基于 LightGBM 模型的信贷违约概率预测研究

黄乐乐，陈林
暨南大学管理学院，广东 广州 510632
DOI:10.61369/ASDS.2025040019

摘　　要： 信用评级是信贷业务的核心，为此各种统计建模方法应运而生．随着大数据时代的到来，收集数据的范围显著扩大，可用于信用评级的特征数量也随之增加．这些带来了特征冗余的风险，因此特征选择是建模过程中至关重要的一步．本文提出了一种两阶段信用评分建模方法．首先对全部特征进行基于 Mean Variance 的独立性检验，进行初步筛选，然后采用基于 LightGBM 的分类模型得到最终的违约概率预测模型．此外，我们构建了一个虚拟特征，用于检测模型中是否仍然存在冗余特征．最后，将该方法应用于实际的在线信贷业务数据，以评估该方法的有效性。

关 键 词： 信用评级；特征冗余；独立性检验；LightGBM

Research on Credit Default Probability Prediction Based on LightGBM Model

Huang Lele, Chen Lin
School of Management, Jinan University, Guangzhou, Guangdong 510632

Abstract: Credit scoring is central to the credit loan business, and various statistical modeling methods have been developed for this purpose. With the advent of the big data era, the scope of collected data has expanded significantly, leading to an increase in the number of features available for credit scoring. However, this also introduces the risk of redundant features, making feature selection a crucial step in the modeling process. This paper presents a two-stage credit scoring modeling method. In the first stage, an independence test based on mean variance is performed on the entire set of features for preliminary screening. In the second stage, classification models based on LightGBM are employed, with further feature selection carried out to refine the final model. In addition, we constructed a virtual feature to help detect whether there are still redundant features in the model. Finally, the proposed method was also applied to real-world online lending data, which yielded promising results to evaluate the effectiveness of the model.

Keywords: credit scoring; redundant features; independence test; LightGBM

引言

随着互联网金融的快速发展，个人消费信贷也变得更加快捷便利。基于信用的网贷业务蓬勃发展，主要的产品形态包括现金贷、现金分期、消费分期、虚拟信用卡等。尽管产品形态各异，但其背后风险管理的核心逻辑依然是对借款人风险的精准评估。目前，业界最常用的信用评估工具是各种信用模型。模型的效果取决于模型的输入，即输入的各种特征变量，以及具体的模型算法，例如逻辑回归模型、决策树模型、随机森林模型、GBDT、XGBOOST、LightGBM[1] 等。

首先，我们关注信用风险模型可用的特征。传统金融机构通常能够获取央行信用报告及相关信用报告产品。从这些报告中提取的各种特征对于逾期付款预测非常有效。由于这些特征是可靠的预测指标，对模型的要求相对较低，使得常用的评分模型能够取得良好的效果。在线贷款平台通过互联网广告吸引客户，鼓励用户下载安装其应用，并完成注册流程。在此过程中，在用户同意的情况下，平台可以访问用户设备中的各种数据，包括手机型号、品牌、内存大小、电池状态、已安装的应用列表以及短信内容等。这些数据与传统金融机构收集的数据存在显著差异。通常，单一的设备信息特征无法有效预测逾期付款。传统模型基于这些特征评估信用风险时，结果往往不尽如人意。因此，这促使我们寻找更合适的模型，将现有特征应用于新的在线业务场景。

接下来，我们将回顾信用评分中特征筛选和主要信用模型的发展历程。首先， 我们关注信用模型。50年代，Bill Fair 和 Earl Isaac 提出基于 Logistic 回归模型的 FICO 信用评分体系，成为业界应用最成熟的信用风险模型。Crook 和 Banasik 等（2004）[2] 对拒绝推断

作者简介：
黄乐乐（1986-），男，河南济源人，博士后；
陈林（1981-），男，广东河源人，教授。

Copyright © This Work is Licensed under A Commons Attibution-Non Commercial 4.0 International License.

方法是否能提高信用风险评估模型的预测效果进行了研究，在重加权方法、插值方法和二元 probit 方法情形下比较了拒绝推断的结果。Chen 和 Åstebro（2012）[3] 考虑了非随机缺失情形下信用风险评估中的拒绝推断问题，并引入了贝叶斯方法。Feng 等（2018）[4] 考虑了在信用风险评估中第一类错误和第二类错误的相对损失，基于柔性概率给出了一种动态集成方法。Dirick 等（2019）[5] 在信用评估中引入了宏观经济相关变量，使用 mixture cure 模型预测借贷者的违约概率。Fang 和 Chen（2019）[6] 指出在建模过程中将目标函数设定为 KS 取到最大值，在综合考虑模型表现、计算复杂度和模型可解释性方面具有明显的优势。Shen 等（2020）[7] 提出一种基于迁移学习的三阶段拒绝推断框架用于信用评分，有效地解决了负迁移问题。Nikita 等 (2022) [8] 为 GBDT 引入了逐步特征增强机制，基于树的嵌入技术简化了特征增强的过程。Mushava 和 Michael（2022）[9] 基于 XGBoost 提出使用广义极值（GEV）分布的分位数函数作为链接函数来增强对罕见情况的检测。He 等（2023）[10] 为了解决利用多源信息进行信用评分的这些挑战，提出了一种基于逻辑回归的去中心化多方方法，使用垂直联邦学习范式制定逻辑回归模型。Chatterjee 等（2023）[11] 从宏观视角对信用评级在信贷市场中的作用进行了研究。此外，关于信用模型的研究成果还有很多，在此无法一一列举。

现在我们来回顾一下特征筛选方面的研究进展。在高维数据集中，预测因子（自变量）的数量巨大，多重共线性、过拟合和计算效率低下会显著影响预测模型的稳定性和可解释性。为了应对这些挑战，特征筛选方法被用作预处理步骤，以识别和消除不相关或冗余变量。通过降低输入空间的维数，这些方法不仅增强了模型性能，还提高了计算效率和可解释性。特征筛选在信用风险建模、基因组学和图像处理等场景中尤为重要，因为这些场景中潜在预测因子的数量通常远远超过观测值的数量。常用的筛选技术包括 LASSO[12]，SCAD[13] 等，以及很多基于独立性检验的统计量用于特征筛选，如 Cui 等 [14] 提出的 Mean Variance (MV) 统计量。陈等 [15] 使用均值方差指数 (MV) 和卡方独立性检验进行变量选择并进行违约概率预测，取得了良好的结果。王等 [16] 使用 MV 统计量对员工离职的影响因素进行了分析，取得了不错效果。

本文探讨了一种两阶段信用评分建模方法，并将其应用于个人消费信贷的真实数据集。本文的其余部分结构如下：第一部分介绍了所使用的变量选择方法、分类器和评估指标。第二部分将分析个人消费信贷相关数据，基于真实数据展示如何使用本文提出的方法建立信用风险评估模型。最后，第三部分将介绍主要结论和讨论。

一、主要模型及方法

（一）MV 方法

MV(mean of variance) 指标是 Cui 等 [14] 提出的一种非参数检验统计量，可用于度量连续型变量和分类变量之间的独立性。设 X 为连续随机变量，Y 为 R 分类的分类变量，记作 $(y_1, y_2, \cdots, y_k)$，则 $MV(X|Y)$ 指标定义为

$$MV(X|Y) = E_X\left(Var_Y\left(F(X|Y)\right)\right),$$

其中 $F(X|Y)=P(X \le x|Y)$ 表示给定 Y 时 X 的条件分布函数。$MV(X|Y)=0$ 等价于 X 与 Y 相互独立。样本形式的 MV 检验统计量为：

$$\hat{MV}(X|Y) = \frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}\hat{p}_r\left(\hat{F}_r(X_i)-\hat{F}(X_i)\right)^2,$$

其中 $\hat{F}_r(x)=\frac{1}{n}\sum_{i=1}^{n}I(X_i \le x, Y_i = y_r)/\sum_{i=1}^{n}I(Y_i = y_r)$，是给定 $Y=y_r$ 条件下 X 的经验条件分布函数，$\hat{F}(x)=\frac{1}{n}\sum_{i=1}^{n}I(X_i \le x)$ 是 X 的无条件经验分布函数，$p_r=\frac{1}{n}$ 为第 r 类的样本比例，$I(\cdot)$ 为示性函数。

（二）LightGBM 模型

LightGBM（Light Gradient Boosting Machine）是一种基于决策树算法的梯度提升模型，通过集成多个弱分类器来提升预测效果，特别适用于大规模和高维度的数据集。与传统的梯度提升方法不同，LightGBM 采用基于叶子的树生长策略， 并通过限制树的深度来控制复杂度。此外，LightGBM 还引入了多种优化技术，如基于直方图的决策树学习、基于梯度的一侧采样（GOSS）等，使其在性能和速度上具有显著优势。凭借其高效性和准确性，LightGBM 已被广泛应用于学术研究和工业实践中。

假设训练数据为：

$$D = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{n},$$

模型预测值为：

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k\left(x_i\right),$$

其中 $f_k \in F$ 是第 k 棵树，F 是 CART 树的函数空间。

每一轮迭代的目标是最小化损失函数的泰勒展开（通常使用二阶展开）：

$$L^{(t)} \approx \sum_{i=i}^{n}\left[g_i f_t\left(x_i\right)+\frac{1}{2}h_i f_t\left(x_i\right)^2\right]+\Omega\left(f_t\right)$$

其中：

$g_i = \dfrac{\partial L\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i}$ 为一阶导（梯度）

$h_i = \dfrac{\partial^2 L\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^2}$ 为二阶导（Hessian）

$\Omega(f)$ 是树结构的正则项，用于防止过拟合

二、实证分析及结果

我们使用某在线借贷平台的实际业务数据进行建模。每一笔借款订单为一个样本，我们观察到期日后30天内用户是否还款，

 Copyright © This Work is Licensed under A Commons Attibution-Non Commercial 4.0 International License.

还款则为好客户，不还则为坏客户。

该批数据共有18444个样本，我们根据时间选择申请时间距今最近的2770个样本作为 OOT 数据集。使用的特征共1433个，主要分为三大类：用户输入的个人相关信息（如性别、年龄）、从用户手机采集的信息以及借贷 app 的埋点信息。其中，以“app_”开头的特征为基于手机安装的 app 列表衍生得到的特征，如“app_up_FINANCE_3d”表示3天内更新过的金融类 app 的数量。以“tag_”开头的特征为基于短信记录衍生得到的特征，如“tag_1_num_60d”表示60天内收取的类别为1的短信的数量。其他特征则为关键字正则匹配得到的相关统计特征，如“bank_num_20d”表示20天内收到的含有银行相关关键字的短信数量。

首先使用 MV 指标对全量特征进行扫描，在99% 置信水平下，有277个特征显著地与因变量不独立，可以作为后续建模的备选特征。经过特征初筛，我们将原始的1433个特征减少到277个特征，在尽可能不损失信息情况下降低了后续模型的复杂度。

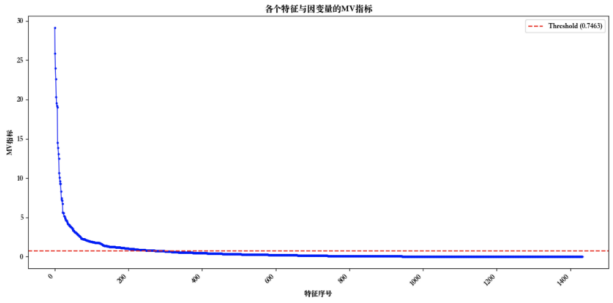

图1. 各个特征与因变量的 MV 指标取值

接着，我们使用277个特征作为备选特征，并引入一个虚拟特征（服从标准正态分布的随机数），共计278个特征，进行 LightGBM 建模。将数据集随机划分为训练集和测试集，并调整参数，得到最终模型。在参数调整过程中，关注虚拟特征是否入选模型。如果虚拟特征入选模型，则继续对参数进行微调，尤其是 L1 惩罚参数。经过多次随机划分训练集和测试集，我们发现经过 MV 指标初筛后再进行建模，虚拟特征基本不会入选模型。经过参数调整及虚拟特征验证我们得到了最终的模型，有181个特征入选模型， 图2展示了前25个重要特征的名称及 gain 取值。模型在 OOT 数据集上具有良好的效果， KS 达到0.34以上（图3）。模型在训练集和测试集上的 KS 均为0.36，整体模型表现平稳。

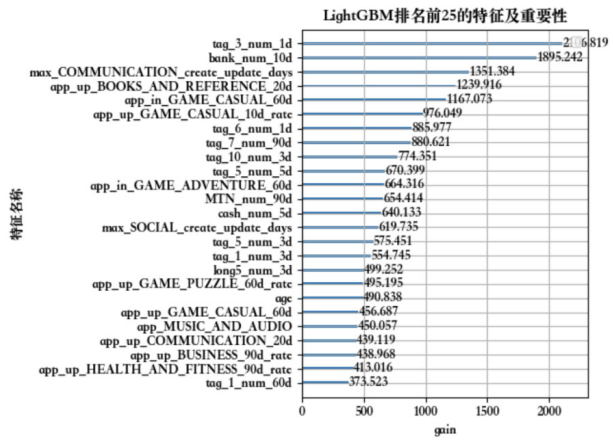

图2. 模型排名前25的重要特征

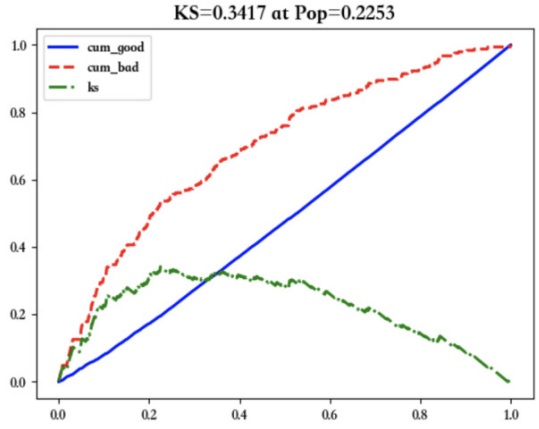

图3. 模型在 OOT 数据集上的效果

三、结束语

本文提出的两阶段信用评分建模方法，通过 Mean Variance 独立性检验的初步特征筛选和 LightGBM 分类模型的精准预测，有效解决了高维特征环境下构建信贷模型的特征冗余问题。此外，创新性地引入虚拟特征检测机制，进一步确保了模型的精简性与可靠性。实际业务数据的验证结果表明，该方法不仅显著提升了信用评分的准确性，同时增强了模型的稳定性，为大数据时代的信贷风险管理提供了切实可行的解决方案。未来研究可进一步探索动态特征更新机制，以适应不断变化的信用风险环境。

参考文献

[1] KE G, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems, 2017, 30: 3146-3154.
[2] BANASIK J, CROOK J, THOMAS L. Sample selection bias in credit scoring models[J]. Journal of the Operational Research Society, 2003, 54(8): 822-832.
[3] CHEN G G, ÅSTEBRO T. Bound and collapse Bayesian reject inference for credit scoring[J]. Journal of the Operational Research Society, 2012, 63(10): 1374-1387.
[4] FENG X, XIAO Z, ZHONG B, et al. Dynamic ensemble classification for credit scoring using soft probability[J]. Applied Soft Computing, 2018, 65: 139-151.
[5] DIRICK L, CLAESKENS G, JERUSALEM G, et al. Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models[J]. Journal of Business & Economic Statistics, 2019, 37(1): 40-53.
[6] FANG F, CHEN Y. A new approach for credit scoring by directly maximizing the Kolmogorov-Smirnov statistic[J]. Computational Statistics & Data Analysis, 2019, 133: 180-194.
[7] SHEN F, ZHAO X, KOU G. Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory[J]. Decision Support Systems, 2020, 137: 113366.
[8] KOZODOI N, JACOB J, LESSMANN S. Fairness in credit scoring: Assessment, implementation and profit implications[J]. European Journal of Operational Research, 2022, 297(3): 1083-1094.
[9] MUSHAVA J, MURRAY M. A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function[J]. Expert Systems with Applications, 2022, 202: 117233.
[10] HE H, ZHANG S, SHEN F, et al. A privacy-preserving decentralized credit scoring method based on multi-party information[J]. Decision Support Systems, 2023, 166: 113910.
[11] CHATTERJEE S, CORBAE D, NAKAJIMA M, et al. A quantitative theory of the credit score[J]. Econometrica, 2023, 91(5): 1803-1840.
[12] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1996, 58(1): 267-288.
[13] FAN J, LI R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.
[14] CUI H, LI R, ZHONG W. Model-free feature screening for ultrahigh dimensional discriminant analysis[J]. Journal of the American Statistical Association, 2015, 110(510): 630-641.
[15] 陈秋华，杨慧荣，崔恒建．变量筛选后的个人信贷评分模型与统计学习 [J]. 数理统计与管理，2020，39(2)：13.
[16] 王冠鹏，秦双燕，崔恒建．员工流失的影响因素分析与预测 [J]. 系统科学与数学，2022，42(6)：1616-1632.1

Copyright © This Work is Licensed under A Commons Attibution-Non Commercial 4.0 International License.