Modern Management 现代管理, 2021, 11(2), 146-151
Published Online February 2021 in Hans. http://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2021.112018

Hans汉斯

数据挖掘技术在零售业务中的运用研究

陈　俊，乔　辉

江苏银行股份有限公司风险管理部，江苏　南京
Email: xaviercj@163.com, qiaohui@jsbchina.cn

收稿日期：2021年1月23日；录用日期：2021年2月3日；发布日期：2021年2月24日

摘　要

近年来，大数据、人工智能技术的高速发展，正深刻改变着当前的金融生态和金融格局。对于大部分商业银行，尽管已经积累了大量的数据，但对数据的利用还不够深入，导致数据对业务的支持力度明显不足。因此，如何利用银行自身积累的数据资源，并从中提取出有益于商业银行经营和决策的信息，是商业银行面临的一个重要挑战。本文通过介绍大数据时代数据挖掘技术的概念、作用及方法，进一步分析数据挖掘技术在客户风险评价和客户关系管理方面的应用，浅析数据挖掘技术在零售业务中的运用价值，以期为商业银行的大数据应用提供参考借鉴。

关键词

大数据，数据挖掘，零售业务，评分模型

Research on the Application of the Data Mining Technology in Retail Business

Jun Chen, Hui Qiao

Risk Management Department of Bank of Jiangsu, Nanjing Jiangsu
Email: xaviercj@163.com, qiaohui@jsbchina.cn

Received: Jan. 23rd, 2021; accepted: Feb. 3rd, 2021; published: Feb. 24th, 2021

Abstract

In recent years, the rapid development of big data and artificial intelligence technology is profoundly changing the current financial ecology and financial pattern. For most commercial banks, although they have accumulated a large amount of data, the use of data is not deep enough, leading to the obvious lack of data support for business. Therefore, it is an important challenge for com-

文章引用: 陈俊, 乔辉. 数据挖掘技术在零售业务中的运用研究[J]. 现代管理, 2021, 11(2): 146-151.
DOI: 10.12677/mm.2021.112018

mercial banks to make use of the data resources accumulated by banks themselves and extract the information that is beneficial to commercial banks' operation and decision-making. By introducing the concept, function and method of data mining technology in the era of big data, this paper further analyzes the application of data mining technology in customer risk assessment and customer relationship management, and analyzes the application value of data mining technology in retail business, in order to provide reference for the application of big data in commercial banks.

Keywords

Big Data, Data Mining, Retail Business, Score Model

1. 数据挖掘技术的概念及作用

数据挖掘是指从大量数据中提取隐蔽在其中，但又有潜在价值的信息和规律的过程。简而言之，数据挖掘是一种决策支持过程，它主要基于人工智能、统计学等基础技术，通过对大数据进行自动化地分析，做出归纳性的推理，从中挖掘出潜在的模式和规律。

与传统的数据分析不同，数据挖掘的目的不在于验证某个假定模式的正确性，而是智能性地在大数据中找到合适的模型，并依此辅助决策，提升决策的准确性和效率性。因此，利用数据挖掘的方法对客户数据进行科学的分析，发现其数据模式及特征、存在的关联关系和业务规律，并根据现有数据预测未来的业务发展趋势，对商业银行的管理、决策制定、提升核心竞争力具有重要的意义和作用。

2. 数据挖掘技术方法

数据挖掘技术方法很多，常用的方法主要包括回归、决策树、聚类、神经网络、关联规则等。

2.1. 回归分析

所谓回归分析就是根据相关关系的具体形态，选择合适的数学模型，来近似的表达变量间的依赖关系。其分析过程根据已有数据进行模型的训练，发现规律和趋势，推测未来目标数据值。在商业银行零售业务中，回归分析可以被应用到各个方面。如通过对零售业务存量客户的历史行为回归分析，对客户后续的行为趋势做出预测并进行针对性的营销改变。又如对零售业务存量客户申请时点的数据和逾期表现数据的回归分析，对客户违约风险进行预测并针对性的设计客户准入策略。

2.2. 决策树分析

决策树分析是一种分类分析方法，其目的是找出数据库中的一组数据对象的共同特点并按照分类模式将其划分为不同的类别。具体可以应用到分类、趋势预测中，如商业银行将客户在一段时间内购买产品的情况划分成不同的类，根据情况向客户推荐关联类的商品，从而增加银行零售产品的销售效率。

2.3. 聚类分析

聚类类似于分类，区别是聚类不依赖于预先定义好的类别。它将数据分组成多个类别，使得同一类别中的数据之间具有较高的相似性，但不同类别之间的数据差异性较大，不同类别的数据关联性很低。

比如通过客户属性对客户进行团体划分，并在此基础上根据团伙属性识别风险团伙，应用于团伙欺诈的识别。

2.4. 神经网络分析

神经网络作为一种先进的人工智能技术，因其自身自行处理、分布存储和高度容错等特性，非常适合处理非线性的以及那些以模糊、不完整、不严密的知识或数据为特征的问题，它的这一特点十分适合解决数据挖掘的问题。但该分析方法存在的主要问题是可解释性较弱，对于银行零售授信业务的应用性不足，但对于欺诈防范是一种好的方法。

2.5. 关联规则分析

关联规则是隐藏在数据项之间的关联或相互关系，即可以根据一个数据项的出现推导出其他数据项的出现。关联规则挖掘技术主要应用于金融行业中对客户需求的预测领域。如各银行在自己的 ATM 机上捆绑顾客可能感兴趣的本行产品信息，供使用本行 ATM 机的用户了解。又比如如果数据库中显示，某个高信用限额的客户更换了地址，这个客户很有可能新近购买了一栋更大的住宅，因此会有可能需要更高信用限额，更高端的新信用卡，或者需要一个住房改善贷款，这些产品都可以通过信用卡账单邮寄给客户。再比如当客户打电话咨询的时候，数据库可以有力地帮助电话销售代表。销售代表的电脑屏幕上可以显示出客户的特点，同时也可以显示出顾客会对什么产品感兴趣。一旦获得了这些信息，银行就可以改善自身营销。

3. 数据挖掘技术在零售客户风险评价方面的应用

近年来，零售业务取得了蓬勃发展，并逐渐呈现线上化趋势。零售业务的线上化一方面提升了业务效率、节约了人工成本，另一方面因为无接触特点随之而来的则是该业务风险激增。然而，相比于对公业务，零售业务客群庞大，积累了大量的业务数据。在此背景下，利用银行积累的业务数据挖掘客户风险信息，并对零售客户风险进行识别成为数据挖掘技术的一个重要应用。

3.1. 数据挖掘技术应用于零售客户风险评价的必要性

为充分利用银行积累的大量零售业务数据，同时有效识别客户风险，数据挖掘技术在零售客户风险评价方面的运用应运而生，基于数据挖掘技术的风险评价结果是对潜在客户的筛选过滤、审批决策的过程。该技术可以根据申请客户的申请信息、征信信息等评估其信用程度[1]。通过数据挖掘技术应用于零售客户风险评价后主要有利于银行解决三大难题：一是解决如何有效地对零售客户进行分层管理的问题；二是解决如何防范向有不良记录的客户授信的问题；三是解决如何提高银行预防和抵抗零售信用风险的能力的问题。因此，在现有大数据、线上化背景下，数据挖掘技术应用于零售客户风险评价具有必要性。

3.2. 数据挖掘技术应用于零售客户风险评价的实现路径

本节将进一步阐述如何实现数据挖掘技术在零售客户风险评价方面的应用。申请评分模型是数据挖掘技术应用于客户风险识别最常见也是最成熟的方法之一，因此以 A 银行互联网贷款申请评分模型为例介绍数据挖掘技术应用的具体实现路径。互联网贷款申请评分模型构建的流程包括：业务访谈与数据提取、数据分析与清洗、模型设计、建模数据准备、变量分组、模型构建、模型评估、模型验证等 8 个步骤。具体如下：

1) 业务访谈与数据提取。一是通过对消费金融与信用卡中心、网络金融部开展访谈，了解业务逻辑

及数据存储逻辑。二是提取建模所需的原始数据，包括：借呗决策、人行征信、借呗月度逾期切片、运营商手机信息、地税信息、腾讯反欺诈结果、腾讯反欺诈明细 7 张原始表，涉及数据 7000 万条，数据提取时间为 2017 年 9 月至 2020 年 1 月。

2) 数据分析与清洗。一是通过对数据主键唯一性、时间分布、缺失值比例、统计分布四个方面的检测，以及对月度逾期数据的分析，对数据进行清洗。二是对清洗后的数据通过身份证件号、查询时间进行关联，形成建模宽表，共包含 650 个变量。

3) 模型设计。通过账龄、逾期表现、迁徙率等分析确定本次建模的观察期、表现期和好坏定义。其中，观察期为 2017 年 9 月至 2018 年 9 月，表现期 15 个月。表现期内从未出现逾期或从未有过提款记录的客户定义为“好”；表现期内出现过 M3 (61~90 天)及以上逾期的客户定义为“坏”。按照上述口径，确定“好客户”293 万，“坏客户”3.4 万。

4) 建模数据准备。将符合上述观察期和表现期要求的样本作为模型开发的全量样本，再采用分层抽样方法进行建模数据准备。抽取方法为：首先，抽取全部 3.4 万个“坏客户”；然后，按照 20:1 的比例随机抽取 68 万个“好客户”；最后，对抽样形成的建模数据按照 7:3 的比例形成开发样本和验证样本，开发样本用于模型训练，验证样本用于模型验证。

5) 变量分组。一是对宽表中的 650 个变量进行分组，计算不同分组的 WOE 值(即证据权重，其数值等于好客户占比除以坏客户占比的商的自然对数)。二是计算经 WOE 值转换后的每个变量的 IV 值(即信息值，信息值越大，对“好坏客户”的区分能力越强)。三是根据 IV 值的大小对变量的区分能力进行排序，筛选 IV 值大于 0.1 的变量作为最终的入模候选变量。本次入选变量数为 45 个。

6) 模型构建。设置变量入模的阈值(即模型变量系数的 P 值小于 0.05)，并通过逐步回归方法，对符合阈值条件的变量进行筛选，形成评分卡模型，共计 24 个初选模型。

7) 模型评估。对模型表现的评估，首要基于 KS 值和基尼(Gini)系数这两个统计值。KS 值和基尼系数(Gini)用来衡量模型区分好坏的能力，也就是模型的预测能力。一般来说，具有更高 KS 值和基尼系数的模型有着更好的预测能力，但同时也要确保模型不会只是为了简单地增加 KS 造成过度拟合。

8) 模型验证。开展模型投产前全面验证工作，分为定性和定量验证。定性验证包括对模型和支持体系相关的治理结构、政策、流程、控制、文档管理等方面的测试；定量验证包括对模型的区分能力、稳定性方面的测试。经验证通过后，模型才可投产应用。

A 银行基于数据挖掘技术，开发互联网贷款申请评分模型，有效填补了 A 银行针对互联网消费贷款产品申请审批决策模型的空缺，为零售网贷业务授信环节的统一规范化风险管理提供了依据，网贷申请评分模型必将对零售业务的健康、快速发展发挥重要的作用。

4. 数据挖掘技术在零售业务客户关系管理中的应用

在大数据时代，商业银行的客户信息来源范围得到大幅度扩展，商业银行可以摆脱过去单一分析客户行为的阶段，将客户行为与特定的时间、地点、生活场景及客户的社会关系背景联系起来，对客户进行全方位的描述。正因为数据范围的大幅度扩展，商业银行能够对客户的分析和细分更加趋于精细。通过运用聚类分析方法，对客户进行多维度的细分。运用关联规则分析等技术手段，充分发现、挖掘客户的需求，进行客户综合价值的分析与预测。结合客户的细分情况、需求分析情况与价值情况进行分析，制定差异化的客户管理策略。与此同时，通过对客户行为进行精准化的分析预测，做好客户的动态管理工作，提升客户的价值[2]。

总体而言，数据挖掘在零售业务客户关系管理中的应用主要包括客户识别、客户营销、交叉销售、产品和服务定价四个方面，具体分析如下：

4.1. 客户识别

客户识别是客户关系管理的初始环节，主要包括目标客户分析和客户细分两个方面。客户识别的具体逻辑是：通过客户的购买记录数据找出客户特征，识别银行的潜在客户以及最具盈利价值的客户，并进行客户细分，将具有相似特征的客户进行分类，从而为银行实施客户关系管理提供基础。

根据客户识别的特点和逻辑，可按照客户的背景资料、消费偏好等信息维度将客户分为不同类型。通过客户分类，可以帮助银行掌握不同客户群的特征，找出客户消费的行为和规律，计算出不同客户对银行的贡献程度等，从而筛选出客户群体的种类。在实际应用过程中，可通过聚类分析的数据挖掘技术实现。

4.2. 客户营销

在识别目标客户后，商业银行需要采取针对性的措施进行客户营销，一般会通过电话、电子邮件等渠道直接向目标客户推销产品。在进行客户营销前，首先需要了解客户的个性化需求和对不同产品的响应程度。基于数据挖掘技术中的客户响应模型即可以实现这一功能，该模型主要根据客户基本信息和历史交易数据，预测目标客户对某一产品和服务的响应程度，从而有针对地进行营销，提升营销响应率，并降低营销成本。

4.3. 交叉销售

交叉营销属于客户营销的范畴，也是对客户营销的进一步拓展，所谓交叉营销就是指银行基于现有客户发现其多种需求，向其提供多种满足其需求的相关产品和服务的销售方式。通过交叉销售银行既可以降低营销成本，维系现有客户资源，还可以实现存量客户价值最大化。

对银行来说，正确地预测客户的需求，并及时组织好匹配的产品和服务以响应客户的需求是一件较难完成的工作，这需要大量历史数据的储存与分析。一般而言，可以采用数据挖掘中的关联规则进行分析，利用关联规则可以分析客户交易行为与客户背景信息(如年龄、性别、收入、教育程度等)之间的关联关系，找出客户交易行为的影响因数，分析客户在购买一项产品和服务的同时，最可能购买的金融产品和服务，确定最优的销售组合。

4.4. 产品和服务定价

借助数据挖掘技术还可以对客户进行市场细分，从而建立更加科学、完善的动态定价系统。具体包括三个方面：一是基于统计特征，将用户按照年龄、性别、职业等特征分配到不同类别中，再针对不同的类别进行有针对性的措施。二是基于客户价值，主要是按照客户资产对银行的贡献度为标准，将客户分为高端客户、中端客户、低端客户。或是按照客户的信用风险程度分为高风险客户、中风险客户和低风险客户。三是基于消费行为，在数据挖掘技术的支持下，可以对客户群体的细分方法做出改进，比如通过客户使用电子银行或购买理财产品的的时间段进行细分。又比如通过对客户进行市场细分，建立起综合服务和信贷差异化定价体系，做到不同产品、不同客户、不同区域的差别化定价，最终实现一户一策的综合化、差异化服务，提升精准营销水平。再比如将零售客户纳入定价系统，进行客户选择，针对不同的服务内容给予客户不同优惠，可以达到差别化定价和最佳客户体验的双重目的[3]。

5. 数据挖掘技术的优点与不足

综上所述，数据挖掘技术在银行零售业务的应用广泛、效果显著，相比较银行进行零售业务的传统管理模式，通过先进的数据挖掘技术分析并建立模型作为决策的依据，不仅有助于银行业务效率的提升，

还有助于银行客户的价值挖掘和创造。总体而言，数据挖掘技术具有明显的技术优势，但同时基于数据挖掘技术构建的模型是对现实的高度抽象化，无法避免模型风险的产生，因此数据挖掘技术同样存在不足之处。具体概括如下：

5.1. 数据挖掘技术存在的优势

1) 数据挖掘技术的全面性强。通过数据挖掘技术，考虑的因素更全面，涉及客户的信息也更丰富，因此可以全方位、多维度的考量客户风险。

2) 数据挖掘技术的适用性强。通过数据挖掘技术，可针对不同客户、不同产品进行细分分析，从而更准确地预测客户的未来表现，提高营销效率或降低坏账率。

3) 数据挖掘技术的效率性高。通过数据挖掘技术，可以实现计算简便、反应迅速的目标，具体地根据客户最新个人信息、行为信息或征信信息，利用计算机自动挖掘分析，加快决策流程，从而降低营运成本，大大提高了效率[4]。

4) 数据挖掘技术的稳定性高。数据挖掘技术拥有良好的稳定性，可以增加决策的客观性和一致性，提高客户满意度，对我行零售业务平稳快速发展有很好的推动性。

5.2. 数据挖掘技术存在的不足

1) 数据挖掘技术构建的模型不能完全替代人工的经验判断。数据挖掘技术是建立在历史数据的基础上，通过发现客户信息数据中的规律，寻找关联关系并找出在客户关系管理中需要关注的重点，为银行的发展零售业务起指导作用。但针对历史的数据的分析预测并不能完全代表未来，且随着零售业务及整个社会信用状况的发展，决定客户需求、信用等的关键因素也会不断变化，构建的模型也可能不再适用。因此，银行应以模型判断为依据，结合人工前瞻性的判断管理零售业务。

2) 数据挖掘技术在解决客户隐私问题方面存在弊端。在大数据时代，银行获得了大量数据，对这些数据的挖掘可以给银行带来客观的收益。但同时，对于政府和商业数据的挖掘，可能涉及到国家安全或商业机密之类的问题，对客户信息数据的挖掘，可能会侵犯客户隐私，引起客户的反感、投诉甚至诉诸法律。如何解决保密、法律和伦理方面的问题对银行也是个不小的挑战。

参考文献

[1] 谭波, 滕光进, 王浩. 基于大数据的客户关联关系及风险预警研究[J]. 清华金融评论, 2017(8): 19-21.

[2] 黄敏. 基于云计算技术视角的大数据挖掘技术分析[J]. 数字技术与应用, 2019(12): 12-19.

[3] 段萍. 大数据时代下数据挖掘在银行中的应用[J]. 科技经济导刊, 2017(6): 21-25.

[4] 杜磊. 银行大数据挖掘、使用和管理的方法[J]. 现代经济信息, 2019(4): 8-11.