COMPGW02 : WEB ECONOMICS PROJECT
Individual Report

Kamakshi Bansal
16114486
(GROUP 22)
kamakshi.bansal.16@ucl.ac.uk
University College London.

Abstract

To have the better online sale, improving the display advertising has become crucial. Real-time bidding has emerged as a new display advertising paradigm. RTB allows advertisers to buy and sell an ad impression online within fraction of seconds till the time user loads the web page. Powerful machine learning tools such as Logistic Regression, AdaBoost Classifier, XGBoost have been used to automate the bidding process. We are provided with a data set with 2697738 rows where each row represents an impression which includes logs of 'bids', 'clicks', payprice,bidid', 'userid', 'IP', 'adexchange', 'domain', 'url', 'urlid', 'slotid', 'slotvisibility', 'slotformat','creative', 'keypage', 'OS', 'browser' for each bid request.
In this paper, statistical analysis of this dataset has been done i.e evaluation of impressions, clicks, cost, CTR (click through rate), CPM(cost per mile) and eCPC (effective cost per click).This paper also discusses the implementation of non linear bidding strategy to optimize the bid so as to maximize the number of clicks.

1 INTRODUCTION

As an ad impression loads in the user's web browser, the information about the user and the page on which user is currently present is passed to the ad exchange. The ad exchange auctions it off to the advertisers who is willing to pay the highest price for that ad impression by sending the bid request. Finally, the winner's ad will be shown to the visitor along with the normal content of the web page. All this process happens in milliseconds till the web page is loaded. Advertisers typically use demand-side platforms to help them decide which ad impressions to purchase and how much to bid on them based on a variety of factors, such as the sites they appear on and the previous behavior of the users loading them. For e.g : ASOS might recognize that a user has previously been on its site looking at a specific pair of shoes, for example, and therefore may be prepared to pay more than Amazon or Best Buy to serve ads to him. The price of impressions is determined in real time based on what buyers are willing to pay, hence the name real-time bidding. We have considered clicks as or major KPI, though CPM and ePCP have also been evaluated. In the second price auctions, the second highest bids are defined as the market price for the winner.

2 RELATED WORK

Online advertising is one of the most fast growing area in IT industry. Over the past 10 years, its revenue has increased from $6.0 billion in 2002 to $36.6 in 2012, with a compound annual growth rate of 19.7%. Revenues from advertising amounted to $12 billion in 2012 which was a 9% increase from 2011. Prior to 2009, the display advertising was mainly dominated by premium contracts (since 1994) comprising of 40% of the impressions and the remaining 60% impressions coming from ad networks (since 1996), which were generally referred to as remnant. However, in premium contracts impressions are guaranteed as compared to ad networks where they are not.[2]

One of the major contributors to this growth has been RTB (real time bidding), which is also sometimes referred to as programmatic buying. This allows the advertisers to make decisions for every impression or auction. The major advantage of RTB is that it uses data from the publishers and the advertisers to gain an overall understanding of the bidding behavior. This results into a more customized campaign which is what is desired by the advertisers.To make RTB successful, Ad exchanges were created to balance the demand and supply in ad networks. The interaction with these ad exchange take place through demand-side platform (DSP) and supply-side platform (SSP) thereby taking full advantage of RTB. [3]

3 STATISTICAL DATA ANALYSIS

In the given data set,the real time bidding feedback log, there are 9 unique advertisers whose logs are given. Figure 1 shows the data analysis for each one of them. Column 1 shows the total number of impressions won by each advertiser. Column 2 shows the total number of clicks that were made by the users on their advertisement. Column 3 shows the total amount of money spent by the advertiser following the second price auction. Column 4 , 5 and 6 evaluates the CTR (click through rate), CPM(cost per mile) and eCPC (effective cost per click) respectively which are are evaluated using the following formulas :

$$CTR = \frac{Total\_number\_of\_clicks}{Total\_no.\_of\_impressions}$$

$$CPM = \frac{Total\_payprice\_spent\_by\_the\_advertiser}{Total\_no.\_of\_impressions}$$

$$eCPC = \frac{Total\_payprice\_spent\_by\_the\_advertiser}{Total\_number\_of\_clicks}$$

Higher the CTR, better is the bidding strategy and in case of eCPC, the more lower, the better it is because our goal is to maximize the number of clicks by the user on the advertisements and minimize the amount spent by each advertiser. Figures 2, 3, 4 and 5 represnts CTR distribution against different features for the advertisers 1458 and 3358.

- Advertiser 1458 received the CTR on Friday while advertiser 3358 received the highest on Thursday.

	advertiser_id	impressions	clicks	cost	CTR	CPM	eCPC
0	1458	177491	451	45149	0.254%	254.37	100.11
1	2259	74055	45	43365	0.061%	585.58	963.67
2	2261	56059	37	43365	0.066%	773.56	1172.03
3	2821	114996	144	43365	0.125%	377.1	301.15
4	2997	4028	251	38497	6.231%	9557.35	153.37
5	3358	133700	233	35357	0.174%	264.45	151.75
6	3386	175635	358	45150	0.204%	257.07	126.12
7	3427	167034	340	35555	0.204%	212.86	104.57
8	3476	131200	175	35554	0.133%	270.99	203.17

Figure 1: *Statistical analysis of the training data set .*

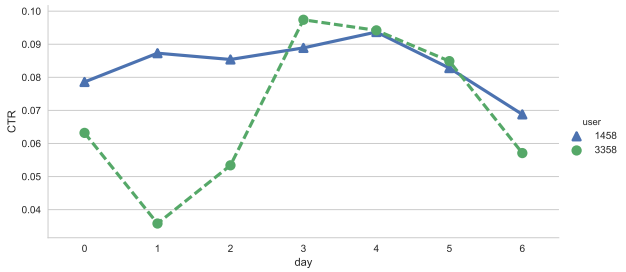

Figure 2: *CTR distribution against weekday for advertisers 1458 and 3358*

- Advertiser 1458 received the CTR on Friday while advertiser 3358 received the highest on Thursday is the morning as compared to evening and night while it is viceversa for the advertiser 3358 with the highest CTR at night.
- The ad CTR for both the advertisers shows erratic behaviour and the trends are different.
- The CTR for advertiser 1458 is higher from 1st adexchange while it is highest from the 3rd adexchange for the advertiser 3358. learning_rate_init=0.001.

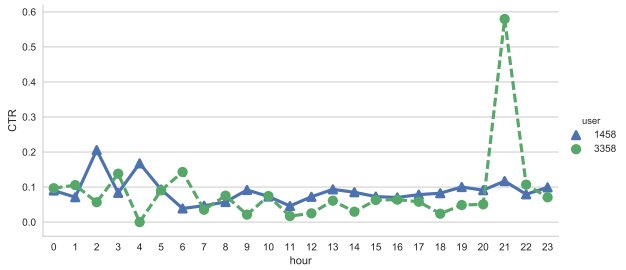

Figure 3: *CTR distribution against hour for advertisers 1458 and 3358*

Figures 6,7,8 and 9 represnts eCPC distribution against different features for the advertiser 3358.

- Advertiser 3358 suffers the highest eCPC on Tuesday while it is much cost effective on Thursday.
- It suffers the highest eCPC during afternoon and varies greatly throughout the day.

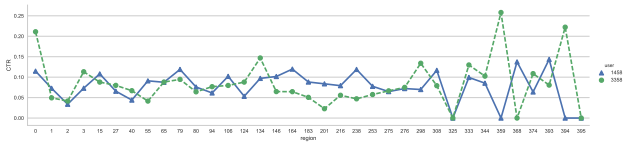

Figure 4: *CTR distribution against region for advertisers 1458 and 3358*

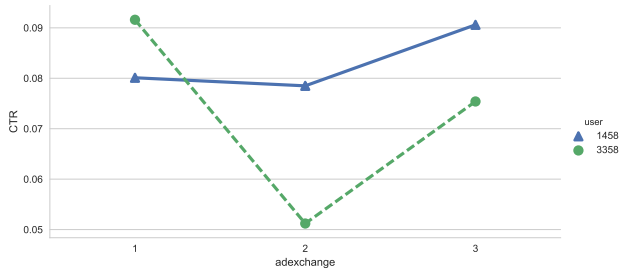

Figure 5: *CTR distribution against adexchange for advertisers 1458 and 3358*

- ECPC distribution against region for advertiser 3358 shows erratic behaviour suffering the most for region 18.
- Advertiser 3358 suffers the highest eCPC from adexcahnge 2.

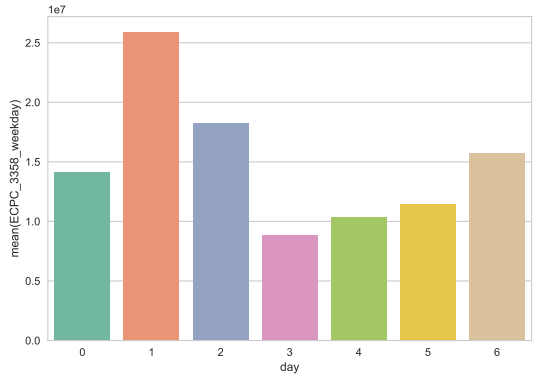

Figure 6: *ECPC distribution against weekday for advertiser 3358.*

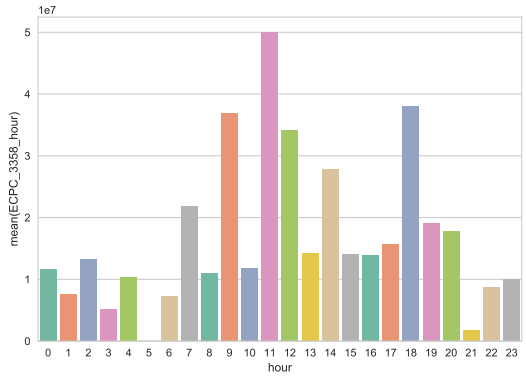

Figure 7: *ECPC distribution against hour for advertiser 3358.*

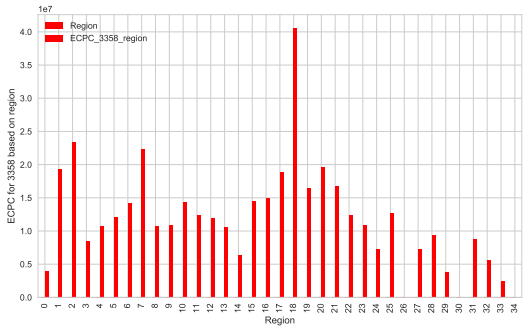

Figure 8: *ECPC distribution against region for advertiser 3358.*

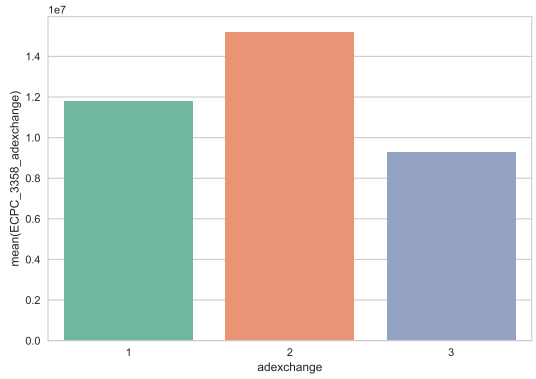

Figure 9: *ECPC distribution against adexchange for advertiser 3358*

In the second price auctions, the second highest bids are defined as the market price for the winner. Figures 10 and 11 shows the market price distribution against different features for advertisers 1458 and 3358.

- Market price has different trends for both the advertisers. The total amount spent by the advertiser 1458 is almost constant throughout the week whereas for advertiser 3358 it's intially low from Monday to Wednesday then rises high on Thursday and finally drops on Friday.
- Market price spent for 1458 is moderately similar throughout the day with the highest spent during the afternoon whereas the amount spent by the advertiser 2258 is highest in the morning and night remaining low for the rest of the day.

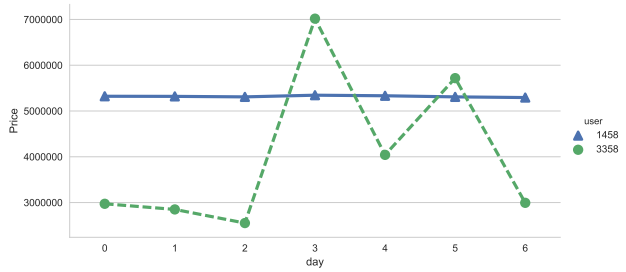

Figure 10: *Market price distribution against weekday for advertisers 1458 and 3358*

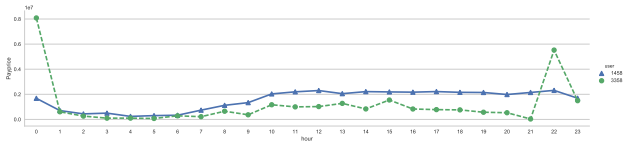

Figure 11: *Market price distribution against hour for advertisers 1458 and 3358*

4 BEST BIDDING STRATERGY

I along with my team members implemented the linear bidding strategy (which can be seen in group report) where the bid value is linearly proportional to the pCTR and inversely proportional to avgCTR. Our goal was to optimize the base bid so as to maximize the number of clicks. It gave 168 as the highest number of clicks for the best bid 104 (refer figure 12).

	bid	bidding_strategy	imps_won	total_spend	clicks	CTR	CPM	CPC
51	104	linear	113686	5547127.0	168	0.1478	48793.4	33018.61
52	106	linear	115334	5655144.0	168	0.1457	49032.76	33661.57
53	108	linear	116974	5766663.0	168	0.1436	49298.67	34325.38
54	110	linear	118570	5873097.0	168	0.1417	49532.74	34958.91
55	112	linear	120180	5977030.0	168	0.1398	49733.98	35577.56
56	114	linear	121715	6084171.0	168	0.138	49987.03	36215.3
57	116	linear	123172	6183731.0	168	0.1364	50204.03	36807.92

Figure 12: *The data analysis for the bid giving the highest number of clicks for the Linear Bidding Strategy.*

4.1 NON LINEAR BIDDING STRATERGY (EXPONENTIAL)

It was decided among our team members that we will all train the model for Linear bidding Strategy and whosoever's model will give the best accuracy score, we will use that model for the further implementation of non linear bidding stratergy.
I implemented two models namely XGBoost and Adaboost to predict the clicks on the Validation set and hence predict the pCTR

values (probability of getting click as 1) which had to be used further to optimize the base bid .With the help of the class DictVectorizer() and LabelEncoder(), all the coulmns with string data was first converted to the categorial data. The vectorized data was then feeded to the machine learning models with the help of scikit learn. Since, the dimension of the data was huge, it took a lot of time to train the model. With XGboost,the performance accuracy was approximately 73 % while with adaboost it was approximately 75%. The literature review for this model has been done in the group report.
Parallely, one of my team member was implementing Logistic Regression using **hot encoding** which performed the best giving the accuracy score of 85% approximately. Hence, I used the pCTR values from his model to further implement my non linear bidding stratery.
To evaluate the the base bid with the constant budget of 6500 CNY fen, I used to the exponential function which performed better than the linear bidding strategy we used earlier. The formula which I used to optimize the base bid is:

$$bid = base\_bid * exp(\frac{pCTR}{avgCTR})$$

4.2 RESULTS

Using the above non linear bidding strategy improved the highest number of clicks to 171 for the best bid 34 (refer figure 14). It can be seen from the exponential graph (figure 13) [1], as the value of x increases, the value of y also increases. Hence, for the values which are near to zero (from right hand side), the value of y is always greater than 1 but closer to 1 (approaching 1).

Since, the value of $(\frac{pCTR}{avgCTR})$ was majorly close to zero (but greater than zero), hence the value of $exp(\frac{pCTR}{avgCTR})$ was majorly close to 1 but greater than 1 i.e there was a slight increase in the value of $(\frac{pCTR}{avgCTR})$ due to which there was slight increase in the bid price for non linear bidding strategy using the formula above(and not significant increase in the bid price) as compared to linear bid price due to which there was an increase in the number of impressions won from 113686 to 138421 (refer figures 12 and 14). Therefore, the number of clicks increased to 168 from 171.

If the bid price (evaluated via exponential function) would have increased significantly as compared to the bid price evaluated with linear bidding strategy, then we might have ended up finishing out budget earlier and hence we would not have achieved good results. It can be seen that this non linear bidding strategy is the best bidding strategy as it is performing better than the linear bidding strategy. It can be also seen from figure 15, that we are getting the highest number of clicks 171 for the CTR value of 0.123.

4.3 CONCLUSION

It can be concluded that this non linear bidding strategy using exponential function is the best bidding strategy as it is performing better than the linear bidding strategy giving 171 clicks as compared to the linear bidding strategy with clicks as 168.

The optimal non linear bid was 34 with 138421 impressions won and 6199575.4 amount spent, getting the number of clicks as 171, with CTR as 0.1235, CPM as 44787.82 and CPC as 36254.82 as can be seen from the figures 14 and 15. Though my team mates implemented other non linear bidding strategies among which the best was gate strategy giving 173 clicks as can be seen from the group report.

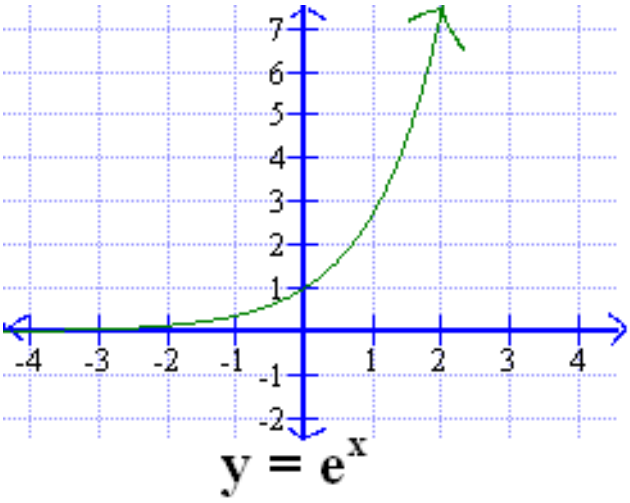

Figure 13: *The non linear exponential graph*

	bid	bidding_strategy	imps_won	total_spend	clicks	CTR	CPM	CPC
16	34	exponential	138421	6199575.0	171	0.1235	44787.82	36254.82

Figure 14: *The data analysis for the bid giving the highest number of clicks for the non linear bidding strategy.*

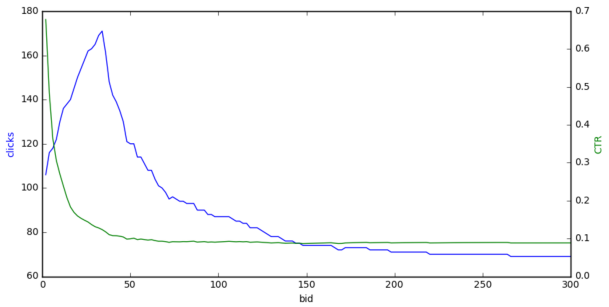

Figure 15: *Distribution of CTR and clicks for the non linear bidding strategy.*

As can be seen from figure 15, CTR (green curve) is high for low bids and decreases as the bid value increases and finally becomes constant. On the other hand, clicks are around 100 for the lower bids going to the maximum of 171 clicks for bid value 34 and then decreasing consistently and finally reaching below 80.

4.4 CONTRIBUTION TO THE PROJECT

4.4.1 Kamakshi Bansal

- **Strength**: Strong analytical ability, Coding Abilities, Hard working and focused . Complete tasks on time efficiently.
- **Key contribution**:
  - Data exploration, analysing eCPC, CTR, market price for different advertisers.
  - Implemented non linear bidding strategy (Exponential

bid strategy).

- Feature extraction.
- Implemented XGBoost for pCTR estimation after converting strings to Categorical data using DictVectorizer.
- Constant and Random Bidding.
- Literature review of my parts.
- Report writing of my parts.

4.4.2 Said Abdullahi

- **Strength**: Coding abilities, background knowledge of NLP, hardworking, commitment to the project.
- **Key contribution**:
  - Data exploration, specifically on user feedback i.e. analyzing CTR per feature.
  - Implemented non linear bidding strategy (Squared bid strategy and ORTB).
  - Feature extraction.
  - Logistic Regression pCTR estimation.
  - Linear bidding strategy.
  - Literature review of his parts.
  - Report writing of his parts.

4.4.3 James Shiztar

- **Strength**: Coding abilities, background knowledge of NLP, creative, patient.
- **Key contribution**:
  - Data exploration.
  - Implemented non linear bidding strategy (Gate bid strategy).
  - Feature extraction.
  - Implemented Naive Bayes and Deep Learning pCTR estimation
  - Combining & analyzing of bidding strategies.
  - Literature review of his parts.
  - Report writing of his parts.

5 APPENDIX

The code for the individual work (Linear bidding strategy and Statistical Analysis of Data) can be found here : `https://github.com/kamakshi22/WebEconomics/blob/master/Web_economics_KAMAKSHI%202.ipynb`

The code for the group and individual work (non linear bidding strategy) can be found here : `https://github.com/SaidAbdullahi/web_econ_coursework/blob/master/web_econ_group_part.ipynb`

References

1. Exponential function. http://www.ask-math.com/exponential-graph.html.
2. Real-time bidding for online advertising: Measurement and analysis. Shuai Yuan, Jun Wang, Xiaoxue Zhao.
3. Optimal real-time bidding for display advertising. Weinan Zhang, Shuai Yuan, Jun Wang.