Relax challenge

The adopted user is 1445 (representing 12, 042% of the all the users 12000 in takehome_users dataset).

Data preprocessing involved replacing missing values, converted invited_by_user feature into binary value(1,0), created new features from creation_time (day, week and day of week). We label encoded categorical features. There was no duplicated value.

Identify which factors predict future user adoption

We used a tree base algorithm to build the predictive model (because are not sensitive to noisy data or outliers and usually provides good performance). we performed the cross validation and split the data into train and test set (75/25). The dataset is over unbalanced (12% over 87%), we considered F1-score as the metric for model evaluation and selection. We also look at the value of area under the curve.

After training and evaluating each model, we found that:

- random forest: F1_score is 0.54, AUC is 0.869
- tune random forest: F1_score is 0.58, AUC is 0.892
- XGBoost : F1_score is 0.55 , AUC is 0.878

The model that best predict rider retention is the tune random forest. The F1_score and the area under the curve is good as compared to other.

The features that influence user adoption are last_session_creation_time, organisation the user belong to, the creation_sources on Personal projects and opted_in_to_mailing_list.