Web-Analytics

Нечеловеческий трафик. Анализ данных веб-аналитики

группа 6

27 февраля 2024

О нас

Смирная София

- выпускница МИФИ (бак)
- моделирую на суперкомпьютерах
- реализовала несколько проектов с ML

Абраменко Александр

- студент ВШЭ, 3 курс
- работал с нейронками
- занимал позицию младшего аналитика

Халявин Олег

- Студент НГУ, 1 курс (маг)
- реализовал несколько проектов по анализу данных

О нас

Анисова Татьяна

- Выпускница ИТМО (бак)
- системный аналитик
- красный диплом музыкалки

Старков Никита

- от курьера до бизнес/системного аналитика
- строил свою онлайн-школу с ИИ
- студент РУДН, 3 курс

Задача

Представьте себя веб-аналитиком сайта www.tinkoff.ru. К вам приходит продуктовый аналитик Tinkoff Black и говорит, что при анализе АБ-теста была замечена подозрительная активность на форме: есть список id посетителей, которые в день генерируют тысячи заявок. Это негативно сказывается на корректности анализа АБ-теста.

На вас поставили задачу: научиться фильтровать НЧТ на данных веб-аналитики.

Обзор

Что такое НЧТ

Нечеловеческий трафик - это тип онлайн-трафика, который генерируется без участия человека.

Бот - программа, созданная для многократного повторения циклической задачи, скрипт для автоматизации веб-процесса.

01 Простые боты

- статический IP-адрес
- ID пользователя
- ID устройства

02 Сложные боты

- имитация движений мыши
- подмена IP
- случайные прокси

03 Ботнеты

- сеть хостов (много IP-адресов)
- могут заражать другие устройства

Цели НЧТ

01 Простые боты

Индексирование сайта для поисковых систем

Автоматизация рутинных задач

Мониторинг работы сайта и анализ активности пользователей

Обновление контента на сайте / его кража

Слив маркетинговых бюджетов

02 Сложные боты

Накрутка поведенческих факторов

Влияние на рекомендательные системы, скликивание конкурентов

Слив маркетинговых бюджетов, спам

Фишинг, кража контента

Выявление уязвимостей, атаки на сайт

03 Ботнеты

DDoS-атаки

Влияние на рекомендательные системы, скликивание конкурентов

Слив маркетинговых бюджетов

Спам

Майнинг

Ссылка на таблицу по типам ботов: https://docs.google.com/document/d/1hfl6WCV444qGbIEDToJc0xEdVTT0bJjXwhbX6IDEosM/edit

Полезные боты

- помогают пользователям взаимодействовать с контентом на сайтах
- предоставляют полезные сервисы

Веб-краулеры

- анализ сайтов
- индексация страниц для создания релевантной выдачи

Поисковые роботы в соцсетях

- сбор информации с сайтов, которыми поделились в соцсетях
- улучшение рекомендаций
- борьба со спамом

Маркетинговые боты

- автоматизация маркетинговых задач
- сканирование веб-сайтов для сбора информации о конкурентном рынке, ценах, популярности товаров

Чат-боты

- автоматизация коммуникации с потенциальными клиентами
- выполнение рутинных задач

Безвредные боты

- не приносят однозначного вреда или пользы
- чаще всего используются для развлечения или изучения возможностей ботов

Сканеры безопасности

- проверяют сайт на защищенность
- выявляют уязвимые места

Сервисы мониторинга

- отслеживают работоспособность сайта
- оповещают о возможных сбоях

Вредоносные боты

- боты, созданные для выполнения нежелательных или опасных действий в интернете
- могут вредить сайту и его посетителям или рекламным кампаниям

Скрейперы

- воровство контента и сбор контактных данных

Клик-боты

- создание мошеннических рекламных кликов

Спам-боты

- рассылка рекламы, вирусов
- ресурс для DDoS-атак.

Брутфорс

- получение доступа уровня Администратор
- кража учетных записей

Лид-боты

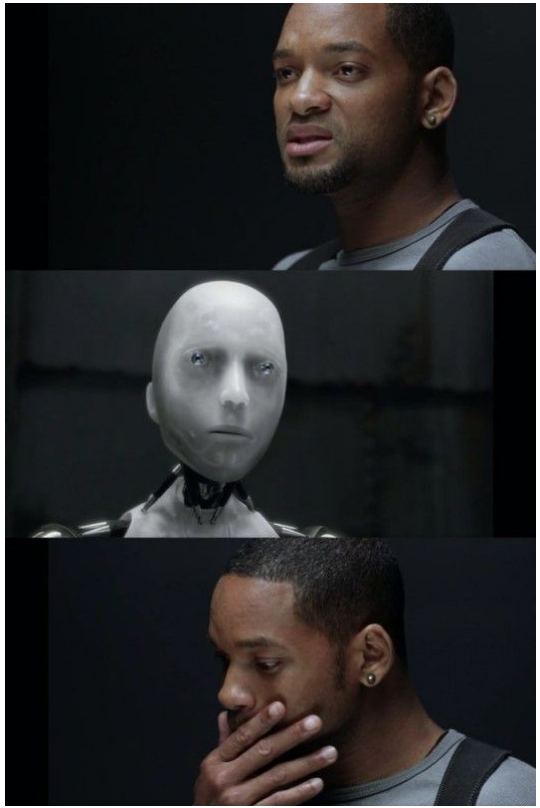
- заполнение контактных форм на сайте
- слив бюджета на увеличении SMS-рассылок
- DDoS-атаки

OTP-боты

- обход систем двухфакторной аутентификации
- кража учетных записей и личных данных

Характерные паттерны поведения

- резкое повышение активности
- однотипность поведения
- IP-адрес (простых ботов легко вычислить по его статичности)
- высокая скорость выполняемых действий
- несовместимость версии ОС и размеров экрана устройства
- аномальное кол-во новых устройств
- один источник
- высокий показатель отказов
- большое число прямых заходов на сайт
- устройства с устаревшими ОС и версиями браузеров



Какой вред НЧТ приносит аналитике

Искажение трафика

Ложные показатели взаимодействия с контентом

Накрутка поведенческих факторов

Усложнение подбора целевой аудитории, таргетинга и распределения рекламного бюджета

Вывод систем из строя

После успешной DDoS-атаки из-за неработоспособности сервиса аналитика вовсе не собирается

Неправильные аналитические выводы

Алгоритмы

Данные

Колонки в таблице:

- **t** - дата и время события
- **event_name** - название события
- **visitor_id** - уникальный id посетителя
- **visit_number** - порядковый номер визита
- **url** - адрес страницы
- **adblock_flg** - флаг наличия блокировщика рекламы
- **os** - платформа
- **os_version** - версия платформы
- **browser** - браузер
- **browser_version** - версия браузера

Присутствуют записи только о четырех **event_name**:

- **page_view** - загрузка страницы
- **autofill** - отработала механика восстановления заявки (посетитель заполнил форму на половину —> ушел с сайта —> вернулся через несколько дней —> информация, которую он заполнял ранее, заполнилась автоматически)
- **short_application** - короткая заявка (посетитель заполнил ФИО + номер телефона)
- **full_application** - полная заявка (посетитель полностью заполнил и отправил заявку)

Алгоритмы выявления вредоносного НЧТ

Бот-активность может быть связана с:

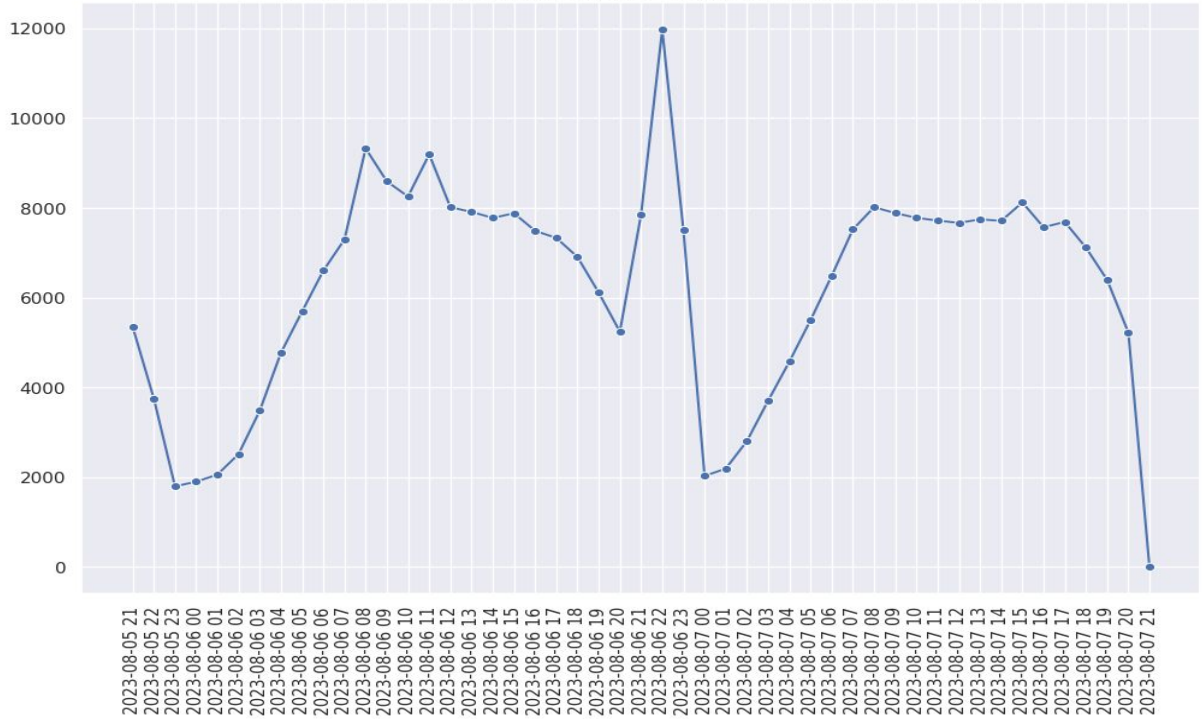
- Увеличением количества уникальных пользователей
- Высокой активностью (подозрительным количеством заявок)
- Старыми версиями устройств и браузеров
- Аномально коротким временем заполнения формы перед отправкой заявки
- Малым номером визита

Алгоритмы фильтрации пользователей:

- Удалять пользователей с высокой активностью по событиям
- Удалять пользователей, использующих старые версии (без блокировщика)
- Удалять пользователей с малым числом визитов и действий

Результаты

Распределение уникальных пользователей по часам

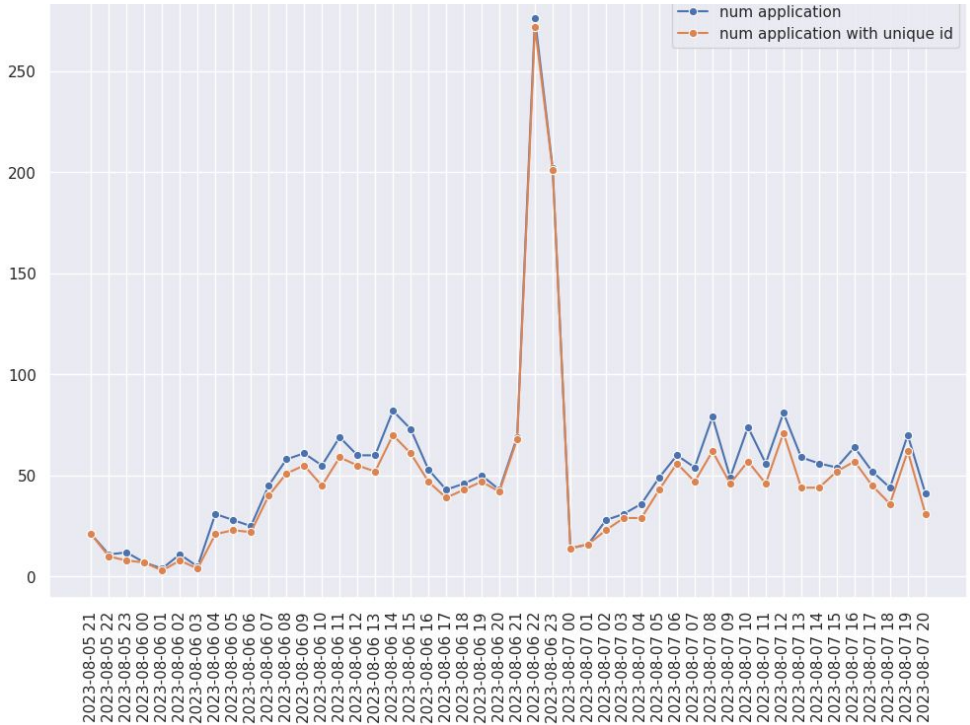


Заметен резкий пик с 20 часов 6 августа до полуночи 7 августа. Увеличение активности в позднее время является нестандартным. Также можно видеть увеличение числа пользователей с 8 до 11 часов 6 августа.

Распределение событий по часам

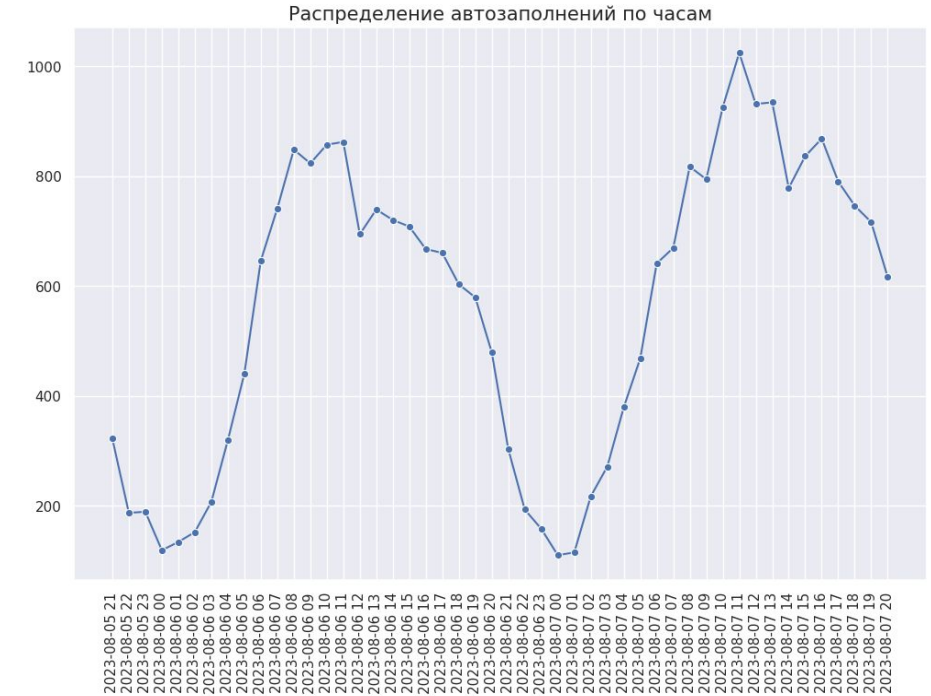
Рассмотрим статистику по каждому из событий, чтобы заметить возможную корреляцию с высоким количеством уникальных пользователей в конкретные временные интервалы.

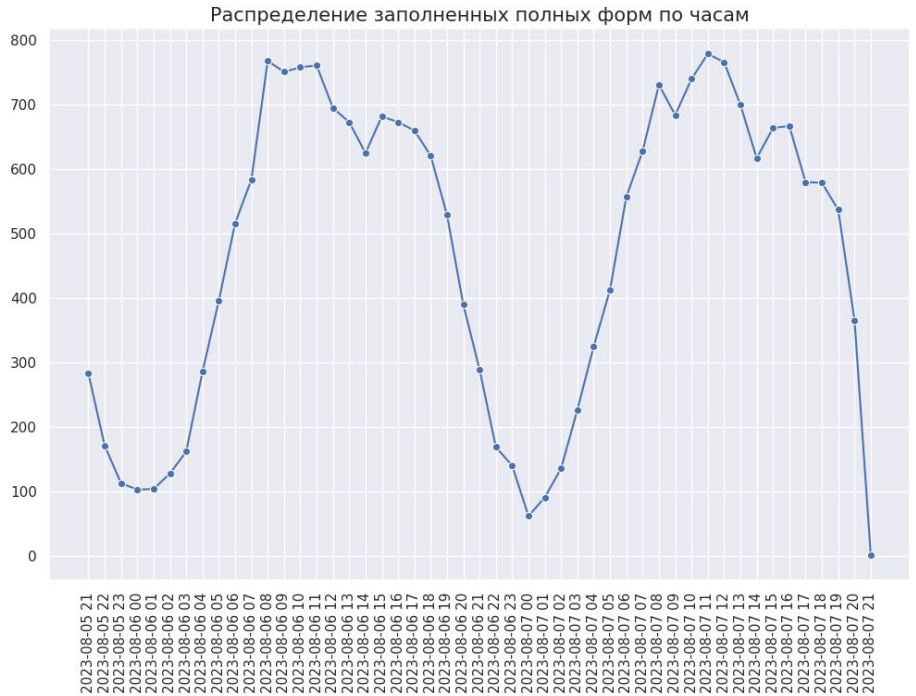
Распределение коротких заявок по часам



Аномальная активность в период с 20 часов 6 августа до полуночи 7 августа сохраняется и для пользователей, заполнивших короткую заявку.
Скорее всего, на сайт была совершена атака лид-ботов.
Атака была произведена либо с использованием смены прокси, либо ботнетом.

Распределение событий по часам





Выводы по графикам

Можно обратить внимание на корреляцию количества заполнений полных форм и автозаполнений в периоды с 8 до 12 часов дня 6 и 7 августа.

В эти часы проходит 69,7% от общего трафика.

		autofills
t	t	
2023-08-06	8	848
	10	857
	11	862
2023-08-07	10	925
	11	1024
	12	931
	13	934
	16	868

		full_applications
t	t	
2023-08-06	8	768
	9	751
	10	758
	11	761
2023-08-07	8	731
	10	740
	11	779
	12	766

Выводы по графикам (продолжение):

Отсутствие пиков на графиках автозаполнений и полных форм с 20 часов 6 августа до полуночи 7 августа, говорит о том, что боты не могли успешно отправить заявку и после заполнения формы не возвращались на сайт. Скорее всего, боты не могли пройти проверку мобильного телефона.

Следовательно, можно сделать промежуточный вывод, что в разные временные интервалы цели деятельности ботов были различные.

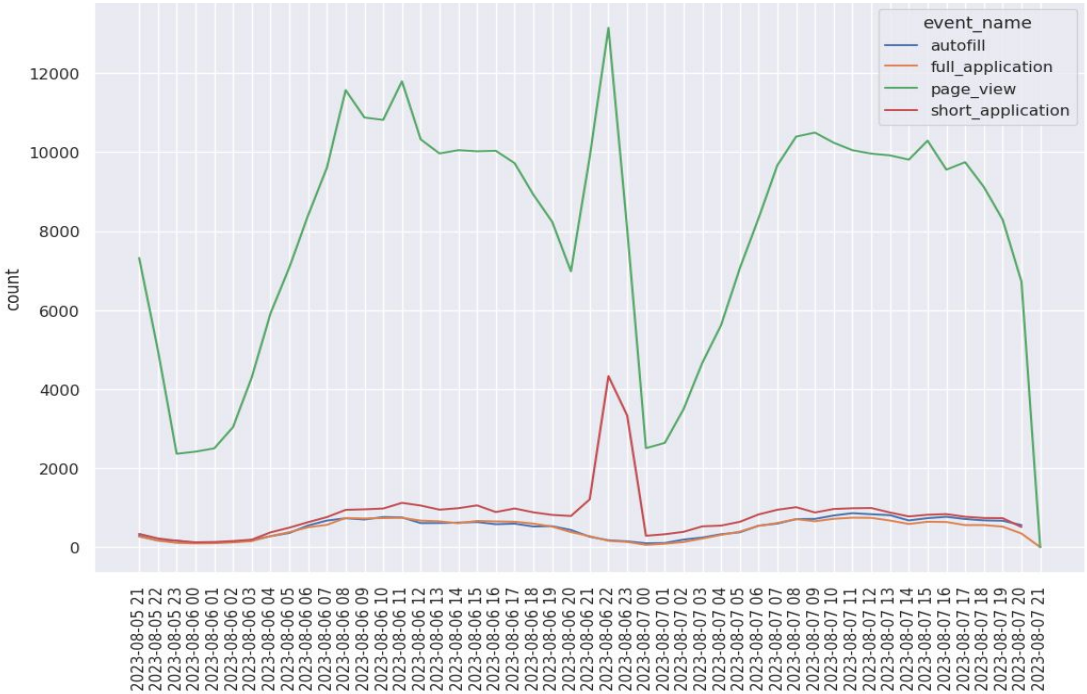
Применим первый алгоритм

Выявим пользователей с подозрительно высокой активностью по всем событиям и удалим их из датасета.

Количество пользователей после этого уменьшилось на 1,1%.

Как можно заметить, значительных изменений это не принесло, график сохранил пики.

Распределение событий для людей с менее чем, 10 визитами



Посмотрим на распределение числа визитов за два дня (с 21 часа 5 августа по 21 час 7 августа), определим сколько пользователей отклоняется на более чем два стандартных отклонения вниз.

Выводы по графикам:

Как можно заметить, поведение графика не меняется. Часть пользователей всё же совершала большое число визитов, но определение их в категорию ботов может быть затруднительно ввиду сложности определения точного количества действий, которое будет являться пороговым. Также посредством стандартных отклонений ни один пользователь не был выявлен как подозрительный, из-за чего следует отказаться от данной гипотезы.

Короткое время отправки заявки

Задав пороговое значение в 20 секунд для совершения действия, удалось выявить 1821 пользователя, что является 0,01% от общего числа уникальных пользователей, что приводит к выводу о слабой гипотезе.

Трудности определения продолжительности заполнения формы пользователем делает невозможным внедрение данного алгоритма для выявления подозрительной активности.

Устаревшие версии браузеров

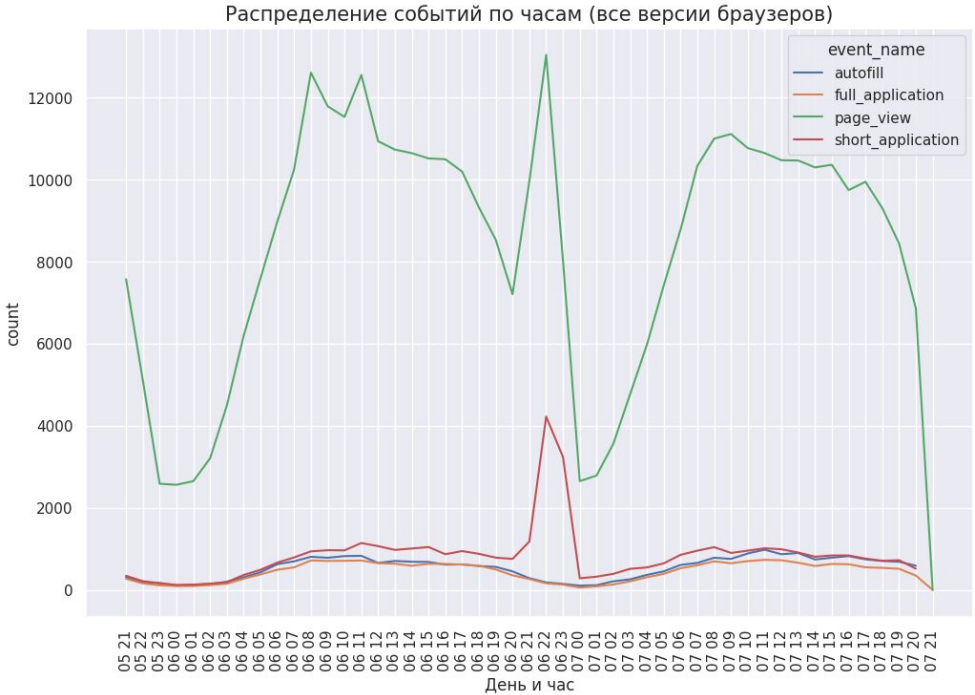
- Веб-браузеры большинства людей обновляются автоматически. Старые версии браузеров редко используются реальными людьми. Поэтому можно заключить, что трафик с устаревших версий, скорее всего, вызван ботами. Таким способом можно фильтровать самых разных ботов.
- Оставим только популярные браузеры (Chrome, Opera, Firefox и др.), это ~95% исходного датафрейма.
- Если оставить только их актуальные версии и старые, но с блокировщиком рекламы, получим 50% исходного датафрейма.
- Данные об актуальных версиях браузеров предоставляются W3Schools. Это бесплатный веб-сайт для обучения программированию, на который заходит более 60 миллионов человек ежемесячно.
- Будем смотреть на число событий каждого типа и конверсии из просмотра страницы в полную / короткую заявку

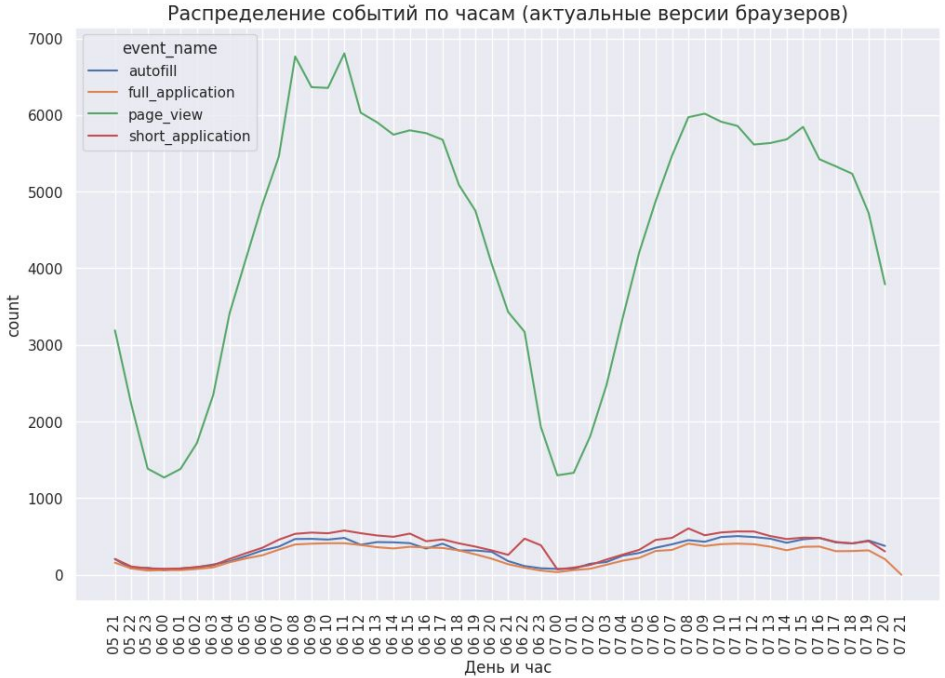
Актуальные версии браузеров

Браузер	Актуальные версии	Время выпуска	Процент устаревших
Chrome	108 и выше	декабрь 2022	4
Opera	94 и выше	декабрь 2022	0.3
Firefox	109 и выше	январь 2023	0.4
Edge	109 и выше	январь 2023	0.3
Safari	15 и выше	сентябрь 2021	0.3
MIUI	15 и выше	июль 2021	< 1
Samsung	15 и выше	июль 2021	< 1
Yandex	21 и выше	февраль 2021	< 1

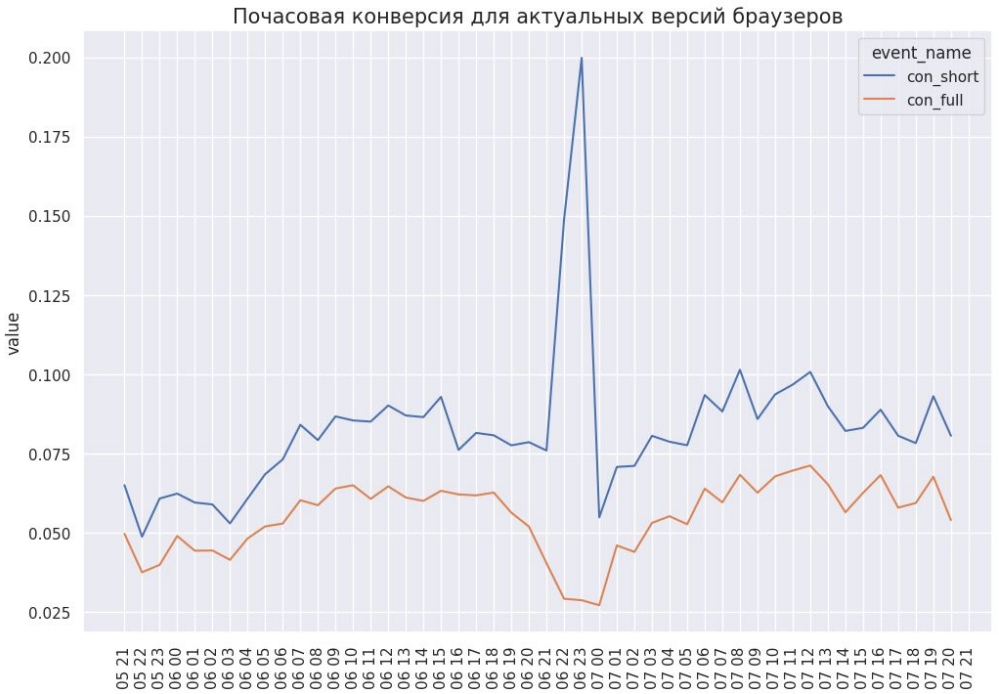
Результаты (Браузеры)

Результаты (Браузеры)

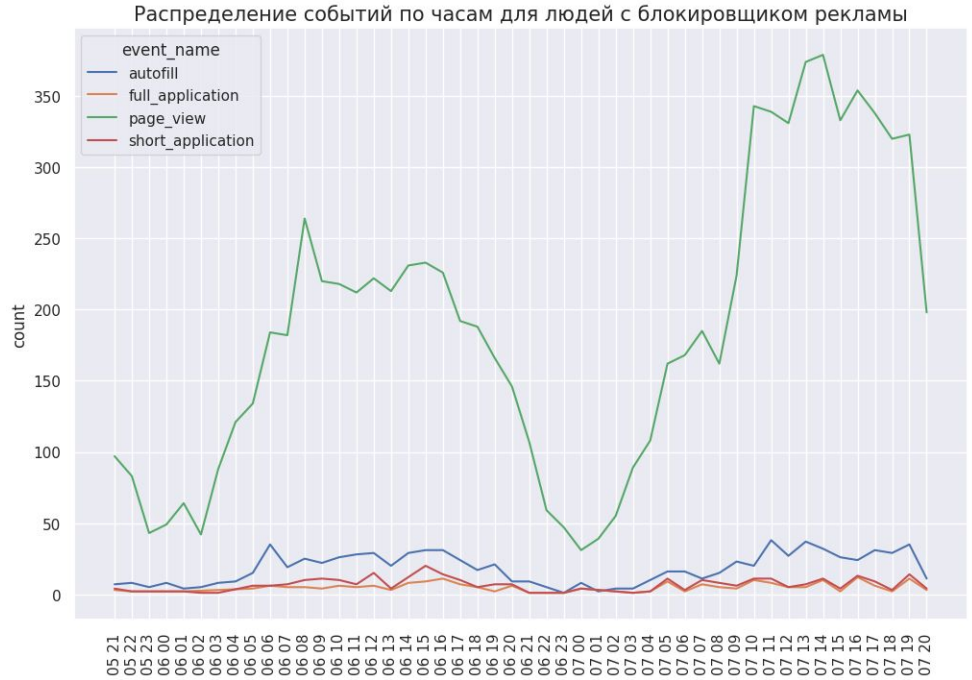




Результаты (Браузеры)



Результаты (Блокировщик рекламы)



Результаты (Блокировщик рекламы)

Таким образом фильтрация людей по блокировщику рекламы тоже может отсекать ботов. Очевидно, они не используют его.

Но блокировщик рекламы стоит у маленького числа людей (~7K и ~10K строк в датасете)

Поэтому такой поход будет не очень эффективен на наших данных. Его можно использовать в качестве вспомогательного. Поэтому люди со старыми версиями браузеров, но с блокировщиком рекламы не считались как боты.

Итоги

Итоги

- Удалось выяснить периоды активности ботов, а также в какое время каким было их контрольное действие
- Предположительно использовались клик-боты и лид-боты, целью которых было испортить статистику событий, уникальных пользователей, а также просмотры страницы
- Алгоритм использует библиотеки pandas и numpy, которые работают с векторами, поэтому время работы оптимальное
- Интегрировать в Tableau будет достаточно просто, так как алгоритмы построены на фильтрации, среднем и стандартном отклонении, которые есть в инструментах визуализации.

Итоги

- При аналитике данных огромную опасность представляют клик-боты, так как они искажают значения конверсии и абсолютные значения, которые очень важны для понимания желаний пользователей, а также актуальности продукта
- Недостатки алгоритмов в том, что пороговое значение для отсечения ботов является трудно определимым. Так как они меняют прокси, то воспринимаются как отдельные пользователи. В абсолютных значениях будет заметен скачок, но при вычислении активности по пользователям, подозрительную активность тяжело зафиксировать. Отсекать ботов можно по конверсии в следующий этап воронки, однако при таком подходе часть обычных пользователей также может быть отнесена к ботам.
- Алгоритмы хороши для определения выбросов по значениям и проверки, являются ли они следствием действий ботов или причина кроется в другом

Потенциальные точки роста алгоритма и данного подхода

При большем количестве данных можно построить линейную модель, взяв за признаки высокие значения активности и старые версии. Сформировать линейную модель для строгого определения пороговых значений, чтобы снизить процент захвата людей или же модель классификации, чтобы сразу отсеивать подозрительных пользователей.
Также можно использовать методы кластеризации и классификации.

Будем рады ответить на ваши вопросы