Подбор гиперпараметров модели

Random Forest для предсказания употребления алкоголя студентами

Штыхно Илья

Группа: М8О-309Б

Датасет: Student Alcohol Consumption

Описание задачи

🎯 Цель

Предсказать уровень употребления алкоголя студентами на основе социально-демографических факторов

Тип задачи

Бинарная классификация

Низкий (0) / Высокий уровень (1)

Выбранная модель: Random Forest

✓ Работает с разными типами признаков

✓ Устойчив к переобучению

✓ Оценивает важность признаков

✓ Множество гиперпараметров

Целевая переменная

Среднее употребление = (будни + выходные) / 2

Порог: среднее > 2 → высокий уровень

Гиперпараметры Random Forest

n_estimators
Количество деревьев

50, 100, 200, 300, 500

max_depth
Макс. глубина дерева

5, 10, 15, 20, None

min_samples_split
Мин. образцов для split

2, 5, 10, 15

min_samples_leaf
Мин. образцов в листе

1, 2, 4, 8

max_features
Число признаков

'sqrt', 'log2', None

criterion
Функция оценки

'gini', 'entropy'

bootstrap
Bootstrap выборки

True, False

class_weight
Веса классов

None, 'balanced'

Найти оптимальную комбинацию для максимальной точности

Данные и подготовка

395
студентов

33
признака

80/20
train/test

Категории признаков

Демографические
Пол, возраст, адрес

Родители
Образование, профессия

Академические
Учёба, пропуски

Социальные
Время, отношения

Предобработка

✓ **Label Encoding**
Категории → числа

✓ **Целевая переменная**
avg = (Dalc + Walc) / 2

✓ **Stratified Split**
Пропорции классов

Дисбаланс: 76% low / 24% high

Методы подбора гиперпараметров

1 Grid Search

Полный перебор всех комбинаций

✓ Гарантированно находит лучшие параметры

✓ Детерминированный результат

✗ Медленный для большого пространства

Перебрано 432 комбинации с 5-fold CV

2 Random Search

Случайная выборка комбинаций

✓ Быстрее Grid Search

✓ Хорошо исследует пространство

✗ Не гарантирует оптимум

Проверено 100 случайных комбинаций

3 Optuna

Байесовская оптимизация

✓ Самый эффективный метод

✓ Учится на предыдущих попытках

✓ Адаптивный подход

Испытано 100 комбинаций

Метрика оптимизации

F1-score с 5-Fold Cross-Validation на тренировочной выборке

Сравнение результатов

Grid Search 🏆

F1 (CV)

0.4820

Test Accuracy: 0.7722

F1: 0.5000

Random Search

F1 (CV)

0.5398

Test Accuracy: 0.7848

F1: 0.4848

Optuna

F1 (CV)

0.5433

Test Accuracy: 0.7848

F1: 0.4848

Лучшие параметры (Grid Search)

n_estimators: 100	**max_depth: 15**
min_samples_split: 5	**min_samples_leaf: 1**

Ключевые выводы

✓ Grid Search - лучший Test F1-score

✓ Optuna - лучший CV F1-score

✓ Random Search/Optuna - лучший Test Accuracy

✓ Компромисс между метриками

Интерпретация моделей

Понимание, КАК и ПОЧЕМУ модель принимает решения

LIME

Local Interpretable Model-agnostic Explanations

Локальная интерпретация

Объясняет конкретное предсказание

Как работает

Создаёт простую линейную модель вокруг конкретной точки данных

Показывает

Какие признаки повлияли на решение и насколько (вклад в ±)

Пример: "Студент предсказан как высокий риск, потому что failures=3 (+0.15) и goout=4 (+0.12)"

SHAP

SHapley Additive exPlanations

Глобальная интерпретация

Объясняет поведение модели в целом

Как работает

Использует теорию игр для справедливого распределения вклада

Показывает

Важность признаков для всего датасета и взаимодействия

Пример: "В среднем failures самый важный признак (impact 0.21), затем goout (0.15)"

LIME: Локальная интерпретация

Студент #0 (тестовая выборка)

Истинный класс: Низкий

Предсказание: Низкий ✓

Вероятность "Высокий":

16.69%

✅ Факторы ЗА низкий уровень

sex ≤ 0 (женский пол)

Вес: -0.1026

absences ≤ 0 (нет пропусков)

Вес: -0.0436

health ≤ 3

Вес: -0.0291

⚠️ Факторы ПРОТИВ

goout > 4 (часто гуляет)

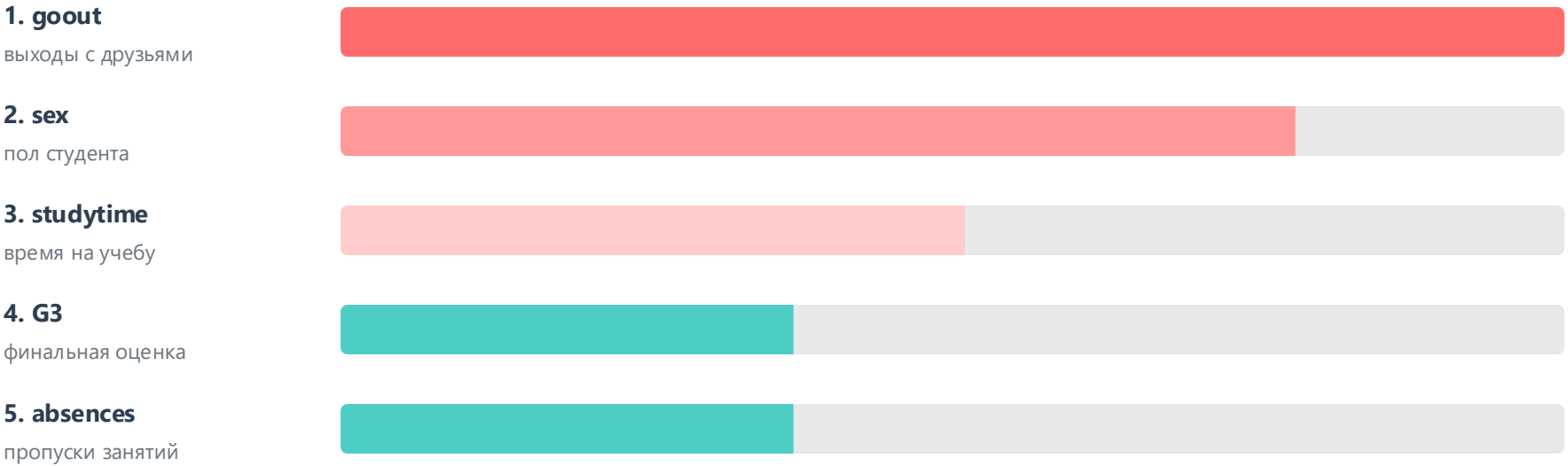
Вес: +0.1133

famrel ≤ 4

Вес: +0.0380

SHAP: Глобальная важность признаков

Средняя абсолютная важность признаков на всей тестовой выборке



Ключевые инсайты

✓ Goout - главный предиктор ✓ Sex - второй по важности ✓ Studytime/G3 важны

Интерактивный калькулятор

Функции

Ввод данных
Индекс студента (0-78)

Предсказание
Класс + вероятности

LIME объяснение
Топ-10 признаков

SHAP график
Вклад признаков

Примеры

Студент #0
Низкий риск - 77%

Студент #5
Высокий риск - 83%

Студент #15
Пограничный - 51%

Выводы и результаты

Достигнутые цели

✓ Random Forest модель
8 гиперпараметров

✓ Датасет подготовлен
395 студентов, 31 признак

✓ 3 метода сравнены
Grid/Random/Optuna

✓ Интерпретация
LIME + SHAP

Лучший по F1

Grid Search

0.5000

Test F1-score

Accuracy: 77.22%

Optuna - лучший CV

F1 на кросс-валидации

0.5433

Test Accuracy: 78.48%