金融场景下大数据建模常见的数据质量问题及应对策略研究

文／马文星[1] 王锋[2] 韦晓[2]
1.工银科技有限公司 2.中国质量认证中心

[摘要] 随着科技的快速发展，数字金融已被视为是科技金融、绿色金融、普惠金融及养老金融发展的核心，其健康发展对于铸造新质生产力具有重要意义。大数据建模是更好、更快地发展数字金融的重要工具和方法，本研究从金融的视角出发，在深入分析金融数据的特征、大数据建模的价值和保证数据质量重要性的基础上，对金融场景下大数据建模过程中常见的数据缺失、数据错误、数据滞后、数据失衡和数据冗余等数据质量问题，提出针对性的应对策略，以充分发挥数据要素价值，助力数字金融发展。

[关键词] 数字金融 大数据建模 数据质量 应对策略
[中图分类号] N945.12 F49 【DOI】10.16691/j.cnki.10-1214/t.2024.07.009

金融作为国家核心竞争力的重要组成部分，其高质量发展是推动中国式现代化强国建设不可或缺的内在要求。党的十八大以来，我国金融事业取得了新的重大成就[1]。习近平总书记在2023年中央金融工作会议上首次提出并强调要做好“五篇大文章”。其中的数字金融依托人工智能、区块链等信息技术手段，推动金融行业在业务模式等方面大力创新，已被视为推动科技金融、绿色金融、普惠金融及养老金融发展的核心，其健康发展对于铸造新质生产力具有重要意义[2]。

大数据建模是发展数字金融的重要工具和方法，通过收集、处理和分析海量数据，帮助金融机构更准确地识别风险、评估价值、制定策略和优化服务[3]。在大数据建模过程中，数据质量是影响模型效能的关键要素之一，以高质量的数据为基础进行建模，可推动金融数据与服务的融合应用，提升数据利用效率，充分发挥数据要素价值[4]。然而，在金融场景下，大数据建模过程较为复杂，所涉及的数据范围较广，容易出现影响大数据建模的数据质量问题。

1 金融场景下大数据建模数据质量的重要性

1.1 金融数据的定义及特征

金融数据在国内外尚未有明确而统一的定义。美国在“非公开信息”的概念中对其有所提及，指“消费者提供给金融机构、来源于消费者任何交易或所获得服务的可识别个人金融信息”；我国则将“个人金融信息”定义为“金融机构通过外部渠道或自身业务开展获取、保存与加工的个人信息的总称”，包括个人的账户、财产、身份等信息。由此可见，金融数据不仅包括金融机构采集的原始数据、加工处理后的数据，还包含具有金融属性的数据，如宏观经济数据、金融市场数据等。

金融数据具有以下特征：一是规模大，涵盖范围

广，通常以大量观测值和交易记录的形式存在；二是更新频率高，要求对数据进行实时或近实时获取和分析；三是多样性，通常来源于不同的渠道；四是关联性，数据之间常存在关联关系；五是敏感性，国家政策变化等情况均会对金融数据造成影响[5]。

1.2 大数据建模的价值

1.2.1 推动产品创新

一是大数据建模能够获取与客户相关且具有价值的信息，从而对金融工具的不同特征予以重新分解和组合，打造出更符合客户需求的金融产品；二是借助大数据建模还能分析挖掘整个市场的交易数据，设计出更合理、高满意度的产品。

1.2.2 加强风控能力

大数据建模技术是智能化技术之一，将其与金融风控相融合，可以量化各种信息，实时监控各类风险，预测客户未来行为，实现对金融风险更好的控制管理。

1.2.3 提高营销效率

金融机构采集用户的相关信息，并进行量化处理，结合多维度的特征，将用户进行合理划分归类，提供精准、有效的金融服务，提高金融服务效率和水平。

1.2.4 促进普惠覆盖

大数据建模技术与信贷业务相融合，突破了传统“关系型信贷”的思维方式，实现大规模、低成本地获客，解决普惠金融发展的多种问题。

1.2.5 优化资产结构

在资产处置中，金融机构可以运用大数据建模技术，基于产品的市场价格、波动趋势或客户特征等进行精准定价和交易。

1.2.6 扩大监管感知

金融市场及金融数据具有多样性、关联性、敏感性，单一因素已不能决定或解释金融市场的变化方向。然而，“大数据建模+监管”的模式引入多元化数据，计算出各类金融指数，建立起科学有效的金融监管体系。

1.3 保证数据质量的重要性

《“十四五”大数据产业发展规划》强调“构建行业数据治理体系”，鼓励企业发挥技术驱动治理的作用。人民银行印发的《金融科技发展规划（2022-2025年）》提出“高质量推进金融数字化转型”。国务院《关于构建数据基础制度更好发挥数据要素作用的意见》提出要“完善治理体系”。还有首次将数据治理与监管评级挂钩的《银行业金融机构数据治理指引》(银保监发〔2018〕22号)，以及行业标准《个人金融信息保护技术规范》(JR/T 0171-2020)等，为数据质量工作提供了一定的指导。目前，中国质量认证中心牵头编制国家认证认可行业标准《数据产品质量评价要求》，为数据产品质量评价提供规范。

数据质量管理是涉及多部门的复杂的系统性工程，要求各部门高度协调一致，加强数据标准化与规范化建设，强化数据全生命周期的质量管控[6]。数据质量能否得到有效保证将直接影响大数据建模的效果，决定数据价值的高低，长远来看将影响整个企业数字化转型[7]。

由于金融数据来源众多，数据结构复杂多样，金融机构不仅要转变数据存储方式，还应开发或引进用于数据质量检测和清洗的智能化工具，立足数据质量视角，综合考虑数据质量主客观维度[8]。因此，保证数据质量，是做好大数据建模，充分发挥数据资产价值、资本价值的基石。

2 金融场景下大数据建模常见的数据质量问题

2.1 数据缺失

金融数据在采集、传输、存储等过程中，可能由于各种原因导致数据缺失或遗漏，会直接影响模型的效果。金融场景下大数据建模常见的数据缺失问题主要包括数据值缺失和数据标签缺失。

数据值缺失包括两个方面：一是金融业务在运营过程中所采集的某些字段数据缺失；二是大数据建模为促进金融业务开展所需但未能在金融业务运营过程中采集的数据，例如工商类、司法类、舆情类等数据。

数据标签缺失是指在数据建模过程中数据标签或目标变量存在缺失或不完整的情况，可能是由于数据标注不全等原因导致的，会影响有监督学习模型训练。

2.2 数据错误

数据准确是金融大数据建模的基石，如果建模数据存在错误或偏差，将会直接导致模型结果错误，进而影响金融决策的有效性。金融场景下大数据建模过程中常见的数据错误问题主要有数据值错误和数据标签错误。

数据值错误是指金融场景下大数据建模所采集的某些字段数据出现了错误，可能存在以下几点原因：一是数据采集错误；二是数据记录错误；三是数据源错误。

数据标签错误是指在数据建模过程中数据标签或目标变量存在不准确的情况，可能是由于人为或机器错误标注数据等原因导致的，会使得有监督学习模型训练结果产生错误，进而影响模型的效能。

2.3 数据滞后

数据时效性是指数据在特定时间范围内保持其有效性的能力。金融市场是一个高度动态的环境，价格、利率、汇率等金融数据变化速度快，需频繁进行获取、处理、传输、存储等操作，流程较为复杂，可能使得数据更新不及时，导致出现数据时效性问题，进而影响模型的训练和预测。

2.4 数据失衡

数据均衡性是指不同类别或者不同标签样本数量的差异程度。样本数量差异越大数据均衡性越差。数据均衡性较差会使得大数据模型在训练时更倾向于关注数量较多的类别或标签样本，从而导致模型对数量较少的类别或标签预测效果不佳，降低模型性能。数据失衡的原因可归纳为以下几个方面。

一是金融业务本质的不均衡性。如金融领域某些事件发生的概率较低，以及客户行为的差异性，都会使得某些类别或标签的样本数量较少。

二是金融机构业务策略调整。金融机构可能由于市场变化、风险控制、国家政策变化等原因改变业务发展方向，使得部分类别或标签的样本数量较少。

三是数据获取难度大。例如，某些高风险客户不愿意或不配合提供部分数据，使得金融机构在业务开展过程中获取此类样本数据难度较大，进而导致此类样本数据量较少。

2.5 数据冗余

数据冗余性是指在数据集中存在重复、多余或者不必要的数据，这些数据不仅增加了数据处理的复杂性和成本，还可能影响模型的准确性和训练效率。数据冗余性问题在金融场景下大数据建模过程中较为常见。金融场景下为了满足业务开展需要，进行大数据建模所需要的数据涉及范围非常广泛，包括金融市场数据、企业数据、宏观经济数据和司法数据、舆情数据、房地产数据、反洗钱数据和反欺诈数据等。这些数据来源多样，有的来源于金融机构内部，有的来源于市场采集，有的来源于外部采购，不同来源的数据可能包含相同或类似的信息，容易导致数据冗余。同时，很多数据之间存在相关性，使得部分数据冗余，例如个人收入与个人支出、上市公司股票价格与市值之间均存在相关性。

3 金融场景下大数据建模常见的数据质量问题应对策略

3.1 数据缺失问题应对策略

金融场景下的数据缺失问题包括数据值缺失和数据标签缺失。对于数据值缺失，可根据具体情况选择删除、插值填充（如均值插值、中位数插值等）、预测补充（如决策树、随机森林等）或寻找替代数据；对于数据标签缺失，可通过调研和分析历史数据补充，若不影响建模也可删除。

3.2 数据错误问题应对策略

金融场景下的数据错误问题包括数据值错误和

数据标签错误。对于数据值错误，需分析原因并采取相应措施，如优化采集设备和技术、加强人员培训、强化数据预处理和清洗、选择高质量数据源等；对于数据标签错误，要及时更正或删除。

3.3 数据滞后问题应对策略

针对金融场景下的数据滞后问题，可以考虑采用以下策略：一是实时采集数据，或与数据供给机构建立实时数据连接；二是流式处理数据，对实时采集到的数据进行处理和分析；三是高性能计算，缩短建模时间，提高数据时效性；四是并行计算，同时处理多个子任务；五是数据压缩和存储优化，采用有效的数据压缩和存储优化方法；六是引入自动化建模工具，优化建模流程，减少建模误差，提高建模效率，从而更好地保障数据时效性。

3.4 数据失衡问题应对策略

针对金融场景下的数据失衡问题，可从以下方面考虑：一是数据增强。也即增加数量较少的类别或标签样本，使得数据保持均衡；二是算法调整和参数设置。通过调整模型对不均衡数据的处理方式，使得模型更加关注数量较少的类别或标签样本；三是特征选择和降维。通过特征选择和降维技术，删除冗余和无关的特征，减少特征的数量，从而提高数据的均衡性。

3.5 数据冗余问题应对策略

金融场景下的数据冗余问题体现在数据集中存在重复、多余或者不必要的数据。针对此类问题，一是需要对数据进行清洗，剔除重复、错误或不需要的数据；二是需要对数据进行整合，避免重复存储；三是需要对数据进行统计分析，检查数据之间的相关性和重复性；四是需要加强数据治理，建立健全数据治理体系。

4 结语

数字金融是“五篇大文章”发展的核心，其健康发展对于铸造新质生产力具有重要意义。工银科技积极发挥在大数据建模领域的优势，依托中国工商银行多年数字金融场景实践经验，开发形成的“工银e数”数据产品围绕数据集成、数据管理和数据应用的核心技术能力，为金融行业在数字化营销、风控、运营场景等方面提供金融级一站式数智集成解决方案。并且，该数据产品荣获了中国质量认证中心颁发的全国首张数据产品质量评价证书，经认证评价，该产品质量等级为AAAA级，表现出较大价值。未来，工银科技将继续探索内外数据在金融行业的价值，开发更多高质量的数据产品，释放数据生产力，促进数字金融的健康发展。❖

[参考文献]

[1] 潘锡泉. 高水平推进金融强国建设：理论蕴意、现实问题及实现机制——做好金融五篇大文章的视角[J]. 西南金融，2024 (04)：36-47.

[2] 张壹帆，陆岷峰. 数字金融对金融新质生产力提升的作用机制研究[J]. 河南社会科学，2024(05)：74-84.

[3] 石旭芳. 数字经济下金融数据治理的困局及对策探究[J]. 金融会计，2023 (06)：48-54.

[4] 田长星. 金融科技在金融数据治理领域的应用[J].金融纵横，2022 (11)：32-36.

[5] Fan L.，Zhang J. The application of big data in finance: International Conference on Computer Science， Communication and Signal Processing[C]. Singapore: Springer-Verlag，2019：667-675.

[6] 杜小勇，陈跃国，范举，等. 数据整理——大数据治理的关键技术[J]. 大数据，2019(03)：13-22.

[7] 张淑芬，尹振涛. 商业银行数字化转型的数据治理问题[J]. 银行家，2021 (02)：116-119.

[8] 胡亚茹，许宪春. 企业数据资产价值的统计测度问题研究[J]. 统计研究，2022，39(09) ：3-18.