Лабораторная работа №6
Регрессия
(максимум – 18 баллов)

По данной работе необходимо подготовить отчет в формате блокнота Jupyter Notebook (.ipynb) либо в текстовом виде (.pdf). В отчете должны быть:

1) исходные коды
2) результаты выполнения.

Основная часть (10 баллов)

В этом задании вам предстоит построить модель для прогнозирования цены недвижимости в зависимости от того, в каком районе Бостона она располагается.

1. Загрузите данные из файла "boston.csv" о недвижимости в различных районах Бостона. Столбцы (признаки) имеют следующий смысл:
    a. CRIM – уровень преступности
    b. ZN – доля жилых земель, разделенных на участки площадью более 25 000 кв.футов
    c. INDUS – доля площадей, не связанных с розничной торговлей
    d. CHAS – наличие реки (1, если граничит с рекой; 0 в противном случае)
    e. NOX – качество воздуха (концентрация оксидов азота)
    f. RM – среднее количество комнат в доме
    g. AGE – доля жилых помещений, построенных владельцами до 1940 года
    h. DIS – взвешенные расстояния до пяти бостонских центров занятости
    i. RAD – транспортная доступность (индекс доступности радиальных автомагистралей)
    j. TAX – налоги (ставка налога на 10 000 долларов США)
    k. PTRATIO – соотношение количества учеников и учителей
    l. B – нормированное значение доли афроамериканцев среди жителей
    m. LSTAT – процент населения с низким социальным статусом
    n. MEDV – медианная цена недвижимости (тыс. $) – это и будет целевой признак
2. Проверьте, что у всех загруженных данных числовой тип.
3. Проверьте, есть ли по каким-либо признакам отсутствующие данные. Если отсутствующие данные есть – заполните их медианным значением.
4. Посчитайте коэффициент корреляции для всех пар признаков. *Подсказка: воспользуйтесь методом corr() для датафрейма, чтобы получить сразу всю корреляционную матрицу.*
5. С помощью одной из библиотек визуализации постройте тепловую карту (heatmap) по корреляционной матрице.
6. Выберите от 4 до 6 признаков (на свое усмотрение), которые в наибольшей степени коррелируют с целевым признаком (ценой недвижимости).
   Справка. Коэффициент корреляции изменяется от -1 до 1. Значение -1 означает точную обратно-пропорциональную зависимость (чем меньше одна переменная, тем больше вторая, и наоборот). Значение 1 означает точную прямо-пропорциональную зависимость. Значение 0 означает полное отсутствие зависимости. Таким образом, чем ближе модуль коэффициента корреляции к 1, тем сильнее прослеживается зависимость между признаками.
7. Для каждого из выбранных признаков в паре с целевым признаком постройте точечную диаграмму (диаграмму рассеяния).
8. Визуально убедитесь, что связь между выбранным признаком и целевым

прослеживается. Если на основе графика считаете, что зависимости нет – исключите этот признак из дальнейшего рассмотрения (но при этом как минимум 3 признака должно остаться в любом случае).

9. Сформируйте список факторных признаков и целевую переменную.
10. Выполните разбиение датасета на обучающую и тестовую выборки в соотношении 8:2. При формировании обучающей и тестовой выборок строки из исходного датафрейма должны выбираться в случайном порядке. *Подсказка: можно воспользоваться функцией train_test_split из библиотеки sklearn.model_selection*.
11. Из набора линейных моделей библиотеки sklearn возьмите линейную регрессию, обучите ее на обучающем наборе.
12. Получите векторы прогнозных значений целевой переменной на обучающей и на тестовой выборках.
13. Посчитайте коэффициент детерминации (R^2) и корень из среднеквадратичной ошибки (RMSE) на обучающей и на тестовой выборках.

Дополнительные задания (8 баллов)

14. (1 балл) Постройте boxplot («ящик с усами») для целевого признака (MEDV). Определите, какие значения можно считать выбросами.
    Указание. Если по диаграмме выбросы определить не смогли, то для выполнения дальнейших действий считайте выбросами значения MEDV=50.0.
15. (2 балла) Отфильтруйте исходные данные, удалив выбросы. Пересоздайте тестовую и обучающую выборки, переобучите модель. Посчитайте показатели R^2 и RMSE. Как они изменились? О чем это говорит?
16. (2 балла) Из набора линейных моделей библиотеки sklearn возьмите гребневую регрессию (Ridge). Обучите модель. Посчитайте показатели R^2 и RMSE.
17. (3 балла) Постройте полиномиальную регрессию с использованием полинома 3й степени. Посчитайте показатели R^2 и RMSE. Сравните все полученные результаты.