*EDA – разведывательный анализ данных.*

Миссия данного проекта - помочь UNICEF в работе по повышению 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.

Сделать это предполагается с помощью модели, которая предсказывала бы результаты госэкзамена по математике для каждого ученика школы.

Чтобы определиться с параметрами будущей модели, мы проведём разведывательный анализ данных и оценим результаты этого анализа.

Для анализа у нас есть набор данных, содержащий параметры, описывающие социально-бытовые условия учеников и оценку за экзамен по математике. В этом проекте подготовим данные для нашей модели.

Итак, наш датасет содержит следующие столбцы :

**1 school** — аббревиатура школы, в которой учится ученик

**2 sex** — пол ученика ('F' - женский, 'M' - мужской)

**3 age** — возраст ученика (от 15 до 22)

**4 address** — тип адреса ученика ('U' - городской, 'R' - за городом)

**5 famsize** — размер семьи('LE3' <= 3, 'GT3' >3)

**6 Pstatus** — статус совместного жилья родителей ('T' - живут вместе 'A' - раздельно)

**7 Medu** — образование матери (0 - нет, 1 - 4 класса, 2 - 5-9 классы, 3 - среднее специальное или 11 классов, 4 - высшее)

8 **Fedu** — образование отца (0 - нет, 1 - 4 класса, 2 - 5-9 классы, 3 - среднее специальное или 11 классов, 4 - высшее)

9 **Mjob** — работа матери ('teacher' - учитель, 'health' - сфера здравоохранения, 'services' - гос служба, 'at\_home' - не работает, 'other' - другое)

10 **Fjob** — работа отца ('teacher' - учитель, 'health' - сфера здравоохранения, 'services' - гос служба, 'at\_home' - не работает, 'other' - другое)

11 **reason** — причина выбора школы ('home' - близость к дому, 'reputation' - репутация школы, 'course' - образовательная программа, 'other' - другое)

12 **guardian** — опекун ('mother' - мать, 'father' - отец, 'other' - другое)

13 **traveltime** — время в пути до школы (1 - <15 мин., 2 - 15-30 мин., 3 - 30-60 мин., 4 - >60 мин.)

14 **studytime** — время на учёбу помимо школы в неделю (1 - <2 часов, 2 - 2-5 часов, 3 - 5-10 часов, 4 - >10 часов)

15 **failures** — количество внеучебных неудач (n, если 1<=n<=3, иначе 0)

16 **schoolsup** — дополнительная образовательная поддержка (yes или no)

17 **famsup** — семейная образовательная поддержка (yes или no)

18 **paid** — дополнительные платные занятия по математике (yes или no)

19 **activities** — дополнительные внеучебные занятия (yes или no)

20 **nursery** — посещал детский сад (yes или no)

21 **higher** — хочет получить высшее образование (yes или no)

22 **internet** — наличие интернета дома (yes или no)

23 **romantic** — (yes или no)

24 **famrel** — семейные отношения (от 1 - очень плохо до 5 - очень хорошо)

25 **freetime** — свободное время после школы (от 1 - очень мало до 5 - очень мого)

26 **goout** — проведение времени с друзьями (от 1 - очень мало до 5 - очень много)

27 **health** — текущее состояние здоровья (от 1 - очень плохо до 5 - очень хорошо)

28 **absences** — количество пропущенных занятий

29 **score** — баллы по госэкзамену по математике

Достаточно много параметров, подавляющее большинство которых – ранговые.

По ходу работы построили тепловую карту пропусков в данных. Из карты видно, что количество пропусков достаточно велико. Большинство моделей не предполагают пропуски в данных, поэтому мы заменим их (пропуски) на значение моды для номинативных и на медианное значение для метрических переменных.

При выполнении проекта мы исследуем каждый столбец, избавимся от выбросов, устраним пропуски в данных, затем проведем корреляционный анализ метрических переменных и тест Стьюдента для номинативных переменных. В результате мы должна получить датасет, который содержит только переменные, влияющие на оценку по математике, без пропусков в данных и не содержащие выбросов.

После выполнение вышеописанных этапов ,в результате EDA для прогнозирования влияния социально-бытовых условий учеников на количество баллов, набранных на экзамене по математике были получены следующие выводы:

В данных достаточно много пропусков, только в столбцах **school**(аббревиатура школы, в которой учится ученик ),**sex**(пол ученика) и **age**(возраст ученика) не было пропусков данных. Все пропуски заменялись на моду, в случае номинативных и на медианное значение, в случае метрических переменных.

Выбросов, требующих коррекции данных, было немного.

В столбце **fedu**(образование отца) – выброс вверх - явная опечатка.

В столбце **famrel**(семейные отношения) – выброс вниз удален.

В столбцах ранговых переменных **fedu**(образование отца), **traveltime**(время в пути до школы), **studytime** (время на учёбу помимо школы в неделю),**failures** (количество внеучебных неудач), **famrel**(семейные отношения),**freetime**(свободное время после школы), **goout**(проведение времени с друзьями), **health**(текущее состояние здоровья) используются цифровые значения.

В таких столбцах значения, которые методом IQR определяются как выбросы, входили в диапазон корректных значений. В таких столбцах мы не изменяли данные.

Положительная корреляция мажду **age** и **absence** может означать, что с увеличением возраста ученики пропускают больше уроков математики, а отрицательная между **age** и **score** , что чем старше ученик, тем хуже он сдает экзамен по математике.

Тест Стьюдента показал, что самые важные параметры, которые предлагается использовать в дальнейшем для построения модели, это **address,medu,mjob,failures,higher,romantic,age и absences.**

Команду , которая собирала данные в школе GP, следует попросить обратить внимание на качество своей работы ☺ .

1. Какова была ваша роль в команде?  
     
   **На данном этапе роль DS трудно переоценить. От того, насколько качественно подготовлены данные для модели, напрямую зависит конечный результат всей команды. Большая ответственность. Максимум профессионализма и внимательности к деталям ☺**
2. Какой частью своей работы вы остались особенно довольны?  
    **Работа с библиотеками PYTHON для анализа данных – это увлекательно .**
3. Что не получилось сделать так, как хотелось? Над чем ещё стоит поработать?  
      
   **На мой взгляд, надо больше уделить внимания Boxplot' ам. Навыки быстрой интерпретации   
    необходимо совершенствовать. Разобраться в том, где PYTHON создает копию датасета,  
    а где изменяет текущий тоже было бы очень полезно.**
4. Что интересного и полезного вы узнали в этом модуле?  
     
   **Полезного – масса. Применение PYTHON для разведывательного анализа данных – это просто и эффективно, много полезных мелочей. Heatmap,например, – это очень наглядно.Разъяснения как устранять пропуски в данных не удаляя их – очень ценны для меня. Удалить просто. Впоследствии получить кривую модель – ещё проще.**
5. Что является вашим главным результатом при прохождении этого проекта?  
     
   **Главным результатом после выполнения проекта считаю отсутствие растерянности при   
   подготовке данных для моделирования. Появился четкий пошаговый план, минимальные инструменты и понимание, что я могу подготовить данные и провести их первичный анализ. Понятно, что в этом деле нужно ещё сильно расти, читать, изучать чужой опыт, однако первый шаг сделан.**
6. Какие навыки вы уже можете применить в текущей деятельности?

**Моя сегодняшняя деятельность, к сожалению, далека от DS, а возможно я просто не вижу возможности для применения полученных знаний и навыков, в силу неопытности. Ведь в жизни всегда есть место Data Science ! ☺**

1. Планируете ли вы дополнительно изучать материалы по теме проекта?

**В обязательном порядке.**