Описание.

Суть проекта — написать модель, которая предсказывала бы результаты госэкзамена по математике для каждого ученика школы. И на ранней стадии выявляла студентов, находящихся в группе риска, путем определения влияния условий жизни учащихся в возрасте от 15 до 22 лет, на их успеваемость по математике.

Чтобы определиться с параметрами будущей модели,проведем разведывательный анализ данных.

Выделим основные шаги:

1)Просмотрим на данные.

2)Проверим данные на пустые значения.

3)Посмотрим на распределение признака для числовых переменных, устраним выбросы.

4)Оценим количество уникальных значений для номинативных переменных.

5)По необходимости преобразуем данные.

6)Проведем корреляционный анализ количественных переменных.

7)Отберем не коррелирующие переменные.

8)Проанализируем номинативные переменные и устраним те,

которые не влияют на предсказываемую величину (в нашем случае — на переменную score).

9)Отберем данные, пригодные для дальнейшего построения модели.

10)При необходимости, повторим цикл, пока модель не заработает как требуется.

В результате Разведывательного анализа данных (EDA),

оценки влияния условий жизни учащихся в возрасте от 15 до 22 лет на их успеваемость

по математике были получены следующие выводы:

1)В данных достаточно много пустых значений.

2)Выбросы найдены только в столбце пропущенных занятий,

что позволяет сделать вывод о том, что данные достаточно чистые.

3)Самые важные параметры, которые предлагается использовать в дальнейшем для построения модели-это mjob,'higher','score'.

Этапы работы над проектом:

Общий обзор датасета, выявление количества пропусков в столбцах.

Просмотр колонок на количество уникальных значений и их количественный состав.

Визуализация данных для лучшего восприятия.

Устранение технических ошибок и приведение данных к общему виду.

Определение диапазонов выбросов.

Заполнение отсутствующих значений, преобразование категориальный признаков через dummy и просмотр результата предсказательной модели. Модель показала посредственные результаты.

Удаление всех строк с отсутствующими данными, преобразование категориальных признаков через dummy и просмотр результата предсказательной модели. Из первоначального количества 395 строк осталось только 97. Модель показала посредственные результаты.

После корреляционного анализа переменных, анализа номинативных переменных и теста Стьюдента было отобрано 2 рабочих признака 'higher','mjob' и рассчитана предсказательная модель с расчетом MAE,MSE и R\_2.

Результат получился лучший из всего вышеперечисленного.

После добавления нормализации результат стал еще лучше.

Во время выполнения проекта были отработаны навыки по ранее пройденным темам.