Predict the Click-Through Rate and Average Cost Per Click for Keywords Using Machine Learning Methodologies

Lihui Shi
Centerfield Media
El Segundo, CA
shilihui@uw.edu

Bo Li
School of Arts
Wuhan Sports University
Wuhan, Hubei, China
82bobo@163.com

Abstract

In search engine optimization (SEO), advertisers bid on tons of keywords on Google, for example, so that their clickable ads can appear in Google's search results. In order for the advertisers to maintain their budget more efficiently and achieve the best performance, Google Adwords provides all the performance data for the advertisers' keywords, but the advertisers need to know what is the best amount they should bid for each keyword, and what would be the keyword performance with that bid. A very important task for advertisers is to use the historical data to predict the click-through rate (CTR) and average cost per click (CPC) for a keyword with a set of features. The CTR and average CPC are two essential metrics to measure the paid search performance on keyword level, so that the advertisers will be able to optimize the bids and achieve the highest profits for their Google Adwords accounts. We try to predict the CTR and average CPC of a keyword using some machine learning methods. We use cross validation to evaluate the results and find the optimal predictors for the CTR and average CPC.

Keywords
Search Engine Optimization, Paid Search, Google Adwords, Click Through Rate, Average Cost Per Click

1. Introduction

1.1 Paid Search and Google Adwords Introduction

In search engine optimization (SEO), usually the pay per click (PPC) model is employed on the paid advertising. PPC is an internet advertising model used to direct traffic to websites, in which an advertiser pays a publisher (typically a website owner or a network of websites, such a Google, Bing, Yahoo, The New York Times, CNN.com, etc) when the ad is clicked. Google AdWords, Bing Ads, Yahoo Gemini are the top three paid search websites on which advertisers bid on certain keywords in order for their clickable ads to appear in the search results. Take Google as an example, there are hundreds of millions of keywords are bid on Google Adwords by tons of advertisers every day, and different advertisers compete with each other on their commonly interested keywords to be more competitive on Google search results page, in order to attract more traffic from the internet users, including impressions, clicks, and eventually conversions. Meanwhile, the advertisers need to pay Google a certain amount of money, whenever their ads are clicked by the internet users. Therefore, the advertisers need to find out the optimal point between the spend and revenue on Google, so that their budgets can be spent as efficiently as possible and they can achieve the highest possible profits.

Google Adwords employs the generalized second-price auction, in which each advertiser does not pay its own bid whenever their ads are clicked, instead, he pays the advertiser's bid right behind him. For example, the highest bidder

pays the price bid by the second-highest bidder, the second-highest pays the price bid by the third-highest, and so on. The auction plays out in an automated fashion every time a visitor triggers the ad spot [1, 2].

Due to the current bidding mechanism, advertisers need to plan how much they are willing to pay Google for each keyword carefully. Advertisers optimize the bid on keywords, adjust or rewrite website content and site architecture, in order to achieve a higher ranking in search engine results pages to enhance pay per click (PPC) listings.

On Google Adwords, when bidding on a keyword in their PPC campaigns, the advertisers need to choose a keyword match type, which tells Google how aggressively or restrictively the advertisers want it to match their advertisements to keyword searches. There are different match types including exact match, broad match, phrase match, modified broad match, or negative match. Different keyword match types have different definitions, and they help control which searches can trigger the ad. For example, in the exact match case, the user query must exactly match the bid terms. In the broad match case, the bid terms can be related more loosely, such as being a subset of the query words. Obviously that broad match usually gets the highest traffic since it allows Google to interpret the keywords the advertisers are bidding on and map the keyword to many search terms, and exact match often gets the lowest since it is the most restrictive one. Most advertisers would bid the highest on exact match and the lowest on broad match. In this paper, we will only focus on the keywords under exact match on Google Adwords, and discuss how we can apply machine learning methodologies to accurately predict the CTR and average CPC on the keyword level.

1.2 Some Metrics on Google Adwords

There are some basic and important metrics on Google Adwords, which can also be shown on Google Adwords User Interface (UI), and their definitions and explanations are given below:

- Impression: an impression is counted each time the advertiser's ad is shown on a search result page on the Google. Each time advertiser's ad appears on Google, it's counted as one impression.

- Click: when someone clicks the advertiser's ad, like on the blue headline of a text ad, Google AdWords counts that as a click.

- Conversion: a conversion happens when someone clicks the advertiser's ad and then takes an action that the advertiser has defined as valuable to his business, such as an online purchase or a call to the advertiser's business.

- Click-through rate (CTR): the number of clicks that the advertiser's ad receives divided by the number of times his ad is shown expressed as a percentage (clicks/ impressions = CTR). A high CTR is a good indication that users find the advertiser's ads helpful and relevant.

- Conversion rate (CR): the average number of conversions per ad click as a percentage (conversions/clicks = CR). The conversion rates vary a lot across different industry on Google Adwords. Landing page optimization is the key factor leading to conversion rates improvement.

- Maximum CPC (bid): a bid that the advertiser sets to determine the highest amount that he is willing to pay for a click on his ad. If someone clicks the advertiser's ad, that click won't cost him more than the maximum CPC bid that he sets.

- Average CPC: the average amount that the advertiser has been charged for a click on his ad. Average CPC is calculated by dividing the total cost of the clicks by the total number of clicks. The advertiser's average CPC is based on his actual CPC he is charged for a click on his ad other than the max CPC he sets.

- Quality score: an estimate of the quality of the advertiser's ads and landing pages triggered by that keyword. Having a high quality score means that Google thinks the advertiser's ad and landing page are relevant and useful to someone looking at his ad. A 1-10 range quality score can be shown for any keywords.

- Average position: a statistic that describes how the advertiser's ad typically ranks against other ads. This rank determines in which order ads appear on the Google search results page. The highest position is "1," and there is no "bottom" position. An average position of 1-8 is generally on the first page of Google search results, etc. The advertiser's rank can change, causing its position on the page to fluctuate as well, so the average position can give the advertiser an idea of how often his ad beats his competitors' ads for position. The most important thing about average position for the advertiser is to find what's profitable for him, which might not be to show in the top position. Ad position of a keyword on Google is determined by a metric called Ad rank.

- Ad rank: a value used to determine the advertiser's ad position and whether his ads will show at all. Ad rank is calculated using the advertiser's max CPC bid, quality score and the expected impact of extensions and other ad formats. Notice that this metric is not shown on Google Adwords UI, and it is more like a mysterious part.

Besides the above metrics provided by Google Adwords, the advertisers have their own metrics such as revenue and profit, as well as revenue per click (RPC), and they are all important performance metrics and are often used by the advertisers to optimize their SEO campaigns and to improve their performance on Google Adwords.

1.3 Bid Optimization for CTR and Average CPC on Google Adwords

For most advertisers, the objective function to be maximized is the profit. For each keyword, we have:
profit = revenue - cost = (RPC - CPC) × clicks = (RPC - CPC) × impressions × CTR.

As can be seen from this formula, both CTR and average CPC are important metrics that can determine the performance of the advertiser's SEO campaigns. Therefore, it is a crucial question on how to accurately predict the CTR and average CPC on each keyword for the advertisers.

Since our goal is to predict the CTR or average CPC of a keyword, we cast it as a predictive modeling problem – that is, to predict the CTR or average CPC given a set of features. However, it is not an easy task for the CTR and average CPC prediction, since both are impacted by many features on Google Adwords.

In general, all the features related to the keyword performance can be categorized into two types: contextual features and historical features. The contextual features represent the current information regarding the context in which an ad is shown, such as the device used by the users, the number of words used in the ad title and in the body of an ad, the length of the URL, the individual words and terms used in the title and the body, etc. On the contrary, the historical features depend on previous performance for the keyword or the ad, for example, the CTR, the average CPC, the average position, etc, during a certain period in the near past. In aggregate, researchers found that those historical features provide considerably more power than those contextual features for CTR prediction.

2. Data Set with Keyword Performance

First, let's take a look at the features that have impacts on a keyword's CTR. It is well known that the CTR of a keyword decreases significantly when its average position gets lower. Since usually a higher bid for a keyword is more likely to earn a higher position for this keyword on Google search result page, so the advertiser's max CPC definitely has an impact on the keyword's CTR. The recent impressions, clicks, cost, average CPC, etc, all have some exploratory power on the CTR prediction.

Second, for the features that have impacts on a keyword's average CPC, we consider all the previous features related to the CTR prediction as well. However, compared with the CTR prediction, the average CPC prediction is even more complicated and challenging. Due to the generalized second-price auction employed by Google Adwords, the CPC an advertiser charged for a click of a keyword is influenced by the bids from other advertisers, i.e., his competitors as well. In general, the details of the relationship between average CPC and maximum CPC are the study of many works on search engine auction models.

Due to the hierarchical structure of Google Adwords account settings, under the same accounts there are different campaigns, and under the same campaign there are different adgroups, then under the same adgroups there are different

keywords. For each keyword, we also have different segments such as devices. We assume that the Google Adwords account is maintained well by the advertiser, so there are no duplicate keywords within the different adgroups or campaigns. Therefore, we only need to look at the data on the keyword level with different dates.

Match type also play an important role for the keyword performance. For example, for the same keyword, the broad match usually attracts much higher traffic volume than the exact match, even with a much lower bid. Due to this reason, it is not easy and fair to compare the performance between the different match types, therefore we will only work on the search exact match types, as mentioned earlier.

It is obvious that all advertisers always continue to adjust the bid for their keywords, either raise or drop, in order to improve the performance. For simplicity, in this date set, if the bid of the keyword was adjusted on that date, we only use the max CPC before it was adjusted in that row. Also, the advertisers sometimes pause the keywords which continuously have a bad performance, and for simplicity, we will only study the current enabled keywords in the advertiser's Google Adwords account.

The data set we use in this paper is assumed to be collected from the same advertiser, and it includes over 6000 keywords on his Google Adwords account which have impressions during the second quarter of 2016. For the same keyword, it appears in multiple rows, and each row represents this keyword's performance on a given date, as long as this keyword has nonzero impressions on that date. In total, there are over 200,000 rows (records) in this data set.

More specifically, each records in the data set includes the following variables:

(1) Date
(2) Keyword
(3) Impression, clicks, average position and max CPC on this day
(4) Impression, clicks, average position, cost, CTR and average CPC during the past 7 days
(5) CTR and average CPC during the next 7 days

3. Some Machine Learning Methods for CTR and Average CPC Predictions

3.1 Regression

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'response variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship. There are a variety of different regression methods such as linear regression, nonlinear regression, logistic regression, nonparametric regression, etc.

In our data set, we found the approximate linear relationship between the CTR and a few other performance metrics, so we use linear regression for simplicity. Due to the same reason, the average CPC prediction is applied by linear regression as well.

3.2 Random Forest

Random forest is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is a collection of decision trees that together produce predictions and deep insights into the structure of the data. When the training set for the current tree is drawn by sampling with replacement, a certain proportion of the cases are left out of the sample. This out-of-bag data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

Random forest corrects for decision trees' habit of overfitting to their training set. Random Forest is one of the most powerful, fully automated machine learning techniques. To implement random forest method, we only need to start with a suitable collection of data including variables we would like to predict or understand and relevant predictors, then with almost no data preparation or modeling expertise, analysts can effortlessly obtain surprisingly effective

models. Since all the variables are continuous in our data set, when random forest is applied for such data, it is a regression tree instead of classification tree.

3.3 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. "Gradient boosting" represents "gradient descent" plus "boosting". In gradient boosting, the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. Gradient boosting is widely used in many applications due to its easy use, efficiency, accuracy and feasibility.

4. Results and Analysis

By checking the data set in details, we calculate that the overall CTR for all the keywords during the second quarter of 2016 is 12.78%, the overall average CPC is $11.49. Both metrics have large variations, with standard deviation of $5.21 for average CPC during the past 7 days, and standard deviation of 12.4% for CTR during the past 7 days. Notice that the calculation of both CTR and average CPC depend on the date range period selected. The CTR does not exist if there is no impression during this period, and the average CPC does not exist if there is no click during this period. Since CTR is not reliable when the number of impressions are very low during the period selected for its calculation, and we did not differentiate the high volume and low volume keywords, so the CTR can be unreliable sometimes in the data set, with a huge standard deviation.

Since there are 200,000 records in our data set, in order to make the computation faster enough for those machine learning algorithms so that we can run a large number of iterations, we randomly select 10% of the whole records to work on, meanwhile most of the keywords still remain in this new data set. Now with a much smaller data set of 20,000 records, in the cross validation for each machine learning algorithm, 10% of those records were randomly sampled as the testing set in each iteration, and the other 90% of the records are used as the training set.

After running each machine learning method with 50 iterations for both CTR prediction and average CPC prediction and observing the numerical results, the findings can be summarized below:

Table 1. Feature Importance on CTR Prediction for Each Machine Learning Methods

	Linear Regression	Random Forest	Gradient Boosting
First feature	CTR during the past 7 days	Max CPC	CTR during the past 7 days
Second feature	Average position during the past 7 days	Average position during the past 7 days	Max CPC
Third feature	Current average position	CTR during the past 7 days	Average position during the past 7 days

Table 1 shows that for the CTR prediction, each machine learning method shows different significance orders for the features. Overall, CTR during the past 7 days plays the most significant role, followed by the max CPC and the average position during the past 7 days. Therefore, the top three important features for the average CPC prediction are pretty clear, as the above three. This result implies that the future CTR for a keyword highly depends on its CTR recent history, and the higher max CPC can improve the average position of the ad, and its future CTR as well.

Table 2. Feature Importance on Average CPC Prediction for Each Machine Learning Methods

	Linear Regression	Random Forest	Gradient Boosting
First feature	Max CPC	Max CPC	Max CPC
Second feature	Average CPC during the past 7 days	Average CPC during the past 7 days	Average CPC during the past 7 days
Third feature	Average position during the past 7 days	Average position during the past 7 days	Average position during the past 7 days

From Table 2 we can see that for the average CPC prediction, each machine learning method shows that the max CPC plays the most significant role in the average CPC prediction, followed by the average CPC and the average position during the past 7 days. Therefore, the top three important features for the average CPC prediction are pretty clear, as the above three. This result implies that the bid itself is the most significant feature to decide its average CPC in the near future charged by Google, and its recent history on average CPC and average position also are very important to predict its average CPC in the near future.

Now we evaluate the accuracies of the three machine learning methods for their CTR prediction and average CPC prediction. Mean squared error (MSE) is used to evaluate the accuracy of the different algorithms, by comparing the predicted CTR or predicted average CPC with the true CTR or average CPC respectively for each record, with the squared loss function. Table 3 and Table 4 below are summaries of the MSEs for the different machine learning methods in each iteration, with their means and standard deviations after 50 iterations.

Table 3. MSEs on CTR Prediction for Each Machine Learning Methods in 50 Iterations

	Linear Regression	Random Forest	Gradient Boosting
Mean	2.05%	2.04%	2.59%
Std Dev	0.21%	0.20%	0.21%

Table 4. MSEs on Average CPC Prediction for Each Machine Learning Methods in 50 Iterations

	Linear Regression	Random Forest	Gradient Boosting
Mean	6.82	5.97	7.25
Std Dev	0.61	0.52	0.44

As can be seen from Table 3, linear regression and random forest plays equally well for the CTR prediction, with gradient boosting the worst. Their prediction accuracy is very high, with only 2% discrepancy from the real CTR on average. However, all three methods almost have the same variation in their prediction errors.

Table 4 shows that random forest provides the best prediction on average CPC, also with gradient boosting the worst, which has the lowest variation in prediction error, interestingly. The prediction accuracy of average CPC is much worse than the CTR prediction, with nearly $6 discrepancy from the real average CPC with the best method, i.e., random forest. This can be explained by the fact that there are other factors impacting the average CPC as well, especially other advertisers' bids and quality scores, which cannot be known.

5. Conclusions and Future Work

In this paper, we investigate the keywords' CTR and average CPC prediction problems on Google Adwords using a wide range of performance features. Different machine learning methods, including regression, random forest and gradient boosting are applied to evaluate the prediction performance on both metrics. We find that random forest turns out to be the best method for both the CTR prediction and the average CPC prediction, while the gradient boosting gives the most inaccurate results.

Though it is interesting to determine the best features, and how much each feature may overlap with other features, we believe that ultimately, the best practice is to include as many features as possible in the final model. In our future work, we would like to collect and explore more features, including the user level data, such as the demographical

features. We will also evaluate other features such as device, the time of the day as well as the day of the week for the keyword's ad appears on Google. We will also investigate the more complicated and dynamic bid scenario with changing bids for the keywords, using the philosophy of more proactive methods in process control and adjustment [3,4].

References

[1] Dembczynski, K., Kotłowski, W., and Weiss, D., Predicting ads' click-through rate with decision rules, *WWW2008*, Beijing, China. April 21–25, 2008.
[2] Kitts, B., Laxminarayan, P., LeBlanc,B.,and Meech, R., A formal analysis of search auctions including predictions on click fraud and bidding tactics, *Proceedings of the First Workshop on Sponsored Search, Sixth ACM Conference on ECommerce*, June 2005.
[3] Shi, L., Kapur, K., A synthesis of feedback and feedforward control for process improvement under stationary and nonstationary time series disturbance models, *Quality and Reliability Engineering International*, vol. 31, no. 3, pp. 343-354, 2015.
[4] Shi, L., Kapur, K., Quasi-feedforward and feedback control for random step shift disturbance models, *Quality Technology & Quantitative Management*, vol. 12, no. 1, pp. 69-82, 2015.

Biography

Lihui Shi is a senior data scientist in Centerfield Media at El Segundo, CA, USA. He earned B.S. in Information and Computational Science from Hebei University of Technology, Tianjin, China, Masters in Statistics from Nankai University, Tianjin, China and University of Washington, Seattle, WA, USA and PhD in Industrial and Systems Engineering from University of Washington, Seattle, WA, USA. He has published papers on journals and conferences, such as Quality and Reliability Engineering International, Quality Technology and Quantitative Management, International Journal of Performability Engineering, etc, on different research areas such as statistical process control, process adjustment, parametric and nonparametric statistics, supervised and unsupervised learning, etc. Dr Shi has been working in data science team of several IT companies and completed research projects on A/B & Multivariate Testing on web analytics, paid search (Google Adwords, PLA and display ads), and bidding models using machine learning methods. His research interests include quality and reliability, applied statistics and machine learning. He is a member of IIE, INFORMS, ASQ (American Society for Quality) and ASA (American Statistical Association).

Bo Li is an Associate professor in School of Arts at Wuhan Sports University, Hubei, China. He holds a Bachelor's degree in Physical Education and a Master's degree in Physical Education and Sports Training, at Wuhan Sports University, Hubei, China. He received his PhD from the National University of Physical Education and Sport of Ukraine, in Kiev, Ukraine. He taught courses in Latin dance, Jazz and physical education. He has research interests in physical education, dance kinesiology and data analysis in dancing performance. He has published over 10 papers in academic journals and international conferences, in English, Russian and Chinese.