Проект от МегаФон

«Предсказание вероятности подключения услуги»

Выполнила – Горохова Анастасия

студентка факультета «Искусственный интеллект», специальность – Data Science

г. Санкт-Петербург 2023г.

Задача

Построить алгоритм, который для каждой пары пользователь-услуга определит вероятность подключения услуги. Метрика качества осуществляется функцией f1 невзвешенным образом.

Особенности

Значительный размер занимаемой памяти признаками пользователей. Размер файла в распакованном виде более 20Гб.

Подготовка данных

В связи с тем, что признаки пользователей занимают значительный размер памяти, подготовка данных осуществлялась с использованием промежуточных файлов.

На первом этапе соединялись предложения услуги по полю id с записями признаков. Если на пользователя приходилось несколько записей признаков, то формировалось декартово произведение записей, предложений с записями признаков по данным пользователям. Для формирования соединения использовалась библиотека Dask.

На втором этапе обрабатывались дублирующийся записи, получившиеся на первом этапе. Были оставлены записи, в которых признаки пользователей действовали на момент предложения услуги.

Выбор модели

Свой выбор я остановила на модели CatBoostClassifier. У данной модели высокая обобщающая способность, гибкость и универсальность подхода. А также показатели CatBoostClassifier на AUC-ROC кривой и Precision-Recall кривой лучше LogisticRegression

Параметры модели

```
cat_params = {
    'loss_function':'Logloss',
    'eval_metric': 'F1',
    'auto_class_weights': 'Balanced',
    'random_state': 42,
    'logging_level': 'Verbose',
    'cat_features': feat_categ,
    'one_hot_max_size': 20,
    'early_stopping_rounds': 50,
}
```

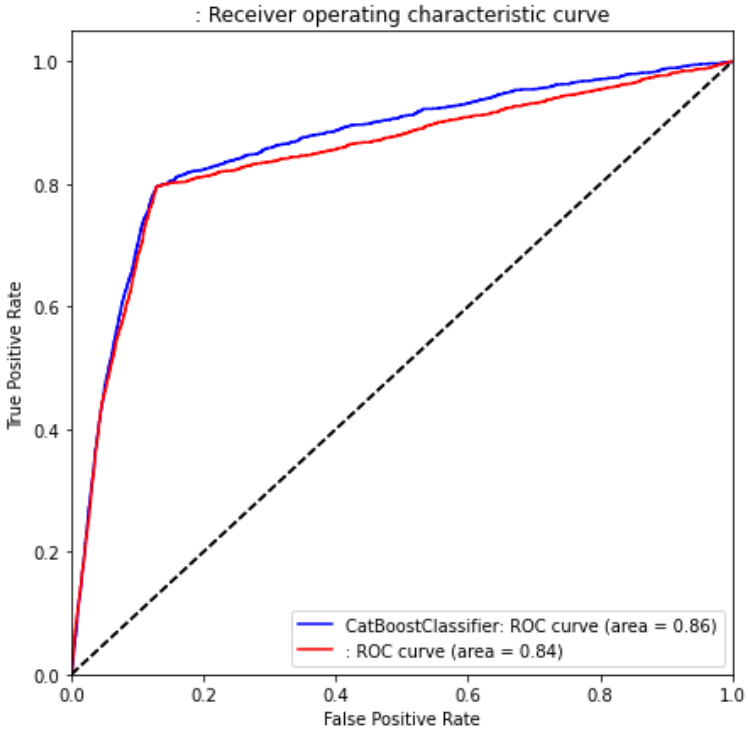
Параметры модели, найденные по сетке кросс-валидации

```
best_cat_params = search_result['params']
best_cat_params
```

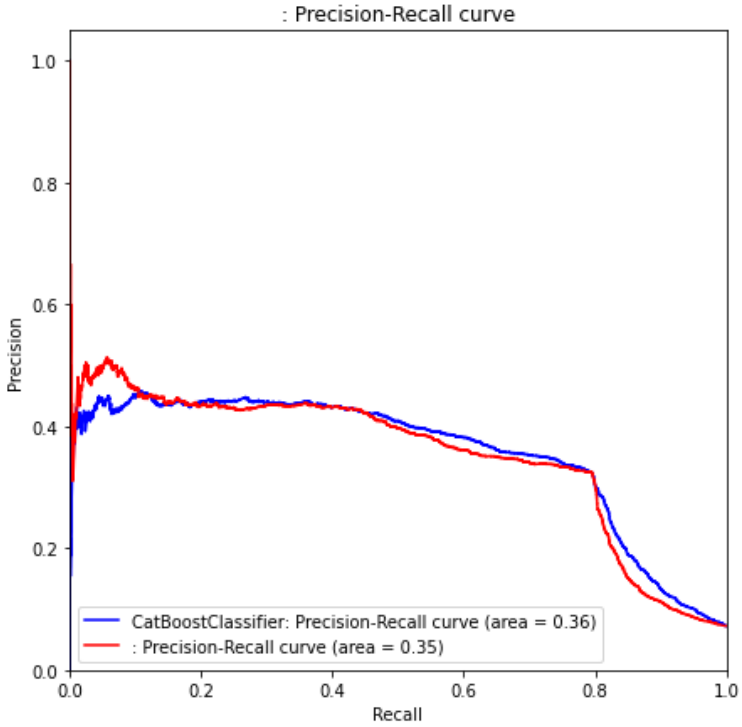
```
{'bagging_temperature': 2,
 'depth': 4,
 'l2_leaf_reg': 5,
 'iterations': 500,
 'learning_rate': 0.03}
```

Перебор параметров настройки моделей на значительных диапазонах не производился в связи с ограничениями по оперативной памяти и производительности технического обеспечения.

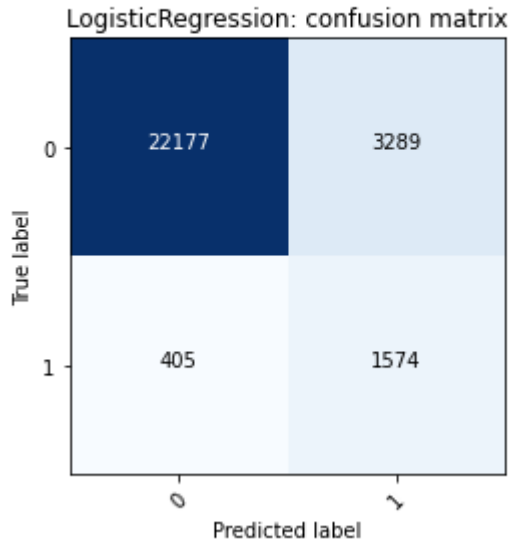
```
CatBoostClassifier: AUC_ROC = 0.863
LogisticRegression: AUC_ROC = 0.845
```



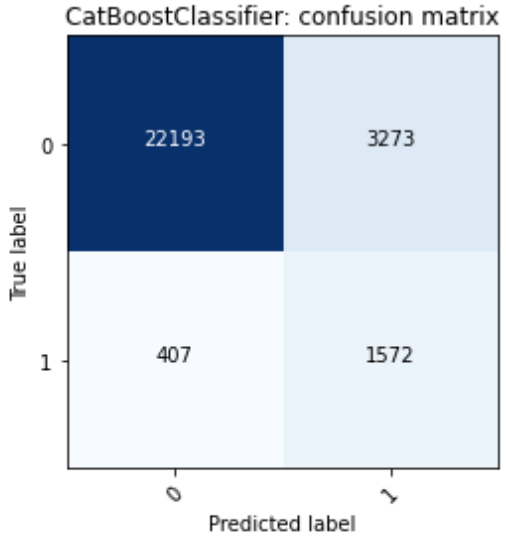

```
CatBoostClassifier: AUC_PR = 0.356
LogisticRegression: AUC_PR = 0.353
```



```
Confusion matrix, without normalization
[[22177  3289]
 [  405  1574]]
```




```
Confusion matrix, without normalization
[[22193  3273]
 [  407  1572]]
```



Матрица смежности показывает

- в левом верхнем углу количество истинных предсказаний класса 0
- в правом верхнем углу количество ложных предсказаний класса 1 (ошибка второго рода)
- в левом нижнем углу количество ложных предсказаний класса 0 (ошибка первого рода)
- в правом нижнем углу количество истинных предсказаний класса 1

Если принять за нулевую гипотезу - положительный отклик клиента на услугу (класс 1), и осуществлять клиентам рассылку предложений по подключению услуги согласно предсказаниям модели, то

- количество ошибок первого рода характеризует сколько клиентов не получили предложения, хотя потенциально они готовы совершить подключение;
- количество ошибок второго рода характеризует сколько клиентов получили предложения, хотя они не собираются совершать подключение.

Оценка модели

```
LogisticRegression
              precision    recall  f1-score   support

         0.0       0.98      0.87      0.92     25466
         1.0       0.32      0.80      0.46      1979

    accuracy                           0.87     27445
   macro avg       0.65      0.83      0.69     27445
weighted avg       0.93      0.87      0.89     27445


CatBoostClassifier
              precision    recall  f1-score   support

         0.0       0.98      0.87      0.92     25466
         1.0       0.32      0.79      0.46      1979

    accuracy                           0.87     27445
   macro avg       0.65      0.83      0.69     27445
weighted avg       0.93      0.87      0.89     27445
```

f-score — среднее гармоническое precision и recall и она держит баланс между метриками. В случае, если мы хотим охватить больше клиентов, то ориентируемся по recall. Ну, а если мы не хотим тратить неэффективно бюджет на FP, то лучше смотреть с большим уклоном на precision (precision обеспечивает максимальное содержание TP в предсказаниях модели)