附件

ICS 35.240.40
CCS A 11

JR

中华人民共和国金融行业标准

JR/T 0221—2021

人工智能算法金融应用评价规范

Evaluation specification of artificial intelligence algorithm in financial application

2021－03－26 发布　　2021－03－26 实施

中国人民银行　发布

目　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国人民银行提出。

本文件由全国金融标准化技术委员会（SAC/TC 180）归口。

本文件起草单位：中国人民银行科技司、中国金融电子化公司、北京金融科技产业联盟、中国工商银行股份有限公司、中国农业银行股份有限公司、中国银行股份有限公司、中国建设银行股份有限公司、中国银联股份有限公司、北京国家金融科技认证中心有限公司、交通银行股份有限公司、中国邮政储蓄银行股份有限公司、招商银行股份有限公司、上海浦东发展银行股份有限公司、中信银行股份有限公司、中国光大银行股份有限公司、华夏银行股份有限公司、中国民生银行股份有限公司、兴业银行股份有限公司、平安银行股份有限公司、北京百度网讯科技有限公司、中国平安人寿保险有限公司、财付通支付科技有限公司、北京银联金卡科技有限公司、华为技术有限公司、腾讯云计算（北京）有限责任公司、上海云从企业发展有限公司、神州数码信息服务股份有限公司、中金金融认证中心有限公司。

本文件主要起草人：李伟、李兴锋、程胜、郭栋、杨波、邱晓慧、谢国斌、黄本涛、关晓辉、高天、刘宝龙、夏知渊、黄炳、强锋、陈建军、刘龙、牛菲菲、王臻、渠韶光、丁平、郭铸、张发波、王大森、邱雪涛、黄勇、赵赟、李曹建、钟亮、贺瑶函、李锋、杨志、张彬、姬冰芳、林冠峰、黄杏、周辉、黎建辉、杨海钦、关成敏、吴永强、徐菁、许海洋、李兴建、李徐泓、符海芳、曹晓琦、李峰、陈明、蒋增增、李军、王进、温昱晖、刘书元、袁曈阳、李冬妮、刘文其、吴宝民、王飞宇。

人工智能算法金融应用评价规范

1 范围

本文件规定了人工智能算法在金融领域应用的基本要求、评价方法、判定准则。

本文件适用于开展人工智能算法金融应用的金融机构、算法提供商、第三方安全评估机构等。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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。

JR/T 0171-2020 个人金融信息保护技术规范

JR/T 0199-2020 金融科技创新安全通用规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

深度学习 deep learning

通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据分布式特征表示的一种学习方法。

3.2

部分依赖图 partial dependency plot；PDP

一种以可视化形式展示一个或两个特征对机器学习模型预测结果的边际效应的方法。

3.3

个体条件期望 individual conditional expectation；ICE

用于显示对于每一个样本实例而言当改变某一个特征值时人工智能算法预测结果如何改变的一种可视化方法。

3.4

全局代理模型 global surrogate model

一种通过训练全局样本对黑盒算法预测的可解释方法。

注：通过解释代理模型得出有关黑盒模型的具有可解释意义的结论。

3.5

代表性样本 prototype

能全面反映真实数据样本的一个数据实例。

3.6

非代表性样本 criticism

不能很好地由代表性样本表示的数据实例。

3.7

有影响力的样本 influential instance

对模型参数或预测值影响程度最大的样本。

3.8

对抗样本攻击 adversarial attack

通过在正常样本上添加难以察觉的微小扰动误导人工智能算法的攻击方法。

3.9

物理对抗攻击 physical adversarial attack

通过物理手段在真实世界构建对抗样本攻击人工智能算法的方法。

3.10

回归算法 regression algorithm

研究因变量和自变量之间关系的预测性建模技术。

3.11

决策树算法 decision tree algorithm

在已知各种情况发生概率的基础上，通过构成决策树来求取目标值的期望值不小于零的概率的一种算法。

注：常用的有ID3、C4.5和C5.0等。

3.12

图算法 graph algorithm

由节点表示随机变量、边表示变量之间依赖关系的图结构算法。

3.13

集成学习算法 ensemble learning algorithm

通过构建并结合多个基学习器完成学习任务的算法。

注：一般分为bagging、boosting和stacking等类别，代表性算法有随机森林、GBDT、XGBoost、LightGBM等。

3.14

模型无关可解释 model-agnostic method

将人工智能算法解释和模型分离，能够应用于任何模型并可实现模型灵活性、解释灵活性和表示方式灵活性的一种方法。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

AUC：ROC曲线下方面积（Area under the Curve of ROC）

C&W：Carlini和Wagner攻击（Carlini and Wagner's Attack）

ETL：数据抽取、转换、装载（Extract-Transform-Load）

FGSM：快速梯度下降法（Fast Gradient Sign Method）

FPR：假阳率（False Positive Rate）

GBDT：梯度提升决策树（Gradient Boosting Decision Tree）

IoU：交并比(Intersection over Union)

IV：信息量(Information Value)

JSMA：基于雅可比矩阵的显着性图攻击（Jacobian-based Saliency Map Attack）

mAP：平均精准率（mean Average Precision）

MAE：平均绝对误差（mean absolute error）

MSE：均方误差（Mean Square Error）

QPS：每秒查询率（Queries-per-second）

RMSE：均方根误差（Root Mean Square Error）

R^2：决定系数（R-Square）

RSD：相对标准偏差（Relative Standard Deviation）

TPR：真阳率（True Positive Rate）

TPS：每秒处理的事务数（Transactions Per Second）

XGBoost：极端梯度提升（Extreme Gradient Boosting）

5 总则

5.1 评价框架

本文件从安全性、可解释性、精准性和性能方面开展AI算法评价，适用对象分为资金类场景和非资金类场景。

AI算法安全性为算法在金融行业应用提供安全保障，是决定AI算法是否可用的基础，只有在满足安全性要求的前提下才能在金融领域开展应用。AI算法安全性评价主要从目标函数安全性、算法攻击防范能力、算法依赖库安全性、算法可追溯性、算法内控等方面提出基本要求、评价方法与判定准则等。

AI算法可解释性是判断算法是否适用的重要依据，可解释性越高算法内在逻辑、技术实现路径、决策过程、预期目标越明晰，算法更易于被理解、匹配、应用和管理。AI算法可解释性评价从算法建模准备、建模过程、建模应用三个阶段提出基本要求、评价方法与判定准则等。

AI算法精准性和性能是评价算法应用效果及目标预期的主要因素，一般而言精准性和性能越高算法应用效果越好。AI算法精准性和性能评价从算法建模过程、建模应用两个阶段提出基本要求、评价方法与判定准则等。

AI算法评价内容框架见下图。

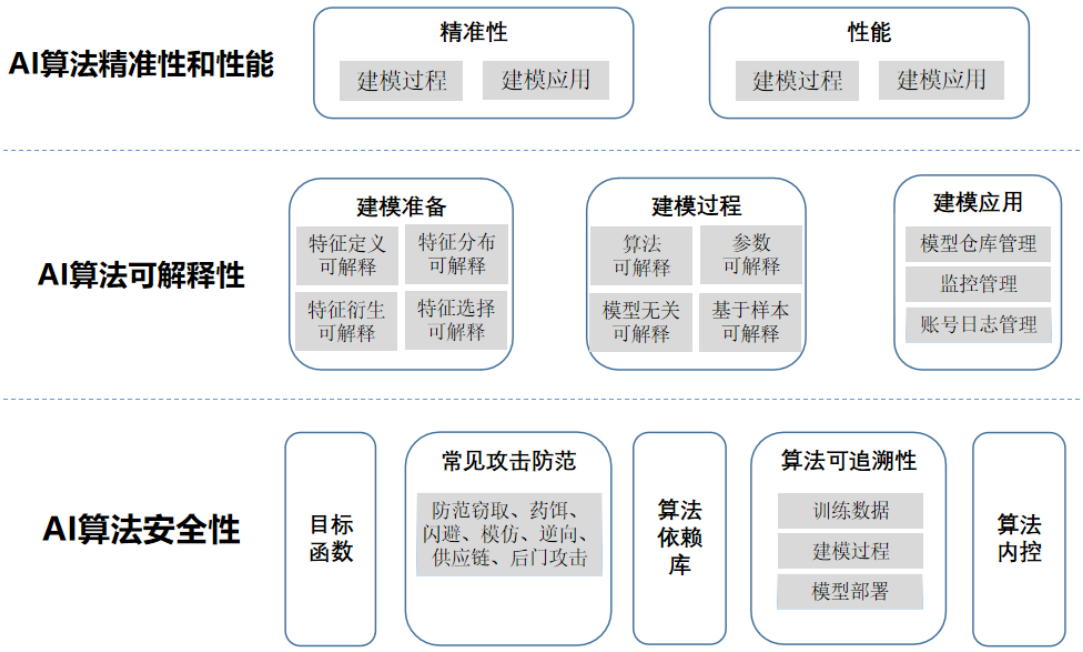

图　AI 算法评价内容框架图

5.2　评价方法

评价方法及说明如下：

a）　查阅材料：查阅审计报告、自查报告、外部评估报告、设计文档、开发文档、用户文档、管理文档、产品检测报告等相关材料。

b）　查看系统：查看系统日志、配置文件、参数设置、产品版本、网络配置等。

c）　访谈人员：与被测系统或产品有关人员进行交流、讨论等活动，获取相关证据，了解有关信息。

d）　系统测试：利用专业工具，通过对目标系统的扫描、探测等操作，使其产生特定的响应等活动，通过分析响应结果，获取证据以证明信息系统的基本要求、性能、安全性是否得以有效实施。

e）　攻击测试：利用专业攻击方法，对 AI 算法进行攻击，分析攻击结果，获取证据以证明系统的安全性是否得以有效实施。

f）　算法测试：基于业务样本数据，通过模型对目标变量进行预测，将预测出的结果和目标变量真实值进行比对并计算相应的 AI 算法评估指标。

g）　查看算法：查看代码、样本数据、训练数据等。

6　安全性评价

6.1　目标函数

人工智能算法金融应用目标函数评价内容见表1。

表 1　目标函数评价内容

序号	基本要求	评价方法	判定准则	备注
1	目标函数不应存在偏见歧视	查阅材料	设计文档有目标函数的说明，目标函数设计上不存在肤色、性别、国籍、年龄、健康等偏见歧视。	资金类场景、非资金类场景全部适用
2	算法表达能力应充分	1)查阅材料	设计文档对算法表达能力进行明确要求，能够适用于不同于训练阶段的全新使用情况。	
		2)系统测试	经测试，系统中算法表达能力与设计文档一致。	
3	目标函数运算成本应符合实施要求	1)查阅材料	设计文档对目标函数运算成本进行了分析，提出了明确要求。	
		2)查看系统	1.系统中未采用某种低计算成本的替代函数。 2.目标函数的运算成本符合JR/T 0199-2020的相关要求。	

6.2　常见攻击防范

6.2.1　窃取攻击防范

人工智能算法金融应用窃取攻击防范评价内容见表2。

表 2　窃取攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	应防止训练数据在传输、存储环节被窃取或篡改	1)查阅材料	管理文档对训练数据传输、存储环节制定了明确的安全管理要求。	资金类场景全部适用，非资金类场景适用第1、2、3、4项要求
		2)查看系统	系统在数据传输、存储方面采取的安全措施与管理文档要求一致。	
2	应防止模型在传输、存储环节被窃取或篡改	1)查阅材料	管理文档中有对模型传输和存储环节的安全管理说明。	
		2)系统测试	系统中算法模型在传输和存储环节不会被窃取或篡改。	
3	应保障训练数据隐私，避免用户敏感信息泄露	1)查阅材料	设计文档制定了安全防护措施，如差分隐私加噪等。	
		2)系统测试	1.系统采取与设计文档一致的安全防护措施，保障训练数据安全，防止用户敏感信息泄露。 2.训练数据的安全与隐私性符合JR/T 0171—2020的相关要求。	
4	应保障AI算法训练步骤安全	1)查阅材料	规划方案和实施方案中制定了AI算法训练过程安全保障措施。	
		2)查看系统	系统具有相应AI算法训练过程安全防护的配置，并有审计日志，保障训练步骤安全，防范训练过程被窃取，避免训练信息泄露。	

序号	基本要求	评价方法	判定准则	备注
5	AI算法应具有辨识度	1)查阅材料	规划方案和实施方案中制定技术保障措施，保证具有对恶意模仿算法的辨识度，如算法模型添加水印等方法。	资金类场景全部适用，非资金类场景适用第1、2、3、4项要求
		2)系统测试	系统能区分AI算法与恶意模仿的攻击算法。	

6.2.2 药饵攻击防范

人工智能算法金融应用药饵攻击防范评价内容见表3。

表 3 药饵攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	训练数据来源应可信、可靠	查阅材料	设计文档中有对训练数据来源的说明，明确训练数据不包含标签错误、记录造假等恶意伪造的药饵数据。	资金类场景、非资金类场景全部适用
2	应防止数据存储、使用等环节被投放药饵数据	1)查阅材料	设计文档或管理文档中有对训练数据存储、使用等环节的安全管理说明，防范训练数据被攻击或被投放药饵数据。	
		2)查看系统	系统中数据存储、使用等环节的安全管理与设计文档或管理文档的要求一致。	
3	训练数据分布应合理	1)查阅材料	设计文档或管理文档明确训练数据分布的具体说明，防止药饵攻击数据点混入，造成模型倾斜等错误状况。	
		2)查看系统	系统中训练数据分布与设计文档或管理文档的要求一致。	
4	宜具备检测数据集并过滤噪声和异常值的能力	1)查阅材料	规划方案或实施方案中明确采用回归分析等方法检测数据集，过滤其中的噪声和异常值。	
		2)系统测试	系统采取了与规划方案或实施方案一致的方法，能有效检测并过滤数据集中的噪声和异常值。	
5	宜具备通过综合多个独立子模型训练结果的方式增强AI算法抗药饵攻击的能力	1)查阅材料	规划方案或实施方案中采用集成分析等方式来提升AI算法抗药饵攻击能力。	
		2)查看系统	系统中有多个子模型，将多个子模型的综合结果作为最终结果。	

6.2.3 闪避攻击防范

6.2.3.1 对抗样本攻击防范

人工智能算法金融应用对抗样本攻击防范评价内容见表4。

表 4　对抗样本攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	算法应有效防御对抗样本攻击	攻击测试	开发文档或设计文档中根据业务应用场景需求制定了对抗样板攻击最低容忍成功率，经专家评估满足业务需求。 攻击测试时，基于FGSM、迷惑深度学习分类模型（DeepFool）、C&W、JSMA等算法生成对抗样本进行攻击测试，|测试攻击的最大成功率-设计的最低容忍成功率偏差|≤0.1。	资金类场景全部适用，非资金类场景适用第1项要求
2	应将对抗样本加入训练数据集	1）查阅材料	设计文档制定了将对抗样本加入训练数据集的设计要求。	
		2）查看系统	系统中记录了添加到训练集的对抗样本来源、数量等信息。	

6.2.3.2　物理对抗攻击防范

人工智能算法金融应用物理对抗攻击防范评价内容见表5。

表 5　物理对抗攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	算法应有效防范物理对抗攻击	攻击测试	AI算法性能最佳的条件下，当测试数据集中加入10%的物理对抗样本进行攻击时，资金类场景：|攻击后算法精准性-原算法精准性|≤0.05，非资金类场景：|攻击后算法精准性-原算法精准性|≤0.07。	资金类场景全部适用，非资金类场景适用第1项要求
2	应在算法使用阶段添加对抗样本检测模块	1）查阅材料	设计文档制定了算法使用阶段添加对抗样本检测模块的方案。	
		2）系统测试	算法包含对抗样本检测模块，且与设计方案一致。	

6.2.4　模仿攻击防范

人工智能算法金融应用模仿攻击防范评价内容见表6。

表 6　模仿攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	算法应有效防范模仿攻击	攻击测试	AI算法性能最佳的条件下，当测试数据集中加入10%的模仿样本进行攻击时，资金类场景：|攻击后算法精准性-原算法精准性|≤0.07，非资金类场景：|攻击后算法精准性-原算法精准性|≤0.09。	

6.2.5 逆向攻击防范

人工智能算法金融应用逆向攻击防范评价内容见表7。

表 7 逆向攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	应防止信息过度反馈	1）查阅材料	设计文档制定了反馈信息的要求，在满足用户需求的前提下，遵循最小够用的原则，避免反馈信息过多，造成逆向攻击。	资金类场景全部适用，非资金类场景适用第1、2项要求
		2）系统测试	系统中AI算法对反馈信息的保护和限制与设计文档一致。	
2	应限制探测频率	1）查阅材料	设计文档对AI算法探测频率进行要求和设计，避免恶意探测获取AI算法信息，逆向构建恶意模型。	
		2）系统测试	系统中AI算法对探测频率的限制和要求与设计文档一致。	
3	应增强对逆向攻击的防御能力	查阅材料	实施方案中采用隐私聚合教师模型等方式增强AI算法防御能力，有效防范逆向攻击。	

6.2.6 供应链攻击防范

人工智能算法金融应用供应链攻击防范评价内容见表8。

表 8 供应链攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	应对开源模型、第三方模型进行安全性风险评价	查阅材料	设计文档中对AI算法使用的开源模型、第三方模型是否存在安全风险进行评价，并给出了评价结果。	资金类场景、非资金类场景全部适用
2	在开源模型、第三方模型基础上做迁移调优时，应进行安全性加固	1）查阅材料	设计文档有在开源模型、第三方模型基础上做迁移调优时进行安全加固的方法介绍。	
		2）查看系统	系统中对AI算法安全性加固的方法与设计文档一致。	
3	基于开源模型、第三方模型做出的模型，应进行安全性风险评价	查阅材料	设计文档中有对基于开源模型、第三方模型做出模型的安全性风险评价，并给出了评价结果。	

6.2.7 后门攻击防范

人工智能算法金融应用后门攻击防范评价内容见表9。

表 9　后门攻击防范评价内容

序号	基本要求	评价方法	判定准则	备注
1	算法应用时，应对输入预处理，过滤掉能触发后门攻击的输入	1）查阅材料	规划方案或实施方案中有输入预处理相关说明，明确过滤能触发后门攻击的输入。	资金类场景、非资金类场景全部适用，满足任意一项要求即可
		2）查看系统	系统中AI算法应用有预处理模块，能对输入进行预处理。	
2	针对深度学习算法应具备采用模型剪枝技术的能力	查阅材料	规划方案或实施方案中明确采用模型剪枝技术，能适当剪除原模型神经元，减小后门攻击发生的可能性。	

6.3　算法依赖库

人工智能算法金融应用依赖库评价内容见表10。

表 10　算法依赖库评价内容

序号	基本要求	评价方法	判定准则	备注
1	应对开源学习框架以及依赖库的安全性进行评估	查阅材料	设计文档有对AI算法所使用的开源学习框架以及依赖库安全性评估的说明。	资金类场景、非资金类场景全部适用
2	应定期开展算法所用开源框架以及依赖库的内部审计	1）查阅材料	管理文档中有开展算法所用开源框架以及依赖库内部审计的相关说明。	
		2）查看系统	系统中记录了算法所用开源框架以及依赖库的内部审计结果。	

6.4　算法可追溯性

6.4.1　训练数据可追溯性

人工智能算法金融应用训练数据可追溯性评价内容见表11。

表 11　算法训练数据可追溯性评价内容

序号	基本要求	评价方法	判定准则	备注
1	应记录训练数据获取时间	1）查阅材料	设计文档中有要求记录训练数据获取时间的相关内容。	资金类场景全部适用，非资金类场景适用第1、2、3、4、5项要求
		2）查看系统	算法训练系统对训练数据获取时间进行了记录。	
2	应记录训练数据来源	1）查阅材料	设计文档中有要求记录训练数据来源的相关内容。	
		2）查看系统	算法训练系统对训练数据来源进行了记录。	
3	应记录训练数据量	1）查阅材料	设计文档中有要求记录训练数据量的相关内容。	
		2）查看系统	算法训练系统对训练数据量进行了记录。	

序号	基本要求	评价方法	判定准则	备注
4	应记录数据存储介质的标识	1）查阅材料	设计文档中有要求记录数据存储介质标识的相关内容。	资金类场景全部适用，非资金类场景适用第1、2、3、4、5项要求
		2）查看系统	算法训练系统对训练数据存储介质的标识进行了记录。	
5	训练数据应有完整签名或校验码	1）查阅材料	设计文档中有训练数据采用完整签名或校验码的说明。	
		2）查看系统	训练数据采用的完整签名或校验码与设计文档一致。	
6	应记录采样方法	1）查阅材料	1.开发文档中有数据采样方法的记录。 2.设计文档中有要求记录数据采样方法的相关内容。	
		2）查看系统	算法训练系统有训练数据采样方法的记录。	

6.4.2 建模过程可追溯性

人工智能算法金融应用建模过程可追溯性评价内容见表12。

表 12 算法建模过程可追溯性评价内容

序号	基本要求	评价方法	判定准则	备注
1	应保存建模过程中的建模脚本	1）查阅材料	设计文档中有要求保存建模过程中建模脚本的相关内容。	资金类场景、非资金类场景全部适用
		2）查看系统	算法训练系统中保存了建模过程中的建模脚本。	
2	应记录建模过程中的软硬件环境	查阅材料	开发文档记录了建模过程中使用的软硬件环境。	
3	应记录建模过程中的操作者	查阅材料	开发文档记录了建模过程的操作者。	
4	应记录建模的起止时间戳和迭代次数	1）查阅材料	设计文档中有要求记录建模的起止时间戳和迭代次数的相关内容。	
		2）查看系统	算法训练系统记录了建模的起止时间戳和迭代次数。	
5	应保存建模过程中参数迭代的相关记录	1）查阅材料	设计文档中有要求保存建模过程中参数迭代记录的相关内容。	
		2）查看系统	算法训练系统保存了建模过程中参数迭代的相关记录。	

6.4.3 算法部署可追溯性

人工智能算法金融应用部署可追溯性评价内容见表13。

表 13　算法部署可追溯性评价内容

序号	基本要求	评价方法	判定准则	备注
1	应记录AI算法部署的操作者	1）查阅材料	设计文档中有要求记录AI算法部署操作者的相关内容。	资金类场景、非资金类场景全部适用
		2）查看系统	系统记录了AI算法部署的操作者。	
2	应标识部署时间及相关结果	1）查阅材料	设计文档中有要求标识部署时间及结果的相关内容。	
		2）查看系统	系统标识了部署时间及相关结果。	
3	应保存AI算法部署过程的相关脚本	1）查阅材料	设计文档中有对保存AI算法模型部署相关脚本的说明。	
		2）查看系统	系统保存了AI算法部署过程的相关脚本。	
4	应记录部署的软硬件环境配置信息	查阅材料	开发文档中记录了部署的软硬件环境配置信息。	

6.5　算法内控

6.5.1　技术管理

人工智能算法金融应用技术管理评价内容见表14。

表 14　技术管理评价内容

序号	基本要求	评价方法	判定准则	备注
1	应建立算法上线应用前的内部评审机制	1）查阅材料	管理文档中明确要求建立包含评审委员会的评审机制，委员会负责算法评审工作，对算法进行评审，确定算法满足要求后才可上线应用。	资金类场景全部适用，非资金类场景适用第1、2、3项要求
		2）访谈人员	具备实现内部评审所需的资质与能力。	
2	应建立算法日常监测体系	1）查阅材料	管理文档中有监测体系的相关说明，监测系统来监测上线算法运行状态，及时反馈算法缺陷。	
		2）查看系统	系统已建立监测体系并与管理文档一致。	
3	应建立算法退出处置机制	查阅材料	管理文档中有明确算法退出处置机制的相关说明，对无法满足需求的算法应在保障安全的前提下停止使用，并及时采取相应措施消除算法退出带来的不利影响。	
4	应全面记录算法开发至退出的全过程	查阅材料	管理文档中有记录算法开发至退出全过程的相关说明。	

6.5.2　风险控制

人工智能算法金融应用风险控制评价内容见表15。

表 15　风险控制评价内容

序号	基本要求	评价方法	判定准则	备注
1	应具备AI算法突发情况应急处理机制	1）查阅材料	管理文档有处理AI算法突发错误状况的说明。	资金类场景全部适用，非资金类场景适用第1、2、3项要求
		2）查看系统	统预留干预措施来处理AI算法突发的错误状况。	
		3）访谈人员	管理员在使用干预功能前需要进行身份验证。	
2	应向用户充分提示AI算法的固有缺陷和使用风险	1）查阅材料	管理文档有对AI算法的固有缺陷和使用风险的说明。	
		2）查看系统	系统或应用已向用户提示AI算法的固有缺陷和使用风险。	
3	应建立风险赔偿机制	查阅材料	管理文档中有建立风险赔偿机制的相关说明，明确因违法违规或者管理不当等造成用户损失，依法承担损害赔偿责任。	
4	应具备AI算法道德风险防范机制	查阅材料	设计文档中有AI算法道德风险防范机制的相关说明。	

7　可解释性评价

7.1　可解释性评价维度

AI算法可解释性评价从算法建模准备、建模过程、建模应用提出基本要求、评价方法与判定准则等。

在建模准备阶段不需要对深度学习、集成学习算法进行要求。

在建模过程阶段的模型无关可解释、基于样本可解释方面，不需要对回归算法、决策树算法、图算法、其他统计学算法进行要求。

7.2　建模准备

7.2.1　特征定义可解释

人工智能算法金融应用特征定义可解释评价内容见表16。

表 16　特征定义可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	特征定义应满足相关业务逻辑和规则	查阅材料	设计文档有特征定义相关描述，能满足相关业务逻辑和规则。	资金类场景全部适用，非资金类场景适用第1、2项要求
2	特征定义应在系统中有明确记录	查看系统	系统中有特征定义明确记录。	
3	特征定义应有可查、详细的ETL指标加工过程的记录	查看算法	代码中有可查、详细的特征定义ETL加工过程的记录。	

7.2.2　特征分布可解释

人工智能算法金融应用特征分布可解释评价内容见表17。

表 17　特征分布可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	应符合日常生活场景和业务规则	查阅材料	设计文档中特征分布需符合日常生活场景和业务规则。	资金类场景全部适用，非资金类场景适用第1、2项要求
2	应确保特征分布合理，并具备对缺失值、异常值检验的能力	1）查阅材料	设计文档中有对缺失值、异常值检验的相关内容。	
		2）查看系统	系统中特征分布正常，同时能对缺失值、异常值等进行检验，与设计文档一致。	
3	应能展示特征的相关统计指标	查看系统	系统中能展示特征的平均值、最大值、最小值、中位数、众数、样本数等指标。	

7.2.3　特征衍生可解释

人工智能算法金融应用特征衍生可解释评价内容见表18。

表 18　特征衍生可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	特征衍生应合理	查阅材料	设计文档中有特征衍生的相关说明。其中，对于资金类场景，只允许离散的特征交叉，不允许复杂、毫无业务意义的特征衍生。	资金类场景全部适用、非资金类场景适用第1、2条要求
2	基于业务的特征衍生，应在系统中有明确记录	查看系统	系统中对于业务类特征衍生有明确记录，并能查询详细的ETL特征衍生加工过程。	
3	基于算法的特征衍生，应在系统中明确记录和展示	查看算法	代码中对于通过算法做特征衍生有明确记录，并能展示算法特征衍生过程和逻辑。	

7.2.4　特征选择可解释

人工智能算法金融应用特征选择可解释评价内容见表19。

表 19　特征选择可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	特征选择过程应有一定的量化统计指标作为决策依据	查看系统	系统中特征选择的过程有一定的量化统计指标作为决策依据。至少根据IV值、集成算法GBDT、XGBoost中的特征重要性方法、变量相关性、变量t显著性检验指标中一项指标进行特征选择。	资金类场景全部适用、非资金类场景适用第1、2项要求
2	特征选择过程不应有歧视性	查阅材料	设计文档中对于特征选择的过程，不能够有明显的歧视性。	

序号	基本要求	评价方法	判定准则	备注
3	特征选择的业务逻辑和算法依据应在系统中明确记录	1）访谈人员	访谈人员能说明特征选择的依据和业务逻辑及使用算法进行特征重要性选择的具体操作过程。	资金类场景全部适用、非资金类场景适用第1、2项要求
		2）查看系统	系统中有特征选择的业务逻辑和算法依据的记录。	

7.3 建模过程

7.3.1 算法可解释

7.3.1.1 回归算法

回归算法可解释评价内容见表20。

表20 回归算法可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	应通过统计指标对算法进行可解释性说明	查看系统	系统通过统计指标对算法进行可解释性说明，包括但不限于线性回归方程的回归系数、t检验值、F检验值。	资金类场景全部适用，非资金类场景适用第1项要求
2	应对算法整体进行可解释性说明	查看系统	系统提供算法整体的可解释性说明，包括但不限于R^2、MSE、RMSE。	

7.3.1.2 决策树算法

决策树算法可解释评价内容见表21。

表21 决策树算法可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	应通过可视化方式对算法进行可解释性说明	查看系统	系统利用可视化的方式展示决策树树形结构的实现过程进而对算法解释。	资金类场景全部适用，非资金类场景适用第1项要求
2	应对决策规则进行说明	查看系统	系统能对决策规则等进行说明，增强算法的可解释性。	

7.3.1.3 图算法

图算法可解释评价内容见表22。

表22 图算法可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	应通过图的表达、传播链条对算法进行可解释性说明	查看系统	系统通过图的表达、传播链条来说明算法的可解释性。	资金类场景全部适用，非资金类场景适用第1项要求
2	系统上宜能展示图算法的一度、二度关系和节点之间的传播关系等指标	查看系统	系统有专门的地方展示图算法的一度、二度关系和节点之间的传播关系等指标。	

7.3.1.4 其他统计学算法

其他统计学算法可解释评价内容见表23。

表 23 其他统计学算法可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	应对所使用的其他统计学算法进行必要性说明	查看系统	系统提供该统计学算法的必要性说明。	资金类场景全部适用，非资金类场景适用第1项要求
2	应通过至少一项具体统计指标对算法可解释性进行说明	查看系统	系统根据具体业务场景和算法，至少定义一项具体统计指标来对算法进行可解释性说明。	

7.3.1.5 集成学习和深度学习算法

集成学习和深度学习算法可解释评价内容见表24。

表 24 集成学习和深度学习算法可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	应对所使用的集成学习和深度学习算法进行必要性说明	查看系统	系统提供业务场景使用集成模型或深度学习模型的必要性说明。	资金类场景、非资金类场景全部适用

7.3.2 参数可解释

7.3.2.1 参数定义

参数定义评价内容见表25。

表 25 参数定义评价内容

序号	基本要求	评价方法	判定准则	备注
1	应对算法参数和超参数应有明确定义	查看系统	系统中每个算法中的算法参数和超参数在算法程序文件中有明确定义。	资金类场景、非资金类场景全部适用
2	基于业务的参数定义应符合业务逻辑要求	查阅材料	设计文档中有对基于业务实现的算法参数明确定义，且定义符合业务逻辑要求。	

7.3.2.2 参数选择

参数选择评价内容见表26。

表 26 参数选择评价内容

序号	基本要求	评价方法	判定准则	备注
1	应明确调参所依据的具体指标	查阅材料	设计文档和系统中需要明确调参所依据的具体指标。	资金类场景、非资金类场景全部适用

序号	基本要求	评价方法	判定准则	备注
2	应具备对不同调参方式的支持能力	查看算法	代码中对于手工调参，选择最常用的参数开始调参。对于自动化调参，有明确的调参工具和代码过程支撑。	资金类场景、非资金类场景全部适用

7.3.3 模型无关可解释

人工智能算法金融应用模型无关可解释评价内容见表27。

表 27 模型无关可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	部分依赖图应满足可解释性的技术要求	查看系统	1.系统定义部分依赖图函数反映特征与 AI 算法预测结果。 2.系统通过部分依赖图直观展示特征与 AI 算法预测结果之间的关系。	资金类场景、非资金类场景全部适用。 资金类场景需满足任意三项要求，非资金类场景需满足任意一项要求。
2	个体条件期望应满足可解释性的技术要求	查看系统	1.系统中需通过中心化个体条件期望图实现可解释性。 2.系统中需通过导数个体条件期望图，观察目标函数相对于特征的导数，实现可解释性。	
3	累积局部效应应满足可解释性的技术要求	查看系统	1.系统中需将单个局部效应进行累积来反映单一特征变量对预测结果的整体影响情况。 2.系统中需利用强相关的两个变量对预测结果的联合效应实现对算法的可解释性。	
4	全局代理模型应满足可解释性的技术要求	查看系统	1.系统中需通过全局代理模型对 AI 算法进行近似，并对其进行解释。代理模型必须为可解释性模型。 2.系统中需全局代理模型与原模型的相似性可通过 R2 进行测量。	
5	局部代理模型应满足可解释性的技术要求	查看系统	系统中需通过局部代理将输入的可解释表示样本集中的每个新样本转换为原黑盒模型可识别的输入矩阵，并获取这些新样本的预测标签，最终以可解释表示样本集及其预测标签集为输入，以输入样本与待解释样本的相似度为度量，以加权的方式训练生成线性模型，从而评估可解释表示样本集中每个样本对该样本预测的影响。	
6	Shapley 值应满足可解释性的技术要求	查看系统	1.系统中需将模型预测类比为多个特征成员的合作问题，将最终预测结果类比合作中的总收益，特征的贡献程度将决定其最终分配到的收益——重要性评估值。 2.系统中建议使用常用的 SHAP(SHapley Additive exPlanations)。	

7.3.4 基于样本可解释

人工智能算法金融应用基于样本可解释评价内容见表28。

表 28 基于样本可解释评价内容

序号	基本要求	评价方法	判定准则	备注
1	反事实解释应满足可解释性的技术要求	算法测试	使用算法测试方法，对特征变量取否，看看是否必定会发生目标变量也取否。	资金类场景、非资金类场景全部适用。 资金类场景需满足任意两项要求，非资金类场景需满足任意一项要求。
2	代表性和非代表性样本应满足可解释性的技术要求	查看系统	1.系统中在样本中要求有能够标识代表性样本或非代表性样本，并对其进行相关说明。 2.系统中代表性样本或非代表性样本能够解释 AI 算法模型。	
3	有影响力的样本应满足可解释性的技术要求	查看系统	1.要求在系统中有一个功能点，可以识别出训练样本中最有影响力的样本。 2.系统中能够给出样本最有影响力的说明。 3.系统中最优影响力的样本能够对模型进行解释。	

7.4 建模应用

7.4.1 模型仓库管理

模型仓库管理评价内容见表29。

表 29 模型仓库管理评价内容

序号	基本要求	评价方法	判定准则	备注
1	对模型进行管理，提供对模型的可回溯方面的功能	查看系统	1.每次建模完成的模型需要导入到系统中的模型仓库进行集中管理。 2.系统中能够从模型仓库中方便的调取、查阅每个模型的要素，包括但不限于模型构建时间、构建人、模型的参数、进入模型的特征变量等。	资金类场景、非资金类场景全部适用
2	对模型版本进行管理，提供对版本可回溯方面的功能	查看系统	1.模型升级完成后，旧版本的模型在系统中保留或存有备份。 2.系统日志中留有模型版本升级的审计记录，保证模型的版本更新可追溯。	

7.4.2 监控管理

监控管理评价内容见表30。

表 30　监控管理评价内容

序号	基本要求	评价方法	判定准则	备注
1	提供算法上线的相关功能	查看系统	1.系统中对每次模型的上线应用，需进行相应的记载。 2.系统中能够从上线管理模块中方便的调取、查阅每个上线模型的一些要素，包括但不限于模型的上线时间、下线时间、对应的数据源、业务场景等。	资金类场景、非资金类场景全部适用
2	提供算法监控的功能	查看系统	1.系统中需对已上线的模型提供必要的监控功能。 2.系统中需监控模型运行的状态、硬件资源耗用情况、异常报警情况、特征变量的波动情况、目标变量预测后的分布情况等。	

7.4.3　账号和日志管理

账号和日志评价内容见表31。

表 31　账号和日志评价内容

序号	基本要求	评价方法	判定准则	备注
1	提供账号管理的功能	查看系统	系统中能够对构建算法模型的账号实现密码验证、账号间权限隔离等安全功能。能够根据账号追踪到具体的建模人员。	资金类场景、非资金类场景全部适用
2	提供日志管理的功能	查看系统	系统中能够提供必要的日志管理功能，对模型构建和操作的一些细粒度过程进行记载。	

8　精准性评价

8.1　精准性评价维度

从建模过程、建模应用两个建模阶段来提出基本要求、评价方法、判定准则。

8.2　建模过程

把建模过程中数据样本分为训练集、验证集和测试集。建模过程的AI算法精准性评价内容见表32。

表 32　建模过程的 AI 算法精准性评价内容

序号	基本要求	评价方法	判定准则	备注
1	二分类算法应通过算法精准性的指标评估要求	算法测试	在测试集上： 1.资金类场景 AUC≥0.75。 2.非资金类场景 AUC≥0.65。	资金类场景、非资金类场景全部适用

序号	基本要求	评价方法	判定准则	备注
2	多分类算法应通过算法精准性的指标评估要求	算法测试	在测试集上： 1.资金类场景微平均F1或宏平均F1≥0.20。 2.非资金类场景微平均F1或宏平均F1≥0.15。	资金类场景、非资金类场景全部适用
3	回归算法应通过算法精准性的指标评估要求	算法测试	在测试集上： 1.资金类场景RSD≤0.1。 2.非资金类场景RSD≤0.2。	
注：1.对于不适用于二分类、多分类、回归算法指标评测的应用场景，精准性可使用ABtest进行评测，应用AI算法后的精准性指标应优于基准模型。 2.金融行业AI算法精准性的相关指标定义见附录。				

8.3 建模应用

建模应用的数据样本需取线上业务应用时的离线或在线样本。建模应用的AI算法精准性评价内容见表33。

表33 建模应用的AI算法精准性评价内容

序号	基本要求	评价方法	判定准则	备注
1	二分类算法的线上预测效果评估要求，和建模过程中，线下训练指标保持一致	算法测试	1.资金类场景 AUC≥0.75。 2.非资金类场景 AUC≥0.65。	资金类场景、非资金类场景全部适用
2	多分类算法的线上预测效果评估要求，和建模过程中，线下训练指标保持一致	算法测试	1.资金类场景 微平均F1或宏平均F1≥0.20。 2.非资金类场景 微平均F1或宏平均F1≥0.15。	
3	回归算法的线上预测效果评估要求，和建模过程中线下训练指标保持一致	算法测试	1.资金类场景 RSD≤0.1。 2.非资金类场景 RSD≤0.2。	
4	二分类算法有一定泛化性，线上预测和线下训练的指标差异应在一定范围内	算法测试	|线上预测AUC - 线下训练AUC|≤0.1。	
5	多分类算法有一定泛化性，线上预测和线下训练的指标差异应在一定范围内	算法测试	微平均F1差异或宏平均F1差异≤0.1。	
6	回归算法有一定泛化性，线上预测和线下训练的指标差异应在一定范围内	算法测试	RSD差异≤0.1。	
注：1.对于不适用于二分类、多分类、回归算法指标评测的应用场景，精准性可使用ABtest进行评测，应用AI算法后的精准性指标应优于基准模型。 2.对于业务冷启场景，在精准性指标评测，可根据实际情况酌情放宽标准。				

9 性能评价

9.1 性能评价维度

从建模过程、建模应用两个建模阶段来提出基本要求、评价方法、判定准则。

9.2 建模过程

建模过程的AI算法性能评价内容见表34。

表 34 建模过程的 AI 算法性能评价内容

序号	基本要求	评价方法	判定准则	备注
1	整个训练过程的总时长应通过指标的评估要求	算法测试	训练过程总时长≤1 天。	资金类场景、非资金类场景全部适用
注：1.在安全可信计算或联邦学习场景下，由于涉及加密和网络通信，模型训练时间较长。具体性能和加密算法、网络状况有关，需单独另行定义。 2.对计算机视觉、语音识别、自然语言处理的金融应用场景不作算法性能方面的评估要求。				

9.3 建模应用

建模应用的AI算法性能评价内容见表35。

表 35 建模应用的 AI 算法性能评价内容

序号	基本要求	评价方法	判定准则	备注
1	实时系统的单条预测响应时间应通过指标评估要求	算法测试	1.资金类场景 单条预测≤10min。 2.非资金类场景 单条预测≤1h。	资金类场景、非资金类场景全部适用
2	实时系统的批量预测 1000 条响应时间应通过指标评估要求	算法测试	1.资金类场景 批量预测 1000 条≤1h。 2.非资金类场景 批量预测 1000 条≤1 天。	
3	实时系统的 QPS 应通过指标评估要求	算法测试	1.资金类场景 QPS≥20。 2.非资金类场景 QPS≥5。	
4	实时系统的 TPS 应通过指标评估要求	算法测试	1.资金类场景 TPS≥20。 2.非资金类场景 TPS≥5。	
注：根据业务需求定义自身 TPS 值、QPS 值。				

附录
（资料性）
金融行业AI精准性的相关指标定义

1 对于分类算法模型

以二分类为例，混淆矩阵，见下表。

表 混淆矩阵

混淆矩阵		真实值	
		正	负
预测值	正	TP	FP
	负	FN	TN
注：1.真正例(True Positive，TP)为被模型预测为正的正样本。 2.假正例(False Positive，FP)为被模型预测为正的负样本。 3.假反例(False Negative，FN)为被模型预测为负的正样本。 4.真反例(True Negative，TN)为被模型预测为负的负样本。			

准确率：

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

精准率：

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

召回率：

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

F1值：

$$F_1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{4}$$

AUC值：

$$AUC = \frac{\sum_{ins_i \in positiveclass} rank_{ins_i} - \frac{M\times(M+1)}{2}}{M\times N} \tag{5}$$

式中：

AUC ——AUC值；

$rank_{ins_i}$ ——第i条样本的序号（概率得分从小到大排，排在第rank个位置）；

M ——正样本的个数和；N为负样本的个数；

$\sum_{ins_i} \in positiveclass$ ——把正样本的序号加起来。

KS（Kolmogorov-Smirnov）值：

$$TPR = \frac{TP}{TP+FN} \tag{6}$$

——TPR为真阳性率，表示当前分到正样本中真实的正样本所占所有正样本的比例。

$$FPR = \frac{FP}{FP+TN} \tag{7}$$

——FPR为假阳性率，表示当前被错误分到正样本类别中真实的负样本所占所有负样本总数的比例。

KS = max|TPR - FPR|

2 对于回归算法模型

MAE 是绝对误差的平均值。

$$MAE = \frac{1}{n}\sum_1^n |(actual_k - predicted_k)| \tag{8}$$

式中：

——$actual_k$ 为真实值；

——$predicted_k$ 为模型预测值。

RMSE 是预测值和实际观测之间平方差异平均值的平方根。

$$RMSE = \sqrt{\frac{\sum_1^n (actual_k - predicted_k)^2}{n}} \tag{9}$$

式中：

$actual_k$ ——真实值；

$predicted_k$ ——模型预测值。

3 对于图像识别模型

IoU：

$$IoU = \frac{area\ \ of\ \ overlap}{area\ \ of\ \ union} \tag{10}$$

式中：

$area\ of\ overlap$——重叠区域面积；

$area\ of\ union$ ——为集合区域面积。

各类别mAP：

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{11}$$

式中：

AP_i——在一定的 IoU 阈值下，某类别的中的 P-R 曲线下的面积，即不同阈值下精度的均值。

4 对于自然语言处理模型

针对分类的每一个类别，评估指标如召回率、准确率、F1值，计算方式和多分类一样。

对单个类别的分类精度，有微平均与宏平均。

微平均从分类器的整体角度考虑，不考虑分类体系的小类别上的分类精度。微平均利用被正确分类标注的文本总数a_{all}、被错误分类标注的文本总数b_{all}，以及应被正确分类标注而实际上却被错误地排除的文本总数c_{all}分别替换“对于分类算法模型”中的TP、FP、FN得到微平均召回率、微平均准确率和微平均F1值。

微平均F1值计算如下式所示：

$$MicroF_1 = \frac{2\times MicroR\times MicroP}{MicroR+MicroP} = \frac{a_{all}}{a_{all}+b_{all}} = \frac{\sum_{i=1}^{P} a_i}{\sum_{i=1}^{P}(a_i+b_i)} \tag{12}$$

式中：

MicroR——为微平均召回率；

MicroP——为微平均精准率。

宏平均是从分类器小类别的整体考虑，首先计算出每一类别的召回率与准确率，然后对召回率与准确率分别取算术平均值得到的宏平均召回率与宏平均准确率。最后根据宏平均召回率与宏平均准确率计算宏平均F_1值。

a）宏平均召回率

$$MacroR = \frac{1}{p}\sum_{i=1}^{p} recall_i \tag{13}$$

式中：

Recalli——类别 ci 的召回率；

p　　——为分类体系类别数目。

b）宏平均精准率

$$MacroP = \frac{1}{p}\sum_{i=1}^{p} precision_i \tag{14}$$

式中：

precision——为类别 ci 的精准率；

p　　——为分类体系类别数目。

c）宏平均F1值

$$MacroF_1 = \frac{2\times MacroR\times MacroP}{MacroR+MacroP} \tag{15}$$

式中：

MacroR——为宏平均召回率；

MacroP——为宏平均精准率。

宏平均考察分类器对不同类别的处理能力。尤其在非平衡数据集上，宏平均能够更好地衡量分类器处理小样本类别的分类能力。换句话说，微平均从文本分类标注正确总数角度衡量分类精度，宏平均从每一类别文本标注正确的角度衡量分类精度。

参 考 文 献

[1] GB/T 22081—2016 信息技术 安全技术 信息安全管理实践指南
[2] GB/T 27910—2011 金融服务信息安全指南
[3] JR/T 0071—2020 金融行业网络安全等级保护实施指引