Q/GFZQXX-00X-2023

广发证券股份有限公司企业标准

Q/GFZQXX-00X-2023

广发证券风险导向的审计模型开发标准

Development Standard for Risk-based Audit Model in GF SECURITIES

2023-0X-XX 发布　　2023-0X-XX 实施

广发证券股份有限公司 发布

目 次

前　言

本标准依据GB/T 1.1-2020《标准化工作导则 第一部分：标准化文件的结构和起草规则》给出的规则起草。

本标准由广发证券股份有限公司提出。

本标准由广发证券股份有限公司归口。

本标准起草部门：广发证券股份有限公司稽核部。

本标准主要起草人：徐佑军，宋弘涛，赵康毅，杨立峰，王文天，降泽一。

引　言

大数据作为新一代的技术与架构，被用于在成本可承受的条件下，通过快速的采集、发现和分析，从大体量、多样化的数据中提取价值；机器学习则对所研究的问题进行模型假设，利用计算机从训练数据中深度学习、得到模型参数，并以此对数据进行预测和分析。审计模型，是上述前沿科技在审计领域应用的核心载体，对于实现审计全覆盖、提升审计线索识别能力具有重要意义。

然而，传统上构建审计模型的方式方法，存在着基础信息来源不够全面、模型逻辑依赖主观经验、模型迭代需要人工推动等一系列制约模型开发和应用的难点，为规范和统一广发证券审计模型的开发方式、方法，建立自我驱动的风险导向工作流程，提高开发效率，提升模型质量，特制定本《广发证券风险导向的审计模型开发标准》。

本标准适用于公司内部所有基于案例的审计模型的开发工作，包括案例归集、案例处理、案例转化、数据工程、模型开发和模型管理等阶段。通过遵循本标准，能够更加高效地开发出高质量、高可靠性的审计模型，为审计团队提供重要审计线索，助力实现内部审计的核心价值目标，进而促进公司战略的达成。

广发证券风险导向的审计模型开发标准

1 范围

本标准规定了广发证券股份有限公司审计模型构建所依赖的案例归集、案例处理、案例转化、数据工程、模型开发和模型管理的规范。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41462-2022 基于文本数据的金融风险防控要求

JR/T 0236—2021 金融大数据术语

JR/T 0176.3—2021 证券期货业数据模型 第3部分：证券公司逻辑模型

JR/T 0200-2020 金融科技创新风险监控规范

3 术语和定义

3.1 违规行为

形成监管处罚、造成公司经济损失或其他不良后果的行为。

3.2 可疑样本

审计模型通过运算后输出的可能存在违规行为的样本集。

3.3 特征变量

案例分析提取的自变量，用于进行审计模型的搭建；或用于训练模型的数据的属性或变量，在监督学习中通常是模型输入的一部分。

3.4 标签

在监督学习中，标签是与输入数据相关联的目标变量或输出。模型的目标是学习如何从特征预测标签。

3.5 训练集

训练集是用于训练模型的数据子集，模型通过学习训练集中的数据来拟合关系。

3.6 验证集

验证集是用于模型调优的数据子集，在训练过程中调整模型的参数和超参数，以提高模型的性能。

3.7 测试集

测试集是用于最终评估模型性能的独立数据子集，它不用于模型的训练或验证。

3.8 超参数

超参数是模型训练过程中的配置参数，它们不是从数据中学习的，而是由开发者手动设置，例如学习率、批处理大小、模型深度等。

3.9 损失函数

损失函数是用于度量模型预测与实际标签之间差异的函数。它是优化算法的

一部分，用于调整模型参数。

4 案例归集

4.1 案例仓库

案例仓库本质上是一种数据库，用于归集和存储公司外部有关的风险事件信息、监管处罚信息以及公司内部审计案例信息、风险函件信息、合规问责信息等一切可以用于构建审计模型的案例及相关的案例标签，是实施风险导向审计的基础设施。

4.2 案例采集

通过技术手段加人工参与，定期检索和采集有关的国内外风险事件报道以及中国证监会、证券业协会、基金业协会、期货业协会、交易商协会、中国人民银行、外汇管理局等监管机构处罚案例；指定专人，定期收集公司内部检查（或审计）发现的典型（或共性）问题、风险函件内容、合规问责信息等。上述内外部案例，根据采集方式的不同，通过自动或手工的方式写入案例仓库，实施标准化管理。

5 案例处理

5.1 案例分类

案例分类旨在通过科学的划分标准，明确案例的主要性质及其模型应用领域，也是后续开展统计分析的重要基础。根据是否属于司法机关刑事处罚、中国证监会及其派出机构行政处罚措施、监管措施等范畴，对案例进行一维分类；结合公司投资银行、财富管理、交易及机构、投资管理四大业务板块以及具体业务线、职能线，对案例进行二维分类。

5.2 案例分级

案例分级的目的，是在有限的开发资源下，优先保障重要程度靠前的案例实现向模型的转化和应用。针对案例的一维分类，根据《证券公司分类监管规定》中的扣分标准，对各分类赋予相应的分数；针对案例的二维分类，依据我司研究型审计模型--Hurricane（飓风）雷达图，计算出各条线的风险程度得分。通过加权平均一、二维分类得分，得到案例中的事件对公司负面影响程度的估计值，进而将案例划分为劣后、普通、优先三级，首先确保优先级案例的转化。

5.3 特征提取

借助自然语言处理（NLP）技术，通过文本处理、命名实体识别、分类和信息抽取等方法，将自然语言文本转化为结构化的信息，即智能化地从案例文本中提取关键信息，如业务线、处罚主体、违规行为描述、处罚形式、处罚金额等，并将这些信息整合成案例对应的特征标签，以更有效地处理和利用大量文本数据，为案例分析和转化提供有力支持。

6 案例转化

6.1 案例分析

案例分析旨在对已完成分类分级和特征提取的案例，进一步抽象关于违规行为事实的描述，从中提取特征变量及判断逻辑，并转化为模型开发需求。本环节采用的主要方法包括特征分析法、专家分析法。

6.1.1 特征分析

结合案例事实，提取案例中违规行为的要点，识别违规行为的特征变量及判断逻辑，并转化为可量化指标的过程。

6.1.2 专家分析

根据案例的二维分类，召集对应条线的业务主管、审计项目组长（如涉及）形成专家组，通过询问、分析专家意见，形成可量化指标的过程。

6.2 特征识别

根据案例性质，简化违规行为事实，提取特征变量及判断逻辑。判断逻辑可参照前述自然语言处理技术提取的特征标签，多为“存在”、“不存在”等谓语词。提取特征变量时需要考虑如下因素：

如案例仅为事实阐述，无定性表述，则选取对应违规行为作为特征变量；

如案例包含定性表述，如“不规范”、“不健全”、“未勤勉尽责”等词语，则可以搜索相关法律法规，结合案例库中同类案例，确定违规行为构成要件，形成特征变量。

6.3 特征转化

选取特征变量中能够客观取值或判断的数值，形成模型变量阈值的输入值：

如特征变量为可以直接从公司数据平台提取的数值，则无需调整；

如特征变量数值无法直接从公司数据平台提取，则需要结合处罚依据、行业特点对特征变量进行调整，确定输入的阈值，必要时可以采用专家分析法确定。

7 数据工程

7.1 事前分析

在开始数据工程项目前，需对对模型所涉及的数据进行分析。了解模型所需处理和分析的数据类型、数据可获得性以及数据将如何用于决策支持，这一环节起着承前启后的作用，非常重要。

7.2 数据采集

收集与模型需求相关的数据。数据可以来自数据库、API、第三方数据提供商、公开数据采集等。

7.3 数据清洗

清洗数据旨在去除错误、缺失值和重复项。这一环节的工作包括数据验证、填充缺失值、处理异常值和标准化数据格式。

7.4 数据转换

在本环节，可能遇到需要对数据进行转换操作的情形，以便将其转为适合分析和建模的形式。转换操作包括数据合并、聚合、筛选、变换等。

7.5 数据打标

对数据集中的样本分配标签或注释，以支持机器学习模型的训练和数据分析任务。

7.6 数据存储

选择适当的数据存储系统，如 MySQL、MongoDB 数据库等；设计数据存储结构，包括表、架构和索引；完成模型所需数据的入库。

7.7 数据集成

如果数据来自不同的源头，需要将这些数据整合到一个统一的数据集中，以便进行跨数据源的分析。

7.8 质量监控

实施数据质量控制措施，确保数据的准确性、一致性和可靠性，包括监测数

据质量、记录数据质量指标和处理数据质量问题。

7.9 自动化调度

对数据工程的整个流程实施自动化，以便定期或实时地处理数据更新。通常使用工作流管理工具来调度数据工程作业。

7.10 安全与合规

采取数据加密、访问控制、合规性规定遵循等措施，确保数据的安全性和合规性。

8 模型开发

8.1 需求分析

模型开发前，需对案例处理阶段形成的案例类型、优先级以及案例转化阶段形成特征变量、判断逻辑，进行进一步分析，同时也会考虑数据的可获得性和模型的可实现性，以确保能够正确选择和规划模型开发。

8.2 模型选择

根据问题类型和开发需求选择适当的模型，本环节需要重点考虑以下因素。

8.2.1 问题类型

确定问题类型，例如分类、回归、聚类等，以决定选择哪种类型的模型。

8.2.2 数据质量

评估数据质量，包括数据的完整性、准确性、缺失值等。数据质量可能影响到模型的性能。

8.2.3 数据量

确保数据量足够大，以支持所选择的模型。复杂的深度学习模型需要大量数据来获得好的性能。

8.2.4 模型复杂性

考虑模型的复杂性和参数数量。在数据量有限的情况下，选择较简单的模型可能更容易泛化。

8.2.5 计算资源

考虑可用的计算资源，包括处理器、内存和 GPU。一些深度学习模型需要大量计算资源。

8.3 特征工程

特征工程旨在对原始数据进行处理和转换，提取、选择或创建更有信息量的特征，以供模型使用。

8.3.1 特征提取

从原始数据中提取数值、文本、图像或其他类型的特征，以便用于模型训练。

时间序列:提取滞后特征以考虑时间序列数据的历史信息。创建滚动统计特征，如滚动平均值或滚动标准差。

图像:使用预训练的卷积神经网络（CNN）模型提取图像特征。

文本：使用词袋模型、TF-IDF、Word2Vec等技术将文本数据转化为数值特征。进行文本分词、去停用词等文本预处理操作。

8.3.2 特征选择

选择最相关的特征，以减少维度和降低过拟合风险。基于统计测试或模型重要性进行特征选择。

8.3.3 特征变换

对数变换、标准化、归一化等操作，以确保特征的尺度一致，有助于模型训练。对类别特征进行独热编码或标签编码，以便模型能够处理。

8.3.4 特征创建

基于领域知识创建新的特征，例如从两个特征派生出一个新特征，或者构建交互特征。

8.3.5 降维

使用主成分分析（PCA）等降维技术来减少高维数据的复杂性。特征选择也可以被视为一种降维方法。

8.4 数据集划分

数据集划分是模型开发的关键步骤之一，涉及将可用数据集划分为训练集、验证集和测试集，以用于模型训练、调优和评估。正确的数据集划分对于获得可靠的模型性能非常重要。本环节需要考虑以下事项。

8.4.1 数据集

训练集：用于模型的训练，模型会根据训练数据来学习模式和关联关系。训练集占总数据集的大部分，通常在60%到80%之间。

验证集：用于模型的调优和选择超参数。在训练过程中，通过验证集来监控模型性能，以便及时调整模型。典型的划分比例是总数据集的10%到20%。

测试集：用于最终评估模型的性能，以模拟模型在实际应用中的表现。模型未见过测试集中的数据。测试集通常占总数据集的10%到20%。

8.4.2 交叉验证

对于较小的数据集，可以使用交叉验证来更充分地利用数据。常见的交叉验证方法包括k折交叉验证和留一交叉验证。交叉验证将数据划分为多个折叠（folds），每个折叠轮流作为验证集，其余折叠用于训练。

8.4.3 分层抽样

如果数据集中的类别不平衡（某些类别的样本数量远多于其他类别），需要采用分层抽样来确保训练集、验证集和测试集中的类别分布相似。

8.4.4 时间序列数据

对于时间序列数据，通常按照时间顺序划分数据集，以确保模型在未来的数据上进行测试。通常将早期数据作为训练集，中期数据作为验证集，最后一段时间的数据作为测试集。

8.4.5 随机化

在进行数据集划分时，可以考虑随机化样本的顺序，以减少可能的偏差。

8.5 模型搭建

模型搭建是模型开发的关键步骤，基于前期的需求分析、特征工程和模型选择，利用各种工具和框架来创建一个契合案例逻辑、适用于解决特定问题的机器学习或深度学习模型。这一过程需要考虑以下方面。

8.5.1 工具和框架

根据项目需求和选择的模型，确定适当的工具和框架。常见的工具和框架包括Scikit-Learn、TensorFlow、PyTorch等。

8.5.2 模型可解释性

如果模型的可解释性对问题至关重要，需要选择易于解释的模型结构。

8.5.3 模型架构

设计和定义模型的架构，例如网络层、神经元数量、激活函数等。模型架构需要与问题和数据相匹配，以便有效地捕获数据的特征。

8.5.4 模型初始化

初始化模型的权重和偏差，通常采用随机初始化的方式。权重初始化策略可以影响模型的训练速度和性能。

8.5.5 正则化

考虑是否需要添加正则化项，如L1正则化或L2正则化，以防止过拟合等问题。

8.5.6 批处理规范化

对于深度神经网路，需要考虑批处理规范化来加速训练和提高模型稳定性。

8.5.7 损失函数

选择合适的损失函数，用于衡量模型的预测与实际标签之间的差异。损失函数的选择与问题类型和模型架构密切相关。

8.5.8 优化算法

选择适当的优化算法，例如随机梯度下降（SGD）、Adam、RMSprop等，以用于调整模型参数以最小化损失函数。

8.5.9 超参数调优

调整模型的超参数，如学习率、批处理大小、正则化参数等，以达到最佳性能。通常使用交叉验证来进行超参数调优。

8.6 模型训练

使用训练集数据来拟合模型参数，监控训练过程，并使用验证集数据来调整超参数，以改进性能。

8.7 模型评估

使用测试集数据来评估模型的性能，考虑指标如准确度、精确度、召回率、F1 分数、均方误差等。

8.8 模型部署

将训练好的模型部署到生产环境中，以用于实际应用。本环节涉及到模型封装、API 创建、容器化等步骤。

8.9 监控和维护

在生产环境中监控模型的性能，定期组织模型使用情况回顾，根据需要重新训练模型以应对数据漂移，并进行必要的维护和更新。

9 模型管理及应用

9.1 模型组构建

模型组是一种逻辑组织结构，用于将相关模型组织在一起，以便更好地管理和检索。可以创建模型组来代表不同的业务板块、项目或应用场景（如：投资银行审计模型、财富管理审计模型、交易及机构审计模型、投资管理审计模型；操纵市场审计模型、洗钱识别审计模型、财务造假识别模型等）。

9.2 模型组细分

在每个模型组内部，根据需要划分模型类别或子组。这有助于更细粒度地组织模型，以便于查找和维护。如果模型组代表不同的业务板块，那么子组可以代表不同的任务或阶段（如：投资银行_IPO 审计模型组、投资银行_并购审计模型组等）。

9.3 模型入库

将审计模型及其相关文件（如：权重、配置文件等）添加到审计分析系统的模型仓库。确保每个模型都有详细的元数据，如描述、性能指标等。如果有需要，可以允许从外部导入模型。确保导入的模型符合规范，并进行合适的验证和测试。

9.4 模型文档

为每个审计模型建立文档或指南，包括模型适用场景、输入输出说明、超参数、数据集来源、部署和性能监控方法等信息，并写入审计工作手册，能够帮助其他审计成员更好地理解模型，思考并创造更多的应用场景。

9.5 模型应用

审计团队在审计项目实施前提起立项申请，经有权人员审批通过后，可以调用模型仓库中对应的模型组；审计团队无权修改模型，且调用行为有日志留痕。模型调用输出的结果，作为审计线索，在实施过程中得到应用，并为构建新的案例提供了重要基础。

10 保障机制

10.1 组织设置

为使风险导向的审计模型开发工作有效落地，审计团队内部设立案例组、模型组两个专责小组，分别指定具备相关资历的人员开展案例归集、处理和分析等案例相关工作，以及数据工程、模型开发和管理等模型相关工作。两个小组相互衔接，紧密配合。

10.2 培训交流

为了让审计团队及时了解新入库的案例和模型，案例组、模型组成员定期在审计例会中通报工作情况，就案例特征、模型使用进行讲解和交流。

10.3 考核安排

为使风险导向的审计模型开发闭环流程形成自我驱动机制，避免陷入对少数

人员经验的依赖或依靠人工介入推动流程运作，考核层面将涉及的案例和模型工作，作为相关人员的考核 KPI 或解锁已实现审计成效的钥匙（前置条件）。

参考文献

[1] GB/T 41462-2022 《基于文本数据的金融风险防控要求》

[2] JR/T 0236—2021 《金融大数据术语》

[3] JR/T 0176.3—2021 《证券期货业数据模型 第 3 部分：证券公司逻辑模型》

[4] JR/T 0200-2020 《金融科技创新风险监控规范》