Конкурс Avito на Boosters.pro. Задача о поиске роботов.

Александра Денисова

Тренировки Machine Learning, 2017

Содержание

Исходные данные

Лог действий пользователей мобильного приложения:

- Группы действий
- Дата и время
- Категория объявлений
- Локация
- Тип приложения (iPhone/Android)
- IP-адрес

Объем данных: **11 млн.** строк, **438 тыс.** уникальных устройств.

Проблема: Непонятно как Avito размечал выборку

1. Гипотезы: аномальное поведение: по количетсву операций либо по пропорциям;

2. Пред валидация: подсказка организаторов о большом количестве роботов в приложении Android;

3. Валидация гипотез по лидерборду.

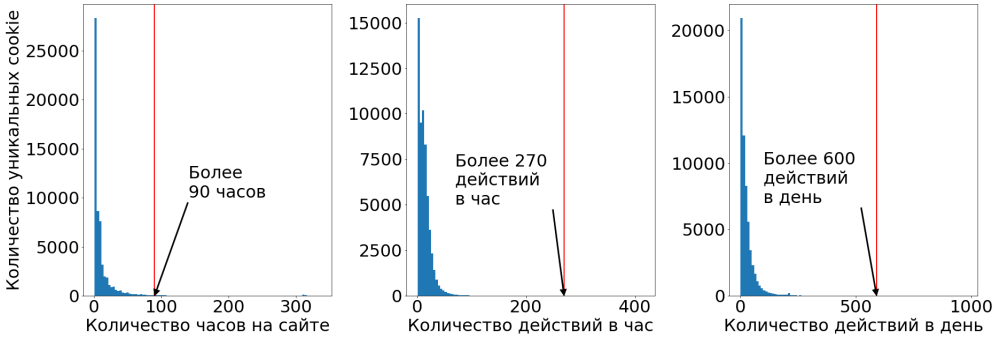
Отсечение по одному действию на устройство

85% кук встречаются 1 раз

99% из них — просмотры Android

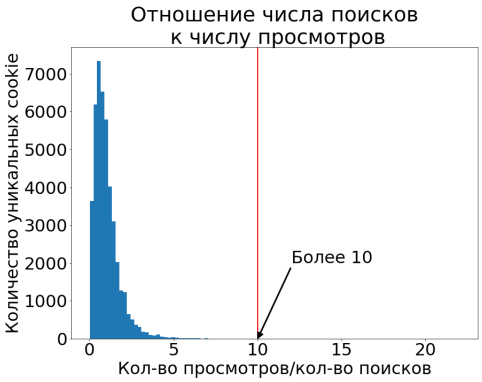
Отсечения по частоте действий

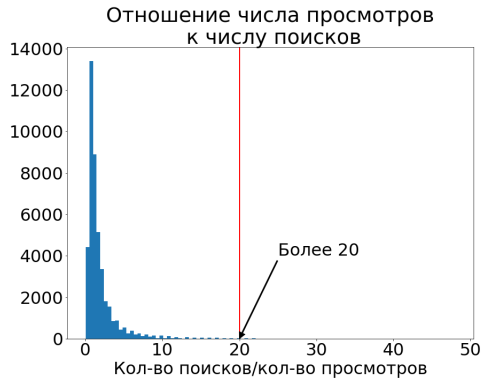
Аномальное количество действий, времени на сайте.



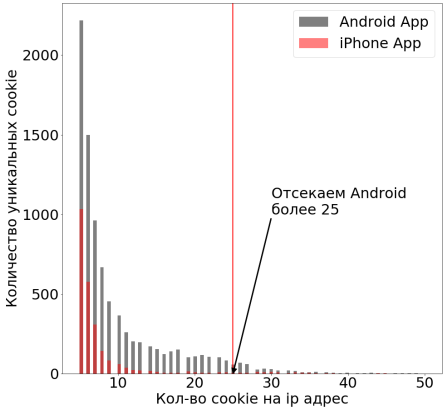
Отношение количества просмотров и поисков

86 % всех действий — поиск и просмотр объявлений.





Большое количество различных устройств Android под одним IP



Отсечение по типам действий

Аномальное поведение с Android устройств.

eventype_name	Android App	iPhone App	difference
Item View	453	8	56.62
User Items,Item View,Items Search	254	7	36.28
User Items,Item View,Items Search,Main Page	320	10.	32
Item View,Item View Phone,Items Search	1406	59	23.83
Main Page,Items Search,User Login,Favorites It...	284	28	10.14
User Login	306	58	5.27
Items Search	1520	363	4.18
Item View,Items Search	7675	1940	3.95
Item View,Item View Phone,Items Search,Main Page	1301	375	3.46
Items Search,Main Page	978	505	1.93
Item View,Items Search,Main Page	6856	3827	1.79
Item View,User Login,Items Search	303	331	0.92

Пробовала, не полетело

- Отсечение по UserAgent — по лидерборду Ruby не робот
- Отсечение по версии клиента Avito — почти все 3 и 4 версии уже отмечены как роботы, 6 и 7 как не роботы.
- Отсечение по маске ip (первые 3 цифры ip)

Идеи кончились - получаем **3** место

Отсечение	Кол-во роботов
1 просмотр Android	360 000
Только просмотры с Android	500
Кол-во времени/действий на сайте	300
Соотношение поиска и просмотров	500
Много устройств с 1 IP	7 000

- Создаем фичи на куку (dummy, count distinct etc.)
- Делим выборку на 5 частей: по очередно на 4х учим forest на 5й предсказываем является ли кука роботом по нашему алгоритму
- Получаем «вероятность» что кука — не робот.
- Вероятность большая, кука робот - считаем не роботом;
- Вероятность маленькая, кука не робот - считаем не роботом.
- Сабмитим
- Пока скор на лидерборде улучшается - повторить операцию (два раза).

Результат

#	Участник	Решений	Результат
1	Павел Свинцов	8	0.9735
2	Александра Денисова	25	0.9695
3	Алексей Рыбалко	9	0.9661
4	Николай Ванаев	1	0.9661

Спасибо!