iPinYou Global RTB Bidding Algorithm Competition Dataset

Hairen Liao, Lingxiao Peng, Zhenchuan Liu, Xuehua Shen
iPinYou Inc.
{hairen.liao, lingxiao.peng, zhenchuan.liu, x}@ipinyou.com

ABSTRACT

RTB (Real Time Bidding) is one of the most exciting developments in computational advertising in recent years. It drives transparency and efficiency in the display advertising ecosystem and facilitates the healthy growth of the display advertising industry. It enables advertisers to deliver the right message to the right person at the right time, publishers to better monetize their content by leveraging their website audience, and consumers to view relevant information through personalized ads. However, researchers in computational advertising area have been suffering from lack of publicly available datasets. iPinYou organizes a three-season global RTB algorithm competition in 2013. For each season, there is offline stage and online stage. On the offline stage, iPinYou releases a dataset for model training and reserves a dataset for testing. The dataset includes logs of ad biddings, impressions, clicks, and final conversions. After the whole competition ends, iPinYou organizes and releases all these three datasets for public use. These datasets can support experiments of some important research problems such as bid optimization and CTR estimation. To the best of our knowledge, this is the first publicly available dataset on RTB display advertising. In this paper, we give descriptions of these datasets to further boost the interests of computational advertising research community using this dataset.

Keywords

Real-Time Bidding(RTB), Demand-Side Platform(DSP), Display Advertising

1. INTRODUCTION

RTB (Real Time Bidding) is one of, if not the, most exciting development in computational advertising in recent years. It drives transparency and efficiency in the display advertising ecosystem and facilitates the healthy growth of the display advertising industry. It enables advertisers to deliver the right message to the right person at the right

time, publishers to better monetize their content by leveraging their website audience, and consumers to view relevant information through personalized ads.

In the RTB-enabled display advertising ecosystem, there are two important players, Ad Exchange and DSP (Demand Side Platform). The Ad Exchange aggregates ad impressions from multiple publishers and sends the bid request of each ad impression to several DSPs. The bid request contains the information about this ad slot such as the size and URL of the ad slot, and audience such as IP address and cookie ID or iOS IDFA of the audience. Receiving the bid request, each DSP will check the advertiser database and user profile database, use bidding algorithms to make two following decisions: 1) Decide whether to bid this ad impression or not. If DSP finds the ad slot and audience match what at least one advertiser currently being served is looking for, DSP will bid this impression; 2) If the DSP decides to bid this impression, decide what is the appropriate bidding price for the ad impression and send the bid response back to the Ad Exchange. The Ad Exchange does the auction after receiving bid responses from all participating DSPs or the deadline is passed. If the DSP wins the auction, the advertiser ad being served by the DSP will show up on the user browser. The real-time bidding process of an ad impression is step-by-step illustrated in Figure 1 in Appendix A.

The goal of the DSP bidding algorithm optimization is to optimize advertisers' KPI (Key Performance Indicator) such as the number of clicks and/or the number of conversions. Given the fixed campaign budget, the advertisers want to get as many as clicks and/or conversions as possible.

The DSP bidding algorithm is essential to a successful DSP. It is where computational advertising meets BigData. The design of a DSP bidding algorithm touches many computer science areas such as machine learning, data mining, statistics, optimization and game theory. It has been of great interest to academic researchers. Computational advertising has become a hot research topic in the past several years. But academic researchers are greatly bothered by the lack of the real dataset and the lack of live evaluation platform.

As the leading DSP in China, iPinYou[1] strongly believes the quality of the DSP bidding algorithm is the core of a DSP company and has invested huge amount of resources on the research and development of DSP bidding platform and algorithm design since its DSP went live in March 2012. By designing a set of the bidding algorithms and using both offline and online experiments to evaluate them, iPinYou has

---

[1] http://www.ipinyou.com

improved campaign performance significantly for a variety of advertisers, which has verified the power of a high-quality DSP bidding algorithm.

In order to further improve the performance of DSP bidding algorithms, stimulate the interest of the research and development of DSP bidding algorithms in the whole research community, and speed up the growth of RTB-enabled display advertising ecosystem, iPinYou organizes the global RTB bidding algorithm competition[2] from April 1, 2013 to December 31, 2013.

iPinYou releases a set of bidding logs, organizes researchers and developers to design the bidding algorithms, evaluates the quality of proposed bidding algorithms through both offline experiments and online advertising campaigns on iPinYou production DSP platform, and set up a prize for the winner.

The participating teams try to optimize the following problem:

$$\max \quad \#click + N \cdot \#conversion \tag{1}$$

$$\text{subject to} \quad a\ fixed\ budget \tag{2}$$

Where N is a preset value to show the relative importance of click and conversion.

Before the competition formally kicks off in April, iPinYou dry runs the competition at Beijing University as a course project from February 2013. There are three seasons for the formal competition. The first season is run from April 1, 2013 to May 30, 2013, the second season is run from June 1, 2013 to September 30, 2013, and the third season is run from October 1, 2013 to December 31, 2013.

For each season, there is offline stage and online stage. During the offline stage, the competition registration website is open and the log data is available for downloading. iPinYou also sets up a leaderboard and the ranking of leaderboard is updated periodically. During the online stage, top ranking teams of the offline stage are selected and their algorithms are integrated with iPinYou production DSP platform to bid real impression and audience for real advertising campaigns. The final ranking of the season is totally decided by the score of the online stage.

On the offline stage of each season, iPinYou releases a new dataset for model training and validation purpose and reserves a dataset for offline testing. After the whole competition ends, iPinYou organizes and releases these three datasets for public use. It is a great asset for computational advertising research community. Researchers in computational advertising community have been bothered by lack of publicly available data for a long time. Researchers and professional practitioners now can use these datasets for academic research, consulting service, and course project. Besides, the dataset is accompanied by very valuable benchmark algorithm performance data from hundreds of participating teams.

In this paper, we give descriptions of these datasets to further boost the interests of computational advertising research community using this dataset. The following sections are organized as follows. First we give an overview of the datasets. Then log format and sample data of the datasets are described. In Appendix A, we give a step by step illustration of ad RTB impression. In Appendix B, we explain

the rules of the competition. We thank all people who contribute to this competition in the acknowledgment section.

2. OVERVIEW OF DATASETS

On the offline stage of each season, iPinYou prepares a set of bidding, impression, click, and conversion logs from some real advertising campaigns run through iPinYou DSP production platform for some period for the training and validation purpose (*training dataset*), and a set of similar logs from the following period from the same set of advertising campaigns for offline testing purpose (*testing dataset*).

The training dataset is released to participating teams at the beginning of offline stage of each season. The complete testing dataset is further randomly divided into two parts: one part is the leaderboard testing dataset to score and rank the participating teams on the leaderboard (*leaderboard testing dataset*), and the other part is withheld for the final offline evaluation (*withheld testing dataset*). The purpose of splitting testing datasets into two parts is to avoid the overfitting of bidding algorithms through repeated submissions. The participant's last offline submission is evaluated using the withheld testing dataset to get a team's final offline score on the leaderboard.

For the dataset that iPinYou releases to public use, it contains all the training datasets and leaderboard testing datasets. The withheld testing datasets are still withheld by iPinYou. If you have special purpose using withheld testing datasets, please contact iPinYou through dsp-competition@ipinyou.com.

Here is some description of the data in the training dataset and the processing that iPinYou has done for a variety of purposes.

For each bidding line, the first column is Bid ID, which uniquely identifies one ad impression opportunity. In the following impression and click log, the Bid ID is same for the same ad impression opportunity. IPinYou applies a hash function to Bid ID before the log is released. The Bid ID can be used to join bidding, impression, and click logs.

2.1 Bidding Log

For each line of the bidding log, it contains information from three dimensions, the user, the ad slot, and bidding information. Timestamp column is when the bid request arrives at the DSP bidding server.

iPinYou ID is the user cookie set by iPinYou. iPinYou applies a hash function to the cookie ID before the log is released for user privacy protection.

User Agent is the user agent of browser through which the user visits the publishers website.

IP address is the first three bytes of user IP address. iPinYou removes the last byte for user privacy protection before the log is released.

Region ID is the region (a.k.a., province or state) ID where the user visits. City ID is the city where the user visits.

Ad exchange column represents which ad exchange this ad impression comes from. Currently, the value can be 1, 2, 3, 4, 5 or 6, which represents Tanx (Alibaba), Adx (Google DoubleClick AdX), Tencent (Tencent), Baidu(Baidu), Youku(Youku) or Amx(Google Mobile) respectively .

URL is where the ad impression will show up. The publisher may ask the ad exchange not to pass this information to DSPs. If so, Anonymous URL ID is the domain page ID set by the ad exchange. Only one of URL and Anonymous

URL ID is meaningful, but not both.

Ad slot ID is the unique ID to show the location where ad impression will show up on the web page. One web page can have multiple ad slots.

Ad slot width and ad slot height show the width and height of the ad slot respectively. Both columns can be an array, which means the publisher allows multiple widths and heights of the ad creative.

Ad slot visibility column represent whether the ad slot is above the fold (1), below the fold (2), or unknown (0). Ad slot format is either fixed ad slot (1) or the popup window (2).

Ad slot floor price is the lowest price at which price the publisher allows DSPs to win the ad slot. If no DSPs bid higher than the floor price, this auction has no winning DSP. iPinYou does a linear scale of bidding price before the log is released. The price unit is RMB Yuan per CPM.

Creative ID is the ad creative of the advertiser for which the DSP bids on the exchange.

Bidding price is what the DSP bids on the exchange. iPinYou does a linear scale of bidding price before the log is released. The price unit is RMB Yuan per CPM.

Advertiser ID is added in Season 2 and Season 3. It represents advertisers.

User Profile IDs column is added in Season 2 and Season 3 in both bidding and impression/click/conversion log files. But User Profile IDs in bidding log file are always set to null for security reasons. They are a set of numbers, which are IDs of DAAT (Digital Audience Advertising Taxonomy) category names. Category IDs and category names of DAAT will be described in Section 2.5 User Profile.

2.2 Impression, Click and Conversion Log

For the impression, click and conversion log, most columns are the same as those of bidding log. Here is the description of additional columns.

Log type indicates the type of the log: 1 (impression), 2(click), or 3(conversion).

Paying price is what the winning DSP actually pays on the exchange. iPinYou does a linear scale of bidding price before the log is released. The price unit is RMB Yuan per CPM.

Landing page URL is the URL the user visits if the user clicks the ad creative. iPinYou applies a hash function before the log is released. In Season 1, there is no Advertiser ID provided, so the Landing page URL is also used to distinguish different advertisers.

2.3 Advertiser Category

In the Season 2 and 3 datasets, the Advertiser ID column can be used to identify the advertiser. In Season 1, there is no Advertiser ID column, but the the Landing page URL can be used as the key to distinguish different advertisers. Table 1 lists all advertisers category.

2.4 Testing Data File

The testing data has different formats from the training data. The testing data set is formatted in the following way:

$$ImpLogColumns\ RelateClicks(C)\ IfHasConv(V) \quad (3)$$

ImpLogColumns may be related with 0, 1 or more clicks (C >= 0), Multiple conversions will be counted only once, so V can only be 0 or 1.

The testing data are generally collected after the training data (for Season 1 and 2 data), but not always so. The testing dataset for the third season is generated from the normal training data.

2.5 User Profile

In Season 2 and Season 3, user profile data is added to the dataset. iPinYou develops Digital Advertising Audience Taxonomy (DAAT) to describe the user information from demographic, geographic, long-term interest and in-market purchase perspectives. iPinYou builds user profile by mining user behavior data through sophisticated algorithms. The user profile is heavily used by iPinYou DSP production platform and iPinYou achieves a great campaign performance by leveraging user profile. In this competition, some participating teams achieved better performance after using the user profile.

The user profile file is a mapping file from DAAT category id to category name. They are in both English and Chinese version.

2.6 Region and City File

iPinYou also provides two separate text files. One is the mapping file from the region id to region name, and the other is the mapping file from the city id to the city name. All region and city values are currently geo locations from the mainland China. Each line of the region file is a pair of the form (region id, region name); each line of the city file is a pair of the form (city id, city name).

2.7 Anti-spam Filtering

iPinYou filters publishers with suspicious activities from the bidding logs. iPinYou uses sophisticated algorithms to detect spams.

2.8 Benchmark

Accompanied with the training and testing dataset, there is a benchmark score file for each season. They are leaderboards of top ranking teams from hundreds of participating teams. Please check the following links for the leaderboards for the three seasons for the benchmark purpose.
Season 1:
`http://contest.ipinyou.com/s1-offline-leaderboard.html`
Season 2:
`http://contest.ipinyou.com/s2-offline-leaderboard.html`
Season 3:
`http://contest.ipinyou.com/s3-offline-leaderboard.html`
Weinan Zhang etc. [1] have also done some benchmarking analysis with this dataset including basis statistics, CTR estimation and comparisons between different bid optimization stategies.

3. LOG FORMAT AND SAMPLE DATA

3.1 Bidding Log

Table 2 is the log format for the bidding data. The column serial number, column name and an example are listed. Columns with ∗ means that data in the column is modified before the log is released.

3.2 Impression,Click and Conversion Log

Table 3 is the log format for the impression, click and conversion log. Most of the columns are the same with

Table 1: Advertiser Category. The Advertiser Key is the Landing Page URL for Season 1 dataset, and the Advertiser ID for Season 2 and 3 datasets.

Advertiser Key	Season	Industrial Category
df6f61b2409f4e2f16b6873a7eb50444	1	Consumer Packaged Goods (CPG)
3a7eb50444df6f61b2409f4e2f16b687	1	Chinese vertical e-commerce
9f4e2f16b6873a7eb504df6f61b24044	1	Vertical online media
1458	2	Chinese vertical e-commerce
3358	2	Software
3386	2	International e-commerce
3427	2	Oil
3476	2	Tire
2259	3	Milk powder
2261	3	Telecom
2821	3	Footwear
2997	3	Mobile e-commerce app install

Table 2: The bidding log data format. Column with ∗ means the data in the column is hashed or modified before the log is released.

SN	Column	Example
*1	Bid ID	c0550000008e5a94ac18823d6f275121
2	Timestamp	20130218134701883
*3	iPinYou ID	35605620124122340227135
4	User-Agent	Mozilla/5.0 (Windows NT 5.1) \ AppleWebKit/535.11 (KHTML, \ like Gecko) Chrome/17.0.963.84 \ Safari/535.11 SE 2.X MetaSr 1.0
*5	IP	119.163.222.*
6	Region ID	146
7	City ID	147
8	Ad Exchange	2
*9	Domain	e80f4ec7f5bfbc9ca416a8c01cd1a049
*10	URL	hz55b000008e5a94ac18823d6f275121
11	Anonymous URL	null
12	Ad Slot ID	973726_9023493
13	Ad Slot Width	300
14	Ad Slot Height	250
15	Ad Slot Visibility	FirstView
16	Ad Slot Format	Na
17	Ad Slot Floor Price	0
18	Creative ID	f80f4ec7f5bfbc9ca416a8c01cd1a049
*19	Bidding Price	573
20	Advertiser ID	2259
*21	User Profile IDs	null

Table 3: The impression, click and conversion log data format. Most columns are the same with the bidding log, except the Log Type, Paying Price and the Key Page URL. Column 23 (Advertiser ID) and 24 (User Profile IDs) are only availabe in Season 2 and 3 datasets.

SN	Column	Example
*1	Bid ID	01530000008a77e7ac18823f5a4f5121
2	Timestamp	20130218134701883
3	Log Type	1
*4	iPinYou ID	35605620124122340227135
5	User-Agent	Mozilla/5.0 (compatible; MSIE 9.0; \ Windows NT 6.1; WOW64; Trident/5.0)
*6	IP	118.81.189.*
7	Region ID	15
8	City ID	16
9	Ad Exchange	2
*10	Domain	e80f4ec7f5bfbc9ca416a8c01cd1a049
*11	URL	hz55b000008e5a94ac18823d6f275121
12	Anonymous URL	null
13	Ad Slot ID	2147689_8764813
14	Ad Slot Width	300
15	Ad Slot Height	250
16	Ad Slot Visibility	SecondView
17	Ad Slot Format	Fixed
18	Ad Slot Floor Price	0
19	Creative ID	e39e178ffdf366606f8cab791ee56bcd
*20	Bidding Price	753
*21	Paying Price	15
*22	Landing Page URL	a8be178ffdf366606f8cab791ee56bcd
23	Advertiser ID	3358
*24	User Profile IDs	123,5678,3456

the bidding log, except Log Type, Paying Price and Key Page URL. Columns with ∗ means the data in the column is hashed or modified before the log is released. The price unit is RMB/CPM.

4. DATASET DOWNLOAD

The dataset can be downloaded from Baidu WebDrive `http://pan.baidu.com/s/1kTwX2mF`. It is also available at `http://data.computational-advertising.org`, a server host by University College London. The updated downloading address list will be maintained on `http://contest.ipinyou.com/data-release.html`. Please read the README file for more information about the dataset. If you have any other questions about the dataset, feel free to email `dsp-competition@ipinyou.com`.

5. ACKNOWLEDGMENTS

platform with source code of participating teams. This needs guts! iPinYou also provides the award prize to top ranking teams and supports for the award ceremony in each season, technical meetups, and bidathons.

We thank our judges of this competition. They give numerous suggestions to the competition and promote this competition at all kinds of venues. They are Professor Yuan Yao of Beijing University, Dr. James Shanahan of Church and Duncan Group, Professor Han Liu of Princeton University, Professor Jun Wang from University of College London, Professor Jian Guo from Harvard University, Dr. Yan Qu from ShareThis, Dr. Jonathan Qiang Li from Alibaba, and Dr. Dou Shen from Baidu.

During the competition, we organize competition salons in Beijing, Shanghai, Hangzhou, San Francisco, New York City, and Silicon Valley. We thank all the people who host the salon such as Vince Li from MediaMath, Quan Yuan from Alibaba, Ping Zhu from LinkedIn, and Jianqiang Shen from Turn, and the invited guests such as Professor Tong Zhang from Rutgers.

In the end, we thank all participating members and we had a wonderful time!

6. REFERENCES

[1] Weinan Zhang, Shuai Yuan, Jun Wang, and Xuehua Shen. Real-time bidding benchmarking with ipinyou dataset. Technical report, UCL, 2014.

APPENDIX

A. ILLUSTRATION OF AN RTB AD IMPRESSION LIFE

The real-time bidding process of an ad impression is step-by-step illustrated in Figure 1.

Step 1: The user uses the web browser to visit a publisher website, say, www.yahoo.com. The publisher Yahoo! shows the content of this website such as news to the user. In the meantime, there are some ad slots on the web page where the publisher shows ads to the user. The publisher's main job is to produce and show high quality content to the user. On the other hand, the publisher has to monetize their content in some way. One major monetization method is the advertising. The publisher can use its own sales force to sell the ad inventory to advertisers. RTB is another, increasingly important, way for the publisher to sell its ad inventory. Here we assume the publisher Yahoo! decides to monetize its top-right 300*250 ad slot through RTB.

Step 2: The publisher Yahoo! passes the information to the ad exchange, say, Google DoubleClick AdX, including the URL where the ad slot is located, vertical of the web page content such as sports, and user cookie id.

Step 3: The ad exchange AdX composes a bid request and sends the bid requests to several DSPs. Lets assume the DSP iPinYou is one of them.

Step 4: When the iPinYou DSP server receives the bid request from the ad exchange AdX, it passes the information to the bidding engine.

Step 5: The bidding engine sends the user ID to the user profile and advertiser database to look up whether there is a match between the user and one of its advertisers.

Step 6: The bidding engine receives the information about the user and advertisers, and makes the decision whether to bid the impression and how much to bid.

Step 7: The bidding engine composes a bid response and passes it to the DSP server. The bid response includes the bidding price and other information such as where to fetch the ad.

Step 8: The iPinYou DSP server sends the bid response to the ad exchange AdX.

Step 9: The AdX does the auction after receiving bid responses from all DSP servers or the deadline is passed. The AdX have 100ms deadline, which measures the time from when the bid request is sent out from AdX to when the bid response is received at AdX. The ad exchange usually runs the Vickery auction, a.k.a., the second-price auction, in which the highest bidder wins the auction and pays the second-highest bid price. Here let us assume that the iPinYou DSP is the winner of this ad impression auction.

Step 10: The AdX notifies the user's web browser the auction winner and the web browser will send the ad impression request to the ad server where the ad creative is located.

Step 11: The ad server sends the creative to the user's web browser.

Step 12: The user views the ad on the publisher Yahoo! website. If the user is interested in the ad, (s)he will click the ad, which leads to the landing page of the advertiser. The user browses the advertiser website and may do some further actions. For example, if the advertiser is an eCommerce company, the user may make a purchase on the eCommerce website. We call this kind of action a conversion. Usually the advertiser decides what kind of action is considered as a conversion.

B. IPINYOU RTB BIDDING COMPETITION RULES

B.1 Registration

Any people except iPinYou employees, ex-employees or employee spouses can register iPinYou DSP bidding algorithm competition. Once the registration is confirmed by the competition organizer, the registrant formally becomes an iPinYou DSP competition participating team.

B.2 DSP Bidding Optimization Task

The participating teams try to optimize the following task.
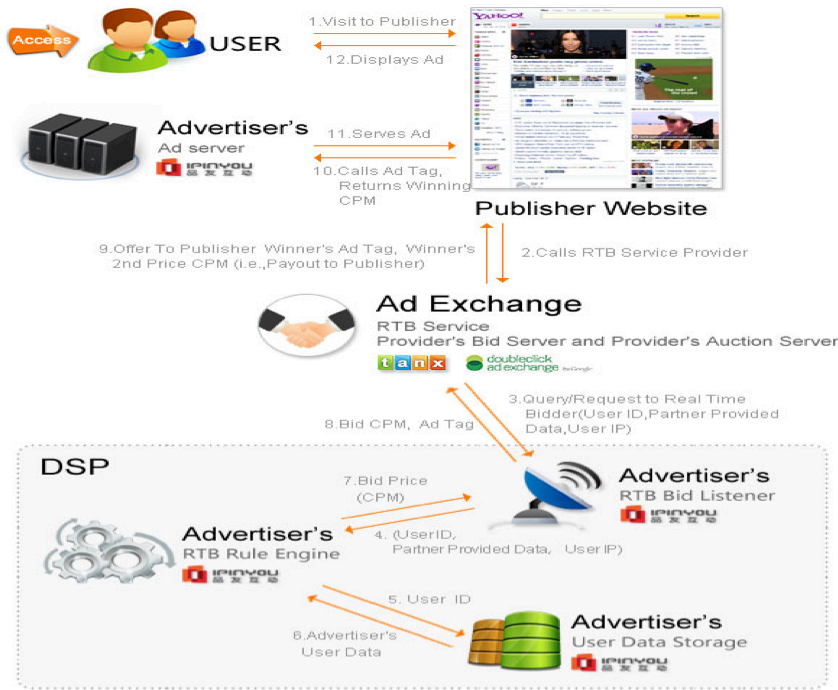
$$\max \quad \#click + N \cdot \#conversion \tag{4}$$

$$\text{subject to} \quad a\ fixed\ budget \tag{5}$$

Where N is a preset value to show the relative importance of click and conversion.

In order to get clicks and/or conversions, the DSP bidding server has to bid on ad exchange to get ad impression first. The DSP bidding server pays ad exchange as long as the ad of the advertiser whom the DSP bid for shows up on the publisher website, i.e. pay by impression. However, one ad impression does not guarantee a click by the user. The user clicks the ad only when (s)he is interested in the ad. Actually its common to see 0.01%∼0.1% Clickthrough Rate(CTR) in desktop display advertising. By the same logic, one ad click does not guarantee a conversion by the user.

Usually the DSP has a fixed ad budget for each advertiser campaign, say, 10,000 RMB. The goal of the advertising campaign is to get as many clicks and/or conversions

Figure 1: A brief illustration of the real-time bidding process: the interaction between user, ad exchange and DSP.

as possible subject to the budget constraint. The scoring function is

$$Score = \#click + N \cdot \#conversion \qquad (6)$$

The weight N shows the relative importance of conversion vs. click. N is set according to characteristics of each season. In Season 1, N is set to 20. In Season 2, N is different for different advertisers. For advertiser 3358, N is set to 2. For advertiser 3476, N is set to 10. In Season 3, N is set to 1. The conversion of Season 3 is actually defined as the landing page arrival.

B.3 Evaluation

There are both offline and online evaluation of DSP bidding algorithms of participating teams.

On the offline stage, iPinYou prepares the testing dataset for the offline evaluation. The testing dataset is a fixed and static dataset. At the beginning of the evaluation, each participant has a fixed budget. If the bidding price of the participant's bidding algorithm is above the paying price in the impression log of testing dataset, the participant's DSP wins this auction and pays the paying price in the impression log for the ad impression. iPinYou evaluation tool automatically consumes the participant's budget and reduces the budget by the paying price. In the meantime, according to the impression and click log, iPinYou evaluation tool automatically decides whether to increase the participant scores or not and how much the score should be increased by.

On the online stage, iPinYou accepts the input file of the participant and run the participant's bidding algorithm for some live advertising campaigns. iPinYou recommends that each participant submit the parameter file for the standard bidding algorithms. But iPinYou also accepts the source code file from the participant and works with the participant to embed the bidding algorithm into the iPinYou production DSP platform. Each participant is given a fixed ad budget, say, RMB 1000 for each advertising campaign, and tries to get the maximal score according to the scoring function.

The evaluation process is the verified by a jury which is composed by a selected group of people.

B.4 Leaderboard and Winner Selection

During the competition, the competition organizer maintains a leaderboard, which is updated periodically.

All participating teams are overall ranked based on the offline testing score. If the scores of two teams are the same, we rank two teams by the submission time, the earlier the submission time is, the higher the ranking is.

For each of the top participating teams who submit their DSP bidding algorithms, we run a set of live advertising campaigns with mixed CPC and CPA goals from different vertical advertisers such as travel, eCommerce, and CPG (Consumer Packaged Goods) and set each campaign to a fixed budget, say, RMB 1000. After the budgets of all live advertising campaigns are consumed or deadline is passed (usually three days), we compute a score also by weighted sum of the click numbers and conversion numbers if there is CPA goal across ads campaigns. We rank the top participating teams by the online stage score and select winners.

B.5 Prize

There is 1,000,000 RMB Yuan Grand Prize for teams which can achieve predefined goal. For each season, there is also a milestone award. Top ranking