Project: Starbucks

Domain Background

This project will focus on the domain of marketing analytics. This is not a new domain, and it is a very active and valuable domain employed by many organizations. McKinsey and Company believes that we are finally reaching a point in advanced analytics to obtain a 5-8x return on investment on marketing spend. SAS describes marketing analytics as a method to "evaluate the success of marketing initiatives", adjust marketing strategies and optimize processes; this accurately describes the goal of our project. I have significant experience in business operations and view this project as an opportunity to increase my skillset.

Problem Statement

This is a classic project of evaluating marketing strategies. A simplification of the problem is a marketing organization that has applied promotions across a wide range of customers without knowledge of whether or not these promotions are successful and without targeting customers with specific promotions. The next logical step is to apply analytics to determine effectiveness and optimize the marketing strategy to positively impact the company's bottom line. Various promotions occurred across different customer segments and analysis on the results can provide valuable insights to employ customer-targeting techniques.

Inputs to this problem are 10 different marketing promotions, basic demographic information of 17,000 customers, and 300,000 interactions of the customers with the app. After clustering the customers into segments, linear regression will determine whether each promotional strategy had a positive, negative, or neutral effect on customer spending within a given segment.

Datasets and Inputs

The following simulated datasets provided by Udacity will be used in performing this analysis: portfolio.json, profile.json, and transcript.json. Portfolio.json is a matrix representing the 10 different types of promotion campaigns. Profile.json is a matrix of 17000 customers. Transcript.json is a matrix of 306,534 events. The features of the dataset are described in the below charts.

Feature	Description
Channels	Email, social media, web, mobile
Difficulty	Effort required by customer to attain (scale 0-20)
Duration	# of days fo validity (3-10)
ID	Unique identifier of promotion
Offer Type	BOGO, discount, or informational

Chart #1: portfolio.json features

Feature	Description
Age	18-118
Date Membership Began	Year-Month-Day for 5 years from 2013-07 to 2018-07
Gender	Female, Male, Other, or None
ID	Unique identifier of customer
Income	Dollar amount from $30,000 - $120,000

Chart #2: profile.json features

Feature	Description
Event	Offer received, offer viewed, transaction, offer completed
Person	Unique identifer of customer - links to profile.json
Time	Hours since test started
Value	Unique identifer of offer - links to portfolio.json

Chart #3: profile.json features

This data will need to be combined into a larger dataset that reflects the transcript file with customer segmentation added in and promotions identified. Each promotion and each customer segment will need to be transformed into separate variables to support linear regression.

Solution Statement

A proposed solution will be to determine whether each promotional strategy has a significant effect (positive or negative) on each customer segment. Based on the datasets, we will cluster the customer data into similar groups using k-means and then separately evaluate efficacy of promotional strategies for each group. Efficacy will be determined through linear regression to determine whether the promotion is statistically significant (rejecting the null hypothesis) compared to no promotion. If it is significant and the coefficient is positive, then the promotion will be considered good. If it is significant but with a negative coefficient then it will be considered bad. If it is not significant then the promotion will be considered neutral. A determination of good, neutral, and bad promotions will become evident for each segment, supporting a potential improvement in overall performance of the Starbucks marketing department.

Benchmark Model

The benchmark for this model will be to naively analyze performance of promotional strategies through the same linear regression models as our complete model, without first doing customer segmentation analysis. This will provide validation of the usefulness of customer segmentation and should reveal that results of regression are more significant for the promotion variable when customer segmentation occurs.

Evaluation Metrics

The clustering algorithm will be measured based on Percentage of Variance Explained (PVE). An elbow graph will determine the appropriate number of clusters in our customer segmentation. For regression, each model will be compared to a null hypothesis to determine whether the variable of the promotion is statistically significant and we will look to prove or disprove this hypothesis.

Outline of Project Design

This project contains 3 different datasets and a significant amount of data. The data must be thoroughly cleaned and explored in order to support our machine learning models. Success in this manner will gain additional insights and drive quality in our final product.

A preliminary workflow will follow these steps:

1. Data cleaning and exploration - missing data in the profile.json file will need to be cleaned and the data in the transcript file will need to be better understood. Transaction.json will be transformed to determine whether a specific purchase was made during a promotional period. Transformation of the features in the data will be required to support regression.
2. Principle Component Analysis (PCA) on profile.json to determine key relationships between variables and rank them in order of importance. Determination of how many components to include in final model.
3. K-means clustering on the PCA dataset to define customer segments. An elbow graph will determine the appropriate number of clusters.
4. Merge the transformed transcript matrix with the newly defined customer segments.
5. Split the merged matrix into separate training and test datasets for each customer segment.
6. Perform linear regression on this data to determine how each customer segment reacts to different promotional strategies vs. the baseline of no promotion.
7. Repeat regression as needed and evaluate results of linear regression to determine if each promotion was good, bad, or neutral for each customer segment.
8. Re-perform analysis by only following steps 1,5,6 and 7, i.e. develop benchmark model. Evaluate benchmark model vs. complete model to observe usefulness of customer segmentation.

Sources

Ariker, M., Díaz, A., & Perrey, J. (2015, November 1). Personalizing at Scale. Retrieved from https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/personalizing-at-scale

What is Marketing Analytics? (n.d.). Retrieved from https://www.sas.com/en_us/insights/marketing/marketing-analytics.html