



大数据“降噪”方法论

部分机构掌握了一定量的客户信息数据，就以为掌握了大数据，忽视对数据分析工具和方法论的研究。在金融业务中，这有可能影响其对风险的识别和防控，并造成风险的积聚和扩散。

文 | 杨凯生

最近,《互联网金融风险专项整治工作实施方案》全文网上曝光,包括第三方支付、P2P网贷、股权众筹、互联网保险、投资理财,以及互联网跨界资管,甚至互联网广告,都将面临一轮集中整治。

分析整治的原因,会发现在法规不够健全、监管不够有力,行业自律较弱,投资者教育欠缺外,还有一个重要的症结——企业和机构对大数据技术的理解和认识存在一定偏差。

大数据技术的发展和进步给人们提供了新的工具,即从更宽视野、更多维度、更全方位来认知问题和分析问题的方法。但部分机构掌握了一定量的客户信息数据,就以为掌握了大数据,忽视对数据分析工具和方法论的研究。在金融业务中,这有可能影响其对风险的识别和防控,并造成风险的积聚和扩散。

因此,在互联网时代,企业和机构对大数据的认识,需要结合正确的方法论、认识论,处理好碎片化的信息和完整性的数据、以及结构性的数据和非结构性的数据的关系。

大小不能以量区分

有人认为,有了大数据,就可以轻视对传统小数据的开发和利用。但大数据有大数据的长处,大数据也有大数据的不足,两者无法相互替代。

尽管迄今为止,并没有对大数据统一而权威的定義。但大数据的重要特征,在于它应该既包括结构性的数据,以及在生成的时候表现为非结构性数据的信息。而小数据,主要是指传统的二维结构性数据。

从技术角度上看,传统的小数据有经典的数理统计分析模型和成型的挖掘技术。而大数据的管理理论、分析方法仍在快速发展和跟进,特别是要采集、挖掘和使用非结构性的数据,仍没有完全成型或者定型。此外,

还有一部分非结构性的数据,在最后使用的时候需要通过技术手段把它转换成结构化数据才能实现。

从处理角度上看,大数据会随着数据量的急剧增加,其中的数据噪音会有快速增长。有时,数据噪音的增长幅度会快于数据量的增长幅度。因此,在大数据领域,对其挖掘、筛选、清洗的成本,将会明显高于小数据。

从相互关系的角度看,大数据通常比较容易反映的是事物的相关关系,而小数据往往容易得出的是事物的因果关系。在很多情况下,相关关系是不能简单地代替因果关系的。小数据它可以抽取世间的一些事物最核心的内容,最基本的内容。因此与大数据相比,小数据的单位信息容量更大,所以大数据的颗粒度和小数据的颗粒度不同。

以银行的数据为例,我们经常定义其为小数据,因为它都是以会计为基础,以计算为方式表现出来,反映了交易活动最核心的内容和最终的结果。比如,客户存款多少、贷款多少、买了黄金多少等等。但是,客户之所以进行这个交易、他的决策过程、行为路径,就无法通过传统的小数据,也就是银行的账本反映出来。

而收集这类信息,却正是大数据的优势。作为一家银行,如果能够通过收集、掌握大数据,了解客户的行为路径,了解客户的决策过程,无疑对提升服务水准、防控金融风险价值很大。所以,只有把小数据方法的完备性、准确性,同大数据分析的多维性、及时性融合起来,才能对管理带来一种质的飞跃。

风险不应自我回避

《巴塞尔协议 III》中,要求银行业在观察客户的违约概率和违约损失率时,数据积累的长度必须长于 5 年或 7 年,甚至更长时间。

在互联网时代,企业和机构对大数据的认识,需要结合正确的方法论、认识论,处理好碎片化的信息和完整性的数据、以及结构性的数据和非结构性的数据的关系。

大数据的优势，在于其可以直接把音频、视频，包括文字非结构化的数据都能数据化，这样的话分析使用起来就会很便捷。但在记录当中，可能掺杂着噪音、埋伏有陷阱。

此外，对于数据清洗还要有严格的流程。巴塞尔委员会之所以做这项规定，就是为了避免因数据的缺陷，而导致在风险识别和计量上出现失误。

但在新晋互联网企业对投资人和客户的宣传中，往往会看到一句话：运用大数据技术。且不论这种对客户群行为数据的保留和采集是否经济、合理和必要，单从数据是否完整上看，就已经把自己得到的数据误以为是全量数据；把自己所拥有的一个样本，认为是具有充分代表性的随机性的样本。

盘点部分“出事”的互联网金融公司，除了一些人为的原因，大多数都存在这种对于数据的片面理解，过分高估了自己的数据处理能力——对自己拥有的这些数据，究竟能不能用，应该怎么样用于风险识别和管控，他们并没有经过反复验证。

但凡了解建模、数据分析和数据挖掘的人，都明白模型越是复杂，纳入的变量越多，就越容易出现这样的问题。这个也证实我们在金融风险的管控当中，必须注意到的模型风险。在看待信用风险、市场风险时，都要借用大量的模型，而模型的质量怎么样，模型是否可靠，实际上最终决定了对信用风险、市场风险、操作风险的识别和计量是否准确。

这就像金融企业面对的客户，客户的个性化、差异化很大，要对他们各自的违约风险和违约损失做出判断，仅靠一些模型的评估可能还不够，有时还需要借助必要的专家判断。

比如审批贷款时，会采用高分段自动进入，低分段自动拒绝、中段分段机器识别以后加以必要的人工干预的方法。这就是为什么有的时候大家经常抱怨银行效率太低，放一个贷款需


要审来审去。首先，高分段经过严格的评估以后，高分段大体占到个人按揭贷款的20%左右，进行自动审核的。而企业贷款，法人贷款，是要经过模型识别通过以后，才能进行第二轮判断。

大数据时代可能更迷茫

大数据的优势，在于其可以直接把音频、视频，包括文字非结构化的数据都能数据化，这样的话分析使用起来就会很便捷。但在记录当中，可能掺杂着噪音、埋伏有陷阱。所以，对于数据信息的不当理解，对于数据分析方法工具的盲目应用，让我们面对茫茫数据时，有可能变得比以往缺乏数据信息的年代更加迷惑。

因此作为数据的使用者，我们应该明确的是，人不能成为机器的奴隶，因为机器和模型都是为我所用的，本身就是人设计的。同时，并非世间万物都是可以数据化的，比如人的情感。

尽管有人说，未来的一切都可以数据化，比如现在，有人已经通过云计算和大数据分析写文章。但我认为，一些数据化较强的分析文章可以写，因为这类文章本身就公式化的。但是，类似《红楼梦》这种充满情感的文字就不太可能，因为通过自动生成的诗词，很难达到较高的艺术水准。

因此，作为互联网时代的现代人，只有学会了怎么样客观地看待数据，怎么审慎地选择方法，才能从这个复杂的社会中提炼出比较有价值的结论。也只有这样，才是真正具备了大数据思维和互联网思维。 

（作者系中国银行业监督管理委员会特邀顾问、中国工商银行原行长）