Итоговый отчёт по проекту

14 мая 2025 г.

Содержание

1 Введение и Актуальность **3**

2 Состав команды и роли **4**

3 Описание заказчика проекта и проблемы **4**

4 Бизнес цели **4**

5 Бизнес модель **6**

6 Оценка экономики проекта **7**

6.1 Затраты на проект 7

6.2 Окупаемость проекта 8

7 Анализ датасета **9**

7.1 Информация о наборе данных 9

7.2 Качество данных 15

7.3 Обогащение данных 16

8 Предобработка данных **18**

8.1 Методы предобработки данных 18

8.2 Подготовка данных для обучения 19

9 ML-решение **19**

9.1 Выбор и разработка модели 19

9.2 Настройка и обучение модели 20

9.3 Система учета экспериментов 20

9.4 Сравнение моделей 21

9.5 Результаты и интерпретация 22

9.6 Подбор гиперпараметров 23

9.7 Важность признаков 23

10 Валидация и тестирование **23**

10.1 Методы валидации 23

10.2 Выбор метрик качества 24

11 Заключение **24**

12 Приложение **24**

1. Введение и Актуальность

В последние годы рынок онлайн-торговли значительно расширился, что привело к увеличению числа продавцов на маркетплейсах. Однако многие из них на старте сталкиваются с трудностями при составлении объявлений. В частности, им часто не хватает знаний о том, какие ключевые слова и визуальные элементы необходимы для того, чтобы увеличить количество просмотров и продаж.

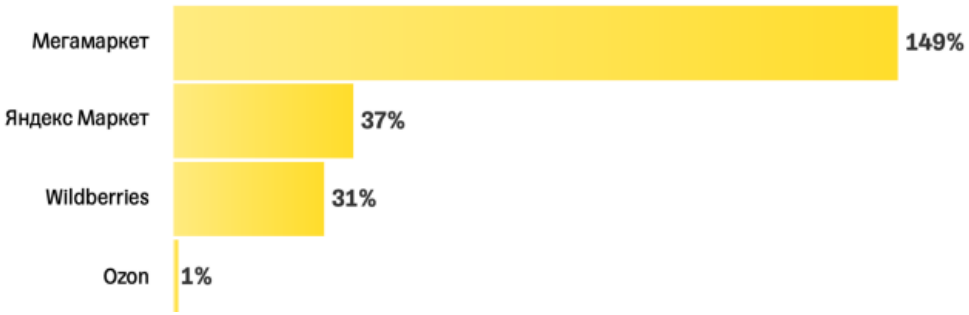

Рис. 1: Прирост числа селлеров по площадкам за 2024 год

Из-за большого количества товаров одной категории на торговых площадках продавцам приходится уделять больше внимания качеству своих объявлений, чтобы выделиться на фоне конкурентов. Простое размещение товара больше не гарантирует продажи — важна каждая деталь: от привлекательных фотографий до убедительного описания.

Покупатели стали более избирательны, сравнивая множество предложений перед покупкой. В таких условиях грамотно оформленное объявление становится ключевым инструментом для привлечения внимания. Высококачественные изображения, которые демонстрируют товар с разных ракурсов, детальные и честные описания характеристик, выгодные условия доставки и акционные предложения помогают повысить вероятность покупки.

Проект направлен на разработку прототипа сервиса, который предоставит продавцам маркетплейсов возможность оценки описаний товаров (включающих текст и фотографии). Этот инструмент позволит даже неопытным пользователям определить, является ли привлекательным и эффективным их объявление.

Система призвана помочь новичкам в продажах, а также снизить нагрузку на специалистов по публикациям на маркетплейсе, упростив процессы размещения товаров. В результате, продавцы смогут быстрее привлекать внимание покупателей, а маркетплейс повысит свою прибыль, удерживая более успешных продавцов и уменьшая отток новых пользователей.

2. Состав команды и роли

ФИО	PM	BU	DS	ML	BA	Итого
Аскарбек Казыбек	0	0.1	0.3	0.5	0.1	1
Дельман Александр	0.1	0.1	0.3	0	0.5	1
Кириенко Владислав	0.75	0.1	0.05	0	0.1	1
Партин Максим	0.15	0.6	0.05	0	0.2	1
Шевченко Кирилл	0	0.1	0.3	0.5	0.1	1
Итого	1	1	1	1	1	

Таблица 1: Состав команды и роли

3. Описание заказчика проекта и проблемы

ProstoMarket - российский интернет маркетплейс. Занимает одну из лидирующих позиций на российском рынке. Его ближайшие конкуренты: Wildberries, Ozon, Avito и другие. Продавцы на платформе ProstoMarket часто испытывают трудности с продажей своего товара. Это влечет к их разочарованию сервисом и прекращению его использования. Для компании ProstoMarket отток продавцов введет к уменьшению их потенциальной прибыли. С каждого продавца маркетплейс берет комиссию от 4% до 18%. В таблице 2 представлены размеры комиссии для основных самых популярных категорий товаров.

Категория товара	Взимаемый процент
Красота и здоровье	7%
Товары для дома	8 %
Одежда, обувь и аксессуары	9 %
Техника	14%
Автомобили	18%

Таблица 2: Процент комиссии, взимаемой маркетплейсом

Для решения оттока клиента компания ProstoMarket наняла команду AdOptimizer, специализирующую на оптимизации объявлений продавцов на маркетплейсах.

4. Бизнес цели

Компанией ProstoMarket была представлена детальная финансовая отчетность за последние 5 лет (2019 - 2023 года).

В частности, компания предоставила следующие данные. Количество продавцов по годам:

Год	**Количество продавцов**
2019	38 тыс.
2020	56 тыс.
2021	90 тыс.
2022	120 тыс.
2023	450 тыс.

Таблица 3: Количество продавцов по годам

По этим данным видно, что за последние 5 лет произошел экспоненциальный рост числа продавцов, связанных с рядом исторических событий (пандемия, внедрение повсеместной удаленной работы и другие). По расчетам аналитиков компании количество продавцов к концу 2024 года должно быть 600 тыс. К 2025 году компания планирует выйти на IPO (первичное размещение акций).

Также компания предоставила детализированную информацию по числу продавцов в каждой из основных категорий на 2023 год (продавцы могут работать в нескольких разных категориях).

Категория	**Процент продавцов**
Красота и здоровье	20%
Товары для дома	50 %
Одежда, обувь и аксессуары	35 %
Техника	5 %
Автомобили	1%

Таблица 4: Процент продавцов по категориям

А также данные по оттоку продавцов с 2020 по 2023 год

По этим данным видно, что компания находится в стадии активно роста, но имеет проблему связи с резким оттоком новых продавцов (рис. 2). Отток продавцов в 2023 году составил целых 30%.

Были поставлены следующие цели команде AdOptimizer достижение, которых ожидает руководство маркетплейса ProstoMarket:

- Уменьшить отток продавцов с маркетплейсов с 30 % в 2023 году к 10 % к 2027 году;
- Увеличение ARPU в расчете на одного продавца. **ARPU** — отношение суммарной выручки с получения комиссии от продавцов к общему количеству продавцов.

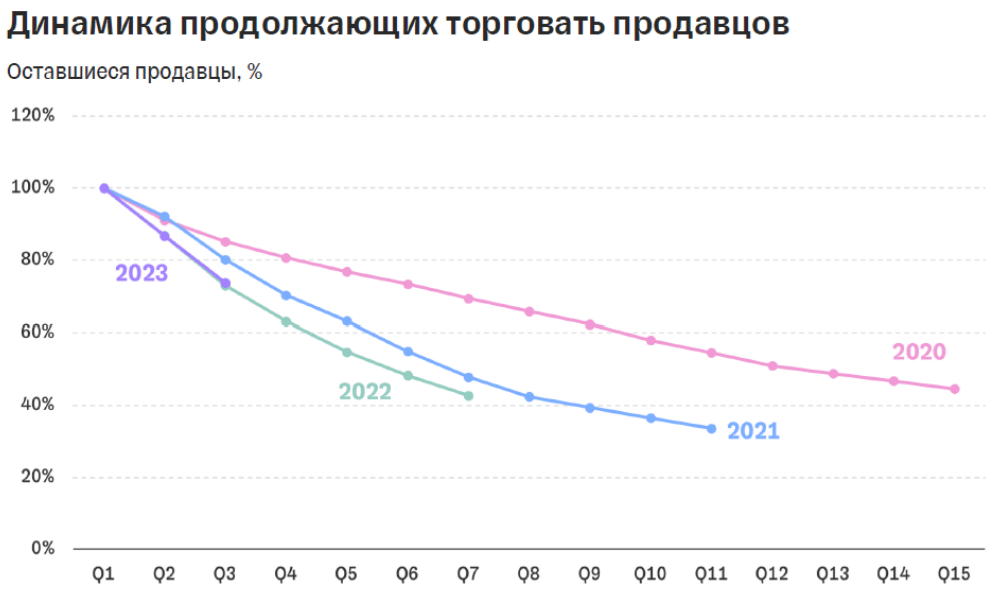

Рис. 2: Динамика продолжающих торговать продавцов

5. Бизнес модель

В результате серии разговоров с топ-менеджерами компании ProstoMarket была составлена следующая бизнес модель:

Основная аудитория

- Индивидуальные предприниматели и небольшие компании, у которых нет ресурсов для найма собственного маркетолога или дизайнера;
- Продавцы без опыта работы на маркетплейсах. Люди, которые только начинают продавать свои товары онлайн и нуждаются в помощи для создания качественных объявлений.

Модель монетизации

- **Ежемесячная подписка**. Основные функции сервиса будут доступны по подписке, которую пользователь сможет в любой момент отменить;
- Проведение **индивидуальных (разовых) услуг** за отдельную плату. К таким услугам может относиться проведение детального исследования рынка для отдельного клиента с целью улучшения работы сервера на конкретном рынке;
- **Процент от продаж**. Дополнительная комиссия за улучшение конверсии объявлений, которая взимается в виде процента от увеличенных продаж после применения рекомендаций сервиса.

Основные функции сервиса

- **Анализ текущих объявлений.** Автоматический анализ текста и изображения товара: оценка качества заголовков, описания товаров, ключевых слов и других элементов объявления;

- **Рекомендации по оптимизации.** Предложения по изменению текста, подсказки по улучшению формулировок. Помощь в подборе релевантных ключевых слов для увеличения видимости и привлечения целевой аудитории. Рекомендации по цветовой гамме, шрифтам, расположению элементов и другим аспектам дизайна, чтобы сделать объявление более привлекательным.

Ключевые преимущества сервиса

Интеграция на российский рынок. В сервис заранее будут заложены особенности российского рынка, которые являются его основным преимуществом по сравнению с зарубежными аналогами

Потенциальные риски

- Наличие потенциальной конкуренции с аналогичными продуктами;

- Необходимость постоянного улучшения имеющего функционала для удержания потенциальных клиентов;

- Необходимость учитывать тренды и сезонность в данных (новогодние распродажи, черная пятница);

- Влияние исторических событий на покупательную способность населения, а соответственно на потенциальное желание новых продавцов выходить на маркетплейс. Примеры таких событий: изменение ключевой ставки, санкции и другие.

6. Оценка экономики проекта

6.1. Затраты на проект

Реализация проекта планируется с 2024-2027 год. Первая полноценная, работающая версия продукта будет представлена в третьем квартале 2025 года, далее планируется поддержка и улучшение продукта для компании ProstoMarket. В таблице 5 представлены операционные расходы, необходимые для реализации проекта. В таблице 6 капитальные затраты. Все расходы были учтены с учетом инфляции в 8 %. Было учтено, что некоторые расходы понадобятся не сначала проекта.

Итоговые расходы на проект за 4 года составят 113037 тыс руб.

Категория	2024	2025	2026	2027
Зарплаты сотрудников	**20204**	**22508**	**25312**	**27843**
Руководитель проекта	3396	3735	4109	4520
ML инженер	4992	5491	6040	6644
Бизнес аналитик	2016	2218	2439	2683
Системный аналитик	2448	2693	2962	3258
Дизайнер	0	504	1109	1219
Маркетолог	1512	1663	1829	2012
Бэкенд разработчик	2880	3168	3484	3833
Фронтенд разработчик	2760	3036	3340	3674
Подрядчики	**3108**	**3356**	**3626**	**3780**
Интернет	108	116	126	136
Аренда офиса	3000	3240	3500	3780
Итоговая сумма	**23312**	**25864**	**28938**	**31623**

Таблица 5: Операционные расходы по МСФО (в тыс. руб.)

Категория	2024	2025	2026	2027
Техническое оборудование	**3300**	**0**	**0**	**0**
Ноутбук с видеокартой	500	0	0	0
Ноутбук без видеокарты	1050	0	0	0
Лицензия на ПО	750	0	0	0
Сервер с видеокартой	1000	0	0	0

Таблица 6: Капитальные расходы по МСФО (в тыс. руб.)

6.2. Окупаемость проекта

Для расчета окупаемости проекта были приняты во внимание следующие факты:

- Поставленная бизнес цель по уменьшению оттока продавцов будет выполнена к 2027 году;
- Соотношение продавцов по основным категориям товаров не изменится;
- Ожидаемое число продавцов к 2026 году будет 820 тыс., к 2027 году 900 тыс;
- Продавец продает хотя бы один товар в месяц;
- В качестве расчетных показателей будем использовать среднюю стоимость покупаемых товаров, представленную аналитиками компании ProstoMarket (таблица 7).

Категория	Стоимость
Красота и здоровье	2.5
Товары для дома	1
Одежда, обувь и аксессуары	5
Техника	50
Автомобили	3000

Таблица 7: Средняя стоимость покупаемых товаров по категориям в тыс. руб

В таблице 8 показано, что компания теряет в связи с оттоком 30 % продавцов по годам (показан не полный расчет).

Категория	2027	2028	2029	2030	2031
Красота и здоровье	103.32	113.4	129.16	145.7	168.91
Товары для дома	118.08	129.6	150.1	172.9	191.41
Одежда, обувь и аксессуары	464.94	510.3	590.4	710.86	921.34
Техника	1033.2	1134	1291.23	1470.74	1699.1
Автомобили	15940.8	17496	19700.23	22300.21	25300.12

Таблица 8: Денежные потери без применения модели по годам

В таблице 9 показано, что компания теряет в связи с оттоком 10 % продавцов по годам (показан не полный расчет).

Категория	2027	2028	2029	2030	2031
Красота и здоровье	34.44	37.8	41.9	46.4	56.1
Товары для дома	39.36	43.2	49.3	58.7	72.1
Одежда, обувь и аксессуары	154.98	170.1	182.3	198.3	213.5
Техника	344.4	378	412.3	456.1	510.3
Автомобили	5313.6	5832	6700.12	8100.743	9500.34

Таблица 9: Денежные потери с применением модели по годам

Расчеты показывают, что проект выйдет на окупаемость к 2031 году, то есть спустя 5 лет после своего полноценного старта. Это считается хорошим сроком окупаемости бизнеса с вложениями 113 млн рублей.

7. Анализ датасета

7.1. Информация о наборе данных

Источник датасета – соревнование Avito Demand Prediction Challenge.

Общий вес – 146.76 GB.

Табличная тренировочная часть содержит 1,503,424 записи по 18 столбцов:

- **item_id** - уникальный идентификатор объявления (тип: категориальный).
- **user_id** - уникальный идентификатор пользователя (тип: категориальный).
- **region** - регион, в котором размещено объявление (тип: категориальный).
- **city** - город, в котором размещено объявление (тип: категориальный).
- **parent_category_name** - главная категория объявления (тип: категориальный).
- **category_name** - подкатегория объявления (тип: категориальный).
- **param_1**, **param_2**, **param_3** - опциональные параметры описания объявления (тип: категориальный).
- **title** - название объявления (тип: категориальный).
- **description** - описание объявления (тип: категориальный).
- **price** - цена товара (тип: числовой).
- **item_seq_number** - последовательный номер объявления пользователя (тип: числовой).
- **activation_date** - дата размещения объявления (тип: категориальный).
- **user_type** - тип пользователя (тип: категориальный).
- **image** - идентификатор изображения, связанного с объявлением (тип: категориальный).
- **image_top_1** - код классификации изображения (тип: числовой).
- **deal_probability** - целевая переменная, вероятность того, что товар будет продан (тип: числовой).

Набор данных также содержит архив с изображениями, которые связанны с некоторыми объявлениями через уникальные идентификаторы.

Количество уникальных значений признаков: На графике представлено количество уникальных значений для каждого признака на логарифмической шкале, что позволяет сравнивать их разнообразие. Признаки, такие как item_id и user_id, имеют наибольшее число уникальных значений — каждый объект и пользователь представлены индивидуальными идентификаторами. Признаки title и description также обладают

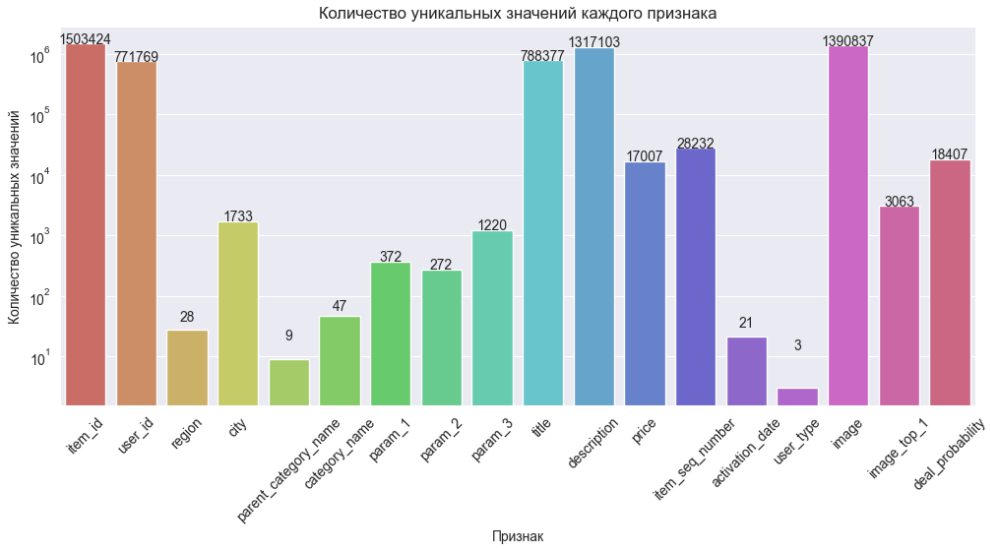

Рис. 3: Количество уникальных значений каждого признака

высокой вариативностью, что говорит о значительном разнообразии текстового описания объявлений. В то же время признаки region, parent_category_name и user_type имеют небольшое количество уникальных значений (от 3 до 28), что указывает на их категориальный характер с ограниченным числом возможных вариантов. Таким образом, признаки в датасете можно разделить на две группы: с высокой вариативностью (для текстов и идентификаторов) и низкой (для категориальных данных), что может влиять на выбор методов предобработки данных при обучении модели.

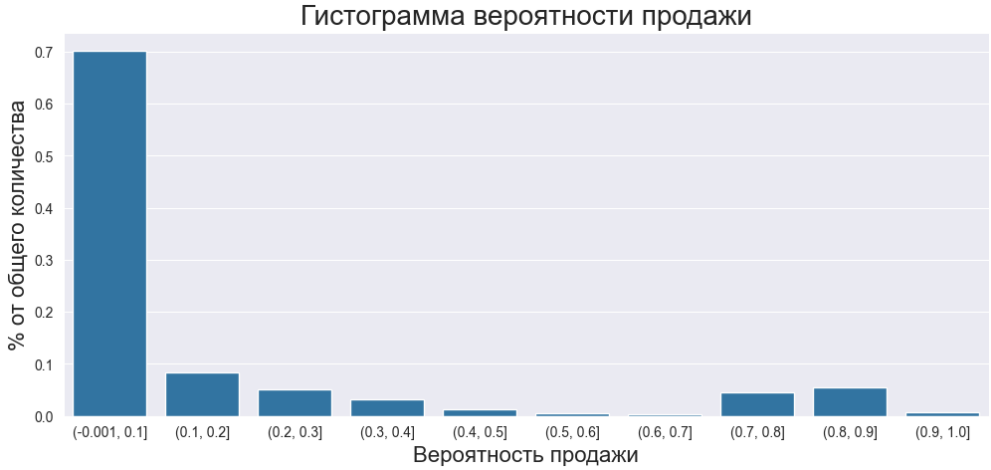

Рис. 4: Гистограмма вероятности продаж

Гистограмма вероятности продажи: Распределение вероятностей продажи сильно смещено влево: большая часть объявлений имеет вероятность продажи менее 0.1. Это указывает на то, что значительная доля объявлений имеет низкий шанс на успешное завершение сделки.

Распределение типов пользователей: Большая часть пользователей (более половины) — частные лица, на втором месте компании, а магазины составляют минималь-

Рис. 5: Распределение типов пользователей

ную долю. Это распределение указывает на доминирование объявлений от от индивидуальных пользователей, что может сказываться на их характере и качестве.

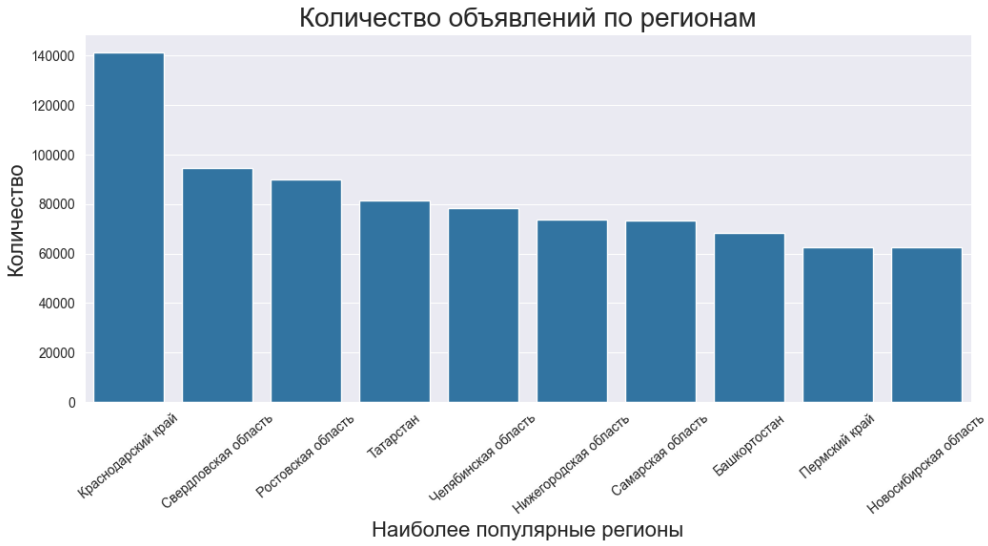

Рис. 6: Количество объявлений по регионам

Количество объявлений по регионам: На графике представлено распределение количества объявлений по регионам, где заметно выделяется Краснодарский край с явным отрывом от остальных. Для других регионов наблюдается с постепенное снижение количества объявлений. Высокая концентрация активности в одном из регионом может быть связана с экономической активностью, численностью населения или уровнем развития цифровой инфраструктуры.

Количество объявлений по городам: Распределение по городам демонстрирует, что Краснодар и Екатеринбург лидируют по количеству объявлений, но в целом

Рис. 7: Количество объявлений по городам

разница между топовыми городами относительно невелика. Это говорит о более равномерном распределении активности среди крупных городов, что может указывать на схожий уровень спроса.

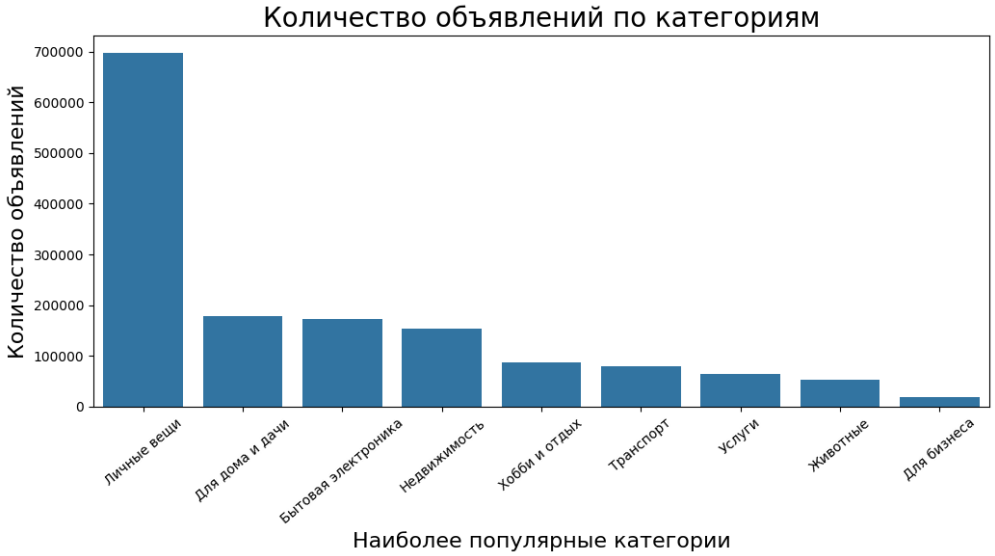

Рис. 8: Количество объявлений по категориям

Количество объявлений по категориям: График показывает значительное смещение распределения количества объявлений в сторону категории "Личные вещи которая содержит почти 700 тысяч объявлений. Это в несколько раз превышает количество объявлений в других категориях, таких как "Для дома и дачи "Бытовая электроника"и "Недвижимость которые занимают места со второго по четвертое с примерно равным количеством записей. В остальных категориях наблюдается резкое снижение числа объявлений, а категории товары "Для бизнеса"представлены минимальным числом записей. Такое распределение указывает на неравномерную активность пользователей по разным категориям, с высокой концентрацией в нескольких ключевых сегментах.

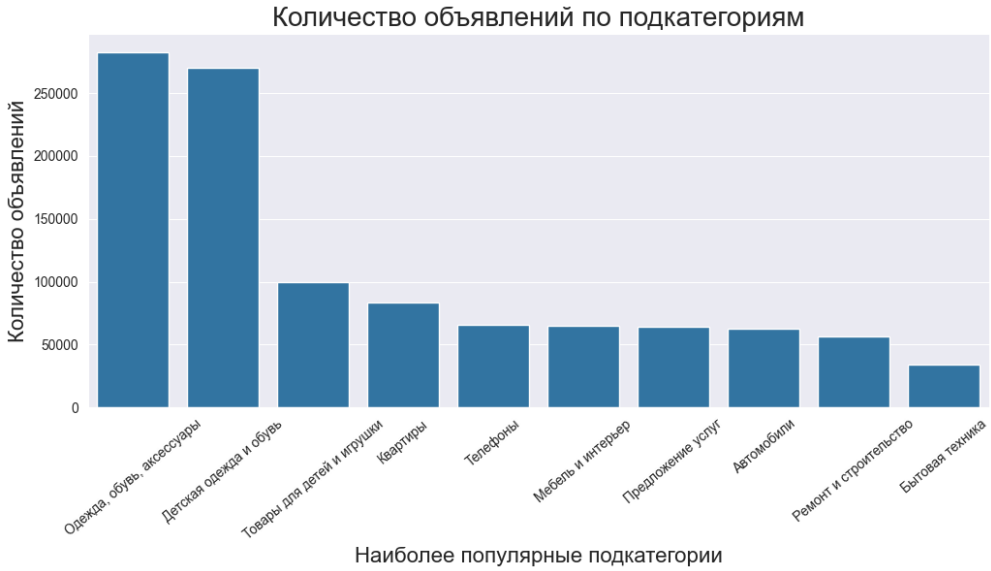

Рис. 9: Количество объявлений по подкатегориям

Количество объявлений по подкатегориям: Распределение количества объявлений неравномерное: две подкатегории "Одежда, обувь, аксессуары"и "Детская одежда и обувь"значительно выделяются, с количеством объявлений более чем в 2.5 раза превышающим остальные. Такое сильное смещение указывает на популярность определённых сегментов и потенциально высокий уровень конкуренции в них.

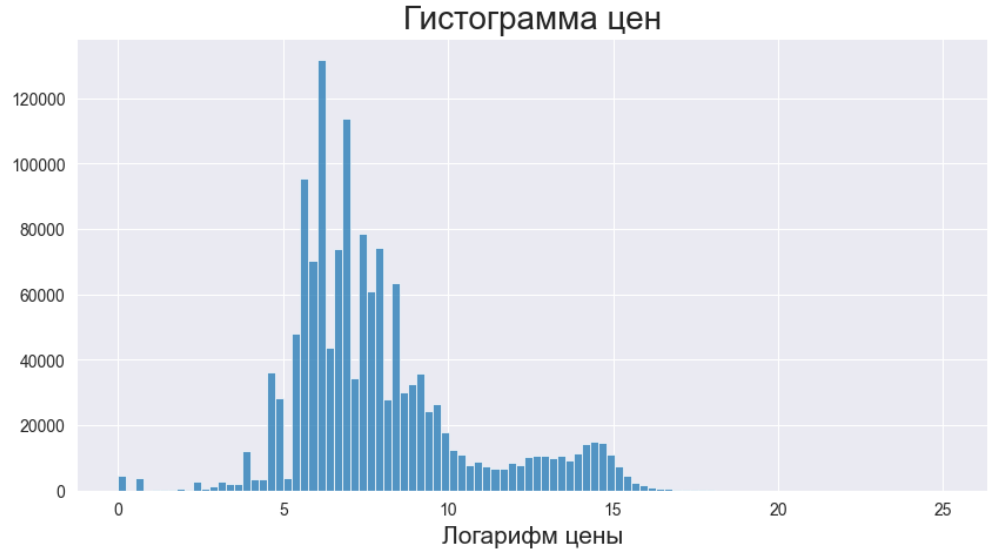

Рис. 10: Гистограмма цен

Гистограмма цен: Логарифмическое распределение цен демонстрирует, что большинство объявлений сосредоточены в диапазоне логарифмических значений 5–8. Это соответствует низким ценам товаров (примерно от 100 до 3000 рублей), что логично для товаров, ориентированных на массового потребителя. На графике также заметна правосторонняя асимметрия — есть длинный хвост, что указывает на небольшое количество объявлений с очень высокими ценами, выходящими далеко за рамки среднего

диапазона. Наличие пиков на определённых значениях может указывать на влияние округления цен, а также на тенденцию пользователей придерживаться общепринятых ценовых точек. Данное распределение подчёркивает, что для большинства товаров ожидается низкая стоимость, а дорогие предложения встречаются значительно реже.

Рис. 11: Вероятности продаж по категориям

Вероятности продаж по категориям: Вероятность продажи значительно варьируется в зависимости от категории товаров. Категории «Услуги», «Животные» и «Транспорт» демонстрируют более высокие медианные значения вероятности продажи, в то время как «Личные вещи» и «Для бизнеса» имеют низкие медианные значения, что говорит о более низкой успешности сделок в этих категориях. Также наблюдаются выбросы вероятностей в некоторых категориях, таких как «Для дома и дачи», «Бытовая электроника» и «Недвижимость», где отдельные объявления достигают высоких вероятностей продажи, несмотря на низкую медиану. Это говорит о том, что в данных присутствуют редкие успешные объявления.

Цены по категориям: Распределение цен сильно варьируется по категориям. Наибольшие медианы наблюдаются у категорий "Транспорт"и "Недвижимость что логично для дорогостоящих товаров. В других категориях, например, "Личные вещи"и "Услуги цены сосредоточены вокруг низких значений с меньшими межквартильными размахами.

7.2. Качество данных

В наборе данных имеются пропуски в нескольких столбцах:

- **param_3** - 57.37% пропусков.
- **param_2** - 43.54% пропусков.

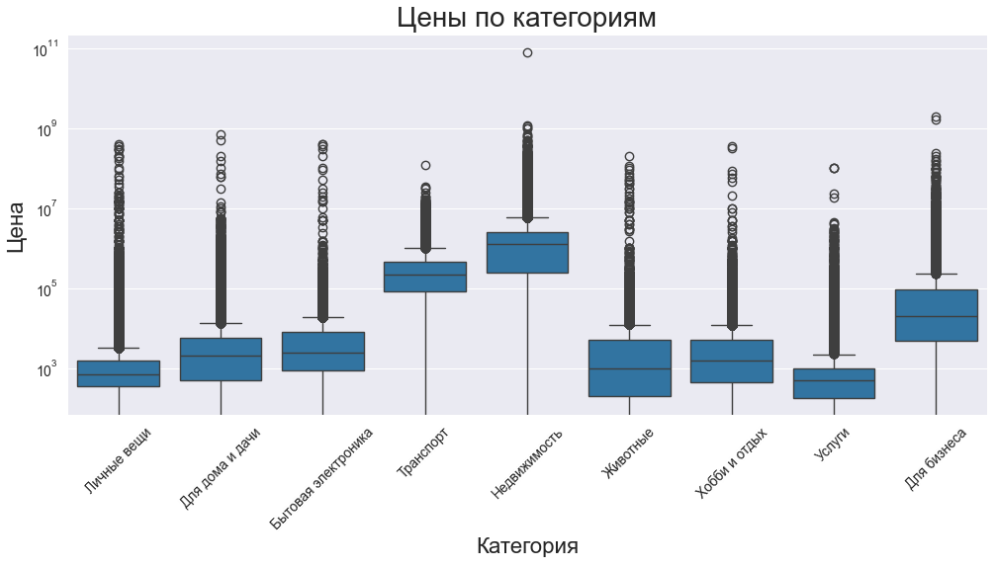

Рис. 12: Цены по категориям

- **description** - 7.73% пропусков.
- **image** - 7.49% пропусков.
- **image_top_1** - 7.49% пропусков.
- **price** - 5.68% пропусков.
- **param_1** - 4.10% пропусков.

7.3. Обогащение данных

Процесс обогащения данных был ключевым этапом в проекте, поскольку исходные признаки содержали разнородные данные (тексты, изображения, категориальные переменные) и не предоставляли достаточно информации для построения высококачественной модели. Этот процесс включал добавление следующих данных:

- **Текстовые статистики для заголовка и описания объявления:**
  - длина текста в символах — отражает объём представленной информации;
  - количество слов — позволяет оценить полноту текста;
  - количество уникальных слов — помогает выявить разнообразие лексики;
  - пересечение уникальных слов — измеряет степень повторяемости слов между заголовком и описанием;
  - разность уникальных слов — показывает, насколько заголовок и описание дополняют друг друга.
- **Статистики изображений:**

  - среднее значение RGB-каналов — характеризует общий цветовой баланс изображения;
  - белизна (доля светлых пикселей) — определяет визуальную яркость изображения;
  - тусклость (обратная насыщенность цветов) — помогает оценить насыщенность цветов;
  - яркость — показывает освещённость изображения;
  - контрастность — измеряет разницу между светлыми и тёмными областями изображения.

- **Косинусная близость между заголовком и описанием,** рассчитанная на основе эмбеддингов, полученных с использованием модели RuBERT. Этот показатель помогает выявить, насколько тесно связаны тексты заголовка и описания.

- **Текстовые эмбеддинги для заголовка и описания объявления:** используются для представления текста в числовом формате, что позволяет моделям лучше учитывать смысловую информацию.

- **Картиночные эмбеддинги:** извлекаются, чтобы представить изображения как векторные признаки и учитывать их визуальные особенности.

Каждое из добавленных обогащений данных было выбрано на основании проведённых экспериментов. Текстовые статистики помогли оценить структуру текста и дали модели возможность учитывать, насколько заголовок и описание являются дополняющими друг друга. Визуальные характеристики изображений помогли модели концентрироваться на информации, играющей ключевую роль в восприятии объявления. Косинусная близость заголовка и описания дала возможность количественно оценить содержательную взаимосвязь, а текстовые и картиночные эмбеддинги предоставили модели дополнительные уровни информации о содержимом текста и изображений, что улучшило способность различать визуально привлекательные объявления и те, которые могли бы быть менее интересны покупателям. Добавленные признаки значительно улучшили качество предсказаний и позволили добиться высокой точности в решении задачи.

В рамках экспериментов были протестированы различные подходы к созданию эмбеддингов:

- **Текстовые эмбеддинги:**

  - **TF-IDF:** в исходной реализации векторы имели размерность 20,000 признаков, что создавало сложности для моделей. После снижения размерности до

300 признаков с помощью SVD удалось добиться баланса между информативностью и эффективностью;

- **FastText:** словные эмбеддинги позволили учитывать семантические связи между словами, что улучшило представление текстов;
- **RuBERT:** контекстные эмбеддинги RuBERT (адаптация BERT для русского языка) обеспечили хорошее представление текстов за счёт учёта грамматических и семантических связей.

- **Картиночные эмбеддинги:**
  - **ResNet:** глубокая свёрточная нейронная сеть, последний полносвязный слой которой заменялся на Average Pooling;
  - **CLIP:** универсальная модель, проецирующая текстовые и визуальные эмбеддинги в общее семантическое пространство;
  - **DINO:** современный метод самообучения, основанный на трансформерах, позволяющий извлекать высокоуровневые семантические признаки из изображений.

8. Предобработка данных

8.1. Методы предобработки данных

Предобработка данных была необходимым этапом перед обучением моделей, так как данные содержали пропуски, категориальные признаки и текстовую информацию, требующую особой обработки. Основные шаги предобработки включали:

- **Заполнение пропусков**;
- **Кодирование категориальных признаков**;
- **Удаление неинформативных признаков**.

Пропущенные значения обрабатывались по-разному в зависимости от типа признаков. Для непрерывных признаков пропуски заменялись медианой соответствующего столбца, поскольку медиана не чувствительна к выбросам. Пропуски в категориальных признаках заменялись новой отдельной категорией – пустой строкой.

Категориальные признаки обрабатывались в зависимости от типа тестируемой модели. Если рассматриваемая модель предполагала наличие под капотом собственных алгоритмов предобработки категориальных признаков, то предпочтение отдавалось им, так как они более умные: анализируют распределение значений и уже на основе этого

распределения применяют подходящий метод кодирования. Для других моделей, которые не умеют в самостоятельную обработку категориальных признаков, применялся Count Encoding, то есть каждая категория заменялась на частоту её появления в данных. Также были протестированы One-Hot Encoding и Target Encoding. Первый подход сильно увеличивал размерность признакового пространства и замедлял обучение моделей. Второй показал результаты чуть хуже, чем Count Encoding.

Также стоит отметить, что для линейных моделей происходила обязательная стандартизация данных, то есть приведение данных к единому масштабу. Невыполнение этого важного шага сильно подрывает способность линейной модели к сходимости, а иногда и вовсе делает этот процесс невозможным.

Некоторые признаки, такие как item_id, user_id, activation_date, title, description были удалены после извлечения из них всех полезных данных, так как они не содержали прямой информации для предсказания вероятности продажи.

8.2. Подготовка данных для обучения

Данные трансформировались в формат, подходящий для каждой модели, с которой проводились эксперименты.

9. ML-решение

9.1. Выбор и разработка модели

В рамках решаемой задачи были рассмотрены следующие подходы:

- **Ridge-регрессия**;
- **LightGBM**;
- **CatBoost**.

Для достижения консистентности экспериментов все модели обучались на одинаково предобработанных данных. Бустинговые модели запускались с равным числом деревьев.

Первая модель, Ridge-регрессия, была выбрана в качестве бейзлайна, от которого можно отталкиваться. Это линейная модель с регуляризацией, обладающая рядом преимуществ, таких как простота реализации, высокая скорость обучения и интерпретируемость полученных результатов. Основным недостатком стала ограниченность модели в виде неспособности учитывать сложные нелинейные зависимости между признаками и целевой переменной. Кроме того, Ridge-регрессия, как и любая линейная модель,

требует стандартизации исходных данных и предварительного кодирования категориальных признаков. В тестировании Ridge-регрессия продемонстрировала стабильные, но относительно слабые результаты, что было ожидаемо для модели такого типа.

Следующим шагом был переход к более продвинутым алгоритмам, а именно бустингам. На данный момент бустинговые модели по-прежнему остаются лучшими для работы с табличными данными. Всё благодаря их производительности, проверенной временем эффективности, интерпретируемости и способности к работе с разнородными данными. Здесь хотелось провести эксперименты с двумя популярными моделями – LightGBM и CatBoost. Первая модель быстро учится даже при огромном числе наблюдений и большой размерности признакового пространства, как было в нашем случае, а благодаря ансамблированию качество значительно превосходит линейную модель.

Финальной моделью для тестирования стал CatBoost, который был выбран за счёт своей способности эффективно кодировать категориальные признаки, гибких настроек, встроенных механизмов борьбы с переобучением и высоким качеством, демонстрируемым во многих задачах.

В результате оценки качества модели на кросс-валидации CatBoost продемонстрировал лучшие результаты по ключевым метрикам качества, а также неплохую производительность, так что дальнейшие эксперименты сводились в попытки улучшения данной модели посредством исследования новых способов предобработки исходных данных, а также создания потенциально важных признаков.

9.2. Настройка и обучение модели

Обучение модели производилось на NVIDIA GeForce RTX 3080 16Gb, 32Gb оперативной памяти. Использование этих мощностей позволило обучать модель на большом размере датасета.

9.3. Система учета экспериментов

Для выбора модели стояла задача проведения множества экспериментов. Для удобства мониторинга каждого такого эксперимента использовалась платформа ClearML. Логирование экспериментов включало сохранение параметров моделей, метрик качества, артефактов (таких как графики важности признаков и файлы с обученными моделями), а также промежуточных результатов обучения. Это позволило отслеживать прогресс обучения, сравнивать различные итерации моделей и легко воспроизводить эксперименты с заданными параметрами.

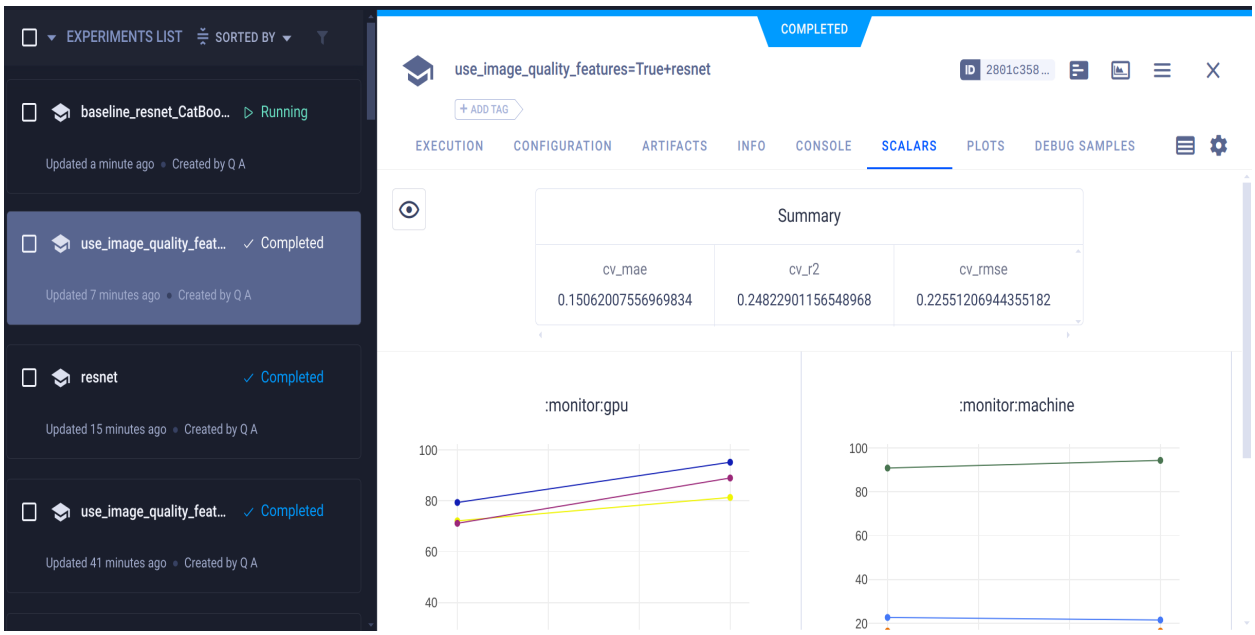

Рис. 13: Пример экрана эксперимента в ClearML

9.4. Сравнение моделей

Общая производительность моделей. Результаты экспериментов представлены в таблице 11.

Модель	RMSE	MAE	R2
Ridge-регрессия	0.256	0.175	0.032
LightGBM	0.227	0.152	0.239
CatBoost	**0.225**	**0.150**	**0.250**

Таблица 10: Сравнение моделей по ключевым метрикам.

Как видно, CatBoost показал наилучшие результаты по всем метрикам. Это свидетельствует о способности данной модели лучше улавливать сложные зависимости в данных и справляться с поставленной задачей предсказания.

Влияние текстовых признаков. Для изучения влияния текстовых признаков были проведены эксперименты с разными подходами обработки текста: базовая модель (без текстовых признаков), TF-IDF, RuBERT и FastText. Эксперименты состояли в использовании эмбеддингов, полученных:

- **только для заголовков (title)**;
- **только для описаний (description)**;
- **одновременно для заголовков и описаний (title и description)**.

Результаты тестирования с использованием текстовых эмбеддингов представлены в таблице 12.

Модель	Признаки	RMSE	MAE	R2
Baseline	–	0.225	0.150	0.250
Baseline + TF-IDF	title	**0.224**	**0.149**	**0.260**
	description	**0.224**	**0.149**	**0.261**
	title + description	**0.223**	**0.149**	**0.263**
Baseline + RuBERT	title	0.227	0.152	0.240
	description	0.227	0.151	0.240
	title + description	0.226	0.150	0.241
Baseline + FastText	title	0.224	0.149	0.260
	description	0.225	0.150	0.253
	title + description	0.224	0.150	0.258

Таблица 11: Результаты тестирования с использованием текстовых эмбеддингов.

Влияние признаков изображений. Для оценки влияния визуальных признаков на качество предсказаний были проведены эксперименты с различными комбинациями признаков. Результаты тестирования:

Модель	RMSE	MAE	R2
Baseline	0.225	0.150	0.249
Baseline + Image Statistics	0.225	0.150	0.252
Baseline + ResNet	0.226	0.151	0.244
Baseline + Image Statistics + ResNet	0.226	0.151	0.248
Baseline + Image Statistics + CLIP	**0.223**	**0.149**	**0.261**
Baseline + Image Statistics + DinoV2	0.224	0.149	0.257

Таблица 12: Сравнение моделей по использованию визуальных признаков.

9.5. Результаты и интерпретация

Использование текстовых эмбеддингов для заголовка и описания объявления действительно улучшает качество модели. Лучшими метриками выделился TF-IDF для описаний и заголовков вместе. FastText также продемонстрировал хорошие результаты. Особенно улучшение качества наблюдалось при включении в модель эмбеддингов для описания объявления. Использование эмбеддингов, полученных с помощью RuBERT, привело к ухудшению качества.

Гипотезы по улучшению качества при внедрении в модель визуальных статистик и картиночных эмбеддингов также подтвердились. Визуальные статистики немного улучшили объясняющую способность модели. Модель CLIP, сочетающая текстовые и визуальные эмбеддинги в общем семантическом пространстве, показала наилучшие результаты среди всех экспериментов.

Анализ результатов экспериментов позволил выделить ключевые факторы, влияющие на качество прогнозирования, и сформировать оптимальную модель, которая

эффективно использует комбинацию текстовых и визуальных признаков. Было принято решение остановиться на подходе, сочетающем текстовые и визуальные статистики, косинусную близость, а также TF-IDF- и CLIP-эмбеддинги.

9.6. Подбор гиперпараметров

Для определения оптимальных гиперпараметров для данной задачи мы воспользовались фреймворком Optuna. В результате процесса оптимизации были получены следующие значения метрик.

Модель	**RMSE**	**MAE**	**R2**
Baseline	0.225	0.150	0.250
Optimized	**0.221**	**0.147**	**0.276**

Таблица 13: Сравнение результатов бейзлайна с оптимальными параметрами.

Найденные оптимальные параметры:

- **iterations** = 1773;
- **depth** = 8;
- **learning_rate** = 0.104;
- **l2_leaf_reg** = 5.33.

9.7. Важность признаков

Анализ важности признаков выявил следующие ключевые переменные: param_1, price, image_top_1, param_2, parent_category_name. Также важно отметить, что большинство признаков, полученных с изображений и текста, полезны для предсказания успешности сделки.

10. Валидация и тестирование

10.1. Методы валидации

Обучение каждой модели проводилось на тренировочном наборе данных с использованием кросс-валидации для оценки качества. Для кросс-валидации использовались методы `KFold` и `StratifiedKFold`. В случае `StratifiedKFold` целевая переменная `y` предварительно масштабировалась и округлялась до двух знаков после запятой, что позволяло добиться сбалансированного разбиения данных.

Таким образом, кросс-валидация позволила оценить устойчивость модели к различным разбиениям и минимизировать риск переобучения.

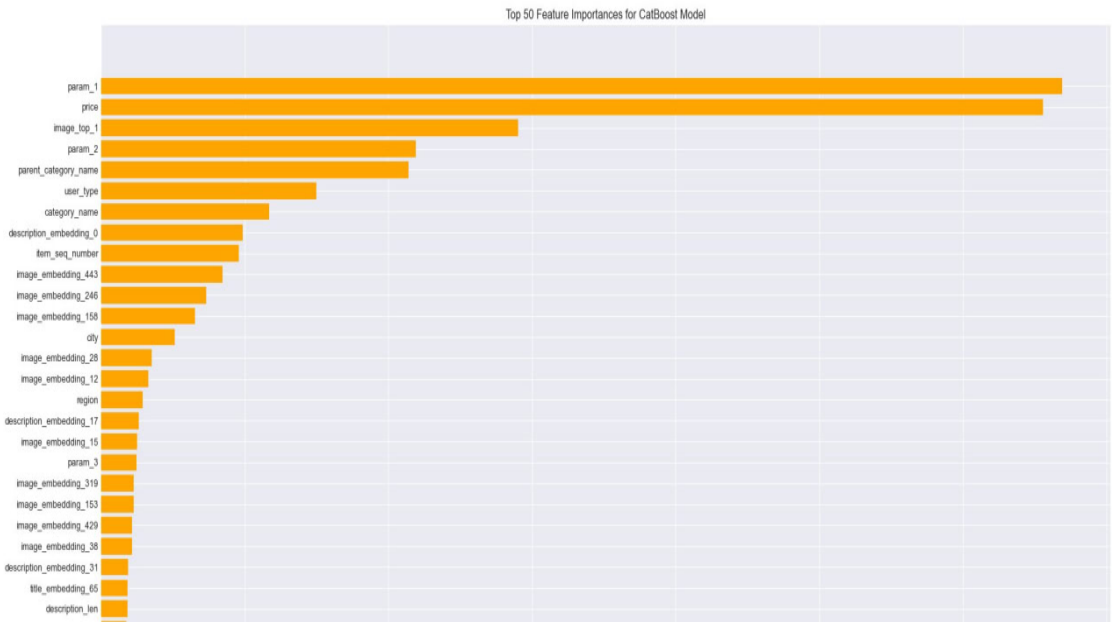

Рис. 14: Важность признаков

10.2. Выбор метрик качества

Сравнительный анализ проводился с использованием следующих метрик:

- **RMSE (Root Mean Square Error)** – основной показатель для регрессионных задач, отражающий среднеквадратичную ошибку модели;
- **MAE (Mean Absolute Error)** – среднее абсолютное отклонение предсказаний от истинных значений;
- **R2 (Coefficient of Determination)** – коэффициент детерминации, показывающий долю объяснённой вариации в данных.

Особое внимание уделялось анализу распределения ошибок по фолдам. Это дало возможность понять не только среднее качество модели, но и её поведение в худших сценариях.

11. Заключение

Проект имеет финансовый потенциал и высокую возвратность инвестиций, поэтому руководство компании ProstoMarket приняло решение инвестировать в разработку проекта.

12. Приложение

1. Github репозиторий: https://github.com/soundwave-77/ML-project

2. Демо модели: https://avito-rate.streamlit.app/

3. Дашборд в datalens: https://datalens.yandex.cloud/8u6m2xcd1ue6v