Relax Take Home Challenge Solution

Dataset and EDA

Dataset consists of engagement records, that show login times of users and users file that represent user characteristics, including last login time and creation time. The preliminary data analysis showed that data sets had some missing values but no outliers. Some records were missing probably because some registered users never logged on to the system. Also there are repeater values for both the user name and e-mail address values. Further exploration revealed that same names were a result of similarities rather than repetitions. However, this was not the case for e-mail addresses, since they must be unique. For the sake of the model, repeater (20x2) email addresses were deleted.

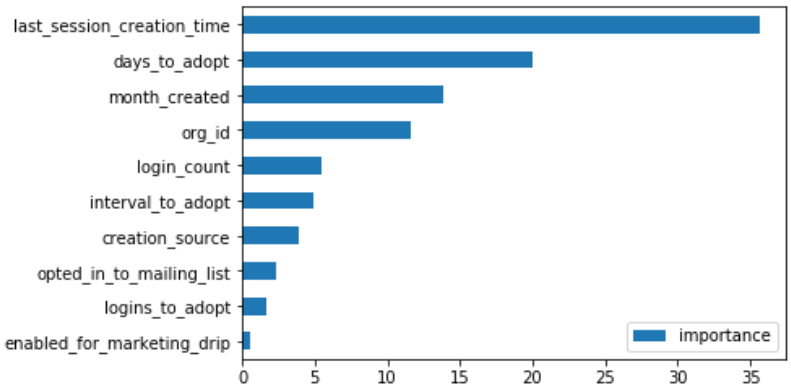
Feature Engineering, DF Manipulation

Initially only feature engineering carried out was to generate a label column for identifying adopt users through login records. All time representing strings were converted to time stamps. There were missing values for some user ids. Since they never logged in, those values were filled with 0 for the label column. Preliminary ML models using columns provided by the given databases only did not produce satisfactory results, since the computer was not able to learn with so few features. Therefore more feature engineering was necessary. Using the login records and creation date value, additional features such as 'login count, days till adoption, login time intervals till adopt and month created' were added. Values were mostly calculated until adoption date (i.e. login counts until adoption) for better modeling. Again missing values were present due to lack of user login, therefore those values were filled with zeros. There was a class imbalance in the label column which was balanced by incorporating class weight.

Machine Learning Models

Train-test split, grid search for hyperparameter tuning, cross validation were used for all models. The incorporated Catboost model has own algorithm for categorical variables, however, all other models needed dummy variables for categorical columns, which were done. A correlation heatmap revealed the pearson correlations for customer adoption, which is an indicator for factors affecting customer adoption. Since accuracy is not adequate enough to describe model's skill, ROC_AUC score, Precision-Recall-auc score, confusion matrix were utilized to measure effectiveness. Randomforest, logistic regression and catboost classifiers were very accurate.

Feature Importances



Many models are able to describe feature importances, which show outlay of dominant features defining the label set. In this project Catboost and Random forest feature importances were used. Important features were last session creation time, days until adoption, month created and logins until adoption.

Please see Jupyter notebook for further details.