



Задача классификации

Логистическая регрессия

Подготовил: студент группы М8О-307Б-23
Бельский Г. Б.

Почему логистическая регрессия?

Подходит для объяснимых и реальных задач: чётко показывает вклад каждого признака, легко интерпретируется и устойчива к мультиколлинеарности, особенно с регуляризацией.

Прекрасно работает на инженерных и табличных данных со смешанными признаками — как в моем датасете

Легко масштабируется и прост в настройке

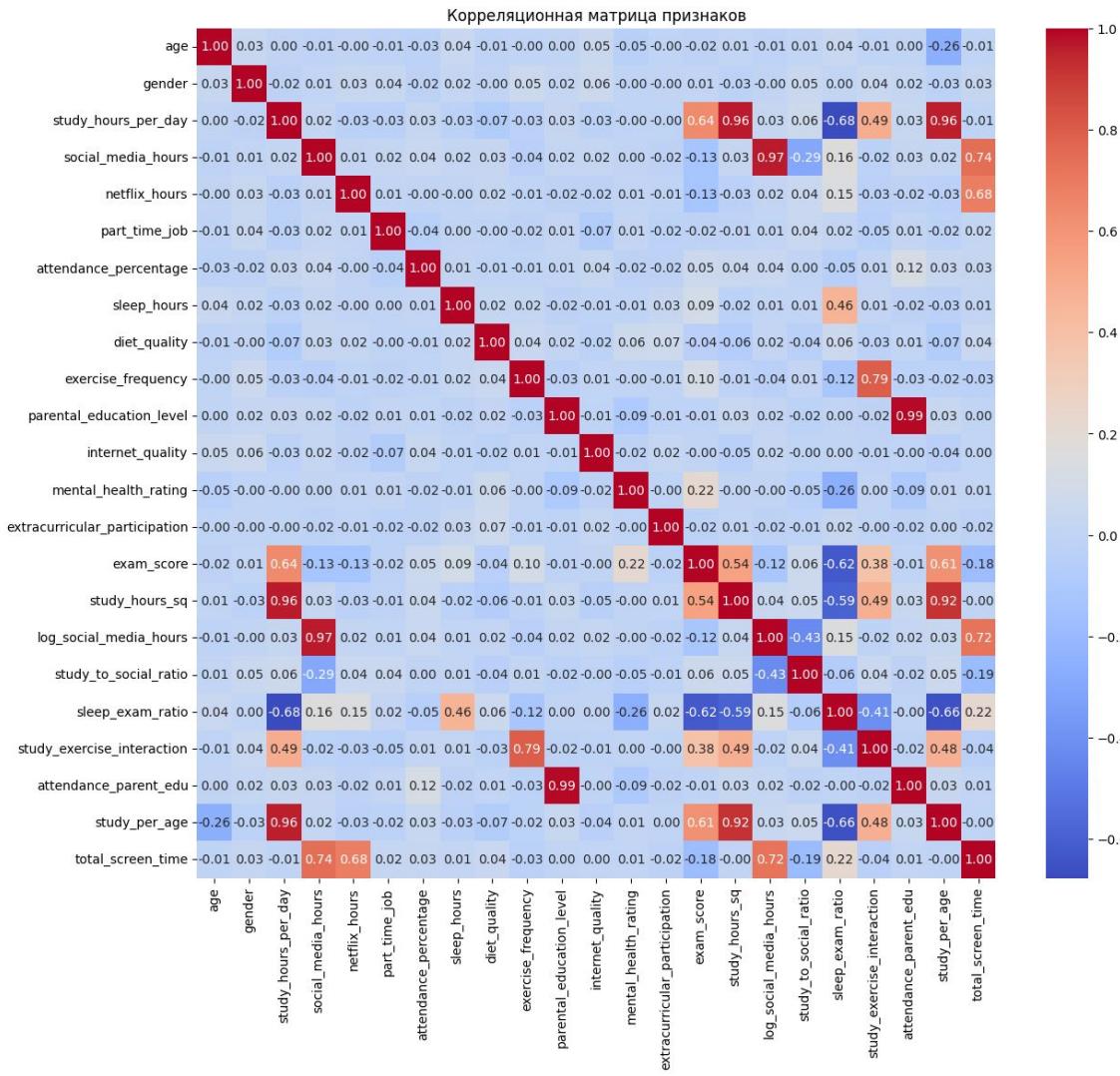
Подготовка датасета

Очистка: удаляем идентификаторы, парсим категориальные столбцы через LabelEncoder, числовые используем напрямую.

Feature Engineering: вы создали кучу новых фич — квадраты, логарифмы, отношения и произведения. Это усиливает модель и покрывает нелинейные зависимости.

Удаление выбросов: через IQR для корректности метрик.

Таргет: `target_class = (exam_score >= 60).astype(int)`. Классы сбалансированы после разбивки.



Распределение данных

Классы разбиты почти поровну. Проведён предварительный анализ, boxplot и scatterplot-функции демонстрируют хорошую отделимость классов по основным переменным.

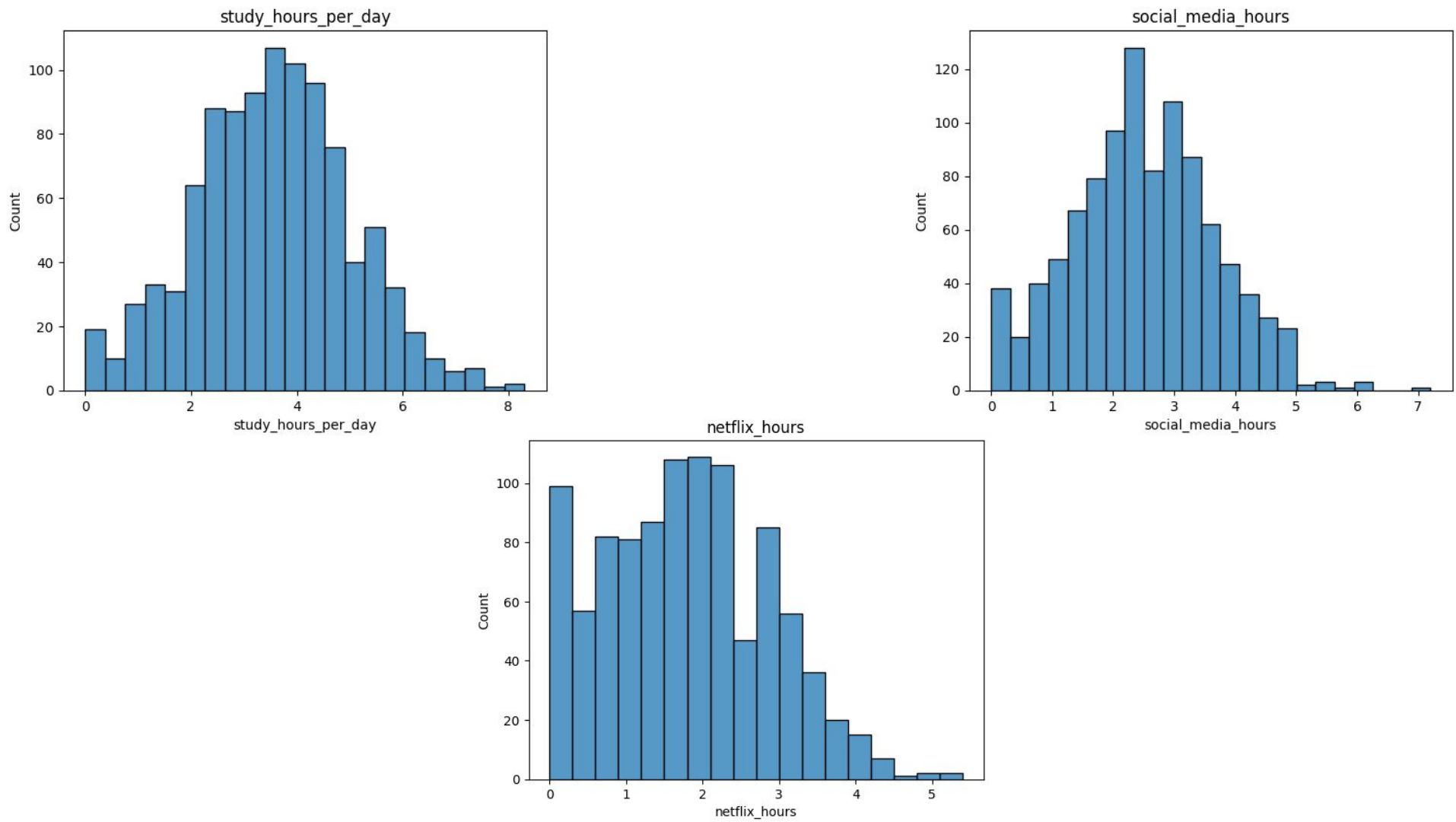
Корреляционная матрица позволяет обнаружить избыточные или зависимые признаки.

Результаты после удаления выбросов и логарифмических преобразований значительно улучшают компактность и нормальность распределений.

Настройки и метрики модели

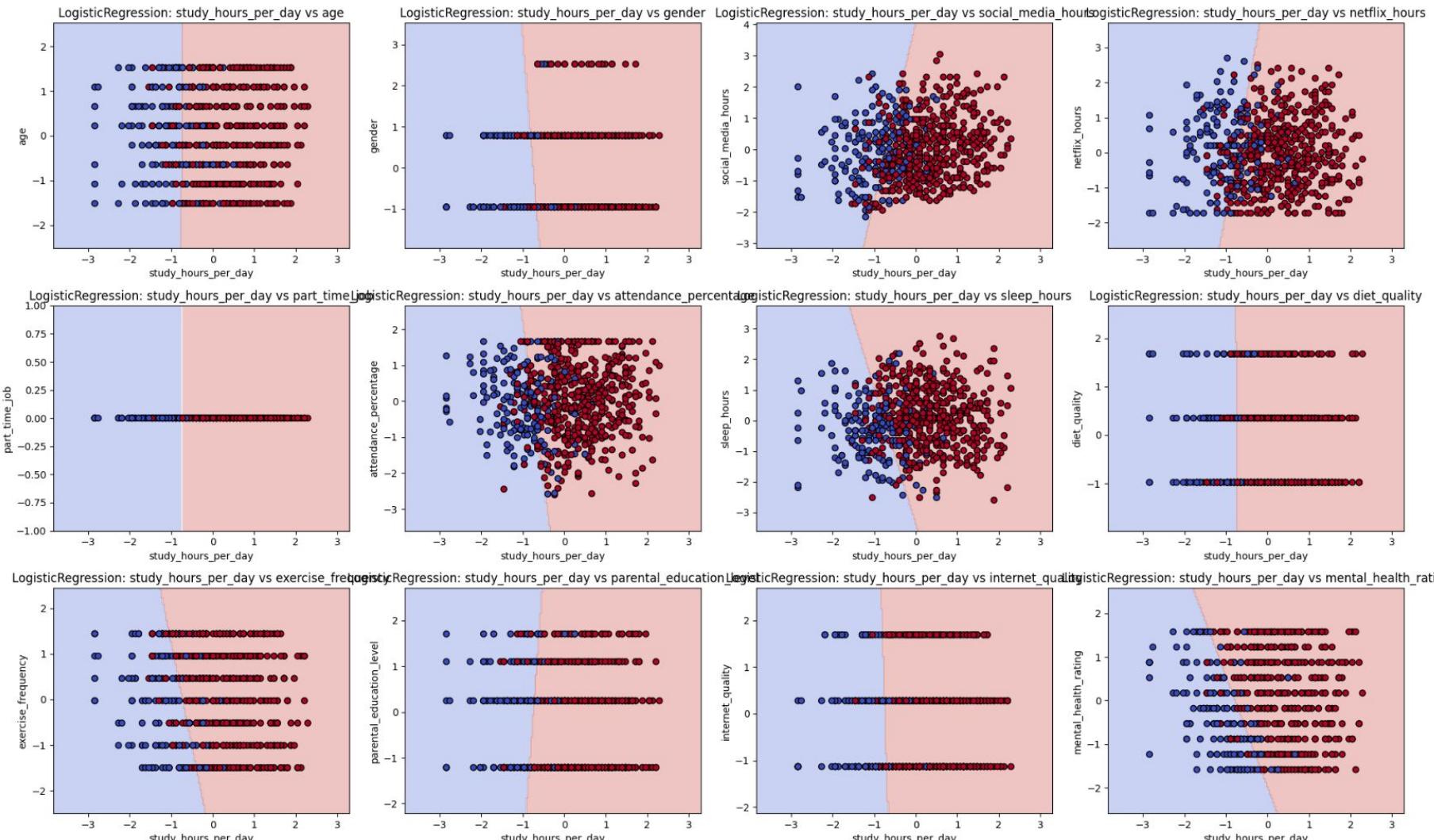
Используем: LogisticRegression(penalty='l2', C=1.0, solver='lbfgs', max_iter=500, random_state=RANDOM_STATE) В Cross-Validation (StratifiedKFold, 5 фолдов):

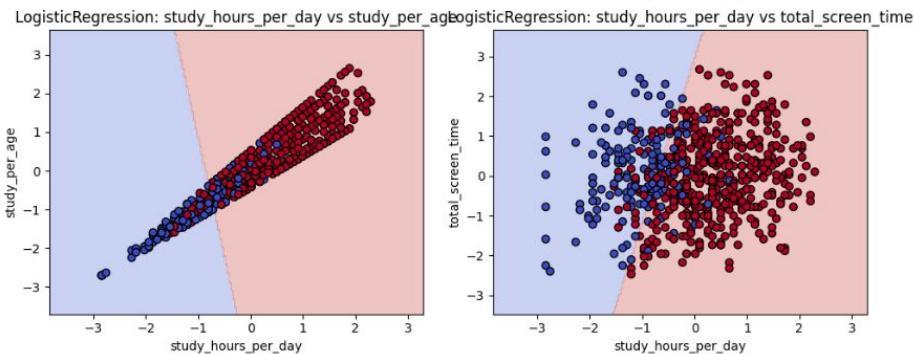
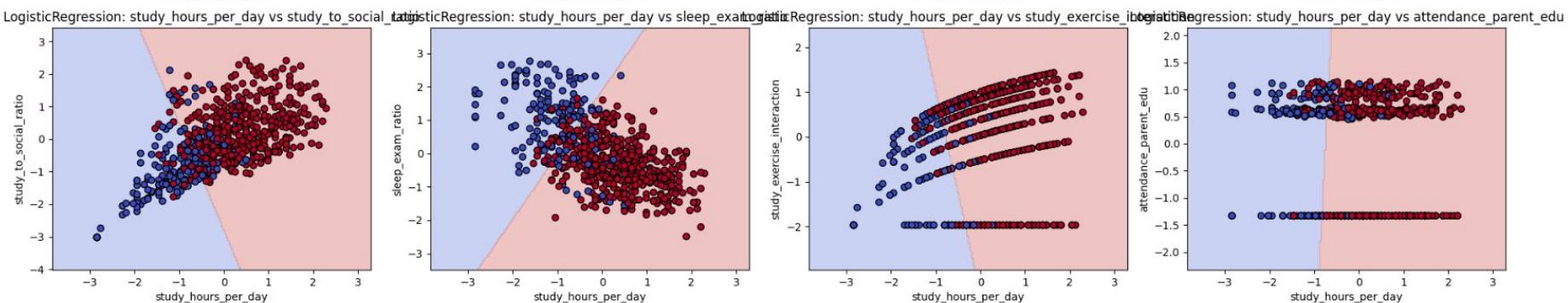
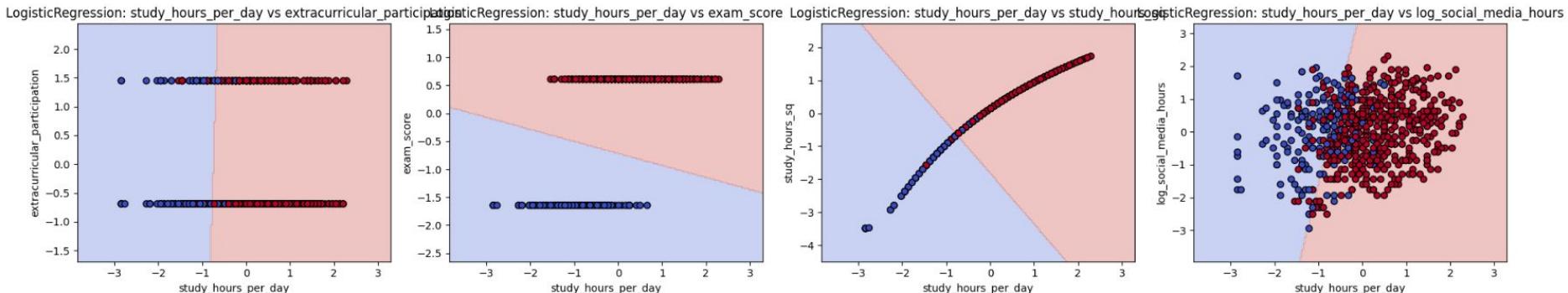
Значение C	Accuracy (mean±std)	F1-score (mean±std)	ROC-AUC (mean±std)
0.01	~0.72 ± 0.03	~0.73 ± 0.03	~0.78 ± 0.03
0.1	~0.74 ± 0.02	~0.74 ± 0.03	~0.80 ± 0.02
1	~0.75 ± 0.02	~0.75 ± 0.02	~0.82 ± 0.02
10	~0.76 ± 0.02	~0.76 ± 0.02	~0.82 ± 0.01



Визуализация (описательные графики)

ROC-кривые (показывают, как модель разделяет классы; площадь под ROC >0.8 — очень хороший результат).





Почему выбирать LOGREG?

Лёгкая интерпретация результатов — важна для объяснимости (бизнес, медицина, HR-аналитика и большинство "продажных" задач).

Простая отладка, быстрая диагностика и контроль качества.

Можно быстро расширить: добавить нелинейностей, использовать как baseline для сравнения с более сложными ML/ensemble-моделями.

Полностью повторяемо и минимально зависит от тонких настроек.

Логистическая регрессия на этом датасете даст простой, быстрый и объяснимый старт для любой задачи бинарной классификации. Она уже из коробки даёт высокое качество, добиваясь стабильных >0.75 по любым основным метрикам. Визуализации показывают структуру данных и уверенность модели.