Finance 金融, 2019, 9(3), 137-141
Published Online May 2019 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2019.93015

Hans汉斯

Research on Credit Risk of P2P Lending Based on CatBoost Algorithm

Hongxiang Li, Hao Huang, Zixuan Zheng

University of International Business and Economics, Beijing
Email: k7huoyingmingren@qq.com

Received: Apr. 8th, 2019; accepted: Apr. 23rd, 2019; published: Apr. 30th, 2019

Abstract

Under the development of Internet finance and data mining technology, the use of machine learning algorithms in the traditional financial field and P2P platform field is of great significance to reduce the default risk of borrowers and ensure the good operation of financial industry and P2P platform. This paper uses the loan data of Australia P2P platform, compares with CatBoost algorithm and traditional machine learning algorithm, and uses AUC value and accuracy as evaluation standard. Empirical research shows that CatBoost algorithm is superior to traditional machine learning algorithm in credit scoring and can achieve better accuracy.

Keywords

CatBoost, Credit Risk, Machine Learning

基于CatBoost算法在P2P借贷信用风险的研究

李鸿祥，黄　浩，郑子旋

对外经济贸易大学，北京
Email: k7huoyingmingren@qq.com

收稿日期：2019年4月8日；录用日期：2019年4月23日；发布日期：2019年4月30日

摘　要

在互联网金融和数据挖掘技术的发展下，运用机器学习算法在传统金融领域和P2P平台领域，对降低借款人的违约风险，保证金融行业与P2P平台良好运营具有重要意义。本文利用澳大利亚P2P平台Ratesetter官网上的贷款数据，通过CatBoost算法与传统机器学习算法作比较分析，以AUC值和准确率

文章引用: 李鸿祥, 黄浩, 郑子旋. 基于 CatBoost 算法在 P2P 借贷信用风险的研究[J]. 金融, 2019, 9(3): 137-141.
DOI: 10.12677/fin.2019.93015

作为评价标准，实证研究显示CatBoost算法在对信用评分中优于传统机器学习算法，能够达到更好的准确性。

关键词

CatBoost，信用评分，机器学习

1. 引言

在经济下行，地方债务高筑，公司债务违约频发，P2P 平台暴雷，金融风险加剧的背景下，党的十九大提出了三大攻坚战，第一大攻坚战防范系统性金融风险。如何控制借款人违约风险，降低贷款逾期，对传统金融企业和新兴 P2P 平台的持续健康运营，对整个金融体系的健全和稳定具有重要意义。在互联网金融和大数据时代下，运用最新的机器学习算法，对现有金融领域的信用评价和贷款逾期风险问题的控制，既能保护金融机构和 P2P 平台的持续运营，也能保护投资者的合法权益。近年来，各方学者运用机器学习算法如逻辑回归，BP 神经网络，支持向量机，随机森林，梯度提升(GBDT)等算法对个人信用风险研究作了许多工作。秦宛顺，石庆焱(2003) [1]运用 Logistic 回归(lr)对个人信用风险进行评估。刘云焘，吴冲，王敏，乔木(2005) [2]运用支持向量机(SVM)用于商业银行信用风险评估。张道宏，张璇，尹成果(2006) [3]将 BP 神经网络算法应用在个人信用评估上。杨海江，魏秋萍，张景肖(2011) [4]基于改进的 AdaBoost 算法构建的信用评分模型。蒋翠清，王睿雅，丁勇(2017) [5]运用随机森林算法(rfs)对 P2P 平台信用逾期风险进行预测。谭中明，谢坤，彭耀鹏(2018) [6]基于梯度提升决策树(GBDT)对 P2P 网贷借款人进行信用风险评测研究。本文基于最新的机器学习 CatBoost 算法构建信用风险模型。

2. 模型介绍

2017 由 Yandex 公司[7]推出的 CatBoost 算法是一种能够处理分类数据的梯度提升(GBDT)算法。CatBoost 运用了一种有效的方式来将分类型数据转化成数值型数据并且防止过拟合[8]。CatBoost 在执行随机排列后能有效地处理分类数据，通过使用多个排列来训练不同模型来防止过度拟合，进而获得对梯度的无偏估计，以减轻梯度估计偏差的影响，提高模型的鲁棒性。这个处理分类数据主要通过以下三步完成：

1) 将初始数据进行随机排列，产生多个随机排列。
2) 将具有浮点或类别的标签值转换为整数。
3) 通过下面的公式将分类变量转换成数值型变量：

$$\text{avg\_target} = \frac{\text{countInClass} + \text{prior}}{\text{totalCount} + 1}$$

其中 countInClass 是具有当前分类特征值的对象标签为 1 出现的次数，totalCount 是具有与当前值匹配的分类特征值的对象总数，prior 是分子的初始值。

Caboost 算法的优点：1) 能够自动处理分类型数据，无须进行 one-hot 编码。2) CatBoost 性能优秀，默认参数也能达到良好的分类效果。CatBoost 主要通过以下主要参数如表 1 所示。

Table 1. List of main parameters
表 1. 主要参数表

参数名称	参数含义
learning_rate	学习率
depth	树的深度
l2_leaf_reg	正则化系数
loss_function	损失函数
one_hot_max_size	对于某些变量进行one-hot编码
leaf_estimation_method	迭代求解的方法，梯度和牛顿

3. 实证研究

3.1. 数据描述

本文数据来自于澳大利亚 P2P 平台 Ratesetter 官网数据。截止到 2018 年 9 月 30 日，澳大利亚 Ratersetter 官网上公布了 26,948 条贷款数据，其中未按时还款的有 1142，已完成还款的有 10,451，正在还款的有 15,391，本文从已完成还款的 10,451 条数据中选择 1200 和发生违约的 1132 条数据作为原始数据进行模型训练。澳大利亚 P2P 数据有 12 个特征变量，6 个为分类型数据，6 个为数值型数据。具体名称和变量数据类型如表 2 所示。

Table 2. Introduction to the characteristics of the P2P data set
表 2. P2P 数据介绍

变量名称	数据类型
利率	数值型
月期限	数值型
目的	分类数据
贷款金额	数值型
在途本金	数值型
工作状态	分类数据
收入	分类数据
是否提前还款	分类数据
年龄	数值型
住房状态	分类数据
现有信贷数目	数值型
工作职位	分类数据

3.2. 实验步骤

1) 原始数据中在途本金这一特征缺失严重将其删除，是否提前还款这一特征也选择了删除，一是该特征与是否发生违约相关度极高，一般发生了违约的借款人通常不存在提前还款的行为，存在提前还款的借款人通常不会违约，但以是否提前还款这一特征作为判断是否违约就已经能做出高达 90%的判断；二是在根据借款人个人基本信息判断该笔贷款是否会违约时，金融机构无法得知该笔贷款是否会提前还款这一信息。

2) 由于各个数值型变量的数字量级又差别，如利率与贷款金额相差几十万倍所以将原始数据中的数值型变量进行归一化处理，使其范围归一到[0,1]之间，对于特征变量 X 来说，以如下公式将 X’进行归一化：

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

3) 使用 k 折交叉检验来检测模型的准确性，防止模型过拟合。k 折交叉检验是将原始数据划分为 k 个均等份子集。在 k 个子集中选择 k−1 个子集作为训练集，剩下一个子集作为测试集。k 折交叉检验会对将每个子集都作为一次训练集，重复 k 次，本文进行 50 * 5 交叉检验，通过对原始数据随机划分 50 次，每次进行 5 折交叉检验，防止模型过拟合。使用 Sklearn 库中的支持向量机，逻辑回归，随机森林，GBDT 和 Yandex 开源的 CatBoost 进行实验。CatBoost 调整后的参数为：

learning_rate：0.01；depth：6；l2_leaf_reg：2；Interations:1000。

3.3. 结果评价

本文以 AUC 值和准确率作为模型评判标准。对于两分类问题，原始类为 positive、negative，分类后得到四个预测结果，真阳(TP)，伪阳(FP)，真阴(TN)，伪阴(FN)。ROC 空间将伪阳性率(FPR)定义为 X 轴，真阳性率(TPR)定义为 Y 轴。AUC 为 ROC 曲线下的面积，一般 AUC 均在 0.5 到 1 之间，AUC 越高，模型的区分能力越好。它们的计算公式如下：

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(FP + TN)$$

实验数据结果如表 3 所示，ROC 曲线如图 1 所示。从表 3 上可以看出，CatBoost 算法的 AUC 值、准确率都优于其他机器学习算法，说明 CatBoost 算法能够提高信用逾期预测准确率。CatBoost 默认参数已经能实现优秀的预测效果，通过对 CatBoost 算法参数的优化能够进一步提高 P2P 平台的信用逾期预测准确率。

Table 3. Comparison of modeling results
表 3. 模型结果对比

评估标准	支持向量机	逻辑回归	随机森林	GBDT	CatBoost	调参的CatBoost
AUC	0.8296	0.8741	0.8817	0.8871	0.8928	0.8941
准确率	0.7921	0.8115	0.8054	0.8141	0.8214	0.8233

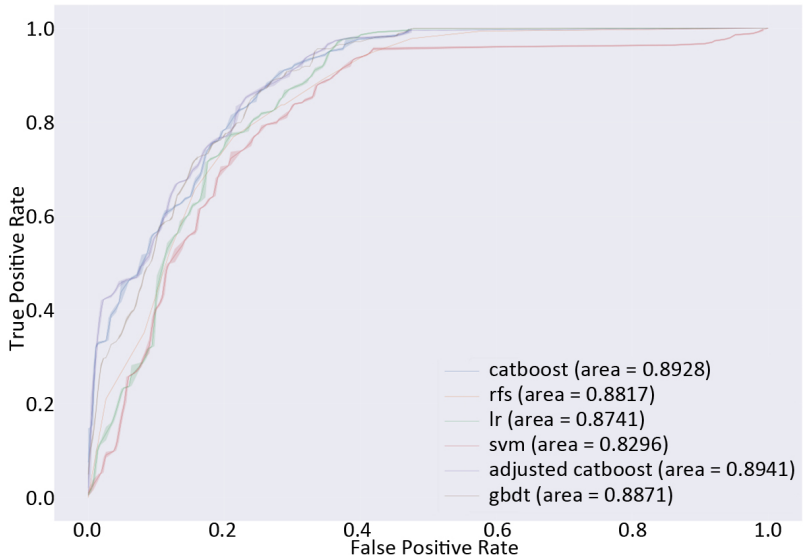

Figure 1. Comparison of ROC curve
图 1. Roc 曲线图对比

4. 总结

本文基于澳大利亚 P2P 平台 Ratestter 真实贷款数据，对数据进行数据清理，归一化。运用 CatBoost 算法和传统的机器学习算法包括逻辑回归，支持向量机，随机森林，梯度提升(GBDT)算法构建信用逾期模型。运用 Roc 曲线，AUC 值，准确率作为评价标准，进行 50 * 5 交叉检验使得实验结果更具有说服力，实验结果表明 CatBoost 算法有更高的 AUC 值和更高的准确性。运用机器学习 CatBoost 算法在贷款信用风险控制中，有一定的学术意义和时间价值，能够为金融行业的风险控制提供新的思路和方法。

本文的研究还能从以下几个方面予以补充：第一，针对样本数据正负比例严重失衡等问题，可采用过采样或通过 KNN 近邻等方式生成相似数据进行数据处理。第二，将其他机器学习算法与 CatBoost 算法融合起来构建信用逾期模型，提升预测准确度。第三，调整和优化 CatBoost 算法参数，例如用网格搜索等方式，使得模型参数更加拟合应用场景，达到更好的信用预测效果。

参考文献

[1] 秦宛顺, 石庆焱. 一个基于 Logistic 回归的个人信用评分模型[J]. 21 世纪数量经济学(第 4 卷), 2003.

[2] 刘云焘, 吴冲, 王敏, 等. 基于支持向量机的商业银行信用风险评估模型研究[J]. 预测, 2005, 24(1): 52-55.

[3] 张道宏, 张璇, 尹成果. 基于 BP 神经网络的个人信用评估模型[J]. 情报杂志, 2006, 25(3): 68-70.

[4] 杨海江, 魏秋萍, 张景肖. 基于改进的 AdaBoost 算法的信用评分模型[J]. 统计与信息论坛, 2011, 26(2): 27-31.

[5] 蒋翠清, 王睿雅, 丁勇. 融入软信息的 P2P 网络借贷违约预测方法[J]. 中国管理科学, 2017, 25(11): 12-21.

[6] 谭中明, 谢坤, 彭耀鹏. 基于梯度提升决策树模型的 P2P 网贷借款人信用风险评测研究[J]. 软科学, 2018(12): 136-140.

[7] Dorogush, A.V., Ershov, V. and Gulin, A. (2018) CatBoost: Gradient Boosting with Categorical Features Support. arXiv:1810.11363

[8] Nguyen, V.K., Zhang, W.E. and Sheng, Q.Z. (2018) Identifying Price Index Classes for Electricity Consumers via Dynamic Gradient Boosting. In: Hacid, H., Cellary, W., Wang, H., Paik, H.Y. and Zhou, R., Eds., *Web Information Systems Engineering WISE* 2018, *WISE* 2018. *Lecture Notes in Computer Science*, Vol. 11234. Springer, Cham, 472-486. https://doi.org/10.1007/978-3-030-02925-8_33