Урок 28. Машинное обучение: датасет и регрессионная модель

9 класс · Урок 28 из 34 · 45 минут

Вся программа 9 класса

Регрессия в машинном обучении на Python. Урок для 9 класса о подготовке данных, прогнозе продаж и проверке того, помогает ли модели новый признак.

Содержание урока

- 01. Цель урока
- 02. Ход урока
- 03. Этап 1. Вступление: Сколько напечатать
- 04. Этап 2. Из наблюдений в модель
- 05. Этап 3. Опыт с прогнозом продаж
- 06. Этап 4. Как выбрать другую модель
- 07. Этап 5. Главное и проверка понимания

01. Цель урока

Научиться готовить данные и строить в Python модель, которая прогнозирует количество продаж, а затем проверять, помогает ли ей новый признак.

02. Ход урока

Этап	Время
1. Вступление: Сколько напечатать	5 мин
2. Из наблюдений в модель	11 мин
3. Опыт с прогнозом продаж	17 мин
4. Как выбрать другую модель	8 мин
5. Главное и проверка понимания	4 мин
Всего	45 мин

03. Этап 1. Вступление: Сколько напечатать 5 мин

ГОВОРИМ УЧЕНИКАМ

Кружок линогравюры продаёт открытки на школьной ярмарке. В прошлый раз мимо стола прошло много людей, но покупки пошли только после того, как ребята поставили стенд с образцами. Перед новой ярмаркой нужно решить, сколько открыток подготовить. Как понять, достаточно ли для прогноза числа посетителей?

Сегодня построим численный прогноз с помощью регрессии в машинном обучении на Python. Затем проверим, как меняется результат, если учесть дополнительное условие.

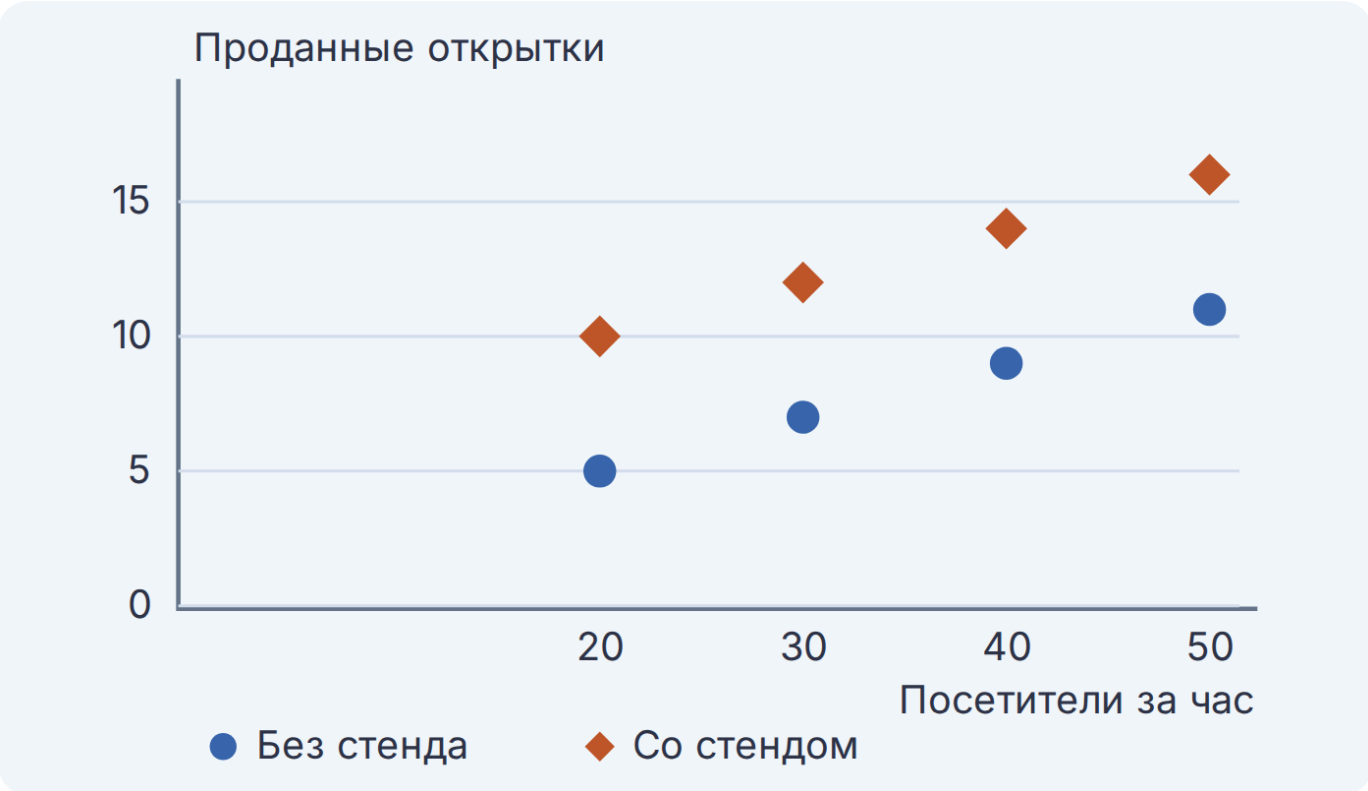
04. Этап 2. Из наблюдений в модель 11 мин

ГОВОРИМ УЧЕНИКАМ

Число посетителей может влиять на продажи, но сами по себе эти данные не говорят, что увидели покупатели. Чтобы проверить пользу стенда, надо записывать, стоял ли он у стола в каждом раунде. Лучше собрать случаи и со стендом, и без него при похожем числе посетителей. Тогда проще сравнить продажи в раундах с похожим числом посетителей и проверить, есть ли разница, когда ставят стенд.

В науке о данных, или Data Science, сначала задают вопрос, затем собирают и проверяют данные, строят модель при необходимости и объясняют, что можно сделать с результатом. Машинное обучение начинается там, где программа подбирает закономерность по примерам. В датасете каждая строка описывает один раунд ярмарки. Посетители и наличие стенда — признаки. В прошлых раундах число посетителей уже известно. Перед новой ярмаркой его придётся оценить, а решение поставить стенд можно принять заранее. Фактические продажи — то, что мы хотим предсказать. Перед обучением проверяем, что строки измерены за равные промежутки, и различаем пропуск в записи и нулевые продажи.

Линейная регрессия ищет линию, которая проходит как можно ближе к точкам по выбранному правилу. С одним признаком наклон показывает, как меняется прогноз при росте числа посетителей. Со вторым признаком модель может добавить поправку, если поставили стенд. Чтобы узнать, насколько эта поправка помогает, заранее откладываем несколько раундов. Считаем среднюю абсолютную ошибку: берём разницу между прогнозом и фактом без знака, а потом усредняем. Для сравнения пригодится простой ориентир, например средние продажи обучающих раундов.



Восемь раундов, на которых обучается модель.

05. Этап 3. Опыт с прогнозом продаж 17 мин

ГОВОРИМ УЧЕНИКАМ

Два стола при одинаковом числе посетителей могут продать разное количество открыток. Посмотрим, видна ли такая разница на графике и сможет ли модель использовать её в прогнозе. Сравним варианты на одних и тех же проверочных строках.

Задание. Проверьте, помогает ли новый признак

1. Откройте код в подготовленном Python с пакетами `scikit-learn` и `matplotlib`. Первые восемь строк используются для обучения, последние четыре — для проверки. Выполните код, сравните прогнозы и ошибки. На графике найдите два раунда с похожим числом посетителей и разными продажами.
2. Придумайте четыре новых учебных раунда с двумя уровнями посещаемости: для каждого уровня задайте один раунд со стендом и один без стенда. Запишите ожидаемые продажи до запуска прогноза. Замените последние четыре строки своими и повторите сравнение. Обучающие строки оставьте прежними.
3. В паре решите, какой прогноз вы взяли бы как ориентир для подготовки открыток. Объясните, какие настоящие данные могли бы изменить ваше решение.

Код для графика и прогноза

```
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error

# Строка: посетители за час, стенд (0 или 1), проданные открытки.
rows = [
    (20, 0, 5), (20, 1, 10), (30, 0, 7), (30, 1, 12),
    (40, 0, 9), (40, 1, 14), (50, 0, 11), (50, 1, 16),
    (25, 0, 6), (25, 1, 10), (45, 0, 10), (45, 1, 16),
]
train, test = rows[:8], rows[8:]

def features(data, with_stand):
    return [[visitors, stand] if with_stand else [visitors]
            for visitors, stand, _ in data]

def sales(data):
    return [sold for _, _, sold in data]

actual = sales(test)
average = sum(sales(train)) / len(train)
print("Ориентир по среднему:", mean_absolute_error(actual, [average] *
len(test)))

for with_stand in (False, True):
    model = LinearRegression().fit(features(train, with_stand),
sales(train))
    predicted = model.predict(features(test, with_stand))
    print("С учётом стенда:", with_stand)
    print("Прогнозы:", predicted.round(1), "Факт:", actual)
    print("Средняя ошибка:", round(mean_absolute_error(actual,
predicted), 2))

for stand, label in [(0, "Без стенда"), (1, "Со стендом")]:
    points = [(x, y) for x, s, y in train if s == stand]
    plt.scatter([x for x, _ in points], [y for _, y in points],
label=label)
plt.xlabel("Посетители за час")
plt.ylabel("Проданные открытки")
plt.legend()
plt.show()
```

Результат: два сравнения ошибок и вывод о том, когда дополнительный признак помог. Укажите, какая проверка относится к учебным данным и какие данные нужны для решения на настоящей ярмарке.

Для учителя

Первая проверка даёт среднюю абсолютную ошибку 2,75 для постоянного среднего, 2,5 для модели с одним признаком и 0,5 для модели со стендом. Класс может получить другой порядок на новых строках. Засчитывайте решение, если пары сравнивают модели по одним и тем же строкам и объясняют, какие данные собрали бы дальше. Для работы в классе пакеты лучше установить заранее. Если доступен только один компьютер, пары предлагают новые строки, а учитель прогоняет два варианта.

06. Этап 4. Как выбрать другую модель 8 мин

ГОВОРИМ УЧЕНИКАМ

Регрессия подходит, когда нужен ответ числом. Если ответ — «успеем продать весь тираж» или «не успеем», можно использовать классификацию. Дерево решений, например, учится задавать вопросы о признаках и выбирать ответ по ветке. Если заранее нет нужных ярлыков и требуется найти похожие дни ярмарки, подойдёт кластеризация. Эти методы используют закономерности в данных, но дают разные результаты.

Обсуждение. Три вопроса организатора

Организатор хочет предсказать, сколько изделий продадут, заранее отметить смены с риском распродажи и выяснить, какие смены похожи друг на друга. Для каждого вопроса предложите тип модели. Выберите один вопрос и объясните, как вы проверили бы полезность результата.

Для учителя

Число — регрессия, категория риска — классификация, группы похожих смен — кластеризация. Для первых двух полезна проверка по отложенным реальным сменам. Для групп — проверка, что они устойчивы, понятны организатору и помогают принять решение, например распределить запасы.

07. Этап 5. Главное и проверка понимания 4 мин

> ГОВОРИМ УЧЕНИКАМ
>
> Новый столбец полезен, когда его значение известно до прогноза и помогает точнее ответить на новых примерах. Эту пользу нужно измерить.

Главное за урок

- Наблюдения нужно измерять одинаковым способом, а различия условий записывать отдельными признаками.
- В регрессии модель учится предсказывать число, а сравнивать её удобно с простым ориентиром.
- Проверочные строки показывают, помогает ли признак на примерах, которых модель не видела при обучении.

Итоговые вопросы

1. Можно ли взять число проданных открыток за тот же час как признак для прогноза этого числа? Почему?
2. Если модель со стендом ошибалась меньше на учебных строках, какой следующий результат нужен для решения о печати?

Для учителя

1. К моменту прогноза ответ ещё неизвестен. Использовать его как признак значит подсказать модели то, что требуется узнать. 2. Сравнение ошибок на отложенных реальных раундах, где были разные уровни посещаемости и условия у стола. Лучший прогноз в этой проверке ещё не доказывает, что стенд сам по себе вызвал рост продаж.

Источники

- scikit-learn. LinearRegression · Проверено 05.11.2026. Линейная регрессия подбирает коэффициенты по методу наименьших квадратов.
- scikit-learn. Mean absolute error · Проверено 05.11.2026. Средняя абсолютная ошибка оценивает размер ошибки регрессии.