frequency Monetary) model is widely used to understand customer behavior. The model refers to the "recency of last purchase, purchase frequency, and monetary value of purchase." Previous research has shown that the greater the *R* and *F* value, the more likely a customer will create a new transaction with the company or business. The *M* value corresponds to the likelihood of a customer conducting repeated interactions with a company again. Previous research suggests each variable could have differing importance based on industries conflicting with Hughes's idea of equality (Wu, et al., 2020).

Customer segmentation is used to provide more "personalized" marketing strategies to customers. Clustering, or grouping a set of objects (customers) in such a way that objects in the same group, is used to achieve customer segmentation. However, there are some issues with using clustering algorithms (Maulina, Surjandari, & Rus).

Customer Segmentation using Centroid Based and Density Based Clustering Algorithms, compares k-means and density-based clustering algorithms (DBSCAN) for customer segmentation. In the article, it is shown that DBSCAN provides more meaningful customer segmentation than k-means (Hossain, 2017).

In the paper by Li, the authors noted that the widely used k-means algorithm is sensitive to the initial set of selected centroids. Thus, the quality of the final clusters can vary widely depending on the centroids' selection. The authors propose an improved version of the k-means algorithm to mitigate the drawbacks of the original k-means algorithm. Their research suggests the improved algorithm improves customer

segmentation accuracy (Li, 2013). However, in our chosen paper, the authors compare centroid-based algorithms with Fuzzy C Means to find the optimum cluster number.

3. Algorithms to be Implemented

Initially we will be using standard clustering algorithm K-Means and K-Medoids and set the one which is performing better as a baseline model. Further we will be using centroid based clustering algorithm compare it with the baseline model.

Apart from a standard clustering algorithm we can also group the customers based on recency, frequency and monetary value with ranking methodology which divides each feature into quartiles and rank each quartile.

Once we have decided on the clusters i.e., the labels we will be using a tree-based algorithm - Random Forest such that feature importance can be extracted. Along with that we intend to explore other algorithms such as SVM, XGBoost and a Stack Model.

A stacked model is an algorithm which utilizes the predictions of the previously implemented algorithms as meta learning data and predicts on top of those predictions.

4. Experiments and Analysis

Dataset

This is 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 (Chen, Online Retail Data Set).

Attribute Information:

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.
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 each transaction was generated.
UnitPrice: Unit price. Numeric, Product price per unit in sterling.
CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country: Country name. Nominal, the name of the country where each customer resides.

Data Analysis

Data preparation is at the heart of the data mining process. Preprocessing data to improve the efficiency and ease of the data mining process has become an important problem.

1. Data Preparation

The real-world data is never consistent so a series of data cleansing steps have to be followed to handle the noise in the dataset.

i. Handling missing values
ii. Handling the outliers
iii. Data type corrections.
iv. Data inconsistency.

2. Feature Engineering
   Derive meaningful features from the exist features which can potentially useful for the machine learning model in identifying the patterns between the customers.

Experiments:

Once the data preparation and feature engineering steps are completed. We will follow a series of experiments as below

K-Means vs K-Medoids vs centroid-based clustering algorithms

To create a baseline model, we will be using K-Means and K-Medoids on the customer's behavioural data such as recency, frequency and monetary value and determine the number of clusters based on the elbow curve plot. Results are to be compared and best performing model is to be selected as baseline model.

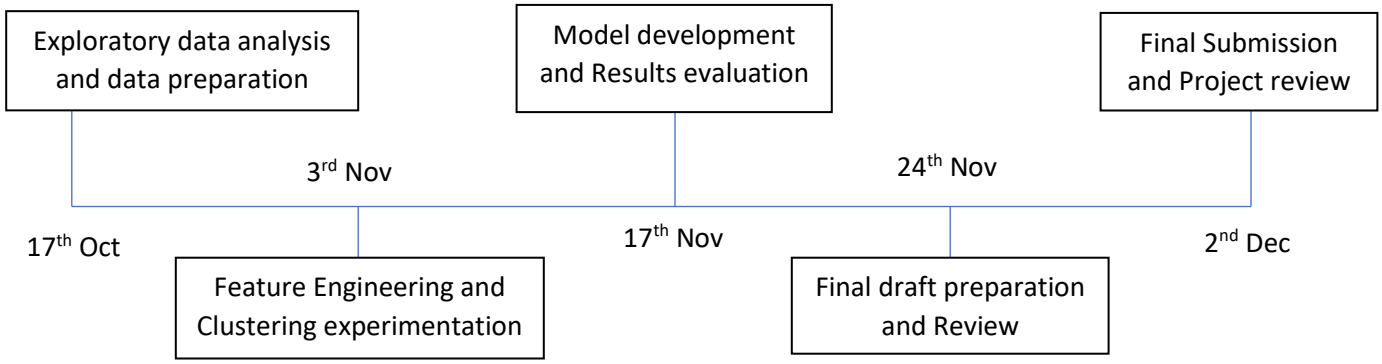
The baseline model is then evaluated along with the centroid based clustering algorithm to decide on the final approach.

RFM Ranking methodology

Each feature in divided into quartiles and a rank is assigned to each quartile such as for recency customers in the first quartile will be ranked higher because of they made purchases recently. In a similar way frequency and monetary value is also ranked and by grouping these ranking features we can extract the final rank for each customer.

We intend to experiment with this approach and compare the results with clustering algorithms.

5. Timeline



References

Chen, D. (n.d.). Online Retail Data Set. London, UK: School of Engineering, London South Bank University.

Chen, D., Sain, S. L., & Guo, K. (n.d.). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of database marketing & customer strategy management, 19*(3), 197-208. doi:10.1057/dbm.2012.17

Du, X.-P. (2006). Data Mining Analysis and Modeling for Marketing Based on Attributes of Customer Relationship.

Hossain, A. S. (2017). Customer segmentation using centroid based and density based clustering algorithms. *3rd International Conference on Electrical Information and Communication Technology (EICT)*, (pp. 1-6). doi:10.1109/EICT.2017.8275249

Kohavi, R., & Parekh, R. (2004). Visualizing RFM Segmentation. *Proceedings of the Fourth SIAM International Conference on Data Mining.* doi:10.1137/1.9781611972740.36

Li, G. (2013). Application of Improved K-means Clustering Algorithm in Customer. *Applied Mechanics and Materials*, (pp. 1081-1084). doi:10.4028/www.scientific.net/AMM.411-414.1081

Maulina, N. R., Surjandari, I., & Rus, A. M. (n.d.). Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering. *2019 16TH INTERNATIONAL CONFERENCE ON SERVICE SYSTEMS AND SERVICE MANAGEMEN.*

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical problems in engineering*, 1-7. doi:10.1155/2020/8884227