Отчёт по лаб №3

实验报告 3

Чжоу Сяосюэ ИУ1И-41м

10 мая 2025г.
2025 年 5 月 10 日

1 Цель работы

Использование методов autoML в наборе данных по кардиологии для решения задачи бинарной классификации

2 Ключевой навык

Применение autoML моделей, расчет точности классификатора

3 Описание методов исследования

AutoML сокращает ручное вмешательство за счет автоматизации выбора
признаков, выбора модели и настройки гиперпараметров. Распространенные фреймворки (такие как H2O, AutoGAMA, AutoML LightAutoML и т. д.) используют сеточный поиск или эволюционные алгоритмы для оптимизации производительности модели.

4 Ход работы

4.1 Обработка исходных данных

В этом эксперименте мы использовали исходный набор данных ЭКГ и выполнили над ним несколько операций предварительной обработки данных.

1 实验目的

用 autoML 方法处理心脏病学数据集以解决二元分类问题

2 关键词

应用 autoML 模型，计算分类器准确率

3 研究方法说明

AutoML 通过自动化特征选择、模型选择和超参数调优，减少人工干预。常见框架（如 H2O、AutoGluon、BlueCast 等）利用网格搜索或进化算法优化模型性能。

4 实验过程

4.1 原始数据处理

在本实验中，我们使用了原始的 ECG 数据集，并对其进行了多项数据预处理操作。首先，使用

Сначала мы использовали pandas для чтения необработанных данных, содержащих первые 5000 строк, и сохранили все числовые столбцы для последующего анализа. Затем столбцы, содержащие пропущенные значения, были удалены для обеспечения целостности данных. Затем, путем фильтрации столбцов, содержащих только несколько уникальных значений, те постоянные столбцы, которые не меняются, удаляются, чтобы не мешать обучению и анализу модели.

Чтобы получить представление о взаимосвязях между признаками, мы рассчитали коэффициенты корреляции между числовыми столбцами и визуализировали эти взаимосвязи с помощью матрицы рассеяния. Кроме того, для удаления возможных выбросов мы отфильтровали данные по 99%-ному квантилю и оставили только те строки, собственные значения которых были меньше этого квантиля, тем самым удалив экстремальные выбросы и еще больше улучшив качество данных.

4.2 Построение моделей

- Используйте H2O AutoML, чтобы установить ограничение времени обучения в 60 секунд;
- Используйте LightAutoML для установки временных ограничений и выполнения многоуровневого слияния моделей;
- Сделайте прогнозы для того же тестового набора и сохраните матрицу ошибок и индикаторы.

4.3 AutoML H2O

H2O AutoML — это мощный инструмент автоматизированного машинного обучения, который упрощает процесс машинного обучения и помогает пользователям быстро создавать и обучать эффективные модели с

pandas 读取了包含前 5000 行的原始数据，并保留了所有数值型列，以便进行后续分析。接着，删除了包含缺失值的列，以确保数据的完整性。随后，通过筛选仅包含多个唯一值的列，去除了那些没有变化的常量列，避免其对模型训练和分析造成干扰。

为了深入了解各个特征之间的关系，我们计算了各数值列之间的 相关系数，并使用散点矩阵（scatter matrix）可视化了这些关系。此外，为了去除可能存在的异常值，我们根据 99% 分位数筛选了数据，仅保留那些特征值小于该分位数的行，从而去除了极端的异常数据，进一步提高了数据的质量。

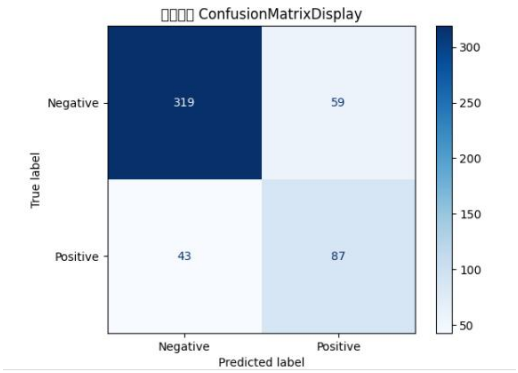
4.2 模型构建与训练

- 使用 H2O AutoML 设置训练时间上限 60 秒；
- 使用 LightAutoML 设定时间限制，执行多级模型融合；
- 对同一测试集分别进行预测，并保存混淆矩阵与指标。

4.3 AutoML H2O

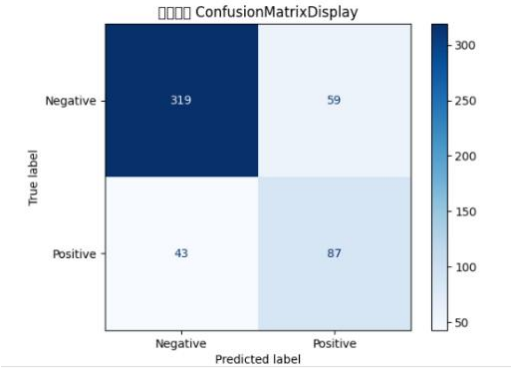
H2O AutoML 是一个强大的自动化机器学习工具，它简化了机器学习流程，通过自动化的方式帮助用户快速构建和训练高效的模型。H2O AutoML 支持多种机器学习算法，能够自动选择最佳的模型和

помощью автоматизации. H2O AutoML поддерживает различные алгоритмы машинного обучения и может автоматически выбирать лучшую модель и гиперпараметры.



Confusion matrix:
[[319 59]
[43 87]]
Accuracy: 0.80
F1-Score: 0.63
Recall: 0.67
Precision: 0.60
ROC-AUC: 0.88

超参数。

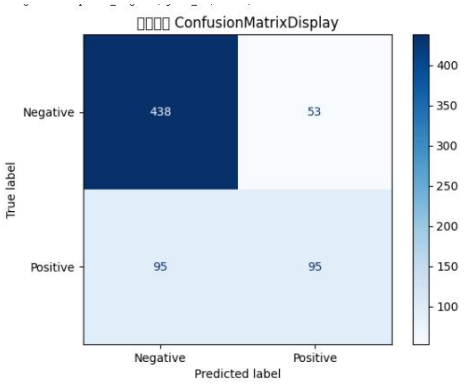


Confusion matrix:
[319 59]
[43 87]]
Accuracy: 0.80
F1-Score: 0.63
Recall: 0.67
Precision: 0.60
ROC-AUC: 0.88

4.4 AutoML LightAutoML

LightAutoML обеспечивает быстрое обучение модели и анализ важности признаков для табличных данных.

Функция потерь с усилением градиента на основе LightGBM.

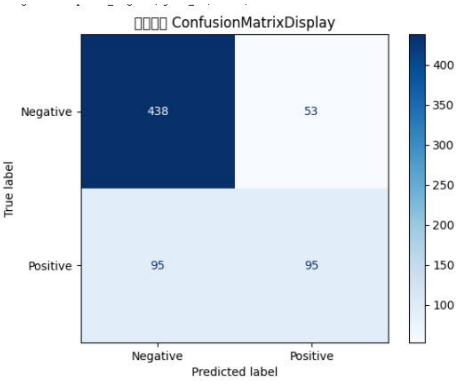


Confusion matrix:
[[438 53]
[95 95]]
F1-Score: 0.56

4.4 AutoML LightAutoML

LightAutoML 针对表格数据，提供快速模型训练和特征重要性分析。

基于 LightGBM 的梯度提升损失函数。



Confusion matrix:
[[438 53]
[95 95]]

Accuracy: 0.78
Recall: 0.50
Precision: 0.64

F1-Score: 0.56
Accuracy: 0.78
Recall: 0.50
Precision: 0.64

4.5 AutoML GAMA

GAMA использует генетический алгоритм для поиска наилучшего конвейера машинного обучения.

Confusion matrix: [[645 102]
[119 169]]
F1-Score: 0.60
Accuracy: 0.79
Recall: 0.59
Precision: 0.62

4.5 AutoML GAMA

GAMA 使用遗传算法搜索最佳机器学习管道。

Confusion matrix: [[645 102]
[119 169]]
F1-Score: 0.60
Accuracy: 0.79
Recall: 0.59
Precision: 0.62

5 Заключение

1) H2O AutoML продемонстрировал наилучшие результаты по метрике Accuracy (0.80) и ROC-AUC (0.88), а также сбалансированные значения Recall и Precision, что делает его наиболее подходящим для задачи бинарной классификации Healthy_Status на основе данных ЭКГ. Это указывает на его способность эффективно классифицировать как положительные, так и отрицательные случаи.
2) LightAutoML показал более низкие результаты по F1-Score (0.56) и Recall (0.50), что ограничивает его пригодность для задач, где важно эффективно находить все положительные случаи.
3) GAMA также показал достойные результаты, но его Accuracy и Recall уступают H2O AutoML.

В итоге, H2O AutoML является лучшим выбором для задачи бинарной классификации Healthy_Status, так как он предлагает наиболее сбалансированное и эффективное решение, с высокой точностью и хорошей способностью к выявлению всех положительных случаев в данных ЭКГ.

5 结论

1) H2O AutoML 在 Accuracy（0.80）和 ROC-AUC（0.88）指标上表现最佳，同时其 Recall 和 Precision 值也较为平衡，这使其成为基于 ECG 数据 的 Healthy_Status 二分类任务的最佳选择。这表明它能够有效地对正负样本进行分类。
2) LightAutoML 的 F1-Score（0.56）和 Recall（0.50）较低，这限制了它在需要有效识别所有正样本的任务中的应用。
3) GAMA 也表现出了不错的结果，但其 Accuracy 和 Recall 不及 H2O AutoML。

总的来说，H2O AutoML 是 Healthy_Status 二分类任务的最佳选择，因为它提供了最平衡和高效的解决方案，具有较高的准确度和较强的正样本识别能力。

5 Ссылки на литературу

References

[1] Bodini M, Rivolta M W, Sassi R. Classification of ECG signals with different lead systems using AutoML[C]//2021 Computing in Cardiology (CinC). IEEE, 2021, 48: 1-4.

5 参考文献

References

[1] Bodini M, Rivolta M W, Sassi R. Classification of ECG signals with different lead systems using AutoML[C]//2021 Computing in Cardiology (CinC). IEEE, 2021, 48: 1-4.