Root Ad Auction Dataset

Written By: David E. Weirich

Date: May 22, 2019

# Overview

This dataset contains one month of auction data from a real time bidding (RTB) agent. Agents are models with the following inputs and outputs:

* INPUT: features describing inventory for which an ad has been requested
* OUTPUT: a bid amount for the inventory (in USD)

Agents are allowed to bid exactly once per auction. Once all bids have been collected, the agent with the highest bid wins the auction and its ad is shown to the user. This is a *second price auction*, so the company then pays the second highest bid.

Once seeing the ad, the user may decide to click on it, and may then decide to install the Root app. This is the most valuable outcome.

This dataset contains only in-app ads. I.e., no website/youtube/etc. ads.

# Problem

The dataset could be used for two potential modeling tasks. First, train a classifier which identifies which impressions are likely to be clicks or installs. This is a highly imbalanced classification problem, especially when considering installs; that is, there are many more impressions which do not result in a click or install than impressions that do.

The second (and far more interesting) problem is to build your own bidding agent. Train a model which predicts how much to bid on each impression. Then compare the predicted bid with the actual historical second price (the “spend” column in the dataset). If your agent bid higher, then you won the auction. Tally up the total spend as well as the total number of clicks or installs your agent won. Try to win as many clicks or installs with as small a budget as possible. Keep in mind that this data is timestamped. Your bidding agent should not use data from the future to make bids on the present.

# Obtaining the Data

The data is available on a Google drive in the form of several csv files (one per day). Access to the drive will be granted to anyone who chooses to work on this project. **Please do not share this data with anyone who has not been granted access.**

# Description of Dataset

|  |  |
| --- | --- |
| **Column\_Name (data\_type)** | **Description** |
| auction\_id (string) | Unique ID for this auction. |
| campaign\_id (string) | ID of the ad campaign making the bid |
| line\_item\_id (string) | ID of the line item making the bid |
| inventory\_source (string) | Name of the exchange where the auction took place |
| app\_bundle (string) | App bundle the ad is displayed in (hashed) |
| category (string) | Comma separated list of IAB categories, self reported by biddermachine/data\_set.py:36: the app developer. (<https://support.aerserv.com/hc/en-us/articles/207148516-List-of-IAB-Categories>) |
| inventory\_interstitial (double) | Whether the ad responsible for the impression accepts interstitial ad (a static image after a video plays) or not; 1 =r yes, 0 = no |
| geo\_zip (string) | ZIP code where the phone is located at the time of the auction |
| platform\_bandwidth (string) | Bandwidth the phone was using at the time of the auction. One of:   * CELL\_3G * WIFI * CONNECTION\_UNKNOWN * UNKNOWN * CELL\_UNKNOWN * ETHERNET * CELL\_4G * CELL\_2G * nan |
| platform\_carrier (string) | Cell service provider. E.G. Verizon |
| platform\_os (string) | Operating system of the user’s phone.   * Android * iOS * -1 (unknown) |
| platform\_os\_version (string) | Operating System Version |
| platform\_device\_make (string) | Phone Manufacturer. E.G. Samsung |
| platform\_device\_model (string) | Model of the phone |
| platform\_device\_screen\_size (string) | Screen size of the phone. One of:   * S * M * L * XL * UNKNOWN * nan |
| rewarded (double) | Whether this impression is rewarded or not (i.e. view the ad for in-game currency); 0 = no, 1 = yes |
| bid\_floor (double) | Minimum bid this auction will accept |
| bid\_timestamp\_utc (timestamp) | Timestamp of the bid, in UTC |
| hour (string) | Hour of the impression, in “HH:00:” format, in UTC |
| day (string) | Calendar date of the impression. Zero padded to two digits |
| day\_of\_week (string) | Name of the day of the week. |
| month (string) | Calendar month of the impression. Zero padded to two digits |
| year (string) | Calendar year of the impression. |
| segments (string) | Comma separated list of audience segments this device belongs to |
| creative\_type (string) | Type of creative. One of   * banner * video |
| creative\_size (string) | Size of the creative, in WIDTHxHEIGHT format. Only for banner |
| spend (double) | Final cost of the impression (2nd price) paid to the exchange, in USD |
| clicks (double) | Whether this impression resulted in a click; 0 = no, 1 = yes. |
| Installs (double) | Whether the impression resulted in an install; 0 = no, 1 = yes |

Note: Blue indicates a dependant variable. Red indicates an extra variable which is neither dependent nor independent but may be useful for some other purpose. All others are independent variables.