TalkingData AdTracking Fraud Detection Challenge

Guozhen Li, Hairu Lang
STA 208 - Statistical Machine Learning - Final Project

June 11, 2018

1 Problem Description

This problem is a data analysis challenge from Kaggle (www.kaggle.com), originally found at `https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection`. The data provider, TalkingData, tracks a large base of mobile device activity, and wants to understand which ad clicks end up with app downloads, and which ones lead to nothing (thus are suspects of fraud clicks). We are provided with a training data set of over 180 million clicks, and the objective is to predict whether a click leads to an app download.

For each click, from the raw training data set we know about:

- IP address
- app of marketing
- device type
- mobile device operating system type
- mobile ad publisher channel
- click time (date and time, down to seconds)
- whether a download occurs at the end
- download time, if a download occurs

This data analysis task poses, at least, the following challenges:

1. Extract useful features from the seemingly dispersed information found in the raw training data.
2. Handling of the extra-large data set (over 180 million samples, as an over 7.5GB text file), which can hardly fit into the memory of a normal personal computer.
3. Efficient and effective learning method to learn from the data.

2 Data Exploration

3 Feature Engineering and Data Preprocessing