Так как записи без цены не дадут нам возможности обучить модель или проверить ее качество, удалим их.

Для предсказания цены нам необходим только год и месяц. Выделим их из данных о дате публикации объявления.

В адресе оставим только улицу, для упрощения обработки категориального признака.

В списке улиц присутствуют неявные дубликаты. Приведем названия улиц к единому написанию.

Добавим столбец с информацией о том, является ли этаж квартиры последним.

Преобразуем типы данных у числовых столбцов.

Модель стоимости жилья Магнитогорске

Импорт библиотек и загрузка данных

Введение

Данные

Загрузка данных из базы данных

подключаемся к базе данных

загружаем данные

закрываем соединение с базой данных

Первичное знакомство с данными

Для трех квартир цена не указана, поэтому удалим эти строки, так как они не подходят ни для обучения модели ни для контроля качества.

контроль размерности

В столбце с типом квартиры хранятся данные о количестве комнат и типе квартиры. Разделим их и сохраним в двух разных столбцах.

Посмотрим количество пропусков в столбце с данными о районе.

Заполним пропуски значением неизвестно.

Построение модели

Подготовка обучающей и валидационной выборки.

Выделим характеристики, которые будем использовать для обучения модели.

не влияет на цену квартиры

может оказывать влияние

вместо него будем использовать более информативные столбцы, сгенерированные на основе данных из этого столбца

целевой признак

id для демонстрации на собеседовании

Разобьем выборки на обучающую и валидационную в отношении 4 : 1.

Предварительна обработка данных

Выделим категориальные и числовые признаки.

Количество комнат попадает в категориальные признаки, потому что содержит значение «многоквартирная», которое нельзя заменить конкретным числовым значением.

Преобразуем категориальные признаки в числа с помощью порядкового кодирования.