Датасет: Факторы успеваемости студентов

ВЫПОЛНИЛА: ГЛЕБОВА МАРИЯ, ГРУППА М8О-307Б-23

Цель: Исследовать методы подбора гиперпараметров для модели Random Forest и сравнить их эффективность

Задачи:

1. Выбрать модель и изучить её гиперпараметры
2. Подготовить датасет для обучения
3. Подобрать гиперпараметры тремя методами (Grid Search, Random Search, Optuna)
4. Сравнить результаты методов
5. Создать калькулятор с интерпретацией модели (LIME и SHAP)

Выбор модели - Random Forest

Причины выбора:

- Хорошо работает для классификации и регрессии
- Устойчив к переобучению
- Обрабатывает разные типы признаков
- Много гиперпараметров для настройки
- Позволяет оценивать важность признаков

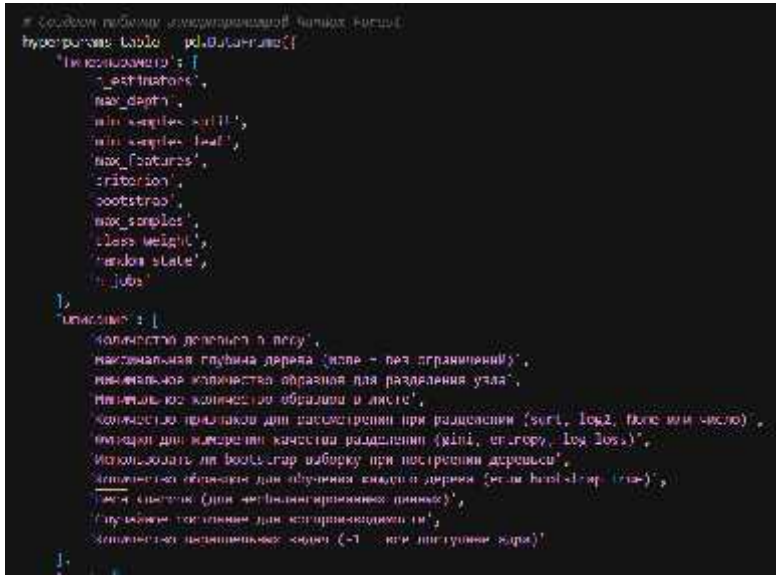


Таблица - гиперпараметров Random Forest:

	Гиперпараметр	Описание	Тип	Значение по умолчанию
0	n_estimators	Количество деревьев в лесу	int	100
1	max_depth	Максимальная глубина дерева (None - без ограни...	int или None	None
2	min_samples_split	Минимальное количество образцов для разделения...	int или float	2
3	min_samples_leaf	Минимальное количество образцов в листе	int или float	1
4	max_features	Количество признаков для рассмотрения при разд...	str, int, float или None	sqrt
5	criterion	Функция для измерения качества разделения (gin...	str	gini
6	bootstrap	Использовать ли bootstrap выборку при построен...	bool	True
7	max_samples	Количество образцов для обучения каждого дерев...	int, float или None	None
8	class_weight	Веса классов (для несбалансированных данных)	dict, list или str	None
9	random_state	Случайное состояние для воспроизводимости	int	None
10	n_jobs	Количество параллельных задач (-1 - все доступ...	int	None

Датасет и подготовка данных

-**Датасет:** Факторы успеваемости студентов

-Размер: 6607 записей, 20 признаков

- **Целевая переменная:** Exam_Score (балл на экзамене)

- **Классы:** 3 категории (Низкая, Средняя, Высокая успеваемость)

- **Предобработка:**

  - Заполнение пропусков

  - Кодирование категориальных признаков

  - Разделение на train/test (80/20)

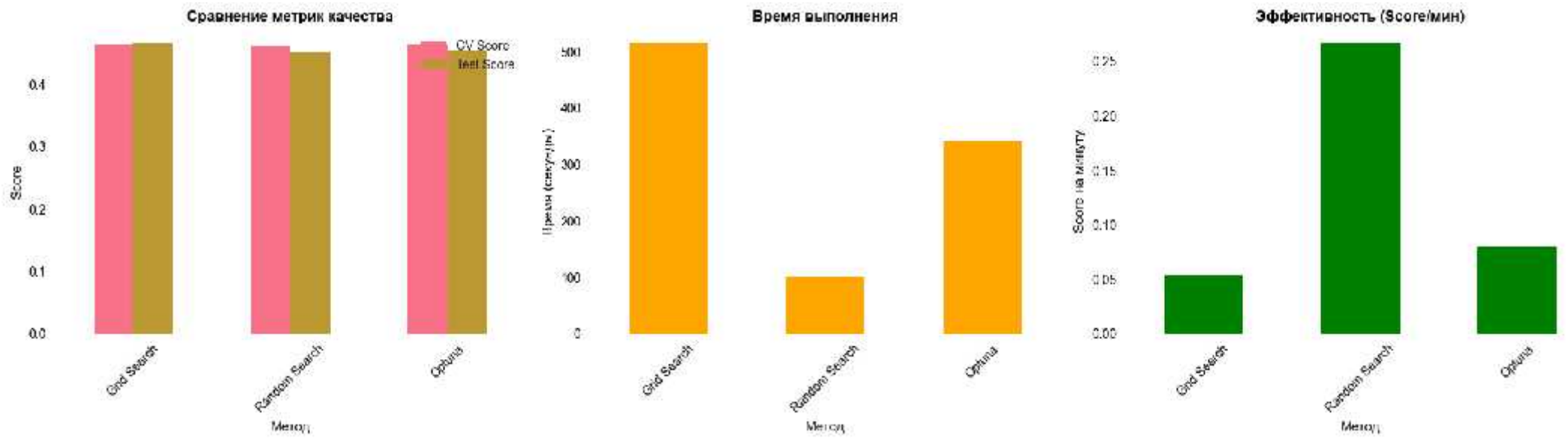
Методы подбора гиперпараметров

Grid Search: Полный перебор всех комбинаций
- Плюсы: гарантированно находит лучшую комбинацию
- Минусы: очень медленный

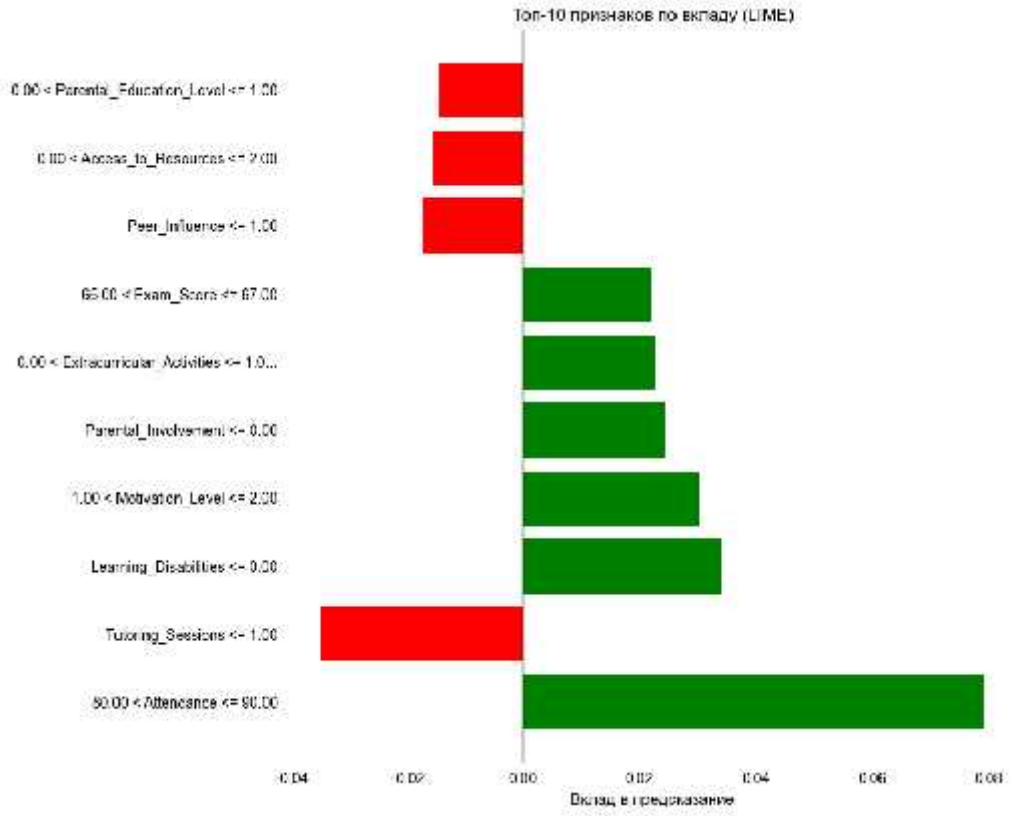
Random Search: Случайный поиск из заданных диапазонов
- Плюсы: быстрее Grid Search
- Минусы: не гарантирует оптимальность

Optuna: Байесовская оптимизация
- Плюсы: умный поиск, учится на результатах
- Минусы: сложнее в понимании

Сравнение результатов методов

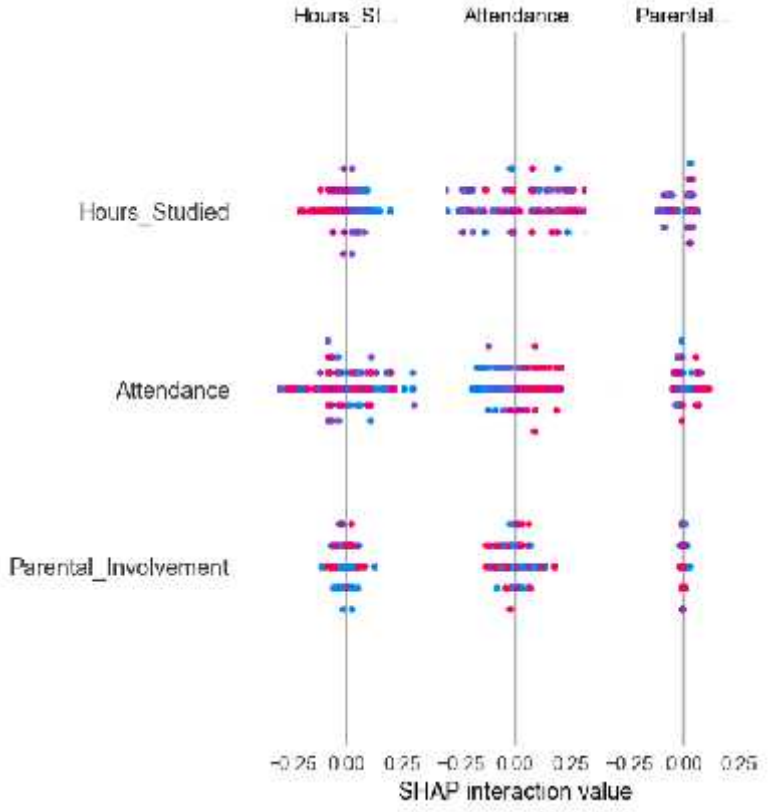


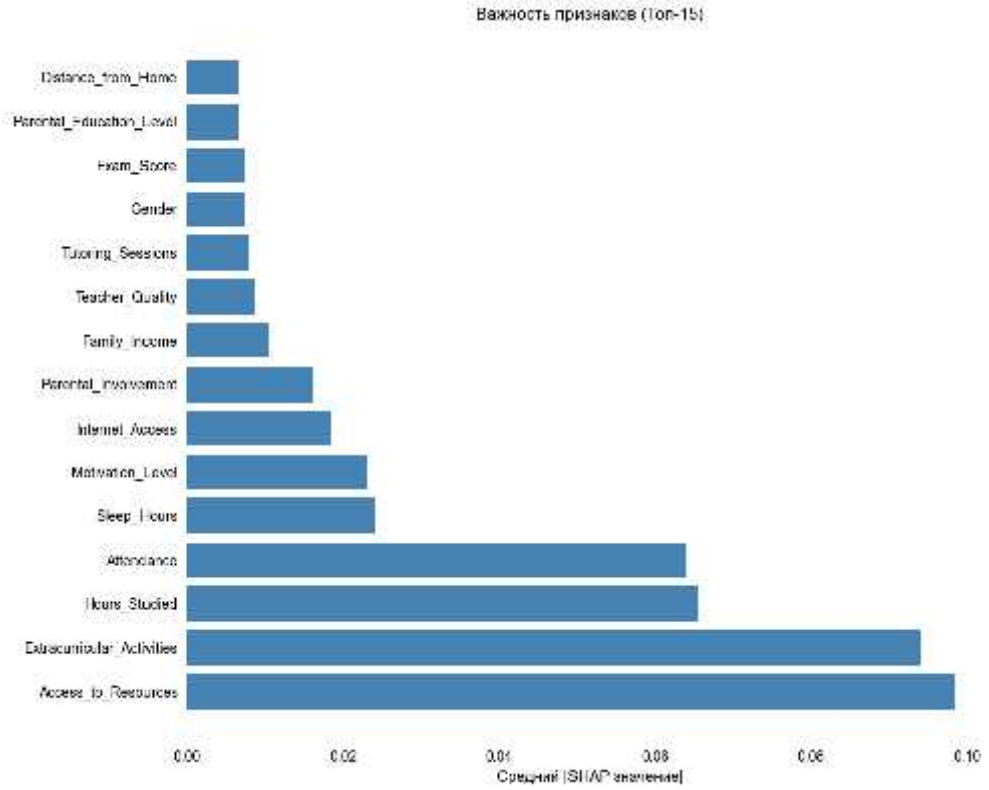
Локальная интерпретация (LIME)




```
LIME  ЛОКАЛЬНАЯ ИНТЕРПРЕТАЦИЯ

⧗ Создаем LIME объяснитель...
✅ LIME объяснитель создан

📊 Анализируем пример #1126 из тестовой выборки:
   Реальный класс: 0 (Низкая успеваемость)
   Предсказанный класс: 1 (Средняя успеваемость)
   Вероятности: [0.17884382 0.41674024 0.40483662]

⧗ Генерируем объяснение LIME...

📋 Таблица вклада признаков (топ 15):
                                  Признак     Вклад
                  80.00 < Attendance <= 90.00  0.079233
                      Tutoring_Sessions <= 1.00  0.046488
                  Learning_Disabilities <= 0.00  0.034104
                  1.00 < Motivation_Level <= 2.00  0.030198
                      Parental_Involvement <= 0.00  0.024319
  0.00 < Extracurricular_Activities <= 1.00  0.022704
                  65.00 < Exam_Score <= 67.00  0.021827
                      Peer_Influence <= 1.00 -0.017369
            0.00 < Access_to_Resources <= 2.00 -0.015394
    0.00 < Parental_Education_Level <= 1.00  0.014488
...
           1.00 < Distance_from_Home <= 2.00  0.012730
                0.00 < School_Type <= 1.00  0.011776
                      0.00 < Gender <= 1.00  0.010857
                      Physical_Activity <= 2.00  0.009126
```

Глобальная интерпретация (SHAP)





Основные выводы:

1. Random Forest показал хорошие результаты для задачи классификации успеваемости
2. Optuna оказался наиболее эффективным методом подбора гиперпараметров
3. LIME и SHAP дополняют друг друга в интерпретации модели
4. Наиболее важные признаки: Previous_Scores, Attendance, Hours_Studied

Практическая значимость:

- Создан калькулятор для предсказания и интерпретации результатов
- Модель может использоваться для анализа факторов успеваемости студентов

Спасибо за внимание!