Y-data Class 21-22
Kaggle Competitions

Daniel Hen

A word about myself

- Senior Data Scientist, working **~6** years @ **Fyber**
- Previously Software Developer, then migrated into the Data Scientist role
- Dealing mostly with tabular data, but also love Computer Vision and developed a crazy app with it, which you all can visit here
- Think that overall - continuously learning is the key for success!
- Linkedin | GitHub | Medium

Competitions

- You will execute 2 competitions during the period of ~6 weeks:
    a. House Prices With Advanced Feature Engineering (Supervised - Regression problem)
    b. Ad Tracking Fraud Detection (Supervised - Classification problem)

The Problem

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad.

But, this competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 70+ explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.

The main goal of this competition is to predict the final price of each home.

House Prices with Advanced Feature Engineering

Skills you are going to practice

- Exploratory Data Analysis (aka EDA) for better understanding the data
- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting
- Model Evaluation

The problem

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, China is the largest mobile market in the world and therefore suffers from huge volumes of fraudulent traffic
- 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.

Ad Tracking Fraud Detection

The Challenge

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, you will observe a dataset covering approximately ~30 million clicks over a range of 4 days.

Skills you are going to practice

- Exploratory Data Analysis (aka EDA) for better understanding the data
- Creative feature engineering
- Advanced Classification techniques, also like random forest and gradient boosting, and maybe some other models?!
- Model Evaluation

The goal is to deliver something every 1-2 weeks (during the upcoming 6 weeks)

- ***EDA - a notebook -*** *deliver can be an .ipynb file, a GitHub link (preferred!) -* ***By Dec 31st, 2021 - 23:59:00 PM***
- *Baseline Model (with a submission!) -* ***By January 14th, 2022 - 23:59:00 PM***
- *Final Model (with a submission!) -* ***By January 31st, 2022 - 23:59:00 PM*** **- also the competition deadline!**

Notes:

- ***You are welcome to discuss and debate each other***
- ***You can reach out with questions any time you have***

Rules

1. *We Trust You* ***- this means everything***
2. *You can team up with your friends for one / both competitions* ***- max. 3 people***
3. *Make sure to plan ahead (so you won't get exhausted to the submission days, whatever these may be)*
4. *Most Importantly - Have Fun!*


Q&A