

Анализ оттока клиентов банка

учебный проект Яндекс.Практикум
тема: «Выпускной проект»

автор: Сергей Васильев
foto.2000@yandex.ru

февраль, 2022

Содержание

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Исследовательский анализ данных

Исследование зависимостей оттока от признаков

Кластеризация клиентов

Итоги исследования

Исследовательский анализ данных

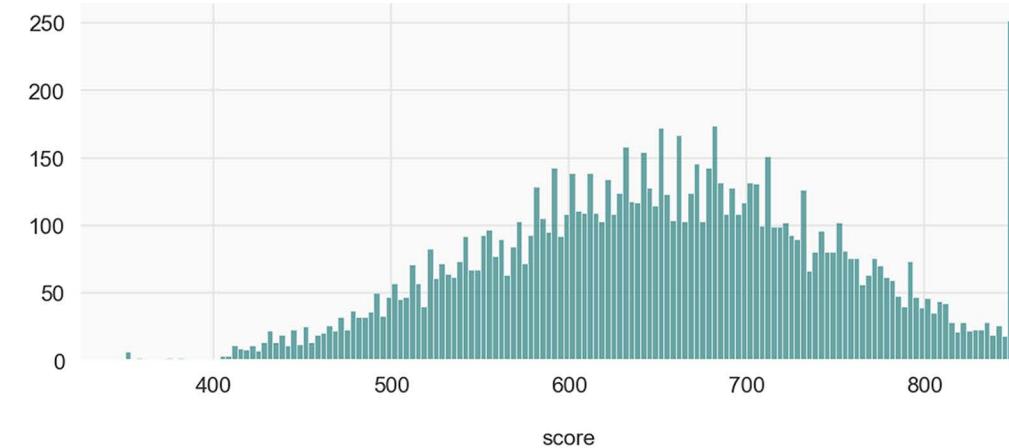
Гистограммы входных данных

Гистограммы

score
balance
objects

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

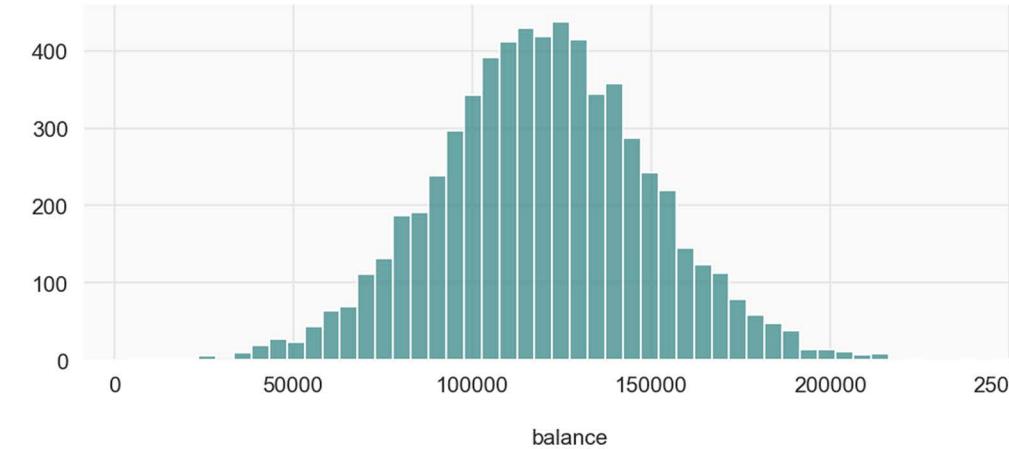
Гистограмма «score»



Распределение `score` выглядит нормальным со смещением вправо.

Обращает на себя внимание пик на максимальном значении, равном 850. Это похоже на искусственное присвоение некоторого значения всем значениям, превышающим заданный максимум.

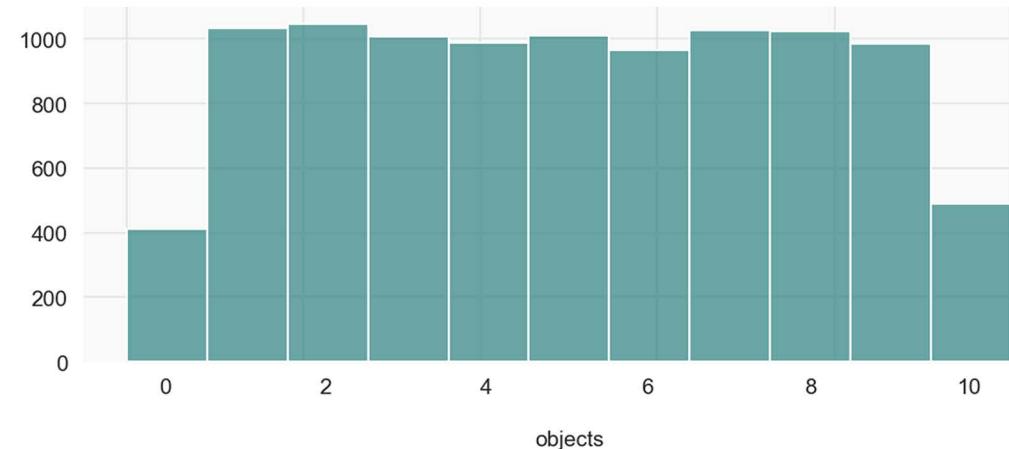
Гистограмма «balance»



Практически идеальное нормальное распределение, что и настораживает. В реальности более вероятно смещение влево и длинный хвост высоких значений.

Стоит отметить, что распределения `balance` и `objects` не согласуются друг с другом. Логично предположить, что чем больше у клиента объектов в собственности, тем вероятнее, что и баланс его будет больше.

Гистограмма «objects»



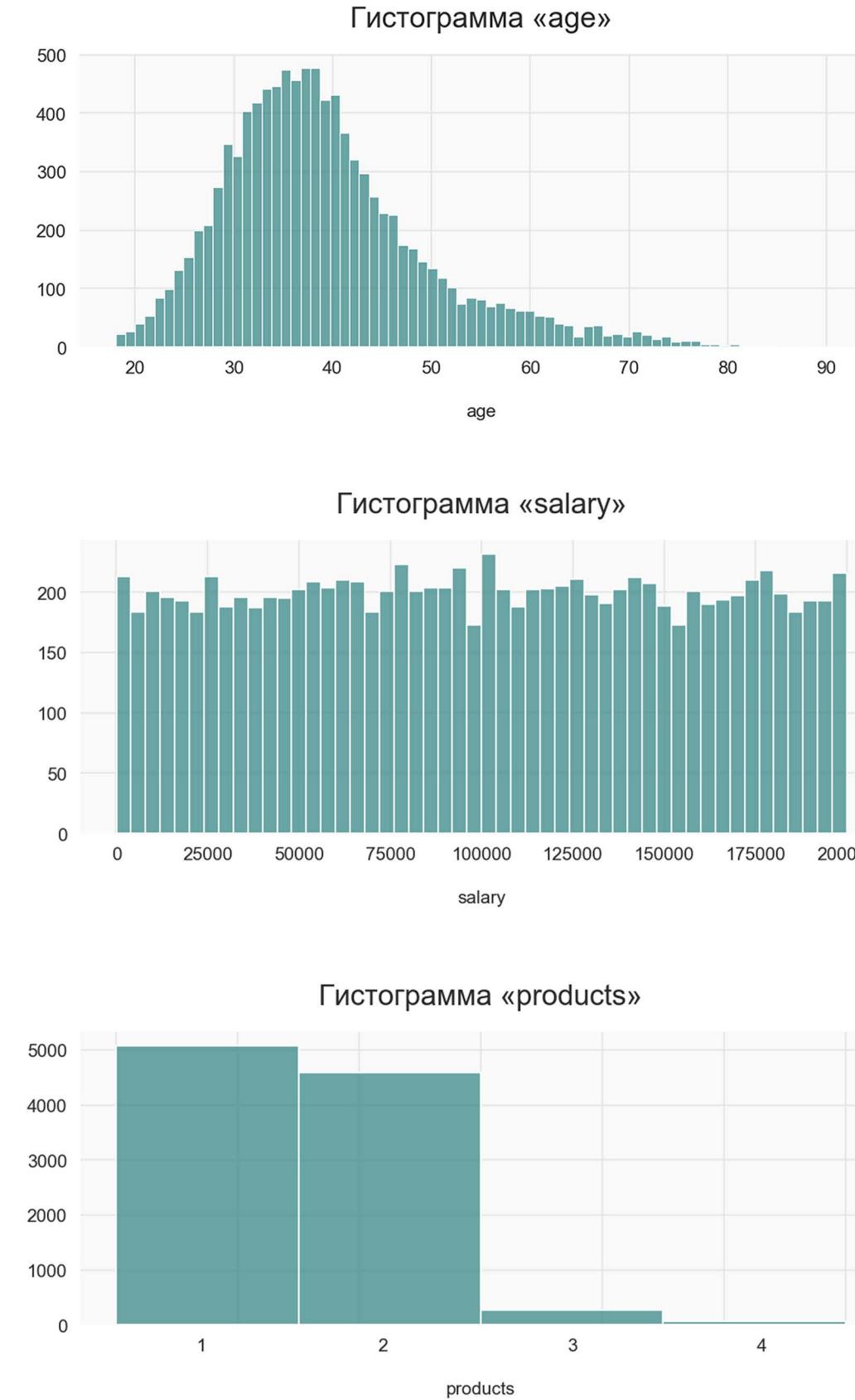
Если отбросить крайние значения (0 и 10 объектов в собственности), для любого другого значения (1-9) количество клиентов примерно одинаково.

Вряд ли такое равномерное распределение возможно в реальности.

Гистограммы

age
salary
products

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



Распределение `age` похоже на логнормальное со смещением влево.

Большая часть клиентов банка находится в возрасте примерно 25-55 лет.

Данные распределены практически равномерно во всем диапазоне.

Предположительно, данное поле содержит не зарплату, а некое число в заданном диапазоне, и задача поля — ранжирование клиентов. Если это действительно так, то стоит привести данное поле в соответствие с его назначением.

Подавляющее количество клиентов используют 1-2 продукта.

Стоит проанализировать в будущем, почему клиенты редко используют более 2-х продуктов. Возможно, неконкурентность некоторых продуктов способствует оттоку клиентов.

Исследование зависимостей оттока от признаков

Диаграммы размаха

Зависимость оттока от города

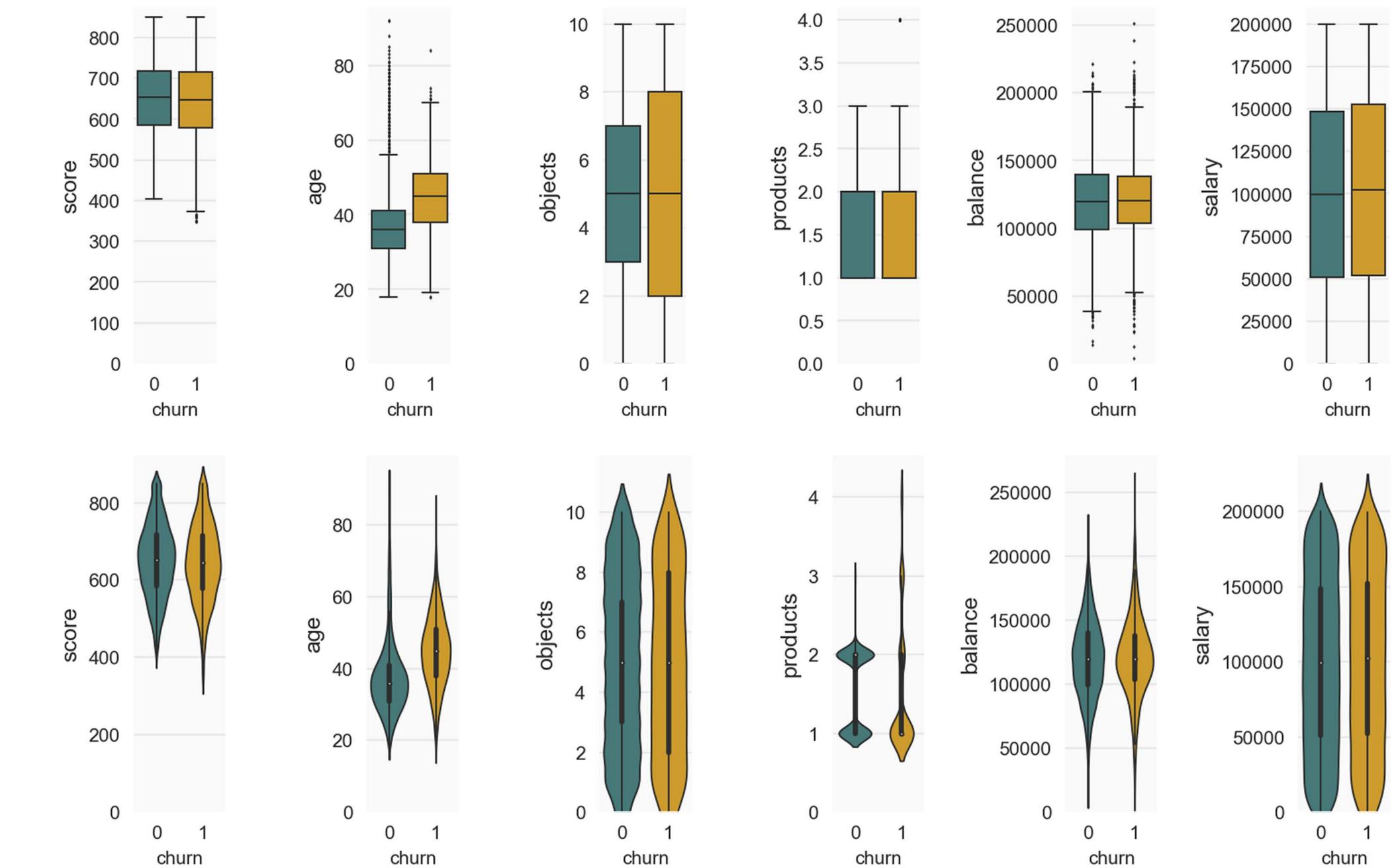
Средние значения

Зависимость оттока от возраста

Диаграммы размаха

score
age
objects
products
balance
salary

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



Хорошо видна зависимость оттока от возраста (age).
Клиенты в оттоке существенно реже используют 2 продукта (products).
Остальные поля практически не влияют на отток.

Доля оттока по городам

Средние значения признаков

Доля оттока в зависимости от города

city	active	churn	churn_%
Ростов Великий	1695	814	32.4%
Рыбинск	2064	413	16.7%
Ярославль	4204	810	16.2%

в Ростове отток существенно выше, чем в двух других городах

Средние значения признаков, сгруппированных по оттоку

	churn	active	churn	ratio
loyalty	0.555	0.361	1.537	
gender	0.573	0.441	1.299	
age	37.408	44.838	1.199	
city_id	197.779	227.750	1.152	
products	1.544	1.475	1.047	
objects	5.033	4.933	1.020	
salary	99738.392	101465.678	1.017	
creditcard	0.707	0.699	1.012	
balance	119535.864	120746.971	1.010	
score	651.853	645.351	1.010	

Сильнее всего средние значения среди оставшихся и ушедших клиентов различаются для поля `loyalty`.

Заметно различаются средние значения для возраста и пола (`age` и `gender`).

По этим трем признакам, а также городу (`city`) и количеству продуктов (`products`) предположительно можно определить клиента, склонного к оттоку.

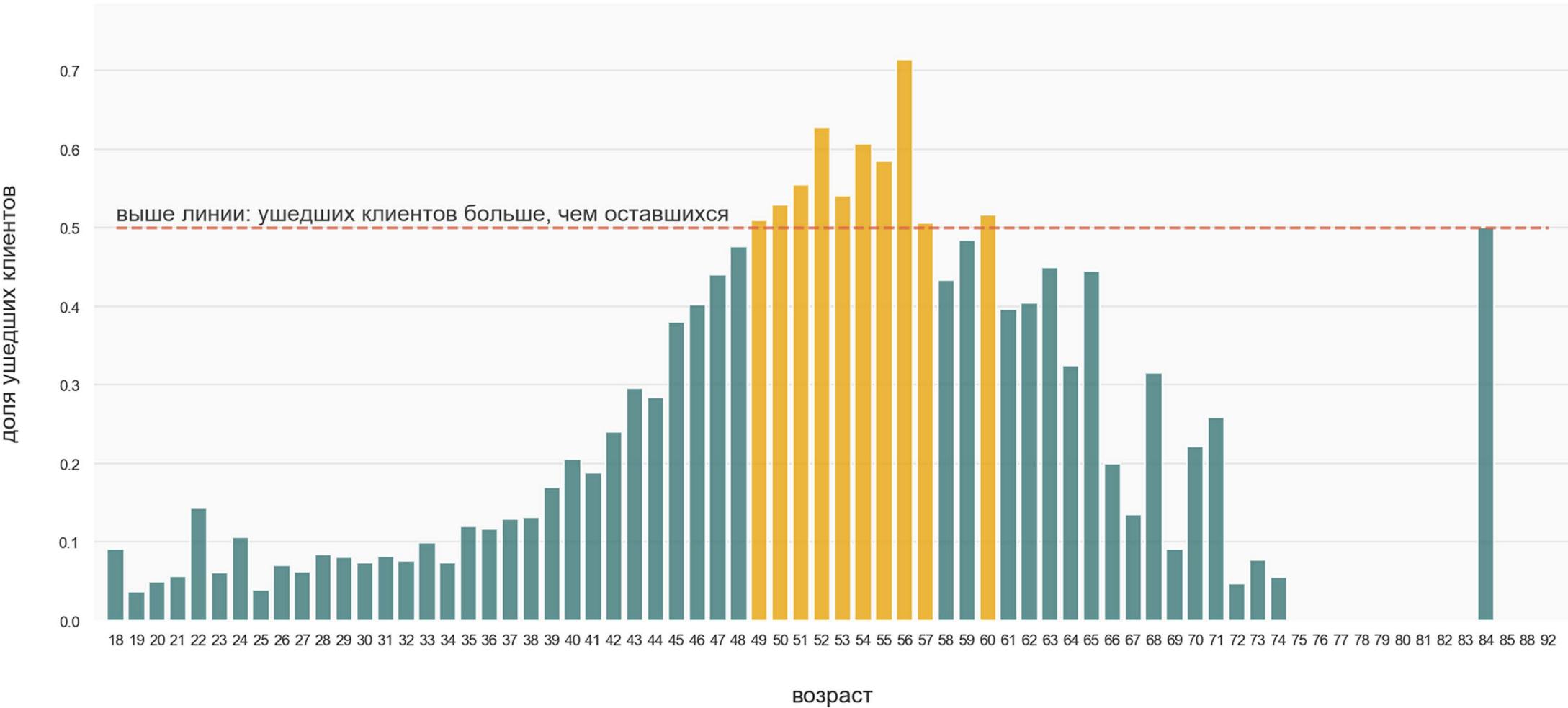
ratio показывает коэффициент различия между средними значениями признаков для Active и Churn, ratio всегда ≥ 1

Зависимость оттока от возраста:

доля
ушедших
клиентов

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Доля ушедших клиентов в зависимости от возраста



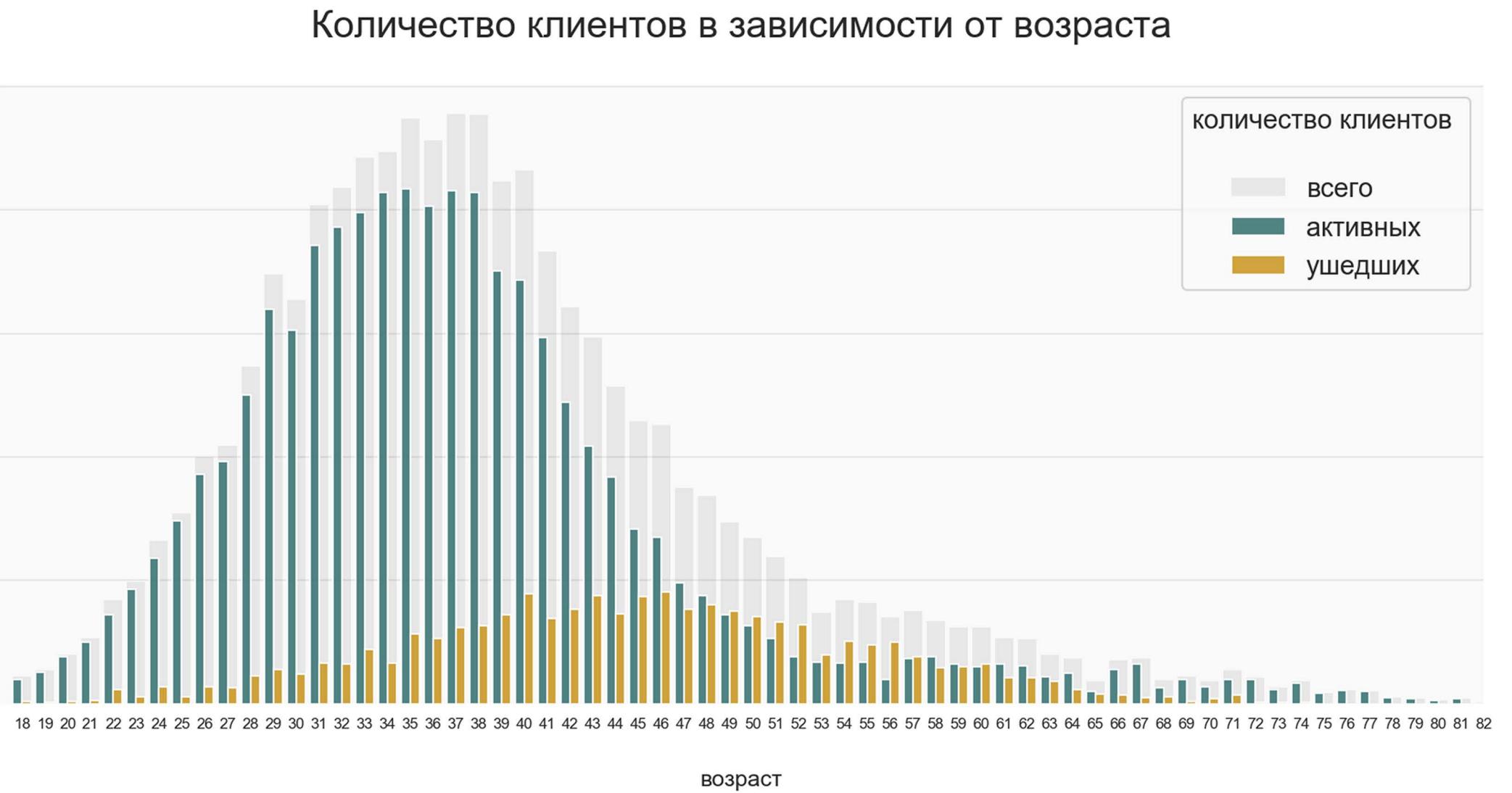
До 35-38 лет доля ушедших клиентов относительно стабильна.
Затем доля ушедших клиентов растет и достигает максимума около 55 лет,
а к 70-72 годам снижается к первоначальным значениям.

Для возрастов 49-57 лет количество ушедших клиентов преобладает над количеством оставшихся.

В возрасте старше 74 лет всего лишь 1 ушедший клиент из 54, который и определил пик на графике в 84 года.

Зависимость оттока от возраста: количество клиентов

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



Наибольшее количество активных клиентов банка находится в возрасте примерно 30-40 лет.
Наибольшее количество ушедших клиентов находится в возрасте примерно 40-50 лет.
В возрасте 49-57 лет количество ушедших клиентов преобладает над количеством оставшихся.
В зоне малых (до 21 года) и больших (от 72 лет) значений количество ушедших клиентов исчезающе мало.

Кластеризация клиентов

Кластеры по доле оттока

Примеры интерпретации кластеров

Кластеры, упорядоченные по доле оттока

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Доля оттока и средние значения признаков

churn	loyalty	gender	age	city_id	products	objects	salary	creditcard	balance	score
75.9%	0	0	50.0388	Ростов	1.46983	4.99569	101447	0.702586	119932	652.841
62.8%	0	1	47.3827	Ростов	1.47292	5.10108	100589	0.718412	120205	634.083
43.4%	0	0	44.5802	Ростов	1.54557	4.9294	100548	0.700899	118733	650.49
36.6%	0	1	45.9286	Ростов	1.46201	5.14438	102532	0.703647	122312	646.926
35.1%	1	1	46.3024	Ростов	1.58065	5.11694	100797	0.681452	121323	655.246
33.3%	0.985185	0.355556	63.4815	Ростов	1.45185	4.75556	97105.4	0.688889	119294	642.919
27.0%	1	0	36.5949	Ростов	1.54641	4.81224	101510	0.696203	119418	655.789
26.8%	0	0	34.1925	Ростов	1.5025	5.1975	104855	0.7375	118021	653.025
24.5%	0.954717	0	62.9245	Ростов	1.55094	4.89811	95666.6	0.698113	119324	648.951
19.3%	1	0	41.8362	Ростов	1.53395	4.93475	102439	0.691079	119973	653.411
17.4%	0	1	32.2571	Ростов	1.52286	5.31143	98911.3	0.74	120685	650.666
16.1%	0.980874	1	61.0546	Ростов	1.50273	5.0082	96335.5	0.688525	116384	653.432
11.4%	0	0	31.1751	Ростов	1.58753	5.03717	100580	0.707434	119655	646.645
11.1%	1	1	40.8271	Ростов	1.51101	4.89207	100020	0.680617	119068	652.357
9.2%	1	1	32.2824	Ростов	1.54962	4.75318	100537	0.720102	119687	657.29
7.6%	0	1	32.4499	Ростов	1.53236	5.12712	99183.9	0.706471	122797	648.161
6.7%	1	0	29.8669	Ростов	1.55479	4.98197	97530.9	0.705964	116822	650.302
3.5%	1	1	29.9857	Ростов	1.54116	5.05598	98914.4	0.729967	120012	652.82

Примеры интерпретации кластеров

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

женщины
из Ростова
в возрасте 50 лет
с loyalty=0
имеют долю оттока 76%

мужчины
из Ростова
в возрасте 47 лет
с loyalty=0
имеют долю оттока 63%

женщины
из Рыбинска или Ярославля
в возрасте 30 лет
с loyalty=1
имеют долю оттока 7%

мужчины
из Рыбинска или Ярославля
в возрасте 30 лет
с loyalty=1
имеют долю оттока 3%

Итоги исследования

Факторы влияния на отток

Факторы влияния на отток

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Лояльность [loyalty]

бинарный признак
ноль означает повышенную склонность к оттоку

Пол [gender]

бинарный признак
женщины более склонны к оттоку

Город [city]

категориальный признак
в Ростове отток двое выше, чем в других городах

Возраст [age]

числовой дискретный признак
в диапазоне 40-70 лет доля оттока повышена
максимальная доля оттока около 55 лет

loyalty = 0

женщины

Ростов

49-57 лет