Видеокурс от Megafon + курсовой проект

Финальный проект

Гасилин М.В.

Информация о модели:

Для решения задачи была выбрана модель CatBoost Classifier со следующими параметрами:

Параметр	Значение
Максимальное количество деревьев, которые решают задачу (n_estimators)	200
Глубина дерева (max_depth)	3
Коэффициент регуляризации L2 (l2_leaf_reg)	5

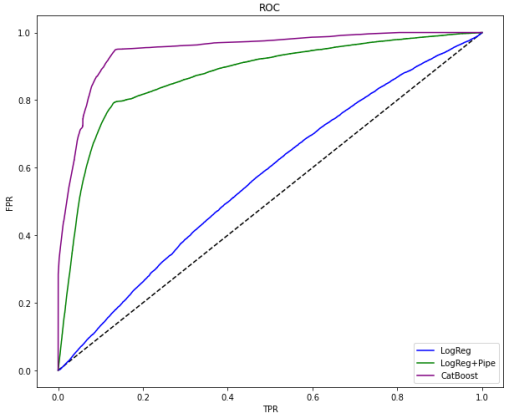
Параметры были подобраны в результате использования метода grid search.

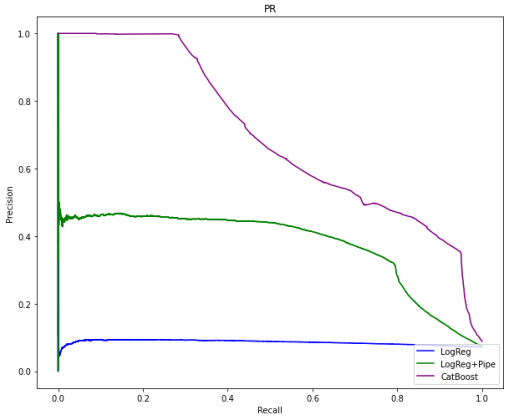
```
In [77]: lg_gsc = run_grid_search(model_catb, X_train, y_train, param_grid, kfold_cv)

Best roc_auc score: 0.94

Best parameters set found on development set:

{'depth': 3, 'iterations': 300, 'l2_leaf_reg': 5}

Grid scores on development set:

0.941 (+/-0.002) for {'depth': 3, 'iterations': 300, 'l2_leaf_reg': 5}
0.941 (+/-0.002) for {'depth': 3, 'iterations': 300, 'l2_leaf_reg': 15}
0.941 (+/-0.002) for {'depth': 3, 'iterations': 300, 'l2_leaf_reg': 25}
```

Обоснование выбора модели:

Сравнительные тесты 3-х моделей Логистической регрессии, Логистической регрессии с использованием пайплайнов и CatBoost.

Из графиков видно, что CatBoost уверенно лидирует при анализе общепринятых показателях (ROC AUC, AUC(recall, precision)) качества моделей на тренировочном наборе данных.

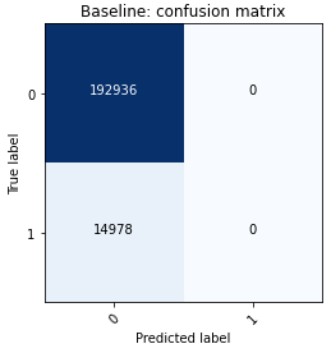




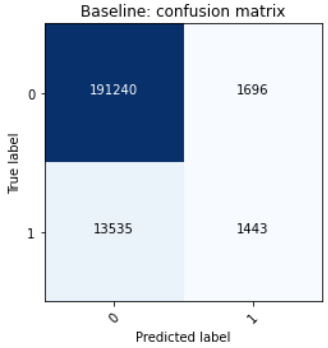
Обоснование выбора модели:

Матрицы ошибок классификации для исследованных моделей:

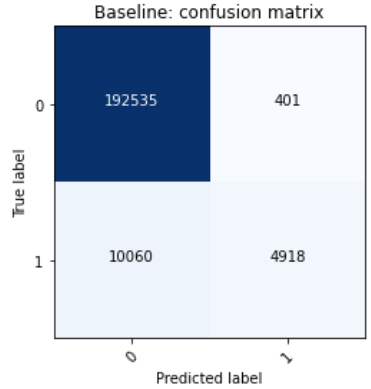
Логистическая регрессия



Логистическая регрессия
с пиплайнами



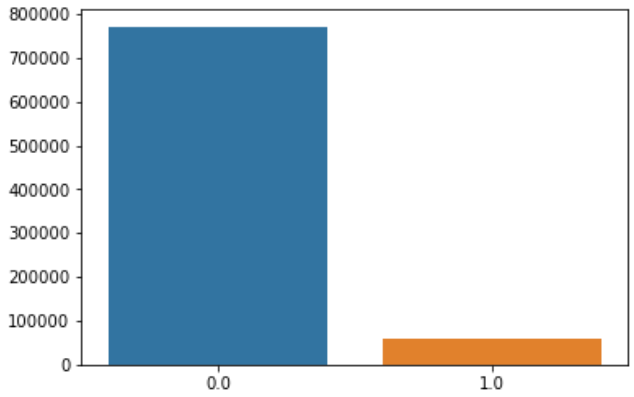
CatBoost



Принцип составления индивидуальных предложений для абонентов:

Поскольку разбиение положительных и отрицательных откликов в тренировочной выборке разбилось на следующие части:

target	Количество	Доля	Доля %
0	771467	0,927631	92,76%
1	60186	0,072369	7,24%
Итого:	831653		



То и полученные ответы в файле answers_test.csv с большой долей вероятности должны быть разделены в такой же пропорции. А для этого порог отделения вероятности должен быть 0,34138984.

И предложить услуги только 4398 абонентам.

target	Количество	Доля	Доля %
0	66833	0,938257219	93,83%
1	4398	0,061742781	6,17%
Итого:	71231		

Не реализованное, но потенциально полезно:

К сожалению, время отведенное на решение задачи в рамках учебного процесса и «домашние» компьютерные мощности не позволили в полной мере исследовать файл features.csv. Как мне кажется, использование статистических гипотез или нейронных сетей смогло бы разбить профили всех пользователей на классы эквивалентности. И динамику перетекания абонентов между этими классами. Что, в свою очередь, позволило бы выявить более тонкие взаимосвязи между абонентами. И, как мне кажется, добавило в модель новые показатели. Которые увеличили бы точность предсказаний.

Но на текущем этапе моего обучения мне потребуется на этот анализ и корректировку модели от 3-х до 6-ти месяцев ежедневной полной занятости. Жаль, что не получится. Т.к. задача очень интересная.