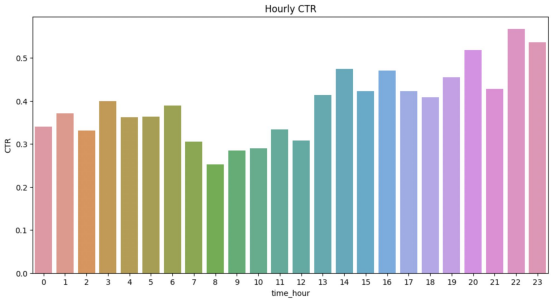
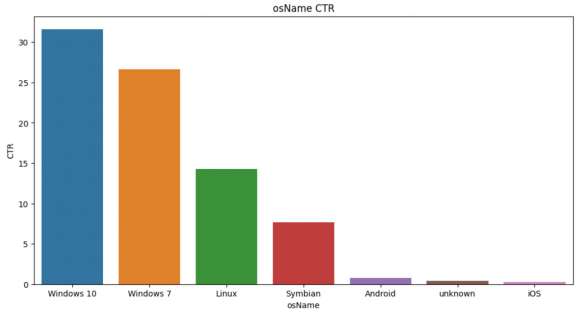
(Human + ML) ~~vs~~ + LLM = ?

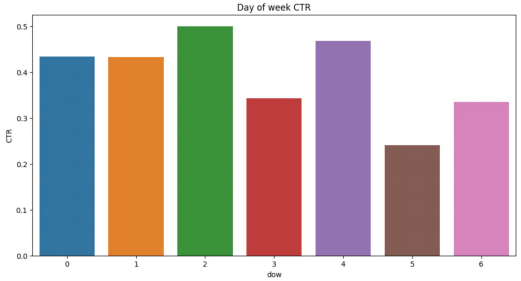
EDA (1)

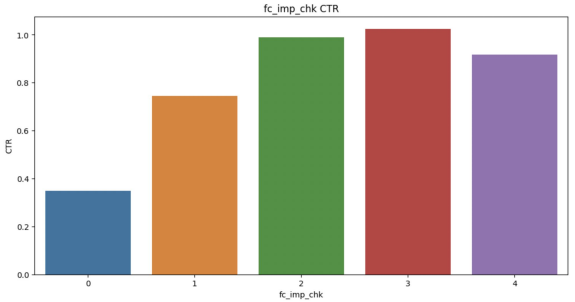
tag_click	count	share_%
0	951231	99.608
1	3747	0.392

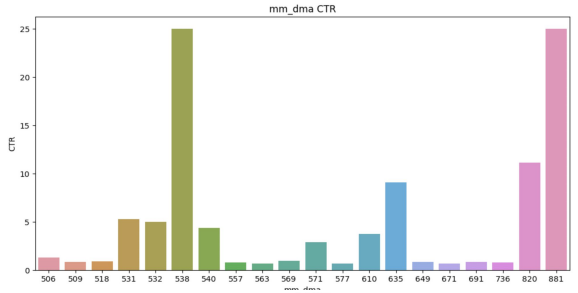
сильная разбалансировка классов

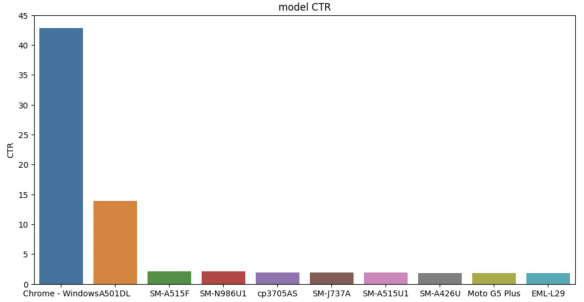


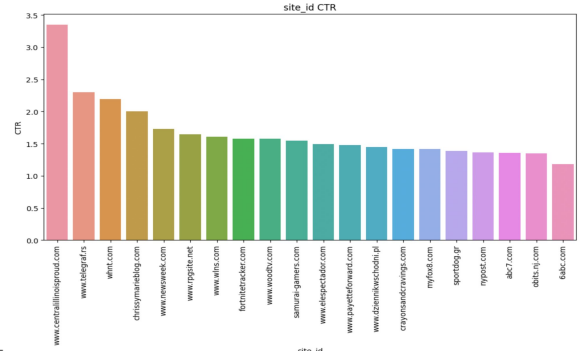




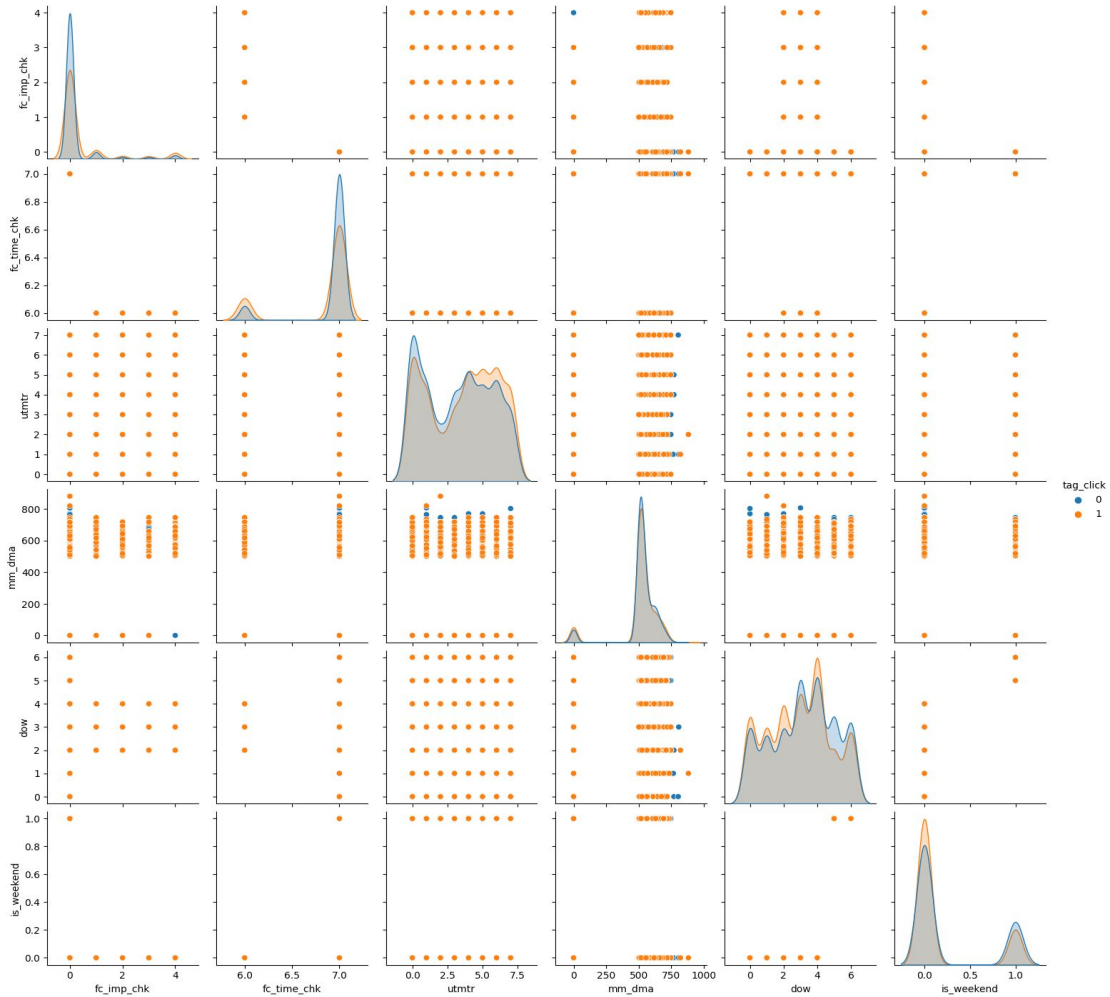








При том, что вероятность клика ниже 0.4%, некоторые факторы увеличивают значение CTR до 40% (но при низкой частоте появления события, поэтому требуется более сложный инструмент анализа - ML). При этом существует дифференциация CTR от дня недели, часа просмотра, региона, сайта просмотра и пр.



EDA (2)

1. Из-за разбалансированности классов сделана RandomUnderSampler (балансировка классов);
2. На графике видно:
   a. смещение распределения целевой метрики относительно дня недели (в будние дни, особенно в пятницу) вероятность клика выше;
   b. в некоторых DMA в зависимости от дня недели вероятность клика снижается (или вообще нет события в определенные дни недели);
   c. частота клика относительно частоты показов смещается к вечернему периоду (18:00 - 00:00);
   d. частота показов первый раз не приводит к соразмерному увеличению частоты клика .

TASK

Identify Important Factors: Find out what factors are most likely to make someone click on online ads. Tell me which factors are the strongest predictors and rank them.

HUMAN + ML = …

1. Разделение датасета на train (611 185 объектов; в том числе распределение по классам: {0: 608784, 1: 2401}), validation (152797 объектов; {0: 152176, 1: 621}), test (190996 объектов; {0: 190271, 1: 725});
2. Балансировка классов на тренировочном датасете с использованием RandomUnderSampler (после балансировки тренировочный датасет: 7203 объектов; {0: 4802, 1: 2401});
3. Features: ["fc_imp_chk", "fc_time_chk", "utmtr", "mm_dma", "osName", "model", "hardware", "site_id", "dow", "is_weekend", "minute_sin", "minute_cos"];
4. Модель (ML): CatBoostClassifier;
5. Оптимизация гиперпараметров: optuna;
6. Метрика качества: PRAUC (Precision Recall AUC).

PRAUC (test) = 0.027

feature	feature_importance
site_id	44.380
dow	8.723
mm_dma	7.994
model	7.248
osName	7.208
utmtr	6.318
fc_imp_chk	5.433
minute_cos	3.849
minute_sin	3.247
hardware	3.060
is_weekend	2.054
fc_time_chk	0.485

	precision	recall	threshold
count	142,075.000	142,075.000	142,075.000
mean	0.010	0.830	0.246
std	0.011	0.237	0.229
min	0.000	0.000	0.020
1%	0.004	0.072	0.028
10%	0.004	0.412	0.040
20%	0.005	0.701	0.067
30%	0.006	0.859	0.082
40%	0.006	0.910	0.098
50%	0.007	0.934	0.117
60%	0.009	0.959	0.175
70%	0.011	0.975	0.325
80%	0.013	0.986	0.547
90%	0.016	0.997	0.622
95%	0.021	0.999	0.664
99%	0.033	1.000	0.759
max	0.750	1.000	0.986

Устанавливая значение трешхолда для модели предсказания, например, 0.759, можно увеличить вероятность клика почти в 10 раз (precission = 3.3%) в сравнение с текущей вероятностью 0.39%

HUMAN + LLM = …

1. Датасеты (train, validation, test) подготовлены для модели;
2. Promt уточнен (расширен):
   a. прописан стандартный план создания и тестирования ML модели;
   b. рекомендована модель ML;
   c. рекомендована метрика качества для тестирования модели;

PRAUC (test) = 0.024

feature	feature_importance
site_id	59.734
osName	7.124
mm_dma	6.050
fc_time_chk	5.542
dow	4.828
model	4.327
utmtr	4.132
minute_sin	2.879
hardware	2.278
minute_cos	1.695
is_weekend	1.030
fc_time_chk	0.379

HUMAN ~~vs~~ + ML ~~vs~~ + LLM

feature	feature_importance_ML	feature_importance_LLM
site_id	44.380	59.734
dow	8.723	4.828
mm_dma	7.994	6.050
model	7.248	4.327
osName	7.208	7.124
utmtr	6.318	4.132
minute_cos	3.849	1.695
minute_sin	3.247	2.879
hardware	3.060	2.278
is_weekend	2.054	1.030
fc_time_chk	0.485	5.542
fc_time_chk	0.485	0.379

1. Полученные сопоставимые результаты в результате реализации двух подходов;
2. Попытки сформировать ответ на поставленный вопрос LLM без “помощи” не привели к значительному эффекту;
3. Низкое значение PRAUC говорит о необходимости включения в модель дополнительных фичей, например, атрибуты пользователя (возраст, пол, образование и т.д.) и его предпочтения (статистика посещения сайтов, статистика покупок и т.д.)