Hans Journal of Data Mining 数据挖掘, 2019, 9(4), 145-152
Published Online October 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.94018

Hans汉斯

Research on Bank Credit Card Default Prediction Based on Machine Learning

Huawei Shan

University of International Business and Economics, Beijing
Email: 793347061@qq.com

Received: Oct. 9th, 2019; accepted: Oct. 17th, 2019; published: Oct. 24th, 2019

Abstract

Credit card business is the core business of Banks. Commercial Banks seize the market and develop customers by issuing credit cards. Although credit card business brings high profits to banks, extensive credit card management leads to high default rate of credit card customers, which brings great risks to banks. Therefore, how to effectively manage the risk of credit card business has become one of the hot issues in the banking industry. This paper uses machine learning related algorithms to construct a bank credit card default prediction model, predicts credit card users' defaults in the next month, and assists banks in risk management. Specifically, this paper constructs credit card default prediction models through logistic regression, decision tree, random forest, adaboost and gradient boosting decision tree, and compares the prediction effects of five models under different feature selection methods through evaluation indexes such as accuracy. In this paper, relevant data of credit card holders of a bank are used for experiments. The experimental results show that different feature selection methods have a greater impact on model performance than algorithm selection. Among them, the filter feature selection is more adaptable.

Keywords

Machine Learning, Credit Card Default, Feature Selection

基于机器学习的银行信用卡违约预测研究

单华玮

对外经济贸易大学，北京
Email: 793347061@qq.com

收稿日期：2019年10月9日；录用日期：2019年10月17日；发布日期：2019年10月24日

文章引用：单华玮. 基于机器学习的银行信用卡违约预测研究[J]. 数据挖掘, 2019, 9(4): 145-152.
DOI: 10.12677/hjdm.2019.94018

摘　要

信用卡业务是银行的核心业务，各大商业银行通过发行信用卡来抢占市场和发展客户。虽然信用卡业务给银行带来了高额利润，但信用卡的粗放式管理导致信用卡客户存在较高的违约率，给银行带来了极大的风险。因此，如何有效针对信用卡业务进行风险管理已经成为银行业的热点关注问题之一。本文采用机器学习的相关算法构建银行信用卡违约预测模型，预测信用卡用户次月的违约情况，辅助银行进行风险管理。具体地，本文通过逻辑回归、决策树、随机森林、自适应增强和梯度提升树这五类算法来构建信用卡违约预测模型并通过准确率等模型评价指标对比不同特征选择方式下五种模型的预测效果。本文使用某银行信用卡持卡人的相关数据进行实验，实验结果表明，相比于算法选择，不同的特征选择方式对于模型性能有更大的影响，其中，过滤式特征选择的适应性更强。

关键词

机器学习，信用卡违约，特征选择

1. 引言

信用卡诞生于十九世纪八十年代末，起初被应用于服装行业。二十世纪五十年代，美国银行正式发行银行信用卡，民众可以使用银行信用卡透支消费并享有一定的免息还款期。信用卡的便捷性为人们的日常生活带来了极大便利，迅速受到社会各界民众的欢迎。随着经济全球化时代的到来，信用卡业务已经成为全球金融服务市场中增速最快的产品之一，为银行带来了巨额利润。然而，由于信用卡的粗放式管理，信用卡客户的违约率逐年增加，为银行带来了极大的风险。因此，如何有效地利用信用卡客户的相关信息对客户违约风险进行管控，已经成为银行业的重点关注问题之一。

本文采用机器学习的相关技术来构建银行信用卡违约预测模型，辅助银行进行风险预判，降低银行的损失。具体地，本文通过逻辑回归、决策树、随机森林、自适应增强和梯度提升树这五类算法来构建信用卡违约预测模型。此外，通过准确率等评价指标来对比不同特征选择方式对模型的预测效果的影响。

2. 相关工作

近年来，机器学习的相关技术被广泛应用于银行信用卡违约预测任务的相关研究中，并取得了不错的成果。方匡南等人[1]使用 Lasso-logistic 筛选评价指标和构建个人信用风险评估模型，提高了信用风险预警的效果。Venkatesh 等人[2]使用随机森林构建信用卡违约预测模型。Yeh 等人[3]在信用卡违约预测研究中使用人工神经网络(ANN)构建模型，并取得了良好的性能。Yang 等人[4]在信用卡违约预测研究中引入集成学习模型。与支持向量机等模型相比，集成学习模型具有更优的性能。为了避免复杂的特征工程，Hsu 等人[5]使用递归神经网络(RNN)作为特征提取器并利用提取的动态特征以及静态特征来训练增强的 RNN 模型(RNN-RF)，以预测信用卡的违约情况。

在传统机器学习算法的基础上，越来越多的学者使用改进的算法进行银行信用卡违约预测的相关研究。朱健[6]在信用卡违约预测研究中使用改进的随机森林算法来构建模型。具体地，通过对自变量进行

混合核内积运算的方式来扩大随机森林的自变量选择空间，提高了原有随机森林算法的分类效果。刘铭等人[7]在传统模糊神经网络的基础上，借助灰狼算法提出了改进型模糊神经网络算法，并将其应用于信用卡违约预测研究。Bahnsen 等人[8]提出了贝叶斯最低风险(BMR)模型，并利用该模型预测信用卡欺诈情况。Xu 等人[9]提出了一种基于 RIPPER 算法的改进模型，利用 RIPPER 算法生成的规则对信用卡违约客户进行预测。

3. 实验过程

3.1. 数据描述

本文使用 Kaggle 提供的 default of credit clients 数据集，包括某银行(现金及信用卡发行)的信用卡持卡人的个人信息和用户信用卡消费还款信息。该数据集共有 3 万条数据和 24 个属性，具体地，包括 23 个特征描述属性和 1 个目标属性(次月是否违约)。具体信息如下表 1 所示。

Table 1. Introduction of data attributes
表 1. 数据属性介绍

属性	说明	取值
ID	编号	1, 2, …, 30,000
LIMIT_BAL	授信金额	[10,000, 1,000,000]
SEX	性别	1, 2
EDUCATION	教育	0, 1, 2, 3, 4, 5, 6
MARRIAGE	婚姻	0, 1, 2, 3
AGE	年龄	21, …,79
PAY_0	9 月的还款情况	−2, −1, 0, 1, …, 8
……	……	……
PAY_6	4 月的还款情况	−2, −1, 0, 1, …, 8
BILL_AMT1	9 月账单数额	[−165,580, 964,511]
……	……	……
BILL_AMT6	4 月账单数额	[−339,603, 961,664]
PAY_AMT1	9 月以前的付款金额	[0, 873,552]
……	……	……
PAY_AMT6	4 月以前的付款金额	[0, 528,666]
Default.Payment.Next.Month	下个月是否支付	0, 1

3.2. 数据预处理

原始数据中往往含有很多噪声数据，为了提高数据质量，需要对原始数据进行数据预处理。本文的数据预处理阶段主要包括数据清洗、特征处理、样本均衡和特征选择四个部分。

1) 数据清洗

在查看数据集属性取值和分布情况的基础上，针对教育(EDUCATION)、婚姻(MARRIAGE)和 n 月的还款情况(PAY_n)这三类属性进行数据清洗。EDUCATION、MARRIAGE 和 PAY_n 均包含未知取值且未知取值在每个属性中占比较小，故而将部分未知取值合并到合理的已知取值中。例如：EDUCATION 有

7 类取值，其中，取值为 0、取值为 5 和取值为 6 表示的含义未知且占比很低，故将这三个类取值归并到取值 4 (其他)中。

2) 特征处理

由于不同模型对输入数据的要求不同，故需要对原始数据进行特征处理以适应模型的需求。特征处理阶段主要包括特征编码和数据标准化两个部分。特征编码是针对数据集中的离散型变量进行独热(one-hot)编码。数据标准化是针对连续型变量进行区间缩放标准化，将取值范围缩放到 0~1 之间。

3) 样本均衡

原始数据集中未违约客户(取值为 0)的人数远高于违约客户(取值为 1)的人数，属于样本不均衡数据集。为了避免样本不均衡对模型的性能产生影响，本文采用 SMOTE 过采样算法为未违约类合成新样本，从而实现样本均衡。样本均衡前后数据集的数据分布情况如表 2 所示。

Table 2. Data distribution before and after oversampling
表 2. 过采样前后数据分布情况

阶段	类分布情况
SMOTE 过采样前	0:0.7788 1:0.2212
SMOTE 过采样后	0:0.5 1:0.5

4) 特征选择

本文采用包裹式特征选择和过滤式特征选择两种方式来探索不同的特征对于模型性能的影响。包裹式特征选择是以机器学习模型和评测性能的指标作为特征选择的准则，每次对若干特征进行筛选。过滤式特征选择是以每个特征和结果的相关性作为评估指标对若干特征进行筛选，最终保留相关性最强的几个特征。

3.3. 研究方法

由于银行信用卡持卡人的违约情况通常包括未违约和违约两类，因此，银行信用卡违约预测可以定义为一个二分类任务。本文拟采用机器学习中监督学习的方式在训练集上训练模型并在测试集上检验模型的性能。下面具体介绍本文使用的分类模型：

1) 逻辑回归模型

逻辑回归模型(Logistic Regression)是常见的机器学习模型，多用于二分类任务。逻辑回归模型首先通过对多元线性回归进行非线性变换得到所需的分类函数，进而通过构建损失函数来表示预测的输出类别与训练数据的实际类别之间的偏差。之后，通过最小化损失函数获取模型的最优参数。具体公式如公式(1)、(2)所示：

$$h_\theta\left(x^{(i)}\right)=\text{sigmoid}\left(\theta_0+\theta_1X_1^{(i)}+\theta_2X_2^{(i)}+\cdots+\theta_nX_n^{(i)}\right) \tag{1}$$

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right)+\left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right] \tag{2}$$

其中，$h_\theta\left(x^{(i)}\right)$ 表示分类函数，$J(\theta)$ 表示损失函数，$X_1^{(i)}$ 表示第 i 个样本的第一个特征，θ_n 表示第 n 个特征的特征权重。

2) 决策树模型

决策树模型[10] (Decision Tree)是通过一系列规则对数据进行分类的模型，主要包括 ID3、C4.5 和

CART 三种基本方法。本文采用 CART 决策树来构建信用卡违约预测模型。CART 决策树以基尼系数作为分类标准，通过基尼系数选择最优特征并决定特征的最优二值切分点。基尼系数越大，样本集合的纯度越低。假设给定样本集合 D，基尼系数的公式如公式(3)所示：

$$\mathrm{Gini}(D)=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{3}$$

其中，C_k 表示样本集合 D 中属于第 k 类的样本子集。

3) 随机森林模型

随机森林模型[11] (Random Forest)是集成学习中基于自助聚集策略(Bagging)的典型代表模型，被广泛应用于各种分类任务。随机森林模型是指集合多个决策树对数据集进行训练并预测的分类模型。具体而言，随机森林模型首先通过有放回的方式从原始样本中进行 n 次随机采样得到 n 个决策树训练所需的样本集。之后，独立训练每个决策树。在此基础上，通过一定的集合策略得到样本最终的输出结果，其中，集合策略主要包括两类方法：加权平均法和投票法。

4) 自适应增强模型

自适应增强模型[12] (AdaBoost)是集成学习中基于提升策略(Boosting)的代表模型之一。自适应增强模型基于相同的训练集训练得到不同的弱学习器，进而集合多个弱分类器得到所需的强分类器。在迭代训练过程中，自适应增强模型通过调整样本和弱分类器的权重筛选权值系数最小的弱分类器组合成一个强分类器。

5) 梯度提升树模型

梯度提升树模型[13] (GBDT)是集成学习中基于提升策略(Boosting)的代表模型之一，被广泛应用于分类任务中，展现了良好的性能。梯度提升树模型采用 CART 决策树作为基分类器，通过 Gradient Boosting 对基分类器进行集成学习。在迭代训练过程中，梯度提升树将每轮迭代训练产生的弱分类器在上一轮分类器的残差基础上不断训练，以逼近真实值。

3.4. 评价指标

本文采用准确率(Accuracy)和受试者工作特征曲线下的面积(AUC)作为模型的评价指标。在机器学习模型的性能度量中，准确率是常见的评价指标，被广泛用于度量二分类或多分类任务模型的性能。准确率是指预测正确的样本数占总样本数的比例，通常而言，准确率越高，模型的性能越好。准确率的具体公式如公式(4)所示：

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

其中，TP (True Positive)代表将正类预测为正类的个数，TN (True Negative)代表将负类预测为负类的个数，FN (False Negative)代表将正类预测为负类的个数，FP (False Positive)代表将负类预测为正类的个数。

AUC 是指横纵坐标分别为假正率(FPR)和真正率(TPR)的 ROC 曲线下方的面积，表示模型预测的正类排在负类前面的概率。AUC 是衡量二分类任务中模型性能优劣的评价指标之一。AUC 值的大小与模型的性能优劣呈正相关，即 AUC 值越大，模型的性能越优。一般而言，AUC 值通常大于 0.5，若小于 0.5，则代表模型预测效果与随机猜想的效果相同。

4. 实验结果与分析

4.1. 盲机学习中不同模型的性能对比

本文首先采用盲机学习的方式训练各类机器学习模型，并将训练好的模型在测试集上预测，检验模

型的性能。为了避免模型参数对模型性能产生较大影响，通过 GridCV 选择最佳参数。各个机器学习模型的具体测试结果如表 3 所示。

Table 3. Performance of different models in blind machine learning
表 3. 盲机学习方式中不同模型的性能

模型	AUC 值	平均准确率	1 类准确率	0 类准确率
逻辑回归模型	0.7000	78.21%	55.36%	84.63%
决策树模型	0.6884	79.25%	50.3%	87.39%
随机森林模型	0.6892	80.04%	49.1%	88.73%
自适应增强模型	0.6878	79.83%	49.1%	88.46%
梯度提升树模型	0.6517	79.3%	39.9%	90.4%

从表 3 可以看出，在盲机学习的方式下，五种模型的稳健性由高到低依次为：逻辑回归模型、随机森林模型、决策树模型、自适应增强模型、梯度提升树模型。从预测准确率而言，随机森林模型、自适应增强模型和梯度提升树模型的性能要优于单一的决策树模型和逻辑回归模型。

对于银行而言，模型能够正确预测违约用户的重要性要高于正确预测未违约用户。从 1 类准确率和 0 类准确率两个评价指标中可以看出，五种模型对于未违约用户的预测准确率均高于违约用户。其中，逻辑回归模型在预测违约用户方面的性能最佳。

4.2. 不同特征选择方式的模型性能对比

从时间复杂度而言，输入数据的特征越少，模型训练的速度越快，时效性越高。因此，本文通过特征选择从所有特征中筛选出重要特征来构建模型，提升模型的训练速度。本文主要采用两种特征选择方式：包裹式特征选择和过滤式特征选择。

4.2.1. 包裹式特征选择

在包裹式特征选择方法中，使用递归特征消除(RFE)策略来筛选特征，其中，基模型为随机森林模型。在此基础上，使用筛选后的授信金额(LIMIT_BAL)、年龄(AGE)等 10 个特征来构建逻辑回归模型、决策树模型、随机森林模型、自适应增强模型和梯度提升树模型，并在测试集上检验模型的性能。具体的模型测试结果如表 4 所示。

Table 4. Performance of different models after wrapping feature selection
表 4. 包裹式特征选择后不同模型的性能

模型	AUC 值	平均准确率	1 类准确率	0 类准确率
逻辑回归模型	0.6108	78.98%	29.2%	92.96%
决策树模型	0.5765	63.07%	48%	67.3%
随机森林模型	0.654	80.54%	38.4%	92.38%
自适应增强模型	0.6747	77.81%	49.06%	85.89%
梯度提升树模型	0.5903	52.67%	70.34%	47.73%

由表 4 可知，相比于盲机学习，五种模型的稳健性均有所下降，拟合效果均比未经过特征选择的模型的拟合效果差。从违约用户的预测准确率而言，与盲机学习相比，逻辑回归模型、决策树模型、随机

森林模型、自适应增强模型的预测性能均有所下降，而梯度提升树模型对于违约用户的预测准确率有显著提升。综上结果可以看出，包裹式特征选择方式的适应性较弱。

4.2.2. 过滤式特征选择

由于数据集中既有离散型属性又有连续型属性，故采用卡方检验来选择特征，并使用筛选后的 10 个特征构建逻辑回归模型、决策树模型、随机森林模型、自适应增强模型和梯度提升树模型。各个模型的测试结果如表 5 所示。

Table 5. Performance of different models after filtering feature selection
表 5. 过滤式特征选择后不同模型的性能

模型	AUC 值	平均准确率	1 类准确率	0 类准确率
逻辑回归模型	0.6934	76.93%	55.83%	82.86%
决策树模型	0.6905	77.01%	54.86%	83.23%
随机森林模型	0.6913	76.15%	56.61%	81.64%
自适应增强模型	0.6916	77.06%	55.09%	83.23%
梯度提升树模型	0.6883	77.09%	54.12%	83.54%

由表 5 可以看出，过滤式特征选择对于五种模型的影响基本相同，从侧面证实了该种特征选择方法的适用性较强。相比于盲机学习，五种模型对于违约用户的预测准确率均有所提升，最低提升 0.5%，最高提升 14%；对于未违约用户的预测准确率有所下降，但仍然维持在一个较高的水平；对于整体的预测准确率均有小幅度的下降。

从以上实验结果可以看出，虽然盲机学习利用了原始数据中的所有特征来训练模型，但与经过特征选择后训练的模型相比，性能并未有显著提升。其中，过滤式特征选择采用 10 个特征训练模型，而模型对于违约用户的预测性能和可解释性均有所提升。

5. 结论

本文采用机器学习的相关技术构建信用卡客户违约预测模型，辅助银行及时发现信用卡违约用户，降低银行的风险，避免不必要的损失。本文从特征和模型两个维度进行实验，对比不同特征选择方式下五种不同模型的性能。从相关分析中可以看出，同一特征选择条件下，集成模型的综合性能往往要优于单一模型，效果更好。因此，银行可以优先考虑使用集成模型来对信用卡客户违约情况进行预测。从特征选择的方式而言，过滤式特征选择方法的适应性比包裹式特征选择方法的适应性强。此外，相比于盲机学习，过滤式特征选择能够稳定提升各种模型对违约用户的预测准确率。银行可以首先考虑采用过滤式特征选择方式对相关数据进行预处理，不仅能够提升模型的训练速度，也能增强模型的可解释性。

参考文献

[1] 方匡南, 章贵军, 张惠颖. 基于 Lasso-logistic 模型的个人信用风险预警方法[J]. 数量经济技术经济研究, 2014, 31(2): 125-136.

[2] Venkatesh, A. and Jacob, S.G. (2016) Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers. *International Journal of Computer Applications*, **145**, 36-41. https://doi.org/10.5120/ijca2016910702

[3] Yeh, I.C. and Lien, C. (2009) The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications*, **36**, 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020

[4] Yang, S. and Zhang, H. (2018) Comparison of Several Data Mining Methods in Credit Card Default Prediction. *Intel-*

ligent Information Management, **10**, 115. https://doi.org/10.4236/iim.2018.105010

[5] Hsu, T.C., Liou, S.T., Wang, Y.P., *et al*. (2019) Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Hof, 12-17 May 2019, 1572-1576. https://doi.org/10.1109/ICASSP.2019.8682212

[6] 朱健. 四种数据挖掘算法的信用卡违约识别对比研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2017.

[7] 刘铭, 张双全, 何禹德. 基于改进型模糊神经网络的信用卡客户违约预测[J]. 模糊系统与数学, 2017, 31(1): 143-148.

[8] Bahnsen, A.C., Aouada, D., Stojanovic, A., *et al*. (2016) Feature Engineering Strategies for Credit Card Fraud Detection. *Expert Systems with Applications*, **51**, 134-142. https://doi.org/10.1016/j.eswa.2015.12.030

[9] Xu, P., Ding, Z. and Pan, M.Q. (2017) An Improved Credit Card Users Default Prediction Model Based on Ripper. 13*th International Conference on Natural Computation*, *Fuzzy Systems and Knowledge Discovery*, Guilin, 29-31 July 2017, 1785-1789. https://doi.org/10.1109/FSKD.2017.8393037

[10] Breiman, L., Friedman, J.H., Olshen, R.A., *et al*. (1984) Classification and Regression Trees. Wadsworth, Belmont.

[11] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[12] Freund, Y., Schapire, R. and Abe, N. (1999) A Short Introduction to Boosting. *Journal—Japanese Society for Artificial Intelligence*, **14**, 1612.

[13] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203450