Relax.Inc Data Science Challenge

Project Objective: Find out the important factors for predicting user adoption

Data: User data which includes general information about users such as account creation date and account creation source as well as data on user activities which include the count of user logins over time

Findings from Exploratory Data Analysis:

- Majority of users did not adopt the app. User adoption rate is around 12%.
- Majority of users did not opt into the mailing list or enable marketing drip. The data shows that users who did not opt in the mailing list or enable the marketing drip are less likely to be adopted later. Marketing campaigns could be an important factor for user adoption.
- Users who created their account in March, April and May are less likely to be adopted users.
- Many users created an account through "organization invite".
- A lot more users created accounts in 2013 than in 2014 or 2012.
- Around half of the users did not get referred by anyone when signing up.
- The user engagement with the software is low. Majority of users only have up to 1 average weekly session over time. Majority of users have up to 2 total sessions for the first 30 days. Users should be expected to use the software at work more often if they have fully adopted the software. The low engagement for 30 days indicates that there's a need to motivate the users to log in more for the first 30-day trial so that they can possibly be retained later.
- Even though there are a lot more users who created account in 2013, the proportion of the user who are not adopted is also high.

Methodology & Findings for Predictive Modeling:

<u>Methodology:</u>

- To make predictions on whether users are going to be adopted or not, I have created several classification models.
- Algorithms used: Random Forest, SVM
- Imbalanced class issue: Since the model has imbalanced class issue (majority of users are not adopted), I have used couple of techniques:
    - Oversample technique to add more copies in the minority group in the training data.
    - Cost-sensitive learning on SVM which penalizes misclassifications of minority class.

<u>Findings:</u>

- I have found the total user sessions for the first 30 days, the average weekly sessions for the first 30 days are the most importance features. This has validated that users' activities for the first 30 days can be a good indicator on the user adoption
- The timing/seasonality of the account creation has become the most important negative factors. For example, people who signed up in 2014 are less likely to be adopted. People who signed up in April and May are less likely to be adopted.
- Organization invite has become a negative factor which means people who get invited by their organization are less likely to be adopted.

See details here: https://github.com/yukaabe/Data-Science-Projects-Portfolio-Repo/tree/master/Relax%20User%20Adoption%20Data%20Science%20Challenge