多源数据信用评级普适模型栈框架的构建与应用①

黄志刚[1] 刘志惠[1,2] 朱建林[3]

（1. 福州大学经济与管理学院；2. 福建商学院金融系；
3. 中国人民大学财政金融学院）

研究目标：研究多源数据普适模型栈在信用评级中的构建与应用。**研究方法：**从我国在线信贷行业实际情况出发，提出一种基于多源数据的普适模型栈评分框架，该框架可以根据各个申请人不同的数据基础，自由选择纳入评分模型数据，生成子评分模型，然后再将子评分模型转换为常见的信用评分卡模型。**研究发现：**基于多源数据的普适模型栈评分框架不但灵活、普适，其评分有效性也比单个 XGBoost 信用评分模型更好。**研究创新：**将机器学习模型与传统评分卡模型进行了完美的融合。**研究价值：**解决了机器学习模型在信用风险管理中可解释性差的问题。

关键词 风控 信用评分 评分模型 机器学习 模型栈框架

中图分类号 F832.39 **文献标识码** A

引 言

借贷过程主要包含资金交互和信息交互（Verstein，2011）。在现实中掌握资金的投资者（贷款人）会因为时间或空间的阻隔，得不到所需要的信息，而掌握信息的借款人又因为交易和谈判成本无法获得所需资金。随着信息技术的发展，网上信贷模式的出现为贷款人和借款人提供了一个信息交流的平台。所谓在线信贷是指贷款人和借款人依靠在线信贷平台实现信贷交易，而不是通过传统金融机构的线下模式（Bachmann 等，2011；Lin，2009），它是一种互联网加小额信贷的新型借贷模式。在线信贷一般会构建第三方在线信贷平台，以信用类贷款的形式将资金借给有需要的借款人，它为借款人与贷款人提供了公开透明的小额信用交易的可能。这种模式是对传统银行借贷市场的有效改进，特别是针对小额借贷者（Herzenstein 等，2008），是脱虚向实的有效手段。传统信贷业务中，为缓解信息不对称引起的逆向选择和道德风险，金融机构一般会要求贷款客户提供抵押或担保。而拥有良好发展趋势的小微企业与具有良好资质和信用的个人贷款者，可能因为缺少抵押或担保，被传统金融机构排斥在服务群之外。这部分客户的信贷需求便能够从在线信贷平台获得满足（Agarwal 和 Hauswald，2008）。由于在线信贷平台对借款人和投资者的要求都不苛刻，在降低交易成本的情况下，还能吸引更多小额投资者和优质借款客户群，所以在线信贷平台具有快速

① 本文获得国家自然科学基金项目（71473039）、国家社会科学基金重大专项（18VDL012）的资助。

发展并成为普惠金融生力军的现实基础。随着在线信贷行业的快速发展，在线信贷场景中信用风险评级理论也有了长足的发展，目前国内外对在线信贷信用风险评级方法的主要研究如下。

国外学者把信用风险计量的研究分为四个阶段：多元判别分析、多元非线性回归分析、人工智能和组合信用风险度量模型。一是多元判别分析阶段。Beaver（1966）提出了单变量的判定模型，并分析了 30 种财务数据，证明现金流量与负债总额的比率和净收入占总资产的比例具有良好的预测能力。Altman（1968）分析了 1946～1965 年破产制造业公司的数据，选取了 5 个财务数据比值，构建了一种多元判别法，即 Z 评分模型。后来，Z-Score 模型被 Altman 等（1977）做了扩展，形成了 ZETA 模型。二是多元非线性回归分析阶段。由于欺诈客户或违约客户与特征因子间有时候不是线性相关的，所以 Z-Score、ZETA 等模型变得效果不是很好。在 20 世纪 80 年代开始使用了多元非线性回归实现信用风险计量，如逻辑回归。逻辑回归的优势在于其适用性比线性回归更广泛，效果也更好；缺点是最大似然估计不能估计线性可分样本的参数。除此以外，信用风险计量也开始使用聚类模型和 KNN 等机器学习模型。三是人工智能阶段。20 世纪 80 年代末，信用风险度量开始使用人工智能方法。Messier 和 Hansen（1988）首次将该专家系统引入金融危机预警领域。Odom 和 Sharda（1990）在金融危机预警研究中引入了神经网络。Altman（1994）也使用神经网络进行了破产回归研究。Angelini（2000）等构造了多元风险回归和神经网络的混合模型，可以有效预测违约概率。四是组合信用风险度量模型。因为信用风险的影响因子复杂多样，一种预测方法往往很难取得较理想的效果，一系列信用风险计量模型逐渐被提出。1993 年，KMV 公司推出了信用监测模型（Credit Monitor Model），后来 Longstaff 和 Schwarz（1995）等又做了改进，使它变成了信用风险度量经典模型之一。1997 年，J. P. Morgan 和 KMV 公司联合开发了基于两阶段方法的 Credit Metrics。Credit Suisse Bank 提出的 CreditRisk＋模式也非常经典，该模型使用方便，能直接求解债券或贷款资产组合的损失率，而且有较高的计算效率，参数也更少。CreditRisk＋模型的不足之处在于其准确性和可信度不好。

我国对信用风险计量模型的研究还处于起步阶段。王春峰等（1998）使用神经网络取得了比判别分析方法更好的预测效果，并将其应用于信用风险评级。张玲（2000）尝试用 II 类线性判别模型做信用风险评估。沈沛龙和任若恩（2002）研究了信用风险计量模型的理论基础，并比较了有代表性的模型，分析了信用风险计量模型的发展趋势。李旭升等（2008）、李太勇等（2013）、杨胜刚等（2013）、熊志斌（2013）等探讨了贝叶斯网络、决策树等机器学习方法在信用风险计量中的效果和应用空间。李从刚等（2015）借鉴了国际通行的骆驼评级法的指标体系，提出了 P2P 在线信贷市场信用风险评估的指标体系，设计了基于 BP 神经网络的信用风险评估模型。李琦和曹国华（2015）用 CreditRisk＋模型评估信贷资产组合的信用风险水平。王锦虹（2015）在逆向选择理论指导下研究了 P2P 信用风险问题。于晓虹和楼文高（2016）采用随机森林实现信用风险预测，取得了较好的效果。

就在线信贷风险计量方法而言，传统方法有 KMV 模型、CreditRisk＋模型、逻辑回归和生存分析等，因为这些方法一般都有严格的假设，所以限制了这些方法的应用场景。随着互联网的普及、各种数据的积累、大数据技术的兴起，以及机器学习算法（如 XGBoost、深度学习、强化学习等）快速发展，基于多来源多类型数据的机器学习信用风险评分模型逐渐被行业认可，并广泛使用。传统信用风险度量方法主要考虑的是借款人特征与违约结果之

间的强因果关系，而机器学习方法却可以从许多硬信息、软信息和行为信息中获得，然后通过模型找出这些特征与违约结果之间深层的关联关系。

一、背景技术

目前，在线信贷行业使用最多的两种信用风险计量模型有 WOE 评分卡和机器学习评分模型。WOE 的全称是 Weight of Evidence，也叫作证据权重，WOE 是对原始自变量的一种编码形式。其计算公式为：

$$WOE_i = \ln\left(\frac{py_i}{pn_i}\right) = \ln\left(\frac{y_i/y_T}{n_i/n_T}\right) \tag{1}$$

其中，py_i 为该组中坏样本占样本中所有坏样本的比例，pn_i 为该组中好样本占样本中所有好样本的比例，y_i 为该组中坏样本的数量，n_i 为该组中好样本的数量，y_T 为样本中所有坏样本的数量，n_T 为样本中所有好样本的数量。

WOE 评分卡类似逻辑回归模型，该模型的好处是模型稳定，不会因为短时大量异常客户的申请，给评分性能带来波动，所以 WOE 评分卡能保证信用评分的稳定性。而机器学习评分模型一般比 WOE 评分卡更准确，目前常用的模型有逻辑回归、随机森林、XGBoost 模型等。就模型性能而言，WOE 评分模型一旦设计好就不会变化，运行速度快，性能和效果很稳定。机器学习评分模型的有效性更好，AUC（Area Under ROC Curve，受试者工作特征曲线下方的面积大小）、KS（Kolmogorov-Smirnov，洛伦兹曲线）等评价指标一般好于 WOE 评分卡；但是，机器学习评分模型中一般近期样本的权重会大于远期样本（这样可以更好地拟合近期风险特征），但是也会因为短期大量异常客户的申请，评分性能产生较大波动。因此，传统统计学习评分卡与机器学习评分模型结合使用，既能保证模型的准确性，又能保证模型的稳定性，是一种更合理的解决方案。

1. XGBoost 集成学习模型

XGBoost 全名叫极度梯度提升（Extreme Gradient Boosting），是基于树的 Boosting 算法。其算法过程如下：

对于一个给定的数据集 $D=\{(x_i, y_i)\}$（$|D|=n$，$x_i \in R^m$，$y_i \in R$），m 是特征，n 是样本容量。集成树模型的预测输出：

$$\hat{y}_i = \phi(X_i) = \sum_{k=1}^{k} f_k(X_i) \qquad f_k \in F \tag{2}$$

其中 $F=\{f(x)=w_{q(x)}\}$（$q: R^m \to T$，$w \in R^T$）是回归树空间，这里 q 表示每棵树的结构，它将一个示例映射到对应的叶索引。T 是树的叶子数量，每个 f_k 对应每棵树的结构 q 和叶子的权重 w。不同于决策树，每一个回归树包含每个叶子的连续得分，用 w_i 代表第 i 个叶子的得分。设目标函数为：

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{3}$$

其中 $\Omega(f)=\gamma T+\frac{1}{2}\lambda||w||^2$，$l$ 是可微凸损失函数，代表预测结果 $\hat{y}_i$ 和实际结果 y_i 之间的误差，为了避免模型的过拟合问题，模型加入了 Ω 正则项。对式（3）中目标函数的优化不能使用传统的优化方法，而是使用另外的附加方法进行训练，每一次都是在保留原模

型的基础上，添加一个新函数到模型中去。y（t）是第 t 次迭代中第 i 次的预测，我们需要添加 f_t 来最小化下面的目标。

$$L^{(t)}=\sigma\sum_{i=1}^{n}l(y_j,\hat{y}_t^{t-1}+f_t(x_i))+\Omega(f_k) \tag{4}$$

式（4）相对于式（3）增加了 f_t 来改善模型，为了更快地优化目标函数，用二阶泰勒展开来近似原来的目标函数。

$$\widetilde{L}^{(t)}\approx\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\Omega(f_t) \tag{5}$$

其中，$\widetilde{L}^{(t)}\approx\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\Omega(f_t)$。

在不考虑常数项的情况下，简化后的目标函数为：

$$\widetilde{L}^{(t)}\approx\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\Omega(f_t) \tag{6}$$

定义 $I_j=\{i\mid q(x_i)=j\}$ 则：

$$\begin{aligned}\widetilde{L}^{(t)}&=\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\gamma T+\lambda\sum_{j=1}^{T}w_j^3\\&=\sum_{j-1}^{T}[(\sum_{i\in I_j}g_i)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^2]+\lambda T\end{aligned} \tag{7}$$

对于给定 $q(x)$ 可以计算出叶子 j 的最优权重 w_j：

$$w_j=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_J}h_i+\lambda} \tag{8}$$

计算相应的最优目标函数值：

$$\widetilde{L}^{(t)}(q)=-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma T \tag{9}$$

式（9）可以用来计算树的结构 q 的得分，但通常情况下不可能列举出所有可能的树的结构 q，利用贪心算法每一次尝试对已有的叶子加入一个分割，假设 I_L 和 I_R 是左右子树分数分割后的节点，令 $I=I_L\cup I_R$，分割后的损失函数如式（10）：

$$L_{split}=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{(\sum_{i\in I}g_i)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma \tag{10}$$

另外，为了防止过拟合，XGBoost 采用了 Shrinkage 方法来降低过拟合的风险，Shrinkage 技术最早由 Friedman（1999）提出。除了利用 Shrinkage 技术来抑制过拟合外，XGBoost 在构造树模型过程中也借用了随机森林中随机选择一定量的特征子集来确定最优分裂点的做法，以达到抑制过拟合的目的。特征子集越大则每个弱分类器的偏差就越小，但

方差就越大，子集的比例可以通过利用交叉验证来确定。

2. 信用评分卡模型

信用评分模型本质上是一种有监督学习模型，也是模式识别中的二分类问题。信用评分模型通常将客户群体分为非违约客户和违约客户，并根据历史上每个客户类别的若干样本，从历史已知数据中分析两类客户群体的特征，并根据分类规则建立数学模型，再利用模型计算贷款人违约风险作为消费信贷的决策依据。

一般来说，我们通常说的评分卡模型一般都是用于贷前审批，也称为贷前评分卡，用于贷前审批阶段对借款申请人的风险评估。逻辑回归模型是最常见的一种信用评分的模型，该模型理论基础比较完善，有很强的解释性，是目前商业银行信用评分卡开发的主要方法。逻辑回归是以贷款人违约与否作为因变量，一般称为“好样本”或者“坏样本”，用 0 或者 1 来表示。然后利用借款人已有的资料建立模型，对借款人违约的概率（$y=1$）进行预测。模型表达式为：

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k \tag{11}$$

通过式（11）得到的是客户违约概率估计，但在实际应用中通常需要转换为简单直观的评分，也就是要将概率值标准化为分数值。评分的校准是指通过线性转换将最终模型评分与好/坏比（odds）建立一定对应关系的过程。odds 为好样本概率 p 与坏样本概率 $1-p$ 的比值，其表达式为：

$$odds=\frac{p}{1-p} \tag{12}$$

评分卡设定的分值刻度可以通过将分值表示为比率对数的现行表达式来定义，公式如下：

$$score_{总}=A+B\times\ln(odds) \tag{13}$$

设置比率为 x（也就是 $odds$）的特定点分值为 P_0，而 PDO 是指 $odds$ 成为原来的两倍需要提高的分数。则比率为 $2x$ 的点的分值为 P_0+PDO。代入式（13）可得到：

$$\begin{cases}P_0=A+B\ln(odds)\\P_0+PDO=A+B\ln(2x)\end{cases} \tag{14}$$

则：

$$\begin{cases}B=\dfrac{PDO}{\ln2}\\A=P_0-B\ln(x)\end{cases} \tag{15}$$

由于 P_0 和 PDO 的值都是已知常数，将计算出的 A、B 值代入式（13），从而得到不同的 x 下的评分卡分值。

二、普适模型栈框架设计

传统银行评分卡使用的变量较少，一般是与贷款强相关的硬信息，可以分为三类：贷款人基本信息、贷款人社会关系信息和个人征信信息。一是贷款人基本信息。基本的客户信息

包括客户的基本情况、工作情况、经济收入情况、社会保障情况以及其他信用情况，主要来自客户贷款（信用卡）申请表等申请材料。基本情况包括性别、年龄、婚姻状况、学历、供养人数、本地居住时间、住房情况等。工作情况包括工作单位的性质、所属行业以及当前单位工作时间。经济收入情况包括个人与家庭工作收入、其他经济来源、家庭资产与负债情况。社会保障情况包括是否参加养老保险、是否参加失业保险、是否缴纳公积金和是否参加医疗保险。其他信用情况包括移动电话缴费情况、水电煤气缴费记录。其他特殊或不良记录包括破产记录、偷逃税记录、公安机关处罚记录等。二是货款人社会关系信息。包括银行业务中的存款和贷款类信息。存款类信息指定期存款、活期存款、证券资金账户、理财账户等信息。贷款类信息指个人贷款、信用卡等。三是个人征信信息。理论上个人征信信息至少应该包括银行信用记录、社会保障数据、如是否参加养老保险等。各类缴费情况如水电煤气缴费情况等。其他不良记录，如破产记录、偷逃税记录等。目前，个人征信报告由人民银行征信局提供，征信报告的数据包括银行信贷、税务、缴费等，而且还在逐步丰富。其中银行信用评级关注的信息包括：银行信贷、信用卡、准贷记卡信息、为他人担保信息和人行征信报告查询次数等。

与传统银行相比，在线信贷企业一般很难获得这么完整、规范的数据。一方面，在线信贷场景面向的客户群资质相对较差，银行信贷使用的很多硬信息这些客户都不具备，所以在线借贷平台使用了很多替代数据，如客户填写的申请表、网上授权爬取数据、第三方数据服务商数据等。另一方面，目前流行的消费金融的消费场景很多，不同的消费场景所拥有的基础数据也不一样，如汽车贷款有汽车的一些相关数据，而医疗美容贷款则包括医疗项目的数据。所以，为了弥补申请评分卡中部分硬信息的缺失和融合不同的消费场景，为各个来源的数据分别训练子评分模型，然后再根据不同的场景融合为综合评分卡，是一种更好的解决方案。

1. 基于多源数据的普适模型栈框架设计思想

为了灵活有效地实现在线信贷中的信用风险评级，本文提出一种普遍适用于多源数据的模型栈框架，基本思想是：首先，根据不同场景、不同客户群的不同数据，将数据分组后分别训练子评分模型，把训练好的子评分模型放入模型栈；其次，根据预测客户具有的数据基础，选择模型栈中的多个子评分模型，把选出的多个子评分模型的评分结果（0～1 之间的小数，表示该客户的违约概率）转换为等分区间，再利用评分卡建模的方法，形成最终的评分模型。

本文中选用某互金平台拥有的数据作为构建普适模型栈框架基础，这些数据可分为 9 种，包括：多头借贷、高风险特征、与身份证手机号关联的信息数量、黑名单关联性、支付宝数据、信用卡数据、高风险行为数据、QQ 好友与群中风险数据、短信中高风险信息。根据该框架构建的基本思想，先将这 9 种数据和违约 Y 值，分别构建 XGBoost 分类预测模型，每个模型分别优化，使其达到最佳分类效果；接着，把训练得到的 9 个子模型放入模型栈。然后，把各个子模型预测的结果整合为一个 9 列的矩阵（将其中各子模型的预测结果转换为具体分数），再通过评分卡建模方法得到最终的信用评分。

在该案例中，模型栈框架根据不同的数据来源，构建了 9 个 XGBoost 子评分模型，而在其他场景下，子模型的数量和选用的算法都可以自由选择。在该案例中也做了很多算法组合的尝试，最终发现还是以 XGBoost 这一集成学习算法作为各个子评分模型能更好地拟合违约风险。各个子模型虽能较好地预测用户的信用风险，但是整合模型的准确率更高，并且

预测效果也更稳定。当面对不同的借贷场景或不同的客户群时，模型可用的数据也不同。这时，先将数据根据来源或客户群分组，然后自由选择入模数据，自由选择模型算法，自由组合入栈子模型，就连最终将子模型预测结果进行融合的逻辑回归也可以根据情况替换为其他算法。因此，该普适模型栈框架既能更灵活地适应各种信贷场景，又能更有效、更稳定地获得预测结果。

2. 模型评价指标

一般来说，机器学习方法分类模型有些重要的评价指标，如混淆矩阵、ROC、AUC、KS值等。其中，混淆矩阵的表达式见表1。

表1　　混淆矩阵

		预测值		合　计
实际值		响应	不响应	
	响应	真正例	假负例	真正例＋假负例
	不响应	假正例	真负例	假正例＋真负例
合计		真正例＋假正例	假负例＋真负例	真正例＋假负例＋假正例＋真负例

注：真正例（True Positive，TP），假正例（False Positive，FP），假负例（False Negative，FN），真负例（True Negative，TN）。

根据表1混淆矩阵的结果可以计算出相应的机器学习模型的评价指标，常用的评价指标如表2所示。

表2　　机器学习常用评价指标

指标名称	计算公式
Accuracy Rate（准确率）	$(TP+TN)/(TP+TN+FN+FP)$
True Positive Rate（TPR）也称 Recall（召回率 R）	$TP/(TP+FN)$
False Positive Rate（FPR）	$FP/(FP+TN)$
Precision（精确率 P）	$TP/(TP+FP)$
F1值	$2PR/(P+R)$

以FPR为横坐标，TPR为纵坐标，把不同的点连成曲线，可以得到ROC曲线。ROC、KS曲线原理见图1、图2。

ROC曲线又称为受试者工作特征曲线（Receiver Operating Characteristic Curve），又称为感受性曲线（Sensitivity Curve）。从图1可以看出ROC曲线是以假阳性概率（False Positive Rate）为横轴，真阳性（True Positive Rate）为纵轴所组成的坐标图。ROC曲线下方与坐标轴围成的面积叫作AUC（Area Under ROC Curve），AUC值越大说明分类器性能越好。

KS曲线也叫做洛伦兹曲线（Kolmogorov-Smirnov Curve），KS值是对模型风险区分能力进行评估的最重要指标，该指标衡量的是好坏样本累计分部之间的差值，好坏样本累计差异越大，KS指标越大，那么模型的风险区分能力越强。另外，从图2可知，KS曲线和ROC曲线的画法类似，在不同的概率阈值下，计算出不同的FPR和TPR值，再分别画出TPR和FPR曲线，则两条曲线之间的距离就是KS值。不同的KS值对应的模型评价标准见表3。

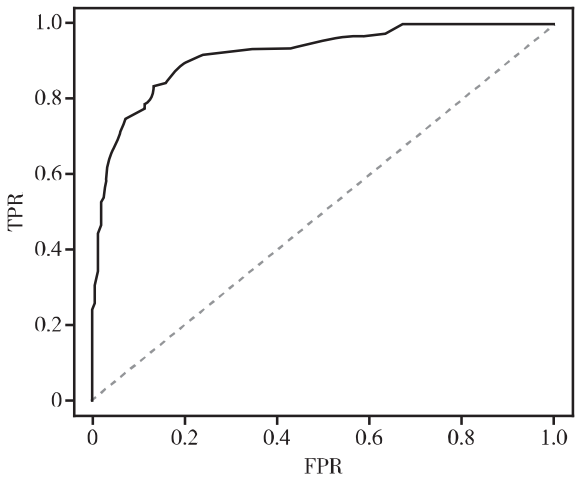

图 1 ROC 曲线

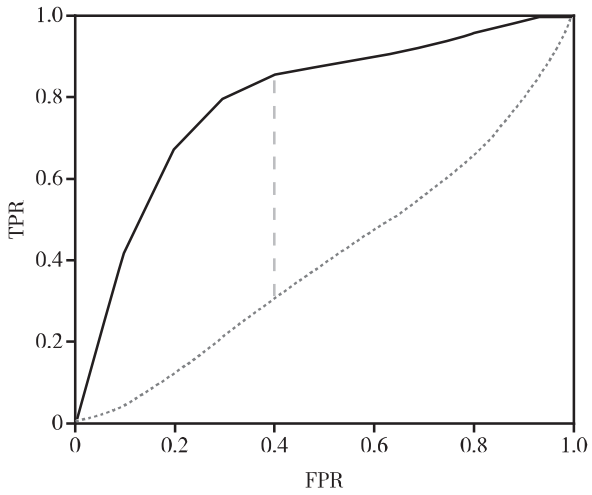

图 2 KS 曲线

表 3 KS 值评价标准

KS 值	评价结果
小于 0.20	模型无鉴别能力
0.20～0.40	模型可以接受
0.41～0.50	模型有较好的区别能力
0.51～0.60	模型有很好的区别能力
0.61～0.75	模型有非常好的区别能力
大于 0.75	模型异常，可能存在问题

三、实证分析

实验数据来自于国内某 P2P 公司 2017 年 4～6 月实际交易 40678 笔小额在线贷款数据，借款人借款周期为 1 个月，借款金额为 1000～3000 元。滚动率根据历史数据计算 M_0（未逾期）递延至 M_1（逾期 1～29 天），M_1 递延至 M_2（逾期 30～59 天），依此类推。根据滚动率模型，发现逾期客户进入 M_2 后还款比例已经很低，故“坏客户”采用逾期进入 M_2 的客户群。本文在实验中将 40678 个贷款样本中 32537 个正常还款样本标记为 0（无逾期），8141 个逾期样本标记为 1（逾期超过 30 天），总体坏样本比为 20.01%。样本信息包括多头借贷、高风险特征、与身份证手机号关联的信息数量、黑名单关联性、支付宝数据、信用卡数据、高风险行为数据、QQ 好友与群中风险数据、短信中高风险信息，共 9 种 547 个变量。

表 4 数据总体情况统计

样本量	好样本 0	坏样本 1	坏样本比例	样本时间窗口
40678	32537	8141	20.01%	2017.04～2017.08

资料来源：某金融信息科技有限公司。

为了比较基于多源数据的模型栈普适框架的有效性，实验中选取将所有特征组合为一个大宽表，再用目前在线信贷行业普遍使用的 XGBoost 评分模型作为对比算法，然后与本文预先设计的 9 个 XGBoost 子评分模型相结合，再将子模型的输出结果作为输入变量，形成最终的评分卡模型，效果如下所述。

1. XGBoost 信用评分模型实证分析

相对于 WOE 评分卡模型，XGBoost 信用评分模型过程更为简洁，无须进行变量筛选，直接将所有 547 个变量输入模型，预测测试集样本信用违约概率。试验采用 XGBoost 算法，通过参数寻找最优方法，学习率设为 0.01，树深度为 5。另外，本次实验考虑了对样本缺失情况进行补缺的分析，并采用中位数补缺的方法进行缺失值填补。在包含缺失值和不包含缺失值的情况下，XGBoost 算法模型的预测结果对比见表 5。

表 5　　包含缺失值的模型预测结果

评价指标	准确率	召回率	精确率	F1
包含缺失值	0.7716	0.8028	0.9006	0.8489
不包含缺失值	0.7677	0.7982	0.8998	0.8382

从表 5 的结果可以看出，包含缺失值的 XGBoost 模型效果要好于利用中位数补缺后的模型，这也说明在对模型进行缺失值补足后可能会带来更多的数据噪音，从而影响实验结果，因此本研究采用不对缺失值进行处理的方法。在不考虑缺失值的情况下，XGBoost 模型 KS 值的计算结果见图 3。

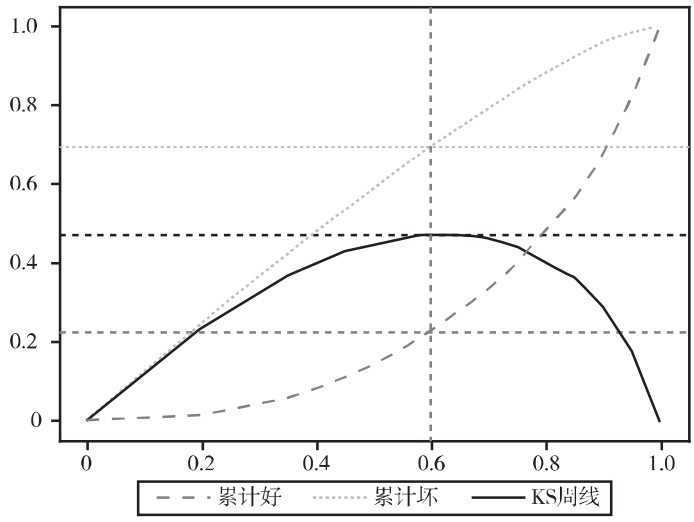

图 3　KS 值计算结果

从图 3 中可以看出，KS 值可以通过累计坏样本与累计好样本之差来实现，在本次实验中，模型的 KS 值为 0.4715。

为了验证 XGBoost 的分类效果，本次实验利用了相同的数据对常用的信用评分建模方法进行了测试，几种模型的计算结果见表 6。

表 6　　不同模型下的试验结果对比

机器学习方法	KS	AUC
逻辑回归	0.413	0.776
随机森林	0.445	0.791
Adaboost	0.438	0.788
GBDT	0.463	0.808
XGBoost	0.472	0.821

从表 6 可以看出，在业界常用的 5 种评分模型中，GBDT 与 XGBoost 的分类效果最好，其中 GBDT 模型的 KS 值为 0.463，AUC 值为 0.808，而 XGBoost 模型获得了 0.472 的 KS 值和 0.821 的 AUC 值，这说明 XGBoost 的预测能力更好。故本论文选择 XGBoost 算法为模型基础方法。

由于企业在实际生产经营环节中，模型的运用需要考虑公司实际经营情况。一般来说，贷款机构特别关注模型将坏客户预测为好客户的情况，而为了减少这样的情况，机构可以通过提高贷款评分，也就是只贷款给那些信用得分较高的客户的方法来实现。但这样操作的结果必然会出现大量申请人被拒绝的情况，也就是对应的样本通过数量的下降，这无疑会降低公司的营业额。根据这种情况，本次试验列出了在不同的预设概率下，XGBoost 模型预测结果（见表 7）。

表 7　　不同预设概率下的 XGBoost 模型预测结果

预设概率	累计好样本	累计坏样本	KS 值	通过率	误放率
(0.50，0.55]	0.1444	0.5895	0.4451	0.7124	0.0994
(0.55，0.60]	0.1818	0.6427	0.4610	0.6595	0.0863
(0.60，0.65]	0.2234	0.6949	0.4715	0.6030	0.0750
(0.65，0.70]	0.2760	0.7441	0.4681	0.5383	0.0644
(0.70，0.75]	0.3317	0.7927	0.4610	0.4723	0.0507
(0.75，0.80]	0.3984	0.8385	0.4401	0.4063	0.0420
(0.80，0.85]	0.4816	0.8800	0.3984	0.3328	0.0295
(0.85，0.90]	0.5594	0.9231	0.3637	0.2514	0.0196
(0.90，0.95]	0.6701	0.9579	0.2877	0.1593	0.0116
(0.95，1.00]	0.8097	0.9854	0.1757	0.0600	0.0041

从表 7 可以看出，模型在预设概率为 0.60～0.65 的条件下（也就是只有当某个借款人被预测为好人的概率大于 0.60 时才通过筛选），实现了最高的 KS 值为 0.4715，且有 60.30%的申请客户通过了贷款申请，但模型也将 7.5%的坏样本误判为好样本。如果从实际业务的角度出发，贷款机构可以通过综合比较三个指标的情况，选择一种最适合公司经营情况的预设概率阈值。另外，为了使得机器学习模型的输出结果更具有可操作性，也可将机器学习模型输出概率值转换为常见的信用评分值。在假设基础分为 500，pdo 设为 10，p_0 设为 50 的情况下，表 8 的结果是转换为传统信用评分的情况。

表 8　　机器学习模型转换为信用评分

输出概率	好样本	坏样本	总计	好样本占比（%）	坏样本占比（%）	总体占比（%）	坏账率（%）
(0.0，0.5]	1286	1071	2357	19.78	65.54	28.97	45.44
(0.5，0.6]	755	209	964	11.61	12.79	11.85	21.68
(0.6，0.7]	843	160	1003	12.97	9.79	12.33	15.95
(0.7，0.8]	1030	110	1140	15.84	6.73	14.01	9.65
(0.8，0.9]	1328	69	1397	20.42	4.22	17.17	4.94
(0.9，1.0]	1260	15	1275	19.38	0.92	15.67	1.18

从表 8 可以看出，机器学习模型也可以通过信用评分卡转换法转换为常见的信用评分，从而使得模型的输出结果更加具有可解释性。

2. 基于多源数据普适模型栈框架实证分析

为了验证模型的有效性，我们对 9 个子信用风险评级模型分别训练，并对预测有效性做了分别统计，又把 9 个子评分模型通过普适模型栈框架，将每个模型的预测概率利用评分卡建模的方法转换为具体评分，再把 9 个评分作为最终的模型输入变量，然后再利用评分卡建模方法做成一个新的完整的评分卡，9 个子模型的结果如表 9 所示。

表 9　　某在线信贷多源数据信用风险评级有效性

数据种类	准确率（%）	召回率（%）	AUC（%）	KS（%）
多头负债	37.19	65.37	66.33	21.38
高风险操作行为数据	45.99	59.21	64.31	20.56
高风险特征	34.52	54.01	59.76	17.52
信用卡数据	26.75	55.38	57.62	15.25
支付宝数据	32.07	57.45	55.08	13.17
与身份证手机号关联信息数	28.03	57.70	54.84	11.67
QQ 好友与群中风险数据	26.19	58.68	53.77	9.34
黑名单关联性	25.33	50.89	52.99	5.98
短信中高风险信息	24.26	51.54	51.51	3.01

从表 9 可以看出，9 个子 XGBoost 机器学习评分模型中，多头负债、高风险操作行为数据用于违约预测效果最好，这也符合小额现金贷客户群共债现象比较严重且行业欺诈黑产盛行的普遍现象。

最后，将 9 个子模型的预测结果作为输入变量融合成一个集成评分卡模型，在不同预设概率下的普适模型栈框架综合模型实验结果见表 10。

从表 10 可以看出，模型在预设概率为 0.60～0.65 的条件下（也就是只有当某个客户预测为好人的概率大于 0.60 才通过筛选），实现了最高的 KS 值为 0.4927，且有 62.54%的申请客户通过了贷款申请，但模型也将 7.64%的坏样本误判为好样本。如果从实际业务的角度出发，贷款机构可以通过综合比较 KS 值、通过率、误放率三个指标的情况，选择一种最适合公司经营情况的预设概率阈值。

表 10　　不同预设概率下的普适模型栈框架实验结果

预设概率	累计好样本	累计坏样本	KS 值	通过率	误放率
(0.50，0.55]	0.1423	0.6227	0.4804	0.7421	0.0905
(0.55，0.60]	0.1811	0.6630	0.4819	0.6675	0.0833
(0.60，0.65]	0.2211	0.7138	0.4927	0.6254	0.0764
(0.65，0.70]	0.2726	0.7473	0.4747	0.5445	0.0613
(0.70，0.75]	0.3301	0.7927	0.4626	0.4950	0.0499
(0.75，0.80]	0.4054	0.8385	0.4331	0.4567	0.0403
(0.80，0.85]	0.4912	0.8801	0.3889	0.3886	0.0279
(0.85，0.90]	0.5674	0.9231	0.3557	0.2876	0.0188
(0.90，0.95]	0.6802	0.9579	0.2777	0.1895	0.0106
(0.95，1.00]	0.8132	0.9854	0.1722	0.0506	0.0032

从表 7、表 9 和表 10 的对比可以看出，将多种来源的不同数据分别做 XGBoost 机器学习评分模型，然后通过将子模型做成一个新的基于评分卡模型的普适模型栈，即比各个子评分模型更有效，也比把所有特征做成一个单独的 XGBoost 有效。同时各子模型可以预先训练完成，然后策略部门根据不同产品的特点自由组合为新的综合评分模型，因此灵活性更好；而基于多源数据的集成机器学习评分模型，因为有更多其他维度的数据，所以违约预测的有效性表现得最稳定。

同样，为了使得普适模型栈框架综合模型的输出结果更具有可操作性，也可将集成机器学习框架模型转换为传统评分卡，表 10 的转换结果如表 11 所示。

表 11　　普适模型栈框架综合模型转换为信用评分

得分区间	好样本	坏样本	总计	好样本占比（%）	坏样本占比（%）	总体占比（%）	坏账率（%）
low≤Score≤570	1311	1064	2375	16.11	13.08	29.19	44.80
571≤Score≤578	779	207	986	9.57	2.54	12.12	20.99
579≤Score≤585	841	157	998	10.34	1.93	12.27	15.73
586≤Score≤595	1072	106	1178	13.18	1.30	14.48	9.00
596≤Score≤615	1277	56	1333	15.70	0.69	16.38	4.20
616≤Score≤high	1254	12	1266	15.41	0.15	15.56	0.95

从表 11 可以看出，基于多源数据的集成机器学习模型可以通过信用评分卡转换法转换为常见的信用评分，这说明普适模型栈框架方法除了预测能力更好之外，还可以与传统方法相互融合，从而构建更加有效的风控体系。

四、结　　论

本文根据我国在线信贷行业的特点，建立了一套基于多源数据信用评级的普适框架，并与业界普遍使用的基于 XGBoost 算法的大数据风控模型做了比较，得出以下结论。

第一，基于多源数据信用评级的普适模型栈框架可以根据不同消费场景和不同客户拥有的资质特征，灵活选择入模数据，并分成数据组，分别构建多个子评分模型。而 XGBoost 评分模型需要所有申请人具有相同的特征变量，这限制了该模型的普适性。第二，基于多源数据信用评级的普适模型栈框架的有效性优于 XGBoost 评分模型。实验中利用 XGBoost 算法，并使用了全部 547 个输入变量构建评分模型，KS 值达到了 0.47；而使用普适模型栈框架将 547 个输入变量按数据类型分为 9 组，分别与 Y 构建 9 个子评分模型，然后再用评分卡建模方法形成最终评分模型，其模型的 KS 高达 0.49。实验表明，基于多源数据信用评级的普适模型栈框架比 XGBoost 评分模型更有效。第三，变量的有效性是决定模型预测能力的最根本因素。从子模型的对比来看，多头借贷数据和风险行为数据的预测远高于黑名单、短信风险等指标维度。但这并非说明黑名单、短信风险数据无效，只能说明该数据公司提供的这些维度的数据质量较差，也变相地说明了特征工程的重要性。在个人隐私保护越来越严格和规范的今天，如何获取合规、稳定和高质量的客户的多维度的数据以及对所获得的原始变量进行衍生和深度加工能力将直接决定一家公司的金融科技能力，否则，无论采用多么先进的算法也无法带来模型预测能力的明显提升。可以预见，在今后很长一段时间，对数据获取和对数据的挖掘能力将成为未来金融科技的主要战场。第四，普适模型栈框架在实际应用

中应注意数据成本问题。在线信贷业务往往采用线上实时决策模式，客户申请借款后通过一个集成模型框架输出一个评分，但对于大型金融机构来说，评分模型往往有数十甚至数百个，如果每个借款申请人都需要查询所有子模型变量，这无疑会大大增加数据查询成本。因此，如何更经济地使用子模型框架模式，还需根据各公司具体业务情况选择经济高效的子模型组合方式，从而设计出最为合理的信贷风控策略。

从本文的研究可以看出，在线信贷的数据具有高维、复杂、非线性等特征，同时由于贷款人的消费场景不同，客户资质不同，信用评分模型能使用的特征变量也不同。因此，根据贷款人拥有的数据基础，灵活选择信用评分模型入参，是在线信贷场景的特殊需求。本文提出的基于多源数据的普适模型栈信用评分框架能更好地解决申请人特征值缺失问题，更灵活地应用于在线信贷评分场景，同时该普适框架也比单一 XGBoost 评分模型更有效，但在实际使用中需要注意数据成本的问题，如何设计合理的信贷风控策略是一个值得进一步研究的内容。

参 考 文 献

[1] Agarwal S., Hauswald R. B. H., 2008, *The Choice between Arm's-Length and Relationship Debt: Evidence from Eloans* [R], Federal Reserve Bank of Chicago Working Paper Series 2008—10.

[2] Altman E. L., 1968, *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy* [J], Journal of Finance, 23 (4), 589～609.

[3] Altaman E. I., Marco G., Varetto F., 1994, *Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks* [J], Journal of Banking and Finance, 18 (3), 505～529.

[4] Altman E. I., Haldeman R. G., Narayanan P., 1977, *ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations* [J], Journal of Banking and Finance, 1 (1), 29～54.

[5] Angelini P., 2000, *Are Banks Risk Averse? Intraday Timing of Operations in the Interbank Market* [J], Journal of Money Credit & Banking, 32 (1), 54～73.

[6] Bachmann A., Becker A., Buerckner D., Hilker M., Kock F., Lehmann M., Tiburtius P., Funk B., 2011, *Online Peer-to-Peer Lending - A Literature Review* [J], Journal of Internet Banking & Commerce, 16 (2), 1～18.

[7] Beaver W. H., 1966, *Financial Ratios As Predictors of Failure* [J], Journal of Accounting Research, 4, 71～111.

[8] Herzenstein M., Andrews R. L., Dholakia U. M., Lyandres E., 2008, *The Democratization of Personal Consumer Loans? Determinants of Success in Online Peer-to-Peer Lending Communities* [R], Boston University, School of Management, Research Paper, 2008.

[9] Lin M., 2009, *Peer-to-Peer Lending: An Empirical Study* [R], AMCIS 2009 Doctoral Consortium.

[10] Longstaff F. A., Schwartz E. S., 1995, *A Simple Approach to Valuing Risky Fixed and Floating Rate Debt* [J], Journal of Finance, 50 (3), 789～819.

[11] Messier W. F., Hansen J., 1988, *Inducing Rules for Expert System Development: An Example Using Default and Bankruptcy Data* [J], Management Science, 34 (12), 1403～1415.

[12] Odom M. D., Sharda R., 1990, *A Neural Network Models for Bankruptcy Prediction* [R], 1990 IJCNN International Joint Conference on Neural Networks.

[13] Verstein A., 2011, *The Misregulation of Person-to-Person Lending* [J], UC Davis Law Review, 45 (2), 445～529.

[14] 李从刚、董中文、曹筱钰：《基于 BP 神经网络的 P2P 网贷市场信用风险评估》[J]，《管理现代化》2015 年第 4 期。

[15] 李琦、曹国华：《基于 Credit Risk+模型的互联网金融信用风险估计》[J]，《统计与决策》2015 年第 19 期。

[16] 李太勇、王会军、吴江、张智林、唐常杰：《基于稀疏贝叶斯学习的个人信用评估》[J]，《计算机应用》2013 年第 11 期。

[17] 李旭升、郭春香、郭耀煌：《扩展的树增强朴素贝叶斯网络信用评估模型》[J]，《系统工程理论与实践》2008 年第 6 期。

[18] 沈沛龙、任若恩：《现代信用风险管理模型和方法的比较研究》[J]，《经济科学》2002 年第 3 期。

[19] 王春峰、万海晖、张维：《商业银行信用风险评估及其实证研究》[J]，《管理科学学报》1998 年第 1 期。

[20] 王锦虹：《基于逆向选择的互联网金融 P2P 模式风险防范研究》[J]，《财经问题研究》2015 年第 5 期。

[21] 熊志斌：《基于非线性主成分分析的信用评估模型研究》[J]，《数量经济技术经济研究》2013 年第 10 期。

[22] 杨胜刚、朱琦、成程：《个人信用评估组合模型的构建：基于决策树—神经网络的研究》[J]，《金融论坛》2013 年第 2 期。

[23] 于晓虹、楼文高：《基于随机森林的 P2P 网贷信用风险评价、预警与实证研究》[J]，《金融理论与实践》2016 年第 2 期。

[24] 张玲：《财务危机预警分析判别模型及其应用》[J]，《数量经济技术经济研究》2000 年第 3 期。

A General Stack Framework of Credit Risk Rating Models Based on Multi Source Data

Huang Zhigang[1] Liu Zhihui[1,2] Zhu Jianlin[3]

(1. School of Economics and Management, Fuzhou University;
2. Department of Finance, Fujian Business University;
3. School of Finance, Renmin University of China)

Research Objectives: General stack scoring framework based on the data from a variety of channels. **Research Methods**: Based on the actual situation of the online lending industry in China, this paper proposes a general stack scoring framework based on the data from a variety of channels. This framework can freely select the data of the score model according to the different data base of each lender, and then the sub-scoring model is transformed into a common credit scoring card model. **Research Findings**: The general stack scoring framework is not only flexible and general, but also has better effectiveness than a single XGBoost credit scoring model. **Research Innovations**: The machine learning model and the traditional scorecard model are perfectly integrated. **Research Value**: The problem of poor interpretability of machine learning in credit risk management is solved.

Key Words: Risk Management; Credit Score; Score Card Model; Machine Learning; Stack Framework Model

JEL Classification: C53; C81; G21

(责任编辑：陈星星)