Анализ привычек студентов и их академической успеваемости

Смирнов Д. В.

Мехманико-математический факультет
Московский государственный университет

2025

Введение

В проекте будет проводиться анализ датасета https://www.kaggle.com/datasets/aryan208/student-habits-and-academic-performance-dataset
Будет сделан разведочный анализ данных (**EDA**). Будут построены и оценены три модели:

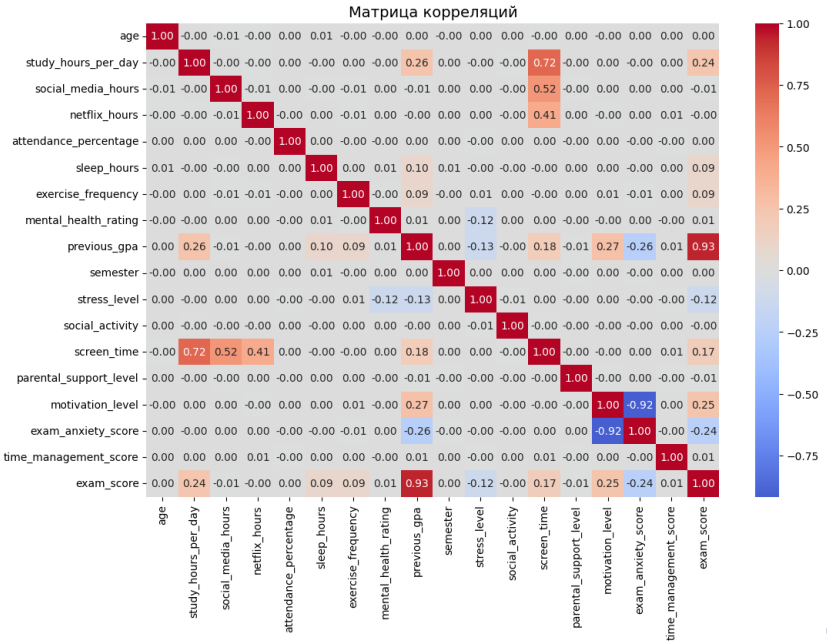
- Линейная регрессия
- k ближайших соседей(kNN)
- Случайный лес(Random Forest)

Оцениваться модели будут с помощью 3 метрик:

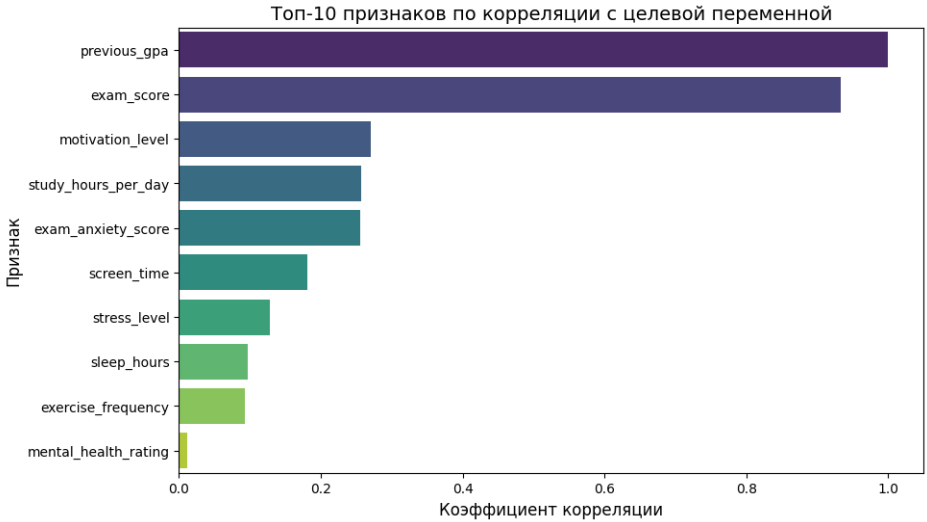
- *MSE* - среднеквадратичное отклонение
- MAE - среднее абсолютное отклонение
- R^2 - коэффициент детерминации

EDA

В датасете всего 31 признак, причём пропущенных значений нет. Есть целевая переменная **previous_gpa** - предыдущий средний балл. Посмотрим на корреляцию числовых признаков:



А теперь на топ-10 признаков по корреляции с целевой переменной:

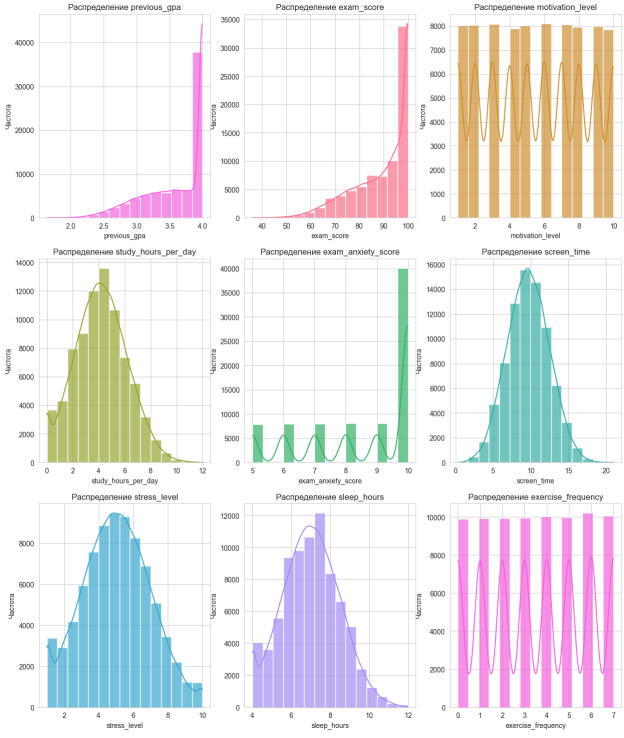


Берём до признака "ментальное здоровье"

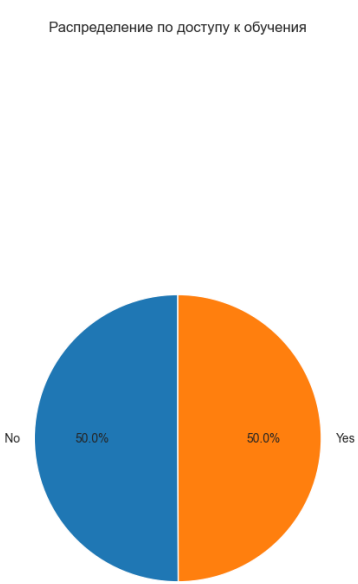
Проделаем то же самое с категориальными(кодировали с помощью LabelEncoder):

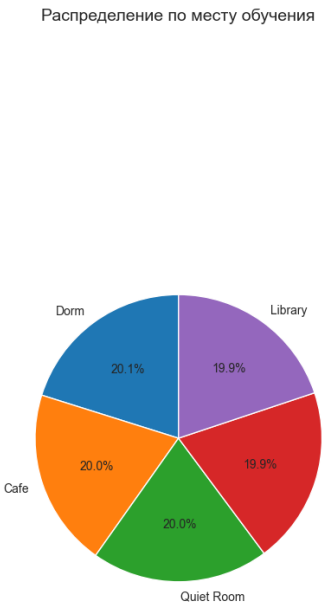
Берём из них: доступ к обучению, обычное место учёбы и риск отчисления

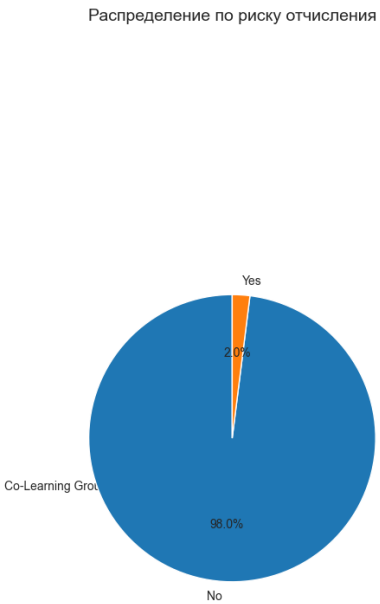
Давайте посмотрим на распределения отобранных признаков, сначала числовые:



Теперь категориальные:







Линейная регрессия

На входе есть датасет (y_i, x_i), где y_i - таргет, а x_i - d-мерный вектор столбец признаков. Модель линейной регрессии:

$$y_i = x_i^1 w_1 + \ldots + x_i^d w_d + w_0 = w^T x_i + w_0$$

Нам нужно предложить какую-то оценку для весов, т.е. для w. Есть хорошая оценка $w = (X^T X)^{-1} X^T y$. В классической регрессии эта оценка являеся наилучшей несмещённой оценкой в классе всех линейных несмещённых оценок(Best unbiased linear estimation). Если предположить, что $w_0 \sim N_d(0, \sigma^2 E)$, то оценка будет оптимальной. В проекте использовалась именно эта модель, без модернизаций в виде регуляризаций, т.к. оказалось достаточно самой базовой линейной регрессии.

Модель показала следующие результаты:

- $MAE = 0.11996900445550267$
- $MSE = 0.026550070321346795$
- $R^2 = 0.8778533376779928$

Для новой точки данных, т.е. для предсказания, алгоритм вычисляет расстояние до всех элементов обучающей выборки, берёт "ближайшие" k соседей и усредняет их: $\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$

Есть вариация взешенного kNN с учётом расстояний: $\hat{y} = \frac{\sum_{i=1}^{k} w_i v_i}{\sum_{i=1}^{k} w_i}$, где $w_i = \frac{1}{d(x,x_i)^2}$ или $w_i = \frac{1}{d(x,x_i)}$. k, d, формула для весов - гиперпараметры модели. В проекте использовалась такая модель, гиперпараметры искались с помощью GridSearch

Модель показала следующие результаты:

- $MAE = 0.13770174959616205$
- $MSE = 0.0331005047452342$
- $R^2 = 0.8477173082079044$

Похуже, чем в линейной регрессии. Но можно лучше, если более точно подобрать гиперпараметры, но, учитывая размер датасета, это совсем не быстро.

Случайный лес(Random Forest)

Алгоритм создаёт M новых обучающих наборов путем выбор с возвращением из N имеющихся наблюдений. На каждом узле дерева смотрим только на случайное подможество признаков($max\_features$). Вычисляется MSE текущего узла, после строится разделение с заданным порогом, вычисляется MSE после разделения и сравнивается с MSE до разделения(выигрыш от разделения). Этот процесс повторяется для каждого из M деревьев с усреднением всех итоговых значений. Модель показала следующие результаты:

- $MAE = 0.1163445245998015$
- $MSE = 0.025033270105707814$
- $R^2 = 0.8848315521048163$

Немного получше чем в линейной регрессии, но скорость работы гораздо меньше, +подбор гиперпараметров.

Заключение

Лучшая модель - RandomForest. Но если важна быстрота, то лучше использовать линейную регрессию