Презентация

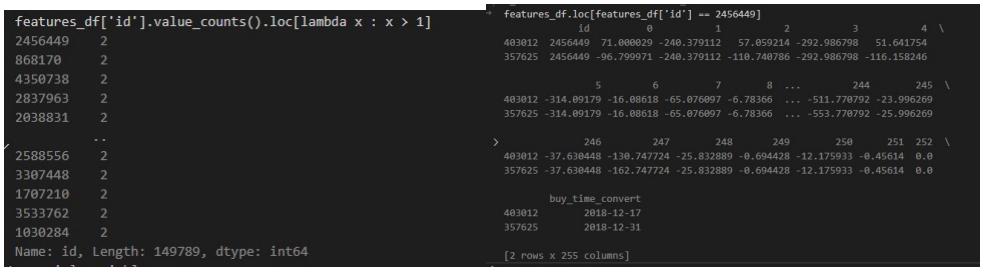
Замечания по проекту

Выборка фичей для пользователя производится по времени buy_time т.е. не берутся данные пользователя (features) старше времени подключение услуги (target) . Т.е features.csv.buy_time должна быть меньше или одного времени data_test.csv.buy_time

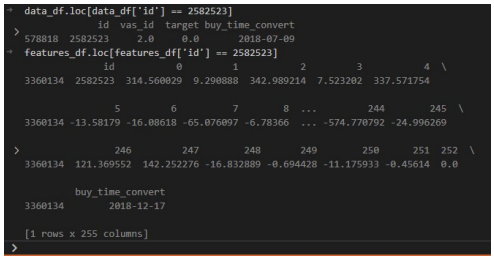
В полученом датасете для некоторых пользователей обучающей выборки отсутствуют фичи

```
missing_features.shape[0]
422929
source_df
source_df.head()
source_df.shape
             id  vas_id  target buy_time_convert          0          1  \
0       2582523     2.0     0.0       2018-07-09        NaN        NaN
1       1292549     2.0     0.0       2018-07-09        NaN        NaN
2       4053116     1.0     0.0       2018-07-09        NaN        NaN
3       4158361     2.0     0.0       2018-07-09        NaN        NaN
4       3754468     4.0     0.0       2018-07-09        NaN        NaN
...         ...     ...     ...              ...        ...        ...
831648   555080     5.0     0.0       2018-12-31 -96.799971 -408.179112
831649  1729471     5.0     0.0       2018-12-31 -86.209971 -397.589112
831650  3676177     2.0     0.0       2018-12-31 -96.799971 -408.179112
831651  2255038     2.0     0.0       2018-12-31 -96.799971   49.450888
831652  3022610     2.0     0.0       2018-12-31 -49.339971 -218.339112

                2           3           4          5  ...         243  \
0             NaN         NaN         NaN        NaN  ...         NaN
1             NaN         NaN         NaN        NaN  ...         NaN
2             NaN         NaN         NaN        NaN  ...         NaN
3             NaN         NaN         NaN        NaN  ...         NaN
4             NaN         NaN         NaN        NaN  ...         NaN
...           ...         ...         ...        ...  ...         ...
831648 -110.740786 -460.786798 -116.158246 -481.89179  ... -977.373846
831649 -100.150786 -450.196798 -105.568246 -471.30179  ... -977.373846
831650 -110.740786 -460.786798 -116.158246 -481.89179  ... -977.373846
831651 -104.390786  219.293202 -109.808246  200.30821  ... -977.373846
831652   73.139214  145.093202   67.721754  123.98821  ... -977.373846

              244        245        246        247        248        249  \
0             NaN        NaN        NaN        NaN        NaN        NaN
1             NaN        NaN        NaN        NaN        NaN        NaN
2             NaN        NaN        NaN        NaN        NaN        NaN
3             NaN        NaN        NaN        NaN        NaN        NaN
4             NaN        NaN        NaN        NaN        NaN        NaN
...           ...        ...        ...        ...        ...        ...
831648 -613.770792 -25.996269 -37.630448 -306.747724 -25.832889 -0.694428
831649 -613.770792 -25.996269 -37.630448 -306.747724 -25.832889 -0.694428
831650 -613.770792 -25.996269 -37.630448 -306.747724 -25.832889 -0.694428
831651 -613.770792 -25.996269 -37.630448 -306.747724 -25.832889 -0.694428
831652 -613.770792 -25.996269 -37.630448 -303.747724 -25.832889 -0.694428

             250      251  252
0            NaN      NaN  NaN
1            NaN      NaN  NaN
2            NaN      NaN  NaN
3            NaN      NaN  NaN
4            NaN      NaN  NaN
...          ...      ...  ...
831648 -12.175933 -0.45614  0.0
831649 -12.175933 -0.45614  0.0
831650 -12.175933 -0.45614  0.0
831651 -12.175933 -0.45614  0.0
831652 -12.175933 -0.45614  0.0
```

В датасете features есть дублируюущиеся user id. Думаю что это один и тот же пользователь но с обновленной информацией о себе.



В обычающей выборке были данные пользователя фичи которого будут описаны в будущем(после подключения услуги). Такие данные тоже будут отброшены.



Формирование DataFrame происходит через класс DataManager. В нем есть возможность как загрузить данные из csv так и загрузить ранее сформированный DataFrame.

Подбор параметров на `RandomizedSearchCV` был долгим и рассчитаны один раз, по необходимости

Некоторые библиотеки я переиспользовал из своих прошлых наработок, так что код может быть кое где избыточен.

Информация о модели, ее параметрах, особенностях и основных результатах.

Модель выбрал `RandomForestClassifier` с параметрами:

n_estimators = 900,

min_samples_split = 28,

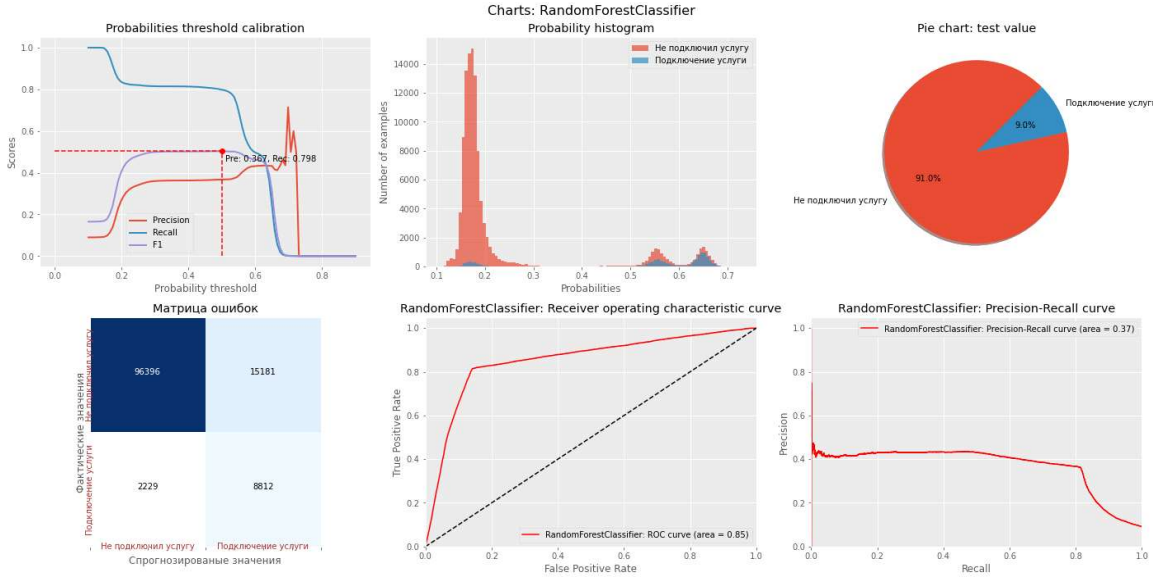
min_samples_leaf = 2,

max_features = 'sqrt',

max_depth = 14,

bootstrap = False,

Основные характеристики модели следующие



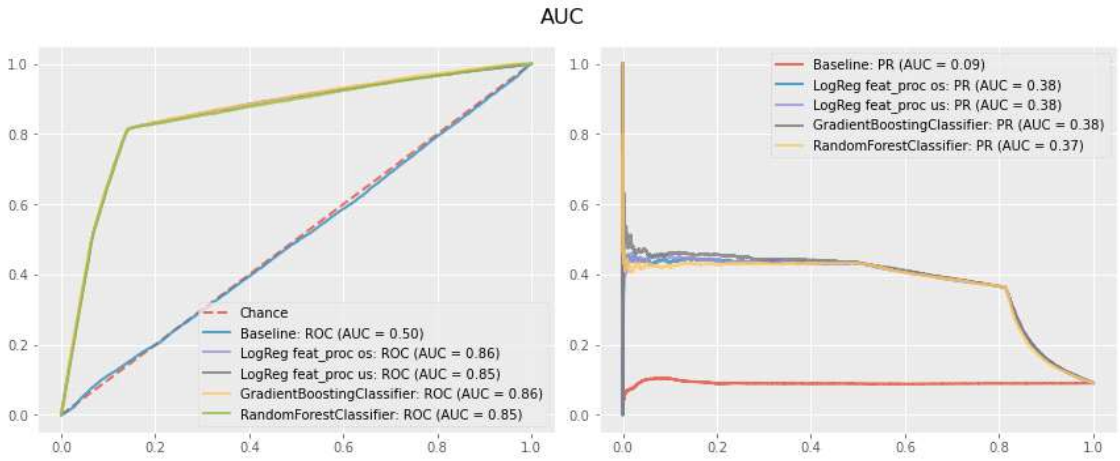
Немного странная ROC кривая, на небольших данных, с выборкой в 6000 записей, графики были более привычны и предсказуемы. Думаю проблемы будут в данных, возможно существуют выбросы или аномалии которые следует обработать.

Обоснование выбора модели и ее сравнение с альтернативами.

Модель выбрал `RandomForestClassifier` т.к. удалось достичь наибольшего параметра f1-score (macro) Treshhold выбрал 0.5 т.к. на этом уровне есть граница баланса(f1) Precision и Recall.

Name model	Threshold	F-Score	Precision	Recall	Roc-AUC	f1-score(macro)
Baseline	0.5	0.155	0.088	0.633	0.496	0.331
LogReg feat_proc os	0.5	0.501	0.362	0.814	0.856	0.708
LogReg feat_proc us	0.5	0.501	0.362	0.813	0.855	0.708
GradientBoostingClassifier	0.5	0.502	0.363	0.812	0.859	0.709
RandomForestClassifier	0.5	0.503	0.368	0.798	0.852	0.71

В принципе можно взять и логическую регрессию если критична скорость обучения, значения не очень сильно отличаются



Интересный график Precision-Recall curve. Максимизируя полноту (Recall) после ~0.82 начинаем получать сильное падение точности (Precision)

Принцип составления индивидуальных предложений для выбранных абонентов.

Этот вопрос я не совсем понял. Если необходимо сделать предсказание vas_id, то его я не делал(не успел). Думаю для решения этой задачи обратиться к технологии рекомендательных систем и предсказывать подключаемую услугу по похожести пользователей (user-user).