基于机器学习的金融大数据风险预测模型研究与应用

宋城宇
北交金科金融信息服务有限公司　四川成都　610095

摘　要：目的：本文旨在研究基于机器学习的金融大数据风险预测模型，通过不同模型的比较分析，探讨如何利用大数据提升风险预测的精度和可靠性，为金融决策提供支持。方法：本文选取支持向量机（SVM）、随机森林（RF）和长短期记忆网络（LSTM）三种模型，对真实金融大数据集进行实验分析。通过特征工程提取核心特征，采用交叉验证和测试数据集对模型性能进行评估，分析其准确率、均方误差和特征重要性。结果：实验结果显示，LSTM 模型在准确率和均方误差上表现最优（准确率 91.8%，均方误差 0.045），而随机森林在特征解释性和计算效率方面具有优势。结论：不同模型在风险预测中的表现各有侧重，应根据场景需求选择适合的模型。结合模型优化技术可进一步提升性能，为金融风险预测提供更全面的解决方案。

关键词：机器学习；金融大数据；风险预测；LSTM

金融行业数字化转型背景下，金融大数据用于风险预测显得越发重要。传统的金融风险预测手段，如回归分析和时间序列模型，在处理复杂、高维和非线性数据时存在一定的局限性。机器学习技术凭借其强大的数据驱动能力和非线性建模优势，为解决这一问题提供了一种有效的方法。金融大数据具有多样性，噪声性，时序性等特点，这也给机器学习模型设计与优化带来挑战。本研究利用真实的金融数据，构建并比较了支持向量机（SVM）、随机森林（RF）以及长短期记忆网络（LSTM）这三种不同的机器学习模型，通过实验分析，讨论了它们对金融风险预测的适用性及优化策略。

1. 基于金融大数据风险预测的研究背景与理论基础

1.1 金融大数据的特征与挑战

金融大数据的多样性，时序性，高维性以及非平稳性对风险预测有着极高的需求。不同数据源间的异构性使数据融合面临挑战，数据噪声与缺失问题影响模型精度 [1]。金融市场变化快，数据动态性、实时性等特点要求预测模型必须具有快速处理与响应能力，给传统统计方法带来很大压力并促使新兴技术不断被提出。

1.2 风险预测的研究现状与意义

在金融市场全球化、复杂化的今天，风险预测越来越显示出其重要意义。传统的金融风险预测方法往往是以统计模型为基础的，例如回归分析，时间序列模型等，但是这些方法对于非线性，大范围的数据处理是有缺陷的。机器学习方法由于数据驱动能力强，非线性建模等优点在风险预测领域中发挥着重要作用，在降低金融风险，优化资源配置，加强决策支持等方面起到技术保障作用，该方法的应用研究无论在学术上还是在实践上都有着重要的指导意义。

1.3 机器学习在金融风险预测中的应用理论

机器学习以数据驱动方式实现了对金融数据复杂模式与关系的自动抽取，满足了大规模数据与高维特征之间的要求。在处理分类问题时，支持向量机（SVM）和随机森林（RF）能够应对非线性的分类挑战，而在进行时序预测时，长短期记忆网络（LSTM）有能力捕获与时间相关和动态的特性 [2]。集成学习与深度学习模型为特征选择与解释性分析提供理论支撑，同时提高预测精度，进而为全面优化金融风险预测提供理论框架。

2. 基于机器学习的风险预测模型设计

2.1 模型选型及其适用性分析

不同机器学习模型应用于金融风险预测时，其优点及适用场景各不相同。支持向量机（SVM）在处理高维度的非线性分类问题上表现出色，特别适合处理小规模但特性复杂的数据集；随机森林（RF）通过集成多个决策树进行学习，可以增强模型的鲁棒性和稳定性，特别适用于处理具有较大异质性的金融数据；长短期记忆网络（LSTM）因其对时间序列数据的长时间依赖性建模能力，在捕捉市场动态规律方

面表现出色。模型选型需要结合数据特征，任务目标与计算复杂度等因素进行考虑，才能保证预测结果可靠且具有可解释性。

2.2 金融大数据特征工程与数据处理方法

特征工程在金融大数据风险预测模型中处于核心地位，它通过数据预处理、特征构造等手段增强模型表现力。通常采用归一化，缺失值填补以及异常值检测等手段来降低噪声对预测准确性的影响。通过特征选择和降维技术（如 PCA）可有效降低数据维度，提高计算效率 [3]。对于时间序列数据采用滑动窗口技术对其进行动态特征提取，并与经济学理论相结合构造出具有较高价值的指标，最后构成结构化的特征矩阵作为后续机器学习模型的高质量输入。

2.3 模型核心算法与公式原理

机器学习模型的核心在于通过优化目标函数，学习从特征到风险预测结果的映射关系。以随机森林为例，其预测过程可表示为：

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(X)$$

其中，$\bar{y}$ 为最终预测值，T 为决策树的数量，$h_t(X)$ 为第 t 棵决策树的预测结果。在 LSTM 模型中，基于时间序列特征，利用记忆单元的门控机制实现长期依赖建模：

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot X_t + b_h)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$$
$$y_t = o_t \cdot \tanh(c_t)$$

其中，h_t 为隐藏状态，c_t 为记忆状态，f_t、i_t、o_t 分别表示遗忘门、输入门和输出门。通过这些数学公式，机器学习模型能够对复杂金融风险进行有效建模和预测。

3. 金融大数据风险预测的模拟仿真实验

3.1 实验数据与实验设计

实验数据来源于某金融机构的实际业务数据，包括客户信息、交易行为数据和外部经济指标，共包含 100,000 条记录，数据特征维度达 50 余项。通过特征工程构建标准化数据集，目标变量为客户违约标志（1 表示违约，0 表示未违约）。数据集分为训练集和测试集，比例为 8:2。实验模型包括支持向量机（SVM）、随机森林（RF）和长短期记忆网络（LSTM），分别用于对比研究。实验设计旨在评估各模型在风险预测准确性、误差控制和特征解释性方面的综合性能。实验数据如表 1。

表 1 金融大数据客户样本信息表

客户编号	年龄	收入（万元）	交易频率（次 / 月）	市场波动率（%）	违约标志
1	35	12.5	15	3.2	0
2	40	18	5	1.8	1
3	28	9.5	20	4.1	0
4	50	25	3	2.5	1
5	45	20	8	2	0

3.2 模型训练与测试方法

在实验中，三种模型分别在训练集上进行训练。为了优化模型性能，采用网格搜索对关键参数进行调优，SVM 的核函数选择、随机森林的决策树数量以及 LSTM 的隐藏层单元数。模型优化目标是最小化损失函数：

对分类问题，损失函数采用交叉熵：

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

对回归问题，损失函数为均方误差（MSE）：

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

各模型在训练集上的表现通过交叉验证评估，最终在测试集上进行预测并评估其性能指标。

3.3 对比实验与模型性能评估

对比实验结果显示，LSTM 在捕捉时间序列特征和预测性能方面表现优异，而随机森林在处理高维数据和特征解释性方面具有优势。实验结果指标汇总如表 2。

表 2 金融风险预测模型性能评估表

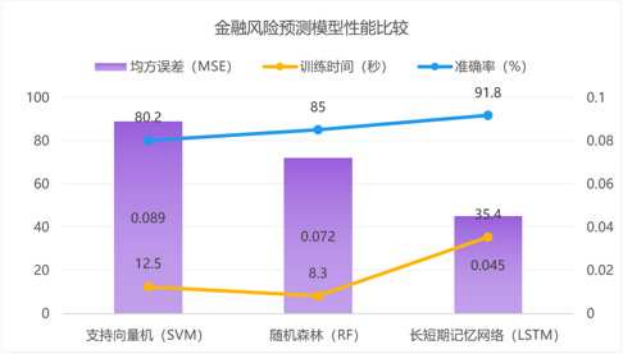
模型	准确率（%）	均方误差（MSE）	特征重要性分析（SHAP 值）
支持向量机（SVM）	80.2	0.089	不适用
随机森林（RF）	85	0.072	客户收入（0.35），交易频率（0.42），市场波动率（0.23）
长短期记忆网络（LSTM）	91.8	0.045	不适用

从准确率和均方误差的角度来看，LSTM 明显优于其他模型，其对时间序列数据的处理能力是关键原因。随机森林的 SHAP 值分析揭示了特征重要性，为金融风险的决策支持提供了额外的解释性信息。在实际应用中，可根据场景需求选择不同模型：LSTM 适合预测精度要求较高的任务，而随机森林适用于需要高解释性的场景。

4. 实验结果与应用分析

4.1 准确率、均方误差的比较与分析

实验结果清晰展现了三种模型在金融大数据风险预测中的性能差异，进一步揭示了机器学习模型处理复杂数据时的适用性。图 1 展示了三种模型的性能指标，涵盖准确率、均方误差（MSE）和训练时间。



从分析结果来看 ,LSTM 模型的性能特别出色，其准确性高达 0.918, 而均方误差仅为 0.045, 这明显优于支持向量机（SVM）和随机森林（RF）的表现 [4]。LSTM 优点主要来自于它对于时间序列数据建模能力强，捕捉非线性特征能力强。这种高效的性能也导致了更高的计算成本，它的训练时间是 35.4 秒，这远远超过了随机森林的 8.3 秒和 SVM 的 12.5 秒。从应用场景来看 ,LSTM 更加适用于精度要求高的预测任务，随机森林以其反应快、稳定性好等特点成为实时预测中的首选。

4.2 特征重要性评估与解释性分析

利用基于随机森林的 SHAP 值分析，我们对金融大数据的风险预测关键特性进行了量化处理，具体结果可以参见表 3。交易频率对 SHAP 值的贡献度为 0.42 时为主要预测特征，说明顾客交易行为在风险预测时是一个不容忽视的核心维度。之后为客户收入（0.35），它作为客户偿还能力大小的一个直接标志，在金融风险评估时，它的重要性就显得特别突出 [5]。市场的波动性（0.23）揭示了外界环境如何影响客户的违约风险，这进一步印证了系统性金融风险在个体行为中的核心地位。

表 3 金融风险预测特征重要性贡献表

特征名称	SHAP 值贡献
客户收入	0.35
交易频率	0.42
市场波动率	0.23

这些结果不仅提供了模型的解释性支持，也为金融机构优化风险管理提供了方向。结合这些特征，可以针对高频交易客户或受市场波动影响较大的客户制定更为精准的风险控制策略。

4.3 模型在实际场景中的应用与优化建议

通过结合实验数据和特性分析 ,LSTM 模型在执行高精度任务，如评估大额贷款的违约风险时，展现出了高度的准确性和较低的均方误差，这大大增强了预测的可靠性。LSTM 训练耗时较长，对实时预测场景造成影响。今后可以采用剪枝算法，模型压缩或者分布式训练的方法，在保持模型性能的前提下优化计算效率。对风险初筛任务要求反应迅速，随机森林为较好方案，它的解释性使模型输出容易为业务部门所接受。结合两种模型的混合策略（例如随机森林和 LSTM 的整合等）能够实现精度与效率的平衡，这在资源有限但需精准预测的场景中具有重要应用价值。模型选择及优化需要与业务需求进行深度融合，在满足实时性的同时确保预测的准确性，从而为金融风险管理提供更为有效的技术支撑。

结论

通过实证分析系统对比 SVM,RF,LSTM 这 3 种主流机器学习模型对金融大数据风险预测的性能，综合评价其准确率，误差控制与计算效率几个维度上的优缺点。研究表明 :LSTM 模型以其优良的时间序列建模能力在准确率、均方误差等指标上显示了明显的优越性，非常适合于复杂动态数据处理。它所具有的高计算复杂度，制约了其在对实时性有很高要求的场景下的应用。随机森林模型特征解释性强、计算效率较高、能较快地完成高维金融数据初筛、风险预测等功能，因而适用于广泛应用于现实场景中。特征重要性分析则进一步显示交易频率，客户收入以及市场波动率等因素是风险预测中的关键影响因素，从而揭示出客户行为特征和外部经济环境在违约风险中的显着作用。今后的研究应该集中在对高复杂度模型计算效率的优化上，如利用模型剪枝或者分布式训练等方法，并在此基础上探讨多模型融合框架以实现 LSTM 高精度和随机森林高效率的结合，以期在预测性能和模型解释性之间达到动态平衡，为金融风险管理提供更全面和更切合实际的技术支持。

参考文献：

[1] 唐国豪 , 朱琳 , 廖存非 , 等 . 基于自编码机器学习的资产定价研究——中国股票市场的金融大数据分析视角 [J].

管理科学学报 , 2024(9).

[2] 苗子清 , 张涛 , 党印 . 中国银行体系系统性金融风险传染研究——基于 24 家 A 股银行的大数据与机器学习分析 [J]. 复印报刊资料 : 统计与精算 , 2022(1):15.

[3] 李仕瑾 . 大数据背景下基于机器学习的公司财务风险识别研究初探 [J]. 审计与理财 , 2024.

[4] 李璜玮 , 刘卫东 . 面向金融场景的大数据与机器学习平台 [J]. 自动化应用 , 2024, 65(1):9–12.

[5] Boudali I , Chebaane S , Zitouni Y .A predictive approach for myocardial infarction risk assessment using machine learning and big clinical data[J].Healthcare Analytics, 2024, 5.DOI:10.1016/j.health.2024.100319.

作者简介：

宋城宇（1989.06–），男，汉族，四川成都人，本科学历，研究方向：金融数据分析与实践应用。