Identify factors predicting future user adoption

Summary: The active duration, which is the time interval in days between a user’s last login time and the time of creating an account, plays the dominant role in predicting an “adopted user”.

Solution process:

1. Generate the target variable (adopted_user) from the usage summary table. In total, 2235 adopted users are labeled, taking a proportion of about 18.6% of all 120,000 users.

2. Conduct exploratory data analysis between variables and the target to check the relation between a single variable or category and the adapted_user. For instance, the numerical variable active_duration in days suggests an adopted user when it is larger than 80. Other categorical variables including email subscribing and invited by a user do not have an obvious correlation with the target.

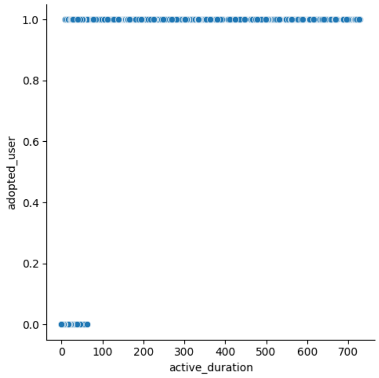

Figure 1 Active_duration versus adopted_user.

3. Build a machine learning model to predict user adoption. To capture more relevant items, recall is used as a metric. Among a series of tested algorithms, the Adaptive Boosting (AdaBoost) Classifier has the highest recall for validation data and is selected as the final model. Moreover, the feature importance analysis and recursive feature selection suggest the active duration is the dominant role in the model.

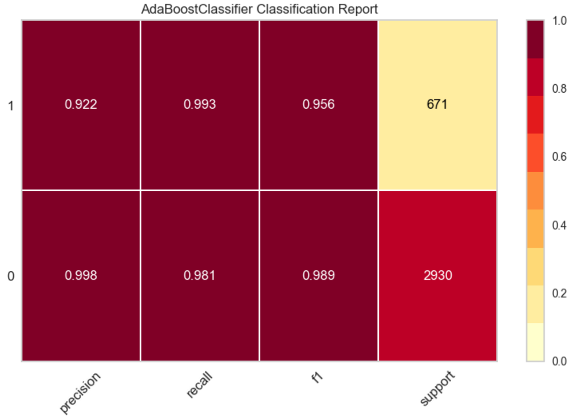

Figure 2 Classification report of AdaBoost Classifier. The cross-validation recall of identifying adopted user is 0.993.

Future work includes deploying the model.