**Краткое описание**

**Описание задания:**

Вас пригласили поучаствовать в одном из проектов UNICEF — международного подразделения ООН, чья миссия состоит в повышении уровня благополучия детей по всему миру.

Суть проекта — отследить влияние условий жизни учащихся в возрасте от 15 до 22 лет на их успеваемость по математике, чтобы на ранней стадии выявлять студентов, находящихся в группе риска.

Для этого нужно построить модель, которая предсказывала бы результаты госэкзамена по математике для каждого ученика школы.

**Цель проекта:**

Провести разведывательный анализ данных и определиться, какие признаки мы оставим для модели.

**Задачи проекта:**

* определить характер признаков (числовые, номинативные), а также характер взаимосвязи результата и признака;
* поработать с пропусками и выявить выбросы, понять, как их можно устранить
* выяснить взаимную корреляцию признаков и корреляцию с целевой переменной;
* из всего множества имеющихся признаков отобрать те, которые подойдут для построения модели, предсказывающей целевую переменную.

**Краткая информация о данных:**

Нам дан датасет, в который занесены данные учеников от 15 до 22 лет. На основании этих данных нам нужно выявить те признаки, которые оказывают наибольшее влияние на результаты экзамена по математике.

У нас 29 признаков, которые нужно исследовать, и целевая переменная. Выборка состоит из 395 учащихся.

**Этапы работы над проектом:**

1. Провести первичную обработку данных.
2. Посмотреть на распределение признаков для числовых переменных, устранить выбросы.
3. Оценить количество уникальных значений для номинативных переменных.
4. По необходимости преобразовать данные
5. Провести корреляционный анализ количественных переменных
6. Отобрать не коррелирующие переменные.
7. Проанализировать номинативные переменные и устранить те, которые не влияют на предсказываемую величину (в нашем случае — на переменную score).
8. Сформулировать выводы относительно качества данных и тех переменных, которые будут использованы в дальнейшем построении модели.

**Ответы на вопросы саморефлексии:**

1. Какова была ваша роль в команде? – нас не делили на команды, я сама себе была и капитаном, и исполнителем.

2. Какой частью своей работы вы остались особенно довольны? – Я осталась довольна тем, что большинство данных мне удалось сохранить (я удалила всего 5 строк из датасета, в которых не было значения целевой переменной), изучила методы работы с пропусками и, как мне кажется, достаточно успешно смогла заменить пропущенные значения.

3. Что не получилось сделать так, как хотелось? Над чем ещё стоит поработать? – Мне не хватило работы с числовыми признаками, выявления их корреляции. Также поняла, что нужно дальше прокачивать свои навыки программирования, т.к. были моменты, когда пришлось обращаться за помощью к сокурсникам и ментору в слак.

4. Что интересного и полезного вы узнали в этом модуле? – очень понравилось работать с данными, искать взаимосвязи, заполнять одни признаки через другие. Также для меня стало открытием, что влияние признаков мы итогово оцениваем статистическими методами, а не при помощи корреляции. Думаю, это ценное знание мне обязательно пригодится в дальнейшей работе

5. Что является вашим главным результатом при прохождении этого проекта? – применение полученных знаний на практике, понимание того, как правильно работать с выбросами и пропусками в данных.

6. Какие навыки вы уже можете применить в текущей деятельности? – Графическое отображение данных, работа с пропусками и выбросами, выявление зависимостей в данных.

7. Планируете ли вы дополнительно изучать материалы по теме проекта? – да, подготовка данных, насколько я поняла, одна из важнейших задач Data Scientist, поэтому я хочу глубоко погрузиться в тему. Буду рада рекомендации дополнительных источников по этой теме.