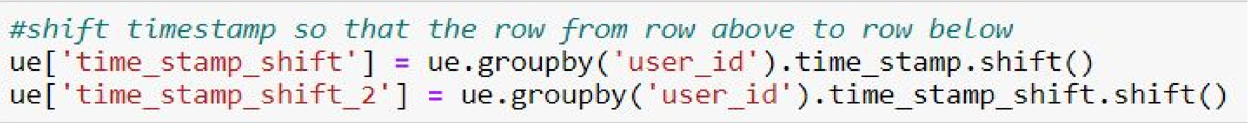
Relax Inc. Data Challenge

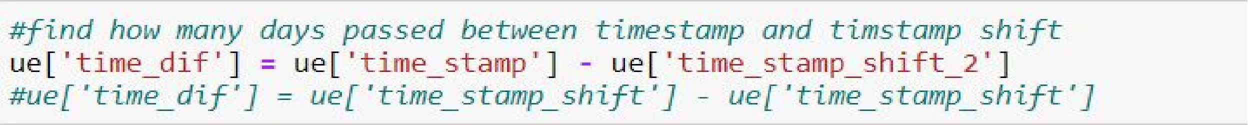
Summary of the Approach

Adopted User Definition

For the purpose of the project, we defined ‘adopted user’ as having more than 3 visits in less than 7 days. Per the code below, user engagement time_stamp is shifted twice and then we take the difference between the first timestamp and the third timestamp. This should give us the number of days between the first time someone was on the site and the third time they were (if applicable).



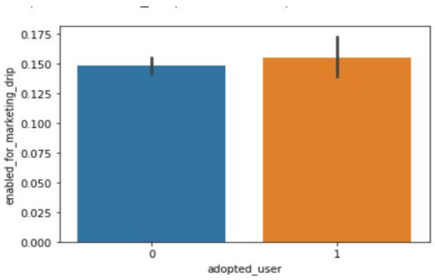
```
#shift timestamp so that the row from row above to row below
ue['time_stamp_shift'] = ue.groupby('user_id').time_stamp.shift()
ue['time_stamp_shift_2'] = ue.groupby('user_id').time_stamp_shift.shift()
```



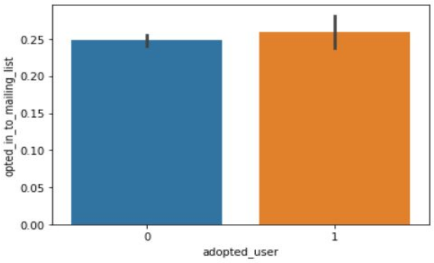
```
#find how many days passed between timestamp and timstamp shift
ue['time_dif'] = ue['time_stamp'] - ue['time_stamp_shift_2']
#ue['time_dif'] = ue['time_stamp_shift'] - ue['time_stamp_shift']
```

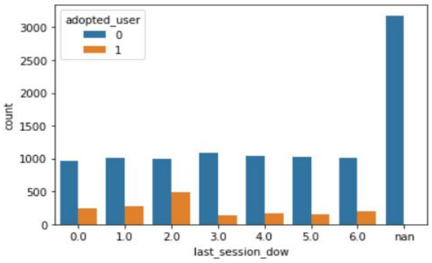
Exploratory analysis

Graph 1

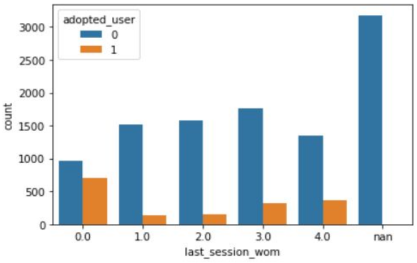


Graph 2





Graph 3



Graph 4

Based on the exploratory analysis, we see that adopted users are a bit more likely to have opted in the mailing list and enabled for marketing drip. You can see that there are more adopted users during the first week of the month relative to other months (Graph 3).There are a high number of users who don't seem to have logged in after creation. On top of that, we see that there are more adopted users during day 1 and 2 (Tuesday and Wednesday) (Graph 4).

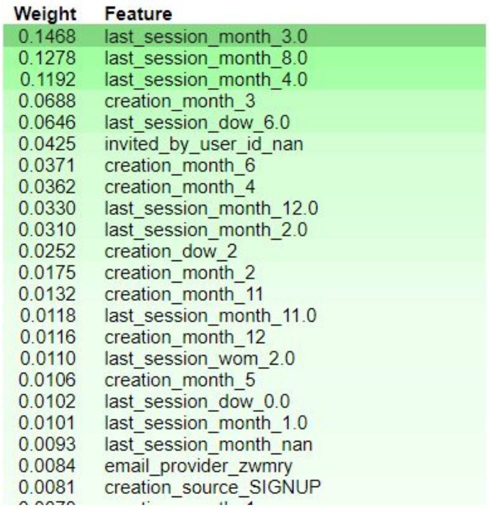
Modeling

This modeling approach was based on implementing a classifier model where 1 are adopted users and 0 as users that are not adopted users. For this analysis, we tried many models such as: Xgboost, Random Forest, AdaBoost, and Logistic Regression. Then, we also split the data into test and train and used cross validation to select the best model.

Results

The best performing model was Xgboost with 89.3% accuracy.

Features



Weight	Feature
0.1468	last_session_month_3.0
0.1278	last_session_month_8.0
0.1192	last_session_month_4.0
0.0688	creation_month_3
0.0646	last_session_dow_6.0
0.0425	invited_by_user_id_nan
0.0371	creation_month_6
0.0362	creation_month_4
0.0330	last_session_month_12.0
0.0310	last_session_month_2.0
0.0252	creation_dow_2
0.0175	creation_month_2
0.0132	creation_month_11
0.0118	last_session_month_11.0
0.0116	creation_month_12
0.0110	last_session_wom_2.0
0.0106	creation_month_5
0.0102	last_session_dow_0.0
0.0101	last_session_month_1.0
0.0093	last_session_month_nan
0.0084	email_provider_zwmry
0.0081	creation_source_SIGNUP

This demonstrates which features are associated with adopted usage.As you see, the most important variables for predicting adopted usage is whether people logged in month 3,4,8 (March, April, and August) are major predictors of adopted usage. Also if people logged in on Sunday and people who weren't invited to join the site by anyone are major positive predictors of adopted usage.

Additional research

We can do a variety of tests to test the model. For example, offers additional promotions say discount during March to promote getting users to try the software. Or, we can prompt people on Sunday to use the software and see if this improves usage.