Факультет
Аналитики Big Data
Финальный проект
Тамбовцев Р.

Прогноз смертности в больницах для пациентов с сердечной недостаточностью (СН), поступивших в отделения интенсивной терапии, основанный на машинном обучении ретроспективный анализ базы данных MIMIC-III ('Medical Information Mart for Intensive Care') Магазин медицинской информации для интенсивной терапии.

Задача разработать и проверить модель прогнозирования госпитальной смертности от всех причин среди пациентов с СН, поступивших в отделение интенсивной терапии.

Целевая переменная — результат:
0 - жив
1 - умер

Описание данных

Используя запросы языка структурированных запросов (PostgreSQL, версия 9.6), данные о демографических характеристиках, основных показателях жизнедеятельности и лабораторных показателях были извлечены из следующих таблиц в наборе данных MIMIC III: ПОСТУПЛЕНИЯ, ПАЦИЕНТЫ, ICUSTAYS, D ICD DIAGNOSIS, DIAGNOSIS ICD, LABEVENTS, D LABIEVENTS, CHARTEVENTS, D ITEMS, NOTEEVENTS и OUTPUTEVENTS.

Основываясь на предыдущих исследованиях 7–9 13–15, клинической значимости и общедоступности на момент презентации, мы извлекли следующие данные: демографические характеристики (возраст на момент госпитализации, пол, этническая принадлежность, вес и рост); жизненные показатели (частота сердечных сокращений (ЧСС), систолическое артериальное давление [САД], диастолическое артериальное давление [ДАД], среднее артериальное давление, частота дыхания, температура тела, насыщение пульса кислородом [SPO2], диурез [первые 24 часа]); сопутствующие заболевания (гипертензия, мерцательная аритмия, ишемическая болезнь сердца, сахарный диабет, депрессия, железодефицитная анемия, гиперлипидемия, хроническая болезнь почек (ХБП) и хроническая обструктивная болезнь легких [ХОБЛ]); и лабораторные показатели (гематокрит, эритроциты, средний корпускулярный гемоглобин [MCH], средняя концентрация корпускулярного гемоглобина [MCHC], средний корпускулярный объем [MCV], ширина распределения эритроцитов [RDW], число тромбоцитов, лейкоциты, нейтрофилы, базофилы, лимфоциты, протромбиновое время [PT], международное нормализованное отношение [INR], NT-proBNP, креатинкиназа, креатинин, азот мочевины крови [BUN], глюкоза, калий, натрий, кальций, хлорид, магний, анионная щель, бикарбонат, лактат, концентрация ионов водорода [pH], парциальное давление CO2 в артериальной крови и ФВ ЛЖ) с использованием языка структурированных запросов (SQL) с PostgreSQL (версия 9.6).

Извлеченные демографические характеристики и показатели жизнедеятельности регистрировались в течение первых 24 часов после каждого поступления, а лабораторные показатели измерялись в течение всего пребывания в отделении интенсивной терапии. Сопутствующие заболевания идентифицировали с использованием кодов МКБ-9. Для переменных данных с несколькими измерениями для анализа было включено рассчитанное среднее значение. Первичным результатом исследования была внутрибольничная смертность, определяемая как жизненный статус на момент выписки из больницы у выживших и невыживших.

Алгоритм действий для решения:

- Импортируем необходимые модули
- Загружаем файл с данными и проверяем его содержание на ошибки
- Используя SimpleImputer:
  - Заменяем пропущенные данные типа float и int средними значениями;
  - Переменной outcome присваиваем наиболее частое значение
- Собираем новый датафрейм для последующей работы
- Оцениваем зависимость различных заболеваний и пола со смертностью
- Создаём корреляционную матрицу для оценки зависимости сопутствующих заболеваний

  со смертностью (максимальная корреляция со смертностью у гипертонии и почечной недостаточностью)
- Создаём корреляционную матрицу для оценки зависимости значений результатов различных анализов со смертностью (максимальное значение обратной корреляции у лимфоцитов и нейтрофилов (врождённая составляющая иммунитета), максимальная прямая корреляция у MCV — mean corpuscular volume — Средний объем эритроцитов и MCH (Mean Cell Hemoglobin) — содержание гемоглобина)
- Значительных величин корреляций нет. Это указывает на отсутствие проблемы мультиколлинеарности в предоставленных данных.

Алгоритм действий для решения:

- Для оценки типа распределения значений переменных, строим графики, выбранных

  случайным методом переменных. Результат говорит о том, что у них преобладает нормальное распределение

- Разбиваем данные на тестовый и тренировочный датасет

  (x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.3, random_state=123))

- Для обучения модели используем XGBoost — библиотеку, реализующая методы градиентного бустинга.

- Оцениваем модель с помощью модуля sklearn:

  from sklearn. metrics import classification_report, accuracy_score, confusion_matrix

- Исходные данные несбалансированные, для решения этой проблемы будем использовать

  модуль SMOTE (Synthetic Minority Oversampling Technique) – Oversampling.

  SMOTE помогает увеличить недопредставленные числа в наборе данных модели машинного обучения. Это лучший метод, который позволяет увеличить число редких случаев вместо дублирования предыдущих.

Алгоритм действий для решения:

- После использования SMOTE, вновь делим данные на обучающую и тестовую выборку в соотношении 70/30 и обучаем на XGBOOST
- При сравнении показатели точности после использования SMOTE были улучшены.

Значения до использования SMOTE

In [80]:
```
print(classification_report(y_test, pred))
```

	precision	recall	f1-score	support
0	0.88	0.98	0.93	301
1	0.71	0.23	0.34	53
accuracy			0.87	354
macro avg	0.79	0.60	0.64	354
weighted avg	0.85	0.87	0.84	354

Значения после использования SMOTE

[79]:
```
print(classification_report(Y_test, pred_b))
```

	precision	recall	f1-score	support
0	0.94	0.92	0.93	302
1	0.93	0.94	0.94	309
accuracy			0.93	611
macro avg	0.93	0.93	0.93	611
weighted avg	0.93	0.93	0.93	611