**德国信用评分系统报告**

前言：此报告是关于凯捷中国JTP德国信用评分系统的数据分析和信用评分系统建立。代码为 100% Jupiter Notebook， 整个项目以在Github上传。项目主要内容和过程由Eric导师指导，实习生邹培炎进行实现。

相关链接： [**https://github.com/peiyan03/Credit-Score-Project**](https://github.com/peiyan03/Credit-Score-Project)

目录：

1.1 数据描述性和关联性分析

1.2 初步指标变量选择

2.1 分箱

2.2 二次指标变量筛选

3.1 模型构造

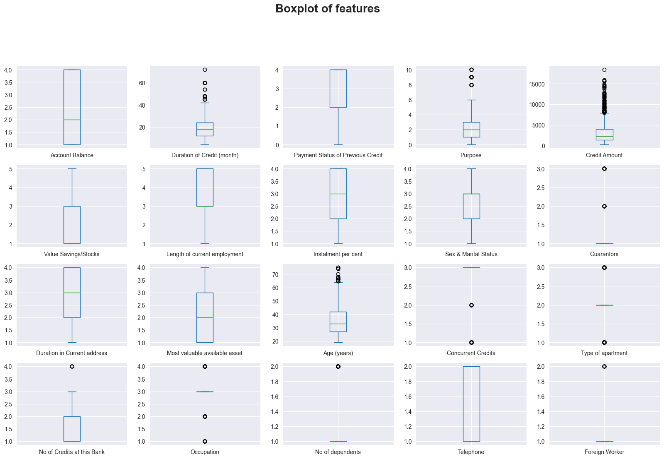
3.2 模型验证

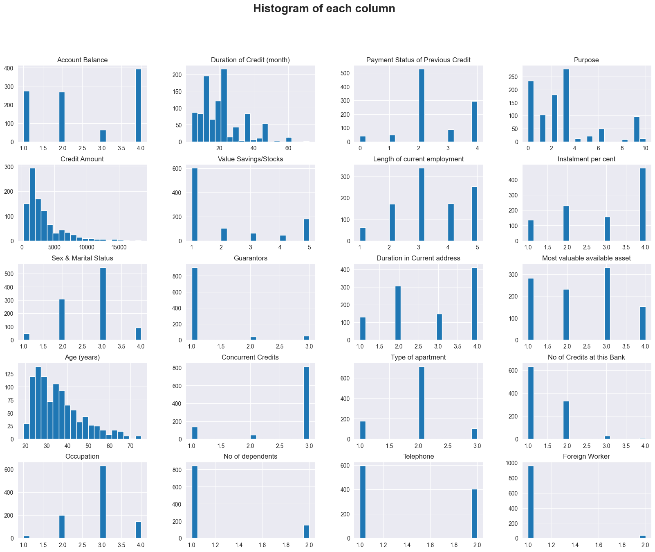
4. 评分卡建立

5. 信用评分数计算

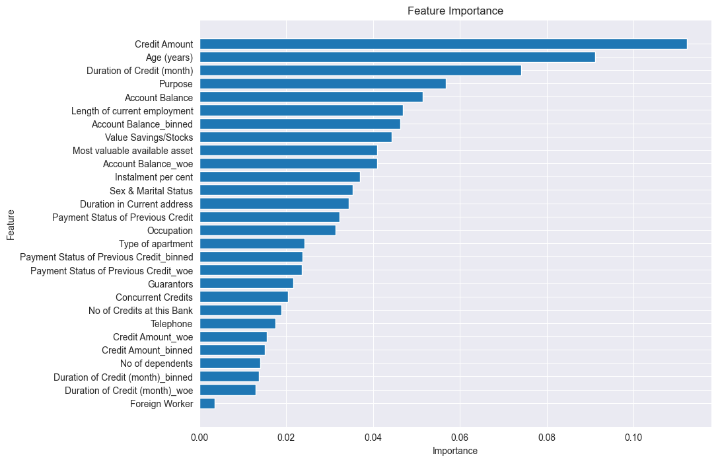
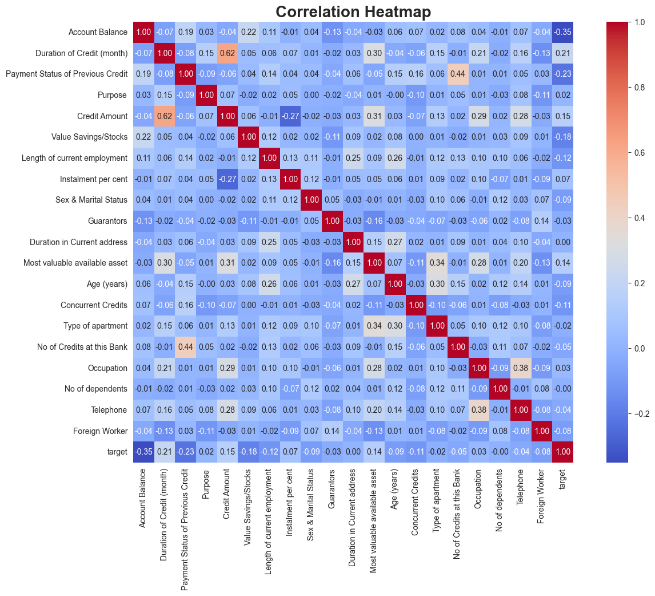
**6. 总结**

1.1 数据描述性和关联性分析

首先，我对初始数据集进行了全面的描述性分析，通过绘制箱形图和直方图来直观地展示每个变量的数据结构和分布情况。这些图表为我提供了对数据集中各变量特征的初步理解，有助于识别数据中的潜在异常值和分布特征。



在此基础上，我生成了数据的相关性热图，以评估各个变量与目标变量“target”之间的相关性。这一分析旨在帮助我们识别出对目标变量具有显著正面或负面影响的特征，为后续的变量筛选和分箱过程奠定基础。

此外，我还应用了随机森林模型（Random Forest Model）来计算每个变量对目标变量的重要性（影响值）。随机森林模型作为一种基于集成学习的方法，能够有效地评估各特征在预测中的贡献度，从而为变量筛选提供科学依据。

最后，将两图的数据中每个影响力值进行排序，筛选出每张图中前十的变量。

1.2 初步指标变量选择

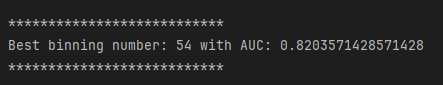
我在这里合并了两组排名前十的变量，以下为最后选择进行下一部分箱的指标：

*['Sex & Marital Status', 'Account Balance', 'Concurrent Credits', 'Value Savings/Stocks', 'Payment Status of Previous Credit', 'Credit Amount', 'Most valuable available asset', 'Duration of Credit (month)', 'Length of current employment', 'Age (years)', 'Purpose', 'Instalment per cent']*

2.1 分箱

为了优化变量分箱的数量，从而提高后续模型构建和训练的准确性，我采用了基础的逻辑回归模型，将数据集按8:2的比例分为训练集和测试集并运用均衡分箱决策（Uniform）。通过训练模型，我评估了不同分箱方案下模型的性能，并记录了对应的AUC值（Area Under the Curve）。AUC是ROC曲线（受试者工作特征曲线）下面积的度量，ROC曲线描述了模型在不同阈值下的真阳性率（True Positive Rate）与假阳性率（False Positive Rate）之间的关系。AUC值越接近1，表示模型区分正负样本的能力越强。

在进行多种分箱方案的比较后，我最终选择了具有最高AUC值（0.82）的54个分箱方案。这意味着每个连续变量将被分成54个宽度相等的区间，从而确保分箱后的变量在模型中具有最佳的区分能力。



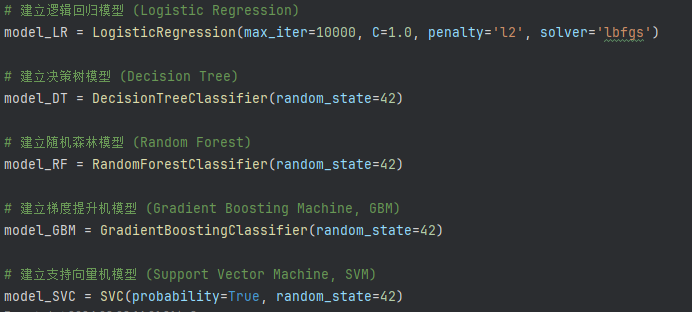
2.2 二次指标变量筛选

在初步分箱的基础上，我进一步对变量进行了二次筛选。这一步的核心是计算每个分箱内好样本与坏样本的比例，并由此计算出每个分箱的WOE（Weight of Evidence）值。WOE值是一种将分类变量转化为连续变量的技术，通过衡量分箱内好坏样本的相对比例，提供了变量对目标变量区分能力的量化指标。

此外，我还基于WOE值计算了每个特征的IV（Information Value）值。IV值用于评估特征的预测能力，通常认为IV值越高，特征的预测能力越强。在实际操作中，我根据IV值对变量进行了排序，最终选择了预测能力最强的前11个变量作为模型构建的主要输入特征。这些变量包括：

*['Duration of Credit (month)', 'Credit Amount', 'Account Balance', 'Age (years)', 'Payment Status of Previous Credit', 'Value Savings/Stocks', 'Purpose', 'Most valuable available asset', 'Length of current employment', 'Concurrent Credits', 'Sex & Marital Status']*

3.1 模型构造



在本次分析中，我选择并应用了五种常见的数据分析模型，包括逻辑回归模型 (Logistic Regression)、决策树模型 (Decision Tree)、随机森林模型 (Random Forest)、梯度提升机模型 (Gradient Boosting Machine, GBM) 和支持向量机模型 (Support Vector Machine, SVM)。这些模型覆盖了广泛的算法类型，能够充分探索数据的特征与模式。

为确保模型的性能最优，我进一步运用了堆叠分类器（Stacking Classifier）和交叉验证技术（Cross-Validation）来训练和评估每一个模型。堆叠分类器通过将多个模型的预测结果进行组合，通常可以提高整体的预测性能。交叉验证则通过多次分割数据集并重复训练与测试，有效减少了模型评估中的偏差和方差。

3.2 模型验证

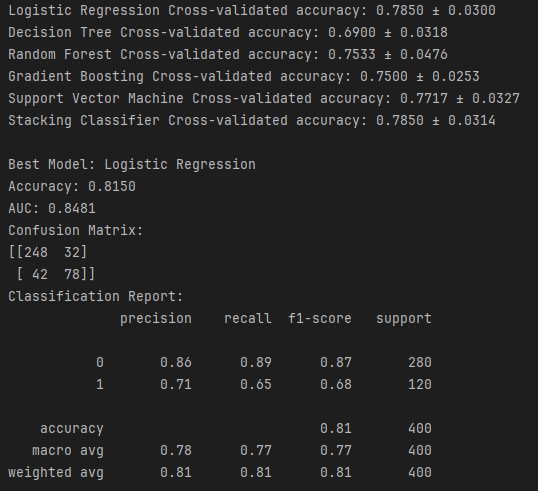
通过这一系列的自动化流程，在尽量减少人工干预的情况下，我最终找到了准确率最高的模型。此过程确保了模型选择的科学性和客观性，使得最终选定的模型在预测能力和泛化性能方面达到了最优。

最后的验证结果：

* 逻辑回归模型 (Logistic Regression)：交叉验证的平均准确率为 78.50%，标准误差为 ± 0.03。这表明模型在不同数据划分上的表现相对稳定。
* 决策树模型 (Decision Tree)：交叉验证的平均准确率为 69.00%，标准误差为 ± 0.0318。该模型的准确率较低，表明其对数据的过拟合问题较为严重。
* 随机森林模型 (Random Forest)：交叉验证的平均准确率为 75.33%，标准误差为 ± 0.0476，显示出良好的预测性能和稳定性。
* 梯度提升机模型 (Gradient Boosting Machine, GBM)：交叉验证的平均准确率为 75.00%，标准误差为 ± 0.0253，性能接近于随机森林模型。
* 支持向量机模型 (Support Vector Machine, SVM)：交叉验证的平均准确率为 77.17%，标准误差为 ± 0.0327，表现出色，特别是在数据边界清晰的情况下。
* 堆叠分类器 (Stacking Classifier)：交叉验证的平均准确率为 78.50%，标准误差为 ± 0.0314，与逻辑回归模型的表现相似。

最终选择的最佳模型是逻辑回归模型，其在测试集上的准确率为 81.50%，AUC 值为 0.8481，表明模型具有较强的区分能力。

总体上，模型的宏平均 (Macro avg) 精确率为 0.78，召回率为 0.77，F1-score 为 0.77，表明模型在各类别之间的表现较为均衡。加权平均 (Weighted avg) 的各项指标均为 0.81，进一步确认了模型的稳健性和良好的预测能力。



4. 评分卡建立

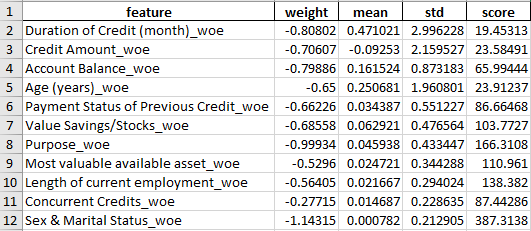
首先，数据通过 StandardScaler 进行标准化处理，这一步骤的目的是确保每个特征的数据在相同的尺度上，以避免特征值的大小对模型的影响。标准化后的数据被用于训练模型，并且在计算评分卡时，也会基于这些标准化的特征。

在这个函数中，我首先提取了模型的系数 coef 和截距 intercept。这些系数代表了每个特征对模型预测的影响大小。

* PDO (Points to Double the Odds): PDO 是一个衡量标准，用于确定在分数增加一定值后，正负样本的比率如何翻倍。在这一步中，B 值通过公式 B = pdo / np.log(2) 计算出来，它会影响分数的变化。
* Base Score (A): 这是模型的基础分数。通过公式 A = base\_score - B \* intercept 计算，它代表了当所有特征值等于均值时模型的起始得分。

对于每个特征，评分卡会计算如下内容：

* Weight: 该特征在模型中的系数，表示其对目标变量的影响方向和大小。
* Mean: 该特征的均值，基于标准化前的原始数据计算。
* Std (标准差): 该特征的标准差，用于表示数据的离散程度。
* Score: 评分卡中的每个特征得分通过公式 score = -weight \* B / feature\_std 计算得到。此分数反映了该特征对最终得分的具体影响。

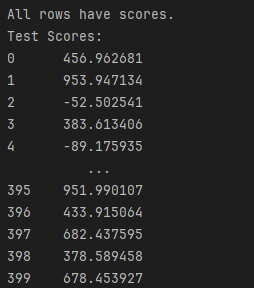


整个评分卡构建过程的目的是将逻辑回归模型的输出转化为一个可解释的评分系统。通过这个系统，可以将每个特征的贡献以分数的形式展示出来，使得最终的模型更易于理解和应用。同时，这个评分系统也能够更清晰地解释各个变量对目标预测的贡献程度。

5. 信用评分数计算

在最终的评分计算中，主要任务是使用之前构建的评分卡，将测试数据中的每个样本计算出相应的信用评分，并将这些分数与原始测试数据结合在一起，最终保存为一个 Excel 文件。

我在代码里先打印了最后的分数结果以确保每一行数据都有对应的分数：



最后通过评分卡函数对每个样本进行打分，并将这些分数与原始数据结合，生成最终的评分表。这一过程确保了评分系统的准确性和实用性。

6. 总结

在本项目中，我成功地构建了一个全面且可解释的信用评分系统。通过多种数据分析技术，包括描述性分析、相关性分析、分箱、以及二次变量筛选，我精心选择了具有高预测能力的特征，为模型构建打下了坚实基础。随后，我采用了多种机器学习模型，并结合堆叠分类器和交叉验证技术，确保了模型的稳健性和最佳性能。最终，我选择了逻辑回归模型作为核心模型，并基于该模型构建了评分卡。

评分卡系统将模型的复杂输出转化为易于理解的信用评分，清晰地展示了每个特征对最终评分的贡献。这一评分系统不仅在预测准确性上表现出色，还具备高度的实用性和解释性，为信用风险管理和决策制定提供了强有力的支持。

通过将最终的信用评分与原始数据结合，我生成了完整的评分表，并将其保存为 Excel 文件，确保结果的可视化和应用便利性。整体而言，本项目的成果为信用评估领域提供了一个高效、可靠的解决方案，具有广泛的应用前景和商业价值。