ИТОГОВЫЙ ПРОЕКТ

Курс: “Профессия Data Scientist PRO”
Часть 1 «Введение в Data Science»
Специализация: “Machine Learning”

# Модель машинного обучения совершения целевого действия сервиса «Сберавтоподписка»

Луцевич Анна

## Цель:

Разработка и внедрение модели предсказания целевого действия (конверсии в действие) сервиса «Сберавтоподписка»

## Исходные данные:

Две исходных выборки активности клиентов на сайте сервиса за 8 мес. (май—декабрь 2021г.)

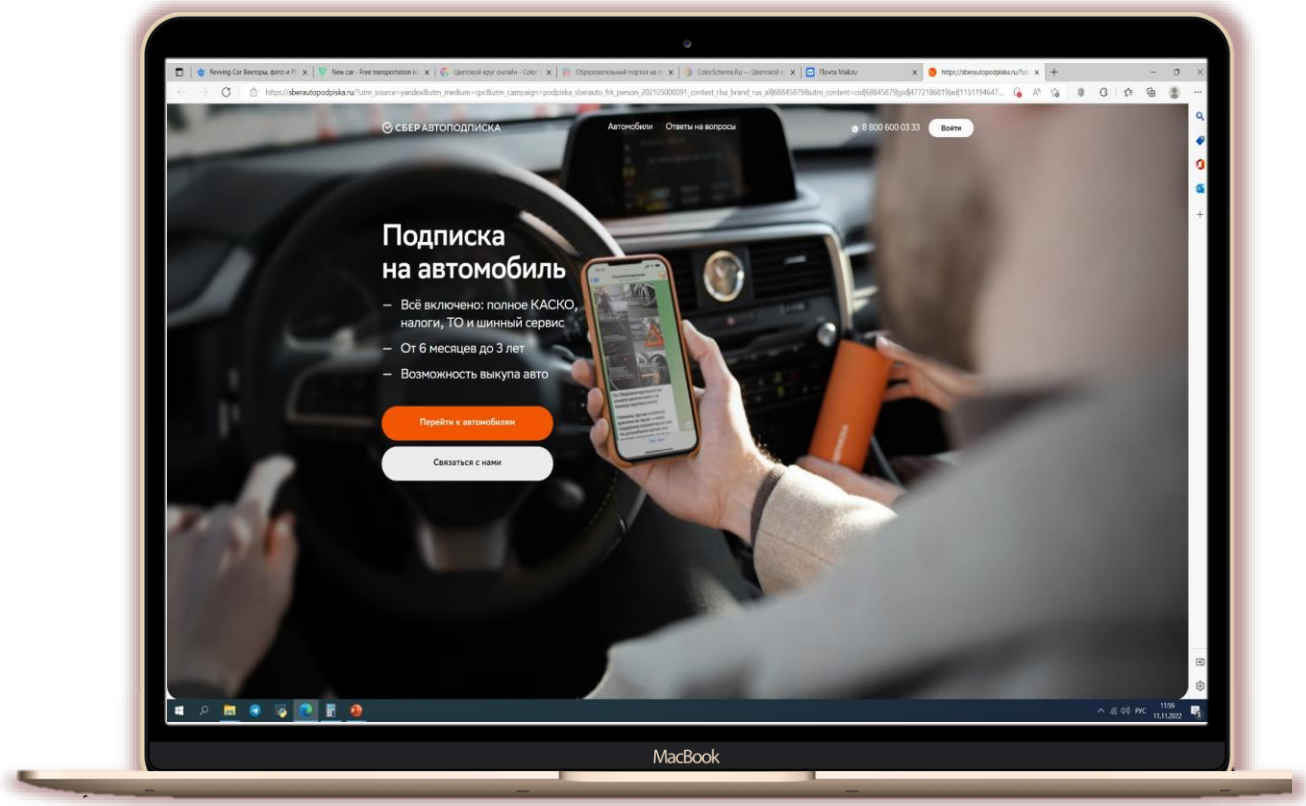
ga_sessions.pkl — характеристики визитов посетителей

ga_hits.pkl — действия, проведенные клиентами в каждый из визитов

## Алгоритм выполнения:

По технологии CRISP-DM:

- Анализ, подготовка данных,
- Тренировка модели машинного обучения
- Разработка модели предсказания совершения целевого действия (значение ROC-AUC ~ 0.65)
- Внедрение в реальный процесс модели, оборачивание в промышленный код с обеспечением их стабильности и качества



## ga_sessions.pkl:

строк > 1, 86 млн.
столбцов 18

основная рабочая база, к которой добавлена конвертация визита в целевое действие из файла ga_hits.pkl (признак CR), а также признаки марок модели авто

## ga_hits.pkl :

строк > 15,7 млн.
столбцов 11

база используется для определения величины CR на визит и для исследования марок автомобилей

## Особенности данных в выборках:

- Все данные — строковые (кроме даты и времени)
- Нет числовых данных
- Часть информации зашифрована, без ключей их раскрытия
- Много пропущенных данных
- Много нулевых данных
- Много признаков

## Выделение целевой переменной:

- Конвертация трафика в целевое действие в базе *ga_hits* (события типа «Оставить заявку» и «Заказать звонок»)

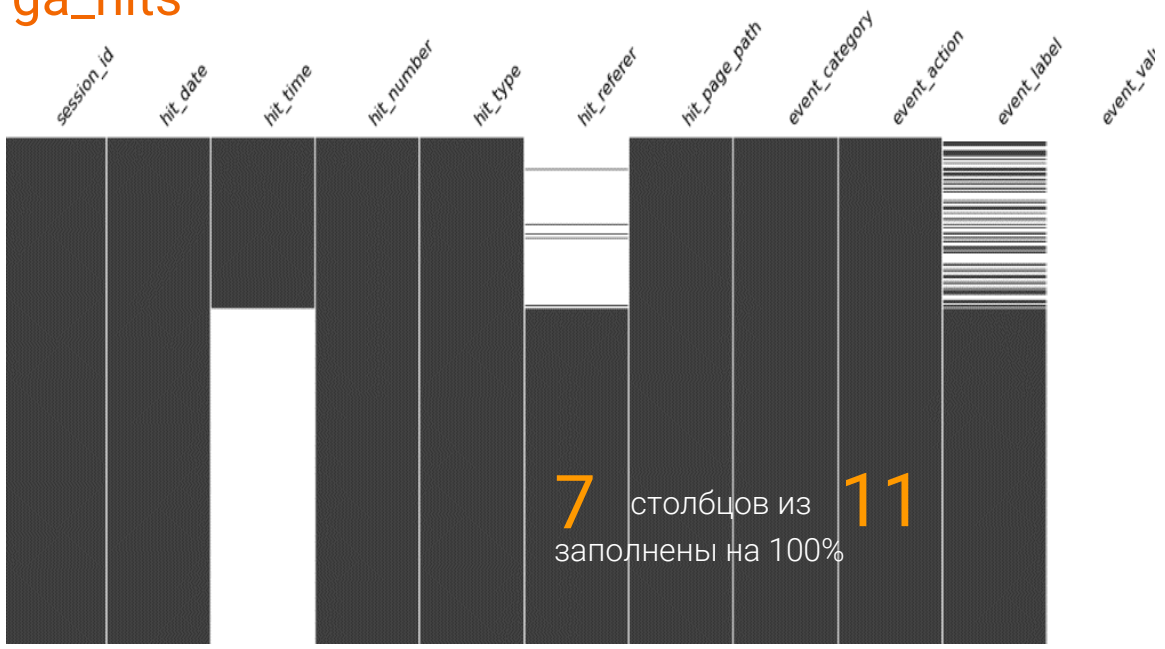
## Дубликаты:

- Дубликаты в обеих базах не установлены

## Выбросы:

- Выбросы и аномалии в данных не установлены

## ga_hits



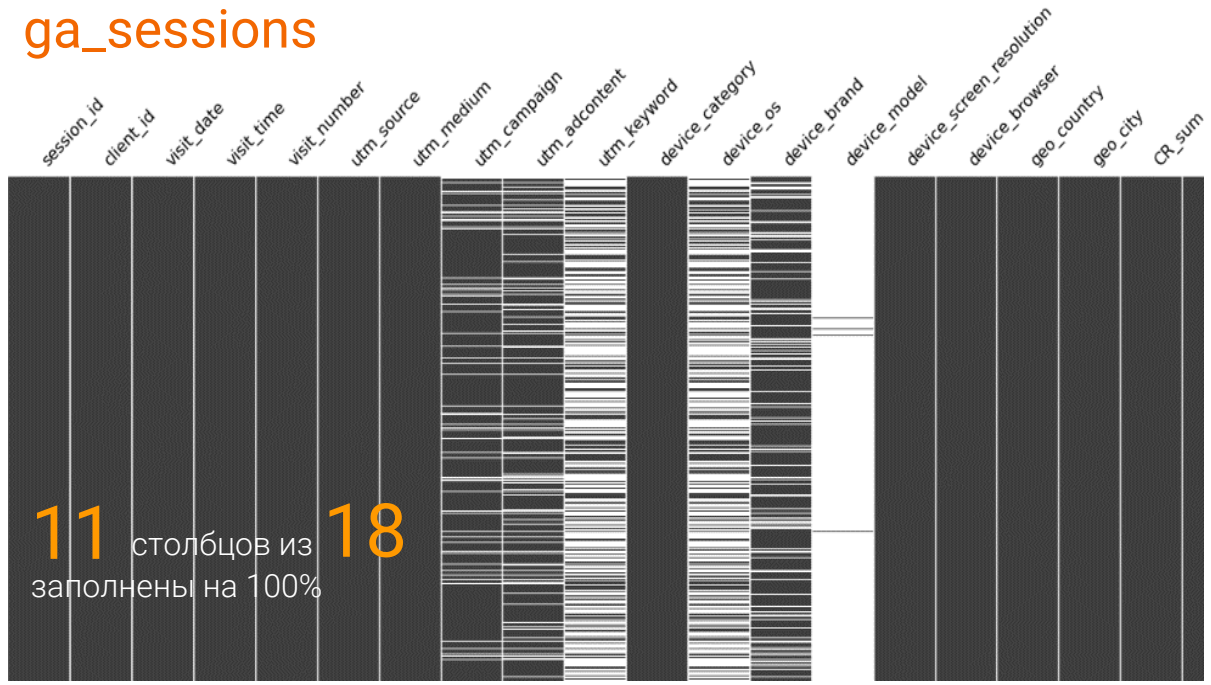

- 4 колонки не заполнены полностью — от 24 до 100% данных отсутствуют
- колонка ‘event_value’ не заполнена совсем

**Сделано:**
- удалены 4 столбца незаполненные >24%
- сформирован столбец CR: целевое действие
- сформирован столбец с маркой автомобиля (148 значений в 3 508 732 строках, т.е. в 22,3% всей выборки)

**Особенности:**
- установлена существенная зависимость конверсии от марки авто, категориальная переменная сгруппирована по сессиям и стандартизована

**Итоги:**
- все значения таблицы заполнены на 100%, без потери количества строк

## ga_sessions



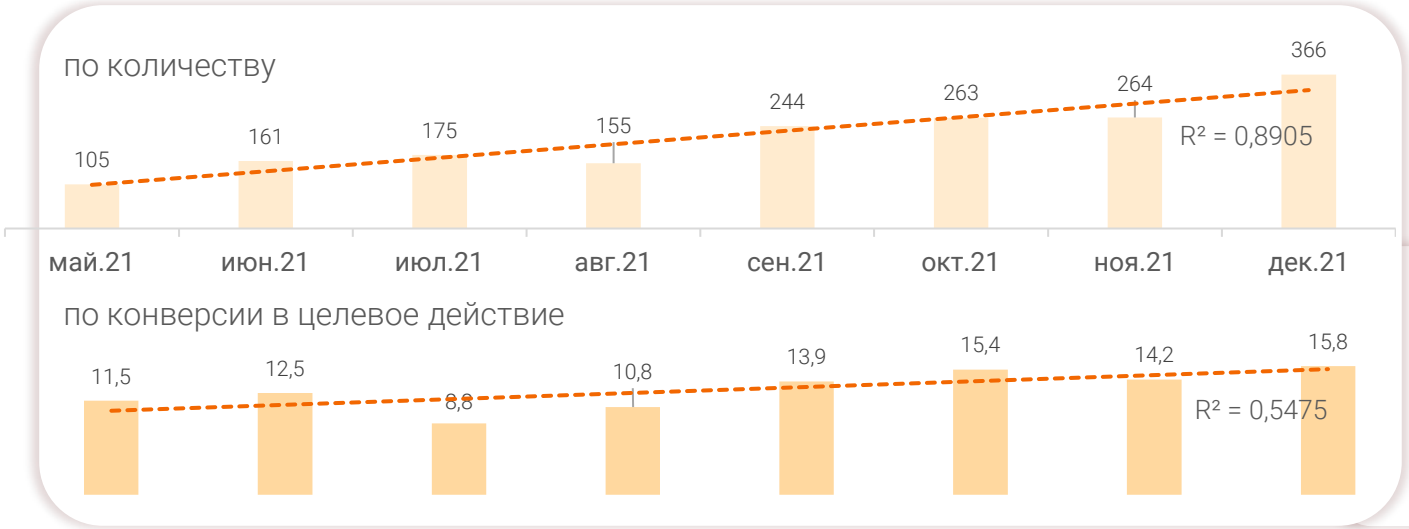

- 7 колонок не заполнены полностью — от 0,5 до 99% данных отсутствуют
- 5 колонок имеют нулевые значения — от 0,006 до 16%

**Сделано:**
- добавлен столбец CR, кол-во строк уменьшилось на 6,7%
- удалены 3 колонки, незаполненные более 58%
- отработаны 4 колонки с отсутствующими данными и 5 колонок с нулевыми значениями (от 0,01% до 16%)

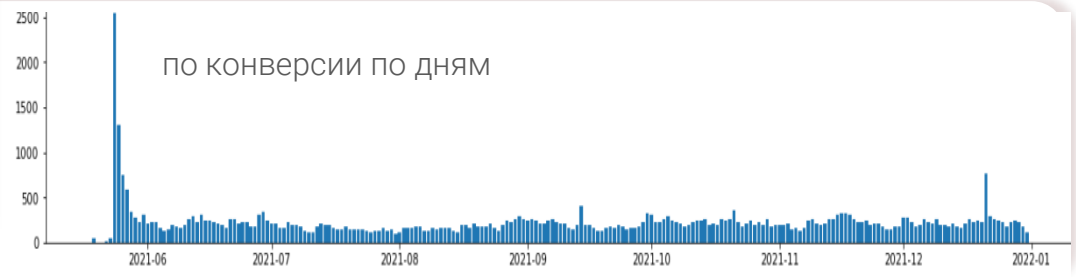
**Особенности:**
- замены отсутствующих и нулевых значений на моды
- 2 столбца – на моды столбцов, 6 столбцов на моды, полученные при группировке

**Итоги:**
- проведено заполнение 8 столбцов, все значения таблицы заполнены на 100%, без потери количества строк
- дальнейшие “докрутки” таблицы нецелесообразны

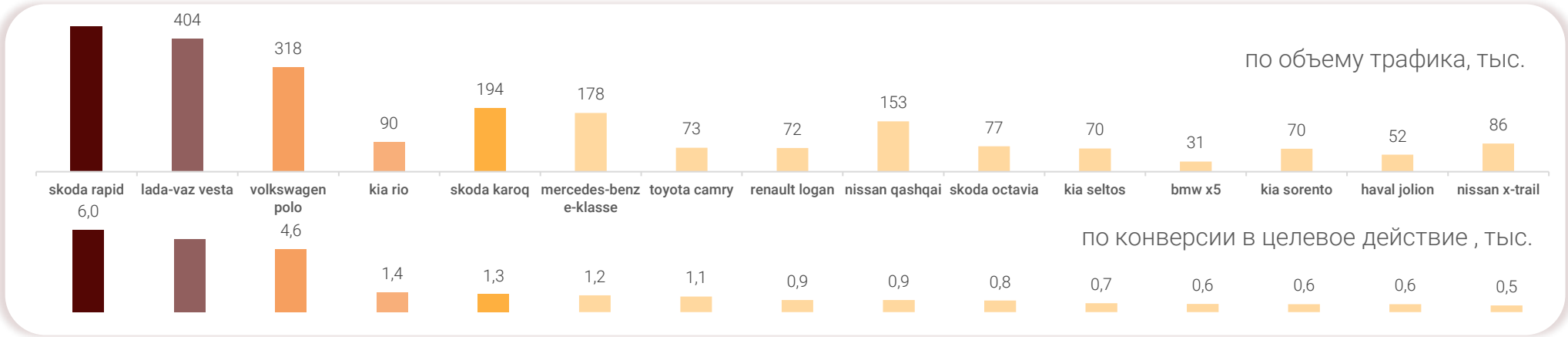
# Исследование распределения интереса клиентов во времени




- По динамике виден практически непрерывный тренд роста общего количества визитов клиентов, но при этом тренд конверсии выражен меньше
- На графике по дням явно видны дни пиков конверсии в результат
- Целесообразно исследовать рекламные кампании в периоды снижения и роста как количества визитов, так и конверсии




# Исследование лидеров по маркам автомобилей




- Skoda, Lada, Volkswagen polo — авто эконом класса имеют наибольший спрос
- В пятерку лидеров также попал mercedes-benz e-klasse, т.е. авто среднего класса стоимости
- Среди лидеров нет ни одной марки авто высокого класса стоимости

# Витрина данных (на базе ga_sessions + CR + (модели авто))

## ✓ Сформированы дополнительные столбцы:

**'org_traffic'**
- информация по органическому трафику

**'advertising_social_NW'**
- информация по рекламе в социальных сетях

**'city_of_presence'**
- информация по городам присутствия

**'CR-result'**
- бинарный признак конверсии в целевое действие

**"month', 'dayofweek**
- месяц и день недели из столбца даты

## ✓ Переформированы категориальные столбцы:

Т.к. лидеры конверсии в основном совпадают с лидерами количества, то при формировании новых признаков взяли лидеров количества

'device_screen_resolution' – 4947 значений

**'device_screen' –** 2 значения по кол-ву символов

| | |
|---|---|
| малое разрешение | 1373517 |
| большое разрешение | 358749 |

'device_browser' – 54 значения

**'device_browser_' –** 4 значения, выделяем основные браузеры и прочие

| | |
|---|---|
| Chrome | 951584 |
| Safari | 436705 |
| other_browser | 220212 |
| YaBrowser | 123765 |

'device_brand' – 199 значений

**'device_brand_' –** 5 значений, выделяем основные бренды и прочие

| | |
|---|---|
| Apple | 867121 |
| Samsung | 311641 |
| Xiaomi | 269251 |
| Huawei | 173828 |
| other_bran | 110425 |

'utm_campaign' – 487 значений

**'utm_campaign_'–** 5 значений, выделяем основные кампании и прочие

| | |
|---|---|
| LTuZkdKfxRGVceoWkVyg | 618120 |
| other_campaign | 424512 |
| LEoPHuyFvzoNfnzGgfcd | 321404 |
| FTjNLDyTrXaWYgZymFkV | 234976 |
| gecBYcKZCPMcVYdSSzKP | 133254 |

'utm_source' – 280 значений

**'utm_source_' –** 6 значений, выделяем основные каналы привлечения и прочие

| | |
|---|---|
| ZpYIoDJMcFzVoPFsHGJL | 552631 |
| other_source | 456078 |
| fDLlAcSmythWSCVMvqvL | 277060 |
| kjsLglQLzykiRbcDiGcD | 245178 |
| BHcvLf0aCWvWTykYqHVe | 110963 |
| bByPQxmDaMXgpHeypKSM | 90356 |

geo_city' – 2365 значений

**geo_city_' –** 11 значений, выделяем основные города и прочие

| | |
|---|---|
| Moscow | 817754 |
| other_city | 444710 |
| Saint Petersburg | 278402 |
| Yekaterinburg | 33555 |
| Krasnodar | 30260 |
| Kazan | 27689 |
| Samara | 23433 |
| Nizhny Novgorod | 20782 |
| Ufa | 20283 |
| Novosibirsk | 20115 |
| Krasnoyarsk | 15283 |

## ✓ Сформирован дополнительный кат. признак:

марки и модели авто – 146 значений → т.к. конверсия между целевой переменной и марками авто близка к 1, оставляем все марки авто без изменений

## ✓ Не учитываем при моделировании:

'geo_country' – 158 значений → т.к. Russia занимает 95% в выборке

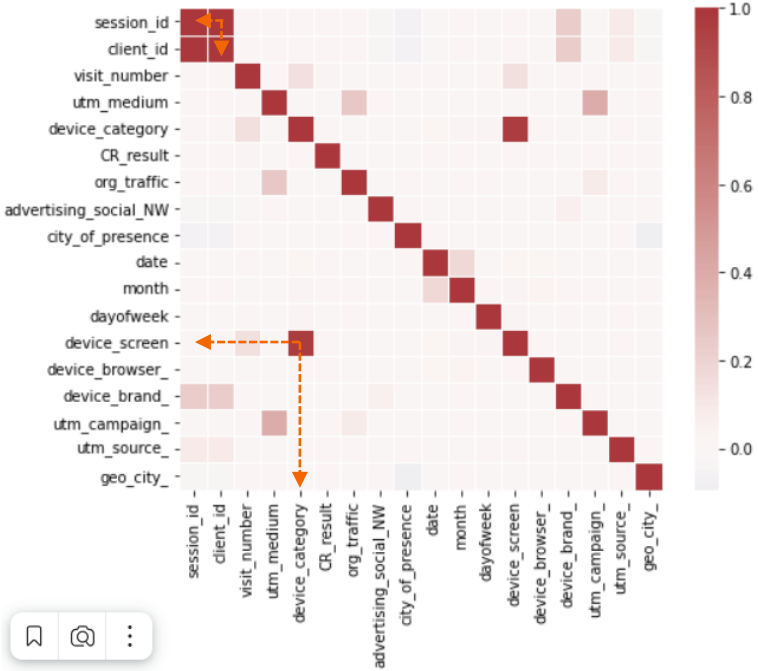
'utm_adcontent' – 281 значенит → т.к. одно из значений занимает 90% всей выборки

# Витрина данных (на базе ga_sessions + CR + (модели авто))

## ✓ Исследование корреляции признаков:

Корреляция близкая к 1 между ‘session_id’ и ‘user_id’, учитывая аналитику от сессии — удаляем user_id

Также близкая к 1 корреляция между ‘device_category’ и ‘device_screen’, удалим разрешение экрана




## ✓ Состав датасета перед моделированием:

- session_id: идентификатор сесии;
- device_browser__*: браузеры;
- device_category_*: категории девайсов;
- device_brand__*: бренды девайсов;
- utm_source__*: каналы привлечения;
- utm_medium_*: типы привлечения;
- utm_campaign_*: рекламные кампании;
- geo_city_*: города;
- city_of_presence_*: города присутствия;
- advertising_social_NW_*: социальная и иная реклама;
- org_traffic_*: вид трафика;
- visit_number_std: номер визита (после стандартизации);
- month_std: месяц (после стандартизации);
- dayofweek_std: день недели (после стандартизации);
- CR_result: конверсия в действие;
- `model_auto_*`: модели авто (добавлены в третью выборку)

## ✓ Стандартизация переменных:

Категориальных через OneHotEncoder

- 'device_browser_', 'device_category', 'device_brand_', 'utm_medium', 'utm_source_', 'utm_campaign_', 'geo_city_', 'city_of_presence', 'advertising_social_NW', 'org_traffic'

Количественных через StandardScaler

- 'visit_number', 'month', 'dayofweek'

Итого в датасете перед моделированием — 1 732 266 строк, 99 столбцов (без моделей авто) / 245 (с моделями авто)

# Результаты обучения моделей

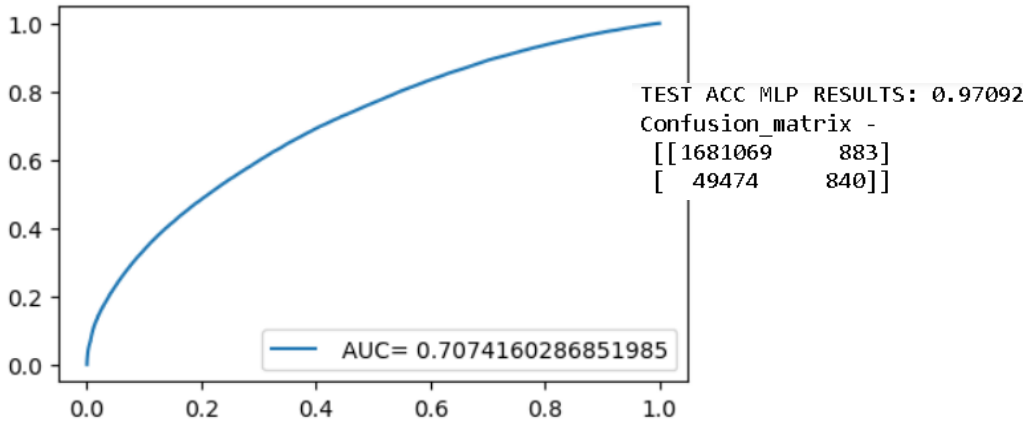
## ✓ Сравнительный анализ методов по выборкам

| | | Выборка 1 | Выборка 2 с сокращением признаков | Выборка 3 с моделями авто и сокращением признаков |
|---|---|---|---|---|
| | | 99 столбцов | 49 столбцов | 121 столбец |
| Логистическая регрессия | accuracy_test | 97,10% | 97,10% | 97,10% |
| | AUC | 0,59 | 0,5893 | 0,6742 |
| | время обучения | 1min 39s | 28,9 s | 24,5 s |
| Случайный лес | accuracy_test | 96,90% | 96,70% | 96,80% |
| | AUC | 0,5478 | 0,524 | 0,6204 |
| | время обучения | 10min 19s | 7min 23s | 9min 40s |
| Многослойный персептрон | accuracy_test | 97,10% | 97,10% | 97,10% |
| | AUC | 0,5952 | 0,5965 | 0,6951 |
| | время обучения | 2min 12s | 2min 6s | 3min 42s |

- Кросс-валидация ни по одному из методов ни одной из выборок не показала ухудшения характеристик (нет признаков переобучения моделей), поэтому в отчете не указаны ее результаты
- Сокращение признаков через их ранжирование методом RFE
- Поведения методов примерно идентичны по выборкам, поэтому их характеристики были взяты по умолчанию (усложнение только значительно увеличивало время обучения)

Лучший результат по ROC-AUC, с превышением заданной границы — по методу многослойного персептрона

## ✓ Итоги обучения модели по всей выборке:




Несмотря на высокие в целом показатели точности модели, предсказанная конверсия ниже реальной почти в 30 раз! Модель требует дальнейшей доработки

# Применение модели

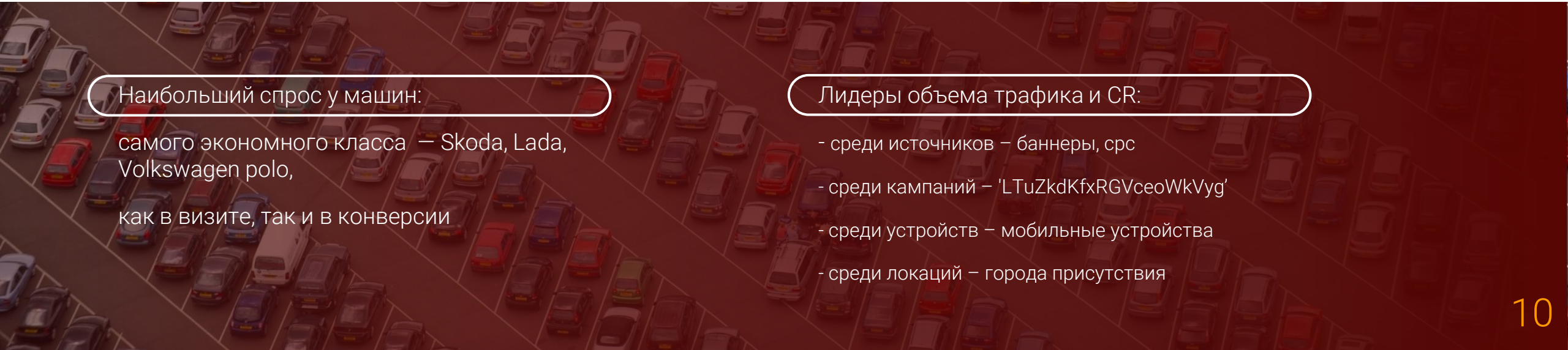
## ✓ Модель сериализована в файл: 'ML_proect_auto_pod_MLP.pickle'

## ✓ Состав датасета для применения модели:

- session_id: идентификатор сесии;
- device_browser__*: браузеры: 'device_browser__Chrome', 'device_browser__Safari', 'device_browser__YaBrowser', 'device_browser__other_browser';
- device_category_*: категории девайсов: 'device_category_desktop', 'device_category_mobile', 'device_category_tablet;
- device_brand__*: бренды девайсов: 'device_brand__Apple', 'device_brand__Huawei', 'device_brand__Samsung', 'device_brand__Xiaomi', 'device_brand__other_brand;
- utm_source__*: каналы привлечения: 'utm_source__BHcvLfOaCWvWTykYqHVe', 'utm_source__ZpYIoDJMcFzVoPFsHGJL', 'utm_source__bByPQxmDaMXgpHeypKSM', 'utm_source__fDLlAcSmythWSCVMvqvL', 'utm_source__kjsLglQLzykiRbcDiGcD', 'utm_source__other_source';
- utm_medium_*: типы привлечения: 'utm_medium_app', 'utm_medium_blogger_channel', 'utm_medium_blogger_header', 'utm_medium_blogger_stories', 'utm_medium_clicks', 'utm_medium_cpa', 'utm_medium_email', 'utm_medium_organic', 'utm_medium_other', 'utm_medium_outlook', 'utm_medium_post', 'utm_medium_referral', 'utm_medium_smartbanner', 'utm_medium_smm', 'utm_medium_sms', 'utm_medium_stories', 'utm_medium_tg;
- utm_campaign_*: рекламные кампании: 'utm_campaign__FTjNLDyTrXaWYgZymFkV', 'utm_campaign__LEoPHuyFvzoNfnzGgfcd', 'utm_campaign__LTuZkdKfxRGVceoWkVyg', 'utm_campaign__gecBYcKZCPMcVYdSSzKP', 'utm_campaign__other_campaign';
- geo_city_*: города: 'geo_city__Krasnodar', 'geo_city__Krasnoyarsk', 'geo_city__Moscow', 'geo_city__Nizhny Novgorod', 'geo_city__Novosibirsk', 'geo_city__Saint Petersburg', 'geo_city__Samara', 'geo_city__Ufa', 'geo_city__Yekaterinburg', 'geo_city__other_city', ;
- city_of_presence_*: города присутствия: 'city_of_presence_Москва+Санкт-П', 'city_of_presence_другие города';
- advertising_social_NW_*: социальная и иная реклама: 'advertising_social_NW_иная реклама', 'advertising_social_NW_реклама в соц.сетях';
- org_traffic_*: вид трафика: 'org_traffic_органический трафик', 'org_traffic_платный трафик';
- visit_number_std: номер визита (после стандартизации);
- month_std: месяц (после стандартизации);
- dayofweek_std: день недели (после стандартизации);
- CR_result: конверсия в действие;
- model_auto_*: модели авто: 'x0_acura', 'x0_asia', 'x0_audi a3', 'x0_audi q5', 'x0_bmw 3-serii', 'x0_bmw 7-serii', 'x0_bmw x3', 'x0_buick', 'x0_dacia', 'x0_fiat', 'x0_gac', 'x0_great-wall', 'x0_haval', 'x0_hawtai', 'x0_honda', 'x0_honda civic-type-r', 'x0_hummer', 'x0_hyundai', 'x0_kia', 'x0_kia k5', 'x0_kia rio', 'x0_kia seltos', 'x0_kia sorento', 'x0_kia sportage', 'x0_lamborghini', 'x0_land-rover', 'x0_land-rover range-rover-velar', 'x0_lexus', 'x0_lincoln', 'x0_mazda', 'x0_mercedes-benz', 'x0_mercedes-benz c-klasse', 'x0_mercedes-benz cls-klasse', 'x0_mercedes-benz e-klasse', 'x0_mercedes-benz gle', 'x0_mercedes-benz gls-klasse', 'x0_mercedes-benz v-klasse', 'x0_mini', 'x0_mini hatch', 'x0_nissan qashqai', 'x0_porsche macan', 'x0_ravon', 'x0_renault', 'x0_renault duster', 'x0_rolls-royce', 'x0_seat', 'x0_skoda', 'x0_skoda kodiaq', 'x0_skoda octavia', 'x0_skoda rapid', 'x0_suzuki', 'x0_toyota', 'x0_toyota alphard', 'x0_toyota camry', 'x0_toyota corolla', 'x0_toyota land-cruiser-prado', 'x0_uaz', 'x0_volkswagen passat-cc', 'x0_volkswagen polo', 'x0_volkswagen tiguan', 'x0_volkswagen touareg'

# Итоговые выводы проекта

- ✓ лучшие результаты моделирования показала модель по методу многослойного персептрона
- ✓ для повышения точности предсказания необходимы признаки моделей авто и сокращение числа признаков
- ✓ характеристики модели – acc = 97,1%, ROC-AUC = 0,7074
- ✓ предсказание конверсии почти в 30 раз меньше реальным показателям
- ✓ модель можно дорабатывать для точности прогноза конверсии



**Наибольший спрос у машин:**

самого экономного класса — Skoda, Lada, Volkswagen polo,

как в визите, так и в конверсии

**Лидеры объема трафика и CR:**

- среди источников – баннеры, cpc
- среди кампаний – 'LTuZkdKfxRGVceoWkVyg'
- среди устройств – мобильные устройства
- среди локаций – города присутствия