

DSC 630 – DATA SELECTION & PROJECT PROPOSAL

Project Milestone - 2



June 16, 2022

Karthikeyan chellamuthu, SUbhashini natarajan

**Objective:**

The business problem that we plan to build the model for is, customer segmentation, for the Brazilian e-commerce company, Olist. Customer Segmentation is an important step in marketing for an organization. The process helps in customizing marketing campaigns, prevent customer churn, prioritizing product development or services.

**Data Source:**

Data for the project is sourced from Kaggle. This is a public dataset of orders made at Olist store. This dataset has several dimensions including, customer, geolocation, order, payment, reviews, product, product category and sellers. For the purposes of customer segmentation, not all dimensions will be considered, but only the following will be used -

* Customer,
* Order
* Customer review

Link to the dataset:

<https://www.kaggle.com/code/marianakralco/brazilian-ecommerce-clustering-rfm-and-kmeans/data>

**Process:**

The process to be followed for the project is laid out below -

1. Data ingestion – Load the multiple data files into panda dataframes.
2. Exploratory Data Analysis – Perform EDA to understand the dataset. Plot essential graphs for the understanding.
3. Data transformation – Based on EDA results, perform necessary data transformation required for modeling.
4. Feature selection – Identify and select the features required for the model building.
5. Build the model.
6. Evaluate the model.
7. Determine if the model can be deployed for solving the business problem.

**Model selection:**

The model that we plan to build for customer segmentation is k-means clustering. The model helps in classifying the customer groups based on purchase history and customer review.

We chose K-means clustering due to the following reasons –

* Scalability for large datasets – The data set we chose has three years sales data.
* Familiarity and relatively simpler to implement.
* Possibility for generalization

**Model evaluation:**

While clustering algorithms have multiple evaluation metrics like, homogeneity score, v-measure, silhouette coefficient, we intend to use, Adjusted Rand Index (ARI), due to the possibility of larger number of clusters with three years data.

**Learning objective:**

Through our prior projects we have learnt, building linear regression for trend analysis, logistic regression for classification problems. By performing this project, we aim to learn the process involved in customer segmentation through clustering. We aim to learn to use elbow method to determine the optimal number of clusters (optimal k-value), determine the clusters, evaluate the model and interpret the various clusters. Also, will learn to plot the clusters for better visualization.

**Risks:**

The three year dataset, if has data quality issues or has imbalanced classes will be a risk to the project. Secondly, if the clustering score is not good, then it implies that k-means will not be the correct clustering algorithm to use for segmentation.

**Contingency plan:**

If the classes are imbalanced then will use SMOTE to balance them. If ARI has close to 0 value, it implies that k-means is not the correct clustering algorithm. In this case will manually determine RFM (Recency, Frequency, Monetary value) and cluster the customers based on RFM score.