基于 LightGBM 算法的 P2P 项目信用评级模型的设计及应用①

马晓君　沙靖岚　牛雪琪

（东北财经大学统计学院）

研究目标：在大数据和互联网金融发展的背景下，依据个人信用，有效控制 P2P 项目的违约率以保证相关金融项目或平台的良好运营。**研究方法**：本文基于美国 P2P 平台 Lending Club 的海量真实交易数据，采用“多观测”与“多维度”两种数据清洗方式，运用 2016 年微软亚洲研究院提出的机器学习算法 LightGBM，兼顾权威性和创新性地对平台内贷款项目的违约风险进行预测，并对不同数据清洗方法的结果进行比较。**研究发现**：基于多观测的 LightGBM 算法的预测结果最好，比 Lending Club 平台历史交易数据算的平均履约率提升了 1.28 个百分点，可减少约 1.17 亿美元的违约借款。**研究创新**：运用不同的数据清洗方式和较为前沿的机器学习算法（LightGBM）预测违约率。**研究价值**：在 LightGBM 算法得出违约率影响因素的结果基础上，可以明确 Lending Club 及广大 P2P 平台的改进内容和各国在该领域内发展完善的方向。

关键词　P2P　信用评级　违约率控制　数据清洗　LightGBM 算法

中图分类号　C812　　**文献标识码**　A

DOI:10.13653/j.cnki.jqte.20180503.001

一、问题的提出

伴随着互联网技术的发展，以及大数据时代的到来，互联网金融得到了快速发展并得到国家高度的重视，做好信用评级相关工作，则是保证互联网金融业健康发展的关键一环。

P2P 是以“普惠”为核心思想的互联金融的典型代表，它利用网络实现了投资者和借款人的直接对接，并使双方都获得收益。Robert 和 Benjamin（2010）发现利用综合的信用系统根据借款人的个人名誉、信用评分等个人信息，可以筛选出高质量的借款人。Iyer 等（2016）研究发现，P2P 借贷市场中的出借人具有很强的信用甄别能力。但 P2P 行业的快速发展，也伴随着层出不穷的问题严重阻碍其发展。P2P 网络平台存在本地偏好（Lin 和 Viswanathan，2016）和羊群效应（Herzenstein 等，2011）。徐硕正（2016）称 P2P 行业虽已声名鹊起，却是毁誉参半，誉者赞其曰“金融创新”“普惠金融”，毁者斥之“非法集资”“庞氏骗局”。发展至今，我国 P2P 网贷行业出现了大规模平台跑路、倒闭潮的情况，为投资者带来了巨大的损失，违约率远高于国外 P2P 平台的违约率，而风险管理控制是降低违约率的关键所在，也是 P2P 行业发展的核心。

① 本文获得国家社科基金项目“高维数据下企业信用评级方法的改进与应用研究”（17BTJ020）、国家自然科学基金项目（71772113）、2017 年度辽宁省哲学社会科学规划基金项目（L17BTJ003）的资助。

当今世界的数据量呈爆炸型增长，P2P 行业也不例外，大数据的出现提供给分析者更多的信息，机器学习作为研究大数据挖掘的重要方法之一，其在各行各业的应用就显得尤为重要。对于 P2P 行业而言，运用机器学习方法，解决项目违约预测问题具有重要的意义。本文希望利用国外较为优秀的 P2P 平台 Lending Club 的真实交易数据，进行违约率和违约项目的预测，进而对我国 P2P 行业的发展起到借鉴作用，为国家宏观调控、把握政策导向提供素材。与此同时，本文是 LightGBM 机器学习方法在 P2P 风险管理控制领域应用的一种尝试，这样的尝试具有一定的学术意义和实践价值，能够为我国的 P2P 行业发展提供有力的技术支持，进而促进中国互联网金融业的长足发展。

二、现有算法的梳理

自 Zopa 在英国诞生以及 Prosper 公开交易数据以来，就吸引了一大批经济学、管理学、社会学和信息技术方向的研究人员致力于 P2P 网络借贷的研究，目前 P2P 网络借贷平台的学术研究正方兴未艾。

朱顺泉（2002）、刘铮铮（2006）、康为勋（2016）运用层次分析法对企业和银行的信用评级进行研究。程琛（2009）、王洪欣（2013）、常轶（2015）运用 BP 神经网络模型对借款人进行信用评价。吕晓丹和范宏（2013）认为决策树对企业的信用风险评估具有较高的准确性。Malekipirbazari 和 Aksakalli（2015）分别运用 K 均值聚类法、支持向量机和随机森林的方法对社会贷款进行评估，结果显示随机森林的效果优于其他方法。Li 等（2016）用 K 均值聚类法检测异常贷款人，研究发现检测出的异常贷款人多数信用得分都不高。

1. BP 神经网络

BP 神经网络（Back Propagation Neural Network，BPNN）是一种多层前馈式反向传播神经网络，常用于分类或回归预测，可以进行误差反向传播学习，通常由输入层、隐含层、输出层三层构成。一个三层 BPNN 的拓扑结构如图 1 所示：

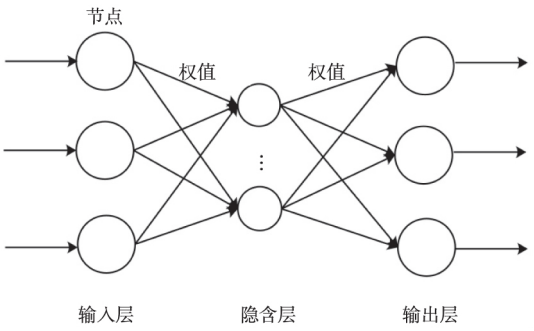

图 1 BP 神经网络结构图

输入层负责输入已有的相关信息，由隐含层的神经元接收来自输入层的信息并且进行处理，最后由输出层对隐含层神经元提供的信息进行信号反馈，得出分类或预测结果。利用误差的反向传播，从而不断调节不同神经层间的阈值、权值，降低预测误差。在这种循环的反向传播中，不断优化模型，直到满足一定退出条件，使得整个神经网络具有较好的分类预测效果。神经网络在预测前首先要训练网络，通过训练使网络具有联想记忆和预测

能力。

BP 神经网络分类的准确度高，可以进行并行分布处理，且分布存储及学习能力强，对噪声神经有较强的鲁棒性和容错能力，充分逼近复杂的非线性关系，具备联想记忆的功能。但是神经网络需要大量的参数，如网络拓扑结构、权值和阈值的初始值，这使得在方法的使用上具有较大的操作难度；同时该方法不能观察内部的学习过程，输出结果难以解释，会影响到结果的可信度和可接受程度；更重要的是，学习时间过长，甚至可能达不到学习的目的。

2. K 均值聚类法

1967 年 MacQueen 首次提出了 K 均值聚类算法（K-means 算法），用来处理数据聚类的相关问题，该算法由于算法简便，广泛地运用于科学和工业领域。

K 均值聚类算法是先随机选取 K 个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。一旦全部对象都被分配了，每个聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。

K 均值聚类算法的描述如下：

步骤 1：任意选择 K 个记录作为初始的聚类中心。

步骤 2：计算每个记录与 K 个聚类中心的距离，并将距离最近的聚类作为该点所属的类。

步骤 3：计算每个聚类的聚集点的均值以及每个对象与折现中心的距离，并根据最小距离重新对相应的对象进行划分。重复该步骤，直到步骤 1 不再有明显的变化。

K 均值聚类法虽然可以实现聚类的平方误差最小且分类明显，但在实践中存在很多问题，例如，算法中选取的 K 个聚类中心的选择十分困难；人为进行的初始划分具有很强主观性，对结果的准确性影响较大；不断地调整中心将会增加计算量，不适合在海量数据的项目中使用。

3. 支持向量机

支持向量机（Support Vector Machine，简称 SVM）是由 Vapnik 带领的 AT&Bell 实验室研究小组于 1995 年首次提出的一种全新的具潜力的机器学习算法，初期的目的在于解决模式识别和分类问题。其核心思想就是将输入变量映射到一个高维特征空间中，然后构造一个能够在原空间中产生非线性类边界的线性模型从而实现对输入数据分类，SVM 即用来寻找该线性模型——最优超平面。

支持向量机采用非线性映射的方法，将特征空间进行分类并最大化分类边际，可以解决小样本情况下的高维、非线性机器学习问题，并且提高泛化性能，避免神经网络结构选择和局部极小点问题。但是由于 SVM 是借助二次规划来求解支持向量，而求解二次规划将涉及 m 阶矩阵的计算（m 为样本的个数），当 m 数目很大时该矩阵的存储和计算将耗费大量的机器内存和运算时间，加之该方法对缺失数据及其敏感，对非线性问题没有通用的解决方案，并不适合在海量数据的项目中使用。

4. 随机森林

随机森林是由美国科学家 Leo Breiman 将其在 1996 年提出的 Bagging 集成学习理论与 Ho 在 1998 年提出的随机子空间方法相结合，于 2001 年发表的一种机器学习算法。随机森林是以决策树为基本分类器的一个集成学习模型，它包含多个由 Bagging 集成学习技术训练得到的决策树，当输入待分类的样本时，最终的分类结果由单个决策树的输出结果投票

决定。

随机森林是以 K 个决策树 $\{h(X,\theta_k),k=1,2,\cdots,K\}$ 为基本分类器，进行集成学习后得到的一个组合分类器。当输入待分类样本时，随机森林输出的分类结果由每个决策树的分类结果简单投票决定。这里的 $\{\theta_k,k=1,2,\cdots,K\}$ 是一个随机变量序列，由随机森林的两大随机化思想决定的：

（1）Bagging 思想。从原样本集 X 中有放回地随机抽取 K 个与原样本集同样大小的训练样本集 $\{T_k,k=1,2,\cdots,K\}$（每次约有 37%的样本未被抽中），每个训练样本集构造一个对应的决策树。

（2）特征子空间思想。在对决策树每个节点进行分裂时，从全部属性中等概率随机抽取一个属性子集通常取 $[\log_2(M)]+1$ 个属性，为特征总数，再从这个子集中选择一个最优属性来分裂节点。由于构建每个决策树时，随机抽取训练样本集和属性子集的过程都是独立的，且总体都是一样的，因此 $\{\theta_k,k=1,2,\cdots,K\}$ 是一个独立同分布的随机变量序列。训练随机森林的过程就是训练各个决策树的过程，由于各个决策树的训练是相互独立的，因此随机森林的训练可以通过并行处理来实现，这将大大提高生成模型的效率。随机森林中每个决策树都按照 Bootstrap 方法抽样后寻找最佳分裂变量和阈值的流程构造决策树。将以同样的方式训练得到 K 个决策树组合起来，就可以得到一个随机森林。当输入待分类的样本时，随机森林输出的分类结果由每个决策树的输出结果进行简单投票（取众数）决定。

随机森林克服了决策树过拟合问题，对部分噪声和异常值有较好的容忍性，对高维数据分类问题具有良好的可扩展性和并行性。此外，随机森林是由数据驱动的一种非参数分类方法，只需通过对给定样本的学习训练分类规则，并不需要分类的先验知识。该方法的缺陷在于：单棵决策树的预测效果很差，且对于有不同级别、属性的数据，级别划分较多的属性会对随机森林产生很大的影响，精确度无法保证。

在前人研究的基础上，本文运用来自美国 Lending Club 的全部真实交易数据，引入当前最新的 LightGBM 算法，进行信用评级的定量研究，该算法在一定程度上缓解了之前算法的不足。最后的实证研究结果显示，LightGBM 算法对违约项目的预测具有较高的准确性，适合广泛运用于中国的 P2P 领域的风险控制。

三、模型设计

LightGBM（Light Gradient Boosting Machine）是 2016 年微软亚洲研究院公布的一个开源、快速、高效的基于决策树算法的提升（GBDT、GBRT、GBM 和 MART）框架，被用于排序、分类、回归等多种机器学习的任务，支持高效率的并行训练。

1. LightGBM 的相关理论基础

（1）Gradient Boosting。Boosting 是用一系列子模型的线性组合来完成学习任务的，它分为两种类型：AdaBoost 和 Gradient Boosting，LightGBM 属于 Gradient Boosting 的一种。

Gradient Boosting 的思想是：一次性迭代变量，迭代过程中，逐一增加子模型，并且保证损失函数不断减小。

假设 $f_i(X)$ 为子模型，复合模型为：

$$F_m(X)=\partial_0 f_0(X)+\partial_1 f_1(X)+\cdots+\partial_m f_m(X) \tag{1}$$

损失函数为 $L[F_m(X), Y]$，每一次加入新的子模型后，使得损失函数不断朝着信息含量次高的变量的梯度减小：

$$L[F_m(X),Y] < L[F_{m-1}(X),Y] \tag{2}$$

（2）决策树。决策树（Decision Tree）是一种分类和回归的方法，实际研究中大多用于分类。决策树的结构呈树形结构，大多运用的是二权树，在每一个叶子节点上，根据某一判断条件，输出“符合条件”和“不符合条件”两类，不断重复向下输出，如图 2。可以把决策树理解成众多 if-then 规则的集合，也可以认为是定义在特定空间与类空间上的条件概率分布。决策树的创建包括 3 个主要步骤：特征选择、决策树的生成和决策树的修剪，该方法具有可读性高、分类速度快的优点。

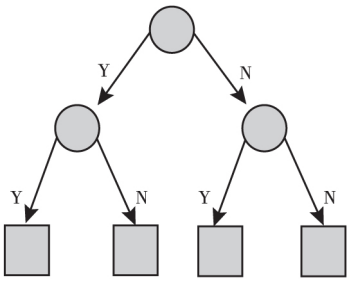

图 2 决策树结构

决策树的分裂方法分为两类，一类是按叶子分裂的学习方法（Leaf-wise Learning）；另一类是按层分裂的学习方法（Level-wise Learning）。

按叶子分裂的学习方法是指在分裂的过程中要不断地寻找分裂后收益最大的节点，对其进行进一步的分裂，其他非收益最大化的结点不再继续分裂，以这样的规则生长这棵树。这样做的优点是可以使算法更加快速有效；缺点是会忽略掉那些被舍弃的叶子上的信息，导致分裂结果不够细化。图 3 描述的就是按叶子分裂的过程。

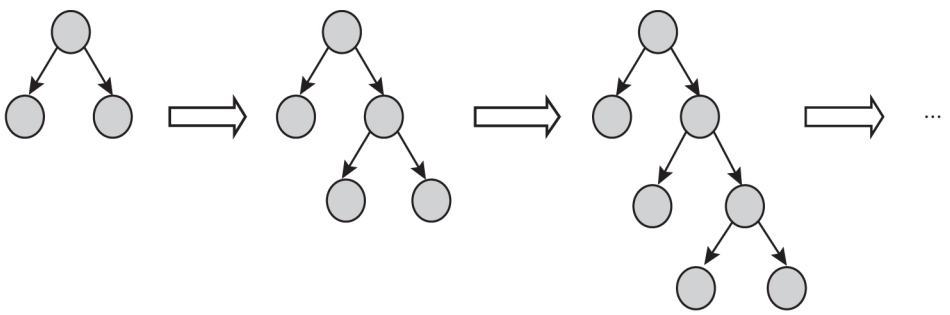

图 3 按叶子分裂的决策树学习过程

按层分裂的学习方法与按叶子分裂的学习方法不同，如图 4，它不需要挑选收益最大化的节点，每一层的每一个结点都要进行分裂，也就是说每次迭代都要遍历整个训练数据的所有数据。优点是每一层的叶子可以并行完成，具有天然的并行性；缺点是这样会产生很多没有必要的分裂，需要更多的计算成本，同时，也会占用较大的运行内存。

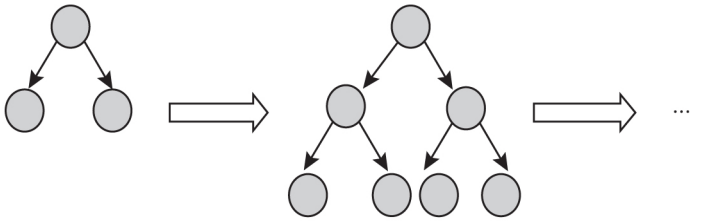

图 4 按层分裂的决策树学习过程

（3）GBDT。GBDT（Gradient Boosting Decision Tree）是机器学习中一个长盛不衰的模型，事实上：

GBDT＝Gradient Boosting＋Decision Tree

即若 Gradient Boosting 中的每一个子模型都是一个 Decision Tree，这样的模型就是 GBDT。

GBDT 拥有着 Gradient Boosting 和 Decision Tree 的功能共同特性，具有训练效果好、不易过拟合等优点。GBDT 的工具主要包括 XGBoost、Pgbrt、Sklearn、R. GBM 等。GBDT 在工业界应用广泛，通常被用于点击率预测，搜索排序等任务。GBDT 也是各种数据挖掘竞赛的致命武器，据统计 Kaggle 上的比赛有一半以上的冠军方案都是基于 GBDT。

2. LightGBM 应用

LightGBM 是 GBDT 的一种，被提出的主要原因是为了解决 GBDT 在海量数据遇到的问题，让 GBDT 可以更好更快地用于实践。

LightGBM 中的决策树子模型是采用按叶子分裂的方法分裂节点的，因此它的计算代价比较小，也正是因为选择了这种分裂方式，需要控制树的深度和每个叶子节点的最小数据量，从而避免过拟合现象的发生。LightGBM 选择了基于 Histogram 的决策树算法，将特征值分为很多个小“桶”，进而在这些“桶”上寻找分裂，这样可以减小储存成本和计算成本。另外，类别特征的处理，也使得 LightGBM 在特定数据下有比较好的提升。

LightGBM 分为三类：特征并行、数据并行和投票并行。特征并行运用在特征较多的场景，数据并行应用在数据量较大的场景，投票并行应用在特征和投票都比较多的场景。

LightGBM 通过以下几个主要的参数实现算法控制与优化：

num _leaves：每棵数的叶子数量

learning _rate：学习率

max _depth：最大学习深度。限制树模型的最大深度，用于控制过拟合现象。当 max _depth＜0 时，没有学习深度的限制

min _data：一片叶子中数据的最小数量，可以用来控制过拟合现象

feature _fraction：选择特征占总特征数的比例，取值在 0 到 1 之间。当 feature _fraction＜0 时，LightGBM 在每一次迭代时会随机选择部分特征，feature _fraction 用于控制选择总特征数的比例。该参数可以用于加快训练速度，并且控制过拟合现象

bagging _fraction：选择数据占总数据量的比例，取值在 0 到 1 之间。与 feature _fraction 类似，但是随机并且不重复选择的是相应比例的观测，注意要将其设置成大于 0 的比例。该参数可以用于加快训练速度，并且辅助控制过拟合现象

LightGBM 算法自 2016 年发布以来，已经广泛运用于大数据机器学习领域，与之前的

XGBoost 并称为当今机器学习的“倚天屠龙”。公开数据的实验表明 LightGBM 能在学习效率和准确率上都表现出比其他已有 Boosting 工具更好的表现，相比 XGBoost 速度更快，内存占用更少，准确率更高。此外，实验也表明 LightGBM 通过使用多台机器进行特定设定的训练，它能取得线性加速效果。因此，总结该算法的优点显著体现在如下五个方面：①更快的训练速度；②更低的内存消耗；③更好的模型精度；④支持并行学习；⑤可以快速处理海量数据。将性能优良的 LightGBM 算法运用于信用评级系统，其可靠性和灵活性将大大促进相关领域的长足发展。

四、模型检验

为减少 P2P 平台的违约率，更好地辨别借款人已公开的信息，本部分基于 LightGBM 的方法，对 Lending Club 平台已完成的借贷数据进行分类与分析，以便于未来预测贷款人是否会发生违约行为。

1. 数据的获取

美国的 Lending Club 是美国最大的 P2P 平台，它为全球的 P2P 出借者和借款者提供了一个良好的平台，基本上占据了全球 P2P 中介行业的大半壁江山。自 2007 年成立以来，每一条交易数据都完整的保存在平台的数据库中并透明公开，所有数据未解压前能达到近百兆的体量。鉴于该平台在行业内的巨大份额和数据的良好获得性，选择 Lending club 的交易数据进行研究具有较强的权威性和实际意义。

本部分使用数据均来自 Lending Club 平台自 2007 年成立以来截至 2017 年 6 月 31 日的所有贷款数据，来源于 Lending Club 官方网站（https：//www. lendingclub. com/info/download-data. action），不再详细标出，若非 Lending Club 官方网站数据，笔者会单独标出。

2. 数据的结构

数据包括两部分：一部分（Loan Data）是 Lending Club 平台通过筛选而预测出的违约率比较低的，从而被允许在平台上发布借款信息，并且得到了投资人投资的所有借款项目相关数据，这些项目包括了违约项目和未违约项目，从 2007 年到 2017 年第二季度，共有 2276819 条观测数据，每条观测涉及 137 个变量。另一部分（Decline Loan Data）是未通过 Lending Club 平台筛选从而没有资格进行借款的被拒绝项目数据，共 8505916 条观测，变量数为 9。

由观测总量数据可知，截至 2017 年上半年，Lending Club 平台贷款申请的平均通过率约为 21. 12%（被拒绝项目数据变量中没有年份，因此无法具体到各年的数据，只能粗略计算总体的申请通过率）。

3. 描述性数据统计分析

Lending Club 目前是美国交易额最大的 P2P 平台。据南方财富网统计，Lending Club 占据美国 P2P 市场 65%的市场份额。

表 1 是 Lending Club 截至 2016 年 6 月 30 日的交易额、交易数量和单均交易额，可以看出除了 2017 年，交易额、交易数量和单均交易额均呈逐年增长的趋势。从图 5 可以看出，除 2016 年交易量的增长率很低，其他年份的交易额和交易量的增长率都在 70%以上。由表 1 可以看出，自 2016 年开始，交易量的增长率呈递减趋势，原因是 LendingClub 创始人雷诺德·拉普兰奇（Renaud Laplanche）由于信息披露丑闻，以及涉嫌利益冲突而被迫离职，严重影响了 LendingClub 的交易量。

表 1 2007～2017 年交易额和交易数量

年 份	年交易额（美元）	年交易数量（笔）	单均交易额（美元/笔）
2007	2152175	251	8574.4
2008	13457075	1562	8615.3
2009	46324425	4716	9822.8
2010	116706400	11536	10116.7
2011	257363650	21721	11848.6
2012	717942625	53367	13452.9
2013	1982759550	134814	14707.4
2014	3503840175	235629	14870.2
2015	6417605175	421095	15240.3
2016	25185441700	434407	57976.6
2017（Q1 & Q2）	2976401550	202230	14717.9
总计	174291924663	13882722	12554.6

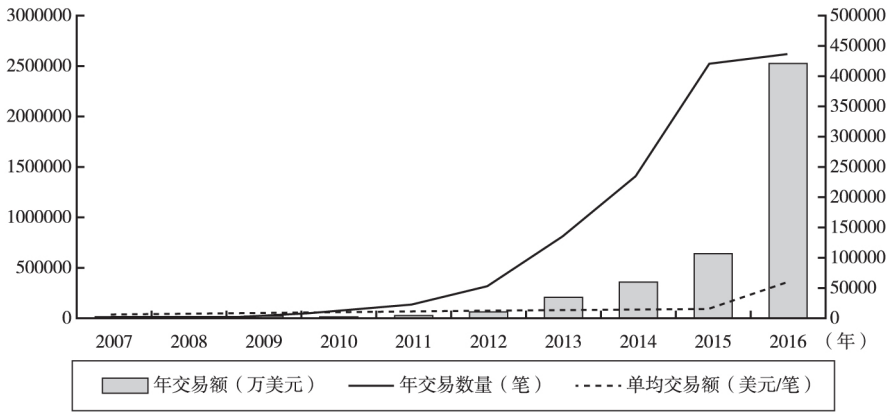

图 5 各季度订单额分布

Lending Club 借款期限有两种贷款期限，分别是 36 个月和 60 个月，从表 2 可以看出有 75.25%的人选择 36 个月期限的贷款，24.75%的借款者选择 60 个月期限的贷款。

表 2 借款期限分布比例

贷款期限	总贷款数量	占总贷款比例
36 个月	531782	75.25%
60 个月	174938	24.75%
总计	706720	100%

根据贷款数据得知，Lending Club 平台借款者的借款利率是由借款者的贷款等级（A～G）来决定的，从 6%到 24%不等，借款人和贷款人没有资格决定利率的大小，每个贷款等级还分五个子等级（1～5），因此所有通过审核的借款人被分成 35 个等级，对应着 35 种利率。

Lending Club 平台的借款利率随着贷款等级的降低而增加，利率根据贷款人的等级来决定，避免了因借款人和投资人决定的利率不合理而导致的违约风险。

图 6 和图 7 分别展示了 36 个月期限贷款和 60 个月期限贷款各等级的占比。可以看出 36 个月期限的贷款等级整体要比 60 个月期限贷款等级高很多。36 个月期限的贷款中，B 等级最多，A、B 等级占了总贷款的 50%以上，几乎没有 G、F 等级的贷款，随着年份的增加，各等级所占比例没有明显变化；60 个月期限贷款中，A 等级的比例极低，C、D 等级居多，随着年份的增加，D 等级贷款明显大幅增加，A 等级略有减少，G、E、F 等级明显减少，其他等级没有明显变化。

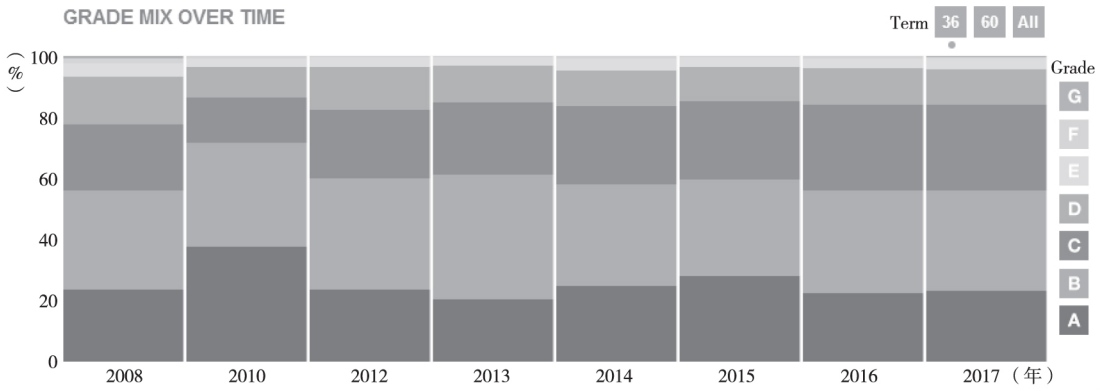

图 6　36 个月期限贷款各等级占比

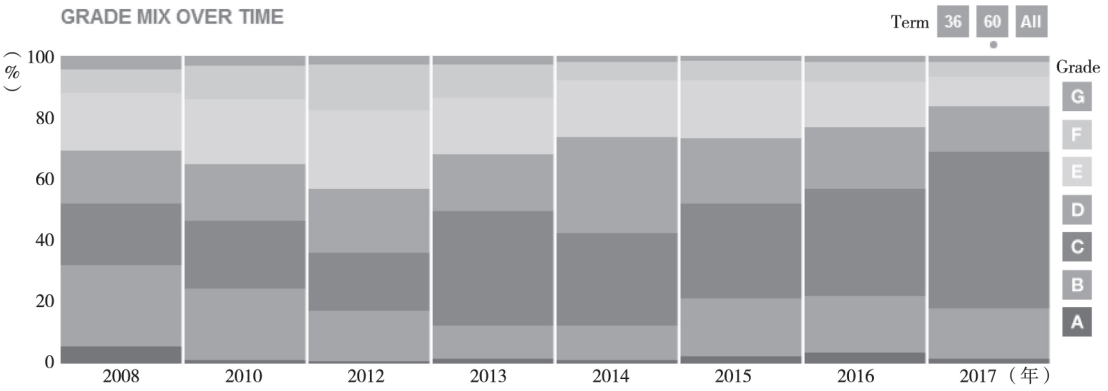

图 7　60 个月期限贷款各等级占比

图 8 是汇总两种期限的所有贷款得到的各等级占比，D 等级的贷款占比逐年增加，G、F、E 等级的贷款比例明显逐年下降。

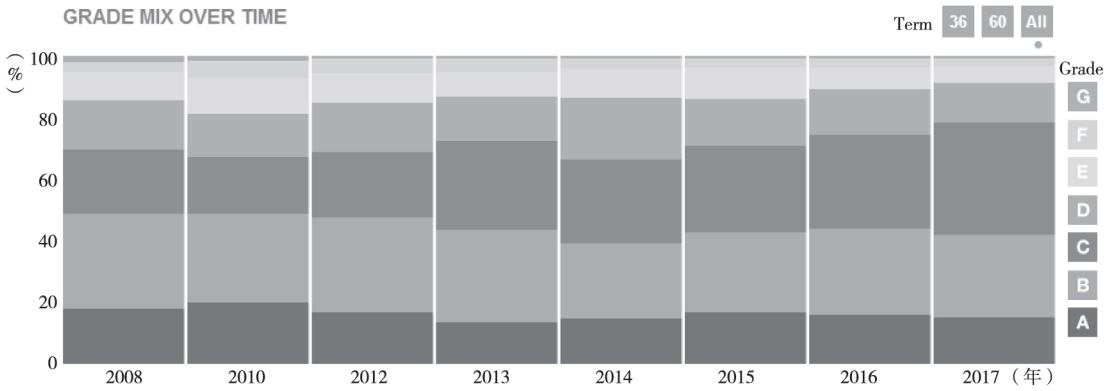

图 8 所有贷款各等级占比

根据 Lending Club 网站数据分析平台的违约率，其中贷款状态分为 8 种。

表 3 贷款状态及含义

Loan-status	含 义
Fully Paid	已还清
Current	进行中
Charge Off	核销
Default	拖欠
In Grace Period	宽限期
Issued	发出
Late（16～30days）	延期 16～30 天
Late（31～120days）	延期 31～120 天

只有 Fully Paid 和 Charge Off 是已经一笔贷款的最终状态，其他 6 种状态都是贷款没有结束而生成的中间状态，贷款到期后都会归为 Fully Paid 或 Charge Off 状态。

图 9 是按贷款等级计算的违约率，可以明显地看出随着贷款等级的下降（从 A 到 G），违约率呈明显的上升趋势，但由于贷款等级越低，所占比例越低，所以整体的违约率在 B、C 量等级的违约率中间，为 21.18%。

表 4 是按照贷款期限计算的违约率，可以明显地看出：贷款期限为“36 个月”的贷款违约率要明显低于贷款期限为“60 个月”的贷款，“60 个月”的贷款的违约率约为“36 个月”贷款的 2 倍。

表 4 按照贷款期限计算的违约率

贷款期限	违约率
36 个月	15.50%
60 个月	32.19%
总计	21.18%

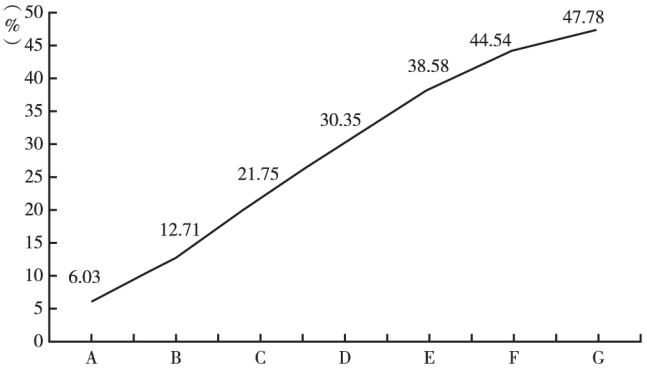

图 9 按贷款等级计算的违约率

根据图 10 从年份来看，违约率总体上呈现出先下降的趋势。2007 年和 2008 年的率很高，一方面是因为 Lending Club 刚成立，风险控制体系还有待更加完善；另一方面，金融危机起到了很大的助推作用。2012 以前的贷款均已到期，2012 年及以后的贷款有一部分是没有到达贷款期限的，这也是违约率逐年降低的原因之一。

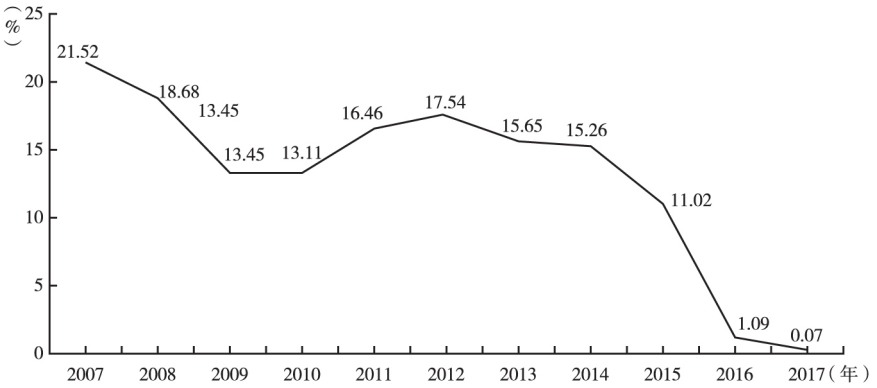

图 10 按年份计算的违约率

4. 基于 LightGBM 算法的二分类模型

(1) 数据的清洗。本文的目的在于在平台原来的基础上降低违约率，即增加违约贷款的识别度，因此，本模型只需要对已通过平台预测的贷款项目进行分析和学习，即上文提到的数据——Loan Data。

该部分数据包括 Lending Club 平台 2007～2017 年第 2 季度的所有借贷数据，共 1521328 条观测，137 个变量。在 Loan_statu 是本研究的被解释变量，其中只有 Fully Paid 和 Charge Off 是贷款到期后确定该贷款是否违约的状态，即：完全还款和坏账状态，变量 loan_statu 中的其他状态都是贷款没到期时生成的中间状态。本节只对到期后的贷款结果进行研究，因此，只保留 Fully Paid 和 Charge Off 的贷款项目，其他的项目均删除。剩余数据情况如表 5。

表 5 筛选贷款状态后数据情况

观测状态	数 量	比 例
Fully Paid	525869	79.30%
Charge Off	137301	20.70%
总 计	663170	100%

由于变量和缺失值都过多，本研究对数据进行两次不同角度的清洗，第一种是本着尽可能保留变量的角度进行清洗；第二种是本着尽可能保留更多观测的角度进行清洗。希望通过比较两种清洗方式的结果，得出更好的数据清洗方式与结果。多变量清洗方式：

①删除以下变量：缺失值达 50%以上的 31 个变量；22 个空变量；因为我们的目的是根据贷款前的借款者信息来预测借款人是否会违约，所以贷款过程中产生的变量不在我们的研究范围内，删去 23 个此类的无关变量，剩余 61 个变量。

②为了研究方便，对以下分类变量进行数值替换（“－”前为替换前内容，“－”后卫替换后数值）。

③删除变量值缺失过多的观测，进行最后的数据清洗与整理，最终以 565227 条观测、61 个变量的数据集“one”进行分析。

表 6 替换变量结果

变 量	替换内容
addr_status	*AK*－1、*AL*－2、*AR*－3、*AZ*－4、*CA*－5、*CO*－6、*CT*－7、*DC*－8、*DE*－9、*FL*－10、*GA*－11、*HI*－12、*IA*－13、*ID*－14、*IL*－15、*IN*－16、*KS*－17、*KY*－18、*LA*－19、*MA*－20、*MD*－21、*ME*－22、*MI*－23、*MN*－24、*MO*－25、*MS*－26、*MT*－27、*NC*－28、*ND*－29、*NE*－30、*NH*－31、*NJ*－32、*NM*－33、*NV*－34、*NY*－35、*OH*－36、*OK*－37、*OR*－38、*PA*－39、*RI*－40、*SC*－41、*SD*－42、*TN*－43、*TX*－44、*UT*－45、*VA*－46、*VT*－47、*WA*－48、*WI*－49、*WV*－50、*WY*－51
application_type	*INDIVIDUAL*－1、*JOINT*－2、*DIRECT*－3
emp_length	＜1 *year*－0.5、1 *year*－1、2*years*－2、3*years*－3、4*years*－4、5*years*－5、6*years*－6、7*years*－7、8*years*－8、9*years*－9、10＋*years*－10
grade	*A*－1、*B*－2、*C*－3、*D*－4、*E*－5、*F*－6、*G*－7
home_ownership	*MORTGAGE*－1、*RENT*－2、*OWN*－3、*ANY*－4、*NONE*－5、*OTHER*－6
initial_list_status	*w*－1、*f*－2
Loan_status	*Fully Paid*－1、*Charge Off*－2
purpose	*car*－1、*credit_card*－2、*debt_consolidation*－3、*educational*－4、*home_improvement*－5、*house*－6、*major_purchase*－7、*medical*－8、*moving*－9、*other*－10、*renewable_energy*－11、*small_business*－12、*vacation*－13、*wedding*－14
sub_grade	*A1*－1、*A2*－2、*A3*－3、*A4*－4、*A5*－5、*B1*－6、*B2*－7、*B3*－8、*B4*－9、*B5*－10、*C1*－11、*C2*－12、*C3*－13、*C4*－14、*C5*－15、*D1*－16、*D2*－17、*D3*－18、*D4*－19、*D5*－20、*E1*－21、*E2*－22、*E3*－23、*E4*－24、*E5*－25、*F1*－26、*F2*－27、*F3*－28、*F4*－29、*F5*－30、*G1*－31、*G2*－32、*G3*－33、*G4*－34、*G5*－35
term	36 *month*－1、60 *month*－2
verification_status	*Source Verified*－1、*Verified*－2、*Not Verified*－3

多观测清洗方式：

①在多变量清洗方式步骤②的基础上继续删除与被解释变量无关的变量，最终留下 24 个变量。

②删除有缺失值的观测。最终以 569339 条观测、24 个变量的数据集“two”进行分析。

（2）基于 LightGBM 的二分类模型。数据集 one 的训练：

参数设置为：

$num\_leaves=25$　　$learning\_rate=0.1$

$max\_depth=8$　　$min\_data=200$

$feature\_fraction=1$　　$bagging\_fraction=0.5$

模型结果：从图 11 可以看出当迭代次数为 67 时，测试集错误率和训练集错误率同步达到最小，训练集的误差率为 19.5538%，测试集的误差率为 19.94%，即测试集的准确率达到 80.06%。可以看出测试集和训练集的错误率与损失值一直在同步下降，因此没有发生过拟合现象，预测结果是有效。

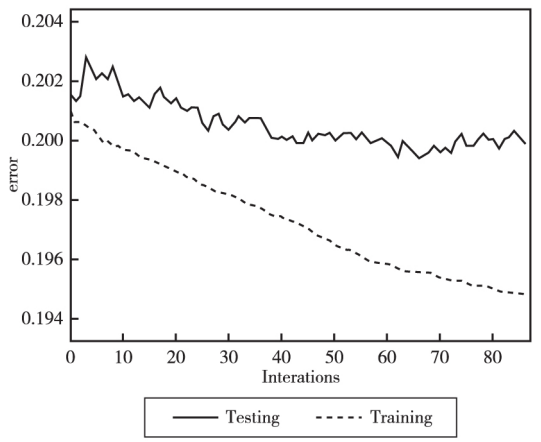

图 11　训练集和测试集的错误率

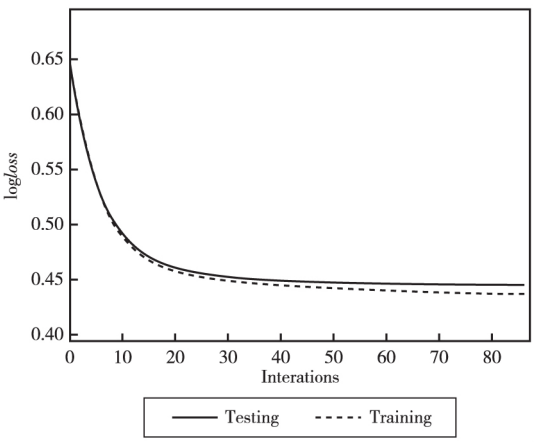

图 12　训练集和测试集的损失值

图 13 列出的是对模型贡献度最高的前十个解释变量。贡献度从高到低排序依次是：借款发生的时间、借款利率、债务收入比、总银行卡最高信用额度、借款金额、借款人所在州、年收入、借款期限、最高总信用额度、开设循环账户距今的月分数。

其中贡献率最高的是贷款发生的时间，达到 24%以上，借款利率、债务收入比的贡献率在 12%左右，其余几个解释变量的贡献率在 8%至 5%之间。

数据集 two 的训练：

模型参数为：

$num\_leaves=18$，　　$learning\_rate=0.1$，

$max\_depth=6$，　　$min\_data=200$，

$feature\_fraction=1$，　　$bagging\_fraction=0.5$

模型结果：从图 14 可以看出当迭代次数为 130 时，测试集错误率和训练集的错误率同步达到最小，训练集的错误率为 19.382%，测试集的错误率为 19.9%，即测试集的准确率达到 80.1%。测试集和训练集的错误率和算是函数同步下降，因此没有发生过拟合现象。

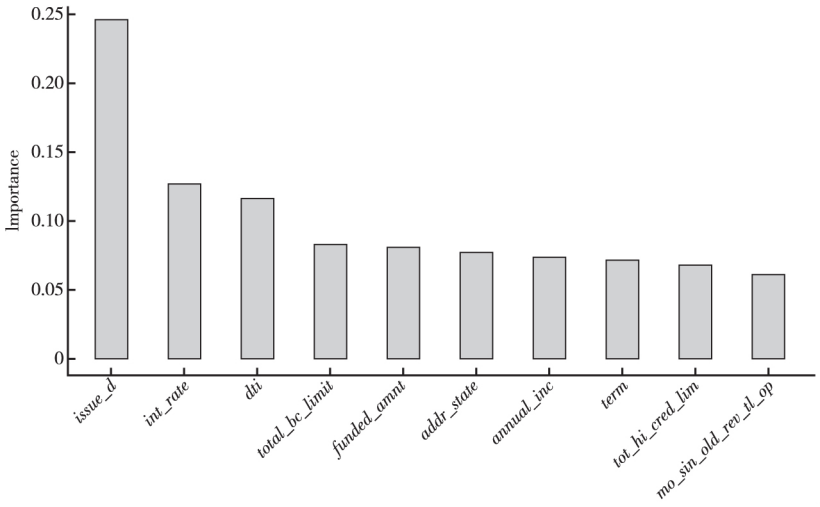

图 13　贡献率最高的前十个解释变量

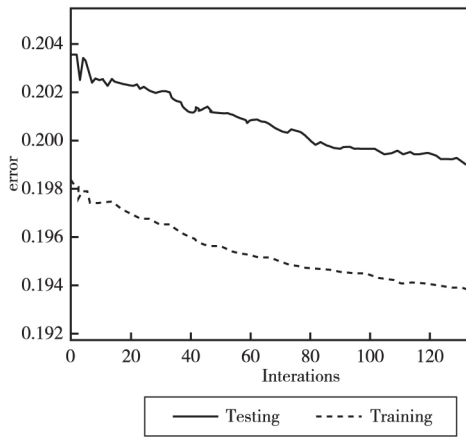

图 14　训练集和测试集的错误率——two 数据集

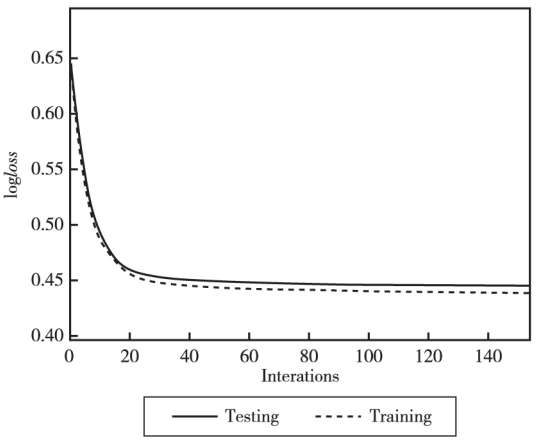

图 15　训练集和测试集的损失值——two 数据集

图 16 列出的是对模型贡献度最高的前十个解释变量。贡献度从高到低排序依次是：借款发生的时间、借款人所在州、循环信用额度、借款利率、年收入、循环额度利用率、贷款子等级、债务收入比、借款金额、每月应还款金额、开通银行卡数量。

其中贡献率最高的也是贷款发生的时间，达到 21%左右，借款人所在州贡献为 16%，循环信用额度的贡献率为 11%，其余几个解释变量的贡献率在 5%～10%。

（3）数据集损益对比分析。截至 2017 年 6 月 31 日，Lending Club 已完成项目 663170 个，共计约 91.11 亿美元。其中，未违约项目 522711 个，占完成总项目数的 78.82%；违约项目 140459 个，占完成总项目数的 21.18%。

基于多观测数据集的 LightGBM 算法具有更高的精度。数据经过两种不同方式的清洗，利用 LightGBM 算法预测得出：one 数据集的违约率为 19.9%，two 数据集的违约率为 19.94%，相比于 Lending Club 历史违约结果都有了明显地减少，且基于多观测数据集预测的结果精度更高。基于多观测数据集的 LightGBM 算法得到的违约率（19.9%）比平台本

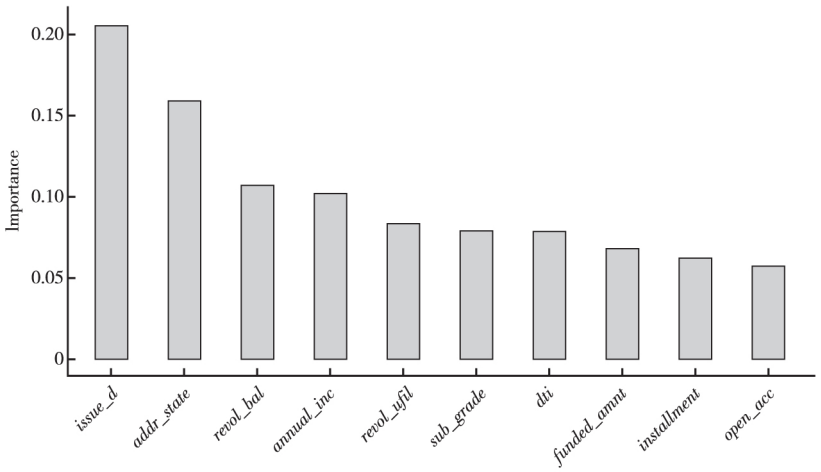

图 16 贡献率最高的前十个解释变量

身的实际违约率 21.18%降低了 1.28 个百分点。

虽然每次的测试集、数据集和变量的筛选顺序在计算机程序中都存在差异，但就预测的影响因素贡献率排序来看，并没有很大的差别。本文综合成以下五个比较重要的因素：①借款金额、贷款利率与信用等级，信用等级越低，贷款利率与借款金额越高，违约风险越大，这是非常重要的参考因素。②年收入、每月还款金额与月负债收入比，年收入代表着一个人的经济实力，能最大程度地显示出借款人是否有能力每月按时偿还债务，其次，每月还款金额与月负债收入比过高，意味着借款人无能力按时还款。③贷款人所在州，居住在不同地区的居民，从某种程度上意味着他的经济实力和背景，可以作为平台和投资者的参考因素。④贷款发生的年份，经济大背景的不同，会影响信贷的违约情况，平台的技术成熟度也是随着时间的推移而提高的，这都会影响平台的违约率，但是这一因素是投资者无法控制的，因此并不作为主要参考因素。⑤其他能够代表借款人经济状况的因素：账户数量、循环信用额度、最高总信用额度、开设循环账户距今的月分数。

五、结　　语

通过上文的分析可以发现，现在多数 P2P 平台广泛使用的违约预测模型有待改进，LightGBM 算法将是未来改进方法的备选之一。本文在利用 LightGBM 机器学习算法进行违约预测后，虽然仅仅降低了 1.28%的违约率，但贷款基数大，粗略估计，如果 Lending Club 平台从成立以来利用此模型进行信贷审核，对于所有 Lending Club 的投资者而言，将避免发生总计 1.17 亿美元的损失。在未来个人信用评级模型的指标选取方面，可以更多地考虑上文提到的五大影响因素，进而设置相关的衡量指标。

参 考 文 献

[1] Arya S.，Eckel C. C.，Wichman C.，2013，*Anatomy of the Credit Score* [J]，Journal of Economic Behavior and Organization，95，175～185.

[2] Berger S. C.，Gleisner F.，2009，*Emergence of Financial Intermediaries in Electronic Markets*：

The Case of Online P2P Lending [J], Business Research, 2 (1), 39～65.

[3] Everett C. R., 2015, *Group Membership, Relationship Banking and Loan Default Risk: The Case of Online Social Lending* [J], Banking and Finance Review, 7 (2), 15～54.

[4] Freedman S., Jin G. Z., 2008, *Do Social Networks Solve Information Problems for Peer-to-Peer Lending ? Evidence from Prosper. Com* [R], NET Institute Working Paper.

[5] Freedman S., Jin G. Z., 2011, *Learning by Doing with Asymmetric Information: Evidence from Prosper. Com* [R], NBER Working Paper No. 16855.

[6] Li H., Zhang Y., Zhang N., Jia H., 2016, *Detecting the Abnormal Lenders from P2P Lending Data* [J], Procedia Computer Science, 91, 357～361.

[7] Herzenstein M., Andrews R. L., Dholakia U., 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.

[8] Herzenstein M., Sonenshein S., Dholakia U. M., 2011, *Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions* [J], Journal of Marketing Research, 48, 138～149.

[9] Iyer R., Khwaja A. I., Luttmer E. F., Shue K., 2016, *Screening Peers Softly: Inferring the Quality of Small Borrowers* [J], Management Science, 62 (6), 1554～1577.

[10] Kupp M., Anderson J., 2007, ZOPA: Web 2.0 Meets Retail Banking [J], Business Strategy Review, 18 (3), 11～17.

[11] Klafft M., 2008, *Peer to Peer Lending: Auctioning Microcredits over the Internet* [A]. in Agarwal A., Khurana R. (eds), *Proceedings of the International Conference on Information Systems* [C], Technology and Management, IMT, Dubai.

[12] Lin M., 2009, *Peer-to-Peer Lending: An Empirical Study* [C], AMCIS 2009 Donsortium.

[13] Lin M., Prabhala N, R,, Viswanathan S., 2011, *Judging Borrowers by the Company They Keep: Social Networks and Adverse Selection in Online Peer-to-peer Lending* [C], Western Finance Association 2009 Annual Meeting Paper.

[14] Malekipirbazari M., Aksakalli V., 2015, *Risk Assessment in Social Lending via Random Forests* [J], Expert Systems With Applications, 42 (10), 4621～4631.

[15] Meng Q., Ke G., Wang T., Chen W., Ye Q., Ma Z., Liu T., 2016, *A Communication-Efficient Parallel Algorithm for Decision Tree* [C], 30th Conference on Neural Information Processing Systems (NIPS 2016).

[16] Ravina E., 2007, *Beauty, Personal Characteristics and Trust in Credit Markets* [C], American Law & Economics Association Annual Meetings.

[17] Renton P., 2012, *The Lending Club Story* [M], Renton Media LLC.

[18] **艾金娣**:《P2P **网络借贷平台风险防范**》[J],《**中国金融**》2012 **年第** 14 **期**。

[19] **巴曙松、侯畅、唐时达**:《**大数据风控的现状、问题及优化路径**》[J],《**金融理论与实践**》2016 **年第** 2 **期**。

[20] **李焰、高弋君、李珍妮、才子豪、王冰婷、杨宇轩**:《**借款人描述性信息对投资人决策的影响——基于** P2P **网络借贷平台的分析**》[J],《**经济研究**》2014 **年第** S1 **期**。

[21] **李悦雷、郭阳、张维**:《**中国** P2P **小额贷款市场借贷成功率影响因素分析**》[J],《**金融研究**》2013 **年第** 7 **期**。

[22] **廖理、张伟强**:《P2P **网络借贷实证研究:一个文献综述**》[J],《**清华大学学报(哲学社会科学版)**》2017 **年第** 2 **期**。

[23] **廖理、李梦然、王正位**:《**聪明的投资者:非完全市场化利率与风险识别——来自** P2P **网络借贷的证据**》[J],《**经济研究**》2014 **年第** 7 **期**。

[24] **廖理、吉霖、张伟强**:《**借贷市场能准确识别学历的价值吗?——来自** P2P **平台的经验证据**》

[J],《金融研究》2015 年第 3 期。

[25] 卢馨、李慧敏:《P2P 网络借贷的运行模式与风险管控》[J],《改革》2015 年第 2 期。

[26] 马若微、唐春阳:《考虑误判损失的 Logistic 违约预测模型构建》[J],《系统工程理论与实践》2007 年第 8 期。

[27] 王会娟、廖理:《中国 P2P 网络借贷平台信用认证机制研究——来自“人人贷”的经验证据》[J],《中国工业经济》2014 年第 4 期。

[28] 微软亚洲研究院:《开源 | LightGBM:三天内收获 GitHub 1000+星》[OL/EB], http://www.sohu.com/a/123480446_133098.

[29] 张昊、纪宏超、张红宇:《XGBoost 算法在电子商务商品推荐中的应用》[J],《物联网技术》2017 年第 2 期。

An Empirical Study on the Credit Rating of P2P Projects based on LightGBM Algorithm

Ma Xiaojun Sha Jinglan Niu Xueqi

(Dongbei University of Finance and Economics)

Research Objectives: In the context of big data and Internet finance development, it is effective to control the default rate of P2P projects to ensure good operation of relevant financial projects or platforms according to personal credit. **Research Methods**: In this paper, based on the P2P platform Lending Club massive real transaction data, we use the 'multi-observation' and 'multi-dimensional' two kinds of data cleaning method, and predict the risk of default by the 2016 Asian Microsoft LightGBM machine learning algorithms of authority and innovative. Then compare the results of different data cleaning method. **Research Findings**: LightGBM algorithm based on multi-observations of predicted results is the best. Lending Club platform historical trading data to calculate the average execution rate of 1.28%, can reduce about MYM117 million in loan default. **Research Innovations**: Using different data cleaning methods and more advanced machine learning algorithm (LightGBM) to predict default rate. **Research Value**: Based on the result of the influencing factors of default rate, not only the suggestions on development of Lending club and P2P platforms are pointed out, but also the direction of national development in this field.

Key Words: P2P; Credit; Control of Default Rate; Data Cleaning; LightGBM Algorithm

JEL Classification: C39

(责任编辑:王喜峰)