# **机器学习与资产定价短期课程：专业报告与教学建议**

## **1. 引言：使课程目标与现代资产定价研究保持一致**

### **1.1. 课程相关性与结构的初步评估**

拟议中的“机器学习与资产定价”短期课程，在当前金融领域机器学习日益关键的背景下，具有高度的相关性 1。该课程专注于为金融专业高年级本科生提供实践应用，满足了市场对实践技能的关键需求。其课程结构——三次为期2.5小时的课程，每次均附有相应作业——设计简洁。然而，如何在有限的时间内平衡广泛的主题（涵盖多种资产类别和机器学习技术）与确保学生有足够深度的学习，将是一个挑战。课程强调以Python代码实现为核心，并且每节课都围绕一篇权威的机器学习资产定价论文进行复现，这是一种卓越的教学方法，与该领域的研究和实践应用方式高度一致 6。这种安排也符合课程“侧重实践案例”和“提供直观体验”的目标。

### **1.2. 与当代金融教育趋势的联系**

从纯理论金融向经验金融和计算金融的转变是金融教育领域的一个重要趋势。本课程通过将“Python代码实现为核心”和“实时操作”作为重点，积极响应了这一趋势。此外，“论文复现”部分直接满足了学生参与前沿研究的需求，正如大量涌现的探讨机器学习在资产定价中应用的工作论文和学术出版物所强调的那样 1。

### **1.3. 初步考量与建议**

该课程以实践为导向，以Python为中心，并结合论文复现的教学方法值得高度肯定，这与当前量化金融教育的趋势完全吻合。然而，课程大纲中列出的“研究对象”范围——包括A股股票、股指、大宗商品、外汇和碳价格——对于一个仅有三次课程的短期项目而言，可能过于宽泛，特别是如果计划对每个资产类别都进行深入的论文复现。

具体来说，课程目标是在实践层面覆盖五种截然不同的资产类别。考虑到每节课仅有2.5小时，并且需要完成一篇论文的复现工作，这意味着在2.5小时内有效涵盖机器学习应用、Python代码实现以及对一篇学术论文的深入理解和复现，本身就是一项非常密集的任务。如果试图在三次课程中，通过每节课处理1-2种资产类别来覆盖所有五种资产，可能会导致对每个主题的处理都流于表面，难以达到预期的学习深度。

因此，建议对课程内容进行聚焦。一种方案是选择少数几种资产类别进行更深入的探讨，确保学生能够透彻理解相关模型和实践方法。另一种方案是将课堂上的论文复现作为示范性案例，学生在课后作业中，将学到的方法应用于其他资产类别。这样的调整，既能保证学习的深度，也符合课程“减少理论讲解，提供直观体验”的初衷。

## **2. 为资产定价应用筛选核心机器学习概念**

### **2.1. 核心机器学习算法及其相关性**

本课程应涵盖一系列在资产定价文献中广泛应用的机器学习模型 1，具体包括：

* **正则化方法 (例如 LASSO, Ridge, Elastic Net):** 这些方法对于处理资产定价中常见的“因子动物园” (factor zoo) 和高维预测变量空间至关重要 1。它们通过在损失函数中加入惩罚项来选择重要特征并防止模型过拟合。
* **降维方法 (例如 PCA, 自动编码器 Autoencoders):** 这些技术有助于从大量的股票特征中提取潜在的风险因子或信息 1。例如，Gu et al. (2021) 和 Uddin and Yu (2020) 的研究展示了自动编码器在学习潜在资产定价因子方面的应用 5。
* **基于树的方法 (决策树, 随机森林 Random Forest, 梯度提升回归树 GBRT):** 这些非参数方法能够有效地捕捉预测变量之间的非线性关系和交互效应，而无需预设严格的参数形式 1。随机森林和梯度提升树在许多研究中表现出优越的性能 12。
* **神经网络 (NNs, 包括深度学习, LSTMs, Transformers):** 神经网络，特别是深度学习模型，能够对复杂的非线性模式和序列数据进行建模 1。长短期记忆网络 (LSTM) 由于其处理序列依赖性的能力，尤其适用于股票价格等时间序列数据 19。

课程还应简要介绍监督学习、无监督学习和强化学习这三大类机器学习范式。监督学习在收益率预测中占主导地位 5。无监督学习（如PCA、聚类）则有助于特征工程和理解数据结构 2。强化学习作为一种新兴方法，在动态投资组合优化领域展现出潜力 2。

在资产定价的实证研究中，研究者们面临着一个被称为“因子动物园”的现象 1，即存在数百个甚至更多的潜在预测因子。传统计量经济模型在处理如此高维的数据时，往往会遇到过拟合和多重共线性等问题，导致模型在样本外表现不佳。机器学习方法，特别是正则化（如LASSO）和降维（如PCA）技术，正是为了应对这类高维挑战而设计的。它们能够从大量特征中筛选出最重要的信息，或者将高维数据压缩到更易于管理的低维空间，从而提高模型的泛化能力和经济解释性。因此，在课程中介绍这些方法时，不应仅仅将其视为一种新的建模工具，更应强调它们是解决实证资产定价中基本问题的有效手段。

### **2.2. 关注Transformer等先进模型**

近期，以Transformer为代表的先进模型在资产定价领域引起了广泛关注。例如，Kelly等人（2025年）在NBER工作论文中提出的“人工智能资产定价模型”（AIPM）便是一个典型代表 15。这类模型将Transformer网络嵌入到随机折现因子（SDF）中，通过跨资产信息共享和非线性建模，显著降低了定价误差。Transformer模型的核心优势在于其“注意力机制”，它使得模型能够灵活地在不同资产之间共享条件信息，从而生成具有“上下文感知”能力的预测 15。这与传统金融机器学习模型主要关注“单一资产预测”（即资产 i 的预测仅依赖于资产 i 自身的变量）形成了鲜明对比。这种跨资产视角意味着，一个资产的最优投资组合权重可能取决于其他资产的特征，模型能够学习和利用资产属性之间的条件相似性。

这种转变预示着未来的资产定价模型将更加关注资产间的网络效应和相互依赖关系。虽然本课程主要面向本科生，侧重于基础机器学习工具，但简要介绍Transformer这类前沿模型，特别是其核心思想（如跨资产信息共享），将极大地开阔学生的视野，使其了解该领域未来的研究方向。即使无法在课程中完整复现复杂的Transformer模型，讨论其概念和潜力也是非常有价值的。

### **2.3. 比较分析方法**

资产定价的机器学习文献中，经常包含对不同机器学习方法性能的“比较分析” 1。课程应引导学生理解，没有任何一种机器学习模型是普适最优的；模型的表现取决于具体问题、数据特性和分析背景。课程计划中每节课复现一篇论文，这自然会引导学生将所复现论文中选择的模型与简单的基准模型或其他模型进行比较，从而培养其批判性思维和模型选择能力。

### **2.4. 资产定价中关键机器学习模型概览表**

为了帮助学生构建清晰的知识体系，建议提供一个关键机器学习模型的概览表。

| **模型类别** | **具体算法示例** | **在资产定价中的常见应用** | **主要优势** | **潜在劣势/挑战** | **相关研究参考** |
| --- | --- | --- | --- | --- | --- |
| 正则化方法 | LASSO, Ridge, Elastic Net | 因子选择、高维数据下的收益率预测、防止过拟合 | 有效处理多重共线性问题，提高模型稀疏性和可解释性 | 可能错误地剔除重要变量，对超参数敏感 | 1 |
| 基于树的方法 | 决策树, 随机森林, 梯度提升回归树 (GBRT) | 收益率预测、波动率预测、捕捉非线性关系 | 能处理非线性关系和特征交互，对异常值不敏感，随机森林和GBRT通常性能较好 | 单个决策树易过拟合，复杂模型可能难以解释 | 12 |
| 神经网络 | 多层感知机 (MLP), LSTM, Transformer | 复杂非线性关系建模、时间序列预测 (LSTM)、跨资产信息建模 (Transformer) | 强大的非线性拟合能力，能自动学习特征表示 | 计算量大，易过拟合，"黑箱"特性导致可解释性差，对数据量和质量要求高 | 6 |
| 降维方法 | 主成分分析 (PCA), 自动编码器 (Autoencoder) | 从大量特征中提取潜在因子、数据可视化、降低模型复杂度 | 减少数据维度，去除冗余信息，提高计算效率 | 可能损失部分信息，主成分的经济意义可能不明确，自动编码器训练可能复杂 | 1 |
| 比较分析研究 | - | 评估不同ML模型在特定资产定价问题上的表现 | 为模型选择提供实证依据，推动方法论发展 | 结果可能依赖于特定的数据集和评估标准，难以一概而论 | 1 |

**价值主张：** 该表格为学生提供了一个关于在资产定价中应使用哪些工具以及用于哪些任务的简明比较摘要。将模型与应用、优势、劣势以及基础研究联系起来，提供了一个全面且易于理解的学习辅助工具，直接支持课程的教学目标。

## **3. 深入探讨：特定资产类别的考量与研究前沿**

### **3.1. A股股票（中国市场）**

* **机器学习应用:** 研究表明，机器学习方法可用于对A股投资组合的超额收益进行建模 11。尽管并非所有案例研究都专门针对A股，但机器学习在股票价格预测方面的原理是通用的 23。
* **数据来源:** Alpha Vantage 声称支持上海证券交易所 (SSE) 和深圳证券交易所 (SZSE) 的数据 26。EODHD 覆盖美国及非美国交易所，可能包括A股市场 28。yfinance 在获取某些中国ETF数据时曾报告过问题 33，但一些专注于中国市场的库，如 AKShare 和 Tushare，可能是更可靠的数据源 34。值得注意的是，一个在韩国市场复现Gu, Kelly, Xiu (2020) 的GitHub项目 (Asset-Pricing-via-ML 9) 指出，与美国市场相比，其数据可得性存在显著局限，这一挑战可能同样适用于A股研究。
* **独特性考量:** A股市场具有一些独特性质，例如散户投资者参与度较高、政策影响显著以及过往的资本管制等，这些因素可能会影响模型的表现和因子的定义。“因子动物园”问题 1 以及数据质量和可得性问题 4 在A股市场可能更为突出。
* **实践启示:** 在非美国市场（如韩国市场 9）进行全面历史数据采购所面临的挑战，直接预示着针对A股的实践练习需要仔细审查数据，并可能需要将研究重点放在数据更易通过国际API获取的近期数据或特定大盘股上。例如，韩国市场的复现研究指出，“Gu论文中使用的近90%的数据都无法获得”，这导致了幸存者偏差和结果稳健性较差的问题。这强烈暗示在尝试复现全面的因子研究时，A股市场也可能面临类似的数据挑战。因此，对于A股的案例研究，课程可能需要简化范围（例如，减少因子数量，专注于特定指数如上证50，或者如果可行，使用预先清洗过的数据集），或者专门安排教学时间讨论该市场特有的数据获取和清洗难题。

### **3.2. 股票指数 (例如 标普500指数)**

* **机器学习应用:** 决策树、随机森林、梯度提升树、支持向量机和神经网络等机器学习模型已被用于标普500成分股的长期预测 12。人工智能增强的因子分析也被用于预测标普500指数的动态 13。此外，随机森林、LSTM和逻辑回归等多种模型在标普500指数预测任务中进行了比较 17。博科尼大学学生投资俱乐部也曾使用随机森林、XGBoost、逻辑回归和SVM进行标普500指数的交易策略研究 22。
* **数据来源:** 标普500指数的历史数据可通过多种API广泛获取，如Alpha Vantage 27、EODHD 28，也可直接从纳斯达克等数据源获取 37。
* **关键发现:** 研究表明，梯度提升树和随机森林在预测标普500成分股方面表现更优，但其预测准确性对经济稳定性较为敏感 12。预测准确性随预测期限的延长而下降。消费者信心指数和净出口等宏观经济因素对预测结果具有显著且随时间变化的影响 12。
* **实践启示:** 标普500机器学习模型性能对“经济稳定性”和“宏观经济因素”的敏感性 12，意味着指数预测模型的特征工程不应局限于价格和交易量数据，而应扩展至纳入宏观经济指标。这直接关联到“市场机制转换”这一挑战。具体而言，既然研究 12 表明标普500机器学习模型的准确性受到经济稳定性和宏观经济变量的影响，而市场机制分析 38 的目的正是识别这些不同的经济阶段，那么，为了实现对标普500指数的稳健预测，机器学习模型理想情况下应包含能够捕捉宏观经济状况的特征，甚至明确地对机制转换行为进行建模。因此，在教授标普500指数预测时，课程应鼓励学生考虑将宏观数据（例如来自美联储 40 或经合组织 40 的数据）作为特征纳入模型，并讨论模型性能在不同市场机制（例如3838 中识别的扩张期、衰退期、动荡期）下可能存在的差异。

### **3.3. 大宗商品**

* **机器学习应用:** 机器学习被用于大宗商品的战略资产配置，其中会考虑均值回归特性 42。利用机器学习预测国际股权风险溢价，可能与依赖大宗商品出口国家的股市表现相关 6。人工智能也被用于大宗商品价格预测，并结合了天气模式和情绪分析等另类数据 43。
* **数据来源:** Commodities API 45、Financial Modeling Prep 46、EODHD 28 和 Alpha Vantage 27 均提供大宗商品数据。
* **独特性考量:** 均值回归是大宗商品价格的一个关键特征 42。套期保值需求对价格有重要影响。PriceVision 是一款专门用于大宗商品预测的人工智能/机器学习工具 43。
* **实践启示:** 人工智能驱动的大宗商品价格预测越来越多地使用另类数据（如天气数据、卫星图像、新闻情绪）43，这突出表明了向更全面、数据更丰富的建模方式发展的趋势。这意味着课程可以设计一个大宗商品的案例研究，演示如何整合至少一种类型的另类数据（例如，如果可行，通过自然语言处理技术分析新闻头条来获取新闻情绪；或者如果API可提供，则使用预处理的情绪评分），从而让学生接触到这种先进的技术。具体而言，传统的大宗商品预测依赖于历史价格和供需基本面。人工智能和机器学习使得处理多样化的非结构化数据源成为可能 43。例如，利用天气数据预测咖啡作物产量，或利用自然语言处理技术从新闻中分析市场情绪。这表明纯粹基于价格的模型可能会遗漏重要的预测信息。因此，一个实践练习可以包括获取大宗商品价格数据，然后用一个简单的另类数据特征（例如，全球经济情绪的代理指标，或者如果能找到历史数据，特定天气事件指标）来增强数据，观察其是否能改进一个基础的预测模型。这将与课程的“实践导向”特性相符。

### **3.4. 外汇 (Forex)**

* **机器学习应用:** 支持向量机、神经网络、随机森林、循环神经网络和LSTM等模型被用于外汇市场的分析和预测 19。包括趋势跟踪、剥头皮交易和套息交易在内的算法交易策略也日益受到人工智能的驱动 47。
* **数据来源:** Alpha Vantage 26、EODHD 28、Financial Modeling Prep 50 和 Polygon.io 51 均提供外汇数据。
* **关键因素:** 利率决策、GDP和通货膨胀等经济指标以及市场情绪对外汇市场有重大影响 47。流动性约束可以解释套息交易的风险溢价 52。
* **实践启示:** 尽管许多机器学习模型被应用于外汇市场 19，但由于外汇市场的高噪声信号比和效率，实现持续盈利的预测极具挑战性。过拟合是一个主要风险 19。因此，外汇案例研究应特别强调稳健的回测、样本外验证，以及过拟合和数据窥探的风险。重点可能更多地在于利用机器学习理解市场动态，而不是追求不切实际的预测准确性。具体而言，外汇市场是一个高度流动且相对有效的市场。许多参与者使用包括机器学习在内的复杂模型 19，这使得寻找持续的套利机会或预测优势变得困难。机器学习模型，特别是复杂的模型，很容易对外汇市场的噪声数据产生过拟合 19。因此，在教授外汇机器学习时，设定切合实际的期望至关重要。论文复现可能侧重于探索市场动态的研究（例如，新闻情绪的影响 48），或者一个非常具体、短期的市场无效性，而不是一个承诺高回报的通用价格预测模型。教学价值将在于模型构建和严格评估的过程。

### **3.5. 碳价格**

* **机器学习应用:** 混合机器学习方法（如SSA分解结合SVR/LSSVR）被用于预测欧盟碳排放交易体系 (EU ETS) 和中国碳交易试点市场的碳价格 14。其他模型包括随机森林、XGBoost、RNN、LSTM和MLP 14。特征提取和非线性整合是关键技术 55。
* **数据来源:** Trading Economics 提供欧盟碳排放配额 (EUA) 数据和API 56。Puro.earth 提供CORC碳清除价格指数 57。通过常用的Python库（如 yfinance）直接获取碳期货（如EUA）的行情代码似乎比较困难 58。EODHD的商品期货列表 32 和其通用API页面 28 均未明确列出碳价格或EUA。ICE提供EUA期货数据，可能通过Python接入 64。
* **关键因素:** 碳价格受到配额、能源价格、宏观经济状况和政策等多重因素的影响 14，并且呈现非线性和非平稳特性 14。
* **实践启示:** 尽管关于碳价格预测的机器学习学术研究日益增多 14，但通过 yfinance 等常用Python库为学生练习获取易于访问、免费且具有长历史周期的碳价格数据似乎存在局限性。可能需要依赖Trading Economics 56 或专门的数据提供商 3。这为课程提供了一个机会，可以强调如何处理不太常见的数据集或API，或者基于研究中描述的特征（非线性、非平稳）使用模拟数据。具体而言，碳定价是课程列出的研究对象之一。机器学习在此领域的应用已有研究 14。然而，通过 yfinance 等标准库寻找现成的、免费的历史碳价格数据似乎具有挑战性 58。EODHD的直接覆盖范围也不明确 28。专门的数据提供商 56 更有可能是数据来源。因此，对于碳价格模块，数据获取将是一个关键的学习点。如果通过免费/低成本来源直接进行API访问对于课堂环境来说比较困难，教师可能需要提供数据集，或者将练习重点放在使用公开可用（即使不易通过API访问）数据集的论文上。或者，根据研究中描述的已知非线性和非平稳特性（依据14）模拟数据，也可能是练习特定机器学习技术的一种选择。

### **3.6. 资产类别与机器学习考量概览表**

为了帮助学生理解不同资产类别在机器学习应用中的特性和数据需求，下表提供了一个概览。

| **资产类别** | **机器学习的关键特征** | **数据可得性说明及主要API提供商 (附文献索引)** | **示例研究重点/论文类型** | **相关研究参考** |
| --- | --- | --- | --- | --- |
| A股股票 | 散户参与度高，政策敏感性强，可能存在独特因子结构 | Alpha Vantage (SSE, SZSE) 26, EODHD (XSHG, XSHE) 29, AKShare/Tushare (中国市场专用) 34。数据深度和广度可能不及美国市场 9。 | 因子有效性检验，市场情绪分析，事件研究 | 9 |
| 股票指数 (例如 标普500) | 受宏观经济影响大，成分股联动性强，流动性好 | yfinance (例如 ^GSPC), Alpha Vantage, EODHD 29 | 指数收益率预测，波动率建模，宏观因子影响分析 | 12 |
| 大宗商品 | 均值回归特性，受供需、天气、地缘政治影响，另类数据应用潜力大 | Commodities API 45, Financial Modeling Prep 46, EODHD 29, Alpha Vantage 35 | 价格预测 (结合另类数据)，波动性建模，套期保值策略 | 42 |
| 外汇 (Forex) | 高流动性，高频交易普遍，噪声信号比高，受利率和经济数据驱动 | Alpha Vantage 35, EODHD 29, FMP 50, Polygon.io 51 | 短期趋势预测，交易信号生成，情绪分析，套息交易建模 | 19 |
| 碳价格 | 政策驱动型市场，非线性、非平稳特征明显，数据源相对较新 | Trading Economics (EUA) 56, Puro.earth (CORC) 57, ICE (EUA期货) 64。通用库数据可能有限。 | 价格预测 (混合模型)，政策冲击分析，与其他能源品联动性研究 | 14 |

**价值主张：** 该表格为学生提供了一个关于不同资产类别特性、数据来源以及机器学习应用侧重点的快速参考。这有助于学生将理论知识与各资产类型的实践联系起来，并了解在何处寻找数据，直接支持课程的“实践导向”特性。

## **4. 加强实践环节：Python、数据与论文复现**

### **4.1. 金融机器学习的Python生态系统**

为了有效地进行机器学习在资产定价中的实践，掌握相关的Python库至关重要。

* **核心库:**
  + **Pandas:** 用于数据处理、时间序列分析以及处理金融数据集的核心库 9。其 DateTimeIndex、resample（重采样）、shift（移位）、rolling（滚动计算）和 expanding（扩展计算）等功能对金融数据分析尤为关键 65。
  + **NumPy:** 提供高效的数值计算和多维数组操作，是许多其他科学计算库的基础 9。
  + **Scikit-learn:** 一个全面的机器学习库，包含各类算法（回归、分类、聚类、降维）、模型选择工具（训练测试集划分、交叉验证）和预处理方法（如特征标准化 StandardScaler）9。
  + **Statsmodels:** 提供用于估计多种统计模型、执行统计检验和数据探索的类和函数，包括ETS、ARIMA模型和时间序列分解等 65。
  + **TensorFlow/Keras:** 用于构建和部署深度学习模型，特别是神经网络和LSTM模型 19。
  + **Matplotlib/Seaborn:** 用于数据可视化，绘制趋势图、分布图和模型结果图 67。
* **专业金融机器学习库:**
  + **PyAnomaly** 77**:** 专注于公司特征和因子生成、分位数投资组合构建以及横截面回归分析。该库覆盖了文献中超过200个公司特征，非常适合复现涉及大量因子的研究。
  + **gingado** 78**:** 一个开源的经济学机器学习库，有助于从官方来源获取时间序列数据，提供基于随机森林的自动基准模型，能够模拟具有因果关系的面板数据，并辅助模型文档的生成。
  + **yfinance** 33**:** 广泛用于从雅虎财经下载历史市场数据。虽然总体稳健，但在处理特定国际ETF时曾报告过一些问题 33。
  + **TA-Lib** 67**:** 广泛用于计算技术分析指标，如RSI、布林带、MACD等。
  + **Backtrader / vectorbt** 67**:** 用于回测交易策略。vectorbt 利用NumPy和Pandas实现高性能计算。

这些专业库，如 PyAnomaly 和 gingado，能够显著简化论文复现过程，特别是对于那些涉及大量公司特定特征或需要整合宏观经济数据的研究。它们通过封装繁琐的数据处理步骤，使学生能够更专注于机器学习建模的核心环节。在课程中引入这些工具，可以让学生接触到更高效的研究方法，从而在有限的课程时间内完成更具挑战性的复现任务。

下表总结了推荐的Python库及其在资产定价中的应用：

| **库名称** | **在课程中的主要用途** | **资产定价中的关键功能/特性** | **给本科生的提示** | **相关研究参考** |
| --- | --- | --- | --- | --- |
| Pandas | 数据处理与操作，时间序列分析 | DataFrame, Series, DateTimeIndex, groupby, merge, resample, rolling | 金融数据分析的基石，务必熟练掌握 | 65 |
| NumPy | 数值计算，数组运算 | array, 线性代数运算, 随机数生成 | Pandas的基础，理解其数组操作对性能优化有益 | 65 |
| Scikit-learn | 通用机器学习算法，模型评估与选择 | 各种回归/分类器, train\_test\_split, cross\_val\_score, Pipeline, StandardScaler | 入门友好，文档完善，覆盖大部分基础ML任务 | 22 |
| Statsmodels | 统计建模，时间序列模型，计量经济学分析 | OLS, ARIMA, VAR, GARCH, 时间序列分解, 统计检验 | 更侧重传统统计模型，与Scikit-learn互补 | 65 |
| TensorFlow/Keras | 深度学习模型构建 (NN, LSTM, Transformer) | Sequential API, Functional API, 各种层 (Dense, LSTM, Conv1D), 优化器 | 适用于复杂非线性建模，学习曲线较陡，从简单MLP入手 | 19 |
| PyAnomaly | 公司特征生成，因子模型，横截面分析 | 自动化生成200+文献特征，因子模型回归，投资组合排序 | 极大简化因子研究的数据准备工作，专注于模型应用 | 77 |
| gingado | 经济时间序列获取，面板数据模拟，自动基准模型 | 从官方源获取数据 (SDMX)，随机森林基准，因果面板数据生成 | 适合需要整合宏观经济数据或进行因果推断的研究 | 78 |
| yfinance | 从雅虎财经下载历史市场数据 | Ticker 对象, history() 方法获取OHLCV数据 | 方便快捷获取常用股票、指数数据，注意API的非官方性质和可能的限制 | 67 |
| TA-Lib | 技术分析指标计算 | RSI, MACD, Bollinger Bands, MA等常用指标 | 快速生成交易策略中常用的技术信号 | 67 |
| vectorbt / Backtrader | 交易策略回测 | 高性能回测引擎，与Pandas/NumPy集成，多种回测指标 | 学习如何客观评估策略表现，避免未来函数等陷阱 | 67 |

**价值主张：** 该表格为学生提供了一个关于金融领域标准机器学习Python工具的快速参考指南。突出显示特定用途和关键功能有助于学生理解在资产定价背景下*为何*以及*如何*使用每个库。“给本科生的提示”可以指出对初学者友好的方面或常见陷阱。这直接支持了课程“以Python代码实现为核心”的目标。

### **4.2. 数据获取与管理策略**

* **主要数据提供商 (具备API访问):**
  + **Alpha Vantage** 26**:** 提供全球股票（包括A股）、外汇、加密货币、大宗商品、经济指标、基本面数据以及新闻/情绪数据。其免费套餐通常有请求频率限制（例如，每天25次请求 26）。支持JSON/CSV格式，可通过HTTP请求（与Python兼容）访问。
  + **EODHD (EOD Historical Data)** 28**:** 覆盖70多个交易所（包括中国A股的上海XSHG和深圳XSHE 32），超过15万个交易代码，以及ETF、共同基金、指数、外汇、债券、基本面和新闻/情绪数据。其免费入门计划通常限制每日API调用次数（例如20次）和历史数据深度（例如过去一年）28。提供JSON/CSV输出和Python示例代码。虽然通用搜索未明确列出碳价格数据 32，但ICE数据（包括EUA期货 64）可能通过合作或高级套餐提供。
  + **Financial Modeling Prep (FMP)** 26**:** 提供股票（免费版侧重美国市场，付费版覆盖全球）、外汇、加密货币和大宗商品的实时及历史数据。免费套餐通常有每日调用限制（例如250次）和历史数据深度限制（例如5年）26。通过REST API提供JSON数据。
  + **Trading Economics** 56**:** 提供广泛的宏观经济数据，以及欧盟碳排放配额、大宗商品和外汇市场数据。提供API和Python包。免费访问权限有限，付费计划提供更多数据。
  + **Puro.earth** 57**:** 专门提供CORC碳清除价格指数（综合指数和生物炭指数）。提供方法论和数据概览。除指数值外，免费公开使用的API访问细节在摘要中未完全明确。
  + **Yahoo Finance (通过 yfinance 库)** 33**:** 广泛使用且免费。覆盖股票、ETF、指数以及部分大宗商品/期货。可通过 .SS（上海）和 .SZ（深圳）后缀获取A股数据 33。摘要中的搜索结果未直接显示 yfinance 可获取碳排放期货（EUA）的常用代码 58。
* **数据质量与预处理:** 强调数据清洗、缺失值处理（如5中讨论的插补技术）、归一化和特征缩放的重要性，因为机器学习模型对这些因素非常敏感 89。课程应涵盖scikit-learn的预处理工具。
* **横截面数据与时间序列数据的区分:** 课程提到了这两种数据类型。横截面数据（例如，特定时间点的公司特征）对因子模型至关重要。时间序列数据（例如，每日价格）用于预测。确保学生理解如何为每种类型的分析构建数据。PyAnomaly 77 非常适合生成横截面特征。Pandas 则是时间序列处理的关键 65。

金融数据提供领域的一个明显趋势是API优先的访问方式，并提供不同等级的免费和付费服务 26。这意味着，对于学生而言，学习如何通过编程方式获取、清洗和管理来自这些API的数据是一项至关重要的实践技能。课程应包含一个关于此内容的环节，例如使用Python的 requests 库或数据提供商的专用库。具体而言，由于课程强调“Python代码实现”和“实时操作”，而现实世界的金融分析依赖于从各种提供商处获取数据，且大多数现代数据提供商都提供API 29。学生需要学习如何与这些API交互以获取模型所需的数据，这包括理解API密钥、请求限制、数据格式（JSON、CSV）以及解析响应。因此，课程应投入时间演示如何使用Python进行API交互，涵盖发出请求、处理潜在错误以及将数据解析为Pandas DataFrame等内容。这对于任何有志于成为量化金融专业人士的人来说都是一项基本技能。

下表总结了部分数据提供商及其特点：

| **数据提供商** | **资产覆盖范围 (A股, 标普500, 外汇, 大宗商品, 碳价格, 宏观)** | **免费套餐可用性及限制 (请求限制, 历史深度, 特定资产)** | **访问方式 (API, Python库等)** | **相关文献索引** |
| --- | --- | --- | --- | --- |
| Alpha Vantage | 股票 (全球, 含A股), 外汇, 加密货币, 大宗商品, 经济指标, 基本面, 新闻/情绪 | 有免费套餐，通常限制每日请求次数 (如25次/天) 和历史数据深度 | HTTP API (JSON/CSV), Python可直接调用 | 26 |
| EODHD | 股票 (70+交易所, 含A股XSHG/XSHE), ETF, 基金, 指数, 外汇, 债券, 基本面, 新闻/情绪 | 有免费入门套餐，通常限制每日API调用 (如20次/天)，历史数据仅过去一年 | HTTP API (JSON/CSV), 提供Python代码示例 | 29 |
| Financial Modeling Prep (FMP) | 股票 (免费版侧重美国), 外汇, 加密货币, 大宗商品 | 有免费套餐，通常限制每日调用 (如250次/天)，历史数据5年 | REST API (JSON) | 26 |
| Trading Economics | 宏观经济数据, 欧盟碳配额 (EUA), 大宗商品, 外汇 | 免费访问有限，付费计划提供更多数据和API访问 | API, 提供Python包 | 56 |
| Yahoo Finance (via yfinance) | 股票, ETF, 指数, 部分商品/期货 (A股代码后缀.SS/.SZ) | 免费 | Python库 yfinance | 33 |
| Puro.earth | CORC碳清除价格指数 (综合和生物炭) | 指数值公开，API细节未详 | 网站数据，方法论可下载 | 57 |

**价值主张：** 实际的资产定价需要数据。学生，尤其是在短期课程中，需要关于可访问数据源的指导。一个关注与课程相关资产以及免费/低成本访问的数据提供商比较表非常实用。详细说明限制有助于为学生的项目和作业设定切合实际的期望。这支持了课程的“实践导向”特性和“实时操作”方面。

### **4.3. 论文复现策略**

* **核心论文：Gu, Kelly, Xiu (2020) (GKX) "Empirical Asset Pricing via Machine Learning"** 6**:** 这篇论文比较了多种机器学习模型在美国股票收益率预测中的表现，影响深远，是课程核心复现内容的绝佳选择。已有R语言 8 和Python语言（针对韩国市场 9；针对德国市场 10）的复现尝试。Tidy Finance的R语言复现 8 提供了详细步骤并讨论了挑战。韩国市场的Python复现 9 则突出了非美国市场数据可得性的问题。
* **备选/补充论文（供复现/讨论）:**
  + **Kelly et al. (2025) "Artificial Intelligence Asset Pricing Models"** 15**:** 介绍了基于Transformer的模型。完整复现可能过于复杂，但讨论其核心概念，或许复现其“线性投资组合Transformer”部分是可行且极具启发性的。目前摘要中未找到公开的Python代码 15。
  + **Qian & Zhang (2025) "Long-term forecasting in asset pricing: Machine learning models' sensitivity to macroeconomic shifts and firm-specific factors"** 12**:** 关注标普500成分股，比较了多种机器学习模型，并分析了对宏观因素的敏感性。这与课程范围高度吻合。目前摘要中未找到公开的Python代码 93。
  + **Zhang et al. (2023) "Carbon trading and COVID-19: a hybrid machine learning approach"** 14**:** 适用于碳价格模块。一个GitHub仓库 (pranavramkumar/wqu\_capstone\_5481 73) 引用了此文，并列出了用于碳价格预测的CEEMDAN-LSTM、XGBoost和Transformer的Python库。
  + **更简洁模型的论文:** 对于第一节课，复现一篇使用较简单模型的论文（例如，针对特定资产类别的惩罚性线性回归或随机森林模型）可能更容易上手，然后再尝试GKX或Kelly et al. (2025)的部分内容。
* **复现重点:** 论文复现不应仅仅是运行代码，更重要的是理解其方法论：数据预处理、特征工程、模型选择、超参数调整以及评估指标（如样本外R平方、投资组合表现等）。

对于本科生而言，从零开始复现复杂的学术论文极具挑战性且耗时。幸运的是，一些关键论文（如GKX (2020)）已有公开的GitHub复现代码（即使是部分复现或使用R语言）8。教师可以利用这些资源作为起点，根据需要将其调整为Python版本，并引导学生逐步理解代码，重点解释每一步骤背后的原因，而不是让学生在有限的时间内从头构建所有内容。例如，Tidy Finance上关于GKX (2020)的R语言复现 8 非常详细，并讨论了复现过程中遇到的挑战。其他针对不同市场的Python版本也存在 9。通过这种方式，课程可以引导学生分析、修改和扩展现有的复现代码。这不仅能教会他们阅读和理解研究代码这一宝贵的实践技能，也使得在课程中完成复杂的复现任务成为可能。教学重点从而从纯粹的编码转向对研究方法论和机器学习应用的理解。

## **5. 应对机器学习在资产定价中的关键挑战**

### **5.1. 过拟合与泛化能力**

* **问题描述:** 机器学习模型，特别是像神经网络或梯度提升树这样的复杂模型，可能会在训练数据上表现得过于优异（低偏差），但在新的、未见过的数据上表现不佳（高方差）1。由于金融数据固有的噪声特性和不断变化的市场动态，这在金融领域是一个臭名昭著的问题。“因子动物园”的存在进一步加剧了这一问题 1。
* **缓解技术:**
  + **正则化 (L1/Lasso, L2/Ridge, Elastic Net):** 通过对模型复杂度施加惩罚来防止过拟合 1。GKX (2020) 的研究中使用了Elastic Net。
  + **交叉验证 (例如 k-折交叉验证, 时间序列感知的划分):** 提供对样本外性能更稳健的估计，并用于超参数调整 96。对于时间序列数据，需要仔细划分数据（例如使用滚动窗口）以防止前视偏差。
  + **特征选择/工程:** 选择相关的预测变量并创建有意义的特征可以减少噪声和模型复杂度 4。
  + **集成方法 (随机森林, Boosting):** 结合多个模型的预测以提高泛化能力 22。
  + **早停 (Early Stopping):** 对于像神经网络这样的迭代模型，当验证集上的性能开始下降时停止训练。
* **课程整合:** 实践练习应始终涉及将数据划分为训练集、验证集（用于调优）和测试集。学生应报告样本内和样本外的性能指标。

### **5.2. 数据窥探 (前视偏差, 选择偏差)**

* **问题描述:** 在模型开发或选择过程中，无意中使用了来自测试集或未来的信息，导致回测结果过于乐观 97。这可能通过在同一数据集上反复进行假设检验或使用完整数据集进行过早的特征选择而发生。
* **后果:** 在回测中看起来表现优异的模型，在实际交易中却可能遭遇惨败 97。
* **缓解技术:**
  + **严格的训练-测试集划分:** 在最终评估之前，始终保留测试数据不被模型训练过程接触 97。
  + **时间序列的按时间顺序划分:** 确保训练数据严格位于验证/测试数据之前的时间段 97。
  + **多重检验校正:** 如果测试了许多模型/假设，则需要进行校正（例如Bonferroni校正、错误发现率控制）97。
  + **样本外测试:** 最终的检验标准是在真正未见过的数据上的表现。
* **课程整合:** 明确讨论数据窥探的概念。在所有编码练习中强调仔细的数据处理规程。

### **5.3. 模型可解释性 (XAI)**

* **问题描述:** 许多机器学习模型（尤其是神经网络和复杂的集成模型）通常被视为“黑箱”，难以理解它们为何做出特定的预测 2。这在金融领域是一个问题，原因包括监管要求（例如解释贷款决策 99）、建立信任以及模型调试。
* **在金融领域的重要性:** 监管合规（如GDPR, MiFID II）和问责制的需求使得透明度至关重要 4。理解特征重要性有助于进行经济解释 1。
* **技术方法:**
  + **特征重要性图 (例如 来自随机森林, GBRT):** 显示哪些变量驱动了预测。
  + **部分依赖图 (PDP):** 可视化一个或两个特征对预测结果的边际效应。
  + **SHAP (Shapley Additive Explanations)** 99**:** 一种基于博弈论的方法，通过将贡献归因于每个特征来解释单个预测。该方法与模型无关。
  + **LIME (Local Interpretable Model-agnostic Explanations):** 通过用可解释模型在局部逼近黑箱模型来解释单个预测。
* **课程整合:** 介绍可解释性的概念，并在案例研究中演示一到两种技术（例如，随机森林的特征重要性，或针对简单模型的基本SHAP分析）。这符合提供“直观体验”和减少“理论讲解”的目标。

### **5.4. 市场机制转换与非平稳性**

* **问题描述:** 由于经济周期、政策转变、危机事件等因素，金融市场表现出不断变化的动态（机制）38。在一个机制下训练的模型在另一个机制下可能表现不佳 12。平稳性假设通常不成立。
* **对机器学习模型的影响:** 如果没有明确设计来适应这些结构性断裂，机器学习模型可能难以适应 4。
* **自适应方法:**
  + **机制识别模型:** 使用机器学习（例如，3838 中带有GARCH的t分布混合模型；3938 中的统计跳跃模型）或统计方法（例如马尔可夫转换模型）来识别当前机制。
  + **机制特定模型:** 为不同机制训练不同的模型，或将机制作为特征纳入模型。
  + **动态重新训练/更新:** 定期用最新的数据重新训练模型。GKX (2020) 每年更新模型参数 8。
  + **时变参数模型:** 一些机器学习方法可以隐式地捕捉时变关系。
* **课程整合:** 讨论市场机制的概念。可以通过案例研究分析模型在不同历史时期（例如，2008年危机前、2008年危机后、COVID-19期间）的表现，以说明机制转换的影响。3838 识别了四种机制（新兴扩张期、稳健扩张期、谨慎衰退期、市场动荡期）并分析了各机制下的资产表现，这可以作为讨论的基础。

在实践中，过拟合、数据窥探、可解释性和市场机制转换等挑战并非孤立存在，而是相互关联的。例如，一个过拟合的模型更不可能在不同的市场机制下保持良好的泛化能力。而一个缺乏可解释性的“黑箱”模型，则使得在市场机制发生转变导致模型失效时，难以诊断其根本原因。数据窥探可能导致对一个实际上已经过拟合的模型的错误信任，使其在面对机制转换时更加脆弱。因此，课程应将这些挑战视为一个相互关联的整体，并强调稳健的金融机器学习需要一个全面的模型开发和验证方法，而不是将每个挑战视为独立的检查项。例如，在讨论用于防止过拟合的交叉验证时，如果数据跨越了这些时期，也应提及它在评估跨越潜在（历史）机制转变的泛化能力方面的作用。在讨论可解释性时，应将其与诊断机制变化期间模型故障联系起来。

同时，全面应对这些挑战通常需要复杂的技术和大量的计算资源，这可能超出了本科短期课程的范围。因此，教学重点应放在让学生*意识到*这些挑战，并*介绍*基本的缓解策略，而不是要求他们掌握先进的解决方案。“提供直观体验”的目标在此至关重要。对于每个挑战，课程可以介绍其概念，用简单的例子解释为何它在金融领域是一个问题，然后演示一到两种基本的Python技术来应对它（例如，针对过拟合使用简单的训练-验证-测试集划分和正则化；针对可解释性使用特征重要性分析；针对机制转换影响比较模型在两个不同历史子时期的表现）。这样可以在不给学生带来过重负担的前提下，培养他们的意识和基础技能。

## **6. 核心学习资源的筛选与组织**

### **6.1. 推荐的资产定价理论本科教材**

虽然本课程旨在最大限度地减少纯理论讲解，但学生仍需具备一定的理论基础。课程大纲已将“基本资产定价理论”列为先决条件。对于需要补充或复习这方面知识的学生，以下是一些参考方向：

* Kerry E. Back 的 "Asset Pricing and Portfolio Choice Theory" 66 和 George Pennacchi 的 "Theory of Asset Pricing" 101 都是该领域的经典著作，内容全面，涵盖效用理论、随机折现因子、因子模型、动态模型及衍生品定价等。但这两本书主要面向博士或金融数量硕士研究生，对于本科生而言可能过于艰深，更适合作为教师的参考用书。
* **建议：** 鉴于本课程的实践导向，除了标准资产定价教材作为背景知识外，更应推荐那些能够将资产定价理论与*实证*方法及机器学习应用联系起来的资源。例如，一些侧重于金融数据科学的Python书籍，或者像 "Tidy Finance with R" （尽管是R语言，但概念可迁移）这样的实践导向教材，可能对学生的编程作业更有直接帮助。课程可以将更高级教材的特定章节列为“拓展阅读”。

### **6.2. 关键综述论文与有影响力的研究文献**

* **综述性论文:**
  + Bagnara (2022) "Asset Pricing and Machine Learning: A critical review" 1: 对机器学习方法（正则化、降维、决策树、神经网络、比较分析）及其在资产定价中面临的计量经济学挑战进行了出色的概述。*这对于构建课程中机器学习技术的框架至关重要。*
  + Ye et al. (2024/2025) "From Factor Models to Deep Learning: Machine Learning in Reshaping Empirical Asset Pricing" 2: 全面回顾了监督学习、无监督学习、强化学习在资产定价中的应用，以及维度降低、缺失数据处理、另类数据融合、去噪等创新方法和面临的挑战。*这对于理解更广阔的背景和未来趋势非常有价值。*
  + Kapetanios & Kempf (2022) "Interpretable Machine Learning Modelling for Asset Pricing" 96: 专注于通过机器学习和雅可比正则化来推广Fama-Macbeth回归，并探讨了可解释性问题。*这对于课程中关于“模型可解释性”挑战的讨论很有帮助。*
* **有影响力的研究论文 (供复现/讨论的范例):**
  + Gu, Kelly, Xiu (2020) "Empirical asset pricing via machine learning" 6: 这是本课程计划复现的核心论文。
  + Kelly, Kuznetsov, Malamud, Xu (2025) "Artificial Intelligence Asset Pricing Models" 15: 介绍了基于Transformer的随机折现因子模型。*这篇论文代表了该领域的前沿研究。*
  + Freyberger, Neuhierl, Weber (2020) 7: 与机器学习在资产定价中的应用相关。
  + Kozak, Nagel, Santosh (2020) 7: 关于因子模型和机器学习的重要研究。
  + 针对特定资产类别的论文，例如：Qian & Zhang (2025) 关于标普500指数的研究 12，Zhang et al. (2023) 关于碳价格的研究 14。

### **6.3. 在线讲座、教程与Python资源**

* **资产定价基础 (在线讲座):**
  + Coursera "Asset Pricing Fundamentals" 103: 涵盖金融市场功能、不确定性下的选择、或有要求权市场、投资组合问题、CAPM等。课程难度为中级，假设学生具备本科水平的数学、统计学和微观经济学知识。*有助于巩固资产定价理论的先修知识。*
  + CEU "Asset Pricing" (ECBS5184) 104: 涵盖股票、债券、衍生品定价，量化工具，并与当前金融新闻联系。*是资产定价概念的另一个良好参考。*
* **金融与机器学习的Python教程/课程:**
  + Quantra "Python for Machine Learning in Finance" 68: 涵盖Python基础 (Numpy, Pandas)，数据导入/可视化，机器学习任务（目标/特征创建, RSI, EMA），机器学习算法概述，训练测试集划分，模型训练/预测，评估指标，回测。*与本课程的Python部分直接相关。*
  + GeeksforGeeks "Time Series Data Visualization in Python" 69: 使用Pandas, Numpy, Seaborn, Matplotlib, statsmodels进行趋势、季节性、ACF、ADF检验、差分、移动平均等分析，并以股票数据为例。*有助于培养时间序列数据处理技能。*
  + DataCamp "Time Series Analysis in Python" 75: 使用Python讲解相关性、自相关性、白噪声、随机游走、ARMA、协整等概念。*对时间序列分析方面很有用。*
  + Tidy Finance 博客/教程 (例如 8): 提供R和Python代码示例，用于各种金融机器学习任务，包括GKX复现和期权定价。*是非常好的实践案例。*
  + Interactive Brokers, Python for ML in Finance 66: 关于使用Python进行股票预测的机器学习集成的文章，提及Pandas, Numpy, Alpha Vantage, scikit-learn。
  + Python for Finance: Time Series Analysis (MLQ.ai 博客) 65: Pandas的DateTime索引、重采样、移位、滚动/扩展计算、布林带，以及Statsmodels (ETS, EWMA, ARIMA)。
  + ProjectPro "Stock Price Prediction using Machine Learning Project" 23: 讨论了LSTM、移动平均、RMSE/MAPE，并使用Keras和NFLX数据作为示例。
  + 关于Python中因子数据和收益率的YouTube教程 106。

尽管目前已有许多优秀的通用Python机器学习资源 68 和针对特定金融任务（如时间序列分析 65）的教程，但在将基础资产定价理论与面向本科生的机器学习实践相结合方面，仍存在一定的教学资源缺口。本课程，特别是其论文复现和以Python为中心的案例研究，恰好能够填补这一空白。因此，在筛选学习资源时，应侧重于那些能够支持这种整合学习的材料。

此外，一些关键论文（如GKX (2020)）的公开GitHub复现代码（即使是部分复现或使用R语言编写）8 是非常宝贵的教学资产。教师可以利用这些代码作为起点，根据需要将其调整为Python版本，并引导学生理解代码的每一步，而不是要求他们在有限的时间内从头开始编写所有内容。这不仅能教会学生阅读和理解研究代码这一宝贵的实践技能，也使得在课程中完成复杂的复现任务成为可能，教学重点从而从纯粹的编码转向对研究方法论和机器学习应用的理解。

## **7. 课程强化的战略性建议**

为了进一步提升“机器学习与资产定价”短期课程的教学效果和学生体验，兹提出以下战略性建议：

* 平衡广度与深度:  
  考虑到课程仅有三次课，若要对所有五个资产类别（A股、股指、大宗商品、外汇、碳价格）都进行专门的论文复现，任务可能过于繁重。
  + **建议方案A (侧重深度):** 选择2-3个具有代表性的、差异化的资产类别进行深入讲解和论文复现（例如，一个主流股票指数如标普500，A股以突显市场差异性，以及大宗商品或碳价格以引入非股权类资产）。其他资产类别可以通过简要的案例研究进行介绍，或者作为课后作业，让学生应用所学方法进行拓展。
  + **建议方案B (兼顾广度与聚焦复现):** 对所有资产类别进行概览式介绍，但选择一至两篇奠基性或方法论上具有普适性的论文（如GKX (2020) 6，或Kelly et al. (2025) 15 的部分内容），其原理可以被阐释并应用于多个资产类别。“每节课一篇论文”可以理解为针对一篇大型论文的不同方面，或者选择规模较小、更聚焦的论文。
* **精心设计论文复现环节:**
  + **建议:** 为每次论文复现提供结构清晰、注释详尽的Python Jupyter Notebook模板。该模板可以预先处理数据加载和基本的数据清洗步骤，使学生能够将精力集中在实现核心的机器学习模型和解读结果上。这通过减少环境配置和数据准备的障碍，更好地支持了课程“实践导向”和“提供直观体验”的目标。
  + 可以借鉴已有的复现成果（例如 8）作为教学起点或教师参考。
* **主动整合挑战性议题的讨论:**
  + **建议:** 在每节课或每个案例研究中，安排专门的片段讨论相关的挑战（如过拟合、数据质量、模型可解释性、市场机制转换等）。例如，在讨论模型评估时，引入样本外R平方并讨论过拟合问题。在展示基于树的模型结果时，演示特征重要性作为一种基础的可解释性工具。
* **优化课后作业设计:**
  + **建议:** 课后作业应要求学生将当堂课学习的机器学习技术和Python代码应用于略有差异的数据集或资产类别，或者修改所复现论文中的参数。这有助于巩固学习成果并鼓励学生独立探索。例如，如果第一节课复现了在标普500指数上使用随机森林的论文，那么课后作业可以是要求学生将随机森林应用于A股数据或其他指数。
* **夯实Python与机器学习基础:**
  + **建议:** 尽管Python基础是先修要求，但提供一个包含课程相关的核心Pandas、NumPy和Scikit-learn命令的“复习型”Jupyter Notebook会很有帮助。同样，在第一节课开始时，对基础机器学习概念（如监督学习与无监督学习、偏差-方差权衡、常用评估指标）进行简明扼要的回顾也是有益的，这与课程大纲中“机器学习基础知识（课程中会简要介绍）”的说明一致。Quantra课程 68 的结构可以作为参考。
* **最终项目/总结性元素 (可选，若时间/课程结构允许扩展):**
  + **建议:** 如果课程可以扩展或具有一定的灵活性，一个小型最终项目——让学生选择一个资产类别和一种机器学习模型进行小型的实证研究——将会是一个非常有价值的总结性体验。
* **制定数据可访问性计划:**
  + **建议:** 优先选择那些为学生提供强大免费套餐或学术许可的数据源（例如 yfinance、Alpha Vantage免费套餐、EODHD免费套餐）。对于免费数据稀缺的资产类别（例如碳价格或深度的A股历史特征数据），教师应提前准备并提供数据集，以确保所有学生都能完成练习。对于需要API密钥的服务（如Alpha Vantage或EODHD），应提供清晰的密钥获取指南。
* **强调经济直觉与机器学习的结合:**
  + **建议:** 尽管课程“侧重实践”并旨在“减少理论讲解”，但将机器学习模型的结果与资产定价理论和经济直觉联系起来至关重要。例如，如果某个特征被发现是重要的，应讨论其在经济学上作为收益率预测因子的合理性。这可以防止机器学习沦为纯粹的“黑箱”操作。Bagnara (2022) 1 的评论中提到机器学习偏离传统计量经济学以及需要经济解释，这一点非常关键。

一个逻辑清晰且循序渐进的课程流程对于这样一个短小精悍的实践性课程至关重要。可以考虑从更成熟、可解释性更强的模型和易于获取的数据入手（例如，在标普500数据上应用线性模型或随机森林），然后逐步过渡到更复杂的模型或数据环境（例如神经网络、A股市场，或大宗商品的另类数据），并在此过程中持续融入对各种挑战和模型解释的讨论。

具体而言：

1. **第一节课** 可以聚焦于：基础的机器学习工作流程，获取主要指数数据（例如使用 yfinance 获取标普500数据），复现一篇使用随机森林或惩罚性回归的论文（例如，简化并调整GKX的方法论）。引入过拟合的概念和基本的模型可解释性方法（如特征重要性）。
2. **第二节课** 可以探索：一个不同的资产类别（例如A股或外汇），并突出其数据获取方面的挑战；或者引入一个更复杂的模型（例如用于时间序列预测的简单神经网络/LSTM）。通过比较模型在不同历史时期的表现来讨论市场机制转换的影响。
3. **第三节课** 可以尝试：一个新兴的资产类别（例如，如果数据可管理，则选择碳价格；或者选择包含另类数据提示的大宗商品），或者讨论一篇近期前沿论文中的更高级概念（例如，Kelly et al. (2025) 15 中提出的跨资产信息共享思想，即使不完全实现其代码）。回顾并强调所有挑战和最佳实践。

这种渐进式的复杂性设计，允许学生逐步建立信心和技能，与课程“提供直观体验”的目标相一致，同时也能够触及高级主题和多样化的资产类别。

## **8. 结论与展望**

本报告对“机器学习与资产定价”短期课程的设立提供了多方面的分析与建议。该课程以其高度的实践导向、对Python编程的重视以及通过论文复现连接前沿研究的教学方法，准确把握了当代金融教育与行业发展的脉搏。

核心建议包括：

1. **聚焦课程内容：** 在有限的课时内，应审慎平衡所涉猎资产类别的广度与每个类别学习的深度。建议选取2-3个核心资产类别进行深入的论文复现和案例分析，其他资产类别则可通过引导学生在课后利用所学方法进行探索。
2. **精选核心算法与前沿视野：** 课程应涵盖正则化方法、基于树的模型、神经网络（包括LSTM）和降维技术等核心机器学习工具。同时，适时引入如Transformer等前沿模型的核心思想，能够有效拓展学生的学术视野。
3. **强化Python实践生态：** 充分利用Pandas、NumPy、Scikit-learn等基础库，并引入PyAnomaly、gingado等金融专用库，可以极大提升教学效率和学生的实践能力。数据获取方面，应指导学生熟悉常用金融数据API（如Alpha Vantage, EODHD, yfinance）的使用，并注意免费额度的限制。
4. **系统应对机器学习挑战：** 教学过程中需系统性地融入对过拟合、数据窥探、模型可解释性（XAI）和市场机制转换等核心挑战的讨论与应对策略介绍，培养学生严谨的量化研究素养。
5. **整合优质学习资源：** 推荐结合经典综述（如Bagnara (2022) 1, Ye et al. (2024/2025) 2）与具有代表性的实证研究论文（如Gu, Kelly, Xiu (2020) 6），并辅以优质的在线教程和Python编程资源。

展望未来，机器学习在资产定价领域的应用将持续深化。本课程通过培养学生的实践操作能力和对前沿研究的理解，将为他们未来在学术研究或金融实务中应用这些强大工具打下坚实的基础。通过精心设计课程内容、教学方法和实践环节，有望使学生真正掌握从机器学习视角参与资产定价实践的核心技能，并获得宝贵的直观体验。

#### Works cited

1. www.econstor.eu, accessed May 22, 2025, <https://www.econstor.eu/bitstream/10419/288177/1/JOES_JOES12532.pdf>
2. From Factor Models to Deep Learning: Machine Learning in Reshaping Empirical Asset Pricing - ResearchGate, accessed May 22, 2025, <https://www.researchgate.net/publication/379289960_From_Factor_Models_to_Deep_Learning_Machine_Learning_in_Reshaping_Empirical_Asset_Pricing>
3. [2403.06779] From Factor Models to Deep Learning: Machine Learning in Reshaping Empirical Asset Pricing - arXiv, accessed May 22, 2025, <https://arxiv.org/abs/2403.06779>
4. (PDF) Intelligent Finance: Emerging Applications and Challenges of Machine Learning in Asset Pricing - ResearchGate, accessed May 22, 2025, <https://www.researchgate.net/publication/388054323_Intelligent_Finance_Emerging_Applications_and_Challenges_of_Machine_Learning_in_Asset_Pricing>
5. arxiv.org, accessed May 22, 2025, <https://arxiv.org/pdf/2403.06779>
6. Empirical Asset Pricing via Machine Learning: The Global Edition, accessed May 22, 2025, <https://www.iimb.ac.in/sites/default/files/inline-files/Empirical%20Asset-Pricing-Machine-Learning.pdf>
7. arXiv:2503.00549v1 [econ.EM] 1 Mar 2025, accessed May 22, 2025, <https://arxiv.org/pdf/2503.00549>
8. Replicating Gu, Kelly & Xiu (2020) – Tidy Finance, accessed May 22, 2025, <https://www.tidy-finance.org/blog/gu-kelly-xiu-replication/>
9. leeway00/Asset-Pricing-via-ML: Empirical asset pricing via ... - GitHub, accessed May 22, 2025, <https://github.com/leeway00/Asset-Pricing-via-ML>
10. kristina969/Empirical-Asset-Pricing-via-Machine-Learning ... - GitHub, accessed May 22, 2025, <https://github.com/kristina969/Empirical-Asset-Pricing-via-Machine-Learning-Evidence-from-the-German-Stock-Market>
11. Automated Machine Learning and Asset Pricing - MDPI, accessed May 22, 2025, <https://www.mdpi.com/2227-9091/12/9/148>
12. Long-term forecasting in asset pricing: Machine learning models ..., accessed May 22, 2025, <https://ideas.repec.org/a/eee/ecofin/v78y2025ics1062940825000634.html>
13. AI-Enhanced Factor Analysis for Predicting S&P 500 Stock Dynamics - arXiv, accessed May 22, 2025, <https://arxiv.org/pdf/2412.12438>
14. Carbon trading and COVID-19: a hybrid machine learning approach ..., accessed May 22, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10127197/>
15. NBER WORKING PAPER SERIES ARTIFICIAL INTELLIGENCE ASSET PRICING MODELS Bryan T. Kelly Boris Kuznetsov Semyon Malamud Teng Andrea, accessed May 22, 2025, <https://www.nber.org/system/files/working_papers/w33351/w33351.pdf>
16. [2503.00549] The Uncertainty of Machine Learning Predictions in Asset Pricing - arXiv, accessed May 22, 2025, <https://arxiv.org/abs/2503.00549>
17. Forecasting S&P 500 Using LSTM Models - arXiv, accessed May 22, 2025, <https://arxiv.org/html/2501.17366v1>
18. Comparative Analysis of Machine Learning Models for S&P 500 Prediction, accessed May 22, 2025, <https://norma.ncirl.ie/7542/>
19. Machine Learning Algorithms for Forex Market Analysis | Market Pulse, accessed May 22, 2025, <https://fxopen.com/blog/en/machine-learning-algorithms-for-forex-market-analysis/>
20. Predicting Forex Currency Exchange Rate using Machine Learning - GitHub, accessed May 22, 2025, <https://cepdnaclk.github.io/e18-4yp-Predicting-Forex-Currency-Exchange-Rate-using-Machine-Learning/>
21. accessed January 1, 1970, <https://arxiv.org/pdf/2501.17366.pdf>
22. Machine Learning Models for S&P 500 Trading: A Comparative ..., accessed May 22, 2025, <https://bsic.it/machine-learning-models-for-sp-500-trading-a-comparative-analysis-of-random-forest-xgboost-and-regression-techniques/>
23. Stock Price Prediction using Machine Learning with Source Code - ProjectPro, accessed May 22, 2025, <https://www.projectpro.io/article/stock-price-prediction-using-machine-learning-project/571>
24. Artificial Intelligence Asset Pricing Models - EconPapers - RePEc, accessed May 22, 2025, <https://econpapers.repec.org/RePEc:nbr:nberwo:33351>
25. Predicting Stock Prices with Machine Learning: A Comparative Analysis of Models and Investment Strategies - NHSJS, accessed May 22, 2025, <https://nhsjs.com/2025/predicting-stock-prices-with-machine-learning-a-comparative-analysis-of-models-and-investment-strategies/>
26. Best Free Finance APIs for Stock & Crypto Market Data in 2025: A Detailed Comparison, accessed May 22, 2025, <https://noteapiconnector.com/best-free-finance-apis>
27. API Documentation | Alpha Vantage, accessed May 22, 2025, <https://www.alphavantage.co/documentation/>
28. Fundamental Data API for 70+ exchanges worldwide, accessed May 22, 2025, <https://eodhd.com/lp/fundamental-data-api>
29. End-of-Day Historical Stock Market Data API: The Best Web Service ..., accessed May 22, 2025, <https://eodhd.com/financial-apis/api-for-historical-data-and-volumes>
30. accessed January 1, 1970, <https://eodhd.com/financial-apis/stock-api-and-fundamental-data>
31. accessed January 1, 1970, <https://eodhd.com/financial-apis/all-in-one>
32. Financial Exchanges API. Get Stock Market Ticker Lists | Free plan, accessed May 22, 2025, <https://eodhd.com/financial-apis/list-supported-exchanges>
33. Fail to fetch history of China ETF · Issue #2274 · ranaroussi/yfinance - GitHub, accessed May 22, 2025, <https://github.com/ranaroussi/yfinance/issues/2274>
34. Awesome-data shows most interesting data-source around the financial world! - GitHub, accessed May 22, 2025, <https://github.com/akfamily/awesome-data>
35. Alpha Vantage: Free Stock APIs in JSON & Excel, accessed May 22, 2025, <https://www.alphavantage.co/>
36. Financial data API marketplace: Indices Historical Constituents Data API, accessed May 22, 2025, <https://eodhd.com/marketplace/unicornbay/spglobal>
37. S&P 500 (SPX) Historical Data - Nasdaq, accessed May 22, 2025, <https://www.nasdaq.com/market-activity/index/spx/historical>
38. www.ssga.com, accessed May 22, 2025, <https://www.ssga.com/library-content/assets/pdf/global/pc/2025/decoding-market-regimes-with-machine-learning.pdf>
39. Regime-Aware Asset Allocation: a Statistical Jump Model Approach - arXiv, accessed May 22, 2025, <https://arxiv.org/html/2402.05272v1>
40. Finance and Credit: Websites - LibGuides at Salt Lake Community College, accessed May 22, 2025, <https://libguides.slcc.edu/finance/websites>
41. Cross-Sectional Data Analysis - Definition, Uses, and Sources - Corporate Finance Institute, accessed May 22, 2025, <https://corporatefinanceinstitute.com/resources/data-science/cross-sectional-data-analysis/>
42. Machine Learning for Asset Management and Pricing - SIAM Publications Library, accessed May 22, 2025, <https://epubs.siam.org/doi/book/10.1137/1.9781611977905>
43. Importance of Artificial Intelligence in Commodity Price Forecasting ..., accessed May 22, 2025, <https://community.nasscom.in/communities/ai/importance-artificial-intelligence-commodity-price-forecasting>
44. AI in commodity price forecasting: how it transforms trading - Agiboo, accessed May 22, 2025, <https://www.agiboo.com/ai-in-commodity-trading-a-series-part-ii-better-price-forecasting-what-ai-actually-delivers/>
45. Commodities API | Real-Time & Historical Prices for Crude Oil, Gold, Silver, Coffee & More., accessed May 22, 2025, <https://commodities-api.com/>
46. Commodity Market Data APIs | Real-Time Quotes | FMP, accessed May 22, 2025, <https://site.financialmodelingprep.com/datasets/commodity>
47. A Deep Dive into Forex Trends and Strategies - Number Analytics, accessed May 22, 2025, <https://www.numberanalytics.com/blog/deep-dive-forex-trends-strategies>
48. AI in Forex Trading: A Game-Changer for 2025 - The5ers, accessed May 22, 2025, <https://the5ers.com/ai-in-forex-trading-a-game-changer/>
49. Top 5 Free Financial Data APIs for Building a Powerful Stock Portfolio Tracker, accessed May 22, 2025, <https://dev.to/williamsmithh/top-5-free-financial-data-apis-for-building-a-powerful-stock-portfolio-tracker-4dhj>
50. Full Historical Forex Chart API | Financial Modeling Prep, accessed May 22, 2025, <https://site.financialmodelingprep.com/developer/docs/stable/forex-historical-price-eod-full>
51. Overview | Forex REST API - Polygon.io, accessed May 22, 2025, <https://polygon.io/docs/rest/forex/overview>
52. Empirical Asset Pricing and Machine Learning Applications - Stockholm School of Economics - SSE Research Hub, accessed May 22, 2025, <https://research.hhs.se/esploro/outputs/doctoral/Empirical-Asset-Pricing-and-Machine-Learning/991001639499206056>
53. Carbon Credits Price Prediction Model (CCPPM) - OPUS at UTS, accessed May 22, 2025, <https://opus.lib.uts.edu.au/rest/bitstreams/461f0803-c620-4550-94c4-d873601d74c9/retrieve>
54. Application of various machine learning algorithms in view of predicting the CO2 emissions in the transportation sector | Science and Technology for Energy Transition (STET), accessed May 22, 2025, <https://www.stet-review.org/articles/stet/full_html/2024/01/stet20240008/stet20240008.html>
55. A Hybrid Model for Carbon Price Forecasting Based on Improved Feature Extraction and Non-Linear Integration - MDPI, accessed May 22, 2025, <https://www.mdpi.com/2227-7390/12/10/1428>
56. EU Carbon Permits - Price - Chart - Historical Data - News, accessed May 22, 2025, <https://tradingeconomics.com/commodity/carbon>
57. CORC Carbon Removal Price Indexes - Puro.earth, accessed May 22, 2025, <https://puro.earth/corc-carbon-removal-indexes>
58. Carbon Emissions Futures Historical Prices - Investing.com, accessed May 22, 2025, <https://www.investing.com/commodities/carbon-emissions-historical-data>
59. EUA Futures Pricing - ICE, accessed May 22, 2025, <https://www.ice.com/products/197/EUA-Futures/data>
60. yfinance - PyPI, accessed May 22, 2025, <https://pypi.org/project/yfinance/>
61. ICE EUA Futures Dec '25 (CKZ25) - Barchart.com, accessed May 22, 2025, <https://www.barchart.com/futures/quotes/CKZ25>
62. ICE Launches EUA 2 Futures To Support EU Carbon Market Expansion - Nasdaq, accessed May 22, 2025, <https://www.nasdaq.com/articles/ice-launches-eua-2-futures-support-eu-carbon-market-expansion>
63. Supported Commodities, accessed May 22, 2025, <https://www.commodities-api.com/symbols>
64. ICE Connect, accessed May 22, 2025, <https://www.ice.com/publicdocs/ICE_Connect_Platform_Overview.pdf>
65. Python for Finance: Time Series Analysis - MLQ.ai, accessed May 22, 2025, <https://blog.mlq.ai/python-for-finance-time-series-analysis/>
66. Python for Machine Learning in Finance | IBKR Quant, accessed May 22, 2025, <https://www.interactivebrokers.com/campus/ibkr-quant-news/python-for-machine-learning-in-finance/>
67. Best Python Libraries for Algorithmic Trading and Financial Analysis - QuantInsti Blog, accessed May 22, 2025, <https://blog.quantinsti.com/python-trading-library/>
68. Python for Machine Learning in Finance Course - Quantra by QuantInsti, accessed May 22, 2025, <https://quantra.quantinsti.com/course/python-machine-learning>
69. Time Series Analysis & Visualization in Python | GeeksforGeeks, accessed May 22, 2025, <https://www.geeksforgeeks.org/time-series-data-visualization-in-python/>
70. Option Pricing via Machine Learning with Python – Tidy Finance, accessed May 22, 2025, <https://www.tidy-finance.org/python/option-pricing-via-machine-learning.html>
71. Asset pricing & factor regressions - Pythonic Finance, accessed May 22, 2025, <https://randlow.github.io/posts/finance-economics/asset-pricing-regression/>
72. What is YFinance library? - GeeksforGeeks, accessed May 22, 2025, <https://www.geeksforgeeks.org/what-is-yfinance-library/>
73. pranavramkumar/wqu\_capstone\_5481: Git repo for WQU ... - GitHub, accessed May 22, 2025, <https://github.com/pranavramkumar/wqu_capstone_5481>
74. yfinance Library - A Complete Guide - AlgoTrading101 Blog, accessed May 22, 2025, <https://algotrading101.com/learn/yfinance-guide/>
75. Time Series Analysis in Python Course - DataCamp, accessed May 22, 2025, <https://www.datacamp.com/courses/time-series-analysis-in-python>
76. 36. Elementary Asset Pricing Theory - Advanced Quantitative Economics with Python, accessed May 22, 2025, <https://python-advanced.quantecon.org/asset_pricing_lph.html>
77. About — PyAnomaly 1.01 documentation - Read the Docs, accessed May 22, 2025, <https://pyanomaly.readthedocs.io/en/latest/introduction.html>
78. gingado: a machine learning library focused on economics and ..., accessed May 22, 2025, <https://www.bis.org/publ/work1122.htm>
79. Empowering Financial Insights: Unlocking the Potential of Yahoo Finance API - SmythOS, accessed May 22, 2025, <https://smythos.com/ai-integrations/api-integration/yahoo-finance-api/>
80. Top 5 Stock Data Providers of 2025: Features, Pricing & More, accessed May 22, 2025, <https://brightdata.com/blog/web-data/best-stock-data-providers>
81. Best Stock Market APIs (Reviewed in 2025), accessed May 22, 2025, <https://www.alphavantage.co/best_stock_market_api_review/>
82. Alpha Vantage | Documentation | Postman API Network, accessed May 22, 2025, <https://www.postman.com/api-evangelist/blockchain/documentation/j4n0jl2/alpha-vantage>
83. Analyzing Apple's Stock Like an Institution with Python - InsightBig, accessed May 22, 2025, <https://www.insightbig.com/post/analyzing-apple-s-stock-like-an-institution-with-python>
84. China De Xiao Quan Care Group Co. Ltd (CDXQ PINK) stock market data APIs, accessed May 22, 2025, <https://eodhd.com/financial-summary/CDXQ.US>
85. User API and Monitoring daily API Limits | EODHD APIs Blog - EOD Historical Data, accessed May 22, 2025, <https://eodhd.com/financial-apis-blog/user-api-and-monitoring-daily-api-limits>
86. Historical Prices and Fundamental Financial Data API, accessed May 22, 2025, <https://eodhd.com/pricing>
87. ICE Launches EU Carbon Allowance 2 Futures - Intercontinental Exchange, accessed May 22, 2025, <https://ir.theice.com/press/news-details/2025/ICE-Launches-EU-Carbon-Allowance-2-Futures/default.aspx>
88. ICE sees traction in new EU emissions futures - FOW, accessed May 22, 2025, <https://www.fow.com/insights/ice-sees-traction-in-new-eu-emissions-futures>
89. Machine Learning for Quantitative Finance: Use Cases and ..., accessed May 22, 2025, <https://roundtable.datascience.salon/machine-learning-for-quantitative-finance-use-cases-and-challenges>
90. Datastream Macroeconomics Analysis - LSEG, accessed May 22, 2025, <https://www.lseg.com/en/data-analytics/products/datastream-macroeconomic-analysis>
91. bkelly-lab/ReplicationCrisis: Code for "Is There a Replication Crisis in Finance" by Jensen, Kelly and Pedersen (2023) - GitHub, accessed May 22, 2025, <https://github.com/bkelly-lab/ReplicationCrisis>
92. Artificial Intelligence Asset Pricing Models - IDEAS/RePEc, accessed May 22, 2025, <https://ideas.repec.org/p/nbr/nberwo/33351.html>
93. Large Language Models for Financial and Investment Management: Applications and Benchmarks, accessed May 22, 2025, <https://www.pm-research.com/content/iijpormgmt/51/2/162>
94. Selected Topics in Time Series Forecasting: Statistical Models vs. Machine Learning - MDPI, accessed May 22, 2025, <https://www.mdpi.com/1099-4300/27/3/279>
95. website/blog/gu-kelly-xiu-replication/index.qmd at main - GitHub, accessed May 22, 2025, <https://github.com/tidy-finance/website/blob/main/blog/gu-kelly-xiu-replication/index.qmd>
96. Interpretable Machine Learning Modelling for Asset Pricing - King's College London, accessed May 22, 2025, <https://www.kcl.ac.uk/business/assets/pdf/dafm-working-papers/2022-papers/interpretable-machine-learning-modelling-for-asset-pricing.pdf>
97. Exploring the Impact of Data Snooping on Machine Learning Models, accessed May 22, 2025, <https://www.numberanalytics.com/blog/exploring-impact-data-snooping-machine-learning-models>
98. 6 Data Snooping Stats to Boost Your Model Performance - Number Analytics, accessed May 22, 2025, <https://www.numberanalytics.com/blog/6-data-snooping-stats-to-boost-model-performance>
99. Explainability in Machine Learning - Take Control of ML and AI ..., accessed May 22, 2025, <https://www.seldon.io/explainability-in-machine-learning-2/>
100. A Quick Guide to Model Explainability - Experian Insights, accessed May 22, 2025, <https://www.experian.com/blogs/insights/model-explainability/>
101. Asset Pricing and Portfolio Choice Theory - Kerry E. Back - Oxford ..., accessed May 22, 2025, <https://global.oup.com/academic/product/asset-pricing-and-portfolio-choice-theory-9780190241148>
102. Theory of Asset Pricing: 9780321127204: Economics Books @ Amazon.com, accessed May 22, 2025, <https://www.amazon.com/Theory-Asset-Pricing-George-Pennacchi/dp/032112720X>
103. Asset Pricing Fundamentals | Coursera, accessed May 22, 2025, <https://www.coursera.org/learn/asset-pricing-fundamentals>
104. (ECBS5184)Asset Pricing - Study guide, accessed May 22, 2025, <https://ceu.studyguide.timeedit.net/modules/ECBS5184>
105. Machine Learning & Deep Learning in Trading – I | Online Courses | Quantra by Quantinsti, accessed May 22, 2025, <https://quantra.quantinsti.com/learning-track/machine-learning-deep-learning-trading-1>
106. Algorithmic Trading – Machine Learning & Quant Strategies Course with Python - YouTube, accessed May 22, 2025, <https://www.youtube.com/watch?v=9Y3yaoi9rUQ>