Kaggle Competition:

# Fraud Detection Challenge

Can you detect fraudulent click traffic for mobile app ads?

# Fraud Detection

## Challenging:

- Imbalanced Classification
- Prediction Speed

## Supervised learning:

- Decision Tree

## Unsupervised learning:

- Clustering

# Dataset Information

## Basic Information

- Total 184,903,890 rows
- Interested viewers: 456,846 rows  (0.24% of total data)
- Not interested viewers: 184,447,044 rows
- Time Period: 11/ 2017

### Undersampling

Random Undersampling aims to balance class distribution by randomly eliminating majority class examples.

- Interested viewers: 5000 rows
- Not interested viewers: 10,000 rows

# Random Forest Model

Feature ranking:
1. app (0.505252)
2. channel (0.202086)
3. ip (0.121593)
4. device (0.100206)
5. os (0.039008)
6. hour (0.031854)

# Grid Search
# For Random Forest Model (1)

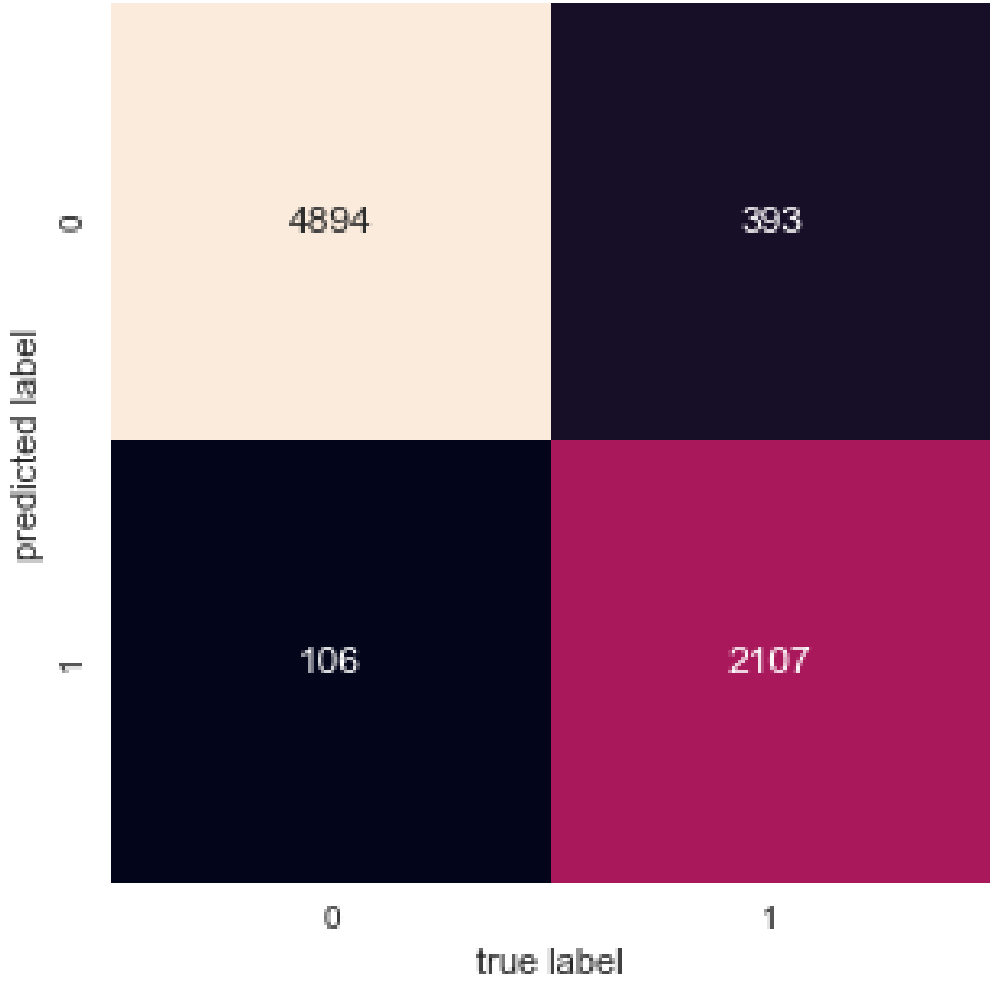
Precision: 0.95

Recall: 0.84

Test Set Accuracy: 0.95

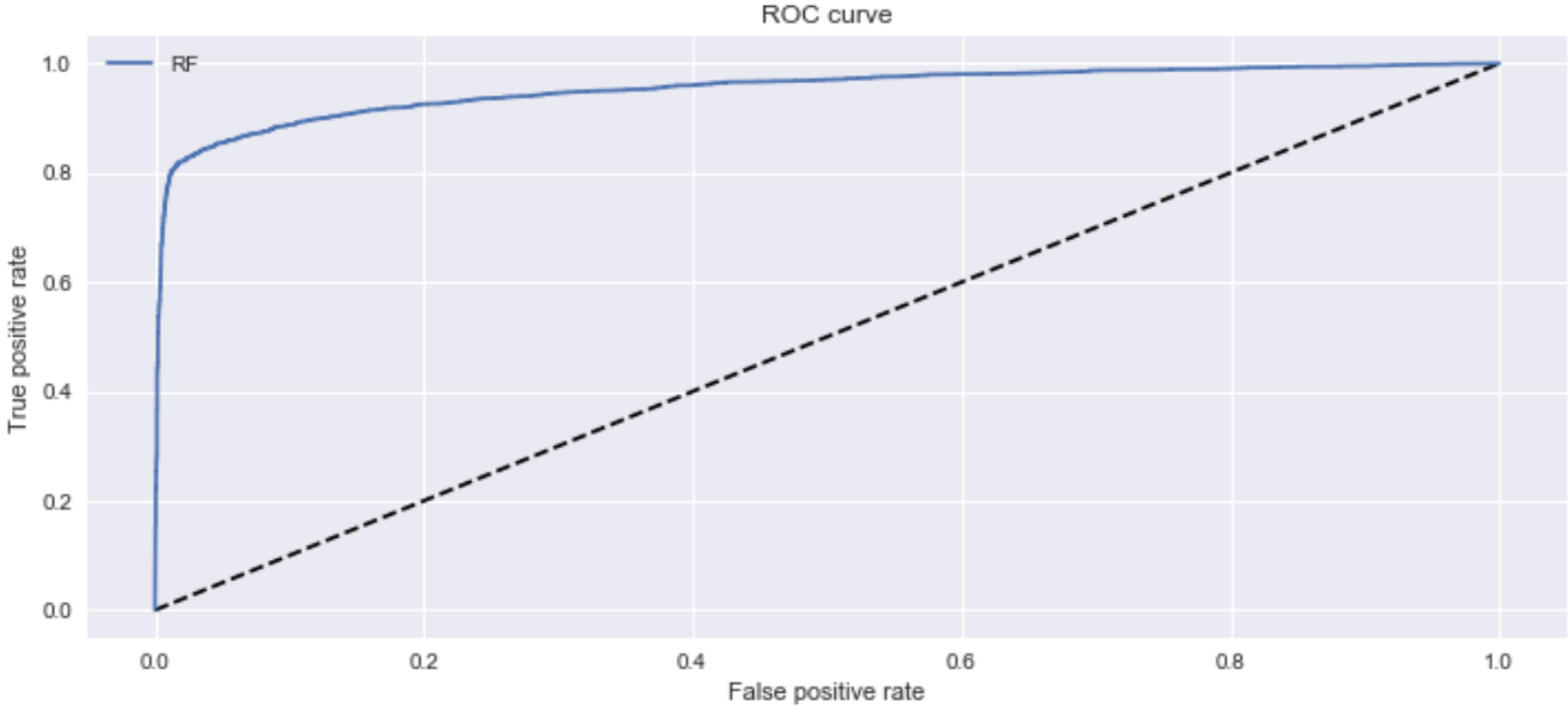
Train Set Accuracy:0.93

F1:  0.89



# Grid Search
# For Random Forest Model (2)

Area Under the Curve (AUC): 0.95



# Kaggle

3 submissions for Yi Chiang

Sort by Most recent

All Successful Selected

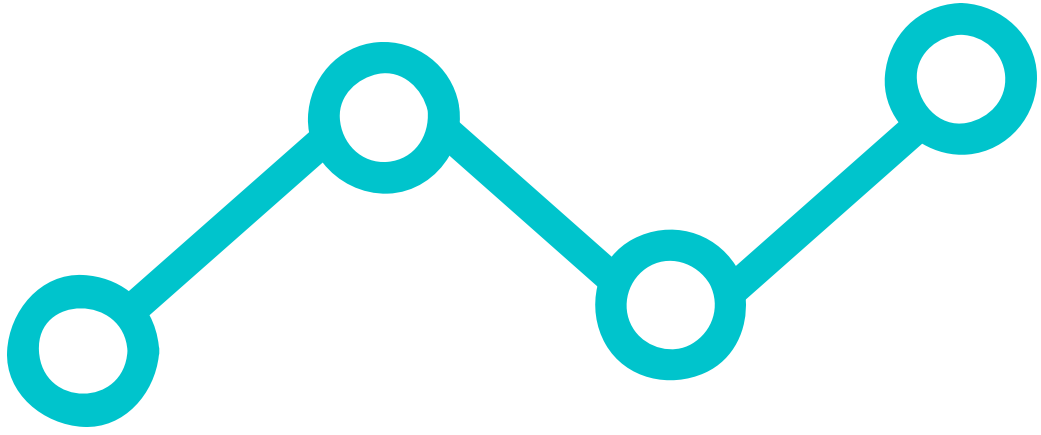
| Submission and Description | Private Score | Public Score | Use for Final Score |
| --- | --- | --- | --- |
| **submission_second.csv**<br>6 hours ago by Yi Chiang<br>catboost | 0.8831411 | 0.8798630 | ☑ |
| **submission.csv**<br>7 hours ago by Yi Chiang<br>change column names | 0.8912362 | 0.8881358 | ☑ |

# DEMO

http://yi-chiang.com
http://metis-3.s3-website-us-east-1.amazonaws.com/

# Future Work

- Apply different imbalanced classes techniques
- Spark
- Unsupervised learning



# End

# Dataset Information

184,903,890 rows × 8 columns

Each row of the training data contains a click record, with the following features.

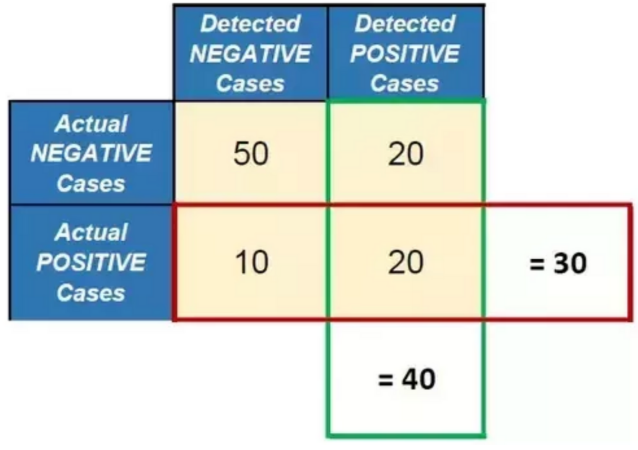
- `ip` : ip address of click.
- `app` : app id for marketing.
- `device` : device **type** id of user mobile phone (e.g., iphone 6 plus, iphone 7, huawei mate 7, etc.)
- `os` : os version id of user mobile phone
- `channel` : channel id of mobile ad publisher
- `click_time` : timestamp of click (UTC)
- `attributed_time` : if user download the app for after clicking an ad, this is the time of the app download
- `is_attributed` : the target that is to be predicted, indicating the app was downloaded

https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/data

# Recall And Precision

Recall: What percentage of people actually infested with virus were detected correctly using your device? = 20/ 30 or 66.67%

Precision: What percentage of people detected positive using your device were actually infested with virus? = 20/ 40 or 50.00%




Source: https://www.quora.com/What-is-the-best-way-to-understand-the-terms-precision-and-recall?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa