COMPGW02/M041: Web Economics Project

Individual report

Said Kassim
Data Science, UCL
ucaksas@ucl.ac.uk

ABSTRACT

Real Time Bidding (RTB) has brought a recent paradigm shift in the way online advertising is done. RTB enables advertising companies to purchase ad impressions in real time after a user visits publisher web pages which have ad slots. By visiting such web pages, a bid request for an ad is triggered and sent by the publisher through their Supply Side Platform (which handles their advertising inventory) to the ad exchange. At the ad exchange the advertisers via their Demand Side Platforms (DSPs) will compete in a second price auction which needs to start and finish in less than a hundred milliseconds [1] in order to maintain user satisfaction. Once the winner is found, they are notified and their ad is displayed to the user.

The DSPs are there to help advertisers manage their ad campaigns and optimize their bidding strategies. Using the data that comes in with the bid request, the DSPs can compute bids for the ad impressions while taking various constraints into consideration such as the market value, budget and campaign lifetime. The DSPs have to maximize their Key Performance Indicators (KPIs) e.g. Click Through Rate (CTR) and Cost per Mille (CPM) at the same time managing the aforementioned constraints.

1. INTRODUCTION

In this report I will explore the RTB display advertising dataset. I will provide a statistical overview of the data provided and with the help of various visualizations will give a deeper understanding of how the RTB auction took place and what were the strategies of the advertisers involved. Then I will explain the bidding strategy that I came up with to optimize on the given KPIs and discuss its merits. I will measure this bidding strategy against the baseline strategy to find how it performs and provide detailed results.

As a group of 3, each member was required to explore the given dataset and to come up with an individual bidding strategy to optimize on the following evaluation metrics: clicks, CTR, spend, CPM and CPC. I came up with my own bidding strategy which performed quite well. Given a budget of 6250 CNY fen, I got 173 clicks and a CTR of 0.1809 % , all while spending 6063 CNY fen of the allocated budget. Further details of the approach and the performance will be described in later sections.

2. RELATED WORK

RTB is a recent development in display advertising with a mileage of just around 8 years given its inception of 2009 [1]. It is an important development and one that will continue to increase spending on display advertising. According to Statista [2], the share of RTB spend in digital display advertising is set to grow to around 29.5% in 2018 and eMarketer says the US spend alone in RTB should hit close to $27 billion this year [3]. This shows the shift towards RTB/programmatic advertising for online display advertising.

With the sensitivity involved in advertising data and the importance of privacy, getting real data of RTB advertising has been hard for researchers in the field. But in 2013, the Chinese advertising technology company iPinYou released the dataset used in its global RTB algorithm competition. With this dataset, research into bid optimization and CTR estimation has progressed and according to [4] this was the first publicly available dataset on RTB display advertising. This has set a precedent and now people outside advertising technology companies can get in on it and devise new and innovative bidding strategies for the RTB space.

Bid optimization is well researched as it is an old problem, however bid optimization for RTB is a different and newer problem. Some of the key works include [5] who show that the bid price has a linear relationship to the predicted CTR (pCTR) for each ad impression being auctioned and [6] who derived simple bidding functions that can be calculated in real time and show that the optimal bid has a non-linear relationship with the click-through and conversion rates, demonstrating better results than linear strategies.

3. APPROACH AND RESULTS

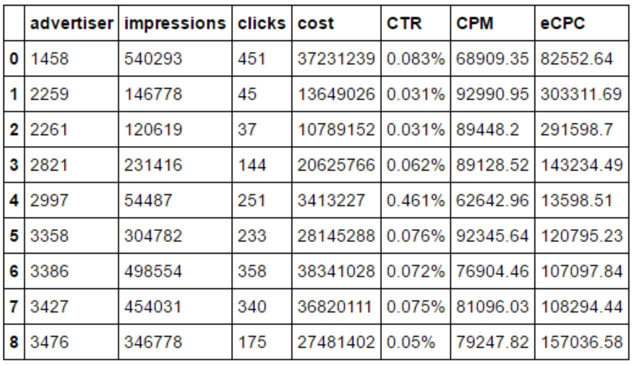
3.1 Problem 1: Data Exploration

The dataset provided for this problem was an RTB dataset of ad impressions. It included a training set, a validation set, and a testing set. The training and validation sets contain the following 26 fields:

click, weekday, hour, bidid, logtype, userid, useragent, IP, region, city, adexchange, domain, url, urlid, slotid, slotwidth, slotheight, slotvisibility, slotformat, slotprice, creative, bidprice, payprice, keypage, advertiser, usertag

The test set did not contain 3 of the 26 fields: bidprice, payprice and click. We would have to generate our own bid prices and submit for evaluation. The training and validation sets were to be used for model building and testing.

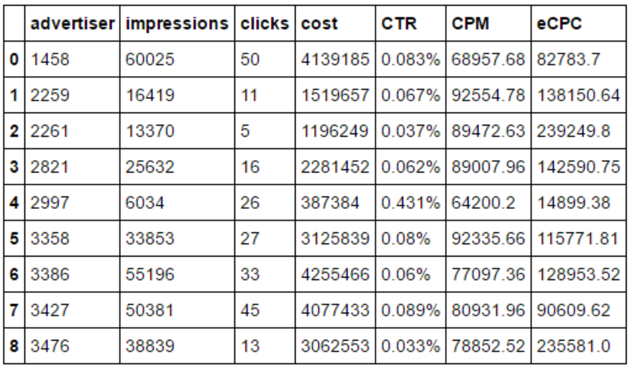
First, I started by finding the basic statistical information of the data per individual advertiser, specifically: the number of impressions won, the number of clicks achieved, the amount of their budget spent, their Click Through Rate (CTR), Cost per Mille (CPM) and effective Cost per Click (eCPC). These statistics for the training data are shown below in figure 1.



	advertiser	impressions	clicks	cost	CTR	CPM	eCPC
0	1458	540293	451	37231239	0.083%	68909.35	82552.64
1	2259	146778	45	13649026	0.031%	92990.95	303311.69
2	2261	120619	37	10789152	0.031%	89448.2	291598.7
3	2821	231416	144	20625766	0.062%	89128.52	143234.49
4	2997	54487	251	3413227	0.461%	62642.96	13598.51
5	3358	304782	233	28145288	0.076%	92345.64	120795.23
6	3386	498554	358	38341028	0.072%	76904.46	107097.84
7	3427	454031	340	36820111	0.075%	81096.03	108294.44
8	3476	346778	175	27481402	0.05%	79247.82	157036.58

Figure 1: Basic Statistics from training data

For the validation data the statistics are shown below in figure 2.



	advertiser	impressions	clicks	cost	CTR	CPM	eCPC
0	1458	60025	50	4139185	0.083%	68957.68	82783.7
1	2259	16419	11	1519657	0.067%	92554.78	138150.64
2	2261	13370	5	1196249	0.037%	89472.63	239249.8
3	2821	25632	16	2281452	0.062%	89007.96	142590.75
4	2997	6034	26	387384	0.431%	64200.2	14899.38
5	3358	33853	27	3125839	0.08%	92335.66	115771.81
6	3386	55196	33	4255466	0.06%	77097.36	128953.52
7	3427	50381	45	4077433	0.089%	80931.96	90609.62
8	3476	38839	13	3062553	0.033%	78852.52	235581.0

Figure 2: Basic Statistics from validation data

From the 2 tables above, it is clear that the training and validation are not in equally split parts. The ratio is about 9:1. Also I observed that the CTR for the advertisers are roughly the same except for advertiser 2997 who seems to have an above average CTR which we will have to further explore to unearth the reason. It may be due to different platforms of advertising i.e. 2997 might be advertising on mobile which usually has higher CTRs due to the smaller screen size making users inadvertently click. Also I observed that although the 9 advertisers had similar CPMs their eCPCs were quite varying. This could be due to different campaign strategies and targeting different markets. This would also partially explain why advertiser 2997 had a high CTR. Maybe the mobile market is less expensive so this advertiser has a lower eCPC but high CTR.

The next part of my exploration into the RTB dataset was to look at user feedback i.e. explore how CTR for the advertisers is affected by various different factors. In particular I look at the CTR performance of 2 advertisers : 1458 and 3358 and how they relate to the day of the week, the time of the day, the region, the ad exchange, the OS used, the browser used and the slot size. Choosing 2 advertisers only made it easier to analyse and to visualize.

3.1.1 CTR - Weekday

Figure 3 shows the effect of the day of the week on the CTR of advertisers 1458 and 3358.

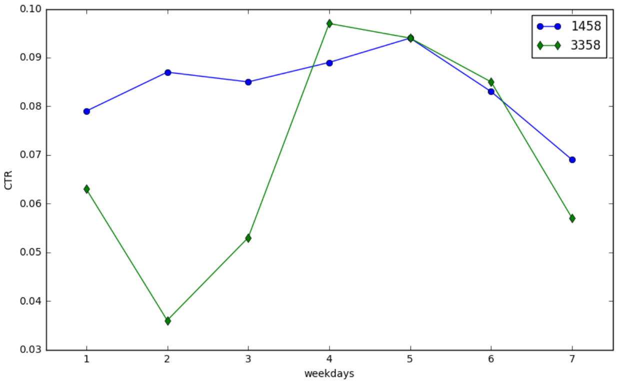


Figure 3: CTR distribution for the day of the week

For advertiser 1458, the CTR is quite even across all days with the least CTR on Sunday, and the highest CTR observed on Friday.

Advertiser 3358 has generally low CTRs with a particularly low CTR on Tuesday and high CTRs on Thursday and Friday.

3.1.2 CTR - time

Figure 4 shows the effect of the the time of the day on the CTR of advertisers 1458 and 3358.

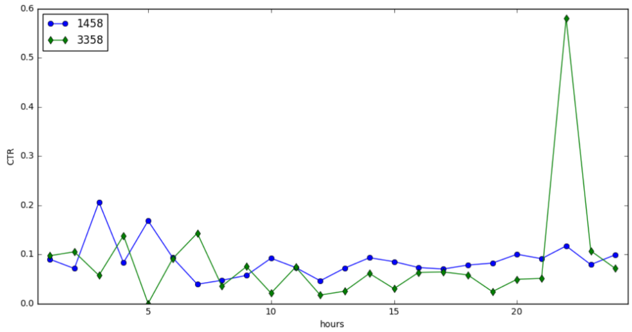


Figure 4: CTR distribution for the time of the day

Advertiser 1458 has its peak CTRs at 0300 and 0500 and lowest CTRs at 0700. Advertiser 3358 has its peak CTRs at 2300, with significant movement at that time from other times and CTRs of 0.0% at 0500.

3.1.3 *CTR - region*

Figure 5 shows the effect of the the location on the CTR of advertisers 1458 and 3358.

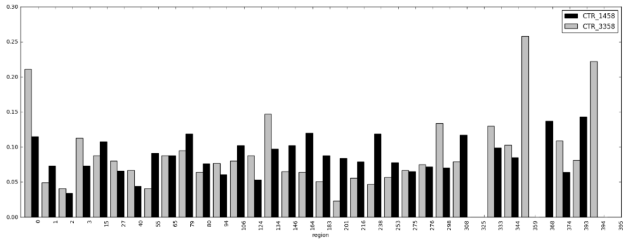


Figure 5: CTR distribution for the region

Advertiser 1458 is generally consistent across all regions except 4 of the 34 regions where it has no activity whatsoever. Advertiser 3358 is more volatile with high activity in some regions such as 359 and 394 and no activity in others.

3.1.4 *CTR - ad exchange*

Figure 6 shows the which ad exchange has the highest CTRs for advertisers 1458 and 3358.

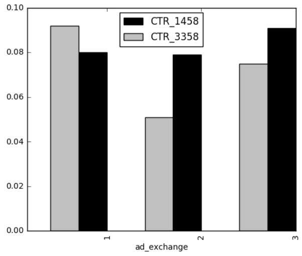


Figure 6: CTR distribution for the ad exchange

Advertiser 1458 is again is consistent across all 3 ad exchanges. Advertiser 3358 is less consistent with more of its CTR from ad exchange 1 and less with ad exchange 2.

3.1.5 *CTR - ad slot size*

Figure 7 shows the effect of the slot size on the CTR of advertisers 1458 and 3358.

The banner (1000x90), the standard (300x250) and the large rectangle(336x280) have generally high CTRs.

Advertiser 1458 has its CTR spread across most ad slot sizes with its highest CTR from the standard (300x250) and its second highest CTR from the leaderboard (728x90). Advertiser 3358 has its highest CTR from the 360x300 and the vertical banner (120x240).

We can also discover that some advertisers perform better in certain regions due to their high CTRs from certain regional ad slot sizes. For example advertiser 3358 has a high CTR in the Slider IAB Rising Star (950x90) which is a Chinese ad slot size, so we can say advertiser 1458 has a Chinese based audience.

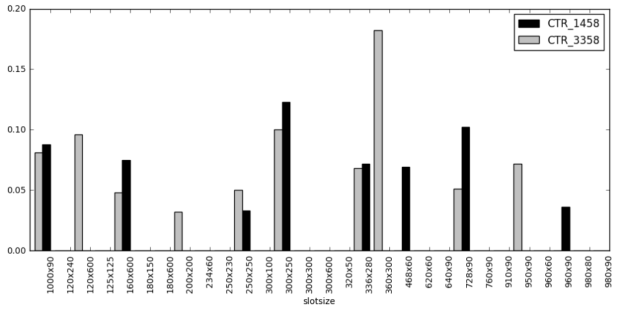


Figure 7: CTR distribution for the ad slot size

Also from the 468x60 slot size in which advertiser 1458 has a high CTR, we can tell that these particular ads were targeted to tablet users as that slot size is meant for tablets. So advertiser 1458 also has a tablet using audience.

3.1.6 *CTR - OS*

Figure 8 shows the effect of the Operating System (OS) on the CTR of advertisers 1458 and 3358.

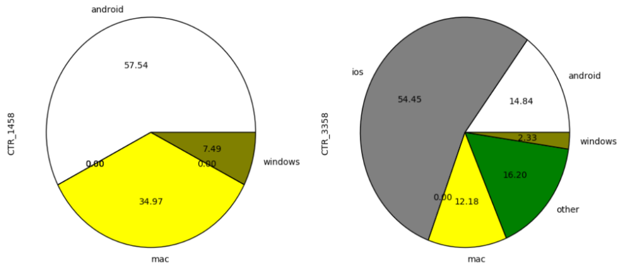


Figure 8: CTR distribution for OS

From the pie charts we can say tha Android users are more likely to click on ads from advertiser 1458, while iOS users tend to click on ads from advertiser 3358. Advertiser 1458 has no CTR for iOS at all meaning that they probably have no ads compatible for iOS devices. Advertiser 1458 seems to have a greater share of the PC users (both Windows and Mac). Advertiser 3358 has a 16.20% CTR from users who don't use a traditional OS.

3.1.7 *CTR - browser*

Figure 9 shows the effect of the browser on the CTR of advertisers 1458 and 3358.

Advertisers 1458 and 3358 have a nearly equal share of Safari users. Opera users mainly click on ads from advertiser 1458 and no Opera user clicks on ads from advertiser 3358. The 2 advertisers have an even share of the IE, Chrome and Firefox browsers as well as other lesser known browsers.

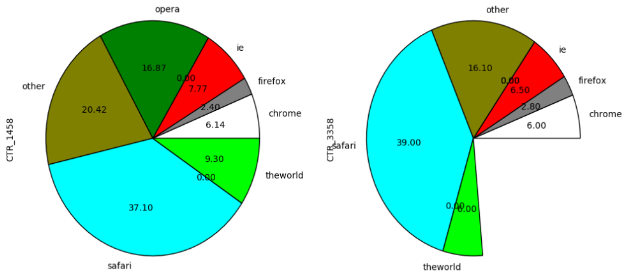


Figure 9: CTR distribution for browser

3.2 Problem 4: My best bidding strategy

For this problem, we had to individually come up with bidding strategies that would improve on the linear bidding strategy which does the following:

$$\mathbf{bid} = \mathbf{base\_bid} * \mathbf{pCTR/avgCTR}$$

For my strategy I decided to square the normalized pCTR term i.e. the pCTR/avgCTR, so my bidding function ended up looking like this:

$$\mathbf{bid} = \mathbf{base\_bid} * (\mathbf{pCTR/avgCTR})^{\mathbf{2}}$$

This seemed intuitive to me as it accentuates the importance of the pCTR on the generated bid. So when the normalized pCTR is greater that 1 squaring it will ensure the bid is high and thus increases the chances of winning that particular impression which is likely to be clicked, whereas if the normalized pCTR is less than 1 then we don't really want to win that impression as it is not likely to be clicked so by squaring it we bid even lower hence lowering our chance of winning that 'bad' impression.

From figures 10 and 11 we can see that the squared function shifts the click distribution up and its lowest clicks is around 79 compared to the low of 16 from the linear function. Also when bidding ends the square function ends with clicks at 139 compared to the 100 clicks from the linear bidding function. It also does the same for the CTR distribution.

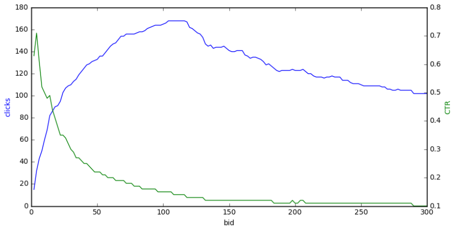


Figure 10: Clicks and CTR distribution for the linear bidding strategy

As is shown in figures 12 and 13 the linear bidding's highest click count is 168 compared to my squared function's 173 and CTR is 0.15 compared to 0.18, proving with the KPIs that indeed my strategy of using the squared of the normalized pCTR does perform better than the linear bidding strategy.

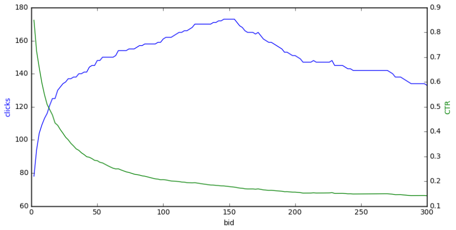
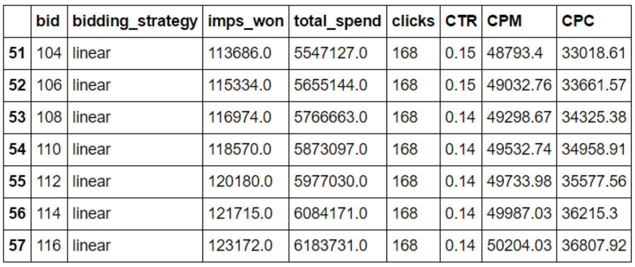
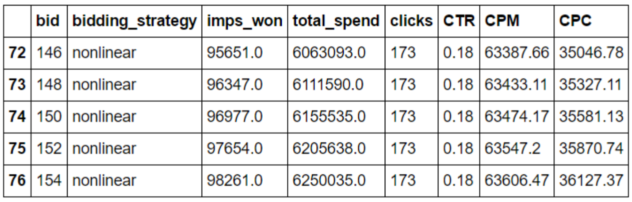


Figure 11: Clicks and CTR distribution for my bidding strategy



	bid	bidding_strategy	imps_won	total_spend	clicks	CTR	CPM	CPC
51	104	linear	113686.0	5547127.0	168	0.15	48793.4	33018.61
52	106	linear	115334.0	5655144.0	168	0.15	49032.76	33661.57
53	108	linear	116974.0	5766663.0	168	0.14	49298.67	34325.38
54	110	linear	118570.0	5873097.0	168	0.14	49532.74	34958.91
55	112	linear	120180.0	5977030.0	168	0.14	49733.98	35577.56
56	114	linear	121715.0	6084171.0	168	0.14	49987.03	36215.3
57	116	linear	123172.0	6183731.0	168	0.14	50204.03	36807.92

Figure 12: Highest clicks for linear bidding strategy



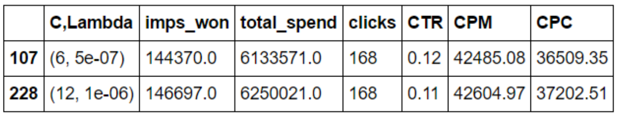
	bid	bidding_strategy	imps_won	total_spend	clicks	CTR	CPM	CPC
72	146	nonlinear	95651.0	6063093.0	173	0.18	63387.66	35046.78
73	148	nonlinear	96347.0	6111590.0	173	0.18	63433.11	35327.11
74	150	nonlinear	96977.0	6155535.0	173	0.18	63474.17	35581.13
75	152	nonlinear	97654.0	6205638.0	173	0.18	63547.2	35870.74
76	154	nonlinear	98261.0	6250035.0	173	0.18	63606.47	36127.37

Figure 13: Highest clicks for my bidding strategy

I also implemented the ORTB bidding strategy [6] to try to see if there was an improvement on the linear bidding strategy. The formula is shown below:

$$w(b(\theta)) = \frac{b(\theta)}{c + b(\theta)} \tag{1}$$

I found the optimal C to be 6 and the optimal λ as 5×10^{-7}. Unfortunately, it performed exactly the same as the linear strategy in terms of click count achieved with a slightly lower CTR. The click count and CTR is shown in figure 14.



	C,Lambda	imps_won	total_spend	clicks	CTR	CPM	CPC
107	(6, 5e-07)	144370.0	6133571.0	168	0.12	42485.08	36509.35
228	(12, 1e-06)	146697.0	6250021.0	168	0.11	42604.97	37202.51

Figure 14: Highest clicks for ORTB bidding strategy

4. CONCLUSION

In this report, I explored the RTB display advertising data and extracted various meaningful insights that informed my decision for downstream tasks such as pCTR estimation and bidding optimization. I also explained my bidding strategy that adds a simple non linearity while improving on the performance of the linear bidding against the set KPIs. In future, I would like to pursue more bidding strategies that take both CTR and spend into consideration so as to increase clicks while maintaining a good spend. I would also like to analyse how various constraints such as budget and campaign lifetime affects my bidding strategy as in this report everything was fixed and thus not as realistic.

In our group we had the following dynamic, we all did our separate data exploration first. Then we came together to discuss how to tackle the group sections of the coursework. We decided to split the group sections amongst us while still keeping an eye out for each other and lending a hand. The constant and random bidding was handled by Kamakshi Bansal. I worked on the feature engineering with generous input from Kamakshi and James Shields. We then each implemented a pCTR estimation model where I implemented the Logistic Regression, Kamakshi implemented the XGBoost and James implemented both Naive Bayes and a Multi Layer Perceptron Neural Network in Tensorflow. For my pCTR estimation model, I did data balancing to help the model with the imbalance in the clicks and did feature selection. I then did the linear bidding strategy part. We each then implemented a non linear bidding strategy where Kamakshi did an exponential function which she will explain in her report, James did his gate function which he will explain and I implemented the ORTB[6] function (mentioned briefly above and explained in detail in the group report section 4.2.2) and my squared function as described in section 3.2 above. Once we had all the strategies set up, James then analysed all our strategies and tried to combine them into our best model. Finally, we all worked together in writing the group report with each doing their parts of the report while James worked on the literature review.

5. APPENDIX

My code for the data exploration part is at the following Github repository:

https://github.com/SaidAbdullahi/webecon_dataexploration

The individual bidding strategies I implemented is within the group repository for bidding strategies at the following address:

https://github.com/SaidAbdullahi/web_econ_coursework
/blob/master/web_econ_group_part.ipynb

6. REFERENCES

[1] S. Yuan, J. Wang, and X. Zhao. Real-time bidding for online advertising: measurement and analysis. In ADKDD, 2013.

[2] https://www.statista.com/statistics/267762/share-of-rtb-in-digital-display-ad-spend-in-the-us/

[3] https://www.emarketer.com/Article/Programmatic-Direct-Takes-Majority-of-Programmatic-Ad-Dollars/1013035

[4] Zhang, Weinan, et al. ”Real-time bidding benchmarking with ipinyou dataset.” arXiv preprint arXiv:1407.7073 (2014).

[5] C. Perlich, B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost. Bid optimizing and inventory scoring in targeted online advertising. In KDD, pages 804-812, 2012.

[6] Zhang, W., Yuan, S., Wang, J. (2014, August). Optimal real-time bidding for display advertising. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1077-1086). ACM. Chicago