第一章

引言：为什么资产定价需要机器学习

第一节　资产定价的核心问题：为什么不同的资产会有不同的收益

一、什么是资产定价

资产定价是金融学中最重要的主题之一，它试图回答资产价格是如何被决定的这一重要问题。为了让大家更好地理解资产定价要解决的问题，我们以一个非常简单的股票资产投资问题为例来解释这个问题。

假设现在是2021年4月30日，你手上有100万元，投资期限为1个月，你希望投资在中国股票市场从而获取投资回报。目前进入你视野的有两只股票，分别是贵州茅台（600519.SH）和长江电力（600900.SH），你的任务是在这两只股票中选择一只来进行投资，投资目标是在1个月后卖出股票，使得你的投资回报最大（为了简化问题，先不考虑风险因素），你需要如何科学地进行投资决策呢？

二、CAPM

首先介绍一下大名鼎鼎的资本资产定价模型（capital asset pricing model，CAPM）。CAPM是由美国学者威廉·夏普（William Sharpe）、林特尔（John Lintner）、特里诺（Jack Treynor）和莫辛（Jan Mossin）等人于1964年在资产组合理论和资本市场理论的基础上发展起来的。为了让大家更容易理解资产定价理论的实际作用，本书不去涉及CAPM的推导和证明过程，直接给出CAPM，资产的预期超额收益由式（1-1）决定：

$$E[R_i]-R_f=\beta_i(E[R_M]-R_f) \tag{1-1}$$

$$\beta_i=\mathrm{Cov}(R_i,R_M)/\mathrm{var}(R_M) \tag{1-2}$$

其中，$E[\cdot]$为期望符号，$E[R_i]$和$E[R_M]$分别为资产i和市场组合的预期收益率；R_f为无风险资产收益率。β_i（贝塔值）刻画了不同资产i对于市场组合的风险暴露程度，即描绘了不同资产i收益

率对市场收益率变化的敏感程度。从 CAPM 中我们知道，不同资产的超额收益率只由市场组合的预期收益率和不同资产对市场风险的暴露程度决定。

让我们回到上面的问题，如果我们相信 CAPM 是对的，现在要做的事情只有两件：①估计未来一个月的市场收益率是多少；②分别获取贵州茅台和长江电力的贝塔值。表 1-1 中给出了实际中两只股票和市场的真实数据，接下来就是我们做投资决策的时候了。

表 1-1　现实生活中的数据

股票名称	贝塔值	市场预期收益率/%	个股预期收益率/%	5 月市场实际收益率/%	5 月个股实际收益率/%
2021 年 4 月 30 日投资情况一					
贵州茅台	1.03	0.64	0.66	4.89	10.63
长江电力	0.22		0.14	4.89	−0.57
2021 年 4 月 30 日投资情况二					
贵州茅台	1.03	**−0.64**	−0.66	4.89	10.63
长江电力	0.22		−0.14	4.89	−0.57
股票名称	**贝塔值**	**市场预期收益率/%**	**个股预期收益率/%**	**6 月市场实际收益率/%**	**6 月个股实际收益率/%**
2021 年 5 月 31 日投资情况一					
贵州茅台	1.07	0.64	0.68	−0.67	−6.46
长江电力	0.24		0.15	−0.67	4.07
2021 年 5 月 31 日投资情况二					
贵州茅台	1.07	**−0.64**	−0.68	−0.67	−6.46
长江电力	0.24		−0.15	−0.67	4.07

注：数据为作者基于 Wind 数据整理得到。

从表 1-1 中能够看到，以 2021 年 4 月 30 日的投资决策为例，贵州茅台和长江电力的贝塔值分别为 1.03 和 0.22。长期平均而言，中国股票市场投资组合收益率为 0.64%，如果预期市场在未来一个月的预期收益率和历史平均值相同，下个月会涨 0.64%（情况一），根据 CAPM，这种情况下我们对两只股票的未来预期收益率就是 0.66%和 0.14%，就应该买入贵州茅台这只股票。实际情况下，

2021 年 5 月市场实际收益率为 4.89%，贵州茅台和长江电力的个股实际收益率分别为 10.63%和－0.57%，因此 4 月底按照 CAPM 买入贵州茅台这只股票就是正确的。

如果时间来到 2021 年 5 月 31 日，我们发现两只股票的贝塔值发生了细微的变化，分别为 1.07 和 0.24。你又要面临相同的投资选择，如果你认为 5 月份市场涨了很多，预计 6 月份市场会有回调，预期市场收益率为－0.64%。根据 CAPM，这种情况下我们对两只股票的未来预期收益率就是－0.68%和－0.15%，就应该买入长江电力这只股票。实际情况下，2021 年 6 月市场实际收益率为－0.67%，贵州茅台和长江电力的个股实际收益率分别为－6.46%和 4.07%，5 月底按照模型买入长江电力这只股票就是正确的。

从上面的案例能够清楚地看到资产定价理论模型对于实际投资的重要指导意义。股票的真实收益率由贝塔值[①]和市场预期收益率两者共同决定，如果预期市场会上涨，则应该买入贝塔值高的股票；如果预期市场会下跌，则应该买入贝塔值低的股票。而从两只股票的真实收益率也能看到，其确实与市场真实收益率和其贝塔值相关。

在上面的案例中，聪明的读者可能会注意到，就算用股票的贝塔值乘以市场实际收益率，也无法完全等于股票的真实收益率。这种区别有两种可能：第一种是 CAPM 不足以完全描绘真实的世界，真实的股票收益率可能还受到其他因素的影响，CAPM 并不完整；第二种是上面的案例只看了一期，而且股票数量不够多，从统计上而言，一期两只股票的预测偏差不能说明任何问题。

---

① 另外一个有意思的问题是，CAPM 描述了承担的风险和收益之间关系，模型告诉我们，如果相信市场收益率平均而言是正的话，就应该买入贝塔值高的股票，这样就能够获得更高的收益。承担更高的风险，从而获得更高的收益，好像并没有什么问题。但是从各种实证结果上，我们都会发现“低贝塔异象之谜”：买入贝塔值高的资产组合，未来的预期收益率会更低，反而买入贝塔值低的资产组合，未来的预测收益率会更高。

三、Fama-French 三因子模型

对于上述问题，在金融学领域早就有聪明的学者也发现了。Fama 和 French (1993)在基于美国股票市场收益率的研究中发现，股票的贝塔值不能完全解释不同股票回报率的差异，并进一步在 CAPM 中加入市值因子(small-minus-big，SMB)、账面市值比因子(high-minus-low，HML)来提高模型的解释力。资产定价模型从 CAPM 进一步变成了式(1-3)中的 Fama-French 三因子模型(Fama-French 3-factor model，FF3)。对比两个模型不难发现，Fama-French 三因子模型下，选股票就不仅仅依赖于贝塔值和市场预期收益率两个变量，同样重要的还有资产在市值因子和账面市值比因子的暴露程度。

$$E[R_i]-R_f=\beta_i(E[R_M]-R_f)+\beta_i^{\mathrm{SMB}}(E[R_{\mathrm{SMB}}])+\beta_i^{\mathrm{HML}}(E[R_{\mathrm{HML}}]) \tag{1-3}$$

尽管 Fama-French 三因子模型经常被攻击缺少理论基础，但是不可否认的是，该模型是资产定价论文中不可绕过的一个重要基准模型，论文引用量也在金融学领域高居前列。

四、因子动物园视角

当然，虽然 Fama-French 三因子模型部分程度上弥补了 CAPM 的不足，但是该模型也并不是完美的，很多能够用来预测股票收益率的其他特征并没有被三因子模型覆盖，如股票的动量、盈利、投资等。在此基础上，各种其他的因子模型不断被提出，其中影响力比较大的模型有 Carhart (1997)的动量四因子模型、Novy-Marx (2013)的四因子模型、Fama 和 French (2015)的五因子模型(FF5)、Stambaugh 和 Yuan (2017)的错误定价四因子模型、Hou-Xue-Zhang (2021)的 q 五因子模型。不同的因子模型作者都基于自己对资本市场的理解和严谨的实证结果为人们展示了他们眼中正确的因子模型。

尽管各种各样的因子模型被提出来，但还是有学者源源不断地发现各种股票特征能够显著地预测股票未来的收益率，并且他们发

现的这些特征的信息并没有被上面的因子模型所刻画，学术上把这些不能被因子模型所解释的股票特征称为异象[①](anomalies)。Hou等(2020)整理了美国资产定价文献中452个学术异象因子，Hou等(2021)也在中国股票市场整理出426个异象，Jensen等(2021)复现了93个国家的406个能够预测股票未来收益率的特征。每一个特征背后都是一篇已经发表的学术文献，这些文章都基于实证数据表明自己发现的股票特征能够对股票未来的预期收益率起到预测作用。

Cochrane在2011年美国金融学年会上的主题演讲中提道，目前实证资产定价学术领域的关键问题是：我们目前拥有了几百个因子，然而在这个因子动物园中，到底哪些特征真的给未来的预期收益率提供了独立的信息？哪些特征是冗余的？上面的问题并不仅仅是学术问题，也是实际投资中需要面对的问题。当期资产的价格反映出投资者对资产未来预期收益率的折现，投资者在观测到资产如此多的特征时，他们是依据什么信息进行投资决策的？在这个信息量爆炸的时代，投资者面临这么多影响变量，到底该如何进行正确的投资决策？因为不管投资者有多少信息，最后投资者必须作出的投资决策就只有这个资产到底该买多少这一个问题而已。因此，如何找到合适的工具来处理几百个特征的高维信息，实现信息的去噪提纯，成为资产定价领域的一个重要问题。

第二节　当资产定价遇到机器学习

一、什么是机器学习

人工智能先驱Arthur Samuel在1959年创造了“机器学习”一词，他将机器学习描述为“使计算机在没有明确编程的情况下进行学习”。他编写了一个西洋棋程序。这个程序的神奇之处在于，编

① 关于股票的某个特征是否能够被称为异象，以及异象的实证检验方法，会在后文中详细阐述。

程者自己并不是个下棋高手，于是就通过编程，让西洋棋程序自己跟自己下了上万盘棋。通过尝试哪种布局（棋盘位置）会赢、哪种布局会输，久而久之，西洋棋程序“学习”了什么是好的布局、什么是坏的布局。“学习”后的西洋棋程序下西洋棋的水平超过了Samuel。

可以把机器学习概念与人的学习行为进行类比。例如，你今天出门，发现天上乌云很多，并且天气非常闷热，蜻蜓都飞得很低，你第一次无视了观察到的这些现象，不带伞出门，结果直接被淋了。通过这次教训，你记住了这个现象和结果。第二天出门时你又发现了类似的情况，这次你就学会带着伞出门了。这个案例如果换成机器学习的话，就是用历史数据去标注天气情况特征（例如天空是否有乌云、湿度、蜻蜓飞的高度等），随后标注当天天气是否下雨作为根据标签训练模型，下次出门的时候，只要给模型输入当天天气情况特征，模型就会预测当天是否下雨了。

机器学习算法有多种多样，目前广泛用于我们的日常生活中，如你平常刷的手机App会记录你的兴趣点，并基于机器学习的推荐算法模型给你推送相关广告，去上班时的人脸打卡识别系统背后就基于卷积神经网络（CNN），翻译软件会使用自然语言处理（natural language processing，NLP）模型，在围棋领域战胜人类从而声名鹊起的AlphaGo系统就基于深度学习。

二、机器学习的相关术语

人们会根据自己的历史经验，归纳总结出规律，并在未来遇到新的问题时，用这个规律预测这个问题的答案。机器根据历史数据训练模型，当输入新的数据时，根据模型发现的规律来预测未来。机器学习与人的学习行为非常类似，图1-1具体展示了两类学习行为之间的异同。下面来介绍机器学习中的相关术语和概念。

进行机器学习的必要前置条件是有历史数据，称为“数据集”，在数据集中每一条记录是关于一个对象的描述，称为“样本”，每一

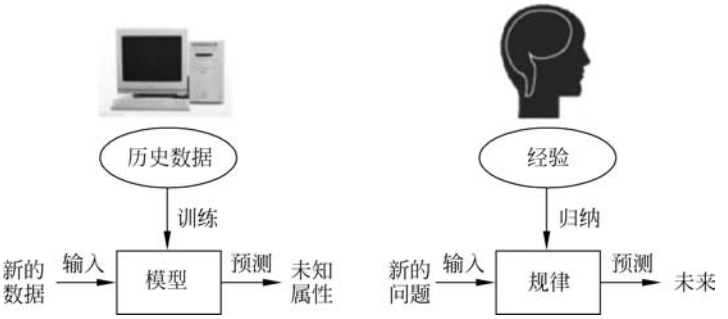

图 1-1　机器学习逻辑与人类学习逻辑示意图

个样本会有很多“特征”来对其进行表现或描述，描述样本的“特征”越多，整个数据的“维度”就越高。从数据集中学得模型的过程称为“训练”，这个过程可以通过执行某种算法来实现。训练过程中使用的数据称为“训练集”，其中每一个样本称为“训练样本”。如果我们的目标是要基于历史数据来进行“预测”，只有样本的特征数据是不够的，还需要获得训练样本结果的信息，称为“标记”。倘若要预测的“标记”是离散值问题，如明天股票是“上涨”还是“下跌”，这类学习任务是“分类”问题；倘若要预测的“标记”是连续值问题，如明天股票的收益率是多少，这个值可以是－20％到＋20％的任意数值，这类学习任务是“回归”问题。当在训练样本中，模型学会了数据集中样本“特征”与“标记”之间的关系之后，就要用训练好的模型在新的数据上进行预测了，这个过程称为“测试”，被预测的样本称为“测试集”。

以表 1-1 这个数据集为例，在 2021 年 4 月 30 日投资情况中，第一条记录就是关于贵州茅台这个“样本”的描述，贝塔值 1.03 这个“特征”描述了贵州茅台在 2021 年 4 月 30 日这天的市场暴露大小为 1.03，贵州茅台在 5 月份个股实际收益率 10.63％就是要预测的“标记”，这显然是一个“回归”问题。与上面“训练集”样本时间段对应的就是 2021 年 5 月 31 日的样本，可以看成“测试集”。

三、机器学习算法用于股票收益率预测问题中的优点

机器学习算法非常适合解决股票收益率预测问题，因为在前面的介绍中我们可以看到，股票收益率预测是一个高维预测问题，而机器学习算法提供了解决高维预测问题的工具。笼统地说，机器学习提供了路径，使得计算机能够从数据中学习到特征。机器学习能够解决高维统计问题，是因为其和传统计量经济学习方法比，有着更好的统计特性，能够更好地应用于样本外的预测问题。

第一，机器学习算法能够优化传统计量经济学中函数形式假定过强的问题。传统计量经济学方法假设因子与股票收益率之间是简单的线性关系，然而现实市场中有些因子和收益率之间的关系并不是严格的线性关系，可能存在非线性的关系。机器学习算法（尤其是神经网络模型）并不需要人为地假设因子与股票收益之间的具体函数形式，而是基于真实的历史数据去拟合两者的关系，这种非参数估计的方法能够更好地用于描述因子与股票收益率之间的关系，从而提高预测的准确性。

第二，机器学习算法能够优化因子过多或因子之间相关系数过高导致估计系数方差过大的问题。在传统计量经济学方法下，当因子数量过多（例如因子数接近或等于样本数）或因子之间相关系数过高时，模型的自由度将会下降，估计系数的方差将会上升。当估计系数方差太大时，基于该估计系数进行的样本外预测结果的方差也会上升，导致预测的结果不佳。在预测问题中的主要目的并不是解释过去，而是在有了新的数据后如何更好地预测未来，因此模型样本外的预测能力才是重要的。为了解决这种样本内预测结果较好、样本外预测结果较差的问题（这种模型的泛化能力较差，在机器学习中称为过拟合问题），机器学习算法引入了惩罚项机制。通过加入惩罚项来压缩变量维度或收缩估计系数方差，可以使模型的样本外预测结果得到改善。这也是机器学习算法在预测问题上强于传统计量经济学方法的原因之一。

第三，机器学习算法能够优化由被解释变量信息含量较低导致

估计系数偏差的问题。现实中，影响股票收益率的因素往往十分复杂，不同类型的股票同一时期的影响因子不一样，同一类型的股票不同时期的影响因子也不一样，这就导致了股票收益率的影响因子很多，噪声更多。传统计量经济学方法无法很好地区分哪些因子是有效因子、哪些因子是噪声。这样会导致传统计量经济学方法获得的估计系数产生偏差。为了应对这种信噪比低的情况，机器学习模型引入特征选择机制，通过剔除噪声，保留更少的有效因子来提升模型的估计准确性。

总结来说，与现有的实证金融和金融计量经济学方法相比，机器学习算法有以下优点：①机器学习算法拥有正则化等优秀的对抗过拟合技术，模型的稳健性能够使其在样本外预测问题上表现优越；②机器学习模型可以直接从高维度的数据中选择最优模型，是完全数据驱动型，而不需要基于经济学假设来预设模型；③机器学习算法可以模拟非常复杂的函数形式，探索数据结构中的各种可能性，而传统计量经济学方法一般都是简单的线性函数。

四、机器学习算法用于股票收益率预测问题面临的挑战

前面已经阐述了机器学习算法的种种好处，但是机器学习算法并不是万能的，机器学习算法的灵活性是把双刃剑，用好了可以非常有效地帮助我们解决问题，但是用得不好将会导致我们得到很多错误的研究结论。因此机器学习算法是否能够在资产定价问题中大展身手是不确定的，特别是在以下方面，金融问题与其他学科问题非常不同，这些方面也是将机器学习算法用于资产定价问题中需要额外注意的。

（一）弱预测变量问题

与其他学科不同，金融领域的预测问题面临的最大挑战之一就是弱预测变量问题，也称为低信噪比问题。这是由于影响股票价格的因素非常多，不同的因素被不同的市场参与者解读的方式也不同，这就导致任意单一因素都会对市场产生一定的影响，但是又无

法产生特别大的影响。股票市场的性质决定了所有能够预测资本市场收益率的特征变量都只是弱预测性。其他学科的预测问题样本外准确性(以回归问题的 R^2 为例)往往可以达到 90%以上,而在金融问题上,预测模型的样本外 R^2 能到 1%就已经非常好了。

尽管在金融学问题中,资产价格的特性就导致其可预测性非常低,但是任何对于模型样本外的预测准确性的提升都能非常显著地转化成巨大的经济意义。以 Campbell 和 Thomson(2008)的研究[①]为例,他们估算,在月度收益率预测问题上,如果样本外预测 R^2 预测准确性能上升 0.5%,对于一个典型的风险厌恶的均值方差投资者而言,其投资组合的超额收益率能够上升 40%。所以尽管资产定价的预测问题面临巨大的挑战,但是其巨大的经济回报还是深深吸引着很多学者和从业人员。

(二)预测变量的不稳定性问题

在其他机器学习问题中,预测特征变量一旦找到,往往会具有比较好的持续性,变量的预测能力一般比较稳定,如在天气预测问题中,气温、气压和水汽这些自然特征都是非常持续稳定的预测变量,当这些变量出现某些特征时,往往就会带来降雨,这是一种自然科学的规律。金融学作为社会科学,其经济规律是由人们的行为来共同决定的,而自然人行为的不稳定,往往就给我们的预测变量带来很强的不稳定性。在金融学领域,同样的预测变量,通常不能在较长时间内持续稳定地发挥预测作用,不同时期对股票收益率有预测能力的变量也会随着时间而发生变化。

导致金融领域预测变量不稳定的原因有以下三点。

1. 金融市场的自我修复机制

从逻辑上来说,资产的收益率取决于资产目前的价格和投资者未来对该资产未来回报的预期。如果目前公开市场上有某个特征能够持续稳定地预测未来资产的收益率,那么市场上只要有人

① 这个估算并没有考虑实际的交易成本和交易滑点等问题。

群发现这个特征，他们就会持续投入资金在该类特征的资产上以获取未来的稳定超额回报。但是当过多的资金涌入某类特征的资产之后，短期必然会对这类资金价格产生一个快速的拉升，当期交易该特征资产的人增多，当期资产价格就会逐渐高于其理性水平，未来资产价格的预期收益率可能会下降，后面再按照该特征买入股票的投资者不一定能够获得盈利。因此，理论上任何因子在被公开后，随着交易该因子的人数增加，必然会出现“因子拥挤”的现象，导致因子逐渐失效。McLean 和 Pontiff（2016）的实证研究支持了上述逻辑，研究发现，美国的异象性因子在公开发表之后，其对股票未来收益率的预测能力会随着交易该因子的投资者增加而逐渐降低。

2. 金融市场受到其经济状态的影响，市场状态的改变可能引起预测变量的失效

以美国资本市场非常有效的动量因子[①]为例，股票的动量特征在很多时候都能正向地预测股票下一期收益率，但是 Daniel 和 Moskowitz（2016）发现，当市场处于恐慌时期，尤其是伴随着波动率高的市场衰退和反弹时，动量因子的预测能力就会失效，甚至反向预测股票未来预期收益率。此外，Rapach 等（2010）发现股票市场时间序列的可预测性与经济周期有关，当市场处于衰退时期，预测变量对于市场收益率的预测能力会上升；而当市场处于繁荣时期，预测变量对于市场收益率的预测能力则会下降。

3. 金融市场受到其经济环境和制度的影响，市场环境或者规则制度的改变可能引起预测变量的失效

金融市场的规律是由其市场环境和制度决定的，不同的投资者结构和市场制度的变化都会引起预测关系的变化。例如 Chu 等

① 股票的动量特征能够用来预测股票的未来收益率由 Jegadeesh 和 Titman（1993）发现，动量是指过去一段时间收益率较高的资产在未来获得的收益率仍会高于过去收益率较低的资产。

(2020)基于美国证监会做空制度改革的准自然随机试验(SHO)[①]研究发现,随着美国股票市场做空制度限制的放松,美国金融市场上 11 个异象性因子的预测能力显著下降。

(三)数据挖掘带来的过拟合问题

过拟合问题是在使用机器学习算法时特别需要注意的问题,关于这个问题的原因和解决方案,会在第三章详细阐述。但是这里还是要强调在金融领域中需要额外注意这个问题的原因。简单来说,有时候自己的模型在训练集上用训练数据拟合得非常好,但在样本外的测试集中却表现得很糟糕。这是因为机器学习算法允许复杂的、非线性的模型很好地拟合数据,但它们也有过拟合数据的风险,让模型错误地记住了一些噪声带来的特征。对抗过拟合的最好方式就是扩大数据样本,但是由于金融市场的数据量有限,特别是在月度股票收益率的预测问题中,样本量受到时间的限制,模型的结果在样本外的表现很难保证。所有的分析都是基于历史数据的回测结果(back-testing),当我们站在今天构建模型的时候,这件事本身就已经用到了未来信息,所以很多时候过拟合问题是不可避免的,只能尽可能地提升模型的稳健性。

第三节　相关学术文献介绍

机器学习模型在降维、惩罚项和泛函数等技术上的突破在解决

---

① 美国证监会实施的 SHO 试点项目是基于随机对照试验思想。其基本思想是通过随机将股票分为两个部分,一部分股票取消规则(这部分称为试点组或实验组),另一部分股票维持规则现状(这部分称为非试点组或对照组)。通过比较实验组和对照组股票在取消规则卖空交易的变化,从而获得"提价交易规则"对卖空交易的影响。2005 年 5 月,美国 SHO 试点项目正式实施,试点项目在罗素 3 000 成分股中随机挑选了 1 000 只股票作为实验组,取消了这些股票卖空约束中的"提价交易规则",使得这些股票能够在任何价格变化时随时被卖空,剩余 2 000 只股票作为控制组,继续维持股票卖空约束中的"提价交易规则"。

以上前两个问题上具有天然的优越性。由于以上方面的优势，机器学习技术已经成为金融领域中的应用前沿之一，特别是在预测金融市场运动、处理文本信息、改进交易策略方面（苏治 等，2017），很多论文探索了不同类型的机器学习算法在股票收益率预测的效果。其模型具体分为以下三类。

第一类是金融学中较为常用的降维类模型，这类模型的优点是能将高维度数据压缩成低维，同时还能保留较多的信息。例如：Rapach 和 Zhou(2018)，Maio 和 Philip(2015)基于主成分分析(principal component analysis，PCA)的方法使用美国宏观变量来预测股票市场未来收益率；Kelly 和 Pruitt(2015)基于最小偏二乘模型使用风格因子(style factors)收益率资产组合来预测股票市场。

第二类是带惩罚项的线性模型，其优点是通过加入惩罚项，降低噪声信息的因子荷载，从而提高预测效果。例如 Chinco 等(2018)基于套索回归(LASSO)分析了一分钟频率的个股收益率预测。

第三类是非线性模型，这类模型的优点在于能够基于历史数据信息拟合预测变量与收益率之间的非线性结构。例如有学者基于随机森林(random forest，RF)、模糊神经网络和长短期记忆神经网络模型等人工智能算法，检验了技术和宏观预测因子在日度股票价格收益率预测的效果外 R^2(Fischer et al.，2018；Sirignano et al.，2018；Bao et al.，2017；Butaru et al.，2016)。Gu 等(2020a；2020b)探索了神经网络模型、自编码机等深度学习模型在个股月度收益率的效果，获得了非常好的样本外预测准确率。

中国股票市场依然处于不断发展和完善的阶段，很多国内学者也尝试结合机器学习技术解释中国股票市场的预期收益率预测问题。姜富伟等(2011)研究了中国市场投资组合和根据公司行业、规模、面值市值比和股权集中度等划分的各种成分投资组合的股票收益的可预测性；陈卫华和徐国祥(2018)发现深度学习预测沪深 300 指数的效果明显好于传统计量经济学模型；李斌等(2017，2019)分别采用了支持向量机、神经网络、Adaboost 等机器学习算法，利用 19 项技术指标预测股价方向，发现基于机器学习算法预测所构建的

投资组合确实能取得更好的投资收益。

第四节　相关业界应用场景

目前人工智能技术已经广泛地应用于金融行业中，在很多业务场景下帮助金融从业人员提升工作效率。

一、量化投资

人工智能在量化投资领域，通过模型建立、数据的输入与处理学习进行预测、选股、择时，不仅可以通过模拟人类的思考模式去捕捉市场信息，更可以挖掘出潜在的信息与模式，更加有效地提供投资决策，强大的学习能力能够不断地积累经验，根据实际市场的反馈信息、市场风格的变化去及时地、自适应地修正调整模型，作出当下最为契合的投资组合，效率更高且避免了人为因素的干扰，最大限度地做到风险和预期收益的可测、可控。在量化投资中，信号发现、信号增强、投资组合优化、交易执行优化、风险管理等几个方面的机器学习、深度学习模型都大有可为。

海内外很多资产管理机构都会借助人工智能算法来进行量化投资，如 D. E. Shaw、Two Sigma、Citadel 等公司。这些公司都十分强调机器学习和分布式运算的运用，而基金大部分员工可能都没有金融背景，直接从理工院校的计算机科学、数学和工程专业毕业生中选聘。国内很多量化规模排在前列的头部量化私募也是如此，如幻方量化对人工智能软硬件研发累计投入近 2 亿元，建立 AI 实验室，并将每年总体营收的大部分投入人工智能领域[①]。

2017 年 10 月 18 日，EquBot LLC、ETF Managers Group 共同推出了全球第一只应用人工智能、机器学习进行投资的 ETF（交易所交易基金）：AI Powered Equity ETF（AIEQ. US）。其投资策略是基于 EquBot 公司开发的量化模型生成，由 IBM Watson 超级计

① 数据来源于官网 https://www.high-flyer.cn/ai_lab.html。

算机提供技术支持，通过人工智能算法进行量化择时、量化选股、因子分析、事件驱动分析决策。图 1-2 显示了 2017 年 10 月到 2021 年 6 月 30 日，该产品和标普 500 指数收益率的表现情况。在近 4 年的时间中，标普 500 资产累计收益率约 72%，AIEQ 产品收益率约 62%，在美国高度有效的资本市场下，AI 管理的资产投资组合与市场投资业绩表现相近。

图 1-2　AIEQ 产品和标普 500 收益表现

资料来源：AIEQ 官网。

二、智能投顾

智能投顾(robo advisory)是指计算机依据投资理论搭建量化交易决策模型，再将投资者风险偏好、财务状况及理财规划等变量输入模型，为用户生成自动化、智能化、个性化的资产配置建议，并可以自动执行交易以及资产再平衡。智能投顾相比传统投顾的核心区别在于对 AI、大数据等技术的应用，通过技术应用极大降低投顾服务门槛，有助于挖掘长尾市场。智能投顾的概念产生于美国。得益于美国市场量化投资和 ETF 的蓬勃发展，自 2008 年起，Betterment、Wealthfront、Future Advisor 等第一批智能投顾公司相继成立，在智能投顾市场深耕细作、稳健增长。根据 Statista 在 2019 年 2 月发布的美国智能投顾市场报告，美国智能投顾管理的资产在

2021 年预计达到 2.2 万亿美元，占全球资管行业的 2.2%。中国智能投顾市场也在高速崛起，2019 年 10 月 24 日，中国证券监督管理委员会（以下简称“证监会”）发布《关于做好公开募集证券投资基金投资顾问业务试点工作的通知》，开启了基金投顾正规化发展时代。截止到 2021 年 7 月 2 日，已有 50 家机构先后获得了基金投顾业务试点资格，包括公募基金 21 家、第三方独立销售机构 3 家、银行 3 家、证券公司 23 家。

智能投顾产品的关键在于数据、模型和算法，数据是基础，模型决定配置比例，算法决定投资方法。图 1-3 为智能投顾服务流程图，首先利用大数据对用户进行画像分析，得到相关的风险偏好信息，再利用人工智能技术依据客户的风险偏好构建合适的投资组合，并根据市场反馈结果进行动态的再平衡。在整个智能投顾流程中，很多业务场景都需要借助 AI 算法的力量来帮助提升服务效率。例如在大类资产配置模型领域，资产协方差矩阵是开展风险管理的基础，而传统方法估计的资产协方差矩阵通常只能基于历史序列的线性信息展开估计，导致评估结果与真实情况发生偏差。针对传统资产配置方法的不足，可以使用生成条件对抗网络（cGAN）模型从贝叶斯学派视角提升对资产风险的估计效率，从而改善大类资产配置的结果。

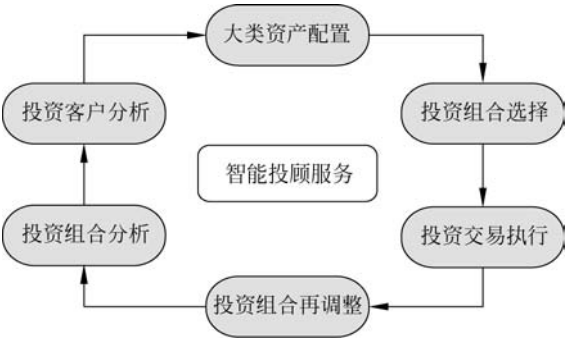

图 1-3　智能投顾服务流程图

参考文献

陈卫华，徐国祥，2018. 基于深度学习和股票论坛数据的股市波动率预测精度研究[J]. 管理世界(1)：180-181.

姜富伟，涂俊，RAPACH D E，等，2011. 中国股票市场可预测性的实证研究[J]. 金融研究(9)：107-121.

李斌，林彦，唐闻轩，2017. ML-TEA：一套基于机器学习和技术分析的量化投资算法[J]. 系统工程理论与实践(5)：1089-1100.

苏治，卢曼，李德轩，2017. 深度学习的金融实证应用：动态、贡献与展望[J]. 金融研究(5)：111-126.

ANG A，HODRICK R J，XING Y，et al.，2006. The cross-section of volatility and expected returns[J]. The journal of finance，61：259-299.

BAO W，YUE J，RAO Y，2017. A deep learning framework for financial time series using stacked autoencoders and long-short term memory[J]. Plos one，12：18-24.

BUTARU F，CHEN Q，CLARK B，et al.，2016. Risk and risk management in the credit card industry[J]. Journal of banking & finance，72：218-239.

CARHART M M，1997. On persistence in mutual fund performance[J]. The journal of finance，52：57-82.

CHINCO A，CLARK JOSEPH A D，YE M，2018. Sparse signals in the cross-section of returns[J]. The journal of finance，74(1)：449-492.

CHU Y，HIRSHLEIFER D，MA L，2020. The causal effect of limits to arbitrage on asset pricing anomalies[J]. The journal of finance，75：2631-2672.

DANIEL K，MOSKOWITZ T J，2016. Momentum crashes[J]. Journal of financial economics，122：221-247.

DANIEL K，MOTA L，ROTTKE S，et al.，2020. The cross-section of risk and returns[J]. The review of financial studies，33：1927-1979.

FAMA E F，FRENCH K R，1993. Common risk factors in the returns on stocks and bonds[J]. Journal of financial economics，33：3-56.

FAMA E F，FRENCH K R，2015. A five-factor asset pricing model[J]. Journal of financial economics，116：1-22.

FAMA E F，FRENCH K R，2020. Comparing cross-section and time-series factor models[J]. The review of financial studies，33：1891-1926.

FISCHER T，KRAUSS C，2018. Deep learning with long short-term memory

networks for financial market predictions [J]. European journal of operational research,270: 654-669.

HARVEY C R, LIU Y, ZHU H, 2015. ... and the cross-section of expected returns[J]. Review of financial studies,29: 5-68.

HOU K, MO H, XUE C, et al., 2021. An augmented q-factor model with expected growth[J]. Review of finance,25: 1-41.

HOU K, QIAO F, ZHANG X, 2021. Finding anomalies in China[R]. Working Paper.

HOU K, XUE C, ZHANG L, 2020. Replicating anomalies[J]. The review of financial studies,33: 2019-2133.

JENSEN T I, KELLY B T, PEDERSEN L H, 2021. Is there a replication crisis in finance[R]. Working Paper.

KAROLYI G A, VAN NIEUWERBURGH S, 2020. New methods for the cross-section of returns[J]. The review of financial studies,33: 1879-1890.

KELLY B, PRUITT S, 2015. The three-pass regression filter: a new approach to forecasting using many predictors[J]. Journal of econometrics,186: 294-316.

LETTAU M, PELGER M, 2020. Factors that fit the time series and cross-section of stock returns[J]. The review of financial studies,33: 2274-2325.

LEWELLEN J, 2015. The cross-section of expected stock returns[J]. Critical finance review,4: 1-44.

LINNAINMAA J T, ROBERTS M R, 2018. The history of the cross-section of stock returns[J]. The review of financial studies,31: 2606-2649.

MAIO P, PHILIP D, 2015. Macro variables and the components of stock returns [J]. Journal of empirical finance,33: 287-308.

MCLEAN R D, PONTIFF J, 2016. Does academic research destroy stock return predictability? [J]. The journal of finance,71: 5-32.

NOVY-MARX R, 2013. The other side of value: the gross profitability premium[J]. Journal of financial economics,108: 1-28.

RAPACH D E, STRAUSS J K, ZHOU G, 2010. Out-of-sample equity premium prediction: combination forecasts and links to the real economy[J]. Review of financial studies,23: 821-862.

RAPACH D, ZHOU G, 2018. Sparse macro factors[R]. Working Paper.

SPIEGEL M, 2008. Forecasting the equity premium: where we stand today[J]. Review of financial studies,21: 1453-1454.

STAMBAUGH R F, YUAN Y, 2017. Mispricing factors[J]. The review of financial studies,30: 1270-1315.