Capstone Proposal
Machine Learning Nanodegree - Udacity
Jonatas Oliveira Lima da Silva
Teresina/PI, July 8st, 2021

Domain Background

Starbuck's Capstone Challenge is about trying to predict how people make purchasing decisions and how those decisions are influenced by promotional offers. It's not just a problem for Starbuck, but it could easily expand to any other company or even another area. According to Stupsend and Arandjelovic [1], every day consumers make decisions about whether or not to buy a product. Some decisions are based on price alone, but in other cases the purchase decision is more complex and many other factors can be considered before the final commitment is made. Companies often introduce additional elements to the offering that aim to increase the perceived value of the purchase. Thus, as in this work, Stupsend and Arandjelovic [1] examine these questions using data-driven machine learning, whether specific goals and readily measurable factors influence customer decisions. Stupsend and Arandjelovic [1] used Naives Bayes and Random Forest to make their predictions, and in this work we can use Artificial Neural Network (ANN) or Random Forest, based on previous knowledge acquired during the course.

Problem Statement

We can dividing the problem in three important classes: receiving offers, opening offers, and making purchases. All of this three event is important to the company, because they can say to the company how engaged the client is with the offer. Besides it, we will try to focus on potential clients to making purchases. We can treat this problem as a binary classifier or multi class classifier. For both kind of problem we can use ANN. ANN will be a good choice because of its characteristic of infer values understandable by humans.

Datasets and Inputs

We have three files: profile.json, portfolio.json and transcript.json.

profile.json : Rewards program users (17000 users x 5 fields)

- gender: (categorical) M, F, O, or null
- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)

portfolio.json: Offers sent during 30-day test period (10 offers x 6 fields)

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive reward
- duration: (numeric) time for offer to be open, in days
- offer_type: (string) bogo, discount, informational

- id: (string/hash)

transcript.json: Event log (306648 events x 4 fields)

- person: (string/hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on event type
    - offer id: (string/hash) not associated with any "transaction"
    - amount: (numeric) money spent in "transaction"
    - reward: (numeric) money gained from "offer completed"
- time: (numeric) hours after start of test

The transcript file has the results of the profile and portfolio data. The attribute event from transcript data return an import information about the customer behavior. So the event could be the class of our prediction. The value and time from transcript data could represent an important trade off to the company, and is important to analyse.

Solution Statement

As the algorithm will be ANN, firstly, to get the solution, the qualitative attributes in the data need to be transformed in quantitative attributes. Included in the preprocess data, a normalisation and a removal of outlier can be applied. So, the ANN algorithm can firstly predict wich of the three class the person correspond. An experiment with the trade-off value/time will also made with some statistical visualisation methods.

Benchmark Model

As Benchmark Model to this project we can cite the paper proposed by Stupsend and Arandjelovic [1], which provide some context to the scenario and where the problem was solved using tow other algorithms: Naıve Bayes and Random forest. The results obtained in this project can be compared with [1], and some of the preprocess approach used in [1] could be used in this work. Besides it, as future work, we can also use Random Forest to solve our problem and use ANN to solve the problem defined in [1].

Evaluation Metrics

Accuracy is a classification metric. With a confusion matrix we can obtaining the percentage of misclassified and correctly classified inputs. So we started our evaluation by examining and comparing the performance with the average classification accuracy.

Project Design

The present project has the follow workflow:

1. First, the problem was studied and a search was made for related works in the literature;
2. Data and files available for the project were analyzed. Based on the characteristics of the problem and the data, the main metric to measure performance was defined: Accuracy;
3. An exploration of the data was made. Some statistical graph will be generated, based on statistical concepts. This visualization was helpful in understanding the data and the problem;

4. After understanding the problem and the data, a pre-processing of the data will be done. The algorithm chosen is Artificial Neural Network because this algorithm recognizes the underlying relationship in a dataset based on the human brain. As ANN operates with discrete attributes, all values need to be changed to numeric values. Some noise from the data will also be removed. These noises can impair accuracy.
5. The ANN implementation will be done with the Pytorch library in AWS SageMaker environment. Pytorch is written in Python.
6. With AWS SageMake, we can easily take full advantage of the AWS infrastructure. In addition, AWS SageMake provides us with a sophisticated logging system, which gives us the details of how the algorithm gets the results and how the algorithm can be improved with some extra hyperparameters.
7. The model created with SageMake can be evaluated and validated after processing some input test data. If necessary, some refinements can be made. In this step we will analyze the metrics and if necessary try to define another type of metrics.
8. Finally, the project conclusion will be written. The conclusion is based on the results, but future work can also be defined.

References

[1] Stubseid, Saavi & Arandjelovic, Ognjen. (2018). Machine Learning Based Prediction of Consumer Purchasing Decisions: The Evidence and Its Significance.