Machine Learning Engineer Nanodegree

Capstone Project Proposal

Artem Shuvalov

16 October, 2021

Overview

Domain Backround

In 2021, the need for the data analysis among the modern corporations is more important than ever before. During the COVID-year, the majority of the companies lost the opportunity to serve the customers in offline-stores, making online services the only connection to their clients. Therefore, the importance of the websites and applications has risen to the levels never seen before.

Starbucks was not an exception, but it was prepared with its state-of-art mobile phones application, which allows users to make orders online, predict the queues and send special offers to the clients.

The reason why I took this project is that I am a CEO of TravelTech startup in Russia, which has an embedded recommendation system. So, the problem stated in this Project is relevant to the realms of Machine Learning I work with.

Problem Statement

The Starbucks would like to analyze how different demographic groups respond best to different types of offers in their application. The task is to combine historical transaction, demographic and offer data to determine the algorithm for mapping of the customer and the most relevant for him\her\them offer type. The evaluation of the performance can be measured using the historical data as well.

Datasets and Inputs

There are 3 chunks of data:

1) Portfolio - consists of the information about the offer provided to the client, including it's type (BOGO/Buy one get one, discount or informational/reward-free), difficulty to achieve for the client, reward, duration and channels.

2) Profile - includes the demographical information about the customers.

3) Transcript - contains the information about the actions with the offer in the customer's application as well as transactions.

Solution Statement

The problem to be solved is to find the optimal model, which will help to get the highest conversion of the clients, i.e. maximize the number of people making the orders among those who saw the each type of offer in the application.

Benchmark Model

It is worth to compare the model's results with the current conversion values for each type of offer. There is no reason to include reward-free offers, as the can increase the client's retention, but would not affect the conversion metrics, we chose the most relevant one.

Evaluation Metrics

The conversion is going to be analyzed using the standard Machine Learning metrics:

1) Precision and Recall
The formulas are the following:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

It means the number of correct results divided by the number of all results.
In terms of the Project, it relates to the number of correctly predicted conversions over the number of actions assessed as conversions.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

It means the number of correct results divided by the number of results that have to be retrieved.
In terms of the Project, it relates to the number of correctly predicted conversions over the number of all conversions.

2) F-1 score - combines Precision and Recall.
The formula is the following:

$$Recall = \frac{2\ \times Precision\ \times Recall}{Precision\ + Recall}$$

It measures of test's accuracy on a dataset, what is basically the harmonic mean of the Precision and Recall.

4) AUC-ROC curve - area under the Receiving Operating Characteristic curve.
   This measure has quite convolved analytical formula to put it into proposal. Basically, it is the integral which measures the relation of the True Positive Rate, or Recall, to the False Positive Rate:

$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + True\ Negative}$$

   It means the probability that the model ranks a random positive example than a random negative example.

Project Design

The standard machine learning project can be divided into the following steps:

1) Data Processing, which includes:
   1.1) Data exploration: analysis of the raw data, creation of the target variables
   1.2) Data Preprocessing and transformation: the data should be standardized during the first steps
   1.3) Creation of the new features: some features can be lacking for the analysis, thus, have to be created
2) Modeling:
   2.1) Creation of the model: the different models for different offer types will be proposed
   2.2) Deploying of the model
   2.3) Hyperparameter tuning
   2.4) Model comparison, using the metrics described above

After these steps, the performance of each model is going to be checked, compared and the best model will be chosen.