Starbucks Capstone Project Proposal

Thien Ngan Doan
Udacity ML Engineer Nanodegree
Jan 2021

Introduction

This Starbucks Capstone project is part of the Udacity Machine Learning Engineer Nanodegree. Udacity partnered with Starbucks provide a real-world business problem and simulated data that mimics their customer behavior.

Domain Background

Starbucks has a reward program that allows customers earning points for purchases. There is also a phone app for their reward program where they send exclusive personalized offers based on customers spending habits.

This project is focused on tailoring personalized offers to the customers who are most likely to use them. The Machine Learning terminology for this is "propensity modeling". Propensity models are "often used to identify the customers most likely to respond to an offer"

With the well-organized related data, machine learning becomes a useful tool to improve revenues for business but also to offer better services. With the smartphone revolution, there are many applications for business services, and those apps collect insightful data about the user behaviors browsing the app that help predicting their needs and provide them the right offers.

This project is a wonderful example, it is a case of study to analyze related data provided by Starbucks and Udacity to make decisions for sending offers to the client.

My goal is analyzing this type of data to apply the same ideas in other similar projects in real productions.

Problem Statement

We want to determine the classes of customers which complete valuable offers to success – in this project we concentrate on two types 'bogo' and 'discount'. Some customers do not want to receive offers and might be turned off by them, some do not view the offers and maybe some fulfill the offer although they never view the offer.

Datasets and Inputs

We use the dataset provided by Starbucks and Udacity. The data consists of 3 files containing simulated data that mimics customer behavior on the Starbucks Rewards mobile app.

1. portfolio.json: information about the offers,
2. profile.json: information about the customers,
3. transcript.json: info about customer purchases and relationships with the offers.

portfolio.json:

In this file, there exist 10 types of offers with related information, such as minimum spend, communicated channels, duration time, amount of reward, …

- 4 types for bogo – Buy One Get One,
- 4 types for discount.
- 2 types for informational.

We will have more research on 2 types of offer "bogo" and "discount" to get the necessary results for our projects.

profile.json

In this file, there exist 17,000 customers with related information, such as: id, gender, age, income, time to become a member.

transcript.json

In this file, there exist 306,534 purchases with information related to the offers and customers.

A customer can receive the offer, view it, and complete it. It is possible that a customer doesn't view the offer or he/she completes it without viewing it.

We process these files to get the whole useful combined data, after that we examine the data to get results of customer classes that complete the most offers of bogo and discount. Next step, we separate the data into two distinct parts for two machine learning prediction models: one for bogo and one for discount.

Solution Statement

We use supervised learning technique to perform our models. Ours are binary classification problems. Many popular models are used to build this kind of model, such as: Logistic regression, Support Vector Machines (SVMs), and Neural Networks, … In this project we use sklearn Logistic Regression to examine in the first step generally, after that we use Autogluon Tabular to get the most powerful prediction of our models.

Benchmark Model

Here we leverage popular and famous benchmark models for our binary logistic regression problems, such as:

1. Sklearn Logistic Regression model: used as a prototype to examine our data for training model. We use only *accuracy* metric as reference point to see the model results with test data.
2. Autogluon Tabular model: very powerful compatible tool for our model. We use six metrics that are completely compatible to our binary prediction models: *accuracy*, *balanced accuracy*, *roc_auc*, *f1*, *precision*, *recall*.

Evaluation Metrics

1. In the first step, as the prototype models for testing our data, we use only '*accuracy score*' metric for sklearn Logistic Regression models.
2. In the second step, as the testing models, with the powerful Autogluon Tabular we have many types of compatible metrics, such as the score of: *accuracy*, *balanced accuracy*, *roc_auc*, *f1*, *precision*, *recall*.
3. In the third step, we implement Autogluon Tabular models on Aws Sagemaker, we use six metrics: *accuracy*, *balanced accuracy*, *roc_auc*, *f1*, *precision*, *recall*.

Project Design

Our project gets data from the 3 json files portfolio.json, profile.json, and transcript.json. We create the working process as following:

I. Part I: Data processing
   1. Examine 3 original json files and logically combine them into a whole raw related data and save in a file "*raw_data.csv*".
   2. From this raw data, we analyze to clean the incompatible and unnecessary data, therefore we get the useful relationship for valuable conclusions. We save this data in a file "*pdata.csv*".
   3. From the data saved in "*pdata.csv*", we refine the data again to get compatible data and separate to the related data for each predicted offer:
      a. *bogo*: we save the data in the file "*bogo.csv*"
      b. *discount*: saved in the file "*discount.csv*";

II. Part II: Modeling
   1. As prototype models, we use Sklearn Logistic Regression to create 2 models to predict:
      a. The success of bogo offers.
      b. The success of discount offers.

The unique metric is used in this model is *accuracy* score.

2. As testing model, we use Autogluon Tabular to create 2 local models to predict:
    a. The success of bogo offers.
    b. The success of discount offers.

    Metrics are used in this model are what attached with the method evaluate of autogluon.tabular.TabularPredictor: *'accuracy'*, *'balanced accuracy'*, *'mcc'*, *'roc_auc'*, *'f1'*, *'precision'*, and *'recall'*.
3. This is the most important part of our project, we implement Autogluon Tabular in AWS Sagemaker to create 2 models to predict:
    a. The success of bogo offers.
    b. The success of discount offers.

    Metrics are used in this model are: *'accuracy'*, *'balanced accuracy'*, *'roc_auc'*, *'f1'*, *'precision'*, and *'recall'*.

References

1. Propensity model.
2. Sklearn Logistic Regression.
3. Autogluon Tabular.
4. Deploying AutoGluon Models with AWS SageMaker.
5. AutoGluon Tabular with SageMaker Examples.
6. Cloud Training with AWS SageMaker - AutoGluon.