新模型，新风险：如何有效管理机器学习与人工智能

现今机器学习模型愈发复杂，新风险丛生。合理调整传统风险管理模式下的验证框架，可更好地管理模型，降低风险。

作者：Bernhard Babel、Kevin Buehler、Adam Pivonka、Bryan Richardson 和 Derek Waldron

利用海量数据所构建的模型，机器学习和人工智能将优化商业决策，提供定制服务，改进风险管理。两者带来的优势也注定会为银行业带来翻天覆地的变化。麦肯锡全球研究院的数据显示，这些技术的应用有望为银行业创造超过 2500 亿美元的价值。

但是，机器学习模型的应用也放大了某些传统模式下的风险。目前，大多数银行在对模型风险进行评估和管理时，采用的都是传统风险管理模式下相对成熟的模型验证框架。这些传统做法虽然能够满足监管合规要求，但仍不足以有效管理与机器学习模型相关的新型风险。

考虑到其管理难度，多数银行都在谨慎前行。比如，它们会试探性地将机器学习模型应用于数字营销等低风险业务中，以测试银行可能会面临的财务、声誉和监管风险。银行害怕自身会不知不觉地触犯反歧视法，从而招致巨额罚款。出于这一担忧，一家银行明令禁止其人力资源部门使用基于机器学习的简历筛选器。鉴于上述情况，如果银行想最大程度地从机器学习模式中受益，更好的、可能也是唯一可持续的办法，就是加强模型风险管理。

目前，监管机构尚未发布任何具体章程，来引导企业如何管理机器学习和人工智能相关模型。在美国，监管机构规定，银行必须负责管理机器学习模型带来的所有风险。与此同时，他们也指出，诸如美联储此前颁布的”模型风险管理指南”（Guidance on Model Risk Management）(SR11-7)等现有监管准则的内容已足够宽泛，可作指导手册而用。

可喜的是，许多银行并不需要通过建立全新的模型验证框架，来应对机器学习模型的风险。它们大可对现有模型管理的验证框架进行一系列补充，以达到同样目的和效果。例如，它们可将新模型纳入模型清单中，并确定相应的风险偏好、风险层级、风险角色、管理职责，以及模型生命周期管理中相关的模型验证技术。

新风险、新选择、新实践

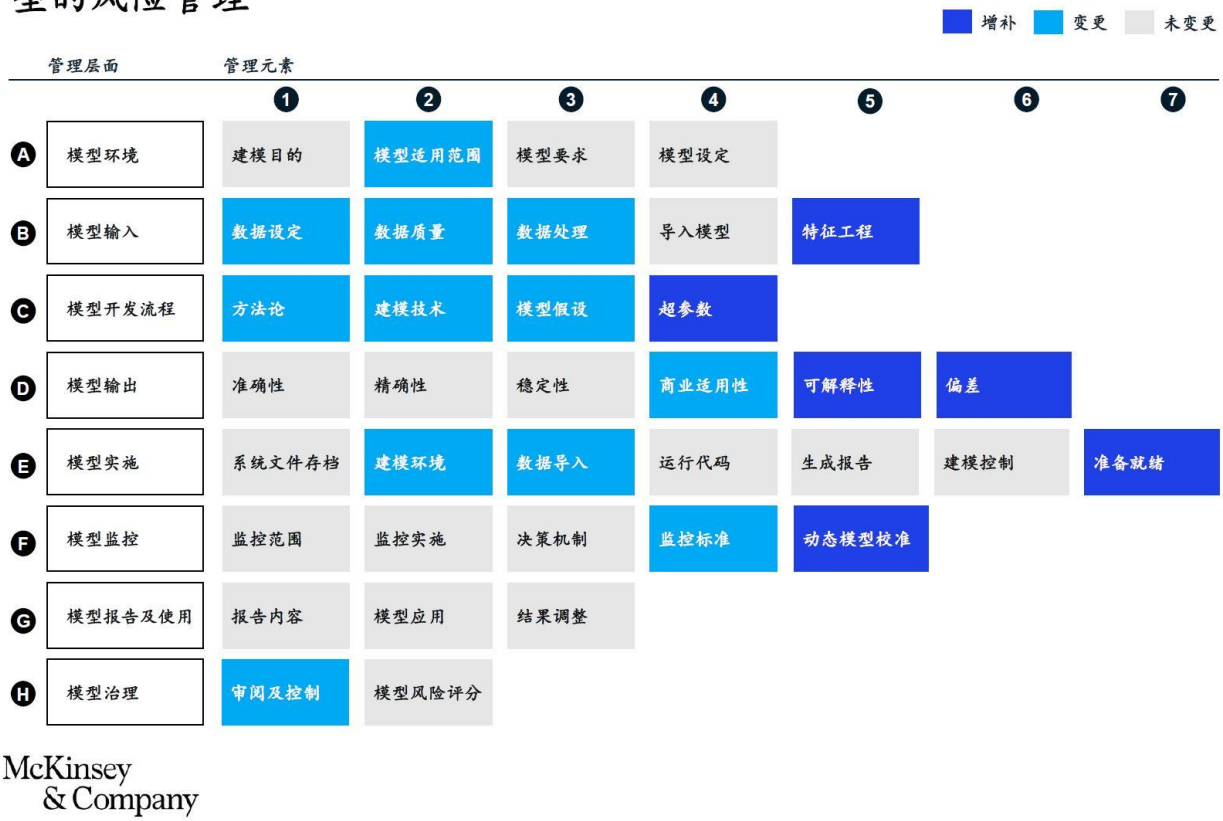
近年来，新兴机器学习模型产生的负面新闻并不少见。2016 年，算法的逆向反馈机制直接导致英镑“闪崩”6%。此外，一辆基于机器学习技术而研发的自动驾驶汽车，也未能正确识别并避让一名推着自行车过马路的行人。

无论机器学习模型被应用于何种行业或应用，这些风险发生的原因，与所有机器学习模型中风险被放大的原因其实都相同：即模型复杂性的大幅增加。机器学习模型通常基于大规模的非结构化数据集（如自然语言、图像和语音信息等），并使用新的软件包和特定的计算基础架构进行构建。这些算法比传统的统计方法要复杂得多，往往需要在测试训练环节开始前，

就做好设计决定。

然而，模型本身的复杂并不意味着我们也要采取过度复杂的应对方式。如下图所示，只要理解得当，银行现有的传统模型验证框架，完全能够有效管理与机器学习模型相关的风险。

对现有传统风险管理模型验证框架进行增改，可实现对机器学习模型的风险管理



从上图中，我们可以清楚地看到，麦肯锡 Risk Dynamics 模型风险验证和管理团队，对模型验证框架和实践方法作出了调整。这一框架覆盖了 SR11-7 的监管要求，曾被用于验证银行业数千个传统模型。它的审验范围涵盖 8 大风险管理层面，共计 25 个风险要素。针对机器学习和人工智能技

术相关模型，模型风险验证和管理团队修改了 12 个已有要素，增补了 6 个新要素，让银行能够借助新模型来有效识别和管理与机器学习相关的风险。

六大新要素

这六大新要素(可解释性、偏差、特征工程、超参数、生产就绪和动态模型校准)代表了对传统验证框架最根本的增改。

可解释性(Interpretability)

受模型架构的牵制，机器学习生成的结果有时会难以理解或作诠释。因而，机器学习又常被称为“黑匣子”。例如，为了帮助业务经理交叉销售，某银行花费数月开发了一个基于机器学习的产品推荐引擎。然而，由于业务经理无法理解模型为何会做此推荐，便决定无视这些建议，甚至对模型采取置之不理的态度。这种忽视会直接带来人力资源的浪费，甚至可能还会错失商业机会。不过，如果一味地听从模型并采取行动，而不深究其背后的原因，可能也会带来严重的后果。

对银行而言，决定机器学习模型的可解释性应到达何种程度，是银行应根据其自身风险偏好而作出的一个政策规定。银行可以规定所有机器学习模型的可解释性都必须保持在统一的高标准，也可以选择根据模型风险的不同而进行具体区分。以美国为例，决定是否批准借贷申请的模型受美国公平信贷法管辖，因此，当模型做出拒绝的决定时，必须提供明确的原

因代码。有些时候，银行可能会认为，机器学习模型做出的某些决策不会对银行带来太多风险——如在特定客户的移动应用上投放产品广告。在这种情况下，了解模型做此决定的原因就没那么重要了。

验证人员还需确保模型符合所选策略。幸运的是，尽管机器学习模型一直有“黑匣子”的别称，但近年来，我们确实在其结果的可诠释性方面取得了重大进展。基于模型类别，我们可从一系列方法中做选择：

偏差(Bias)

一般来说，模型主要会受到四种偏差的影响：样本偏差、测量偏差、算法偏差，以及对特定人群偏见的偏差。在机器学习模型中，后两种类型（即算法和偏见）的偏差可能会被放大。

具体来看，随机森林算法倾向于采用价值更为明确的输入值，但这样会增加决策欠佳的风险。例如，某银行开发了一个随机森林模型，以期识别潜在的洗钱活动。他们发现，该模型倾向于采用具有大量分类值的字段（如职业）。但事实上，某些分类值较少的字段(如国家)则能更好地预测洗钱的风险。

为解决算法偏差，我们应更新模型验证过程，以确保在任何给定情况下，都能选择出合适的算法。当然，有时候也存在一些技术解决方案，比如随机森林模型的特征选择。如果没有技术解决方案，便可换种思路，比如建立“挑战者”模型，即用其他算法来对标该算法的表现。

想要解决针对特定人群的偏见偏差，银行必须首先确定，公平的评判标准是什么。以下四个评判标准最广为人知，但具体的使用情况还要视模型的选择而定:

模型验证者需要确认开发者已经采取了必要的措施来保证公平。在模型开发的各个阶段，验证者可对这些模型进行公平性测试，在必要的情况下，会对从模型设计到模型性能监控的各个阶段进行修正。

特征工程(Feature engineering)

相较于传统模型，机器学习模型的特征工程更为复杂。原因有以下几点：第一，机器学习模型可以容纳海量的信息。第二，机器学习模型基于非结构化的数据源（如自然语言），而这些非结构化数据通常在数据集训练前就需要特征预处理。第三，现在已有越来越多的商业机器学习包都在提供所谓的自动机器学习（AutoML），自动机器学习可以生成大量的复杂特征来测试多种数据转换。使用这些特征产生的模型可能会非常复杂，从而导致过度拟合。比如说，某机构使用了一个商业自动机器学习（AutoML）平台搭建模型，结果发现，一款产品应用程序中的特定字母序列会被视作