DSC 630 – Predictive Analytics

Course Project

Kimberly Cable

(Individual)

Fall, 2022

Data Selection and Project Proposal

Overview

Olist is a Brazilian e-commerce company that connects small businesses to a larger marketplace. It gives these small businesses a way to manage their products, shipping, and online payments. They have approximately 200,000 users in about 180 countries.

Business Problem

With any online retailer, retaining customers is key but knowing who may leave and who stays could be guesswork. Understanding how and why customers stay and shop and why they leave is pivotal to a company's business.

Knowing who your customers are and how and why they shop or do not return plays a big part in customer service which ultimately enhances a customer's satisfaction level. A high satisfaction level could lead to overall better reviews and with these good reviews, their sellers will see more new sales.

This study will hope to find similar traits among its customers. This will help the marketing team know who best to send offers to or to whom might they need to send a discount coupon because it looks like they haven't purchased in a while.

Data

The data has been sourced from Kaggle and has 100,000 online orders from 2016 to 2018. The data consists of records with products, customers, and review information for each transaction provided by Olist.

The data consists of the following datasets (see Figure 1):

- Customers: This dataset has information about the customer and their location.
- Geolocation: This dataset has information about the Brazilian zip codes and their latitude/longitude coordinates.
- Order Items: This dataset has information about the items purchased within each order.
- Order Payments: This dataset has information about the order payment options.
- Order Reviews: This dataset has information about the reviews made by the customers.
- Orders: This dataset has information about all customer orders.
- Products: This dataset has information about all the products sold by Olist.
- Sellers: This dataset has information about the sellers that fulfilled the orders made at Olist.
- Category Name Translation: This dataset has English/Portuguese translations for all products sold at Olist.

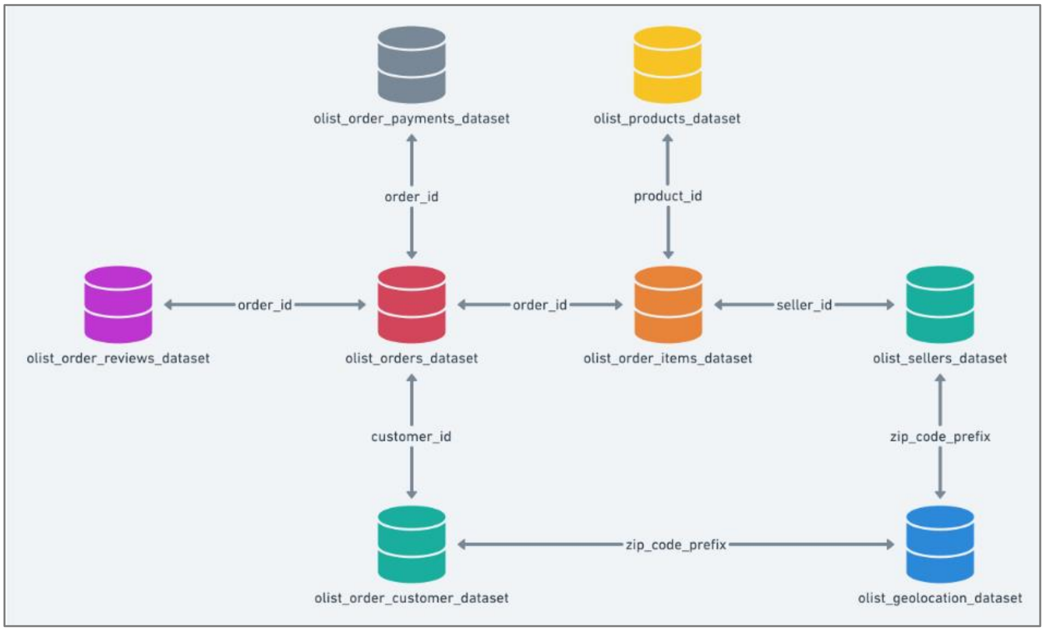

Figure 1: Data Schema

Model Selection

I plan to use the k-means clustering method. K-means is used to help identify clusters or groups within your dataset. This model will help group customers based on previous purchases and reviews that they made. I will also investigate RFM (recency, frequency, monetary) analysis. It is a marketing technique used to rank and group customers based on the number of times the customer has purchased, the last time the customer has made a purchase, and the total dollar amount the customer has spent.

Model Evaluation

To make sure my model is performing correctly, I will use the elbow method and/or the silhouette coefficient to make sure I am using the correct number of clusters. As there may be other ways to evaluate the k-means model, I will research more options.

Learning Objective

I hope to learn more about the customers who shop with Olist. This will help sellers target different types of individuals by grouping them into different categories to know whom they need to target with promotions and whom they need to go after to bring them back.

Risks

As with any dataset, there may be inaccuracies with the data, data may be missing or invalid. Also, the model I have chosen may not be the correct choice and another may be more suited for this data. Further analysis may need to be done.

Contingency Plan

If k-means clustering is not showing good accuracies for clustering, I may investigate another method for segmentation. Re-examining the data may also help. Removing outliers, narrowing the feature selection, etc. may help with model building and evaluation.

References

Arvai, K. (n.d.). *K-Means Clustering in Python: A Practical Guide*. Retrieved from Real Python: https://realpython.com/k-means-clustering-python/

Begam, S. (2021). *Customer Profiling and Segmentation – An Analytical Approach To Business Strategy In Retail Banking*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/03/customer-profiling-and-segmentation-an-analytical-approach-to-business-strategy-in-retail-banking/

Kaggle. (n.d.). *Brazilian E-Commerce Public Dataset by Olist*. Retrieved from https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Pagan Research. (n.d.). Retrieved from olist: https://paganresearch.io/company/olist

Selvaraj, N. (2022). *How to Build Customer Segmentation Models in Python?* Retrieved from 365 Data Science: https://365datascience.com/tutorials/python-tutorials/build-customer-segmentation-models/

Wright, G. (n.d.). *RFM analysis (recency, frequency, monetary)*. Retrieved from TechTarget: https://www.techtarget.com/searchdatamanagement/definition/RFM-analysis