

# Презентация

## Замечания по проекту

Выборка фичей для пользователя производится по времени buy\_time т.е. не берутся данные пользователя (features) старше времени подключения услуги (target) . Т.е. features.csv.buy\_time должна быть меньше или одного времени data\_test.csv.buy\_time

В полученном 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	
...	...	...	...	...	...	...	...
0	NaN	NaN	NaN	NaN	NaN	NaN	243 \
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	-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	-109.888246	200.30821	...	-977.373846	
831652	73.139214	145.093202	67.721754	123.98821	...	-977.373846	
...	...	...	...	...	...	...	...
0	NaN	NaN	NaN	NaN	NaN	NaN	248 \
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	
...	...	...	...	...	...	...	...
0	NaN	NaN	NaN	NaN	NaN	NaN	252 \
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	-12.175933	-0.45614	0.0	0.0	0.0	0.0	
831649	-12.175933	-0.45614	0.0	0.0	0.0	0.0	
831650	-12.175933	-0.45614	0.0	0.0	0.0	0.0	
831651	-12.175933	-0.45614	0.0	0.0	0.0	0.0	
831652	-12.175933	-0.45614	0.0	0.0	0.0	0.0	

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

```

features_df['id'].value_counts().loc[lamba x : x > 1]
2456449 2
868170 2
4350738 2
2837963 2
2038831 2
2588556 2
3307448 2
1707210 2
3533762 2
1030284 2
Name: id, Length: 149789, dtype: int64

features_df.loc[features_df['id'] == 2456449]
   id  0  1  2  3  4 \
403012 2456449 71.000029 -240.379112 57.059214 -292.986798 51.641754
357625 2456449 -96.799971 -240.379112 -110.740786 -292.986798 -116.158246
      5  6  7  8 ... 244 245 \
403012 -314.09179 -16.08618 -65.076097 -6.78366 ... -511.770792 -23.996269
357625 -314.09179 -16.08618 -65.076097 -6.78366 ... -553.770792 -25.996269
      246 247 248 249 250 251 252 \
403012 -37.630448 -130.747724 -25.832889 -0.694428 -12.175933 -0.45614 0.0
357625 -37.630448 -162.747724 -25.832889 -0.694428 -12.175933 -0.45614 0.0
      buy_time_convert
403012 2018-12-17
357625 2018-12-31
[2 rows x 255 columns]

```

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

```

data_df.loc[data_df['id'] == 2582523]
   id  vas_id  target  buy_time_convert
> 570818 2582523 2.0 0.0 2018-07-09
features_df.loc[features_df['id'] == 2582523]
   id  0  1  2  3  4 \
3360134 2582523 314.560029 9.290888 342.989214 7.523202 337.571754
      5  6  7  8 ... 244 245 \
3360134 -13.58179 -16.08618 -65.076097 -6.78366 ... -574.770792 -24.996269
      246 247 248 249 250 251 252 \
3360134 121.369552 142.252270 -16.832889 -0.694428 -11.175933 -0.45614 0.0
      buy_time_convert
3360134 2018-12-17
[1 rows x 255 columns]
>

```

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,`

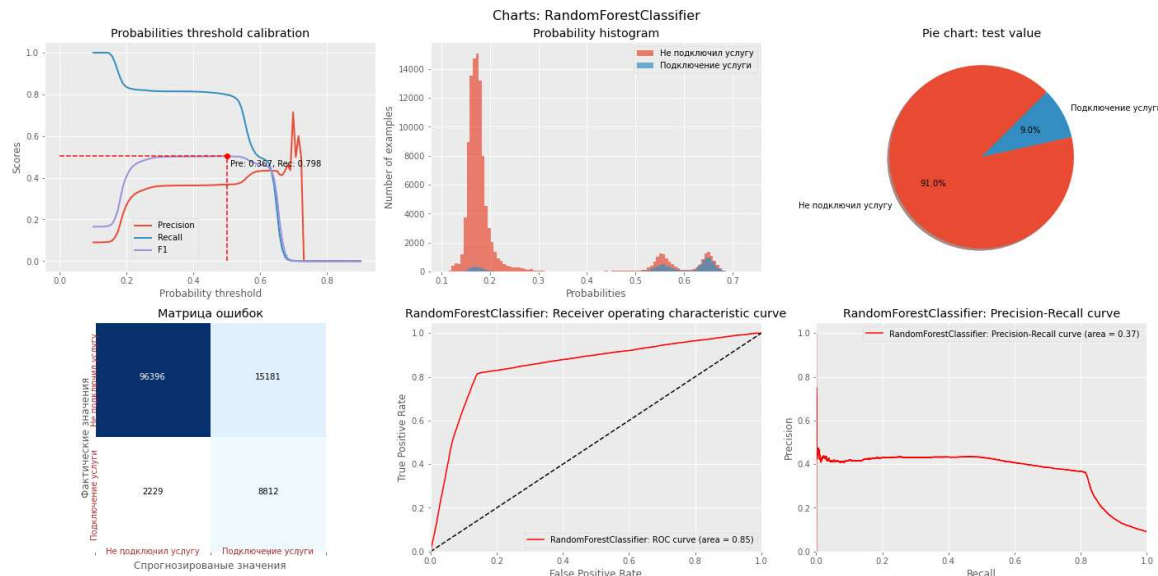
`min_samples_leaf = 2,`

`max_features = 'sqrt',`

`max_depth = 14,`

`bootstrap = False,`

Основные характеристики модели следующие



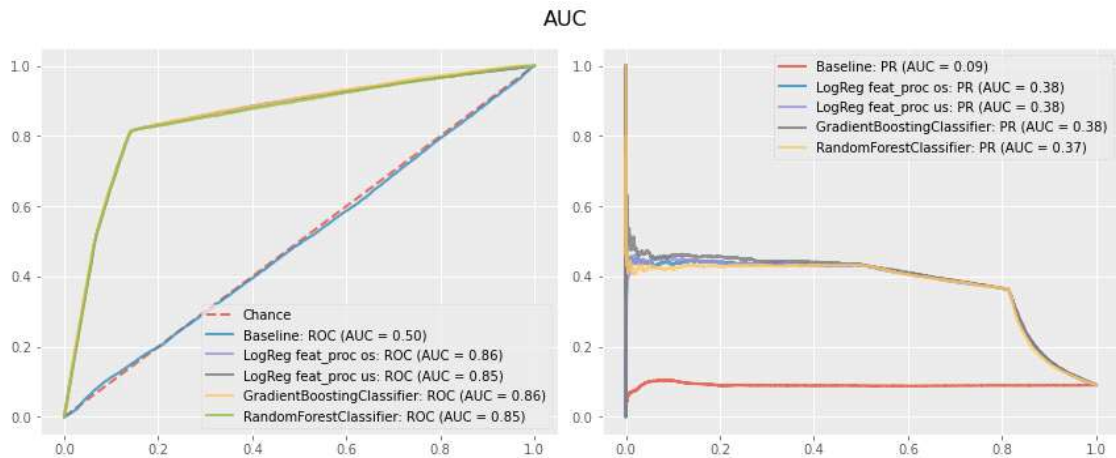
Немного странная ROC кривая, на небольших данных, с выборкой в 6000 записей, графики были более привычны и предсказуемы. Думаю проблемы будут в данных, возможно существуют выбросы или аномалии которые следует обработать.

## 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).