文章编号：2095-2163(2022)07-0052-08 中图分类号：F830 文献标志码：A

# 基于文本挖掘技术的信贷欺诈研究

刘娟娟[1]，梁龙跃[1]，蔡铉烨[2]
（1 贵州大学 经济学院，贵阳 550025；2 中央财经大学 统计与数学学院，北京 102206）

**摘 要**：有效识别贷款申请欺诈倾向是维护借贷双方利益的首要前提，是金融借贷市场一直以来关注的重点。随着文本挖掘技术的发展，贷款申请人提供的贷款描述，使其传达的信息受到更多关注。研究中利用贷款描述文本对欺诈行为进行识别，有助于拓宽非结构化文本数据在金融市场日常交易中的应用。利用深度学习模型 Transformer 对文本信息进行提取，再用自动编码器对文本信息进一步抽取，最终得到文本信息测度。基于 17 个指标构建基准机器学习模型，进一步加入文本信息测度作为新的预测变量。样本外预测结果显示，文本信息测度有助于提升模型拟合效果，在不同模型中提升精度介于 0.68%-1.42%之间，表明结果具有稳健性；特征重要性结果也表明，文本信息测度在模型预测结果的贡献度中位于前 4。验证了文本信息在欺诈识别中的作用。

**关键词**：文本挖掘；反欺诈；Transformer；自动编码器

# Research on credit fraud based on text mining technology

LIU Juanjuan [1], LIANG Longyue[1], CAI Xuanye[2]
(1 School of Economics, Guizhou University, Guiyang 550025, China;
2 School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China)

**【Abstract】** Effective identification of fraudulent tendencies in loan applications is the primary prerequisite for safeguarding the interests of both borrowers and lenders, and has always been the focus of the financial lending market. With the development of text mining technology, the information conveyed by loan descriptions provided by loan applicants has received more attention. The use of loan description texts to identify fraudulent behaviors in the research helps to broaden the use of unstructured text data in daily transactions in the financial market applications. We use the deep learning model transformer to extract the text information, the autoencoder to further extract the text information, and finally get the text information measurement. A benchmark machine learning model is constructed based on 17 indicators, and text information measures are further added as new predictor variables. The prediction results show that the text information measure helps to improve the model fitting effect, and the improvement accuracy is between 0.68% and 1.42% in different models, indicating that the results are robust. The feature importance results also show that the text information measure is in the top 4. Empirical results validate the role of textual information in fraud detection.

**【Key words】** text mining; anti-fraud; transformer; autoencoder

## 0 引 言

信贷欺诈识别不仅是国家有关部门关注的重点，亦是对金融市场日常交易中的严峻挑战。中国金融市场发展起步较晚，金融体系尚不完善，有效识别信贷欺诈问题，有利于互联网金融的创新发展和传统金融业的数字化转型升级。然而，仅靠年龄、学历、房产状况等“硬信息”识别欺诈行为具有一定局限性。大数据背景下，文本数据是经济学中应用较多的非结构化数据，其中蕴含着丰富的信息，被广泛应用于度量经济政策的不确定性、股价预测、波动率等[1]，以及将文本数据运用于违约预测[2]。

借贷申请人所提供的文本数据承载了申请人的意愿、倾向，该类文本数据是指其在申请贷款时所填写的贷款用途、贷款原因等文本，因此具有独特的价值意义。了解客户的资信状况是授信过程中十分关键的环节，是决定是否授予贷款的前提和基础，为此相关平台人员必须综合客户的有关信息（资信状况、还款意愿等），识别客户真伪信息[3]。文本数据的引入拓宽了了解客户信息的渠道，为全面评估客户、减少损失提供了保障。

在信贷欺诈识别模型中，机器学习算法是主流算法之一，与统计、计量分析方法（如：Logit 模型）相比，具有更高的识别效率和准确率[4]。利用机器学

**基金项目**：国家自然科学基金资助项目（52000045）；贵州大学研究生创新人才计划项目（CJ202169）。

**作者简介**：刘娟娟（1996-）女，硕士研究生，主要研究方向：信用风险、机器学习；梁龙跃（1986-）男，博士，讲师，硕士生导师，主要研究方向：数量经济学、机器学习；蔡铉烨（1994-）男，博士研究生，主要研究方向：金融数学、机器学习。

**通讯作者**：梁龙跃 Email：lyliang@gzu.edu.cn

**收稿日期**：2022-01-10

习进行欺诈数据检测主要分为 3 条路径：

(1)根据不平衡样本集，使用机器学习模型预测。如：文献[5]中构建决策树与布尔逻辑函数的融合模型，对金融消费行为进行分析，并在此基础上使用聚类方式区分正常交易与非正常交易，以此判断持卡人交易是否符合规范。文献[6]基于数据挖掘技术，设计信用卡欺诈检测系统，该系统使用贝叶斯分类器对客户数据进行识别，判断客户是否存在欺诈行为。文献[7]提出模糊二范数二次曲面支持向量机模型，用于信贷违约预测。实证结果表明，相比二次曲面支持向量机模型、二次核的加权二范数支持向量机模型等 4 个支持向量机变体模型而言，该模型评估效果得到显著提升。

(2)使用神经网络模型进行预测。文献[8]在 BP 神经网络基础上，融合遗传算法(GA)评估德国信用卡消费行为风险。该研究结果表明，混合模型效果优于单一的 BP 神经网络模型。

(3)平衡样本数据之后进行预测。由于欺诈数据往往具有样本分类不平衡的问题，SMOTE 算法平衡数据被广泛应用于欺诈检测。文献[9- 11]研究结果表明：样本平衡后能有效提升模型预测性能。

虽然贷款申请人所提供的文本数据蕴含丰富信息，但如何从该类文本数据中获取有效信息仍存在一些需要解决的问题。为此，相关人员做了大量的研究工作。文献[12]中指出，在传统的词频统计、词典法等方法中，由于选词及词典本身的限制，往往会存在信息遗漏问题。为了能够充分获取文本信息，自然语言处理技术已广泛应用于文本挖掘。如 CNN、LSTM、RNN、注意力机制等深度学习模型被广泛用于文本信息提取。文献[13]使用了几种典型的 CNN 模型，用于文本分类中的特征提取，获取文本信息的向量。随着人工智能技术的发展，文献[14]中提出了一种完全基于 Attention 机制的 Transformer 模型，打破了人们使用 RNN 与 CNN 做自然语言处理的局限。文献[15]使用多种方式提取文本特征作为新特征变量，用于构建信用违约模型(如：LDA、CNN、Transformer 等)。研究对比发现：加入 Transformer 模型提取的文本特征对模型性能提升效果高于其它文本提取方式。此外，使用深度学习模型所提取的文本信息存在高维问题，一般降维方式为 PCA、LASSO、核 PCA 等方法，但由于经由模型提取后的数据为非线性高维数据，一般降维方法不能有效解决非线性问题，为保证降维效果，需选取合适的降维方法。

本文致力于解决信贷文本信息的提取及降维，并将其运用于信贷欺诈识别。考虑到英文单词具有大小写之分，为降低其重复性，使用 Snowball 对英文进行词干还原，并在此基础上使用 Transformer 提取文本信息，有效获取了文本信息。其次，使用自动编码器(AE)对提取的文本信息进行非线性降维，成功获取文本信息测度指标。最后，利用多个机器学习模型(如：随机森林、XGBoost、GBDT 等)与数据均衡算法(SMOTE、TomekLinks 欠采样等)相结合，作为信贷欺诈识别基准模型。在其基础上引入文本信息测度作为新的预测变量，根据模型预测性能及特征重要性分析，研究贷款申请人所提供的文本数据对信贷欺诈识别的判断能力。

## 1　信贷文本信息提取建模

### 1.1　文本特征模型理论

#### 1.1.1　自动编码器(AE)

自动编码器(AE)是一种基于神经网络的数据降维方法[16]，主要包括编码(Encoder)和解码(Decoder)两部分，其网络结构如图 1 所示。当网络输入确定后，利用输出等于输入来训练自动编码器网络，使得输出尽可能地逼近输入。其中，隐层单元数量的选取要小于输入数据的维度。在数据降维中，AE 只需使用 Encoder 部分的编码操作，将高维度的输入数据映射到低维度的特征编码，达到降低数据维度的目的，且该方法相比于主成分分析(PCA)方法能以非线性方式解决多重共线性问题。

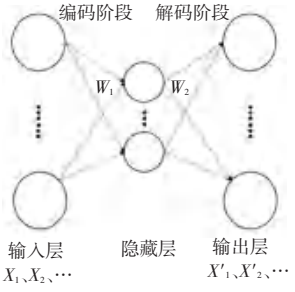


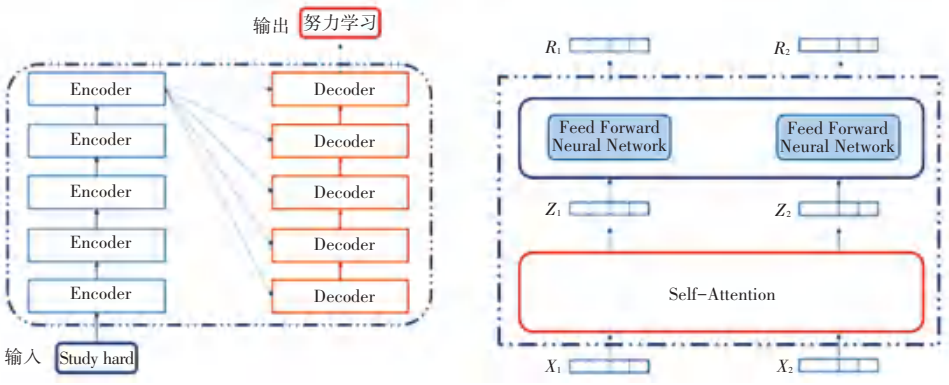

**图 1　自编码结构**

**Fig. 1　Autoencoder structure**

#### 1.1.2　Transformer

Transformer 由 Vaswani 等，在 2017 年提出，其开创性的放弃了基于 RNN、LSTM、GRU 等循环神经网络结构，取而代之使用了 Attention 层和全连接层构建网络，解决了语义长期依赖问题。位置编码器的引入解决了词语顺序的问题，并且由于没有了循环神经网络的递归结构，网络求解过程可以并行完成，大大提高了效率。该模型由一个完整的

Encoder-Decoder 框架构成，如图 2 所示。其中，Encoder 部分功能比较单一，仅用于从原始句子中提取特征，而 Decoder 则功能相对较多，除特征提取功能还包含语言模型功能。




（a）Encoder 框架　　（b）Decoder 框架

**图 2　Transformer 结构**

**Fig. 2　Transformer structure**

### 1.2　信贷文本信息获取及处理

#### 1.2.1　信贷文本信息获取

本文所使用的数据集，来源于美国大型信贷平台 Lending Club 所提供的 2007~2018 年贷款申请人信息，数据集中贷款申请人提供的“贷款描述”即是本文所使用的“文本信息”。该文本主要表现为贷款申请人的贷款目的、贷款理由自述及贷款类别。由于原始数据中并非所有样本均含有贷款描述，经数据预处理后总共获取有效文本信息 51 820 条，其中文本长度 90%以上少于 50 个单词，表明文本数据均为短文本。

#### 1.2.2　信贷文本信息处理

由于原始文本较短且英文单词无需进行分词，故本文在对原始文本进行去除无意义字符、词干还原及转化词向量后，基于 Python 软件构建 Transformer+AE 的融合模型对文本特征进行提取。由于该模型所提取的文本特征维度高达 68 维，为降低维度及便于后期衡量文本信息对模型贡献度，本文使用 AE 将文本信息降维至 1 维，获取最终的文本信息测度（文本特征）。实现流程如图 3 所示。

文本信息测度提取的主要步骤为：

**Step 1**　使用“正则表达式”，剔除无意义字符（如：日期、特殊符号等）。

**Step 2**　使用 Snowball 词干还原，获得原始单词后，通过词袋法对单词出现次数进行排序，选取出现次数排列前38 000的词，获得文本向量。

**Step 3**　将文本向量输入 Transformer 模型，训练并使用编码层获取文本特征（其中包括：位置编码层、Transformer 层以及全连接层），由此可得到多维度的文本特征。

**Step 4**　使用 AE 对高维文本特征进行非线性降维，最终获得一维文本信息测度。

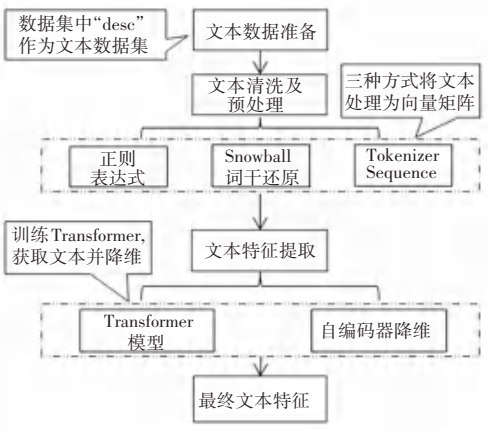



**图 3　文本特征获取流程**

**Fig. 3　Text feature acquisition process**

## 2　变量选取与模型构建

### 2.1　信贷欺诈数据收集及选取

与信用风险客户相比，欺诈风险客户主要表现之一为没有还款意愿，其目的是找到风控系统的漏洞或通过伪造信息等欺诈方式获得利益，是一种主观上的恶意欺诈、拖欠等行为[17]。从定义出发确定欺诈样本，将好样本标签以数字 1 表示，坏样本以数字 0 表示，便于后期模型拟合使用。

本文选取的原始数据集中共有 150 个特征变量，为了客观、全面判断借款人是否有欺诈意图，通过数据特征工程，选取以下 18 个指标构建反欺诈评估体系，各指标含义见表 1。

**表 1　部分特征介绍**

**Tab. 1　Introduction to part of features**

| 指标 | 释义 |
|---|---|
| term | 贷款期限,值以月为单位 |
| loan_amnt | 借款人所申请贷款的列明金额 |
| int_rate | 贷款的利率 |
| emp_length | 工作年限以年为单位 |
| annual_inc | 借款人在登记时自报年度收入 |
| verification_status | 收入是否通过信用证核实 |
| desc | 借款人提供的贷款描述 |
| dti | 每月还款总额/申报的每月收入 |
| Delinq_2yrs | 借款人过去 2 年逾期 30 天以上的违约次数 |
| mo_sin_old_rev_tl_op | 最古老的循环帐户已开立数月 |
| mo_sin_rcnt_rev_tl_op | 最近的循环帐户开立以来的几个月 |
| num_tl_op_past_12m | 过去 12 个月开立的账户数量 |
| total_bc_limit | 信用卡总信用/信用额度高 |
| total_il_high_credit_limit | 最高信用/信用额度 |
| fico | 借款人发放贷款时 fico 均值 |
| incom_load_rato | 申报收入与借贷数比值 |
| home_ownership | 借款人在登记时房屋所有权状况 |
| addr_state | 借款人在贷款申请中提供的国家 |

## 2.2　数据描述性统计

经数据预处理及特征工程后,最终剩余 51 820 个样本,样本集描述性统计结果见表 2。

**表 2　定量指标描述性统计**

**Tab. 2　Descriptive statistics of quantitative indicators**

| 指标名称 | 样本量 | 均值 | 方差 | 最小值 | 最大值 |
|---|---|---|---|---|---|
| target | 51 820 | 0.17 | 0.37 | 0.00 | 1.00 |
| loan_amnt | 51 820 | 17 216.63 | 8 303.89 | 1 000 | 35 000 |
| term | 51 820 | 44.31 | 11.42 | 36 | 60.00 |
| int_rate | 51 820 | 14.58 | 4.50 | 6 | 26.06 |
| emp_length | 51 820 | 6.43 | 3.56 | 0.00 | 10.00 |
| annual_inc | 51 820 | 81 390.03 | 61 562.43 | 8 400 | 7 141 778 |
| verification_status | 51 820 | 1.00 | 0.00 | 1 | 1.00 |
| desc | 51 820 | −0.01 | 0.00 | −0.02 | 0.00 |
| dti | 51 820 | 18.12 | 7.74 | 0.00 | 36.82 |
| delinq_2yrs | 51 820 | 0.28 | 0.76 | 0.00 | 18.00 |
| mo_sin_old_rev_tl_op | 51 820 | 188.11 | 86.40 | 5.00 | 760.00 |
| mo_sin_rcnt_rev_tl_op | 51 820 | 14.60 | 16.92 | 0.00 | 264.00 |
| num_tl_op_past_12m | 51 820 | 1.84 | 1.55 | 0.00 | 25.00 |
| total_bc_limit | 51 820 | 22 778.49 | 20 654.88 | 0.00 | 560 800 |
| total_il_high_credit_limit | 51 820 | 39 544.68 | 40 246.05 | 0.00 | 1 214 546 |
| fico | 51 820 | 698.18 | 29.73 | 662 | 847.50 |
| incom_load_rato | 51 820 | 5.86 | 5.98 | 2.00 | 481.74 |

根据数据描述性统计结果,数据集方差差异显著。为提高模型拟合结果,需对数据进行归一化处理,针对分类变量 home_ownership、addr_state 进行 One-Hot 编码。归一化处理公式为:

$$X_{normalization} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

## 2.3　基准模型介绍

### 2.3.1　随机森林模型

随机森林(Random Forest,RF)算法是一种经典的装袋法(Bagging)模型,其基本原理是先在原始数据集中随机抽样,构成 $n$ 个不同的样本数据集,然后根据这些数据集搭建 $n$ 个不同的决策树模型,最后根据这些决策树模型的投票情况获取最终结果。随机森林具有拟合速度快,方便处理大规模数据、易于实现、可以避免过拟合等优点。

### 2.3.2　GBDT 模型

GBDT(Gradient Boosting Decision Tree)属于提升(Boosting)集成算法中的一种。Boosting 集成算法的构建过程,是不断加强之前弱学习器判别错误的样本权重,保证之后的弱学习器在错误样本上判别正确。GBDT 算法将损失函数的负梯度作为残差的近似值,不断使用残差迭代和拟合树,使残差沿着最大梯度的方向下降,最终生成强学习器。

### 2.3.3　XGBoost 模型

XGBoost(eXtreme Gradient Boosting)是在 GBDT 的基础上,引入正则化损失函数来实现弱学习器的生成。加入了正则化的损失函数,不仅可以降低过拟合的风险,且 XGBoost 模型利用损失函数的一阶导数和二阶导数值进行搜索,通过预排序、加权分位数、稀疏矩阵识别及缓存识别等技术,大大提高了 XGBoost 模型性能。XGBoost 通过最小化下面的正则化目标函数来实现:

$$L(\varnothing) = \sum_i l(\hat{y}, y) + \sum_k \Omega(f_k) \tag{2}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3}$$

其中,$l$ 是损失函数;$\Omega$ 是模型复杂程度的惩罚项;$\gamma$、$\lambda$ 分别是 $L_1$、$L_2$ 的正则化系数。

### 2.3.4　LightGBM 模型

LightGBM 算法在原理上与 GBDT 和 XGBoost 算法类似,都采用损失函数负梯度作为当前决策树的残差近似值,去拟合新的决策树。只是对框架进行了优化(重点对模型训练速度的优化)。其二叉树的分裂增益公式为:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)}{H_L+H_R+\lambda}\right] - \gamma \tag{4}$$

其中，$\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)}{H_L+H_R+\lambda}\right]$ 是指不考虑其它因素，通过分裂得到的增益。但实际上每次引入新叶子节点，都会带来复杂度的代价，即 $\gamma$。

$$G_j = \sum_{i\in I_j} g_i,\ i = 1,2,\cdots,n;\quad j = 1,2,\cdots,T \tag{5}$$

$$H_j = \sum_{i\in I_j} h_i,\ i = 1,2,\cdots n;\quad j = 1,2,\cdots,T \tag{6}$$

其中，$G_j$ 为该叶子节点上样本集合中数据点在误差函数上的一阶导数和二阶导数。

2.3.5 Extra-Trees 模型

极端随机树（Extra-Trees，ET）算法与随机森林算法十分相似，都是由许多决策树构成。ET 算法在节点划分时，选择的特征及对应的特征值不是搜索比较所得，而是随机抽取一个特征，再从该特征中随机抽取一个特征值，作为该节点划分的依据。当子模型的准确率大于 50%，并且集成的子模型数量足够多时，整个集成系统的准确率达到合格。这样做的优点是：提供额外的随机性、抑制过拟合，并且具有更快的训练速度，缺点是增大了偏差（bias）。

2.3.6 ANN 模型

人工神经网络（ANN）是由大量神经元模型组成的信息响应网络拓扑结构，其可以分为几个“层”，如：输入层、隐藏层和输出层。其中，输入层和输出层功能较为单一，隐藏层功能较多。隐藏层可以由多层神经网络层构成，其主要作用是对输入层输入的数据进行计算转换，并将得到的结果传递给输出层。整个神经网络中，每层内部的神经元没有连接，连接只设置在层与层之间。此外，每个连接都具有一个权重值。

## 3 实证分析

本文使用 Python 软件展开实证分析，构建欺诈检测模型，将 51 820 个样本按 9∶1 的比例划分训练集和测试集。由于数据样本的不均衡性，会对模型拟合效果评价产生较大影响，本文选取不同的欠采样、过采样方式对数据集进行均衡采样，探索不同采样方式下模型性能的表现。同时，多元化采样方式有助于增强模型结果稳健性。实证结果表明，在不同采样方式下，加入文本特征后模型性能均有一定提升。实证过程中，将样本集分为两组，一组不加入文本特征指标，另一组加入文本特征指标。

### 3.1 实验结果评价

3.1.1 评价指标

3.1.1.1 真正例率（*TPR*）和假正例率（*FPR*）

在反欺诈模型中，其目的是为了检测出欺诈样本。由于传统的准确率（*Accuracy*）指标无法准确评价该模型实际欺诈检测准确率，为此模型评价采用 *AUC* 指标，并绘制出模型的 ROC 曲线。

对于一个二分类任务，可将所有的样例根据其真实所属类别与模型结果组合分为真正例（*TP*）、假反例（*FN*）、假正例（*FP*）、真反例（*TN*）4 种情况，见表 3。

**表 3 混淆矩阵**

**Tab. 3 Confusion matrix**

| | 预测（正例） | 预测（反例） |
|---|---|---|
| 真实（正例） | 真正例（*TP*） | 假反例（*FN*） |
| 真实（反例） | 假正例（*FP*） | 真反例（*TN*） |

根据表 3 可定义真正率（*TPR*）和假正率（*FPR*）为：

$$TPR = \frac{TP}{TP+FN} \tag{7}$$

$$FPR = \frac{FP}{FP+TN} \tag{8}$$

3.1.1.2 ROC 曲线和 *AUC* 值

受试者工作特征曲线（Receiver Operating Characteristic Curve，ROC）以 *FPR* 为横轴，*TPR* 为纵轴绘制，当其越靠近左上角，表明模型的性能越好，如图 4 所示。但当存在多条 ROC 曲线很难进行比较时，可使用 *AUC* 值对模型性能进行评估。*AUC* 是 ROC 曲线和 $x$ 轴（*FPR* 轴）之间的面积，其值能直接反映出模型拟合结果的优劣。

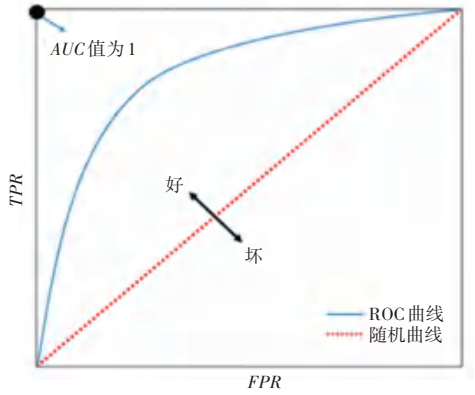



**图 4 ROC 曲线**

**Fig. 4 ROC curve**

3.1.2　实验结果评价

本文选用随机森林、GBDT、XGBoost、LightGBM、ET 以及全连接神经网络(ANN)共 6 个机器学习模型，验证在不同模型上文本信息测度对预测结果贡献的稳健性。

对全样本分别进行邻域欠采样、Tomek Links 欠采样、随机欠采样、随机过采样以及 SMOTE 过采样。为了降低模型过拟合及更多的获取数据信息，研究中将训练集数据随机划分为 10 份进行交叉验证，每次选取其中一份作为校验集，其余部分作为训练集用于模型训练。

3.1.2.1　加入文本数据前预测模型实验结果

根据表 4 可知，除 SMOTE 采样下，LGBM 模型表现最好以外，其余采样方式下最好模型均为 GBDT；在邻域欠采样下，所有模型评价结果明显高于其它采样方式。从总体评价结果来看，GBDT 模型拟合结果最佳。

**表 4　未加文本特征 *AUC* 值**

**Tab. 4　*AUC* value without text feature**

| 模型 | 邻域欠采样 | TomekLinks 欠采样 | 随机过采样 | 随机欠采样 | SMOTE 过采样 |
|---|---|---|---|---|---|
| ET | 0.798 8 | 0.680 2 | 0.683 0 | 0.681 7 | 0.676 1 |
| GBDT | 0.821 9 | 0.710 6 | 0.706 6 | 0.699 6 | 0.693 2 |
| LGBM | 0.816 4 | 0.705 4 | 0.701 1 | 0.694 8 | 0.701 4 |
| RF | 0.809 5 | 0.694 2 | 0.684 5 | 0.695 2 | 0.678 6 |
| XGBoost | 0.805 1 | 0.683 5 | 0.680 2 | 0.675 8 | 0.690 1 |
| ANN | 0.785 6 | 0.673 6 | 0.685 3 | 0.676 2 | 0.662 8 |

3.1.2.2　加入文本数据后预测模型实验结果

从采样方式看：邻域欠采样下所有模型评价结果均高于其他采样方式，其中 SMOTE 过采样方式下除 LightGBM 模型外，其它模型结果均表现欠佳。由此可知，领域欠采样方式是最优采样方式，对提高模型评价结果具有一定意义。从模型角度看，除 SMOTE 过采样方式，其余采样方式下最佳拟合模型为 GBDT 模型，其 *AUC* 值高于其它模型。

**表 5　加入文本特征后 *AUC* 值**

**Tab. 5　*AUC* values after adding text features**

| 模型 | 邻域欠采样 | TomekLinks 欠采样 | 随机过采样 | 随机欠采样 | SMOTE 过采样 |
|---|---|---|---|---|---|
| ET | 0.805 6 | 0.704 5 | 0.705 3 | 0.697 2 | 0.687 8 |
| GBDT | **0.832 0** | **0.724 6** | **0.723 5** | **0.711 2** | 0.693 0 |
| LGBM | 0.825 7 | 0.721 5 | 0.717 1 | 0.705 4 | **0.707 7** |
| RF | 0.823 7 | 0.709 7 | 0.705 7 | 0.705 4 | 0.692 1 |
| XGBoost | 0.814 0 | 0.708 6 | 0.693 3 | 0.689 7 | 0.692 9 |
| ANN | 0.793 2 | 0.691 2 | 0.689 6 | 0.685 8 | 0.681 4 |

对比无文本特征模型的 *AUC* 值，含文本特征模型 *AUC* 值均有显著提升，最高提升效果为 1.42%(随机森林模型)，最差提升效果为 0.68%(ET 模型)，GBDT 模型作为 *AUC* 值最高模型，其提升效果为 1.01%。因此，加入文本特征对模型性能具有提升效果，该特征对预测结果有贡献作用。

## 3.2　模型特征重要性分析

特征重要性可以查看特征变量对目标变量的作用，且按作用大小进行排序。本文选取了提升表现较好的 4 个模型进行特征重要性分析，提取欺诈检测模型中排名前 10 的特征，并观察文本特征在前 10 重要特征中的位置，结果如图 5 所示。

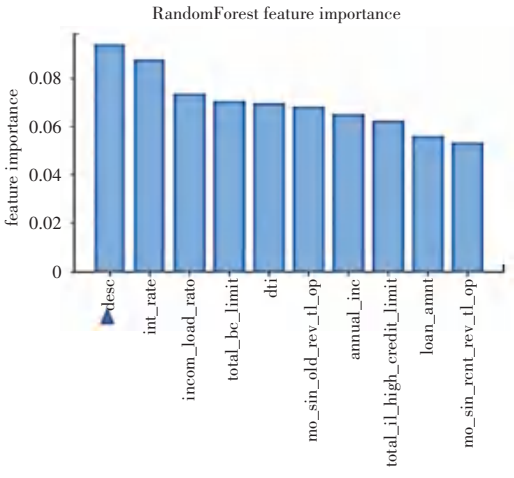
图 5(a)表明，在随机森林模型中，最重要的特征变量为“desc”(文本特征)。可以看出加入文本信息特征会对模型预测的结果造成较大影响，证明文本信息特征能有效改变模型预测结果；而在硬特征中，贷款利率(int_rate)占有重要影响地位。

图 5(b)显示在 GBDT 模型中，最重要的特征变量为 int_rate，次重要特征为 desc，可看出文本特征对模型预测结果的影响程度较为显著。

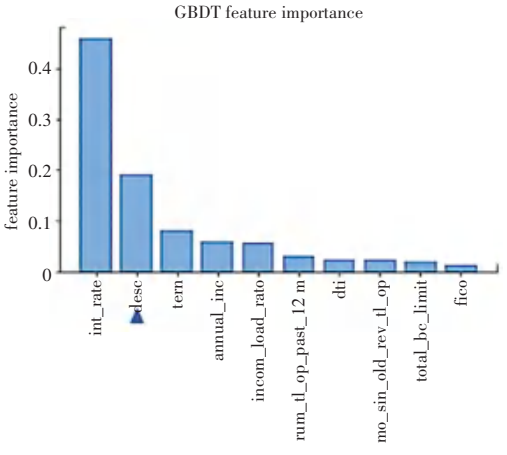
图 5(c)显示文本信息特征“desc”重要性位列第四，展示了加入文本信息特征的作用。除此之外，int_rate 及 term 重要性表现出一致性，且位列第一、第二。

图 5(d)的 LightGBM 模型中，文本(desc)特征重要性排位第一，且重要性显著高于其它特征。除去文本特征外，前 4 个特征的重要性基本一致。

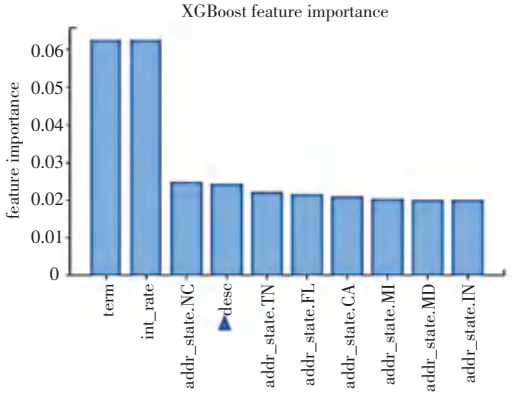
由特征重要性图示可知,文本特征指标在各模型中均是重要特征,在大部分模型中位列第一和第二,其重要性相比硬特征处于重要位置,对模型的预测结果贡献较大。从而验证了加入文本特征后,反欺诈模型风险识别能力得到提升,文本特征的引入具有一定意义。



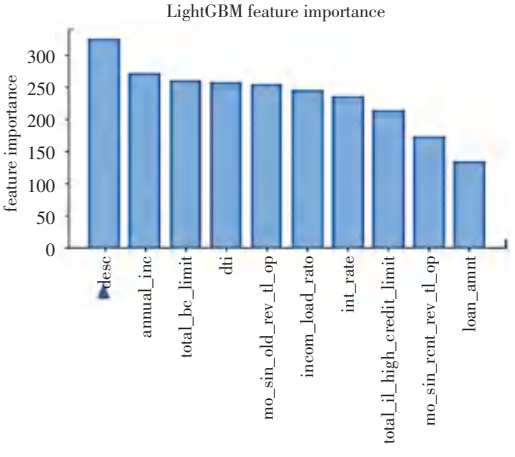

(a) 随机森林




(b) GBDT




(c) XGBoost




(d) LightGBM

注:图中带三角形标号表示文本特征(desc)

**图 5 特征重要性结果图**

**Fig. 5 Feature importance results**

## 4 结束语

本研究中引入文本信息作为新的影响因子,探索了贷款文本信息对欺诈识别的作用,拓宽了非结构化数据在金融交易中的应用。此外,将 Transformer 与 AE 相结合,有效降低了文本信息维度,同时也保证了信息的全面性。

研究结果表明,以贷款利率、借款人年收入、最早循环帐户已开立月数及文本特征为主的 10 个指标与客户欺诈行为相关性最高。在反欺诈预测模型中,文本信息的引入,能够明显提升模型对欺诈客户的识别性能,提升结果介于 0.65%- 1.42%之间。启示有关金融机构平台,在审核贷款申请人信息时,可要求贷款申请人提供必要的文本“软信息”,获取更丰富的贷款人信息,更为全面评估是否授予贷款,维护双方利益,减少不必要损失。

在未来工作中,除基础自编码器外,还可使用其它编码器进行数据降维,也可尝试使用其他新算法构建反欺诈模型,探索更多欺诈检测方式。文本挖掘技术的发展日新月异,新兴的文本挖掘技术也可用于提取文本特征,亦是今后可以挖掘的方向。由于文本特征的特殊性,其对目标变量的影响机制有待进一步挖掘,未来可探究文本特征可解释性分析。

## 参考文献

[1] 沈艳,陈赟,黄卓. 文本大数据分析在经济学和金融学中的应用:一个文献综述[J]. 经济学(季刊),2019,18(4):1153-1186.

(下转第 68 页)