Machine Learning Engineer Nanodegree Program

Capstone Project Proposal

Starbucks Capstone Challenge

AHMED SAFWAT EWIDA

DEC ,2019

Table of Contents

01

Domain Background

Domain Background

Machine learning (ML) has become an increasingly important part of IT today. This effect is seen both in how IT leverages machine learning to improve operations and in how IT supports and enables the lines of business (LOBs). Still, organizations have limited understanding on its effective use and have made limited progress in associating it with business outcomes.

Admittedly , The Companies which will lead in the future are those who will be interested in implementing the machine learning algorithms on the enormous amount of data base which they have , they will be the pioneers in their field.

STARBUCKS is one of flagship Worldwide companies which has been established since 31st March 1971 and have worldwide coffeehouse chain, and has a tremendous database of users , that is why I am interested in implementing my capstone project for STARBUCKS Capstone Challenge as I believe that I can implement a good Machine Learning Model for one of the most Worldwide prestigious companies.

Customers' Concerns are the goal for all companies all over the world , what people like? , how much they want to pay?, when do they are capable to pay? , what is the gender and age of those people who are interested and capable to pay? are very important questions and the answer comes from Historical data which we have to implement a deep learning algorithms to it , and building machine Learning Algorithms according to those Historical data to maximize Companies s' profits.

02

Problem Statement

Problem Statement

The Problem Statement as mentioned in **Starbucks Capstone Challenge ,analysing the data set for STARBUCKS Customers and building a Model t**hat predicts whether or not someone will respond and complete to an offer.

We have an enormous number of users , some of them are making transaction either the received or not received an offer ,others are just viewing the offers without completing it , others responding to specific type of offers and completing it.

We have to make analysis for those who are receiving , viewing and make transaction within the offer period and those customers are our target.

Analysing the demographic feature for the above mentioned customers , their gender , age , income , the membership period and the type of offers which they are interested are the most important step before building our Model to stand on the Features which we will use in our Model.

The customers who are not influenced by the offers , or they purchase without having received an offer or seen an offer are NOT in out target.

Cleaning , analysing and Visualizing the data consumes 90 % of the efforts to build a good model.

Problem Statement

The Below flow chart is our target for the users Whom received and viewed and make transaction within the offer period

03

Datasets and Inputs

Datasets and Inputs

We have three JSON 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

portfolio.json: shape (10 rows x 6 columns)

- 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)

Datasets and Inputs

profile.json:shape (2175 rows x 5 columns) with 17000 unique users.

- 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

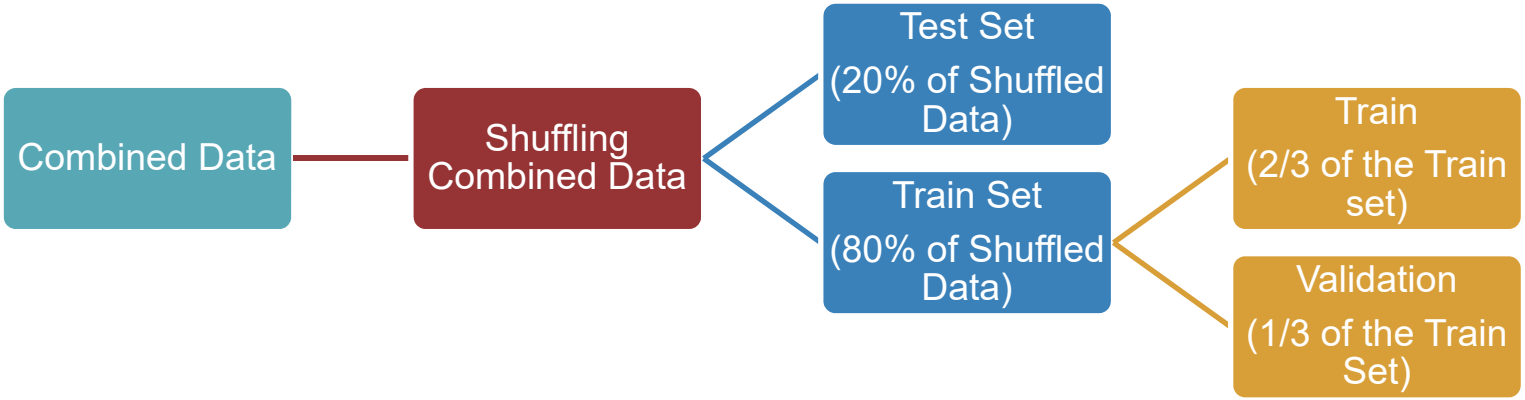
transcript.json: (306534 rows x 4 columns)

- 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

Datasets and Inputs

Combined Data after Cleaning and preparation the Data sets discussed :shape (66501 rows x 40 columns)

- Input features : 39 Columns
- Output Label: 1 Column (“success”) It will be either(“1”) which means that the user react to the offer and successfully completed and transaction happened within the offer period or (“0”) Which means that the user doesn’t react to the offer .



04

Solution Statement

Solution Statement

We will Follow the below process in our Problem Solution:

Solution Statement

- **Fetching the Data:**
  The Data sets mentioned in the previous slide to be converted to CSV Files , and to be ready for next step.
- **Clean /preparation Data:**
  1.Wrangle data and prepare it for training
  2.remove duplicates, correct errors, deal with missing values, normalization, data type conversions, …etc.)
- **Data Visualizing and analysis:**
  1.Visualize data to help detect relevant relationships between variables.
  2.Split into training and evaluation sets
- **Taring Model:**
  The goal of training is to make a prediction correctly as often as possible.
- **Evaluating the Model:**
  1.Uses some metric or combination of metrics to measure the performance of model.
  2.shuffling the data and selecting 20/80 ratio for test/train data set.
  3.Hyper-parameter tuning, which is a corner stone for Model efficiency and performance improvement.
  4.Using test set data which have to predict the output.

05

Benchmark Model

Benchmark Model

- We will use Logistic regression model as a Benchmark in which to compare our models ‘s performance to , because it is fast and simple to implement.
- We will implement the AUC , Precision and Recall Metrics to Compare other Models ‘s Results.

06

Evaluation Metrics

Evaluation Metrics

Our Problem is Classification Problem , that will lead us to use the following Metrics:

ROC_AUC: Area Under Receiver Operating Characteristics curve

Precision : The proportion of positive cases that were correctly identified.

Recall : The proportion of actual positive cases which are correctly identified.

07

Project Design

Project Design

- **Programming Languages , tools and libraries:**
  1. Amazon Sage maker Machine Learning Services
  2. Amazon Sage maker XGBoost built in Algorithm.
  3. sklearn.tree
  4. sklearn.neighbors
  5. sklearn.ensemble
  6. sklearn.linear_model
  7. sklearn.utils
  8. sklearn.metrics

Models Used:

1. Amazon Sage maker XG-Boost built in Algorithm.
2. Random Forest Regressor
3. Decision Tree Classifier
4. K-neighbors Classifier
5. Logistic Regression
6. Cat-Boost Model
7. Light-GBM Model