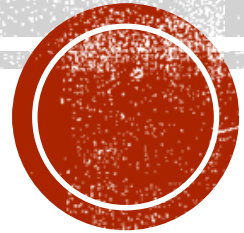
ОТЧЕТ ПО ПРОЕКТУ «ИДЕНТИФИКАЦИЯ ИНТЕРНЕТ-ПОЛЬЗОВАТЕЛЕЙ»



Проскурякова Ольга

ОПИСАНИЕ ПРОЕКТА

В проекте решалась задача идентификации пользователя по его действиям в интернете. Алгоритм анализировал последовательность из нескольких веб-сайтов, посещенных подряд одним и тем же человеком, и определял, Элис это или взломщик (кто-то другой).

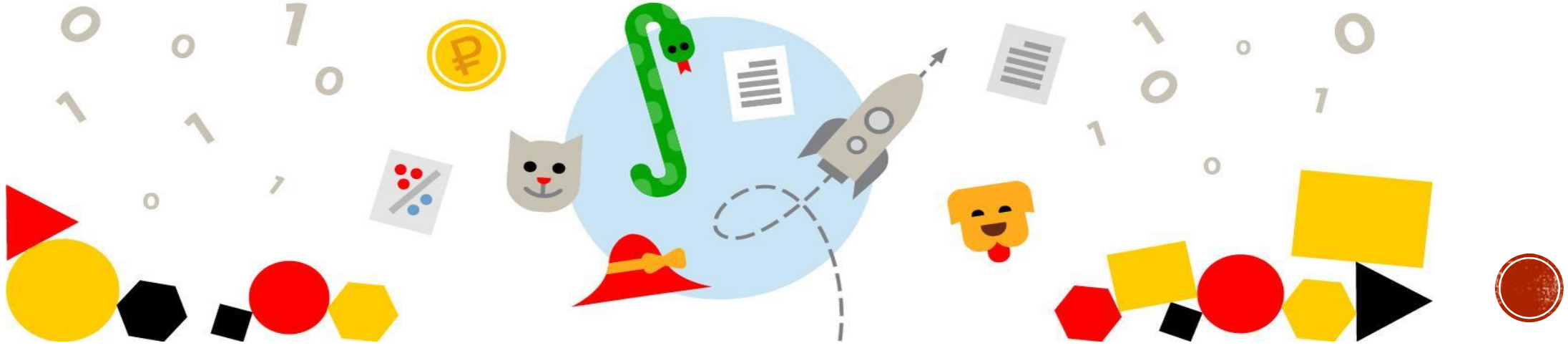
Идея заключалась в том, что каждый пользователь в своем роде уникален по тому, в какой последовательности он посещает сайты, сколько времени он проводит на очередном сайте, в какое время он чаще всего онлайн и.т.д. И из этих соображений можно выделить такие особенности поведения пользователя, которые отличают его от остальных и позволяют идентифицировать.

ИСХОДНЫЕ ДАННЫЕ

В обучающей выборке были следующие признаки:

- **Site1** - индекс первого посещенного сайта в сессии;
- **Time1** - время посещения первого сайта в сессии ;
- …;
- **Site10** - индекс 10ого посещенного сайта в сессии ;
- **Time10** - время посещения 10ого сайта в сессии;
- **user_id** – ID пользователя.

Сессии пользователей выделены таким образом, что они не могут быть длиннее получаса или 10 сайтов. То есть сессия считается оконченной либо когда пользователь посетил 10 сайтов подряд, либо когда сессия заняла по времени более 30 минут.



ПОДГОТОВКА ОБУЧАЮЩЕЙ ВЫБОРКИ

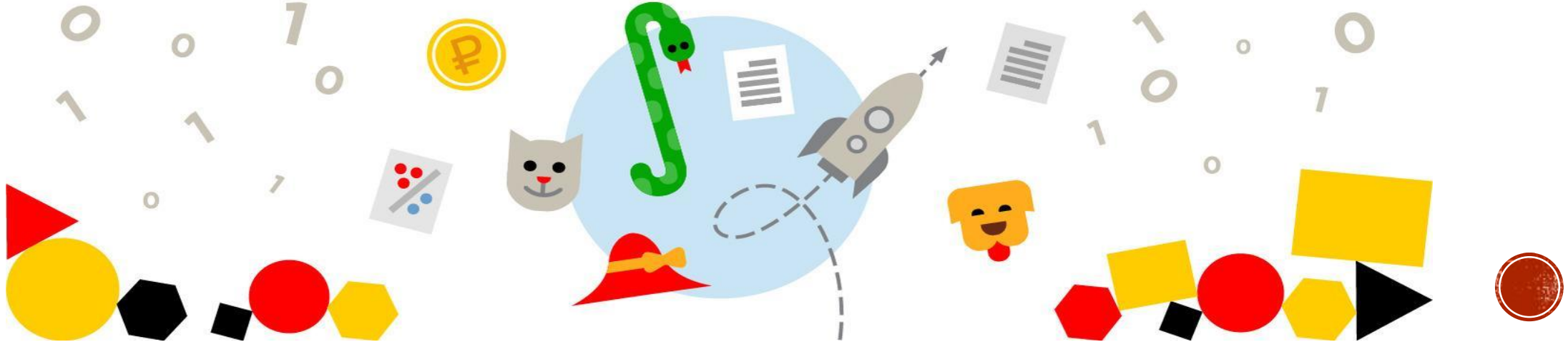
Исходные данные содержали пропуски в тех строках, где сессии оказались меньше 10 сайтов. Для обучающей выборки из исходных данных выделялись только сайты, поэтому пропуски в данных были заполнены нулями.

Распределение классов в обучающей выборке:

Класс	Частота появления в выборке
0	251264
1	2297

Заметен сильный дизбаланс классов, поэтому использовать как метрику долю правильных ответов будет непоказательно. В качестве метрики будет использован **roc_auc_score*.**

***roc_auc_score** – площадь под **ROC**-кривой, выражает долю пар объектов вида (объект класса 1, объект класса 0), которые алгоритм верно упорядочил



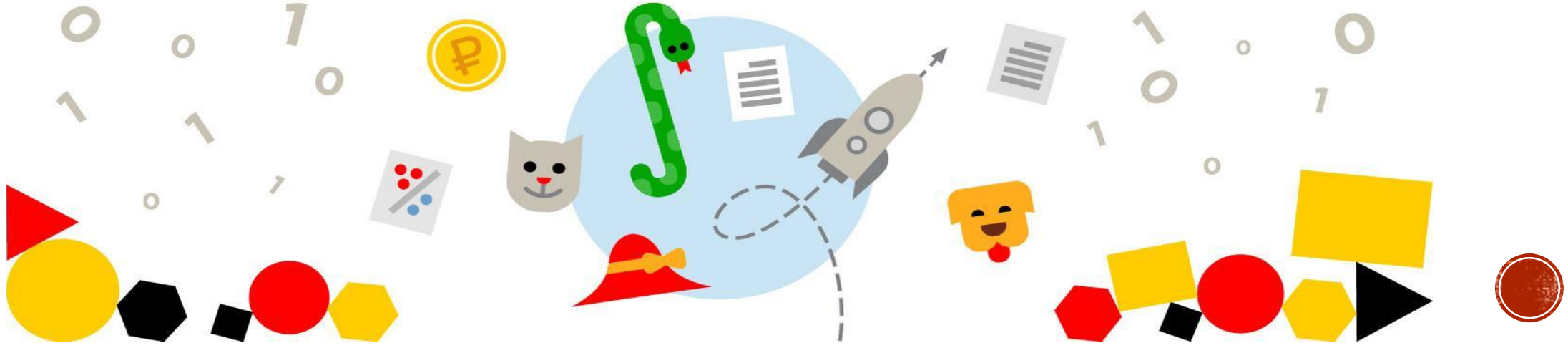
ПОДГОТОВКА ОБУЧАЮЩЕЙ ВЫБОРКИ

В получившейся выборке в качестве признаков были сайты, что не вполне понятно и интерпретируемо. Для построения качественной выборки была использована идея «мешка слов».

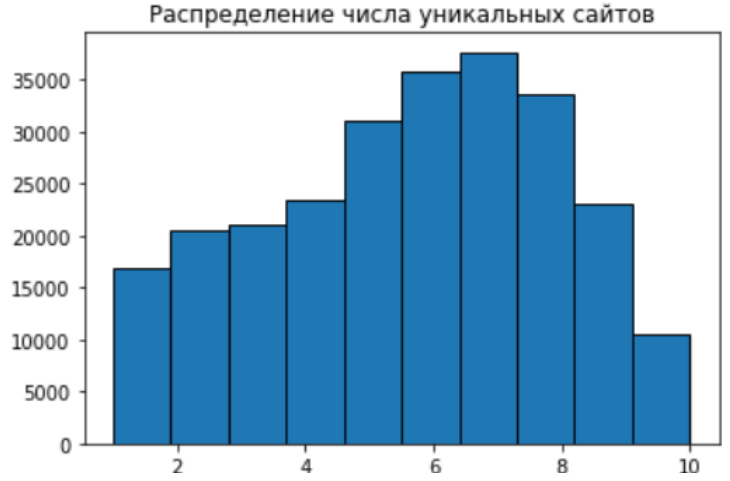
Для этого были созданы новые матрицы, строки которых соответствовали сессиям из 10 сайтов, а столбцы – индексам сайтов. На пересечении i-ой строки и jого столбца было число **nij** – количество раз, когда j-ый сайт встретился в i-ой сессии. Матрицы были созданы с использованием разреженных матриц **Scipy.**

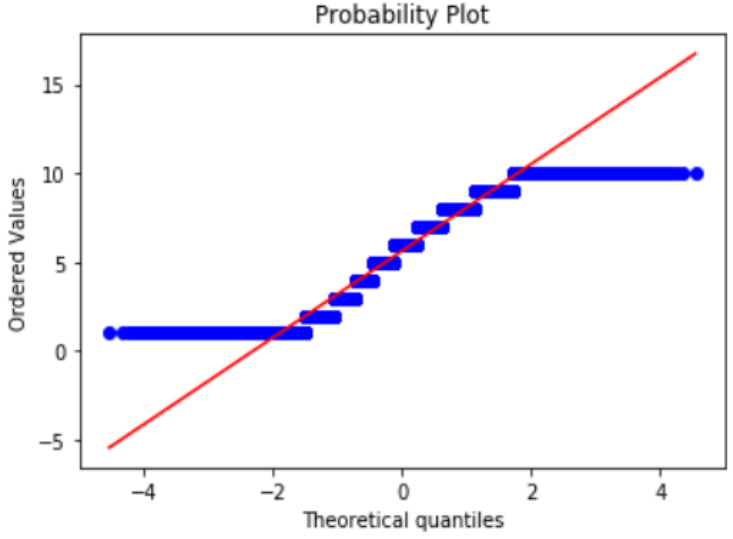
Помимо разреженной матрицы признаков были добавлены дополнительно следующие:

- **day_of_week -** день недели
- **time_of_day -** время дня начала сессии: утро, день, вечер, ночь
- **unique_sites -** количество уникальных сайтов в сессии
- **year_month -** месяц и год начала сессии



ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ

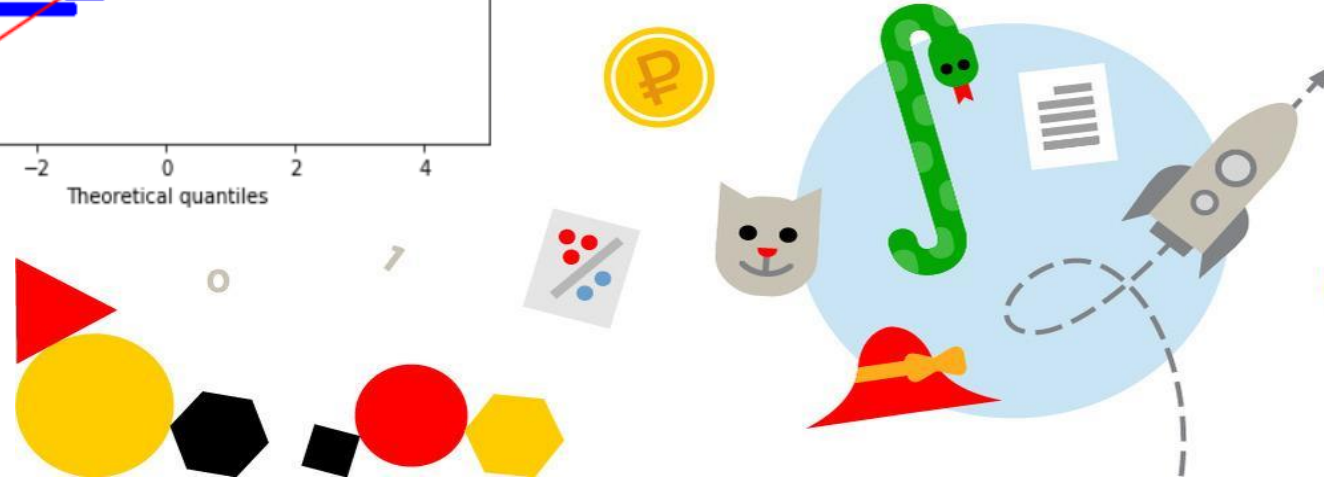


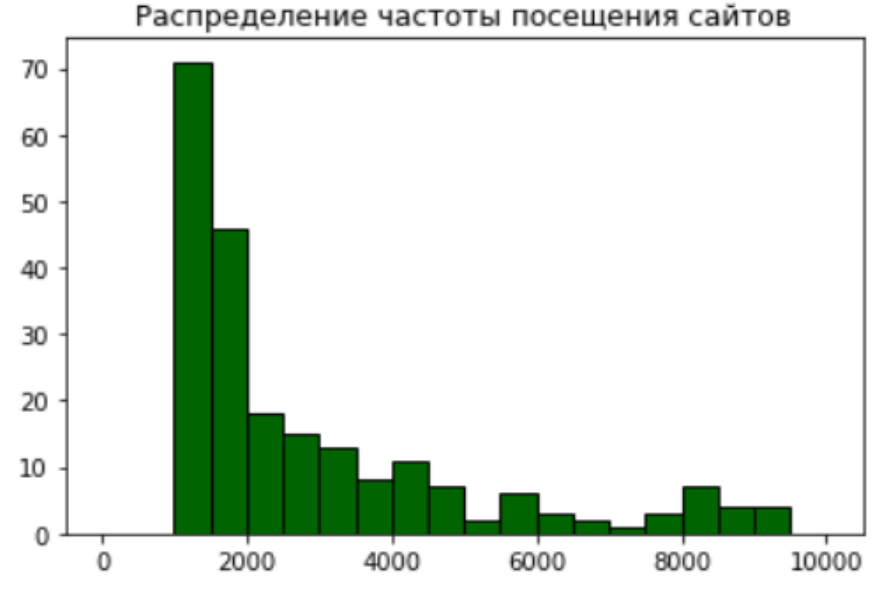


Было построено распределение числа уникальных сайтов. С помощью критерия Шапиро-Уилка и **qq-plot** было определено, что распределение не является нормальным.

С помощью биномиального критерия для доли проверили, что доля случаев, когда пользователь повторно посетил какой-то сайт (то есть число уникальных сайтов в сессии < 10) велика: больше 95%.

Также было построено распределение частоты сайтов, которые были посещены не менее 1000 раз.
95% интервал для средней частоты появления сайта в выборке:
[1830.6, 2948.1]



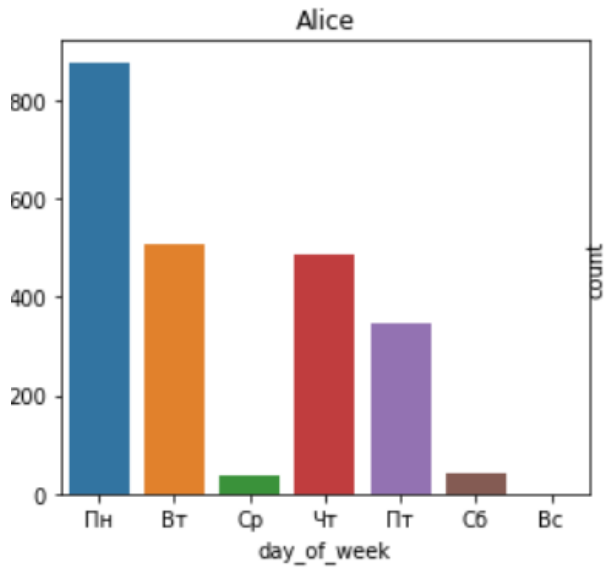


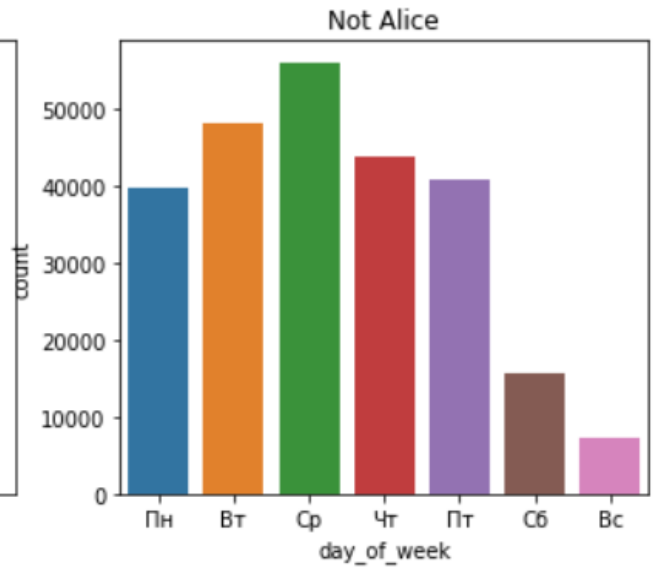
ВИЗУАЛЬНЫЙ АНАЛИЗ ПРИЗНАКОВ

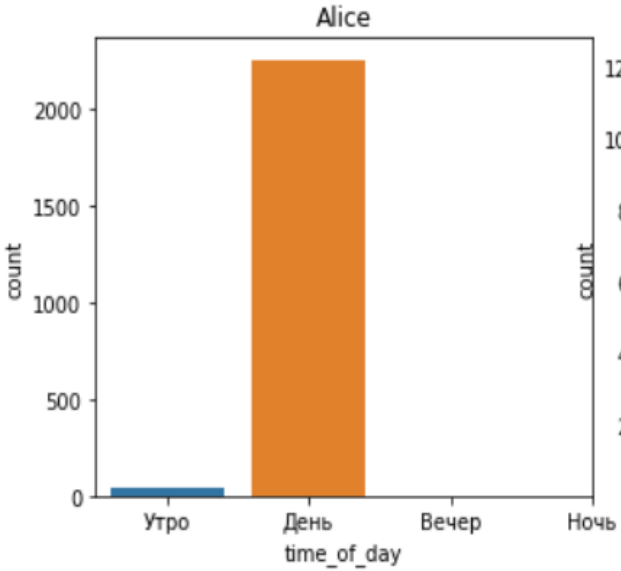
Для визуального анализа были построены распределения признаков.

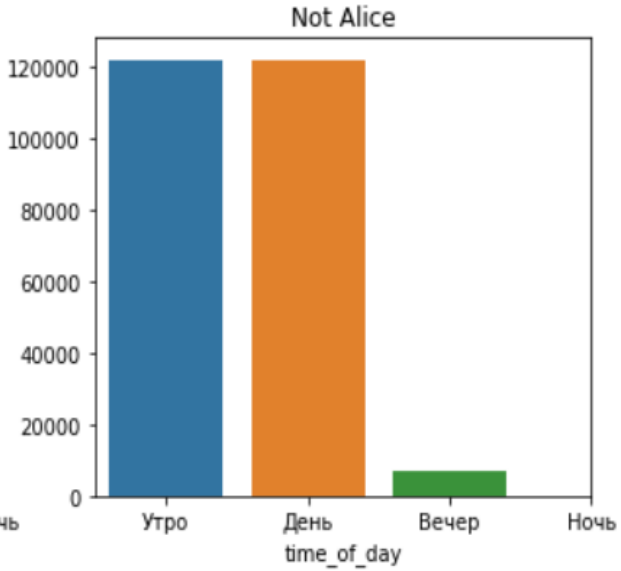
Из распределения дня недели начала сессии видно, что Элис чаще всего бывает в сети в понедельник, реже во вторник, четверг и пятницу, в то время как остальные пользователи чаще бывают в среду. Распределение признака значительно отличается, следовательно он может быть полезен при классификации.

Распределение времени дня также может помочь разделить выборки: Элис чаще всего онлайн днем, почти не бывает утром и вообще отсутствует вечером, в то время, как остальные пользователи проявляют одинаковую активность утром-днем и есть нечастые вечерние сессии.





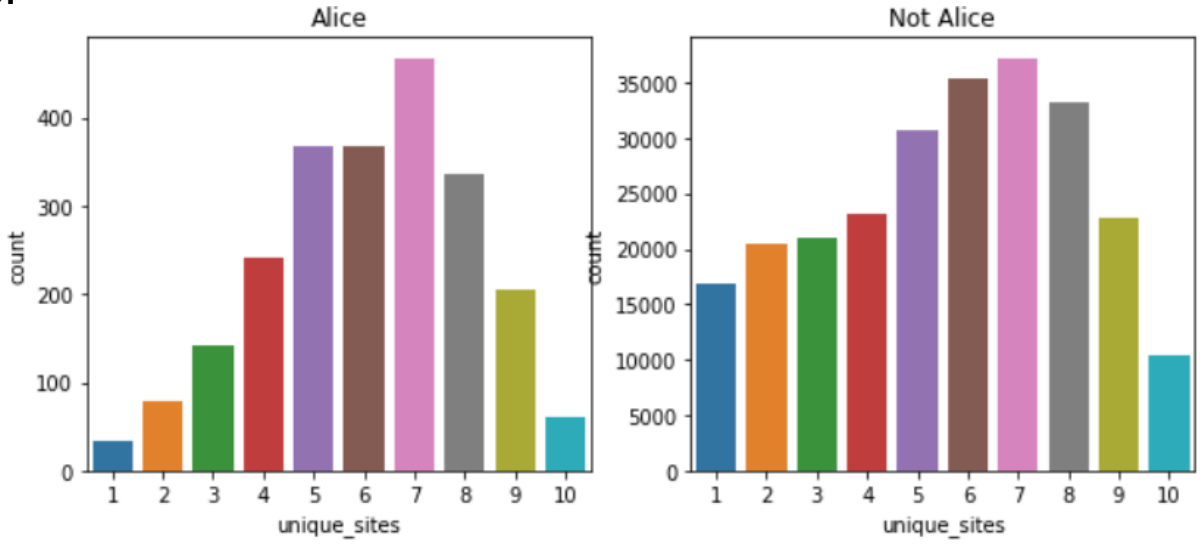


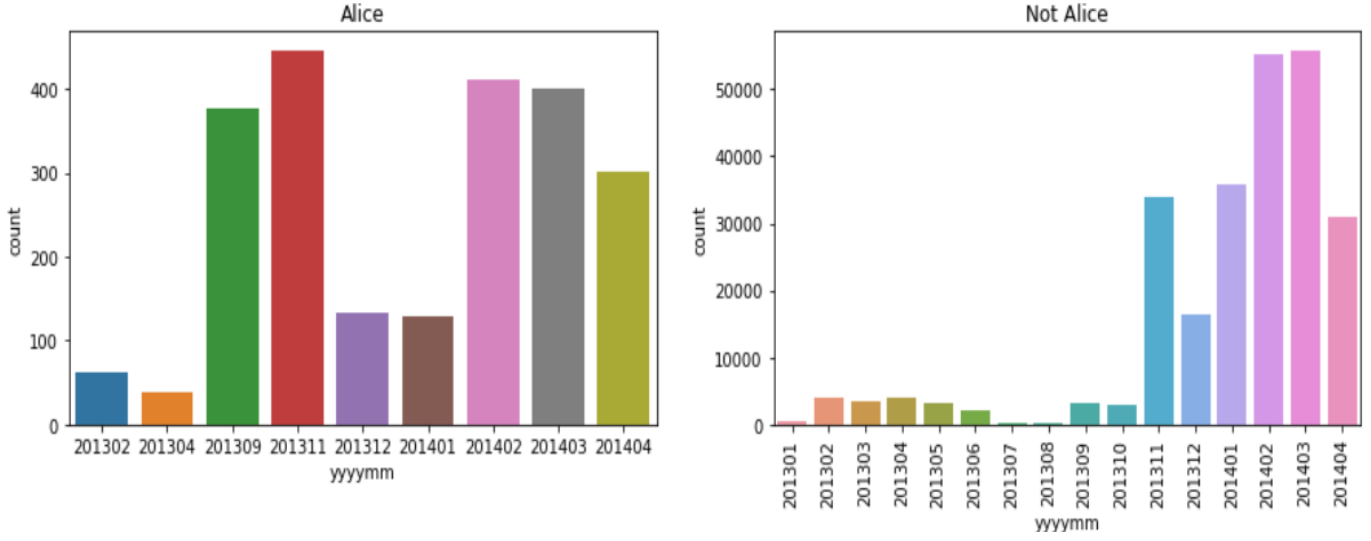


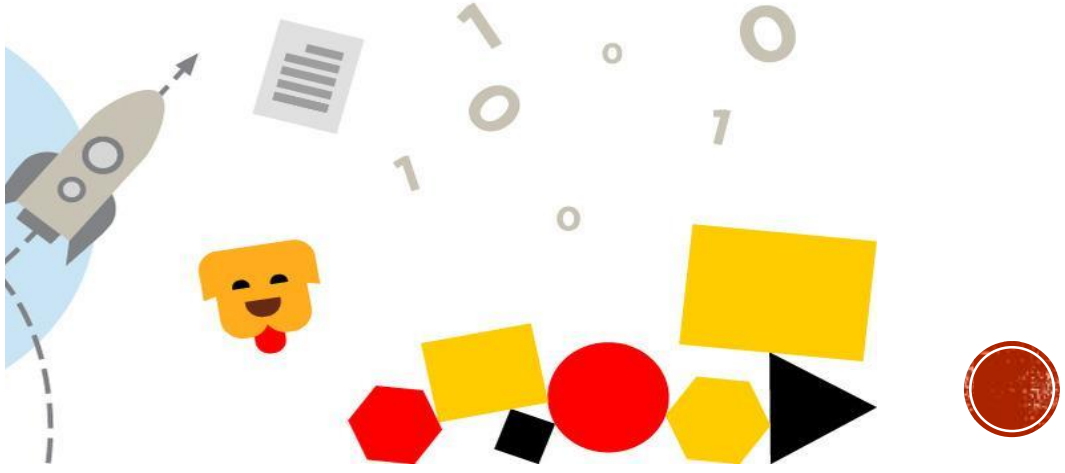
ВИЗУАЛЬНЫЙ АНАЛИЗ ПРИЗНАКОВ

Распределение количества уникальных сайтов в сессии хоть и имеет схожий вид, напоминающий по форме нормальное распределение, однако у распределения Элис более легкие хвосты, в то время как распределение остальных пользователей скошено влево.

Неожиданно неплохо отделяющим выборки признаком стал признак год-месяц начала сессии. У Элис самым активным был ноябрь 2013, в то время как у других пользователей этот месяц был менее популярен для посещения интернета.





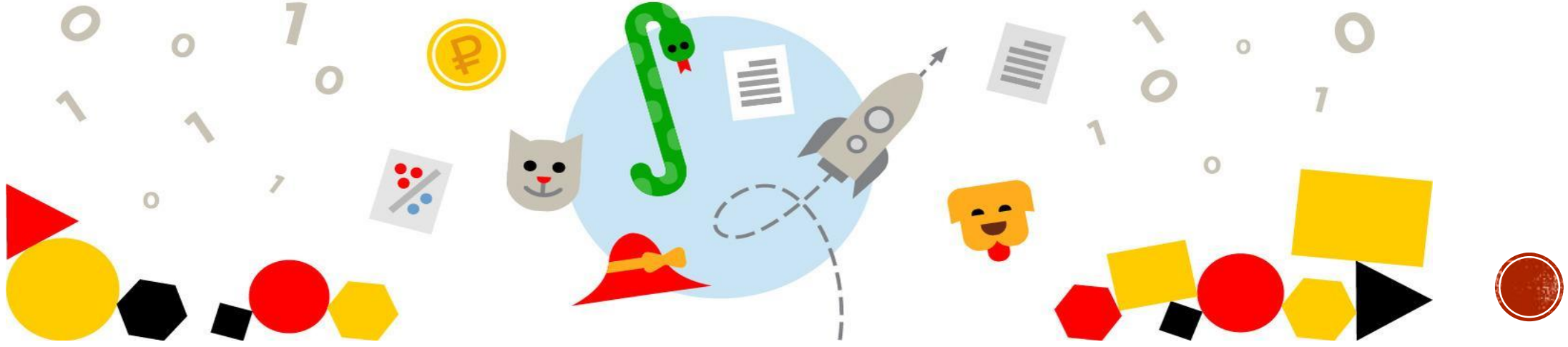


СРАВНЕНИЕ АЛГОРИТМОВ КЛАССИФИКАЦИИ

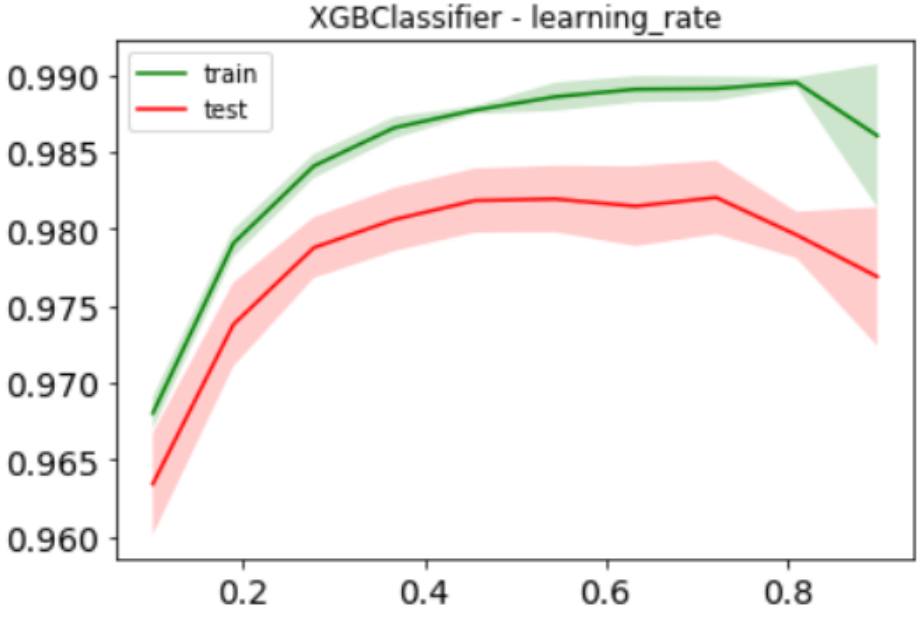
Для сравнения качества работы классификаторов было выбрано и протестировано несколько вариантов.

Классификатор	roc_auc_score на отложенной выборке
XGBClassifier	0.983
SGDClassifier	0.974
LogisticRegression	0.985
LinearSVC	0.668

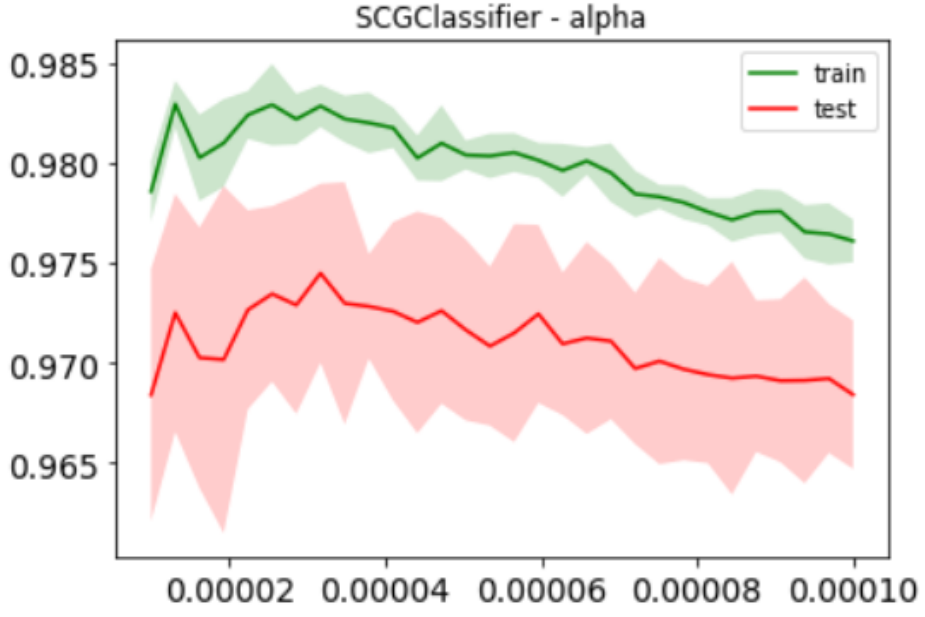
Для некоторых классификаторов был проведен подбор гиперпараметров. Подбор параметров производился по сетке с помощью **GridSearch** из **Sklearn**. Для этого задается набор значений параметров классификатора, классификатор, выборка и метрика для оценки .Производится оценка и выбирается лучшее значение параметра для данного алгоритма и выборки.



ПОДБОР ПАРАМЕТРОВ



В частности, осуществлялся подбор параметра регуляризации **learning_rate** для **XGBClassifier** в промежутке [0.1;0.9]. Из графика зависимости **roc_auc** от параметра видно, что лучшее значение около **leraning_rate=0.7**.



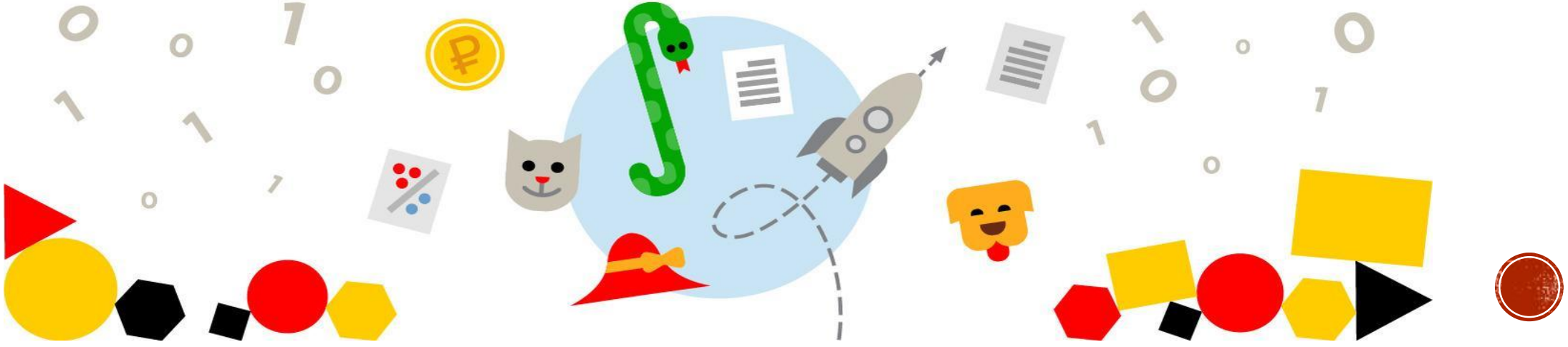
Также осуществлялся подбор параметра регуляризации **learning_rate** для **SGDClassifier** в промежутке [10e-5;10e-3]. В данном случае значение параметра, обеспечивающее большее значение метрики находится около **0.00003**.

ВЫБОР МОДЕЛИ

Изначально был использован алгоритм, использующий стохастический градиентный спуск, реализованный в библиотеке **Sklearn – SGDClassifier.** Алгоритм показал хорошее качество на тестовой выборке без настройки параметров **roc-auc** 0.91. Кроме этого, алгоритм показал быстродействие, которого не было у предыдущих протестированных классификаторов.

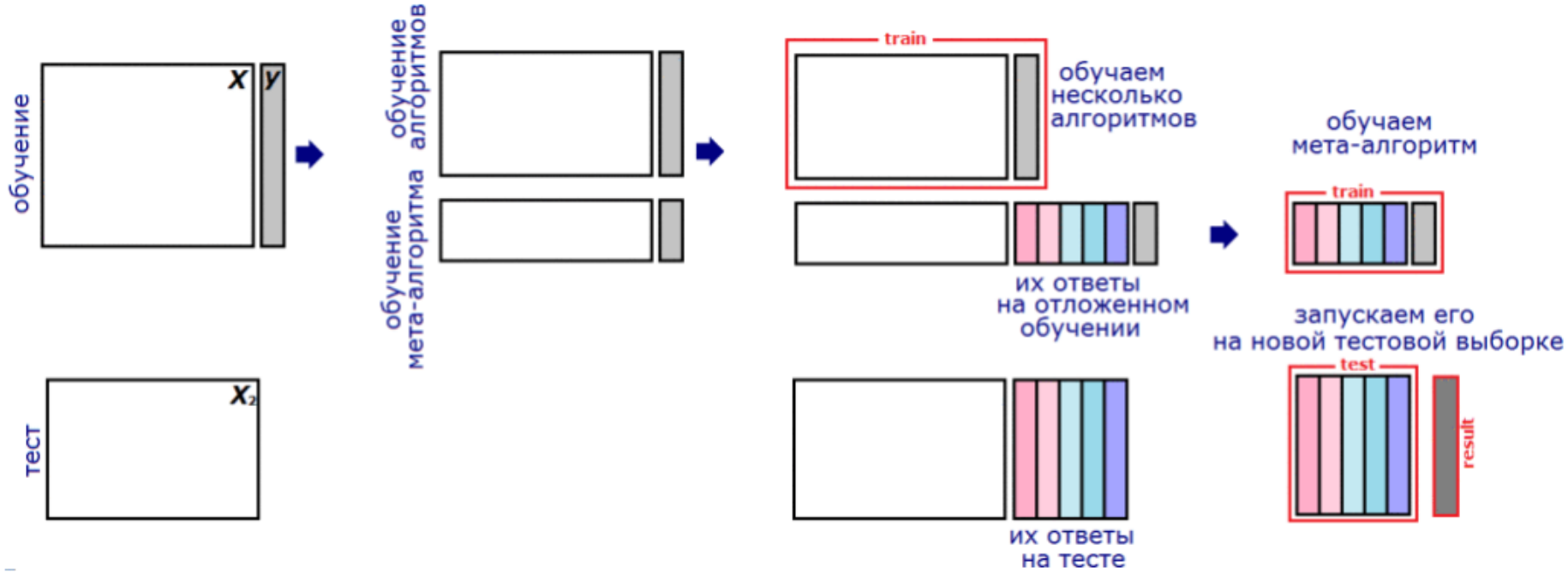
Для улучшения результата был также сделан блендинг классификаторов: **SGDClassifier**, **XGBClassifier** и **LogisticRegression.**

Для блендинга обучающая выборка была разделена на две части. На первой обучены базовые алгоритмы: **LogisticRegression** и **XGBClassifier.** Далее были получены их ответы на второй части и на тестовой выборке. Ответ каждого алгоритма представлял собой новый признак(метапризнак). На метапризнаках второй части обучения был настроен метаалгоритм - **SGDClassifier**. На метапризнаках теста получен конечный ответ.



БЛЕНДИНГ

Схема блендинга:



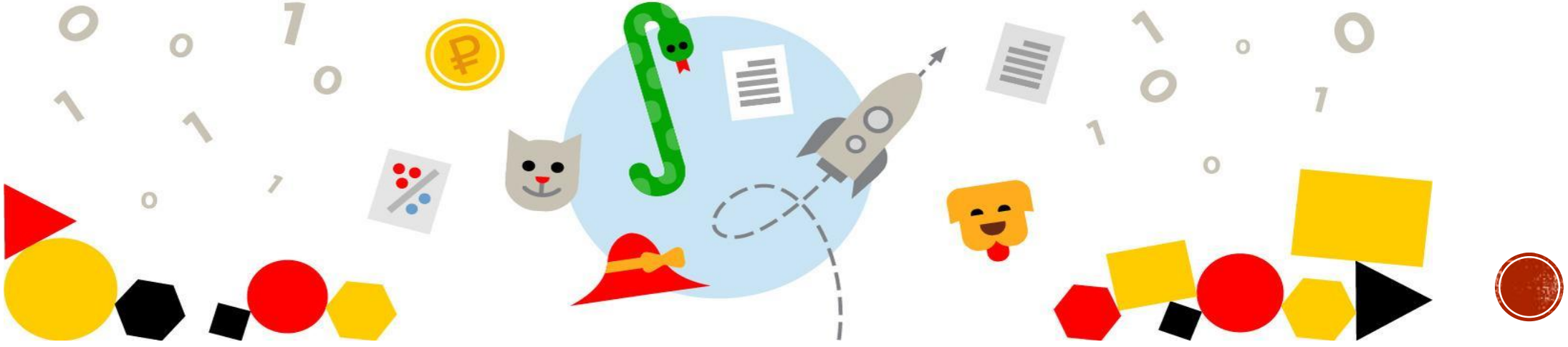
В результате **roc-auc** поднялся до **0.95**, **615** место на лидерборде – топ-30.

564	▾ 5	**kpp9966261**		0.95163	14	10mo
565	new	**Olga Proskuryakova**		0.95161	29	now

ВЫВОДЫ

По итогам соревнования:

- Первая модель **LogisticRegression** показала сразу неплохой результат, даже без подбора параметров и дополнительных признаков **roc_auc = 0.91**;
- С настройкой параметров результат **LogisticRegression** улучшился, но поднять его выше **0.93** не удавалось;
- Для улучшения результата был сделан блендинг 3 алгоритмов, показавших лучшее качество на отложенной выборке:**SGDClassifier**, **XGBClassifier** и **LogisticRegression**;
- Также для улучшения качества были добавлены новые признаки: день недели, время суток и число уникальных сайтов;
- На данный момент получен результат **roc_auc =0.95** , однако есть куда его улучшать;
- Для этого можно: поискать дополнительные признаки, рассмотреть варианты блендинга других алгоритмов(случайный лес, линейный **SVM**) и настройки их гиперпараметров.



ВЫВОДЫ ПО ПРОЕКТУ

В ходе проекта мы:

- познакомились с интересной задачей идентификации пользователей в интернете по их активности;
- Научились обрабатывать «сырые» данные, создавать разреженные матрицы;
- Узнали каким образом можно видоизменять исходные данные и дополнять;
- Научились придумывать новые признаки, которые помогают в решении конкретной задачи;
- Познакомились с новой библиотекой **VowpalWabbit**, быстродействие которой впечатляет;
- Получили опыт участия в соревновании **Kaggle**;
- Узнали, как повысить качество работы алгоритма.

Хочется отдельно отметить, что мне действительно помог проект в решении моей рабочей задачи, удалось прямо взглянуть на нее по-новому. Иначе говоря, если раньше я об этом знала и немного использовала, то сейчас активно применяю все полученные знания на практике, чему очень рада ☺