COMPGW02: Web Economics Project

Farhan Essa

I. Introduction

Over the past decade the most significant revolution in the advertising space for both mobile and digital advertisements has been the mass adoption of RTB.

RTB refers to real time bidding, which its self is a technology that allows for firms representing both buyers and sellers to bid on a price to be paid for an advert every time a banner is loading.

What this means is that when a user opens a webpage, thousands of bids are made by advertisers to display their ad to the user. The amount that is bid for these adverts is determined by the way in which the firms algorithms are programed. The widespread use of algorithms has led to an ever increasing number of ads now being sold on a single impression basis as opposed to the bulk purchases that were seen during the time of Ad Networks.

For this assignment we were initially asked to conduct in-depth exploratory analysis on a size able data set that had been provided for us. As part of this study we were asked to identify and evaluate statistical metrics such as number of impressions, number of clicks, cost, CTR, average CPM and eCPC.

Following the exploration and commentary of the data set we were required to develop a CTR prediction model, which when coupled with a self-defined bidding strategy would assist advertisers.

In order to best predict the CTR various models were tried and tested by myself, details of which can be found in the approach section of this report. It was identified through several metric comparisons that the best model that was implemented was a tuned implementation of the Random Forest Classifier which returned an AUC (area under the curve) value of 0.85. After conducting research into the various bidding strategies, I opted to implement an Optimal Real Time Bidding Model which enabled me to achieve a higher number of clicks.

II. Related Work

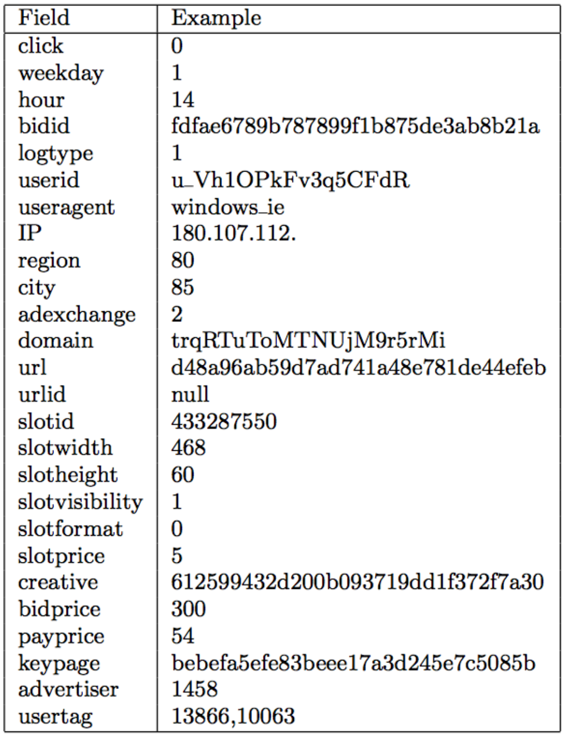
In order to assist and aid the success of the project I made use of various papers that have been referenced at the end of the document.

The first of these papers was the work of Weinan Zhang, Shuai Yuan, and Jun Wang [1], the paper outlined a detailed example of exploratory analysis which was used as the basis for the analysis that was conducted for this assignment.

Another paper that proved to be significantly useful was again the work of Weinan Zhang, Shuai Yuan, and Jun Wang [2]. This paper was used to further understand how non-linear models were implemented conceptually. The formulas contained within this paper allowed me to build an Optimal Real Time Bidding Strategy.

Data set	Training	Validation	Test
Columns	26	26	23
Rows	2697738	299750	299750

Fig. 1.



Field	Example
click	0
weekday	1
hour	14
bidid	fdfae6789b787899f1b875de3ab8b21a
logtype	1
userid	u_Vh1OPkFv3q5CFdR
useragent	windows_ie
IP	180.107.112.
region	80
city	85
adexchange	2
domain	trqRTuToMTNUjM9r5rMi
url	d48a96ab59d7ad741a48e781de44efeb
urlid	null
slotid	433287550
slotwidth	468
slotheight	60
slotvisibility	1
slotformat	0
slotprice	5
creative	612599432d200b093719dd1f372f7a30
bidprice	300
payprice	54
keypage	bebefa5efe83beee17a3d245e7c5085b
advertiser	1458
usertag	13866,10063

Fig. 2.

III. Data Exploration

A. Data Format

Three separate data sets were provided as part of this assignment, this includes the test, training and validation set. Details of these data sets can be found in figure 1.

The feature description and an example of each column within the training data set has been presented in Figure 2.

Whilst the training and validation sets contain all the columns mentioned in figure 2 the test set however defers from this as it does not contain the bidprice, payprice and click fields as this is the data that the final model will be evaluated against.

It should be noted that all monetary values such as the bid price make use of the RMB currency with the unit of Chinese fen * 1000, corresponding to the commonly adopted CPM pricing model.

Adv	Cost	Clicks	Imp	CTR	CPM	CRC
1458	37231.2	451	540293	0.083	68.92	82.55
3476	27481.4	175	346778	0.050	79.25	157.04
3427	36820.1	340	454031	0.075	81.10	108.29
3358	28145.3	233	304782	0.086	92.36	120.80
2259	13649.0	45	146778	0.031	92.99	303.31
2821	20625.8	144	231416	0.062	89.13	143.23
3386	38341.0	358	498554	0.072	76.90	107.10
2997	3413.2	251	54487	0.461	62.64	13.60
2261	10789.2	37	120619	0.031	89.45	291.60

Fig. 3.

B. Basic Statistics

In order to carry out the desired statistical analysis the following metrics had to be calculated from the data, these are:

- Impressions - Calculated by summing each occurrence of the same advertiser
- Clicks - Calculated by accumulating the number of clicks for each impression for each individual advertiser.
- The Click Through Rate (CTR) - Calculated by dividing the number of clicks by the number of impressions.
- The Cost Per Mile (CPM) - Achieved by dividing the total cost accumulated by the number of impressions.
- The Cost Per Click (CPC) - Accomplished by dividing the total cost accumulated by the number of impressions.

Data grouped by advertisers with the metrics mentioned above has been presented in figure 3.

From figure 3 it can be identified that that the advertiser with the highest number of clicks and impressions was advertiser 1458, whilst advertiser 2261 suffered from the lowest click rate and advertiser 2997 suffered from the lowest impression count.

The figure indicates that advertiser 2997 achieved the highest CTR value, this value can be seen to be significantly higher when equated to other advertisers which only achieved values of under 0.1 percent.

The lowest CTR value was achieved by both advertisers 2259 and 2261. The difference between the highest and lowest CTR value is significant at around 1387 percent.

IV. User Feedback

When comparing the CTR distribution throughout the week it can be seen that advertiser 1458 only witnessed slight fluctuations with only a significant drop at the end of the week. Whilst advertiser 3358 also exhibited this drop, the CTR values throughout the week varied greatly with a sudden dip on the second day of the week where the CTR value almost fell to 0 followed by an abrupt rise on the third day.

This variation between both advertisers was however not seen when analysing CTR distribution throughout the day as both reported similar values. The only key difference being that advertiser 3358 exhibited a noteworthy spike at 9pm. It is therefore safe to assume that throughout the day the CTR remains fairly constant.

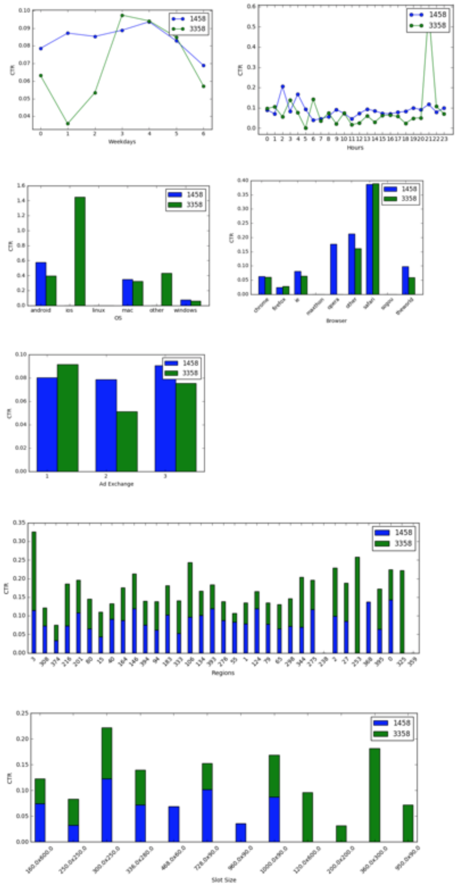


Fig. 4.

When analysing the operating system usage, it became apparent that the advertisers in question had been targeting similar platforms on the whole as fairly similar values for Android, Mac and Windows were attained. The only real difference being that advertiser 1458 witnessed a zero CTR for IOS whereas advertiser 3358 achieved its highest CTR.

After having analysed the values for OS usage it was safe to predict that due to the higher reported rates of CTR for iOS and MacOS that Safari would achieve the highest CTR in comparison to other browsers. This is exactly the behaviour that was observed by both advertisers. What surprised us was that the second highest CTR value was achieved by theworld browser as opposed to Internet Explorer or Google Chrome.

When observing the ad exchange distribution between the two advertisers it was spotted that advertiser 1458 achieved a greater CTR with exchanges two and three than advertiser 3358 that only achieved a higher value at exchange one. The largest difference occurring at exchange two where a difference of 24 percent was witnessed between the two advertisers. This variation is likely due to different advertisers opting to select different exchanges based on specific demands.

CTR distribution across different regions remained fairly matched for both advertisers however it was witnessed that advertiser 3358 achieved a significantly higher CTR in region 3 and in region 253 and 325 it essentially had a monopoly as little to no CTR was reported for advertiser 1458. What was shocking to see was that both advertisers failed to gain CTR in regions 238 and 359.

The CTR distribution for slot sizes can be observed to be fairly constant between the two advertisers. Advertiser 3358 achieved its highest CTR for the 300 by 300 slot whereas advertiser 1458 achieved its highest CTR for the 300 by 250 slot. It can be seen that both advertisers preferred squarer slots.

V. Best Bidding Strategy

A. Feature Engineering and CTR Prediction

Having identified the CTR prediction is a classification based data mining problem, I set my self a target of implementing two separate models to accurately predict the CTR as well as making use of feature engineering that was carried out by all of us in the team. The feature engineering section of the model is based around one hot encoding the features since this allows for better results when running classification models. In order to predict the CTR value the inbuilt predict proba function which is part of most Sklearn classification models was used.

The first model that was implemented was the SGDClassifier. The initial implementation using only the mandatory parameters returned an AUC of 0.65, however after fine tuning it by doing a grid search on alpha to obtain an optimum value of 0.00001 the model returned an accuracy of 0.72, which can be seen in figure 5.

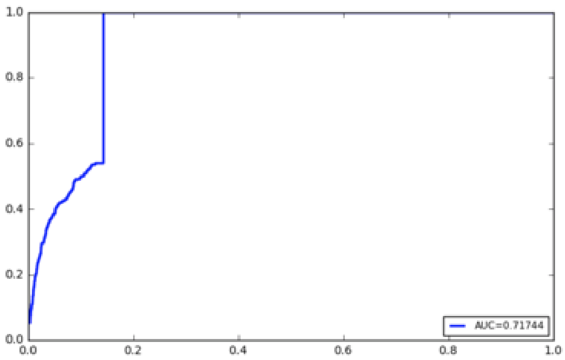


Fig. 5.

The second model that was implemented was the random forest classifier model. During its initial run it achieved an AUC of 0.73, however after fine tuning the model by conducting a grid search to find the optimum values for the number of estimators, minimum sample leaves and random state the model returned and accuracy of 0.85, which can be seen in figure 6.

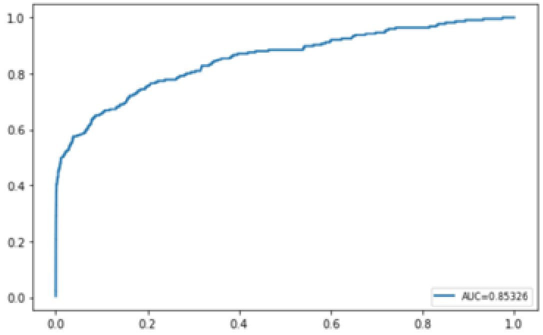


Fig. 6.

B. Bidding Strategies

Having been satisfied with the results obtained from the random forest classifier, the focus was shifted onto developing an optimal bidding strategy. After doing substantial research into the various bidding strategies such as eCPC and ORTB bidding, I decided to implement optimal real time bidding which is based on the works of Weinan Zhang, ShuaiYuan, and Jun Wang.

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

The above formula shows a mathematical representation of ORTB in order to calculate a bid price where c is a user defined constant and λ is a Lagrangian multiplier.

VI. Results and Comparison

Post CTR prediction model and strategy implementation both the linear and ORTB approaches were tested against the validation set with a fixed budget of 6250, which is a value that had been defined in the assignment brief. The results of these can be found in figure 7. It should be noted that the constant was set to 28 and the lambda value was set to $5e * 10 - 7$ for the ORTB model in this experiment.

Strategy	Clicks	CTR	CPC	CPM
Linear	117	0.2889	10.5963	30.6107
ORTB	157	0.1026	39.8083	40.8552

Fig. 7.

From figure 7 it can be seen that the number of clicks achieved through ORTB was about 34 percent higher than that value achieved by making use of the linear strategy. It can however also be seen that the CTR value for ORTB is almost one third of the value that was obtained for the linear bidding strategy.

Additionally the ORTB strategy returned higher values for both the CPC and CPM. Whilst the difference in CPM is not significant the difference in CPC is noteworthy, with the linear bidding strategy returning a value that was one fourth of that which was returned by ORTB.

In order to measure the performance of our strategies we all made use of a formula that was provided by one of our teammates and can be found below.

$$score = \frac{2 \times CTR \times clicks}{cpm \times cpc \times costs}$$

The rationale behind this formula is such that a good bidding strategy is determined by achieving higher values for CTR and clicks whilst minimising the values for CPC and CPM. The significance of the 2 in the formula is that it acts as a weight for the CTR, as this is the most important parameter. When this formula was applied to both strategies the values returned were 3*10-6 for linear bidding and 5*10-6 for ORTB as a result it is safe to assume ORTB is the better performing of the two strategies.

VII. Conclusion

In conclusion this assignment has allowed me to learn and demonstrate my ability to conduct a detailed exploratory analysis on a sizable online advertising data set. It has also given me a chance to exhibit my ability to both feature engineer through the use of one hot encoding and to implement and tune various classification models which include SGD and Random Forest with the aim of predicting CTR as accurately as possible. The code for which can be found at https://github.com/fessa94/WebEcon-UCL-UCABFFE.

Whilst the project ran relatively smoothly and all team members contributed equally it should be mentioned that the timing of the project severely hampered the rate of progress as team members were constantly busy with both interviews for graduate schemes and traveling back and fourths from the UK. Each group member did however deliver on creating a CTR prediction model and implementing a number of bidding strategies.

Whilst I feel that the project has been a success, there were however ways in which it could have been improved. The first of these I feel is that given more time and computing power a better learning algorithm could have been identified. A lot of my time went into tuning the two models that have been described in this paper and whilst I attempted to implement SVM, memory constraints meant that it was nearly impossible to compute and hence had to be looked over. I have added the code for my attempt in my repository.

The other way in which I believe I could have performed better is by spending less time tuning the models and more time actually implementing the different possible bidding strategies as time constraints meant that I was only able to implement ORTB and Linear bidding.

My individual contribution to the project was carrying out feature engineering, creating the most accurate CTR prediction model and implementing both the ORTB and linear bidding strategies. It should be noted that I was not the only one in the team to have implemented the ORTB strategy.

VIII. References

[1] Bid optimizing and inventory scoring in targeted online advertising, 2012.

[2] Optimal Real-time Bidding for Display Advertising, KDD 14. ACM, 2014.

[3] T. Graepel, J. QuiA sonero Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoftaA Z s bing search engine. In Proceedings of the 27th International Conference on Machine Learning ICML 2010, Invited Applications Track (unreviewed, to appear), June 2010.

[4] X.He,J.Pan,O.Jin,T.Xu,B.Liu,T.Xu,Y.Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. Candela. Practical lessons from predicting clicks on ads at facebook. In ADKDD@KDD, 2014.

[5] R. J. Oentaryo, E. P. Lim, D. J. W. Low, D. Lo, and M. Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. ACM Conference on Web Search and Data Mining, 2014.

[6] W. Zhang, S. Yuan, and J. Wang. Real-time bidding benchmarking with ipinyou dataset. CoRR, abs/1407.7073, 2014.