Анализ данных

Кейс компании DOC+

Мы предлагаем вам пройти тестовое задание на знание методов обработки данных. Задание никак не связано непосредственно с медициной и необходимо для демонстрации ваших знаний и умении применять их на практике.

Задание основано на открытом датасете https://archive.ics.uci.edu/ml/datasets/automobile , содержащем информацию о параметрах различных автомобилей (подробное описание датасета по ссылке).

Датасет содержит всего 205 строк, очевидно этого недостаточно для создания полноценной стабильной модели. Цель данного кейса не решить конкретную задачу, добившись максимальной точности, а продемонстрировать ход ваших мыслей и владение инструментами и методами анализа данных. Задание рассчитано на 3-4 часа.

Пожалуйста, проанализируйте представленные данные (dataset.csv) и ответьте на несколько вопросов:

1. Придумайте и опишите наибольшее число вариантов использования этих данных. Для решения каких практических задач их можно применить?

2. Выберите одну из описанных вами задач и реализуйте ее решение на Python / R при помощи оптимальной на ваш взгляд модели. Опишите, почему вы выбрали именно эту модель.

3. Опишите ваш подход к предварительному анализу (в т.ч. визуализации) и обработке данных, работе с признаками, кросс валидацией, настройкой модели и ее оценкой. Объясните выбор того или иного решения и, по возможности, сравните с альтернативами.

4. Опишите, как бы вы улучшили модель из пункта 2, если бы у вас было больше времени и существенно больший датасет.

Желаем успеха!

С уважением,

команда DOC+