Финальный проект.
Анализ оттока банковских клиентов.

Давыдова Анастасия Сергеевна

Яндекс Практикум
2023

Оглавление:

Ссылка на дашборд (Tableau)

Задача (исходная версия от "заказчика"):

- Проанализируйте клиентов регионального банка и выделите сегменты клиентов, которые склонны уходить из банка.
- Проведите исследовательский анализ данных, определите все значимые признаки отточности (интервалы значений характеристик, которые связаны с повышенным оттоком, сравните портреты типичных клиентов, которые склонны и не склонны уходить из банка и т.д)
- Сформулируйте и проверьте статистические гипотезы:

1. Проверьте гипотезу различия дохода между теми клиентами, которые ушли и теми, которые остались.
2. Сформулируйте и проверьте статистическую гипотезу относительно представленных данных, которая поможет внести ясность в исследование

- Объединяя признаки отточности, сформируйте сегменты, отберите из них лучшие и дайте по ним рекомендации

Описание данных.

Датасет содержит данные о клиентах банка «Метанпром». Банк располагается в Ярославле и областных городах: Ростов Великий и Рыбинск.

Колонки:

- userid — идентификатор пользователя,
- score — баллы кредитного скоринга,
- City — город,
- Gender — пол,
- Age — возраст,
- Objects — количество объектов в собственности / equity — количество баллов собственности
- Balance — баланс на счёте,
- Products — количество продуктов, которыми пользуется клиент,
- CreditCard — есть ли кредитная карта,
- Loyalty / last_activity — активный клиент,
- estimated_salary — заработная плата клиента,
- Churn — ушёл или нет.

Предобработка.

На этапе предобработки мы установили корреляционную взаимосвязь между столбцом balance, содержащим пропуски и столбцами с баллами недвижимости (equity), количеством продуктов (products), скорингом (score), зарплатой (est_salart), фактом оттока (churn) и кредитной картой (credit_card). Мы сгруппировали данные по этим переменным и заполнили пропуски медианным значение баланса для каждой группы.

Исследовательский анализ данных.

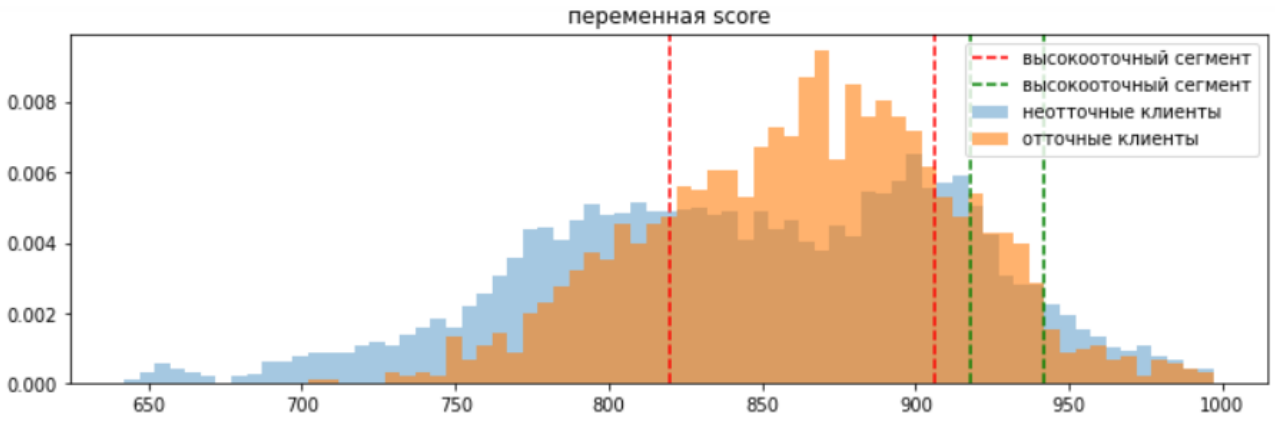
На начальном этапе исследования мы установили, что:

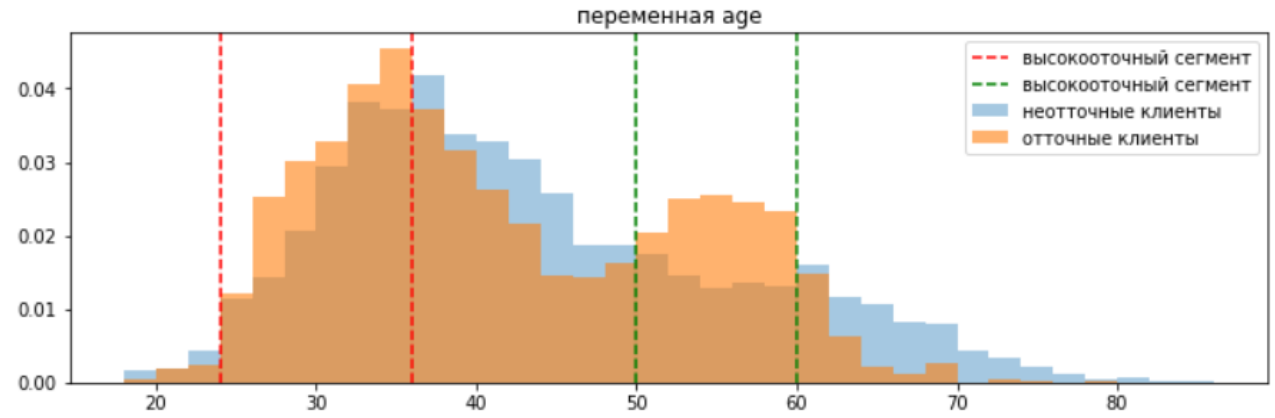
- разброс возраста клиентов - от 18 до 86, медиана - 40 лет,
- разброс скоринга - от 642 до 1000, медиана - 853 балла,
- количество недвижимости колеблется от 0 до 9, медиана - 3 объекта,
- количество банковских продуктов колеблется от 0 до 5, медиана - 2,
- минимальная зарплата клиентов (est_salary) 2546, максимальная - 1 395 064, медиана - 119 626,
- целевая переменная (churn) несбалансированна - неотточных клиентов больше, чем отточных (~80/20),
- клиенты, имеющие кредитную карту, составляют треть от имеющейся выборки,
- активных и неактивных клиентов примерно поровну. Согласно условию, клиенты промаркированные как активные (last_activity = 1) явно сообщают нам о желании покинуть банк, но пока не уходят.

Мы воспользовались коэффициентом корреляции phik, который может обнаружить нелинейные зависимости и обнаружили значимую корреляцию (больше 0.1) между целевой переменной churn и:

- products (0.44),
- equity (0.35),
- last_activity (0.26),
- score (0.23),
- gender_m (0.22),
- credit_card (0.20)
- и age (0.18)

Визуализация переменных: числовые значения – score и age.





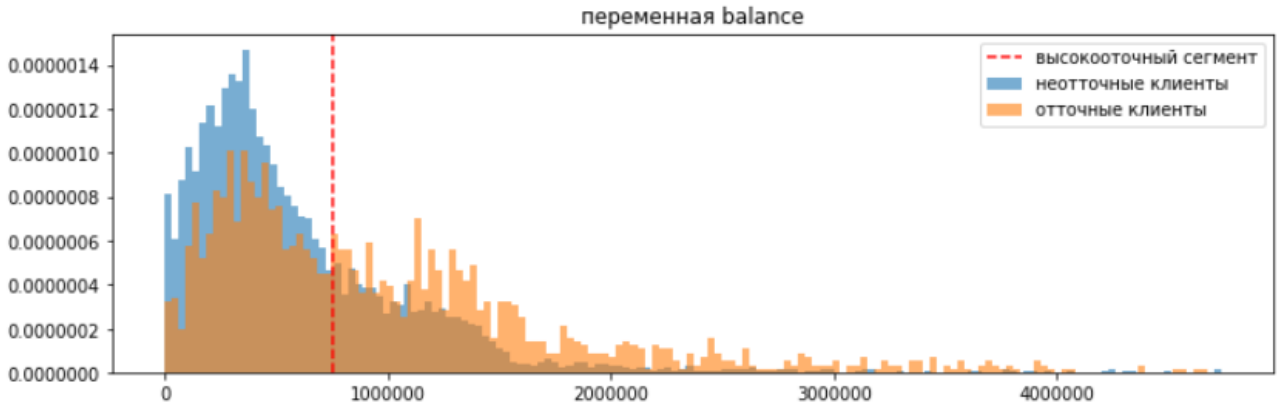
Мы сравнили две группы клиентов (отточные и неотточные), визуализировали распредление их признаков, нашли на графиках участки с повышенным оттоком и подсчитали отток на разных участках. На графике, отражающем баллы кредитного рейтинга (score) мы наблюдаем два участка, где отточные клиенты преобладают над неотточными - 820 – 906 и 918 – 942. Чтобы подтвердить свое предположение, мы посчитали отток на этих участках и соседних отрезках:

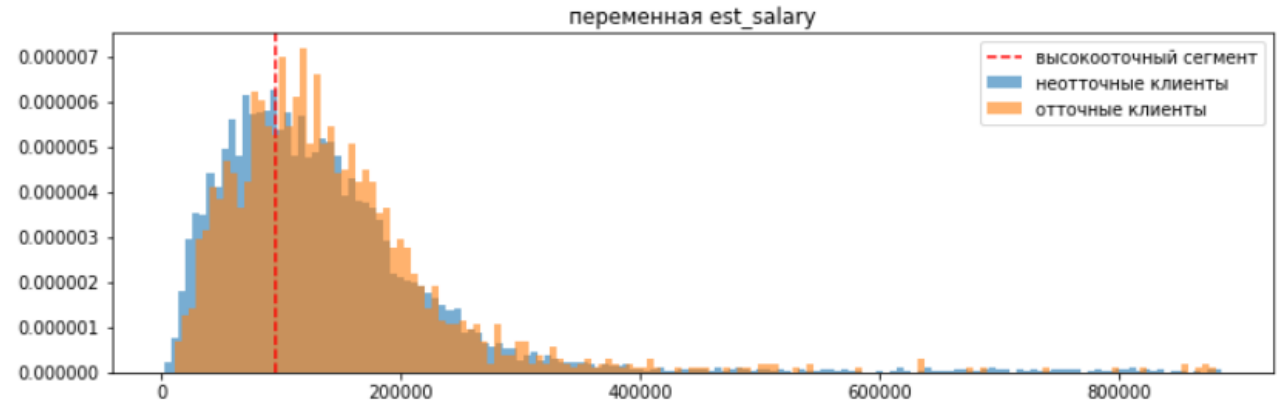
скоринг < 820	10.74 %
группа (820 – 906)	24.25 %
группа (907 – 917)	16.83 %
группа (918 – 942)	20.54 %
скоринг > 942	12.19%

Эту же операцию мы проделали с переменной возраст (age) - мы выделили участки графика "на глаз" и подтвердили предположения математически:

возраст < 24	12.24 %
группа (24 – 36)	20.74 %
группа (37 – 49)	15.22 %
группа (50 – 60)	26.47 %
возраст > 60	6.49 %

Визуализация переменных: числовые значения – balance и est_salary.





Идентичная работа проведена с оставшимися числовыми переменными – баланс (balance) и зарплата (est_salary).

Для зарплаты пороговым значением стали 95 000 - если мы разделим наших пользователей по зарплате, то пользователи, зарабатывающие больше порога чаще нас покидают:

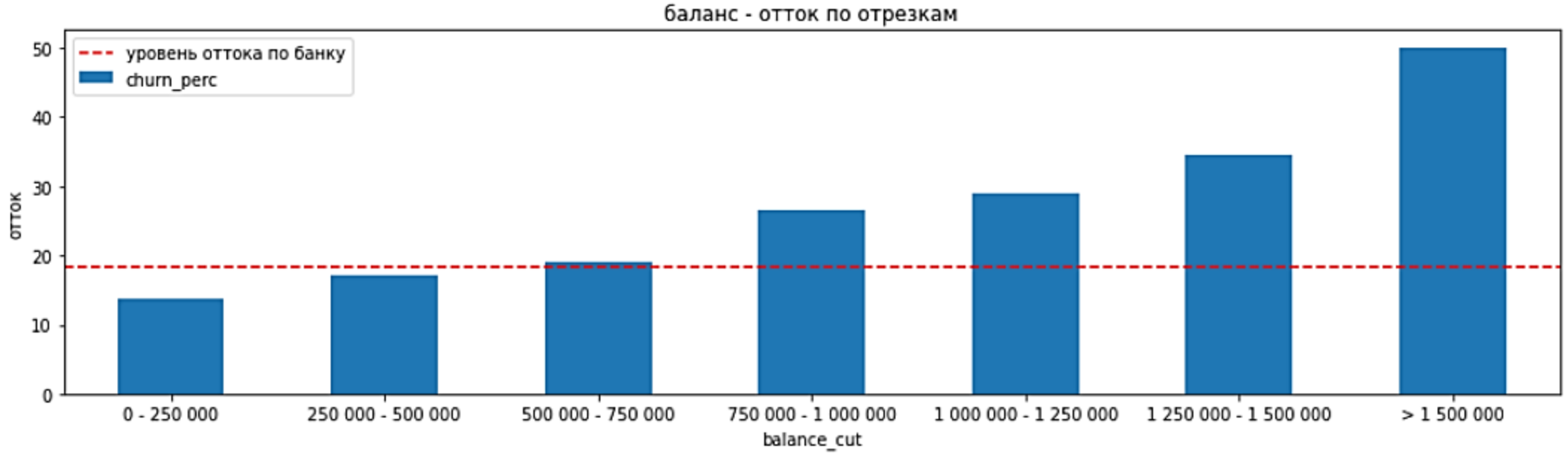
зарплата < 95 000	15.81 %
Зарплата > 95 000	19.62 %

Порогом для баланса (balance) стали 750 000:

баланс < 750 000	16.58 %
баланс > 750 000	34.84 %

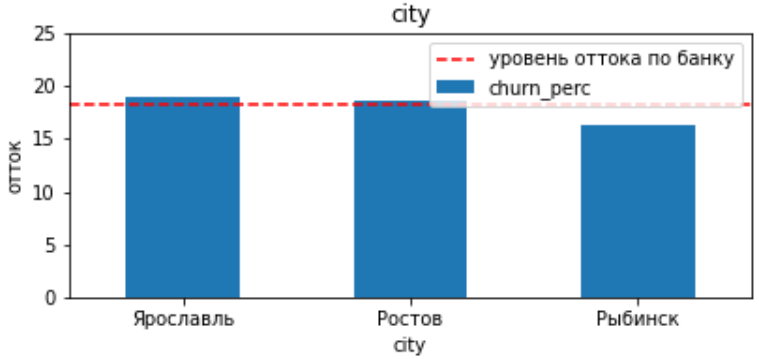
Отток в разрезе баланса выглядит неоднородным, было решено разбить числовую прямую на "корзины" по 250 000 и посчитать отток в каждой.

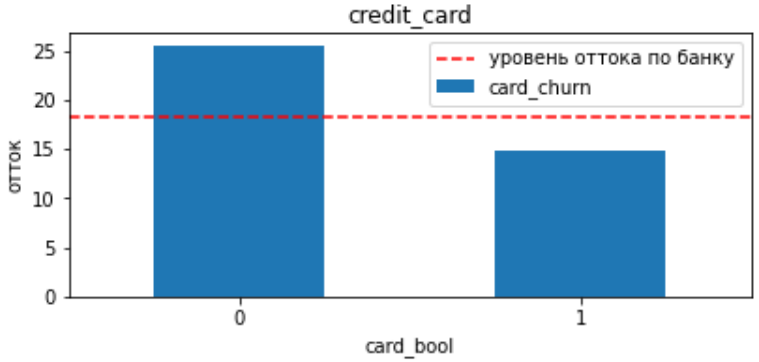
Визуализация переменных: числовые значения – balance.

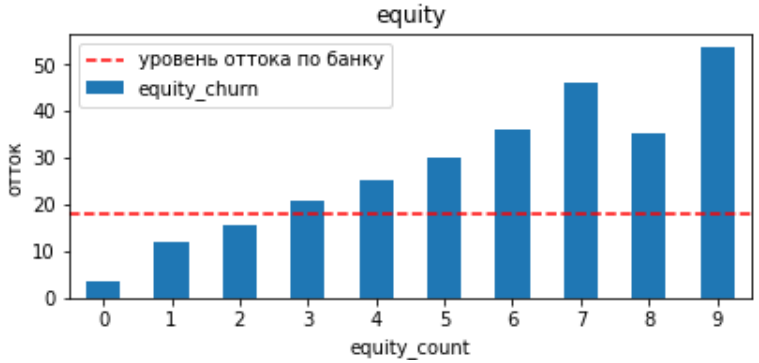


Наше предположение подтвердилось: процент оттока растет с ростом баланса клиентов, пороговым значением является 750 000 – после этого значения отток начинает увеличиваться.

Визуализация переменных: категориальные переменные – city, credit_card, equity.







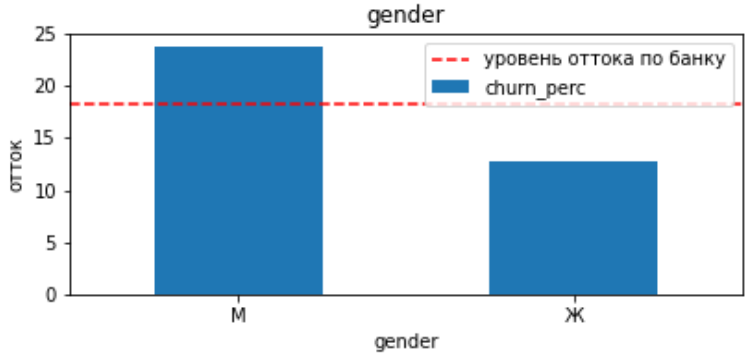
Мы посчитали и визуализировали процент оттока для категориальных переменных. Дополнительно, на графики мы нанесли прерывистую горизонтальную линию - средний отток по банку.

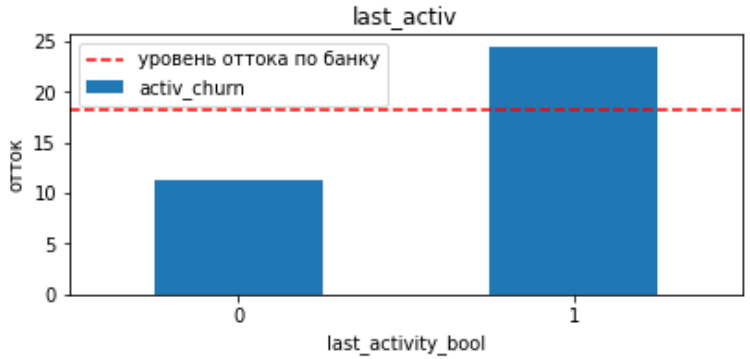
Процент оттока по городам (city) одинаковый.

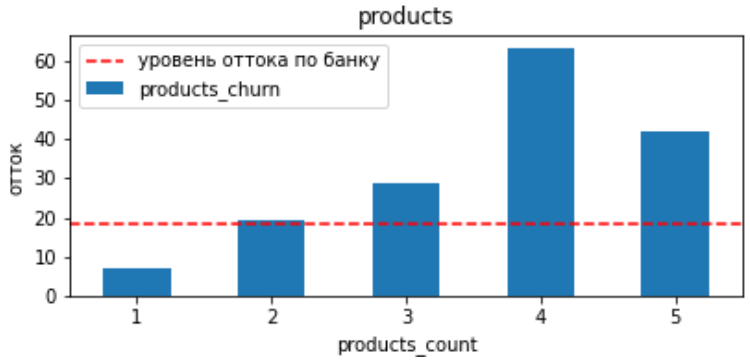
Клиенты без кредитной карты покидают банк в два раза чаще, чем клиенты, имеющие кредитку.

Рост процента оттока зависит от количества баллов недвижимости - чем больше баллов, тем выше отток.

Визуализация переменных: категориальные переменные – gender, last_activ, products.







Мужчины в два раза чаще покидают банк, чем женщины.

Клиенты, проявлявшие в последнее время активность, уходят в два раза чаще (last_activ = 1).

Рост количества банковских продуктов (products) связан с ростом оттока: чем больше продуктов, тем выше отток (пороговое значение – 2 продукта).

Потенциальные признаки отточных клиентов:

- клиенты заявляют о своем желании уйти (last_activ = 1),
- пол клиентов – мужской (gender = M)
- чаще уходят клиенты с кредитным рейтингом ~820-906 и ~918-942,
- с зарплатой более 95 000,
- с балансом более 750 000,
- наиболее отточные клиенты имеют четыре банковских продукта,
- наиболее отточные возрастные группы - 24-36 и 50-60.
- у отточных клиентов чаще нет кредитной карты,
- с ростом количества недвижимости растет процент оттока.

Исследование портретов отточных и неотточных клиентов: абсолютная разница между портретами.

Мы составили портреты двух групп пользователей (отточные и неотточные), сравнили их и нашли значимую разницу в некоторых значениях. Эти переменные могут стать маркерами для сегментации:

- баланс,
- количество баллов недвижимости,
- отметка об активности,
- пол и количество продуктов.

Сегментация по факту наличия кредитной карты – опционально.

Промежуточный итог -

мы собрали в одну таблицу значения корреляции (phik_corr), портретную разницу (diff_mean_abs) и отточные интервалы, обнаруженные на графиках, и вынесем вердикт, какие переменные мы будем использовать в качестве сегментации.

	phik_corr	diff_mean_abs	корреляция_с_целевой	отточный_интервал	сравнение портретов	вердикт
balance	0.03	59.87	есть связь	> 750 000	есть связь	используем
equity	0.35	57.98	есть связь	> 2 - растет с увеличением количества	есть связь	используем
last_activity	0.26	45.83	есть связь	активный клиент (1)	есть связь	используем
gender_m	0.22	38.30	есть связь	пол = мужской	есть связь	не используем
products	0.44	35.23	есть связь	4 продукта	есть связь	используем
credit_card	0.20	22.54	есть связь	нет кредитки (0)	есть связь	используем
city_rybinsk	0.04	14.29	нет связи	-	есть связь	не используем
city_yaroslavl	0.03	5.17	нет связи	-	нет связи	не используем
age	0.18	3.65	есть связь	24-36 и 50-60	нет связи	под вопросом
score	0.23	2.13	есть связь	820-906 и 918-942 баллов	нет связи	под вопросом
est_salary	0.05	0.42	нет связи	< 95 000	нет связи	не используем

Для сегментации нам подоходят переменные balance, equity, last_activity, products и credit_card.
Перспективным кажутся переменные score и age, мы обнаружили корреляцию с целевой переменной, но не обнаружили разницу между портретами - можно проверить ее дополнительно.

Дополнительно проверим статистическими тестами значения баланса и скоринга в двух группах: если есть различия – мы можем использовать их для сегментации.

Формулирование и проверка статистических гипотез.

В ходе исследования нами были выдвинуты две гипотезы:

- **гипотеза 1** – о равенстве доходов между отточными клиентами и теми, которые остались,
- и **гипотеза 2** – о равенстве баллов кредитного скоринга между отточными клиентами и теми, которые остались.

Соответственно, нулевая и альтернативные гипотезы звучат так:

- H0: между двумя группами статистически значимой разницы обнаружить не удалось,
- H1: между группами есть статистически значимая разница.

Обе гипотезы мы проверяли двумя статистическими тестами – t-тестом Стьюдента и непараметрическим критерием Манна-Уитни. В тестах нас интересует значение p-value, которое покажет вероятность того, что отличия в двух выборках случайны. Чем меньше это значение и ближе к нулю - тем лучше.

t-тест сравнивает средние значения двух выборок и сообщает, отличаются они или нет.

Требования t-теста:

- нормальное распределение данных (строгое требование),
- дисперсия двух выборок равна,
- данные являются независимыми и непрерывными.

t-тест также чувствителен к выбросам.

Критерий Манна-Уитни не такой строгий, основная его идея - проранжировать две выборки по порядка от меньшего и большему и сравнить ранги одних и тех же наблюдений, попавших в обе выборки.

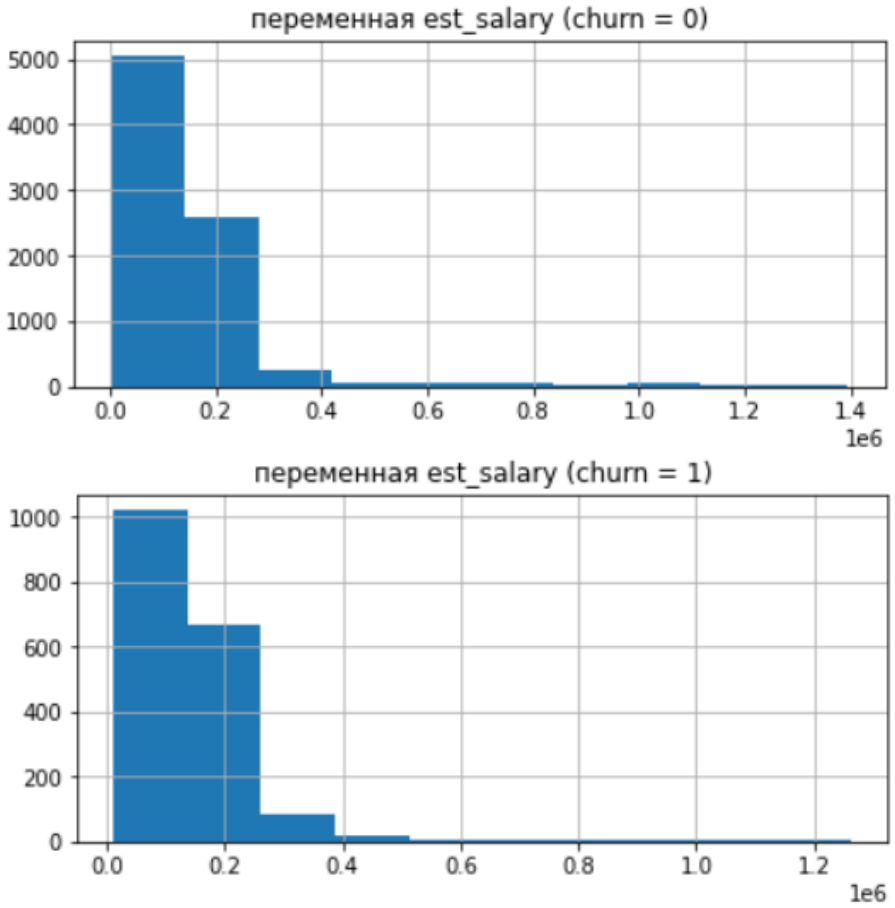
В случаях ненормальных распределений t-тесту доверять нельзя,, стоит опираться на результаты критерия Манна-Уитни.

Для исследования мы установили порог статистической значимости равный 5% - это порог означающий, происходит событие случайно или нет.

Гипотеза 1 – "проверьте гипотезу различия доходов между теми клиентами, которые ушли и теми, которые остались".

H0: различий дохода между двумя группами нет.

H1: доход в группах ушедших клиентов и оставшихся различается.



Дисперсия (разброс) в двух выборках различается, распределение признака- ненормально, оно сильно скошено вправо, присутствует большое количество выбросов.

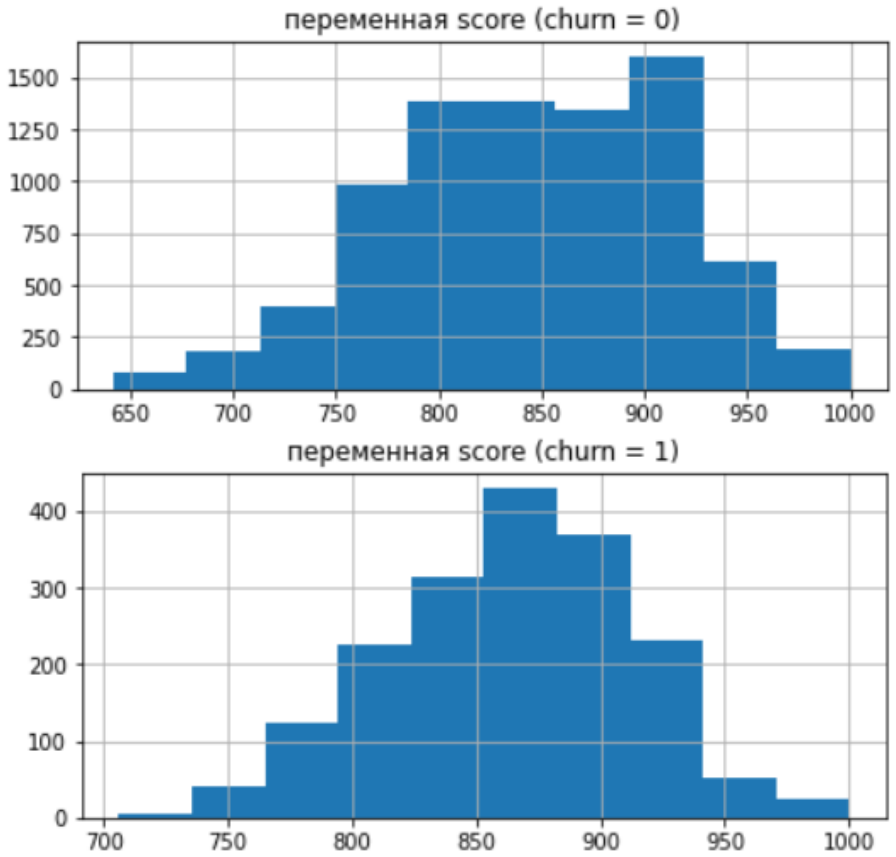
Сравним значений p-value с порогом в 5%: если полученное значение больше порога – мы говорим, что разницы статистики в двух группах нет, если меньше – разница есть.

Тесты показали значение p-value равное ~0.852 (t-тест) и ~0.0002 (u-критерий Манна-Уитни).

Распределение данных не удовлетворяет требованиям t-теста, поэтому мы воспользуемся результатами u-критерия и ***отвергнем нулевую гипотезу*** - доход в группах ушедших клиентов и оставшихся различается.

Гипотеза 2 – "проверьте гипотезу различия скоринга между теми клиентами, которые ушли и теми, которые остались".
H0: различий скоринга между двумя группами нет.
H1: скоринг в группах ушедших клиентов и оставшихся различается.

Дисперсия в двух выборках отличается, но распределение данных близко к нормальному, колоколообразному.



Мы можем довериться результатам t-теста: проверяя вторую гипотезу мы также сравним значений p-value с порогом в 5%.
И t-тест и u-критерий Манна-Уитни показали значения, близкие к нулю, полученные значения меньше порога, значит мы можем ***отвергнуть нулевую гипотезу*** - скоринг в группах ушедших клиентов и оставшихся различается.

Между двумя группами есть разница в значениях баланса и скоринга, мы можем использовать эти переменные для сегментации.

Портреты высокооточных сегментов.

Мы проанализировали различные комбинации из 2-3 отточных признаков, из них отобрали сегменты с наивысшим оттоком и удовлетворяющие следующие требования:

- сегменты должны быть оптимальные по размеру (от 1 000 до 3 000 клиентов),
- в сегмент должно было попадать от 1/3 до 2/3 всех отточных пользователей по банку,
- так же уровень оттока в сегменте должен минимум в 1.5 раза превышать средний по банку.

На основе полученных результатов, мы составили портреты типичных представителей из высокооточных сегментов и рекомендации по их удержанию.

Портрет первый:

- клиенты сообщали о своем желании уйти,
- клиенты владеют более 2 объектов недвижимости,
- довольно высокий скоринг таких клиентов - 820 - 906,
- опционально - если клиенты мужчины, то шанс, что они уйдут – повышается.

Портрет второй:

- клиенты сообщали о своем желании уйти,
- клиенты владеют более 2 объектов недвижимости,
- клиенты – мужчины,
- опционально - зарплата более 95 000 повышает шанс оттока.

Портрет третий:

- клиенты сообщали о своем желании уйти,
- клиенты владеют более 2 объектов недвижимости,
- клиенты не заводят кредитную карту,
- опционально - клиенты молодого возраста (24 - 36).

Рекомендации по удержанию:

Во всех терх сегментах есть общие черты - клиенты владеют более двух объектов недвижимости и заявляют о своем желании уйти. Мы можем предположить, что они не находят интересных банковских продуктов под свои нужды.

Один из вариантов удержания - сделать ставку на банковские продукты, связанные с недвижимостью - рефинансирование ипотеки (при наличии высоких баллов кредитного скоринга - подходит для клиентов из первой группы), льготные программы ипотек для определенной возрастной группы (для молодых - клиенты из третьей группы), программы страхования недвижимости и кредиты на ремонт на выгодных условиях (такой кейс есть у Сбербанка).

Вторая группа состоит из мужчин с недвижимостью и высокой заработной платой. Мы можем предложить этим клиентам более выгодные условия вкладов и капающие проценты на остаток по счету.

Третья группа интересна тем, что в ней молодые клиенты отказываются заводить кредитные карты: увеличенный кредитный лимит и льготный период могут привлечь эту группу. Так же удержать эту группу помогут продукты "по интересам" - карты с повышенным кешбеком на компьютерные игры, путешествия и отдых, рестораны или траты, связанные с машиной. Или (вернемся к сбербанку) выпуск дебетовых "молодежных карт" с бесплатным обслуживанием.