Государственное образовательное учреждение высшего профессионального образования

«Московский государственный технический университет имени Н.Э. Баумана»

(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ_____Информатика и системы управления_____

КАФЕДРА _____Системы обработки информации и управления_____

Р А С Ч Ё Т Н О - П О Я С Н И Т Е Л Ь Н А Я З А П И С К А

к научно-исследовательской работе:

_____Предсказание стоимости жилья_____

Студент группы ____**ИУ5-31М**_____ _____ _____**Байбарин Р.Г.**__
(Подпись, дата) (И.О.Фамилия)

Руководитель _____ ___**Гапанюк Ю.Е.**___
(Подпись, дата) (И.О.Фамилия)

Москва, 2021

Оглавление

Постановка задачи

Целью научно-исследовательской работы является построение модели машинного обучения для предсказания стоимости жилых помещений (квартир и домов) по набору признаков. В качестве обучающей выборки использовался датасет стоимостей жилья разных городов и штатов США.

Анализ

Рассмотрим разведочные характеристики полученного датасета. На рисунке 1 приведены статистические распределения значимых признаков.

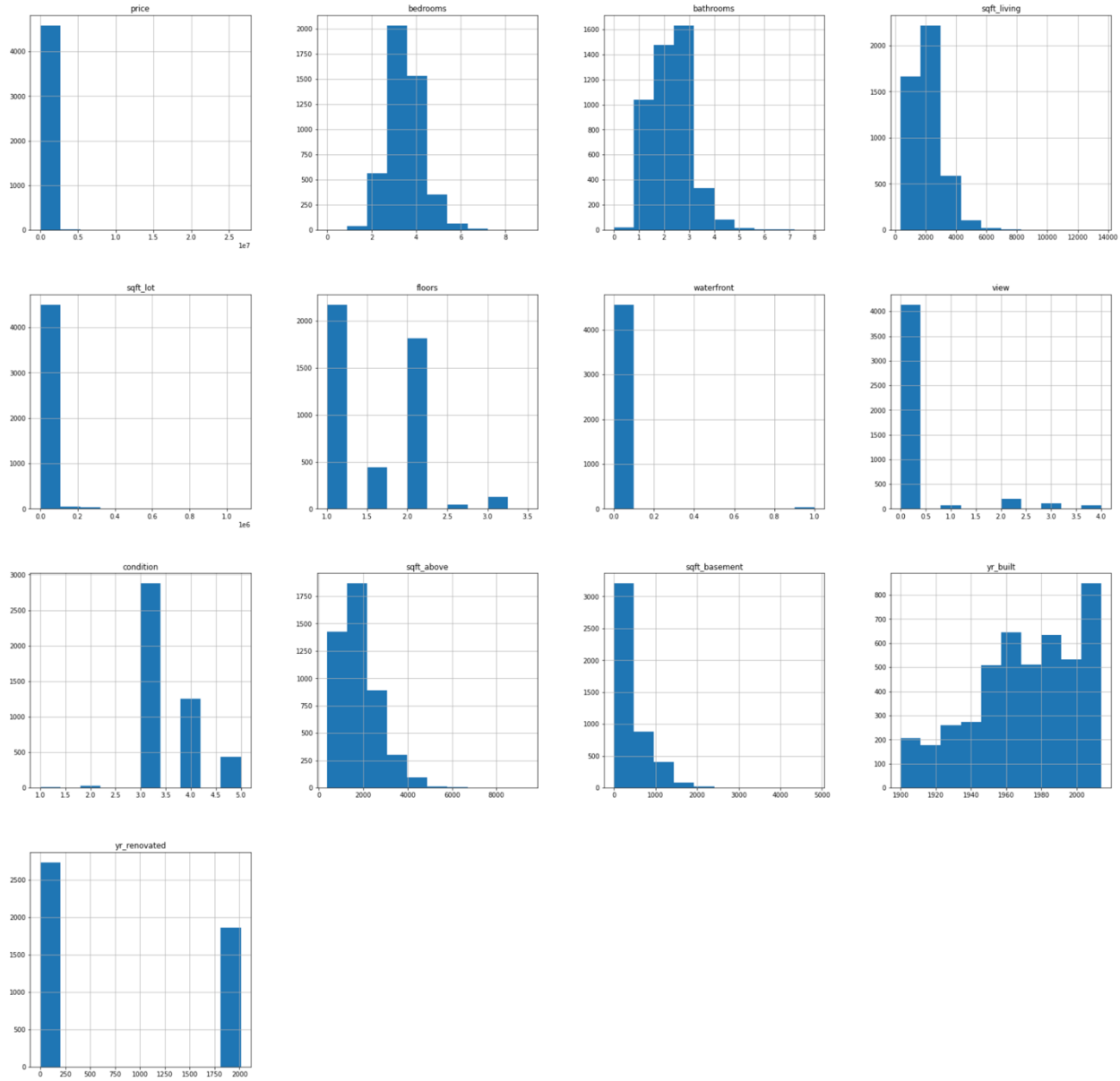

Рис. 1. Распределение значимых признаков

Можем заметить, что ряд признаков (такие как стоимость, жилая площадь) содержат выбросы, которые будут отрицательно влиять на итоговую модель.
Оценим наличие пропусков в данных. На рисунке 2 приведены данные о количестве выбросов в значимых признаках датасета.

date	0
price	0
bedrooms	0
bathrooms	0
sqft_living	0
sqft_lot	0
floors	0
waterfront	0
view	0
condition	0
sqft_above	0
sqft_basement	0
yr_built	0
yr_renovated	0
street	0
city	0
statezip	0
country	0

Рис. 2. Количество пропусков в значимых признаках датасета

Как видно из рисунка 2, данные не страдают от наличия пропусков. Шаг заполнения пропусков может быть пропущен.
Также стоит заметить из рисунка 1, что различные признаки содержат значения разного порядка. Для устранения влияния данного фактора на итоговую модель следует произвести нормализацию некоторых признаков.

Очистка выбросов

На рисунке 3 приведена диаграмма «ящик с усами» признака «стоимость» исходного датасета.

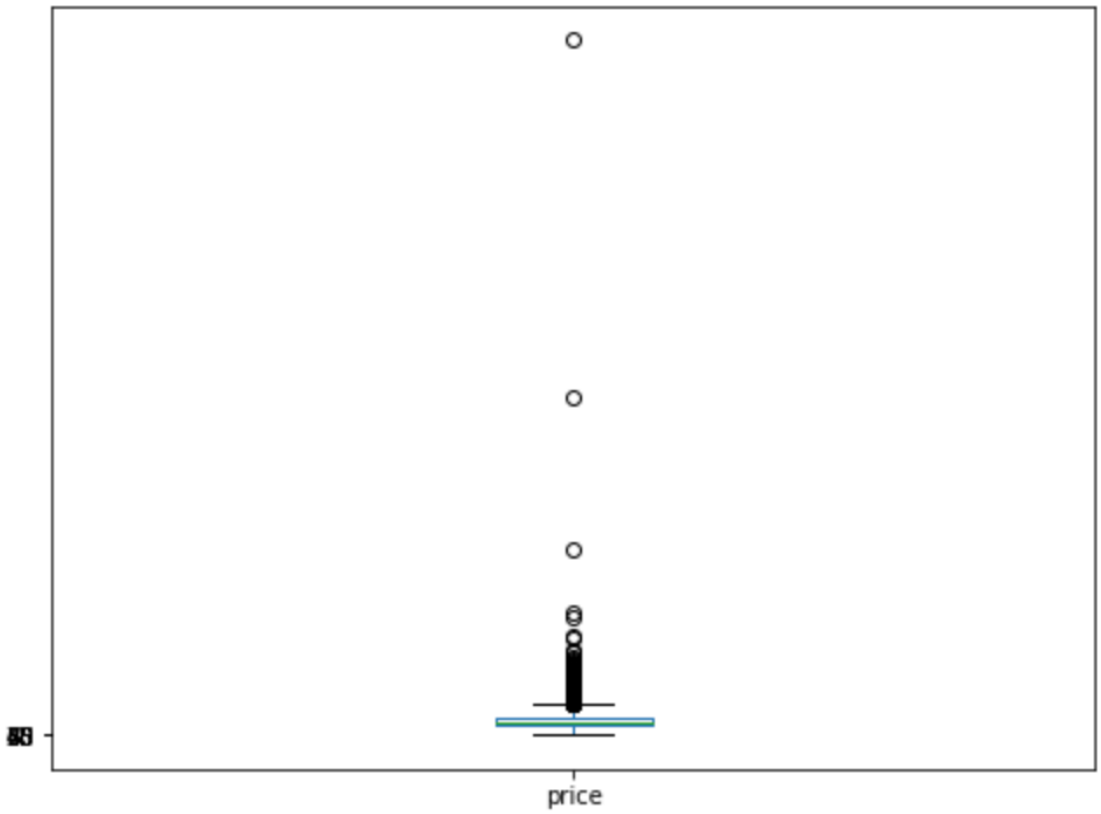

Рис. 3. Распределение стоимости жилья

Эмпирически избавимся от выбросов. При очистке было удалено 5% выборки. Проведем аналогичную операцию с колонками «sqft_lot».

Обратим внимание на распределение значений в данных колонки «waterfront». Данный признак является бинарным, причем распределение классов является крайне несбалансированным. Данный признак следует не учитывать в построении модели.

Также следует избавиться от признаков, которые сложно использовать в предсказании стоимости, а именно: дата, адрес и прочие.

Нормализация

Так как данные признаков имеют серьезный разброс, следует произвести нормализацию признаков для уменьшения абсолютных значений дисперсии параметров, таким образом повышая качество итоговой модели. Ввиду того, что все признаки, которые требуется нормализовать, имеют неотрицательную область значений, в качестве нормализатора будем использовать класс

MaxAbsScaler. Данное преобразование следует применить к признакам стоимости и жилой площади.

Кодирование категориальных признаков

Произведем кодирование категориальных признаков методом OneHotEncoding таких признаков, как: город, код штата. Данное преобразование увеличило количество признаков до 130.

Построение модели предсказания стоимости

Для начала построим модель линейной регрессии для определения начальной точки развития модели машинного обучения. На рисунке 5 приведено итоговое распределение «тестовое против предсказанного».

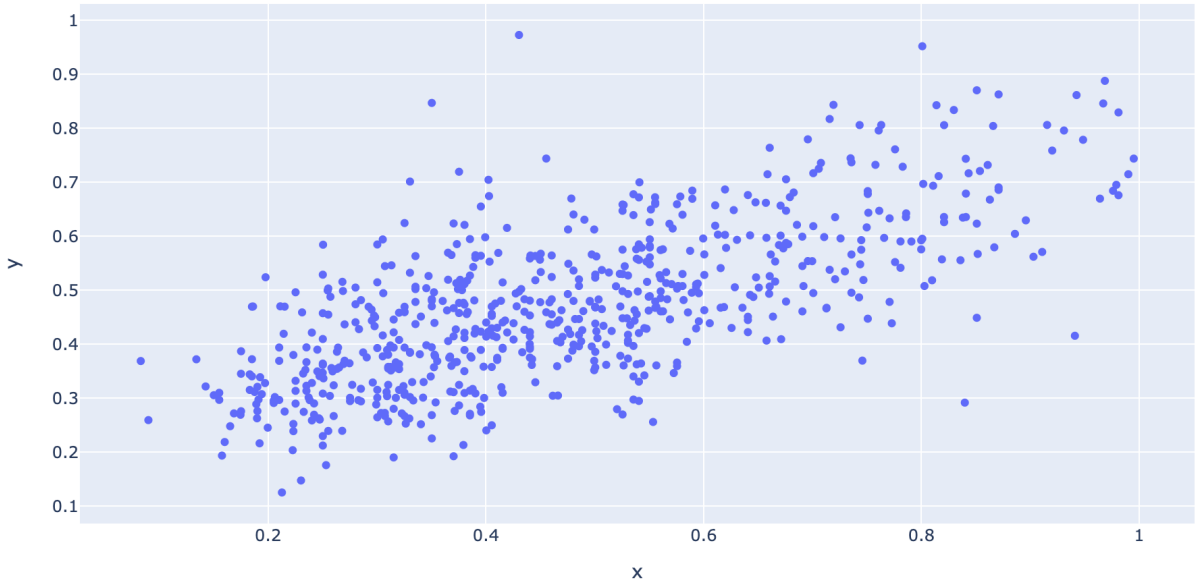

Рис. 5. Тестовое против предсказанного

Можем обратить внимание, что данные лежат достаточно кучно, достаточно близко в прямой y=x. Итоговые метрики имеют следующий вид:

MSE: 0.02

R2 Score: 0.48

Рассмотрим вариант обучения на базе модели случайного леса. В качестве гиперпараметров будем использовать n_estimators=100, а также max_features ='sqrt'. На рисунке 6 приведено распределение «тестовое против предсказанного».

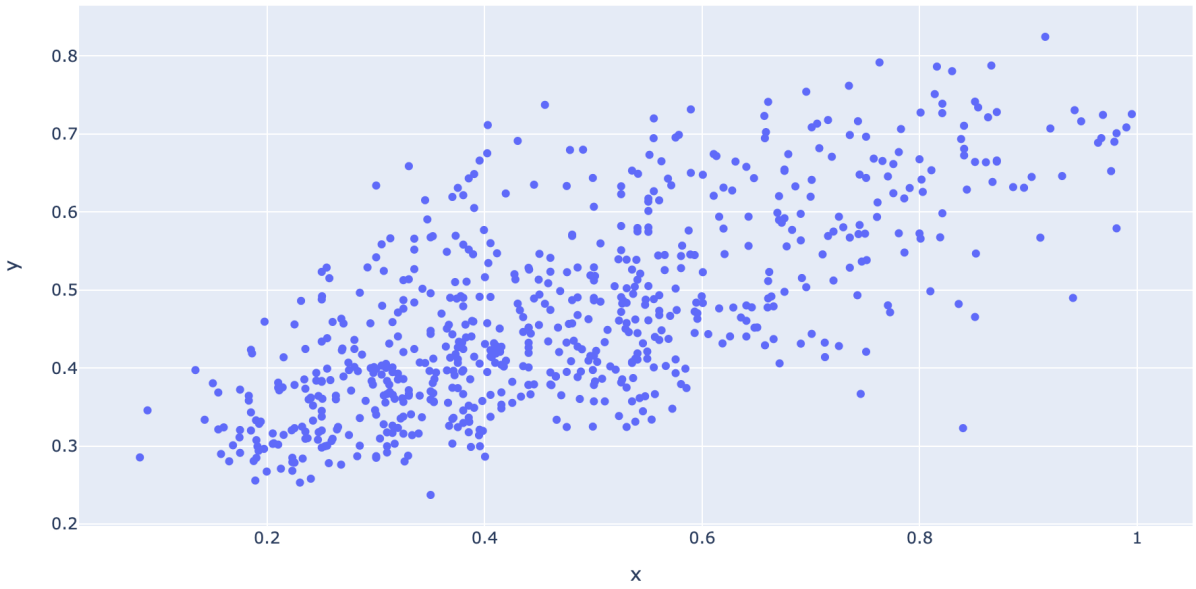

Рис. 6. Распределение результатов случайный лес

Итоговые метрики имеют следующий вид:

MSE: 0.02

R2 Score: 0.50

Вывод

В рамках выполнения научно-исследовательской работы были выполнены операции разведочного анализа данных, очистки данных, подготовки к анализу методами машинного обучения, обучение модели линейной регрессии и случайного леса.

Список использованной литературы

1. Методические указания по курсу «Методы машинного обучения», Гапанюк Ю.Е. МГТУ им. Н.Э. Баумана