20250242-0: 人工智能基础

第二次编程作业：信用数据分析问题

主讲老师：江 瑞　　负责助教：王子安

1 题目介绍

背景介绍：德国信用数据集（German Credit Data）是金融风险分析领域的经典数据集，用于评估贷款申请人的违约风险。该数据集包含 1000 个样本，每个样本代表一位贷款申请人的信用记录，通过 20 个特征（7 个数值型，13 个类别型）预测申请人属于良好信用（Good）或不良信用（Bad）两类。数据特征涵盖人口统计信息（如年龄、性别）、财务状况（如信用金额、存款状态）以及历史信用行为（如信用历史、还款状态）。本次作业中你要作为一个数据分析师对该数据进行建模，分析和讨论。

题目叙述：

1. 参看课件，推导用随机梯度下降法求解二元 Logistic 回归的过程。
2. 对于德国信用数据集进行数据预处理，使最终输入数据包含 20 个特征。采用独立测试集划分的方式，划分数据集为训练集、验证集（可选）和测试集。使用 Logistic 回归算法求解该问题，在测试集上评估申请人的信用状况。
3. 采用与第 2 问中相同的数据集划分方式，设计一个深度神经网络用于解决该信用评估问题，在测试集上预测申请人是否信用良好。选择两个超参数进行实验分析，讨论超参数选择对模型性能的影响。推荐使用优化策略（如 early stopping、正则化等）以提升模型的泛化能力和预测性能。
4. 在测试集上使用多种评估指标（如 Accuracy、Sensitivity、Specificity、Recall、Precision、F1、auROC 等）展示模型性能，并比较 Logistic 回归与深度神经网络的最终结果。针对不同方法在各指标上的表现，提出合理假设来解释为何一种方法的指标较高而另一种较低。
5. （选做）机器学习或深度学习模型可能在预测时表现出模型偏见（Bias），例如对特定群体、类别或特征的系统性倾向，从而影响其客观性和公平性。这种偏见可能源于训练数据、算法设计或现实世界的偏差。请分析你训练的两个模型是否在年龄特征上存在偏见，并通过实验和代码支持你的观点，确保论证有理有据。

2 作业要求

（1） 本次报告要求使用 Python 语言实现，所有代码应写在一个 Jupyter Notebook 文件中（.ipynb 格式）。

（2） 请统一使用 Python 的 ucimlrepo 库导入数据。该库的使用方法请参考该链接。

（3） 报告中需对所实现算法的核心代码进行解释和说明。

（4） 问题的分析与思考是本题的重点考核内容，建议在报告中适当展开。选做题可获得附加分（总分不超过 100 分）；此外，若设计的神经网络模型预测准确率显著优于默认参数下的 Logistic 回归模型，将酌情给予额外加分（常见方法的准确率参考如图1所示）。

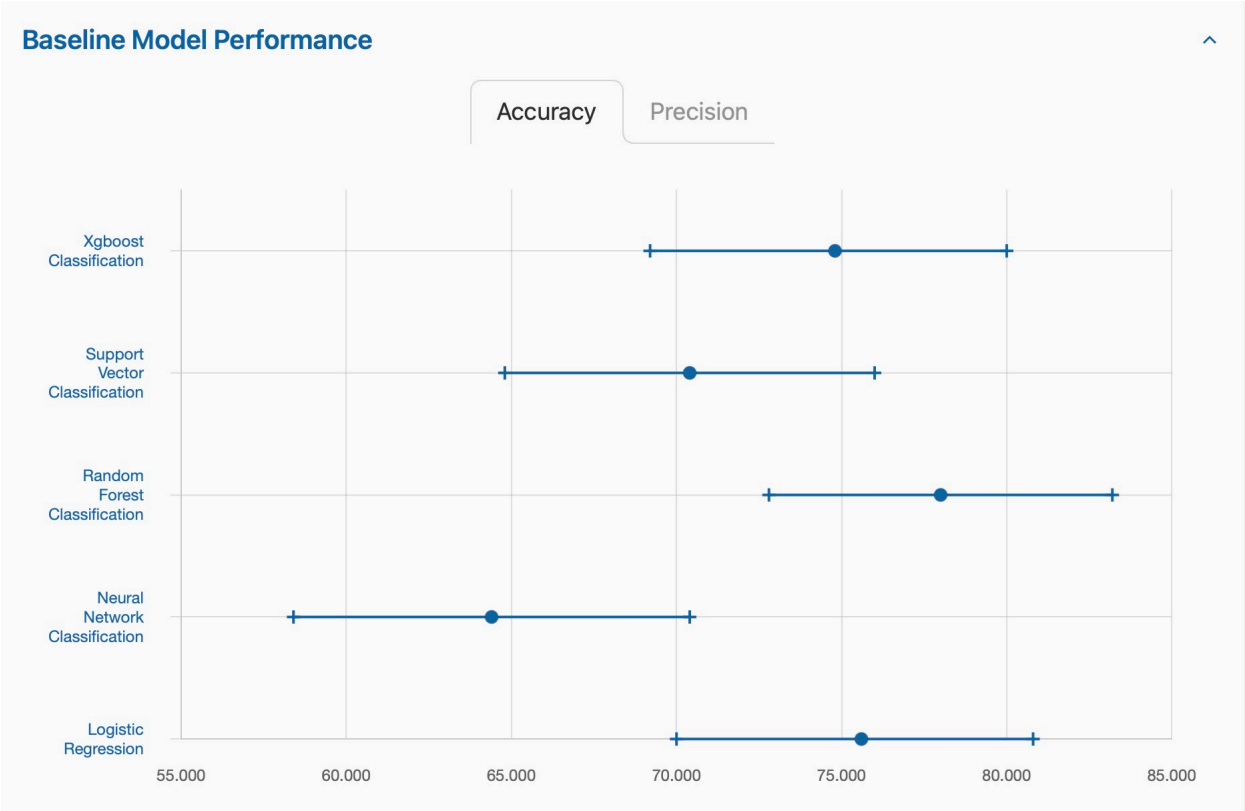

图 1: 常见方法在德国信用数据集上的预测准确率

3 提交说明

你需要写出代码，并在报告中对上述问题进行回答。提交文件格式及命名要求如下（如不按照命名规范提交会扣除少量分数）:

- 编程作业2_学号_姓名.rar（.zip）
  - hw2_学号.ipynb（代码）
  - report2_学号_姓名.pdf（pdf版报告）

本次作业截止日期: 2025 年 4 月 30 日晚 12 点（三周后）