

信贷产品风险冷启动 阶段评分模型的构建思路

宋捷 王方春

(徽商银行信用卡中心 安徽 合肥市 230601)

摘要:在利率市场化和互联网金融以及强监管的背景下,银行业的信贷业务面临着发展与风险的双重压力,一方面进行产品创新实现规模发展,一方面强化产品的风险管控能力。实践表明数据驱动下的风险评分模型能够较为准确的识别客户综合风险。文章讨论在无历史数据无标记样本情况下,创新产品投产初期构建申请评分模型的几种方法,为商业银行新产品准入风险控制提供解决思路。

关键词:信贷;风险控制;冷启动;风险评分模型

中图分类号:F830.5 **文献标识码:**A **文章编号:**1007-4392(2020)02-0051-05

一、研究背景

当前在利率市场化和互联网金融竞争影响下,银行业的发展面临巨大挑战。“金融科技”、“互联网+”等创新路径为金融机构的高质量发展提供契机,以商业银行为代表的金融机构在渠道、产品和运营模式上进行自主创新以不断适应新的金融环境,陆续推出消费贷、联合贷、现金贷、联名信用卡等纯信用类贷款产品。风险管理是商业银行经营信贷产品的基础条件,尤其在金融监管逐步趋严的形势下,监管部门以及金融市场对商业银行的风险控制能力要求逐步提高。可以说,风险管理能力是商业银行发展信贷业务的必备条件之一,也是其在市场化竞争中得以生存发展的核心竞争力。

商业银行要实现高质量发展,必须坚持

走业务创新之路,同时不断促进科技驱动风控能力提升,一方面要坚持业务创新,一方面要强化风险控制能力,为商业银行经营模式带来新挑战。在创新信贷业务场景下,每投产一个新产品,风控系统就需要对客户进行风险评估,从而达到控制实现预设利润率的目标。如何应对新产品投产初期的未知风险是商业银行业务创新过程中面临的新挑战。

二、风险控制流程及冷启动阶段面临的挑战

信贷产品的风险防控主要是能够精准识别客户的信用风险和欺诈风险,从而对于高风险客户予以拒绝准入,对于低风险高收益客户予以优先准入。《新巴塞尔协议》认为商业银行面临的主要风险为信用风险、市场

风险和操作风险三类。信用风险是由于某些原因导致贷款者无法或者不愿按照合约还款导致交易对手遭到损失的可能性。客户信用状况变差或者失去偿还本息的能力,就会产生信用风险,信用风险是商业银行信贷业务面临的主要风险,其造成的损失比重很大。申请欺诈风险是操作风险的一种,包括虚假申请、恶意申请,指欺诈分子利用被盗取的个人身份信息或伪造的身份信息或有意图的组织团伙申请信贷产品,手法较为隐蔽,银行在资质审查时较难识别。整个信贷业务审批流程经过客户申请、信息录入、信息初审、准入验证、欺诈调查、信用审查、初始授信、风险定价等步骤;风控系统审批流程(见图1)主要围绕信用和欺诈这两大风险展开。

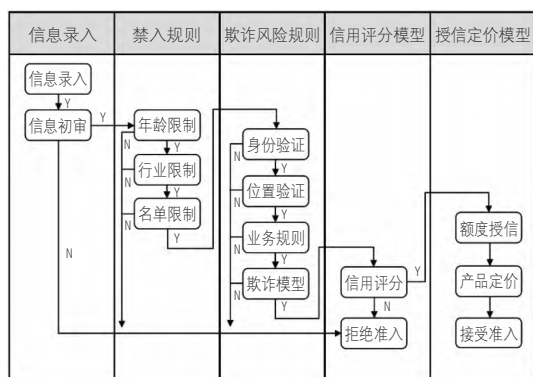


图1 风控系统审批流程图

随着大数据及金融科技在风控领域的场景化应用,银行通过构建大数据风控模型替代原有的人工审批,以“风控规则+风控模型”的模式识别客户准入风险。在欺诈环节,利用人脸识别、NLP、社交网络、知识图谱、地理位置识别等金融科技提高对欺诈客户的识别能力,通过分析欺诈群体的特征构建反欺诈模型精准识别欺诈申请。信用风险控制方面通过构建申请评分模型预测客户未来

一段时间的违约概率,结合规则集控制客户的信用风险。

风险控制的核心在于大数据风控,而大数据风控的核心是构建量化评分模型。量化评分模型是有监督分类的模型,用于预测客户是“好”与“坏”的概率。根据《新巴塞尔协议》,构建量化评分模型需要一定量的历史数据为训练样本,其中至少需要800-1000个“坏客户”标记样本。在创新产品投产初期面临着缺少甚至没有历史标记样本的问题,这时候无法建立数据驱动下的规则决策和有监督(有标记样本)的量化评分模型,这个困境被称之为风控系统冷启动问题。

三、冷启动阶段量化模型的几种设计思路

以往解决风控系统冷启动问题的方法多依赖于建模人员(或者业务人员)的业务理解和经验积累。即在风控系统启动之初,基于过往在其他相似业务场景中的经验积累,确定一定量的业务规则来帮助风控系统完成多个要求,并经过一段时间的业务量的积累,在满足模型建模的最低样本量需求后,建模人员才能正式开始进行风控模型第一个版本的更新。因为该过程多依赖建模人员自身的业务经验积累,所以可能出现两方面问题:一是建模人员自身业务理解能力不够深入,从而建模效果与实际需求出现偏差;另一方面,建模人员理解不统一而产生分歧。此外,单从贷前风险控制角度来看,既没有做到风险分散化,又没有做到审批精细化,同时多个产品共用一套风控策略,风险集中度非常高,因此需要新建风险模型及策略。

根据是否有样本数据,将风控系统的冷

启动分为两个阶段,以下重点讨论两个阶段的量化模型构建思路。第一阶段在新产品投产之前,此时没有任何样本,即无样本可依下的风险模型构建阶段;第二阶段是有一定量的样本积累,但因表现期不够导致无标记样本或标记样本不足,为无标记样本下的风险模型构建阶段。

(一)无样本可依下的风险模型

1.基于外部评分的风险控制。对于银行或者持牌金融机构,在选择数据时,会考虑数据来源、数据接口是否稳定、覆盖度、查得率等,最终选择性价比最高的数据源。随着数据服务商及征信市场洗牌,只有合规的企业才能成为银行的数据服务商。对于商业银行而言,首选数据是中国人民银行征信报告,其次是稳定的外部征信数据,如百行征信以及政府大数据服务机构。

稳定合规数据衍生而成的风控规则和风控模型,是商业银行构成其风险控制的基础。外部评分包括人行征信报告中的数字解读(即征信 1000 分)、外部公司信用分及行内的零售客户评分(或 CRM 星级)。直接用外部评分对客户进行信用评级操作方便,但是需要一定的成本,由于数据孤岛存在,不同评分都依托特定行业场景,针对较强,其中人行征信报告汇总银行业等金融机构信贷及担保数据和公共信息(包括住房公积金、养老保险、欠税欠费、行政处罚等信息),覆盖人群超过 9.5 亿人,其数字解读是商业银行应用最为广泛的外部评分。在使用这些评分时,需要经过充分的分析论证划定准入线,原因在于:一是新产品面临的目标客户群不同;二是产品的定位、风险容忍和审批标准不一样。目前,大多数银行在新产品投

产前均使用外部评分来控制准入风险,以求能快速上线、快速优化迭代。

2.基于德尔菲法的模型构建。德尔菲法,也称专家调查法,即“头脑风暴策略”,是从一组专家中取得共识的策略。具体流程是由每个专家组成员单独、匿名表达各自观点,组织者对观点进行分析和综合,并再次下发给专家组成员,专家组成员重新做出答复,反复该过程,从而逐渐对事项达成共识。

基于德尔菲法的模型适合于没有数据或者数据量极其有限的情况。影响客户信用的因素复杂多样,审批专家就是这些领域知识、经验长期积累的集中代表。由审批专家确定相互独立的规则变量,这些变量不能作为强规则确定申请是否被批准准入或者拒绝,如年龄、学历、借贷次数、欠款金额、交易流水等,但是可以对这些变量作加权求总分,对客户进行综合判断,基本可以构建专家级的风险评级模型,从而基于该模型做进一步决策。该方法操作时需要注意两点,一是需要有一定数量的专家成员;二是观点匿名性,避免受“名人”、“领导”的影响。这种专家评分模型是量化评分模型引入前常用的方法,在风控冷启动阶段也是一种好的选择。

3.基于综合评价方法的模型构建。综合评价方法是将多维度指标通过某种数学变换转化成一个综合指标,目前该方法已经运用在很多领域,如 CPI、综合经济效益指数编制等,可以反映事物的综合水平。量化评分模型中最经典的方法是回归模型,该模型的一般表示为各评分项的线性函数,其系数(权重)是通过训练样本计算得出。综合评价方法基本公式也可以理解为评分项的线性

函数,只是其权重是通过其他方式获得的。一旦确定评分项和权重,当客户申请贷款时,就可以应用该模型计算相应评分,根据评分值与标准线进行比较与评判。

综合评价方法的关键点在于评价指标体系的设计和各指标体系权重的确定。在指标设计过程中,商业银行可以组织风险领域专家,借鉴同类产品评分变量确定综合评价指标体系,所选指标要准确、全面且相互独立,然后选择综合评价方法确定各指标的权重,权重确定方法有很多,在无样本数据情况下可以用如相邻指标比较法、层次分析法、功效系数法等,如果有数据(冷启动第二阶段内容),可以使用PCA、DEA、模糊数学等方法确定权重,然后加权计算出每个申请人的综合风险评分。这种方法的优点是操作简单,以独特的定性与定量相结合的方法,有着专家评分法无法替代的优点;缺点是指标体系建设需要专家级的信审人员,现实中专家级人员往往不够多。为了模型能够产生决策效果,需要将客户的“好”与“坏”区分开来,为了达到这一目标,需要对客户进行资料收集和分析,最后根据评分区间归纳出准入或者拒绝标准线,所需工作量巨大。

(二)无标记样本下的风险模型

1.基于半监督算法的模型构建。对于分类问题而言,根据样本是否被完全标记将分类算法分为无监督、半监督和监督算法。半监督学习是一种不仅使用标记数据也使用非标记数据来进行学习的机器学习方法和技术,特别是在只有一组小量标记数据和大量非标记数据的情况下,半监督学习的方法对模型效果有比较大的提升。通常,非标记数据易获得,而标记数据却很难获得或成本

耗费很高。很多机器学习研究者发现把非标记数据与小量的标记数据混合使用可以有效地提高学习准确性。半监督学习技术的目标就是共同利用标记和非标记两组数据来构建比单独使用任一组数据更好的学习器。

半监督学习工作原理将未标记样本的分布与标记样本做相联系的假设。最常见的是聚类假设,基于该假设的算法是生成式分类器,该分类器是以概率视角求解的半监督算法,将未标记样本属于每个类别的概率视为一组缺失参数,采用EM算法来进行标记和模型参数的估计。另一个是流形假设,先根据训练样本及相似度量衡量创建图,图上的结点对应于每个样本,而边则代表样本间相似度。然后定义所需优化的目标函数,使用决策函数在建立的图上的光滑性作为正则项来求取最优模型参数。无论是聚类假设还是流形假设下的算法,其目的就是在未标记样本上获得最优泛化性能。

在产品投产初期,商业银行可以开放入口允许一部分客户申请,但是此时由于没有足够长的表现期,无标记样本或者标记数据量小导致样本数据分布刻画不完整。首先通过专家人工标记构建训练数据集,快速的构建初始半监督模型来对信贷主体进行风险和信用评估,后期通过不断迭代优化模型以达到最优效果。该模型是在专家标记的基础上,利用金融机构的现有数据对模型进行验证,并根据验证结果对该模型进行调整。半监督评分模型综合了专家经验和数据检验结果的模型方式,适合于有一定业务历史、具备一定业务规模,但数据尚不充分,如标记“坏样本”不够800-1000个或不具备拒绝样本的场景。

2.基于迁移学习的模型构建。深度迁移学习是一种处理标注样本数据少、模型训练成本高而产生的方法。它是从其他数据源训练得到的模型,经过一定的修改和完善,在类似领域进行复用,从而缓解数据源不足的问题。

迁移学习的基本思路就是利用预训练模型,即已经通过现成的数据集训练得出所需模型,在其中找到能够输出可复用特征的层次,然后利用该层次的输出作为输入特征来训练那些需要参数较少、规模更小的神经网络。总体思路可以概括为最大限度地利用有标注的领域的知识,来辅助目标领域的知识获取和学习。通过找到源样本和目标样本之间的相似性,加以合理利用。度量相似度的目标有两点:一是很好地度量两个样本的相似性,不仅定性地确认两个样本是否相似,更需定量地给出相似度水平;二是以度量为准,通过所要采用的学习手段,增大两个领域之间的相似性,从而完成迁移学习。

在新产品风控线建立过程中,风控系统完成了最初的从0到1的阶段,在有较少数据积累的情况下,尝试利用深度迁移学习技术综合使用平台上现有的其他类似业务线的数据,与当前业务线累计的少量样本共同作为建模的备选样本,以期提高风控模型性能,更早完成风控模型的第一个版本的更新,帮助新业务线快速走上正常业务流程。

四、结语

大数据风控是未来风险控制系统发展的必然模式,那些过度依赖人工而忽略数据驱动重要性的金融机构,已经开始慢慢显露其潜在风险。在强监管趋势下,风控流程的

精细化运作应该引起商业银行等金融机构的重视,“风险规则+量化模型”是风险控制系统的基本配置。

在风控冷启动阶段,以上方法有其各自的优缺点,共同的特点是所有方法都需要依托风控领域专家的经验,业务场景或客群的细微差别都会造成经验的失效,并且只能采用非常有限的维度做简单的评估,而无法对信贷主体进行全面的刻画和评价,这也是风控冷启动面临的主要问题,但风控冷启动阶段建模方法依旧是现阶段利用科学构建量化模型的主流选择。后期,随着业务人员和建模人员的业务经验积累,针对具体规则、数据、算法作出逐次的优化和迭代,并根据产品所处的不同生命周期主动性运用最优模型。

参考文献:

- [1]王方春,宋洁.初创期信贷产品申请评分建模方法综述[J].中国信用卡,2019(09):73-75.
- [2]陈建.信用评分模型综述[J].中国信用卡,2005(01):46-49.
- [3]冯振涛,冯梦媛.基于AHP的信用卡评分模型研究[J].金融理论与实践,2016(01):74-77.
- [4]王远江.迁移学习在现金贷违约预测中的应用研究[D].上海师范大学,2019.
- [5]叶夏菁.基于BP神经网络与半监督学习的网贷平台信用评估模型[D].浙江大学,2015.

责编:王博 校对:赵明晓