Capstone Project
Machine Learning Engineer Nanodegree

Rukmini Deshpande
February 17th, 2022

Starbucks Project – Customer behaviour, demographic traits analysis and Prediction

Content:

1. Domain Background.
2. Problem Statement.
3. Datasets and Inputs.
4. Solution Statement.
5. Benchmark Model.
6. Evaluation Metrics.
7. Project Design.
8. References

1. **Domain Background:**

This project is about Starbucks offers and how their customers interact with these offers. Data set given contains simulated data that mimics customer behaviour on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. Starbucks sometimes sends individual messages that contain offers related to its products. These offers can be an advertisement for a new product or an offer such as:

BOGO (buy one get one): User needs to reach to a certain amount to earn a reward of the same amount.
Discount: Discount is a fraction based on order price.

Not all users receive the same offers, sending offers to customers who are not likely to buy their products. On the other hand, we need to attract the new customers, so it is necessary to identify who are the people with a higher probability to respond to that specific offer we send to them. Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only a week. In this project, we will combine transactions data, demographic data and offers to train a model that given a customer and an offer tries to predict whether that customer will respond to that offer. Responding to an offer means viewing the offer and then making the required transactions to complete it while it is active.

2. **Problem Statement:**

Starbucks invests money in offers expecting to have higher profits in back. To combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products. So, selecting the most relevant offer to the correct customers is important to achieve a successful marketing campaign. However, some customers don't buy anything. Other ones don't even see the offer that was sent to them, which may be a problem with the channel chosen.
So, we have four categories:
1. Offer will not be seen nor ordered.
2. Offer will not be seen but accidently ordered.
3. Offer will be seen but not ordered.
4. Offer will be seen and ordered.

What might be a problem with the
1. offer type sent
2. The customer is not the one to be considered as a target.

There are other cases where the customers choose the offer for themselves, which leads them to try new products or spend more money than usual. Those are the situations to be identified and pursued. The problem this project proposes to solve is to predict whether a customer will respond to an offer.

3. **Datasets and Inputs:**

Our datasets here are stored in 3 ". json" files, and they are as follow:

1. portfolio.json:
    1. id (string) - offer id
    2. offer type (string) - type of offer is BOGO (4 records), discount (4 records), informational (6 records)
    3. difficulty (int) - minimum required spend to complete an offer
    4. reward (int) - reward given for completing an offer
    5. duration (int) - time for offer to be open, in days
    6. channels (list of strings)
    7. This dataset contains 10 rows and 6 columns

2. profile.json:
    1. age (int) - age of the customer
    2. became_member_on (int) - date when customer created an app account
    3. gender (str) - gender of the customer (M or F) (14825 M, 14825 F and 14825 not specified) so the data here in balanced
    4. id (str) - customer id
    5. income (float) - customer's income.
    6. This dataset contains 17000 rows and 5 columns.

3. transcript.json:
    1. event (str) - record description (ie transaction, offer received, offer viewed, etc.)
    2. person (str) - customer id
    3. time (int) - time in hours since start of test. The data begins at time t=0
    4. value - (dict of strings) - either an offer id or transaction amount depending on the record
    5. This dataset contains 306534 rows and 4 columns

We will upload these data into S3 bucket to use effectively with Sagemaker. Dataset will be feed to model for training purpose, while test dataset will be used for validation during the training.

4. Solution Statement:

This project will be done using XGBoost algorithm which is one of the best algorithms in Machine Learning Algorithms. The whole project will be done using the Amazon Web Services (AWS) specially Amazon Sage Maker service that I have learned to use during this nanodegree.

Our objectives for this paper are:

1. Determine what demographic traits, if any, influence purchasing behaviour.
2. Determine how well we can predict consumer behaviour after they've viewed a coupon.

5. Benchmark Model:

The benchmark here is logistic regression. Logistic regression: Amazon calls their linear regression and logistic regression algorithms from Linear Learner Class.

6. **Evaluation Metrics:**

   The evaluation metric can be used to quantify the performance of both the benchmark model and the solution model. In this project I will use ROC-AUC.
   Higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

7. **Project Design:**

   This project will be completed with the following steps:

   1. Data processing and exploration:
      a. First will create a Sagemaker notebook instance with fair performance.
      b. Some Exploratory Data Analysis should be applied on the data. This helps us deleting all null values to avoid any errors that we might face after. Also, EDA helps us to find patterns in the data.
   2. Data Feature Engineering: with this step, we do some statistical features for every data point.
   3. Data Splitting: at this step we will split the data into three datasets:
      a. Train Data: to train the model on it.
      b. Validation Data: to validate the model and increase its performance.
      c. Test Data: this dataset is used to calculate the accuracy of the model on data that the model has not seen before.
   4. Train the model: Using sagemaker tuner, we will tune model with best possible hyperparameters and when it is obtained, and when it is obtained will train model to a sagemaker estimator. Here will need to calculate the most suitable hyperparameters for tuning the model.
   5. Test the model: this will be the last step in our project.

8. **References:**

[1] https://www.researchgate.net/publication/256048420_Machine_Learning_for_Targeted_Display_Advertising_Transfer_Learning_in_Action

[2] https://medium.com/@AmishWarlord/predicting-starbucks-customer-behavior-119fc3a43480