Toggle code

Data Scientist Challenge by Peggy Fan

Executive Summary

Overall, the likelihood of a user viewing has increased over the three-month period, with most of the increase happening after the first month.

Detailed Assessment

T-tests show that there was statistically significant increase in the number of clicks between January and February and a decrease in the clicks between February and March that is not statistically significant. The average of click-through-rate (CTR) of all users on a daily basis stays relatively consistent throughout the 3-month period, around 16%. There are no notable differences across months in terms of clicks by hour or by weekday. Logistic regression shows that odds of click are associated with emails that are sent earlier in the day, earlier in the week (noted: there were no emails sent on the weekend), and in the earlier month.

Most of the clicks come from 60 out of the 105 topics, and the majority of the topics are related to finance. Top 20 topics are show below:

0 [Public Finance] 1 [Advertising] 2 [Entrepreneurship] 3 [Growth Hacking] 4 [Consumer Behavior] 5 [Web Marketing] 6 [Unemployment] 7 [Budgeting] 8 [Industry: Retailing] 9 [Insurance] 10 [Motivational] 11 [Personnel Management] 12 [Economic History] 13 [Planning & Forecasting] 14 [Business Ethics] 15 [Banks] 16 [Mortgages] 17 [Management Science] 18 [Exports & Imports] 19 [Financial Accounting]

Tools used

Python (pandas), Sqlite

Techniques tried

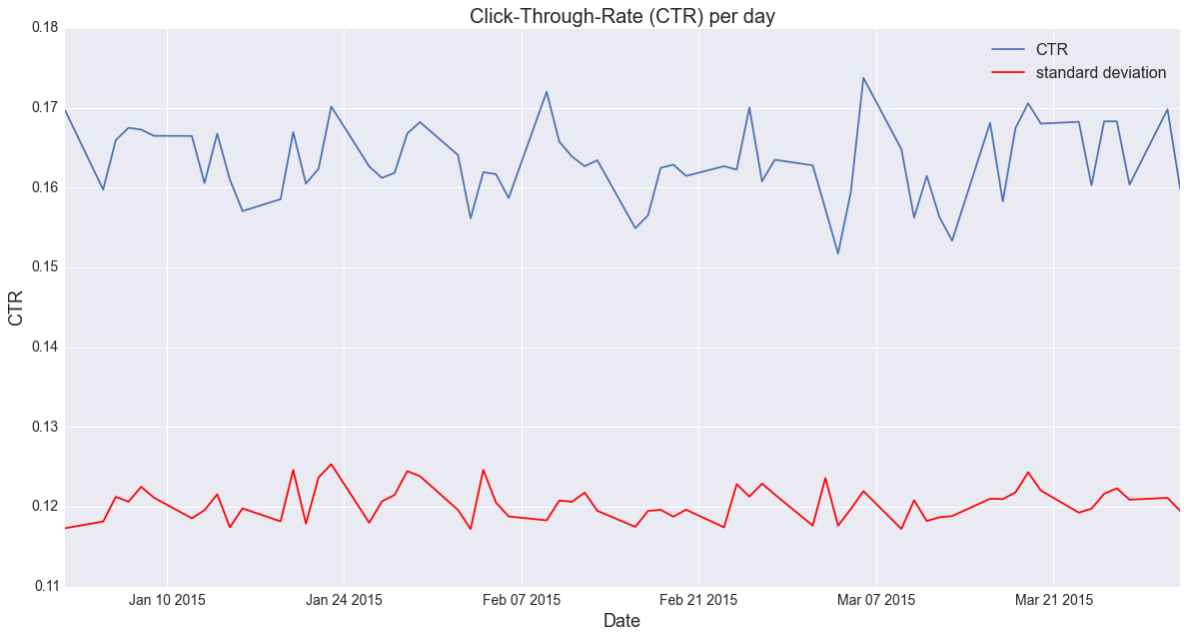
T-test, logistic regression

Three plots that best explain the data

The first plot shows average CTR per day over the 3-month period. It seems fairly consistent, fluctuating between 0.15 and 0.17, with an overall average of 0.16.

	user_id	send_date	click	email_id	CTR
0	1	2015-01-02	1	9	0.111111
1	1	2015-01-05	2	9	0.222222
2	1	2015-01-07	1	12	0.083333
3	1	2015-01-08	0	6	0.000000
4	1	2015-01-12	4	13	0.307692

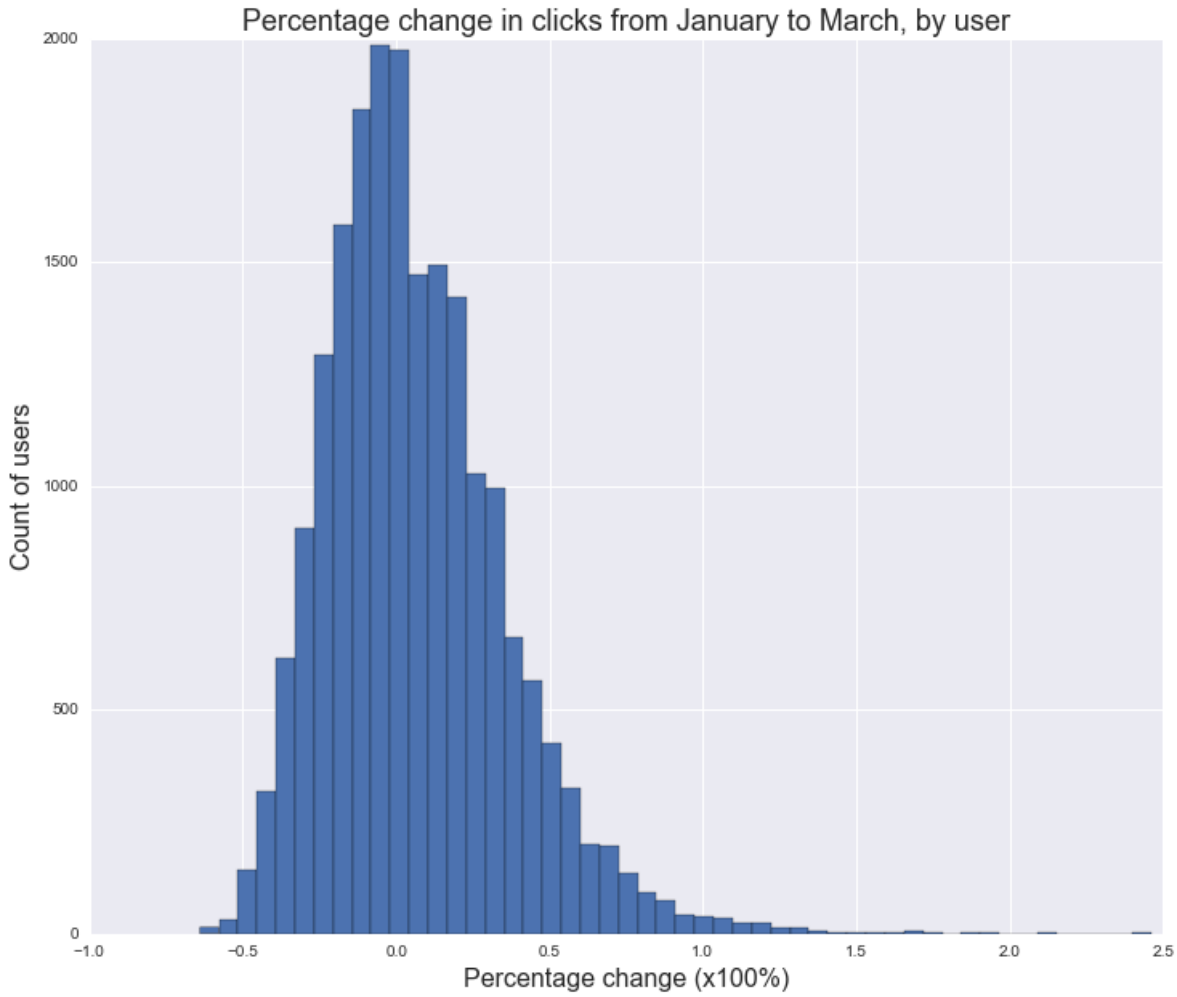
```
<matplotlib.legend.Legend at 0x12868cc50>
```




```
/Users/datascientist/anaconda/lib/python2.7/site-packages/pandas-
0.17.0+76.gdb884d9-py2.7-macosx-10.5-x86_64.egg/pandas/core/fram
e.py:1948: UserWarning: Boolean Series key will be reindexed to ma
tch DataFrame index.
  "DataFrame index.", UserWarning)
```

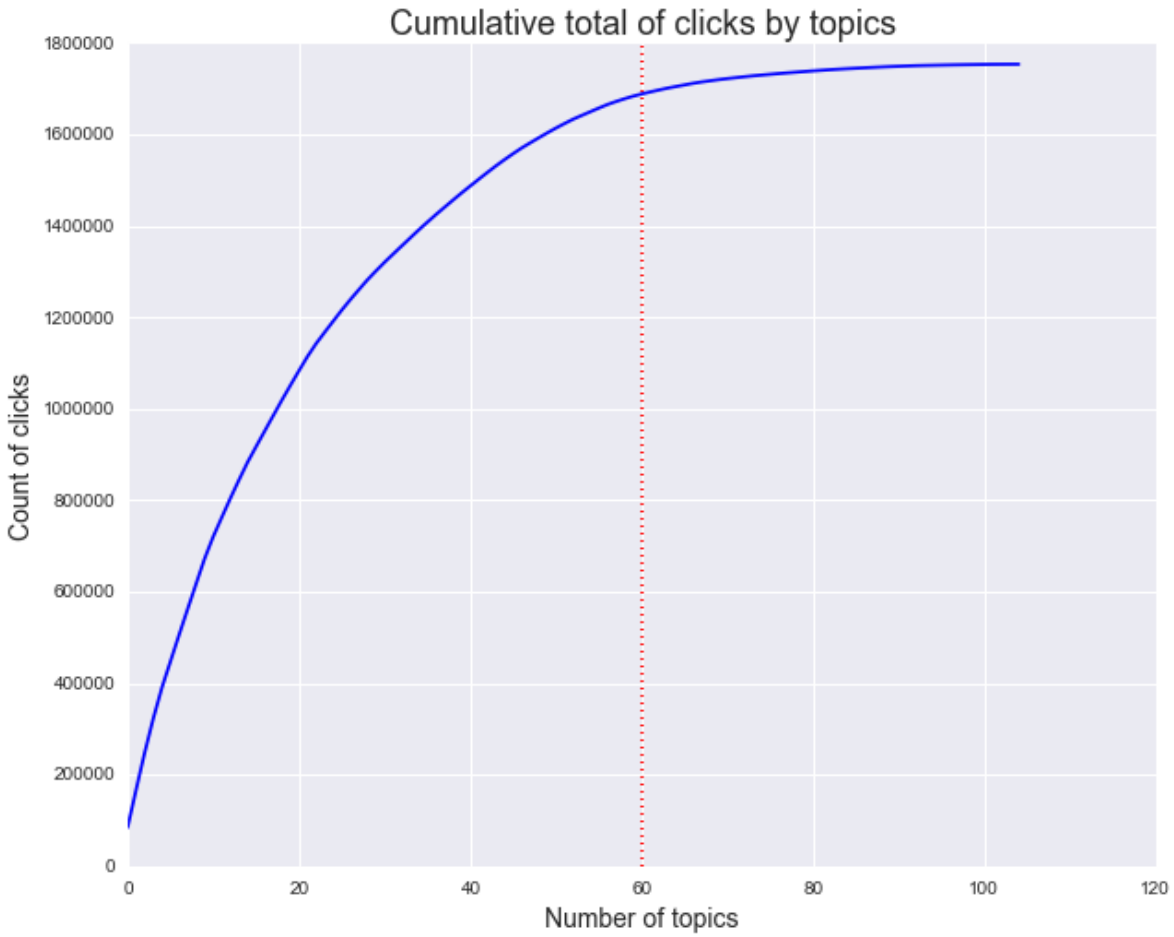
This plot represents the changes in the numbers of clicks between January and March across all users. About 51.2% of users clicked more times in March than in January. But looking at the distribution, positive percentage changes have higher magnitude (up to more than 150% increase).

```
<matplotlib.text.Text at 0x1248960d0>
```



The majority of the clicks came from articles belonging to about 60 out of the 105 available types. A sample of topics is shown in question 2.

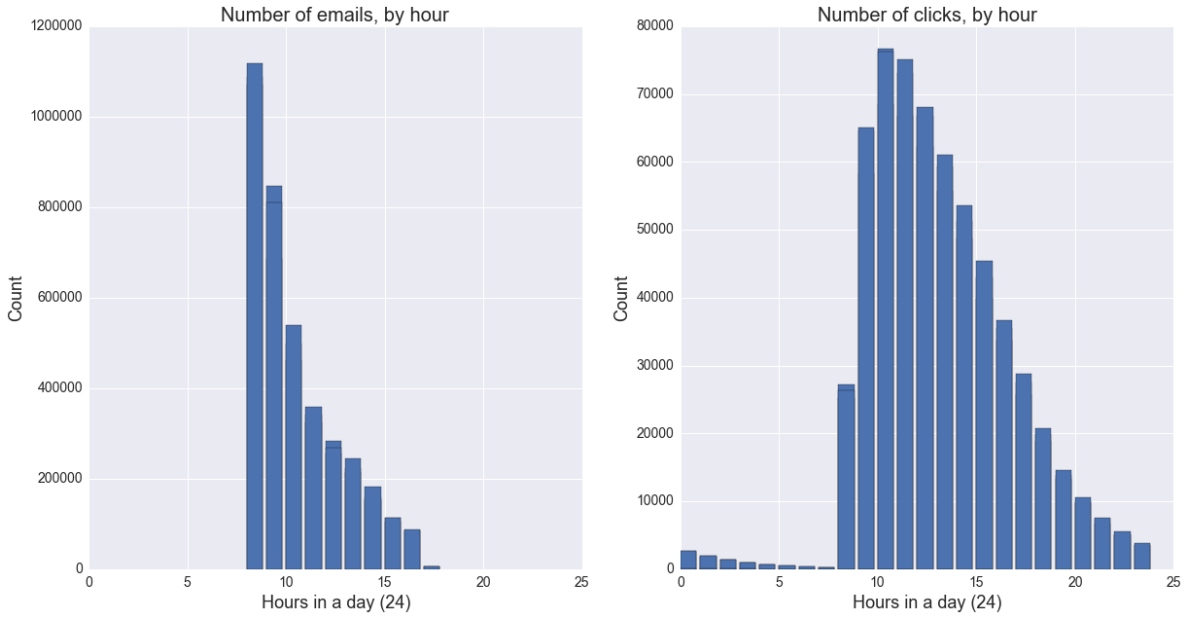
```
<matplotlib.text.Text at 0x111757790>
```



Recommendations

To generate clicks, it can focus on articles on topics that have generate more clicks. It can also extend the time period for sending emails after 5PM. People are still reading and clicking into as late as midnight. That might generate more clicks. There is also a lag between email send time and user click time. Perhaps the emails should start going out later in the day as well.

```
<matplotlib.text.Text at 0x1274d4b50>
```



Other data and questions

The standard deviation of CTR for each user each day is very high overall. I would want to look into email templates and perhaps email click data to see if people open the emails at all, and once they open, how likely they are to click on an article.