What do you expect from a Data Science project?

From many people’s point of view, the data science project is like a magic which turn those boring numerical data into sexy knowledge without much effort. However, the truth is that most of the time, we need to be careful in order to make those thoughtful conclusions. Since there could be so many way to interpret data, it’s very likely you will get lost without a clear understanding of your data.

In this paper, I would like to share with you about my experience with TalkingData Fraud click detection - a Kaggle competition in 2018.

The goal of the project

The structure of this whole story

First part:

Describe how we deal with the fraud click classification:

1. Naive approach
2. Consider the time series property
3. The problems behind the data preprocessing
4. conclusion

Second Part:

Another way, that we deal with block list construction:

1. preprocessing data
2. possible use case

Naïve approach:

Problem statement:

As China’s largest independent big data service platform, TalkingData deal with 3 billion clicks data per day, of which 90% are potentially fraudulent. In this competition, they provided dataset cover four days with about 200 million clicks. They would like to build an algorithm to predicts where a user will download an app after clicking a mobile app ad or not.

While successful, they want to always be one step ahead of fraudsters and have turned to the Kaggle community for help in further developing their solution. In their 2nd competition with Kaggle, you’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!

Dive into dataset:

We first plot the data vs. classes to see how the data is distributed over different attributes value. We converted those time series values into category like hours and days. It is clear that most of the clicks are coming from small group of devices and OS. We also see the distribution of the fraud click didn’t seems correlated with the days in a week() but correlated with the hours in a day. Thus, we could focus on those data in a day to start building our shallow learning model. To avoid the influence of imbalanced dataset, AUC value seems to be a good choice for performance measurement.

Compare with the models:

After some magic spell, we could get two models. One model is build on the Random Forest and the other one is build on Extreme Gradient Boosting. Both of those two models give us a pretty good AUC value around 0.98. The accuracy is about 99% for the majority class and 86% for minority class. Well, this seems to be a pretty good results for most of people. Did you see it? there are two main problems with this naïve approach.