Machine Learning Engineer Nanodegree
Capstone Proposal
Deasy Indrawati
January 14th, 2021

Proposal

Domain Background

Giving offers is one of marketing techniques for promoting new products or services. Starbucks gives rewards to their customers to learn their buying behavior. The rewards are in 3 forms. There are advertisements, buy one get one and discounts. Customers get different rewards. Not all customers get the rewards. The rewards itself have validity, for example BOGO expired in 5 days. They also give demographic data contains gender, age and income. Through this data I need to clustering demographic groups that give the best response to each reward. To do that I will use K-Means Clustering algorithms. K-Means Clustering solve clustering tasks to get better understanding of customer[1].

Problem Statement

By using the data, we need to find the best demographic group that gives the best response to specific rewards. Giving reward to the right customer could be effective ways to promote the new products and increase sales. The way to get that is by clustering the data provided. The data provided is simplified version of Starbucks apps and for one product. Actually, this condition makes us more focus. Later, the solution might be help for their others product to find the effective promotion based on demographic.

Datasets and Inputs

For this project, we are provided with 3 datasets: portfolio, profile and transcript. Portfolio files contain offer id and meta data about each offer (duration, type, etc). Profile files contain demographic data of each customer. Transcript files contain records for transactions, offers received, offers viewed, and offers completed. From portfolio files we can get detailed information about difficulty to get reward from the offer.

From the profile files, there are 17000 data and 5 columns. There are 2175 data with status N/A. Amounts of data with "null" income also 2175 data. The data with N/A value will be cleaned. Beside that outlier data inside the profile files also be removed. Data from profile files will be clustered the customer. After that, clustered data is useful for the Starbucks to give the right offer to customers who have big chance to complete the offer.

Solution Statement

To get higher percentage of customer who complete the offer, we will cluster the data. First thing that need to be done is cleaning the data. We need to remove null data and outliers from the lists. After the data clean, we will merge the files profile and transcript with the key customer id. Data ready to be clustered. The clustered data will contain group of customers based on their demographic who give higher percentage to complete the offer.

This data will be helpful for the Starbucks to give offer to the right customer that had higher chance to complete the offer.

Benchmark Model

There are other clustering algorithms that could be used to do customer segmentation. One possible to use is Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. But K-Means is the most compatible for clustering because DBSCAN is used to clustering different shapes and sizes from large amount of data. Beside that DBSCAN is good for data that is containing noise and outlier.

Evaluation Metrics
(approx. 1-2 paragraphs)

Evaluation metrics will be used is Silhouette coefficient. Through this evaluation we will find out how sense and separate the data clustered. The higher the coefficient, the better the result. To count this coefficient, we need to have:
a: The mean distance between a sample and all other points in the same cluster.
b: The mean distance between a sample and all other points in the next nearest cluster.

Project Design

Starbucks is big company and sell famous coffee in the world. They also have a lot of offer for their customers. But they need to find out the most effective rewards for their customers based on their demographic. Cleaning data is must have procedure before processing it. Outliers and null value will be removed. Profile and portfolio files will be merged with the key customer id. After data ready, we will cluster them. There will be around 15000 data clustered. Clustered data must have special demographic criteria. They will be sorted based the question, "will they have higher chance to complete the offer?"

K-Means algorithm will be used to clustering customer. This method is the most compatible to identify customer who might complete the offer. First thing to do is specify the number of clusters. Second step is initializing centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement. Last, keep iterating until there is no change to the centroids.

There are needs to evaluation for this algorithm, too. Silhouette coefficient will be used to find how highly dense the clustered data. To get the coefficient we need to formulize the mean distance between two nearest clustered data. The result will be between -1 to +1. The score -1 show us that the algorithms used is wrong. The score +1 show us the algorithms used is absolutely correct.

References:
Customer Segmentation Using K Means Clustering
DBSCAN Clustering Algorithm in Machine Learning