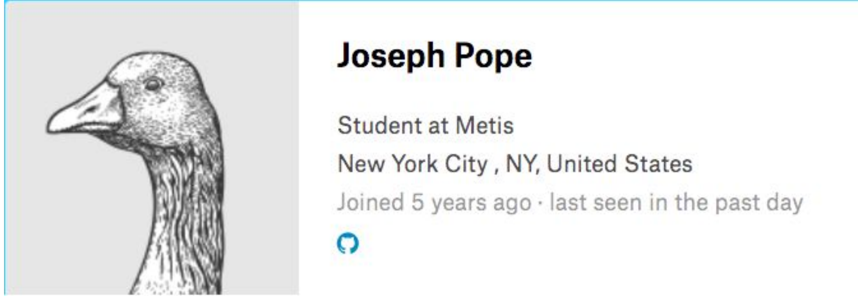
TalkingData

Mobile·Data·Value

Joseph Pope

The Problem



Joseph Pope

Student at Metis

New York City , NY, United States

Joined 5 years ago · last seen in the past day

TalkingData - China’s largest independent big data service platform, largest mobile marketer in the WORLD

1 billion smart mobile devices in active use each month.

Covers 70% of active mobile users devices in China.

3 billion clicks a day, **90% are potentially fraudulent**!!!!

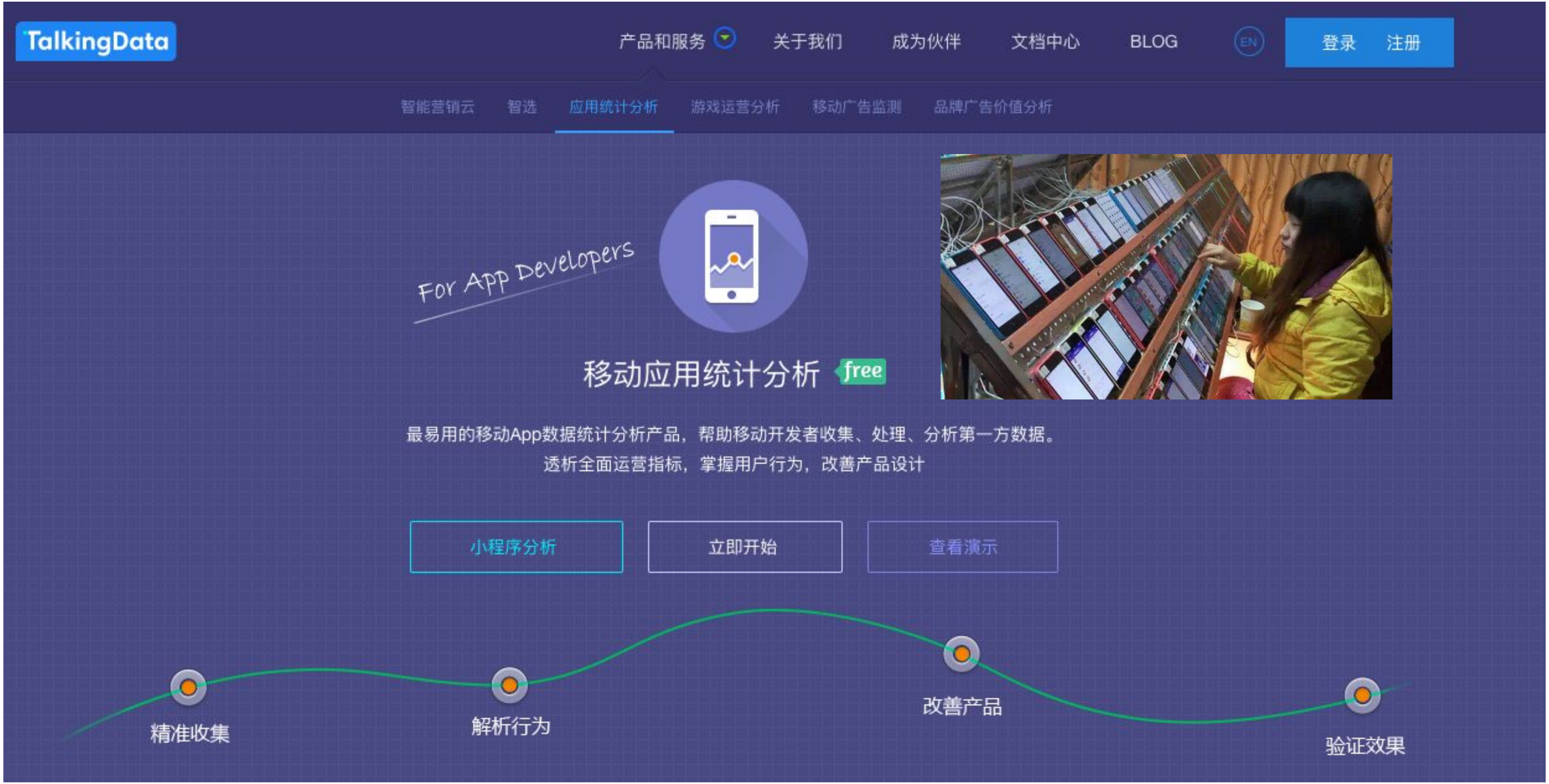
KAGGLE CONTEST - Given a click ID and a few features, predict a download. Don’t find fraud, find the potential app downloaders (humans).

THE DATA

Train, test csv files, containing clickstream data:

- Encoded data for ip addresses, devices, apps, channel and operating systems.
    - Device: iPhone 6 = ‘157’
    - Operating System: Mac High Sierra = ‘22’
    - Lack of clarity on exact definitions
- Time of click, time of attributed download, where applicable. No download date field in the test data.
- Target = ‘is_attributed’ field (1 or 0). Indicates an app download after ad click.

200 million clicks/records over 4 days.

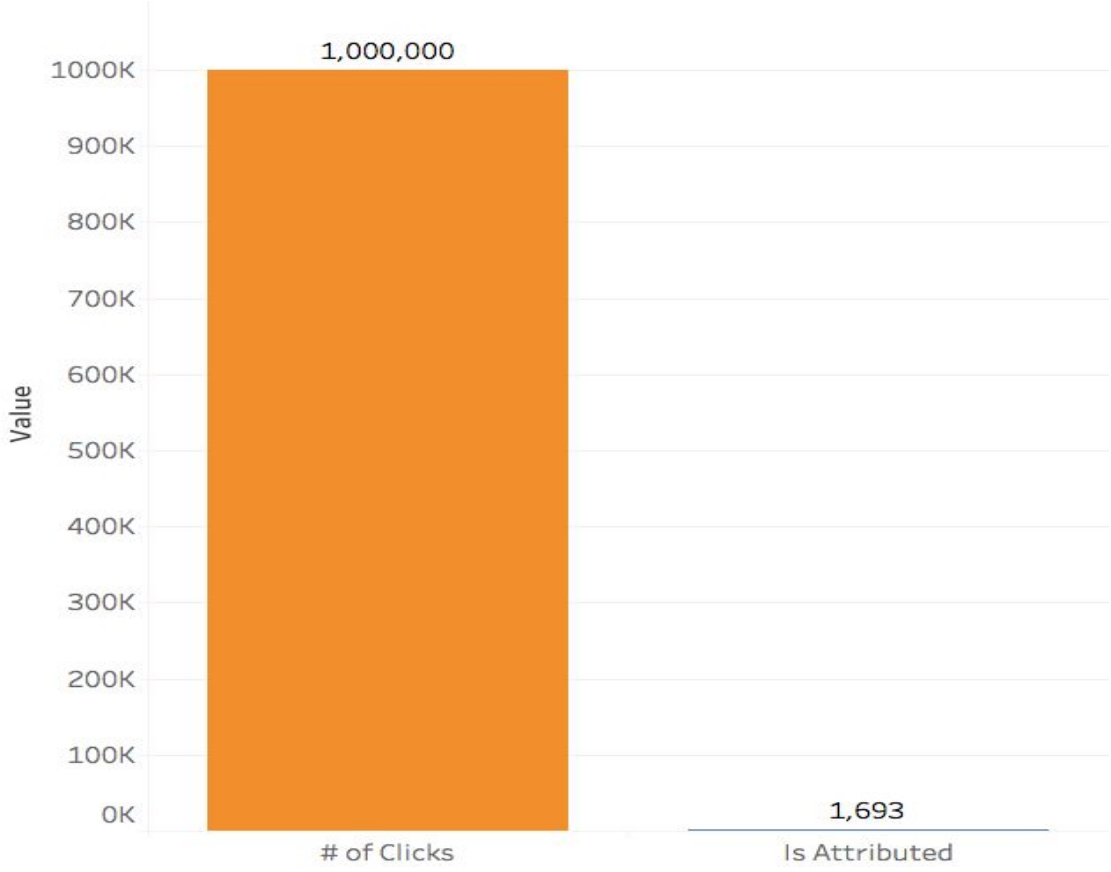


移动应用统计分析

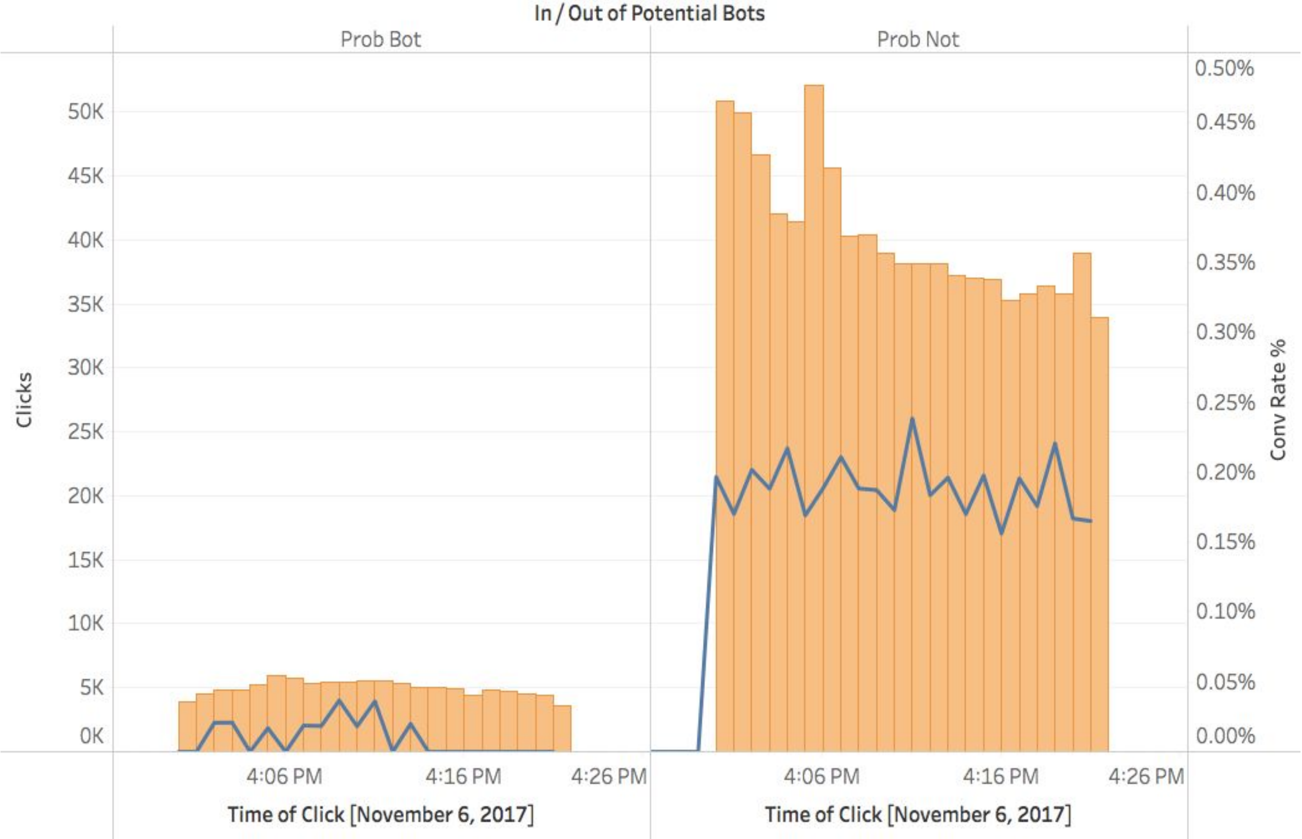
最易用的移动App数据统计分析产品，帮助移动开发者收集、处理、分析第一方数据。
透析全面运营指标，掌握用户行为，改善产品设计

Class Imbalance

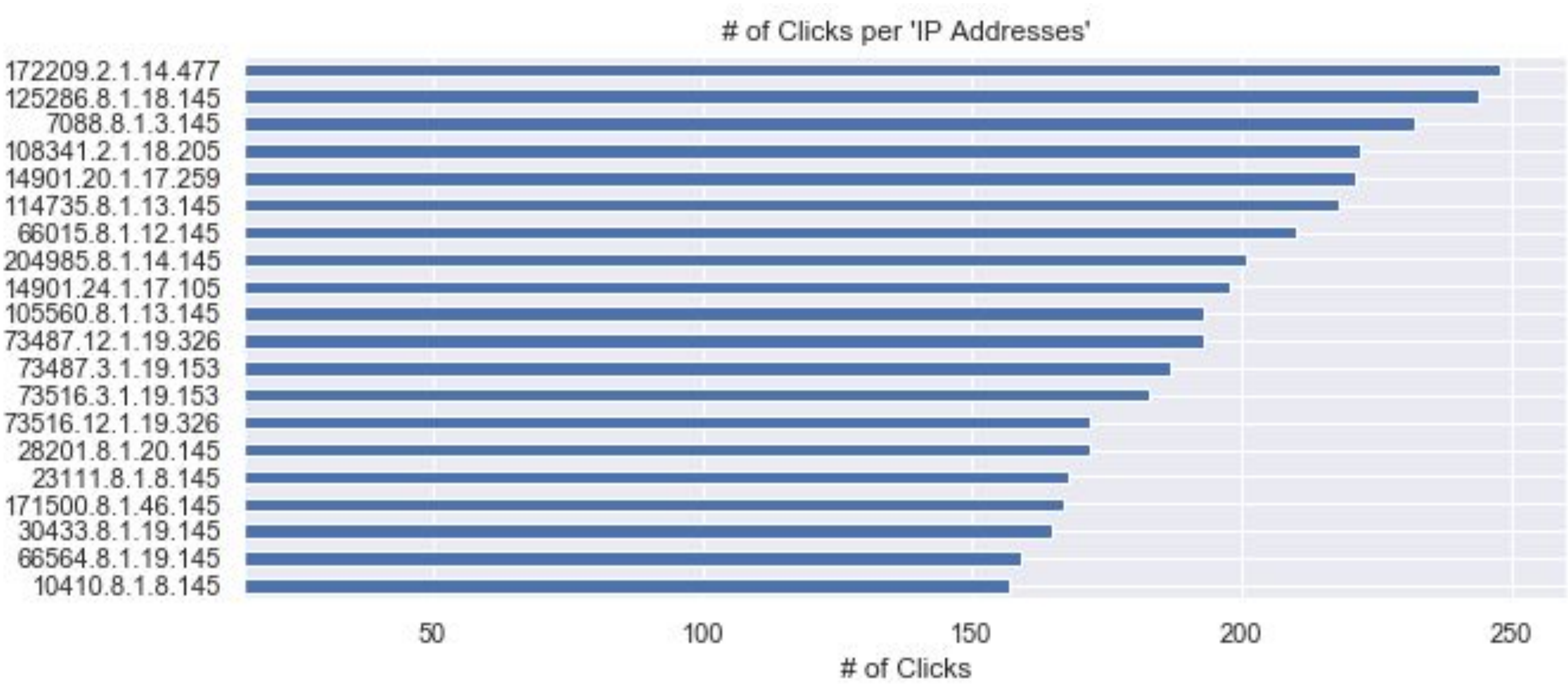
Conversion Rate ~ 0.23%
(downloads / clicks)



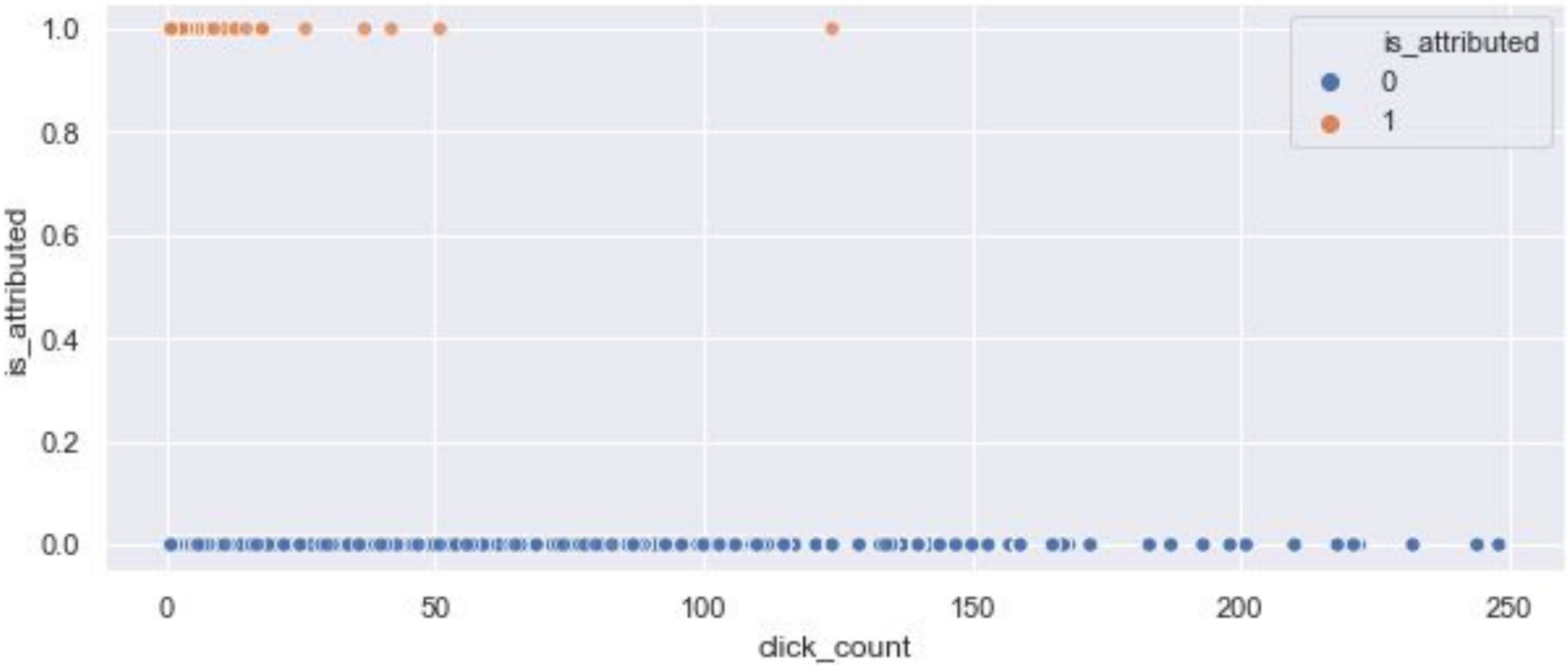
Trends


In / Out of Potential Bots
Prob Bot
Prob Not
Clicks
50K
45K
40K
35K
30K
25K
20K
15K
10K
5K
0K
Conv Rate %
0.50%
0.45%
0.40%
0.35%
0.30%
0.25%
0.20%
0.15%
0.10%
0.05%
0.00%
4:06 PM
4:16 PM
4:26 PM
4:06 PM
4:16 PM
4:26 PM
Time of Click [November 6, 2017]
Time of Click [November 6, 2017]

Ghosts in the Machine



Ghosts in the Machine



Feature Engineering

Concatenate all dimensions into one field, I called ‘User-Agent’. Group and add ‘Click Count’ as new feature.

Review patterns for bot trends. Flag bots to help guide model.

Considered various combinations of features, like device-os counts, app-channel.

Many unique values in each category, so one-hot encoding/dummy variables likely unwise.

Dropped ‘IP’ field from analysis.

Model Selection

Baseline, Logistic Regression

ROC Score: **0.7842**
Recall Score: **0.0000**
Accuracy Score: **0.9984**
Misclassification Rate: 0.001615
Precision Score: **0.0000**
F1 Score: **0.0000**

All scores based on test data, from train/test split.

Final, XGBoost

ROC Score: **0.9548**
Recall Score: **0.7426**
Accuracy Score: 0.9896
Misclassification Rate: 0.010413
Precision Score: **0.1165**
F1 Score: **0.2013**

Model did better with more data.

Model Refinement

- XGBoost (xgbtree) provides a number of parameters to handle overfitting, missing values, imbalanced data and more.
  - eta - Learning rate/shrinkage factor, a form of regularization that can greatly reduce overfitting.
  - scale_pos_weight - handling imbalanced data
  - max delta step - handling imbalanced data.

Next Steps

Kaggle Results:

Private Score

0.9518413

Public Score

0.9528130

Further Feature Engineering. Temporal and device/app.

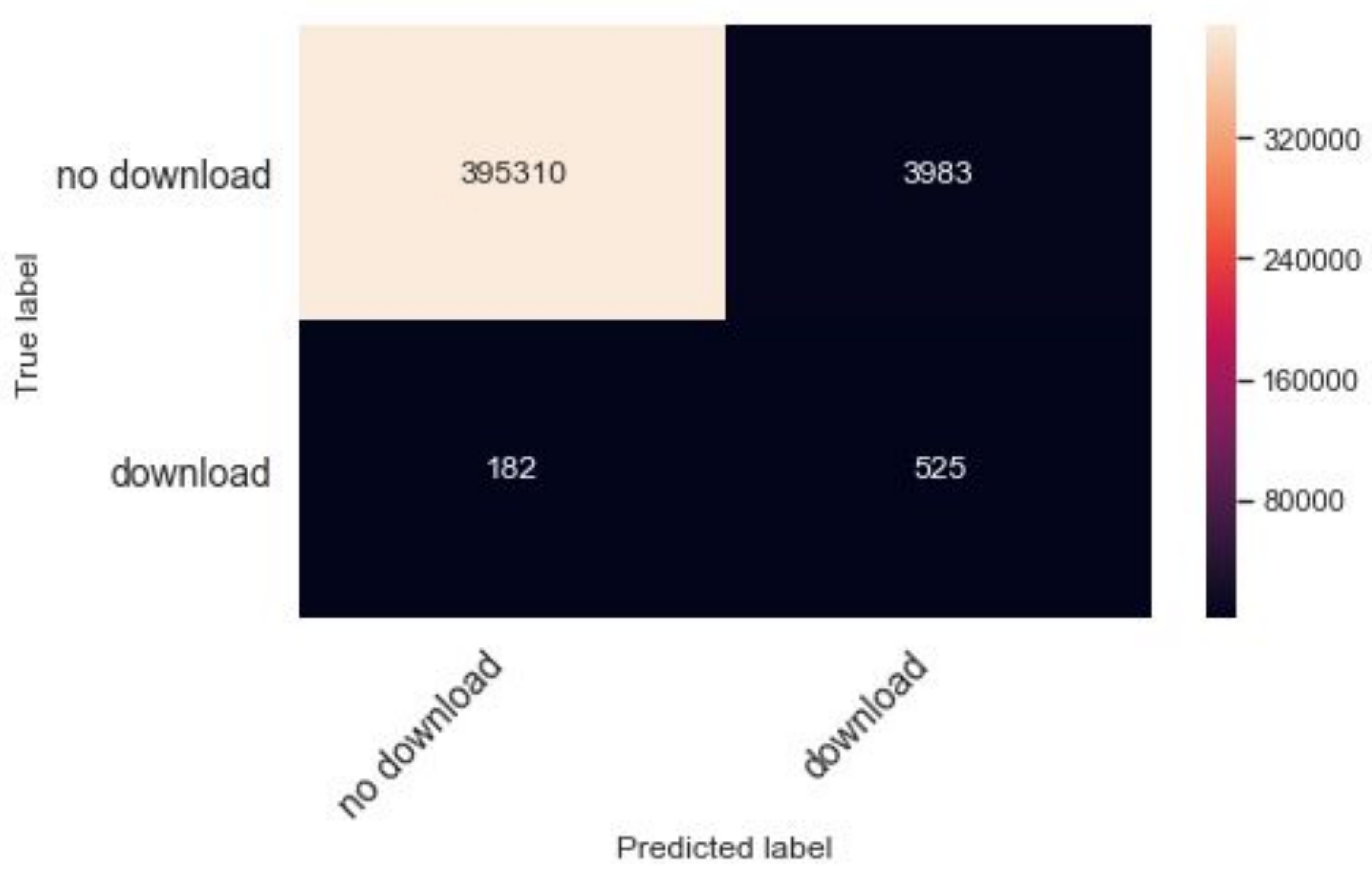
Weight of Evidence approach? Used in binary classification with imbalanced data.

LightGBM

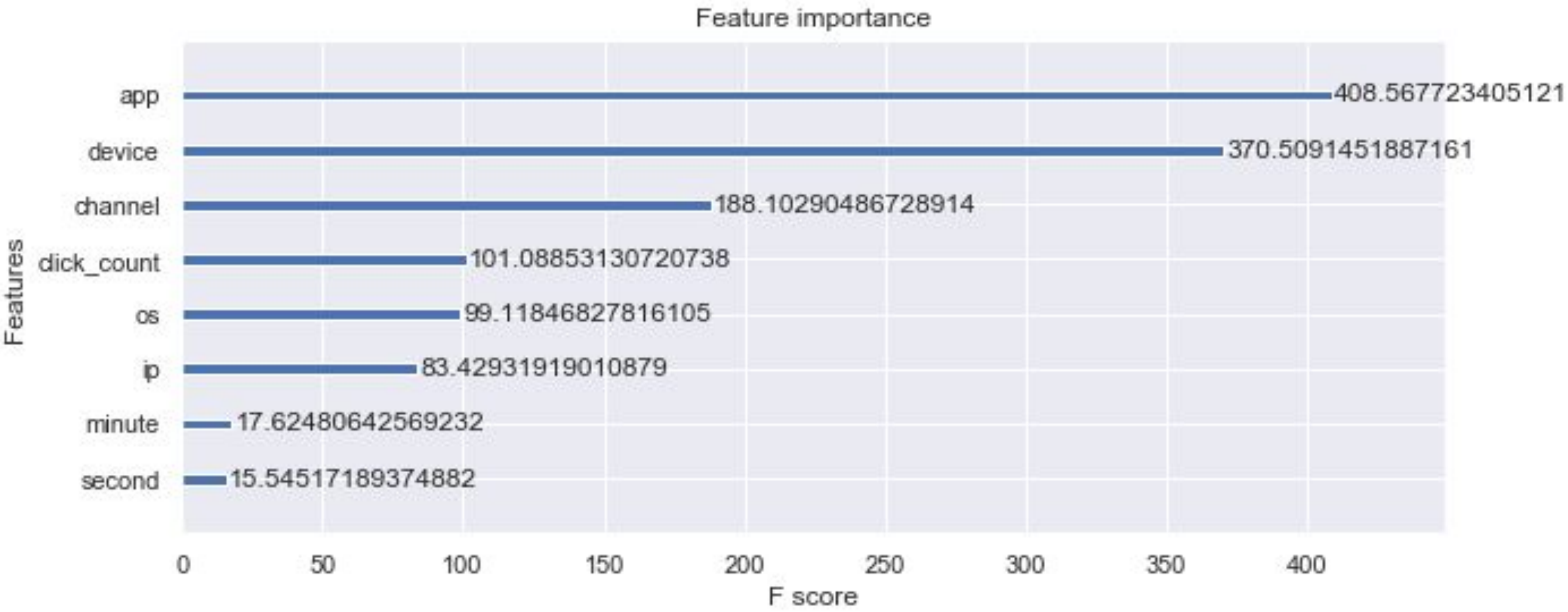
Focus on increasing Precision and F1 score.

SMOTE, CV...

THANK YOU


395310
3983
182
525
no download
download
True label
no download
download
Predicted label
320000
240000
160000
80000

Feature Importance - Gain (XGBoost)


Feature importance
app
device
channel
dick_count
os
ip
minute
second
Features
F score