 DOI: https://doi.org/10.12345/fm.v3i5.17472

Building an Interpretable Credit Scoring Framework: Integrating Machine Learning with Domain Expertise

Lusi Yang Hailong Zhang

School of Finance, Renmin University of China, Beijing, 100872, China

Abstract

In the digital finance era, the management of credit risk is fundamentally reliant on the predictive accuracy of credit assessment models, crucial for the risk control strategies within financial institutions. However, traditional credit scoring methods encounter significant challenges when processing the ever-increasing data volumes. Tree-based machine learning algorithms, such as LightGBM, are popular due to their efficient data processing and excellent predictive capabilities. Nonetheless, their intrinsic "black box" nature hampers model interpretability, conflicting with the trend towards transparency and explainability that financial regulatory bodies are increasingly advocating for. With regulatory focus sharpening on the prevention of algorithmic biases and the assurance of equitable processing, it is imperative to develop credit scoring models that achieve both predictive accuracy and interpretability.

Keywords

credit scoring model; large-scale data; model predictive performance; interpretability

构建可解释的信贷评分框架：整合机器学习与领域专业知识

杨路思 张海龙

中国人民大学财政金融学院，中国·北京 100872

摘 要

在数字金融时代背景下，信贷风险管理依赖于信用评估模型的高预测精度，这对金融机构的风险控制起着至关重要的作用。然而，面对日益增长的数据量，传统的信贷评分方法在处理大规模数据集时受到挑战。基于树的机器学习算法，如 LightGBM，因其高效的数据处理能力和优异的预测性能备受青睐，但其固有的“黑盒”特性却影响了模型的可解释性，这与金融监管机构对模型透明度和解释性要求的趋势相违背。鉴于防止算法偏差和保证处理过程的公平性已成为监管关注的焦点，开发的同时满足预测性能和可解释性标准的信贷评分模型成为当下风险控制领域的迫切需求。

关键词

信贷评分模型；大规模数据；模型预测性能；可解释性

1 引言

在数字金融时代的背景下，信用评估模型已经成为金融机构进行风险控制的关键工具。这些模型的预测准确性是影响信贷业务风险管理效率和质量的决定因素。然而，由于数据量的显著增加以及特征维度的快速扩展，传统的信贷评分方法在处理大规模数据集时面临着显著的挑战。基于树的机器学习算法，如 LightGBM，由于其在处理庞大数据集方面的高效性和卓越的预测能力而备受推崇。但是，这些算法的“黑盒”特性限制了模型决策过程的透明度，使其难以满足金融信贷场景中对解释性的高要求。因此，在提升模型的预测性能的同时增强其可解释性，成为风险控制领域面临的重要挑战。

【作者简介】杨路思（1992–），女，中国天津人，本科，从事风控研究。

2022 年 1 月，原银保监会在发布的《关于银行业保险业数字化转型的指导意见》（银保监办发〔2022〕2 号），对模型和算法风险管理提出了明确要求。根据该指导意见的第二十六条，金融机构必须建立全面的模型和算法风险管理框架，制定相应管理制度，进行数据的准确性和充足性的交叉验证与定期评估。同时，必须审慎设定客户筛选和风险评估等模型参数，并在压力情境下验证这些参数。金融机构还需定期评估模型的预测能力及在不同场景下的局限性，以确保模型的可解释性和可审计性，并保障模型管理的核心环节能够自主控制，同时加强消费者权益保护以防止算法歧视 [1]。

随着中国相关法律法规的实施，监管机构越来越注重模型的透明度和可解释性，尤其是在防止模型偏差和确保处理过程的公平性方面。金融服务提供者被要求透明地向借款人解释贷款被拒绝的原因。因此，开发能够满足性能指标并符合解释性要求的信贷评分模型变得尤为重要。传统的逻辑

回归模型由于其参数估计透明及模型结构直观，更易于解释和理解，特别是通过变量的权重证据（WOE）转换及模型参数的直接解释，突出了其在可解释性方面的优势。这种解释性便于说明各独立变量对最终信用评分的具体影响。不过，逻辑回归在处理特征众多且存在复杂非线性关系的大型数据集时，可能无法捕捉其中的复杂性，这在某些情况下可能会降低模型的预测能力。

针对上述问题，本研究开发了一个结合机器学习技术与领域专业知识的创新型信贷评分框架。该框架旨在通过精确特征选择和机器学习算法的应用，优化模型在处理大规模数据集的预测性能，同时确保评分结果具备可解释性。在模型的早期构建阶段，我们运用信息价值（Information Value，IV）分析，筛选出了在不同时间段均表现出高稳定性和预测能力的特征，同时排除了那些具有高度相关性的变量，确保模型聚焦于对信贷风险评估最为重要的因素。

在构建信贷评分体系的过程中，我们没有直接采用逻辑回归模型作为评分卡的基础。相反，我们进行了详尽的权重证据（Weight of Evidence，WOE）分布分析，并结合行业专家的专业知识和经验，精选出既具有明确商业含义又与信贷风险密切相关的特征集。通过 WOE 转换，本研究量化了每个特征与违约率之间的关系，为构建一个功能强大且易于理解的评分模型打下了基础。此方法不仅提升了模型的预测精度，还增强了其可解释性，使模型能直观展示各变量的相对重要性，并满足监管机构对信贷决策透明度的要求。

为了深入挖掘数据中的复杂结构并提升预测精度，我们选用了 LightGBM 算法，这是一种基于梯度提升的高效机器学习算法。结合领域知识和 WOE 分析，我们不仅增强了模型的整体性能，而且维持了对特征影响力的可解释性，平衡了预测精确度与模型可解释性之间的关系。

2 相关工作

信贷评分模型是信用风险评估的关键手段，通过综合借款人的个人信息和历史违约行为来量化其风险水平，以预测借款人未来偿债能力的概率。在信贷评估模型被银行业所接纳之前，信贷决策往往依赖于银行贷款业务人员的主观判断，这种方法可靠性差，容易受个人偏见影响，且难以应对大批量申请。

信贷评分作为一种信用评估工具，在过去半个多世纪得到广泛应用。最早的信贷评分系统应用于信用卡领域，据 Anderson（2007）所述 [2]，美国的第一个信用卡信贷评分模型诞生于 1941 年左右，依据的评分参数包括申请人的职务、当前职位的工作年数、现居住地址的年数、银行账户和人寿保险政策的详细信息、性别及月还款额。

随后，在 1965 年，费尔、艾萨克公司（FICO）成立，旨在提升消费信贷评估的效率。FICO 评分系统，作为美国个人信用评估的事实标准，广泛应用于信用评价领域。该系统由五大主要影响因素组成，包括客户的信用偿还历史、持有的信用账户数量、使用信用的时长、正在使用的信用类型和新开立的信用账户，这五项在评分过程中被赋予不同的权重 [3]。

随着信贷评分模型及其统计学基础的研究日益深入，学者们对其应用原理和局限性有了更清晰的认识。在这方面，Bolton（2009）对线性回归模型的基本假设和它在信贷评分中的不足进行了详尽的挖掘，揭示了为何线性回归不适宜构建信贷评分模型。此外，该研究还描述了从简单逻辑回归到更为复杂的多变量逻辑回归的转变过程，并证明了在小数据量下逻辑回归的有效性。同时解释了证据权重（WOE）这一概念，即将与选择相关的风险转化为线性度量，大大简化了人脑对风险的评估与理解，从而为基于概率的决策提供了强大的解释性支持 [4]。

随着大数据时代的到来及先进 GPU 技术的发展，为了应对大规模数据处理这一挑战。微软研究团队开发了 LightGBM，这是一种基于树的学习算法的梯度增强框架。LightGBM 专为高效率的分布式计算设计，具备如下显著特点：训练速度快，效率高，对内存的需求低，模型精度高，并且支持并行处理、分布式计算以及 GPU 加速学习。这些特性使得 LightGBM 能够有效处理大规模数据集，为现代数据分析和机器学习领域提供了有力的工具 [5]。

然而，机器学习模型因其“黑盒”的性质，在提供可解释性和可信性方面面临挑战。尤其在受到密集监管的金融信贷行业，模型的解释性受到严苛的监管要求。监管机构强调，模型不仅需要展现出精确的预测能力，还必须能够阐明其决策逻辑。

在应对机器学习模型可解释性的挑战时，学术界已经提出了多种方法。为了深化对复杂模型预测结果的理解，Su-In Lee 已经开发了名为 SHAP（Shapley Additive Explanations）的解释框架 [6]，它为每个特征分配一个针对某一特定预测的重要性值，以此判定各特征在模型决策中所起的作用。SHAP 的创新之处在于提出了一类新的加性特征重要性度量，并从模型内部机制的角度出发，极大地促进了模型解释性的理解。尽管如此，SHAP 框架尚未完全满足金融监管在模型可解释性方面的严格要求。

本研究构建并推进了前人的工作，提出了一种创新的分析框架，该框架融合了传统信贷评分模型的高度可解释性与机器学习模型的出色预测准确度。通过采用信息价值（IV）和证据权重（WOE）对特征进行评估和转化，本研究进一步运用 LightGBM 算法进行了高效的模型构建。为了增强模型的可解释性，论文详细提供了 WOE 值以确保每项信贷决策都能够对监管机构和借款人进行透明化解释。此方法不仅遵守了监管规定，同时也实现了高性能的预测能力。

3 模型方法

3.1 组件说明

3.1.1 证据权重（WOE）

证据权重（Weight of Evidence，WOE）是一种在金融决策和信用评分中常用的技术，其作用在于量化特征变量对于风险的相对贡献，且易于理解和解释。计算公式如下：

$$WOE_i = \ln(N_i/P_i) - \ln(\sum_{i=1}^{n} N_i/\sum_{i=1}^{n} P_i) \quad (1)$$

其中，n 表示将一个特征变量划分成 n 组，n 一般小于 10。N_i 表示第 i 组中好客户样本数（未违约），P_i 表示第 i 组中坏客户样本数（违约）。

证据权重（WOE）的解释力源于其直接以对数比例的形式表现风险，将各特征组内好与坏的结果之比转化为数值，直观地揭示了特征对目标事件（违约）预测能力的强弱。正的 WOE 值说明该属性下好的结果占比较大，反之，负的 WOE 值则暗示坏的结果占主导，这为决策者比较不同特征区间的风险大小提供了直观的依据。

3.1.2 信息价值（IV）

信息价值（Information Value，IV）是评估特征变量在模型中的预测能力，因而可以用来筛选变量。其计算公式如下：

$$IV = \sum_{i=1}^{n} [(\frac{N_i}{\sum_{i=1}^{n} N_i} - \frac{P_i}{\sum_{i=1}^{n} P_i}) * WOE_i] \quad (2)$$

其中，n 表示将一个特征变量划分成 n 组，n 一般小于 10。N_i 表示第 i 组中好客户样本数（未违约），P_i 表示第 i 组中坏客户样本数（违约）。$WOEi$ 表示第 i 组的 WOE 值（按照公式 1 计算）。

信息价值（IV）可以理解为计算好客户（未违约）和坏客户（违约）分布差异来得到，其值越高意味着该变量对模型的预测效果越好。

3.1.3 特征相关性

pearson 相关性系数是评估两个变量的线性相关性，计算逻辑如下：

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \quad (3)$$

其中，x，y 表示不同的特征变量，$cov(x, y)$ 表示两个变量的协方差，σ_x 表示变量 x 的标准差，σ_y 表示变量 y 的标准差。$\rho_{x,y}$ 在 [-1, 1] 之间，$\rho_{x,y}$=—1 表示完全线性负相关，$\rho_{x,y}$=1 表示完全线性正相关，$\rho_{x,y}$=0 表示线性不相关。

一般，当 $|\rho_{x,y}| > 0.8$ 时，两个变量被认为存在较强的线性相关性，因此在模型构建过程中建议考虑剔除其中一方，以避免多重共线性的问题。剔除相关特征有助于降低模型复杂度，并可能提高模型的泛化能力。

3.2 模型框架

在本研究中，我们提出了一个综合的方法论框架，以构建一个既具有高预测性又具备良好解释性的信贷评分模型。我们采用的方法涵盖了从特征选择到模型评估的全流程见图 1，以下是详细的步骤描述。

3.2.1 第一步：信息价值（IV）筛选

我们对所有潜在的候选特征进行了信息价值（IV）的评估。信息价值被广泛认为是评定特征预测能力的重要指标，尤其在信贷评分领域中具有关键作用。为了确保选定特征在不同时间段保持预测能力的稳定性，我们对每项特征按照公式（2）计算出不同时间段的 IV 值。并评估其预测能力和稳定性，筛选出预测能力强且稳定的特征用于后续的分析。

3.2.2 第二步：剔除相关性高的特征

对第一步筛选后的数据集，实施皮尔逊相关性分析，以识别并剔除数据集中高度相关的特征变量。通过这种方法，可以降低模型输入的维度，有效减少了模型的复杂性，并提升了计算效率。具体来说，我们按照公式（3）计算变量之间的皮尔逊相关性系数，当两个变量的相关系数超过设定的阈值时，我们会移除 IV 低的变量。

3.2.3 第三步：基于 LightGBM 的特征重要度评估

经过基于信息价值（IV）的初步筛选和剔除高度相关性的特征，本研究随后采用了 LightGBM 模型来进一步评估特征的相对重要性。依据 LightGBM 模型输出的特征重要性，选取了重要性得分高于零的特征进行后续分析。这种做法有效地将特征候选集减少到一个人工审核和分析的规模。

3.2.4 第四步：WOE 分组计算

经过前面三步的筛选后，对特征重要度大于 0 的特征，进行离散化分组，并按照公式（1）计算每组的 WOE 值。

3.2.5 第五步：依据行业规则进行筛选

结合行业经验和专业知识，对特征的 WOE 值进行深入分析，以筛选出既符合行业经验又具有良好可解释性的变量。通常，具有可解释性特征的 WOE 值应该呈现出单调性或 U 形。

3.2.6 第六步：构建 LightGBM 信贷评分卡模型

在经过行业经验和专业知识筛选后，选择出预测能力强且具有可解释的特征。将这些特征作为 LightGBM 算法的入模特征，用以训练和构建先进的信贷评分卡模型。

3.2.7 第七步：模型评估

对于构建的信贷评分卡模型，本研究采用 Kolmogorov-Smirnov（KS）等统计指标进行综合评估和优化。通过这一过程，我们能够识别并选择出具有较高预测能力的模型。确保选定的评分卡模型在实际应用中具备优秀的风险鉴别能力。如图 1 所示。

总体而言，本研究的方法框架在保持模型预测准确度的同时，着重关注模型特征的解释性，以满足金融风控领域对模型监管的要求。通过这些精心设计的步骤，我们能够构建出一个既符合业务需求又便于监管审核的信贷评分模型。

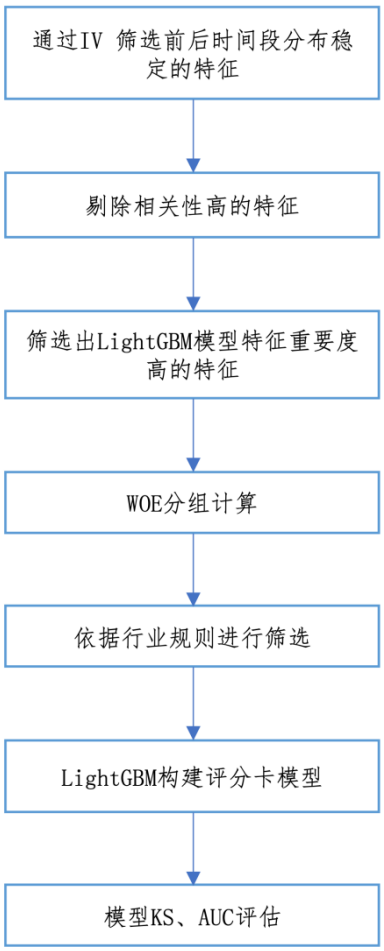

图 1 信贷评分框架

4 实验

4.1 数据

在构建风控评分卡模型时，数据的选择和处理对于模型的准确性和有效性至关重要。风控评分卡模型的数据来源主要包括申请人信息、征信报告数据、多头借贷数据、应用程序（App）行为数据以及第三方数据等。

申请人信息：这些信息通常涵盖申请人的个人基本信息，如年龄、性别、婚姻状况、教育程度、职业信息、收入状况等。申请人提供的这些信息有助于初步评估其信用状况和偿还能力。

征信报告数据：征信报告提供了申请人过去的信用历史，包含信用账户信息、信用卡使用、历史还款记录、贷款申请次数、逾期信息等。这些数据可以帮助金融机构了解申请人以往的信用行为和偿债记录。

多头借贷数据：这类数据反映了借款人在不同金融机构的借贷情况，包括贷款数量、贷款总额、不同类型的贷款情况等。多头借贷信息有助于评估借款人的债务负担和潜在的财务风险。

App 行为数据：金融科技应用程序中的用户行为数据，如登录频次、页面浏览习惯、交易行为等，这些数据可以揭示用户的金融活动习惯和偏好。

第三方数据：这些数据可能来自公共记录、其他金融服务提供商或行业数据库，提供更为全面的背景信息，如用户在电子商务平台的购买记录、社交媒体活动等。

在实际的风控评分卡开发过程中，将这些原始特征通过多种方式进行特征衍生，是提升模型预测性能的关键。通过统计分析、数学变换、分类和分段，以及复杂的算法模型，可以将原始特征变换和组合成更具预测力的高维特征。例如，可以通过计算申请人历史的平均贷款额，或者构建一个涵盖过去一年逾期频率的时间序列特征。

特征衍生的结果是，最终得到的特征空间非常庞大，有时甚至可达到 10 万维度。这就需要通过特征选择和降维技术去除噪声和无关特征，从而精简特征集合并提升模型的泛化能力。

论文实验采用的数据集为某公司内部现金贷数据，数据按照用户首次交易的时间划分为 2022 年 11 月到 2023 年 8 月，样本量为 20 万左右，衍生后的特征维度在 8 万左右，涵盖了用户基本信息、征信报告数据、多源数据和 App 行为数据。这些庞大的数据集合经过严格清洗和处理后，用于训练风控评分卡模型。由于涉及信息安全考虑，只展示部分特征样例数据，如表 1 所示。

表 1 特征样例表

类型	feature	说明
用户基本信息	特征1	学历
用户基本信息	特征2	年龄
用户基本信息	特征3	薪资
征信	特征4	历史贷额度金额平均值
征信	特征5	近6个月贷款金额超过5千机构数占比
征信	特征6	过去一年逾期金额
app	特征7	安装银行类app个数
app	特征8	安装投资理财类间隔最大值
多头	特征9	身份证维度机构查询账龄平均值
多头	特征10	手机号维度近2个月机构查询次数

4.2 特征筛选可解释变量

本次实验将用户首次交易时间在 2022 年 11 月到 2023 年 6 月的数据作为训练集，将用户首次交易时间在 2023 年 7 月的数据作为测试集，将用户首次交易的时间在 2023 年 8 月的数据作为验证集。

通过计算训练集、测试集、验证集特征的 IV，筛选出 IV 稳定且分布一致的特征；将筛选出的特征按照皮尔斯相关性小于 0.8 进行去重（相关性可以作为超参数自由设置）过滤相关性高的特征；将过滤后的特征放入 LightGBM 或者其他机器学习模型，筛选出特征重要度大于 0 的特征。部分特征结果如表 2 所示。

表 2 筛选后的变量

特征重要度	feature	train_iv	test_iv	valid_iv
5,004.55	学历	0.141	0.107	0.110
4,618.17	历史在贷额度金额平均值	0.222	0.204	0.200
1,119.83	身份证维度机构查询账龄平均值	0.049	0.048	0.052
508.56	近2个月安装app个数占比	0.140	0.145	0.147

经过上面的筛选，特征从 8 万多特征筛选出 120 多个特征。在通过 WOE 可解释分析，得到 58 个入模特征。

表 3 是特征“历史在贷额度金额平均值”的 WOE 值，从 WOE 值看出历史在贷金额平均值，随着金额的增加 WOE 值变大，违约率变低，符合业务逻辑，银行会对信用好的客户提供更高的放款额度。

表 3 历史在贷额度金额平均值的 WOE 和 IV

特征	分组	WOE	IV
历史贷额度金额平均值	(0.0, 623.733]	-0.647	0.035
历史贷额度金额平均值	(623.733, 1095.0]	-0.401	0.015
历史贷额度金额平均值	(1095.0, 1573.126]	-0.100	0.014
历史贷额度金额平均值	(1573.126, 2105.903]	-0.280	0.007
历史贷额度金额平均值	(2105.903, 2717.159]	-0.191	0.003
历史贷额度金额平均值	(2717.159, 3400.0]	-0.108	0.001
历史贷额度金额平均值	(3400.0, 4250.0]	-0.067	0.000
历史贷额度金额平均值	(4250.0, 5406.452]	0.174	0.003
历史贷额度金额平均值	(5406.452, 7434.25]	0.176	0.025
历史贷额度金额平均值	(7434.25, inf]	0.997	0.118

4.3 损失函数

评分卡模型是一个广泛应用于金融行业的分类模型，主要用于预测用户的逾期概率，即预测用户是否会在约定的还款期限内违约。为了训练这样的评分卡模型，通常会采用交叉熵损失函数（也称为对数损失）。

在二分类问题中，交叉熵损失函数特别适合衡量实际输出与预期输出之间的差异。具体来说，该损失函数会惩罚那些预测概率与真实标签大相径庭的预测。换句话说，如果模型对于一个真实标签为 1 的样本预测其逾期概率为 0，交叉熵损失会是很大的。同样，如果模型预测准确，损失就会减小。因此，使用交叉熵损失来优化模型可以有效地提高分类预测的准确性。

交叉熵损失函数如下，其中 y 为真实标签，$ypre$ 为模型预测结果：

$$loss = -ylog(y_{pre}) - (1-y)\log(1-y_{pre}) \quad (4)$$

4.4 模型结果对比

通过对比 LightGBM 和逻辑回归的评估指标，显示 LightGBM 在 KS 和 AUC 上都优于逻辑回归，LightGBM 测试集上的 KS 和 AUC 分别为 25.5 和 0.683，逻辑回归测试集上的 KS 和 AUC 为 21.4 和 0.649，这一结果展示出 LightGBM 在预测性能方面的优越性。

5 结语

本研究提出的信贷评分框架成功地解决了在大数据环境下如何构建一个既有高预测精度又符合监管要求的可解释模型的问题。我们通过一系列细致的特征筛选和分析过程，从初步的 8 万多个特征中筛选出了 58 个具有高预测能力和良好解释性的特征。这一过程不仅涉及了信息价值（IV）和相关性分析，还包括了基于 LightGBM 的特征重要性评分和基于 WOE 分布的业务规则和经验分析，确保了模型的高预测性和特征的可解释性。

在模型构建方面，我们采用了 LightGBM 算法，凭借其高效的计算速度和优异的处理大数据集的能力，展现出了较传统逻辑回归模型更加卓越的性能。模型评估结果显示，LightGBM 在关键的统计指标上，如 KS 和 AUC，均优于基线的逻辑回归模型，证明了我们方法的有效性。此外，我们特别注重模型的可解释性，通过可视化 WOE 值和提供特征贡献度解释，使得模型决策过程透明化，满足了金融监管对模型透明度和公平性的要求。

本研究的实践意义在于它为风控领域提供了一个结合最新机器学习技术和传统信贷评分模型优势的新框架。这一框架不仅提高了信贷评分模型在处理大规模数据时的预测能力，也确保了模型决策的可解释性和透明度，为金融机构在当前复杂的监管环境下做出更准确和合理的信贷决策提供了支持。未来的研究可以在此基础上进一步探索如何融合更多的机器学习技术和专业知识，以及如何在保证模型预测性能的同时进一步增强模型的可解释性和可信度。

参考文献

[1] https://www.gov.cn/zhengce/zhengcewenjianku[EB/OL].

[2] Anderson R. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation[J]. OUP Catalogue, 2007.

[3] 姜琳.美国FICO评分系统述评[J].商业研究,2006,37(20):81-84.

[4] Christine Bolton. Logistic regression and its application in credit scoring,2009.

[5] https://lightgbm.readthedocs.io[EB/OL].

[6] Scott M, Lundberg, Su-in Lee. A Unified Approach to Interpreting Model Predictions, 2016.