Лабораторная работа №4

Данные для лабораторной работы доступны по ссылке:
https://drive.google.com/open?id=1QHZPgETP_cet-Nm1Srt6Bk4d6c35h1VE

Задание 1 (3 балла) Загрузить данные из файла, вычислить для них основные статистики: математическое ожидание, дисперсию, среднеквадратическое отклонение, медиану, моду. Построить графики зависимости целевой переменной от нескольких признаков (на выбор).

Задание 2 (3 балла) Разделить все наблюдения 3 части случайным образов в следующих пропорциях: 80% для тренировки модели, 10% для валидации, 10% для тестирования.

Задание 3 (20 баллов) Произвести выбор признаков для модели многомерной линейной регрессии 2 способами**: forward** и **backward** feature selection**.** Качество модели оценивается по 10% данных для валидации, обучение модели производится по 80% (из предыдущего задания). В качестве метрики минимизируется сумма квадратов ошибок. Сравнивать качество моделей будем по коэффициенту детерминации R^2. (Качество моделей оценивается как в ЛР2 и ЛР3).

Задание 4 (10 баллов) Для наилучшей модели, выбранной в пункте 3, выполнить кросс-валидацию (K-fold cross validation), в качестве параметра K взять следующие три значения: 3, 5, 10.

Задание 5 (3 балла) Построить графики для фактических значений целевой переменной и для прогнозируемых с помощью лучшей модели.

Задание 6 (3 балла) Нормализовать все признаки путём вычитания соответствующего математического ожидания и деления на соответствующее среднеквадратическое отклонение.

Задание 7 (10 баллов) Попробовать выбрать значимые переменные при помощи регуляризации, а именно использовать RidgeRegression. Параметр для регрессии выбирается вами произвольно, типовые значения 0.001, 0.01, 0.1 и т.д. Качество модели оценивается как в ЛР2 и ЛР3.

Задание 8 (10 баллов) Попробовать выбрать значимые переменные при помощи регуляризации, а именно использовать LassoRegression. Параметр для регрессии выбирается вами произвольно, типовые значения 0.001, 0.01, 0.1 и т.д. Качество модели оценивается как в ЛР2 и ЛР3.

Задание 9 (10 баллов) Для моделей, полученных в заданиях 7 и 8, произвести подбор оптимального параметра с помощью `GridSearchCV`.

Задание 10 (3 балла) Построить графики для фактических значений целевой переменной и для прогнозируемых с помощью лучшей модели для заданий 7 и 8.

Задание 11 (5 баллов) Повторить задание 4 для моделей из заданий 7 и 8.

Задание 12 (15 баллов) Сделать вывод о том, какая модель из построенных является наиболее подходящей. Проверить стоит ли добавлять в модель нелинейность (признаки в квадрате, произведения между признаками). Провести эксперименты. Вычислить ошибки и коэффициенты детерминации.