Анализ успеваемости студентов

Кокорев Артём БПИ235

Поиск данных

Поиск данных

https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

Student Performance Factors

Что внутри?

6607 строк и 20 значений в каждой

Hours_Studied - Количество часов, потраченных на обучение (int)
Attendance - Процент посещенных занятий (int)
Parental_Involvement - Уровень участия родителей в образовании студента (Low, Medium, High)
Access_to_Resources - Доступность образовательных ресурсов (Low, Medium, High)
Extracurricular_Activities - Участие в дополнительных занятиях (Yes, No)
Sleep_Hours - Среднее количество часов сна за ночь (int)
Previous_Scores - Средние оценки предыдущих экзаменов (int)
Motivation_Level - Уровень мотивации студента (Low, Medium, High)
Internet_Access - Доступ в интернет (Yes, No)
Tutoring_Sessions - Количество занятий с репетитором в месяц (int) метрическая
Family_Income - Уровень дохода семьи (Low, Medium, High)
Teacher_Quality - Качество преподавания учителями (Low, Medium, High), 78 пропусков
School_Type - Тип школы (Public, Private)
Peer_Influence - Влияние сверстников на академическую жизнь (Positive, Neutral, Negative)
Physical_Activity - Среднее количество часов физической активности в неделю (int)
Learning_Disabilities - Наличие проблем с обучаемостью (Yes, No)
Parental_Education_Level - Образование родителей (High School, College, Postgraduate), 90 пропусков
Distance_from_Home - Расстояние от дома до школы (Near, Moderate, Far), 67 пропусков
Gender - пол студента (Male, Female).
Exam_Score - Оценка за экзамен (int)

Цель

Выявить ключевые факторы, влияющие на академическую успеваемость студентов, и определить их относительную важность.

Задачи:

1. Определить влияние временных затрат на обучение на итоговые результаты

2. Выявить роль социально-экономических факторов в успеваемости

3. Оценить влияние качества преподавания и доступности ресурсов

4. Исследовать взаимосвязь между физической активностью и академическими результатами

5. Определить влияние родительского участия на успеваемость

Гипотезы

H1: Существует положительная корреляция между количеством часов обучения и итоговой оценкой

H2: Существует положительная корреляция между уровнем родительского участия и итоговой оценкой

H3: Существует положительная корреляция между доступностью образовательных ресурсов и итоговой оценкой

H4: Существует разница в успеваемости между государственными и частными школами

H5: Существует положительная корреляция между количеством часов сна и итоговой оценкой

H6: Существует положительная корреляция между количеством занятий с репетитором и итоговой оценкой

H7: Студенты с хорошими оценками в прошлом экзамене показывают лучшие результаты

H8: Существует корреляция между влиянием сверстников и итоговой оценкой

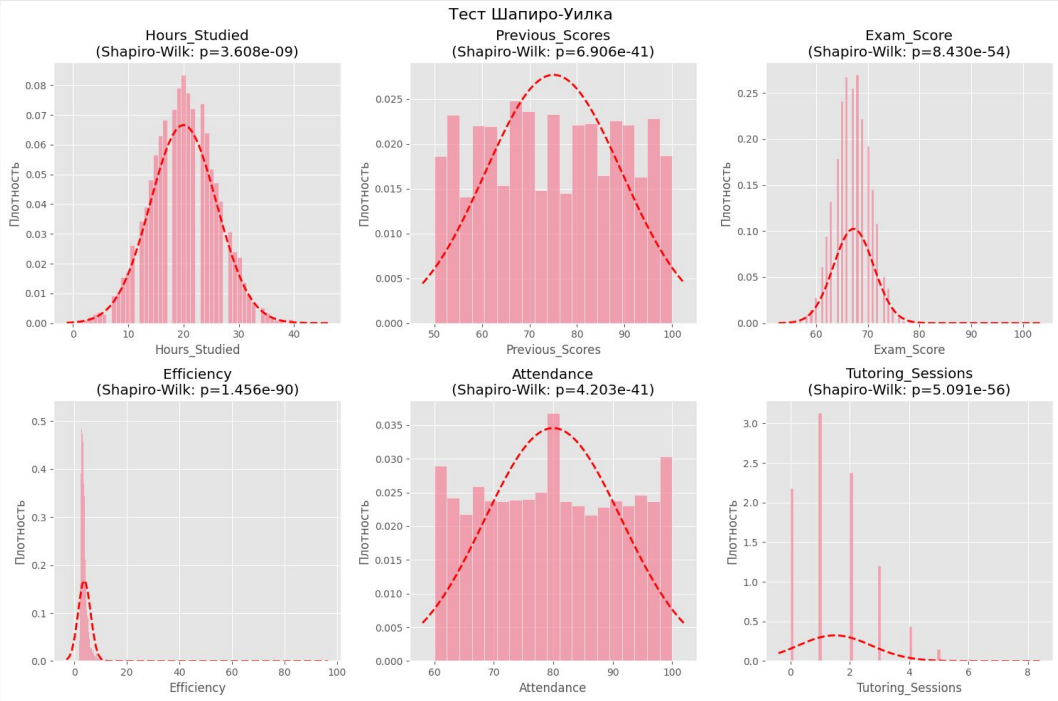
H9: Существует корреляция между родительским образованием и итоговой оценкой

H10: Существует корреляция между уровнем мотивации и итоговой оценкой

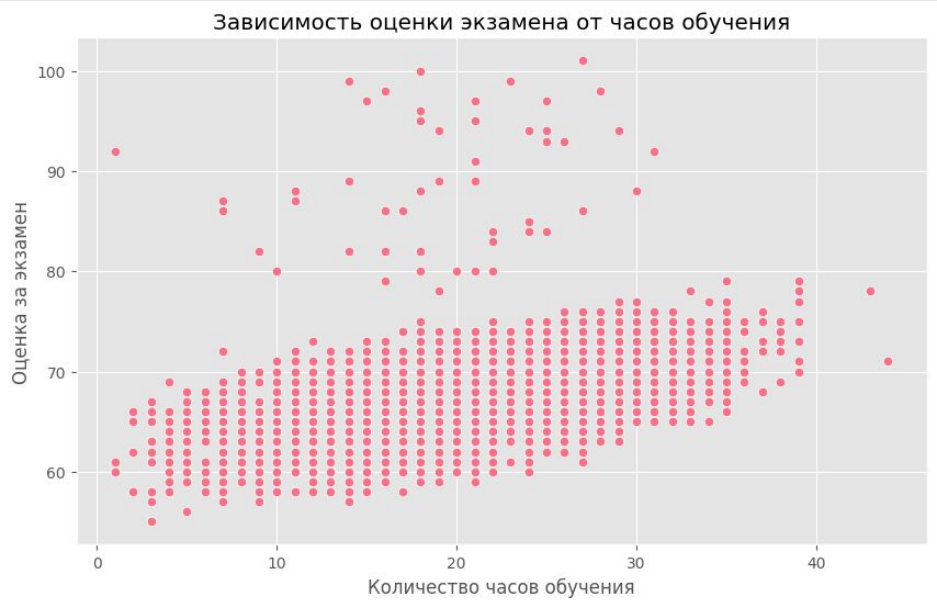
Ps Везде я буду рассматривать влияние на оценку за экзамен, поэтому зависимая переменная будет метрической. Если вторая переменная номинальная, то я буду использовать однофакторный дисперсионный анализ (ANOVA), если бинарная, то t-тест для независимых выборок. Если метрическая, то корреляционный анализ Пирсона.

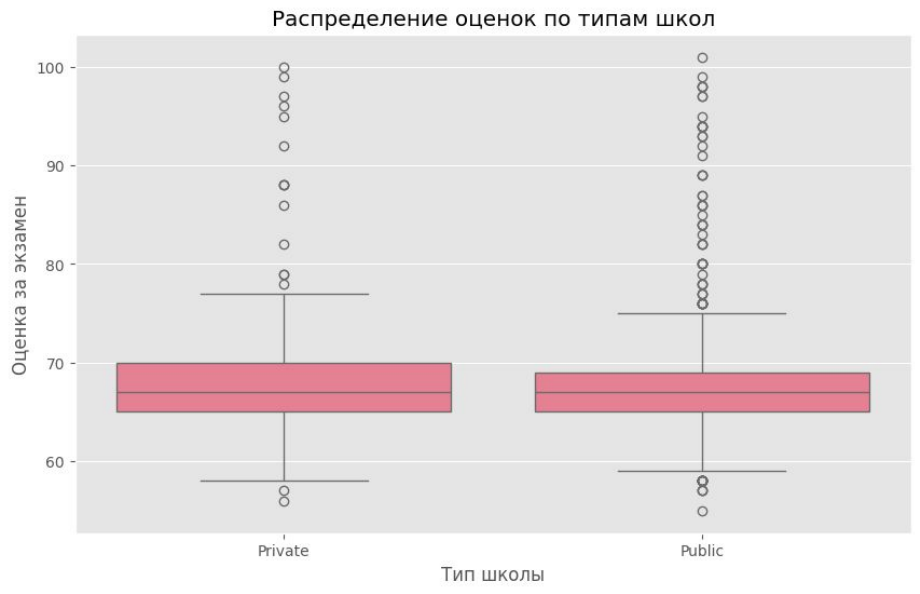
Тест Шапиро-Уилка

Ни одного нормального распределения нет ни визуально ни по коэффициентам, поэтому нужно использовать коэффициенты корреляции Кендалла(не Спирмена, т.к. много повторов и присутствуют выбросы)

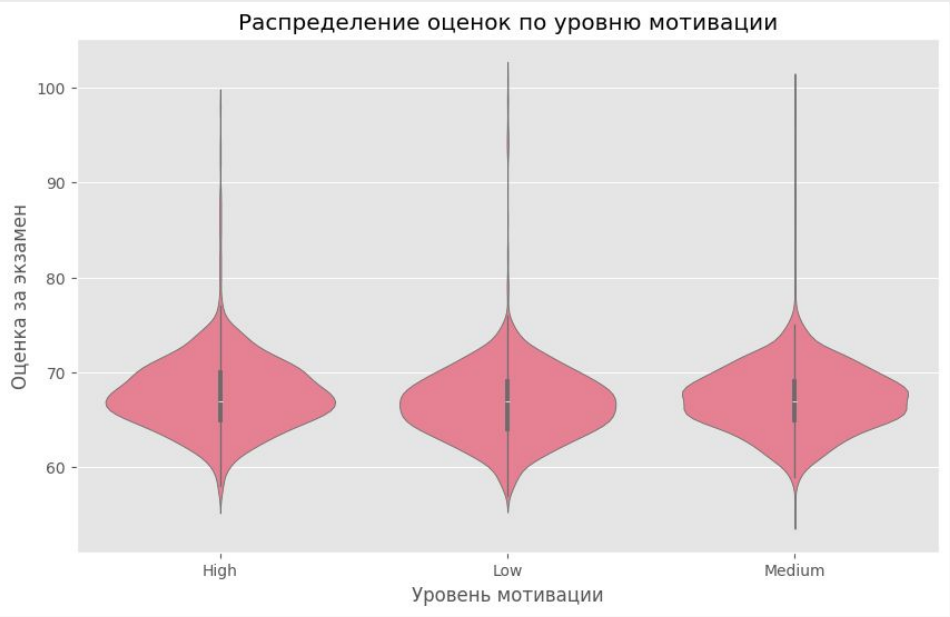


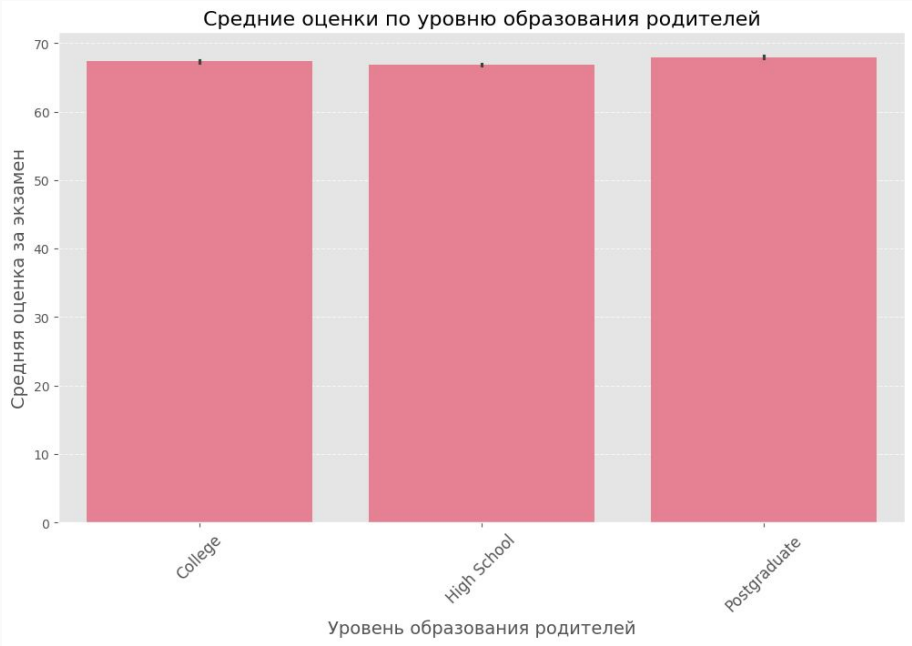
Интересные графики





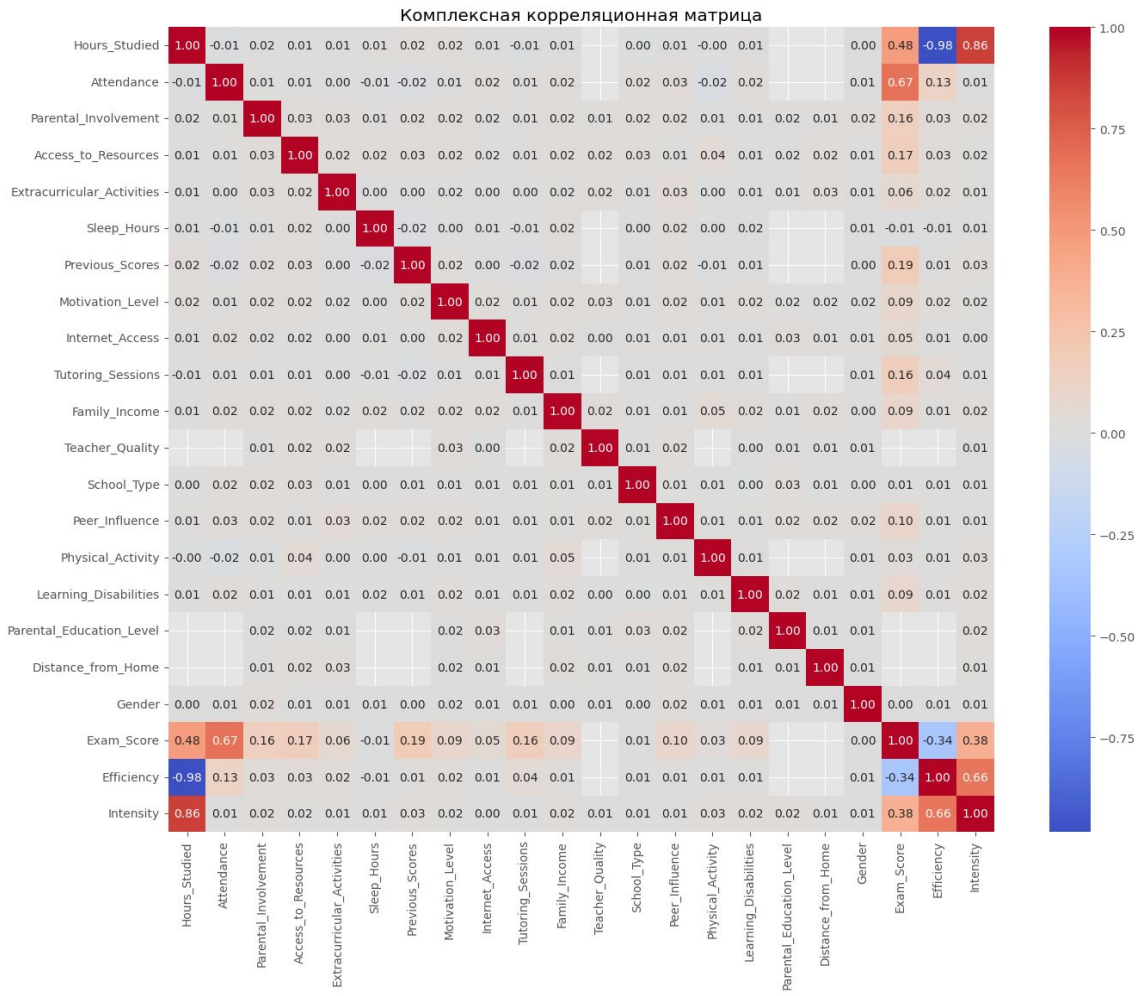
Интересные графики





Корреляция

Видно, что если не брать в расчет составные переменные, самые большие корреляции всё так же у часов обучения и посещаемости, а категориальные переменные между собой практически не коррелируют.



Линейная регрессия

Будем предсказывать успех на экзамене в баллах

Коэффициенты модели:
Hours_Studied: 0.2891
Attendance: 0.1988
Previous_Scores: 0.0483
Tutoring_Sessions: 0.5102
Physical_Activity: 0.1507

Коэффициент детерминации (R^2): 0.6422
RMSE: 2.2488

Часы обучения: за каждый час обучения оценка за экзамен увеличивается на 0.2891 балла. Сильная положительная зависимость.
Посещаемость: за каждый процент посещаемости оценка за экзамен увеличивается на 0.1988 балла. Сильная положительная зависимость.
Прошлые оценки: за каждый балл средней прошлой оценки оценка за экзамен увеличивается на 0.04383 балла. Слабая положительная зависимость.
Количество занятий с репетитором: за каждое занятие оценка за экзамен увеличивается на 0.5101 балла. Сильная положительная зависимость.
Физическая активность: за каждый час физической активности оценка за экзамен увеличивается на 0.1507 балла. Сильная положительная зависимость.

R^2 = 0.6422 - 64% предсказывается моделью, неплохое значение.
RMSE = 2.2488 - среднее отклонение предсказаний от фактических значений.

Бинарная регрессия

Будем предсказывать шанс студента стать отличником и набрать 70+ баллов

Коэффициенты модели:
Hours_Studied: 0.3005
Attendance: 0.1982
Previous_Scores: 0.0492
Tutoring_Sessions: 0.5218
Physical_Activity: 0.1552

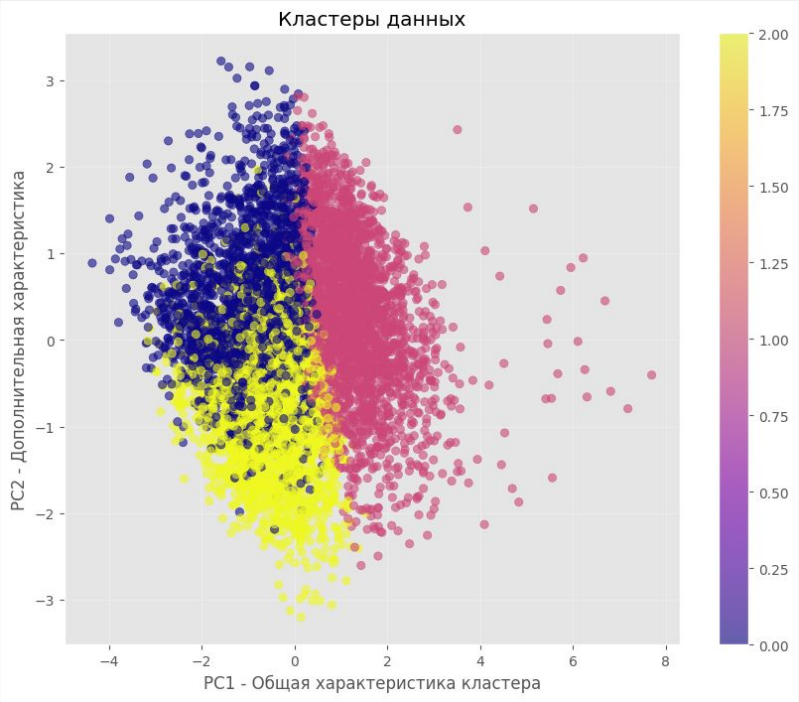
Метрики качества модели:

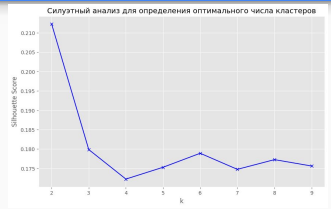
	precision	recall	f1-score	support
0	0.93	0.96	0.95	1090
1	0.79	0.66	0.72	232
accuracy			0.91	1322

Каждый из предикатов положительно влияет на вероятность стать отличником. Также репетиторство и посещаемость имеют больший вес, чем другие предикаты.

У модели хорошая точность 0.91. Модель хорошо предсказывает более успешных студентов, это должно быть связано со спецификой данных, в которых преобладают более высокие оценки.

Кластеры





У меня вышло три кластера:

0 кластер - бездельники с низким посещением, временем на учёбу и плохими старыми оценками, но высокой физической активностью.
1 кластер - отличники, посещающие занятия, изучающие материал и имеющие хорошие оценки.
2 кластер - скатывающиеся вниз студенты, с низким посещением, временем на учёбу, но высокими старыми оценками.

Вернемся к гипотезам

H1: количество часов обучения влияет на оценку за экзамен как главный предикат в обеих регрессиях.

H2: родительское участие в образовании практически не влияет на оценку за экзамен, это было доказано в графиках и сводных таблицах.

H3: влияния доступности образовательных ресурсов не было выявлено.

H4: разницы между государственными и частными школами практически нет(десятые доли), это было доказано в графиках и сводных таблицах.

H5: влияния количества сна не было выявлено.

H6: количество занятий с репетитором очень сильно влияет на оценку за экзамен, это был главный предикат в регрессиях.

H7: влияние прошлых оценок не было выявлено.

H8: влияние сверстников очень слабо влияет на оценку за экзамен, сводные таблицы дали разницу чуть больше 1 процента.

H9: родительское образование не влияет на оценку за экзамен, это было доказано в графиках.

H10: мотивация практически не влияет на оценку за экзамен, это было доказано в графиках.

Выводы

Цель исследования явно достигнута, регрессионные модели показывают хорошие результаты и подсвечивают главные предикаты.

Мне показалось неожиданным, что практически ничего кроме посещения и услуг репетиторства глобально не влияет на оценку за экзамен.

Все социально экономические факторы вместе влияют меньше, чем репетиторство.

Отсутствие влияния родительского участия и образования меня очень удивило, но я думаю эти факторы сопряжены с использованием услуг репетиторства.

Анализ успеваемости студентов

Кокорев Артём БПИ235