

ML. ЛАБОРАТОРНАЯ РАБОТА №3.

ПОДБОР ГИПЕРПАРАМЕТРОВ МОДЕЛИ.

СТУДЕНТ: ЛЕУ ИВАН АЛЕКСАНДРОВИЧ

ГРУППА: М8О-309Б-23

ЦЕЛЬ И ЗАДАЧИ

ЦЕЛЬ И ЗАДАЧИ РАБОТЫ:

1. ВЫБРАТЬ МОДЕЛЬ ДЛЯ ОБУЧЕНИЯ
2. ПОСТРОИТЬ БИНАРНУЮ МОДЕЛЬ, ПРЕДСКАЗЫВАЮЩУЮ REFUSED_FLAG (ЕСТЬ ЛИ ОТКАЗ ЗА НЕДЕЛЮ В БОЛЬНИЦЕ);
3. ПОДОБРАТЬ ГИПЕРПАРАМЕТРЫ С ПОМОЩЬЮ ТРЁХ МЕТОДОВ: GRIDSEARCH, RANDOMIZEDSEARCH, OPTUNA;
4. ВЫПОЛНИТЬ ЛОКАЛЬНУЮ И ГЛОБАЛЬНУЮ ИНТЕРПРЕТАЦИЮ МОДЕЛИ (LIME, SHAP).

ДАТАСЕТ И ДАННЫЕ

Датасет: Медицинские коляски. Датасет предназначен для моделирования реальной деятельности больницы среднего размера с акцентом на укомплектование персоналом, прием пациентов и распределение коек между отделениями.

Источник: `services_weekly.csv` (недельная агрегация по отделениям)
Размер: 208 строк, 10 столбцов

КЛЮЧЕВЫЕ ПОЛЯ: WEEK, MONTH, SERVICE, AVAILABLE_BEDS, PATIENTS_REQUEST, PATIENTS_ADMITTED, PATIENTS_REFUSED, PATIENT_SATISFACTION, STAFF_MORALE, EVENT

ЦЕЛЕВАЯ ПЕРЕМЕННАЯ И БАЛАНС КЛАССОВ

ЦЕЛЕВАЯ ПЕРЕМЕННАЯ: $\text{REFUSED_FLAG} = (\text{PATIENTS_REFUSED} > 0)$

ОПРЕДЕЛИМ ТАК: ОНА ОТВЕЧАЕТ НА ВОПРОС “БЫЛ ЛИ ХОТЯ БЫ ОДИН ОТКАЗ В ГОСПИТАЛИЗАЦИИ ЗА ЭТУ НЕДЕЛЮ ДЛЯ ДАННОЙ БОЛЬНИЦЫ?”

БОЛЬНИЦА ЗА НЕДЕЛЮ НЕ ОТКАЗАЛА НИ ОДНОМУ ПАЦИЕНТУ (ТО ЕСТЬ ВСЕХ ПРИНЯЛА) - 0,
БОЛЬНИЦА ЗА НЕДЕЛЮ ОТКАЗАЛА ХОТЯ БЫ ОДНОМУ ПАЦИЕНТУ - 1

БАЛАНС:

148 — 1 (ЕСТЬ ОТКАЗ),

60 — 0 (НЕТ ОТКАЗА)

ПРИЗНАКИ. ГИПЕРПАРАМЕТРЫ RANDOMFOREST

Категории: service, event → OneHotEncoder

Числовые: week, month, available_beds, patients_request, patients_admitted, patient_satisfaction, staff_morale

Pipeline: ColumnTransformer (ONE) → RandomForest

Гиперпараметры RandomForest (перечень).

ПОДБОР:

n_estimators — количество деревьев, max_depth — глубина, min_samples_split, min_samples_leaf, max_features, criterion

МЕТОДЫ ПОИСКА ГИПЕРПАРАМЕТРОВ

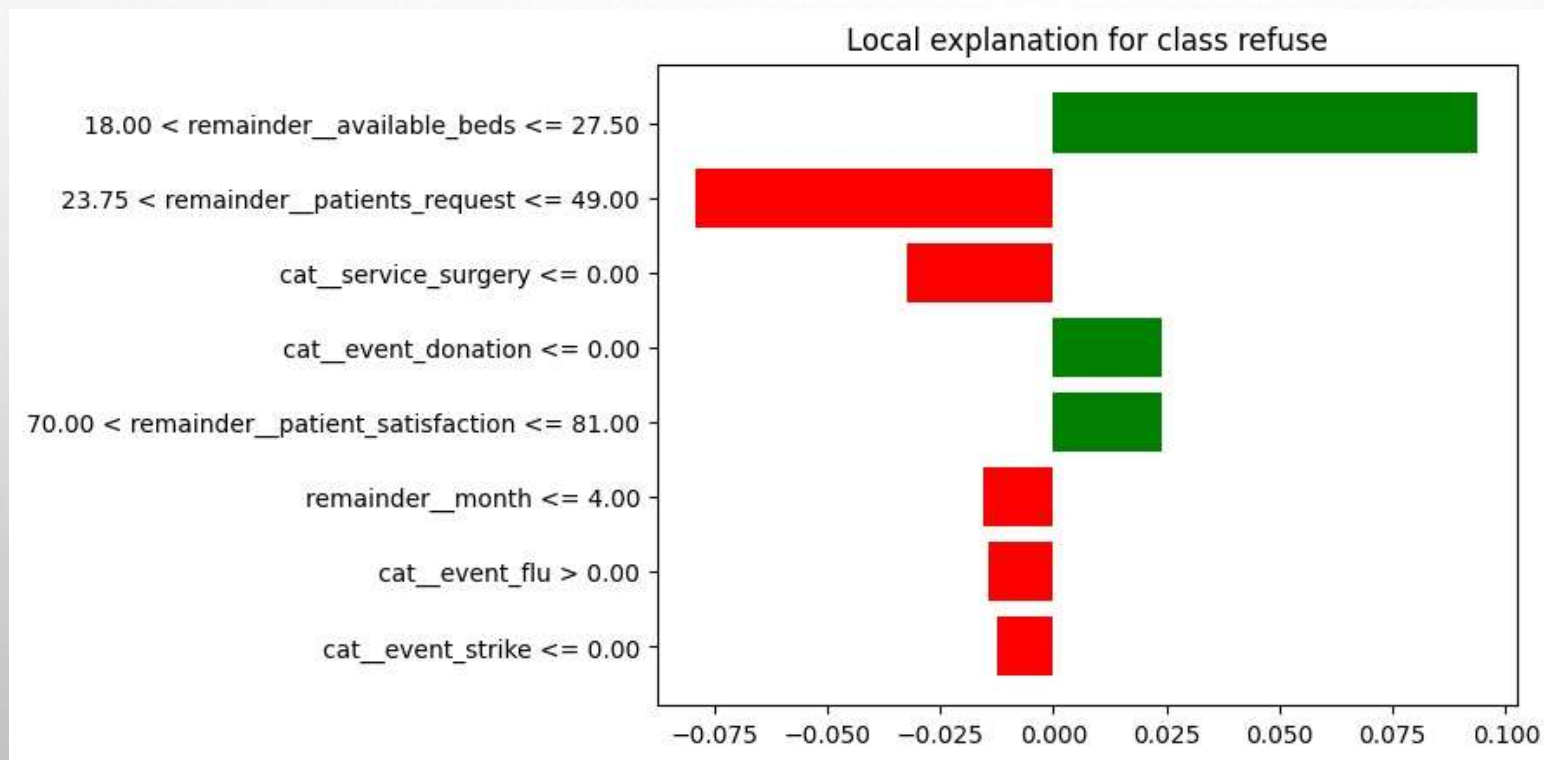
1. GRIDSEARCHCV — ПОЛНЫЙ ПЕРЕБОР (ДЕТЕРМИНИРОВАН). ПЕРЕБИРАЕТ ВСЕ ВОЗМОЖНЫЕ КОМБИНАЦИИ УКАЗАННЫХ ГИПЕРПАРАМЕТРОВ.
2. RANDOMIZEDSEARCHCV — СЛУЧАЙНАЯ ВЫБОРКА КОМБИНАЦИЙ (БЫСТРЕЕ). ВЫБИРАЕМ СЛУЧАЙНЫЕ КОМБИНАЦИИ ГИПЕРПАРАМЕТРОВ ИЗ ЗАРАНЕЕ ОПРЕДЕЛЁННЫХ ДИАПАЗОНОВ.
3. OPTUNA — ПОСЛЕДОВАТЕЛЬНАЯ ОПТИМИЗАЦИЯ (ИНТЕЛЛЕКТУАЛЬНЫЙ ПОИСК). СТРОИМ ВЕРОЯТНОСТНУЮ МОДЕЛЬ ФУНКЦИИ ПОТЕРЬ И ИСПОЛЬЗУЕМ ЕЁ, ЧТОБЫ ВЫБИРАТЬ НОВЫЕ ГИПЕРПАРАМЕТРЫ «УМНЕЕ», ЧЕМ СЛУЧАЙНО.

GRID — ГАРАНТИРОВАННО ПЕРЕБЕРЕТ СЕТКУ, RANDOMIZED — ЭКОНОМИТ ВРЕМЯ, OPTUNA — ЧАСТО ДАЁТ ХОРОШИЕ РЕЗУЛЬТАТЫ ПРИ ОГРАНИЧЕННОМ БЮДЖЕТЕ.

РЕЗУЛЬТАТЫ

- GRIDSEARCHCV – THE BEST! -> ВЫБИРАЕМ КАК ФИНАЛЬНУЮ МОДЕЛЬ
TEST ACCURACY: 0.9337970478998298, {'CRITERION': 'ENTROPY', 'MAX_DEPTH': 5, 'MAX_FEATURES': NONE, 'MIN_SAMPLES_LEAF': 2, 'MIN_SAMPLES_SPLIT': 2, 'N_ESTIMATORS': 150}
- RANDOMIZEDSEARCHCV
TEST ACCURACY: 0.9337970478998298, {'CRITERION': 'ENTROPY', 'MAX_DEPTH': 20, 'MAX_FEATURES': NONE, 'MIN_SAMPLES_LEAF': 3, 'MIN_SAMPLES_SPLIT': 2, 'N_ESTIMATORS': 185}
- OPTUNA
TEST ACCURACY: 0.8959156785243741, {'N_ESTIMATORS': 61, 'MAX_DEPTH': 20, 'MIN_SAMPLES_SPLIT': 2, 'MIN_SAMPLES_LEAF': 4, 'MAX_FEATURES': 'SQRT', 'CRITERION': 'GINI'}

ЛОКАЛЬНАЯ ИНТЕРПРЕТАЦИЯ: LIME



ЧТО ДЕЛАЕТ: ОБЪЯСНЯЕТ ПРЕДСКАЗАНИЕ ДЛЯ ОДНОЙ СТРОКИ (ЛОКАЛЬНО).

ПРИМЕР: ПОКАЗАН ВЫВОД LIME для 7-й НЕДЕЛИ: PATIENTS_REQUEST и AVAILABLE_BEDS ДАЛИ НАИБОЛЬШИЙ ВКЛАД В СТОРОНУ ОТКАЗА.

ГЛОБАЛЬНАЯ ИНТЕРПРЕТАЦИЯ: SHAP

(SHAP - SHapley Additive exPlanations). Что делает: показывает средний, “глобальный” вклад признаков в предсказания модели.

Результат (топ-признаки):

patients_request 0.3207 – **самый сильный**
available_beds 0.1262 – **самый сильный**
patients_admitted 0.0039
week 0.0031
patient_satisfaction 0.003
staff_morale 0.0012
month 0.0011
event_flu 0.0006
service_emergency 0.0006

ГЛОБАЛЬНАЯ ИНТЕРПРЕТАЦИЯ: SHAP. ВИЗУАЛИЗАЦИЯ

