Machine Learning Engineer Nanodegree

Capstone Proposal

Mostafa Osama

April 10th, 2019

Proposal

Domain Background

Mobile ad fraud is one of the biggest challenges the mobile marketing industry is currently facing. Fraud risk is everywhere, but for companies that advertise online, click fraud can happen at an overwhelming volume, resulting in misleading click data and wasted money. Ad channels can drive up costs by simply clicking on the ad at a large scale. With over 1 billion smart mobile devices in active use every month.

A new report from Juniper Research has found that advertisers will lose an estimated $19 billion to fraudulent activities next year, equivalent to $51 million per day.

Problem Statement

TalkingData, China's largest independent big data service platform, covers over 70% of active mobile devices nationwide. They handle 3 billion clicks per day, of which 90% are potentially fraudulent. Their current approach to prevent click fraud for app developers is to measure the journey of a user's click across their portfolio, and flag IP addresses who produce lots of clicks, but never end up installing apps. With this information, they've built an IP blacklist and device blacklist.

In this Kaggle Competition we're challenged to build an algorithm that predicts whether a user will download an app after clicking a mobile app ad. To support your modeling, they have provided a generous dataset covering approximately 200 million clicks over 4 days!

Datasets and Inputs

TalkingData have provided a generous dataset covering approximately 200 million clicks over 4 days. Data sources are divided into training and test datasets, and the training dataset has the following features, IP address of the click, App ID for marketing, device type ID of the user, OS version and the target that is to be predicted. Regarding the test datasets are having the click ID and not included the target that is to be predicted.

Solution Statement

This challenge is a classification model that responsible for identifying to predicts whether a user will download an app after clicking a mobile app ad or not.

We will use Pandas, Numpy and Scikit-Learn framework – also I am on trying Neural Networks as well.

Benchmark Model

From the public leaderboard the winner has Area under Receiving Operating Characteristics Curve score equal to 0.98349 – very amazing score.

Evaluation Metrics

A model in this competition is graded based on the area-under-the-ROC-curve score between the predicted class probability and the observed target, measured on the test data.

Project Design

To develop and manage a production-ready model, you must work through the following stages – recommended by Google:

- Source and prepare your data.
- Develop your model.
- Train an ML model on your data:
  - Train model
  - Evaluate model accuracy
  - Tune hyperparameters
- Deploy your trained model.
- Monitor the predictions on an ongoing basis.
- Manage your models and model versions.