Analysis of the A/B Test:

Description:

This case focuses on an improvement to Yammer's core "publisher"—the module at the top of a Yammer feed where users type their messages. To test this feature, the product team ran an A/B test from June 1 through June 30. During this period, some users who logged into Yammer were shown the old version of the publisher (the "control group"), while other users were shown the new version (the "treatment group").

Problem:

On July 1, you check the results of the A/B test. You notice that message posting is 50% higher in the treatment group—a huge increase in posting. Now we need to figure out if this feature is a real deal or too good to be true.

Hypothesis on the results:

There are a number of factors that could have impacted the results. A few of them are mentioned here:

1. Wrong calculation of the A/B tests.
2. The users in control and treatment groups might not be totally isolated from each other, in other words, the experiment might not have been random.
3. Posting rates are not a good measure to judge the performance of the new feature, it should be investigated if valuable content Is created here

Investigating the results:

For the sake of this case study we have access to the following four tables:

Table1: Users : This table includes one row per user and relevant user account information

Table2: Events : This table includes one row per event, here event can be login events, message events, search events, …

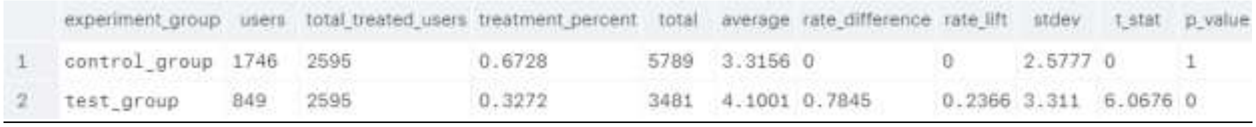
Table 3:Experiments: This table shows which groups users are sorted into for experiments

Table4: Normal distribution: A lookup table for Z-scores and values

First we can argue that the method chosen to carry-on the A/B test was not the most suitable one. As also mentioned in the problem description it might make a difference to consider a 1-tailed or 2-tailed t-test. I however did not explore this here. User separation could also be done in a much smarter way in that, users new to the feature, would the ones who are also familiar with the previous version and their usage of the new feature would be influenced by their prior experience.

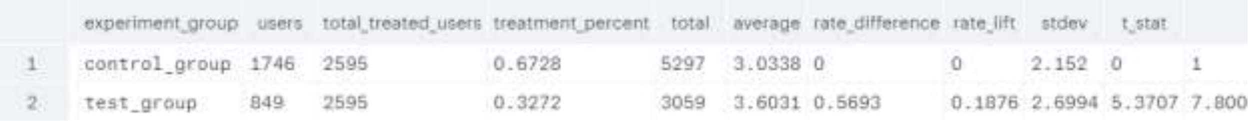
Here using information from the above tables we can investigate the values of other metrics to make sure that they will also score high after introducing the new feature and thereby do not rely only the messaging rate.

Using the login information from Table 2 (event_name='login'), for example we notice that users are logging more into the system on average:



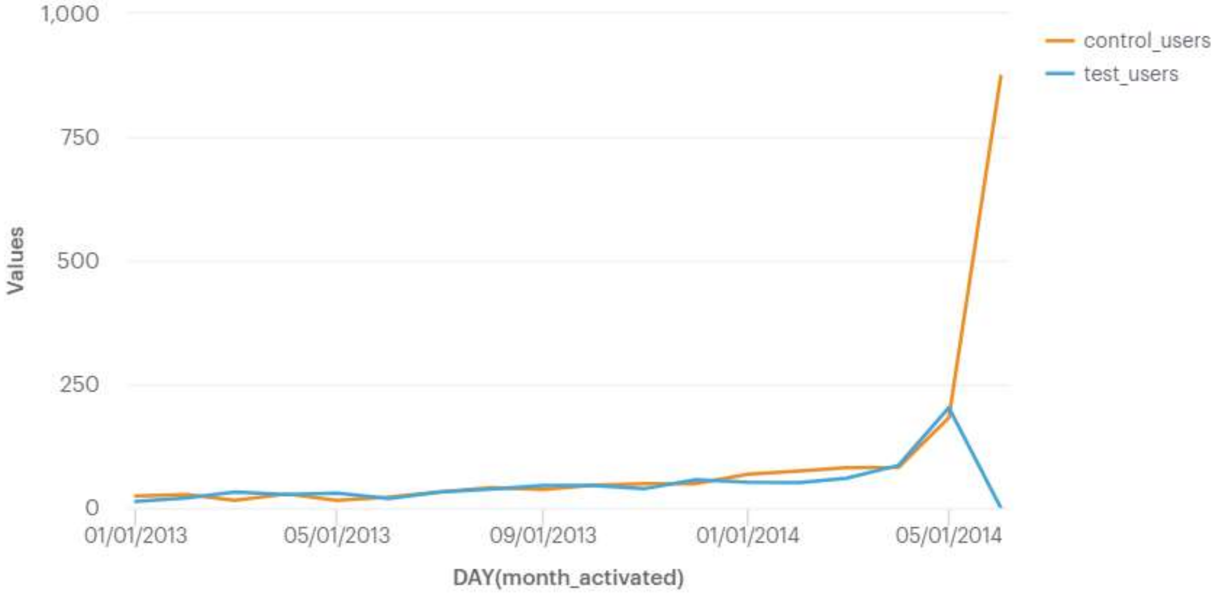
	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	stdev	t_stat	p_value
1	control_group	1746	2595	0.6728	5789	3.3156	0	0	2.5777	0	1
2	test_group	849	2595	0.3272	3481	4.1001	0.7845	0.2366	3.311	6.0676	0

In the above table the average column refers to the average logins per user for the control and test groups on more days (through counting distinct days using the 'occure_at' measure) as the following query results implies (just to make sure there are not login problems:



	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	stdev	t_stat	
1	control_group	1746	2595	0.6728	5297	3.0338	0	0	2.152	0	1
2	test_group	849	2595	0.3272	3059	3.6031	0.5693	0.1876	2.6994	5.3707	7.800

Using the information in the experiments and users table and Mode charts functionality we observe that



New users were assigned to the control group which consequently means that they have less time to post. Therefore, here we have identified one potential problem that might explain the outcome of the A/B test. It would also be worth to evaluate the content posted by the users in the test group. One explanation can be the existing users are purely "evaluating" the new feature and do not actually produce any content of value.