**ДИПЛОМНАЯ РАБОТА**

по курсу «Аналитик данных»

**«Прогноз оттока посетителей из фитнес-клуба»**

Чекуновой Анны Сергеевны



Ноутбук с кодом доступен по ссылке: <https://github.com/annachekunova/analyst_repository/blob/main/Diploma.ipynb>

Постановка задачи.

* Провести EDA представленного датасета.
* Построить модели классификации для прогноза оттока клиентов из фитнес-клуба на примере следующего месяца.
* Построить модели кластеризации для выделения групп среди клиентов.
* Проанализировать основные признаки, влияющие на отток.
* Сформулировать меры по снижению оттока.

Описание данных:

* 'Churn' — факт оттока в текущем месяце (целевая переменная);

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).

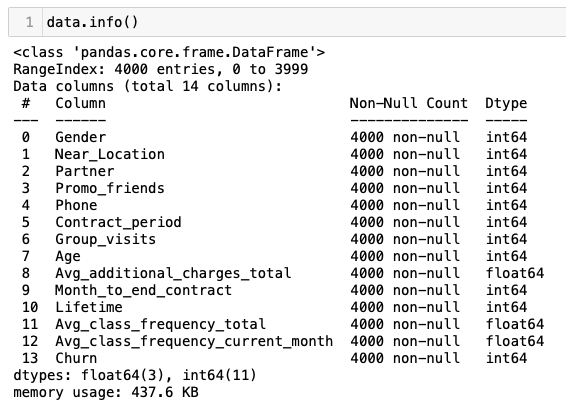
Информация на основе журнала посещений, покупок и информация о текущем статусе абонемента клиента:

* 'Contract\_period' — длительность текущего действующего абонемента (месяц, 3 месяца, 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, спорт-товары, косметический и массажный салон.

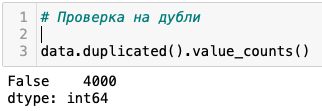
**Exploratory Data Analysis (EDA)** или разведочный анализ данных. EDA играет важнейшую роль после получения набора данных и ставит своей целью выяснить, как с ним работать и получить требуемый результат.

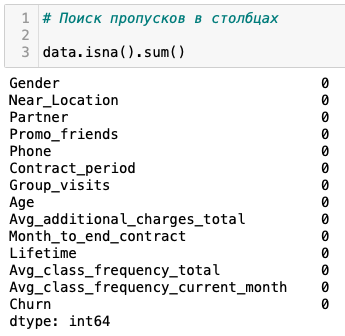
Начать необходимо с импорта необходимых библиотек и загрузки данных.

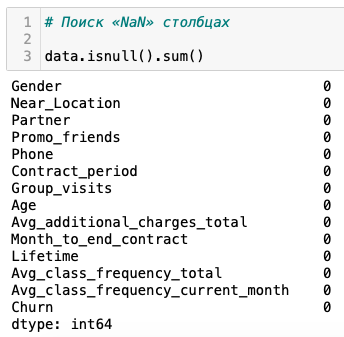
Как видим, в датасете 4000 записей и 14 столбцов (3 поля содержат данные типа float, остальные - int).



Далее необходимо проверить датасет на дубли, пропуски, NaN.







Выводы:

Дублей, пропусков, NaN нет. Это хорошо, значит данные чистые.

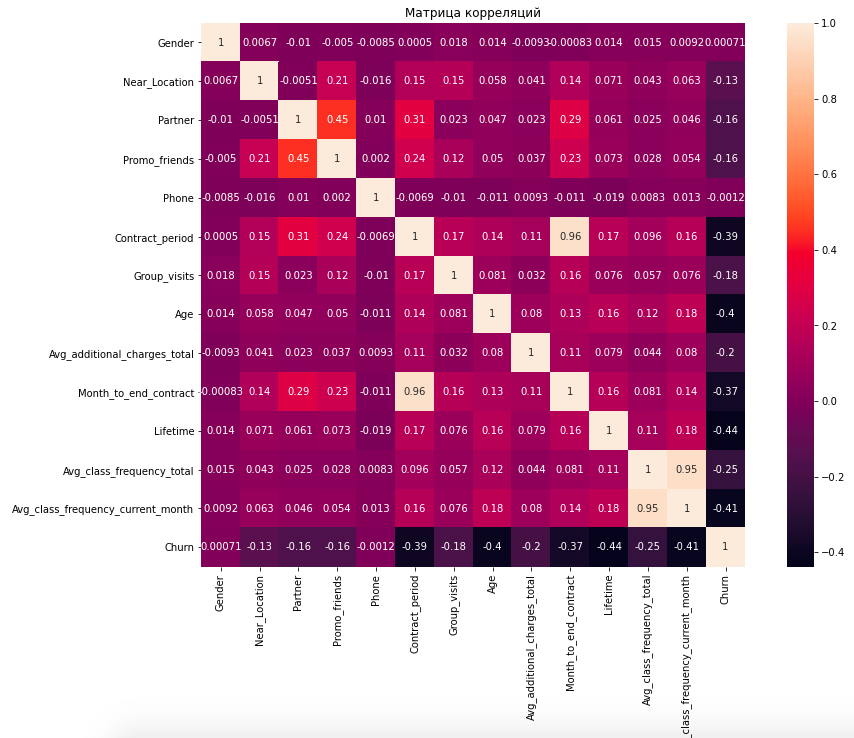
Далее, выводим базовые статистики и проводим исследование распределений по признакам. Из приведенных визуализаций в ноутбуке можно наблюдать следующее:

1. В описании датасета явно не был указан признак пола (мужчина/женщина). При построении столбчатой диаграммы распределения по полу оказалось, что распределение приблизительно поровну 50 на 50. В противном случае, пришлось бы запрашивать дополнительные данные у заказчика.
2. Оценка распределения клиентов по признаку “Near\_Location” показывает, что почти 85% клиентов живут или работают в районе расположения фитнес-центра.
3. Оценка распределения клиентов по признаку “Partner” показывает, что чуть меньше половины клиентов (49%) являются сотрудниками компаний-партнеров фитнес-клуба.
4. Анализ количества клиентов, пришедших по акции “Promo\_friends”, составляет 31%.
5. У 90% клиентов известен номер телефона, что в дальнейшем позволит нам сообщать почти всем клиентам об акциях и спецпредложениях, а также напоминать о тренировках, тем самым, влияя на удержание.
6. Оценка распределения клиентов по признаку “Contract\_period” показывает, что длительность действующего абонемента у 55% клиентов составляет 1 месяц, у 21% - 6 месяцев и у 24% - 12 месяцев.
7. Из всех клиентов фитнес-центра менее половины (41%) посещают групповые занятия.
8. Средний возраст клиентов 29 лет.
9. Средний срок действия абонемента составляет 3,7 месяцев;
10. Среднее расходы на клиента за дополнительные услуги фитнес-центра составляют 147 единиц.
11. Средний срок до окончания текущего действующего абонемента составляет  5 месяцев.
12. Средняя частота посещений в неделю за все время с начала действия абонемента - 1,87 раз в неделю.
13. Средняя частота посещений в неделю за предыдущий месяц - 1,77 раз в неделю.
14. Доля оттока клиентов составила 27%.

После анализа целиком датафрейма делаем выводы:

* На отток могут влиять посещение групповых занятий, срок абонемента, частота посещений, расположение фитнес-клуба..
* На лояльность клиентов влияют наличие промокода от друга,  посещение групповых занятий, партнерские программы.

Чтобы обнаружить точные взаимосвзяи (корреляции) признаков, строю scatter plot matrix и тепловую карту (матрица корреляции), по которым визуально можно оценить, есть ли связь между признаками и на сколько она сильная.

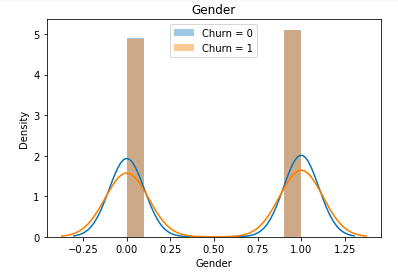


Корреляция средней частоты посещений в неделю за все время с начала действия абонемента (Avg\_class\_frequency\_total) и средней частоты посещений в неделю за предыдущий месяц (Avg\_class\_frequency\_current\_month) не информативна, поскольку очевидна и не дает данных для гипотез.

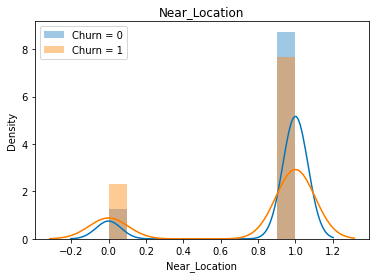
Визуальный анализ тепловой карты не показал очевидной зависимости факта оттока от признаков. Между признаками прослеживается слабая линейная связь. Корреляционная зависимость прослеживается у следующих признаков:

* Month\_to\_end\_contract и Contract\_period (сильная зависисмоть - 0,96);
* Promo\_friends и Partner (слабая зависимость - 0,45).

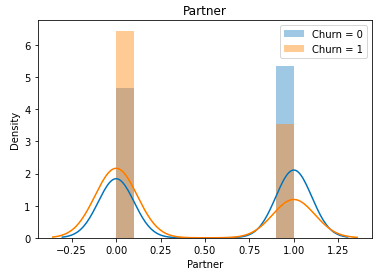
**Далее, визуально сравниваем показатели между двумя группами клиентов: в оттоке и оставшиеся.**



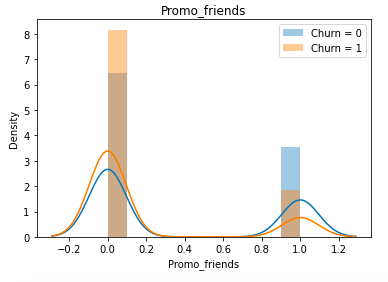
Распределение клиентов по полу среди оставшихся и в оттоке не отличается.



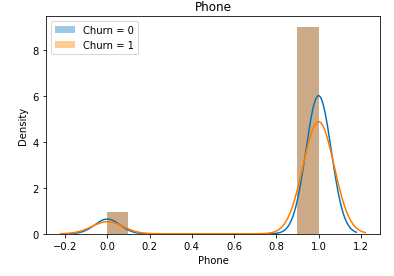
В отток попадает больше клиентов, живущих или работающих в другом районе.



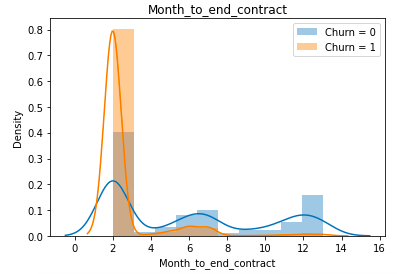
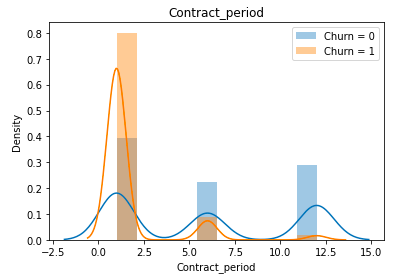
Клиенты компаний-партнеров фитнес-центра реже попадают в отток.

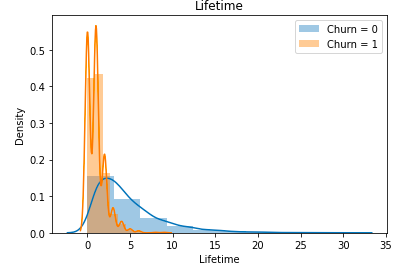


Клиенты, пришедшие по промокодам от знакомых, в два раза реже попадают в отток.

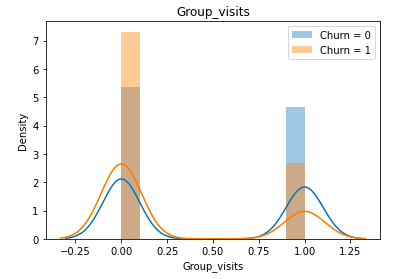


Наличие контактного телефона в базе не влияет на отток.

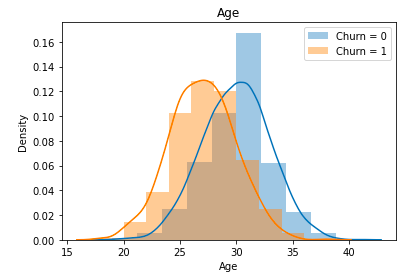




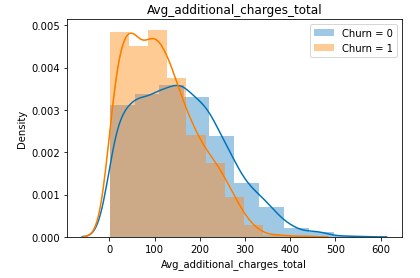
Большая часть клиентов в оттоке с длительностью абонемента 1-2 месяца. Следовательно, срок окончания абонемента будет выше у клиентов, которые остались.



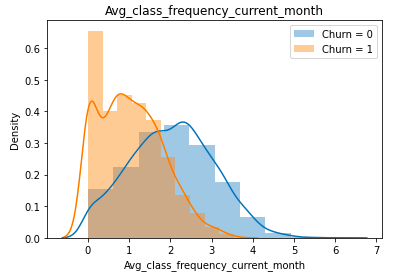
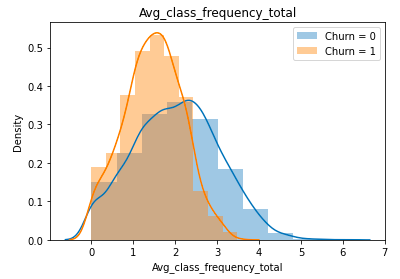
Клиенты, попавшие в отток, реже посещали групповые занятия.



Средний возраст клиентов, попавших в отток, 26 лет. Средний возраст оставшихся клиентов - 31 год.



Средняя суммарная выручка от дополнительных услуг выше от клиентов, которые остались.



Средняя частота посещений в неделю больше у оставшихся клиентов.

**Построение модели прогноза оттока в следующем месяце**

Необходимо апробировать несколько моделей классификации, для оценки производительности посмотреть матрицу ошибок, потом к каждой модели применить перебор параметров через Grid SearchCV, найти лучший результат.

В начале надо разбить данные на обучающую (train) и валидационную (test) выборки.

Обучить модель на train выборке логистической регрессией, случайным лесом, Decision Tree и XGBoost.

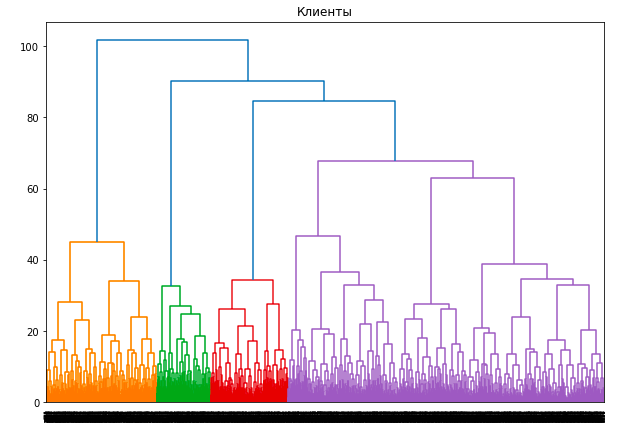
Далее, оценить все модели с помощью матрицы ошибок (confusion\_matrix), к каждой модели подбираем параметры с помощью Grid SearchCV и сравниваем модели.

Модель XGBoos показала себя лучше. Доля правильных прогнозов незначительно, но больше.

## Кластеризация пользователей

Проведем кластеризацию клиентов. Для этого сначала стандартизируем данные с помощью StandardScaler ().

Методом иерархической кластеризации построим дендограмму.



Оптимально разделить клиентов на 4 кластера.

С помощью модели кластеризации на основе алгоритма K-Means спрогнозируем кластеры клиентов.

**Описание кластеров:**

**Кластер 0:** У этих клиентов нет телефонных номеров в базе, в среднем посещают фитнес около 2 раз в неделю, средний срок действия абонемента составляет 5 месяцев. Большой процент клиентов пришли по промокоду от друзей или являются работниками партнерских организаций.

- 86% живут или работают рядом с фитнесом;

- 47% являются работниками партнерских организаций;

- 30% пришли по промокоду от друзей;

- Нет телефонных номеров в базе;

- Средний срок абонемента 4,7 месяцев;

- Групповые занятия посещают 43% клиентов;

- Средний суммарный расход на дополнительные слуги 130 единиц;

- В среднем 5 месяцев до окончания действия абонемента;

- Среднее время действия абонемента 3,9 месяцев;

- Средняя частота посещений в неделю за все время действия абонемента 1,85 раз;

- Средняя частота посещений в неделю за предыдущий месяц 1,72 раз;

- Отток 26,6%.

**Кластер 1:** Самый неуспешный кластер клиентов. Самый большой отток, реже всех посещают фитнес и групповые занятия, самый короткий срок абонемента. Меньше всего клиентов живут или работают рядом с фитнесом.

- 76% живут или работают рядом с фитнесом;

- 38% являются работниками партнерских организаций;

- 19% пришли по промокоду от друзей;

- Средний срок абонемента 1,9 месяца;

- Групповые занятия посещают 29% клиентов;

- Средний суммарный расход на дополнительные слуги 130 единиц;

- В среднем 2,7 месяца до окончания действия абонемента;

- Среднее время действия абонемента 2,2 месяца;

- Средняя частота посещений в неделю за все время действия абонемента 1,2 раза;

- Средняя частота посещений в неделю за предыдущий месяц 0,95 раз;

- Отток 55,7%.

**Кластер 2:** Клиенты этого кластера чаще всех посещают фитнес и групповые занятия. Но срок абонемента не длительный.

- 86% живут или работают рядом с фитнесом;

- 35% являются работниками партнерских организаций;

- 21% пришли по промокоду от друзей;

- Средний срок абонемента 2,4 месяца;

- Групповые занятия посещают 44,5% клиентов;

- Средняя выручка от других услуг 157,73 единиц;

- В среднем 3,1 месяца до окончания действия абонемента;

- Среднее время действия абонемента 4,7 месяцев;

- Средняя частота посещений в неделю за все время действия абонемента 2,7 месяцев;

- Средняя частота посещений в неделю за предыдущий месяц 2,7 месяцев;

- Отток 9%.

**Кластер 3:** Самый успешный кластер по оттоку. Больше всего клиентов живут или работают рядом с фитнесом, много клиентов являются работниками партнерских организаций. Больше половины клиентов пришли по промокоду от друзей. Больше всех посещают групповые занятия. Самые длинные абонементы.

- 94% живут или работают рядом с фитнесом;

- 77% являются работниками партнерских организаций;

- 57% пришли по промокоду от друзей;

- Средний срок абонемента 10,8 месяцев;

- Групповые занятия посещают 54% клиентов;

- Средняя выручка от других услуг 160,73 единиц;

- В среднем 10 месяцев до окончания действия абонемента;

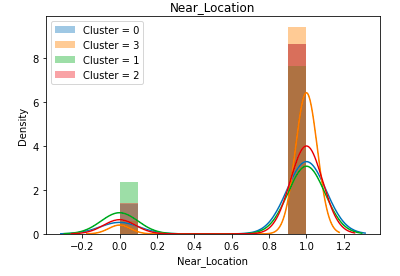
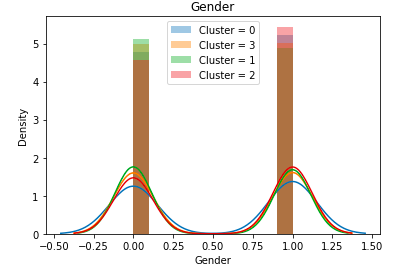
- Среднее время действия абонемента 4,7 месяцев;

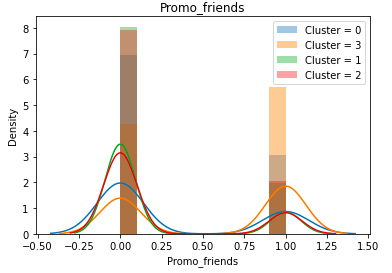
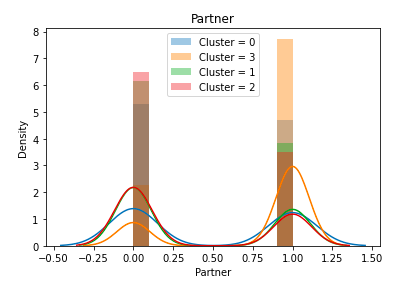
- Средняя частота посещений в неделю за все время действия абонемента 1,9 месяца;

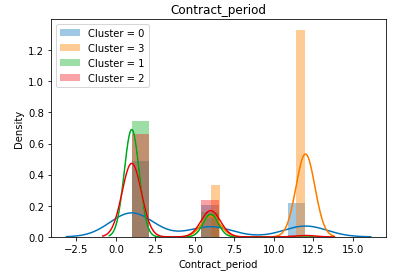
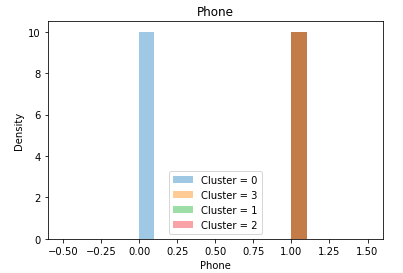
- Средняя частота посещений в неделю за предыдущий месяц 1,9 месяца;

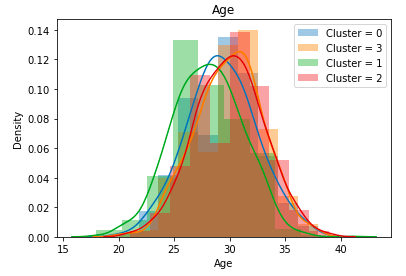
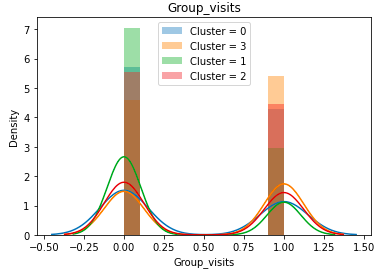
- Отток 3%.

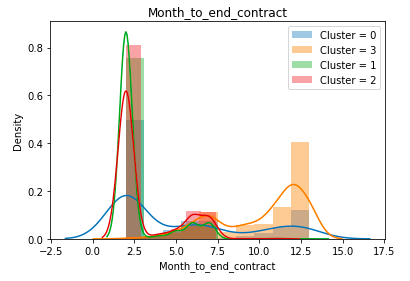
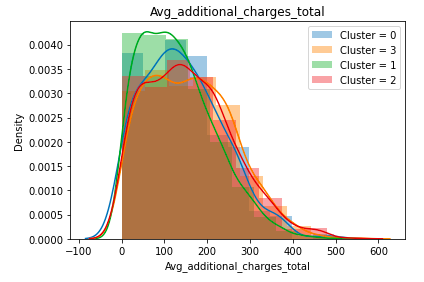
**Далее сравниваю распределения признаков между кластерами.**

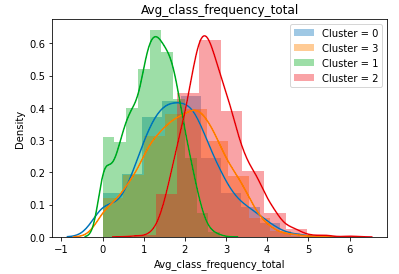
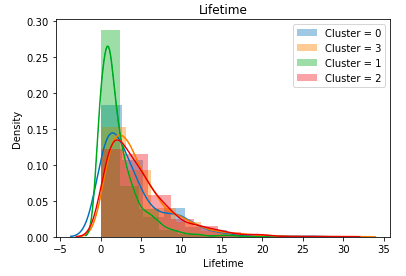


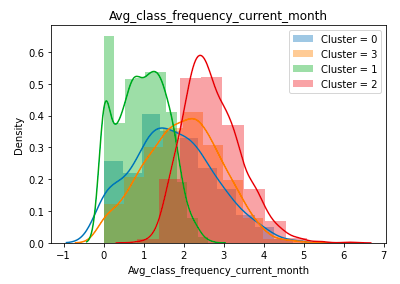












Самые весомые признаки, влияющие на отток:

- Near\_Location;

- Promo\_friends;

- Contract\_period;

- Partner.

**Рекомендации по удержанию клиентов в проблемных кластерах:**

Кластер 0 - необходимо знать номера телефонов клиентов. Предложить дополнительные скидки на групповые занятия или дополнительные услуги при втором посещении в неделю (т.к. в этом кластере средний срок посещения в неделю 1,72).

Кластер 1 - можно попробовать использовать скидки, как и для кластера 0 + для тех клиентов, которые живут далеко от фитнеса, предлагать дополнительные варианты совмещения посещения нескольких занятий (например, зал + сауна, зал + бассейн). Не продавать абонементы с коротким сроком действия - минимум 6 месяцев. Так как в этом кластере самый низкий процент прихода клиентов по акции "приведи друга" предложить скидки для самих клиентов и друзей, в случае, если клиент привел друга.

Кластер 2 - нет проблем с посещением, но короткие абонементы. Возможно, стоит попробовать сделать скидку на продление абонемента до года. Также имеет смысл аналогично кластеру 1 сделать скидку по акции "приведи друга".