

Кофе, цифры и прогноз

Домашнее задание по линейной регрессии в Orange

ВРЕМЯ 70–90 минут	ИНСТРУМЕНТ Orange Data Mining	ФИНАЛ Модель + бизнес-вывод
----------------------	----------------------------------	--------------------------------

ВАШЕ ЗАДАНИЕ

Вы — аналитик сети городских кофе-точек. Команда выбирает режим работы новой точки и хочет заранее оценивать дневную выручку. Ваша цель — исследовать данные, построить линейную регрессию и объяснить результат руководителю без магии и без языка «модель сказала».

Бизнес-вопрос

Можно ли достаточно точно прогнозировать **дневную выручку кофе-точки** по потоку людей, загрузке офисов, доставке, погоде и рекламному бюджету — и какие факторы дают самый заметный вклад?

Что должно получиться

- аккуратный Orange-workflow без лишних переменных;
- короткий разведывательный анализ данных (EDA);
- модель Linear Regression и первое знакомство с метриками качества;
- понятные выводы: какие факторы связаны с выручкой, как читать коэффициенты и где модели нельзя слепо доверять;
- короткая рефлексия о том, как вы использовали искусственный интеллект.

ВАЖНО

Данные синтетические, но устроены как реальная рабочая таблица. У них есть известная скрытая закономерность. Ваша задача — не угадать формулу, а пройти честный путь аналитика.

Что лежит в CSV

Одна строка — один день работы кофе-точки. Цель прогноза — revenue_ils.

Поле	Тип	Смысл	Роль
record_id	Text	Уникальный номер строки	Meta
date	Date	Дата наблюдения	Meta
source_batch	Categorical	Пакет выгрузки из кассовой системы	Meta
pedestrian_flow	Numeric	Людей прошло рядом за день	Feature
office_occupancy_pct	Numeric	Заполненность ближайших офисов, %	Feature
delivery_orders	Numeric	Количество заказов доставки	Feature
temperature_c	Numeric	Средняя температура, °C	Feature
rain_mm	Numeric	Осадки за день, мм	Feature
promo_budget_ils	Numeric	Рекламный бюджет за день, ₴	Feature
revenue_ils	Numeric	Дневная выручка, ₴	Target

Перед моделированием ответьте

- Что является объектом наблюдения?
- Какая переменная целевая?
- Почему record_id, date и source_batch не стоит использовать как обычные признаки?
- Какие признаки вы ожидаете увидеть со знаком «плюс», а какие — со знаком «минус»?

МИНИ-ГИПОТЕЗА

До построения графиков запишите два предположения. Например: «больше пешеходов — выше выручка» или «сильный дождь уменьшает выручку». Потом вернитесь к ним.

Маршрут в Orange

Сначала роли переменных и качество данных. Модель строим только после этого.

1. Загрузите данные

Перетащите File на холст, выберите coffee_point_revenue_orange.csv и подключите Data Table. Проверьте: 240 строк, 10 столбцов, revenue_ils без пропусков.

2. Назначьте роли

Подключите Select Columns. Перенесите revenue_ils в Target, шесть числовых факторов — в Features, a record_id, date и source_batch — в Meta Attributes.

3. Проверьте пропуски

Подключите Feature Statistics и Data Table. Найдите поля с пропусками и посчитайте, насколько проблема велика. Не удаляйте всю строку автоматически.

4. Обработайте пропуски

Подключите Impute и выберите простую стратегию для числовых признаков (Average/Mean). В отчёте объясните, почему для этого учебного набора такой выбор разумен.

Минимальная схема

Источник	Подготовка	Дальше
File → Data Table	Select Columns → Impute	EDA и Linear Regression

КОНТРОЛЬНАЯ ТОЧКА

До EDA у вас должно остаться ровно 6 Features, 1 Target и 3 Meta. Если source_batch попал в признаки, остановитесь и исправьте роли.

Посмотрите на данные глазами

Задача EDA — не создать как можно больше графиков, а заметить форму связей, пропуски и странности.

Обязательные проверки

Распределения. Distributions: изучите revenue_ils и каждый фактор. Есть ли сильная асимметрия, много нулей, необычные значения?

Числовая сводка. Feature Statistics: минимум, максимум, среднее, медиана и пропуски.

Связи с целью. Scatter Plot: revenue_ils по оси Y; по очереди выберите минимум 4 фактора по оси X.

Связи между признаками. Correlations: убедитесь, что пары факторов не дублируют друг друга почти полностью.

Подозрительные строки. Data Table: откройте строки, которые выглядят необычно на графиках, и проверьте их значения.

Что записать в EDA-отчёт

- диапазон и типичное значение выручки;
- сколько пропусков вы нашли и как их обработали;
- два признака с наиболее заметной связью с выручкой;
- один слабый или неочевидный эффект;
- есть ли пары признаков с очень высокой корреляцией (подробно этот риск разберём позже);
- один скриншот самого информативного графика с подписью: что именно на нём видно.

НЕ ПЕРЕПУТАЙТЕ

Корреляция признака с выручкой показывает связь пары переменных. Коэффициент регрессии показывает ожидаемое изменение выручки при изменении этого признака, когда остальные признаки модели зафиксированы.

Постройте линейную регрессию

Теперь у вас есть подготовленные данные и гипотезы. Пора проверить, насколько модель умеет прогнозировать.

5. Добавьте Linear Regression

Соедините Impute с Linear Regression. Выберите No regularization и изучите коэффициенты: так их легче интерпретировать.

6. Первое знакомство с качеством

Select Columns подключите к Test & Score (Data), а Linear Regression — к Learner. Выберите 10-fold Cross Validation; запишите R^2 , RMSE и MAE. Не используйте Test on train data. Почему данные идут до Impute, разберём на следующей неделе.

7. Посмотрите прогнозы

Impute подключите к Predictions (Data), а Model — к Predictors. Покажите Difference и Absolute difference. Это диагностика на данных обучения, а не независимая оценка.

8. Найдите крупные ошибки

Отсортируйте строки по Absolute difference. Найдите три крупнейшие ошибки. Что у этих дней общего и какого признака, возможно, не хватает?

Как читать метрики

Метрика	Как читать	Что лучше
R^2	Доля изменчивости, воспроизведённая моделью в проверке	Ближе к 1
MAE	Средняя абсолютная ошибка в шекелях	Меньше
RMSE	Ошибка в шекелях; сильнее наказывает крупные промахи	Меньше

БРИТВА ОККАМА И КОЭФФИЦИЕНТЫ

Не добавляйте технические поля «на всякий случай»: начните с шести содержательных факторов. И не сравнивайте величины коэффициентов без учёта единиц измерения. Объясните минимум три коэффициента в исходных единицах.

Расскажите историю результата

Отчёт должен быть понятен человеку, который не открывал Orange и не обязан знать формулы.

Ответьте по существу

- Запишите R^2 , RMSE и MAE. Можно ли использовать модель как первый ориентир?
- Какие три признака имеют самый заметный практический эффект? Укажите знак и интерпретацию коэффициента.
- Совпали ли знаки коэффициентов с вашими предварительными гипотезами?
- Какой фактор оказался неожиданным? Почему, по-вашему, получился такой результат?
- Какие дни модель прогнозирует хуже всего?
- Что вы осторожно порекомендуете менеджеру или предложите проверить экспериментом?
- Какого важного фактора нет в таблице, хотя в реальной жизни он мог бы влиять на выручку?
- Почему хороший R^2 ещё не доказывает причинно-следственную связь?

Формат вывода

ШАБЛОН

«Модель объясняет примерно ____% изменчивости дневной выручки. Средняя ошибка составляет около ____ ₴. При прочих равных увеличение ____ на ____ связано с изменением прогноза примерно на ____ ₴. Поэтому менеджеру стоит _____. Однако модель не учитывает _____, а наблюдаемая связь не обязательно является причинной».

Рефлексия об ИИ

- Использовали ли вы ИИ? Если да, на каком этапе анализа?
- Назовите один ответ ИИ, который вы проверили по данным, а не приняли на веру.
- Что вы сделали сами, без ИИ?
- ИИ сделал вас внимательнее к данным или только ускорил оформление? Приведите пример.

ПРАВИЛО КУРСА

ИИ можно использовать как напарника, но финальный вывод должен быть вашим. Если фраза не подтверждается графиком, метрикой или таблицей, она не становится верной от того, что звучит уверенно.

ФИНИШ

Что сдавать

Соберите результат так, чтобы другой человек мог воспроизвести ваш анализ.

Комплект сдачи

- файл Orange Workflow (.ows);
- короткий отчёт на 2–4 страницы или 6–8 содержательных слайдов;
- минимум 2 подписанных визуализации;
- короткая таблица с R^2 , RMSE и MAE — без углублённой оценки;
- интерпретация минимум 3 коэффициентов;
- бизнес-рекомендация, ограничение модели и рефлексия об ИИ.

Критерии

Блок	Баллы	Что проверяется
Подготовка данных	15	Роли, Meta, пропуски, понятный workflow
EDA	25	Графики, наблюдения, проверка связей
Модель	25	Регрессия, коэффициенты и корректный запуск Test & Score
Интерпретация	25	Коэффициенты, рекомендация, ограничения
Рефлексия об ИИ	10	Конкретность, проверка ответа ИИ, личный вклад

Финальный самочек

- ☐ Я не использовал Meta как признаки.
- ☐ Я обработал пропуски до моделирования.
- ☐ Для метрик я использовал Cross Validation, а не Test on train data.
- ☐ Я назвал метрики вместе с единицами измерения.
- ☐ Я отделил связь от причинности.
- ☐ Каждый важный вывод подтверждён числом или графиком.

Удачи. Пусть модель ошибается меньше, чем бариста в вашем имени ☕