



Курсовой проект от компании Мегафон

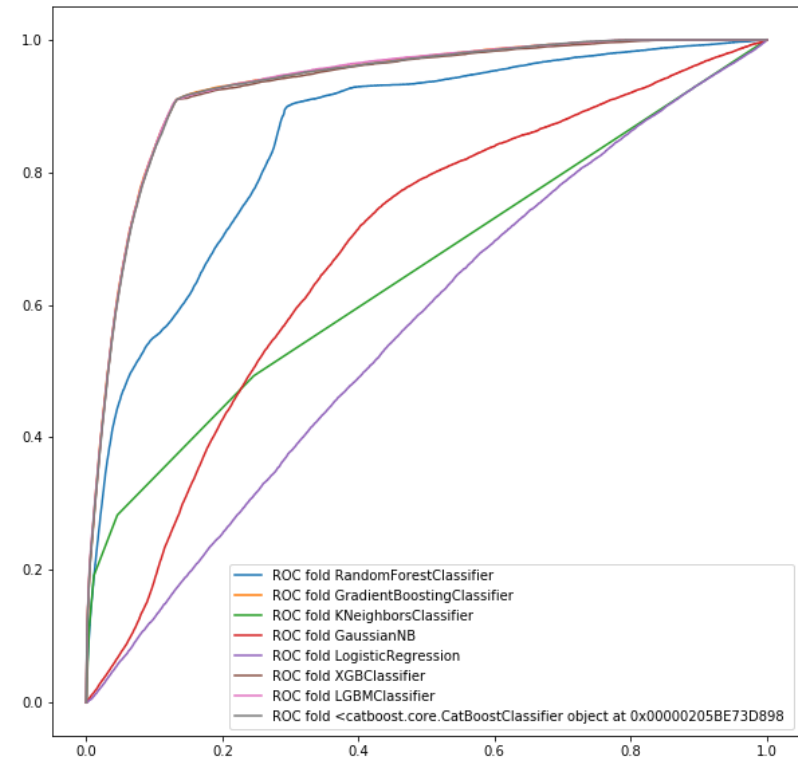
Описание кейса

Структура подготовки данных

- Входными параметрами в модели служат данные о предоставляемой 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 потере до 90% обучающих данных.
- - На входном временном отрезке профиль меняется не значительно. Поэтому в модели использованы данные об активности профиля для известной даты.

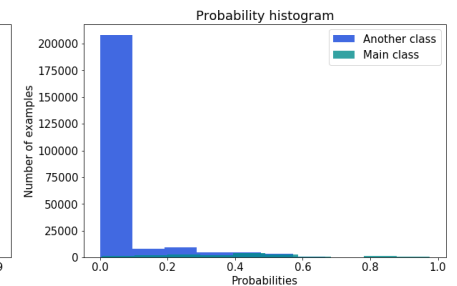
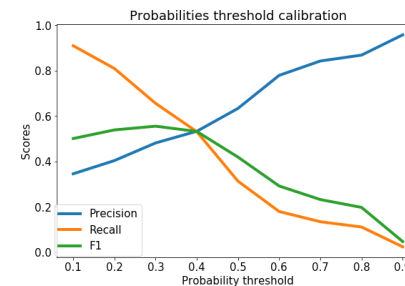
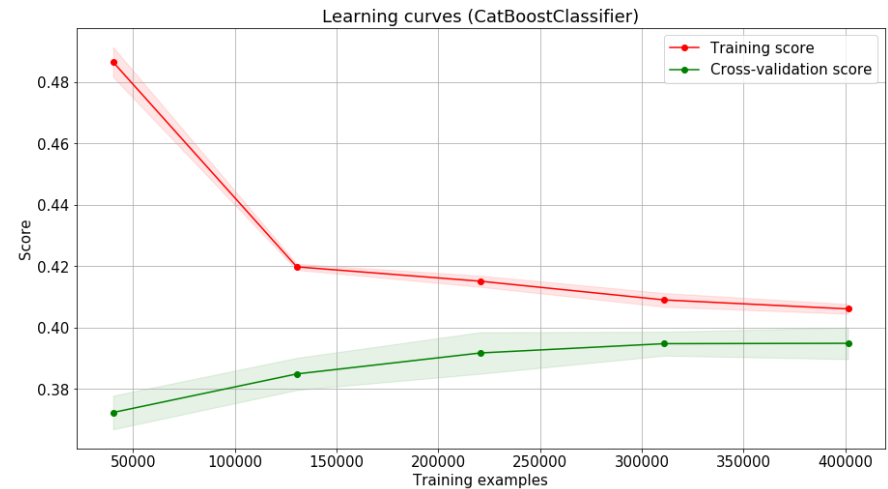
Выбор модели и подбор параметров

- Подготовленные данные опробированы на **8** классических ML-алгоритмах с **дефолтными** параметрами.
- Лучшую точность показали ансамбли. На основе отчетов из `classification_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 тренировочного датасета меньшего размера (100к записей). Подбор параметров улучшил предсказательную способность модели примерно на 5 %.



Результаты подбора параметров

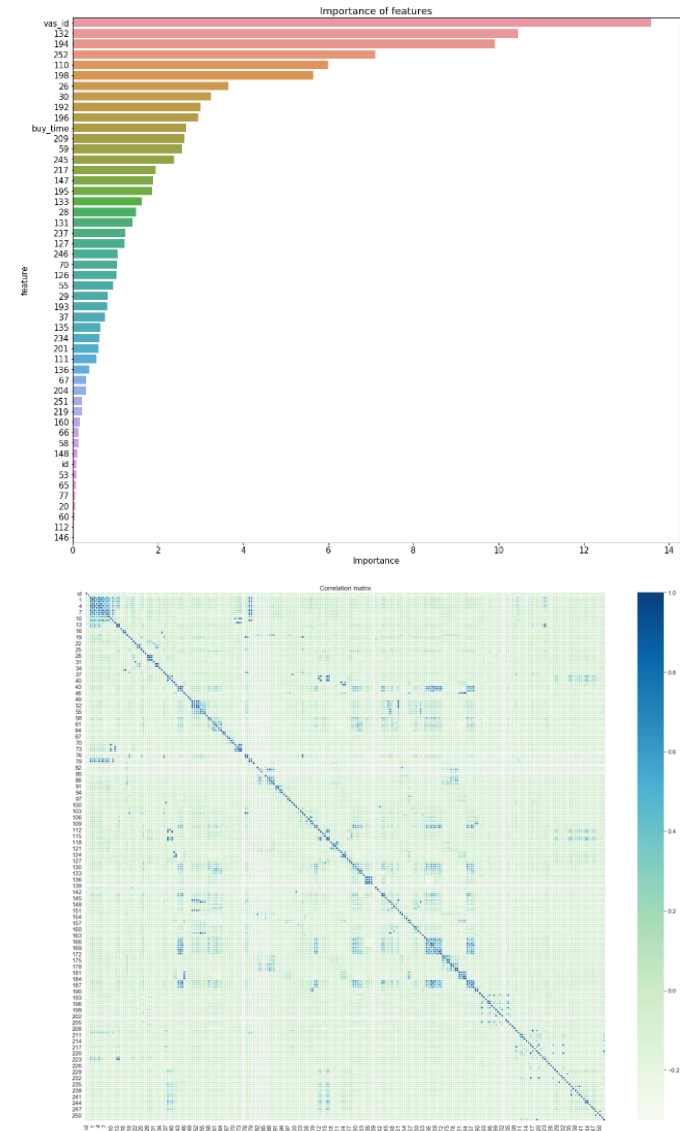
- Для улучшения модели использована балансировка минорного класса посредством **oversampling**. Для этого использовался алгоритм **ADASYN**, в котором заполняемые данные лежат на отрезках между узлами k-ближайших соседей с добавлением случайной дисперсии. Подбором коэффициента балансировки можно улучшить 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

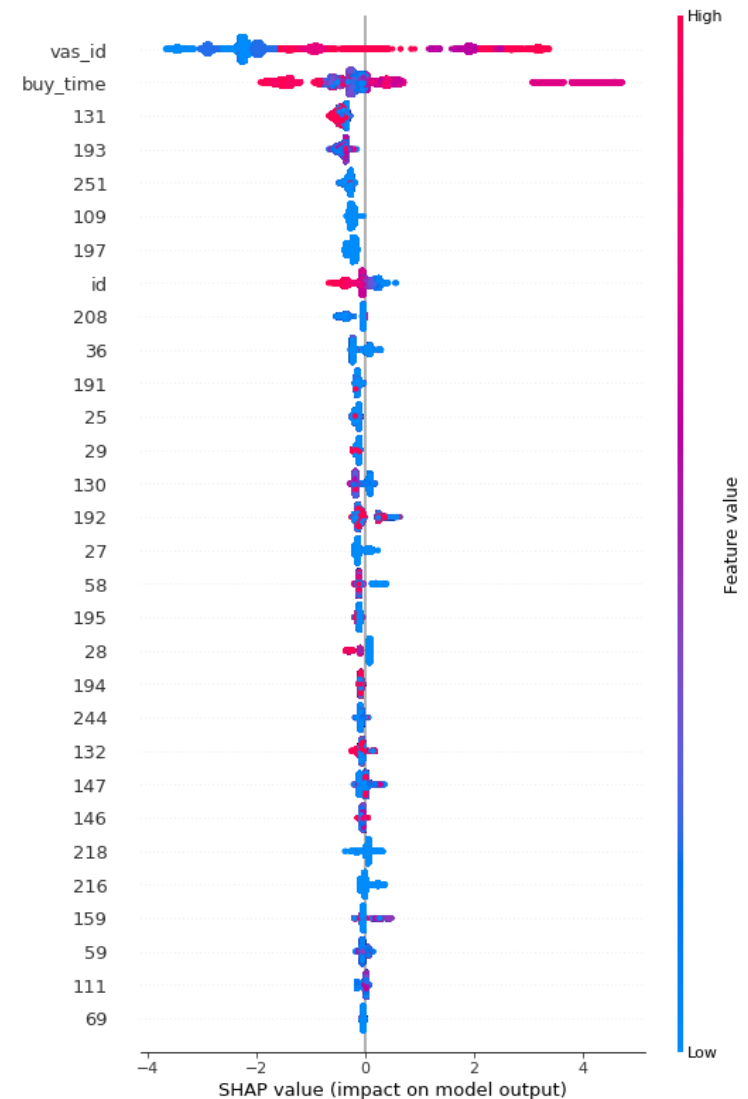
Влияние признаков на работу модели

- 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.



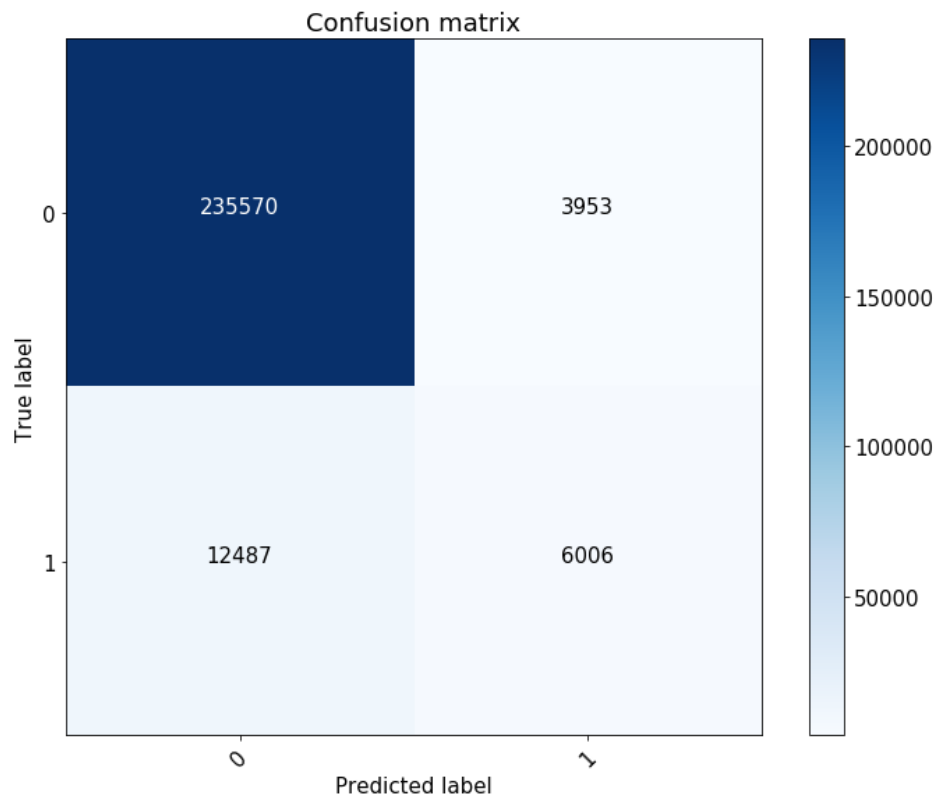
Анализ признаков на работу модели

- **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

- Индивидуальное предложение представляется, как задача мультиклассовой классификации, где для абонента рассчитывается вероятность приобретения той, или иной услуги. Признаками будет служить профиль потребления абонента для отчетной даты формирования предложения.
- Для этого использованы размеченные данные, для которых заведомо известен факт покупки услуги.
- Обучение проводилось алгоритмом **Cat boost**. В итоге получен датасет с вероятностями приобретения той, или иной услуги для каждого пользователя.
- Данная гипотеза требует дополнительного тестирования и возможной настройки.

	id	service # 4	service # 5	service # 2	service # 6	service # 1	service # 9	service # 8	service # 7
1402	4311111	0.103472	0.077442	0.363613	0.037183	0.390437	0.003477	0.004818	0.019559
2184	4148312	0.094584	0.075362	0.314463	0.025551	0.459302	0.004190	0.006338	0.020210
514	1851789	0.088550	0.067129	0.355028	0.027702	0.440361	0.003239	0.006121	0.011869
2340	1669570	0.096643	0.101587	0.332239	0.031339	0.409522	0.003517	0.007632	0.017521
68	2198594	0.072745	0.090245	0.406188	0.033856	0.375705	0.003336	0.002788	0.015137
...
2085	2194580	0.088293	0.078142	0.369767	0.030300	0.410913	0.002754	0.005550	0.014281
1981	1974802	0.090028	0.093871	0.328476	0.023004	0.437929	0.002540	0.004057	0.020095
1468	4016560	0.101412	0.086076	0.336380	0.035369	0.412228	0.005154	0.007119	0.016263
11	2350392	0.093845	0.091121	0.416337	0.027189	0.349841	0.002591	0.004056	0.015021
1301	710027	0.106569	0.075685	0.344350	0.034625	0.415774	0.002381	0.003266	0.017350