

Задание «Оптимизация онбординга в первую сессию»

Компания — SaaS бизнес, платформа по созданию чат ботов. Целевая аудитория - маркетологи и владельцы бизнесов. Компания недавно запустила интеграцию с Instagram - теперь в этом канале существует возможность запускать ботов. В следствии этого на платформу пришло много нового трафика, в том числе пользователей, которые сталкиваются с чат ботами впервые.

Вы — продуктовый аналитик компании и ваша цель помочь продакт менеджеру повысить метрику активации в первой сессии. Под активацией мы понимаем набор ботом N контактов за M первых дней. Эта метрика берется исходя из предположения, что активированные пользователи могут получить больше ценности от платформы и будут иметь больше мотивации оплатить платную подписку.

В качестве исходных данных вам дали события, которые генерировали пользователи в продукте в период с 2021-09-18 — 2021-11-23.

Задача

Вы вместе с продакт менеджером решили копнуть в поведение пользователей в первую сессию, для того, чтобы сделать обучение и/или понять с кем лучше проводить качественные исследования. Для этого вам нужно сделать следующие шаги:

1. Разбить ивенты на сессии (сессия = множество ивентов пользователя, между которыми проходит менее 30 минут)
2. Выделить сегменты на основе данных поведения пользователей в первой сессии (сделать кластеры)
3. Описать сегменты из бизнес смысла.
 - Пример, как мы представляем описание сегментов
 - сегмент 1: пользователи, которые после подключения канала уходят из продукта и у них нет первой сессии в первый день
 - сегмент 2: мотивированные пользователи, которые в первую сессию активно проходят онбординг
4. Сформулировать рекомендации: какие сегменты стоит брать в первую очередь и почему.

Что важно в решении

- Определить сегменты с которым стоит работать в первую очередь. Выделить кластеры.
- Корректно обработанные и свести данные. Достоверность данных в отчёте важна, красота изложения не важна: более красивый отчёт не лучше менее красивого, но с верными данными
 - Использование библиотеки <https://github.com/retentioneering/retentioneering-tools> Для решения этой задачи не обязательно хорошо знать алгоритмы ML. Мы рекомендуем использовать библиотеку retentioneering.

Исходные данные: [Dataset for Career Factory 22032022.zip](#) (зеркало)

Название полей	Описание	Категория
event_name	событие	Данные по поведению юзеров
event_dttm	дата и время события	Данные по поведению юзеров
client_id	уникальный идентификатор пользователя	Свойства пользователей
client_creationdate	дата создания пользователя	Свойства пользователей
admin_status	статус администратора пользователя, 1st_admin-пользователь, который зарегистрировал аккаунт	Свойства пользователей
country_group	регион пользователя	Свойства пользователей
support_provided	была ли предоставлена поддержка пользователю	Свойства пользователей
traffic_source	источник трафика пользователя	Свойства пользователей
first_device_type_of_user	устройство, с которого была произведена регистрация	Свойства пользователей
hear_source	откуда пользователь узнал про платформу	Свойства пользователей
account_creationdate	дата создания аккаунта	Свойства ботов

account_id	уникальный идентификатор аккаунта	Свойства ботов
industry	сфера деятельности аккаунта	Свойства ботов
initial_channel	первый зарегистрированный канал у аккаунта	Свойства ботов
first_account_for_user	является ли аккаунт первым у пользователя	Свойства ботов
current_account_status	платежный статус аккаунта на данный момент	Свойства ботов
ig_followerscount	количество подписчиков в Instagram у аккаунта	Свойства ботов
ig_enable_dttm	дата подключения Instagram канала	Свойства ботов
ig_activation_dttm	дата активации Instagram канала	Свойства ботов
first_ig_flow_publish_dttm	дата публикации первого чат-бота в Instagram	Свойства ботов
trial_dttm	дата подключения trial версии платформы	Свойства ботов
paid_dttm	дата первой оплаты подписки	Свойства ботов
pro_dttm	дата доступа к pro-функциональности	Свойства ботов
ltv_predicted	новое поле от ML отдела, предсказанный LTV	Свойства ботов

Решение

Обзор данных

Из ручного исследования данных в Power BI видно, что соотношение account_id бота и client_id пользователя – многие-к-одному. При этом client_id не определяет admin_status и hear_source – это по сути свойства бота. Свойство account_id бота определяет все поля, кроме данных по поведению пользователей.

Данные об активации не содержатся в исходных данных. Из 7 дат ботов (17 560) и клиентов можно сгенерировать события 7 типов (вех воронки) и обогатить данные. Некоторые сгенерированные события по времени опережают соответствующие события в интерфейсе и не всегда имеют однозначное соответствие, однако можно составить примерную картину соответствия временных меток ботов:

1. После client_ пользователя в веб-версии обычно следует web.page_questionnaire.todo_list_personalization.shown.
2. Подключение Instagram канала примерно соответствует todo_list_item.connect_ig_account.done
3. Прохождение обучения - web.page_view.tutorial.show
4. Публикация канала очень важная веха, без которой активация невозможна и которая с очень высокой вероятностью приводит к активации. Не имеет однозначного события и ассоциирована с минимум тремя событиями web.builder.add_trigger.click, web.builder.publish_success, web.builder.publish_attempt
5. Дата активации ig_activation_dttm не имеет ассоциированных событий, так как набор контактов в общем случае не зависит от действий на форме (возможно, влияет загрузка контактов самим пользователем – этот сценарий необходимо исследовать дополнительно). Активации действительно с высокой вероятностью приводят к платежам.
6. Другие выявленные важные события: прохождение интерактивных подсказок - онбординга (web.onboarding.run), обучения (todo_list_item.start_trial.done).
7. В случае подключения пробной версии регистрируются два события – подключение pro, а за ним сразу (в течение 4 секунд) trial.

Нас в первую очередь интересует адаптация новых пользователей к платформе, а не старых пользователей к инстаграму, поэтому старых пользователей (дата создания до 2021-09-18 00:00) отбрасываем.

Разбиение ивентов на сессии

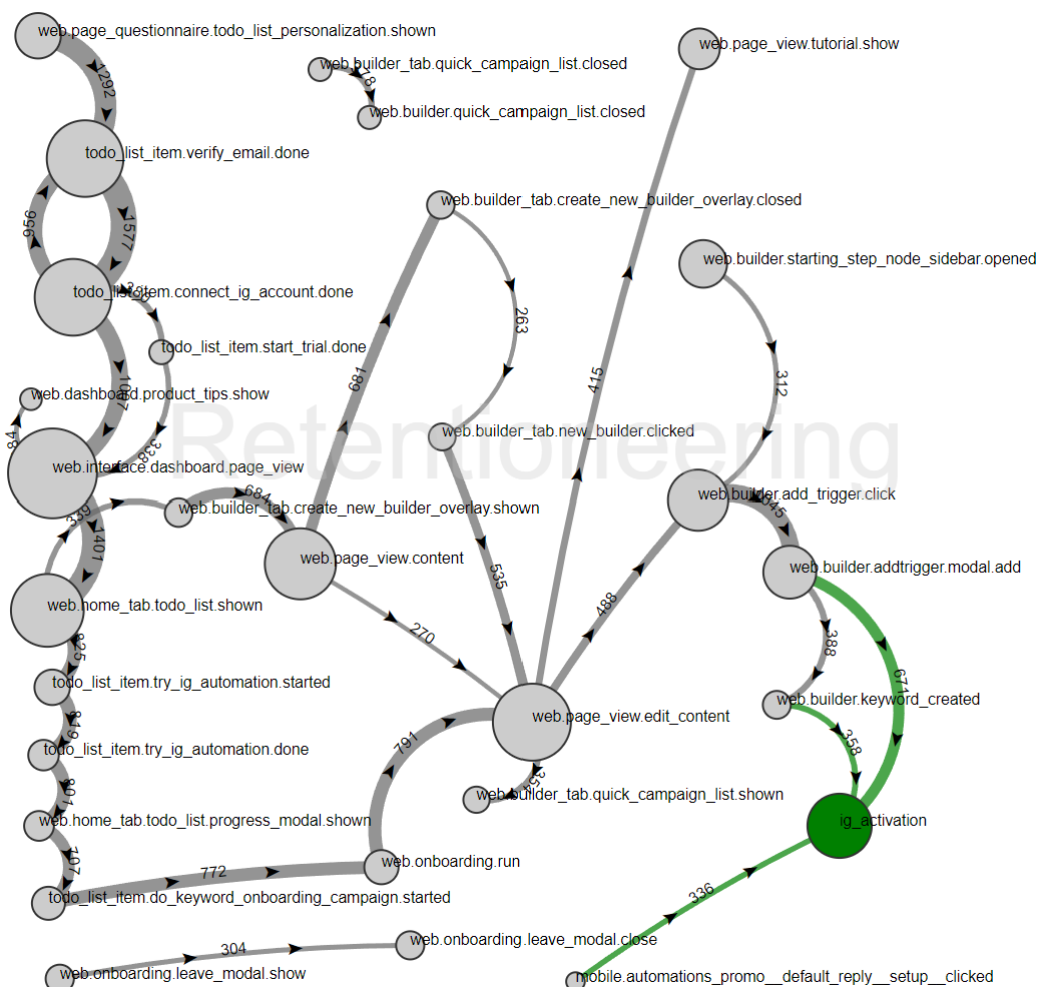
70% всех активаций происходит во 2-4 сессиях, во второй более 40%, а в первой всего 1,58%. Кроме того, видно, что мобильным пользователям хуже даются активации, особенно в первой сессии.

first_device_type_of_user	1	2	3	4	5	6	7	8	9	10
Android		5,14%	2,38%	1,72%	0,77%	0,78%	0,53%	0,13%	0,44%	0,03%
Desktop	1,40%	29,49%	11,49%	8,31%	5,02%	3,41%	3,44%	1,45%	1,41%	1,09%
iOS	0,18%	7,52%	2,84%	2,63%	1,29%	1,15%	0,20%	0,25%		
Bcero	1,58%	42,14%	16,71%	12,66%	7,07%	5,34%	4,17%	1,83%	1,85%	1,12%

Разница в доле активаций в первую и вторую сессию настолько большая, что в дальнейшем целесообразно сделать упор на поиск различий поведения пользователей, которые добились активаций в первую и вторую сессию, также исследования поведения пользователей, добившихся активации во второй сессии, в первую сессию.

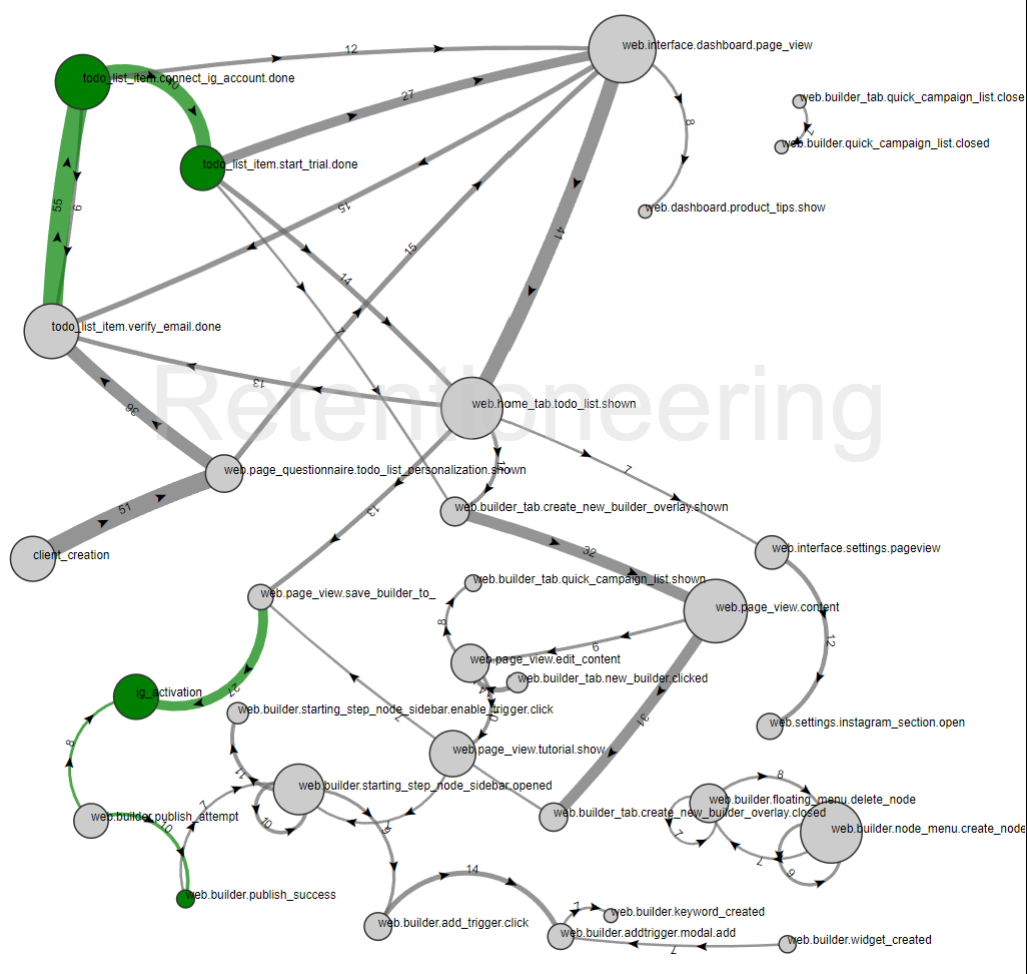
Рассмотрение поведения пользователей в сессиях

Рассмотрим всех пользователей, которые добились активации в первую и вторую сессию и посмотрим на их поведение в обеих сессиях:



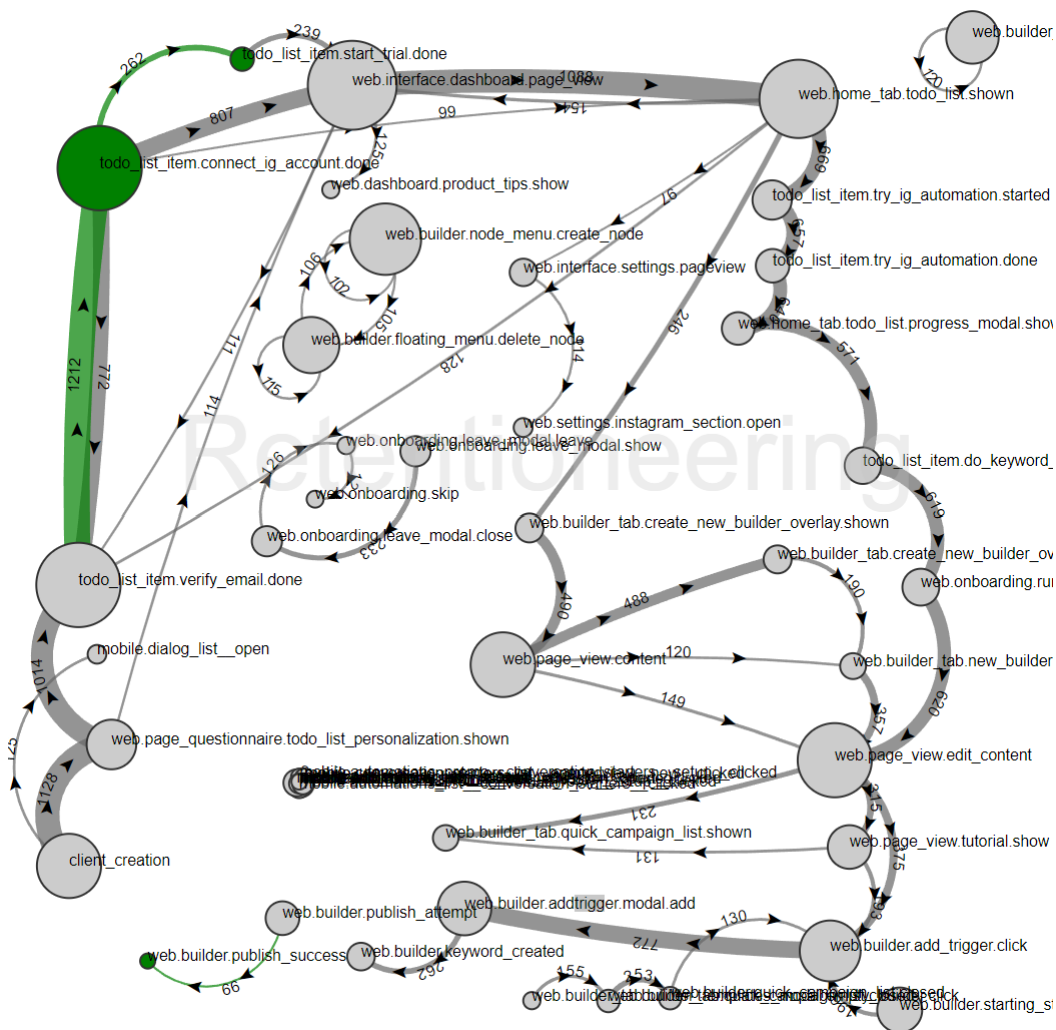
Видно, что часть пользователей прошла онбординг, а часть пользователей сократила путь и обошлась без него, также видна часть пользователей, которая сразу подключила триал-версию и виден уход пользователей в чтение обучения.

Рассмотрим первую сессию пользователей, которые в ней же добились активации:



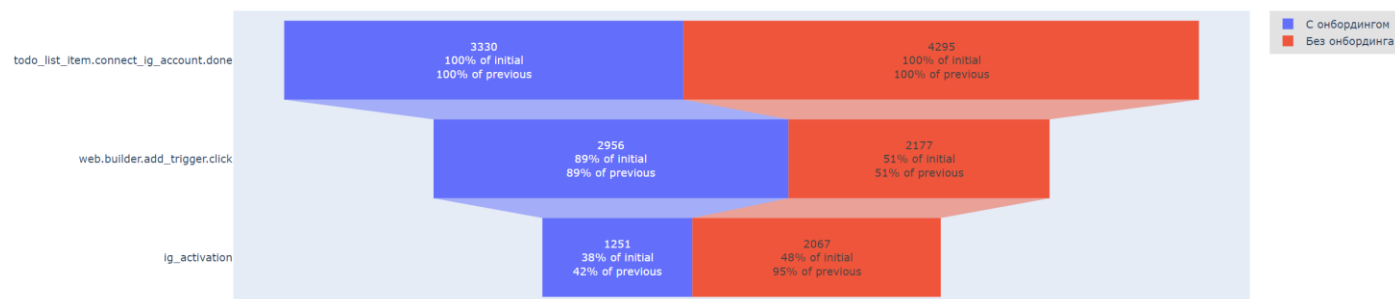
Большая часть пользователей подключает триал-версию сразу после подключения канала, затем сразу начинают работать с билдером, многие успевают посмотреть tutorial и разбираются как через триггер опубликовать канал. Эти пользователи смогли сразу разобраться в платформе и не используют онбординг.

Посмотрим, на что потратили время в первой сессии те, кто добился активации во второй:

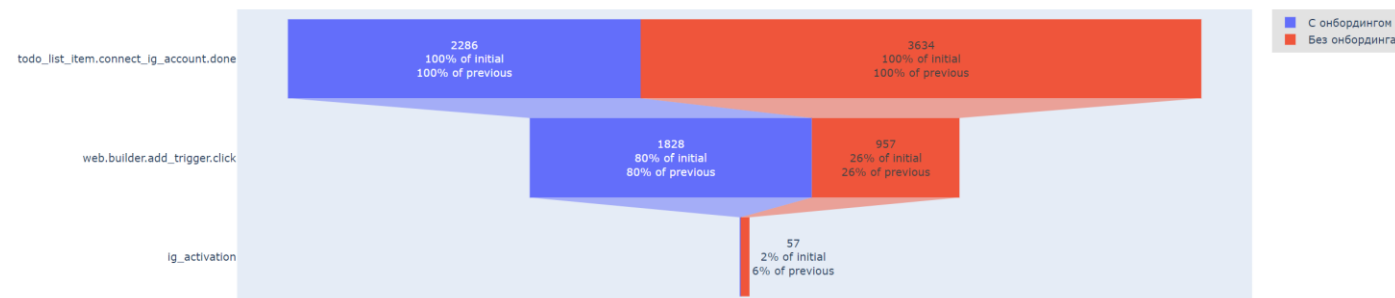


В основном пользователи всю первую сессию тратят на онбординг, при этом половине всё равно нужен tutorial, единицы добираются до триггера и успевают опубликовать канал.

В общей воронке довольно много пользователей проходит онбординг к моменту активации:



Однако в первой сессии их почти нет:

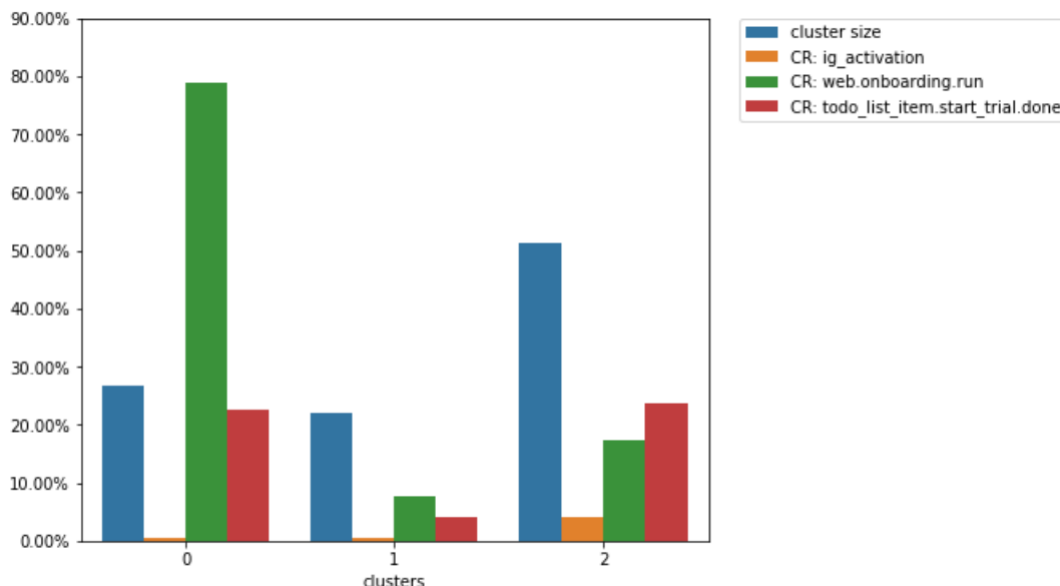


Среднее время от онбординга до попытки публикации канала составляет более 26 минут.

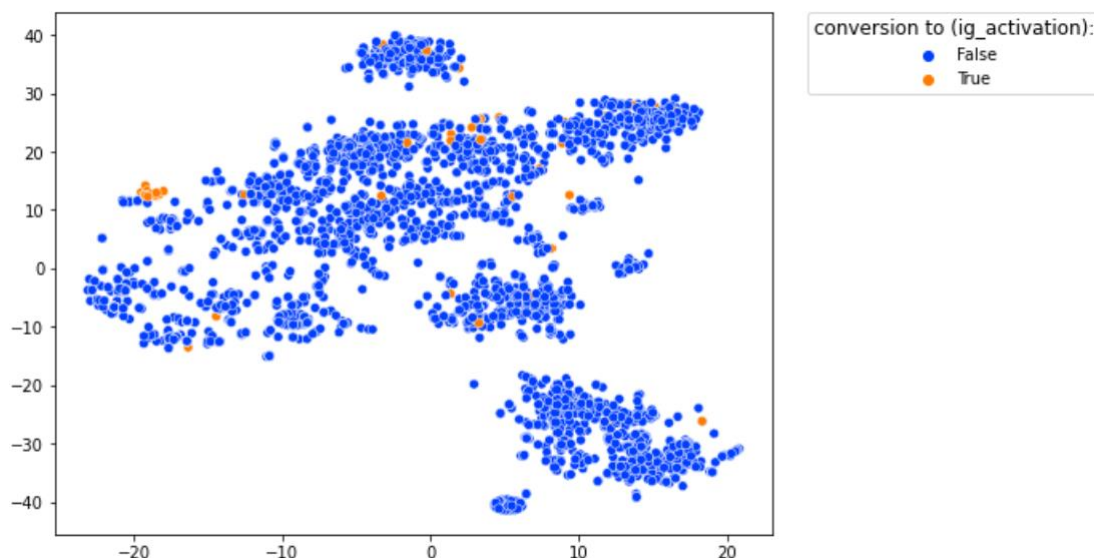
Выделить сегменты на основе данных поведения пользователей в первой сессии (сделать кластеры)

Выделим два сегмента: пользователей проходящих онбординг и пользователей, которые сразу подключают пробную версию.

Сегменты очень плохо поддаются кластеризации (другими словами, трудно найти общий кластер у сегментов):

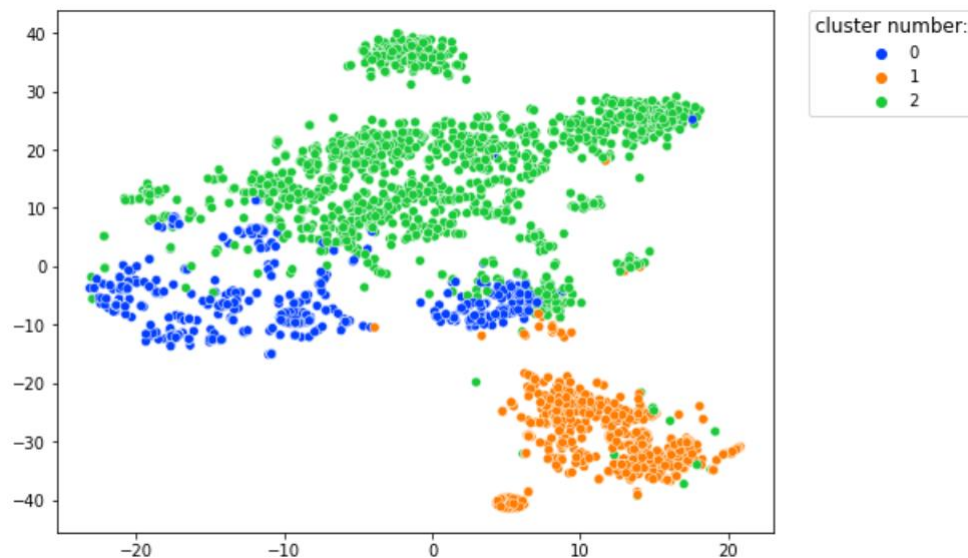


Это видно на визуализации близости поведения пользователей:

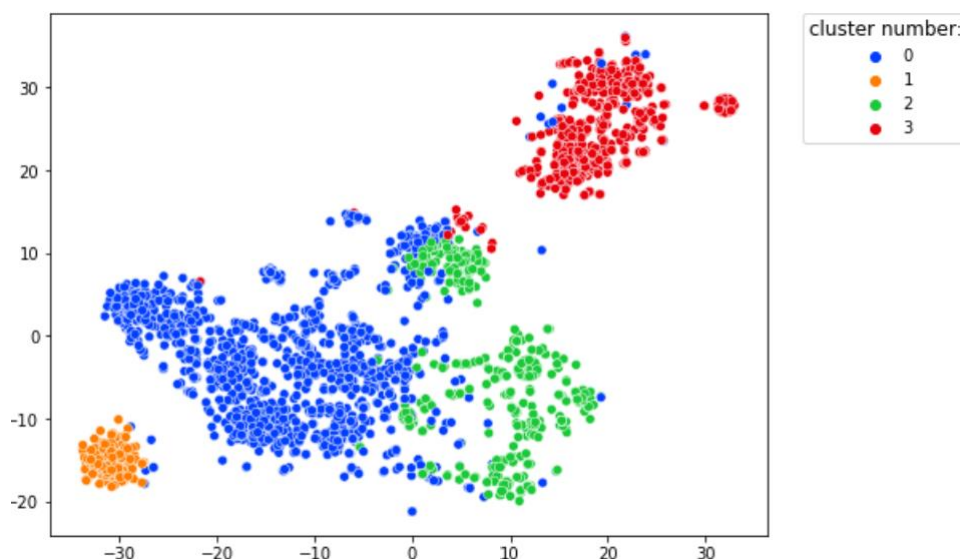
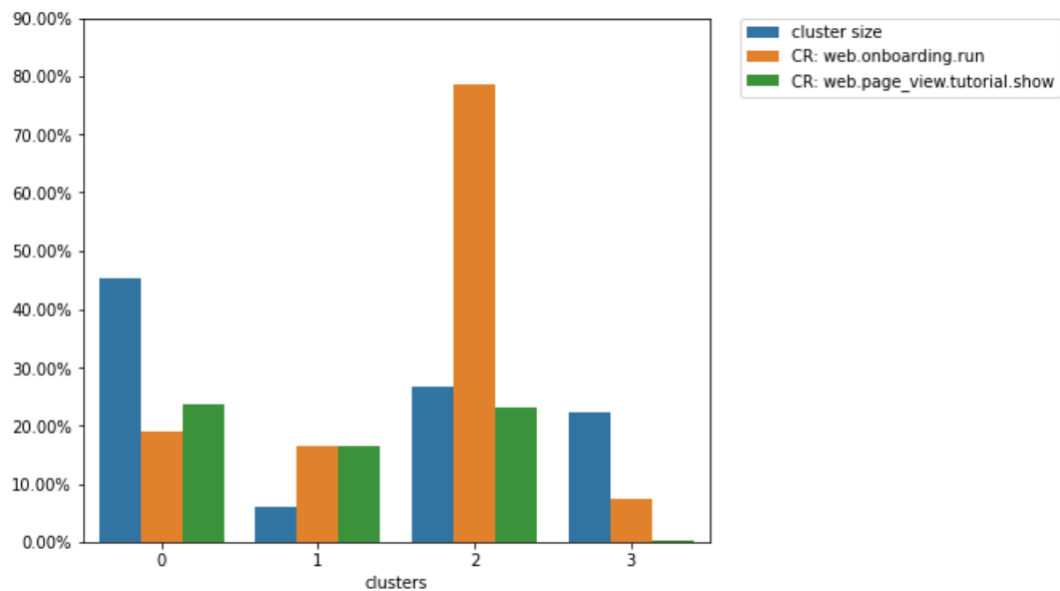


Пользователи, которые добились активации сделали это по-разному, у них нет единого пути. Это действительно опытные одиночки с собственным путём и поведением, найти в их поведении общую систему на данном объёме нельзя.

Найденные кластеры просто достаточно обширные, чтобы включить разное поведение:



Посмотрим на кластеры первой сессии пользователей, которые добиваются активации во второй сессии:



Здесь уже видны кластеры по поведению: 1 – смотрят и онбординг и tutorial. Онбординг им явно непонятен. Кластеру 3 онбординг понятен, но они и в целом имеют немного вопросов по платформе, так как меньше проходят онбординг.

Сформулировать рекомендации: какие сегменты стоит брать в первую очередь и почему

В первую очередь необходимо взять сегмент пользователей, которые проходят онбординг и дают активации во второй сессии: они дадут информацию о том, на какие шаги онбординга у них ушло много времени, что осталось непонятным и потребовало чтения tutorиала. Для цели повышения активаций в первой сессии мало смысла привлекать новых пользователей или уменьшать отток, так как только единичные опытные пользователи дадут прирост активаций в первой сессии, остальные потратят время на онбординг, дальнейшее чтение tutorиала и в лучшем случае дадут прирост во второй и далее сессиях.

Код: <https://www.dropbox.com/s/gica1tek3mv79g8/czhard.ipynb?dl=1>