ФЕДЕРАЛЬНОЕ АГЕНСТВО СВЯЗИ

Ордена Трудового Красного Знамени федеральное государственное бюджетное образовательное учреждение высшего образования

Московский технический университет связи и информатики

(МТУСИ)

Кафедра «Математическая кибернетика и информационные технологии» Дисциплина «Интеллектуальные системы»

Лабораторная работа № 3

Регрессионная модель изменения цен на дома в Бостоне

Выполнил: \_\_\_\_\_\_\_ Д.С. Калининский

Проверил: \_\_\_\_\_\_\_ А.В. Шереметьев

Москва 2022

**Оглавление**

[1 Цель работы 3](#_Toc119580633)

[2 Задачи 3](#_Toc119580634)

[3 Выполнение работы 3](#_Toc119580635)

[4 Результаты 6](#_Toc119580636)

[5 Вывод 9](#_Toc119580637)

# **1 Цель работы**

Реализовать предсказание медианной цены на дома в пригороде Бостона в середине 1970-х по таким данным, как уровень преступности, ставка местного имущественного налога и т.д.

Данный набор содержит относительно немного образцов данных: всего 506, разбитых на 404 обучающих и 102 контрольных образца. И каждый признак во входных данных (например, уровень преступности) имеет свой масштаб. Например, некоторые признаки являются пропорциями и имеют значения между 0 и 1, другие — между 1 и 12 и т. д.

Не путайте регрессию с алгоритмом логистической регрессии. Как ни странно, логистическая регрессия не является регрессионным алгоритмом — это алгоритм классификации.

# **2 Задачи**

Основные задачи лабораторной работы:

1) Ознакомиться с задачей регрессии

2) Изучить отличие задачи регрессии от задачи классификации

3) Создать модель

4) Настроить параметры обучения

5) Обучить и оценить модели

6) Ознакомиться с перекрестной проверкой.

# **3 Выполнение работы**

Набор данных присутствует в составе Keras (рисунок 1).

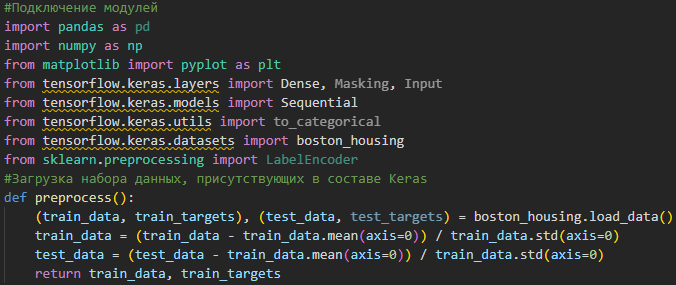


Рисунок 1 – Загрузка набора дынных

404 обучающих и 102 контрольных образца, каждый с 13 числовыми признаками. Цены в основном находятся в диапазоне от 10 000 до 50 000 долларов США. Было бы проблематично передать в нейронную сеть значения, имеющие самые разные диапазоны. Сеть, конечно, сможет автоматически адаптироваться к таким разнородным данным, однако это усложнит обучение. На практике к таким данным принято применять нормализацию: для каждого признака во входных данных (столбца в матрице входных данных) из каждого значения вычитается среднее по этому признаку, и разность делится на стандартное отклонение, в результате признак центрируется по нулевому значению и имеет стандартное отклонение, равное единице.

Определим функцию build\_model() (рисунок 2).

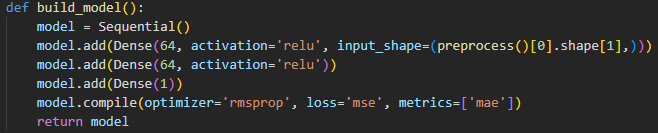


Рисунок 2 – Функцию build\_model()

Сеть заканчивается одномерным слоем, не имеющим функции активации (это линейный слой). Это типичная конфигурация для скалярной регрессии (целью которой является предсказание одного значения на непрерывной числовой прямой). Применение функции активации могло бы ограничить диапазон выходных значений: например, если в последнем слое применить функцию активации sigmoid, сеть обучилась бы предсказывать только значения из диапазона между 0 и 1. В данном случае, с линейным последним слоем, сеть способна предсказывать значения из любого диапазона.

Обратите внимание на то, что сеть компилируется с функцией потерь mse — mean squared error (среднеквадратичная ошибка), вычисляющей квадрат разности между предсказанными и целевыми значениями. Эта функция широко используется в задачах регрессии. Также добавлен новый параметр на этапе обучения: mae — mean absolute error (средняя абсолютная ошибка). Это абсолютное значение разности между предсказанными и целевыми значениями. Например, значение MAE, равное 0,5, в этой задаче означает, что в среднем прогнозы отклоняются на 500 долларов США.

Чтобы оценить качество сети в ходе корректировки ее параметров (таких, как количество эпох обучения), можно разбить исходные данные на обучающий и проверочный наборы, как это делалось в предыдущих примерах. Однако так как у нас и без того небольшой набор данных, проверочный набор получился бы слишком маленьким (скажем, что-нибудь около 100 образцов). Как следствие, оценки при проверке могут сильно меняться в зависимости от того, какие данные попадут в проверочный и обучающий наборы: оценки при проверке могут иметь слишком большой разброс. Это не позволит надежно оценить качество модели.

Хорошей практикой в таких ситуациях является применение перекрестной проверки по K блокам (K-fold cross-validation). Суть ее заключается в разделении доступных данных на K блоков (обычно K = 4 или 5), создании K идентичных моделей и обучении каждой на K—1 блоках с оценкой по оставшимся блокам. По полученным K оценкам вычисляется среднее значение, которое принимается как оценка модели. В коде такая проверка реализуется достаточно просто (рисунок 3).

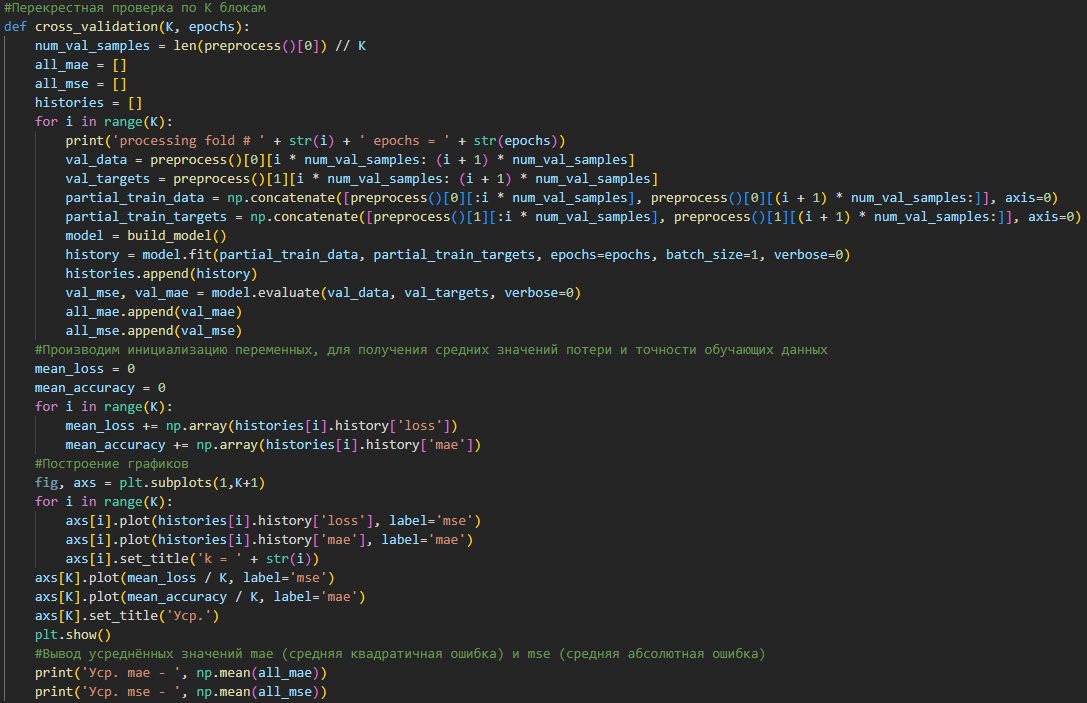


Рисунок 3 – Процедура перекрёстной проверки

Разные прогоны действительно показывают разные оценки, от 2,6 до 3,2. Средняя (3,0) выглядит более достоверно, чем любая из оценок отдельных прогонов, — в этом главная ценность перекрестной проверки по K блокам. В данном случае средняя ошибка составила 3000 долларов, что довольно много, если вспомнить, что цены колеблются в диапазоне от 10 000 до 50 000 долларов.

Необходимо уменьшить или увеличить количество эпох обучения и проанализировать полученные результаты, применить перекрестную проверку по K блокам при различных K (рисунок 4).

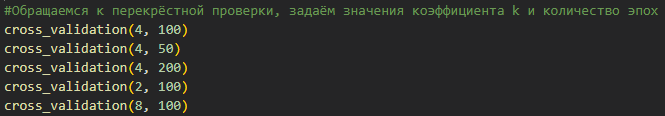


Рисунок 4 – Коэффициент K и кол-во эпох для тестирования

# **4 Результаты**

Для получения визуальных результатов на рисунках 5-9 представлены графики, используем библиотеки matplotlib и numpy.

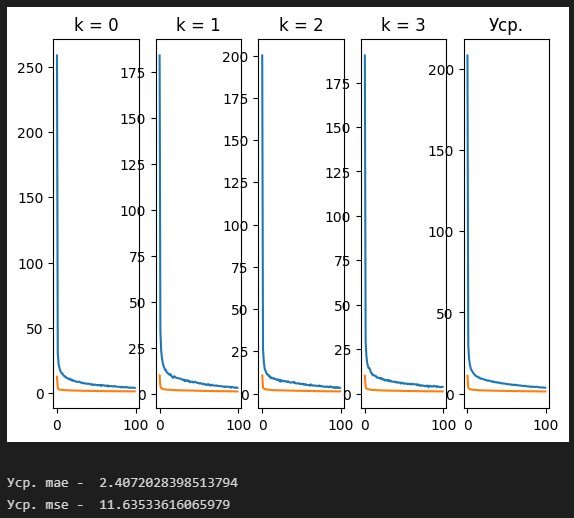


Рисунок 5 – Результаты тестирования модели K=4 и 100 эпох

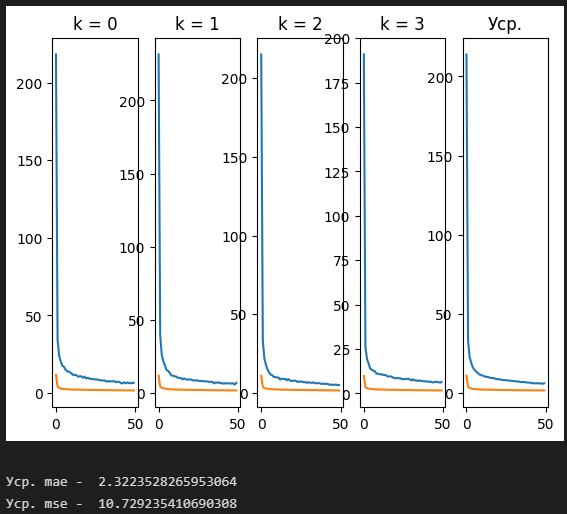


Рисунок 6 – Результаты тестирования модели K=4 и 50 эпох

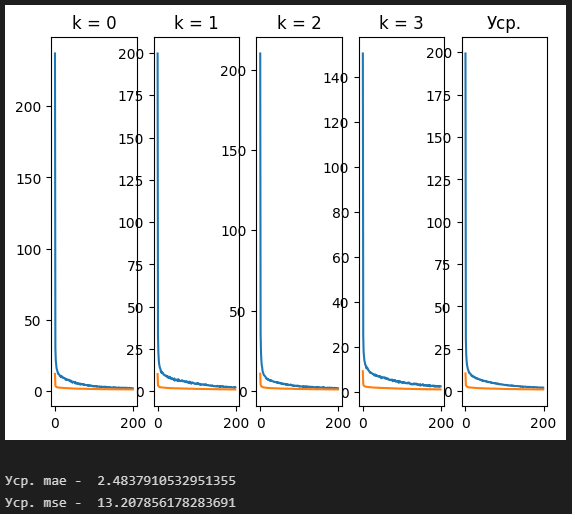
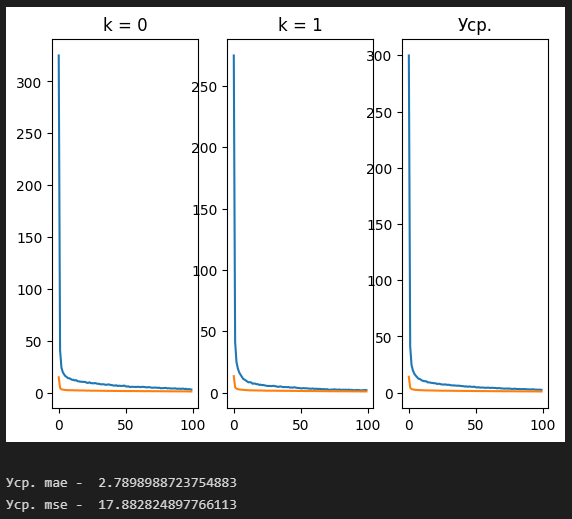


Рисунок 7 – Результаты тестирования модели K=4 и 200 эпох

  
Рисунок 8 – Результаты тестирования модели K=2 и 100 эпох

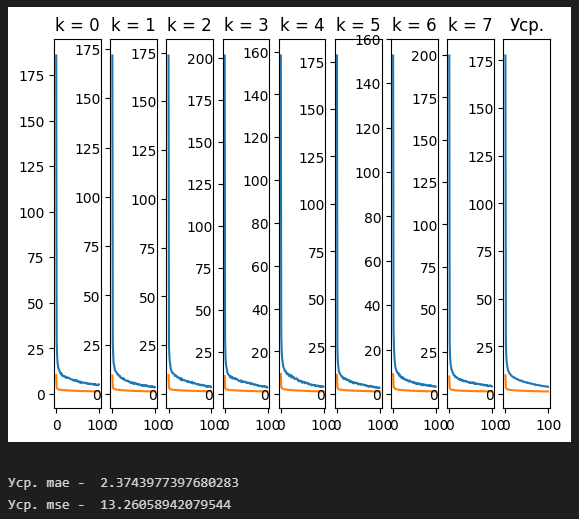


Рисунок 9 – Результаты тестирования модели K=8 и 100 эпох

# **5 Вывод**

В данной лабораторной работе, реализовали предсказание медианной цены на дома в пригороде Бостона в середине 1970-х по таким данным, как уровень преступности, ставка местного имущественного налога и т.д., взятых с библиотеки Keras. Были совершены тестирования с разным кол-вом эпох и разными коэффициентами перекрёстной проверки K, в результате:

1) При базовом значении эпох 100 было сделано два дополнительных тестирования с значениями эпох 50 и 200, в итоги из полученных результатов можем сделать вывод, что при увеличении количества эпох результат ухудшается и наоборот.

2) При базовом значении коэффициента перекрёстной проверки K = 4 было сделано два дополнительных тестирования с значениями 2 и 8, в итоге из полученных результатов можем сделать вывод, что при увеличении коэффициента K показатель улучшается и наоборот.