



> Конспект > 2 урок > Построение модели предсказания цены в зависимости от факторов

> Оглавление

- > [Оглавление](#)
- > [Данные и ноутбук](#)
 - [Описание датасетов](#)
 - [Этапы анализа](#)
- > [Метрики для оценки регрессии](#)

> Данные и ноутбук

Материалы лекции можно [скачать тут](#) вместе с ноутбуком в `ipynb` формате.
Отдельно лежит [html версия](#).

Описание датасетов

По названиям переменных обычно можно догадаться, какие данные они содержат. И все же поясним используемые в лекции переменные.

Основные данные `*_main_data.csv`:

КУПЛЕНО НА
SKLADCHIK.COM

- **id** — идентификатор квартиры
- **timestamp** — дата совершения покупки
- **full_sq** — общая площадь
- **life_sq** — жилая площадь
- **floor** — этаж
- **max_floor** — количество этажей в доме
- **build_year** — год постройки
- **num_room** — количество комнат
- **kitch_sq** — площадь кухни
- **apartment condition** — состояние квартиры
- **sub_area** — территория (например, город)
- **price** — цена квартиры (целевая переменная)

Дополнительные данные `*_additional_data.csv`:

- **healthcare_facilities** — количество ближайших объектов здравоохранения
- **university_num** — количество университетов
- **culture_objects_facilities** — количество культурных объектов

Этапы анализа

Напомним примерный план анализа данных:

1. **EDA** — исследуем закономерности
2. **Подготовка данных** — добавляем новые фичи, приводим данные в нужный вид
3. **Моделирование** — обучаем модели
4. **Пост анализ** — выбираем модель или решаем, нужно ли дальше ее тюнить

> Метрики для оценки регрессии

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 — \text{среднеквадратичное отклонение}$$

- Чем оно выше, тем хуже модель.
- MSE идеальной модели равно 0.
- Сильно штрафует за большие ошибки (чувствительна к выбросам) — можем думать, что модель хуже, чем она есть на деле, но при маленьких ошибках (< 1) обратная ситуация.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} = \sqrt{MSE} — \text{корень из MSE}$$

- Корень введён для перехода к размерности целевой переменной.
- Метрика во многом аналогична MSE .

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| — \text{средняя абсолютная ошибка}$$

- Штрафует за большие ошибки меньше, чем $MSE / RMSE$ (менее чувствительна к выбросам).
- Требуется приближение для вычисления градиента.

$$R^2 = 1 - \frac{MSE(model)}{MSE(ocm)} — \text{коэффициент детерминации}$$

- Недостаток — увеличивается при добавлении объясняющих переменных.
- Максимальное значение равно 1, минимальное значение зависит от данных, на которых произведён расчёт.

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% — \text{средняя абсолютная ошибка в процентах}$$

- Взвешенная метрика MAE .
- Оптимальное решение — взвешенная медиана целевых значений.

$$RMSLE = \sqrt{MSE(\log(y_i + 1), \log(\hat{y}_i + 1))} — \text{среднеквадратичная логарифмическая ошибка}$$

- $RMSE$ в логарифмическом масштабе.
- Внутри логарифмов добавляются константы.
- Больше штрафует за заниженную оценку целевой переменной.