TONGJI UNIVERSITY

中小微企业信贷决策分析与建模

课题名称	统计分析与建模
学　　院	计算机科学与技术学院
专　　业	软件工程
指导教师	高珍
成　　员	2354100 郝哲逸 2351454 黄文 2451736 李明泽
日　　期	12 月 6 日

基于 R 的中小微企业信贷决策分析与建模

（统计分析与建模 · 期末大作业）

---

1. 项目背景与数据来源

1.1 业务背景

中小微企业是国民经济的重要组成部分，但由于规模小、缺乏抵押资产，长期面临“融资难、融资贵”的问题。商业银行在实际业务中，通常依据企业的**票据信息（发票流水）**、**信誉评级**和**上下游交易能力**来评估信贷风险。

本项目旨在模拟银行的决策过程：首先利用企业历史数据量化信贷风险（预测违约概率），然后结合利率与客户流失率的关系，制定能使银行收益最大化的信贷策略（包括是否放贷、贷款额度及利率）。

1.2 数据来源

本项目数据来源于**2020 年高教社杯全国大学生数学建模竞赛（C 题）**。数据集包含三个部分：

1. **附件 1**：123 家有信贷记录企业的历史数据（包含进出项发票、信誉评级及违约记录）。
   Sheet1（企业信息）：

企业代号	企业名称	信誉评级	是否违约
E1	***电器销售有限公司	A	否
E2	***技术有限责任公司	A	否
E3	***电子(中国)有限公司***分公司	C	否
E4	***发展有限责任公司	C	否
E5	***供应链管理有限公司	B	否
E6	***装饰设计工程有限公司	A	否
E7	***家电有限公司***分公司	A	否
E8	***科学研究院有限公司	A	否
E9	***生活用品服务有限公司***分公司	A	否
E10	***建筑劳务有限公司	B	否
E11	***建设工程有限公司	C	否
E12	***建筑劳务有限公司	B	否
E13	***汽车贸易有限公司	A	否
E14	个体经营E14	C	否
E15	***劳务有限公司	A	否
E16	***建筑劳务有限公司	A	否
E17	***消防工程有限公司	A	否
E18	***消防工程有限责任公司	A	否
E19	***科技有限公司	A	否
E20	***贸易有限公司	B	否

   Sheet2（进项发票信息）：

企业代号	发票号码	开票日期	销方单位代号	金额	税额	价税合计	发票状态
E1	3390939	2017/7/18	A00297	-943.4	-56.6	-1000	有效发票
E1	3390940	2017/7/18	A00297	-4780.24	-286.81	-5067.05	有效发票
E1	3390941	2017/7/18	A00297	943.4	56.6	1000	有效发票
E1	3390942	2017/7/18	A00297	4780.24	286.81	5067.05	有效发票
E1	9902669	2017/8/7	A05061	326.21	9.79	336	有效发票
E1	40826107	2017/8/8	A05991	170.94	29.06	200	有效发票
E1	4420531	2017/8/9	A03142	37735.85	2264.15	40000	有效发票
E1	4420532	2017/8/9	A03142	4716.98	283.02	5000	有效发票
E1	15040454	2017/8/11	A02994	46153.85	7846.15	54000	作废发票
E1	40829320	2017/8/14	A05991	162.39	27.61	190	有效发票
E1	2032326	2017/8/16	A00314	4614.12	276.84	4890.96	有效发票
E1	14678366	2017/8/21	A03346	13846.15	2353.85	16200	有效发票
E1	167875	2017/8/23	A01714	4854.37	145.63	5000	有效发票
E1	167876	2017/8/23	A01714	4854.37	145.63	5000	有效发票
E1	167877	2017/8/23	A01714	4854.37	145.63	5000	有效发票
E1	167878	2017/8/23	A01714	4854.37	145.63	5000	有效发票
E1	13428924	2017/8/23	A13557	970.87	29.13	1000	有效发票
E1	167879	2017/8/24	A01714	485.44	14.56	500	有效发票
E1	13690002	2017/8/24	A07155	601.89	36.11	638	作废发票
E1	13690003	2017/8/24	A07155	601.89	36.11	638	有效发票
E1	10769077	2017/8/25	A01798	969.9	29.1	999	有效发票
E1	15462965	2017/8/27	A03775	466.02	13.98	480	有效发票
E1	18308605	2017/8/27	A08485	619.81	37.19	657	有效发票
E1	9279217	2017/8/28	A01709	283.02	16.98	300	有效发票

Sheet3（销项发票信息）：

企业代号	发票号码	开票日期	购方单位代号	金额	税额	价税合计	发票状态
E1	11459356	2017/8/4	B03711	9401.71	1598.29	11000	有效发票
E1	5076239	2017/8/9	B00844	8170.94	1389.06	9560	有效发票
E1	5076240	2017/8/9	B00844	8170.94	1389.06	9560	有效发票
E1	5076241	2017/8/9	B00844	4085.47	694.53	4780	有效发票
E1	5076242	2017/8/9	B00844	4085.47	694.53	4780	有效发票
E1	5076243	2017/8/9	B00844	15042.73	2557.27	17600	有效发票
E1	11459357	2017/8/9	B03700	-2290.6	-389.4	-2680	有效发票
E1	11459358	2017/8/9	B10763	-12307.69	-2092.31	-14400	有效发票
E1	5076245	2017/8/10	B00713	12307.69	2092.31	14400	有效发票
E1	5076244	2017/8/10	B03518	2290.6	389.4	2680	有效发票
E1	11459359	2017/8/10	B03700	4000	680	4680	有效发票
E1	5076247	2017/8/16	B03199	889884.62	151280.38	1041165	有效发票
E1	5076248	2017/8/16	B03199	986793.16	167754.84	1154548	有效发票
E1	5076249	2017/8/16	B03199	885641.02	150558.98	1036200	有效发票
E1	5076250	2017/8/16	B03199	882703.42	150059.58	1032763	有效发票
E1	5076251	2017/8/16	B03199	965032.49	164055.51	1129088	有效发票
E1	5076252	2017/8/16	B03199	706110.25	120038.75	826149	有效发票
E1	5076246	2017/8/16	B08483	494087.2	83994.8	578082	有效发票
E1	11459360	2017/8/16	B10401	-38958.97	-6623.03	-45582	有效发票
E1	11459361	2017/8/16	B10401	-43213.68	-7346.32	-50560	有效发票
E1	11459362	2017/8/16	B10401	-90003.42	-15300.58	-105304	有效发票
E1	11459363	2017/8/16	B10401	41695.73	7088.27	48784	有效发票
E1	11459364	2017/8/16	B10401	12800	2176	14976	有效发票
E1	11459365	2017/8/16	B10401	34839.32	5922.68	40762	有效发票
E1	11459366	2017/8/17	B02814	-3059.83	-520.17	-3580	有效发票
E1	11459367	2017/8/17	B02814	-5153.85	-876.15	-6030	有效发票
E1	11459368	2017/8/17	B02814	-8598.29	-1461.71	-10060	有效发票
E1	11459369	2017/8/17	B02814	-6282.05	-1067.95	-7350	有效发票
E1	11459370	2017/8/17	B02814	-7230.77	-1229.23	-8460	有效发票

2. **附件 3**：银行贷款年利率与客户流失率关系的统计数据。

贷款年利率	客户流失率		
	信誉评级A	信誉评级B	信誉评级C
0.04	0	0	0
0.0425	0.094574126	0.066799583	0.068725306
0.0465	0.135727183	0.13505206	0.122099029
0.0505	0.224603354	0.20658008	0.181252146
0.0545	0.302038102	0.276812293	0.263302863
0.0585	0.347315668	0.302883401	0.290189098
0.0625	0.41347177	0.370215852	0.34971559
0.0665	0.447890973	0.406296668	0.390771683
0.0705	0.497634453	0.458295295	0.45723807
0.0745	0.511096612	0.508718692	0.492660433
0.0785	0.573393087	0.544408837	0.513660239
0.0825	0.609492115	0.548493958	0.530248706
0.0865	0.652944774	0.588765696	0.587762408
0.0905	0.667541843	0.625764576	0.590097045
0.0945	0.694779921	0.635605146	0.642993656
0.0985	0.708302023	0.673527424	0.658839416
0.1025	0.731275401	0.696925431	0.696870573
0.1065	0.775091405	0.705315993	0.719103552
0.1105	0.798227368	0.742936326	0.711101237
0.1145	0.790527266	0.776400729	0.750627656
0.1185	0.815196986	0.762022595	0.776816043
0.1225	0.814421029	0.791503697	0.784480512
0.1265	0.854811097	0.814998933	0.795566274
0.1305	0.870317343	0.822297861	0.820051434
0.1345	0.871428085	0.835301602	0.832288422

1.3 项目目标

1. 构建一个**高可解释性的违约概率预测模型**。
2. 拟合利率与客户流失率的函数关系。
3. 在信贷总额（预算）固定的约束下，优化**信贷额度与利率分配策略**，实现期望收益最大化。

项目代码仓库： https://github.com/SOLDIER-627/Credit*Risk*Analyzer

2. 数据预处理与特征构造

项目最大的难点在于：发票数据是半结构化的流水，而非直接的财务指标。我们在代码中主要进行了以下清洗逻辑：

2.1 发票清洗：剔除噪点

原始数据中存在大量干扰项，处理逻辑如下：

- **作废发票**：直接在计算财务指标前剔除，但保留“作废比例”作为风险特征。
- **负数发票**：这是企业的“红冲”（退货）操作。我们在计算总营收时，将负数金额的绝对值从总额中扣除，而不是简单相加，以还原真实营收。
- **零值处理**：对发票数量为 0 或营收极低的企业进行标记。

2.2 构造企业财务特征

从原始发票流中提取了以下核心特征（Feature Engineering）：

- **规模指标**：总营收、总支出、运营规模（营收+支出）。
- **盈利能力**：毛利润（营收-支出）、利润率。
- **稳定性指标**：发票金额变异系数（CV），用于衡量经营波动性。
- **风险行为指标**：**作废发票比例**、**负数发票比例**。
- **信誉评级**：将 A/B/C/D 评级进行有序编码（Ordinal Encoding）。

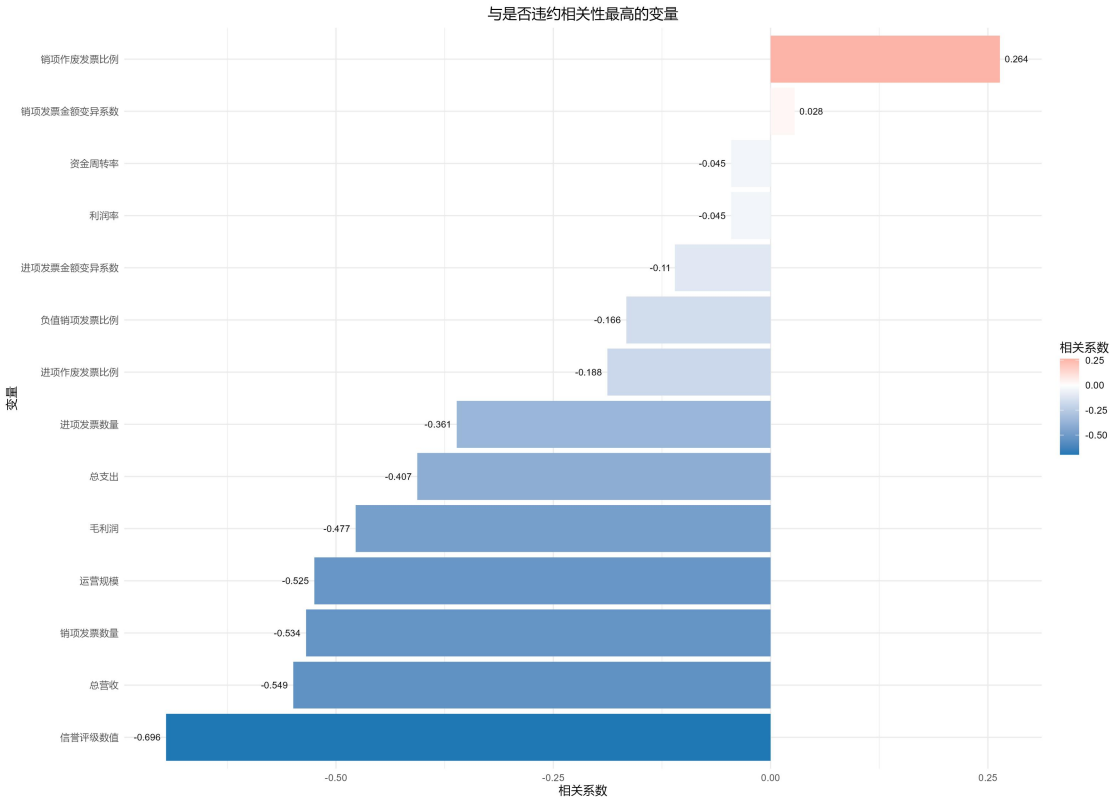
处理后的数据展示：

企业代号	企业名称	信誉评级	是否违约	总支出	进项发票数	进项发票金	总营收	销项发票数	负值销项数	销项发票金	进项作废数	销项作废金	毛利润	运营规模	利润率	资金周转率	是否违约.1	信誉评级.1
E1	***电器销	A	否	7.79E+09	3249	1.845638	5E+09	7886	0.026912	0.701706	0.055798	0.02762	-2.8E+09	1.28E+10	-0.55748	0.642061	0	3
E2	***技术有	A	否	1.65E+08	31435	4.794861	6.67E+08	11665	0.03069	0.737427	0.022422	0.082002	5.02E+08	8.31E+08	0.753151	4.051055	0	3
E3	***电子(牛	C	否	54859923	4367	4.557192	1.13E+09	23688	0.186001	2.382309	0.042535	0.015993	1.07E+09	1.18E+09	0.951342	20.55154	0	1
E4	***发展有	C	否	2.63E+08	521	3.865292	2.16E+09	2041	0.00441	0.227316	0.066308	0.085164	1.9E+09	2.43E+09	0.87824	8.212866	0	1
E5	***供应链	B	否	2.3E+08	2084	3.551957	2.36E+08	1005	0.00995	1.349968	0.039189	0.051887	5968151	4.65E+08	0.025339	1.025998	0	2
E6	***装饰设	A	否	3.27E+08	10814	3.031632	4E+08	913	0.025192	0.869623	0.046132	0.132129	73174165	7.27E+08	0.182884	1.223816	0	3
E7	***家电有	A	否	78273474	12643	3.793057	1.05E+09	8032	0.236429	1.623023	0.034664	0.014358	9.76E+08	1.13E+09	0.92576	13.46987	0	3
E8	***科学研	A	否	1.72E+08	21777	8.115413	4.11E+08	8359	0.02931	0.807848	0.033679	0.113103	2.4E+08	5.83E+08	0.582468	2.395028	0	3
E9	***生活用	A	否	26569889	4199	3.783026	3.91E+08	5760	0.021181	0.542132	0.021896	0.024721	3.64E+08	4.17E+08	0.932	14.70578	0	3
E10	***建筑劳	B	否	5771271	3860	2.663741	3.54E+08	516	0.003876	0.509616	0.057847	0.091549	3.48E+08	3.59E+08	0.983679	61.27009	0	2
E11	***建设工	C	否	1.45E+08	2715	1.752491	1.63E+08	1051	0.000951	1.388764	0.042666	0.059087	17557800	3.08E+08	0.107884	1.12093	0	1
E12	***建筑劳	B	否	95980113	1757	4.405994	2.45E+08	263	0.003802	0.229908	0.021715	0.077193	1.49E+08	3.41E+08	0.608249	2.552644	0	2
E13	***汽车贸	A	否	1.04E+08	13255	4.159253	2.53E+08	6808	0.010723	3.330203	0.015888	0.149532	1.48E+08	3.57E+08	0.586921	2.420847	0	3
E14	个体经营(	C	否	1.3E+08	6527	3.627771	2.47E+08	3097	0.013562	2.013187	0.077717	0.071643	1.17E+08	3.77E+08	0.473039	1.897673	0	1
E15	***劳务有	A	否	4803944	92	0.934292	2.2E+08	2260	0	0.143078	0	0.063018	2.15E+08	2.24E+08	0.978124	45.71261	0	3
E16	***建筑劳	A	否	333748.4	289	2.113973	2.14E+08	390	0	0.564599	0.003448	0.111617	2.13E+08	2.14E+08	0.998437	639.8732	0	3
E17	***消防工	A	否	1.5E+08	7143	2.307584	1.75E+08	561	0.023173	1.079783	0.056158	0.170118	25247222	3.25E+08	0.144349	1.1687	0	3
E18	***消防工	A	否	1.41E+08	5455	2.724685	2E+08	345	0.005797	0.565103	0.053773	0.094488	58748135	3.41E+08	0.29404	1.416511	0	3
E19	***科技有	A	否	2.12E+08	1484	1.679938	2.19E+08	2821	0.023041	0.440199	0.026247	0.089706	6282598	4.31E+08	0.02872	1.029569	0	3

3. 探索性分析（EDA）

在建模前，我们对特征与“是否违约”的关系进行了深入分析。

3.1违约相关性分析



Spearman 相关性分析显示：

1. 核心负相关特征（风险保护因素）：

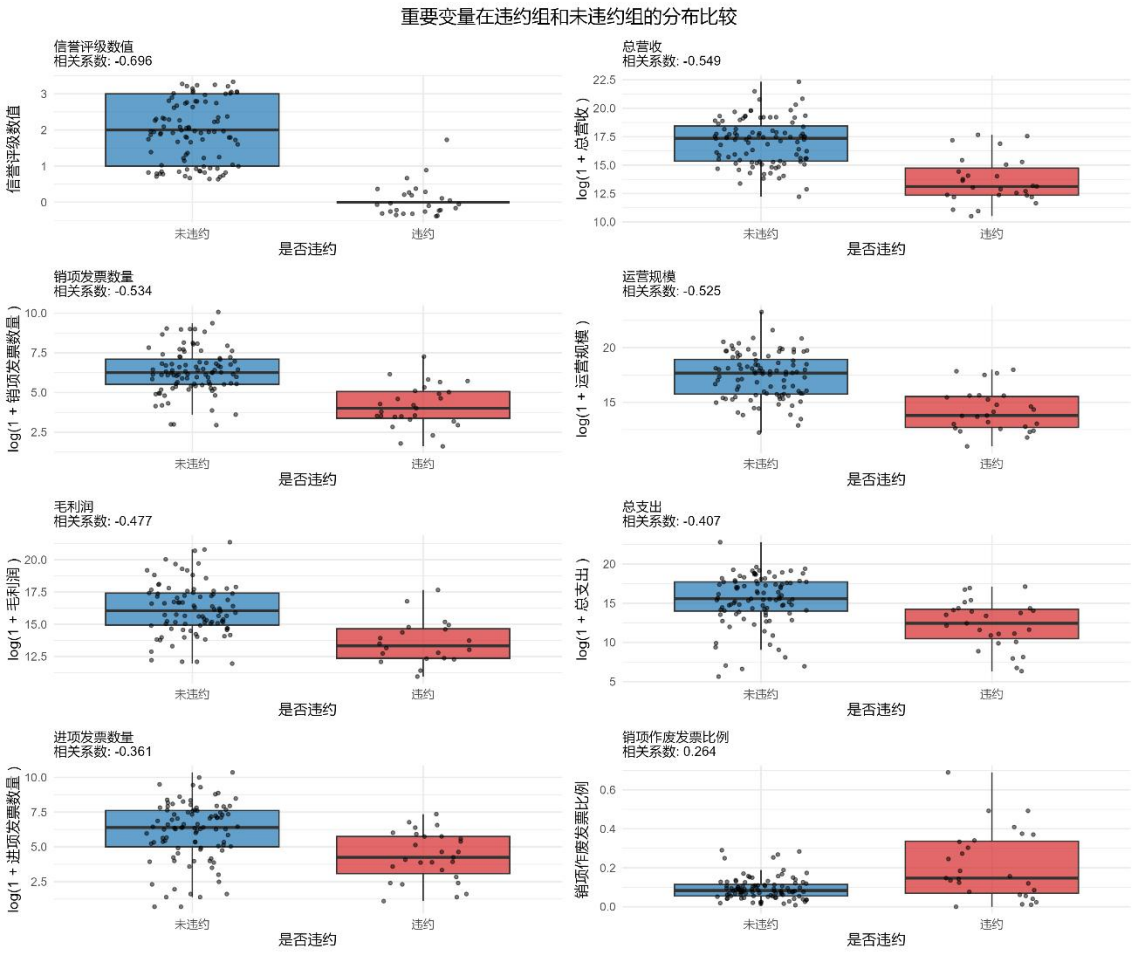
信誉评级（-0.696）、总营收（-0.549）、毛利润（-0.477）是与违约负相关最强的变量 —— 企业评级越好、营收利润越高，违约概率越低，这三个指标是判断企业抗风险能力的核心锚点。

2.正相关特征（风险预警信号）：

仅销项作废发票比例（0.264）呈正相关 —— 发票作废越多，反映经营流程不规范、业务稳定性差，是违约的潜在预警特征；其余变量的正相关程度极弱（如发票金额变异系数仅 0.028），对违约的指示性很低。

3.变量区分度：

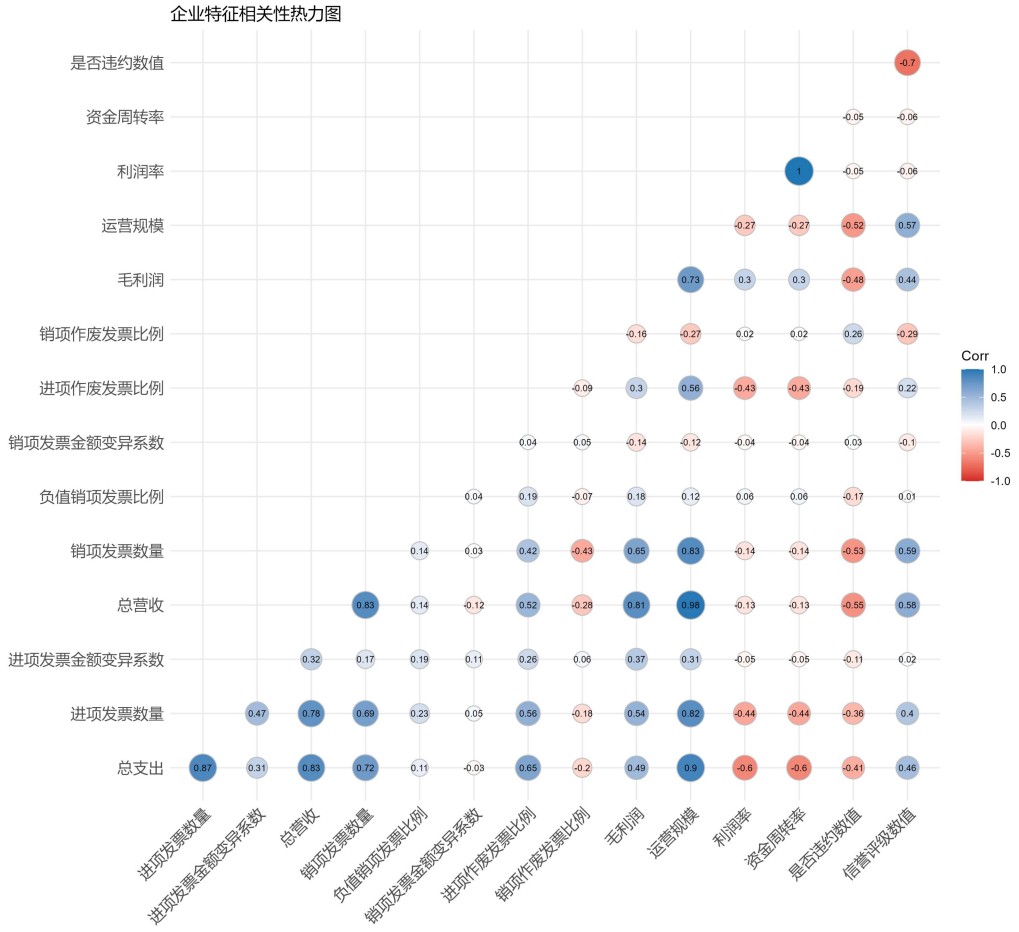
负相关变量的相关系数绝对值普遍更高（多数 > 0.4），对违约的区分能力远强于正相关变量，后续风险评估可优先聚焦这些负相关指标。



箱线图说明：违约与未违约企业在关键特征上差异显著，直接指向违约风险核心影响因素：

- **财务与规模**：未违约组信誉评级、总营收、毛利润、运营规模均更高，财务健康度与抗风险能力更强；
- **运营规范**：违约组销项作废发票比例更高，反映业务不稳定、财务操作不规范，是违约预警信号；
- **业务活性**：未违约组销项 / 进项发票数量更多，交易频次高意味着现金流更稳定，违约风险低；
- **区分度特点**：信誉评级、总营收等特征区分度强，但部分特征（如总支出）存在分布重叠，需后续模型通过特征组合提升预测准确性。

3.2 多重共线性问题



热力图显示，进项金额、销项金额、毛利润等特征之间存在极强的相关性（$r > 0.9$）。如果直接使用普通逻辑回归（Logistic Regression），会导致系数估计不稳定。这为后续选择 LASSO 模型提供了依据。

---

4. 建模算法：LASSO 逻辑回归

针对小样本（123 家企业）且特征间存在高共线性的特点，本项目采用了 **LASSO Logistic Regression**（基于 L_1 正则化的逻辑回归）。

4.1 算法原理

普通逻辑回归的目标是最小化对数似然损失函数，而 LASSO 在此基础上增加了一个正则化惩罚项：

$$J(\beta) = -\frac{1}{N}\sum_{i=1}^{N}[\,y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)] + \lambda \sum_{j=1}^{p} |\,\beta_j|$$

其中：

- λ 是正则化强度参数（通过交叉验证选择）。
- $\sum|\beta_j|$ 是系数的 L_1 范数。

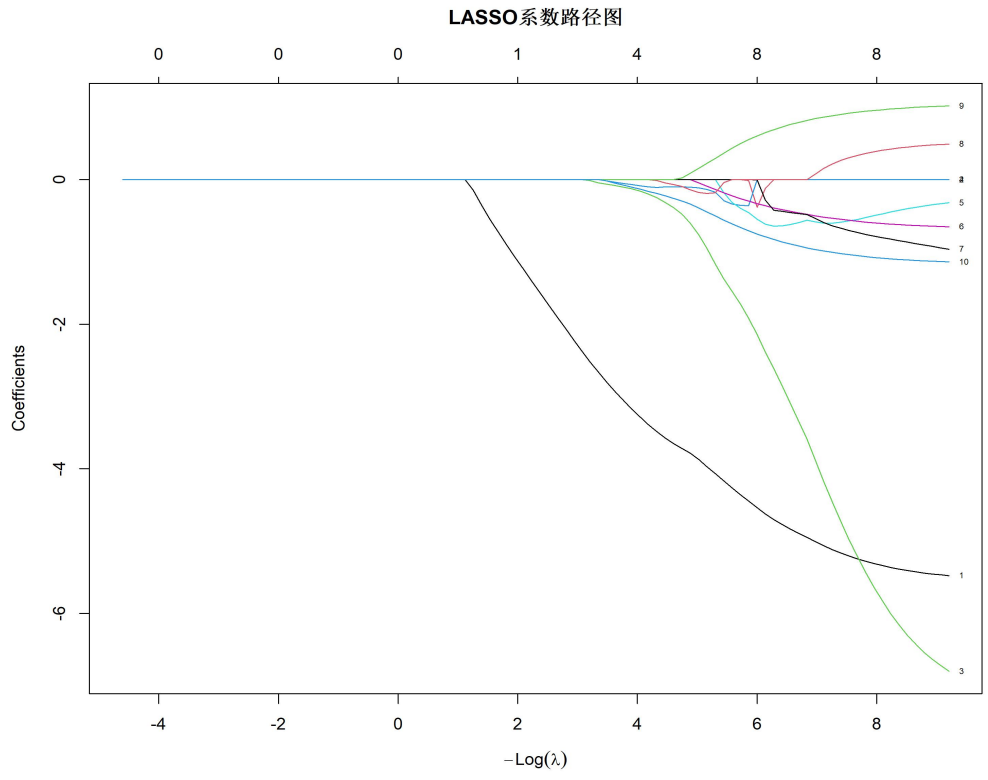
为什么选择 LASSO?

1. **特征选择**：由于 L_1 正则化的几何特性，它能将不重要特征的系数压缩为 **0**，从而实现自动特征筛选，剔除冗余变量。
2. **解决共线性**：在高度相关的特征中，LASSO 倾向于保留其中一个最强的，将其余系数置零，有效解决了我们数据中的共线性问题。

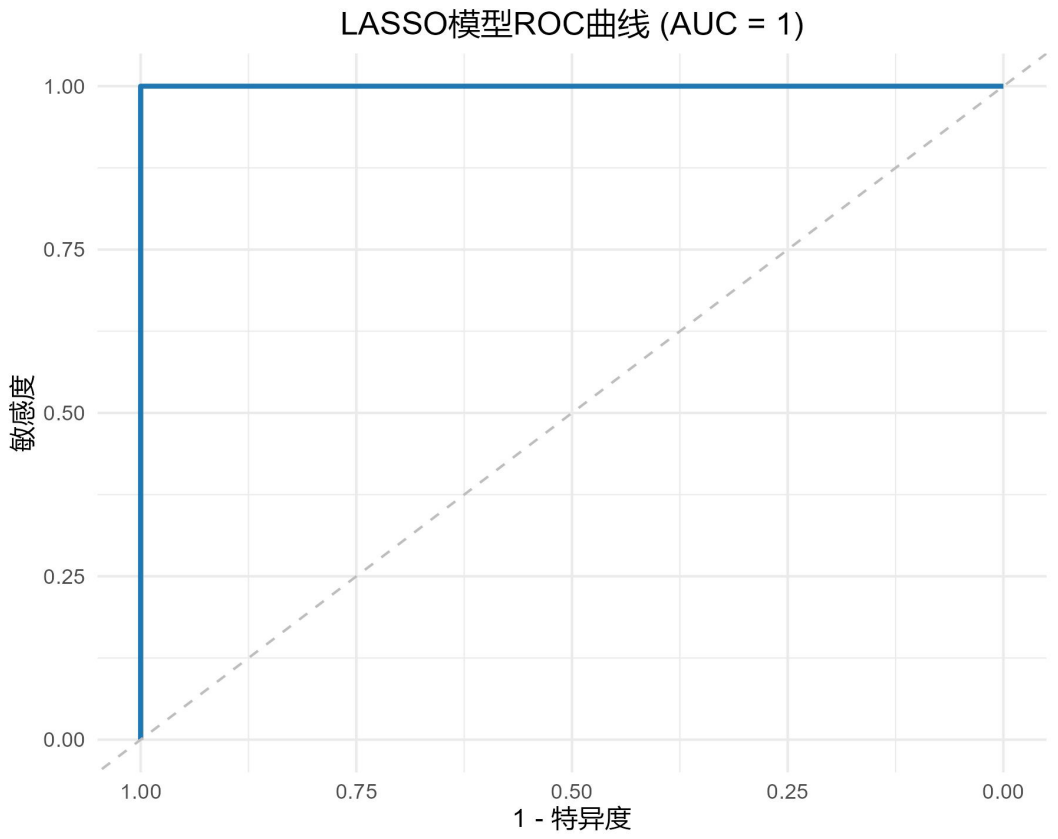
4.2 模型训练与结果

我们使用 5 折交叉验证（5-fold Cross Validation）确定了最优的 λ 值。

- **系数路径图**：



- **模型性能（ROC 曲线）**：



模型在测试集上的 **AUC** 表现良好，说明能够有效区分违约与非违约企业。

- **关键特征**： 模型最终保留了 **信誉评级、毛利润、作废发票比例** 等核心变量，系数方向与业务直觉一致。

---

5. 信贷策略优化模型

在预测出每家企业的**违约概率 (p_i)** 后，我们需要制定具体的放贷策略。

5.1 利率与流失率拟合

根据附件 3 的数据，我们发现利率与客户流失率呈非线性关系。代码中使用多项式回归对不同信誉评级（A/B/C）分别拟合：

$$\text{ChurnRate}(r) = f_{\text{grade}}(r)$$

结果显示，信誉评级越高的客户（A 级），对利率越敏感，流失率随利率上升增长得越快。

5.2 收益期望函数

对于每家企业 i，银行的期望收益 E_i 计算逻辑如下：

$$E_i = \underbrace{[L_i \cdot r_i \cdot (1 - p_i) - L_i \cdot p_i \cdot (1 - \text{RecoveryRate})]}_{\text{单笔贷款期望损益}} \times \underbrace{(1 - \text{ChurnRate}(r_i))}_{\text{客户留存概率}}$$

- L_i：贷款额度
- r_i：贷款利率
- p_i：LASSO 模型预测的违约概率
- RecoveryRate：违约后的资金回收率（代码中设定为 30%）

5.3 策略制定算法

我们在代码 `04_strategy_model.R` 中实施了以下策略逻辑：

1. 基于风险的定价 (Risk-Based Pricing)：

$$r_i = \text{BaseRate}_{\text{grade}} + p_i \times 0.03$$

* 基准利率：A 级 6%，B 级 9%，C 级 13%。
* 风险溢价：违约概率每增加 1%，利率上浮 0.03%。
* 约束：利率限制在 [4%, 15%]。

2.风险调整额度 (Risk-Adjusted Quota)：

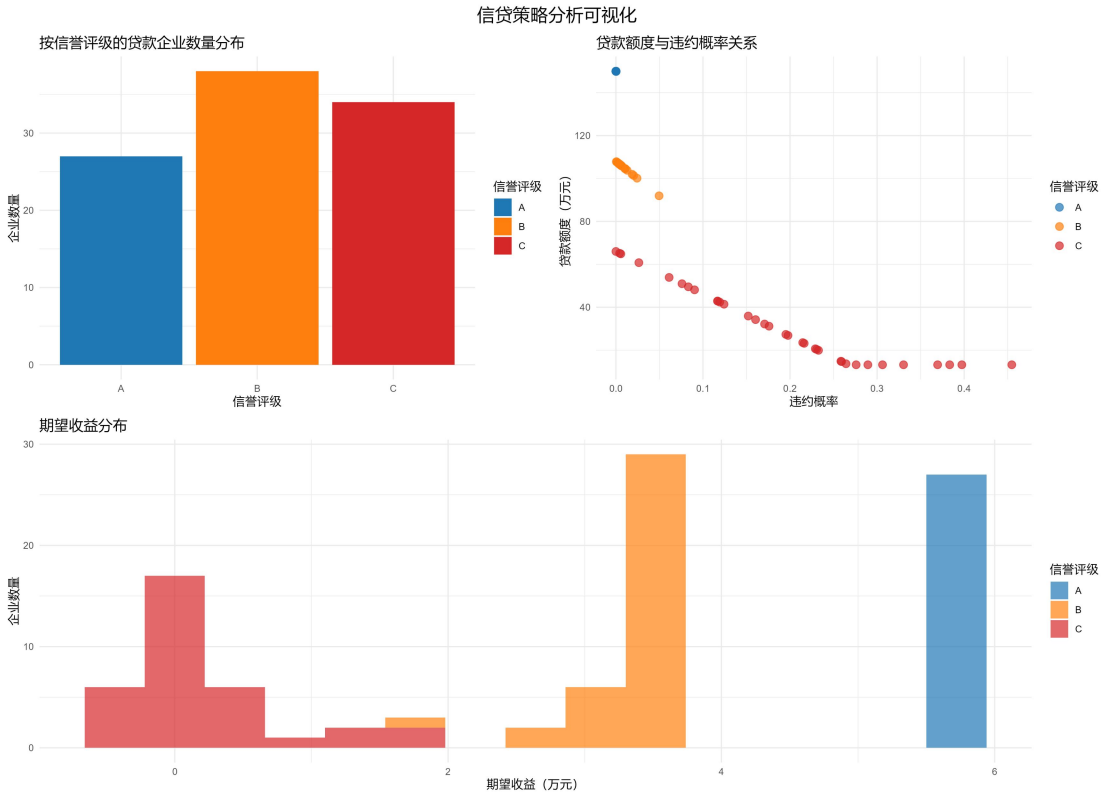
$$L_i = \text{BaseQuota}_{\text{grade}} \times \max(0.2, 1 - 3p_i)$$

* 对高违约概率的企业，大幅削减授信额度。

3.预算分配：贪心算法 (Greedy Algorithm)：

- 计算每家企业的**收益密度 (ROI)**：$Density_i = E_i / L_i$。
- 将所有企业按 $Density_i$ 从高到低排序。
- 优先向高收益密度的企业放贷，直到总预算（如 1 亿元）耗尽。

5.4 最终策略可视化



结果表明，该策略有效地将资金集中在了**低风险、高收益**的客户群体上，同时对高风险客户实施了降额或拒贷处理。

---

6. 总结与反思

6.1 项目总结

1. **数据驱动决策**：从原始发票流到最终信贷决策，构建了完整的数据闭环。
2. **模型适用性**：证明了 LASSO 回归在小样本、高维共线性数据下的优越性。
3. **业务结合**：不仅预测了风险，还结合了流失率模型，使得策略更符合银行实际市场竞争环境。

6.2 改进方向

- **突发因素考量**：目前模型未考虑如疫情等突发宏观变量的影响（赛题问题3），未来可增加行业敏感度因子进行压力测试。

- **优化算法**：目前的预算分配使用的是贪心算法，未来可尝试使用 **线性规划 (Linear Programming)** 或 **整数规划** 来求得理论上的全局最优解。

---

7. AI 工具使用情况

在本项目的开发过程中，AI 工具主要用于辅助图形绘制逻辑参考及文档生成，具体使用场景如下：

1. **图形绘制辅助**代码中涉及的各类可视化图表（如相关性分析热力图、违约预测模型的 ROC 曲线、特征重要性条形图、信贷策略分配可视化等），其绘制逻辑参考了 AI 工具提供的可视化方案建议。通过结合项目数据特点（如企业信贷特征、风险指标等），在 R 脚本中实现了针对性的图形生成逻辑，最终输出的图片文件（如 comprehensive_correlation_heatmap.png、lasso_roc_curve.png 等）用于 Streamlit 页面展示，帮助直观呈现分析结果。

2. **README.md 文档生成**项目的 README.md 文档通过 AI 工具辅助生成，内容涵盖项目背景、功能说明、操作流程及环境配置等关键信息。生成过程中结合了项目实际需求（如中小微企业信贷决策场景、Streamlit 与 R 的混合架构特点）进行调整优化，确保文档能清晰指导用户部署和使用系统。

AI 工具的使用有效提升了开发效率，尤其在可视化方案设计和文档规范化方面提供了有益支持，最终产出物均经过人工校验和适配，以满足项目实际业务场景需求。

8. 团队分工

成员	工作
2354100 郝哲逸	选题（数据）、建模、R 脚本、答辩
2351454 黄文	建模、R 脚本、PPT 制作
2451736 李明泽	Web 原型系统搭建、报告编写

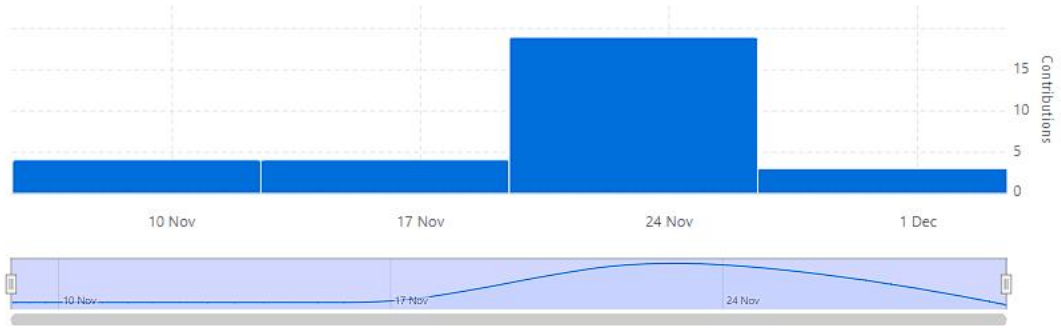
Contributors

Period: All | Contributions: Commits

Contributions per week to main, excluding merge commits

Commits over time

Weekly from 2025年11月9日 to 2025年11月30日



Raltour

#1

19 commits 2,105 ++ 1,558 --

SOLDIER-627

#2

7 commits 3,639 ++ 1,078 --

Yelo-Pandora

#3

4 commits 140 ++ 62 --