

【理论研究】

美国 ZestFinance 公司大数据征信实践

刘新海^{1 2}, 丁 伟³

(1. 中国人民银行 征信中心, 北京 100031; 2. 北京大学 金融智能研究中心, 北京 100871;
3. 中国联通 网络技术研究院, 北京 100032)

摘 要: 美国金融科技公司 ZestFinance 利用大数据进行信用风险评估引起国内互联网金融和征信领域的关注。对 ZestFinance 的商业模式和大数据挖掘技术进行进一步解读, 概述信用风险管理的发展历史和背景知识, 深入分析 ZestFinance 的商业理念, 详细介绍 ZestFinance 的技术路线; 总结对刚刚起步的中国征信业的一些借鉴意义, 希望能够对中国未来的大数据征信有一些启示。

关键词: 征信; 大数据; 信用评分; 机器学习; 数据挖掘

中图分类号: F832.332; TP391.4

文献标识码: A

文章编号: 1674-747X(2015)08-0027-06

一、信用评估技术的进展

征信 (Credit reporting 或者 Credit reference) 是指依法收集、整理、保存、加工自然人、法人及其他组织的信用信息, 并对外提供信用报告、信用评估、信用信息咨询等服务, 帮助客户判断、控制信用风险, 进行信用管理的活动。征信的主要目的是解决交易双方信息不对称的问题, 进行信用风险管理。征信业与信息技术联系密切, 在欧美甚至被归类为信息产业。信息技术的进步驱动了信用风险管理水平的提高, 促进了信用的可获得性。

以消费信贷为例, 美国的个人征信 (或者是个人消费者的信用风险管理) 的发展可以分为三个阶段 (如图 1 所示)。

第一个阶段 (1950 年以前): 定性信用风险决策。1950 年之前, 审贷过程是根据放贷人的经验和对借贷人的了解, 靠人工来完成的。这是一种定性

的决策, 效率比较低, 没有一个客观统一的标准, 放不放贷款完全依赖于信贷员个人的经验。审贷的标准因人而异, 不具备客观性和科学性。从图 1 可以看出在这一阶段信贷市场的发展缓慢, 总的信贷金额规模较小。

第二个阶段 (1950—1980 年代): 局部风险量化分析。1956 年信用评估数据挖掘公司费埃哲 (FICO) 成立, 来自斯坦福的工程师 William Fair 和数学家 Earl Isaac 进行了数学分析模型和工程实现的整合, 两年之后 FICO 公司卖出了它的第一个评分系统。利用 FICO 评分模型, 银行等信贷机构可以进行自动化的批量审贷, 增加了信贷的可获得性, 减少了违约率, 促成了消费信贷的第一次革命^[1]。如图 1 所示, 在该阶段, 消费信贷市场得到了快速的发展。但由于银行只是利用内部的数据进行 FICO 个人评分建模, 因此 FICO 模型在解决信贷交易双方信息不对称问题上的作用仍有局限。

收稿日期: 2015-06-17

基金项目: 国家自然科学基金青年基金 (61105058); (教育部) 留学回国人员科研启动基金 (2015-48) 和国家社会科学基金 (13CJY011)

作者简介: 刘新海 (1976—), 男, 河南南阳人, 博士, 博士后研究员, 主要研究方向为征信、数据挖掘、金融风险管理和大数据; 丁伟 (1972—), 男, 辽宁抚顺人, 博士, 中国联通网络技术研究院高级工程师, 中国会计学会高级会员, 中国计算机学会会员, 中国优选法统筹法与经济数学研究会理事, 财政部财科所硕士生导师, 云南大学兼职教授、博士生导师, 主要研究方向为数据挖掘与征信大数据。

第三个阶段(1980年以后):全局量化风险分析。1980年之后,征信机构 Experian 收集了不同信贷机构个人消费者的信贷信息,形成了消费者的全局信息。不同的信贷结构之间共享全局信息,可以全面地了解个人消费者的信用状况,更好地解决了信息不对称问题,促成了信用风险管理的又一次革命。从图1可以看出,自1980年之后,消费信贷市场得到飞速的扩张。但在该阶段,由于征信机构不能为缺少信贷信息的少量个人消费者提供全局信用记录,因此这类消费者无法享受信贷机构的正常服务。

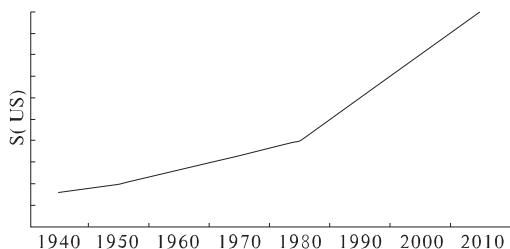


图1 美国消费信贷市场的发展过程

目前,IT信息技术的突飞猛进使人类社会进入了大数据时代,更多维度的数据和不同层次的数据都可以用来挖掘和分析。挖掘数据中大的价值,使其为我们的工作和生活服务,需要有更好的挖掘和分析能力。机器学习的算法和模型可以实现对大数据的深入挖掘和分析^[2]。ZestFinance公司的总裁 Douglas C. Merrill 先生认为大数据技术加上机器学习算法使个人消费信贷的信用风险管理进入了新的阶段。在这个阶段,可以从大数据中挖掘出更多的消费者信用信息;即使是传统征信体系不能覆盖的一些缺乏信贷记录人群的信用信息,也能从大数据中挖掘出来。

二、ZestFinance 的基本商业理念

作为一个创新型的科技金融公司,ZestFinance 自创立以来,由于其独特的商业理念不断受到投资界和互联网金融领域的瞩目^[3]。

(一) 为弱势群体制造信用的普惠金融思想

ZestFinance 认为,它的使命是为每一个人创造公平而且透明的信用信息。ZestFinance 最初的服务对象是只能使用高利贷的人群(称为借贷日贷款人群),通过大数据挖掘出他们的信用信息,帮助他们

享受正常的金融服务。ZestFinance 假定每一个消费者都是“好”人,希望通过搜集证据,证明信贷信息不完整人群的真正的信用状况,进而帮助他们实现享受正常金融服务的权利。

(二) 非传统征信视角看信用

个人征信的目的是通过搜集数据来挖掘出消费者已有和潜在的信用信息,数据和具体的技术只是征信的手段和工具。消费者的信用状况与其信贷记录联系最紧密,是强相关的,所以传统征信是从消费者的信贷记录挖掘其信用状况。但是消费者的信用状况与其信贷记录并非是一一映射的关系。消费者的信用状况还会与消费者的其他信息和行为有联系,尽管这种联系可能较弱。

图2展示了 ZestFinance 公司的大数据征信的视角。与传统征信相同,大数据征信对消费者的信用评估也是基于两个基本面的信息:消费者的还款能力和消费者的还款意愿。所不同的是,传统征信中,数据依赖于银行信贷数据,而大数据征信的数据不仅仅包括传统的信贷数据,同时也包括了与消费者还款能力、还款意愿相关的一些描述性风险特征,这些相关性描述风险特征的抽取与筛选是 ZestFinance 的技术核心。相比于传统征信数据的强相关性,这些大数据征信的数据与消费者的信用状况相关性较弱,ZestFinance 就利用大数据技术搜集更多的数据维度来加强这些弱相关数据的描述能力。这样就使大数据征信不依赖于传统信贷数据,可以对传统征信无法服务的人群进行征信,实现对整个消费者人群的覆盖。

(三) 优于银行的大数据挖掘技术

ZestFinance 的优势在于其强大的数据挖掘能力,能开发出新颖的信用评估模型,从大数据中挖掘出更多不易被发现的消费者信用信息。尽管 ZestFinance 公司自运营以来取得了不错的业绩,但 Merrill 先生还是具有专业学术人士特有的谦逊严谨的风格,他阐述了一些 ZestFinance 的发展弱势,即公司规模还比较小,目前直接服务的人群只有10万人左右;不拥有银行丰富的个人消费者数据,但其本身所有数据却可以被银行通过正常手段和渠道获得。

(四) 对于大数据的理解和利用

不同的应用对大数据有着不同的理解,目前对大数据没有一个统一的概念。Merrill 先生认为“数

据多”并不是“大数据”,能够利用 IT 先进技术将碎片化的信息整合起来才能形成真正有用的大数据。这种对大数据的理解特别适用于征信业,因为征信的基本过程也是将分散于不同信贷机构看似用处不大的局部信息整合成为可以完整描述消费者信用状况的全局信息。

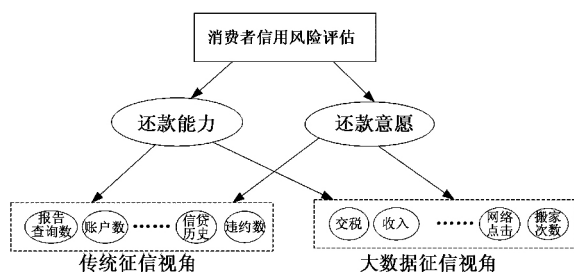


图2 大数据征信视角与传统征信视角的比较

值得指出的是,ZestFinance 虽然利用大数据技术进行征信,但是主要利用的还是结构化的大数据,对于复杂的大数据类型例如文本数据和社交网络数据利用得比较少,主要原因是由于这些复杂的大数据和 ZestFinance 所服务的消费者的信用风险相关性太弱了。这种现象其实也被其他互联网金融的征信实践所验证,例如最大的互联网 P2P 公司 Lending club 最早从 Facebook 平台登录信贷市场,希望通过对社交网络的数据进行信贷审批,结果效果很差,难以为继,最后只好回归传统的信贷手段发展。

三、ZestFinance 的技术分析

ZestFinance 的核心业务是消费信贷审批,主要客户是次级贷消费者,主要的竞争对手是银行或典当行。ZestFinance 的核心竞争力在于其强大的数据挖掘能力和模型开发能力,将机器学习领域比较成熟的技术创造性地应用于传统的信贷风险管理领域。

(一) 采用了基于多角度学习的评分预测模型

传统的信用评分模型一般拥有 500 个数据项,从中提取 50 个变量利用一个预测分析模型做出信用风险量化评估。而在 ZestFinance 的新模型中,往往要用到 3500 个数据项,从中提取 70 000 个变量,利用 10 个预测分析模型进行集成学习或者多角度学习,进而得到最终的消费者信用评分。如图 3 所示,ZestFinance 的数据源是大数据,可以生成数以万计的风险变量,然后分别输入不同的预测模型中,例如欺诈模型、身份验证模型、预付能力模型、还款能

力模型、还款意愿模型以及稳定性模型等。每一个子模型都从不同的角度预测个人消费者的信用状况,克服了传统信用评估中单个模型考虑因素的局限性,使预测更为精准。

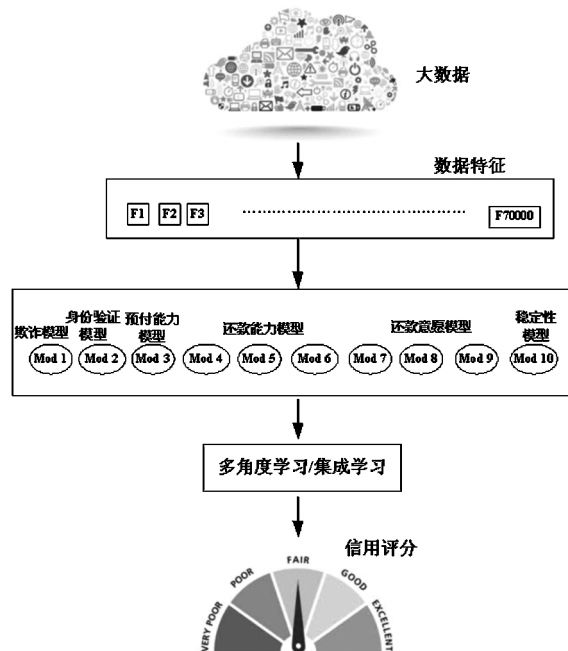


图3 ZestFinance 的信用评估模型解读

机器学习方法在生产、科研和生活中有着广泛应用,而集成学习则是机器学习最近的热门研究方向。集成学习是使用一系列算法模型进行分析预测,并使用某种规则把各个模型分析结果进行整合从而获得比单个算法模型更好的预测效果的一种机器学习方法^[4]。如果把单个模型比作一个决策者的话,集成学习的方法就相当于多个决策者共同进行一项决策。由于融合多种信息和综合多种决策机制,经过集成学习得到的分析预测要明显优于单一模型。不同角度的信息存在关联,各自包含互补信息,多角度学习的过程,相当于一个不断搜集证据的过程,加强互补信息,进行信息融合。例如,两个独立的评分模型对利润提升的能力分别是 16.9% 和 9.4%,传统信用评估中,第二个模型可能被弃用,但如果发现这两个模型分别包含互补信息,将这两个模型的结果进行融合,可以将利润提高至 38%。

这些多角度学习中每一个子模型也没有采用传统的逻辑回归,而是机器学习中的其他的预测模型(该模型的细节对于 ZestFinance 是核心机密)。ZestFinance 的信用评估中之所以没有利用逻辑回归模型的原因是大数据征信的数据比较零碎,而且变量过

多,以及所抽取的风险变量不一定和风险预测变量存在线性关系等。

(二) 评分模型不断更新

当 ZestFinance 不断地进行数据搜集和增加新数据源的同时,评分模型也在不断更新。从 2012 年到目前,差不多每一个季度就会推出一个新的信用评估模型。而且模型是以每一位不同的开发者命名,目前已经有 14 个模型。ZestFinance 评分模型的改进也提高了其信用风险评估水平。虽然这些新的模型仍然会遇到数据充足性和数据可得性的挑战,但是模型的持续改进还在进行中。

(三) 评分模型不断细化

ZestFinance 最早从事的是信贷审批,仅有信贷审批评分模型,随后不断细化其评估模型来支持不断推出的新的信用风险业务。例如 2013 年第一季度推出了催收评分,2014 年第二季度推出了市场营销评分,2014 年初推出了汽车贷款和法律催收。目前已经开发出八类信用评估模型,用于不同信用风险评估服务。

(四) 挖掘丢失数据,变废为宝

丢失数据(Missing data)指的是机器学习中某一些数据项因为种种原因造成了缺失,较多的丢失数据会给建模过程带来挑战。由于 ZestFinance 利用多维度的大数据,所以丢失数据的现象更加突出。ZestFinance 处理丢失数据也有其独特之处:首先 ZestFinance 不断改进其评分模型,增强其处理丢失数据的能力,最新的评分模型可以处理超过 30% 的丢失数据。其次 ZestFinance 另辟蹊径,充分利用丢失数据之间的关联,和正常数据的交叉以及探寻数据丢失的原因。通过这样的深耕细作,获得了一些有用的消费者信用信息。当然这种做法要求使用于特定环境,要结合信贷业务,并对消费者的行为模式有深入的理解。

(五) ZestFinance 和竞争对手的比较

根据调查,80% 左右的信贷风险来自信贷审批环节,一旦消费者获得信贷,后续的管理只能控制 20% 的风险,由此可见科学的信贷审批管理十分重要。开发高质量的信贷审批评分模型,进行科学的审批风险管理可以大幅降低坏账率,并取得比较好的经济效益^[1]。虽然 ZestFinance 并未透漏其坏账率,但在实际应用的过程中,和其竞争对手银行或典

当行的放贷者相比,在信贷审批管理方面还是取得了一些不错的结果。

1. 获得贷款顾客的成本是竞争者的 25%

根据 JMP2012 的行业报告,对于一个在线的借贷者来说,获取一个顾客的平均花销应该在 250 美元和 500 美元之间。如图 4 所示,2014 年,通过 ZestFinance 的服务获得顾客的成本稳定在 100 美元左右。

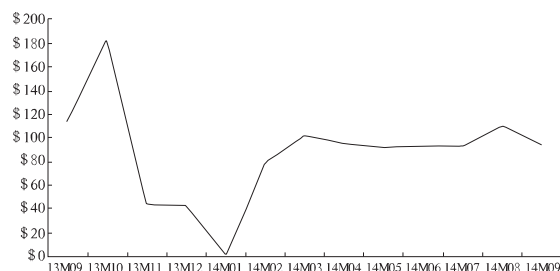


图4 ZestFinance 获得顾客的成本

数据来源: ZestFinance 公司资料。

2. 首次还贷违约率(first pay default rate , FPD Rate) 低于竞争者

如图 5 所示,最下端曲线为 ZestFinance 的首次还贷违约率,其余的三条直线表示其他竞争对手的情况(对于银行放贷者来说,FPD 已经比较稳定)。从图 5 可以看出,从 2012 年初起,ZestFinance 的首次还贷违约率还在不断波动,有时高于第三个竞争对手,随着模型不断改进,到 2013 年已基本稳定,明显低于三个竞争对手。

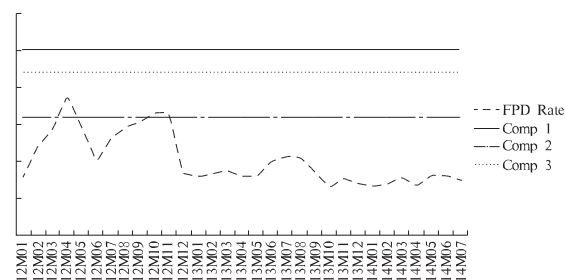


图5 ZestFinance 的首次还贷违约率和其竞争对手比较

数据来源: ZestFinance 公司资料。

3. 不断提高客户的投资回报率

ZestFinance 最初的投资回报率在 100% 左右,随着模型的不断改进,目前的客户投资回报率达到了 150% 以上,如图 6 所示。

4. 投资回报率高于竞争对手

图 7 展示了 ZestFinance 在不同时间和其竞争对手

手在投资回报率方面的比较,可以看出 ZestFinance 的投资回报率明显高于行业标准和其竞争对手。



图6 ZestFinance 不断提高客户投资回报率

数据来源: ZestFinance 公司资料。

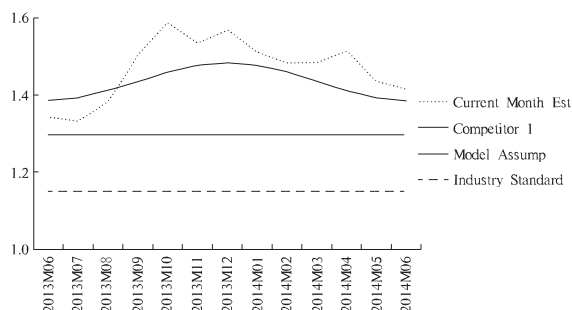


图7 ZestFinance 的投资回报率高于竞争对手

数据来源: ZestFinance 公司资料。

四、ZestFinance 的最新动态

2015年6月底,京东集团宣布投资 ZestFinance, 双方同时还宣布成立名为 JD-ZestFinance Gaia 的合资公司。此举意味着京东将在自身消费金融体系中进一步引入 ZestFinance 的技术和经验,打造创新消费者信用评估体系。京东金融副总裁姚乃胜表示:“目前我国在信用评估方面的体系还不成熟,由于中国消费者通常没有完备的信用记录,个人信用难以预估,这就成了我国消费信贷的发展瓶颈,也是与发达国家普及消费信贷存在明显差距的主要原因。”这次的合作最初将为京东上的分期付款购物行为评估信贷风险,最终的战略目标是为中国各地的企业客户提供信用分析服务。

关于 ZestFinance 的模式,在风险评估领域还存在争议。吹捧者认为,数据勾勒出来的面貌,应该可以让风险分析变得更加精准,因此有助于以更低的成本把消费者信贷提供给额外的人,而其中涉及的人数以百万千万计。然而,一些公共政策专家认为,人们对贷款新模式的热情跑在了实证分析的前面。

美国联邦贸易委员会(Federal Trade Commission)前任律师阿隆·里克(Aaron Riek)称,大数据信用评估技术的准确性和公正性尚未经过证实。面对中国广阔的消费者征信市场,同时信用数据基础非常薄弱,ZestFinance 和京东联手能否用科技金融的手段带来成功的创新,还需要拭目以待。

五、对中国征信业的启示

ZestFinance 通过技术创新和金融风险管理的创新,带来了新颖的商业理念和技术路线,对于正在起步的中国征信业以及互联网金融中的风险管理有一定的启示作用。

(一) 征信勿以“惩罚性”为目的

ZestFinance 以大数据征信技术,假定每一个消费者都是“好”人,不随便设黑名单进行“有罪推理”。ZestFinance 为所有的消费者挖掘信用,用科技的力量推动普惠金融的发展,打破信贷机构为富人服务的怪圈。相比而言,目前国内的信用风险管理,惩罚性太强,一种类似“有罪推理”的思路大行其道,这种方式可能简单有效,但是并没有把征信的作用全面发挥出来。征信的真正作用不仅仅是惩戒失信,更重要的是褒扬诚信。

(二) 定位于特定的服务人群

ZestFinance 的主要服务对象是约占人口 5% 的,信用评分在 500 分以下的次级贷人群。通过对这部分消费者的深入理解,筛选其大数据描述信息,所开发的也是针对这部分人群有效的信贷审批模型。在机器学习领域没有特别通用的分析模型,但是往往存在针对特定范围内有效的模型。所以合理地定位服务人群和深入理解服务对象是开发征信分析模型乃至开展征信服务业务成败的关键。

(三) 加强信用分析模型的研发

ZestFinance 优于其竞争对手和传统信贷机构的一个重要的原因是强大的信用评分模型的开发能力:基于多角度学习的预测模型,模型及时更新而且不断细化。相比而言,中国的信贷审批,或信用风险管理,层次不齐,从定性判断到简单的量化决策都有,总的来说量化分析不足,而且征信机构的信用评分还未推出^[5]。只有加强对量化信用风险分析技术的研发投入,才能真正实现对消费信贷的专业风险管理。同时,值得强调的是,大数据时代,没有现成

的免费午餐,数据和模型需要提炼,需要数据科学家的人工参与,即使把 ZestFinance 的模型拿到中国来,也不能直接用。对数据和消费者的理解和数据挖掘技术的掌握都是建模过程中不能省掉的功课。

同时我们也要对大数据征信以及 ZestFinance 有一个全面的认识:

(一) 征信数据是天然的大数据,无论是基于 Gartner 提出大数据 4V 理论还是本文中 Merill 先生对于大数据的理解

大数据征信不仅仅包括利用信用评估,还包括征信大数据的采集,征信大数据的管理,征信大数据的安全以及征信大数据的可视化等^[6]。同时国外传统的征信机构和新兴的互联网金融机构也都在探索大数据征信的不同方式。例如 FICO 公司的研究表明,将社交媒体和电商网站拥有的在线数据、移动运营商的手机使用数据与传统征信数据结合用于风险建模,提高了模型对客户的区分度,在降低拒贷率的同时,提高了风险预测能力。FICO 目前正在与少数金融机构合作,在小范围内使用这种模型。美国的一家新的评分公司更是利用非传统的数据和心理测量学的分析和建模方法来进行信用评估。

(二) ZestFinance 并非一个征信机构,其主要业务也是征信产品中的信贷申请评分

在进行信用评估时,ZestFinance 并没有直接从征信机构获取数据,建模中的 30% 数据来自于不同的数据代理商,这些数据和征信数据相似。而且 ZestFinance 并没有向征信机构报送数据,主要的原因可能是 ZestFinance 所用的数据项维度太多,面向

特定人群,时效性强,不符合征信机构的数据的通用、成熟、稳定和信用相关性强的特点。

(三) 对于征信大数据的利用并不仅限于机器学习

人工智能的其他分支,如复杂科学、心理测算学都可能是未来利用大数据建模进行信用评估的理论基础^[7]。

(四) Zest Finance 也面临着合规性挑战

由于从事的是金融业务,ZestFinance 就需要像其竞争对手(例如传统银行)一样,接受监管。目前美国借贷的监管环境非常复杂,联邦法律下还有各州的法律并行。所以 ZestFinance 还在继续研究监管机构和其他因素所扮演的角色,在开展大数据征信的同时努力满足合规性的要求。

参考文献:

- [1] 陈建. 信用评分模型技术与应用[M]. 北京: 中国财政经济出版社 2007: 16-57.
- [2] 刘新海,丁伟. 大数据征信应用与启示[J]. 清华金融评论 2014(10): 93-98.
- [3] Mitchell, T. M. *Machine Learning* [D]. The McGraw-Hill Companies, INC. 1997.
- [4] Liu X. *Learning from Multi-view Data* [D]. K. U. Leuven. 2011.
- [5] 刘荣,崔琳琳. 大数据技术在中小企业信用体系建设中的应用[J]. 征信 2015(5): 40-43.
- [6] 袁新峰,赵强,甘瀛. 美国消费者征信行业的经验及启示[J]. 征信 2015(3): 63-68.
- [7] 王俊山. 推动征信标准实施 规范数据库用户管理——《金融信用信息基础数据库用户管理规范》解读[J]. 征信 2015(7): 8-10.

(责任编辑: 姚文)

Big Data Credit Reporting Practices of ZestFinance in the United States

LIU Xinhai^{1,2}, DING Wei³

- (1. Credit Reference Center, the People's Bank of China, Beijing 100031, China;
2. Financial Intelligence Research Center of Peking University, Beijing 100871, China;
3. China Unicom Network Technology Research Institute, Beijing 100032, China)

Abstract: That the American financial science-tech company(ZestFinance) makes assessment of credit risk by using big data has aroused wide concern in the domestic Internet finance and credit reporting fields. Based on a further explanation of the business model and data mining technology of ZestFinance, the paper outlines the development course and background knowledge of credit risk management, makes in-depth analysis on the business philosophy of ZestFinance, and introduces the details of the technical routes; and summarizes the guide to China's credit reporting industry which is on the start line, hoping to give some implications for China's big data credit reporting in the future.

Key words: credit reporting; big data; credit scoring; machine learning; data mining