**Задача**

Сеть фитнес-центров «Культурист-датасаентист» разрабатывает стратегию взаимодействия с клиентами на основе аналитических данных.

Распространённая проблема фитнес-клубов и других сервисов — отток клиентов. Как понять, что клиент больше не с вами? Можно записать в отток тех, кто попросил закрыть договор или удалил аккаунт. Однако клиенты не всегда уходят демонстративно: чаще перестают пользоваться сервисом тихо.

Индикаторы оттока зависят от специфики отрасли. Когда пользователь редко, но стабильно закупается в интернет-магазине — не похоже, что он «отвалился». А вот если две недели не заходит на канал с ежедневно обновляемым контентом, дела плохи: подписчик заскучал и, кажется, оставил вас.

Для фитнес-центра можно считать, что клиент попал в отток, если за последний месяц ни разу не посетил спортзал. Конечно, не исключено, что он уехал на Бали и по приезде обязательно продолжит ходить на фитнес. Однако чаще бывает наоборот. Если клиент начал новую жизнь с понедельника, немного походил в спортзал, а потом пропал — скорее всего, он не вернётся.

Чтобы бороться с оттоком, отдел по работе с клиентами «Культуриста-датасаентиста» перевёл в электронный вид множество клиентских анкет. Ваша задача — провести анализ и подготовить план действий по удержанию клиентов.

А именно:

* научиться прогнозировать вероятность оттока (на уровне следующего месяца) для каждого клиента;
* сформировать типичные портреты клиентов: выделить несколько наиболее ярких групп и охарактеризовать их основные свойства;
* проанализировать основные признаки, наиболее сильно влияющие на отток;
* сформулировать основные выводы и разработать рекомендации по повышению качества работы с клиентами:
  + 1) выделить целевые группы клиентов;
  + 2) предложить меры по снижению оттока;
  + 3) определить другие особенности взаимодействия с клиентами.

**Инструкция по выполнению проекта**

**Шаг 1. Загрузите данные**

«Культурист-датасаентист» предоставил сведения в csv-файлах. Заказчик подготовил данные, которые содержат данные на месяц до оттока и факт оттока на определённый месяц. Набор данных включает следующие поля:

* Данные клиента за предыдущий до проверки факта оттока месяц:
  + 'gender' — пол;
  + 'Near\_Location' — проживание или работа в районе, где находится фитнес-центр;
  + 'Partner' — сотрудник компании-партнёра клуба (сотрудничество с компаниями, чьи сотрудники могут получать скидки на абонемент — в таком случае фитнес-центр хранит информацию о работодателе клиента);
  + Promo\_friends — факт первоначальной записи в рамках акции «приведи друга» (использовал промо-код от знакомого при оплате первого абонемента);
  + 'Phone' — наличие контактного телефона;
  + 'Age' — возраст;
  + 'Lifetime' — время с момента первого обращения в фитнес-центр (в месяцах).
* Информация на основе журнала посещений, покупок и информация о текущем статусе абонемента клиента:
  + 'Contract\_period' — длительность текущего действующего абонемента (месяц, 6 месяцев, год);
  + 'Month\_to\_end\_contract' — срок до окончания текущего действующего абонемента (в месяцах);
  + 'Group\_visits' — факт посещения групповых занятий;
  + 'Avg\_class\_frequency\_total' — средняя частота посещений в неделю за все время с начала действия абонемента;
  + 'Avg\_class\_frequency\_current\_month' — средняя частота посещений в неделю за предыдущий месяц;
  + 'Avg\_additional\_charges\_total' — суммарная выручка от других услуг фитнес-центра: кафе, спорттовары, косметический и массажный салон.
* 'Churn' — факт оттока в текущем месяце.

Путь к файлу: /datasets/gym\_churn.csv. [Скачать датасет](https://code.s3.yandex.net/datasets/gym_churn.csv)

**Шаг 2. Проведите исследовательский анализ данных (EDA)**

* Посмотрите на датасет: есть ли в нем отсутствующие признаки, изучите средние значения и стандартные отклонения (пригодится метод describe());
* Посмотрите на средние значения признаков в двух группах — тех, кто ушел в отток и тех, кто остался (воспользуйтесь методом groupby());
* Постройте столбчатые гистограммы и распределения признаков для тех, кто ушёл (отток) и тех, кто остался (не попали в отток);
* Постройте матрицу корреляций и отобразите её.

**Шаг 3. Постройте модель прогнозирования оттока клиентов**

Постройте модель бинарной классификации клиентов, где целевой признак — факт оттока клиента в следующем месяце:

* Разбейте данные на обучающую и валидационную выборку функцией train\_test\_split().
* Обучите модель на train-выборке двумя способами:
  + логистической регрессией,
  + случайным лесом.
* Оцените метрики accuracy, precision и recall для обеих моделей на валидационной выборке. Сравните по ним модели. Какая модель показала себя лучше на основании метрик?

Не забудьте указать параметр random\_state при разделении выборки и задании алгоритма.

**Шаг 4. Сделайте кластеризацию клиентов**

Отложите в сторону столбец с оттоком и проведите кластеризацию объектов (клиентов):

* Стандартизируйте данные.
* Постройте матрицу расстояний функцией linkage() на стандартизованной матрице признаков и нарисуйте дендрограмму. Внимание: отрисовка дендрограммы может занять время! На основании полученного графика предположите, какое количество кластеров можно выделить.
* Обучите модель кластеризации на основании алгоритма K-Means и спрогнозируйте кластеры клиентов. Договоримся за число кластеров принять n=5, чтобы ваши результаты можно было сравнивать с результатами остальных студентов. Однако, конечно, в жизни никто не скажет вам правильный ответ, и решение остаётся за вами (на основании изучения графика из предыдущего пункта).
* Посмотрите на средние значения признаков для кластеров. Можно ли сразу что-то заметить?
* Постройте распределения признаков для кластеров. Можно ли что-то заметить по ним?
* Для каждого полученного кластера посчитайте долю оттока (методом groupby()). Отличаются ли они по доле оттока? Какие кластеры склонны к оттоку, а какие — надёжны?

**Шаг 5. Сформулируйте выводы и сделайте базовые рекомендации по работе с клиентами**

Сформулируйте основные выводы и предложите рекомендации для стратегии взаимодействия с клиентами и их удержания.

Не описывайте стратегию детально: достаточно 3-4 важных принципов и примеров их реализации в виде определённых маркетинговых действий.

**Как будут проверять мой проект?**

Перед вами критерии оценки проекта. Прежде чем приступить к решению кейса, внимательно изучите их.

На что обращают внимание, когда проверяют проект:

* Как вы готовите данные к анализу?
* Как вы производите разделение данных на обучающую и валидационную выборку?
* Как вы обучаете модели?
* Как вы оцениваете метрики accuracy, precision и recall для обеих моделей?
* На основе какой выборки вы оцениваете метрики?
* Какой вывод вы делаете на основе метрик?
* Как вы стандартизируете данные перед кластеризацией?
* Как вы строите матрицу расстояний и визуализируете её с помощью дендрограммы?
* Как вы обучаете модель кластеризации?
* Какие выводы вы делаете на основе средних значений признаков для кластеров?
* Как вы определеяете какие кластеры склонны к оттоку? Какие выводы делаете на основе определения склонности кластеров к оттоку?
* Какие рекомендации предлагаете для улучшения стратегии удержания клиентов?