Relax Inc Report

1. Initial EDA & Cleaning

The users dataset contained 12,000 entries and 9 features. The 12,000 entries corresponded to each user.

The engagement dataset consisted of 207,912 rows and a time stamps of each visit for each user. The data was sorted was by user_id. Each entry corresponded to a successful login by the user.

Null values were found in the following features:

a) last_session_creation_time, null values here were filled by the creation time from the users data.
b) Invited_by_user_id: Null values here were filled with 0's.

2. Finding out the adopted users

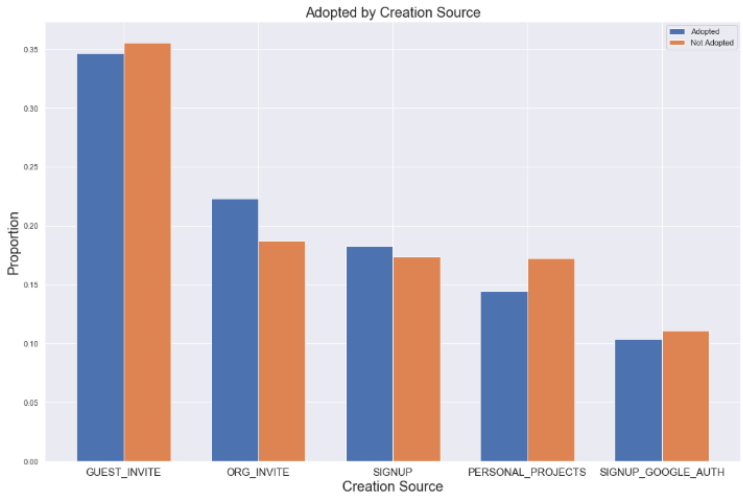
To mark the adopted users, the following steps were taken:

1. Assume that once a user is adopted, they are adopted forever.
2. Set up an initial start date variable: 31-05-12 (The first day in the dataset) and setting an end date variable 7 days: 06-05-12.
3. Sum the visits (Logins) in the 7 day period in step 2.
4. If the sum is greater than 3, the 'adopted' flag will be set to '1' and the user_id will be added to a list of adopted users.
5. Once a user is adopted, they will not be looped though anymore.
6. Add 1 day to the start date in step 2 and repeat steps 3-4.
7. The loop will run until 6th June 2014 (The last day of the dataset).

To save on computation time, users that had **total** visits of less than 3 were not looped through and automatically marked as 0.

The results showed that 1,656 users were adopted out of 12,000 users. This is 13.8%.

3. EDA



EDA has shown that being invited from an organisation had a positive effect on whether a user was adopted or not.

Opting in mailing list and enabled for m marketing drip had no effect when visualized on whether a user was adopted or not.

4. Preprocessing

The following steps were taken to preprocess the data for modeling:

1. Drop the name, email and adopted column out of the training set.
2. Transform the datetime columns to Unix timestamps so they may be used in modelling.
3. Using pd.get_dummies(), we transformed categorical features to numerical features.,
4. Divided he data into train/test splits using a 33% test size.

5. Modeling

The following models were used to find the differences in feature importance.

1) XGB Classifier
2) Random Forest Classifier

The models were simple models run with no cross validation or parameter tuning.

6. Results and further steps

XGB showed that creation_time was the most significant in predicting adoption. The second most important feature was organisation id.

RFC showed that creation_time & last_session_creation_time were the most significant in predicting adoption. The second most important features were organisation id and user_id.

The results of the analysis proved that organisation id and user id made difference in predicting whether a user will be adopted or not.

Meaning that:

1. If a person was invited within his organisation, he was more likely to be adopted.
2. If a person was invited by specific user ids, they were more likely to be adopted.
3. Even though creation time and last_session_creation_time were significant but they do not provide any value to business analysis.

Some further steps to perform would be to:

1. Analyze which organisations and which users led to have more adopted users.
2. Perform some statistical analysis and confirm that creation source by organisation ID was significant in having a positive effect on adoption.
3. Add some cross validation and parameter tuning to model selection.
4. Once organisation and user id's significance is confirmed, potential targeting of these features to enhance the business model can be discussed