Курсовой проект от Мегафон

Федоткин Олег
GeekBrains, факультет искуственного интеллекта

Постановка задачи и исходные данные

Необходимо построить алгоритм, который для каждой пары пользователь-услуга определит вероятность подключения услуги.

Исходные данные:

data_train.csv, data_test.csv, features.csv,

Метрика:

Скоринг осуществляется функцией f1, невзвешенным образом, аналогично функции sklearn.metrics.f1_score(..., average='macro'.

Этапы решения

1. Загрузка данных
2. EDA
3. Построение новых признаков
4. Отбор признаков
5. Построение модели
6. Прогнозирование на тестовом датасете

Отбор признаков и выбор модели

Из 254 признаков были отобраны 100: удалены константные, преобразованы категориальные через кодирование к целевой переменной, вещественные признаки были проверены с помощью VIF на мультиколлинеарность и 110 признаков из них не были в дальнейшем использованы.

Обучены следующие модели: KNeighborsClassifier, XGBClassifier, CatBoostClassifier, LGBMClassifier

Лучшая модель CatBoostClassifier

	model	test_f1_score
0	model_knn	0.506396
1	model_xgb	0.695987
2	model_lgbm	0.707675
3	model_catb	0.717952

Используемая модель

```
final_model = catb.CatBoostClassifier(n_estimators=100, max_depth=10, silent=True,
                                      learning_rate=0.1, random_state=21)
final_model.fit(X_train, y_train)
```