Capstone Proposal

Starbucks Capstone Challenge

Domain Background

- Within the business strategy of the Starbucks company, we focus on the permeability of its offers for a limited group. The datasets provide us with information on both the demographic characteristics of consumers and their receptivity to different types of offers.

Problem Statement

- Using the information provided in the datasets, we intend to infer which way a specific client will respond to a certain type of offer.
- On one hand, We are able to establish a demographic segmentation based on the different categorical fields such as age, gender, income
- On the other, we can link these demographic segments to the type of offer and their associated receptivity

Datasets & Inputs

The data is contained in three files:

- **portfolio.json** - containing offer ids and meta data about each offer (duration, type, etc.)
- **profile.json** - demographic data for each customer
- **transcript.json** - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

- **portfolio.json**
  - *id* (string) - offer id
  - *offer_type* (string) - type of offer ie BOGO, discount, informational
  - *difficulty* (int) - minimum required spend to complete an offer
  - *reward* (int) - reward given for completing an offer
  - *duration* (int) - time for offer to be open, in days
  - *channels* (list of strings)

- **profile.json**
  - *age* (int) - age of the customer
  - *became_member_on* (int) - date when customer created an app account

    - *gender* (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
    - *id* (str) - customer id
    - *income* (float) - customer's income

- **transcript.json**
    - *event* (str) - record description (ie transaction, offer received, offer viewed, etc.)
    - *person* (str) - customer id
    - *time* (int) - time in hours since start of test. The data begins at time t=0
    - *value* - (dict of strings) - either an offer id or transaction amount depending on the record

Solution Statement

- The objective will be to build a machine learning model that allows identifying, based on the different demographic segments, the result of the offers offered to customers.
- To model the predictions about this problem we required supervised Machine learning algortims. we will use classification algorithms
    - Logistic regression
    - Support Vector Machine
    - K-Nearest Neighbors
    - DecisionTree
    - random forest

Evaluation metrics

	Decision Tree	**Random Forest**	**Logistic Regression**	**Support Vector Machine**	**Naive Bayes**	**K-Nearest Neighbors**
Accuracy	100	100	100	86.7	100	83.8
Precision	100	100	100	91.6	100	83.1
Recall	100	100	100	72.9	100	73.7
F-Measure	100	100	100	81.2	100	78.1

Project design

- Data Cleanning
- Exploratoriy Data Analysis
- Master table consolidation
- Machine learning preprocessing

- ML train&fit the model