Курсовой проект от компании Мегафон

Описание кейса

Yury Nikiforov, 03.2020

https://github.com/Nickel-nc/GU_Megafon_Case/blob/master/YN_Megafon_Project.ipynb

Структура подготовки данных

- Входными параметрами в модели служат данные о предагаемой услуге и профиль потребления абонента (**train.csv/test.csv** и **features.csv**)

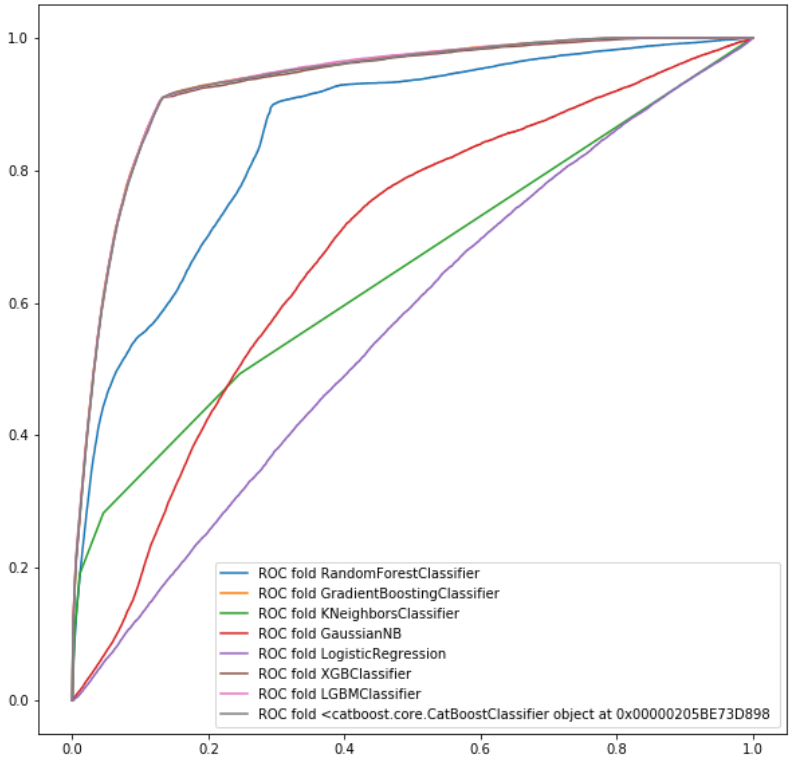
- Слияние датасетов проходит по пересечению **id** пользователя в массивах. Такой вариант не является лучшим, но приемлем в виду особенностей полноты входных данных:

- - Динамика изменения профиля потребления абонента непостоянна. Далеко не для всех размеченных данных доступно изменение профиля потребления. Логично использовать пересечение по полям **id** и **buy_time**, но в текущей ситуации это ведет к потери до 90% обучающих данных.

- - На входном временном отрезке профиль меняется не значительно. Поэтому в модели использованы данные об активности профиля для известной даты.

Выбор модели и подбор параметров

- Подготовленные данные опробированы на **8** классических ML-алгоритмах с **дефолтными** параметрами.
- Лучшую точность показали ансамбли. На основе отчетов из classifitation_report выбрана модель с наибольшей метрикой f1 score avg='macro' - **Cat Boost**, которая использована в бейзлайне. Так как категориальные признаки в чистом виде в данных отсутствуют, подойдут и другие алгоритмы, основанные на градиентном бустинге.
- Основная задача доработки модели состоит в том, чтобы повысить recall в минорном классе, который является целевым.
- В целях экономии времени подбор параметров проходил на батче тренеровочного датасета меньшего размера (100к записей). Подбор параметров улучшил предсказательную способность модели примерно на 5 %.

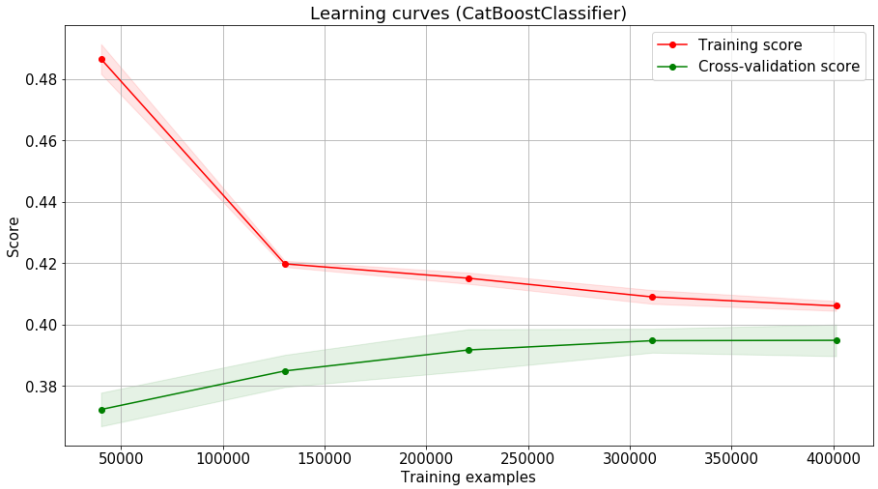


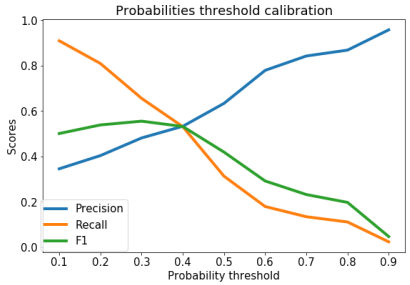
Результаты подбора параметров

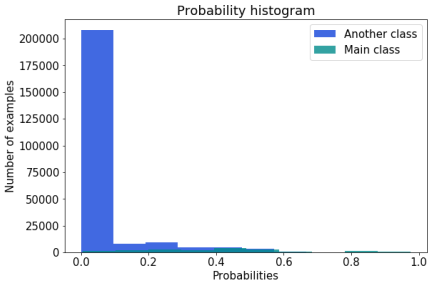
- Для улучшения модели использована балансировка минорного класса посредством **oversampling**. Для этого использовался алгоритм **ADASYN**, в котором заполняемые данные лежат на отрезках между узлами k-ближайших соседей с добавлением случайной дисперсии. Подбором коэффициента балансировки можно улучшить качество метрики до **1%.** Балансировка показала лучший результат на соотношении **0.7.**

- Масштабирование признаков не показала видимого прироста качества метрики.

- В результате порог вероятности для целевой переменной сместился **с 0.2 до 0.4**, метрика **f1-score** увеличилась до **0.69**, а recall – до **0.32.**







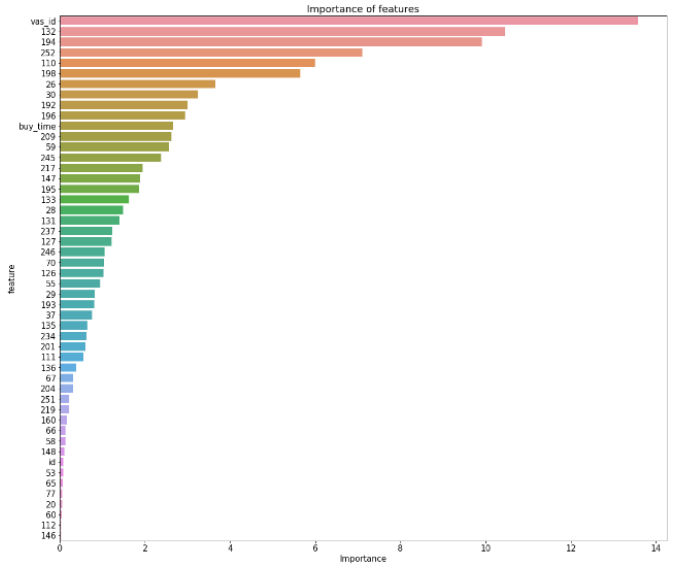
f1	precision	recall	probability
0.556	0.482	0.656	0.3
0.539	0.404	0.809	0.2
0.532	0.533	0.531	0.4
0.501	0.346	0.91	0.1
0.419	0.634	0.313	0.5
0.292	0.779	0.18	0.6
0.233	0.843	0.135	0.7
0.198	0.868	0.112	0.8
0.048	0.958	0.024	0.9

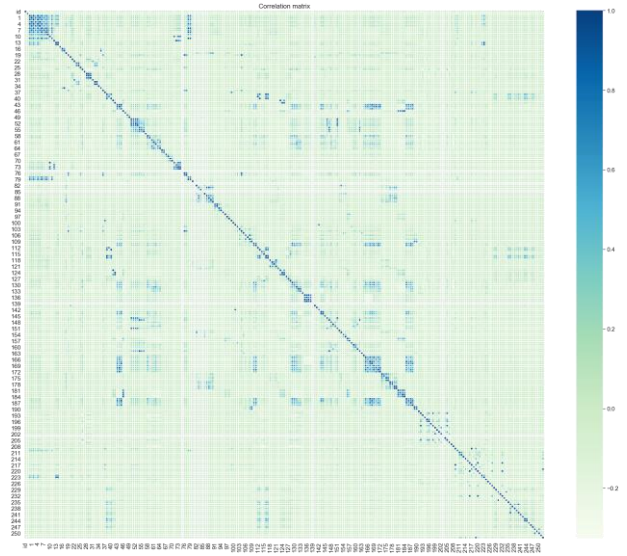
Влияние признаков на работу модели

- Оценка работы модели показала нулевую значимость большинства признаков на результат. На графике представлен рейтинг 50 признаков по значимости. Это значит, что остальные признаки создают излишнюю вычислительную нагрузку на модель и могут быть исключены.

- Можно отметить значимость времени предложения и покупки услуги. Это говорит о статистической значимости наличия полноты истории профиля на каждую отчетную дату для формирования лучшей предсказательной способности модели.

- Коррелограмма базовых признаков между собой высокую линейную зависимость между некоторыми группами данных. Их можно объединить в кластеры.



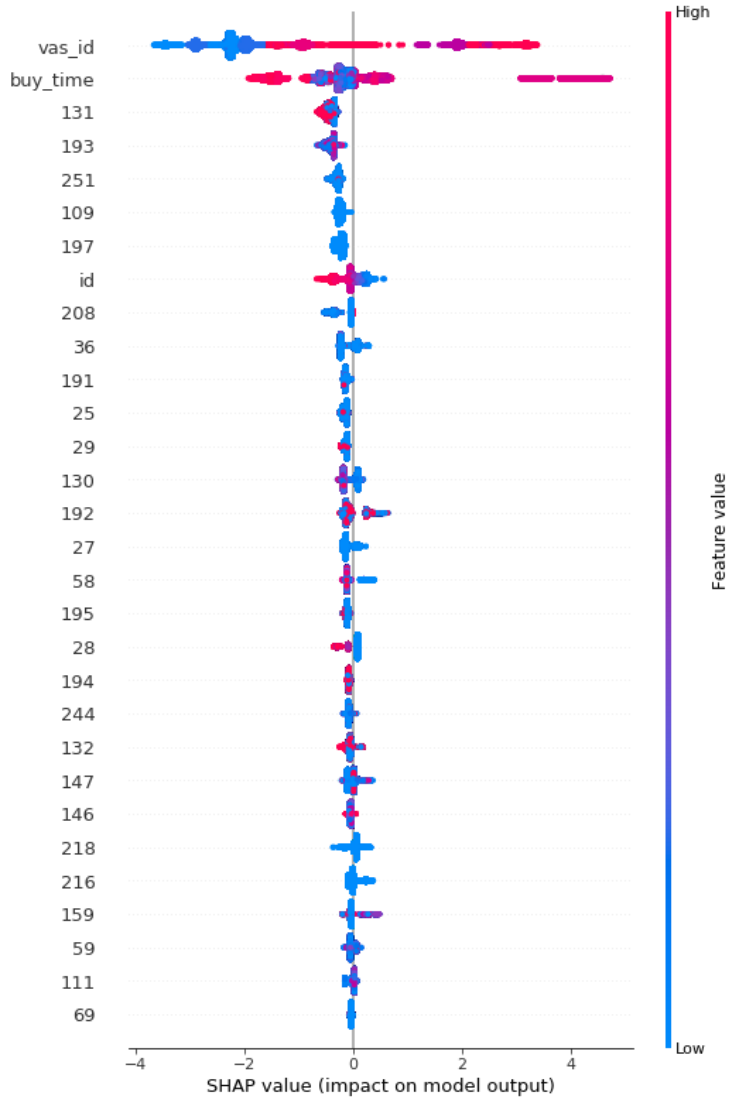


Анализ признаков на работу модели

- **Shap** объясняет важность предикторов при построении модели. Значения слева от центральной вертикальной линии — это negative класс (0), справа — positive (1). Чем толще линия на графике, тем больше таких точек наблюдения. Чем краснее точки на графике, тем выше значения фичи в ней.

- Анализ показал, что на покупку услуги в наибольшей степени влияет характер самой услуги и её своевременное предложение (buy_time). Можно отметить ряд признаков из профиля потребления, которые влияют на вероятность покупки предлагаемых услуг.

- Зная природу обезличенных данных, можно сделать более интересные выводы.



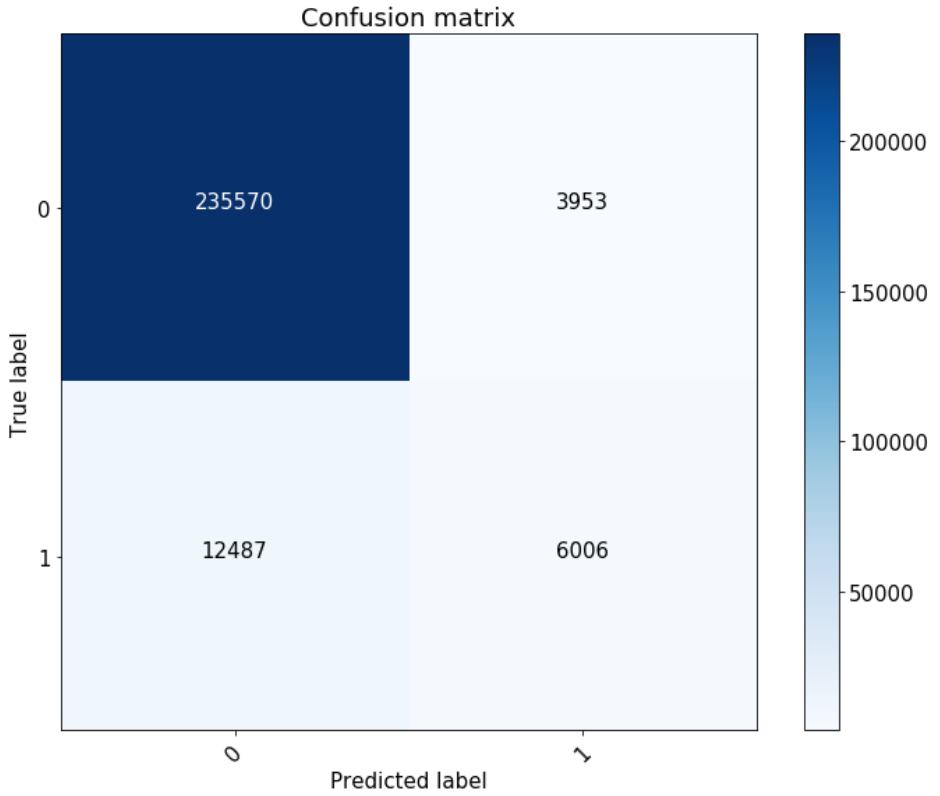
Выводы

- В результате проведенных операций модель научилась верно определять около половины случаев приобретения предлагаемых услуг, и с высокой точностью идентифицировать абонентов, которые не приобретут ту, или иную услугу.

- Стоить отметить наличие вероятностной составляющей при формировании входных данных, таких, как неравномерность предложения услуг по времени, их количества.

- Использование такой модели позволит повысить монетизацию деятельности компании при своевременном предложении целевой услуги абонентам при определенном состоянии их профиля.

- Возможным качественным улучшением модели может послужить сбор данных об актуальности той или иной услуги среди населения в пределах заданного региона



Составление индивидуального предложения

- В основе формирования индивидуального предложения лежит идея предсказания вероятности приобретения абонентом одной, или нескольких услуг при их предложении по текущему состоянию его профиля потребления.
- Идея такого подхода заключается в том, чтобы модель обучилась на размеченных данных предсказывать вероятность покупки поочередно каждой из доступных услуг, при используя лишь данные о профиле потребления. Таким образом, каждая модель имеет свои веса в признаках для каждой услуги

	id	vas_id	buy_time	service # 1	service # 2	service # 4	service # 5	service # 6	service # 7	service # 8	service # 9
0	1054895	2.0	1546808400	0.022448	0.024107	0.267785	0.026638	0.476896	0.011383	0.022510	0.200303
1	1055365	2.0	1548018000	0.020135	0.017459	0.408024	0.018402	0.479087	0.008909	0.046440	0.220383
2	1055795	2.0	1548018000	0.017917	0.021060	0.291930	0.022518	0.442719	0.015466	0.018273	0.295368
3	3770386	2.0	1546808400	0.010058	0.011548	0.165711	0.006890	0.340082	0.014103	0.004199	0.029354
4	2763963	1.0	1546808400	0.011542	0.016133	0.206276	0.025435	0.418571	0.012567	0.004923	0.082420
...	...	...	...	...	...	...	...	...	...	...	...
190	809752	2.0	1547413200	0.014348	0.013524	0.297016	0.014099	0.445570	0.011143	0.021406	0.191518
191	112936	2.0	1547413200	0.026948	0.045427	0.312387	0.063967	0.364001	0.011161	0.009571	0.093627
192	765439	2.0	1547413200	0.068519	0.011458	0.282078	0.015099	0.372811	0.014847	0.059866	0.253805
193	158241	4.0	1548018000	0.018379	0.029254	0.385354	0.018353	0.447022	0.030675	0.038469	0.372930
194	158241	6.0	1547413200	0.030080	0.026325	0.371186	0.016387	0.437630	0.012019	0.015303	0.340582

Комментарии:

- Для каждой услуги вероятность самостоятельна и их сумма может превышать 1, т.к. клиент может приобрести несколько услуг .
- На текущем этапе целесообразно использовать вероятностное представление ответов, при настройке порога вероятности возможен учет весов, как выгоды для компании от факта подключения той или иной услуги.
- В целом, метрики качества в модели показали себя посредственно. Можно отметить дисбаланс в обучающих данных - при детализации до конкретной услуги сказывается недостаток размеченных данных.