ПОДБОР ГИПЕРПАРАМЕТРОВ

Кривошапкин Е. Б. М8О-309Б-23

ЦЕЛЬ И ЗАДАЧИ РАБОТЫ

1. Выбрать модель для обучения (Desicion tree, Random forest, SVM, KNN, Boosting)
2. Показать, какие гиперпараметры есть у выбранной модели
3. Выбрать датасет для обучения и подготовить данные для соответствующей модели
4. Подобрать гиперпараметры для модели и сравнить лучшие подборы для GridSearch, RandomSearch, Optuna
5. Для самого лучшего обучения показать локальную интерпретацию LIME, и глобальную SHAP

ИСПОЛЬЗУЕМАЯ МОДЕЛЬ

Random Forest - Ансамблевый метод **ML**, комбинирующий множество деревьев решений для повышения точности и устойчивости.

- Работа: Деревья на случайных подвыборках; голосование/усреднение.
- Плюсы: Устойчив к переобучению; высокая точность; **feature importance**.
- Минусы: Медленный; менее интерпретируемый.
- Применение: Классификация/регрессия в финансах, медицине, анализе данных

ДАТАСЕТ

- Название: **Pokemon.csv**
- Размер: 800 покемонов, 13 колонок.
- Числовые статы (средние): **HP** ~69, **Attack** ~79, **Defense** ~74, **Sp. Atk** ~73, **Sp. Def** ~72, **Speed** ~68, **Total** ~435.
- **Legendary** (целевой признак классификации): 92% обычные, 8% легендарные.
- Топ-5 типов (**Type 1**): **Water** (112), **Normal** (98), **Grass** (70), **Bug** (69), **Psychic** (57).
- Поколения: 1 (166), 5 (165), 3 (160), 4 (121), 2 (106).
- Пропуски: Только **Type 2** (386), остальные 0.

ПОДГОТОВКА ДАННЫХ

- Загрузка с **pd.read_csv**; выбор релевантных колонок (статы + **Legendary**).
- Преобразование **Legendary** в 0/1, а также категориальных признаков в числовые.
- Переименование колонок на русский для удобства.
- Разделение: **train_test_split** (80/20, **stratify=y** для баланса классов).
- Без сложной обработки (масштабирование в пайплайне); **Random Forest** не чувствителен к шкале.

НАСТРОЙКА МОДЕЛИ

Основные параметры:

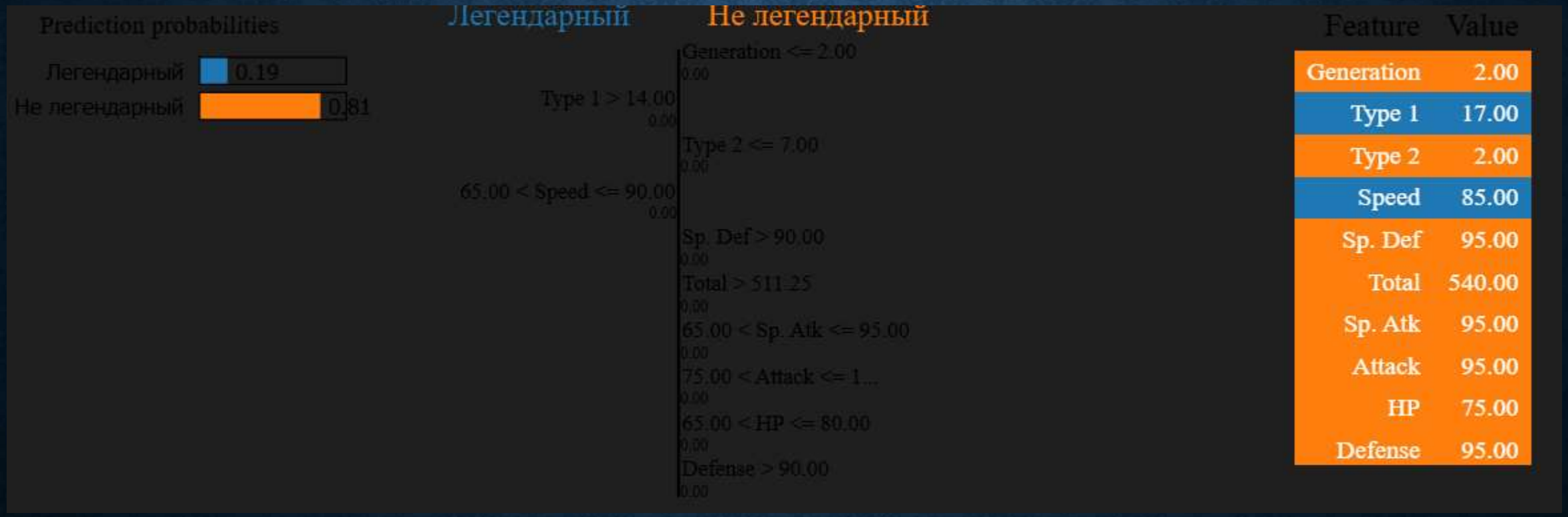
- n_estimators: 100, 200, 300, 400, 500
- max_depth: None, 10, 20, 30, 40, 50
- min_samples_split: 2, 5, 10
- min_samples_leaf: 1, 2, 4
- bootstrap: True, False
- class_weight: None, 'balanced'

СРАВНЕНИЕ МЕТОДОВ ТЮНИНГА

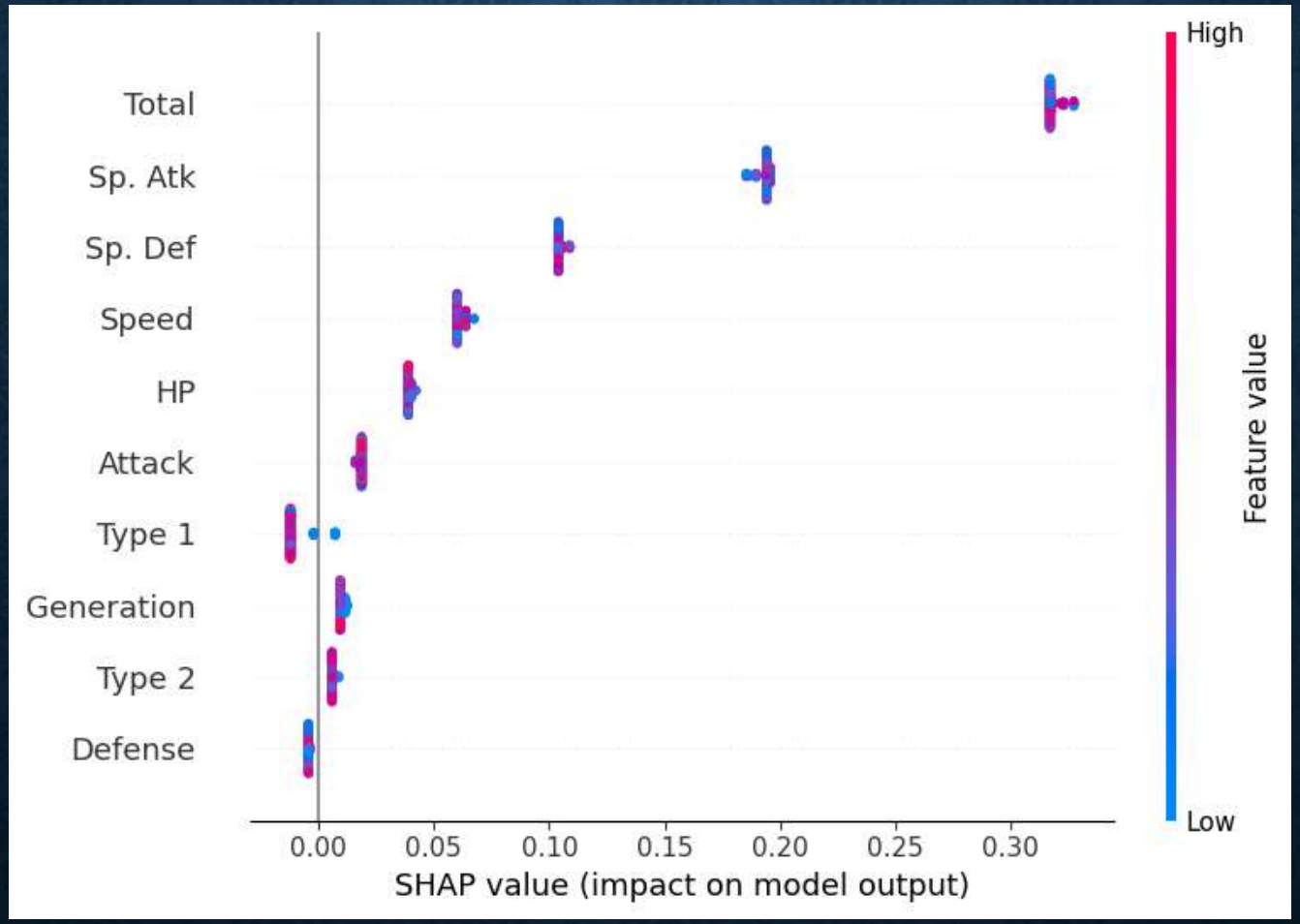
- **Grid Search:** Полный перебор, точный, но медленный.
- **Random Search:** Случайный, быстрее, для грубой оценки.
- **Optuna:** Байесовский, эффективный, балансирует поиск.

	Метод	Лучшие параметры	CV Accuracy	Test Accuracy
0	Grid Search	{'classifier__bootstrap': True, 'classifier__c...	0.965625	0.93125
1	Random Search	{'classifier__n_estimators': 500, 'classifier_...	0.962500	0.95000
2	Optuna	{'classifier__n_estimators': 101, 'classifier_...	0.957812	0.93125

ЛОКАЛЬНАЯ ИНТЕРПРЕТАЦИЯ LIME



ГЛОБАЛЬНАЯ ИНТЕРПРЕТАЦИЯ SHAP



Спасибо за внимание!