Can Small Banks Compete on Local Data in Online Retail Loans? The Example of Baotou Rural Commercial Bank

Chen, Yunxiang

UNIVERSITÉ DE GENÈVE GENEVA FINANCE RESEARCH INSTITUTE

“市民贷”地方中小银行线上零售贷款业务应用研究

Dissertation Submitted to

The University of Geneva

in partial fulfillment of the requirement

for the professional degree of

Doctorate of Advanced Professional Studies in Applied Finance, with Specialization in Wealth Management

by

陈云翔

(FCO N° 58501)

Dissertation Supervisor: Professor Harald HAU
University of Geneva
Associate Supervisor: Professor FAN Wenzhong
Tsinghua University

September, 2020

Disclaimer

I declare that I have read the plagiarism information and prevention documents issued by the University of Geneva.

I certify that this work is the result of personal work and has been written independently. The work is the responsibility of the author, in no way does the work engage the responsibility of the University of Geneva, nor of the supervising Professors.

I declare that all sources of information used are cited in a complete and accurate manner, including sources on the Internet. Other individuals and groups that have contributed to the research work involved in this paper have been clearly identified in the paper.

I am aware that the fact of not citing a source or not quoting it correctly is plagiarism and that plagiarism is considered a serious fault within the University, punishable by penalties.

In view of the above, I declare on my honor that the present work is original.

Signature：陈云翔 Date：2020/9

致谢

感谢 Harald HAU 教师、FAN Wenzhong 教师，对本文的研究工作所做的耐心细致而又非常专业的指导。

感谢包头农村商业银行股份有限公司对本文所做的研究所提供的的支持。

感谢包头农村商业银行王姗姗、王斌、段治龙、王实等对本文研究所提供的支持。

感谢朱璨提供的真诚支持。

摘要

当前在中国的 2000 多家地方中小银行在零售贷款方面面临着线上互联网巨头和线下同质化银行的激烈竞争。但这些中小银行并不是没有机会，而是缺乏能力。随着互联网科技的快速发展，中小银行可以充分利用相对容易获取的本地有效大数据资源获得相对的竞争优势。本文通过包头农商银行的一个线上“市民贷”个人信贷产品的应用实践，对中小银行所能获得的各类数据融合和建模进行了深入研究和分析，利用 XGBoost 和“关系图谱”等工具创新建立了本地化“市民贷”风控模型。在这个实践过程中，本文深入比较分析了基于公共数据集和本地数据集（叠加前者）的模型的优劣，得出了令人激动的结论—即基于本地数据集的优化 Local 模型具有更好的表现。这个基于本地数据集的“市民贷”风控模型于 2019 年在包头农商银行进行了实际的部署和应用，并且在客户获益及银行获益两方面均有巨大的提升。这样的结果充分表明，本文的的研究成果可以帮助中小型银行找到了一条切实可行的，通过金融科技赋能实现零售战略转型的成功之路。

关键词：市民贷款；信用风险模型；数据特征工程；XGBoost；知识图谱

目录

Disclaimer 1
致谢 2
摘要 3
List of Tables 6
List of Figures 7
“市民贷”—地方中小银行线上零售贷款业务应用研究 8
1. 介绍 8
1.1. 包头农商银行所解决的困难和问题 8
1.2. 文献综述 9
1.2.1. 市民贷款信用风险特征研究 9
1.2.2. 对市民贷款信用风险评价方法的研究 11
1.2.3. 综合评述 12
2. 方法论 13
2.1. 2019 年之前包头农商银行零售贷款的困难和问题 13
2.2. 可以利用的资源 16
2.3. 数据源 17
2.4. 利用外部公共大数据和本地数据建模 20
2.5. 发现 21
3. 包头农商银行建模之前零售贷款分析 23
4. 分析 24
4.1. Public 数据集和 Local 数据集数据准备 24
4.2. Public 数据集和第一个优化评分模型 26
4.2.1. XGBoost 建模方法 26
4.2.2. 数据样本 26
4.2.3. 特征加工 27
4.2.4. Public 模型建立与评估 31
4.3. Local 数据集和第二个更加综合的模型 34
4.3.1. 本地数据 34
4.3.2. Local 模型建立与评估 39
4.4. 本地信息带来的增值价值 42
4.4.1. 通过率-坏账率分析 42
4.4.2. Lift 分析 42
4.5. 其它竞争优势 44

4.6. 2019 年 Local 模型放款和财务分析.....45
4.6.1. 2019 年新产品运行基本情况.....45
4.6.2. 2019 年不良与 2018 年不良的对比分析.....46
4.6.3. 新的产品在财务上的表现.....47
4.6.4. 市民贷新产品在客户时效性的表现和对比.....48
4.6.5. 对老客户的分析.....49
4.6.6. 将来的进一步研究.....49
4.6.7. 小结.....49
5. 结论.....50
参考文献.....52
作者简历.....55

List of Tables

表 1：外部互联网网络银行产品分析表（2018 年）.....13
表 2：本地其它银行产品分析表（2018 年）.....14
表 3：原有作为风控判断的数据列举（2018 年）.....15
表 4：包头农商银行未引入外部数据前，零售贷款累计的不良表现（2018 年）.....15
表 5：包头农商银行不同机构个人贷款不良表现（由于授信决策标准不统一，体现在同一机构的不良表现忽高忽低）.....15
表 6：展业成本分析（2018 年）.....16
表 7：展业拓客时间成本分析，客户经理拓客和办手续的时间对比（2018 年）.....16
表 8：各类大数据数据数量分析表（2018 年）.....19
表 9：数据源 1 字段分析表（2018 年）.....19
表 10：数据源 2 字段分析表（2018 年）.....20
表 11： 2 个不同数据源建模成效对比分析表.....21
表 12：2016,2017,2018 包头农商银行个人信贷数据表.....23
表 13：2016,2017,2018 包头农商银行个人信贷成本和办理时效数据表.....23
表 14：训练集测试集好客户及坏客户占比及详情.....32
表 15：模型中前 10 的关键变量，以及变量的数据来源和相对重要程度.....32
表 16：Public 模型在测试集上，用户区分能力分析表.....34
表 17：147 个本地特征举例表.....35
表 18：10 个没有价值的特征表.....36
表 19：经过知识图谱抽取后的 14 个有效特征表.....38
表 20：Local 模型中前 10 的关键变量.....39
表 21：Public, Local^{-14}, Local 模型 KS，AUC 分析表.....41
表 22：Local 模型在测试集上，用户区分能力分析表.....41
表 23：Public 模型与 Local 模型 Lift 表.....43
表 24：营销模型运营分析表（2018 年）.....44
表 25：基于 Local 模型的市民贷产品要素表（2019 年）.....45
表 26：2019 年市民贷投放情况.....46
表 27：各机构不良表现更均衡.....46
表 28：各机构不良整体下降.....46
表 29：全行不良整体下降（2019 年较 2018 年下降了 62.8%）.....47
表 30：信用类贷款两种模式的成本对比（10 万元为例）.....47
表 31：抵押类两种模式的成本对比（10 万元为例）.....48
表 32：2019 年新增贷款成本下降分析表.....48
表 33：市民贷新产品在客户时效性的表现和对比.....48

List of Figures

图 1：数据准备过程 25
图 2：年龄 IV 值 29
图 3：近 4 个月的申请次数 IV 值 30
图 4：存款月均余额 IV 值 31
图 5：Public 模型在测试集上，KS，AUC 图 33
图 6：家庭收入 IV 值 35
图 7：本地农户知识图谱 37
图 8：直系亲属的家庭净收入 IV 值 38
图 9：Local 模型在测试集上，KS，AUC 图 40
图 10：Local^{-14} 模型在测试集上，KS，AUC 图 40
图 11：通过率-坏账率分析图 42
图 12：Lift 分析图 43
图 13：Public 模型与 Local 模型累计提升对比分析图 44
图 14：营销模型建模流程图 45

“市民贷”—地方中小银行线上零售贷款业务应用研究

1. 介绍

1.1.包头农商银行所解决的困难和问题

当前在中国，类似包头农商银行的地方中小型银行有 2000 多家（李杨 2018），这些银行在零售贷款方面，同时面临着线上和线下的激烈竞争（孙国峰 2019）。

在线上，新兴互联网巨头推出的如借呗、微粒贷等零售贷款产品，借助于自身积累的海量线上交易大数据（程华 2018），具有贷款申请速度快（一般 20~30 分钟完成），金额小（平均 5000 元），手续简单、体验良好等特点，造成传统客户，尤其是年轻客户不断流失。

在线下，银行间同质化竞争压力也在不断增强，由于没有更好的依赖于大数据的风险控制手段，中小型银行只能集中于大金额，基于身份的优质人群（如企事业单位白领人群)，依赖于传统的网点，无法形成独有的竞争优势和竞争壁垒，零售贷款只能打价格战，业务发展非常缓慢，且资产质量无法保证（卞维林 2017）。

但中小银行并不是没有机会，而是缺乏能力。随着互联网科技的快速发展，中小银行可以利用本土独立法人、有效数据获取相对容易等优势，把握以下机遇，实现新的发展：1）可以获取在本地公共企事业单位（如社保、个税、公积金、民政、通信运营商等）聚集大量的本地公共数据；2）客户尤其是年轻客群能够普遍熟练使用各种手机智能应用，同时这些应用产生大量的个人数据也能够与本地数据融合产生新型的金融价值；3）存在大量的本地互联网应用场景，例如，医院 APP，社区 APP，本地 B to C 电商，一方面能够给中小型银行提供应用场景服务，另一方面本地中小银行也能够获取这些应用所留存的大量的本地场景数据；4）实际调查表明，即使是互联网金融快速普及，线上业务对本地客户的渗透率仍然低于 40%[1]，本地中小银行仍然还有参与的机会。

因此，对这些地方中小银行而言，如何面对线上和线下的激烈竞争，以及如何

[1] 根据包头农商银行在本地的调查数据得出。

抓住在本地出现的可以利用的大数据建模机会构成了本文的研究背景。

而本文的研究问题则为：本地银行如何通过创新利用本地大数据资源建立本地化“市民贷”[2]风控模型，从而战胜金融科技行业的竞争对手。

本文论述了包头农商银行于 2019 年开展的一个线上“市民贷”个人信贷产品的应用实践，对中小银行所能获得的各类数据融合和建模进行了深入研究和分析，利用 XGBoost 和“关系图谱”等工具创新建立了本地化“市民贷”风控模型。

在研究问题和实践过程中，本文深入比较分析了基于公共数据集和本地数据集（叠加前者）的模型的优劣，得出了令人激动的结论—即基于本地数据集的优化 Local 模型具有更好的表现。这个基于本地数据集的“市民贷”风控模型于 2019 年在包头农商银行进行了实际的部署和应用，并且在客户获益及银行获益两方面均有巨大的提升。

这样的结果充分表明，本文的的研究成果可以帮助中小型银行找到了一条切实可行的，通过金融科技赋能实现零售战略转型成功之路。

1.2.文献综述

1.2.1. 市民贷款信用风险特征研究

贷款特征对本次研究至关重要，建模首先就需要从特征的选取和优化开始。本文已经注意到已经存在一些对不同特征集合的研究（戴宇 2018），一些通用的划分方式包括：从贷款的基本特征、内在特征和外部延伸特征三个大的方面来划分。其中贷款基本特征是指贷款金额、期限、利率、用途和 还款方式等贷款要素。内在特征是指市民的年龄、性别、住房、职业和工作等自 身内在的影响因素。外部延伸特征是指外部记录、放款机构、担保和政策等其 他外部的影响因素。通过对不同的特征集合进行划分，使得市民贷款信用风险评估的研究工作能够更加直观和便于理解。

[2] 这里面的市民是指在包头地区固定居住 1 年以上的 18 岁到 60 岁之间的成年人，职业、性别等均不限。

1.2.1.1.对基本特征的研究包括

Jose A.G Baptista (2006)研究得出得出农户贷款信用风险影响指标还有以下几种:贷款数额、贷款期限、贷款用途、贷款人有 无违法记录、经营理念、经营水平。RubananMahjabeen (2008)通过分析孟加拉国 小额贷款，认为会对贷款风险产生影响的有贷款总额、贷款周期、贷款人拥有的耐用商品价值等要素。

吕京娣，吕德宏(2011)在对农户还贷因素的研究中得出贷款额度显著影响农户贷款的还款行为，并起到正向作用。张润驰、杜亚斌、荆伟等(2017)研究发现： 信用水平指标与农户的真实违约情况关联不显著，意味着当地信贷机构对贷款农户的贷前内部信用评级不能有效地预测农户的信用风险：利率、性别、婚姻状况、职业、教育等微观指标对信用风险有较大影响。

1.2.1.2.对内在特征的研究包括

市民的内在因素相对贷款要素而言基本上是属于同一个层面，但已经体现出了市民的一些特性。所以从研究层面来看，稍微会比贷款要素指标的研究略微的少一些。 而国内外的研究基本上符合总的结构。

20 世纪 30 年代杜兰德构建了包括年龄、性别、居住稳定性、职业、工作稳定性 等 9 因素消费信贷评分体系（大连理工大学迟国泰课题组. 2010，崔健. 2005）。James copestake (2007)通过对金融机构进行问卷 调查，得到农户贷款风险的影响因素主要由贷款人简况状况、贷款人性别、贷款人年 龄、家庭劳动人数、家庭净资产等。SJha 和 KSBawa (2007)通过研究印度小额 贷款的案例，认为影响贷款风险的有文化程度、家庭收入、法律约束、固定资产等。

陈良维(2008)运用决策树算法，并从自然情况、家庭情况、信用情况、分别选取年龄、性别、家庭收入、贷款历史、司法记录等 15 个指标家里农户贷款评价体系 。崔军扬(2011)从吉林省农村信用社抽取 6000 份农户样本，对农户贷款违约影 响因素进行了二元回归分析，结果发现户主性别、年龄，受教育程度，家庭收入与支 出对农户违约行为起显著作用 。郑兰祥，万雪(2014)通过 logit 分析，从安徽省肥西县农村小额贷款公司资信等级评定表等档案，选取了劳动力、年龄、文化水平等 13 个指标。在确定最优变量指标之前，先进行正态性假设检验、异方差检验及多重 共线性

检验，通过检验确定模型可用的指标。夏萌、赵邦宏、王俊芹(2015)通过 相关分析法研究结果显示,农户的受教育程度、家庭劳动力数量、经营状况、信誉状 况等对农户贷款是否违约有较大影响。

1.2.1.3.对外部延伸特征的研究包括

市民贷款外部延伸指标的研究受到各方面因素的影响，相对贷款要素和内在因素 而言都明显少很多，而且出现的会迟一些。一方面受外延指标的获取影响，另一方面也受到指标有效性的影响。

Hartarska.V 和 DenisNadolnyak (2007)对世界银行发放的小额贷款的情况进行 了分析，结果表明贷款人的技术和能力，有无担保等方面是贷款风险的主要影响因素．李正波，高杰，崔卫东(2006)根据实地调查的资料，提出了对应的年龄、性别、 劳动力人数、信用社信誉、服务等指标在内的 19 个评价指标。迟国泰(2009)通 过相关性分析、关联分析、聚类分析建立了一套包含经济、生态、社会、人的全面发 展和科学技术 5 个方面共 79 个指标的科学发展评价指标体系。魏强等(2016)结 合 P2P 网贷平台网络信息领域的特殊风险因素，考虑借款人主要为个人和小微企业，总共筛选了 18 个指标构建网贷指标体系。

1.2.2. 对市民贷款信用风险评价方法的研究

信用风险评价方法主要分为三大类型，主要是按照基础学科进行分类，分别为统 计学方法、运筹学方法和数据挖掘方法。本文应用到的 XGBoost 算法也属于数据挖掘 方法的一种。XGBoost(eXtreme Gradient Boosting)算法是 GradientBoosting 算法的高效实现版本，因其在应用实践中表现出优良的效果和效率，因而也被工业界广为推崇。

XGBoost 算法是 Chen 等(2016)改进梯度提升决策树模型提出的一种集成学习模型，该算法中的决策树具有先后关联，当前预测以上一轮的预测误差为基础，利用各轮预测误差迭代构建模型，提升预 测的准确性。

常用数据挖掘方法研究层面，Dutta and Shekhar (1988)将神经网络预测模型运用于债券信用评级方面。温涛(2004)建立了基于 BP 神经网络的农户信用评级 模型。

并对重庆市 150 个农户的信用状况进行了实证研究。徐佳娜，西宝(2004) 将人工神经网络信用风险评估技术与层次分析法相结合，建立了商业银行信用风险评 估AHP-ANN 模型。蔡丽艳(2011)等采用决策树方法对某信用社的农户小额贷 款信用风险进行评价。张涛(2017)以农户贷款意愿为目标，农户户主特征、家 庭特征等 8 个因素为自变量建立决策树分类模型，以为农村金融机构提供有价值的农 户分类规则。结果表明：农户家庭人口、土地面积、户主受教育程度等 7 个因素的不 同水平取值可以组成 10 个不同属性的分类规则。

其他数据挖掘方法研究层面，王春峰等（2005）运用蚁群算法（modified ants algorithm.MAA）对商业银行的实际案例进行了实证研究。韦艳玲（2009）采用模糊聚类方法，将贷款农户划分成具有不同信用特征的群组，农村信用社进一步深入分 析其信用风险提供理论支持。

知识图谱(Knowledge Graph)又称为科学知识图谱，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。知识图谱的概念演化经过了语义网络、本体论、Web、语义网、链接数据等阶段，并由 Google 在 2012 年提出的，Google 希望通过知识图谱构建下一代的搜索引擎，从而优化搜索结果。雷丰羽（2018）在 2018 年，通过知识图谱高效直观刻画拟授信主体间的关联网络,借助知识图谱技术创新、全维度对主体进行画像,立体复现主体的真实状况.借助知识图谱可以实现之前传统风险管理方案诸多难题和不足,更好帮助金融机构进行风险管理。于小洋（2018）于 2018 年，通过知识图谱技术的通用构建框架，提出了以实体获取和实体关系抽取为主要手段的关系图谱构建方案。设计并实现了一个以资本市场人物和企业关系图谱为主要数据模型的信息系统。

1.2.3. 综合评述

本章主要从信用风险评价特征选取、风险评价方法两个方面对个人贷款信用风险评价相关的文献进行了研究和分析，对目前国内外的研究有了较为深入的了解。应该说前人的研究取得了较多宝贵的成果和总结了丰富的经验。从不同研究视角和研究方法方面丰富了研究的成果，对个人信用风险评价的研究和应用都起到了重要的参考作用。

从文献研究的内容中不难发现，当前个人风险评价指标研究主要集中在个人贷款基本要素、内在因素的研究。评价信息来源局限、评价维度少和针对性不足的问题普遍存在。而随着互联网技术的不断发展，每个人数字化生存的不断拓展，针对个人客户的数据来源愈加广泛，种类愈加完整。因此也就能够以更加客观的方式，在更好的数据环境下对个人贷款信用风险评价进行更深一步研究。而本文应用外部大数据资源和本地大数据资源来对个人信用风险评价进行深一步研究，也正是基于这个大的前提。

2. 方法论

2.1.2019 年之前包头农商银行零售贷款的困难和问题

在 2019 年本文实践之前，包头农商银行零售贷款主要面临的问题有以下几个方面：

1、外部互联网网络银行开始渗透本地市场。例如微粒贷、借呗、好人贷，经市场调研，这些产品在授信人群、审批效率、利率、件均、客户便宜程度方面均有较好的表现，使本地农商行在小微授信和个人消费贷款方面收到较大的挑战。

机构	产品名称	授信人群	审批效率	年利率	件均	客户便宜程度	优势
招联	好期贷	芝麻信用分较高的消费人群	2 分钟	15%-36%	12,000	★★★	填写内容简单；2 分钟内给出额度。
新网	好人贷	客群较广，普通工薪、白领。	2 分钟	10.8%-28%	3,500	★★★	每步均有营销动作；填写内容较简单，多为选择性的选项；2 分钟内出额度。
平安普惠	新一代	线下推广的小微企业主	首次出额 2 分钟提高额度 30 分钟	7.5%-24%	135,000	★ ★	金额较大，利率区间大，客户经理全程服务，维护.。

表 1：外部互联网网络银行产品分析表（2018 年）

2、本地其它银行抢占市场。目前本地各类商业银行抢占小微授信人群，由于资金成本低，对客贷款利率偏低，从而抢占大部分的头部客户，例如建行快贷、邮储极速贷、蒙商银行商赢宝、金谷村镇银行农户产品，在授信人群、审批效率、利率、件均、客户便宜程度方面均有较好的表现。使得农商获得的客户质量相对较低，存量市场被逐步蚕食，以农区信贷市场为例，包头农商行的市场占有率由过去的 65%（2015 年数据）年降至 48%（2019 年数据）；

机构	产品	授信人群	审批效率	年利率	件均	客户便宜程度	优势
建行	快贷	在建行有金融资产及代发客户	5 分钟	4.3%-5%	5,000	★★★	审批速度较快，利率优势十分明显，线上完成操作。
邮储	极速贷	小微商户	15 分钟	7.5%-10%	35,000	★★	额度较高，最高可至 20 万，审批速度较快。
蒙商银行	商赢宝	线下推广的小微企业主	七个工作日	7.5%-12%	155,000	★★	金额较大最高授信可达 100 万，产品灵活，客户经理全程服务，维护.。
金谷	农户贷款	农区客户	三个工作日	7%-9%	85,000	★★	金额高，审批速度快，利率低。

表 2：本地其它银行产品分析表（2018 年）

3、外部数据获取不足，坏账较高。2018 年之前，包头农商银行没有引入大数据资源进行风险评估，单一依靠央行的征信记录，作人为主观的判断，且标准不统一。造成信贷资产质量参差不齐，且无法评判关键影响因素。在发现风险时，不能步调一致的采取有效的风控措施。不良率较高，难以有效的控制。依靠人工调查，主观经验审批，零售贷款不良率表现为 3.2%。

人行数据	本行自有数据
近 3、6、12 个月逾期累计次数	客户评级
命中近 2 年特殊交易类型存在担保人代还、以资抵债	AUM 余额、存款余额、理财余额
信用卡近 3、6、12 个月逾期累计次数	资产信息
近 6、12、24 个月出现 1、2、3+的次数	负债信息
当前贷款、信用卡逾期金额	经营信息
最近 2、3、4、6 个月内的贷款审批、信用卡审批查询次数	收入信息
未结清房贷月供和	家庭成员信息

表 3：原有作为风控判断的数据列举（2018 年）

时间	表现
2018 年第一季度	2.85%
2018 年第二季度	3 .20%
2018 年第三季度	3.11%
2018 年第四季度	3.20%

表 4：包头农商银行未引入外部数据前，零售贷款累计的不良表现（2018 年）

时间	A 支行	B 支行	C 支行	D 支行	E 支行	F 支行
2018 年第一季度	1.70%	3.45%	3.25%	2.36%	1.11%	1.89%
2018 年第二季度	2.28%	1.76%	1.16%	3.13%	3.32%	2.28%
2018 年第三季度	1.14%	3.36%	1.27%	1.44%	3.17%	1.12%
2018 年第四季度	2.85%	1.37%	2.54%	1.55.1%	1.77.5%	3.55%

表 5：包头农商银行不同机构个人贷款不良表现（由于授信决策标准不统一，体现在同一机构的不良表现忽高忽低）

4、展业成本较高，人力资源占用高。小额高频的贷款，是包头农商银行个人信贷的主要定位，在发放贷款时，依靠劳动密集型的方式进行作业，人员投入大，贷款操作成本较高。目前客户经理人均管户数达到 300 笔，但人均管户金额仅为 1890 万，管理能力有限，不利于新增贷款的扩展及贷后管理工作。增加人员，又会提高展业成本，在成本控制和业务扩展方面陷入两难境地。

项目	调查成本	资料成本	办手续的成本	其它费用（公证、评估）	合计（10 万元为例）
信用、保证贷款	150 元/笔	20 元/笔	0	授信金额的 0.3%	470 元
抵押贷款	250 元/笔	20 元/笔	80 元	授信金额的 1.3%	1650 元

表 6：展业成本分析（2018 年）

项目	客户经理拓客	办手续的时间
信用、保证贷款	5 个工作日	3 个工作日
抵押贷款	10 个工作日	10 个工作日

表 7：展业拓客时间成本分析，客户经理拓客和办手续的时间对比（2018 年）

5、线下采集数据较多，没有充分的利用。身份信息、教育信息、工作信息、家庭信息、社会关系等申请表信息是农商行的数据优势，零散在 14 个不同的系统里，没有整合和量化的能力，过去的坏账客户的特征没有量化提炼，不具备对个人信贷投放工作的指导作用。

2.2.可以利用的资源

解决问题的关键依然是大数据，大数据毕竟是一切分析和建模的基础。因此对包头农商银行而言除了尽量获得与外部竞争对手同样的大数据资源外，尽量挖掘本地大数据资源，在本地形成竞争优势就是一条可行的道路。

幸运的是，包头农商银行毕竟是一家在本土深耕多年的银行，在与本地各类机构客户（例如：本地社保、公积金、通讯记录、农户建档信息等）进行长期合作过程中，可以通过技术手段合法利用这些机构所具有的各类针对个人客户的数据资源，并且使其量化和标签化，这些分散在本地各个机构的客户信息可以通过身份证号，手机号等关键字段进行关联以构成本地大数据资源。

根据在本地的调研情况，包头农商银行的外部竞争对手和本地的竞争对手都尚未对本地大数据资源进行大规模实际应用。

本地大数据具有如下价值：

第一，本地特色的外部大数据，例如本地电商交易、社交数据、移动通讯信息等，在本地客群的应用会更加精准。

第二，具有本地特色的场景大数据，例如本地公积金数据、本地小微企业纳税记录、医疗、教育等方面的数据的获取，本土银行更有优势。包头农商银行在这些各类场景中能够深挖客户并控制风险，这是线上互联网信贷机构所不具备的优势。

第三，本地机构利用网点优势，能够通过人工采集获取大量数据，并且这些数据能够沉淀下来，这是其他机构所不具备的。包头农商行经营历史较长，在当地客户根基较好，特别是农户数据收集和建档较为完整、全面。特别是，在对村民社会关系的掌握方面，例如家庭和谐程度、父子、兄弟姻亲关系等情况，包头农商银行具有天然的优势。

同时借助于国内这些年快速发展的各类全国性的大数据公司，包头农商银行也可以像外部互联网金融机构一样，充分获得本地客户的外部大数据资源（例如：外部多头借贷信息、电商交易记录、社交信息等）。

这样本地大数据资源+外部大数据资源，构成了本文可以利用进行建模的数据资源，使得本次论文研究有了坚实基础，并且在此基础上对所建立的风控模型进行持续的优化和评估，最终的目标是针对竞争对手在本地形成竞争优势。

2.3.数据源

本文研究的目标客群是指在内蒙古自治区包头市，这个城市是中国一个比较具有典型意义的 4 线城市，工业、农业基础较好，类似这样的城市全国有 300 多座，涉及人口 3~4 个亿。在这个城市中，居住时间超过一年的常住人口，年龄从 18 岁到 60 岁，职业性别不限，有 260 万人左右，其中有农户 40 万人左右。本地的市民（包括农户）是农商银行服务的主要对象，特别是农商行依靠自身的网点优势，在本地农区占有较大份额的市场，在农户数据获取方面具有天然的优势。

得益于整个国家和包头市多年来外部大数据环境的发展，包头农商银行能够获得的，并且能够持续获得 3 年以内的数据来源，主要分为以下六大类：

1、银行内部数据源（包括 114,356 个人，具体有：姓名、身份证号、电话号码、开户信息、存款表现、贷款表现、是否黑名单等）；这些数据约 96,516,464 个字段。

2、外部金融机构数据源（包括本地可获取的 114,356 个人的央行征信信息，具体有：个人基本信息，在各个持牌正规金融机构的贷款表现等）；这些数据约 77,533,368 个字段。

3、互联网外部大数据源（包括本地约 114,356 人，具体包括：黑名单和多头借贷信息、电商信息、社交信息、移动通信和应用信息等）；这些数据约 552,568,192 个字段。

4、本地外部大数据源（包括本地约 64,357 人，具体有：客户黑名单和多头借贷信息、客户信用分、社保、个税、公积金、通信运营商、民政等）；这些数据约 2,188,138,000 个字段。

5、本地场景大数据源（包括本地约 59,453 人，具体有：医疗、教育、交通出行、电商等）；这些数据约 1,367,419 个字段。

6、农民大数据源（包括本地约 12,468 人，具体有：家庭关系、邻里关系、土地亩数、年收入等）；这些数据约 10,522,992 个字段。

项目	整体数据			本行数据		
	户数	单户信息量	信息总条数	户数	单户信息量	信息总条数
行内数据	612,543	18	11,025,774	114,356	844	96,516,464
外部金融数据（征信）	825,436	156	128,768,016	114,356	678	77,533,368
互联网大数据	1,414,356	4,832	6,834,168,192	114,356	4,832	552,568,192
本地外部大数据	2,403,552	34,000	81,720,768,000	64,357	34,000	2,188,138,000

本地场景大数据（POS电商数据）	1,825,088	23	41,977,024	59,453	23	1,367,419
农民大数据	63,564	156	9,915,984	12,468	844	10,522,992
合计	合计		88,746,622,990	合计		2,926,646,435

表 8：各类大数据数据数量分析表（2018 年）

（备注：整体数据是指基于本地市场该数据的总量，本行数据是指基于本行的存量客户数据，可直接获取的数据）

上述这6部分数据，根据本文建模的需要主要被划分为2大部分：

- 数据源1（公共大数据）：银行内部大数据；外部金融机构数据；互联网外部大数据；

这部分数据源对任何一家外部银行的竞争对手而言，都可以获得，相对来说是公共资源；包头农商银行的主要外部竞争对手都可以获得这样的数据。

序号	数据源 1	字段描述
1	行内数据	开户日期、存款表现、经营数据...共 55 个字段。
2	外部金融数据（征信）	近 3、6、12 个月逾期累计次数；命中近 2 年特殊交易类型存在担保人代还、以资抵债；信用卡近 3、6、12 个月逾期累计次数；近 6、12、24 个月出现 1、2、3+的次数；当前贷款、信用卡逾期金额；最近 2、3、4、6 个月内的贷款审批、信用卡审批查询次数、未结清房贷月供和、未结清贷款笔数、金额；信用卡月还款额...共 156 个字段。 ...
3	互联网外部大数据	通过身份证查询法院执行人黑名单、通过身份证查询非银(含全部非银类型)高风险、按手机号查询，近 6 个月在非银机构-持牌网络小贷机构申请机构数、按手机号查询，近 6 个月在非银机构申请次数...共 80 个字段。 ...

表 9：数据源 1 字段分析表（2018 年）

- 数据源2（本地大数据）：本地外部大数据；本地场景大数据；农民大数据；

包头农商银行所获得这部分数据源，相对数据源1而言，是本地化数据，这些数据具有鲜明的包头本地客户特征，而外部互联网银行无法或者很难获得。这部分数据源是本次研究的重点基础。

表 2.9 数据源 2 字段分析表（2018 年）

序号	数据源 2	字段描述
1	本地外部大数据	个人运营商信息、疑似养卡（异常使用号码）标识、近 3 个月账单平均金额、手机号码归属地市标签、手机号码近 6 个月总停机频次标签（元）...共 50 个字段。 公积金缴存基数、公积金月缴存额等。 个税缴纳基数、个税缴纳额。 个人汽车拥有情况。 个人婚姻状况。 个人户籍状况。 ...
2	本地场景大数据	销售金额、客单数、客单价、产品编号、交易时间...共 14 个字段。 ...
3	农民大数据	基本情况、所在村组、家庭主要成员、拥有房产情况、拥有土地情况、拥有农机具情况...共 156 个字段 ...

表 10：数据源 2 字段分析表（2018 年）

2.4.利用外部公共大数据和本地数据建模

本次研究的关键就在于:

1、利用外部公共大数据（数据源 1）进行建模，构建 Public 模型，实现与竞争对手的同样的表现。

2、在外部公共大数据建模的基础上，引入本地数据源（数据源 2），两个数据源结合后，进行联合建模，构建 Local 模型，并利用知识图谱等工具创新解决联合建模过程中特征选取的困难和问题，目标是使得联合建模调优后的结果显著优于前者。

3、对联合建模进行成本收益分析，判定联合建模在经济上的可行性。

选取的建模方法：

本次研究选取的建模方法是 XGBoost，主要是基于如下理由：

1、保证算法结果稳定的同时还可高效处理大规模数据。

2、容易工程化，在实际应用中容易实现。

3、算法效果好，解决实际问题能力强。

4、对数据的缺失和错漏容忍度高。

2.5.发现

在本文的研究中，第一方面：通过对两个数据源，针对已有的数据集（30774 条包头农商银行 2018 的样本数据）利用 XGBoost 进行建模对比，建模的成效十分显著：

项目	数据源 1 （Public 模型）	数据源 1+数据源 2 （Local 模型）
通过率一致情况下 （通过率=80%）	坏账=1.9%	坏账=1.1%
坏账一致的情况下 （坏账<=1%）	通过率=30%	通过率=75%
lift 分析 （Decile=10）	3.62	5.56

表 11： 2 个不同数据源建模成效对比分析表

在同样的通过率条件下，Local 模型的坏账率要低于 Public 模型；在同样的坏账率条件下，Local 模型的通过率要高于 Public 模型。如果把坏账率控制在 1%以内，Local 模型的通过率能达到 75%，而 Public 模型的通过率只能达到 30%；如果把通过率控制在 80%，Local 模型的坏账率为 1.1%，Public 模型的坏账率为 1.9%。

从 lift 分析来看，Local 模型的增益优于 Public 模型。

在本文的研究中，第二方面，利用新的 Local 模型，建立市民贷产品，在 2019 年一年进行了运行，实际获得的收益非常的显著：

系统上线前（2018 年）和系统上线后（2019 年），比较如下：

客户获益分析：（1）一笔小额贷款的获取时间，从 3 天减少到 3 个小时；（2）笔数：笔数从 2018 年的 73788 户增至 2019 年 197233 户，客户覆盖面更广；（3）利率：客户的平均利率从年化 10.96%降到年化 8.72%；

不同类型客户分析：（1）能够覆盖本地 70%的人群，只有是本地人，没有不良记录都可以借贷 8000 元以上；（2）其中能够覆盖 80 万原来没有借贷的白户人群；（3）原有的 80 万有征信客户的平均借贷金额可以从 3000 元调整到 20000 元；

银行获益分析：（1）放贷余额 2019 年达到 15.35 亿；（2）额度：客户的平均额度达到了 43728 元，风险进一步分散，显著优于过去（96683 元），向互联网银行靠近（5000 元）且优于本地其它银行（72187 元）；（3）当前不良和预计不良低于 1.2%；（4）预计净利润可以达到年化 4%，相对于传统银行零售业务净利润提升 203%；（5）成本：成本快速下降，降低了 60.04%；

在本文的研究中，第三方面，在具体的研究创新上，本文主要做了两个重要创新：

1、在数据源 1 和数据源 2 的基础上，应用 XGBoost 建模工具建模，并取得了非常好的成效。

2、相对于数据源 1，本地数据源存在着较大的缺失、错漏问题。本文利用知识图谱对数据源 2 的本地信息进行清洗和通过权重的调整增强领域知识，经过清洗和调整后，数据质量得到了明显改善，模型的效果进一步得到了优化。

在知识图谱的应用过程中，知识图谱标准化了不同数据源数据的分析框架和预测过程，并对模型的建立有三点补充和完善：

（1）可以对知识图谱中的任一节点进行预测和分析。例如，要研究家族的诚信情况，只需选取与一个人家庭亲戚有关联“关系”的所有节点作为初始变量集；

（2）可以将知识图谱中的结构信息带入量化模型中。例如，使用主成分分析降维，可以考虑把每个包含关系下的分项变化汇总为一个主成分等;

（3）可以在传统量化模型的基础上叠加领域行业观点。例如，农产品价格好，那么必然客户的还款意愿和能力高。

利用知识图谱，例如，一个显著的发现和获益是：经过计算，“直系亲属的家庭净收入”的 IV 值为 0.342。也就是说，“直系亲属的家庭净收入”这个维度具有很高的建模价值。

3. 包头农商银行建模之前零售贷款分析

在 2016 年、2017 年、2018 年，包头农商银行的个人信贷审批没有依托大数据，主要依靠人工的模式进行风险判断。这三年的信贷投放额度、信贷投放增量、贷款笔均、平均利率、贷款笔数、展业成本、办理速度、不良率等具体数据如下:

年	贷款余额（万元）	贷款增量（万元）	贷款户均（万元）	贷款平均利率	贷款户数	不良率
2016 年	677,651.04	14192.2	13.019	11.52%	52086.	3.31%
2017 年	684,257.61	6606.57	11.54	11.32%	59656	2.98%
2018 年	704,018.62	19761.01	9.60	10.96%	73788	3.20%

表 12：2016,2017,2018 包头农商银行个人信贷数据表

年	单笔展业成本		办理速度（工作日）			
	信用类	抵押类	拓客		办理手续	
			信用类	抵押类	信用类	抵押类
2016 年	450 元	1590 元	3 个	8.5 个	2 个	10 个
2017 年	450 元	1620 元	4.5 个	9 个	2.5 个	10 个
2018 年	470 元	1650 元	5 个	10 个	3 个	10 个

表 13：2016,2017,2018 包头农商银行个人信贷成本和办理时效数据表

截至 2018 年底，包头农商银行存量个人信贷共 73788 笔，贷款笔均 9.6 万元，贷款平均利率 10.96%，不良率 3.20%，这些客户分布在，全行 16 个零售产品中，主要结构为抵押类贷款占比 36%，信用贷款占比 64%。

观察从 2016 年到 2018 年三年的各项数据，可以看出：

1. 从 2016 年到 2018 年三年间，贷款余额共增长了 2.64 亿元，三年的平均增量为 1.35 亿，增长较为乏力；
2. 从 2016 年到 2018 年三年间贷款笔均逐年降低，贷款笔数逐年升高，单笔展业成本逐年增加，整体展业成本增长幅度超过投放规模的增长幅度，成本居高不下，但是贷款平均利率受市场因素的影响逐年降低，利润空间逐年降低；
3. 从 2016 年到 2018 年三年间，业务办理速度处于较长的办理周期，客户的满意度降低，外部市场的信贷竞争较为激烈，优质客户流失严重。
4. 从 2016 年到 2018 年三年间，零售贷款的不良率始终居高不下，展业成本提高、贷款利率下降，不良率处于高位，更加挤压了利润空间，个人信贷业务处于一个不可持续的发展状态。

因此，从 2018 年开始，包头农商银行引入了各类大数据集，并且从 2018 年的 73788 个人信贷客户中筛选出 30774 户数据完整、结构清晰且数据价值较高的客户作为样例，建立新的风控模型，推出市民贷新产品，采用新模式进行贷款的投放和管理。

4. 分析

4.1.Public 数据集和 Local 数据集数据准备

通过三类接口方式，包头农商银行构建了自己的数据集，接口与数据来源的关系如下图所示：

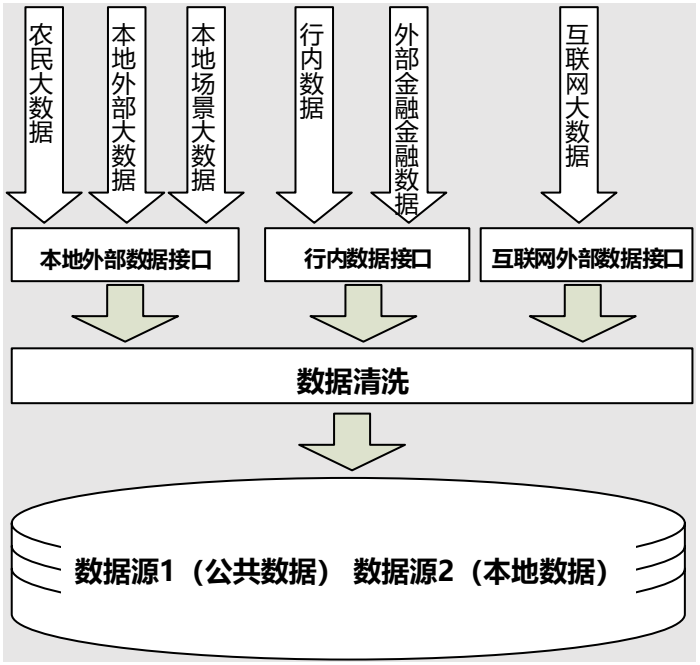

图1：数据准备过程

在建模工作中，对数据样本的探索性数据分析和数据清洗工作占到 70%的工作量，所以风险评价模型的构建主要工作还是对数据集中各字段标签的构建和确定。标签指标的确定过程中，通过变量基本信息分析、变量分布分析、变量频率分析等工作剔除极端值、非正态分布标签；通过数据清洗对数据进行重新审查和校验，目的在于删除重复信息、纠正存在的错误，并保证数据一致性。其中最重要的 2 个步骤如下，以实现对有效标签体系的构建工作，为建模做好准备。

4.1.1. 一致性检查

根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发现超出正常范围、逻辑上不合理或者相互矛盾的数据。简单的一致性检查如性别，年龄等，例如：同一身份证号码下，根据身份证号码的逻辑，校验并纠正数据源中的性别和年龄字段；对一些数据在不同来源值不相同的情况下，采用投票机制设立一致性检查，例如客户家庭人口，值有（2 人，3 人，3 人）三种情况，根据投票，选取 3 人。

4.1.2. 无效值和缺失值的处理

由于调查、编码和录入误差，数据中可能存在一些无效值和缺失值，需要给予适当的处理。常用的处理方法有：估算，整例删除，变量删除和成对删除。例如在对 POS 流水进行处理时，消费金额是必需数据，对无法获取该字段的数据，进行整列删除，对单笔金额超过 10 万元的数据，置换成特殊码（通常是 9、99、999 等）代表无效值和缺失值，同时保留数据集中的全部变量和样本。但是，在具体计算时只采用有完整答案的样本，因而不同的分析因涉及的变量不同，其有效样本量也会有所不同。这是一种保守的处理方法，最大限度地保留了数据集中的可用信息。

4.2.Public 数据集和第一个优化评分模型

4.2.1. 样本目标定义

4.2.1.1.贷前审批模型

“市民贷”产品的贷前审批流程，需要对申请用户的信用资质进行评估，并预测其未来逾期或者坏账的概率，以决定是否给该用户发放贷款。其中贷前审批模型是贷前审批的主要依据。

贷前审批模型是以申请人未来是否逾期作为目标变量，以申请人提交的信息和银行采集的其他信息为预测变量，构建的二分类模型。

4.2.1.2.贷前审批模型目标变量 Y 的定义

通过对存量样本的滚动率分析，确定目标变量定义为：

1）历史最大逾期天数小于 30 天为好客户，即 Y=0；

2）历史最大的逾期天数大于等于 30 天为坏客户，即 Y=1。

4.2.2. 数据样本

为使贷前审批模型具备较好的预测能力，数据样本选取及准备尤为重要。充足且有代表性的数据样本是模型建立的前提。数据样本选取了个人贷款业务在 2018 年发放贷款的 30774 例客户，其中坏客户 1002 例，占比 3.26%，好客户 29742 例，占比 96.74%。

4.2.3. 特征加工

根据特征的数据属性和衍生方法的不同，将特征大致分为两类。一类是基于业务逻辑的业务特征，另一类是利用数学变换、算法衍生、特征交叉与组合等方法生成的衍生业务特征。

4.2.3.1.业务特征

在本文 1.2.1 章节对业务特征的前期研究进行了综述。本文所研究的业务特征来源于实际业务场景中的数据，通过这些数据往往可以构造出大量的反应业务特点的特征。将常见的业务特征分为基本属性特征、基于各类场景数据的特征和关联信息特征三大类。

基本属性特征主要是对研究对象固有的性质和特点的描述，主要涉及身份信息、教育信息、工作信息等申请表信息，通过解析这类记录类信息得到可用于量化描述或分类的特征。

对于各类场景数据的特征，从各类场景得到的关于研究对象的详细的数据，这些量化数据可以用于分类建模。

关联信息主要是通过社交数据建立人与人之间的联系，借助知识图谱的方法，对群体或节点的关联路径深度、关系类型、关系权重、关系密集度、关联节点属性等指标进行计算提取，将复杂的关系网络可视化。

4.2.3.2.衍生业务特征

对业务特征进行数学变换、算法衍生、特征交叉与组合，衍生出具有新的含义的特征，更利于模型计算，从而提升模型的预测能力。根据衍生前后特征数量的变化将衍生方法分为 1-to-1 特征衍生、1-to-N 特征衍生和 N-to-N 特征衍生。

（1）1-to-1 特征衍生

单变量的函数变换：常用的变换函数有绝对值变换、平方变换、对数变换、指数变换以及倒数变换。

分箱：主要应用于对连续变量的离散化和多分类值离散变量的合并。离散化后的特征对异常数据有较强的鲁棒性，不易受极端值的影响；且能避免特征中无意义的波动对模型造成的影响，模型会更稳定。分箱方法主要有等距划分和等频划分，

其中等距划分：将变量的取值范围分为 k 等份，每一份为一箱；等频划分：将变量的观测值个数分为 k 份，使得每份包含大致相同的实例数量。

WOE 转换：WOE 转换是一种有监督的编码方式，将预测类别的集中度的属性作为编码的数值。通俗来讲就是特征在某一取值范围的时候对违约概率的一种映射。

（2）1-to-N 特征衍生

1-to-N 衍生方法指对单个特征进行处理输出多个新特征，主要方法有 One-Hot 编码和均值编码两种，它们都是用于对分类变量进行处理。

One-Hot 编码：主要应用于无序的分类变量，由于分类器往往会将此类数据默认为连续的有序变量进行处理，所以不能直接使用。使用 one hot 编码可以对类别进行“二进制化”操作，然后将其作为模型训练的特征。

均值编码：均值编码是针对高基数的类别特征进行处理，当类别特征的实例值过多时进行 One-Hot 编码容易引起维度灾难，使得模型效果降低。均值编码在贝叶斯的架构下，利用所要预测的目标变量，有监督地确定最适合这个定性特征的编码方式。它最大的特点是基于经验贝叶斯方法利用已知数据估算先验概率和后验概率，通过对先验概率和后验概率做加权平均计算最终的特征编码值。

（3） N-to-N 特征衍生

N-to-N 衍生方法指对多个特征进行处理输出多个新特征，主要方法有多项式变换和决策树算法衍生特征。

多项式的变换:主要是对现有特征进行多项式特征组合形成新的特征矩阵，形如，对 X=(x1,x2)进行 2 阶变换，输出结果为：(x1,x2,x1^2,x1*x2,x2^2)，常用于线性模型中达到非线性的效果。

决策树算法衍生特征:在决策树的系列算法中，每个样本都会落入一个叶子结点上，将叶子结点作为新的特征用于训练模型。树模型本身并不能产生特征，但可以利用其算法的特性产生特征组合。该算法在一定程度上弥补了人工组合特征费时费力的缺陷。

4.2.3.3.特征选取

对于上述的各类业务特征，在评估是否将某一特征纳入模型变量中时，需要综合考虑以下几方面因素，按优先级排序：

（1）符合逻辑且可解释；

（2）有较强的预测能力；

（3）与其他变量相关性较低；

（4）稳定且便于获取；

（5）合规，无法律或伦理的限制；

（6）与申请者相关，且不是金融机构的策略；

（7）去掉后会导致信息损失较大。

对于 30774 个建模样本，本地数据总共 5006 个字段，如何在这些字段中提取出模型要使用的特征变量比较关键。

本文采取的特征筛选方式有：

（1）缺失率太高的变量直接剔除，按 70%的阈值来剔除的；

（2）数值变量中所有值接近常量的变量剔除；

（3）按业务逻辑完全不可解释的变量直接剔除；

（4）分类变量中 unique 值大于 20（不包括 20）的直接剔除；

（5）IV 值小于 0.02 的变量直接剔除；

（6）若两个变量的相关度大于 0.7，剔除 IV 值较小的那个变量。

这里需要对 IV 值进行说明， IV 值衡量的是一个变量的信息量，其值的大小反映了变量对于目标变量的影响程度。一般来说，IV 值小于 0.02 的变量预测能力非常差，且有可能影响模型的稳定性，因此 IV 值低于 0.02 的自变量都会被直接剔除。

经过上述过程的筛选，剩下的特征字段为 284 个，其中比较有代表性的特征有：

（1）年龄

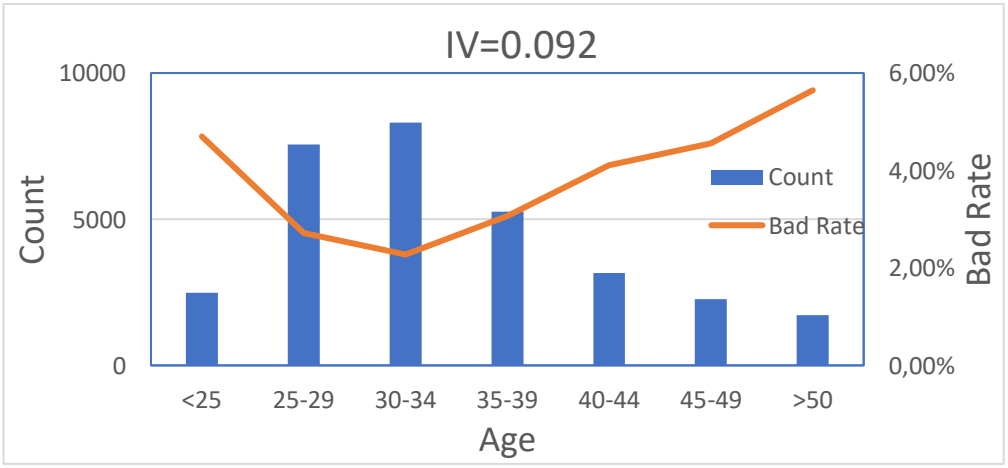

图 2：年龄 IV 值

年龄这个特征来自于用户申请表。对其 IV 值进行分析，首先对年龄做了离散化分箱处理，如上图所示，将其划分为 7 个分组，并计算每个分箱中样本的数量和坏账率。从图中可以看出，“市民贷”的客户主要年龄段为 30-34 岁，且这部分客户的坏账率最低，为 2.28%。整体来看，随着年龄的增加，坏账率呈现先降低后增加的“V”形。经过计算，年龄这个特征的 IV 值为 0.092。

（2）近 4 个月的申请次数

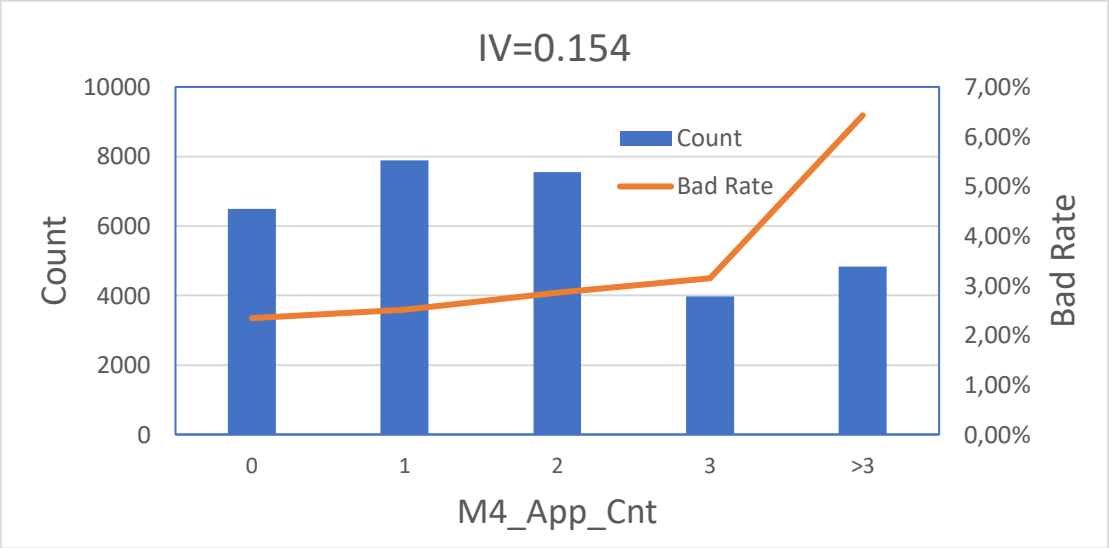

图 3：近 4 个月的申请次数 IV 值

“近 4 个月的申请次数”是互联网大数据类的特征，描述了申请人在近 4 个月内申请其他贷款的次数。从图中可以看出，随着“近 4 个月的申请次数”的增加，坏账率逐渐提升，特别是如果申请人“近 4 个月的申请次数”大于 3，则其坏账率远高于其他申请人。“近 4 个月的申请次数”的 IV 值为 0.154。

（3）存款月均余额

“存款月均余额”来自于行内自有数据。首先做离散化分箱处理，如下图所示，将其划分为 6 个分组，分别并计算每个分箱中样本的数量和坏账率。从图中可以看出，“存款月均余额”小于 0.5k 的用户坏账率最高，随着“存款月均余额”增加，坏账率呈现逐渐降低。经过计算，“存款月均余额”的 IV 值为 0.241。

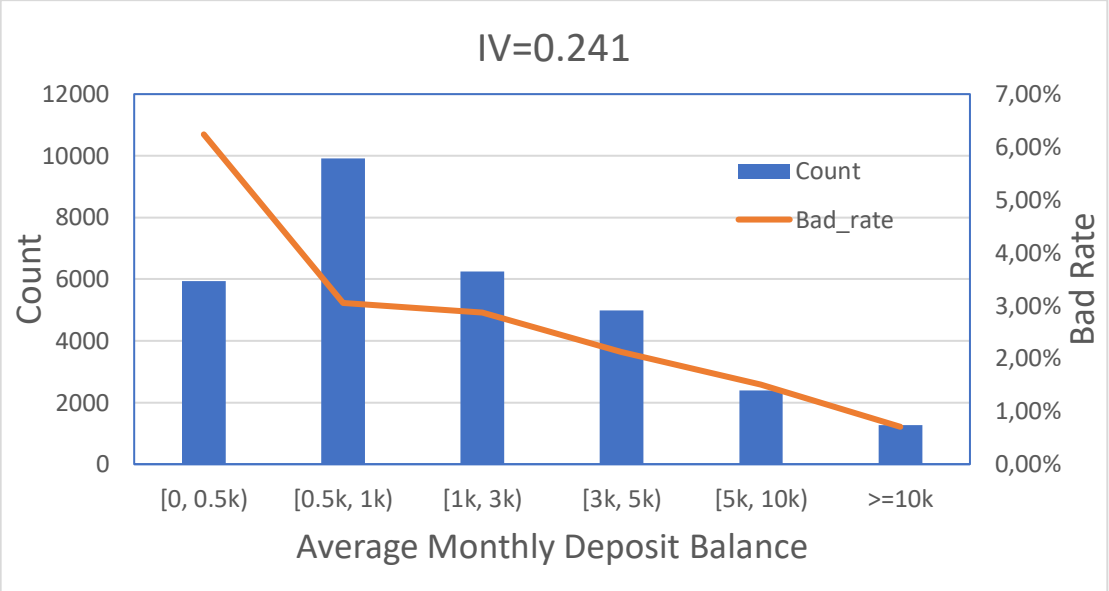

图 4：存款月均余额 IV 值

4.2.4. Public 模型建立与评估

4.2.4.1.XGBoost 建模方法

XGBoost(eXtreme Gradient Boosting)算法是 Gradient Boosting 算法的高效实现版本，因其在应用实践中表现出优良的效果和效率，因而也被工业界广为推崇。2014 年，陈天奇博士提出了 XGBoost 算法，XGBoost 可认为是在 GBDT 算法基础上的进一步优化。首先，XGBoost 算法在基学习器损失函数中引入了正则项，控制减少训练过程当中的过拟合；其次，XGBoost 算法不仅使用一阶导数计算伪残差，还计算二阶导数可近似快速剪枝的构建新的基学习器；此外，XGBoost 算法还做了很多工程上的优化，例如支持并行计算、提高计算效率、处理稀疏训练数据等等。

XGBoost 算法源起于 Boosting 集成学习方法，在演化过程中又融入了 Bagging 集成学习方法的优势，通过 Gradient Boosting 框架自定义损失函数提高了算法解决通用问题的能力，同时引入更多可控参数即可针对问题场景进行优化，最后通过工程实现方面细节优化，在保证算法结果稳定的同时还可高效处理大规模数据，可扩展支持不同编程语言。这些因素共同使它成为了主流机器学习算法之一。

4.2.4.2.Public 模型建立

在模型建立之前，需要将样本数据集随机划分为两份：样本的 70%作为训练集，用来对模型进行训练；样本的 30%作为测试集，用来对训练好的模型进行度量评估。特别的，在样本切分的时候，要确保两份数据中的坏账率基本一致。

以下为训练集测试集好客户及坏客户占比及详情：

	客户类别	数量	占比
训练集样本数量：21520 例	坏用户	709 例	3.29%
	好用户	20811 例	96.71%
测试集样本数量：9224 例	坏用户	293 例	3.18%
	好用户	8931 例	96.82%

表 14：训练集测试集好客户及坏客户占比及详情

在模型训练过程中，采用 5 折交叉验证的方法，对 XGBoost 算法的超参数 num_round、max_depth 等进行网格搜索调参，选择 roc 最大的一组参数作为模型的最终参数，完成模型训练。

变量名	数据来源	相对重要性
存款月均余额	行内数据	5.6%
近 6 个月逾期累计次数	外部金融机构数据（征信）	5.3%
近 4 个月的申请次数	外部金融机构数据（征信）	4.5%
偿债压力指数高	互联网大数据	3.9%
年龄	申请表	3.7%
命中账户状态为逾期	外部金融机构数据（征信）	3.1%
多次命中非银拒绝	外部金融机构数据（征信）	2.8%
近 3 个月夜间申请次数	互联网大数据	2.6%
理财余额	行内数据	2.5%
客户评级	行内数据	2.3%

表 15：模型中前 10 的关键变量，以及变量的数据来源和相对重要程度

上表列出了模型中前 10 的关键变量，以及变量的数据来源和相对重要程度。

4.2.4.3.Public 模型评估

模型评估用来评测模型的好坏。一般是在训练集上完成模型训练后，运用模型对测试集中的样本进行预测，通过计算模型预测的准确程度，来评测模型的好坏程度。对于二分类模型，常用的评价指标有 KS 和 AUC。

在测试集上，将模型预测的概率从小到大排列，分别以这些概率值为阀值，计算出不同的(FPR,TPR)值，然后以 FPR 为横坐标、TPR 为纵坐标，绘制出 ROC 曲线。KS 就是(FPR,TPR)二值组中，TPR 与 FPR 差的最大值；AUC 就是 ROC 曲线下的面积大小，能够量化地反映基于 ROC 曲线衡量出的模型性能。KS 和 AUC 越大，

说明分类器越可能把真正的正样本排在前面，分类性能越好。

经过计算，Public 模型在测试集上，KS=0.3637，AUC=0.7317。

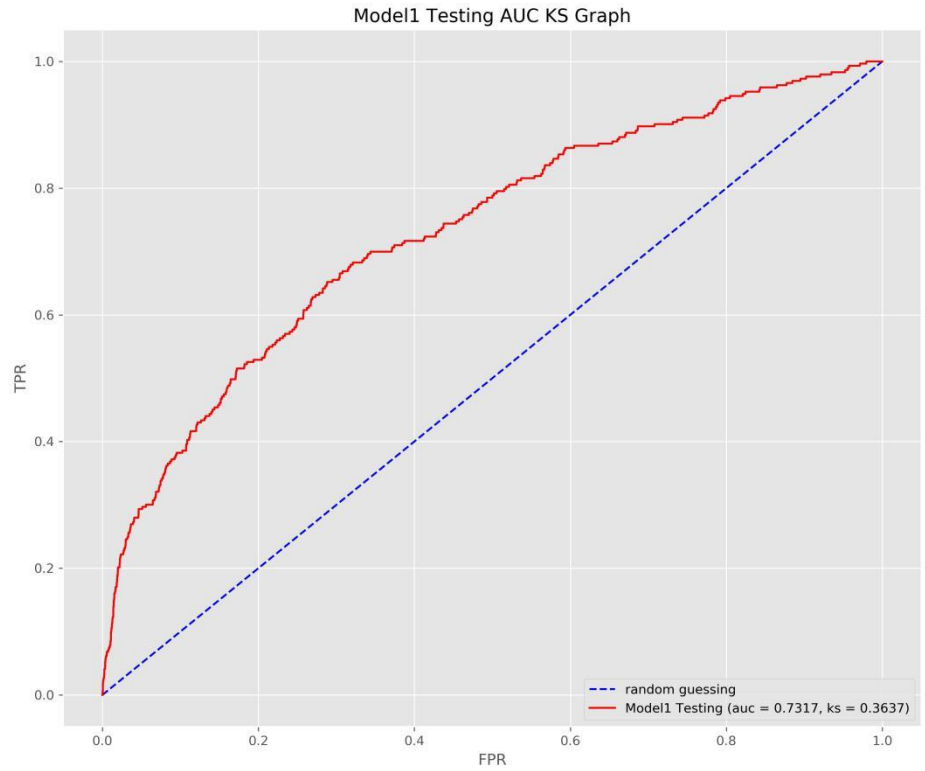

图 5：Public 模型在测试集上，KS，AUC 图

对于测试集中的样本，按照模型预测的概率从小到大排序后，等频划分为 10 组，每组包含 10%的样本。然后分别统计各组的用户数、好用户数、坏用户数，计算出各组的坏账率，如下表所示。从表中可以看出从第 1 组至第 10 组，坏账率呈单调递增的趋势。并且模型筛选出来的最优质的第 1 组客户，坏账率为 0.5%，远低于测试集的总体坏账率 3.2%；而在模型筛选出来最差的第 10 组，坏账率高达 11.5%，远高于总体坏账率 3.2%，说明模型具有较好的区分好坏用户的能力。

另外表中还计算了累计指标，包括累计用户数、累计好用户数、累计坏用户数、累计坏账率、累计好用户的占比、累计坏用户占比，可以反映出模型使用后的结果。假如要求模型的通过率为 80%，则需要对模型预测的用户风险从低到高排序，风险较低的 80%核准通过，其余的 20%拒绝。从表中可以看出，通过部分的坏账率为 1.9%，比不使用模型的总体坏账率 3.2%降低了 1.3%。在被拒绝的 20%用户中，误杀了 1-81.1%=18.9%的好用户，获得的收益是拒绝了 1-47.4%=52.6%的坏用户。

Decile	Population	Current	Overdue	Bad_Rate	Cum_Pop	Cum_Current	Cum_Overdue	Cum_Bad_Rate	Cum_Current_Rate	Cum_Overdue_rate
1	923	918	5	0.5%	923	918	5	0.5%	10.3%	1.7%
2	922	912	10	1.1%	1845	1830	15	0.8%	20.5%	5.1%
3	922	910	12	1.3%	2767	2740	27	1.0%	30.7%	9.2%
4	923	911	12	1.3%	3690	3651	39	1.1%	40.9%	13.3%
5	922	906	16	1.7%	4612	4557	55	1.2%	51.0%	18.8%
6	922	902	20	2.2%	5534	5459	75	1.4%	61.1%	25.6%
7	923	896	27	2.9%	6457	6355	102	1.6%	71.2%	34.8%
8	922	885	37	4.0%	7379	7240	139	1.9%	81.1%	47.4%
9	922	874	48	5.2%	8301	8114	187	2.3%	90.9%	63.8%
10	923	817	106	11.5%	9224	8931	293	3.2%	100.0%	100.0%

表 16：Public 模型在测试集上，用户区分能力分析表

4.3.Local 数据集和第二个更加综合的模型

4.3.1. 本地数据

4.3.1.1.本地特征

本地数据包括本地外部大数据、本地场景大数据和农民大数据。其中本地外部大数据是包头农商行跟本地电信部门合作授权获取的用户电信详单记录以及用户间的亲密程度级别数据；本地场景大数据主要来自于包头当地的支付平台（电商 POS 流水），是用户在线下消费支出类数据；农民大数据主要是包头农商行对本地农户调查获取的数据，包括农户收入、支出、资产、负债、社会关系等。

对于 30774 个建模样本，本地数据总共 34179 个字段，经过数据清洗、指标加工、特征衍生、特征筛选，有效的字段 147 个。下表是这些特征的举例：

序号	变量名称	数据来源	相对重要性
1	家庭收入	实地调查	4.90%
2	近 6 个月月均消费金额	实地调查	3.60%
……			
11	房屋面积	实地调查	3.32%
……			
15	电商平均消费单价	本地电商 POS 系统数据	3.11%
……			
23	是否供养大学生	实地调查	2.84%

……			
54	牲畜数量	实地调查	2.44%
……			
73	宅基地现值	村委会记录	1.88%
……			
88	通讯费近 3 个月账单平均金额	运营商数据	1.72%
……			
95	农机具现值	实地调查	1.32%
……			
111	家庭关系程度	背靠背评议	0.98%
……			
123	邻里关系	背靠背评议	0.72%
……			
147	保额	保险公司信息批量导入	0.38%

表 17：147 个本地特征举例表

“家庭收入”是农民大数据的特征，描述了申请人所在家庭过去 1 年的总收入。从图中可以看出，随着“家庭收入”的增加，坏账率逐渐降低，特别是“家庭收入”大于 10 万的申请人，其坏账率远低于其他申请人。“家庭收入”的 IV 值为 0.317。

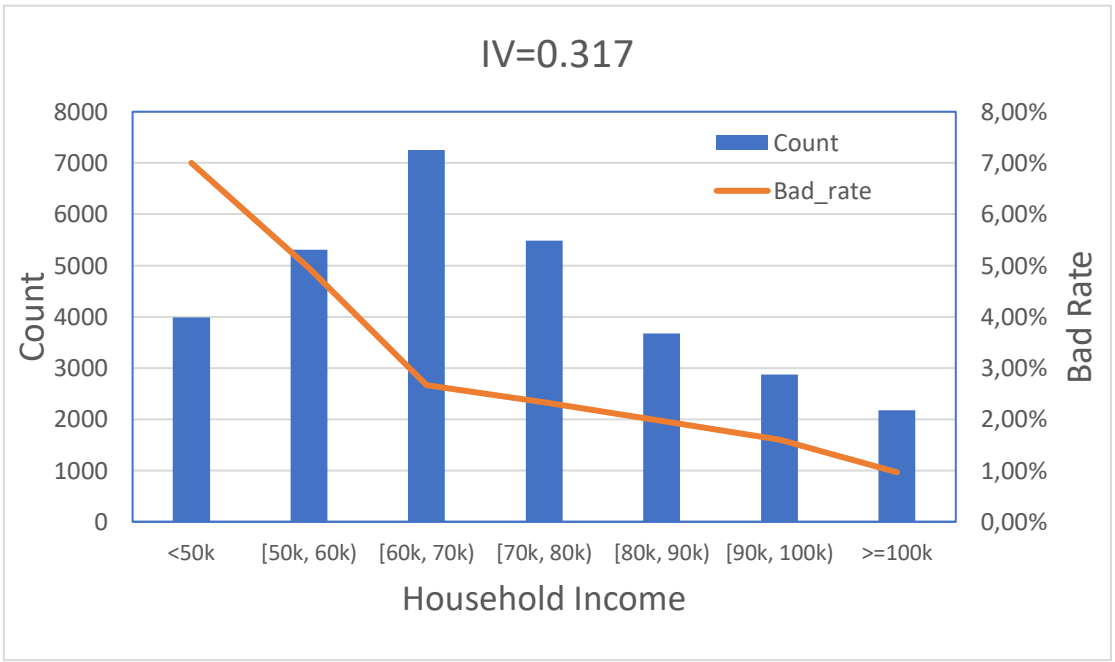

图 6：家庭收入 IV 值

需要说明的是，有一些特征并没有价值，如下表所示：

序号	变量名称	数据来源	IV 值
1	家庭总人口数	实地调查	0.019
2	房屋变现难易程度	实地调查	0.017
3	村委会职务	村委会数据	0.016
4	农机具类型	实地调查	0.014
5	手机在网时长标签	运营商数据	0.013
6	家庭持有产品数	行内数据	0.009
7	手机号码近 6 个月总停机频次	运营商数据	0.006
8	保费金额	保险公司数据	0.005
9	农机具变现难易程度	实地调查	0.003
10	车辆现值	实地调查	0.003

表 18：10 个没有价值的特征表

4.3.1.2.知识图谱关联特征

知识图谱本质上是一种叫做语义网络（semantic network）的知识库，即具有有向图结构的一个知识库，其中图的结点代表实体（entity）或者概念（concept），而图的边代表实体/概念之间的各种语义关系，比如两个实体之间的相似关系等。

目前，随着智能信息服务应用的不断发展，知识图谱已被广泛应用于智能搜索、个人化推荐及金融大数据等领域。知识图谱技术提供一种从海量文本和图像中抽取结构化知识的手段，是大数据分析的关键。关于知识图谱的构建，主要包括三部分：知识获取、数据融合和知识计算及应用。

（1）知识获取：在本地数据中存在着大量的结构化及非结构化数据。在处理非结构化数据方面，首先要对用户的非结构化数据提取正文，其中主要通过自然语言处理（NLP）技术进行信息抽取，从非结构化或半结构化文本中提取指定类型的信息，实现对非结构化信息有效识别和消歧。

（2）知识融合：当知识从各个数据源下获取时需要提供统一的术语将各个数据源获取的知识融合成一个庞大的知识库。

（3）知识计算及应用：知识计算主要是根据图谱提供的信息得到更多隐含的知识；链接预测则可预测实体间隐含的关系；同时使用社区计算的不同算法在知识网络上计算获取知识图谱上存在的社区，提供知识间关联的路径；通过不一致检测技

术发现数据中的噪声和缺陷。

由于本地农民大数据中有大量农户的社会关系信息，为了挖掘出有效的特征，提高贷前审批模型的效果，建立了本地农户知识图谱，如下图所示。

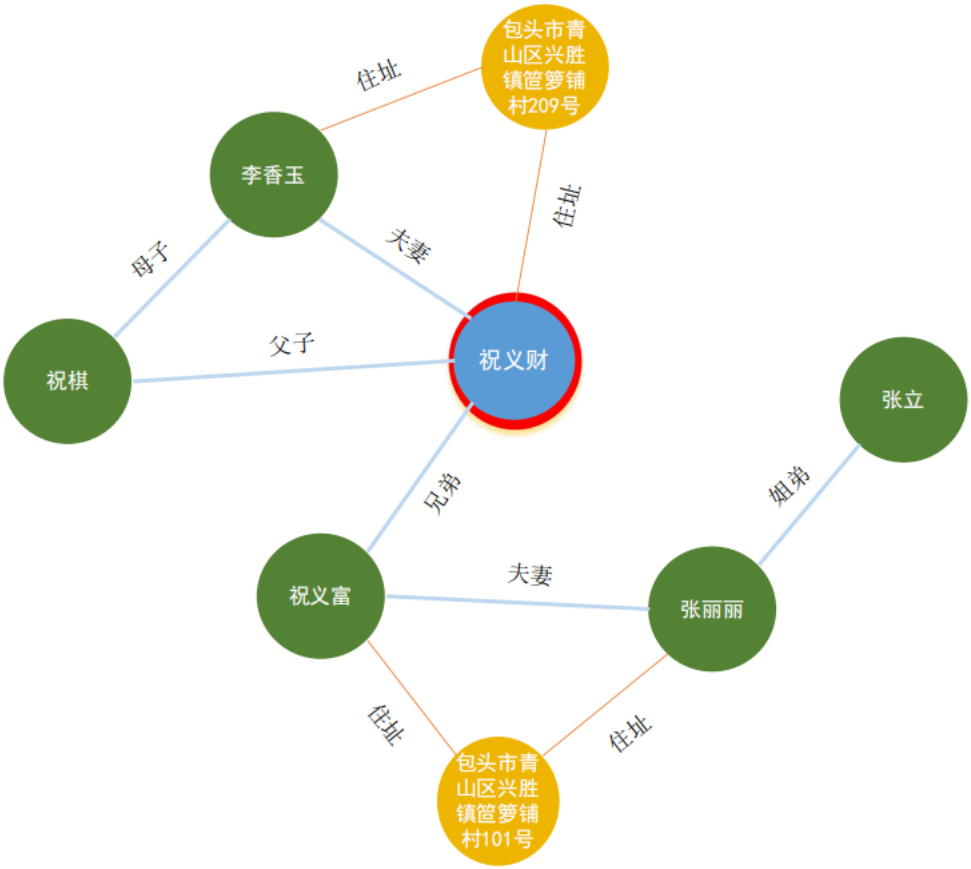

图7：本地农户知识图谱

在知识图谱中，本文主要以个人和住址为节点，人与人之间的关系为边，在个人节点上，还有年龄、婚姻状况、资产、收入、负债等属性。

利用知识图谱，首先进行了信息的校验，如夫妻二人是否居住在同一地址、兄弟二人是否有相同的父亲或者母亲等，来提高采集的信息的真实性。其次，从知识图谱中抽取一些特征，用于建立本地模型，如一度关联人的资产总值、一度关联人总负债、二度关联人的借贷总次数等关联特征。下表是这些特征的举例：

序号	变量名称	数据来源	相对重要性
1	直系亲属的家庭净收入	实地调查+知识图谱	5.40%
2	二度关联人的借贷总次数	征信数据+知识图谱	1.90%
……			
6	一度关联人的资产总值	实地调查+知识图谱	0.93%
7	一度关联人电商平均消费单价	本地电商 POS 系统数据+知识图谱	0.71%
……			
11	二度关联人通讯费近 3 个月账单平均金额	通讯数据+知识图谱	0.68%
……			
13	一度关联人总负债	征信数据+知识图谱	0.55%
14	二度关联人邻里关系	背靠背评议+知识图谱	0.42%

表 19：经过知识图谱抽取后的 14 个有效特征表

“直系亲属的家庭净收入”就是从知识图谱中抽取出来的一个有效特征。首先做离散化分箱处理，如下图所示，将其划分为 6 个分组，分别并计算每个分箱中样本的数量和坏账率。从图中可以看出，“直系亲属的家庭净收入”小于 10k 的用户坏账率最高，随着“直系亲属的家庭净收入”增加，坏账率呈现逐渐降低。经过计算，“直系亲属的家庭净收入”的 IV 值为 0.342。

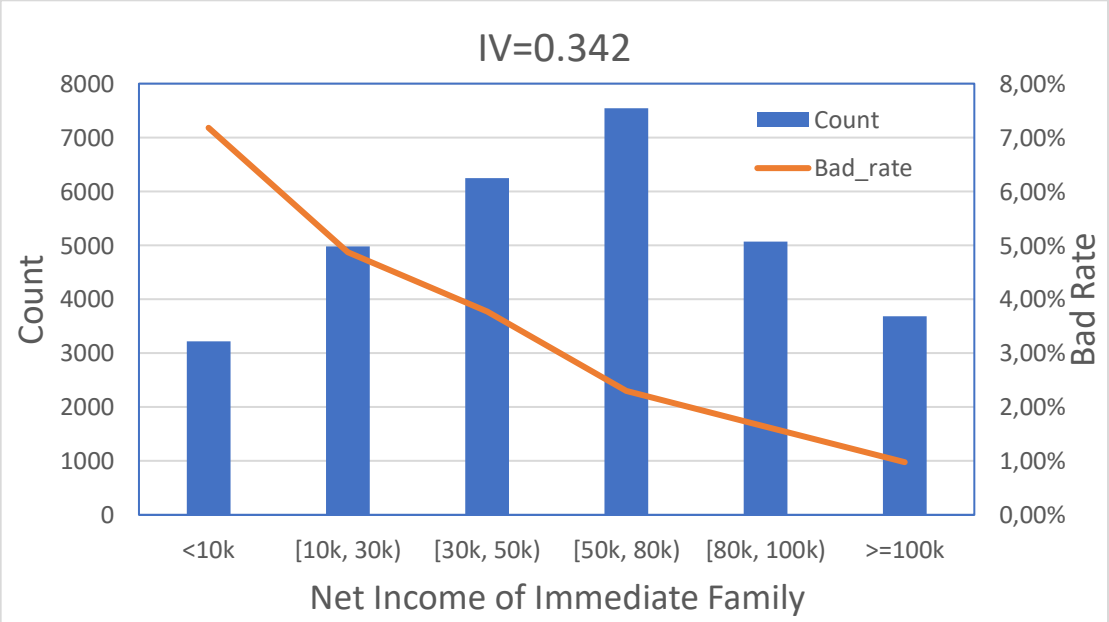

图 8：直系亲属的家庭净收入 IV 值

4.3.2. Local 模型建立与评估

4.3.2.1.Local 模型建立

Local 模型也是采用了 XGBoost 算法，训练集与 Public 模型的训练样本一致，21520 例，但是模型采用了特征，既包含了 Public 数据筛选的 284 个特征，也包含了本地数据的 147 个特征，以及从知识图谱中衍生的 14 个特征，共 445 个特征。

在模型训练过程中，采用 5 折交叉验证的方法，对 XGBoost 算法的超参数 num_round、max_depth 等进行网格搜索调参，选择 roc 最大的一组参数作为模型的最终参数，完成模型训练。

变量名	数据来源	相对重要性
直系亲属的家庭净收入	本地知识图谱	5.4%
家庭收入	本地数据	4.9%
存款月均余额	行内数据	4.7%
近 6 个月逾期累计次数	外部金融机构数据（征信）	4.1%
近 6 个月月均消费金额	本地数据	3.6%
近 4 个月的申请次数	外部金融机构数据（征信）	2.9%
偿债压力指数高	互联网大数据	2.5%
年龄	申请表	2.2%
命中账户状态为逾期	外部金融机构数据（征信）	2.1%
二度关联人的借贷总次数	本地知识图谱	1.9%

表 20：Local 模型中前 10 的关键变量

上表列出了 Local 模型中前 10 的关键变量，以及变量的数据来源和相对重要程度。在这 10 个关键特征中，有 4 个来自于本地数据，其余的来自于 Public 数据，这部分也是 Public 模型的关键特征。

4.3.2.2.Local 模型评估

在同样的测试集进行 Local 模型进行了评估，其 ROC 曲线如下图所示，模型的 KS=0.5310，AUC=0.840。与 Public 模型的 KS 和 AUC 指标比对来看，Local 模型的效果要远优于 Public 模型。

图 9：Local 模型在测试集上，KS，AUC 图

如果不考虑知识图谱所衍生的 14 个变量，建立 Local^{-14} 模型在同样的测试集进行了评估，其 ROC 曲线如下图所示，模型的 KS=0.4917，AUC=0.8027。与 Public 模型和 Local 模型的 KS 和 AUC 指标比对来看，参照下表，Local^{-14} 模型效果介于 Local 模型和 Public 模型之间，仍然优于 Public 模型。

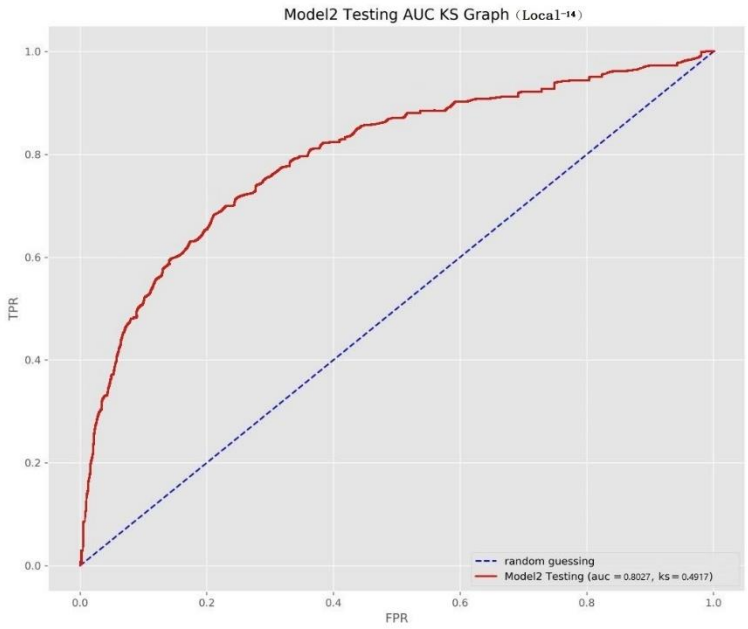

图 10：Local^{-14} 模型在测试集上，KS，AUC 图

	Public 模型	Local^{-14}模型	Local 模型
KS	0.3637	0.4917 对照 Public 模型约提升 35.19%	0.5510 对照 Public 模型约提升 51.5% 对照 Local^{-14}模型约提升 12.06%
AUC	0.7317	0.8027 对照 Public 模型约提升 10%	0.8401 对照 Public 模型约提升 14.8% 对照 Local^{-14}模型约提升 4.7%

表 21：Public, Local^{-14}, Local 模型 KS，AUC 分析表

因此在实践和下述分析中不再单独考虑 Local^{-14}模型。

对于测试集中的样本，按照 Local 模型预测的概率从小到大排序后，等频划分为 10 组，每组包含 10%的样本。然后分别统计各组的用户数、好用户数、坏用户数、坏账率、累计用户数、累计好用户数、累计坏用户数、累计坏账率、累计好用户的占比、累计坏用户占比等，如下表所示。从表中可以看出从第 1 组至第 10 组，坏账率呈单调递增的趋势。并且 Local 模型筛选出来的最优质的第 1 组客户，坏账率仅为 0.1%，而在模型筛选出来最差的第 10 组，坏账率高达 17.7%，说明 Local 模型具有较好的区分好坏用户的能力，且其区分能力优于 Public 模型。

假如要求模型的通过率为 80%，则需要对模型预测的风险较低的 80%核准通过，其余的 20%拒绝。从表中可以看出，通过部分的坏账率为 1.1%，比不使用模型的总体坏账率 3.2%降低了 2.1%。在被拒绝的 20%用户中，误杀了 1-81.7%=18.3%的好用户，获得的收益是拒绝了 1-28.7%=71.3%的坏用户。

Decile	Popu-lation	Cur-rent	Over-due	Bad_ Rate	Cum_ Pop	Cum_ Current	Cum_ Overdue	Cum_ Bad_ Rate	Cum_ Current_ Rate	Cum_ Overdue_ Rate
1	923	922	1	0.1%	923	922	1	0.1%	10.3%	0.3%
2	922	919	3	0.3%	1845	1841	4	0.2%	20.6%	1.4%
3	922	916	6	0.7%	2767	2757	10	0.4%	30.9%	3.4%
4	923	915	8	0.9%	3690	3672	18	0.5%	41.1%	6.1%
5	922	912	10	1.1%	4612	4584	28	0.6%	51.3%	9.6%
6	922	910	12	1.3%	5534	5494	40	0.7%	61.5%	13.7%
7	923	904	19	2.1%	6457	6398	59	0.9%	71.6%	20.1%
8	922	897	25	2.7%	7379	7295	84	1.1%	81.7%	28.7%
9	922	876	46	5.0%	8301	8171	130	1.6%	91.5%	44.4%
10	923	760	163	17.7%	9224	8931	293	3.2%	100.0%	100.0%

表 22：Local 模型在测试集上，用户区分能力分析表

4.4.本地信息带来的增值价值

4.4.1. 通过率-坏账率分析

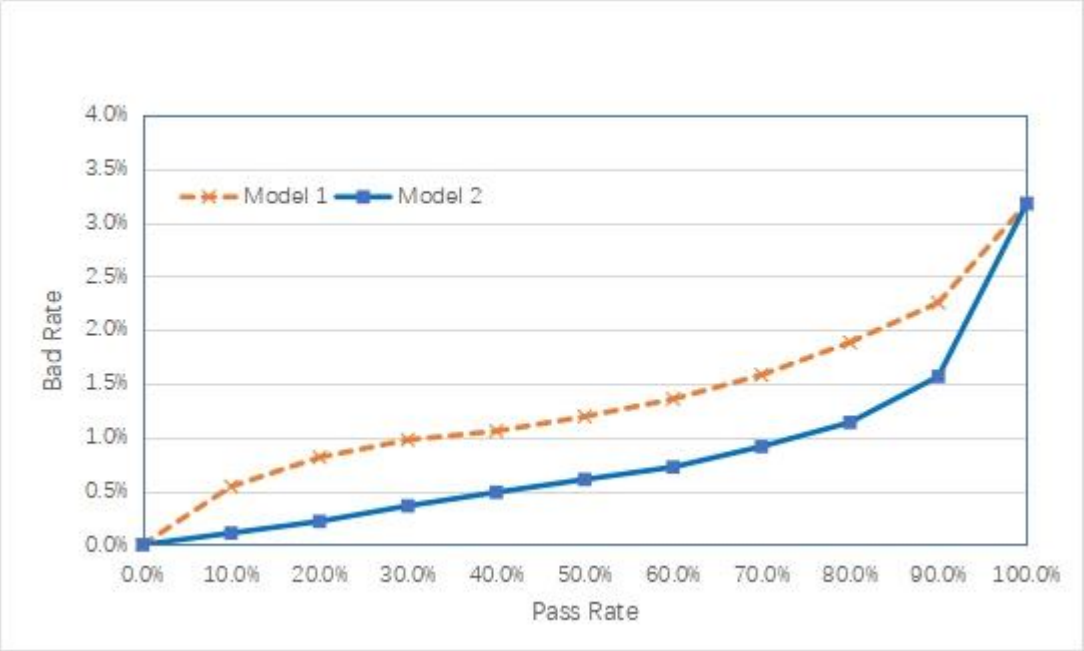

图 11：通过率-坏账率分析图

上图是以通过率为横坐标，坏账率为纵坐标，对比了 Public 模型与 Local 模型效果。从图中可以看出，Local 模型位于 Public 模型的下方，也就是说，在同样的通过率条件下，Local 模型的坏账率要低于 Public 模型；在同样的坏账率条件下，Local 模型的通过率要高于 Public 模型。如果把坏账率控制在 1%以内，Local 模型的通过率能达到 75%，而 Public 模型的通过率只能达到 30%；如果把通过率控制在 80%，Local 模型的坏账率为 1.1%，Public 模型的坏账率为 1.9%。

4.4.2. Lift 分析

Lift 是评估一个模型是否有效的一个度量，它衡量的是一个模型预测能力优于随机选择的倍数，以 1 为界线，大于 1 的 Lift 表示该模型比随机选择捕捉了更多的坏用户，等于 1 的 Lift 表示该模型的表现独立于随机选择，小于 1 则表示该模型比随机选择捕捉了更少的坏用户。

通常会根据模型的预测值将测试样本按分数从低到高排序，等频分为 10 等分，然后计算各组坏样本数在总体坏样本中的占比。

下图中，按照 3 种排序方式进行了分析：1）随机排序；2）按照 Public 模型的预测值排序；3）按照 Local 模型的预测值排序。

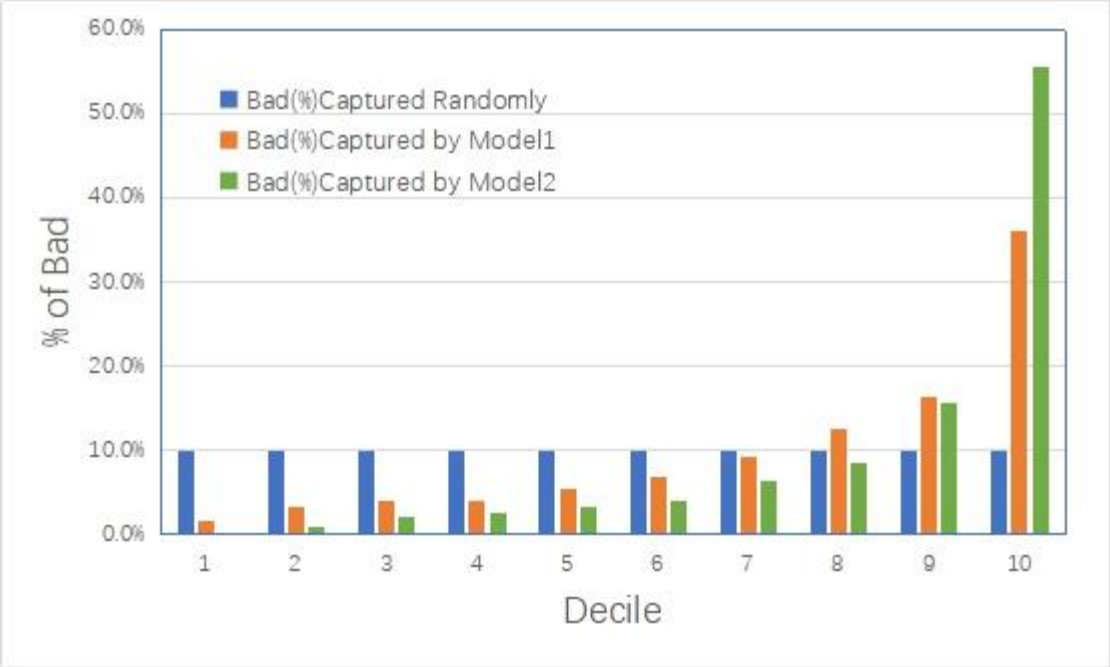

图 12：Lift 分析图

按照从坏到好逐步进行累计计算，取其与随机分组累计计算的比值，得到提升度 Lift，即该评分卡抓取坏客户的能力是随机选择的多少倍。Public 模型与 Local 模型的 Lift 如下表所示，累计提升图见下图。

Decile	Lift_Model1	Lift_Model2
1	1.00	1.00
2	1.09	1.11
3	1.19	1.23
4	1.30	1.38
5	1.44	1.56
6	1.62	1.81
7	1.86	2.16
8	2.17	2.66
9	2.63	3.57
10	3.62	5.56

表 23：Public 模型与 Local 模型 Lift 表

累计提升图能直观地去比较不同模型带来的区分能力增益程度。从图中可以看出，Local 模型的增益优于 Public 模型。

图 13：Public 模型与 Local 模型累计提升对比分析图

4.5.其它竞争优势

模型的另一个重要应用领域是在筛选建立营销白名单方面。

包头地区的电信运营商具有大量的数据（超过 200 万客户），并且对这些客户有超过 2000 个标签可以进行标识。经过协商，包头农商银行可以在电信运营商处部署基于 Local 模型而衍生出的营销模型，对 200 万客户和 2000 个标签进行建模，提前筛选目标客户（风险小，对产品也感兴趣）建立营销白名单，之后通过短信发送营销信息以获取客户。

所建立的模型在应用中会受到 2 个限制：1）原始数据不能从运营商获取，只能获取标签；2）未经客户授权不能调取人行征信。因此，原始的 Local 模型做了调整适应了这两个因素后，生成了一个新的预授信模型部署在运营商那里以生成营销白名单。

这种营销方式的优势在于，准确、高效触达到目标客户，将贷款信息推送给真正需要的人，减少了对无效客户（风险高，并且对产品不感兴趣）的打扰，可以有效地提高营销的成功率。

营销模型运营分析表（2018 年）如下：

项目	结果
原始客户数	2371456
经过模型筛选的客户数	128475
短信发送量（每个客户发送 20 条短信）	2569500
客户被触达后进件	8264
成功率	94.23%

表 24：营销模型运营分析表（2018 年）

具体的建模流程参考下图所示：

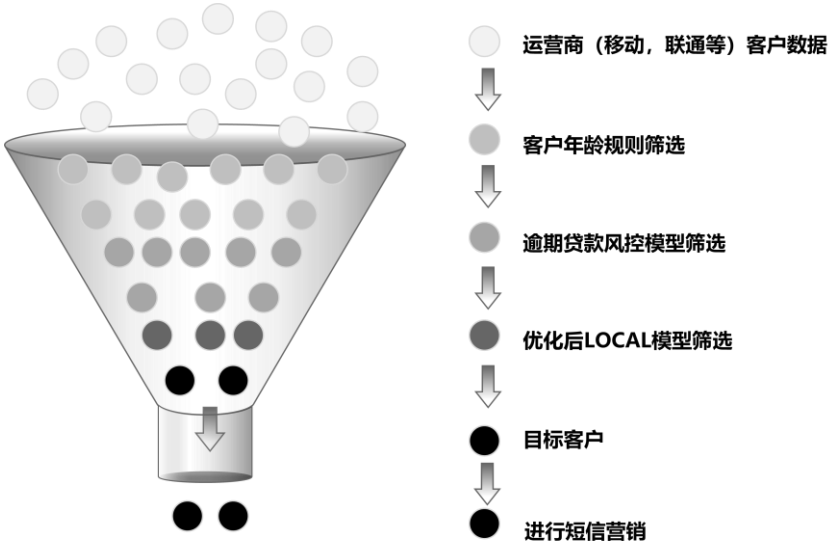

图 14：营销模型建模流程图

4.6.2019 年 Local 模型放款和财务分析

4.6.1. 2019 年新产品运行基本情况

在 2019 年基于 Local 模型，包头农商银行创立了市民贷产品，产品的具体形态如下：

产品要素	描述
产品名称	市民贷
客群	市民百姓、农户、小微企业主等
额度	2000~1000000
利率	年化 6%~18%；平均利率 8.72%
期限	12，24，36
还款方式	等额本息、先息后本
担保方式	纯信用、抵押
件均	43728 元

表 25：基于 Local 模型的市民贷产品要素表（2019 年）

基于这样的产品，在 2019 年全年贷款的放款规模和效率都得到了较大的提升，且成本得到了较好的控制。2019 年每月新增额逐月递增，且递增趋势均逐月增大，2019 年全年新增贷款累计 15.35 亿。

项目	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
新增额	3855	4910	6360	8146	10491	12946	15506	17083	20171	23354	26809	30597
当月累计放款量	3855	7710	12620	18980	27126	37617	50563	66069	83152	103323	126677	153486

表 26：2019 年市民贷投放情况

4.6.2. 2019 年不良与 2018 年不良的对比分析

2019 年引入模型后，在贷款规模逐步攀升后，由于模型的外部筛查及风控策略发生作用，2019 年的不良率呈下降趋势，且由于各机构的风控标准较为统一，各分支机构每季度的不良分布情况较为平均。通过模型的监测，发现风险能够快速、统一的进行调整，有效的控制了各机构的不良情况。

表：各机构不良表现更均衡

季度	A 支行	B 支行	C 支行	D 支行	E 支行	F 支行
2019 年第一季度	1.25%	0.83%	1.52%	1.28%	0.92%	1.68%
2019 年第二季度	1.28%	0.75%	1.54%	1.25%	0.94%	1.72%
2019 年第三季度	1.24%	0.78%	1.42%	1.13%	0.88%	1.66%
2019 年第四季度	1.23%	0.72%	1.45%	1.20%	0.84%	1.65%

表 27：各机构不良表现更均衡

表：各机构不良整体下降

支行	2018 年	2019 年
A 支行	2.85%	1.23%
B 支行	1.37%	0.72%
C 支行	2.54%	1.45%
D 支行	1.55%	1.20%
E 支行	1.77%	0.84%
F 支行	3.55%	1.65%

表 28：各机构不良整体下降

全行不良整体下降（2019 年较 2018 年下降了 62.8%）

	2018 年	2019 年
全行	3.20%	1.19%

表 29：全行不良整体下降（2019 年较 2018 年下降了 62.8%）

4.6.3. 新的产品在财务上的表现

每月投放量累计逐步增大，但贷款平均成本显著下降，以一笔 10 万元的贷款为例，保证类贷款由原来的 470 元，降至 319.42 元，单笔成本降低 150.58 元；抵押贷款由原来的 1650 元，降至 425.42 元，单笔降低 1224.58 元。这里特别值得一提的是由于线上评估的引入，评估费用由原来人工评估的 1000 元降至仅支付评估数据的 26 元。

2019 年新增的 15.35 亿元贷款，通过新模式共节约 676.7+147.93=824.63 万元，贷款成本降低了 60.04%。

信用类贷款两种模式的成本对比（10 万元为例）			
传统模式		引入大数据，Local 建模以后	
人工	150 元	大数据查询（数据源 1+数据源 2）—特指获得一个有效客户	14.20 元
公证	300 元	公证	300 元
资料	20	短信服务	0.15 元/次
		CFCA	2.7 元/次
		身份证 OCR 识别	0.06 元/次
		活体检测	0.45 元/次
		银行卡 OCR 识别	0.06 元/次
		人证核身	1.8 元/次
合计	470 元	合计	319.42 元

表 30：信用类贷款两种模式的成本对比（10 万元为例）

抵押类两种模式的成本对比（10 万元为例）			
传统模式		引入大数据，Local 建模以后	
人工	250 元	大数据查询（数据源 1+数据源 2）—特指获得一个有效客户	14.20 元
公证	300 元	公证	300 元
评估费	1000	评估费	26
办理抵押手续费用	80	办理抵押手续费用	80
资料	20	短信服务	0.15 元/次
		CFCA	2.7 元/次
		身份证 OCR 识别	0.06 元/次
		活体检测	0.45 元/次
		银行卡 OCR 识别	0.06 元/次
		人证核身	1.8 元/次
合计	1650 元	合计	425.42 元

表 31：抵押类两种模式的成本对比（10 万元为例）

	新增贷款（亿元）	新增笔数	2018 单笔成本（元）	2018 成本（元）	2019 单笔成本（元）	2019 年成本（元）	2019 较 2018 单笔降低了（元）	2019 年较 2018 年节约多少成本（元）	2019 年较 2018 年成本降低了
信用类	9.824	9824	470	4617280	319.42	3137982.08	150.58	1479297.92	32.04%
抵押类	5.526	5526	1650	9117900	425.42	2350870.92	1224.58	6767029.08	74.22%
合计	15.35	15350	2120	13735180		5488853		8246327	60.04%

表 32：2019 年新增贷款成本下降分析表

4.6.4. 市民贷新产品在客户时效性的表现和对比

贷款方式	传统模式		线上模式（市民贷新产品）	
	客户经理拓客	办手续的时间	线上拓客营销	办理手续的时间
信用、保证贷款	5 个工作日	3 个工作日	随时	3 个小时
抵押贷款	10 个工作日	10 个工作日	随时	1 个工作日

表 33：市民贷新产品在客户时效性的表现和对比

4.6.5. 对老客户的分析

2019 年引入新模型后，对存量老客户进行分析筛查，通过对存量老客户的分析，共计对 32452 户老客户进行提额，同时由于引入了新的分析方法，极大的降低了调查及作业成本，老客户平均提额幅度在 0.98 万元/笔，共计提额 3.18 亿元，调查及作业成本降低 115.41+21.22=136.63 万元。同时新增的 3.18 亿元的贷款，不良率由 2018 年的 3.20%降至 1.1%，降低了 2.1%，节约损失 667.9 万元。调查及作业成本与风险控制节约损失共计 804.53 万元。

参考 2019 年充分挖掘老客户的经验，在 2020 年一季度继续对老客户进行提额降费，2020 年一季度，挖掘老客户 9245 户，新增贷款 8782 万元，节约调查作业成本 37 万元，不良率下降到 0.13%，节约不良损失 85.19 万元，效果显著。

4.6.6. 将来的进一步研究

Local 模型和市民贷财务分析表明，目前的研究是成功的，下一步的研究主要放在如下 3 个方面：

1、进一步的完善本地大数据，引入更多的数据源。目前本地一些比较重要的大数据集合，例如：普通市民的房产拥有情况等仍需要接入，以完善本地数据集合。

2、进一步提高 XGBoost 模型的能力。之前的模型主要是基于银行 2019 年前既有的 30774 贷款客户，随着 2019 年客户数量的提升和客户表现的明确，模型需要在更大规模数据的基础上，持续迭代升级。

3、进一步提升知识图谱的能力。在包头本地，本地人占比超过 83%[3]，各类人群具有各种各样的千丝万缕的关系，知识图谱需要进一步挖掘这些关系以提升模型的辨识能力。

4.6.7. 小结

在本文的研究中，利用新的 Local 模型，建立市民贷，在 2019 年一年进行了运

[3] 根据包头政府相关文件得出。

行，实际获得的收益非常的显著：

系统上线前（2018 年）和系统上线后（2019 年），比较如下：

客户获益分析：（1）一笔小额贷款的获取时间，从 3 天减少到 3 个小时；（2）笔数：笔数从 2018 年的 73788 户增至 2019 年 197233 户，客户覆盖面更广；（3）利率：客户的平均利率从年化 10.96%降到年化 8.72%；

不同类型客户分析：（1）能够覆盖本地 70%的人群，只有是本地人，没有不良记录都可以借贷 8000 元以上；（2）其中能够覆盖 80 万原来没有借贷的白户人群；（3）原有的 80 万有征信客户的平均借贷金额可以从 3000 元调整到 20000 元；

银行获益分析：（1）放贷余额 2019 年达到 15.35 亿；（2）额度：客户的平均额度达到了 43728 元，风险进一步分散，显著优于过去（96683 元），向互联网银行靠近（5000 元）且优于本地其它银行（72187 元）；（3）当前不良和预计不良低于 1.2%；（4）预计净利润可以达到年化 4%，相对于传统银行零售业务净利润提升 203%；（5）成本：成本快速下降，降低了 60.04%；

从这些数据充分说明，客户获益及银行获益两方面均有巨大的改善和提升。

5. 结论

本文通过包头农商银行的一个线上“市民贷”个人信贷产品的应用实践，对中小银行所能获得的各类数据融合和建模进行了深入研究和分析，利用 XGBoost 和“关系图谱”等工具创新建立了本地化“市民贷”个人信贷产品的风控模型。在这个实践过程中，本文深入比较分析了基于公共数据集和本地数据集（叠加前者）的模型的优劣，得出了令人激动的结论—即基于本地数据集的优化 Local 模型具有更好的表现。这个基于本地数据集的“市民贷”风控模型于 2019 年在包头农商银行进行了实际的部署和应用，并且在客户获益及银行获益两方面均有巨大的提升（例如：客户获取贷款的时间最快到 3 个小时；客户的平均利率从年化 10.96%降到年化 8.72%；2019 不良率年较 2018 年下降幅度为 62.8%；新增的 15.35 亿元贷款的成本较 2018 年下降幅度为 60.04%）。这样的结果充分表明，本文的的研究成果可以帮助中小型银行找到了一条切实可行的，通过金融科技赋能实现零售战略转型成功之路，具有

普遍的应用价值。

本文的价值在于能够帮助中国的 2000 多家地方中小型银行在零售贷款方面与线上互联网巨头和线下同质化银行进行竞争，并取得竞争优势。

未来，随着对本地大数据的进一步搜集扩展和 Local 模型的进一步优化完善，模型和市民贷新产品在包头农商银行还会发挥更大的价值。

参考文献

Arjunwadkar, P. Y. 2018. FinTech: The Technology Driving Disruption in the Financial Services Industry [M]. Auerbach Publications.

卞维林. 2017. 转型致胜：银行零售业务效能提升转型策略 [M]. 江苏人民出版社.

Baptista, José A. G., Joaquim J S RamalhoJ Vidigalda Silva. 2006.Understanding the Micro Enterprise Sector to Design a Tailor-made Microfinance Policy for Cape Verde [R]. Portuguese Economic Journal, 5(3): 225-241.

Batiz-Lazo, B, Maixé-Altés, C., and Thomes, P. 2010. Technological Innovation in Retail Finance: International Historical Perspectives [M]. Routledge.

蔡丽艳，冯宪彬，丁蕊. 2011. 基于决策树的农户小额贷款信用评估模型研究 [J]. 安徽农业科学, 39(2)： 1215-1217.

曹磊，钱海利. 2016 FinTech 金融科技革命. 商周文化出版社.

陈良维. 2008. 决策树算法在农户小额贷款中的应用研究 [J]. 计算机工程与应用，44(31)：242-248.

Chen T, and Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System [C]. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785-794.

陈春宝，徐筱刚，田建中. 2017. SAS 金融数据挖掘与建模: 系统方法与案例解析 [J]. 机械工业出版社.

迟国泰，王卫.2009.基于科学发展的综合评价理论、方法与应用 [M]. 科学出版社，(9) : 27-79.

Copestake, J. 2007. Mainstreaming Microfinance Social Performance Management or Mission Drift [J]. World Development, 35(10): 1721 -1738.

Croxford, H., Ahramson, F, and Jablonowski, A. Translated by 赵瑞安. 2007. 零售银行做强法则 [M]. 经济科学出版社.

崔军扬. 2011.吉林省农村信用社农户贷款违约的影响因素分析 [D]. 吉林：吉林大学 .

崔健. 2005.商业银行个人信用风险评价 [D]. 天津：天津大学.

大连理工大学迟国泰课题组. 2010. 中国邮政储蓄银行农户小额贷款信用风险决策评价系统研究结 项报告 [R]. 大连理工大学.

戴宇. 2018. 农户贷款信用风险评价的研究基于临安区 A 银行的实例 [M]. 浙江农林大学：3-7.

丁振辉. 2014. 大数据背景下的小微企业信用评级研究 [J]. 征信 (11)： 1674-747.

DiVanna, J. A. 2005. 零售银行业的未来 [M]. 中国金融出版社.

Dutta, S., and Shashi, S. 1988. Bond Rating: A Non-Conservative Application of Neural Networks [J]. IEEE International Conference on Neural Networks: 443-450.

范文仲. 2014. 互联网金融理论、实践与监管 [M]. 中国金融出版社.

Harper, R., Randall, D., and Rouncefield, M. 2012. Organisational Change and Retail Finance: An Ethnographic Perspective [M]. Routledge.

Hartarska.V, and Nadolnyak, D. 2007.Does Rating Help Microfinance Institutions Raise Funds? Cross-Country Evidence [J]. International Review of Economics and Finance, 5: 1-14.

Han, J. and Kamber, M., Translated by 孟小峰. 2007. 数据挖掘：概念与技术 [M]. 机械工业出版社出版.

黄子健，王龑. 2015. 大数据、互联网金融与信用资本，破解小微企业融资悖论 [J]. 金融经济学研究院, 30(1).

何平平，车云月.2017. 大数据金融与征信 [M]. 清华大学出版社.

Jha, S., and Bawa, K. S. 2007. The Economic and Environmental Outcomes of Microfinance Projects: An Indian Case Study [J]. Environment, Development and Sustainability, (9): 229-239.

雷丰羽. 2018. 知识图谱在金融信贷领域的应用 [J]. 现代商业 : 89-90

李保旭, 韩继炀, 冯智. 2018. 互联网金融创新与风险管理 [M]. 机械工业出版社.

李杨,孙国峰,朱烨东,伍旭川. 2018. 中国金融科技蓝皮书：中国金融科技发展报告 [R]. 社会科学文献出版社.

李庆萍. 2017. 零售之道 [M]. 中信出版集团.

李正波，高杰，崔卫杰. 2006. 农村信用社农户贷款的信用风险评价研究 [J]. 北京电子科技学院学 报, 14(1)：69-74.

吕京娣，吕德宏.2011.欠发达地区小额信贷还款率影响囚素实证分析 [J]. 广东农业科 : 245-247.

Mahjabeen, R..2008. Microfinancing in Bangladesh: Impact on households, Consumption and Welfare [J]. Journal of Policy Modeling.

Mayer-Schönberger, V., and Cukier, K. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think [M]. Houghton Mifflin Harcourt.

Mayer-Schönberger, V., and Cukier, K. Translated by 周涛. 2013 .大数据时代 [M]. 浙江人民出版社.

Nicoletti, B. Translated by 程华. 2018. 金融科技的未来：金融服务与技术的融合 [M]. 人民邮电出版社.

Peter Goldfinch.2018. A Global Guide to FinTech and Future Payment Trends .Routledge.

Tan, P.-N., Steinbach, M., and Kumar, V. 2013. Introduction to Data Mining. [M]. Pearson New International Edition. Pearson.

孙国峰. 2019. 金融科技时代的地方金融监管 [M]. 中国金融出版社.

王春峰，赵欣，韩冬.2005.基于改进蚁群算法的商业银行信用风险评估方法 [J]. 天津大学学报，7(2)：81-85.

魏强，王迪光.2016.P2P 网贷平台借款人信用风险评价指标体系的研究 [J]，商 (14)：157.

温涛，冉光和，王煌宇，熊德平.2004.农户信用评估系统的设计与运用研究 [J]. 运筹与管理, 13(4)：82-87.

韦艳玲. 2009. 基于模糊聚类的农户信用信息分析 [J]. 广西民族大学学报(自然科学版) , 15(1)：78-80.

谢绚丽. 2018. 科技赋能：中国数字金融的商业实践 [M].中国人民大学出版社有限公司.

徐佳娜. 2004. 基于 AHP・ANN 模型的商业银行信用风险评估 [J]. 哈尔滨理工大学学报, 9(3)：96-99.

夏萌，赵邦宏，王俊芹. 2015.基于相关分析法的农户贷款信用影响因素分析 [J]. 科技通报，31(10)：266-268.

杨涛，贲圣林，杨东，宋科. 2018. 中国金融科技运行报告（2018）[R]. 社会科学文献出版社.

姚志勇. 2010. SAS 编程与数据挖掘商业案例 [M].机械工业出版社

余丰慧. 2018. 金融科技：大数据、区块链和人工智能的应用与未来 [M]. 浙江大学出版社.

于小洋. 2018. 企业法人知识图谱的构建及应用研究 [M]. 青岛大学.

张润驰，杜亚斌，荆伟等. 2017. 农户的小额贷款违约因素影响研究 [J]. 西北农林科技大学学报 (社 会科学版), 17(3)： 67-75.

张涛. 2017. 基于决策树模型的农户产权抵押贷款分类的实证研究 [J]. 陕西农业科学, 63(4)： 91-94.

郑兰祥，万雪. 2014. 基于 Logit 法的我国农村小额贷款公司信用风险评分模型构建研究 [J]. 安徽农 业大学学报, 23(4)： 49-54.

作者简历

陈云翔，男，汉族，1970 年 9 月出生于内蒙古自治区巴彦淖尔市，共产党员，硕士研究生学历，高级经济师、中级审计师。

学习经历：

2016 年 6 月至今，清华大学五道口商学院，全球金融 GFD 金博三期班。

2014 年 6 月-2020 年 6 月，新加坡国立大学经管学院，EMBA 课程。

2008 年 4 月-2010 年 7 月，内蒙古大学经管学院，MBA 课程；

1988 年 9 月-1990 年 7 月，内蒙古农行学校农牧金融专业；

工作经历：

2015 年 10 月至今，包头农村金融研究院，任理事长

2014 年 2 月至今，包头农村商业银行股份有限公司，任党委书记、董事长。

2012 年 7 月-2014 年 1 月，包头市郊区农村信用联社股份有限公司，任党委书记、董事长；

2011 年 6 月-2012 年 6 月，包头市郊区农村信用合作社联合社，任党委书记、理事长；

2006 年 5 月-2011 年 5 月，包头市郊区农村信用合作社联合社，任党委副书记、监事长（期间于 2005 年 10 月至 2007 年 1 月在自治区农村信用联社稽核监察部调训）；

2001 年 4 月-2006 年 4 月，巴彦淖尔市临河区农村信用联社，任监事长；

1998 年 3 月-2001 年 3 月，磴口县农村信用联社，任监事长；

1996 年 4 月-1998 年 2 月，临河市农村信用联社曙光信用社，任信用社主任；

1995 年 2 月-1996 年 3 月，临河市农村信用联社八一信用社，任信用社主任；

1992 年 10 月-1995 年 1 月，临河市农村信用联社小召信用社，任信用社主任；

1990 年 9 月-1992 年 9 月，临河市黄羊信用社，先后任信用社信贷员、主管会计、副主任；