Лабораторная работа №1

Дано
датасет LR_insurance.csv, содержащий данные о клиентах страховой компании
- age — возраст
- sex — пол
- bmi — body mass index — индекс массы тела
- children — количество детей
- smoker — курящий/некурящий
- region — регион проживания
- charges — финальная стоимость страховки

Два столбца является категориальным значением — sex,region
Smoker можно интерпретировать двояко: как категорию и как число (0 или 1)

Задание
1) Проанализировать распределения фичей. При необходимости избавиться от выбросов
2) Разбить датасет в соотношении 80% - тренировочный, 20% - тестовый
3) На тестовых данных: спрогнозировать стоимость страховки, натренировав линейную регрессию (age, bmi, children, [smoker]) → charges, отдельно с/без smoker
4) Исходя из качества полученных моделей сделать вывод о целесообразности интерпретации поля smoker как категориального.
5) Рассчитать ключевые характеристики качества модели на тестовых данных.
6) Проанализировать важность фичей с помощью permutation importance.
7) [BONUS] Проанализировать влияние на результат категориальных полей. Цель — улучшить предсказываемость модели в терминах среднеквадратичного отклонения

Система оценивания
Задание будет оценено по 5-балльной шкале. По итогам курса за каждую лабораторную работу будет стоять оценка. На основе средней оценки будет рассчитана оценка за практическую часть курса. Данная оценка войдет с весом около 0.5 [окончательная развесовка будет уточнена с преподавателем по теории] в финальную оценку за курс.
За позднюю сдачу задания будет снят 1 балл.
Выполнение бонусной части опционально. Оно повлияет на личную лояльность преподавателя к студенту во время непосредственно экзамена.

Формат выполнения
.ipynb файл, без ошибок прогоняемый сверху вниз с любого компьютера
Файл выслать преподавателю посредством мессенджера Telegram : @evgeny_zavalnyuk

Срок выполнения [время фиксируется исходя из отправки решения преподавателю]
17 апреля 2024г. до 23:59:59 МСК