Курсовой проект
“Вероятность подключения тарифа”

Заказчик: МегаФон

Автор: Сергеев Константин Олегович

Постановка задачи

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

Данные

features.csv.zip: id, …
data_train.csv: id, vas_id, buy_time, target
data_test.csv: id, vas_id, buy_time

target - целевая переменная, где 1 означает подключение услуги, 0 - абонент не подключил услугу соответственно.
buy_time - время покупки.
id - идентификатор абонента
vas_id - подключаемая услуга

Метрика

sklearn.metrics.f1_score(…, average='macro').

Формат представления результата

Работающая модель в формате pickle, которая принимает файл data_test.csv из корневой папки и записывает в эту же папку файл answers_test.csv. В этом файле должны находится 4 столбца: buy_time, id, vas_id и target. Target можно записать как вероятность подключения услуги.

Код модели можно представить в виде jupyter-ноутбука.

Презентация в формате .pdf, в которой необходимо отразить:

Информация о модели, ее параметрах, особенностях и основных результатах.
Обоснование выбора модели и ее сравнение с альтернативами.
Принцип составления индивидуальных предложений для выбранных абонентов.
Рекомендуемое количество слайдов – 5 – 10.

	id	vas_id	buy_time	target
415826	2120971	5.0	1546203600	0.0
347850	4075619	2.0	1546203600	0.0
347990	2446060	2.0	1546203600	0.0
347967	212414	2.0	1546203600	0.0
347934	1175153	2.0	1546203600	0.0

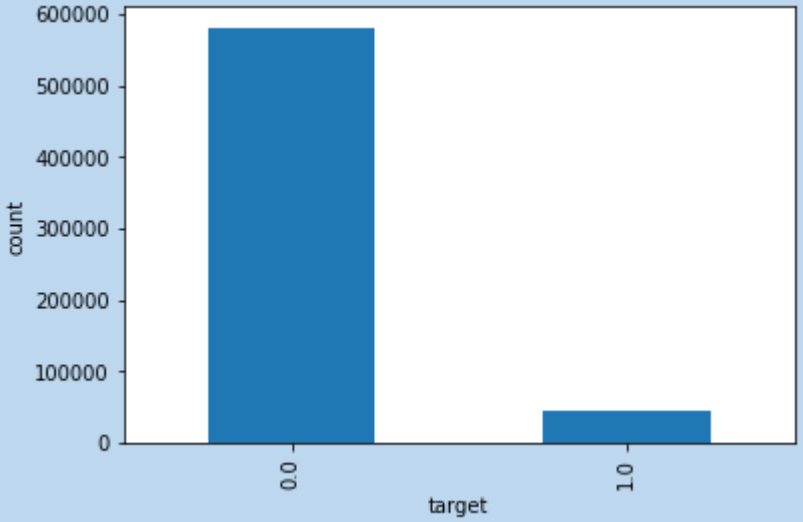
df_train

Анализ данных

Наблюдается сильный дисбаланс классов

```
0.0    580393
1.0     43346
Name: target, dtype: int64
```

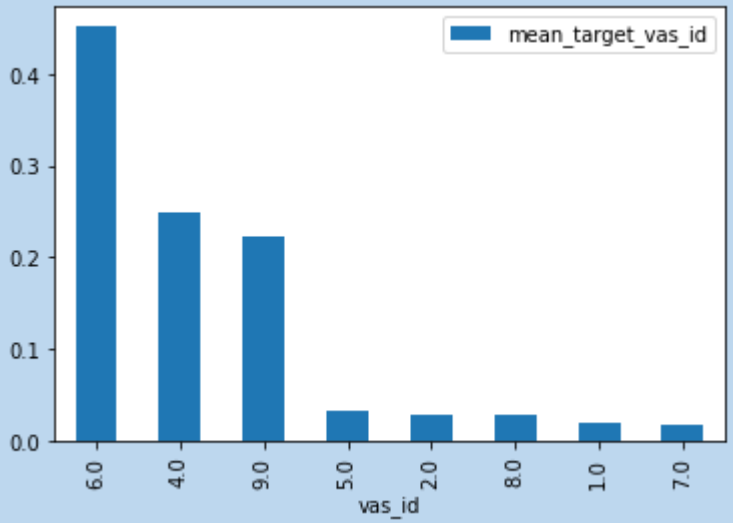
$$\text{disbalance} = \frac{\sum(target == 0)}{\sum(target == 1)} \approx 13.38977$$



Всего у нас 8 тарифов

Средняя подключаемость тарифов:

vas_id	mean_target_vas_id
6.0	0.452483
4.0	0.248161
9.0	0.222981
5.0	0.033415
2.0	0.028479
8.0	0.028118
1.0	0.020216
7.0	0.016899



Построение baseline

1). Пускай абоненты соглашаются только на самый подключаемый тариф, причём всегда

2). Добавим к предыдущему правилу следующий по популярности тариф

3). Добавим к предыдущему правилу следующий по популярности тариф

4). Пускай абоненты соглашаются только на те тарифы, которые они уже подключали

5). Совместим 2). и 4). алгоритмы

Наилучший результат на валидационной выборке показал 2 алгоритм

	precision	recall	f1-score	support
0.0	1.00	0.87	0.93	191074
1.0	0.40	0.99	0.57	16840
accuracy			0.88	207914
macro avg	0.70	0.93	0.75	207914
weighted avg	0.95	0.88	0.90	207914

SCORE: 0.7497796420417978

Примем его результат за базовый

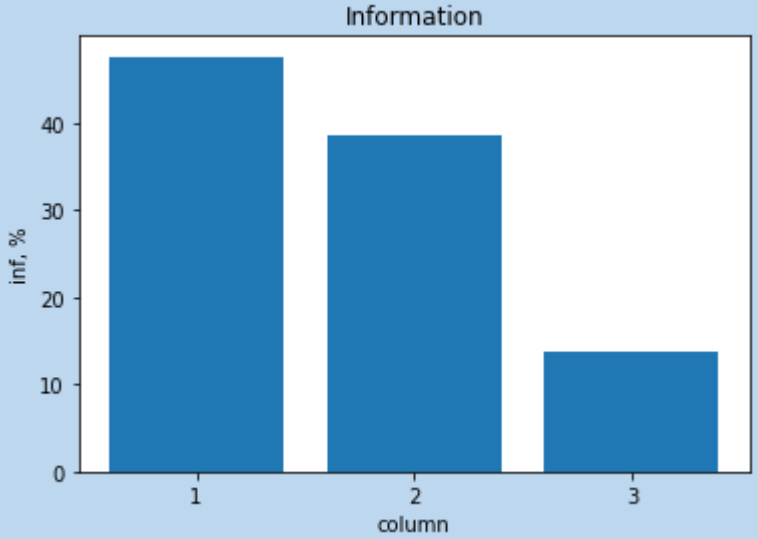
Обработка данных перед обучением

1). Добавим вспомогательный признак mean_target_vas_id, который будет явно показывать модели среднюю подключаемость тарифов

2). Уменьшим количество анимированных признаков. Во-первых уберём те, которые не изменяются, так как они не несут информации. Во-вторых воспользуемся PCA.

const_features = ['75', '81', '85', '139', '203']

Возьмём новое количество анонимированных признаков равным 3 и посмотрим результат.



Потерянная информация оказалась равной 0.0167376 %.
Что относительно немного.

Информативность получившихся столбцов

Результаты обучения

Походу обучения подбирался threshold для каждой модели, путём перебора на тренировочной выборке.

Видно, что 2 алгоритм baseline очень хорошо себя показывает на валидационной выборе и его смог превзойти только XGB с тщательно подобранными гиперпараметрами в ручную.

LogisticRegression показал себя очень плохо, значит наши классы не являются линейно разделимыми.

Для тщательной настройки параметров была выбрана XGB, так как это изначально более сильная модель при хорошо заданных параметрах. Зато для первичного результата очень хорошо подходит CatBoost, так как он более автономный, а как видно результат тоже очень неплохой. При этом две эти модели одного класса, что нам подсказывает один и тот же найденный лучший threshold=0.83.

	score
XGBoost with PCA	0.750921
baseline 2 (top 2)	0.749780
baseline 5 (top 2 + last buy)	0.749041
baseline 1 (top 1)	0.747702
CatBoost	0.747347
CatBoost with PCA	0.747317
baseline 3 (top 3)	0.743471
Random Forest with PCA	0.731184
baseline 4 (last buy)	0.479148
Logistic Regression with PCA	0.448972

Анализ лучшей модели

Выбираем модель XGB с такими гиперпараметрами:

learning_rate = 0.1 – шаг обучения

colsample_bytree = 0.3 – соотношение подвыборок столбцов при построении каждого дерева

max_depth = 7 – максимальная глубина деревьев

n_estimators = 100 – количество деревьев

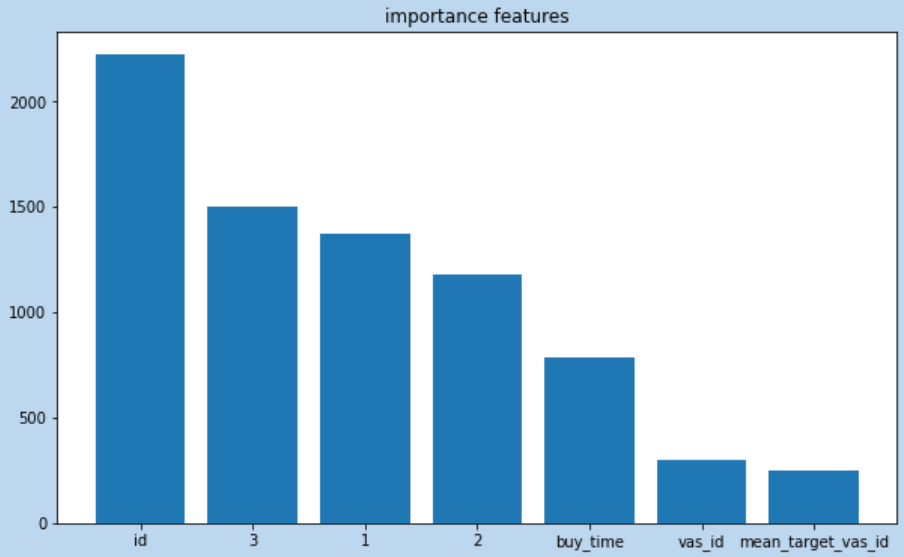
reg_lambda = 4 – регуляризация L2

scale_pos_weight = disbalance – балансировка классов target

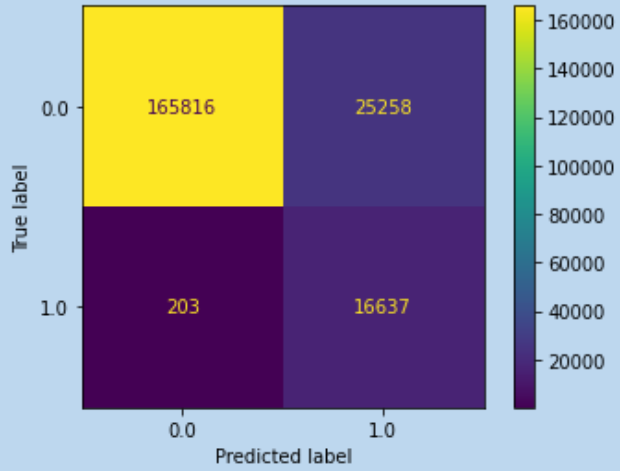
seed = 42 – начальное случайное число

На обучающей выборке был найден лучший threshold = 0.83, методом перебора с шагом 0.01.

Importance features



Confusion Matrix (val)



Видно, что наибольшую значимость имеет id, значит выбор тарифа зависит в большей степени от истории абонента, которому предлагается тариф, от чего он отказывался и на что он соглашался. Так же большую роль играют анонимированные признаки, так как они содержат большую разнородность данных собранную с 200+ столбцов. А вот созданный новый признак имеет малую значимость, модель и без него учитывает среднюю подключаемость тарифов.

Итоговое решение