1 Идентификация пользователей по посещенным веб-страницам

Проделанная работа

Собраны данные по всем пользователям, данные разбиты на сессии фиксированной длины(длина сессии 15, с шагом 10). Далее параметр длины сессии не менялся и не подбирался, сам параметр подбирался основываясь на экспериментах на усеченных выборках, подбор длины сессии на полных данных был бы слишком затратным вычислительно. После выбора длины сессии был построен мешок сайтов(аналогично bag of words), редковстречающиеся сайты были убраны. Часть данных была визуализирована, с помощью временных меток были построены новые признаки. Далее построены эмбеддинги для сессий(были испробованы word2vec и one-hot, ohe-hot показали себя лучше). На эмбеддингах обучены 3 модели: логистическая регрессия, SVM, Catboost. Далее, используя лучшую модель, был идентифицирован конкретный пользователь.

Результаты

Лучше всего себя показал SVM, далее для SVM был подобран гиперпараметр C. Логистическая регрессия показала результат лучше, чем Catboost.

	Logistic Regression	SVM	CatBoost
Accuracy	0.212	0.303	0.198
Balanced Accuracy	0.118	0.2284	0.115
F1	0.131	0.2287	0.112

2 Dota 2

Проделанная работа

Все данные хранятся в формате jsonl, в качестве исходных данных дана подвыборках всех имеющихся. Сначала были собраны все основные признаки из jnosl файлов, по каждому игроку собраны дополнительные статистики.

Так как на результат всей команды не влияет порядок игроков(а в данных игроки упорядочены), то было принято решение перейти от всех индивидуальных признаков и создать по каждому индивидуальному отдельный командный признак. Так, по признакам игроков были агрегированы командные признаки, описывающие силу команды по данной фиче. Из индивидуальных признаков остались лишь номера персонажей, вместо отдельных номеров были созданы one-hot вектора всей команды, для обеих сторон.

На этих данных были обучены 3 модели : Logistic Regression, Random Forest, CatBoost. Лучше всего себя показал CatBoost, случайный лес показал худший результат. Для всех моделей были подобраны гиперпараметры.

Всего в данных получилось около 500 признаков, были проведены эксперименты с PCA. Как оказалось, количество признаков можно сильно уменьшить, без особой потери качества.

В конце были посчитаны важности признаков для CatBoost

Результаты

CatBoost

Logistic Regression

Random Forest

Лучшие результаты на тесте дали CatBoost и Logistic Regression, попробуем взять взвешенное среднее их предсказаний, у CatBoost возьмем вес 0.7, у логистической регрессии 0.3.

Получили такой скор