Отчёт

报告

Джин Хэ
金赫

12 апреля 2025г.
2025 年 4 月 12 日

1 Цель работы

Целью данного эксперимента является классификация состояния здоровья (Healthy_Status) с использованием данных электрокардиограммы (ЭКГ) с использованием подхода автоматизированного машинного обучения (AutoML). Состояние здоровья — это проблема бинарной классификации, разделенная на здоровое (1) и нездоровое (0), и цель состоит в том, чтобы разработать эффективную модель прогнозирования для определения состояния здоровья человека.

Исследуй фреймворки AutoML, обоснуй выбор лучшего. Используй его для для бинарной классификации. Постройте матрицу ошибок (confusion matrix) и рассчитайте F1-метрику для оценки обученного классификатора по признаку Healthy_Status на основе данных параметров ЭКГ.

2 Метод

(1) Предварительная обработка данных;

(2) Разработка функций;

1 目标

本实验的目标是利用心电图（ECG）数据，通过自动化机器学习（AutoML）方法对健康状态（Healthy_Status）进行分类。健康状态是一个二分类问题，分为健康（1）和非健康（0），旨在开发一个高效的预测模型，以判断个体的健康状况。

研究不同的 AutoML 框架，论证选择最佳框架的理由。使用选定的框架进行二元分类。构建混淆矩阵（confusion matrix），并根据心电参数数据，针对 Healthy_Status 标签计算分类器的 F1 分数以评估模型性能。

2 方法

(1) 数据预处理；

(2) 特征工程；

(3) Модели классификации: фреймворк AutoML, включая AutoML H2O, AutoML AutoGluon, AutoML BlueCast, AutoML Fedot, AutoML LightAutoML, AutoML GAMA и AutoML PyCaret;

(4) Индикаторы оценки

3 Обсуждение

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

3.1 AutoML H2O

H2O AutoML создает высокопроизводительные модели путем интеграции нескольких алгоритмов машинного обучения (таких как машина градиентного бустинга GBM, XGBoost, глубокое обучение и т. д.) и объединения автоматической настройки гиперпараметров и выбора модели.

Для задачи бинарной классификации модель оптимизирует потерю перекрестной энтропии:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Среди нихy_i — истинная метка, а $\hat{y}_i$ — предсказанная вероятность.

метод:

Инициализируйте среду H2O и преобразуйте данные в H2OFrame;

Установите максимальное время выполнения и случайное начальное значение

(3)分类模型：AutoML 框架，包含 AutoML H2O、AutoML BlueCast、AutoML Fedot、AutoML LightAutoML、AutoML GAMA、AutoML PyCaret；

(4)评估指标。

3 操作

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。

3.1 AutoML H2O

H2O AutoML 通过集成多种机器学习算法（如梯度提升机 GBM、XGBoost、深度学习等），结合自动化超参数调优和模型选择，构建高性能模型。

对于二分类问题，模型优化交叉熵损失：

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

其中，y_i是真实标签，$\hat{y}_i$是预测概率。

方法：

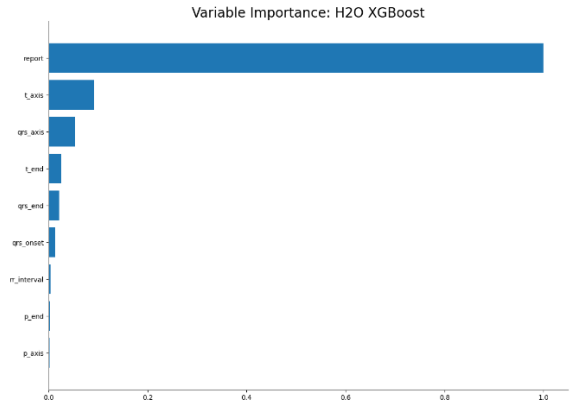
初始化 H2O 环境，转换数据为 H2OFrame；

设置最大运行时间和随机种子，使用

для обучения модели с помощью H2OAutoML;

Выберите лучшую модель из таблицы лидеров и рассчитайте матрицу ошибок и ROC-AUC.

Performance:
MSE: 0.007917322630118222
RMSE: 0.08897933822027573
MAE: 0.03818402237376397
RMSLE: 0.06496635667983598
Mean Residual Deviance: 0.007917322630118222
R^2: 0.9569257047267056
Null degrees of freedom: 514
Residual degrees of freedom: 512
Null deviance: 95.56748634029051
Residual deviance: 4.077421154510884
AIC: -1022.4249345605642



Confusion matrix:
[[390 0]
[3 122]]
F1-Score: 0.99
Accuracy: 0.99
Recall: 0.98
Precision: 1.00

3.2 AutoML BlueCast

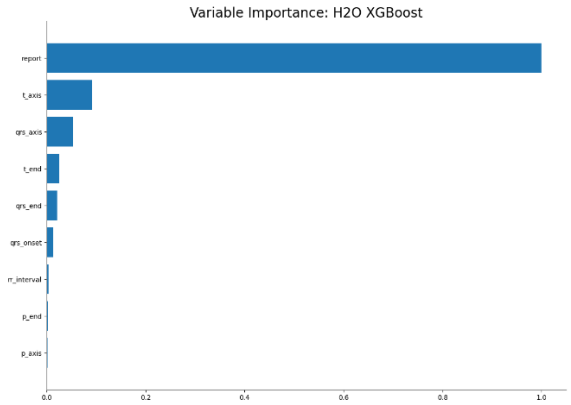
BlueCast фокусируется на задачах классификации и обеспечивает автоматизированное проектирование признаков и выбор моделей на основе XGBoost.

XGBoost оптимизирует целевую функцию:

H2OAutoML 训练模型；

从 leaderboard 中选择最佳模型，计算混淆矩阵和 ROC-AUC。

Performance:
MSE: 0.007917322630118222
RMSE: 0.08897933822027573
MAE: 0.03818402237376397
RMSLE: 0.06496635667983598
Mean Residual Deviance: 0.007917322630118222
R^2: 0.9569257047267056
Null degrees of freedom: 514
Residual degrees of freedom: 512
Null deviance: 95.56748634029051
Residual deviance: 4.077421154510884
AIC: -1022.4249345605642



Confusion matrix:
[[390 0]
[3 122]]
F1-Score: 0.99
Accuracy: 0.99
Recall: 0.98
Precision: 1.00

3.2 AutoML BlueCast

BlueCast 专注于分类问题，提供自动化特征工程和基于 XGBoost 的模型选择。

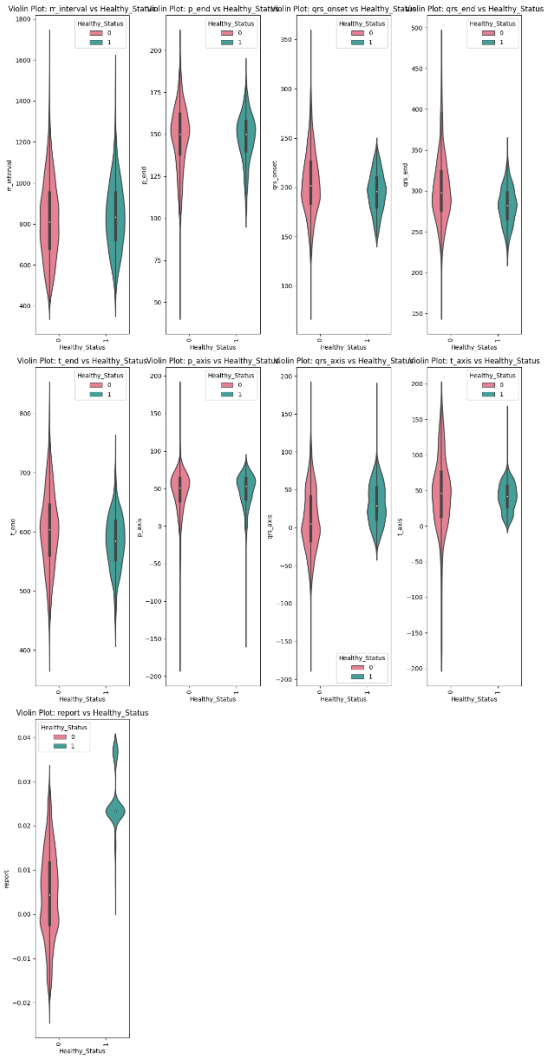
XGBoost 优化目标函数：

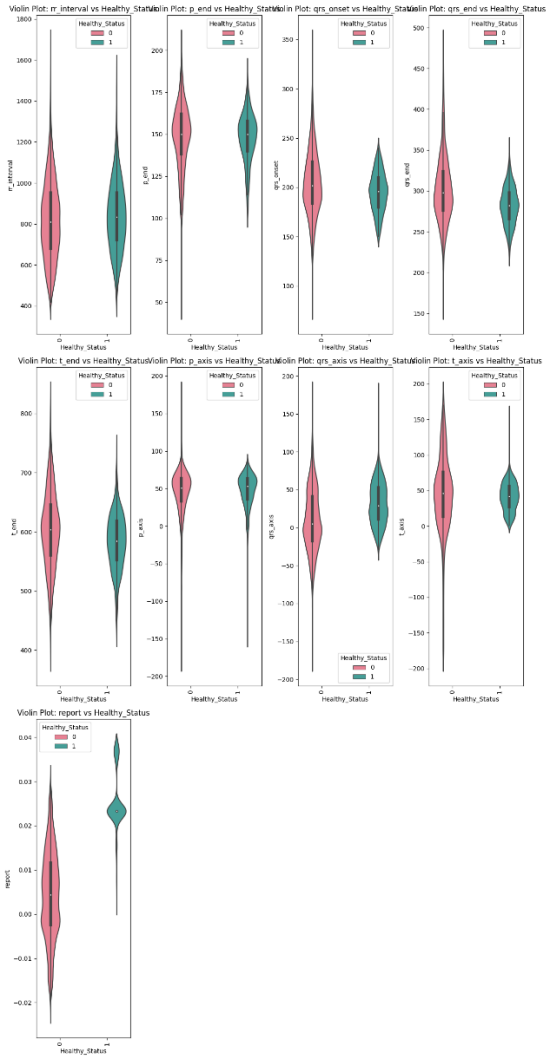
$$Obj = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

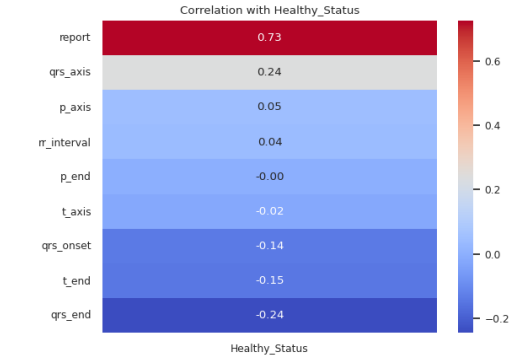
Среди них l — функция потерь, а Ω — член регуляризации.

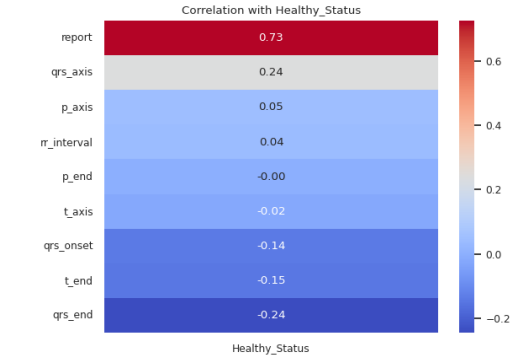
$$Obj = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

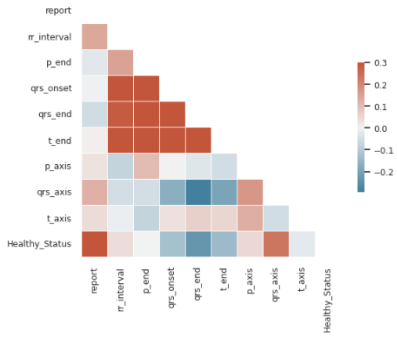
其中，l是损失函数，Ω是正则化项。

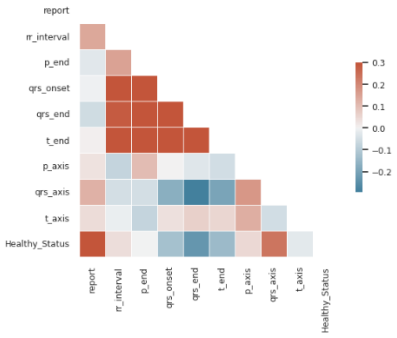


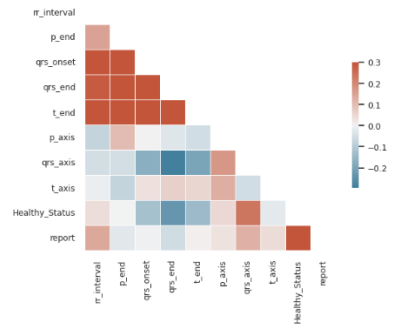




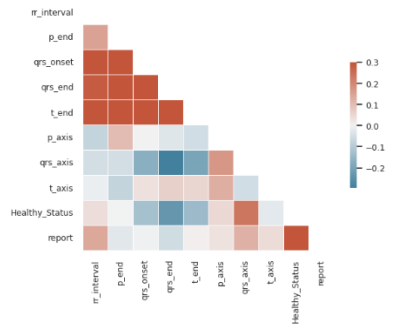




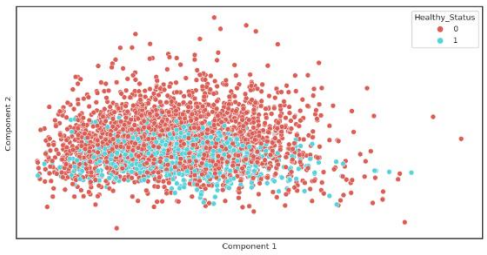




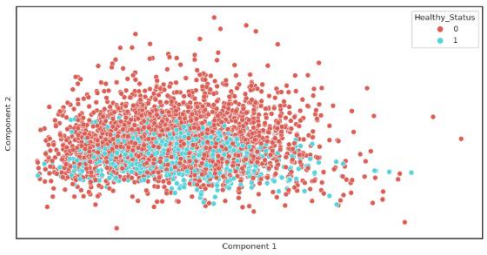
PCA
explained variance :0.88



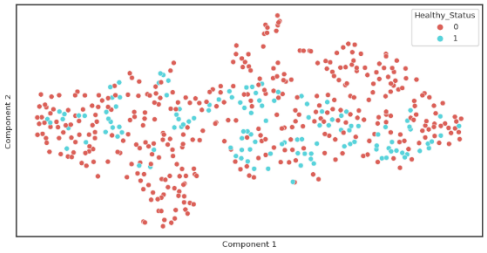
PCA
explained variance :0.88

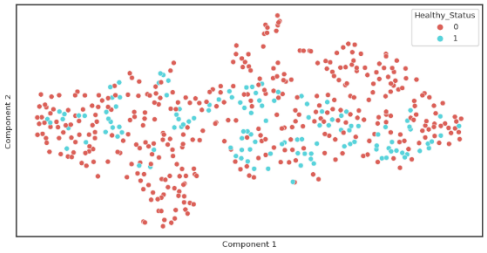


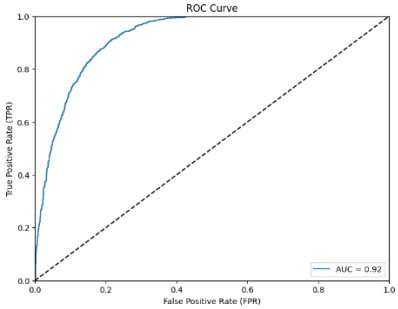
t-SNE

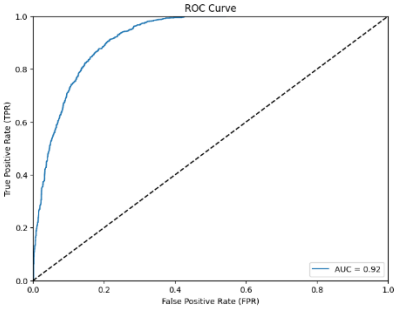


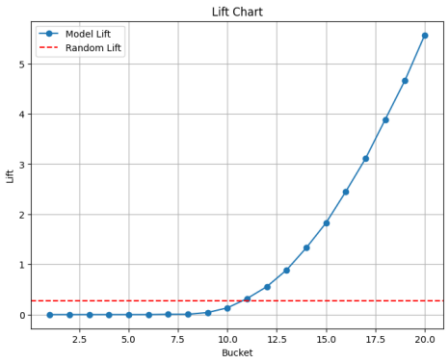
t-SNE

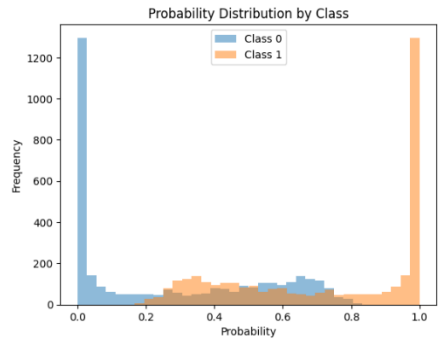










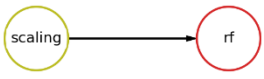


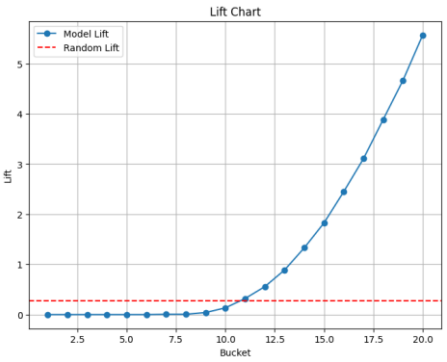
Confusion matrix: [[334 40]
[35 109]]
F1-Score: 0.74
Accuracy: 0.86
Recall: 0.76
Precision: 0.73

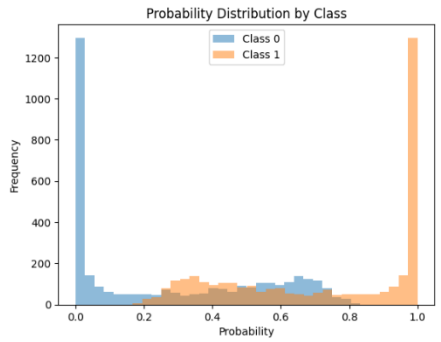
3.3 AutoML Fedot

Fedot использует генетическое программирование для автоматического построения конвейеров машинного обучения.

Генетические алгоритмы оптимизируют функцию приспособленности (например, точность классификации).





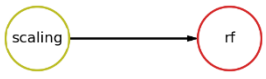


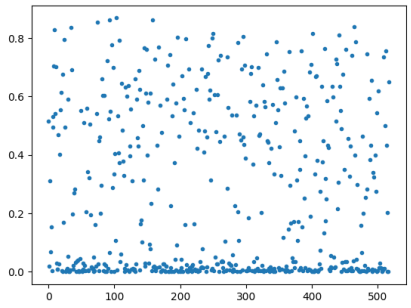
Confusion matrix: [[334 40]
[35 109]]
F1-Score: 0.74
Accuracy: 0.86
Recall: 0.76
Precision: 0.73

3.3 AutoML Fedot

Fedot 使用遗传编程（Genetic Programming）自动构建机器学习管道。

遗传算法优化适应度函数（如分类准确率）。





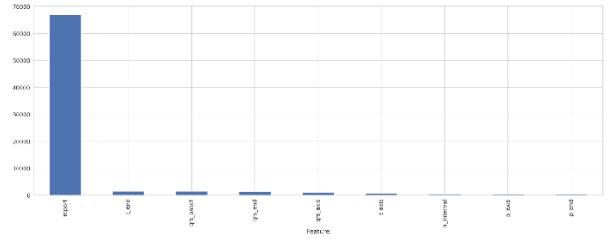
Confusion matrix: [[353 21]
[25 119]]
F1-Score: 0.84
Accuracy: 0.91
Recall: 0.83
Precision: 0.85

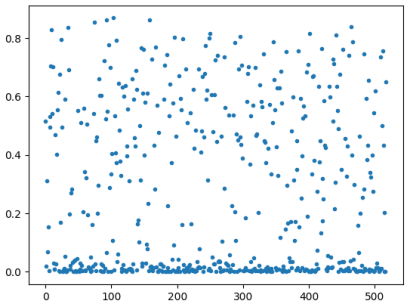
3.4 AutoML LightAutoML

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

Функция потерь с усилением градиента на основе LightGBM.

Confusion matrix: [[457 41]
[85 107]]
F1-Score: 0.63
Accuracy: 0.82
Recall: 0.56
Precision: 0.72





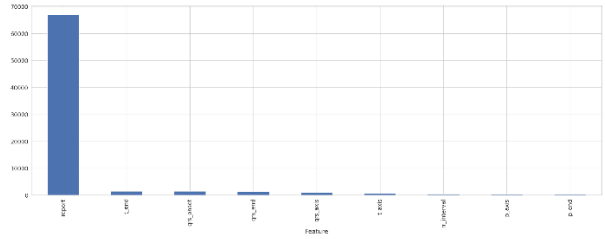
Confusion matrix: [[353 21]
[25 119]]
F1-Score: 0.84
Accuracy: 0.91
Recall: 0.83
Precision: 0.85

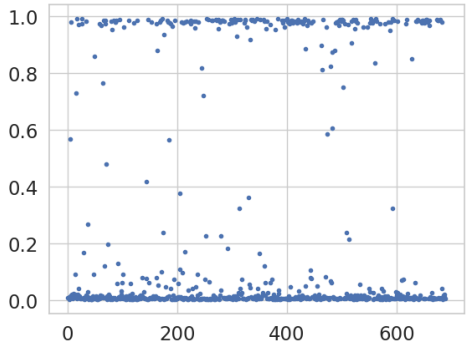
3.4 AutoML LightAutoML

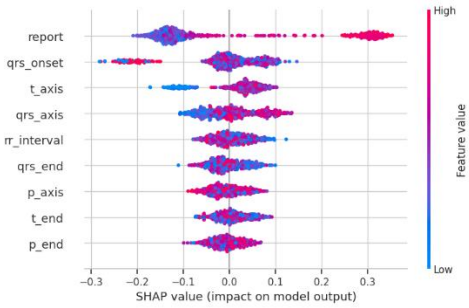
LightAutoML 针对表格数据，提供快速模型训练和特征重要性分析。

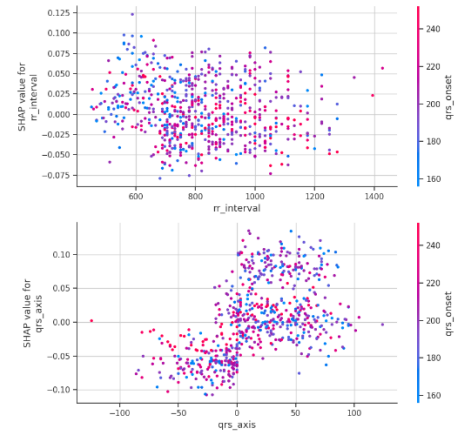
基于 LightGBM 的梯度提升损失函数。

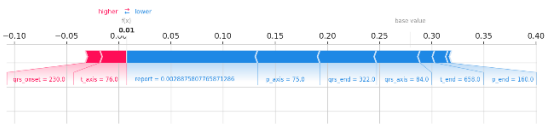
Confusion matrix: [[457 41]
[85 107]]
F1-Score: 0.63
Accuracy: 0.82
Recall: 0.56
Precision: 0.72











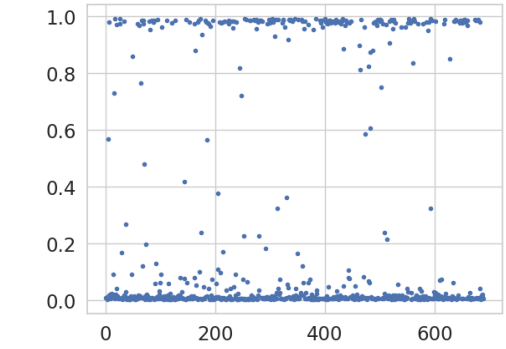
3.5 AutoML GAMA

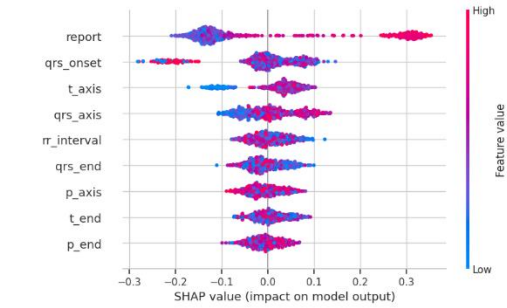
GAMA использует генетический алгоритм для поиска наилучшего конвейера машинного обучения.

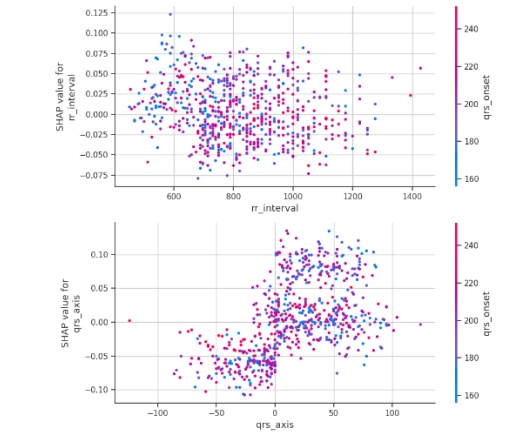
Оптимизация потерь журнала:

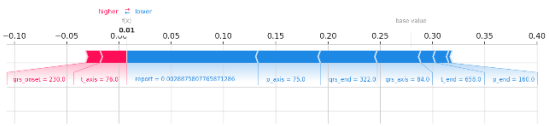
$$LL = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right)$$

accuracy: 0.9777777777777777









3.5 AutoML GAMA

GAMA 使用遗传算法搜索最佳机器学习管道。

优化对数损失：

$$LL = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right)$$

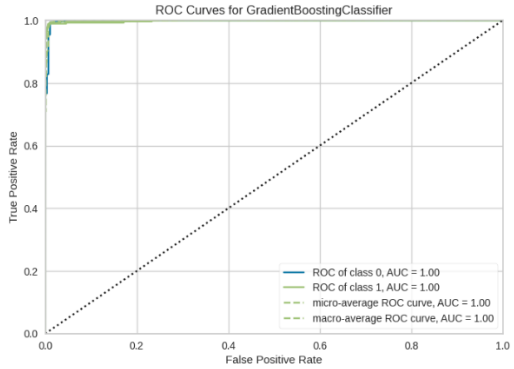
accuracy: 0.9777777777777777

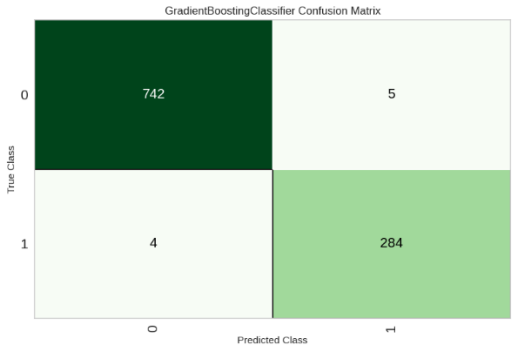
log loss: 0.08609400351663057
log_loss 0.08609400351663057
Confusion matrix: [[645 102]
 [119 169]]
F1-Score: 0.60
Accuracy: 0.79
Recall: 0.59
Precision: 0.62

3.6 AutoML PyCaret

PyCaret — это библиотека машинного обучения с низким объемом кода, которая оптимизирует производительность посредством сравнения моделей и настройки гиперпараметров.

Оптимизация индикатора на основе перекрестной проверки.





4 Ссылки на литературу

References

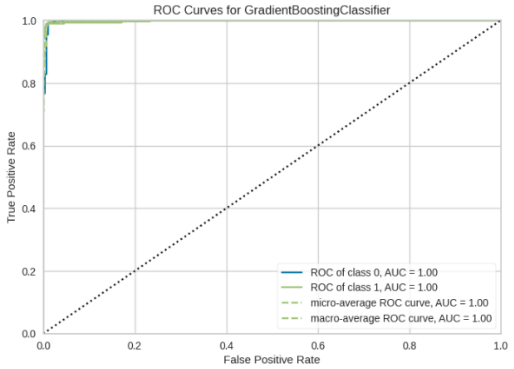
[1] Bodini M, Rivolta M W, Sassi R. Classification of ECG signals with different lead systems using AutoML[C]//2021 Computing in Cardiology (CinC). IEEE, 2021, 48: 1-4.

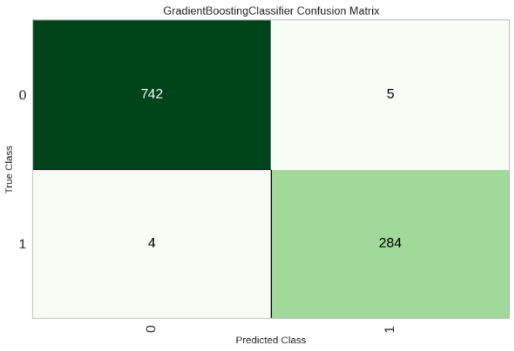
log loss: 0.08609400351663057
log_loss 0.08609400351663057
Confusion matrix: [[645 102]
 [119 169]]
F1-Score: 0.60
Accuracy: 0.79
Recall: 0.59
Precision: 0.62

3.6 AutoML PyCaret

PyCaret 是一个低代码机器学习库，通过模型比较和超参数调优优化性能。

基于交叉验证的指标优化。





4 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.

[2] Shevchenko A D, Bukhov A K, Skvortsova M A, et al. Analysis of ECG Data Using AutoML Frameworks to Predict the Classification of Some Cardiovascular Disease Features[C]//2025 7th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE). IEEE, 2025: 1-6.
[3] Shin S, Park D, Ji S, et al. Medical Data Analysis Using AutoML Frameworks[J]. Journal of Electrical Engineering & Technology, 2024, 19(7): 4515-4522.
[2] Shevchenko A D, Bukhov A K, Skvortsova M A, et al. Analysis of ECG Data Using AutoML Frameworks to Predict the Classification of Some Cardiovascular Disease Features[C]//2025 7th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE). IEEE, 2025: 1-6.
[3] Shin S, Park D, Ji S, et al. Medical Data Analysis Using AutoML Frameworks[J]. Journal of Electrical Engineering & Technology, 2024, 19(7): 4515-4522.