Анализ данных: финальный проект

Идентификация интернет-пользователей

Байкалов Р.А.

Введение

В данном отчете представлены результаты работы над финальным проектом специализации «Машинное обучение и анализ данных».

В проекте решалась задача по идентификации интернет-пользователей. Примером такой задачи может быть, например, идентификация взломщика почты по его поведению. Под поведением пользователя может пониматься история просматриваемых страниц, стиль работы с электронными письмами (чтение, удаление, установка флажков) и даже движение мышкой.

В данном проекте ставилась следующая задача: идентифицировать пользователя по последовательности посещенных им сайтов. Предполагается, что пользователи по-разному переходят по ссылкам и это помогает их идентифицировать. Кто-то сначала в заходит в почту, потом на спортивный сайт, затем читает новости, заходит в социальные сети, потом начинает работать, кто-то сразу начинает работать.

Полученные в ходе реализации проекта подходы были опробованы в соревновании Kaggle «Identify Me If You Can».

Проект выполнен с использованием языка и библиотек python 2.7 в Jupiter Notebook.

Исходные данные

В качестве исходных данных используются данные с прокси-серверов Университета Блеза Паскаля. Данные имею следующий вид. Для каждого пользователя заведен csv-файл с названием user****.csv (где вместо **** 4 цифры, соответствующие ID пользователя). Посещения сайтов в файле записаны в формате: ***timestamp, посещенный веб-сайт.*** Первые пять строк одного из файлов приведены на рисунке 1

	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

Рисунок 1 – Формат файлов исходных данных

Данные имеются по 3000 пользователей. Из-за требованиям по вычислительным ресурсам проект ограничивается 150 пользователями.

Подготовка данных

Для работы над проектом формируются выборки из 3, 10, и 150 пользователей. Первые две выборки используются для иллюстрации работы и отладки кода, рабочая версия проекта реализуется на выборке из 150 пользователей.

Для решения поставленной задачи необходимо сформировать обучающую выборку. Объектами в выборке будут сессии пользователей, состоящие из сайтов и других признаков. Сессией может быть последовательность из ***n*** подряд посещенных сайтов. Длина сессии ***n*** является параметром, которую необходимо подбирать при обучении прогнозных моделей. Также сессии из ***n*** сайтов могут формироваться с использованием скользящих окон размера ***k***, который также является параметром. В таком случае сессии будут пересекаться.

На рисунке 2 приведен DataFrame для выборки из трех пользователей с длиной сессии 10 сайтов.

	site1	site2	site3	site4	site5	site6	site7	site8	site9	site10	user_id
0	4	2	2	11	2	1	8	5	9	7	1
1	4	1	1	1	0	0	0	0	0	0	1
2	4	2	6	6	2	0	0	0	0	0	2
3	3	1	2	1	2	1	1	5	10	3	3
4	3	1	2	0	0	0	0	0	0	0	3

Рисунок 2 – DataFrame для выборки из трех пользователей с длиной сессии 10 сайтов.

В столбцах **site*** записаны индексы посещенных сайтов. Меньшие индексы соответствуют более популярным сайтам. В случае когда количество сайтов пользователя некратно длине сессии появляются короткий сессии, в которых недостающие сайты заменяются нулями. В столбце **user_id** записан id пользователя.

Ниже приведён словарь индексов и соответствующих частот сайтов для выборки из 3 пользователей.

```
{'accounts.google.com': (8, 1),
 'apis.google.com': (9, 1),
 'football.kulichki.ru': (6, 2),
 'geo.mozilla.org': (11, 1),
 'google.com': (1, 9),
 'mail.google.com': (5, 2),
 'meduza.io': (3, 3),
 'oracle.com': (2, 8),
 'plus.google.com': (7, 1),
 'vk.com': (4, 3),
 'yandex.ru': (10, 1)}
```

В выборке из 150 пользователей 137019 уникальных сессий и 27797 уникальных сайтов. 10 самых популярных сайтов это:

1. www.google.fr
2. www.google.com
3. www.facebook.com
4. apis.google.com
5. s.youtube.com
6. clients1.google.com
7. mail.google.com
8. plus.google.com
9. safebrowsing-cache.google.com
10. www.youtube.com

Формат признаков ***site1, site2…site n*** не получится использовать в задаче классификации, поэтому применяется идея мешка слов из анализа текстов. Данные преобразуются в разреженные матрицы частот сайтов, в которых строки соответствуют сессиям, а столбцы индексам сайтов. На пересечении строки *i* и столбца *j* стоит количество раз, которое сайт с индексом *j* встретился в сессии *i.*

Ниже приведена матрица частот сайтов для выборки из 3 пользователей длиной сессии 10 сайтов без пересечений. В матрице 5 строк и 11 столбцов, что соответствует 5 сессиям и 11 уникальным сайтам.

```
matrix([[1, 3, 1, 0, 1, 0, 1, 1, 1, 1, 0],
        [3, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 2, 1, 0, 0, 2, 0, 0, 0, 0, 0],
        [4, 2, 0, 2, 1, 0, 0, 0, 0, 0, 1],
        [1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0]])
```

Ниже приведена матрица частот сайтов для выборки из 3 пользователей длиной сессии 10 сайтов и шириной окна 5 сайтов. В матрице 12 строк и 11 столбцов, что соответствует 12 сессиям и 11 уникальным сайтам.

```
matrix([[0, 3, 1, 0, 0, 0, 1, 0, 0, 0, 0],
        [1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0],
        [0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0],
        [3, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0],
        [2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 2, 1, 0, 0, 2, 0, 0, 0, 0, 0],
        [0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0],
        [2, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0],
        [3, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0],
        [1, 0, 0, 2, 1, 0, 0, 0, 0, 0, 1],
        [1, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0],
        [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]])
```

Первичный анализ данных. Проверка гипотез

Первичный анализ данных проводится для выборки из 10 пользователей с длиной сессии 10.

Сессии пользователя может характеризовать количество уникальных сайтов в сессии. Диаграмма распределения количества уникальных сайтов в сессии приведена на рисунке 3. Из диаграммы видно, что в сессиях из 10 сайтов чаще 6-8 уникальных сайтов.

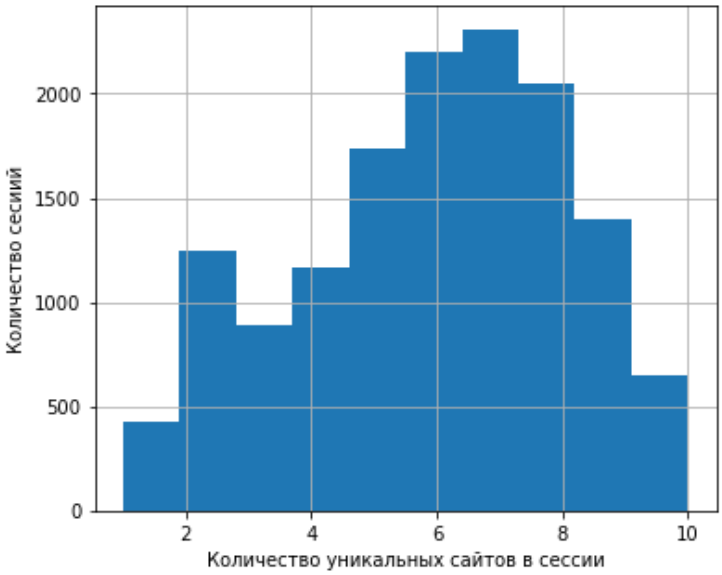

Рисунок 3 – Диаграмма распределения количества уникальных сайтов в сессии

На рисунке 4 приведен Q-Q график для распределения числа уникальных сайтов в сессии. Критерий Шапиро-Уилка отвергает гипотезу о нормальности распределения числа уникальных сайтов в сессии с достигаемым уровнем значимости ***p-value***≈0.

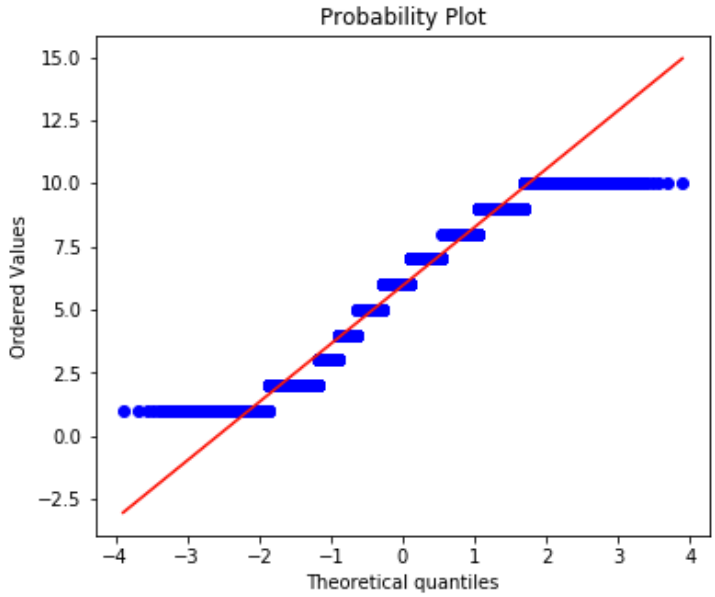

Рисунок 4 – Q-Q график для распределения числа уникальных сайтов в сессии

На Q-Q графике видны тяжелые хвосты, что также говорит против гипотезы о нормальности. Можно сделать вывод, что распределение числа уникальных сайтов не является нормальным, что говорит о том, что количество уникальных сайтов в сессии отличается от сессии к сессии не из-за наличия большого количества случайных факторов, а из-за специфики пользователей.

Большинство сессий состоит из повторяющихся сайтов. Проверим гипотезу о том, что пользователь хотя бы раз зайдет на сайт, который он уже ранее посетил в сессии из 10 сайтов. Биноминальный критерий для доли отвергает гипотезу о том, что доля сессий, в которых пользователь посетил меньше 10 разных сайтов меньше 95% с достигаемым уровнем значимости ***p-value***≈0,022. Доверительный интервал Уилсона для доли сессий, в которых пользователь повторно зашел на сайт, $0{,}9501 \leq \boldsymbol{p} \leq 0{,}9571$.

Доверительный интервал, построенный с помощью бутстрепа, для средней частоты появления сайта в выборке $22{,}515 \leq \boldsymbol{p} \leq 35{,}763$. Самый популярный сайт в выборке из 10 пользователей появился 8300 раз.

Из предварительного анализа данных видно, что сессии из набора сайтов отличаются у разных пользователей, а, следовательно, их можно использовать как признаковое описание в задачах классификации.

Визуальный анализ данных и построение признаков

С целью улучшения качества классификации помимо уже созданных «мешков» слов были созданы следующие основные признаки:

- ***session_timespant –*** продолжительность сессии;
- ***#unique_sites*** – число уникальных сайтов в сессии;
- ***start_hour –*** час начала сессии;
- ***day_of_week –*** день недели сессии.

Дополнительные признаки:

- ***top_site_share –*** доля сайтов в сессии из топ-30 самых популярных сайтов;
- ***max_diff –*** максимальное время проведенное на сайте в сессии;
- ***top_site_time_share*** – доля времени в сессии, проведенного на сайтах из то-30;
- ***num_of_small_time*** – количество сайтов с временем просмотра от 5 до 20 секунд.

Сгенерированные основные признаки для выборки из 10 пользователей с длиной сессии 10 сайтов для первых 5 сессий приведены на рисунке 5.

	session_timespan	#unique_sites	start_hour	day_of_week	user_id
0	7998	8	9	4	1
1	60	2	12	4	1
2	7935	3	9	4	2
3	7998	5	9	4	3
4	1471	3	12	4	3

Рисунок 5 – Основные признаки

Сгенерированные дополнительные признаки для выборки из 10 пользователей с длиной сессии 10 сайтов для первых 5 сессий приведены на рисунке 6.

	top_site_share	max_diff	top_site_time_share	num_of_small_time	user_id
0	0.200000	20	0.030303	1	1
1	0.000000	163	0.000000	1	1
2	0.285714	242	0.054264	1	1
3	0.000000	25	0.000000	0	1
4	0.111111	1	0.142857	0	1

Рисунок 6 – Дополнительные признаки

Ниже представлен визуальный анализ построенных основных признаков для выборки из 10 пользователей. Для наглядности пользователям даны имена

```
{1: 'Mary-Kate',
 2: 'Ashley',
 3: 'Lindsey',
 4: 'Naomi',
 5: 'Avril',
 6: 'Bob',
 7: 'Bill',
 8: 'John',
 9: 'Dick',
 10: 'Ed'}
```

На рисунке 7 приведена диаграмма распределения длины сессии в секундах (ограничена 200 секундами). Из диаграммы следует, что преобладают короткие сессии до 20 секунд и сессии от 20 до 75 секунд.

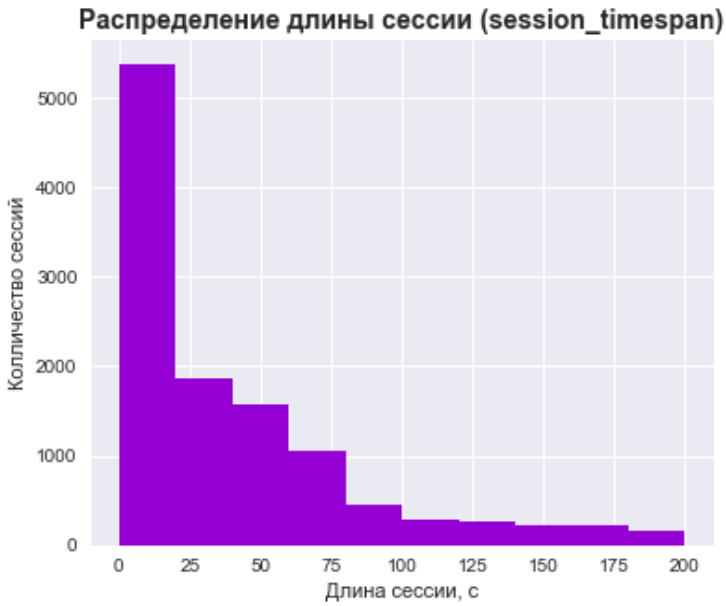

Рисунок 7 – Диаграмма распределения длины сессии в секундах

Распределение числа уникальных сайтов в сессии для каждого из пользователей приведено на рисунке 8.

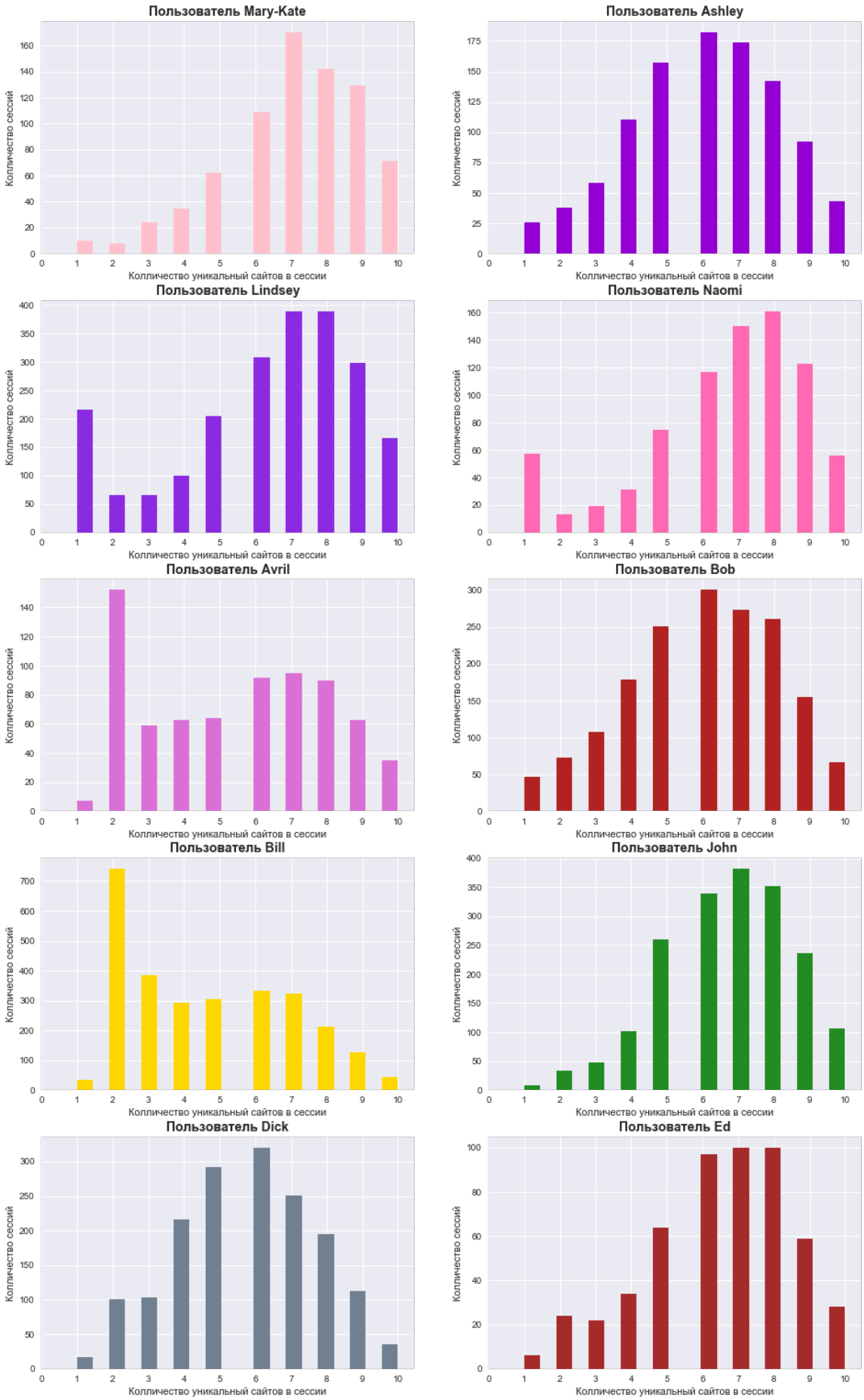

Рисунок 8 – Распределение числа уникальных сайтов в сессии

Распределение часа начала сессии приведено на рисунке 10.

Рисунок 9 – Распределение часа начала сессии

Распределение дня недели начала сессии приведена на рисунке 11.

Рисунок 11 – Распределение дня недели начала сессии

Распределение часа начала сессии для каждого из пользователей приведено на рисунке . Распределение дня недели начала сессии приведено на рисунке

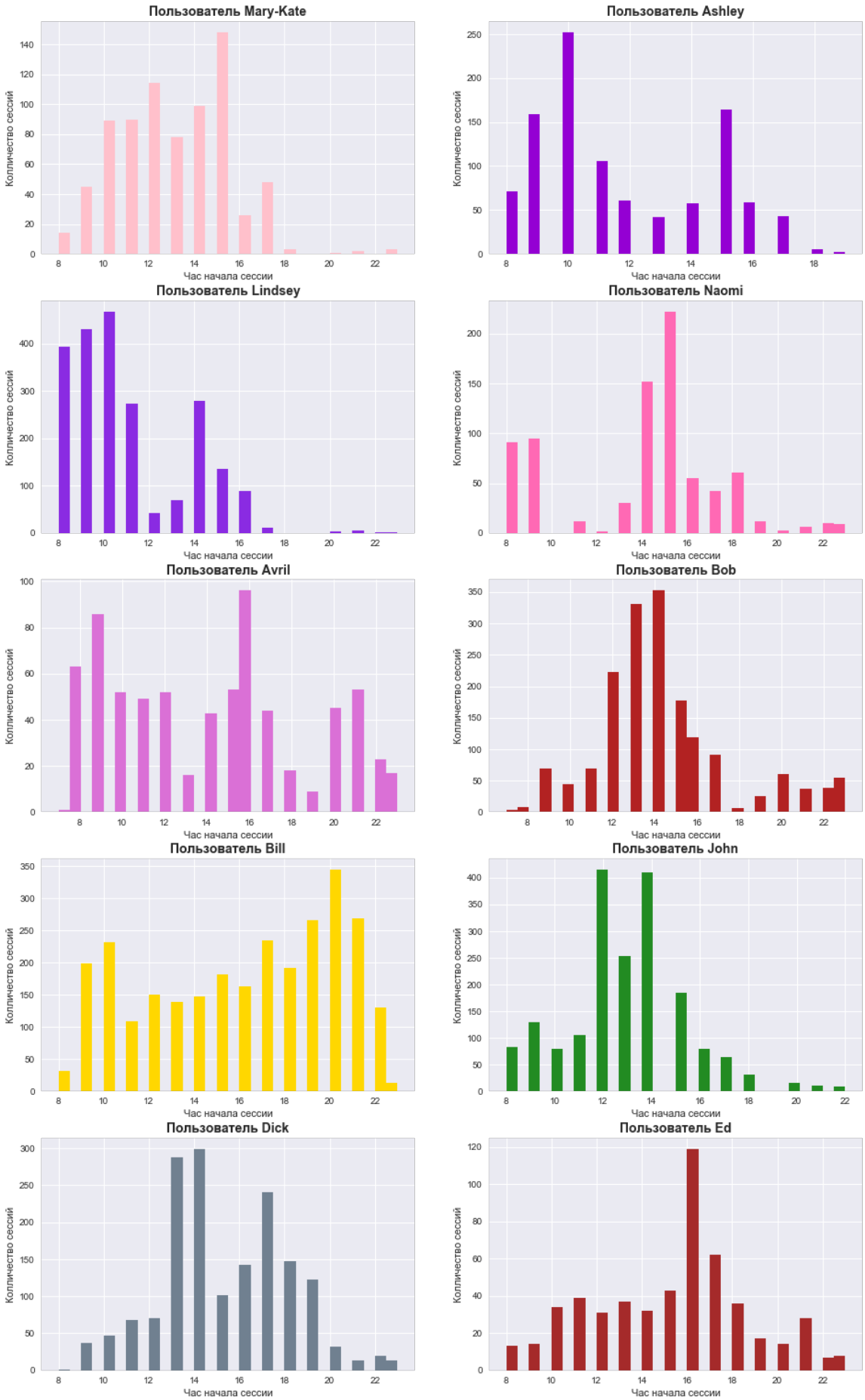

Рисунок 12 – Распределение часа начала сессии

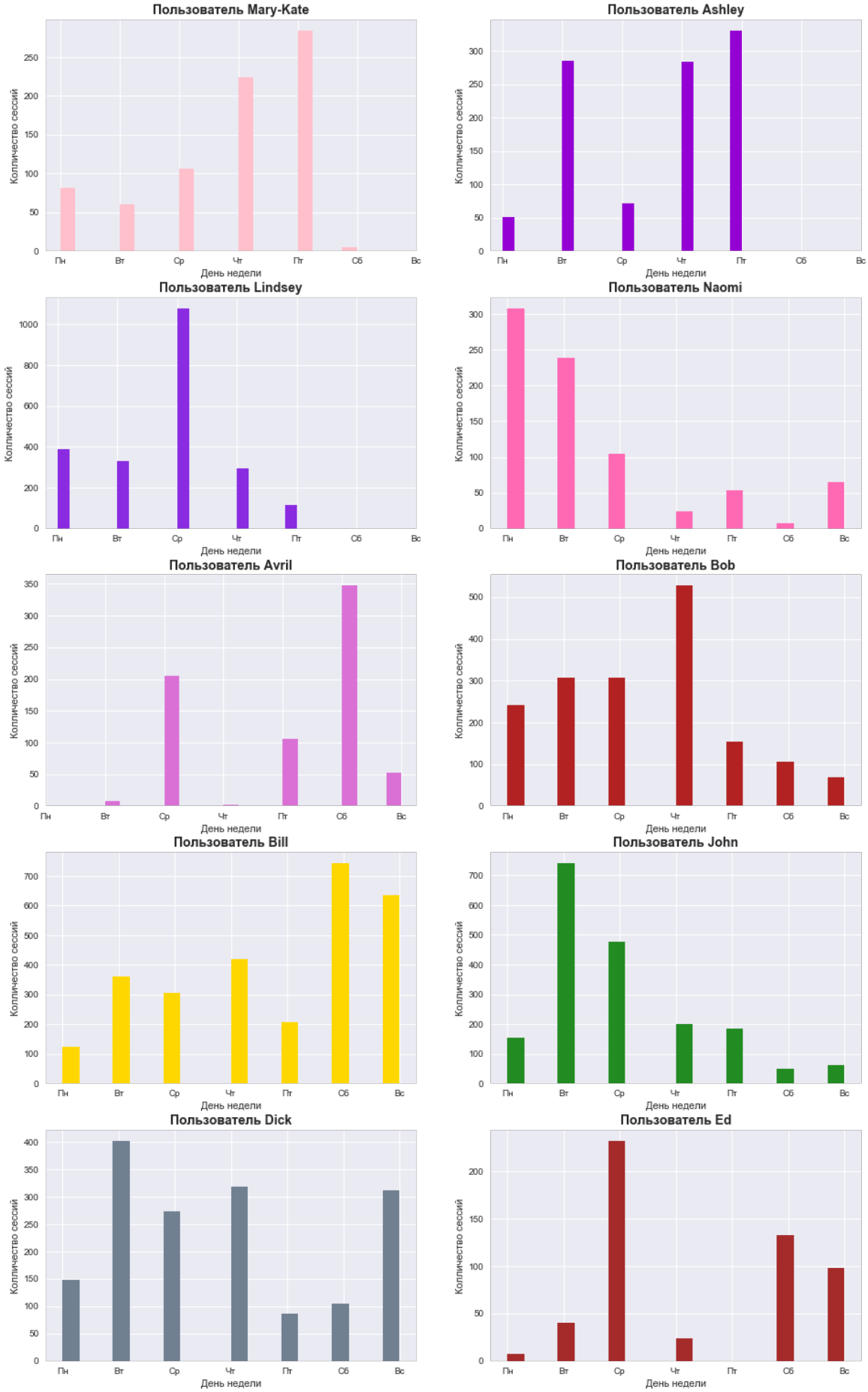

Рисунок 13 – Распределение дня недели начала сессии

Из диаграмм распределения основных признаков можно сделать следующие краткие выводы для каждого из пользователей:

1: 'Mary-Kate'

- У пользователя чаще 7-9 уникальных сайтов в сессии из десяти сайтов;
- Сессии пользователя чаще начинаются с 10 до 15, выделяются 12 и 15 часов;
- Сессии пользователя чаще начинаются во второй половине рабочей недели, редко или практически никогда в выходные.

2: 'Ashley'

- У пользователя чаще 5-8 уникальных сайтов в сессии из десяти сайтов;
- Сессии пользователя чаще начинаются в 9-10, и в 15 часов;
- Выделяются вторник, четверг, пятница, в выходные сессий не бывает.

3: 'Lindsey'

- У пользователя чаще 7-8 уникальных сайтов в сессии из десяти сайтов;
- Большая часть сессий начинается до 10 часов;
- Выделяется среда, в выходные сессий не бывает.

4: 'Naomi'

- У пользователя чаще 7-8 уникальных сайтов в сессии из десяти сайтов;
- Часто сессия начинается в 15 часов;
- Сессии пользователя чаще начинаются во первой половине рабочей недели, бываю сессии в выходные.

5: 'Avril'

- У пользователя чаще 2 уникальных сайта в сессии из десяти сайтов, пользователь сильнее выделяется по этому признаку от большинства;
- Сессии пользователя чаще начинаются в 9 и в 15 часов;
- На рабочей недели сессии часто бывают в среду и пятницу, большинство сессий начинается в субботу.

6: 'Bob'

- У пользователя чаще 6-8 уникальных сайтов в сессии из десяти сайтов;
- Сессии пользователя чаще начинаются в 13-14 часов;
- Чаще всего сесии начинаются в четверг.

7: 'Bill'

- У пользователя чаще 2 уникальных сайта в сессии из десяти сайтов, пользователь сильнее выделяется по этому признаку от большинства;
- Сессии пользователя чаще начинаются до 10 часов и после 17, чаще в 20 часов;
- Дни начала сессии на рабочей неделе распределены примерно равномерно, чаще всего сессии начинаются в выходные.

8: 'John'

- У пользователя чаще 6-8 уникальных сайтов в сессии из десяти сайтов;
- Сессии пользователя чаще начинаются в 12-14 часов;
- Выделяются вторник, среда.

9: 'Dick'

- У пользователя чаще 5-6 уникальных сайтов в сессии из десяти сайтов, пользователь несколько выделяется по этому признаку от большинства;
- Сессии пользователя чаще начинаются в 13-14 часов и в 17 часов;
- Сессии редко начинаются в пятницу, субботу;

10: 'Ed'

- У пользователя чаще 6-8 уникальных сайтов в сессии из десяти сайтов;
- Сессии пользователя чаще начинаются в 16 часов;
- Сессии пользователей чаще начинаются в среду и в выходные.

В целом по построенным диаграммам можно сделать вывод, что основные признаки хорошо разделяет некоторых пользователей, некоторых хуже, но должны помочь улучшить качество классификации.

Ниже приведен визуальный анализ дополнительных признаков для выборки из 10 пользователей.

На рисунке 14 приведено распределение доли сайтов из топ-30 в сессии для каждого из пользователей. Из диаграмм видно, что некоторые пользователи очень хорошо разделяются по этому признаку, например, пользователи ***Bill*** и ***Bob***.

На рисунке 15 приведено распределение доли времени проведенного на сайтах из топ-30. Из диаграмм видно, что пользователи не очень сильно разделяются по этому признаку. Вероятно его не стоит включать в модель.

На рисунке 16 приведено распределение количества сайтов в сессии с временем посещения 5-20 секунд. Из диаграмм видно, что большинство пользователей слабо разделимы по этому признаку.

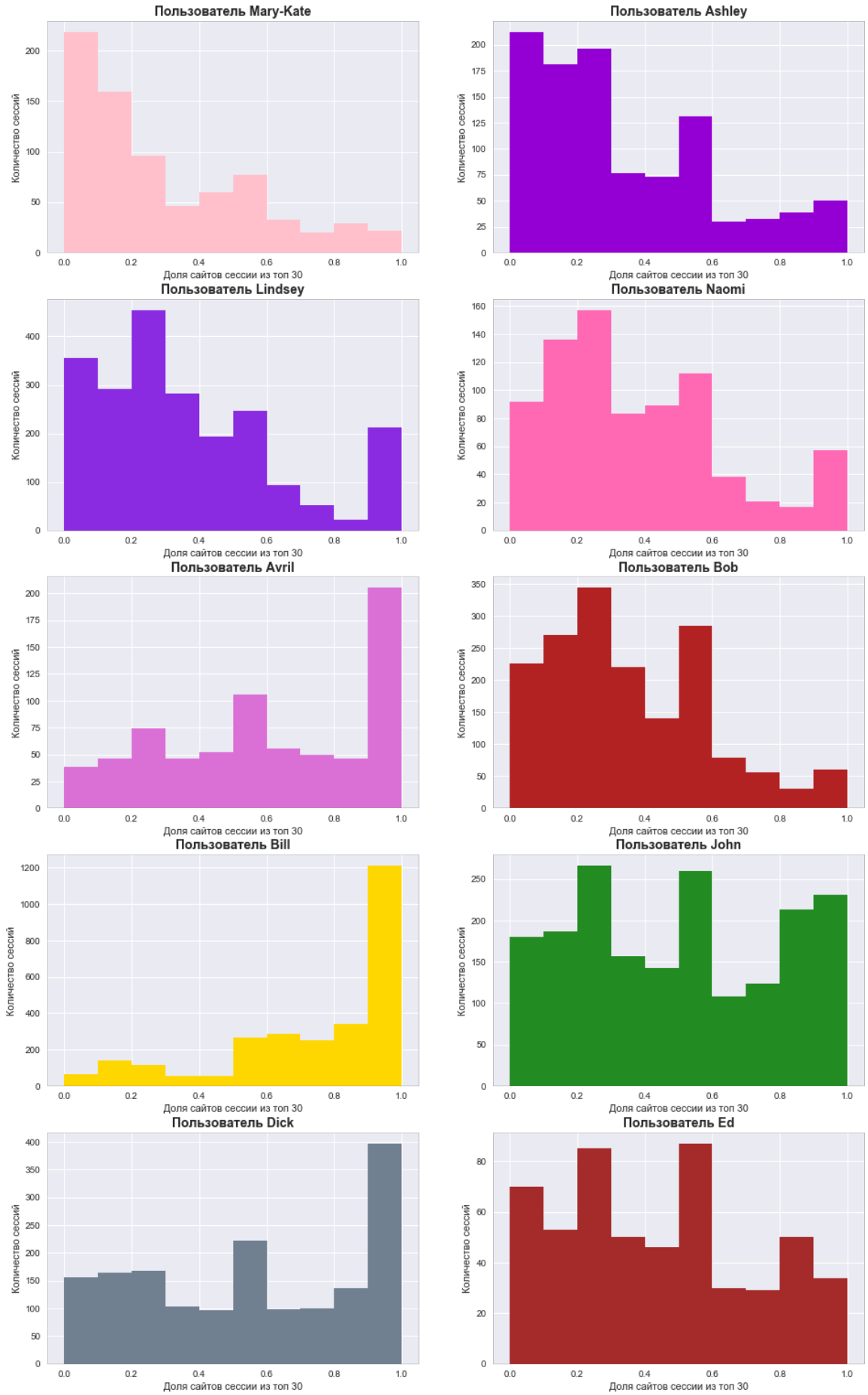

Рисунок 14 – Распределение доли сайтов из топ-30

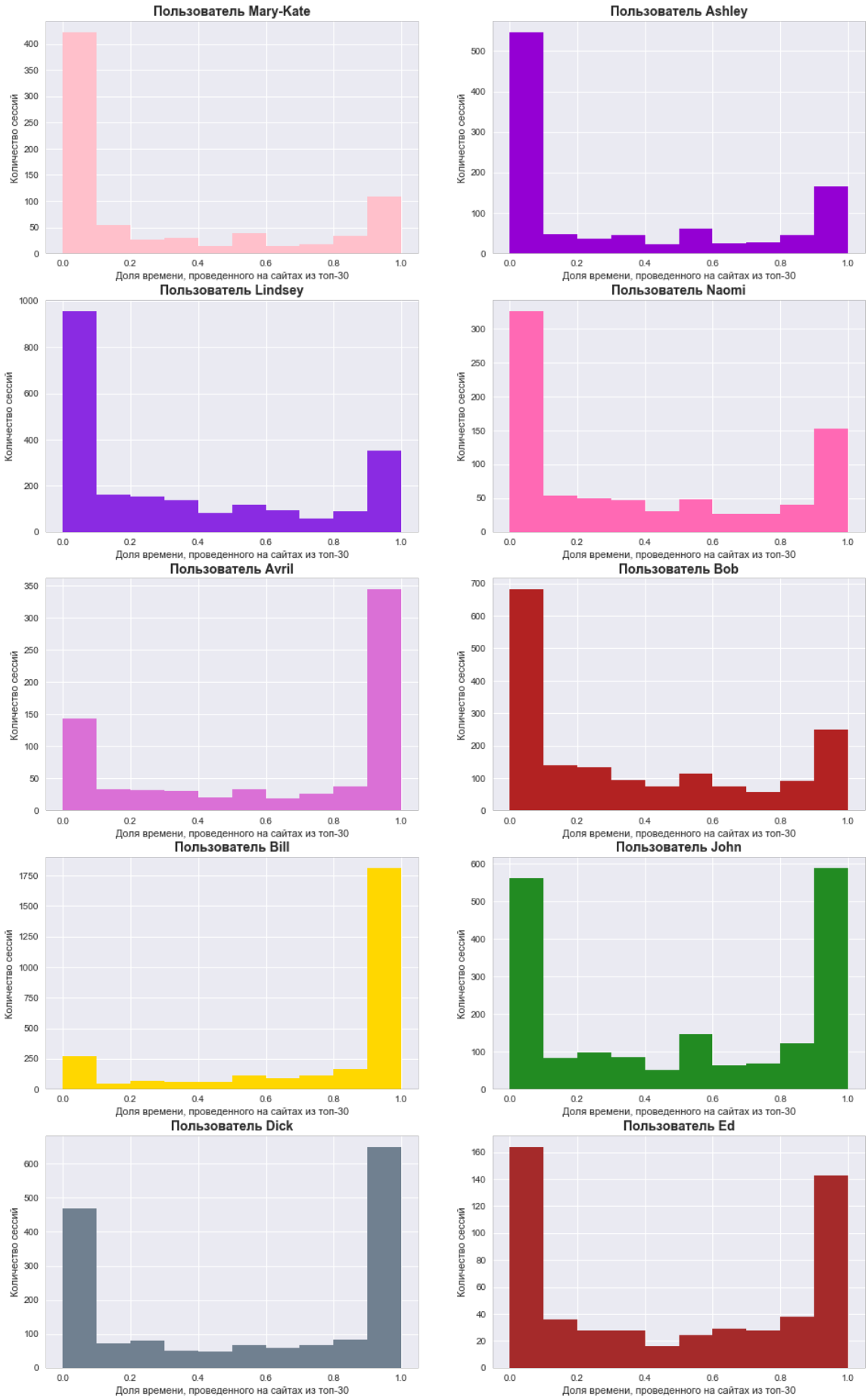

Рисунок 15 – Распределение доли времени на сайтах из топ-30

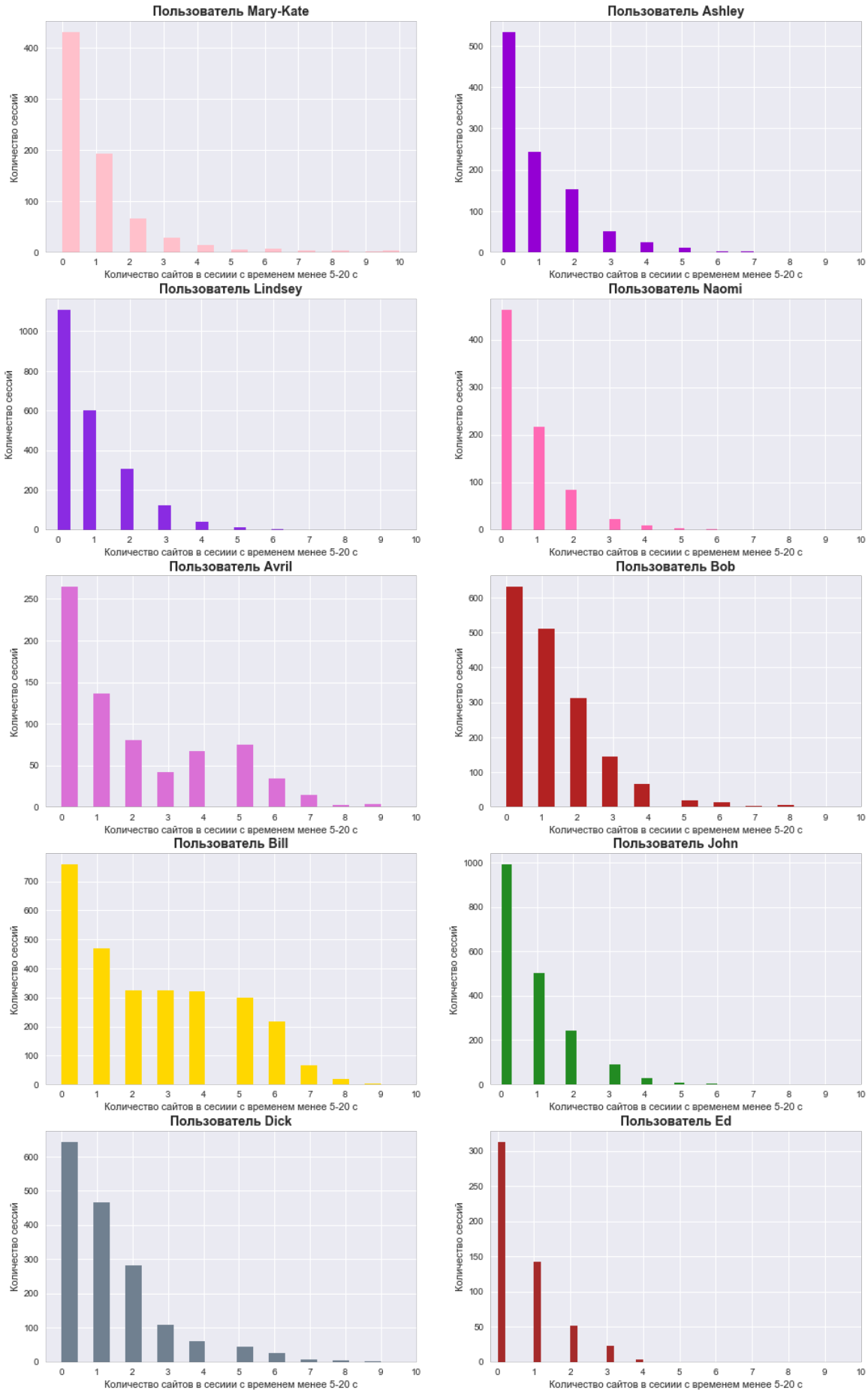

Рисунок 16 – Количество сайтов в сессии с временем 5-20 секунд

Сравнение алгоритмов классификации

В данном разделе проведено сравнение различных алгоритмов на кросс-валидации и на валидационной выборке для выборке из 10 пользователей и длиной сессий 10 сайтов. Выборка для кросс-валидации и валидационная выборка имеют соотношение 70% и 30% соответственно от обучающей выборки.

В качестве стратегии кросс-валидации выбрана стратегия с сохранением соотношения классов ***StratifiedKFold***, перемешиванием и разделением на три фолда.

Первый выбранный алгоритм классификации – алгоритм ближайших соседей ***KNeighborsClassifier*** со 100 соседями. Доля правильных ответов на кросс-валидации и на валидационной выборке соответственно 0,562 и 0,584.

Следующий алгоритм – случайный лес из 100 деревьев ***RandomForestClassifier***. Для случаного леса есть возможность проверить качество модели на объектах не попавших в обучающую выборку (OOB-оценку). Оценка OOB и доля правильных ответов на валидационной выборке соответсвтенно 0,717 и 0,731. Качество несколько улучшилось по сравнению с методом ближайших соседей.

Следующий алгоритм – многоклассовая логистическая регрессия ***LogisticRegression*** с коэффициентом регуляризации ***C*** по умолчанию. Доля правильных ответов на кросс-валидации и на валидационной выборке соответственно 0,761 и 0,782. Данный алгоритм показывает самое высокое качество. С целью улучшения качества с помощью ***LogisticRegressionCV*** был проведен поиск по сетке коэффициента регуляризации ***C*** 40 значений в диапазоне 0,1…7. Качество улучшить не удалось. График зависимости доли правильных ответов от коэффициента ***C*** приведен на рисунке 17.

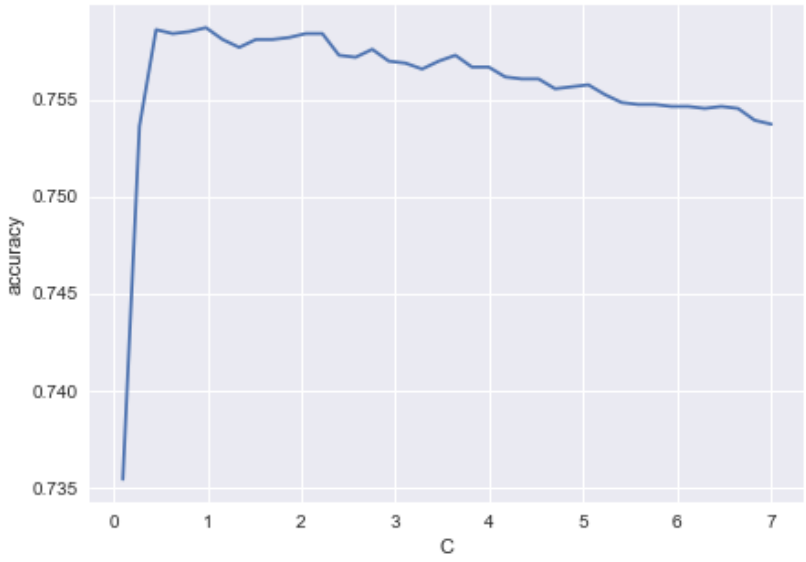

Рисунок 17 – Зависимость доли правильных ответов на кросс-валидации от параметра ***C*** для логистической регрессии

Последний опробованный алгоритм – метод опорных векторов ***LinearSVC*** с коэффициентом регуляризации ***C*** =1. Доля правильных ответов на кросс-валидации и на валидационной выборке соответственно равны 0,753 и 0,777. С помощью ***GridSearchCV*** был осуществлен перебор 30 параметров ***C*** в диапазоне 0,001…1 Доля правильных ответов на кросс-валидации и на валидационной выборке для лучшего ***C*** = 0,104 соответственно равны 0,764 и 0,781. Качество классификации сопоставимо с логистической регрессией, но кросс-валидации данный метод показывает несколько лучший результат . На рисунке 18 приведены кривые зависимость доли правильных ответов на обучающей и тестовой выборках.

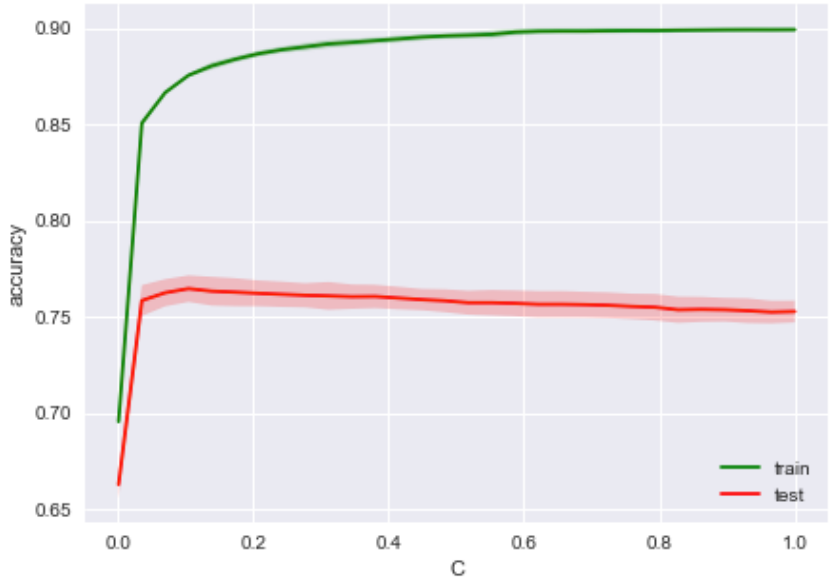

Рисунок 18 – Зависимость доли правильных ответов на кросс-валидации от параметра ***C*** для метода опорных векторов

Выбор параметров длины сессии и ширины окна

Для алгоритма показавшего лучшее качество на валидационной выборке (***LinearSVM*** с ***C*** = 0,104) был проведен перебор параметров длины сессии и ширины окна для выборки из 10 пользователей. Ниже приведены результаты перебора, где первое число – ширина окна, второе – длина сессии, третье – доля правильных на обучающей выборке, треть – доля правильных на валидационной выборке, четвертое – время выполнения операции.

```
10 15 (0.82382149552781048, 0.84048352690210948, 3.6159920692443848)
10 10 (0.76468532445509807, 0.78075373311211183, 2.0581350326538086)
7 15 (0.84794037698440983, 0.85432221669155473, 5.302098989486694)
7 10 (0.79701614637346518, 0.80736684917869583, 2.5995490550994873)
7 7 (0.75298911148303416, 0.76173884187821472, 1.481153964996338)
5 15 (0.86800137255434695, 0.87529634898055952, 6.806030035018921)
5 10 (0.81592472053180387, 0.82456140350877194, 3.4274730682373047)
5 7 (0.77405867456322597, 0.78532479848269321, 2.043919801712036)
5 5 (0.72528306503988282, 0.73624940730203892, 1.3519740104675293)
```

Из полученных результатов следует, что лучшее качество классификации соответствует ширине окна ***k*** = 5 и длине сессии ***n*** = 15. Данным параметрам соответствуют доля правильных ответов на кросс-валидации и валидационной выборке 0,868 и 0,875 соответственно.

Ниже приведены результаты перебора длины сессии и ширины окна для выборки из 150 пользователей.

```
5 5 (0.40858659509908596, 0.42171606560568453, 351.39571595191956)
7 7 (0.43638649409423974, 0.45295840855673264, 292.91506695747375)
10 10 (0.46125889994279129, 0.48362769425388019, 243.30162501335144)
5 15 (0.61379723589019186, 0.63598092178906895, 930.2592360973358)
```

Для выборки из 150 пользователей наблюдается значительное снижение качества классификации.

Идентификация конкретного пользователя и кривые обучения

Многоклассовая классификации для выборки из 150 пользователей показала невысокое качество. Однако, конкретного пользователя можно идентифицировать достаточно хорошо. Для идентификации каждого пользователя на выборке из 150 пользователей было обучено 150 алгоритмов ***LogisticRegression*** «Один-Против-Всех» (***multi_class = "ovr"***).

Доли правильных ответов на кросс-валидации для первых 25 пользователей приведены ниже.

```
User 1, CV score: 0.984400797997
User 2, CV score: 0.995702116127
User 3, CV score: 0.994404631141
User 4, CV score: 0.98471778703
User 5, CV score: 0.987946735067
User 6, CV score: 0.994456243162
User 7, CV score: 0.992539516337
User 8, CV score: 0.983914244184
User 9, CV score: 0.997124909282
User 10, CV score: 0.993829612998
User 11, CV score: 0.994146618008
User 12, CV score: 0.992045590176
User 13, CV score: 0.996918500661
User 14, CV score: 0.994861704994
User 15, CV score: 0.996350846397
User 16, CV score: 0.996196033819
User 17, CV score: 0.990077258838
User 18, CV score: 0.994153990356
User 19, CV score: 0.994743752234
User 20, CV score: 0.990438488505
User 21, CV score: 0.976999274257
User 22, CV score: 0.996011741281
User 23, CV score: 0.987216904347
User 24, CV score: 0.990003530735
User 25, CV score: 0.995945393034
```

Доля правильных ответов получилось очень высокой, но это часично объясняется дисбалансом классов. Для алгоритма более показательной является метрика ***ROC-AUC.*** Среднее по всем классам значение этой метрики на валидационной выборке составляет 0,955.

На рисунке 19 приведены кривые обучения для одного из пользователей. Как видно из рисунка доля правильных ответов на валидационной выборке не перестает растти с увеличением выборки, а, следовательно, дополнительные размеченные данные помогли бы улучшить качество.

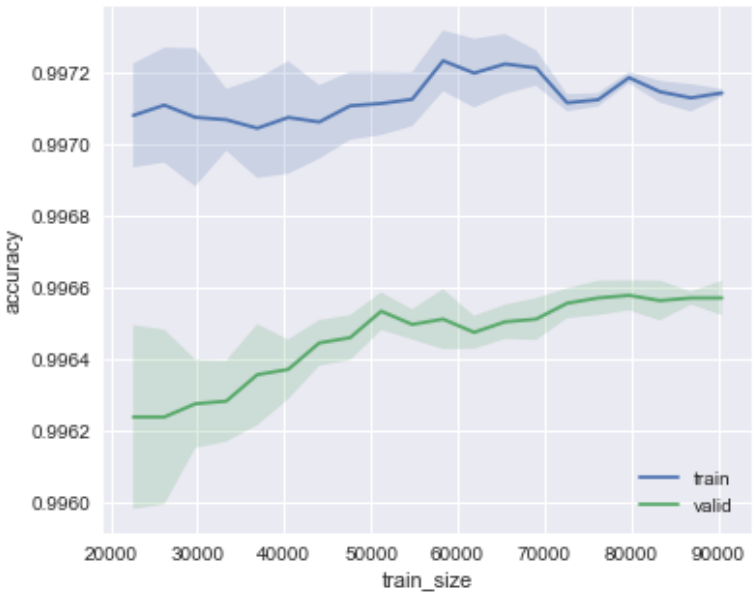

Рисунок 19 – Кривые обучения для одного из пользователей

Улучшение модели

В этом разделе приведены мероприятия направленные на улучшение модели. За алгоритм классификации был принята логистическая регрессия ***LogisticRegression.***

Матрица частот слов была преобразована в формат TF-IDF. Качество классификации снизилось. Средняя метрика ***ROC-AUC*** составила 0,946.

К матрице частот слов были добавлены сгенерированные ранее основные признаки. Качество классификации. Средняя метрика ***ROC-AUC*** составила 0,89. После нормализации основных признаков качество повысилось. Средняя метрика ***ROC-AUC*** составила 0,96.

К обучающей выборке, состоящей из матрицы частот слов и основных признаков, были добавлены нормализованные дополнительные признаки. Качество модели не изменилось. Средняя метрика ***ROC-AUC*** составила 0,96.

Последним шагом на пути улучшения качества классификации для выборки из 150 пользователей стало обучение на выборке с параметрами ширины окна ***k*** = 5 и длины сессии ***n*** = 15. Качество модели существенно выросло. Средняя метрика ***ROC-AUC*** составила 0,988.

Результаты классификации пользователей для выборки из 150 пользователей приведены в Jupiter Notebook, приложенном к отчету.

Участие в соревновании Kaggle «Identify Me If You Can»

Подходы рассмотренные ранее в этом отчете были применены в соревновании Kaggle «Identify Me If You Can».

В соревновании ставилась задача сделать прогноз для сессий из тестовой выборки и определить принадлежат ли они пользователю (Элис). В обучающей выборки были доступны сессии из 10 сайтов или длиной не более 30 минут и время захода на каждый сайт из сессии. Обучающая выборка отсортирована по времени, тестовая выборка также отделена по времени от обучающей. Целевая метрика в задании - ***ROC-AUC.***

В результате была построена модель классификации на основе логистической регрессии ***LogisticRegression.*** Модель обучалась на обучающей выборке составленной из матрицы частот слов и признаков аналогичных рассмотренным в выше в данном отчете. С помощью ***LogisticRegressionCV*** подбирались способ регуляризации (***l1, l2***) и коэффициент ***C.***

Метрика ***ROC-AUC*** на валидиционной выборке составила 0,9583, а на тестовой существенно ниже - 0,9257.

Результаты приведены в Jupiter Notebook, приложенном к отчету.

Заключение

В данном отчете были рассмотрены результаты полученные при работе над финальным проектом специализации «Машинное обучение и анализ данных» по идентификации интернет пользователей.

В качестве итогового алгоритма классификации пользователей была выбрана логистическая регрессия ***LogisticRegression*** с регуляризацией ***l2,*** обученная на выборке из 150 пользователей, состоящей из разреженной матрицы частот сайтов (“мешка слов”) и 8 признаков, с шириной окна 5 сайтов и длиной 15 сайтов.

Средняя доля правильных ответов по всем пользователям порядка 0,99, средняя метрика ***ROC-AUC*** на валидационной выборке 0,988.