

互联网金融风控预测

研究背景与概括

大数据风控的探索

- 互联网金融风控模型成为行业焦点，大数据风控与传统风控相比，通过数据变现商业模式，如精准营销和大数据风控，用于评估用户信用等级和识别欺诈用户
- 传统金融风控依赖信用属性强的金融数据，通过评分系统识别客户还款能力与意愿，而互联网金融风控在传统基础上丰富数据维度，利用行为数据进行风险评估

贷款业务风险控制

- 贷款业务带来营收同时伴随违约风险，借贷方需对贷款进行风险控制，建立有效审核系统代替人工审核，以平衡营收与风险
- 建立机器学习模型自动审核贷款人信息，通过特征筛选和描述性统计分析了解特征与违约关系，对特征变量进行改变以提高模型有效性

数据分析与清洗

数据采集与预处理

- 研究基于Python语言，使用Jupyter Notebook和相关库进行数据分析，包括数据采集、清洗预处理、特征工程、建立机器学习模型和模型融合
- 数据来自信贷平台，包含120万条记录，47列变量信息，抽取80万条作为训练集，20万条作为测试集A和B，进行脱敏处理

数据分析与清洗步骤

- 查看训练集数据量、特征数量、缺失情况和数据类型，对缺失值进行分析和填充，处理文本型数据，删除无用列和重复值，分析特征间相关性

特征工程与模型建立

- 特征工程是机器学习中重要部分，包括特征选择、新特征制造和无量纲化处理，对模型建立起到决定性作用
- 采用随机森林、逻辑回归和梯度提升树分类器作为初步模型，通过交叉验证进行模型调参，发现异常值处理和特征工程有利于建立更精确模型

模型预测与评估

基于模型的预测

- 使用训练集对新测试集进行预测，得到预测数据，其中predict_proba返回所有结果概率，predict返回预测值
- 模型评估采用AUC评估，定义为ROC曲线下与坐标轴围成的面积，AUC越接近1.0，模型真实性越高