Fraud detection.

A case assignment.

Truecaller’s aim is to build trust in digital communications, and the battle against fraudsters is never-ending

We try to make information open and readily available, but scammers, scripters, and scrapers are trying to break our rules, and we are always looking for ways to improve detection of anomalous and suspicious activities

This assignment will test your ability to organize, analyze, and interpret data, generating insights and conclusions about potential scraper patterns. Scrapers are users trying to harvest our contact data.

Truecaller.

Time expectation: max 3-4 days (1h/
day) Expected outcomes:

1. 15mins presentation summarizing the data,techniques you used for exploring data and a synopsis of any fraudulent-looking activity you may have found. The presentation that you submit alongside with the code should be prepared for a Product/Business oriented audience, and not restricted to scientists/engineers.

2. Please include the final list of the suspected user IDs in a separate file.

The Data

Data files are provided from four different event types triggered by the app

1. *signup_data.csv*
   - Triggered when users sign up for Truecaller
   - 26000 different users with country code and signup timestamp in millisecond

2. *search_data.csv*
   - A phone number -> name search (i.e. caller ID) performed on our backend servers
   - Triggered by a call, SMS, manually by user from inside app
   - One call/SMS can trigger a few search events in cases of failures
   - 867131 search records with timestamp

3. *call_data.csv*
   - Timestamps of when users made/recieved calls
   - 1223795 calling records with timestamp

4. *message_data.csv*
   - Timestamps of when users sent/recieved SMS
   - 105785 messaging records with timestamp

truecaller

The Problem Statement

- Explore the data using your preferred methods (e.g. Python/Pandas/Jupyter). Some Guidelines:
  - What does average user behavior look like?
  - What might activity from fraud/scrapers look like?
    - These scrapers are interested in harvesting our search data over time
  - Anything else you think is important for making a decision about fraud?
    - There is no certainty about the definition of fraud -- you need to establish criteria

- Try to find fraudulent behavior
  - The data contain real & imagined examples of scrapers following different patterns
  - Some scrapers behave in the similar manner (as with bots), some are independent
  - Timestamps are important in some cases
  - There are somewhere between 200--2000 fraudulent user_ids mixed with regular users

- Given more time, what approaches would do you think would be promising?
- How does your approach scale if we want to detect fraud with >100M users?

truecaller

Feedback

- We respect the candidate’s time and do not expect a perfect solution
  - We are interested in your approach, ideas, and communication style
- Feel free to ask for clarity by email
- About how long did you spend on each phase of the case assignment?
- Any changes you recommend making to the case assignment?

truecaller


truecaller