Predict result of recommendation for Megafon

Final Project

Mishin Dmitrii 17.07.21

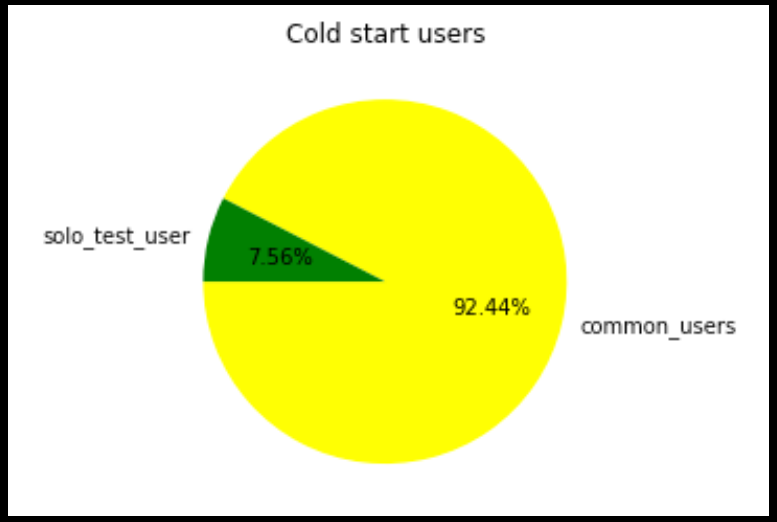
Define problem

- I got a data sets with 860052 clients of Megafon Co. and theirs transaction or behavior history or client infos. Based on this data I need to create a model for predict a result of Megafon’s recommendation for other group in test data set.

Exploratory data analysis

Users

- 7.5% users in test data are different with users from train data.
- Id column have low and high values, so it would have affect for model's good performance.
- I'm going to make model without using id column in.



Exploratory data analysis

Client purchase history

- 92% of Clients don’t buy cervices after recommendation.
- Hight disbalanced target.
- Model needs to handle with a lot of noise in data for do high performance.

Exploratory data analysis

Ures's services

- Most of users were recommended one service.
- Few users were recommended 2-3 services.
- buy_time in train data set and features data set is different. I don't know why and I don't want to fight with this, so I decided to drop buy_time after merging train and features sets.
- Megafon recommends 8 services. 65% of all recommendation are service 1 and service 2.
- Service 4 and service 6 are most connect by user after recommendation.

Model Planning

Features

- Totally 252 Features (Too much for my MacBook Pro M1).
- Use SelectKBest with f_classif to leave only 50 features for the best performance.
- Categorical features were left : 'vas_id', '132', ‘195’.
- Constant features were left : '132', ‘195’.
- Other features seems to be discrete quantity: '0', '1', '3', '5', '6', '7', '20', '22', '24', '40', '41', '44', '61', '70', '110', '114', '116', '117', '118', '119', '126', '127', ‘136', '137', '138', '140', '141', '142', '146', '147', '148', '167', ‘168', '169', '190', '209', '210', '212', '226', '227', '228', '237', ‘238', '240', '242', '243', '247'

Model

- I don’t use Pipeline because it always run out of my MacBook memory. In reason of i do not use what i can’t test, I didn’t use it in my work.
- I safe a python file which would downloads pickle model, transforms test df, makes prediction and safes it to answers_test.csv.
- Model and Parameters: CatBoostClassifier(iterations=1000, depth=10, l2_leaf_reg=5, loss_function='Logloss', learning_rate=0.1, metric_period=100, cat_features=cat_feat)

Scoring

- f1_score(predict, y_pred, average=‘macro') = 0.7362833591500052 .
- Highest f1 score is the result of predict_proba with threshold for 0.4 .
- Run get_predict.py for predicting. (See Addition)
- Thank you for your attention. For more information see my jupyter notebook.

Addition:

- get_predict.py require data_test.csv, model.pkl, features.csv be in the same folder with it.
- get_predict.py create answers_test.csv is a result of predict.