
Вы – менеджер компании Olist Store, которая всегда была на острие передовых технологий в области e-com торговли. Но вы замечаете, что отзывы ваших клиентов на платформе -хуже чем у клиентов, а товары часто доставляются с опозданием

Ваша цель – взять реванш и доказать, что ваша компания не зря славится своими инновационными решениями. Вы решаете провести масштабное исследование исторических данных о продажах, товарах, селлерах и клиентах, чтобы предоставить вашим клиентам лучший сервис. Вы докажете, что вы главные в E-com

В невероятной атмосфере высоких ставок и огромной ответственности за результат вам придется:

1. Провести анализ данных (EDA), чтобы разобраться, какие факторы влияют на калорийный расход.
2. Разработать машинную модель, которую трижды улучшить, доказывая, что вы движетесь в правильном направлении.
3. Создать интерактивный дашборд в DataLens для наглядной демонстрации операционных метрик руководству, здесь не нужно использовать результаты ML модели. Графики могут дублироваться с частью EDA, но давать оперативную информацию для принятия управленческих решений

Вы можете выбрать одну из проблем и построить ее решение, другой займется ваш коллега, не тратьте время на решение обоих.

Структура данных

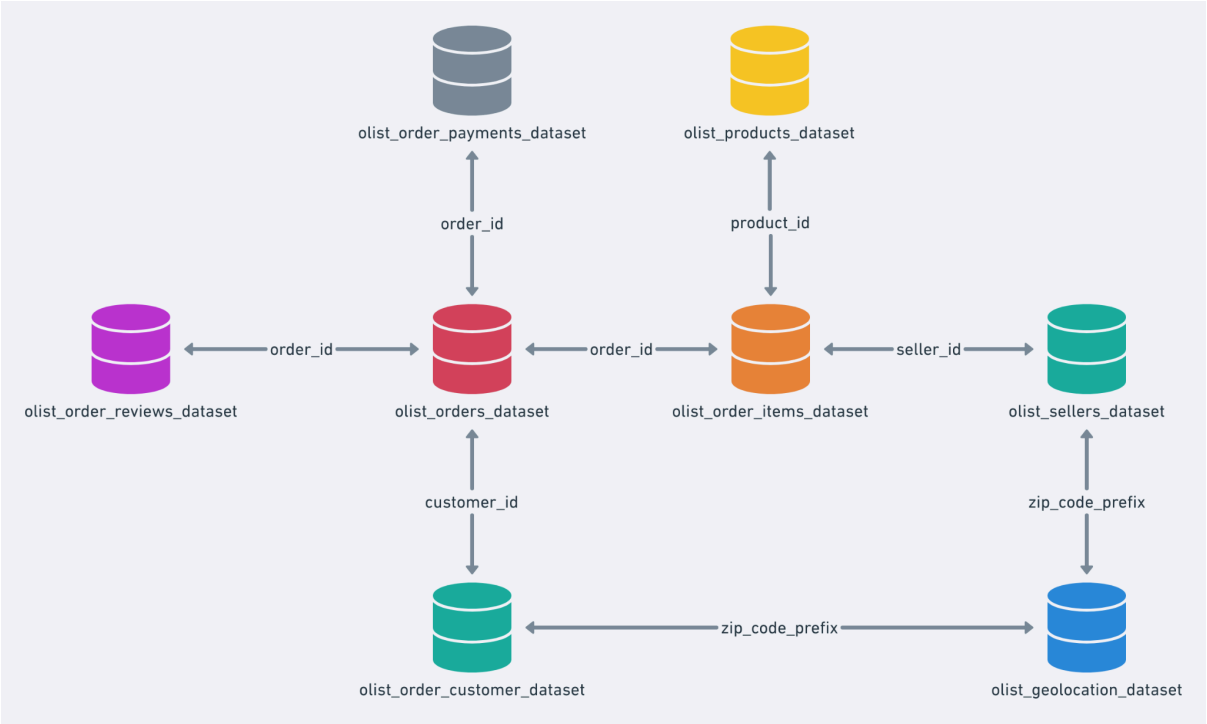
Данные для задания состоят из нескольких таблиц, из которых предстоит подготовить фичи для анализа (признаки по селлерам, товарам, особенностям оплаты).

Таблицы придется объединить по соответствующем id в соответствии с вашим представлением о необходимом результате

Данные доступны по ссылке
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data?select=olist_orders_dataset.csv

Таблицы:

1. olist_orders описывает заказы товаров
2. olist_order_reviews описывает отзывы на заказы
3. olist_order_payments описывает способы оплаты и суммы в заказах
4. olist_order_items описывает состав заказов, кто продал, нужен для связи таблиц воедино
5. olist_sellers описывает признаки селлеров
6. olist_customer описывает признаки покупателям
7. olist_order_customer для связки покупателя и заказа
8. olist_products описывает товары
9. olist_geolocation содержит данные о локациях по зип_кодам доставки



Техническое задание

Вариант 1

Часть 1. Exploratory Data Analysis (EDA)

Цель исследования

Выявить особенности заказов, на которые пользователи ставят низкие оценки в отзывах (review_score 1-2 звезды).

Задачи

1. **Проверка данных**:
    - Определите наличие пропущенных значений, выбросов и аномалий.
    - Примите решение, как корректировать или учитывать эти пропуски и аномалии при дальнейшем анализе.

2. **Анализ распределений**:

    - Постройте графики распределения для числовых переменных
    - Изучите распределение по категориальным переменным.

3. **Корреляционный анализ**:
   Исследуйте взаимосвязи между признаками.
    - Используйте карту корреляций или аналогичный инструмент для визуализации.

4. **Гипотезы и инсайты**:
    - Сформулируйте предположения, какие характеристики присущи заказам с низкими отзывами.
    - По возможности опишите сценарии оптимизации для сокращения количества негативных отзывов на платформе.

Ожидаемый результат

- Читаемые и понятные графики, таблицы и статистика.
- Аналитический отчёт с основными выводами, гипотезами и рекомендациями для руководства, оформленный так, чтобы было ясно, что вы провели глубокий анализ и готовы к дальнейшим шагам.
- Должен содержать минимум 5 значимых для бизнеса вывода

Наводящие бизнес-вопросы

- Какие факторы (скорость доставки, география, категория товара, цена, способ оплаты и т.п.) чаще всего связаны с оценками 1–2?
- Чем «хорошие» заказы (4–5 звёзд) отличаются от «плохих» (1–2)?

- Для каких сегментов (штаты / категории / продавцы) проблема особенно остра?
- Какой профиль «рискованного» заказа, который с высокой вероятностью получит низкую оценку?

Часть 2. Моделирование и Машинное Обучение

Задача

Создать инструмент для выявления заказов, по которым пользователи потенциально могут поставить низкие оценки (1-2) на раннем этапе

Обязательные требования

- Ясное описание выбранной метрики качества (бизнес и ML метрикой) и почему с точки зрения бизнеса это важно.
- Документирование всех этапов с пояснениями – от идей до финальных результатов.
- Код в Jupyter Notebook или аналогичном формате с комментариями / проект в Orange с описаниями, куда может заглянуть любой коллега и понять логику.
- Опишите в отчете, на какие метрики после запуска модели вы будете смотреть, контролируя, что модель функционирует нормально

Часть 3. Создание Интерактивного Дашборда

Задача

Создать интерактивный дашборд для оперативного контроля ситуации руководством. Можете здесь вывести те метрики из EDA, которые посчитаете необходимым мониторить наравне с целевой метрикой % / количества заказов с низкой оценкой

Требования к дашборду

1. **Выведите ключевые метрики и их динамику**

2. **Фильтры для сегментации данных**:.

3. **Визуализации**:

- Графики, диаграммы, таблицы – всё должно быть представлено наглядно и доступно даже для пользователей без технического бэкграунда.

Ожидаемый результат

- Полноценный дашборд в Yandex DataLens.

Вариант 2

Часть 1. Exploratory Data Analysis (EDA)

Цель исследования
Выявить особенности заказов, которые доезжают с задержкой

Задачи

1. **Проверка данных**:
   - Определите наличие пропущенных значений, выбросов и аномалий.
   - Примите решение, как корректировать или учитывать эти пропуски и аномалии при дальнейшем анализе.

2. **Анализ распределений**:

   - Постройте графики распределения для числовых переменных
   - Изучите распределение по категориальным переменным.

3. **Корреляционный анализ**:
   Исследуйте взаимосвязи между признаками.
   - Используйте карту корреляций или аналогичный инструмент для визуализации.

5. **Гипотезы и инсайты**:
   - Сформулируйте предположения, какие характеристики присущи заказам с задержкой.
   - По возможности опишите сценарии оптимизации для сокращения количества заказов с задержкой.

Ожидаемый результат

- Читаемые и понятные графики, таблицы и статистика.
- Аналитический отчёт с основными выводами, гипотезами и рекомендациями для руководства, оформленный так, чтобы было ясно, что вы провели глубокий анализ и готовы к дальнейшим шагам.
- Должен содержать минимум 5 значимых для бизнеса вывода

Наводящие бизнес-вопросы

- В каких связках «продавец → регион клиента» чаще всего бывают задержки?
- Как на задержку влияют типы товаров, вес/размер, стоимость, время покупки?
- Есть ли продавцы или регионы, которые системно нарушают сроки?
- Можно ли заранее предсказать, что заказ будет доставлен позже обещанного срока?

Часть 2. Моделирование и Машинное Обучение

Задача

Разработать модель, позволяющую предсказать, что на этот товар потенциально может доехать с задержкой, чтобы команда могла принять меры по ее предотвращению / скорректировать срок доставки

Обязательные требования

- Ясное описание выбранной метрики качества (бизнес и ML метрикой) и почему с точки зрения бизнеса это важно.
- Документирование всех этапов с пояснениями – от идей до финальных результатов.
- Код в Jupyter Notebook или аналогичном формате с комментариями / проект в Orange с описаниями, куда может заглянуть любой коллега и понять логику.
- Опишите в отчете, на какие метрики после запуска модели вы будете смотреть, контролируя, что модель функционирует нормально

Часть 3. Создание Интерактивного Дашборда

Задача
Создать интерактивный дашборд для оперативного контроля ситуации руководством. Можете здесь вывести те метрики из EDA, которые посчитаете необходимым мониторить наравне с целевой метрикой % / количества заказов с задержкой

Требования к дашборду

1. **Выведите ключевые метрики и их динамику**

2. **Фильтры для сегментации данных**:.

3. **Визуализации**:

    - Графики, диаграммы, таблицы – всё должно быть представлено наглядно и доступно даже для пользователей без технического бэкграунда.

Ожидаемый результат

- Полноценный дашборд в Yandex DataLens.

Формат сдачи и дополнительные указания

1. **Единый отчёт**

    - Все этапы проекта (EDA, модель, дашборд) описываются последовательно в одном документе или ноутбуке.
    - Должна быть логика повествования: от постановки задачи до итогового представления решения.

2. **Презентация и короткое видео**

    - На базе итогового отчёта подготовьте презентацию.
    - Запишите видео длительностью не более 10 минут, чтобы кратко показать основные инсайты анализа, дашборд, объяснить основные шаги проведенного анализа

- Указать обязанности каждого члена команды

3. **Документирование**

   - Каждый график должен иметь подпись, поясняющую, что на нём изображено и какие выводы можно сделать.
   - Код должен сопровождаться комментариями и объяснениями ключевых шагов.
   - Если используете что-то сверх пройденного курса (дополнительные библиотеки, математические приёмы и т. д.), дайте ссылки или краткое описание.

4. **Из чего складывается оценка**:
   - **40%:**: отчет (презентация с EDA, описанием шагов, расчетные файлы)
   - **60%:**: защита проекта (отчетное видео с презентацией и дашбордом)

5. **Критерии оценивания**

   - **До 3 баллов**: за качество EDA (глубина анализа, визуализации, проверка гипотез).
   - **До 3 баллов**: за ML-пайплайн (масштабирование, feature engineering, подбор гиперпараметров) и демонстрацию улучшения метрик (3 итерации).
   - **До 2 баллов**: за интерактивный дашборд (доступность, структурированность).
   - **До 2 баллов**: за потенциальную пользу для бизнеса, применимость предложений по улучшению
   - **Итого максимум 10 баллов**.

Удачи!