**数据仓库与数据挖掘**

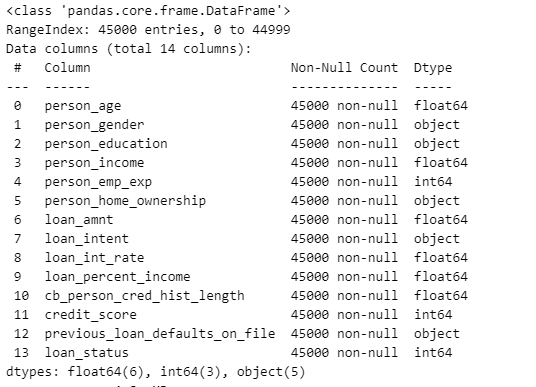
**大作业报告**

**组 员： 王延吉、曾柏文、杨晨、杨雨阳、张梓良**

**提交日期： 2024-12-26**

# 1分析数据集概述

**（一）属性名称与属性类型**



person\_age：贷款申请人的年龄。

person\_gender：贷款申请人的性别。

person\_education：贷款申请人的教育背景。

person\_income：贷款申请人的年收入。

person\_emp\_exp：贷款申请人的工作年限。

person\_home\_ownership：贷款申请人的住房所有权状况。

loan\_amnt：贷款金额。

loan\_intent：贷款用途。

loan\_int\_rate：贷款利率。

loan\_percent\_income：贷款金额占收入的比例。

cb\_person\_cred\_hist\_length：贷款申请人的信用历史长度。

credit\_score：贷款申请人的信用评分。

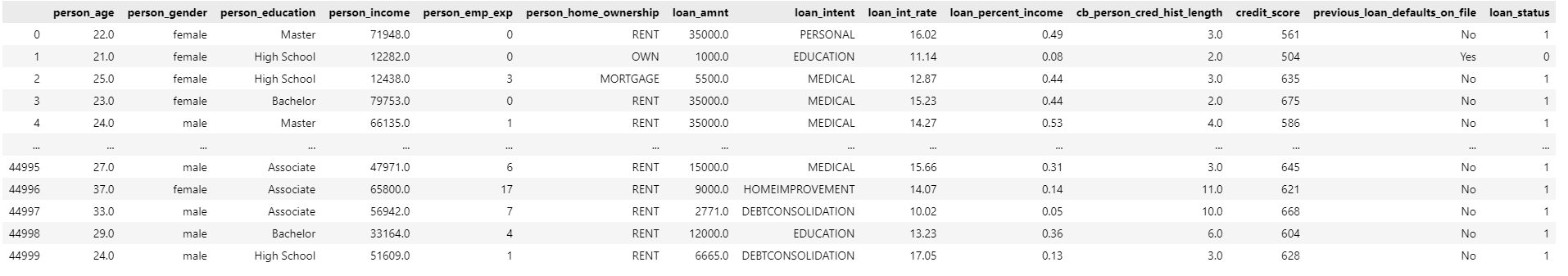
previous\_loan\_defaults\_on\_file：贷款申请人历史违约记录的数量。

loan\_status：贷款审批结果。

**（二）数据规模**

数据集总共有45000条数据

**（三）数据样例**



# 2分析目标

**数据分析预期挖掘目标**

**1.数据预处理：**

检查和处理缺失值：通过填充或删除缺失值来确保数据的完整性。

数据清洗：对异常值进行处理，确保数据质量。

**2.特征工程：**

对类别变量（如 person\_gender、person\_education、person\_home\_ownership）进行编码，以便模型能够处理。

对数值变量（如 person\_age、person\_income、loan\_amnt 等）进行标准化或归一化处理，以提高模型性能。

**3.模型选择与训练：**

选择合适的机器学习模型进行训练。使用了逻辑回归 (LogisticRegression) 、支持向量机 (SVC) 、XGBoost、LightGBM、CatBoost等算法进行建模。

使用 train\_test\_split 方法将数据集分为训练集和测试集，以评估模型的泛化能力。

**4.模型评估：**

使用多种评估指标对模型进行评估，包括准确率 (accuracy\_score)、混淆矩阵 (confusion\_matrix)、分类报告 (classification\_report)、ROC 曲线 (roc\_auc\_score 和 roc\_curve) 等。

根据评估结果调整模型参数，优化模型性能。

**5.模型解释与应用：**

解释模型预测的结果，理解哪些特征对贷款状态有重要影响。

将模型应用于新的数据集，预测贷款状态，为决策提供支持。

**挖掘什么样的模型**

**识别关键特征：**

分析各特征（如收入水平、信用评分、贷款金额等）对贷款审批结果的影响，识别关键特征。通过特征重要性分析，筛选出对模型预测贡献最大的特征。

**处理类别不平衡：**

解决贷款通过和拒绝样本比例不均衡的问题，提升模型对少数类（贷款拒绝）的识别能力。

**构建高性能模型：**

构建一个能够准确预测贷款审批结果的分类模型，优化模型的 ROC-AUC、准确率、召回率和 F1 分数。

**提升模型泛化能力：**

通过模型融合和参数优化，提升模型在新数据上的泛化能力，降低过拟合风险。

**增强模型可解释性：**

提供特征重要性分析和可视化结果，增强模型的可解释性，支持决策制定。

# 3分析流(Python程序)设计与实现

**1.导入的库**

**1matplotlib.pyplot**: 用于数据可视化，绘制静态图表。

**plotly.express**: 提供一个高级接口，用于创建交互式图表。

**seaborn**: 基于matplotlib，提供更美观的统计图形。

**pandas**: 用于数据操作和分析，提供数据结构和数据分析工具。

**numpy**: 用于科学计算，提供多维数组对象和各种派生对象。

**sklearn.model\_selection**: 提供数据集划分工具，如train\_test\_split。

**sklearn.linear\_model**: 提供线性模型，如LogisticRegression。

**sklearn.svm**: 提供支持向量机模型，如SVC。

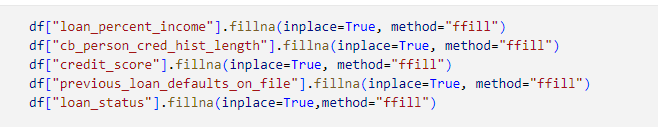
**sklearn.preprocessing**: 提供数据预处理工具，如StandardScaler和MinMaxScaler。

**sklearn.metrics**: 提供模型评估工具，如accuracy\_score、confusion\_matrix等。

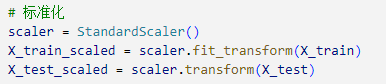
**warnings**: 用于忽略警告信息。

**2.数据预处理**

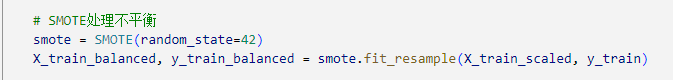
**缺失值填充**



**特征缩放**

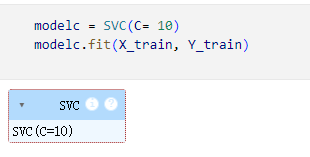


**处理类别不平衡**



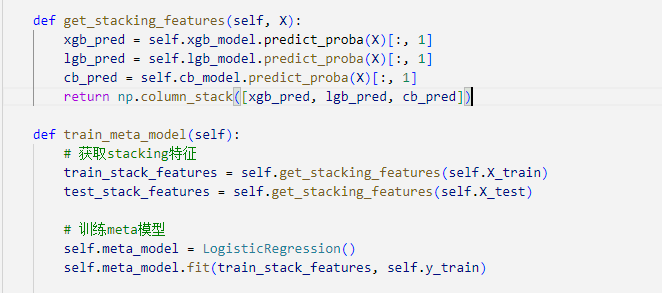
1. **模型训练**

**基础模型训练**



# **复合模型训练**

# 

会导入以下库：  
xgboost、lightgbm、catboost：用于构建基模型。

sklearn.linear\_model：用于构建元模型。

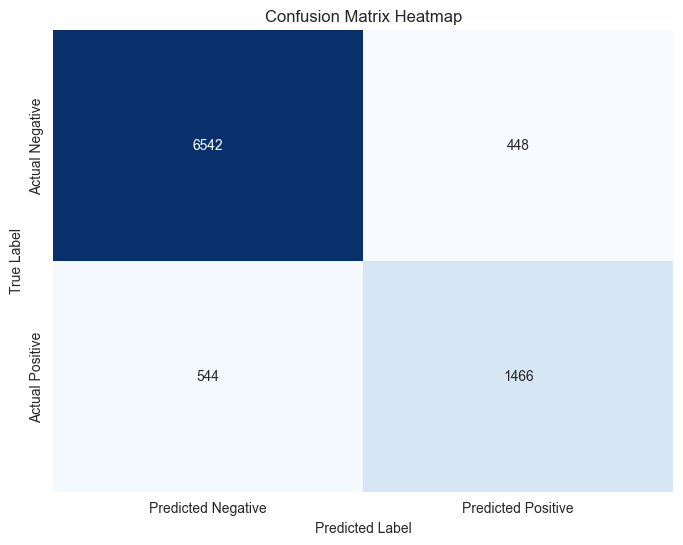
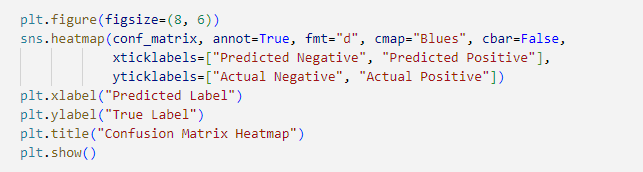
sklearn.ensemble：用于模型融合。

构建基模型（XGBoost、LightGBM、CatBoost）和元模型（逻辑回归）。

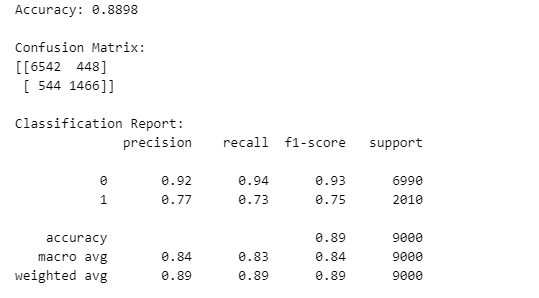
使用 Stacking 模型融合策略，结合基模型的优势。

**4.模型评估**

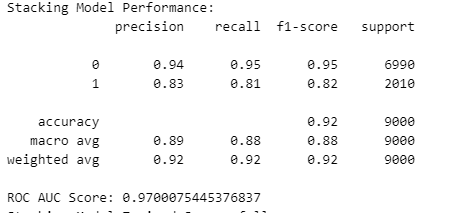
**绘制混淆矩阵**



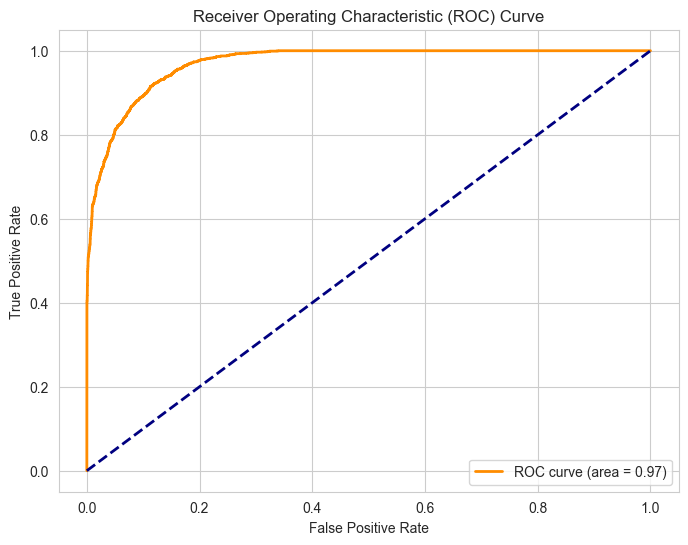
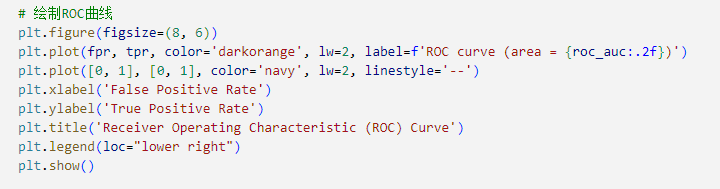
**基础模型性能**



**复合模型性能**

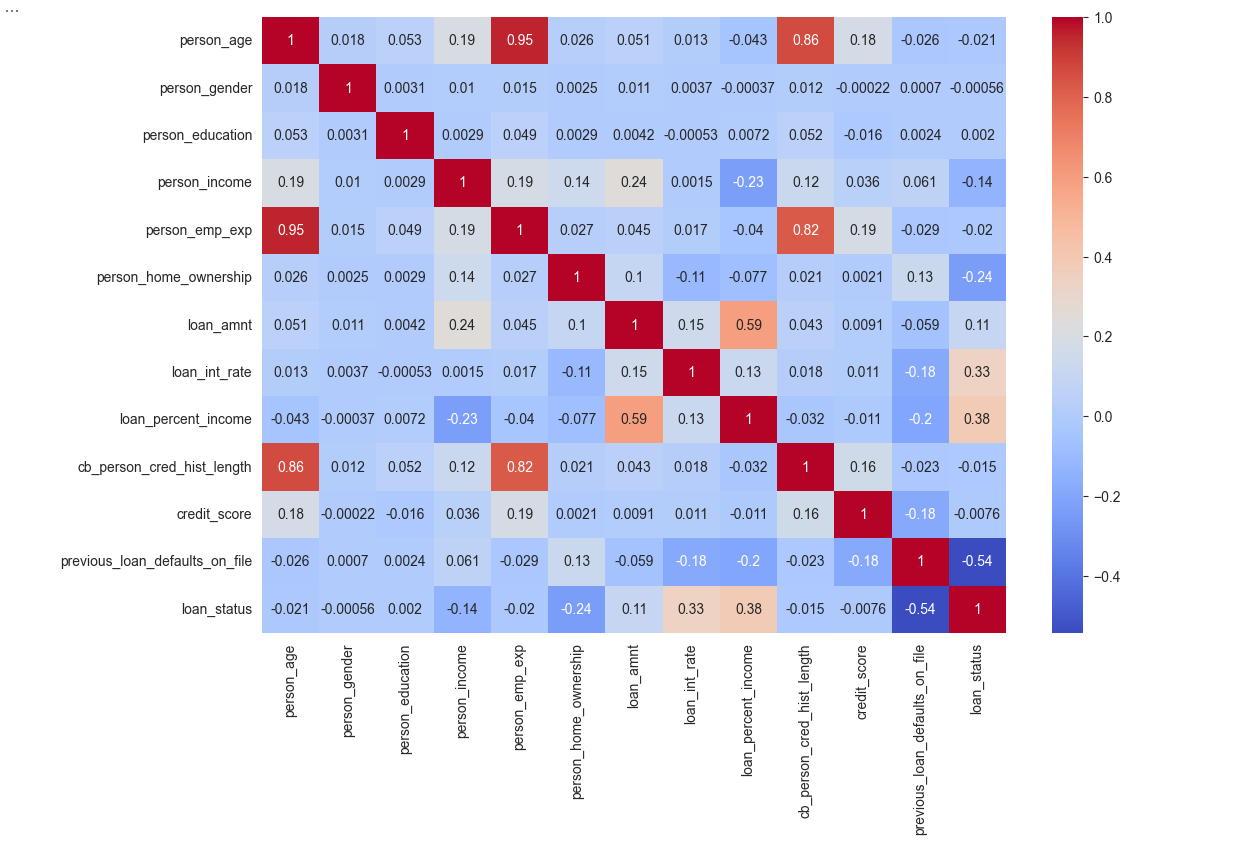


**绘制ROC曲线**

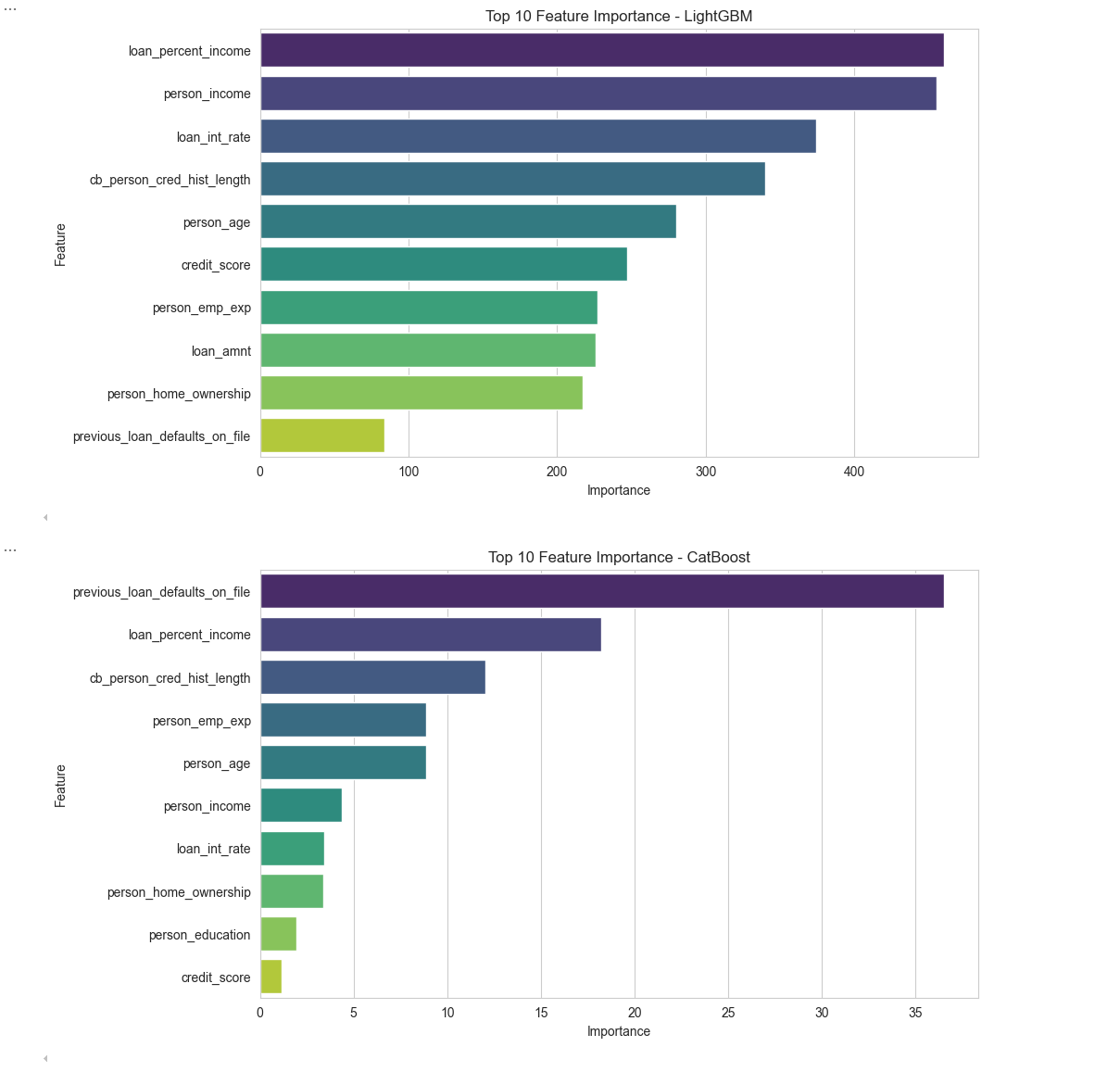


**5.数据分析**

**相关性分析**



**特征重要性分析**



# 4数据分析总结

**1.感悟**

**数据预处理是关键：**

数据质量直接影响模型性能。通过处理缺失值、标准化特征和解决类别不平衡问题，模型的表现显著提升。这让我深刻认识到数据预处理在数据分析中的重要性。

**特征工程是提升性能的核心：**

通过构造新特征（如贷款收入比）和筛选关键特征，模型性能得到了显著提升。这让我明白，特征工程不仅需要技术能力，还需要对业务背景的深刻理解。

**模型融合的优势：**

单一模型的性能有限，而通过 Stacking 模型融合策略，我们成功结合了多个基模型的优势，显著提升了模型的泛化能力和稳定性。这让我认识到，模型融合是解决复杂问题的有效方法。

**可视化增强可解释性：**

通过绘制 ROC 曲线、特征重要性图和混淆矩阵，我们直观展示了模型的表现和决策依据。这不仅帮助我更好地理解模型，也为项目的汇报和展示提供了有力支持。

**持续学习与改进：**

数据科学是一个不断学习和改进的过程。无论是新算法的尝试，还是对业务知识的深入理解，都需要持续学习和实践。这让我更加坚定了不断学习和提升自己的决心。

**2. 不足**

**特征工程的局限性：**

在特征工程中，我们主要依赖于已有的特征和简单的特征构造方法，未能充分利用外部数据或更复杂的特征工程技术。未来可以探索更多特征构造方法，如时间序列特征或外部数据融合。

**模型泛化能力的验证：**

虽然模型在测试集上表现优异，但在实际业务场景中的表现仍需验证。未来可以将模型部署到实际业务中，验证其在新数据上的泛化能力。

**模型可解释性的提升：**

尽管我们进行了特征重要性分析，但模型的决策过程仍不够透明。未来可以尝试使用 SHAP 值等可解释性工具，进一步提升模型的可解释性。

**类别不平衡问题的进一步优化：**

虽然我们使用 SMOTE 处理了类别不平衡问题，但模型对少数类的识别能力仍有提升空间。未来可以尝试其他过采样或欠采样技术，进一步优化模型性能。

**总结**

通过这次贷款审批分类模型的项目，我们不仅提升了技术能力，还对数据科学的应用有了更深刻的理解。我深刻体会到，数据科学不仅是技术的堆砌，更是对业务问题的深入思考和创造性解决。尽管项目取得了一定的成果，但仍有许多不足之处需要改进。未来，我将继续努力，不断提升自己，为更多实际业务问题提供高效的解决方案。

# 5分工说明

小组分工合理，每人都完成了几乎相同的代码量。

# 6 参考文献

注: 如果分析方法和步骤参考某些文献,也请列出:作者;文献名称;出处;年