avito.tech

AVITO ML CUP

Поиск дублей

MISIS Neychev Loss

Top 2 Public LB, Top 3 Private LB

Наша команда

Рыжичкин Кирилл

ex CV RnD @ SBER AI

Аксеновский Максим

CV Engineer @ X5

Груздев Александр

Data Scientist @ GPB

Герасин Тимофей

NLP Researcher @ Huawei

Начальный подход

1 Категориальные признаки

Мэтч по категориям 1-4 уровня, частичный мэтч по 4 уровню, полнота столбца.

2 Текстовые признаки

Частичные мэтчи по названиям и описаниям, BM25, LCP и LCS. Сходства для строк и списков. IOU по n-gram. Антислова. Отношения и разницы длин для всего подряд (l, r)

3 Атрибутивные признаки

Совпадения для словаря атрибутов, топ-атрибуты по категориям

CV	0,48739
LB	0,33978

Неприятный баг

Думали будет удобно...

Соединили трейн и тест для препроцессинга

Ошибка новичка

Неправильно смержили обратно и раздули трейн в 2 раза

Пофиксили

Исправление ошибки дало значительный прирост

CV	0,52389
LB	0,34861

Небольшие изменения

GroupKFold → StratifiedGroupKFold

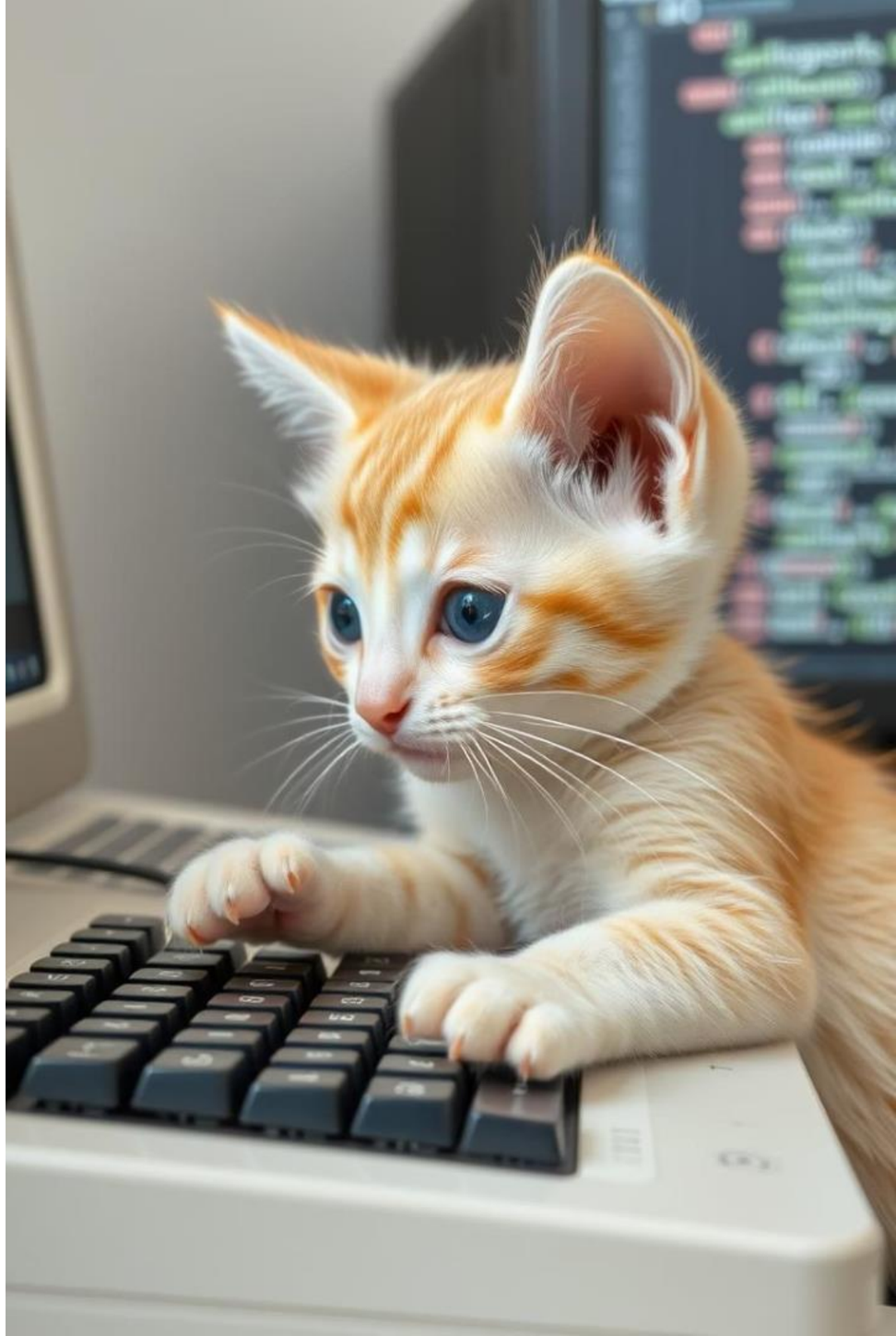
CV: 0,5249 | LB: 0,3495

Отказ от весов классов в лоссе

CV: 0,5418 | LB: 0,3552

Ручная настройка lr=0.1

CV: 0,5498 | LB: 0,3598



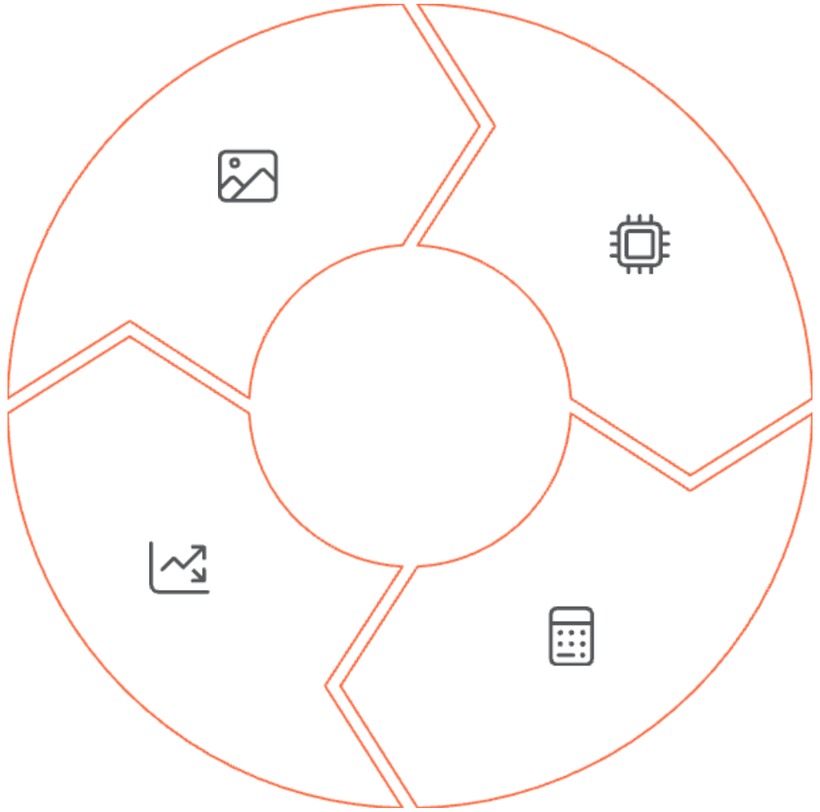
Докидываем картинки

Эмбеддинги

Использовали претрейн openai/clip-vit-large-patch14

Обработка

Заняло одну ночь на локальной 2060



Результат

Значительное улучшение метрик

Косинусное сходство

Рассчитали сходство между изображениями товаров

CV / LB	0,5659 / 0,3774

LightAutoML и категории

LightAutoML

Заменили соло CatBoost на LAMA с 5 фолдами.
Использовали lgb, lgb_tuned, cb, cb_tuned

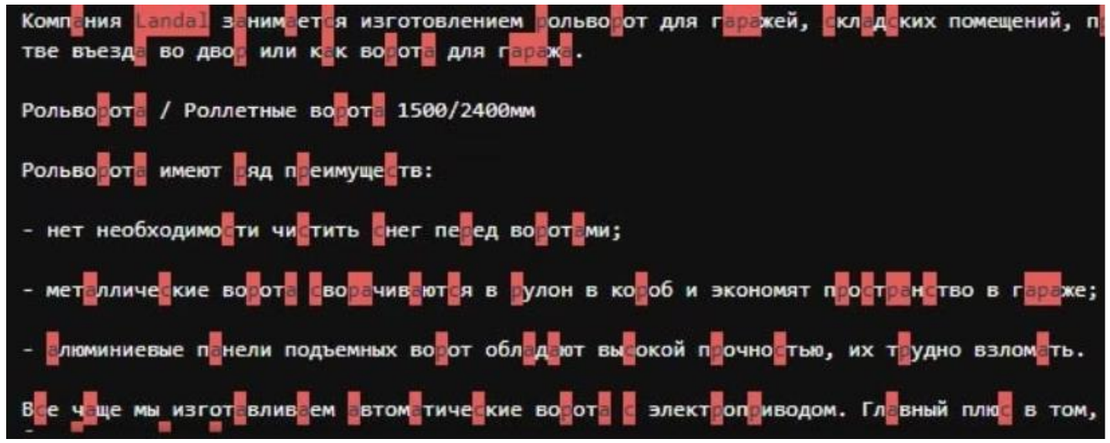
CV	0,5699
LB	0,3824

Категории как признаки

Тут мы вспомнили, что забыли подать сами категории товаров как фичи… Исправляемся

CV	0,5691
LB	0,3920

Чисти вилкой, чисти



Это печально

Выявили, что 97% слов в датасетах были с подменой букв — схожие по написанию русские и английские буквы заменялись друг другом.

Вполне ок...

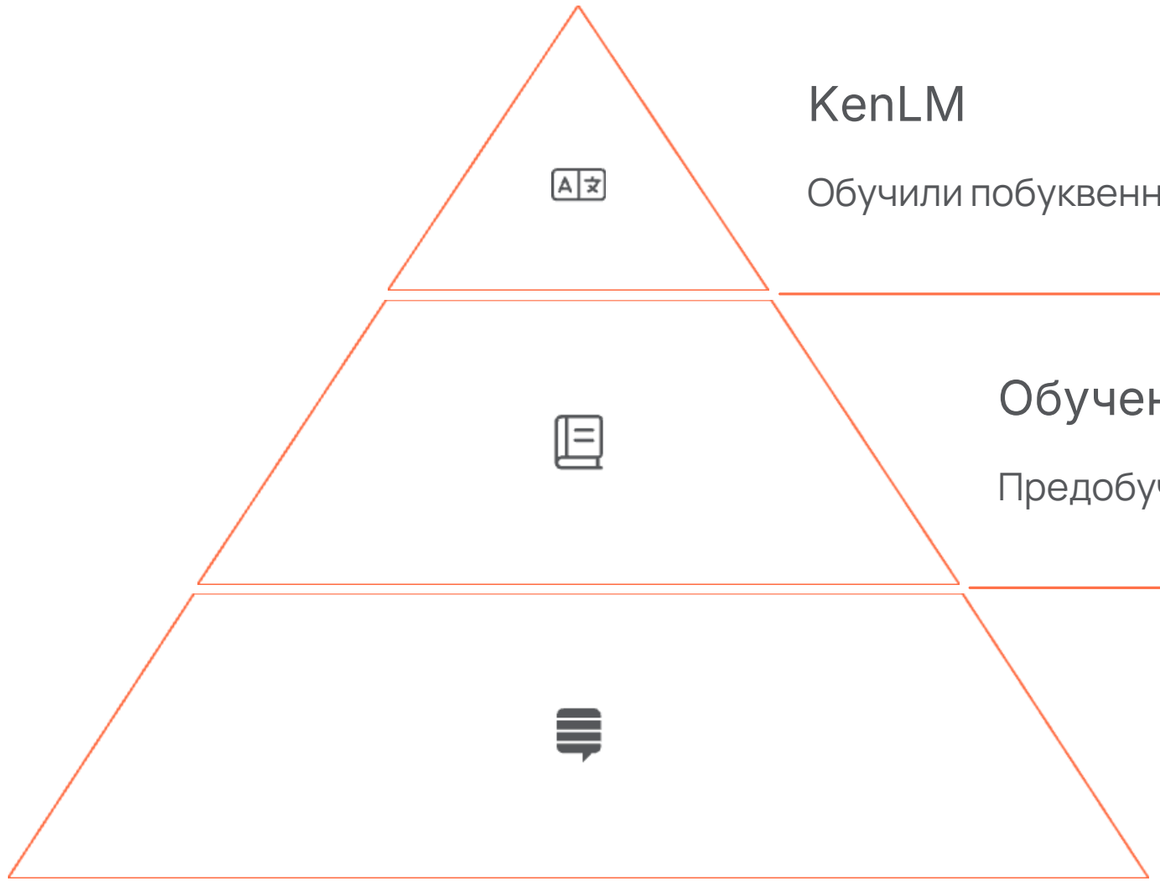
Определять язык слова по мажорирующему языку:

если больше n% букв в нем русские – слово русское.

Иные подходы

Фикс с мультиязычными токенизаторами и перплексией мультиязычного берта оказались менее эффективными.

Братишка, как я kenlm-ом буду чистить?



KenLM

Обучили побуквенную n-граммную модель для фикса букв!

Обучение

Предобучение на русской и английской Википедии

Качество

Данный подход показал себя лучше прошлого наивного фикса по порогу.

На инференсе считали перплексию для двух вариантов: en→ru и ru→en. Выбирали лучший вариант только когда в слове смешаны разные языки.

Новые фичи

- **Расширение текстовых признаков:** Добавили n-грамм IOU (1-7) для описаний товаров.
- **Улучшение метрик схожести:** Добавили token set/sort ratios и dice для описаний товаров.
- **BM25 и TF-IDF:** Применили их к названиям, описаниям и конкатенированным атрибутам.
- **Взвешенное совпадение атрибутов:** Учитывали совпадения ключей и значений для повышения точности.
- **Битые слова:** Включили отношения и абсолютные разницы процентов некорректных слов.

Новая картиночная модель

Взяли обученный под домен претрейн. Также тут добавлены фичи с прошлого шага + фикс букв.

0.589	0.401	1
Скор CV	Скор LB	Модель Marqo-SigLIP
Невероятно!	Эвано как!	Зато какая!

DL для текстов

Мы использовали предобученную модель sergeyzh/BERTA.

Косинусное сходство рассчитывалось для строк, объединяющих название, категорию и описание товара.

0.60

Скор CV

0.41

Скор LB

Еще одна картиночная модель

0.601

Скор CV

+0.01!

0.413

Скор LB

+0.03!

1

marqo-ecommerce-embeddings-L

Зато какая!

FastText

Мы обучили модель FastText на объединенных описаниях товаров. TTA заметно поднял скор.

Также помимо ABAB, ABBA инференса пробовали BAAB и BABA, но прироста не получили.

Метрика	Без TTA	С TTA
CV	0.6096	0.6111
LB	0.4198	0.4235

Еще одна картиночная модель

В этот раз обратились к соревнованию Kaggle 'Shopee - Price Match Guarantee'. Топы там учили картиночные модели на arcface_loss на группы товаров. Только лишь веса топ-5 команды уцелели под грузом времени.

0.6129

Скор CV

0.4252

Скор LB

Немного текстовых претрейнов

Добавили multilingual-e5-large-instruct, снова небольшой прирост

0.6156

Скор CV

0.4273

Скор LB

Будущие эксперименты с userbge-m3 также дадут нам дополнительный небольшой прирост.

Модели по категориям

Мы обучили специализированные модели для каждой из семи категорий первого уровня.

7

Категорий

С отдельными моделями.

0.5 + 0.5

Веса блендинга

Общая и категориальная.

0.436

Скор LB

Финальный результат на лидерборде.

Вспоминаем про LAMA

Заменяем общий CatBoost на LAMA. Категориальные модели CatBoost остались без изменений.

0.4420

Скор LB

Учим картиночную модель

Для более точного сопоставления изображений мы сосредоточились на дообучении своей модели.

Выбор Модели

Дообучили **timm/resnet50.a1_in1k**, быструю и производительную модель.

Метод Обучения

Использовали **Contrastive loss** для эффективного обучения признакам. (OCL оказался хуже)

Оценка Эффективности

0.19 PRAUC на валидации, что выше всех прежних картиночных претрейнов!

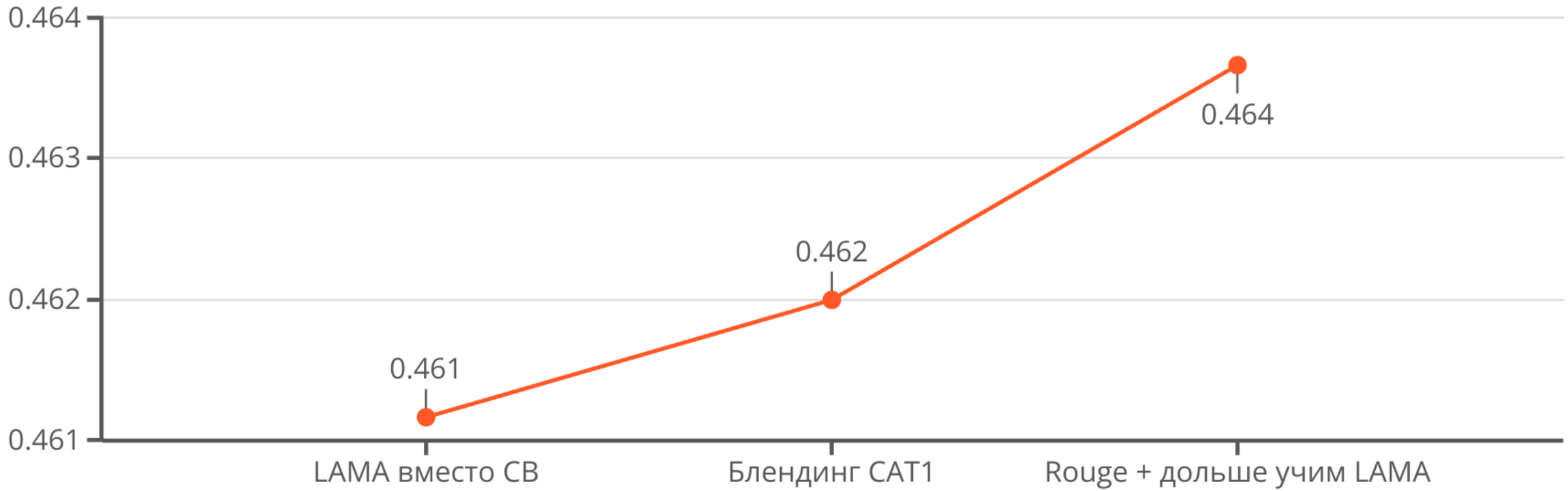
Свой DL на текстах

Дообучили rubert-base на конкатах товаров. Только соло CatBoost как мета-модель!

Картиночная модель resnet с прошлого шага была также интегрирована.

0.626	0.446	0.455
CV Скор	LB Скор	LB Скор с TTA

Снова LAMA



1. Заменили соло CB на LAMA, а также преблендили соло катбусты по категориям и получили
2. Бленд *0.5 * lama_full + 0.5 ** (0.7 * *cb_cat1 + 0.3 **lama_cat1)
3. Добавили rouge по текстам как фичу и также поучили LAMA подольше

Постпроцессинг результатов

После алаймента вероятностей	0.465
После ручной корректировки (оверфит под лб...)	0.471

$$p_1^{adj} = \frac{p_1\pi_1\rho_0}{p_1\pi_1\rho_0 + (1-p_1)\pi_0\rho_1}$$

- Сначала выполнили алаймент категориальных моделей под общую модель LAMA.
- Потом руками домножили вероятности: для животных и электроники на 1.2, для транспорта на 1.3.

Что не удалось

- **TabM:** Отлично показал себя соло (0,4222 на LB), но не смог улучшить общий бленд моделей.
- **Ограничения моделей:** Жирные текстовые и картиночные модели не поддавались обучению :(
- **Постпроцессинг:** Дополнительная обработка (id1,id2) и (id2,id1) не принесла прироста к скору.
- **LAMA по категориям:** Обучение LAMA для отдельных категорий не улучшило общую метрику.
- **Псевдолейблинг:** Применение псевдометок также не привело к ожидаемому повышению метрики.
- **CL > OCL:** Online Contrastive Loss работал хуже обычного Contrastive Loss.
- **Модели по категориям 2 уровня:** Модели по категориям второго уровня тянули скор вниз.
- **Отбор фичей:** Отбор фичей оказался для нас бесполезен (+долгий), в итоге использовали все 470.
- **Разметка новых категорий:** Разметка трех новых категории с помощью LLM/VLM тянула вниз.

Что не успели

1. **Расширение датасета:** Планировали использовать транзитивные цепочки для обогащения данных.
2. **Кросс-энкодер на атрибутах:** Давал 0.2 prauc в соло, не успели интегрировать как фичу.
3. **Модели по кластерам:** Была идея обучать модели по кластерам пользователей, а не только по категориям.
4. **Обучение LLM/VLM:** Трейн 3b 4bit модели требовал 800 часов на 150k сэмплов :(

Корреляция метрик

Важным аспектом было сопоставление результатов на лидерборде с внутренней валидацией. Этот график демонстрирует динамику ключевой метрики PRAUC в разных итерациях.

Корреляция Пирсона	0,978277994

Спасибо за внимание!

Мы были рады поделиться нашим опытом и результатами!