

OZON ozon{ech

Cats DS

Задача 05 «Поиск одинаковых товаров на маркетплейсе»



КОМАНДА «Cats DS»



**Иван
Черных**

- DS / проджект
- @iceman_o_o
- +79263351353



**Толстова
Ольга**

- Аналитик / DS
- @otolstova
- +79032284070



**Степан
Куткин**

- Аналитик / DS
- @step203
- +79129413005



**Егор
Геращенко**

- Аналитик / документация
- @egorgera
- +79995379098

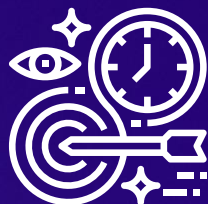


**Дмитрий
Саханенко**

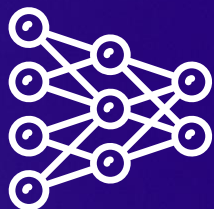
- Капитан / презентация
- @DimirSDV
- +79650359512

Описание задачи

1. АКТУАЛЬНОСТЬ



- Ozon — ведущая мультикатегорийная платформа электронной коммерции и одна из крупнейших интернет-компаний в России. На площадке представлено более 150 млн товарных наименований в 20 категориях: от книг и одежды до продуктов питания и товаров для здоровья.
- Сейчас более 90% ассортимента площадки формируют партнеры маркетплейса, в некоторых случаях предлагающие одинаковые товары по разной стоимости и с разными сроками доставки.
- Ozon нужно постоянно совершенствовать алгоритм определения одинаковых товаров, чтобы клиенты лучше ориентировались в предложениях продавцов



3. ОПИСАНИЕ ИТОГОВОГО ПРОДУКТА

- ML-модель

2. ОПИСАНИЕ ЗАДАЧИ

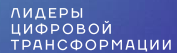


- Разработайте ML-модель, способную определить идентичность товаров по названиям, атрибутам и изображениям.
- Модель должна находить среди пар-кандидатов как можно больше одинаковых товаров с точностью >95%

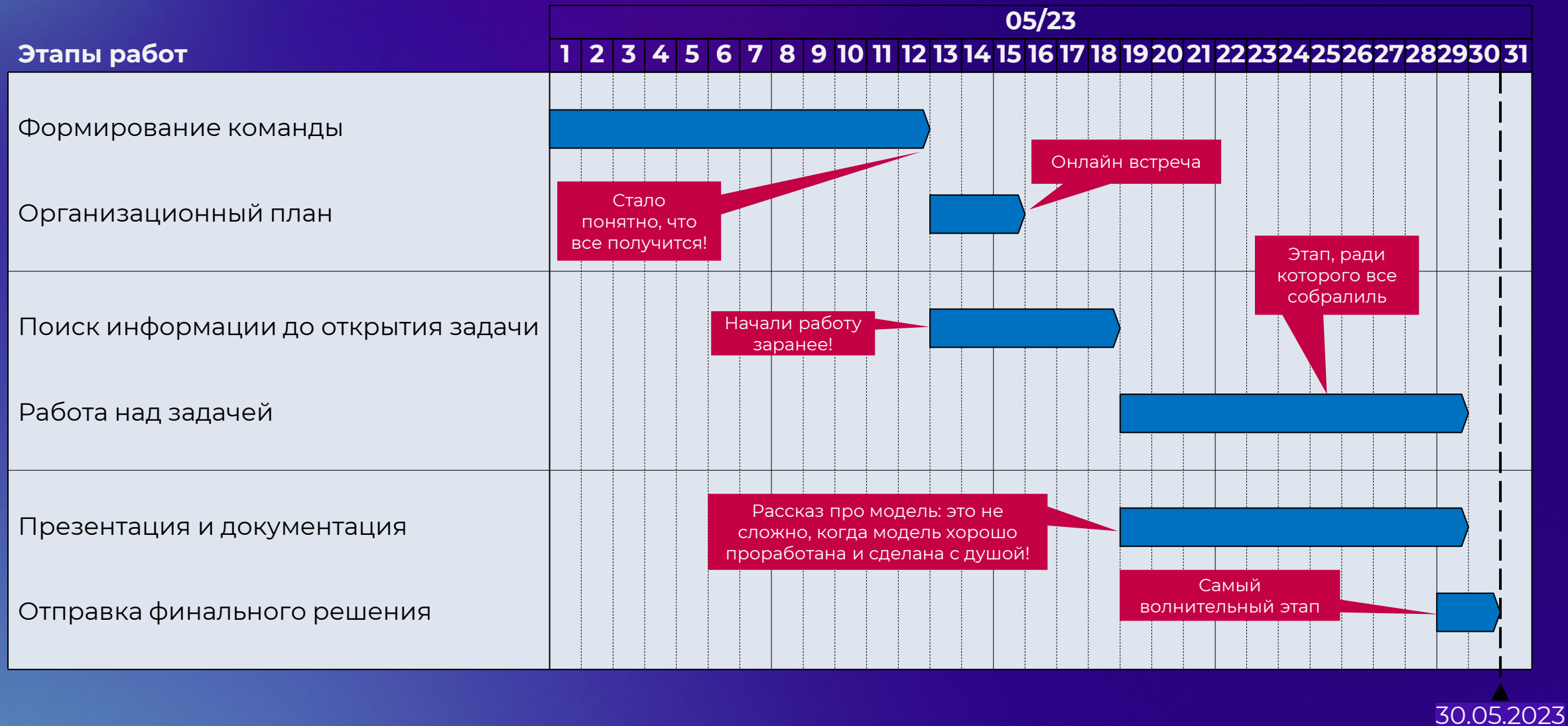
4. РЕСУРСЫ



- Тренировочная выборка: пары одинаковых и различных товаров;
- Тестовая выборка: пары товаров без разметки (выборка для формирования лидерборда);
- Дополнительные данные: названия, атрибуты, векторные представления картинок (эмбединги) товаров



Организационный план





Подходы к решению задачи: названия, цвета и новые признаки



Основные проблемы:

1. Названия товаров на платформе не унифицированы (каждый продавец пишет, как хочет)
2. Похожие картинки могут относиться к совершенно разным товарам



Варианты решения проблем:

1. Выделить признаки из названий: вид товара, марки, модели, характеристики
2. Сделать основой решения описания, а картинки использовать как вспомогательный атрибут



Решение команды Cats DS

1. Работа с названиями:
 - словарь цветов
 - выделение категорий
 - словарь «АнтиСлов»: различия по ключевым словам (Pro, Max и т.д.)
2. Формирование новых признаков

Алгоритм решения задачи

Ключевые этапы решения задачи





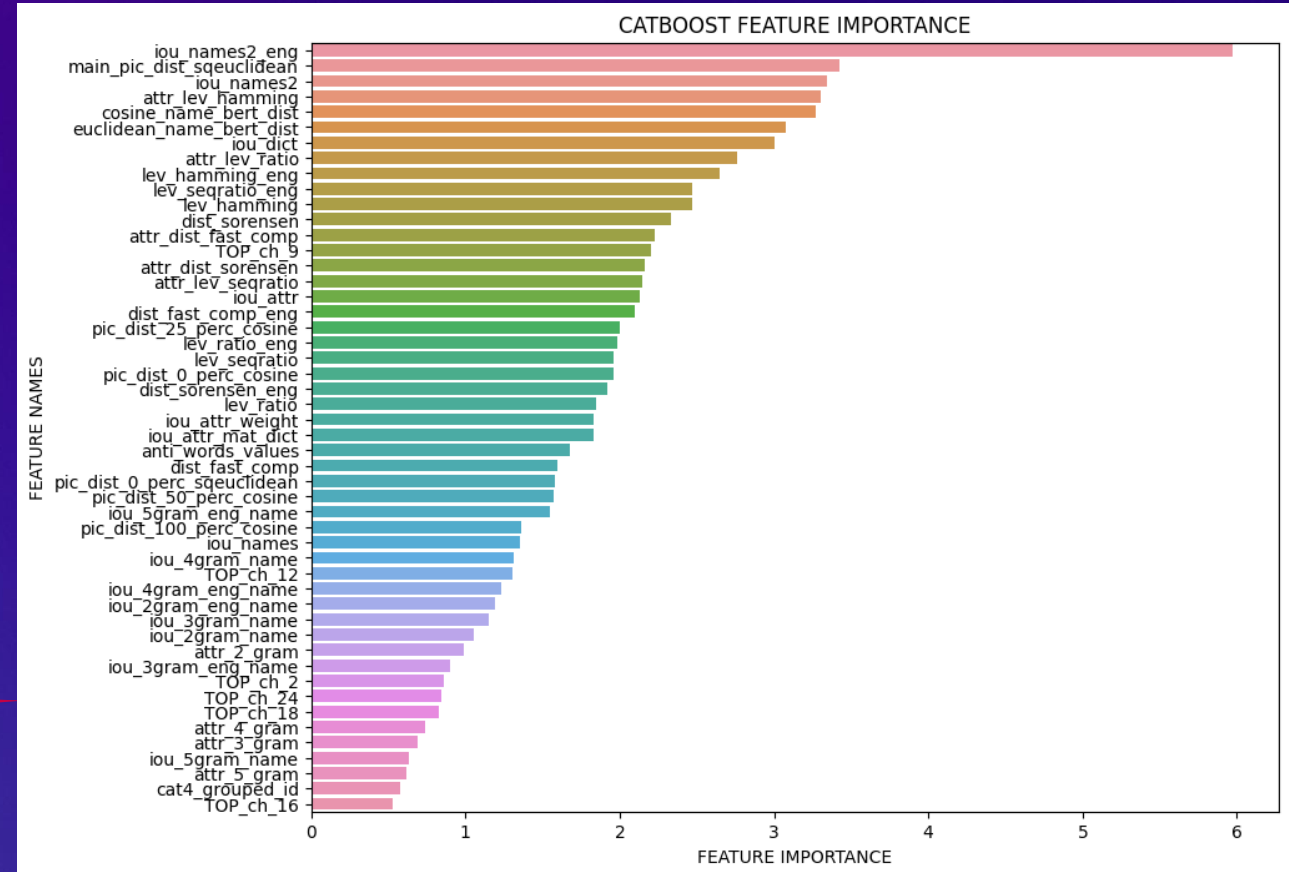
Анализ важности сформированных признаков

ТОП-5 важных признаков

- Наиболее важные признаки:
1. iou_names2
 2. iou_names
 3. pic_dist_0_perc_nan_bert_dist
 4. euclidean_name_bert_dist
 5. pic_dist_100_perc_cosine

Признаки из низшей части
таблицы менее важны для
обучения модели

Распределение признаков по важности





История решения задачи

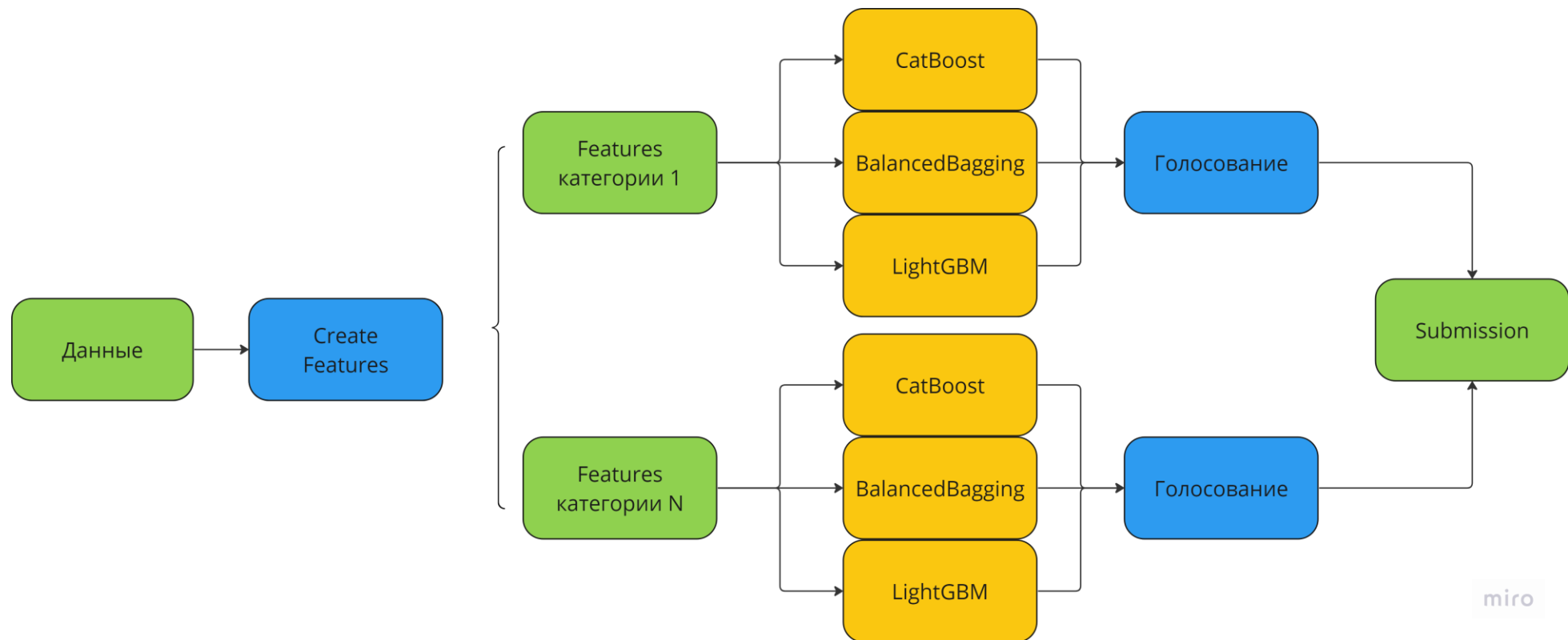
Решения, которые увеличили метрику PR-AUC



Архитектура решения

Особенности архитектуры:

- Формирование новых признаков
- Использование ансамбля моделей для каждой категорий товаров





Ключевые особенности решения

Формирование новых признаков

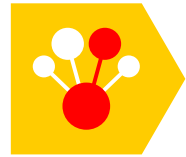
1. Использование IoU (Жаккара, бинарная мера сходства) над названиями и характеристиками
2. N-граммы (биграммы и триграммы) над названиями и характеристиками
3. Словарь «Антислова»: улавливает различия товаров по ключевым словам (например, Pro, Max)
4. Разложение цвета на простые цвета и кодирование как категориальный признак
5. Расчет важности каждой характеристики внутри каждой группы: определяется ТОП характеристик, сравнение по которым подается как отдельный признак

Ансамбль моделей

1. Общая модель, результат которой подается в качестве дополнительных признаков
2. Отдельная модель для каждой товарной категории
3. Использование Under-сэмплинга и Over - сэмплинга
4. Ансамбль моделей: усреднение результатов от CatBoostClassifier, BalancedBaggingClassifier, LGBMClassifier



Используемые библиотеки



CatBoost

Бустинг отлично подходит для создания ML на табличных данных



LightGBM

Центральная библиотека для ML



Пожалуй, лучшее, что создали для работы с таблицами

Визуализация данных

matplotlib

И снова работа с графиками



seaborn

tqdm: отображение прогресса обучения



BalancedBagging Classifier





SciPy: высокая скорость для сложных вычислений





Итоговые метрики и ранкинг



ЛЦТ. Поиск одинаковых товаров на маркетплейсе

2 000 000 Р - Призовой фонд

21 мая 2023 г. - 30 мая 2023 г. Участвую

Обзор

Данные

Турнирная таблица

Команда

Загрузить решение

Комментарии

Публичная турнирная таблица

Приватная турнирная таблица

Место	Название	Решения	Pr_Auc_Macro	Time (s)	Награда
11	WhileTrue	15	0.38478		
12	Laboratory	20	0.36741		
13	ML Rocks	2	0.32277		
14	ГоЛо	3	0.3063		
15	flow	10	0.2904		
16	ВОСТОК1	21	0.29038		
17	Cats DS	15	0.28616		
18	Сквозь турникеты в ML	33	0.28468		
19	fffrt	9	0.27245		
20	Салават Купере	16	0.20938		

<

1

2

3

4

5

...

8

>

10 / стр. ▾

Финальная метрика
Pr_Auc_Macro

0,28616

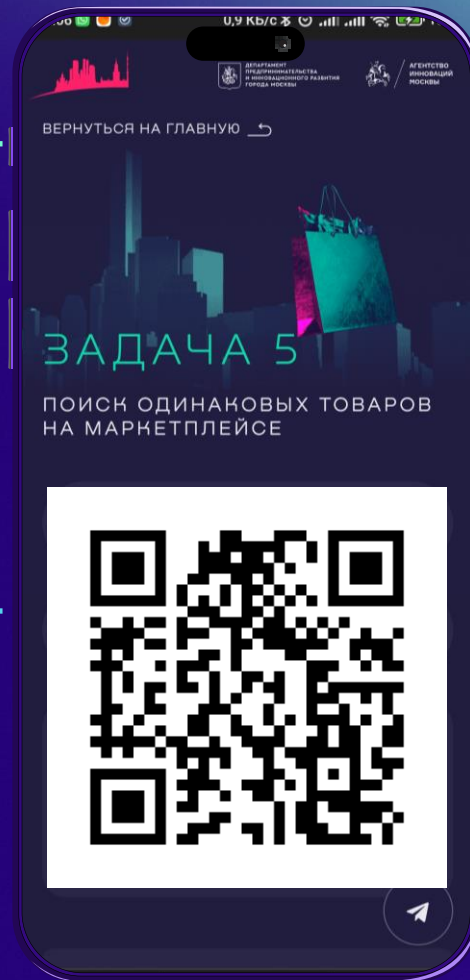
Ранг в публичной
турнирной таблице

17

Ссылка на репозиторий GitHub

- Решение в формате .ipynb

- Презентация



- Сопроводительная документация

- Пример csv-файла