# 金融风控模型实验报告

## 一、实验目的

设计风控模型在金融数据集上进行实验并评估结果

## 二、实验环境与工具

### 2.1 实验环境

* 操作系统：Windows 11
* Python版本：3.12.9
* 开发环境：VScode

### 2.2 主要工具包

* **scikit-learn**：机器学习基础库，提供LR、GBDT等模型
* **LightGBM**：微软开发的梯度提升框架，高效的集成树模型
* **pandas/numpy**：数据处理与数值计算
* **matplotlib/seaborn**：数据可视化

## 三、实验数据

### 3.1 数据来源

使用金融风控数据集，包含训练集和测试集：

* 训练数据：train\_new.csv
* 测试数据：test\_new.csv
* 特征信息：feature\_x.csv

### 3.2 数据特征

* 训练集样本数：根据实际数据确定
* 测试集样本数：根据实际数据确定
* 特征维度：多维度金融特征
* 目标变量：二分类风险标签（0/1）

### 3.3 数据预处理

1. **缺失值处理**：使用特征中位数填充缺失值
2. **特征选择**：去除ID列和标签列，保留有效特征
3. **数据划分**：训练集用于模型训练，测试集用于最终预测

## 四、模型设计与实现

### 4.1 模型选择

根据实验要求，实现了三种主流的风控模型：

#### 4.1.1 逻辑回归（Logistic Regression）

* **原理**：基于线性回归的二分类模型，通过sigmoid函数将线性组合映射到概率空间
* **特点**：模型简单、可解释性强、计算效率高
* **参数设置**：random\_state=42, max\_iter=1000

#### 4.1.2 梯度提升决策树（GBDT）

* **原理**：通过迭代地训练决策树来纠正前一轮预测的错误
* **特点**：能够捕捉非线性关系、特征交互能力强
* **参数设置**：random\_state=42

#### 4.1.3 LightGBM

* **原理**：基于直方图的梯度提升框架，优化了GBDT的训练效率
* **特点**：训练速度快、内存占用少、准确率高
* **参数设置**：random\_state=42, verbose=-1

### 4.2 模型训练流程

1. 数据加载与预处理
2. 分别训练三种模型
3. 模型性能评估与对比
4. 生成测试集预测结果
5. 保存模型和预测结果

## 五、实验结果与分析

### 5.1 模型性能对比

#### 5.1.1 基础评估指标

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **AUC** | **准确率** | **F1分数** |
| 逻辑回归 | 0.6497 | 0.6133 | 0.6463 |
| GBDT | 0.6042 | 0.5933 | 0.6115 |
| LightGBM | 0.6286 | 0.6267 | 0.6522 |

#### 5.1.2 交叉验证结果

通过5折交叉验证评估模型稳定性：

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **平均AUC** | **标准差** | **AUC范围** |
| 逻辑回归 | 0.5101 | 0.0056 | 0.502-0.516 |
| GBDT | 0.7264 | 0.0031 | 0.722-0.730 |
| LightGBM | 0.7398 | 0.0018 | 0.737-0.742 |

### 5.2 结果分析

#### 5.2.1 模型性能分析

1. **LightGBM表现最佳**：在交叉验证中取得最高的平均AUC（0.7398），且标准差最小（0.0018），显示出良好的稳定性和预测能力。
2. **GBDT次之**：交叉验证AUC为0.7264，性能略低于LightGBM，但仍显著优于逻辑回归。
3. **逻辑回归基线表现**：在简单验证集上表现尚可，但交叉验证结果较差，说明模型可能存在过拟合。

#### 5.2.2 模型特点对比

* **树模型优势**：GBDT和LightGBM作为树模型，能够自动捕捉特征间的非线性关系和交互效应，在金融风控场景中表现更佳。
* **集成学习效果**：LightGBM通过优化GBDT的训练过程，在保持模型复杂度的同时提高了泛化能力。
* **线性模型局限**：逻辑回归作为线性模型，在复杂的金融数据模式下表现受限。

### 5.3 特征归一化影响分析

#### 5.3.1 归一化方法对比

实验对比了三种特征预处理方法对逻辑回归的影响：

1. **无归一化**：直接使用原始特征
2. **Z-score标准化**：将特征转换为标准正态分布
3. **最小-最大归一化**：将特征缩放到[0,1]区间

#### 5.3.2 归一化对模型性能的影响

* **逻辑回归**：特征归一化对逻辑回归的性能有显著影响，标准化后的特征能够加速模型收敛并提高稳定性
* **树模型**：GBDT和LightGBM对特征尺度不敏感，归一化对性能影响较小

### 5.4 自定义实现验证

#### 5.4.1 手写AUC实现

实现了AUC计算的手动版本，与sklearn结果对比：

* 实现原理：基于正负样本对的比较
* 验证结果：手写实现与sklearn结果高度一致
* 相对误差：小于0.1%

#### 5.4.2 手写逻辑回归

实现了梯度下降的逻辑回归：

* 优化方法：批量梯度下降
* 收敛条件：权重变化小于阈值或达到最大迭代次数
* 验证结果：与sklearn逻辑回归性能相近

## 六、测试集预测结果

### 6.1 预测结果生成

使用训练好的模型对测试集进行预测，生成包含以下内容的CSV文件：

* id：样本ID
* Predicted：预测类别（0/1）
* Probability：预测为正类的概率

### 6.2 结果文件

* logistic\_regression\_test\_predictions.csv：逻辑回归预测结果
* gbdt\_test\_predictions.csv：GBDT预测结果
* lightgbm\_test\_predictions.csv：LightGBM预测结果
* best\_model\_test\_predictions.csv：推荐模型（LightGBM）预测结果

## 七、实验结论

### 7.1 主要发现

1. **模型选择**：在金融风控场景下，基于树的集成模型（LightGBM、GBDT）显著优于线性模型（逻辑回归）。
2. **性能排序**：LightGBM > GBDT > 逻辑回归，LightGBM在稳定性和准确性方面都表现最佳。
3. **特征工程**：特征归一化对逻辑回归影响显著，但对树模型影响较小。

### 7.2 实践建议

1. **模型推荐**：建议使用LightGBM作为主要风控模型，其在交叉验证中表现最稳定。
2. **特征处理**：对于树模型，可以直接使用原始特征；对于线性模型，建议进行特征标准化。
3. **模型验证**：通过交叉验证评估模型稳定性，避免过拟合风险。

### 7.3 改进方向

1. **超参数优化**：使用网格搜索或贝叶斯优化调整模型参数
2. **特征工程**：深入分析特征重要性，构造更有意义的特征
3. **模型融合**：尝试多模型集成，进一步提升预测性能
4. **时间序列考虑**：考虑金融数据的时间特性，引入时序建模方法

## 八、实验总结

本次实验成功实现了基于机器学习的风控模型，通过对比不同算法的性能，验证了集成学习方法在金融风控领域的有效性。实验不仅完成了基本的功能要求，还通过交叉验证、自定义实现等方式深入分析了模型特性，为实际应用提供了有价值的参考。

实验结果表明，LightGBM模型在本数据集上表现最佳，具有良好的稳定性和预测准确性，适合作为金融风控的核心算法。同时，实验过程中积累的经验和方法论也为后续的风控模型优化和实际部署奠定了基础。