

大学生创新创业训练计划项目 结项申请表

项目名称： P2P 网贷中信贷用户违约预测模型的构建
——基于 stacking 算法的多模型融合

项目类别： 大学生创新训练项目

项目级别： 国家级

立项年度： 2018

项目主持人： 郝建锋

依托单位： 统计与数学学院

指导教师： 潘蕊

中央财经大学教务处制表

填表日期： 2019 年 4 月 8 日

项目成员 (排名)	姓名	所在学院	专 业	班 级	联系电话
	郝建锋	统计与数学学院	经济统计学	经济统计学 16	1305189963
	李小高	统计与数学学院	经济统计学	经济统计学 16	1560090274
	王燕坪	统计与数学学院	经济统计学	经济统计学 16	1881137337
	施霁珂	统计与数学学院	经济统计学	经济统计学 16	1881091931
	陈步	统计与数学学院	统计学	统计学 16	1899008925
指导教师 (排名)	姓名	所在单位	职 称	联系电话	
	潘蕊	统计与数学学院	副教授	62288184	
项目成果	成果具体信息				
结项报告	名称:P2P 网贷中信贷用户违约预测模型的构建——基于 stacking 算法的多模型融合				
发表论文	论文题目		发表日期	刊物名称(期数)	
竞赛获奖	奖项名称		获奖日期	颁发单位	
产品研发	产品名称		应用		
公司注册	公司名称		业务领域	组织机构代码	
项目目标(400 字左右)					
<p>实施目的、拟解决的主要问题:</p> <p>1. P2P 网贷高违约率现状</p> <p>课题小组通过分析国内 P2P 网贷平台的整体发展,以及观察多个 P2P 网贷平台运营情况,发现各平台都不同程度的存在借款人违约行为,而过多的违约现象可能会导致平台出现资金链断裂的情况,影响互联网金融的健康发展。因此,减少网贷违约行为变得极为重要。为了解决上述问题,课题小组从人人贷平台上获取 2015 至 2016 年的相关借贷数据,包括借款人的年龄、职业、借款金额、借款利率、历史借款总额等多方面信息。分析我国 P2P 网贷平台的总体状况以及所获取的数据总结出其违约率较高的原因,并针对这些原因,基于人人贷等有代表性的网贷平台数据,构建更加合理有效的违约预测模型。</p> <p>2. 完善 P2P 网贷领域指标体系</p> <p>传统金融机构进行信用评估时侧重于对借款人还款能力的评估和信贷记录的考核,然而这种评估时效性差,且中国尚未建立全民的信贷记录数据库,采集有一定困难,并且从小组实际采集的数据来看,该类指标覆盖率较低,因此传统的指标体系不适合完全应用于 P2P 网</p>					

贷领域。通过对 P2P 平台上采集的数据的特征分析,发现借款信息和历史信用信息都是 P2P 平台自身运营过程中用户行为所产生的数据,这些信息与平台运营高度相关却未被纳入传统体系考虑范围。因此,本研究在传统金融机构信用评估的基础上,广泛参考专家研究成果,从个人基础信息、借款信息、资产信息、历史信用信息四个方面构建 P2P 网贷信用评估指标体系。

3. 传统违约预测模型

首先考虑传统的信用评估算法。目前违约预测方面多使用分类模型算法,主要包括 Logistic 回归方法、决策树法等统计模型,也包括神经网络等非统计模型。课题小组将结合已经选定的 P2P 网贷信用评估指标体系,分别使用 Logistic 回归方法、决策树法、K 最近邻法对收集的数据进行拟合,得出违约预测模型。

模型构建后,利用准确率和 AUC 值比较各模型的性能,直观比较各模型预测正确的样本占全样本的比率,以评价模型的预测效果。

4. Stacking 集成算法

从相关文献以及课题小组对单一模型的构建过程中发现,众多单一模型在分类准确率、稳健性和可解释性等方面各有偏重。如 Logit 方法易于理解和实现,而准确率偏低,且一般不能直接对不均衡样本使用;决策树方法有较好的拟合精度和可解释性,但假如训练数据中存在噪音数据或缺少代表性数据时,决策树会出现严重的过拟合现象;神经网络模型能充分逼近复杂的非线性关系,分类准确率高,但学习时间长且不能观察中间学习过程,输出结果难以解释。

为了克服单一模型的缺点,研究者们提出了综合多个单一模型有点的集成模型方法,Stacking 集成算法由 Wolpert 于 1992 年提出。为了解决单个模型可能产生的欠拟合和过拟合问题,同时提高预测精度和稳健性,课题小组使用 Stacking 算法将 Logit 模型、Random Forest 模型、Gradient Boosting Machine 模型融合在一起构建新的违约预测模型。

集成模型构建后,同样利用准确率和 AUC 值评估模型性能,并与单一模型进行比较,准确率越高、AUC 值越大,说明模型性能越好。

5. 实证分析与建议

通过混淆矩阵分别比较所构建的单一模型和集成模型的查全率、查准率和 F1 等指标,来反映课题小组构建的模型分类预测在实际应用中的表现,并通过模型输出结果分析变量重要性,观察变量对分类结果的影响大小,重点研究对模型分类效果影响较大的变量,并对平台的信息收集及披露提出可行性建议。

项目成果（1000 字左右）

理论成果（观点结论、政策建议）、实践成果（项目实践与应用成果、产品/服务运营成果）：

目前 P2P 网贷平台仍然存在许多问题，尤其在 2018 年，P2P 平台的暴雷现象尤其这严重，这给众多的投资者带来了不小的损失。其中信用风险较高便是 P2P 网贷平台问题频出的重要原因之一，然而目前行业内并没有形成统一的信用风险衡量标准，国家也对于 P2P 网贷平台的正常健康发展高度重视，如何能够有效地减少信用风险的发生是 P2P 网贷平台面临的一大难题。

本文利用 stacking 方法融合了 KNN 算法、随机森林、xgboost 和 logistic 模型结合探索性分析结果，对人人贷平台的借款人的违约情况进行了预测。通过本文的分析过程可以发现以下结论：

1. 地域差别明显

从借款人的违约率的地理分布来看，西部以及西北部的省份的违约率稍高于东部和中部分省份，同时东部成交量大大高于西部。这说明，东部沿海城市等发达城市信用经济发展较为活跃，适合 P2P 平台这类互联网金融部门拓展业务，开拓市场。

2. 高回报伴随高风险

利率较高、还款周期较短的标的的借款人违约率大大高于利润平缓、还款周期较长的标的的借款人，这充分表明了看似回报值较高的资产配置往往蕴藏着巨大风险，平台应该合理设置较为稳定的平台利率与适当长度还款周期，这有利于降低违约风险。对于借款标的利率设置偏高且又周期较短的用户重点监控，防止其发生违约行为。

3. 历史信用信息和是否认证对违约预测结果影响较大

从模型输出变量重要性来看，历史信用信息和是否认证对借款人违约情况有重大影响，逾期次数越多的借款人更可能违约，实地认证和信用认证之后的借款信用风险更小。平台应该更加重视信用历史数据的收集，同时积极配合线下认证，降低标的的违约风险。

4. 本文依据人人贷平台的数据建立了预测借款人是否违约的模型，模型的精度和泛化能力都表现十分良好，对于借款人的信用风险评估可以起到一定帮助作用。

项目创新（300 字左右）

内容、方法、成果等方面的创新与特色：

1. 切合时代热点

近年来互联网金融飞速的发展为我国的经济增添了新的动力，但互联网金融与传统金融服务有很大的不同。互联网金融借助于互联网的特性使其风险相对于传统金融更大，监管也更为困难，而我国目前的征信系统的服务对象还仅限于线下的传统金融，无法满足 P2P 网贷等的互联网金融的需求。由于准入门槛低、监管缺失及缺乏合适的征信系统，P2P 网贷等互联网金融问题频发，这也引起了投资者的担忧和政府相关部门的重视。政府在 2015 及 2016 年相继发布了《网络借贷信息中介机构业务活动管理暂行办法》、《网络贷款信息中介机构业务活动管理暂行办法》等法规，进一步强调对互联网金融的监管和网络个人征信系统的建立。本小组的研究切合此金融热点，构建了 P2P 网络借贷平台的个人信用评估体系，为网络借贷平台的使用者提供借鉴，能在一定程度上减少违约事件的发生。

2. 使用一手数据

数据可以分为一手数据和二手数据两种，二手数据虽然比一手数据更为简单清晰，但由于二手数据是经过加工的数据，难免会代入加工者的观念，影响数据本身的样貌和其反映的信息。而一手数据往往能代表数据最真实的情况，充分反映研究对象的信息，因此一手数据的价值往往高于二手数据。本小组在研究过程中采用直接从人人贷平台爬取的一手数据，从最真实的数据出发，使得研究模型能够更好地拟合研究对象，构造出的信用评分模型能够较为准确地反映个人的信用状况。

3. 多种方法构建模型

目前违约预测模型的构建多使用分类模型算法，主要包括单一算法如 Logistic 回归、决策树法、K 最近邻方法等统计模型和神经网络等非统计模型，也包括如 stacking 算法等的集成算法。其中每一种单一算法在稳健性及预测精度等方面各有优点，但是在实际应用中单一的模型总是有精度不足、稳健性不足等风险，使得所构建的预测模型存在风险。针对以上问题，小组成员在查阅相关研究资料的基础上决定选用集成算法 stacking 算法进行模型的构建。在具体实践中将 Logit 模型，Random Forest 模型，Gradient Boosting Machine 模型融合在一起，既克服了单个模型容易产生的欠拟合和过拟合问题，也在一定程度上提高了模型的预测精度和稳健性。

4. 符合平台特点

网络借贷平台与传统信贷平台的有一定的相通性，但与传统信贷平台相比，网络借贷平台有隐蔽性强、匿名性、监管难等特点，因此传统借贷平台的信用评估体系不能直接套用于网络借贷平台。但由于传统的金融机构信贷平台拥有多年的历史，其在借款人信用评估方面有丰富的经验，故本研究在传统信贷平台的信用评估指标体系的基础上，广泛借鉴专家研究成果，充分考虑网络借贷平台自身特点，综合考虑各种影响因素，最终确定了包含个人基础信息、借款信息、资产信息及历史信用信息四个大方面的个人信用评价指标。这些指标能尽可能全面的反映网贷平台使用者的各方面信息，符合网贷平台的特点。

5. 具有实际应用价值

目前，各 P2P 网络借贷平台上也已经出现平台自己计算出的借款者信用等级评分供投资者参

考，但各大平台评估个人信用所使用的方法不同，评估结果难免存在偏差。本小组构建的个人信用评估模型可为平台使用者提供另一种借鉴使平台使用者能综合考虑借款者可能存在的风险，更加慎重的做出选择。同时，本小组提供的个人信用评估模型的构建方法也可为各网贷平台改进自身信用评估模型提供借鉴。
项目不足（300 字左右）
项目实施过程和成果尚需深化改进之处： 1.本文建立的模型是基于现有的网贷数据所建立，一旦有新的网贷数据添加进来，无法进行动态调整，只能重新建立相关的信用风险预测模型。 2.本文模型的预测结果的输出是一个 0-1 变量，只能预测借款人是否违约，而无法对借款人进行信用打分和评估借款人的信用额度等。 3.鉴于时间和能力的因素，本文涉及的模型多为一些经典模型，并且没有尝试利用其它模型进行对比分析，无法证明本文的模型对网贷平台信用风险的预测一定是最优的。由此可见，本文仅做了网贷平台信用风险一小方面的研究，还有更广阔的天地值得探索。
指导教师评价意见
对项目过程、成果、创新与不足等方面的评价，是否同意参加结项验收： 项目关注实际问题，对时事热点问题进行了深入研究，研究有所创新。 同意申请结项。 <div>指导教师（签字）： <div>年 月 日</div></div>
依托单位意见
项目成果（研究报告\创业计划书）重复率检测： ____% 是否同意参加验收： <input type="checkbox"/> 同意 <input type="checkbox"/> 不同意 <div>依托单位（签章）： <div>年 月 日</div></div>