ОТЧЕТ ПО ПРОЕКТУ
"ИДЕНТИФИКАЦИЯ ИНТЕРНЕТ-ПОЛЬЗОВАТЕЛЕЙ"

Цель проекта

Идентификация пользователя по последовательности из нескольких веб-сайтов, посещенных подряд.

Описание данных

Исходные данные
Имеются данные с прокси-серверов Университета Блеза Паскаля. Их вид: ID пользователя, время захода, посещенный веб-сайт. Для целей проекта взяты выборки по 3, 10 и 150 пользователям, размещенные в отдельных файлах следующего вида:

10users/user0031.csv

	timestamp	site
0	2013-11-15 08:12:07	fpdownload2.macromedia.com
1	2013-11-15 08:12:17	laposte.net
2	2013-11-15 08:12:17	www.laposte.net
3	2013-11-15 08:12:17	www.google.com
4	2013-11-15 08:12:18	www.laposte.net

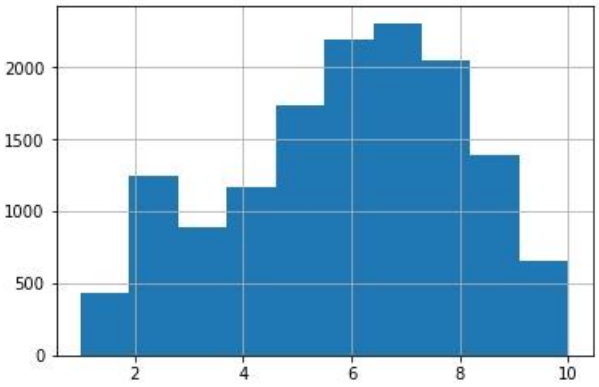
Обработанные данные и признаки
Исходные данные преобразуются в условные сессии следующим образом: для каждого пользователя, начиная с нулевой записи (i=0), берется L последовательных записей (сессия длины L). Данные этих записей последовательно заносятся в новую таблицу (DataFrame): адрес и время захода на первый в сессии сайт, адрес и время захода на второй и т.д. Затем происходит переход на ширину окна W (т.е. на запись i+W) и вновь выбирается L последовательных записей для новой таблицы данных. Адреса сайтов заменяются их индексами (для сопоставления адреса и индекса создается словарь). Если достигнут конец файла, а L последовательных записей не набралось, ставим на оставшиеся в сессии позиции нули. В результате получается подобная таблица сессий:

Предобработка данных

Перебрав несколько возможных вариантов параметров длины сессии и ширины окна, остановимся на L=10 и W=10. На основе обработки данных с этими параметрами преобразуем последовательности сайтов по принципу «мешка слов». Создадим новые разреженные матрицы, в которых строкам будут соответствовать сессии из 10 сайтов, а столбцам – индексы сайтов. На пересечении строки i и столбца j будет стоять число n_{ij} – сколько раз сайт j встретился в сессии номер i. Первый столбец (0, нет сайта) удалим.

Первичный анализ признаков

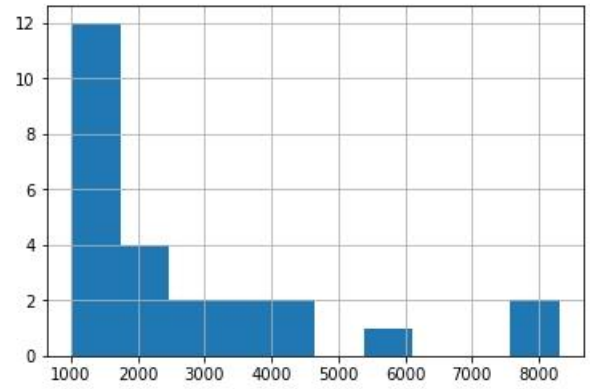
Проанализируем данные по посещенным сайтам для выборки из 10 пользователей.
Распределение числа уникальных сайтов в каждой сессии из 10 посещенных подряд сайтов



И визуальный анализ, и проверка по критерию Шапиро-Уилка отвергают гипотезу о нормальности этого распределения.

Биномиальный критерий для доли не отвергает гипотезу о том, что пользователь хотя бы раз зайдет на сайт, который он уже ранее посетил в сессии из 10 сайтов.

Распределение частоты посещения сайтов (сколько раз тот или иной сайт попадается в выборке) для сайтов, которые были посещены как минимум 1000 раз.



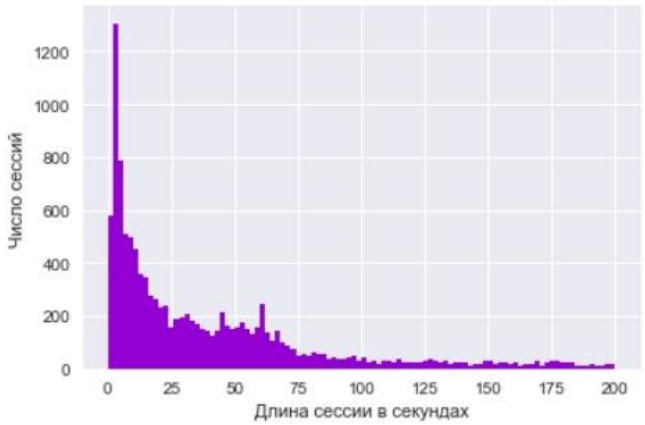
Создание дополнительных признаков

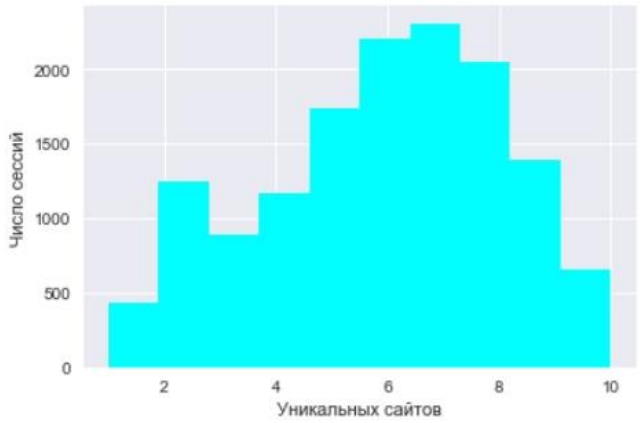
Построим дополнительные признаки на основе таблицы сессий.

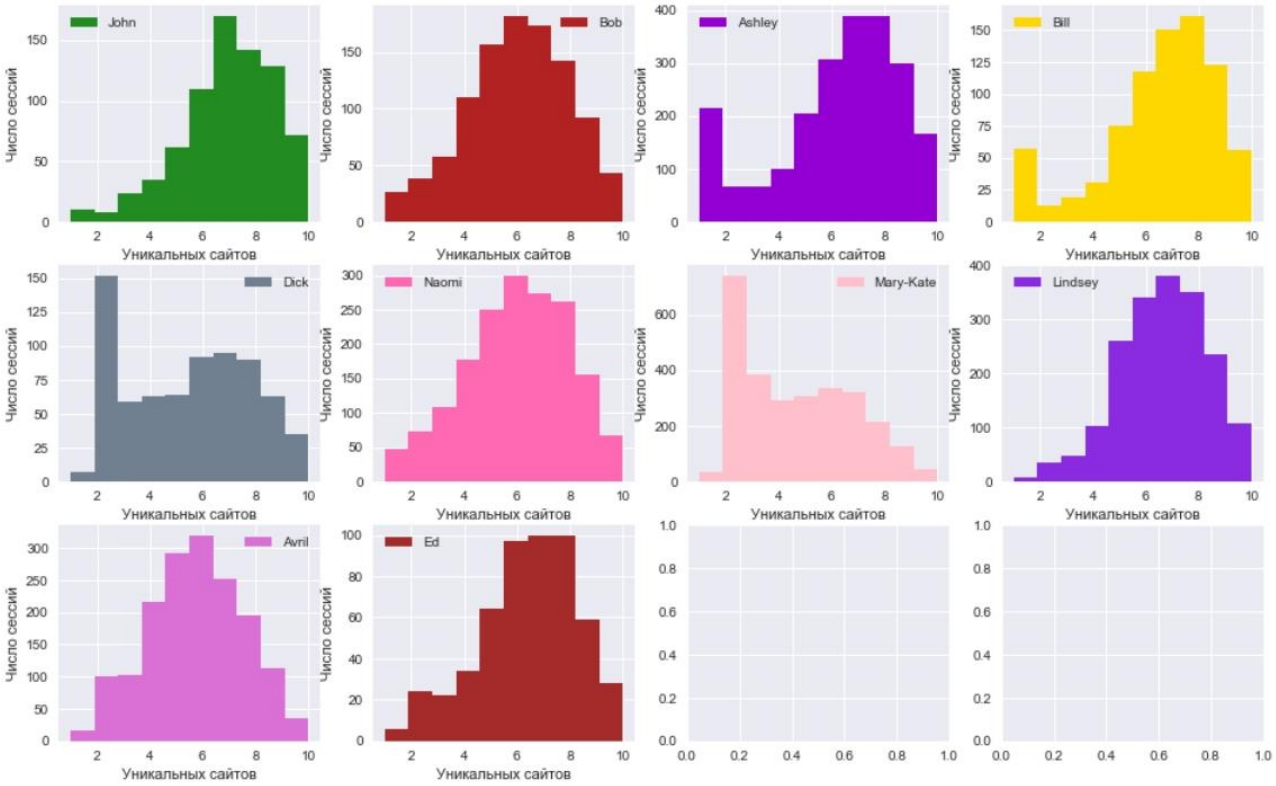
Дополнительные признаки:
- длительность сессии
- число уникальных посещенных сайтов
- день недели
- час начала сессии
- число сайтов в сессии из 30 самых посещаемых
- признак того, попадает ли время начала сессии в т.н. «рабочее время» - с 9:00 до 20:00

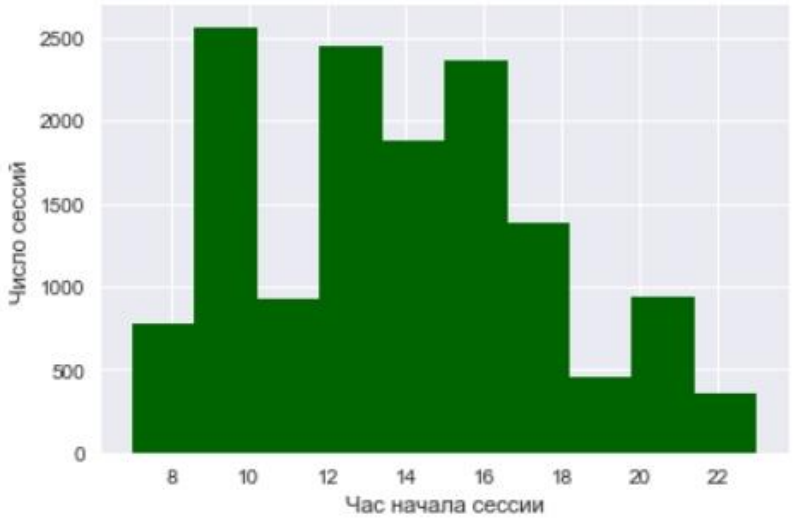
Первичный визуальный анализ признаков

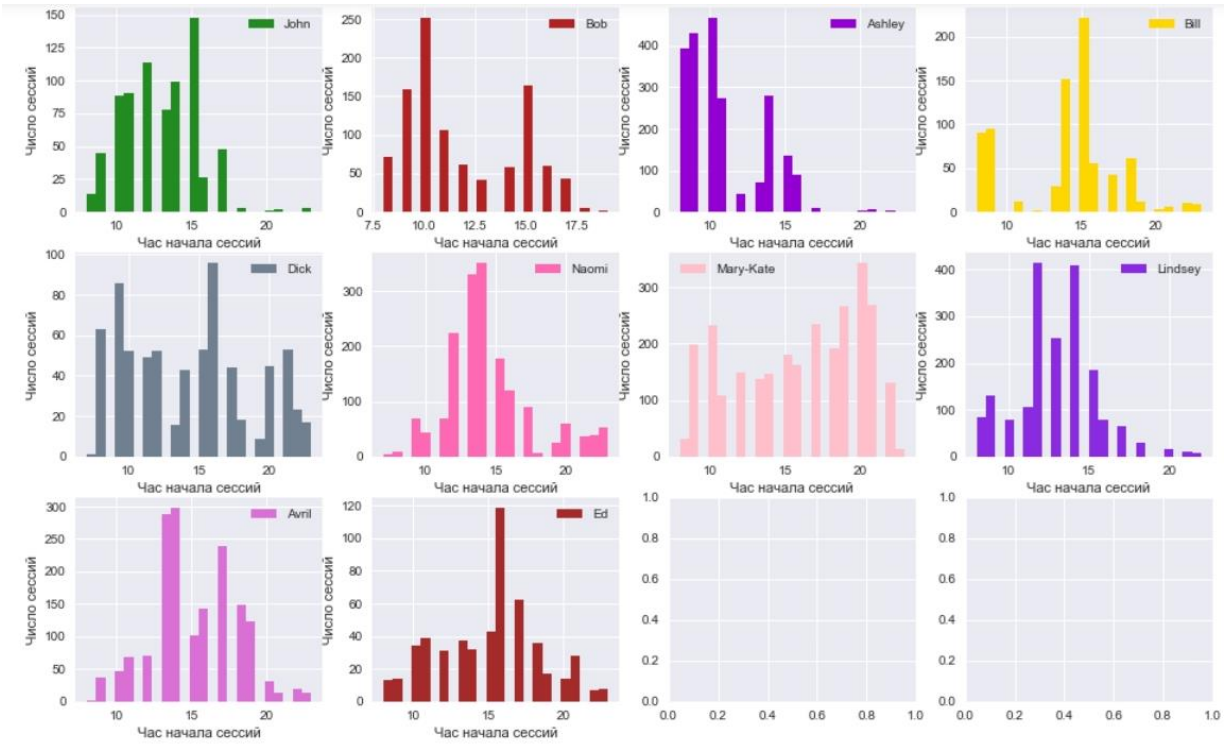
Построим гистограммы распределения признаков из выборки по 10 пользователям, в частности, в разрезе пользователей.

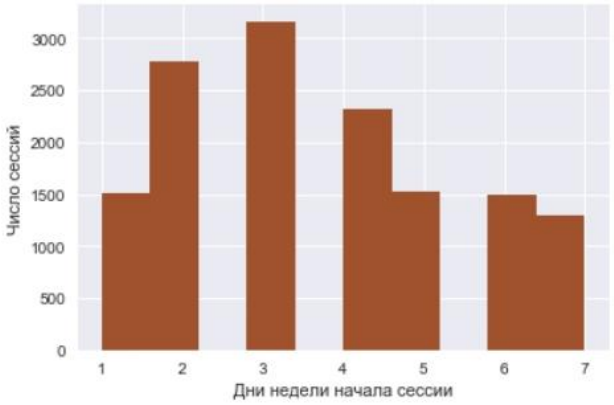


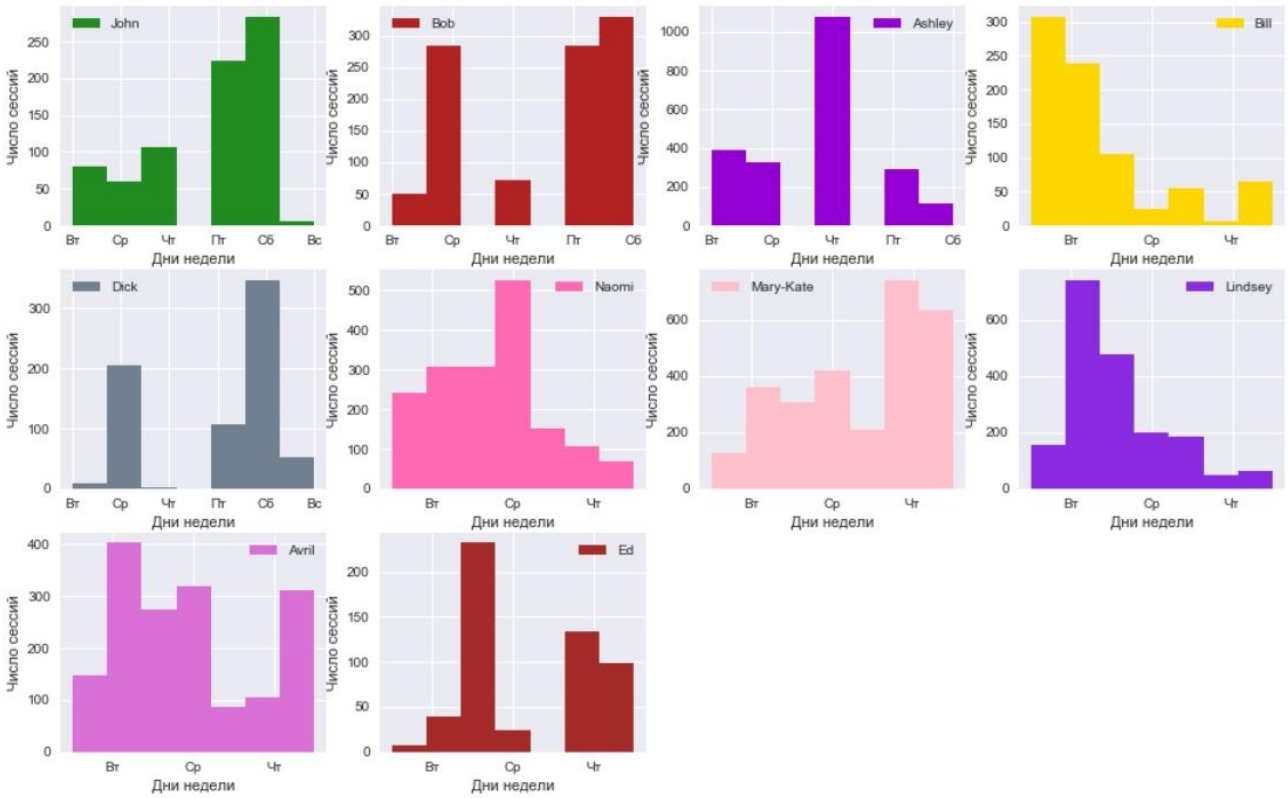


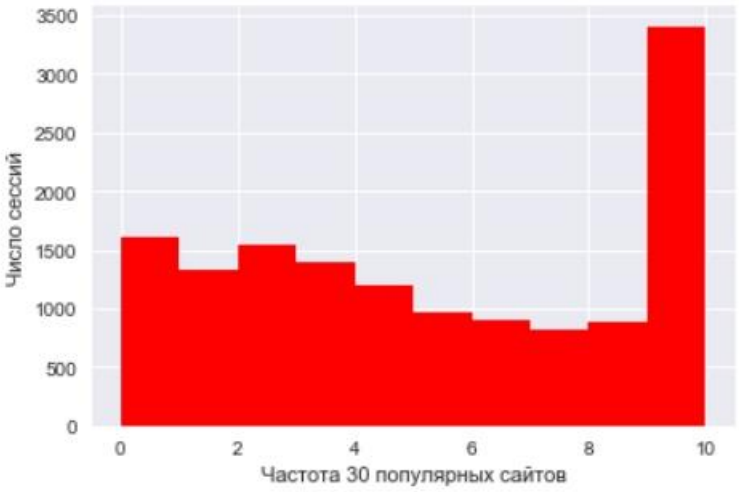


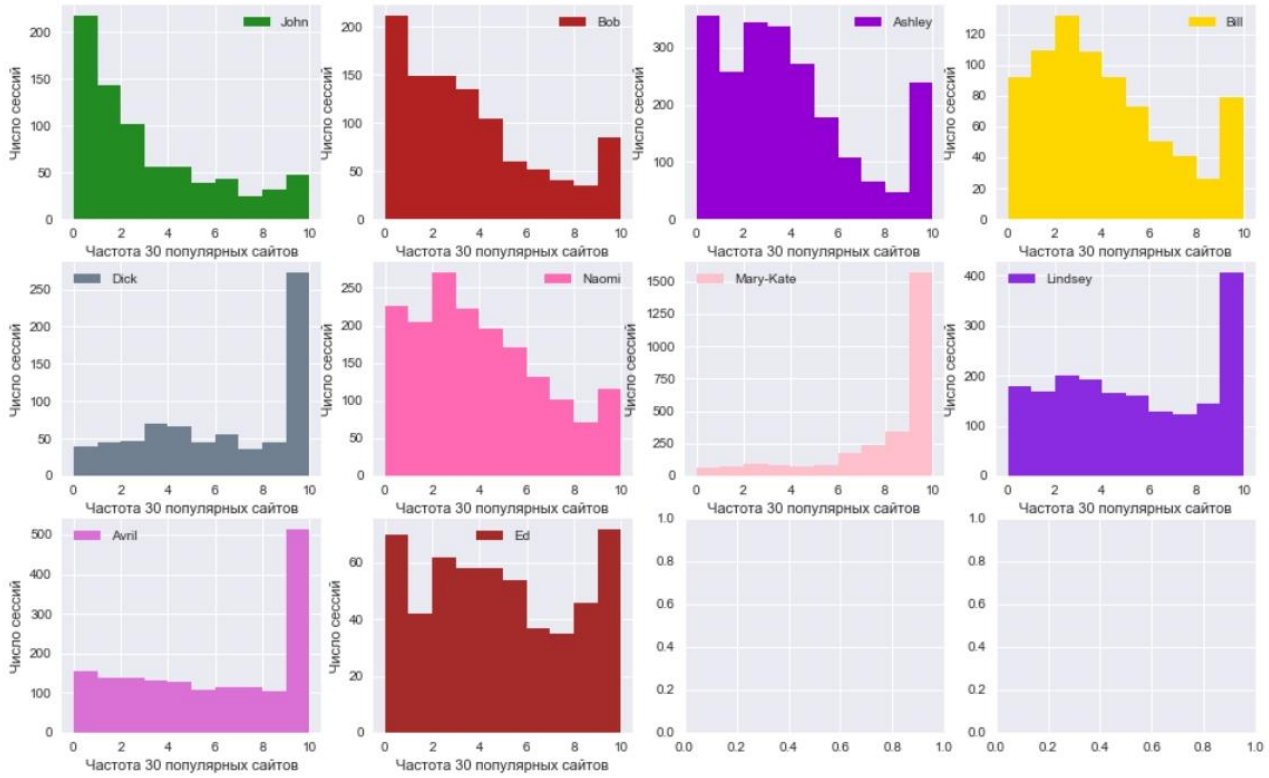

Число сессий
Час начала сессии

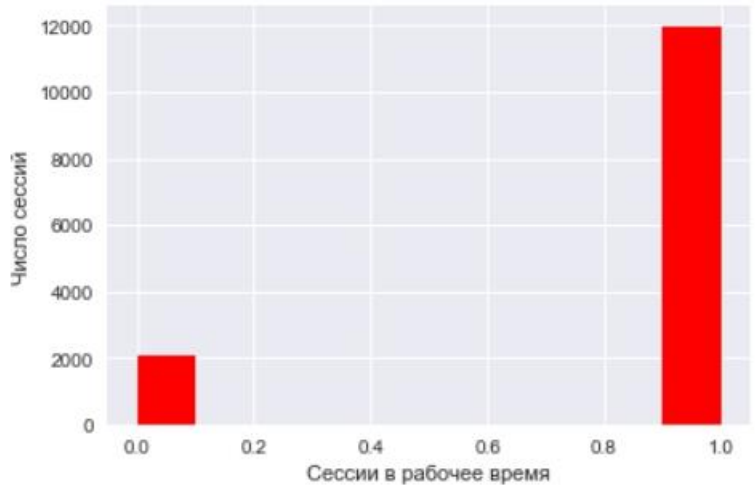

John
Bob
Ashley
Bill
Dick
Naomi
Mary-Kate
Lindsey
Avril
Ed
Число сессий
Час начала сессий

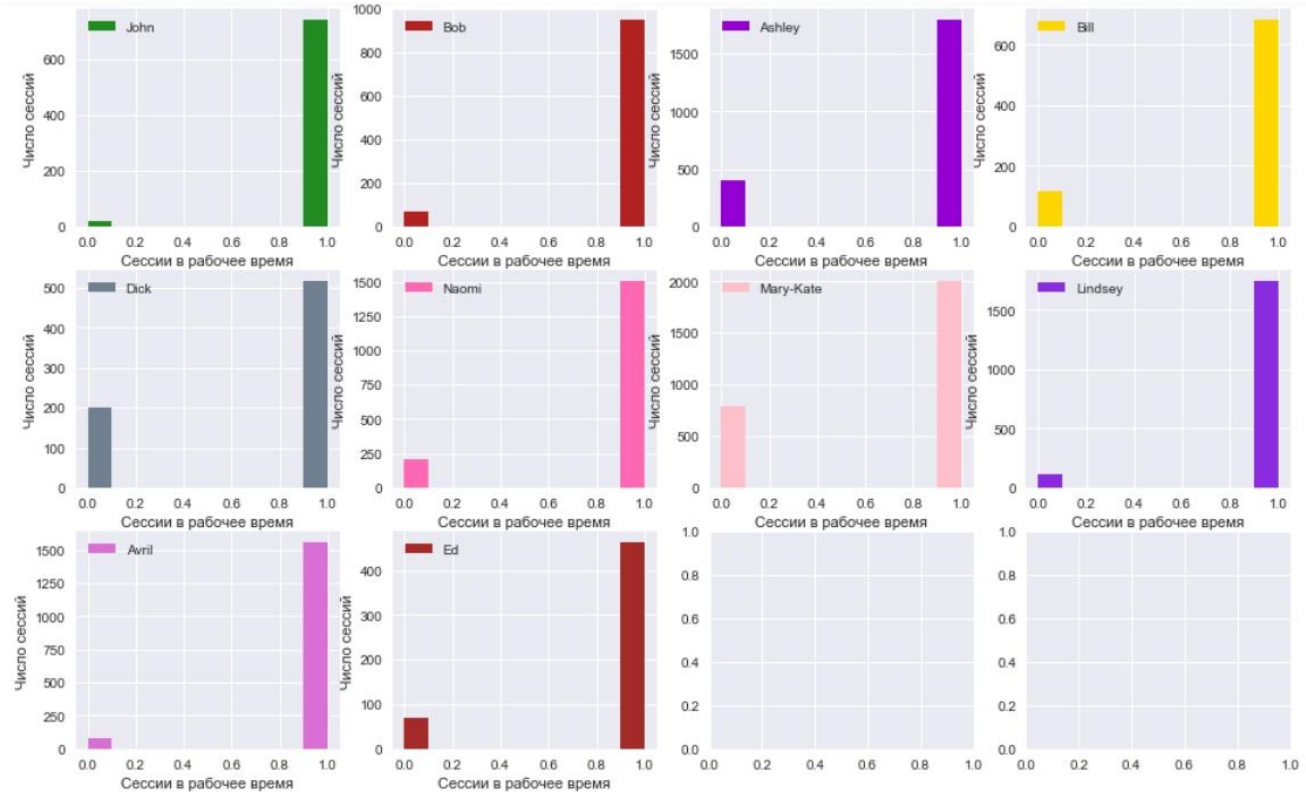

Число сессий
Дни недели начала сессии


John
Bob
Ashley
Bill
Dick
Naomi
Mary-Kate
Lindsey
Avril
Ed
Число сессий
Дни недели
Вт
Ср
Чт
Пт
Сб
Вс


Число сессий
Частота 30 популярных сайтов


John
Bob
Ashley
Bill
Dick
Naomi
Mary-Kate
Lindsey
Avril
Ed
Число сессий
Частота 30 популярных сайтов

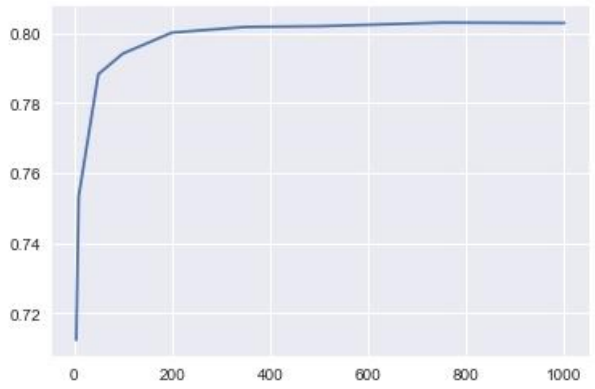




Визуальный анализ позволяет наблюдать для выбранных новых признаков отличия в распределениях для разных пользователей, что говорит о целесообразности включения новых признаков в модель.

Кросс-валидация

В результате перебора следующих моделей с параметрами по умолчанию: метод к ближайших соседей, случайный лес, логистическая регрессия и линейный SVM – был выбран метод случайного леса. Для него была проведена кросс-валидация с поиском по сетке числа используемых деревьев. Для кросс-валидации выборка перемешивалась и разбивалась на 3 фолда. На параметрах числа деревьев от 5 до 1000 наилучшее качество по метрике «число совпадений прогноза (accuracy)» показал случайный лес с 750 деревьями:



Оценка модели

Выбранная метрика accuracy показывает число точных совпадений предсказания модели и разметки. Она применима, поскольку выборка не сильно разбалансирована. Обучение модели на всей выборке для 10 пользователей дало качество на валидационной подвыборке (30% от всей выборки) – 82.63%.

Выводы

Построенная модель может применятся для выявления подозрительной активности на аккаунтах пользователей, например, в корпоративной электронной почте. Для улучшения модели можно было бы рассмотреть большее количество комбинаций «ширина окна»-«длина сессии» и выбрать наилучшую по кросс-валидации.