PRELIMINARY ANALYSIS

Ad portfolio analysis - Optimizing conversions

Balázs Dömény

ABSTRACT
I have found this simulated dataset on my computer from about a year ago, (I know it is sloppy but I'm not sure what it was used to illustrate as an example). The use I thought of it for now is to deliver a preliminary analysis of this dataset - as if it was real google analytics data, which it was supposed to represent anyway - and point towards directions for a more extensive analysis. Though I try to keep my code clean and readable, it probably is helpful to summarize the work done in present document.

1. Exploring the data - first look

The data consists of 43450 records (rows) with 20 attributes for each record. The first 3 are the assigned ID-s, in total 1205 unique keywords grouped in 184 groups and used across 64 campaigns - quite the summary. Summarizing the number of keywords contained in the individual groups, we get 1245 total keywords, which means there are only 40 instances of a keyword already present in another group. One can make the first observation here, namely that there is barely any overlap ($< 5\%$) present among the sample's campaigns. In building a first - and most naive - approximation, one would not make a huge mistake by assuming no overlap at all. The distribution of keywords among the groups is irregular, with mean of 6.8 word/group, but the median is at only 2/group.

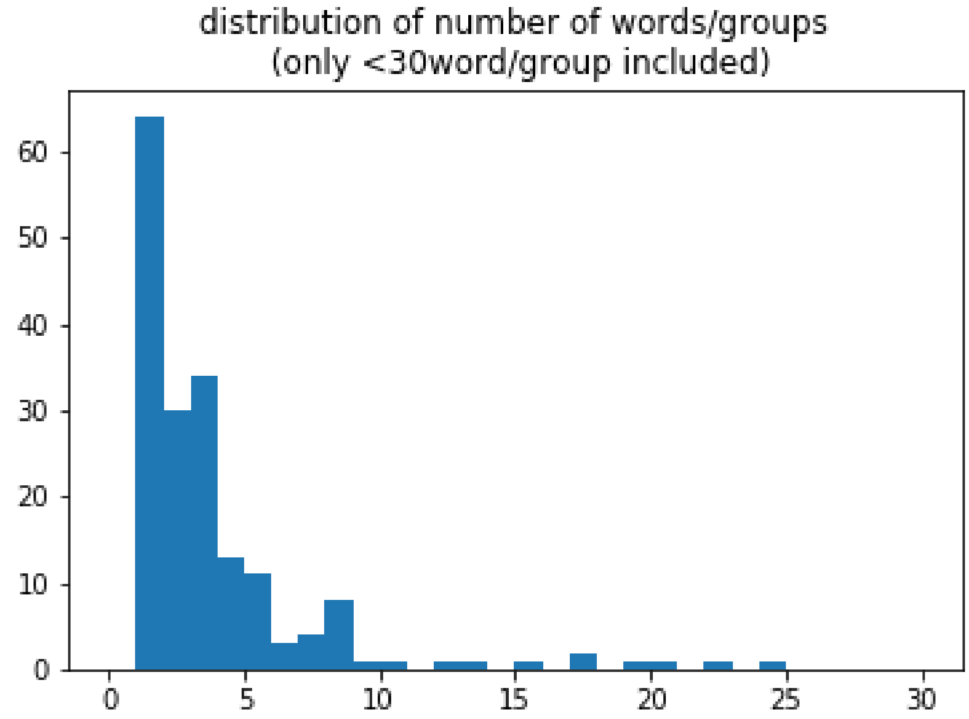
Among other columns the Campaign_type has 2 unique values ('Search only' and 'Search Network with Display Select'), the latter with only 561 instances. The two population does not seem to show significant differences, and considering the small size of the latter (compared to both the first type and the whole sample) I think it is safe to not use this distinction.

I've removed the Campaign subtype field, since it contained the same value ('All features')for every record.

I've also removed Campaign 9 from the sample, since it was a campaign with 0 total impressions during its 4 month period.

1.1. *Time, Effectiveness*

The question would rise whether the key attributes (Impressions, Clicks, conversions and attributes connected to these) change during the length of a campaign or the time of the year for any trend to be identifiable. Since we have data from a period of 371 days - just a little bit longer than a year - associating 'spikes' with specific times of year should be handled with caution or even suspicion. Fortunately there is only one such spike is found in the data, in and around December - this is in all honesty not a



huge surprise.

There are no other significant spikes in either the time of the year nor the length of the campaigns.

After looking through the time-series belonging to most meaningful aspects, I can confidently say that the 'date' column can also be excluded from further investigation until maybe a very refined state of the model.

I have decided to make the individual campaigns the 'building blocks' of the statistics, mainly because this way I will be less likely to introduce bias inside the individual blocks with the time factor (If an adgroup is used in a short and a long campaign, it might be harder to handle due to the discrepancy in the number of records)

To determine the success of an ad campaign I have decided to use two quantities:

- Effectiveness $= \frac{\text{Income}}{\text{Investedamount}}$
- and total profit (income - invested amount)

I can do this, since according to the dataset, every single conversion means income, so there is no instance of success measurable solely in name-recognition, subscription or some other more elusive gain.

After describing the basic properties of the data and populations, I chose to examine two approaches to the task at hand.

2. Differences in campaign-populations

In the first approach I aimed to divide the full sample into sub-populations with statistically different distributions to hint towards the connections between different

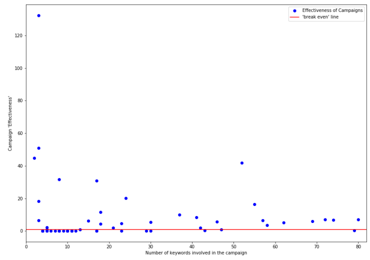


Figure 1. Effectiveness of campaigns - Lower wordcounts can achieve higher values, but have a larger chance of falling flat - As one would expect, the larger the wordcount the smaller the relative deviation is (I suspect $\sigma \sim \sqrt{N}$ or something similar holds true with these type of scenarios.)

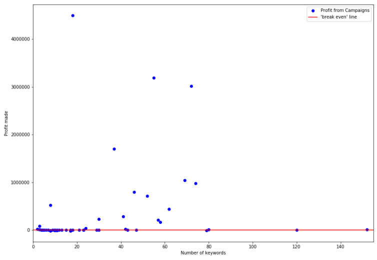


Figure 2. Income from campaigns - Interestingly most (if not all) of the profit is brought in by campaigns with intermediate wordcount

properties. I've divided the sample into three populations by the following criteria:

- **Pop1**: this is the largest population, campaigns which made profit.
- **Pop2**: Population of campaigns which didn't manage to make a profit and lost money but don't belong to Pop3.
- **Pop3**: Campaigns with 0 successful conversion.

2.1. *1D distributions*

After viewing different distributions of the above mentioned populations, one can conclude that they are not completely without overlap, but their distributions (Fig. 3) are distinguishable, which can hint towards a good starting point.

2.2. *2D distributions*

Of course, connections are easier to find and show if one uses both axes available: Some values are bound to be connected and this should be visible with the correct methods. Here I should mention, that in some cases it is justifiable to leave Pop3 out from the plots for the sake of a cleaner picture

I have explored additional relations (most of the plots are present in the code), but these two are the ones I chose to include.

Interesting takeaway after this section is the fact that Pop3 did make so far 30896 (of the used currency) disappear and result in 0 converted clicks (7384 total clicks). With the average CR of pop1, one would expect ∼74 converted clicks out of a similar situation, but even with that of pop2, on average 5.5 clicks would have been accumulated. It seems to be an obvious first choice to reallocate the funds of these campaigns and hopefully using them to make pop2 profitable.

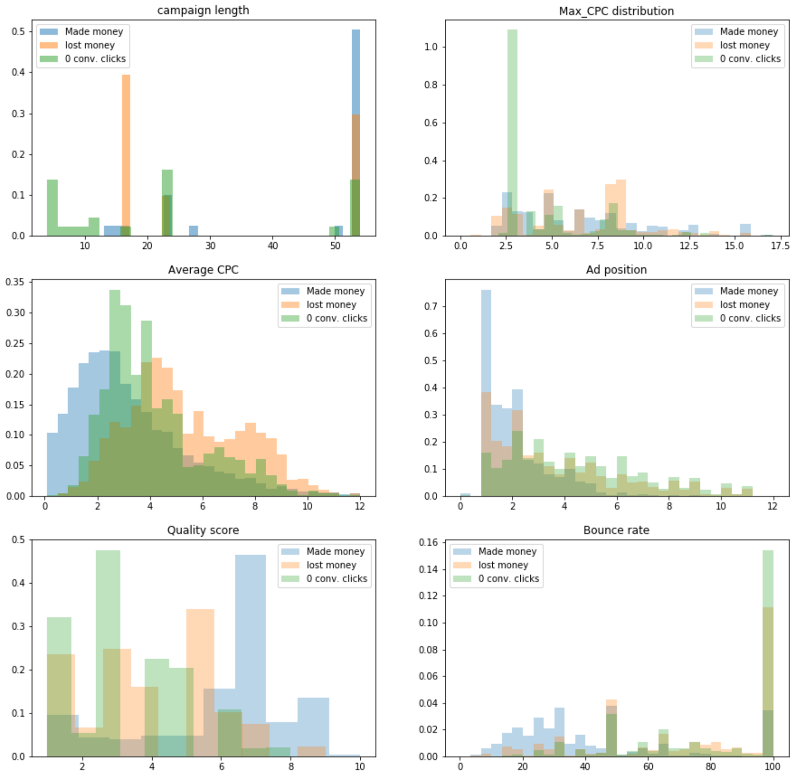


Figure 3. Some relevant distributions of the populations

3. Building a model

After reviewing the populations present in the sample, I have decided to try to create a very simplistic and naive model, which on the other hand should be able to give an idea - 0^{th} approximation - to what values should be optimized. I have started with a few basic assumptions to ensure that the model doesn't become too complex:

- Every change should be written with one equation, where the previous value is transformed into the next one.
- these equations have the shape of $A \cdot f(x) = B$, where A and B are the values transformed, and f(x) is a *friendly* function with exactly one variable (I did deviate from the one variable rule later).
- At first I choose this one variable arbitrarily based on my understanding of the relations, but if there seems to be no correlation, I'll choose another.
- I call *friendly* those functions which are either low-order polynomials (< 4) or are commonly recognisable by shape (e.g. Gaussian, e^x...).

Note: I use $f_n()$ to describe the unknown function in each equation, this does not mean that these functions have to in any way be similar to each other

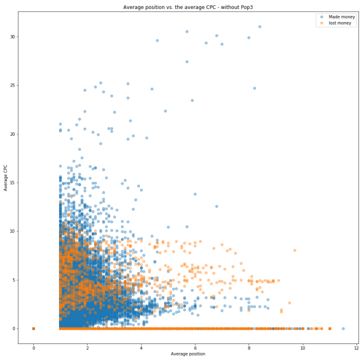


Figure 4. The two different populations are distinguishable. An interesting feature is the subpopulation of high CPC and not-so-good ad position, which despite this seems to be profitable. In their cases there are arguments for both increasing and decreasing the invested amount.

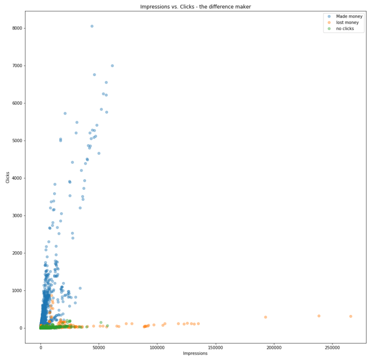


Figure 5. In the case of Impressions and clicks, there is an interesting trend, pop1 and pop2 are almost 'mirroring' each other on the different axes. It is worth noting that while Pop1 looks like as if it could be approximated with a linear equation, the case is not this simple - just the perspective of the stretched x-axis results in the assumption.

3.1. *Starting equations*

We start from a keyword which - by default - has a Quality score. This and the max. CPC together determines the rank, which in turn should give the number of Impressions

$$f_0(max.CPC, Quality) = Impressions.$$

This f_0 likely has other, hidden parameters (such as the average frequency of searches regarding the keyword) which could likely be obtained either through Google analytics or other means, but for now they are unavailable.

$$Impression \cdot f_1(Avg\_position) = Clicks.$$

The f_1 in the second equation is the CTR (*Click Through Rate*), which is recorded in the dataset (and if it wouldn't be, it could be easily calculated) At this step the Clicks can be multiplied by the CPC (*Cost Per Click*) (which is also a function of the Max. CPC and the rank), to get the invested amount.

From the clicks one step further is

$$Clicks \cdot f_2(X) = Converted\_clicks,$$

Where we want to maximize this latter value. X is supposed to represent the likelihood that the searcher did find their goal on the site. A fine indicator for this the quality score, but I'd also factor in the match type (Exact, Phrase, Broad) too, although not sure about the execution.

3.2. *Constructing the functions*

3.2.1. Impression

It became quickly obvious that there is a variable that outranks both variables of f_0, which is supposed to be the keyword Id. Of course, the number of impressions of a word is ultimately controlled by the number of searches for the given word, and the 'auction' for ad space just starts after the fact. This is a data I don't have, which means f_0 has to stay unresolved for now.

3.2.2. CTR

The CTR however seems to be a clean shape which could have an envelope of $c \cdot \frac{1}{x}$ (Fig. 7) . Indeed, plotting CTR as a function of $\frac{1}{Avg.position}$ gives a triangle shape, and if one includes the number of impressions too (as a 2nd variable), it determines the slope of the envelope. This though means we can conclude a general shape of

$$f_1(Avg.position, Impressions) = CTR \leq p \cdot \frac{1}{Impressions \cdot Avg.position}.$$

This is something that can be approximated and the p parameter can be fitted for keywords or groups. Once it has been fitted, the optimum can be found numerically or analytically.

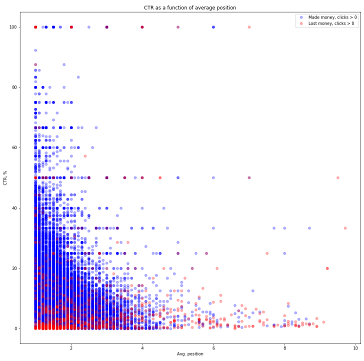


Figure 6. Shape of $f_1(Avg.position)$. Important to note here, that the Pop2 campaigns (red) didn't have even one instance of > 1 conv. clicks on a week. (/keyword). So the 100% or 50% CTR in their case means 1, or 2 total clicks respectively

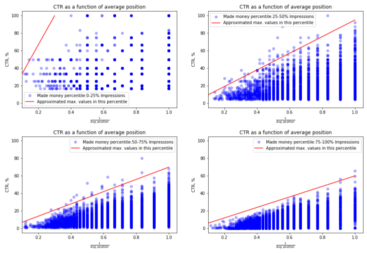


Figure 7. Transformed f_1 plot, with slopes of max. values present. The percentile's effect on the slope is visible. *Note: the envelope slopes in this case weren't made by numerical fitting.*

3.2.3. CR

The third equation once again didn't show easily identifiable shape, but after several plots, CR for large enough click numbers (in the regime where statistical errors didn't matter as much) does seem to be a constant CR=3% with $\sigma = 2.5$ deviation. This conclusion however is not definitive, it might be a more complicated function of multiple things.

3.3. *Takeaway*

Out of the three equations I've found a simple analytical shape for 1 which can be used to start the optimisation and redistribution of sources for a better outcome.

4. Final words

I have started two approaches to start and pick directions for a more comprehensive investigation, and to build a naive model which can be used as first approximation to the problem. I think while both more populations (in the first case) and more complex functions can be introduced to achieve a greater degree of success. I also intend to fit these naive functions and see how they would work in reality, and probably work something out regarding the re-allocation of funds.