



计算金融与仿真

课程论文

论文题目：计算金融与仿真

学生姓名：李晶晶 张璐 朱冯婧

指导老师：邓志斌

1. 投资组合选择介绍

2. 模型建立

决策变量

$$x_i \in \{0, 1\}, \quad i = 1, 2, \dots, N$$

其中,

- $x_i = 1$ 表示选择资助第 i 个贷款对象;
- $x_i = 0$ 表示不选择。

参数说明

- A_i : 第 i 个贷款的金额;
- r_i : 第 i 个贷款的利率;
- P_i : 第 i 个贷款的违约概率;
- B : 总投资预算;
- R_{\max} : 允许的最大违约风险;
- G_k : 第 k 个信用等级的贷款集合;
- α_k : 第 k 个信用等级的最大投资比例;
- m : 最多选择的贷款数量 (Top- m)。

模型形式

$$\begin{aligned}
 & \max_{x_i \in \{0, 1\}} \quad \sum_{i=1}^N x_i A_i [r_i (1 - P_i) - P_i] \\
 & \text{s.t.} \quad \sum_{i=1}^N x_i A_i \leq B \quad (\text{预算限制}) \\
 & \quad \quad \sum_{i \in G_k} x_i A_i \leq \alpha_k B, \quad \forall k \quad (\text{信用等级比例约束}) \\
 & \quad \quad \sum_{i=1}^N x_i A_i P_i \leq R_{\max} \quad (\text{风险控制}) \\
 & \quad \quad \sum_{i=1}^N x_i \leq m \quad (\text{Top-}m \text{ 选择})
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 & \max_{x_i \in \{0,1\}} \sum_{i=1}^N x_i A_i [r_i(1 - P_i) - P_i] \\
 & \text{s.t.} \quad \sum_{i=1}^N x_i A_i \leq B \quad (\text{预算限制}) \\
 & \quad \sum_{i \in G_k} x_i A_i \leq \alpha_k B, \quad \forall k \quad (\text{信用等级比例约束}) \\
 & \quad \sum_{i=1}^N x_i A_i P_i \leq R_{\max} \quad (\text{期望风险控制}) \\
 & \quad \sum_{i=1}^N x_i \leq m \quad (\text{Top-}m \text{ 选择}) \\
 & \quad \sum_{i=1}^N x_i A_i \cdot \tilde{P}_i^{(s)} - \eta \leq \mathcal{M} z_s, \quad \forall s = 1, \dots, S \quad (\text{VaR 限制}) \\
 & \quad \sum_{s=1}^S z_s \leq (1 - \beta) S \\
 & \quad z_s \in \{0, 1\}, \quad \eta \in \mathbb{R}
 \end{aligned} \tag{2}$$

$$\begin{aligned}
 & \max_{x_i \in \{0,1\}} \sum_{i=1}^N x_i A_i [r_i(1 - P_i) - P_i] \\
 & \text{s.t.} \quad \sum_{i=1}^N x_i A_i \leq B \quad (\text{预算限制}) \\
 & \quad \sum_{i \in G_k} x_i A_i \leq \alpha_k B, \quad \forall k \quad (\text{信用等级比例约束}) \\
 & \quad \sum_{i=1}^N x_i A_i P_i \leq R_{\max} \quad (\text{期望风险控制}) \\
 & \quad \sum_{i=1}^N x_i \leq m \quad (\text{Top-}m \text{ 选择}) \\
 & \quad \xi_s \geq \sum_{i=1}^N x_i A_i \cdot \tilde{P}_i^{(s)} - \eta, \quad \forall s = 1, \dots, S \quad (\text{场景损失}) \\
 & \quad \eta + \frac{1}{S(1 - \beta)} \sum_{s=1}^S \xi_s \leq \text{CVaR}_{\max} \quad (\text{CVaR 限制}) \\
 & \quad \xi_s \geq 0, \quad \eta \in \mathbb{R}
 \end{aligned} \tag{3}$$

3. 算法设计

Algorithm 1: 启发式算法：贷款组合优化（含 CVaR 控制）

Input: 贷款数据 $\{A_i, r_i, P_i\}_{i=1}^N$; 预算 B ; Top- m 限制; CVaR 上限 CVaR_{\max} ; 置信水平 β ; 模拟场景矩阵 $\tilde{P}_i^{(s)} \in \{0, 1\}^{S \times N}$

Output: 最优选择向量 $x^* \in \{0, 1\}^N$

```

1  $x_{\text{best}} \leftarrow 0$ ; // 初始为空解
2 根据评分  $\text{score}_i = A_i[r_i(1 - P_i) - P_i]$  降序排列贷款
3  $\text{total\_budget} \leftarrow 0, \text{total\_selected} \leftarrow 0, x_{\text{current}} \leftarrow 0$ 
4 foreach  $i$  in 排序后的贷款列表 do
5     if  $\text{total\_budget} + A_i > B$  或  $\text{total\_selected} + 1 > m$  then
6         continue;
7      $x_{\text{current}}[i] \leftarrow 1$ ;
8      $\text{total\_budget} \leftarrow \text{total\_budget} + A_i$ ;
9      $\text{total\_selected} \leftarrow \text{total\_selected} + 1$ ;
10 if  $\text{Feasible}(x_{\text{current}})$  then
11      $x_{\text{best}} \leftarrow x_{\text{current}}$ ;
12 Function  $\text{Feasible}(x)$ :
13     for  $s \leftarrow 1$  to  $S$  do
14          $L_s \leftarrow \sum_{i=1}^N x_i A_i \cdot \tilde{P}_i^{(s)}$ ;
15      $\eta \leftarrow \beta$  分位点的  $\{L_s\}$ ;
16     for  $s \leftarrow 1$  to  $S$  do
17          $\xi_s \leftarrow \max(L_s - \eta, 0)$ ;
18      $\text{CVaR}_\beta(x) \leftarrow \eta + \frac{1}{S(1-\beta)} \sum_{s=1}^S \xi_s$ ;
19     return 是否满足  $\text{CVaR}_\beta(x) \leq \text{CVaR}_{\max}$  且满足其他约束;
20 return  $x_{\text{best}}$ ;
    
```

4. 案例研究

4.1. 数据集描述

本研究使用的数据集来自 Lending Club 平台，原始数据由 Kaggle 网站<https://www.kaggle.com/datasets/wordsforthewise/lending-club>公开提供。Lending Club 是美国最大的网络借贷平台之一，提供了详尽的个人借款申请及其还款情况的数据，广泛应用于学术界和工业界进行信贷风险评估、违约预测及投资组合优化等研究。

该数据集包含了从 2007 年至 2018 年的借款记录，共计数百万条样本。每条记录对应一笔贷款申请，涵盖了包括贷款金额、利率、借款人信用等级、债务收入比、贷款期限、还款状态、就业年限、收入、地址状态、房屋所有权、FICO 评分区间等在内的多维度信息。

在本研究中，我们主要筛选并保留以下变量用于建模分析：

- **loan_amnt**: 借款人申请的贷款金额, 作为 A_i ;
- **int_rate**: 借款合同中约定的年利率, 用于计算收益率 r_i ;
- **grade**: 借款人的信用等级 (A 至 G), 用于分组限制;
- **loan_status**: 实际贷款的还款状态 (如 Fully Paid、Charged Off), 用于推断违约情况;
- **annual_inc**: 借款人年收入;
- **dti**: 债务收入比, 用于辅助风险刻画;
- **term**: 贷款期限 (如 36 months 或 60 months);
- **emp_length**: 借款人工作年限;
- **addr_state**: 借款人所在州;
- **fico_range_high, fico_range_low**: 借款人 FICO 信用评分区间;

为了满足模型中对违约概率 P_i 的需求, 我们将 “loan_status” 字段中状态为 “Charged Off” 的贷款视为违约样本, 其余如 “Fully Paid”、“Current” 等状态作为非违约样本, 并基于历史频率法估算每一类贷款的违约概率。

此外, 为模拟贷款违约的风险场景, 我们以借款人的信用等级、FICO 评分和历史违约频率为依据, 构建了 S 个 Monte Carlo 风险场景, 用于后续 CVaR 优化模型的风险评估。

通过上述数据处理步骤, 最终形成了一个结构规范、信息完备、适用于组合优化问题的数据集, 为后续实证分析与建模提供了坚实的数据基础。

4.2. 使用机器学习预测违约概率 P_i

尽管本研究使用的数据集中所有贷款均已获得实际资助, 但在资金有限、需进行优选配置的情境下, 我们仍需要对这些已发放贷款的还款风险进行再评估。为此, 我们引入机器学习方法, 对每笔贷款的未违约概率 P_i 进行预测建模, 以作为后续优化模型中的风险输入参数。

建模目标 预测函数的目标是: 对于每笔已发放贷款 i , 根据其已知特征向量 \mathbf{x}_i , 估计其在未来发生违约的概率 $P_i = \mathbb{P}(y_i = 1 | \mathbf{x}_i)$ 。其中, $y_i = 1$ 表示贷款最终发生违约 (如状态为 Charged Off), $y_i = 0$ 表示贷款最终还清 (如状态为 Fully Paid)。

特征构造 我们基于借款人基本属性、贷款合同信息以及信用评级等信息构建预测特征集, 涵盖但不限于以下变量:

- **loan_amnt**: 贷款金额;
- **term**: 贷款期限;
- **int_rate**: 贷款利率;
- **grade** 和 **sub_grade**: Lending Club 信用评级;
- **emp_length**: 工作年限;
- **home_ownership**: 住房类型;
- **annual_inc**: 年收入;
- **dti**: 债务收入比;

- **purpose:** 贷款用途;
- **fico_range_high / low:** 信用评分;

建模方法 考虑到目标变量仍是二元状态（违约 / 未违约），我们采用监督学习的二分类方法进行建模。尝试的模型包括逻辑回归（Logistic Regression）、随机森林（Random Forest）、梯度提升树（GBDT）、极端梯度提升（XGBoost）等。

由于样本中违约样本占比相对较小，我们在训练过程中采用类别加权、欠采样等方式处理类别不平衡问题。

训练与评估 我们将全部已发放贷款随机划分为训练集（70%）与测试集（30%），使用交叉验证调优参数，并基于测试集评估模型表现。评价指标包括准确率（Accuracy）、AUC 值（Area Under the ROC Curve）、F1 分数等。

最终，我们选择 AUC 表现最优的模型用于对所有贷款样本生成违约概率预测值 \hat{P}_i ，作为后续优化模型中的输入。

说明 虽然原始数据中每笔贷款都已实际放款，但我们的建模任务是为现实中的“再选择”提供依据。即在预算受限、资源不足时，如何在这些真实已放款的贷款中优先挑选违约概率低、预期收益高的子集，构建一个更稳健的投资组合。因此， \hat{P}_i 的预测并非用于决定放款与否，而是作为组合优化的“风险估计量”，用于构建期望收益与 CVaR 等风险指标。

4.3. 使用算法求解原始模型、VaR 模型与 CVaR 模型

在本节中，我们分别构建并求解三类贷款筛选优化模型：原始模型、VaR（Value at Risk）模型和 CVaR（Conditional Value at Risk）模型。三者均以最大化投资收益为目标，并在此基础上逐步引入风险控制手段，以模拟实际投资场景中对风险的不同管控需求。

原始模型（期望损失约束） 原始模型以最大化期望净收益为目标，同时设置多个线性约束，确保预算控制、信用等级平衡与风险控制等现实要求。该模型为典型的 0-1 整数规划问题，可使用 Gurobi 等商业求解器在中小规模下获得最优解。

VaR 模型（引入分位损失约束） VaR 模型进一步考虑极端情境下的最大可能损失，将分位数损失 VaR_β 作为新的风险控制手段。通过构造 Monte Carlo 风险场景，我们对贷款可能发生的违约路径进行模拟，并引入辅助变量 $z_s \in \{0, 1\}$ 表示某个场景是否超过 VaR 阈值 η 。

由于 VaR 对应的非凸约束结构和 0-1 变量使得模型求解复杂度较高，通常只适用于中小规模问题，且求解结果可能存在不连续性。

CVaR 模型（引入条件期望损失约束） 为克服 VaR 模型不可微、不可凸等缺点，CVaR 模型通过引入辅助变量 $\xi_s \geq 0$ 和 VaR 估计值 η ，以线性结构实现对尾部风险的精确控制。其约束具有良好的可解性与可扩展性，广泛应用于实际金融优化问题中。

该模型可转化为混合整数线性规划（MILP）形式，使用 Gurobi 等求解器可以在合理时间内获得精确解，特别适合场景规模较大的优化问题。

求解策略 在三类模型中，原始模型为最简形式，适用于风险可控或追求高收益的情境；VaR 模型在表达风险容忍边界方面具有直观优势；而 CVaR 模型则提供了更稳健的风险控制能力和更优的优化特性。

考虑到实际求解效率与解的稳定性，本研究在实现中主要采用 CVaR 模型进行主模型求解，并通过 Gurobi 求解器实现精确建模与最优解求取。对于大规模场景（如 $S > 1000$ ），可结合启发式算法进行初解生成与变量预选，从而提升整体求解效率。

4.4. 结果分析

5. 结论

参考文献

- [1] 彭桥, 肖尧, 杨宇茜, and 杨沛瑾. 中国新质生产力发展水平测度、动态演化与驱动因素研究. 软科学, 39(04):25–34, 2025. ISSN 1001-8409. doi: 10.13956/j.ss.1001-8409.2025.04.05. URL <https://link.cnki.net/urlid/51.1268.g3.20241125.1056.010>.
- [2] 王珏 and 王荣基. 新质生产力: 指标构建与时空演进. 西安财经大学学报, 37(1):31–47, 2024. doi: 10.19331/j.cnki.jxufe.20231124.001.
- [3] 李光勤 and 李梦娇. 中国省域新质生产力水平评价、空间格局及其演化特征. 经济地理, 44(8): 116–125, 2024. doi: 10.15957/j.cnki.jjdl.2024.08.014.
- [4] 曹东勃 and 蔡煜. 新质生产力指标体系构建研究. 上海财经大学学报, (4):50–62, 2024.
- [5] 张海, 王震, and 李秉远. 新质生产力发展水平、空间差异及动态演进. 统计与决策, (24):11–26, 2024. doi: 10.13546/j.cnki.tjyj.2024.24.002.
- [6] 胡佳霖 and 徐俊. 中国新质生产力: 区域差距、动态演进与跃迁趋势. 统计与决策, (21):5–16, 2024. doi: 10.13546/j.cnki.tjyj.2024.21.001.
- [7] 简新华 and 聂长飞. 中国新质生产力水平测度及省际现状的比较分析. 经济问题探索, (10): 3–20, 2024.
- [8] 冉戎, 花磊, 陈烨靖, and 夏艺嘉. 新质生产力发展潜力测度、时空差异及战略着力点研究. 重庆大学学报(社会科学版), 2024. doi: 10.11835/j.issn.1008-5831.jg.2024.10.002. URL <https://link.cnki.net/urlid/50.1023.C.20241030.1300.006>. 网络首发.
- [9] 颜克高, 王馨悦, and 吴心怡. 中国新质生产力发展的水平测度与区域差异研究. 湖南大学学报(社会科学版), 39(1):10–21, 2025. doi: 10.16339/j.cnki.hdxbskb.2025.01.002.
- [10] 徐波, 王兆萍, 余乐山, and 刘柯. 新质生产力对资源配置效率的影响效应研究. 产业经济评论, (4):35–49, 2024. doi: 10.19313/j.cnki.cn10-1223/f.20240417.001.
- [11] 杨智晨, 涂先青, and 王方方. 我国新质生产力发展的理论基础、时空特征及分异机理. 经济问题探索, (1):50–66, 2025.
- [12] 马大晋, 吴旭辉, and 张博文. 中国新质生产力发展水平区域评价与空间关联网络特征. 财经论丛, pages 1–14, 2025. doi: 10.13762/j.cnki.cjlc.20250103.001. URL <https://doi.org/10.13762/j.cnki.cjlc.20250103.001>.
- [13] 程赛楠 and 冯珍. 数实融合对新质生产力的影响研究. 北京理工大学学报(社会科学版), 26(6): 15–27, 2024. doi: 10.15918/j.jbitss1009-3370.2024.1491.
- [14] 张龙, 申瑛琦, and 张伟琦. 新质生产力的原创价值、统计测度与培育方向. 暨南学报(哲学社会科学版), (11):126–144, 2024. doi: 10.11778/j.jnxb.20240975.
- [15] Younes Ataei, Amin Mahmoudi, Mohammad Reza Feylizadeh, and Deng-Feng Li. Ordinal Priority Approach (OPA) in Multiple Attribute Decision-Making. *APPLIED SOFT COMPUTING*, 86, January 2020. ISSN 1568-4946. doi: 10.1016/j.asoc.2019.105893.