Relax Inc. Take-Home Challenge

The **approach** taken was to first respect the time ordering of the data set and to split the data into a training set covering the first 50% of the engagement table and a test set for the remaining 50%. So the key dates are: start date: 2012-05-31, training/test cutoff: 2013-12-03, end date: 2014-06-06. By training data on users that became adopted users between the start date and the cutoff date, we can test predictions on users that joined after the cutoff date. One limitation of this approach is that the second period is shorter and users joining after the cutoff date have less time to become adapted users.

Testing for **adapted users** involved converting login times for each user into a list and testing for 3 consecutive logins that fulfill the conditions: second 12-36 hours after the first, third 12-36 hours after the second, and third under 8 days from the first. The label is defined as a True/False value, True applying to users that fulfill the condition at least once in their designated period. This was merged onto the training and test data separately on the user_id field, creating two separate data sets with no overlapping users.

Features used in the data are the following:

- Creation time, measured as the length in days from registration until the cutoff date (for training) and from registration until the end date (for test).
- Creation source as given in the source data, with each category converted to an integer
- Opted to mailing list, enabled for marketing drip, organization ID and invited by user ID are all kept as in the source data
- Second login indicates the lapsed time in days between registration and the user's second login; users which did not log in for a second time are given a high numeric value

Results of the random forest classifier and MLP classifier on the data are shown below. As the proportion of adapted users is low (14% of the training set and 3% of the test set), the precision and recall are the important scores for determining the predictive power of a model.

Classifier metrics

Random forest

```
Trining score:  1.0
Test score:  0.968593168205
Precision:  0.41095890411
Recall:  0.165745856354
```

MLP

```
Trining score:  0.868796153186
Test score:  0.958394042415
Precision:  0.291208791209
Recall:  0.292817679558
```

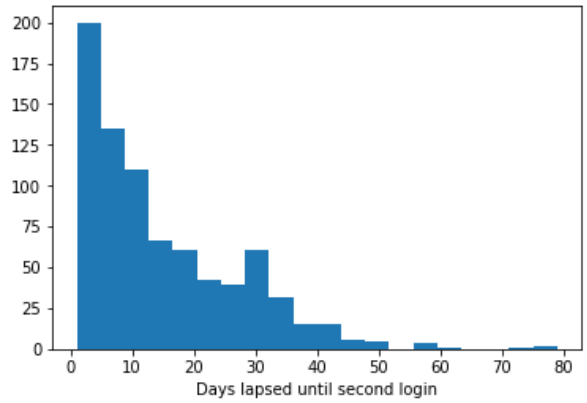
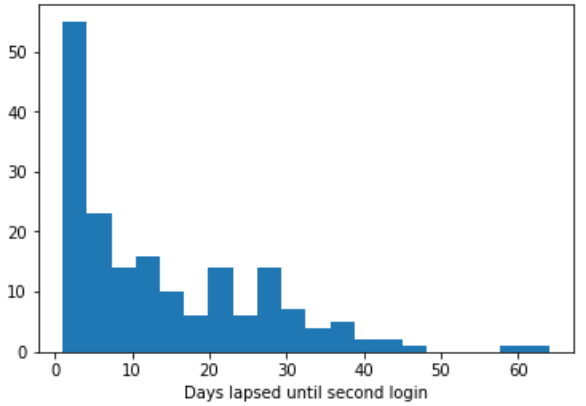
Feature importance on random forest classifier

	feature	feature_importance
0	creation_time	0.213425
1	creation_source	0.024867
2	opted_in_to_mailing_list	0.012504
3	enabled_for_marketing_drip	0.009850
4	org_id	0.150519
5	invited_by_user_id	0.086745
6	second_login	0.502090

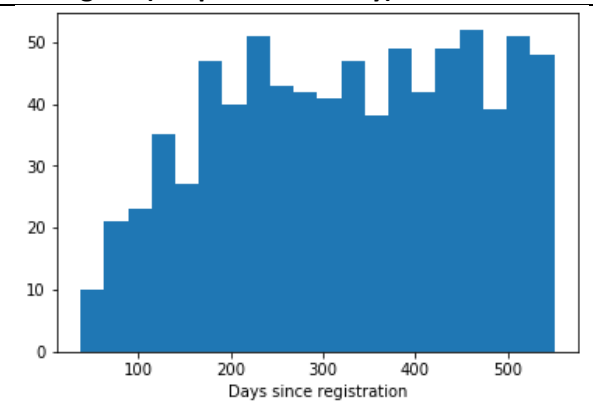
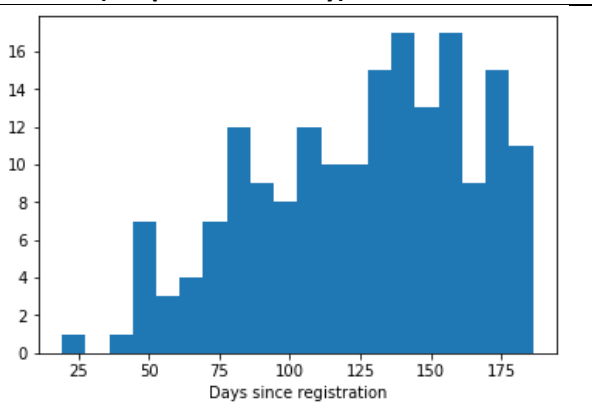
Feature importance on the random forest classifier indicate the second login time is a key predictor of future adoption. For a user to become an adapted user, a second login is a must and provides a clear divider between the two classes. Lapsed time since registration is also an important feature, in that the probability of a user meeting conditions increases with more time given. The third important feature in

organization ID, followed by the ID of the user who invited them. However these two features may become less relevant over time as new users cannot keep coming from the same organizations and the same invitee users. Creation source, email list and marketing drip have very little effect, hence any prediction not involving login frequency is likely to be very inaccurate.

The **histogram of lapsed time before second login** for adapted users (below) clearly shows that adapted users most commonly make a second login within the first few days of registration, and this is the strongest predictor. However more users become adapted over time and the model must pick up those as well.

Training set (adapted users only)	**Test set (adapted users only)**
	

The histogram of lapsed time since registration for adapted users also shows the build-up of adapted users over time, with approx. the first 250 days being critical to determine of a user will become adapted.

Training set (adapted users only)	**Test set (adapted users only)**
	

Potential improvements:

- Equalizing the time span of the the training and test sets may result in better predictions
- Extracting more features from the logins, e.g. no. of logins in the first week/month