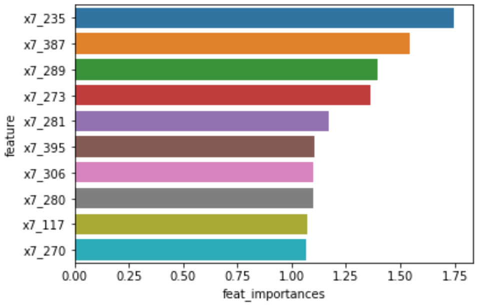
Brief Findings of Relax Challenge

From our analysis, Organization (or group) they belong to is the most fundamental factor to predict the adopted users. Organization 235, 387, 289, 273 and 281 are the top 5 organizations which have more impact the prediction.



Our approach is explained in detail as follows:

Data Wrangling:

- Labelled the target. We grouped and counted the visit time of user in the user engagement summary dataset based on the condition of three separate days in at least a week. Users whose visits are equal or larger than three are labelled as 1, otherwise they would be labelled as 0.
- Merged Data. The targeted dataset is merged with user dataset.

Data Cleaning and Visualization:

- Drop irrelevant columns. Some irrelevant columns such as User Names, UnixTimestamp are dropped.
- Fill missing values. Missing values of invited_by_user_id are filled by mean value
- The dataset has imbalance. The proportion of adopted user is 16.9%. Org_id has 412 category variables.

Feature Engineering, Model Building and Performance Evaluation:

- Numerical inputs are scaled and categorical inputs are converted to number by onehot code
- The dataset is split into train and test dataset. We use the Logistic Regression for modelling and feed the train dataset to the model
- We have high precision and low f1 score. Apparently, the model can recognize the True Negative better than True Positive. Due to imbalance, the model does not work well.

Future work might add more data to balance org_id and integrate more features to improve the prediction accuracy.