Correlation One Exercise

Ruben Bloom (rmb042@gmail.com)

What is your one sentence executive summary?
Changes in overall engagement by click-through rate have been minor notwithstanding variations in content delivered, however differences in user topic preferences are large and it is worth investing in a well-designed personalization system.

What is your detailed assessment for a technical audience? Please quantify, use technical jargon.

Change in click-through rate over three months
The overall click-through rate (CTR) for links sent out, measured week-by-week, exhibited a u-shape over three month period January - March 2015. CTR started at 17%, dropped to a low of 16% in late February, and rose again to 16.6% by late March. This variation is all within 1%, where a 1% change CTR is equivalent to a change of ~7500 links viewed. I do not have the information required to judge if these variations have practical significance, it will depend on the value of a link-viewing. See plot 1) below.

Content Variation
Although it was stated the company makes frequent changes to the content sent to users, this was not readily apparent at the global level. Plots showing the proportion of links by topic and link by content type showed only minor variation over time. See Plot 3) below.
It is possible that there were more significant adjustments in the content to sent to individuals.

Some topics do have much larger variation in the number of articles with that topic sent to each user, i.e. some users receiving vastly more of that topic type and other less. This suggests at least some personalization. I did not examine how this changed over time.

Engagement
In this dataset all users appear to be engaged. Mean CTR for users is 16.4%, and with lower quartile engagement of 15.3%, upper quartile of 17.5%.

User Preference for Topics
The popularity of topics, measured by CTR for each topic, vary significantly. See plot 2) below. The most popular topics have CTRs above 30%, whereas the lowest have CTRs in the single digits. However this result is conflated by differences in the volume of links of each topic sent out. There is a negative correlation of -0.54 between mean CTR and number of a given topic links sent out.

It may be that the CTR for topics always decreases as more are sent out. Perhaps users will only read so many articles on a given topic.

A similarly strong correlation holds between number of links of given topic sent out and the standard deviation of CTR between users. (Is this just noise due to few emails per user?).

In any case, the CTR for any given topic is large, with IQRs of 10-20%. However this is a quick pass and has not be adjusted for different total CTRs between users (although these were checked as mentioned and found to vary only by a small amount on average).

Looking directly at users, there is a difference of 30% between a user's CTR for their favour topic and their average CTR. The average difference is 50% between highest and lowest.

The above all suggests that users have strongly differing preferences among topics, and it is will worth constructing a well-designed personalization system.

User Preference for Content-Types

Types do not show much variation as topics in CTR. See plot in iPython notebook. The average user CTR for each content type are close to the same, regardless of the volume sent out. Content-type does not seem to matter, perhaps because the content type is not salient before users click, while topic is. Personalization of content should not worry about type.

Correlations between Topic Click-Through Rates

User click-through rates for different topics did not show any correlations, i.e. users who tended to read Entrepreneurship articles were not more likely to read Business Development articles. This is surprising and makes me doubt the calculation.

Final Notes

This is a quick pass. Really need control for 1) variation in number of links by topic sent out, 2) difference in overall engagement of users. The latter was checked for and in small, but might make more of a difference.

What techniques did you try?

The approach was primarily descriptive statistics and visualisation. The focus was on appropriate aggregation of data and comparisons.

Some correlations were checked for.

What three plots did you make to explain the data?

See plots at end of document.

1) Plot of overall click-through rate over the three months.

2) Bar Plots of user CTR by topic. Shows both variation in popularity of topics overall, and variation in popularity of topics by user.

3) Normalised area plot of topic proportions over time. Reveals no obvious changes in topic content at a global level which might have affected click-through rates.

What is your commercial recommendation for business unit heads who are non-technical?

As user preferences for different topics vary significantly, in order to maximise engagement it is well-worth investing in a well-designed email personalization system.

What other data would you like to see about the platform? What questions would this additional data help you answer?

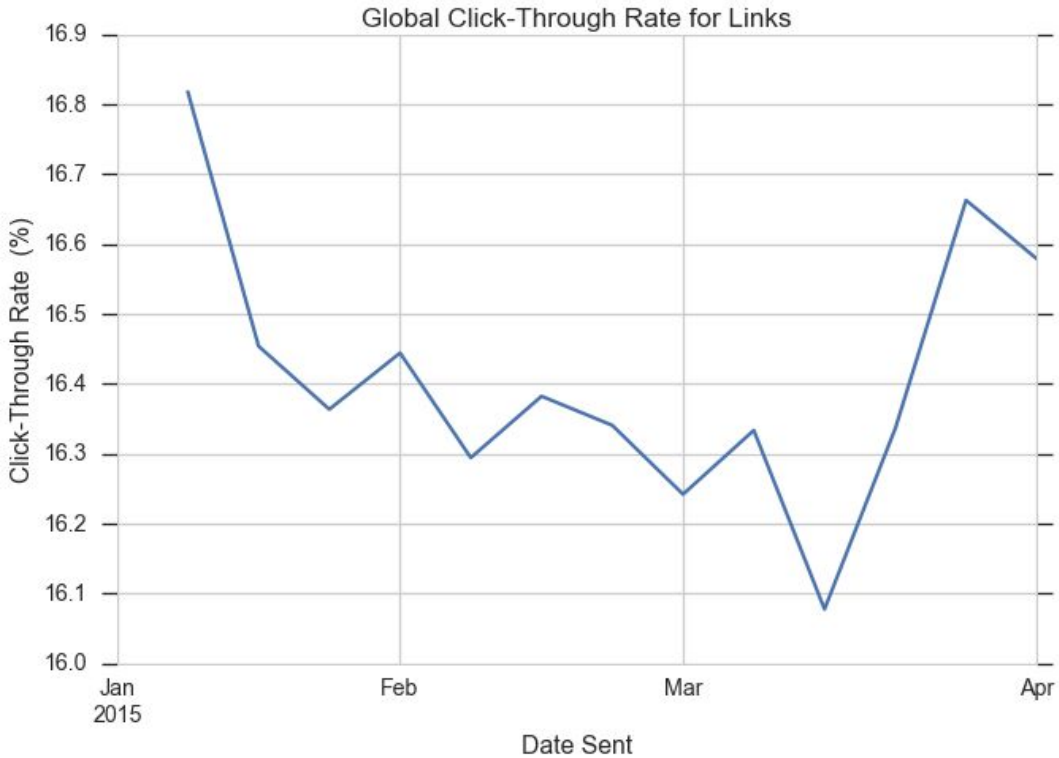
1) Demographic information about users. How did they sign up for the newsletter? Do they have registred topic preferences on the website?

This would aid in answering questions about subpopulations within the users with different interests.

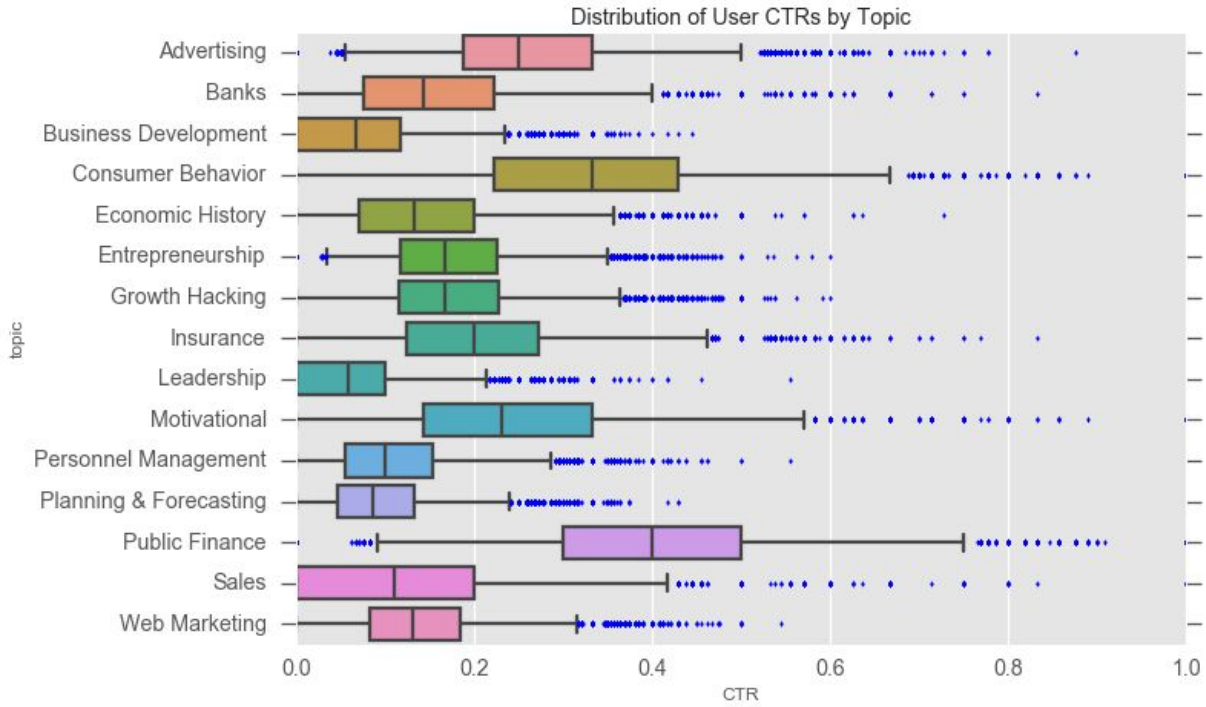
2) Session length after clicking. How long do you users stay on a link after clicking through? This provides information about whether they liked an article they thought they would like.

Plots

1) Plot of overall click-through rate over the three months.



2) Bar Plots of user CTR by topic. Shows both variation in popularity of topics overall, and variation in popularity of topics by use.



3) Normalised area plot of topic proportions over time. Reveals no obvious changes in topic content at a global level which might have affected click-through rates.