**C题 信用风险识别问题**

信用风险识别在金融行业和个体借贷过程中扮演着至关重要的角色。其核心任务是通过分析借款方的个人和财务信息等信息，评估其偿还贷款的能力和意愿，从而降低贷款机构的风险暴露。信用风险评价的准确性直接关系到贷款机构的资产质量和经济健康。因此，建立准确可靠的信用风险评价模型对于金融机构和借款方都具有重要意义。

在大数据背景下，信用风险评价研究中“信用风险评价指标筛选→信用风险得分测算→信用风险等级划分”各步骤均有诸多难题亟待解决。首先，在指标筛选过程中，如何建立恰当的指标筛选模型在大量可能的指标中筛选与信用风险相关性最高的指标，在确保数据全面性和准确性的同时，克服多重共线性、过度拟合等问题是难点之一；其次，实践中个体信用评价往往存在违约样本少、非违约样本多等现实情况，因而在信用得分测算过程中，如何选择适合的信用评分模型，解决模型对违约样本识别不足，并进一步平衡模型预测准确性与可解释性是又一难点；最后，在信用等级划分中，如何在确保等级划分的鲁棒性、普适性的条件下，选择恰当的阈值、聚类模型、非线性规划模型等将信用得分映射到信用等级，求解满足“信用等级越高、信用风险越低”的信用等级划分结果是又一难点。

附件1给出了UCI公开的德国信用数据集，其中编号X1-X24表示个体的个人及财务等信息指标。附件2给出了UCI公开的澳大利亚信用数据集，其中编号X1-X14表示个体的个人及财务等信息指标。请根据附件1、附件2和实际情况建立数学模型解决以下问题：

**问题1** 高维数据往往会为信用风险评价带来评价指标反应信息冗余等问题，请选择合适的模型对德国信用数据集进行指标筛选，以达到提升信用风险评价准确性及可解释性的目标，并给出理由。

**问题2** 传统线性加权方法无法准确刻画评价指标与违约风险之间的非线性关系，加之个体信用评价数据存在非违约样本多、违约样本少的不均衡分布特征，容易造成评价模型对非违约样本识别过度、违约样本识别不足的问题。请选择合适的信用评分模型揭示评价指标与个体违约风险之间的联系，求解德国信用数据集中个体信用得分，并给出理由。

**问题3** 随着Chat GPT等智能AI算法的发展，为信用风险准确识别带来了新的机遇。为对信用评分模型的合理性、准确性进行判别，请分别利用附件1中的德国信用数据集与附件2中的澳大利亚信用数据集，自建信用评分模型。可将所建模型与决策树(Decision Tree, DT)、K最近邻(K-Nearest Neighbor, KNN)、随机森林(Random Forest, RF)、支持向量机(Support Vector Machine, SVM)等多种现有分类模型进行对比，分析模型的判别性能。对比上述多种分类方法对不同数据集的分类效果，并将结果填入表1、表2（至少选择3类模型，3个评价准则进行对比分析）。

**表1 德国信用数据集分类方法对比结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 评价准则 | | | | |
| Accuracy | AUC | Type1-error | Type2-error | … |
| 你们的模型 |  |  |  |  |  |
| DT |  |  |  |  |  |
| KNN |  |  |  |  |  |
| RF |  |  |  |  |  |
| SVM |  |  |  |  |  |
| … |  |  |  |  |  |

**表2 澳大利亚信用数据集分类方法对比结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 评价准则 | | | | |
| Accuracy | AUC | Type1-error | Type2-error | … |
| 你们的模型 |  |  |  |  |  |
| DT |  |  |  |  |  |
| KNN |  |  |  |  |  |
| RF |  |  |  |  |  |
| SVM |  |  |  |  |  |
| … |  |  |  |  |  |

**问题4** 请以“信用等级越高、信用风险越低”为信用等级划分标准，构建非线性规划模型，在德国信用数据集上划分个体信用等级，并说明模型选择理由。

附件1 德国信用数据集

附件2 澳大利亚信用数据集