ML. Лабораторная работа №3.

Подбор гиперпараметров модели.

Студент: Велиев Рауф Рамиз оглы

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: refused_flag = (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 (OHE) → 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 acccuracy: 0.880952380952380 9, {'criterion': 'entropy', 'max_depth': 5, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}

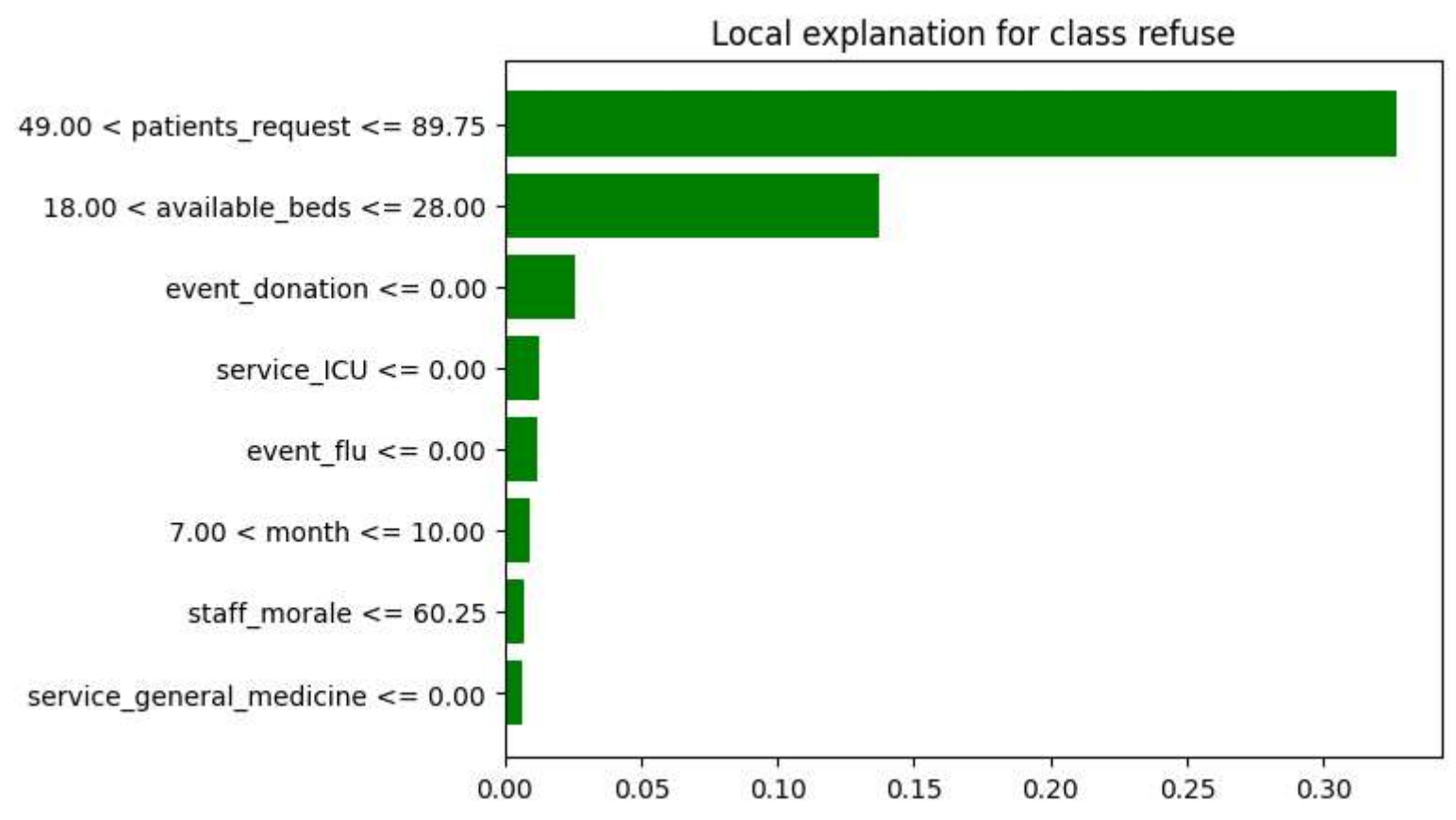
- RandomizedSearchCV

  test accuracy: 0.8571428571428571, {'criterion': 'entropy', 'max_depth': 20, 'max_features': None, 'min_samples_leaf': 3, 'min_samples_split': 2, ‘n_estimators': 185}

- Optuna

  test accuracy: 0.8095238095238095, {'n_estimators': 86, 'max_depth': 20, 'min_samples_split': 8, 'min_samples_leaf': 1, 'max_features': 'log2', 'criterion': 'entropy'}

Локальная интерпретация: LIME



Что делает: объясняет предсказание для одной строки (локально).

Пример: показан вывод LIME для 38-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 0.315 – **самый сильный**
available_beds 0.126 – **самый сильный**
week 0.0078
patients_admitted 0.0047
service_surgery 0.003
month 0.0024
patient_satisfaction 0.0024
staff_morale 0.0021
service_general_medicine 0.0007
event_strike 0.0002

Глобальная интерпретация: SHAP. Визуализация