DSO 530: Statistical Learning Methods
Group Project
Spring 2018

In this group project, you will be given a vaguely defined business problem and a dataset (posted on `Blackboard`) associated with it. You will figure out more details about this business problem and its significance. And you will pick some statistical learning approaches to address certain questions about this problem. The project will be completed in groups of (up to) 5 people each. You will be given some instructions on how to proceed data, but you are on your own to tell a coherent business story. `This project is supposed to be an independent project, after the initial guidence on feature engineering`.

Description of 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. 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.

So, in this project, you are challenged to build an algorithm that predicts whether a user will download an app after clicking a mobile app ad.

This project includes:

1. A final report **due at 5 pm on April 28th**, including
   - i) Name and student ID of every member of the group. Also indicate the contact person and his/her email.
   - ii) Your understanding of the bueinss problem.
   - iii) Description of the data.
   - iv) Review of some of the approaches that you tried or thought about trying and why you think those approaches worked or did not work in the end.
   - v) Summary of the final approach (best) you used and why you chose that approach.
   - vi) Summary of the results, including Area Under ROC Curve (AUC).
   - vii) Business interpretation of your results.
   - viii) Any statistical learning criteria that you think interesting or useful in this problem.
   - ix) Any other problems or questions beyond the original problem that you can leverage the current data set for business purposes.

Note: The final report should be **no more than 7 pages** (more lenghty report incurs a penalty) and submitted in **pdf file** to the following link: https://www.dropbox.com/request/U0n3cy55wtBvsE0gdqvA. `The grader reserves the right to ask for your R codes if he sees inconsistance or other problems in your pdf report.`

2. A 10 min presentation would be in class in **the second to the last week**.

   i) Summary of the topics in your final report (as listed above).

   ii) Graphics that are usually useful in communicating the results.

   iii) All members of the group should appear on the stage, but presentation can be shared among a subset of the members. On the first slide, indicate the contact person and his /her email.

   iv) Q&A. The audience, including instructor, can raise questions to any one of the group members

   Note: In preparing the presentation you should take the audience as a data-savvy manager who is smart with statistical training to the level of multiple linear regression but not beyond.

Grading criteria:

The final project counts 12 out of 100 in the final grade caculation. And the points are decomposed as follows:

1. Presentation (6 points), based on clarity, organization, statistical analysis, response to questions, etc.
2. Final report (6 points), based on based on writing, organization, statistical analysis, insight /actionable information, etc.