Proposal

Domain Background

Starbucks was founded in Seattle in 1971. Starbucks main focus is on selling coffee and it became a brand in itself that it started becoming the main coffee brand for most people where Starbucks exist.

Problem Statement

Starbucks wants to analyze data about their app usage made by the customers and discover which are the groups in there and what are the offers that are more appealing to them than others in order to develop an algorithm that responds differently to different customers to maximize on their return. We are trying to build a machine learning model that is fed by different set of information and we

can predict how the customer will respond to the offer (by receiving, viewing or completing the offer)

Datasets and Input

This data set consists of three different files

profile.json

users

gender: (categorical) M, F, O

age: (numeric) missing value encoded as 118

id:

became_member_on: (date) format YYYYMMDD

income: (numeric)

portfolio.json

Offers sent

reward: money awarded for the amount spent

channels: web, email, mobile, social

difficulty: money required to be spent to receive reward

duration: time for offer to be open, in days

offer_type: bogo, discount, informational

id:

transcript.json

Event

person:

event: offer received, offer viewed, transaction, offer completed

value: different values depending on event type

offer id: not associated with any "transaction"

amount: money spent in "transaction"

reward: money gained from "offer completed"

time: hours after start of test

Solution Statement

Using the data provided, The solution statement I am aiming to discover the main drivers of offer effectiveness, and explore if we can predict the likelihood of viewing, receiving or completing the offer by encoding 'event' data to numerical 'offer received':1, 'offer viewed':2, 'offer completed':3.

Benchmark Model

The benchmark model I would use would be the Naive Bayes Classifier. The naive bayes classification algorithm (Gaussian) is a simplified assumption which tries to establish simple heuristics based on the data, which also requires small amount of training data to estimate the necessary parameters. Despite their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering.

Evaluation Metrics

An evaluation metric that would be used on this problem would be the F1-score. Our problem to solve is not that sensitive which requires very high F1 score, so

the scores are good & sufficient and can be used for the classification purpose to predict whether a customer will respond to an offer.

Project Design

The workflow for the project would go by this:

1. Understanding the Data (Data Exploration) : Big key to solve this problem is to fully understand the data. Profile and portfolio seem simple to understand

2. Analyze the Data (Exploratory Visualization) : The data would be analyzed and visualizations constructed to carefully understand what data is inconsistent and where work could be done to handle imbalance

3. Algorithms and Techniques: We prepare the data in a way that helps our model and its purpose.

4. Data Preprocessing: merge all three datasets and change categorical variables and normalize all the dataset

5. Implementation and Metrics: We will use the F1 score as the model metric to assess the quality of the approach and determine which model gives the best results. It is the weighted average of the precision and recall

6. Get our benchmark model: Test a naive classifier as our base Mode

7. Model Evaluation and Validation: We got 62 % for our benchmark naive so we need to do better for our model

8. Justification: The test dataset is used to evaluate the model. Our model is better than the benchmark. The score is created by the Decision Tree Classifier model, as it validate F1 score is 90.46, which is much higher than the benchmark. Our problem to solve is not that sensitive which requires very high F1 score, so the scores are good & sufficient and can be used for the classification purpose to predict whether a customer will respond to an offer.