Wilfrid Laurier University

BU495Q – Business Analytics

Final Report – Bidding Strategy for Search Engine Marketing

Prepared for Michael Pavlin

5th April, 2015

Prepared by:

Jarome Leslie - 203663930

Jumana Bahrainwala - 203849630

Table of Contents

1.0 Definition of Business Problem

Marketing today has evolved from traditional advertisements on radio and television to an online marketplace where placement of an ad on a particular webpage is prime property. Companies engage in Search Engine Marketing (SEM) to compete against each other for the positioning of their advertisements on search engine results on the basis of particular keywords searched by users. The following analysis seeks to calculate the maximum bid price with respect to a targeted keyword or group of keywords in order to breakeven on the cost of an advertisement. The information is examined from the point of view of a company currently engaged in search engine marketing. The data sets were generated based on internal information and information from a search engine provider.

2.0 Description of the Data

Two datasets were used in the course of this analysis. The first dataset, "SEM_DAILY_BUILD.csv", is comprised of advertisement performance data across a range of dates. The descriptions of the columns in this dataset can be found in the legend in Exhibit 1. Most of the data is in categorical format. The data includes information on number of clicks per ad, the number of applications per ad, the number of times ad shows up in search (impressions), the revenue received from each ad. The second dataset, "SEM_KEYWORD_SUMMARY_S.csv" contains keyword specific information across a range of dates. The descriptions of the columns are summarized in Exhibit 2 and include key fields such as keyword quality score, estimated first page bid and estimated top page bids. These two datasets were concatenated into one dataset based on unique values in both sets and called SEM_DAILY_BUILD_AUGMENT.csv (build dataset). Our models and assumptions were based on the historical value of the build dataset. Once we finished vetting our models, we then applied our prediction results to the validation dataset "SEM_DAILY_VALIDATION.csv".

3.0 Analysis

The analysis begins with a brief description of the plan of action. This is followed by a description of the data transformations performed, the modelling techniques employed, and the model performance.

3.1 Plan of Action

Our goal is to determine the breakeven cost of an advertisement, where breakeven cost is the maximum dollar amount a company is willing to pay for an ad with certain keywords in it. The maximum bid price is a function of three inputs. The three inputs are as follows:

- **Conversion Rate** – percentage of clicks that become applications
- **Approval Rate** – the percentage of product applications that translate to sales of any product
- **Revenue** – the dollar amount received from an ad.

From these inputs breakeven cost is calculated using the following formula:

$$\textbf{Maximum bid price} = \textbf{Revenue} \times \textbf{Conversion Rate} \times \textbf{Approval Rate} \qquad \textbf{(1)}$$

For example: If Revenue is \$150, Approval rate is 60% and Conversion Rate is 10% , the maximum bid amount would be 150 * 10% * 60% = \$9.

The conversion rate and revenue models are based on the build dataset. These are then used to predict the conversion rate and revenue of the validation dataset. We calculate max bid for the validation data.

3.2 Data Transformations

In order to prepare the datasets for the modelling stage, several transformations were performed. These transformations are summarized in Table 1 below and are provided in the accompanying zip file.

Transformations 1 and 2 were performed in MS Access and resulted in the creation of two new datasets “SEM_DAILY_BUILD_AUGMENT.csv” and “SEM_DAILY_VALIDATION_AUGMENT.csv.”

Transformations 3 and 4 were performed in R and resulted in the addition of new columns to the dataset.

Table 1: Summary of Data Transformations

Transformation	Description
1	A unique key based on common variables was created between the two datasets:

	REF: [WEB_ACTVY_DT] & [ENGN_ID] & [CMPGN_NM] & [KEYWD_ID]
2	A unique key was made to create commonality between the build, validation and keyword datasets:
3	A "conversion" field was added to the build dataset as: Conversion = # of Applications / # of Visits
4	Total revenue was calculated as the sum of the revenue of the 6 products: Totalrev = $\sum^{6}_{i=1}$ revenue of product i
5	The maximum bid price was calculated for the build and validation datasets using equation (1)

3.2 Modelling

3.2.1 Modelling for Conversion Rate

Given that Conversion Rate calculates the odds of a click turning into an application, logistical regression was selected as the method of choice. After trying many iterations, the model we used is the one that minimized the residual deviance and null deviance, in comparison to our other iterations. The model comprised of the following variables: CMPGN_NM, IMPRESSIONS, VISITS, QUALITY_SCORE. The model was not an ideal fit for the data as the Residual deviance and Null deviance are still considerably large as shown in Exhibit 3. This shows that the parameters added only marginally improve the model.

3.2.2 Constant Approval Rates

Approval Rate is defined as the number of applications that were approved for a given advertisement. We assumed the approval rate was 60% for any given product based on the case details given to us.

3.2.3 Modelling for Revenue

Linear regression was used to model the revenue generated from an ad. We used linear regression as revenue can be modelled as linear combination of things like visits, impressions etc. The following inputs were used in this model: VISITS, CONDITIONAL_IMPRESSIONS, IMPRESSIONS, IMPRESSION_TOTAL_RANK, DVIC_ID, TOTAL_QUALITY_SCORE, QUALITY_SCORE, and FIRST_PAGE_BID. This model yielded an adjusted R-squared value of 0.6701 meaning that approximately 67% of the variation observed is explained by the model.

3.2.4 Predicting the Maximum bid price

After building and training the model using the build dataset, the conversion rate and revenue models were applied to the validation dataset to make predictions. These values were used to calculate the maximum bid price according to the formula in equation (1).

3.3 Problems with Our Model and Assumptions

Given the high residual deviance value produced by the conversion rate model and the adjusted R-squared value produced by the revenue model, there definitely exists room for the models to be improved. It is also important to note that approval rates most likely vary in reality and would therefore add another layer of complexity to this problem.

4.0 Results

The results of this analysis are contained in the accompanying file, “Validation_Results.csv.” This file contains a column denoting the maximum bid price for a particular advertisement and the keywords it targets. Thus, each record of this file contains a prediction and the overall file corresponds to a bidding strategy. The column corresponding to the bid amount for each row indicate what the maximum price the company must bid for keywords within a certain adgroup.

5.0 Recommendations

Equipped with this bidding strategy for search engine advertising, the company should ensure that its bids never exceed the projected maximum bid price. This will ensure that the company is able to maximize its marketing investments.

There is space to improve the model and the fit for the model is there is more data. Based on different bid amounts, one can also run a clustering algorithm to find out if a certain range of bids map to certain keywords.

Exhibit 1: SEM_DAILY_BUILD Data Dictionary of Relevant Fields

Field Name	**Explanation**
WEB_ACTVY_DT	The date on which the SEM data was collected
ENGN_ID	Search Platform the keyword search was performed on
LANG_ID	Language setting of Search Platform (e.g. English / French)
DVIC_ID	Device Accessed On (i.e. desktop, tablet or mobile)
CMPG_NM	Unique Key for Keyword and Ad group
AD_GRP_NM	Information on message, language and keywords
KEYWD_ID	ID assigned to a the string of keywords used in search
MTCH_TYPE_ID	Match type ID: broad, exact
TOTAL_BID_AMOUNT	Bid Amount for KEYWD_ID
AD_ID	An indicator representing the unique combination of Ad Content
IMPRESSIONS	Number of time Ad was shown in search
CONDITIONAL_IMPRESSIONS	Number of Ad with non-zero quality score
TOTAL_QUALITY_SCORE	Sum of quality scores of each keyword in ad
CLICKS	Number of clicks on an Ad
SPEND	Amount spent on a click
IMPRESSION_TOTAL_RANK	Sum of the Rank for a keyword where keyword has impression
VISITS	Number of website visits tracked
APPLICATIONS	Number of applications
APP_APPROVED	Number of applications approved
APP_DECLINED	Number of applications declined
APP_PENDING	Number of applications pending
APPS_PROD_[1-6]	Number of applications submitted for PRODUCT [1-6]
PROD_[1-6]_APPROVED	Number of PRODUCT [1-6] applications approved
PROD_[1-6]_REVENUE	Total Revenue from all approved PRODUCT [1-6] bookings
QUALITY_SCORE	Keyword Quality Score
FIRST_PAGE_BID	Bid Amount needed to land on google first page
TOP_PAGE_BID	Bid Amount needed to be on top of first page

Exhibit 2: SEM_KEYWORD_SUMMARY_S Data Dictionary of Relevant Fields

Field Name	**Explanation**
WEB_ACTVY_DT	The date on which the SEM data was collected
ENGN_ID	Search Platform the keyword search was performed on
KEYWD_ID	ID assigned to a the string of keywords used in search
CMPGN_NM	Unique Key for Keyword and Ad group
AD_GRP_NM	Information on message, language and keywords
KEYWD_TXT	Keyword Text - String of Words that were used in the Search; separated by a "+"
KEYWD_STATUS	Keyword status: active or paused
MTCH_TYPE_ID	Match type ID: broad, exact
QUALITY_SCORE	Keyword quality score
FIRST_PAGE_BID	*(Google estimated) Bid amount required to place ad in the first page of the users search result
TOP_PAGE_BID	*(Google estimated) Bid amount required to place ad at the top of the first page of the users search result
DESTINATION_URL	The URL address of the webpage that prospects reach when they click on the Ad

Exhibit 3: Conversion rate model Summary

```
summary(convmodel)

Call:
glm(formula = conversion ~ data$CMPGN_NM + IMPRESSIONS + data$VISITS + 
    data$QUALITY_SCORE, family = binomial, data = data)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.17749  -0.60225  -0.51572   0.09358   2.70883  

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -1.495e+00  3.605e-02 -41.471  < 2e-16 ***
data$CMPGN_NMGS_CMPGN1_LANG3  -2.067e-03  6.678e-02  -0.031 0.975311    
data$CMPGN_NMGS_CMPGN10_LANG2 -4.223e-01  8.089e-02  -5.220 1.79e-07 ***
data$CMPGN_NMGS_CMPGN10_LANG3 -2.146e+00  1.972e-01 -10.880  < 2e-16 ***
data$CMPGN_NMGS_CMPGN11_LANG2 -6.513e-01  7.316e-02  -8.902  < 2e-16 ***
data$CMPGN_NMGS_CMPGN11_LANG3 -7.770e-01  1.373e-01  -5.657 1.54e-08 ***
data$CMPGN_NMGS_CMPGN2_LANG2  -8.306e-01  5.162e-02 -16.091  < 2e-16 ***
data$CMPGN_NMGS_CMPGN2_LANG3  -5.993e-01  6.193e-02  -9.678  < 2e-16 ***
data$CMPGN_NMGS_CMPGN4_LANG2   2.654e-01  4.357e-02   6.093 1.11e-09 ***
data$CMPGN_NMGS_CMPGN4_LANG3   3.457e-01  5.525e-02   6.256 3.95e-10 ***
data$CMPGN_NMGS_CMPGN5_LANG2  -8.861e-03  4.613e-02  -0.192 0.847673    
data$CMPGN_NMGS_CMPGN5_LANG3   1.906e-01  5.406e-02   3.527 0.000421 ***
data$CMPGN_NMGS_CMPGN6_LANG2  -4.545e-01  4.216e-02 -10.781  < 2e-16 ***
data$CMPGN_NMGS_CMPGN6_LANG3  -1.188e-01  4.697e-02  -2.529 0.011448 *  
data$CMPGN_NMGS_CMPGN8_LANG2   2.052e-01  7.884e-02   2.603 0.009243 ** 
data$CMPGN_NMGS_CMPGN8_LANG3  -8.932e-01  1.788e-01  -4.997 5.84e-07 ***
data$CMPGN_NMGS_CMPGN9_LANG2  -2.346e-01  4.074e-02  -5.759 8.44e-09 ***
data$CMPGN_NMGS_CMPGN9_LANG3  -1.114e-01  4.601e-02  -2.421 0.015495 *  
data$CMPGN_NMYB_CMPGN1_LANG2   1.306e+01  1.970e+02   0.066 0.947127    
data$CMPGN_NMYB_CMPGN1_LANG3   8.024e-01  1.225e+00   0.655 0.512567    
data$CMPGN_NMYB_CMPGN10_LANG2  1.494e+00  1.415e+00   1.056 0.290817    
data$CMPGN_NMYB_CMPGN2_LANG2   1.033e+00  1.615e-01   6.398 1.57e-10 ***
data$CMPGN_NMYB_CMPGN2_LANG3   1.229e+00  3.293e-01   3.732 0.000190 ***
data$CMPGN_NMYB_CMPGN3_LANG2   1.306e+01  1.970e+02   0.066 0.947127    
data$CMPGN_NMYB_CMPGN3_LANG3   1.496e+00  1.415e+00   1.058 0.290252    
data$CMPGN_NMYB_CMPGN4_LANG2   4.637e-01  5.222e-01   0.888 0.374556    
data$CMPGN_NMYB_CMPGN4_LANG3  -1.007e+01  9.848e+01  -0.102 0.918552    
data$CMPGN_NMYB_CMPGN5_LANG2   1.306e+01  1.970e+02   0.066 0.947128    
data$CMPGN_NMYB_CMPGN6_LANG2   7.980e-01  6.134e-01   1.301 0.193298    
data$CMPGN_NMYB_CMPGN6_LANG3   1.306e+01  1.970e+02   0.066 0.947131    
data$CMPGN_NMYB_CMPGN9_LANG2   1.516e-01  3.869e-01   0.392 0.695252    
data$CMPGN_NMYB_CMPGN9_LANG3  -2.963e-01  1.081e+00  -0.274 0.783985    
IMPRESSIONS                    5.893e-05  3.027e-05   1.947 0.051533 .  
data$VISITS                   -7.307e-04  4.673e-04  -1.564 0.117837    
data$QUALITY_SCORE             1.864e-01  6.830e-03  27.292  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61474  on 109339  degrees of freedom
Residual deviance: 59275  on 109305  degrees of freedom
AIC: 90390

Number of Fisher Scoring iterations: 10
```

Exhibit 4: Revenue model Summary

```
summary(revmodel)

Call:
lm(formula = data$TotalRevenue ~ data$VISITS + data$CONDITIONAL_IMPRESSIONS +
    data$IMPRESSIONS + data$IMPRESSION_TOTAL_RANK + data$TOTAL_QUALITY_SCORE
+
    data$QUALITY_SCORE + data$DVIC_ID + data$FIRST_PAGE_BID)

Residuals:
    Min      1Q  Median      3Q     Max
-2546.0    -9.4     0.3    10.2  4748.3

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.526e+01  2.810e-01 -54.304  < 2e-16 ***
data$VISITS                   5.003e+00  1.666e-02 300.196  < 2e-16 ***
data$CONDITIONAL_IMPRESSIONS  1.683e-01  2.544e-03  66.159  < 2e-16 ***
data$IMPRESSIONS             -1.487e-01  4.719e-03 -31.517  < 2e-16 ***
data$IMPRESSION_TOTAL_RANK   -4.476e-03  1.142e-03  -3.921 8.82e-05 ***
data$TOTAL_QUALITY_SCORE     -3.332e-03  5.123e-04  -6.503 7.89e-11 ***
data$QUALITY_SCORE            2.422e+01  2.128e-01 113.802  < 2e-16 ***
data$DVIC_IDM                -5.351e+00  4.509e-01 -11.867  < 2e-16 ***
data$DVIC_IDT                -2.903e+00  4.896e-01  -5.928 3.07e-09 ***
data$FIRST_PAGE_BID           4.573e+00  3.562e-01  12.836  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.08 on 109330 degrees of freedom
Multiple R-squared:  0.6701,   Adjusted R-squared:  0.6701
F-statistic: 2.468e+04 on 9 and 109330 DF,  p-value: < 2.2e-16
```