Анализ привычек студентов и их академической успеваемости

Смирнов Д. В.

Мехманико-математический факультет
Московский государственный университет

2025

Введение

В проекте будет проводиться анализ датасета:
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)

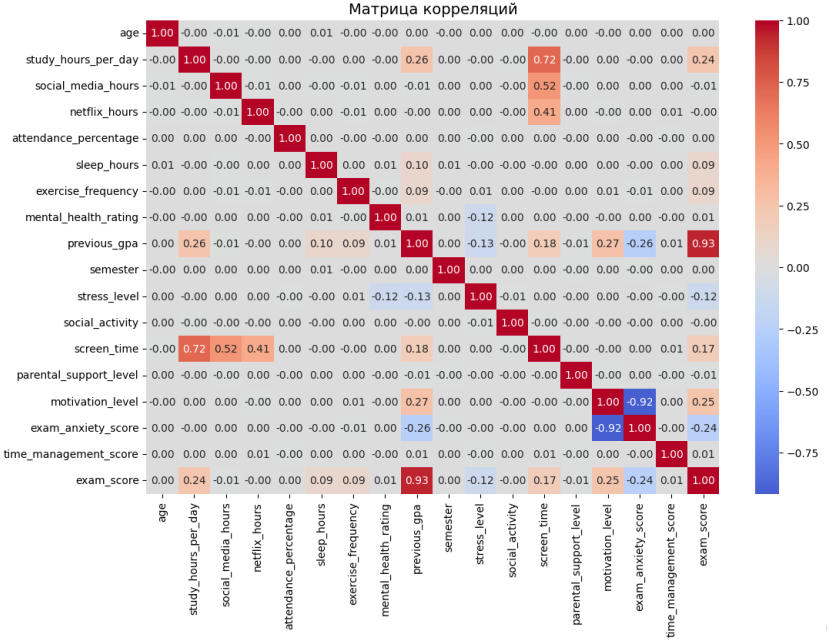
Оцениваться модели будут с помощью 3 метрик:

- $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$-среднеквадратичное отклонение
- $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$ - среднее абсолютное отклонение
- $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$ - коэффициент детерминации

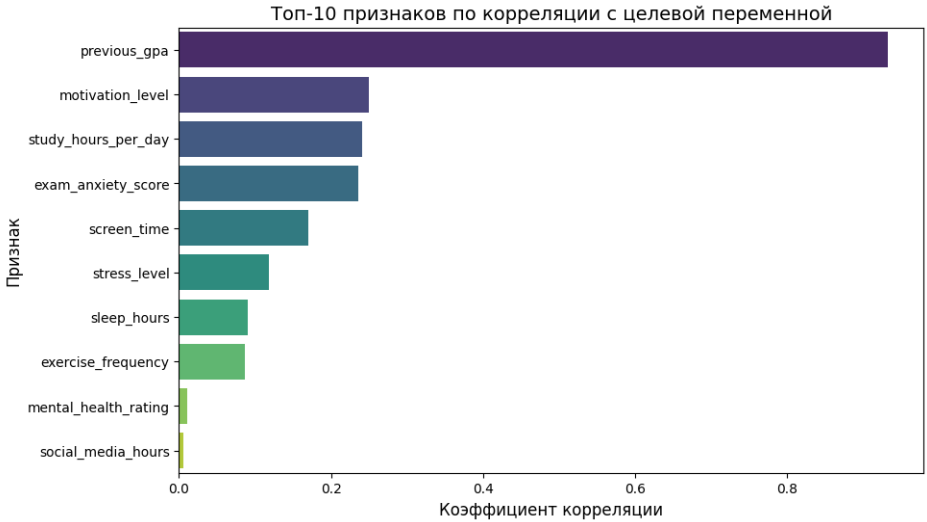
$y_i, \hat{y}_i, \bar{y}$ - истинное значение, предсказанное значение, среднее целевой переменной
N - количество наблюдений в тестовой выборке

EDA

Размер датасета 80000×31, причём пропущенных значений нет. Есть целевая переменная **exam_score** - балл за экзамен. Посмотрим на корреляцию числовых признаков:



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

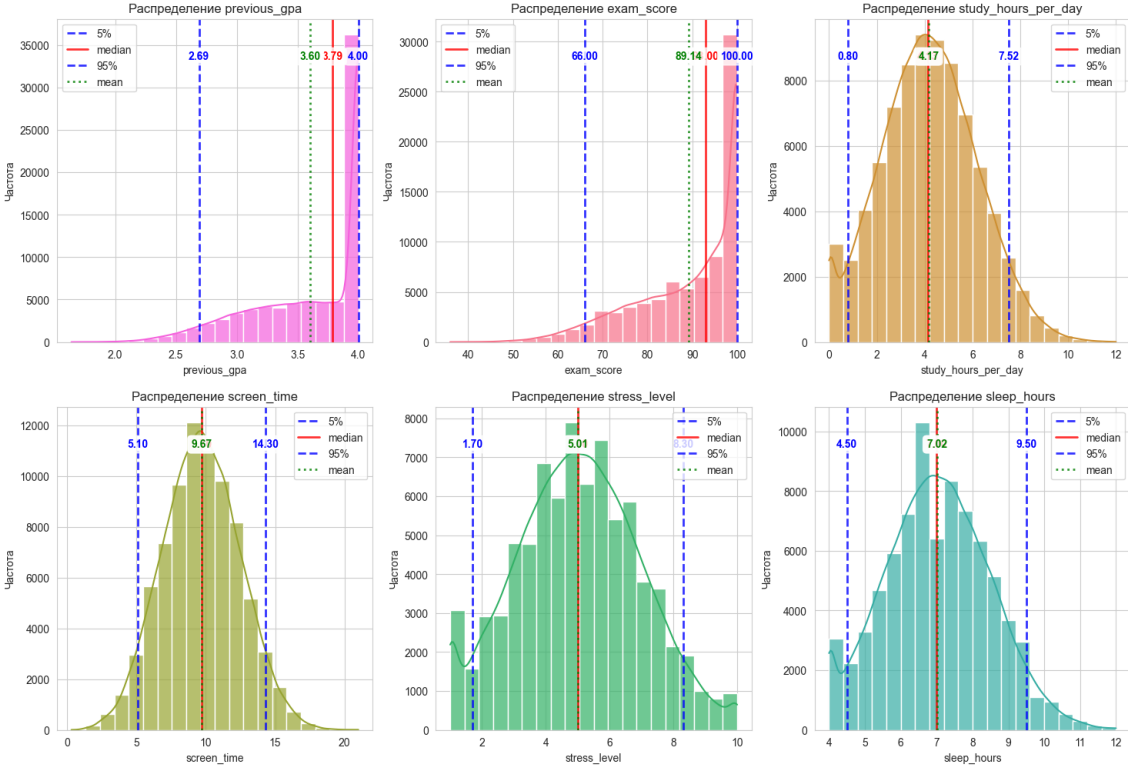


Берём до признака "ментальное здоровье"

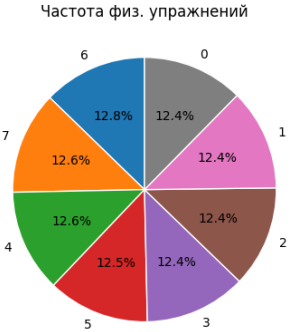
Проделаем то же самое с категориальными(кодировали с помощью LabelEncoder):

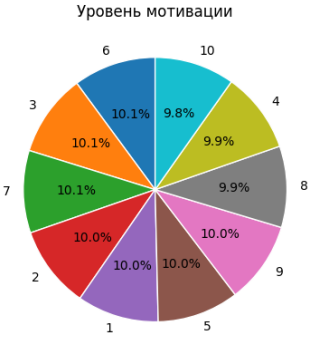
Берём из них: доступ к обучению, обычное место учёбы и риск отчисления

Распределение числовых признаков



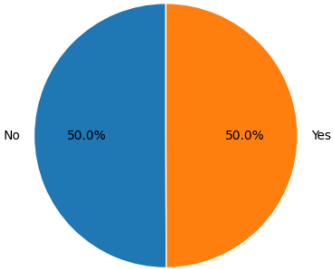

Тревожность на экзамене
5
7
6
8
9
10
9.8%
9.9%
10.0%
10.1%
10.1%
50.0%


Частота физ. упражнений
0
1
2
3
4
5
6
7
12.4%
12.4%
12.4%
12.4%
12.5%
12.6%
12.6%
12.8%

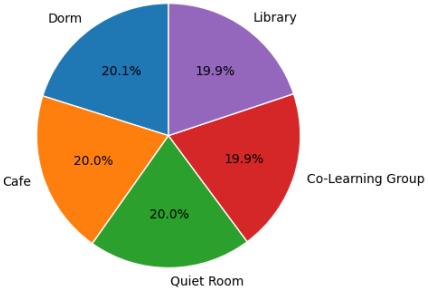

Уровень мотивации
10
4
8
9
5
1
2
7
3
6
9.8%
9.9%
9.9%
10.0%
10.0%
10.0%
10.1%
10.1%
10.1%

Распределение категориальных признаков

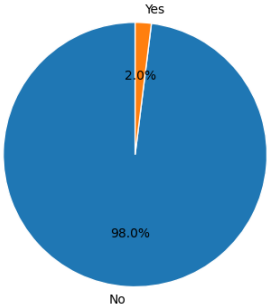
Распределение по доступу к обучения



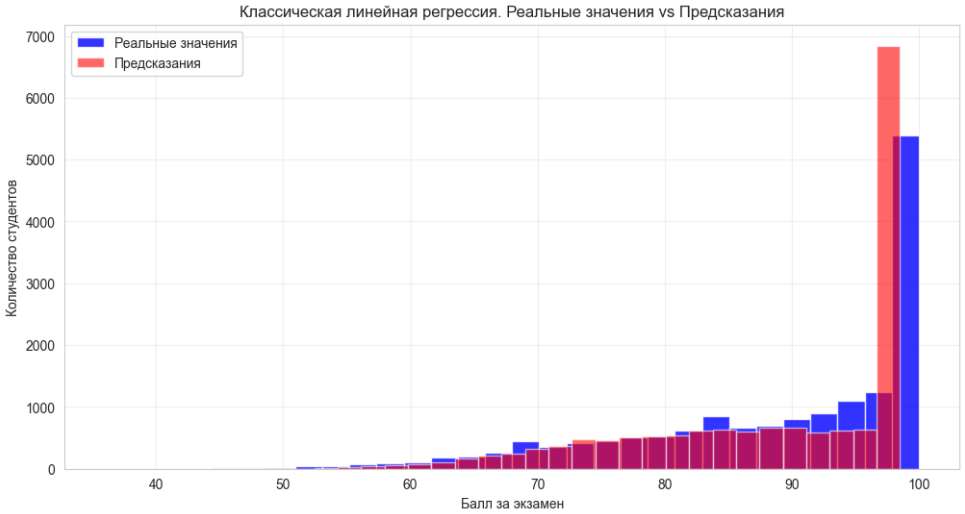
Распределение по месту обучения



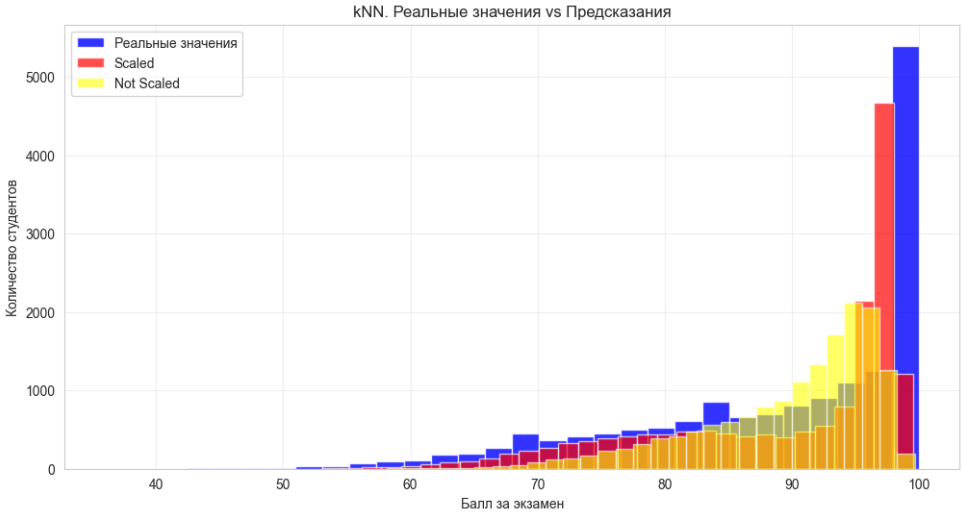
Распределение по риску отчисления



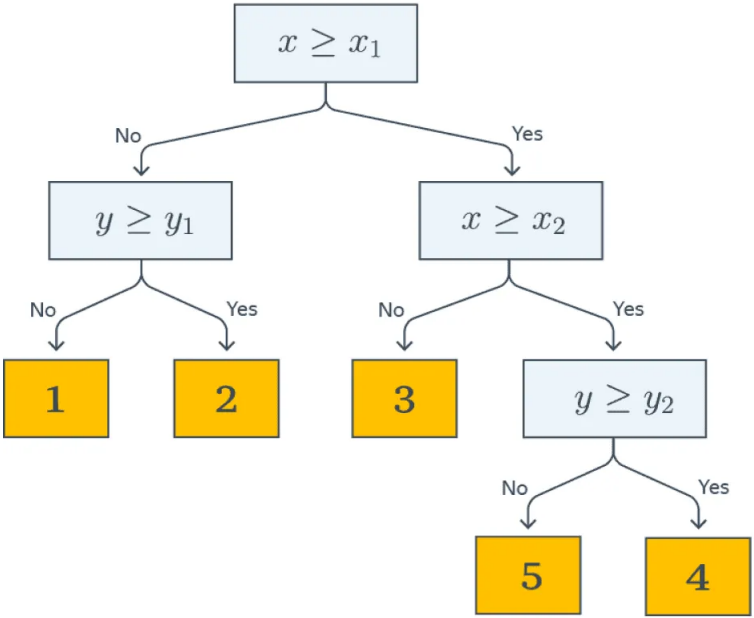
Сравнение качества моделей				
Модель	Scaled	*MAE*	*MSE*	R^2
Linear Regression	Yes	3.1919	17.3130	0.8714
	Not	3.1919	17.3130	0.8714
L1-Regularization	Yes	3.1926	17.3138	0.8714
	Not	3.1944	17.3143	0.8714
L2-regularization	Yes	3.1926	17.3138	0.8714
	Not	3.1944	17.3143	0.8714


Классическая линейная регрессия. Реальные значения vs Предсказания
Реальные значения
Предсказания
Количество студентов
Балл за экзамен
0
1000
2000
3000
4000
5000
6000
7000
40
50
60
70
80
90
100

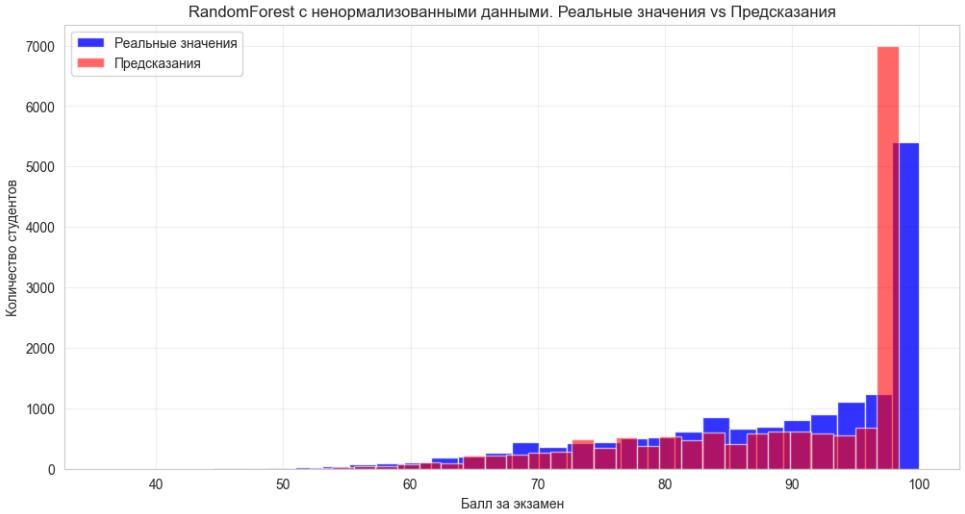
Сравнение качества моделей				
Модель	Scaled	*MAE*	*MSE*	R^2
kNN	Yes	3.6	21.5	0.84
	Not	5.7	52.3	0.61


kNN. Реальные значения vs Предсказания
Реальные значения
Scaled
Not Scaled
Количество студентов
0
1000
2000
3000
4000
5000
40
50
60
70
80
90
100
Балл за экзамен

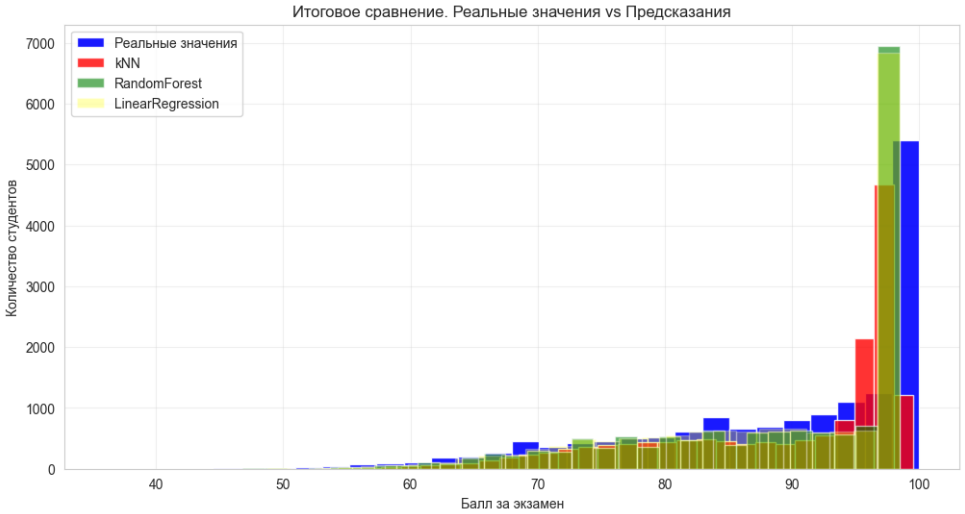
Решающее дерево

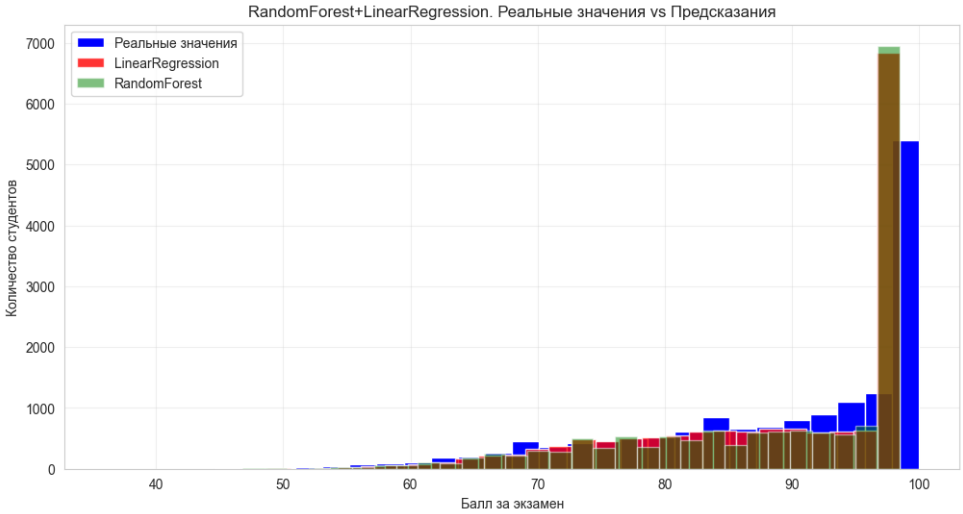


Сравнение качества моделей				
Модель	Scaled	*MAE*	*MSE*	R^2
RandomForest	Yes	3.22	17.1	0.8729
	Not	3.221	17.09	0.873


RandomForest с ненормализованными данными. Реальные значения vs Предсказания
Реальные значения
Предсказания
Количество студентов
0
1000
2000
3000
4000
5000
6000
7000
40
50
60
70
80
90
100
Балл за экзамен

Заключение


Итоговое сравнение. Реальные значения vs Предсказания
Реальные значения
kNN
RandomForest
LinearRegression
Количество студентов
0
1000
2000
3000
4000
5000
6000
7000
40
50
60
70
80
90
100
Балл за экзамен


RandomForest+LinearRegression. Реальные значения vs Предсказания
Реальные значения
LinearRegression
RandomForest
Количество студентов
Балл за экзамен
0
1000
2000
3000
4000
5000
6000
7000
40
50
60
70
80
90
100

Сравнение качества моделей			
Модель	*MAE*	*MSE*	R^2
LinearRegression	3.1919	17.3130	0.8714
kNN	3.6	21.5	0.84
RandomForest	3.2214	17.1	0.873