Отчет о выполнении проекта EDA.

1. **Цели и задачи проекта.**

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.

Целью отчетного этапа проекта является подготовка с помощью разведывательного анализа данных предложенного датасета к построению предсказательной и рекомендательной модели.

Данный этап проекта предполагает решение следующих задач:

- первичная обработка данных;

- устранение выбросов;

- анализ номиативных и количественных переменных;

- проведение корреляционного анализа для номиативных и количественных переменных;

- устранение данных не влияющих на предсказываемую величину.

1. **Краткое описание данных.**

Исследуемых датасет состоит из следующих переменных:

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 4)

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 — в романтических отношениях (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

1. **Основные этапы работы над проектом.**
   1. Первичная обработка данных. Данный этап предполагает изучение типов переменных , пропусков, отрицательных и нулевых значений предложенного датасета.
   2. Устранение выбросов для соответствующих переменных.
   3. Оценка частотного распределения номиативных переменных.
   4. Коррелиационный анализ числовых переменных.
   5. Коррелиационный анализ номиативных переменных.
   6. Определение выводов на основании проведенного анализа и в соответствии с поставленными целями и задачами.
2. **Полученные знания в результате выполнения проекта.**

Выполнение данного проекта потребовало погружение в следующие дисциплины:

- мат. статистика;

- методы работы Python с графическими классами, с функциями статистического анализа.

При исполнении проекта были получены новые знания в указанных областях, позволяющие применять на практике разведывательного анализа данных следующие методы работы:

- подготовка датасета к коррелиационному анализу;

- проведение коррелиационного анализа для числовых и номиативных переменных с помощью методов графических функций и T распределения Стьюдента.

1. **Качественный анализ данных с использованием Boxplot.**
   1. Таблица корреляции score без сегментации по доп. признаку.

|  |  |  |
| --- | --- | --- |
| Столбец | Степень корреляции | Комментарий |
| Medu | Слабая | Заметен повышающий тренд при росте показателя образования матери |
| Higher | Сильная | Оценки заметно выше у мотивированных к получению образования учеников |
| goout | Слабая | Крайний показатель показывает заметное снижение успеваемости учеников |

Остальные признаки, на мой взгляд, в данном случае не выявляют очевидной корреляции. Перейдем к анализу корреляции score с сегментацией по доп. признаку .

* 1. Выберем столбцы для сегментации при построении Boxplot:

famsize, Pstatus , Mjob , Fjob , studytime, paid , activities, higher , romantic , famrel , goout , health, absences

Таблица корреляции score с сегментацией по доп. признаку.

|  |  |  |  |
| --- | --- | --- | --- |
| Столбец | Признак сегментации(hue) | Количество коррелируемых признаков | Комментарий(учитывается логика возможной взаимосвязи событий) |
| Traveltime, Medu,Mjob,Fjob, studytime, Higher, famrel , freetime , goout | romantic | 9 | Замечена четкая тенденция усиления корреляции 9 признаков со score в случае влюбленности ученика |
| guardian | famsize | 1 | Замечена не значительная тенденция усиления корреляции 1 признака со score |
| Sex, Mjob, Fjob, freetime | Pstatus | 4 | Замечена не значительная тенденция усиления корреляции 4 признаков со score в неполной семьи |
| Medu, Pstatus, health | Mjob | 3 | Замечена заметная тенденция усиления корреляции 3 признаков со score в случае если статус работы матери – athome |
| - | Fjob | 0 | Какой либо признак усиливающий тенденцию к корреляции не найден |
| Traveltime | studytime | 1 | Слабая тенденция к усилению корреляции |
| - | paid | 0 | Какой либо признак усиливающий тенденцию к корреляции не найден |
| - | activities | 0 | Какой либо признак усиливающий тенденцию к корреляции не найден |
| Почти все признаки | higher | Более 10 | Замечена четкая тенденция усиления корреляции более 10 признаков со score в случае не желания ученика получать ВО |
| Medu, Fedu, Mjob, Fjob, guardian, freetime | famrel | 6 | Замечена четкая тенденция усиления корреляции 6 признаков со score |
| Medu, Fedu, Mjob, Fjob, | goout | 4 | Замечена четкая тенденция усиления корреляции 4 признаков со score |
| Mjob, Fjob,  Guardian, studytime | health | 4 | Замечена четкая тенденция усиления корреляции 4 признаков со score |

* 1. **Выводы**

Анализ графики Boxplot с использованием сегментации раскрывает очевидные тенденции к усилению корреляцию в ряде случаев. Таким образом оценка данной графики позволяет с уверенностью говорить о корреляции со score следующих признаков:

Medu, Fedu, Mjob, Fjob, studytime, Guardian, studytime, freetime, Higher, famrel , freetime , goout, health, romantic, Pstatus . Признаки не вошедшие в указанный список корреляцию не приносят и их можно исключить.

Найдены статистически значимые различия для колонки company

Найдены статистически значимые различия для колонки specific\_origin

Найдены статистически значимые различия для колонки company\_location

Как мы видим, серьёзно отличаются три параметра: company,  specific\_origin и  company\_location. Оставим эти переменные в датасете для дальнейшего построения модели. Итак, в нашем случае важные переменные, которые, возможно, оказывают влияние на оценку, это: company, specific\_origin, company\_location, review\_date, cocoa\_percent.