Лабораторная работа №3
Подбор гиперпараметров модели

Люзина Мария М8О-307Б-23

Описание датасета Water Potability

Датасет Water Potability представляет собой бинарную задачу классификации: предсказание пригодности воды для питья (Potability: 0 — не пригодна, 1 — пригодна). Ключевые характеристики на основе описания датасета:

- Размер: 3276 экземпляров (строк).
- Признаки: 9 числовых фич (ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity). Нет категориальных переменных.
- Целевая переменная: Potability (бинарная: 0/1).
- Распределение классов: Сбалансированное — примерно 1998 (61%) для класса 0 и 1278 (39%) для класса 1 (лёгкий дисбаланс, но не критический).
- Есть пропуски (missing values) в нескольких фичах (ph, Sulfate, Trihalomethanes — около 10–15% на фичу). Рекомендуется предобработка (импутация, удаление).
- Общий тип: Табличные данные для supervised binary classification. Метрики оценки: accuracy, precision, recall, F1-score (учитывая дисбаланс).

Выбор модели

Для данного датасета лучшей моделью будет Boosting, например, XGBoost. Причины такого выбора:

- Высокая точность: Boosting excels в табличных задачах с числовыми фичами, особенно с лёгким дисбалансом.
- Устойчивость: Хорошо справляется с пропусками (нативная поддержка в XGBoost) и шумом в данных (ph, Turbidity — химические метрики с вариацией).
- Практичность: Быстрое обучение на 3k+ сэмплах, встроенная feature importance для интерпретации (например, Turbidity и Organic_carbon часто топ-фичи).

Гиперпараметры XGBoost

	Гиперпараметр	Значение	По умолчанию	Описание
0	min_child_weight	None	Да	Мин. сумма весов в листе
1	gamma	None	Да	Мин. снижение функции потерь для сплита
2	scale_pos_weight	None	Да	Баланс классов (positive/negative)
3	enable_categorical	False	Да	Автоопределение категорий (False → безопасно с NaN)
4	n_estimators	500	Нет	Количество деревьев
5	learning_rate	0.020000	Нет	Скорость обучения (eta)
6	max_depth	6	Нет	Максимальная глубина дерева
7	subsample	0.800000	Нет	Доля объектов для каждого дерева
8	colsample_bytree	0.800000	Нет	Доля признаков для каждого дерева
9	reg_alpha	0.100000	Нет	L1-регуляризация
10	reg_lambda	1.000000	Нет	L2-регуляризация
11	random_state	42	Нет	Сид для воспроизводимости
12	n_jobs	-1	Нет	Количество ядер CPU (-1 = все)
13	tree_method	hist	Нет	Алгоритм построения деревьев

Подготовка датасета

- Импорт необходимых библиотек
- Импорт датасета и просмотр первых 3 строчек
- Просмотр информации о датасете
- Проверка на пропуски: удалять не будем, модель с методом tree_method = ‘hist’ сама с ними поработает
- Проверка балансов классов, построение распределение целевой переменной
- Разделение на обучающую и тестовую выборку

Подбор гиперпараметров

Общие пространства поиска

```
'n_estimators'     : [400, 600, 800],
'learning_rate'    : [0.05, 0.08, 0.1],
'max_depth'        : [5, 6, 7],
'min_child_weight' : [2, 4],
'subsample'        : [0.8, 1.0],
'colsample_bytree' : [0.8, 0.9],
'reg_alpha'        : [0, 0.1],
'reg_lambda'       : [1.0, 1.5]
```

Сравнение методов

Метод	ROC-AUC	Время	Лучшие параметры
Optuna	0.6553	1м 53с	{'n_estimators': 639, 'learning_rate': 0.011789246568012307, 'max_depth': 7, 'min_child_weight': 1, 'gamma': 0.45252667438341654, 'subsample': 0.7090309952094581, 'colsample_bytree': 0.9775303301213244, 'reg_alpha': 1.3565034673705125, 'reg_lambda': 1.1807309979428942}
RandomizedSearchCV	0.6442	1м 59с	{'subsample': 1.0, 'reg_lambda': 1.0, 'reg_alpha': 0, 'n_estimators': 400, 'min_child_weight': 2, 'max_depth': 6, 'learning_rate': 0.05, 'colsample_bytree': 0.8}
GridSearchCV	0.6350	13м 38с	{'colsample_bytree': 0.9, 'learning_rate': 0.05, 'max_depth': 7, 'min_child_weight': 2, 'n_estimators': 400, 'reg_alpha': 0.1, 'reg_lambda': 1.5, 'subsample': 0.8}

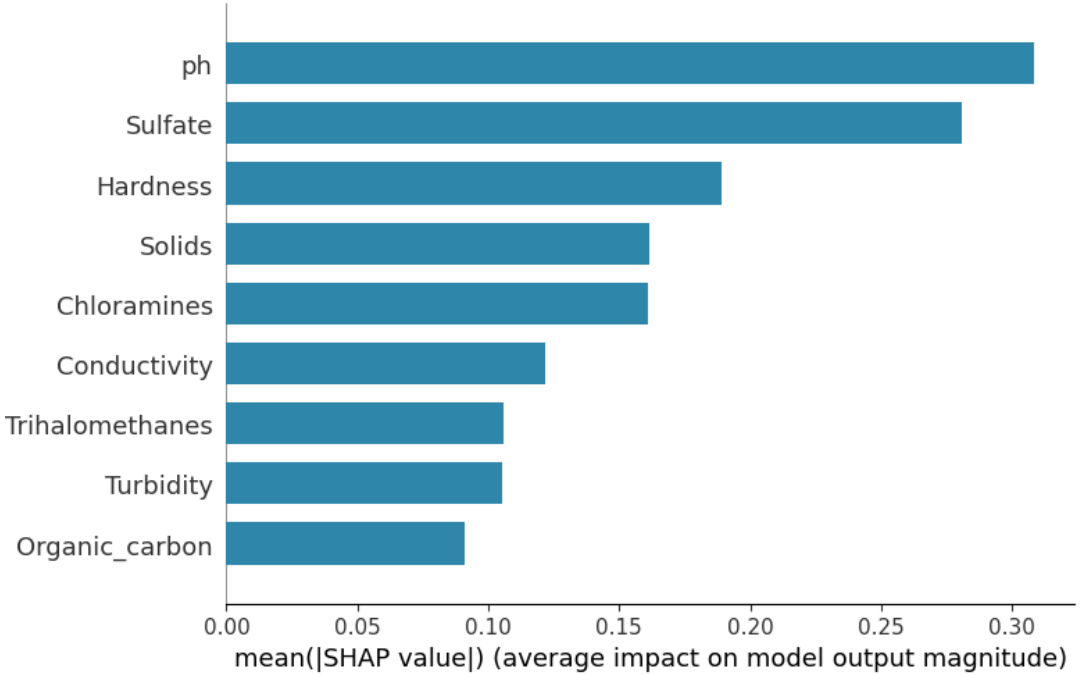
Optuna

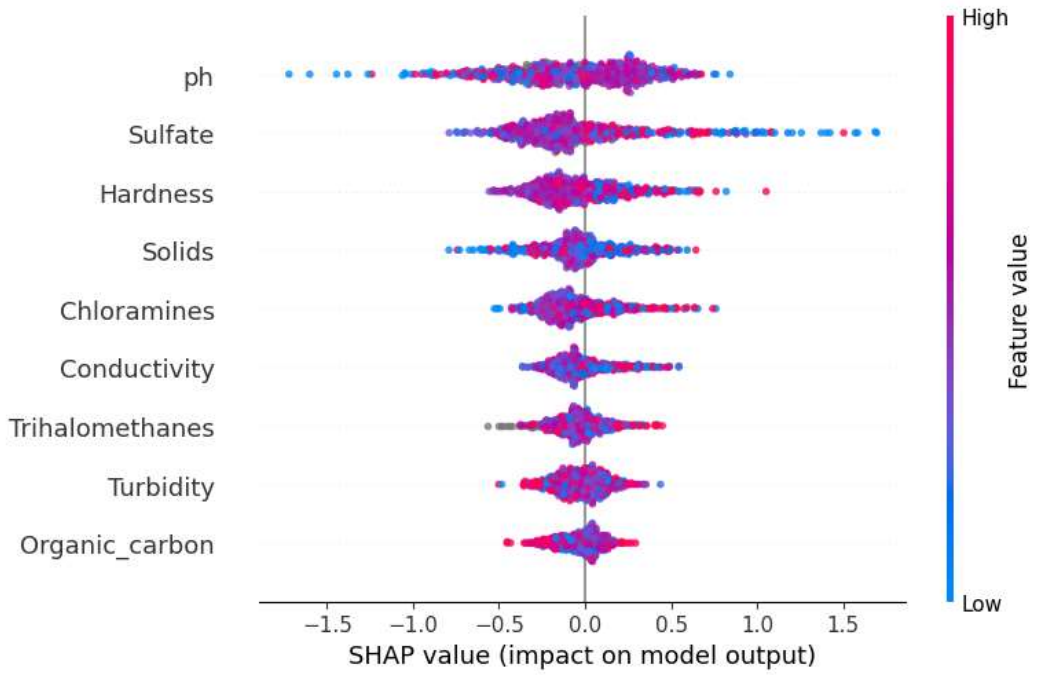
Модель Optuna оказалась лучше всех по следующим причинам:

- Самое широкое и гибкое пространство поиска
- Байесовская оптимизация умнее случайного поиска
- Нашла более тонкую настройку: learning_rate ≈ 0.0118 — очень маленькое, но зато n_estimators = 639 → модель медленно, но точно учится. gamma ≈ 0.045, reg_alpha ≈ 1.18, reg_lambda ≈ 1.36 — сильная регуляризация → меньше переобучения

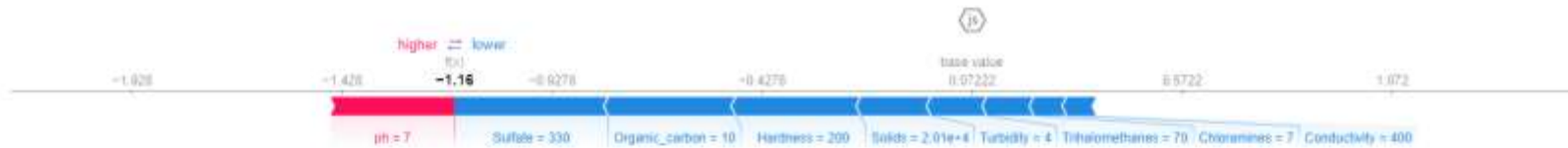
Это классический «осторожный», но очень эффективный ансамбль на шумных данных вроде Water Potability.

Глобальная интерпретация (SHAP)





Калькулятор с интерпретациями



Feature	Value
ph	7.00
Hardness	200.00
Organic_carbon	10.00
Solids	20100.00
Sulfate	330.00
Conductivity	400.00
Trihalomethanes	70.00
Turbidity	4.00
Chloramines	7.00