МегаФон

# Курсовой проект от компании Мегафон

## Описание кейса

кодил: Юрий Никифоров

# Структура подготовки данных

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

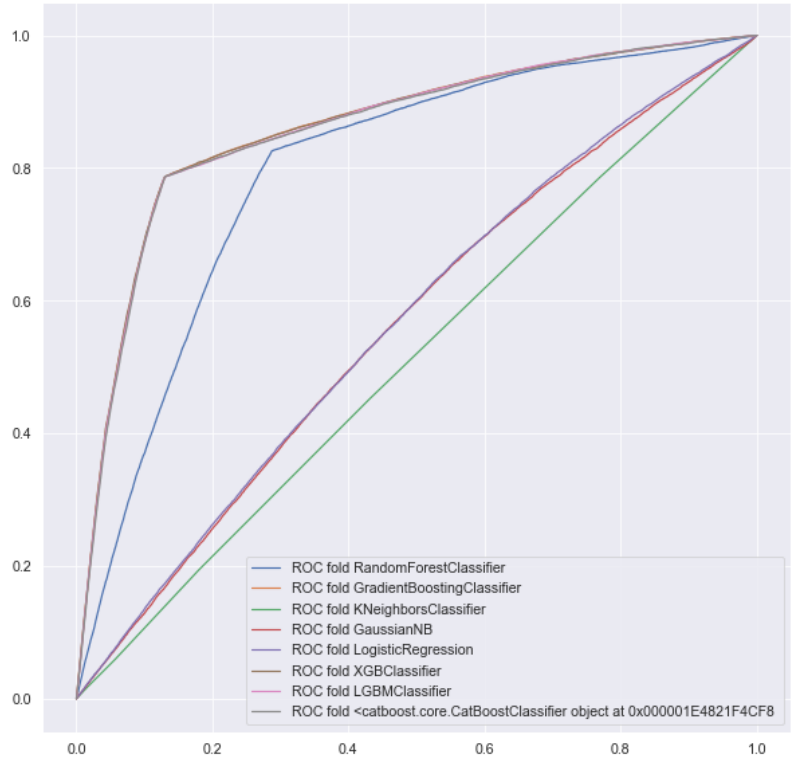
- Слияние датасетов проходит по пересечению **id** пользователя в массивах. Такой вариант не является лучшим, но приемлем в виду особенностей полноты входных данных:

- - 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

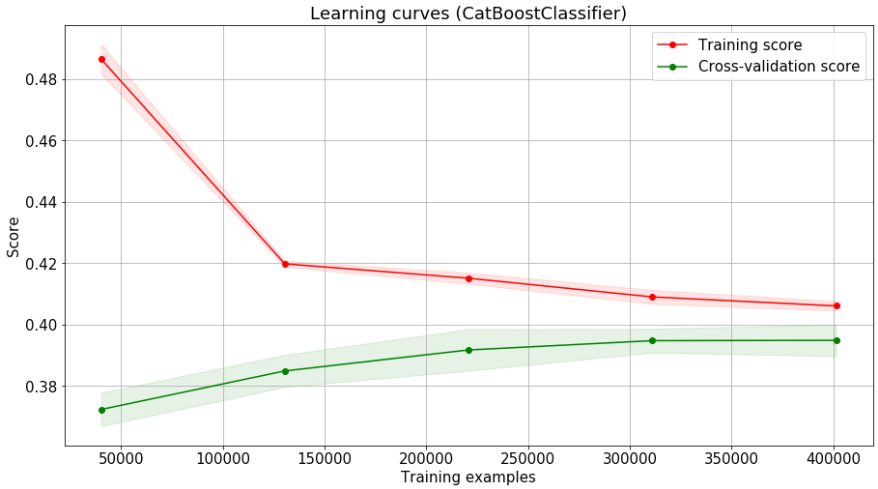
- Подготовленные данные опробированы на **8** классических ML-алгоритмах с **дефолтными** параметрами.
- Лучшую точность показали ансамбли. На основе отчетов из classifitation_report выбрана модель с наибольшей метрикой f1 score avg='macro' (**Cat Boost**), которая использована в бейзлайне.
- В отличие от общей точности предсказания, каждая модель имела крайне низкий recall в минорном классе, который является целевым (базовый recall catb на тесте = **0.04**). Модель требует доработки.
- В целях экономии времени подбор параметров проходил на батче тренеровочного датасета меньшего размера (100к записей). Подбор параметров улучшил предсказательную способность модели примерно на 5 %.

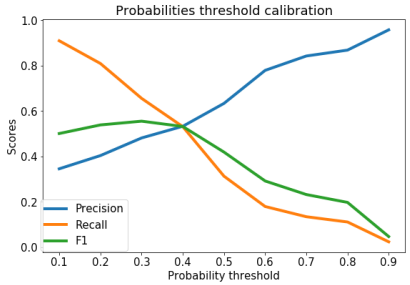


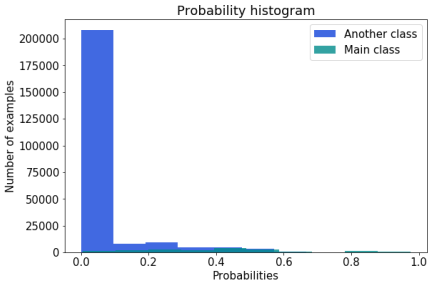
# Результаты подбора параметров

- Наибольший вклад в улучшение модели внесла 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. Подбор коэффициента балансировки показал лучший результат на соотношении **0.7.**

- В результате порог вероятности для целевой переменной сместился **с 0.2 до 0.4**, метрика **f1-score** увеличилась до **0.69**, а recall – до **0.31.**



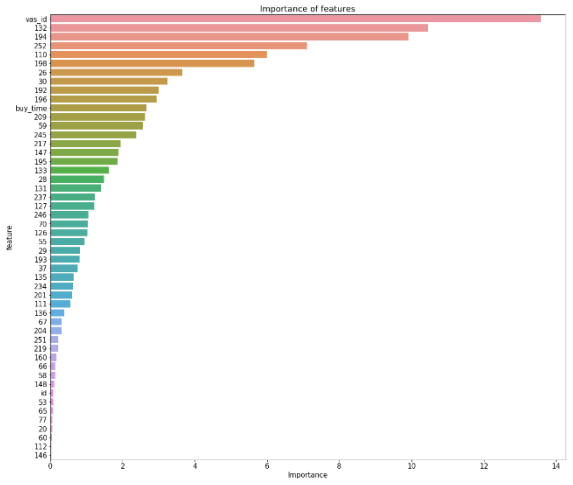

| f1 | precision | recall | probability |
|---|---|---|---|
| 0.556 | 0.482 | 0.656 | 0.3 |
| 0.539 | 0.404 | 0.809 | 0.2 |
| 0.532 | 0.533 | 0.531 | 0.4 |
| 0.501 | 0.346 | 0.91 | 0.1 |
| 0.419 | 0.634 | 0.313 | 0.5 |
| 0.292 | 0.779 | 0.18 | 0.6 |
| 0.233 | 0.843 | 0.135 | 0.7 |
| 0.198 | 0.868 | 0.112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

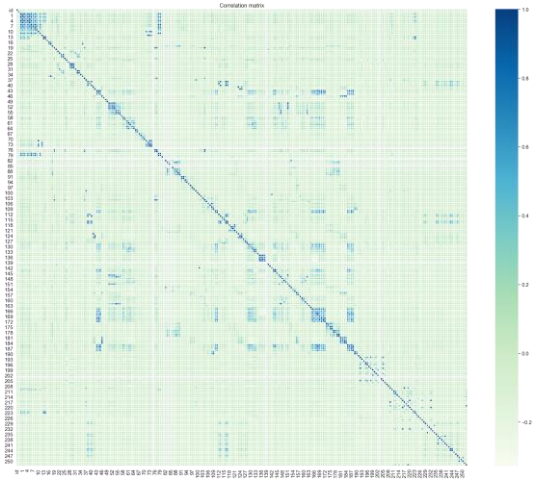
# Влияние признаков на работу модели

- 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.



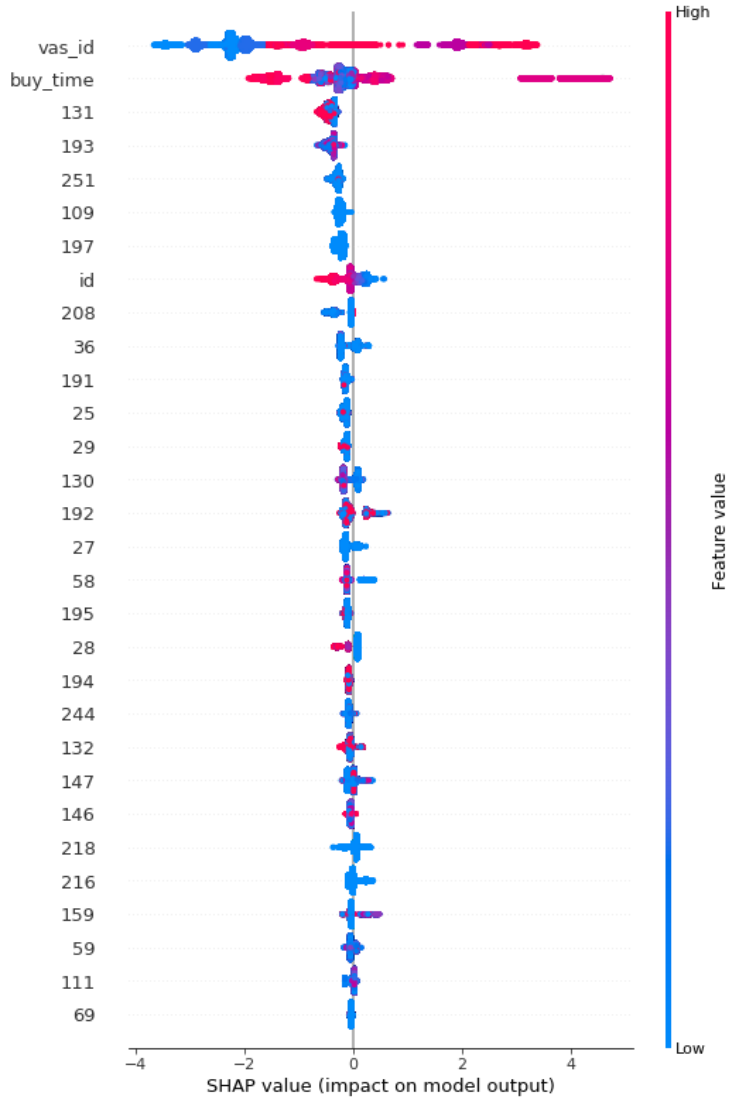



# Анализ признаков на работу модели

- **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