Web Economics Individual Report

Qiuru Dai
University College London
uczlqda@ucl.ac.uk

1 INTRODUCTION

The real time bidding has been a popular research problem in the computer science area. In this report, first the data used for the real time bidding project will be explored. Then, a Xgboost model is implement to predict the Click Through Rate (CTR) of impressions. After that, the Optimal Real-Time Bidding strategy (ORTB) is applied to build the bid price generation model. Finally, a log pCTR threshold technique which can improve CTR is proposed.

The code of our project can be found in :
https://github.com/BlaBlaPer/WebEconomics

2 LITERATURE REVIEW

Different ways to build real time bidding models and evaluation metrics like CPC and CTR is summarized in [1], especially the Xgboost CTR estimation model will be implement in the following report. It is common to have large amount of imbalance data in the real time bidding area. The negative down sampling is a good method to solve this problem [2].

In bidding strategies, many advanced strategies can be found from the literature. The Optimal Real-Time Bidding strategy (ORTB) [3] including a approximation of winning probability and calculate the expectation of the total click number under the budget constraint as the objective function which would be maximized. The authors proposed a mathematical presentation of the real time bidding problem, which make this problem computable. Usually, the real time bidding problem is divided into two part: utility estimation (for example, CTR estimation) and prediction of market value. In [4], these two parts is combined as a whole to find the better global optimization. Moreover, a reinforcement learning method [5] is used in the real time bidding area, which can deal with budget constraints naturally.

3 APPROACH AND RESULT

3.1 Data Exploration

The dataset being used for this coursework contains the impressions log from iPinYou platform including training set, the validation set and the test set. Features of impressions are provided in the dataset, such as advertiser information, time information and so on. In the training set and the validation set, more information can be found including the click status which shows if an impression is clicked and the pay price which is the price at which an impression was sold. The following analysis will mainly focus on the training set and the validation set due to the lack of information in the test set.

The statistical metrics Click Through Rate (CTR), average Cost Per Mille (CPM) and average Cost Per Click (CPC) are used for the basic analysis. These metrics are explained in detail in the group report.

Table 1: Basic Statistics of Training Set

Adv.	Imps	Clicks	Cost	CTR	CPM	CPC
3427	402806	272	30458.71	0.000675	75.62	111.98
2821	211366	131	18828.04	0.000620	89.08	143.73
1458	492353	385	33968.74	0.000782	68.99	88.23
2259	133673	43	12428.24	0.000322	92.97	289.03
3386	455041	320	34931.82	0.000703	76.77	109.16
3358	264956	202	22447.23	0.000762	84.72	111.13
3476	310835	187	23918.78	0.000602	76.95	127.91
2261	110122	36	9873.78	0.000327	89.66	274.27
2997	49829	217	3129.27	0.004355	62.80	14.42
total	2430981	1793	189984.61	0.000737562	78.15	105.96

The basic analysis results for training and validation sets are shown in the table 1 and table 2. We can see that the basic statistics of training set and the validation set is similar. The advertiser 2997 has the highest CTR and the lowest CPC which are 0.004355 and 14.42 in the training set respectively. These values are significantly different from other advertisers'. This could suggest that the impressions of advertiser 2997 are outstanding in terms of cost effectiveness.

In terms of user feedback, the advertiser 3427 and 1458 who have the largest number of impressions are chosen to be analyzed. From figure 1, 2, 3, 4 we can see that the CTR distribution on different features.

In figure 1 ,we can see both advertise 3427 and 1458 achieve the highest CTR when the region is 395 while the total training set achieve the highest CTR when the region equals to 344. Each different number represents different regions.The region seems have larger impacts on advertiser 3427 and 1458. The advertise 1458 has the highest CTR on Tuesday and advertiser 3427 has the highest CTR on Sunday, while the total dataset achieve the highest on Wednesday, but the difference is small. It seems that slot price do not have much impacts on advertiser 3427 and 1458, but shows large impacts in general. The total CTR becomes quite high when the slot width equals to 336, which may suggest a optimal slot width.

In figure 5 and 6, the pay price distribution is shown. However, the difference between the pay price distributions of impressions with clicks or not seems small.

3.2 Bidding Model

3.2.1 *CTR estimation model.*

Predicted Click Through Rate (pCTR) is crucial for real time bidding problem. To have better CTR estimation, the gradient boosting regression tree (GBRT) is implemented. Being different from the logistic regression (LR), GBRT is a non-linear model which could learn the non-linear features, which is difficult to use the feature

Table 2: Basic Statistics of Validation Set

Adv.	Imps	Clicks	Cost	CTR	CPM	CPC
3427	50183	37	3776.74	0.000737	75.26	102.07
2821	26503	23	2394.90	0.000868	90.36	104.13
1458	62353	49	4294.60	0.000786	68.88	87.64
2259	16715	2	1568.81	0.000120	93.86	784.40
3386	56665	28	4350.79	0.000494	76.78	155.39
3358	32939	23	2794.02	0.000698	84.82	121.48
3476	38841	11	2993.75	0.000283	77.08	272.16
2261	13550	3	1214.88	0.000221	89.66	404.96
2997	6176	26	388.78	0.004210	62.95	14.95
total	303925	202	23777.27	0.000665	78.23	117.71

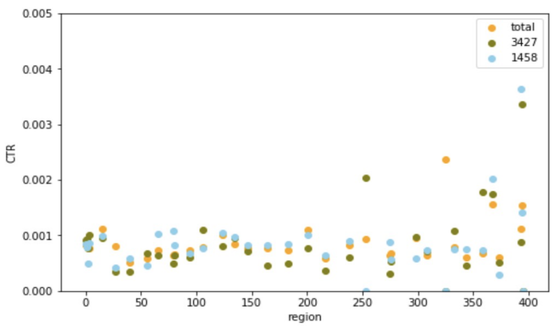

Figure 1: CTR Distribution on Region

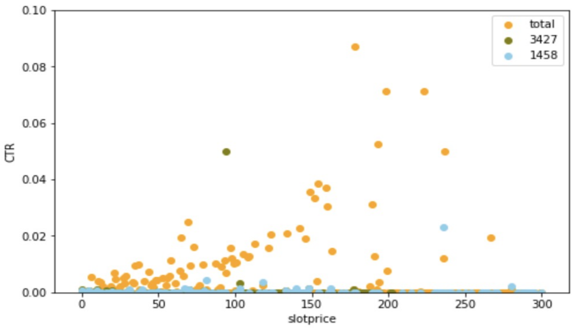

Figure 2: CTR Distribution on Slotprice

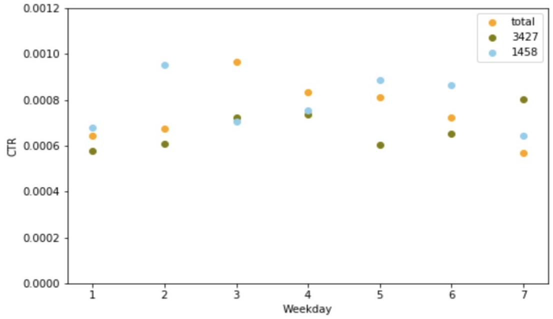

Figure 3: CTR Distribution on Weedday

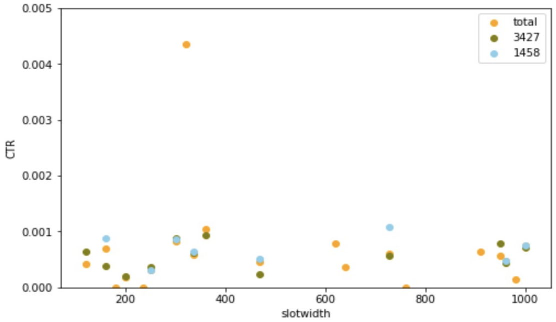

Figure 4: CTR Distribution on Slotwidth

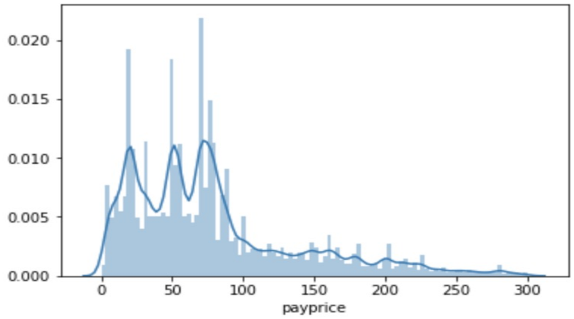

Figure 5: Payprice Distribution of Impressions with no Click

engineering of LR to achieve. Moreover, once the training process is finished, GBRT will abandon most features and only keep the most important features for the prediction [1]. In this project, Xgboost, which is an open-source package providing gradient boosting framework, is used for the GBRT impletation.

As illustrated before, the training set is very large which has 2430981 records, but the number of positive records is very small which is only 1793. To speed up the training process as well as to solve the class imbalance, the negative down sampling technique is used with 0.1 down sampling rate. Since data preprocessing and feature engineering process is shared in the group, it is described in detail in the group report. In total, 127 features are used to train the Xgboost model.

To achieve better model performance, the hyper parameters of the Xgboost model are tuned manually as well as using the grid search method provided from scikit-learn package. The final setting of parameters is showed in table 3. In the figure 7, top 20 important features of the trained Xgboost model are shown. These features have large impacts on our prediction results.

Table 4 shows the evaluation metrics of Xgboost model comparing with the LR model from the linear strategy part of this project. We can see that the Xgboost model has better performance than LR model in general. Especially, the AUC value is much higher

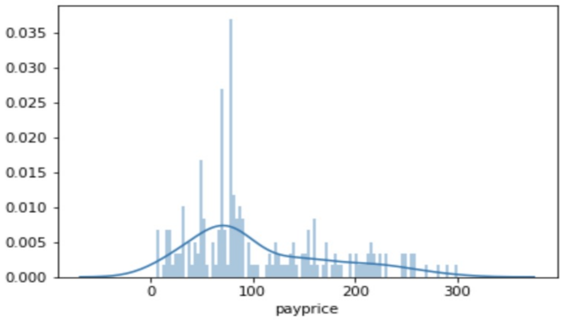

Figure 6: Payprice Distribution of Impressions have Clicks

Table 3: Xgboost Parameters Setting

learning_rate	0.1
n_estimators	100
max_depth	6
min_child_weight	7
gamma	0
subsample	0.8
colsample_bytree	0.8
reg_alpha	0.01
objective	binary:logistic
nthread	4
scale_pos_weight	1
seed	27

Table 4: Model Evaluation Metrics

Metrics	Xgboost	LR
TP	26	1
TN	303717	303719
FP	6	4
FN	176	201
Accuracy	99.940117%	99.932549%
Recall	12.871287%	0.495050%
Specificity	99.998025%	99.998683%
Precision	81.250000%	20.000000%
F1 score	22.222222%	0.966184%
AUC	0.885329	0.794

than the LR one, which shows the Xgboost model has much higher probability to give positive impressions higher pCTR than negative impressions [6].

3.2.2 *Bidding Strategy.*

The Optimal Real-Time Bidding strategy (ORTB) [3] is implemented and analyzed in this project. Comparing with the previously used linear strategy whose bid price is linearly related with the pCTR, ORTB includes the win function and budget constraint, which enables us to calculate the expectation of the total click number under the budget constraint. Since we want to maximize to the total click number in this project, the expectation of clicks is used as the objective function to be maximized. Theoretically, this expectation function should be a good mathematical representation of our object, which means we could expect a better performance.

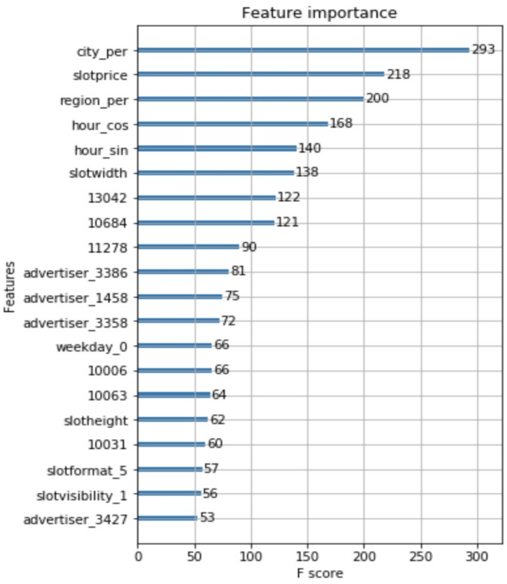

Figure 7: Top 20 Important Features

It could be hard to extract the exact win function from the raw data, so some win functions are designed as approximation of the real win functions. According to [3], the win function observed on the real data tends to be concave, and two win functions have been proposed by the authors. The c in the formulas represent a constant, which would be tuned as a hyper parameter.

$$winrate = \frac{bidprice}{c + bidprice} \tag{1}$$

$$winrate = \frac{bidprice^2}{c^2 + bidprice^2} \tag{2}$$

Based on these win functions above and some reasonable assumptions from the authors, the formulas of bid price are presented below. The λ is the Lagrangian multiplier generated from the calculation process, but it would be treated as another hyper parameter need to be tuned in this project.Both formulas are implemented and tested in the same way. The ORTB strategy using formula (3) to generate the bid price will be called ORTB1 and the strategy using formula (4) will be called ORTB2 in the following report.

$$bidprice = \sqrt{(\frac{c}{\lambda}pCTR + c^2)} - c \tag{3}$$

$$bidprice = c\left[\left(\frac{pCTR + \sqrt{c^2\lambda^2 + pCTR^2}}{c\lambda}\right)^{\frac{1}{3}} - \left(\frac{c\lambda}{pCTR + \sqrt{c^2\lambda^2 + pCTR^2}}\right)^{\frac{1}{3}}\right] \tag{4}$$

To tune the hyper parameters c and λ, first, for both formulas, the parameters are tried in large ranges using for loop to find the optimal ranges of parameters that could maximize the total click model achieves on the validation set. c is tried in the range of 10 to 500 with an interval of 10, and λ is tried in the range of 0 to 1.5e-05 with an interval of 5e-09. Although this method is time-consuming, the optimal ranges for c and λ can be found automatically. For both formula, the optimal range of c is found to be 30 to 180, and the ranges of λ are 1.5e-06 to 2.1e-06 and 2.2e-06 to 3.3e-06 for the ORBT1 and ORBT2 respectively. In the following analysis, we only try the parameter on the much smaller optimal ranges, which would improve our efficiency.

Moreover, the maximum click numbers are plotted to help us to tune the parameters. For example, figure 8 shows the maximum click number that achieved by λ with the c in the optimal range. We can see that when the λ is set to around 3.0e-06 the highest number of clicks is achieved at 166. Then, we can fix the λ to 3.0e-06 and plot the number of click verses the value of c to find the best c value for the given λ.

The table 5 shows the evaluation results of different strategies on the validation set. The optimized Xgboost model mentioned before is used by the strategies to estimate pCTR. We can see that the ORTB1 and ORTB2 have very similar performance, while the linear strategy shows slightly better performance than ORTB. This is different from our expectation that ORTB is expected to have better performance. However, in the following research, we find that it seems ORTB have higher requirement regarding pCTR estimation because it shows more improvement than linear strategy when our pCTR estimation becomes better. On the other hand, all the three models have much better performance than the linear model from the third part of the project (whose clicks is 142, see group report), which suggest that the Xgboost pCTR estimation model is much better than the LR model.

The threshold strategy is a novel solution that could increase CTR significantly. By comparing the pCTR distributions of impressions that are clicked and not clicked, I find that if the log pCTR of an impression is smaller than -9, this impression hardly can be clicked, so the bid price of impressions whose log pCTR is less than -9 is set to 0. This technique can be applied on every bidding model as long as a threshold can be found from its pCTR distribuion. The threshold model shown in table 5 is modified by the ORTB2 model. Comparing this with the ORTB2, we can see that the threshold model has the same clicks but higher CTR and lower CPC.

4 CONCLUSION

In this report,the impression dataset was analysed first. Then, Xgboost pCTR estimation was found to have better performance than LR model. After that, the ORTB and linear strategies are compared. The results showed that linear strategy have better performance with the given Xgboost CTR estimation model. Finally, a threshold technique is proposed to improve the model performance. In the future, better CTR estimation model should be explored, which may be able to fully maximize the performance of ORTB. Alternative win functions of ORTB should be experimented. Further exploration about data and the prediction results should be done to find new real time bidding solutions.

The table 6 describe the task allocation of our group.

REFERENCES

[1] Weinan Zhang, Shuai Yuan, Jun Wang, and Xuehua Shen. Real-time bidding benchmarking with ipinyou dataset. arXiv preprint arXiv:1407.7073, 2014.

[2] He X, Bowers S, Candela JQ, et al. Practical Lessons from Predicting Clicks on Ads at Facebook. Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining - ADKDD14. 2014. doi:10.1145/2648584.2648589.

[3] Weinan Zhang, Shuai Yuan, and Jun Wang. Optimal real-time bidding for display advertising. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1077âĂŞ1086. ACM, 2014.

[4] Ren K, Zhang W, Chang K, Rong Y, Yu Y, Wang J. Bidding Machine: Learning to Bid for Directly Optimizing Profits in Display Advertising. IEEE Transactions on Knowledge and Data Engineering. 2018;30(4):645-659. doi:10.1109/tkde.2017.2775228.

[5] Cai H, Ren K, Zhang W, et al. Real-Time Bidding by Reinforcement Learning in Display Advertising. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM 17. 2017. doi:10.1145/3018661.3018702.

[6] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-874. doi:10.1016/j.patrec.2005.10.010.

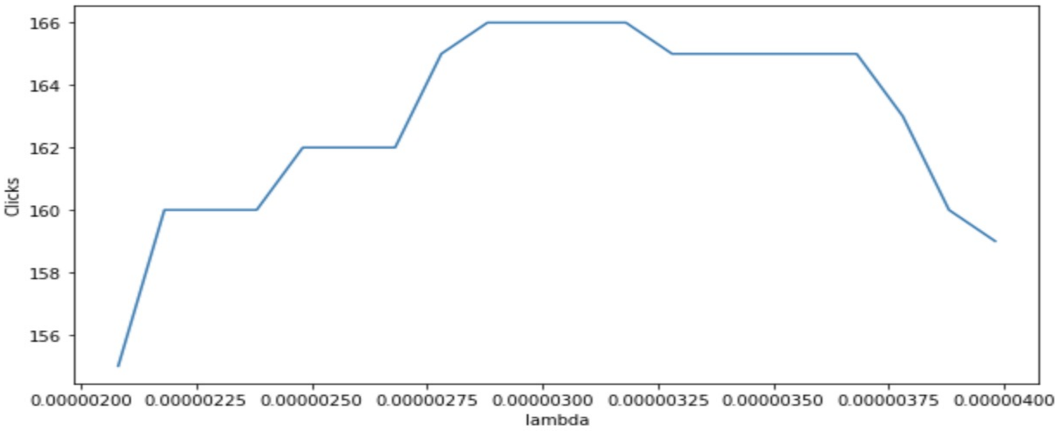

Figure 8: Maximum Clicks of ORBT2

Table 5: Strategy Performance Metrics

	Linear	ORTB1	ORTB2	Threshold
Parameters	131	100, 1.971e-06	110, 3.081e-06	110, 3.081e-06
Imps	137251	144224	142981	136592
CPM	45.44	42.97	42.44	43.76
CTR	0.001217	0.001151	0.001161	0.001215
Clicks	167	166	166	166
Cost	6236.66	6196.82	6068.39	5977.66
CPC	37.35	37.33	36.56	36.01

Table 6: Task Allocation

Tasks	Weisi	Qiuru	Boyang
Data Exploration	X	X	X
Feature Extraction	X		
Down Sampling			X
Constant Bidding	X		
Random Bidding		X	
CTR Estimation Model	X	X	X
Logistic Regression	X	X	X
Factorisation Machine	X		
XGBoost		X	
LGBM			X
DNN			X
Bidding Strategies	X	X	X
Linear	X	X	X
ORTB		X	
PRUD	X		
Combined Strategy	X	X	X
Multiagentt		X	