

课程安排

oooooooooooooo

什么是金融人工智能模型

oooooooooooooooooooo

机器学习模型的优势

oooooooooooo

经济与金融行业就业需求

oooooooooooooo

学习与实战资源

oooooooooooo

# 人工智能与 Python (金融)

Class 1: Why AI & Machine Learning for Economists

葛雷

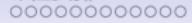
中国人民大学经济院

2025 年 2 月 19 日

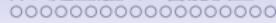


中國人民大學  
RENMIN UNIVERSITY OF CHINA

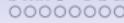
课程安排



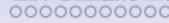
什么是金融人工智能模型



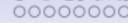
机器学习模型的优势



经济与金融行业就业需求



学习与实战资源



## 课程安排

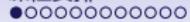
## 什么是金融人工智能模型

## 机器学习模型的优势

## 经济与金融行业就业需求

## 学习与实战资源

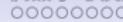
课程安排



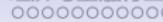
什么是金融人工智能模型



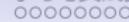
机器学习模型的优势



经济与金融行业就业需求



学习与实战资源



## 课程安排

什么是金融人工智能模型

机器学习模型的优势

经济与金融行业就业需求

学习与实战资源

## 教学团队：英语班

- 葛雷 (中国人民大学数量经济)
  - 莫湘源 (中国人民大学经济学院)
  - 徐一宸 (中国人民大学信息学院)

教学团队：数学班

- 葛雷 (中国人民大学数量经济学)
  - 王晨曦 (中国人民大学数学学院、金融学院)
  - 崔天佑 (中国人民大学数学学院、金融学院)

# 作为经济学家为什么要量化建模？

- 未来经济学家必备工具
  - 人工智能算法远胜人工分析

## Example: 人工智能精度与速度都超过房产分析师

- 涉房不良资产包预测的误差仅 **3%** (分析师误差 10%)
- 单个房产价值预测的精准度约为 5% (分析师误差 8%)
- 每小时评估几万处房产 (分析师 1 户/小时)

# 成果期望

1. Critical Thinking (模型化的推理与逻辑方式)
2. Python 处理金融与经济数据的能力 (Data Scientist)
3. Python 建立数量模型分析与预测的能力

# 修课前需要的能力

1. 经济：大家从大一起经过了 3 年的严格训练
2. 数学：大家从高中起经过了 6 年的严格训练
3. 英语：大家从初中起经过了 9 年的严格训练

Bottomline：量化建模 = 经济 + 数学 + 英语 + Python。量化建模是大家十年苦练，一展抱负与能力的舞台。

## 课程安排

课程考核模拟公司经济量化部门实战商务，锻炼同学们的编程能力、团队协助解题能力 (Free Rider? How to deal with)、实战商务项目能力、经济学展示与表达能力。这些能力对于做学术的同学们同样适用。

- 作业 16%：编程作业，以小组为单位完成编程任务 (No Free Rider) (5 次模型作业每次 4%)
- 研讨交流 8%：课堂 cold call，模拟真实 business challenges
- 期中考试 16%：编程项目期中考评，商务量化模型项目考评
- 期末考核 60%：编程项目期末考评，商务量化模型项目考评

# 主要教材

- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (网上有免费版)
- Hands-On Large Language Models (Git Source)
- Quantecon.org (founded by Thomas Sargent)

# 延伸教材

- Greene, Econometric Analysis, 8th Edition
- The Elements of Statistical Learning (Trevor Hastie , Robert Tibshirani , Jerome Friedman)
- Deep Learning, Goodfellow
- 西瓜书

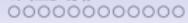
# 课程大纲

- 第 1 周：什么是经济学机器学习模型，简绍人工智能机器学习模型在经济金融行业的应用了解相关行业
- 第 2-3 周：机器学习建模流程，机器学习建模流程与怎样用经济学知识解释量化模型, Python 安装相关程序，掌握 python
- 第 4 周：实战金融数据分析，利用爬虫获取网络数据，利用 API 下载正式金融数据，真实金融大数据分析（其中期末量化项目开题）
- 第 5-6 周：线性机器学习模型训练, 线性机器学习模型 (OLS, Lasso, Ridge, Elastic Net)，分类预测

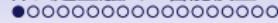
# 课程大纲

- 第 7 周或第 8 周 (期中考试): 实战机器学习线性模型建模设计, 利用真实金融数据, 设计线性机器学习模型
- 第 9-10 周: 集成学习, 模型框架, 实战金融模型文献学习
- 第 11-15 周: 深度学习模型、CNN 模型、大语言模型
- 第 16-17 周 (个人期末考试): 金融学机器学习建模, 量化模型展示、改进、讨论与点评

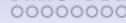
课程安排



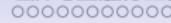
什么是金融人工智能模型



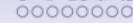
机器学习模型的优势



经济与金融行业就业需求



学习与实战资源



## 课程安排

## 什么是金融人工智能模型

## 机器学习模型的优势

## 经济与金融行业就业需求

## 学习与实战资源

# 我们作为经济学本科生为什么要学习人工智能模型？

- 我们作为经济学本科生为什么要学习机器学习等量化方法？
- 学会这些经济学量化方法对我们未来的工作有什么帮助？
- 该怎样入门学习？

# 机器学习模型 (Machine learning model is dominating in various areas )

机器学习模型 (Machine Learning) 是人工智能领域的算法支持与核心。因为机器学习模型的强大预测能力与准确程度，近些年来被用于生活中各种领域。比如：

1. 文本分析 (ChatGPT、Bert)
2. 图像识别 (人脸识别)
3. 语音识别 (语音输入发)
4. 推荐算法 (淘宝、京东、Amazon)
5. 医药 (大大加快新冠疫苗开发进程) 等等当然
6. 我们的**金融学与经济学**中的很多分析与预测，也在近几年逐渐被机器学习算法所占领。机器学习算法在精准程度与效率上远超人工分析师。

# 为什么近些年机器学习成为金融经济市场的宠儿

So, why now machine learning in economic & financial market?

- 高度竞争的经济金融市场需求
- 机器学习模型不但可以准确分析价值，还可以分析风险
- 市场风险加大，各大企业需要对投资与风险进行准确分析  
(如：房地产公司的投资)

# AI vs Machine Learning vs Deep Learning

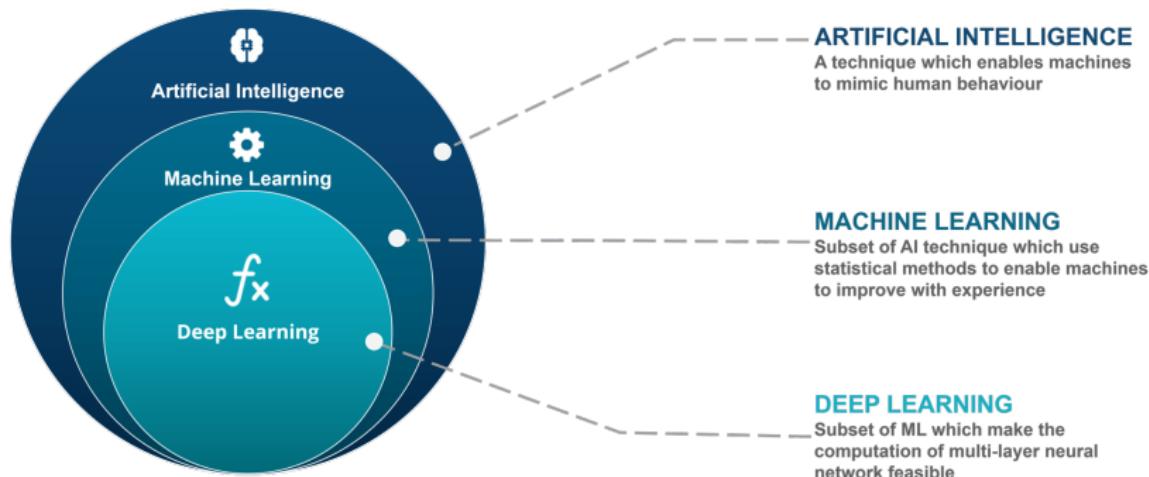


图: AI vs Machine Learning vs Deep Learning

source: <https://www.edureka.co/>

# 机器学习在经济学上的常用算法

在经济学数量建模上常用的 Machine Learning 算法主要有：

1. **线性算法**: OLS, 2SLS, Logit 等。重要，是我们经济学的基本功，也被称为 BLUE, Best Linear Unbiased Estimation, 是作为无偏因果分析的重要工具。作为经济学专业的同学们，学好计量经济学与统计学是非常重要的，也是其他机器学习模型学习的重要基础。其他的常用线性算法：LASSO, Ridge, Elastic, MARS (Multivariate adaptive regression spline)
2. **集成学习 (Ensemble Learning) :**  
Xgboost, Catboost, Adaboost, Random Forest
3. **深度学习 (Deep Learning) :** Artificial Neural Network(ANN), CNN (Convolutional Neural Network), RNN (Recurrent Neural Network)
4. **其他算法:** Bayesian Estimation, KNN, K-means, PCA, Causal Forest (Susan Athey), Double Machine Learning (Victor Chernozhukov) 等等

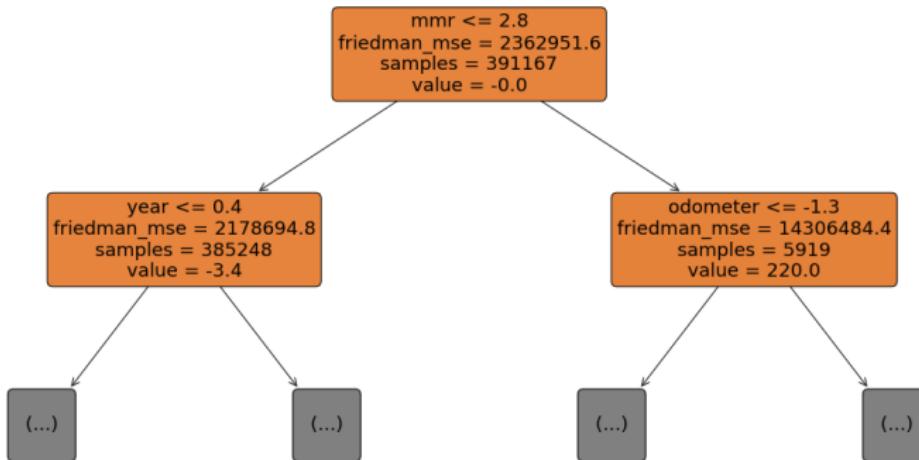
# Machine learning is not a black box

- 在 python 环境下，构建机器学习模型非常简单与直白
- 比如，TensorFlow 将深度学习模型的构建简单化，Keras 有进一步将 TensorFlow 简单化
- Xgboost, Catboost, Adaboost 将集成学习简单化
- 下面我们展示两个简化的机器学习模型（为了学习需要模型结构和参数都为最简单设定）

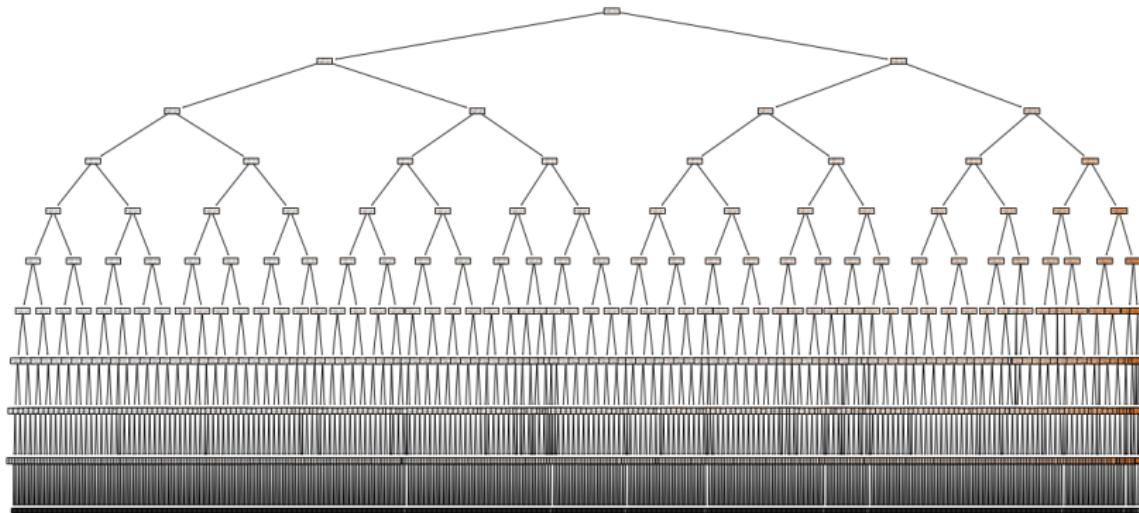
# 集成学习 (Gradient Boosting 模型构建 + 预测)

```
hyperparameters_gb={'n_estimators':200,'learning_rate':0.1,  
reg_gb=GradientBoostingRegressor(**hyperparameters_gb)  
reg_gb.fit(xtrain_ann,ytrain_ann)  
ypred_gb = reg_gb.predict(xtest_ann)
```

# 集成学习 (Gradient Boosting 框架)



# 集成学习 (Gradient Boosting 框架)



# 深度学习 (ANN 模型构建)

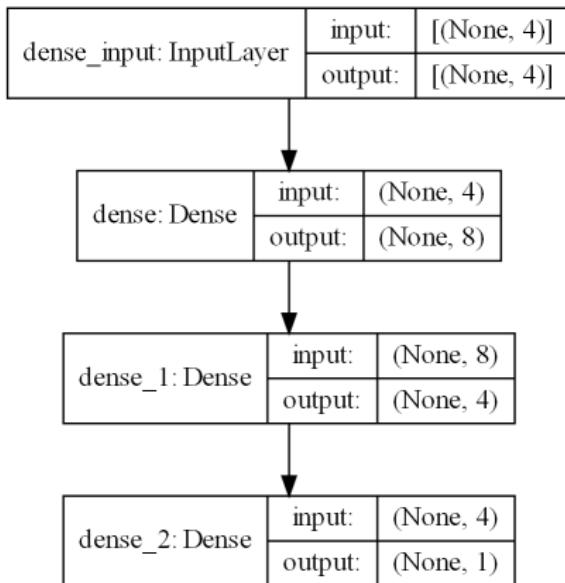
```
K.clear_session()
epochs=10
batch_size=128
model_ann = Sequential()
optimizer = keras.optimizers.Adam(lr=0.001)
model_ann.add(Dense(8,activation = 'relu',
                    input_dim = len(xtrain_ann.columns)))
model_ann.add(Dense(4,activation = 'relu'))
model_ann.add(Dense(1,activation = 'linear'))
model_ann.compile(optimizer = optimizer,loss = 'mse',metrics = ['mae'])
model_ann.summary()
```

# 深度学习（ANN 模型训练与预测）

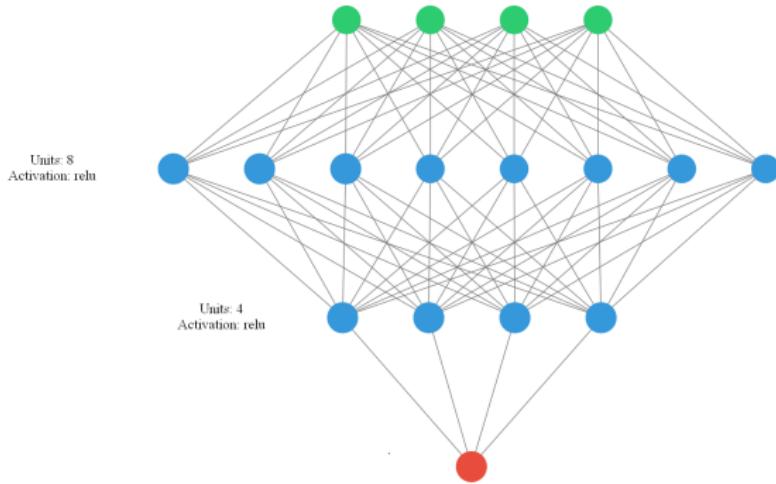
```
%%time
with tf.device('/gpu:0'):
    model_ann.fit(xtrain_ann,ytrain_ann,
                   batch_size=batch_size,epochs=epochs,
                   verbose=0)
ytest_ann = model_ann.predict(xtest_ann,batch_size = 32)
```

Wall time: 1min 29s

# 深度学习 (ANN 模型框架)



# 深度学习 (ANN 模型框架)



# 本地部署大模型

```
from transformers import AutoTokenizer, AutoModelForCausalLM

# Load the tokenizer and model
tokenizer = AutoTokenizer.from_pretrained("gpt2")
gpt_model = AutoModelForCausalLM.from_pretrained("gpt2")
# Example input text
input_text = "what is python"
# input_text = "Python example of using bert for word embedding"

# Tokenize the input text
inputs = tokenizer(input_text, return_tensors="pt")

# Generate text
outputs = gpt_model.generate(**inputs, max_new_tokens=50)

# Decode the generated text
generated_text = tokenizer.decode(outputs[0], skip_special_tokens=True)
print(generated_text)

Setting `pad_token_id` to `eos_token_id`:None for open-end generation.
what is python?

Python is a programming language that is used to create and manipulate data.
```

## Short history of machine learning

1. 线性模型的创立: Legendre (1805) and Gauss (1809) used the linear regression for the prediction of planetary movement.
2. 深度学习的雏形: Warren McCulloch and Walter Pitts (1943) opened the subject by creating a computational model for neural networks.

source: History of the AI from Wikipedia

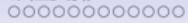
# 机器学习算法战胜人类的案例

1. AlphaGo (微软): AlphaGo research project was formed around 2014 and beat world champion Lee Sedol 9th Dan in 2016.
2. Github Copilot (微软): 人工智能算法按照要求自动写代码, 完成初级程序员的工作。  
<https://copilot.github.com/> (点击链接注册可用)
3. AlphaCode (谷歌): 人工智能算法按照要求自动生成算法, 取代高级程序员的工作。AlphaCode achieved an estimated rank within the top 54% of participants in CodeForces.  
<https://alphacode.deepmind.com>
4. ChatGPT (微软) : 人工智能算法按照要求自动生成文章、算法, 取代金融分析、经济分析、程序员
5. 未来: 人工智能算法 → 人工智能算法 → 人工智能算法  
→...

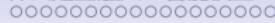
# 那什么是人工智能取代不了的哪？

- 独立思考的智慧 (Independent Thinking)
- 创造能力 (Creativity)
- 人工智能时代的竞争，核心是有独立思考精神与创造性精神的年轻人才的竞争（就像在座的各位青年学子）

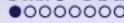
课程安排



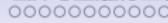
什么是金融人工智能模型



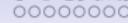
机器学习模型的优势



经济与金融行业就业需求



学习与实战资源



## 课程安排

## 什么是金融人工智能模型

## 机器学习模型的优势

## 经济与金融行业就业需求

## 学习与实战资源

# 什么是学习：人类社会的学习过程 (Human Society in Learning )

我们可以对比机器学习与人类的学习：

1. 学习并不是一个新的概念
2. 人类与人类社会的演进也是一个学习的过程
3. 人类学习 = 尝试 + 改进 + 不断重复

# 人类的学习是一个缓慢积累的过程 (Human Learning is a long and slow process)

人类从历史学到的唯一的教训，就是人类没有从历史中吸取任何教训。 (We learn from history that we do not learn from history.)

——黑格尔



图: Georg Wilhelm Friedrich Hegel

# 人类的学习是一个缓慢积累的过程 (Human Learning is a long and slow process)

人类历史上漫长与痛苦的学习过程：

1. 商有九世之乱：用了整整九个王，学习到了父死子继比兄终弟及更利于封建王朝的稳定统治 (中丁、外壬、河亶甲、祖乙、祖辛、沃甲、祖丁、南庚、阳甲)
2. 斯图尔特王朝：用了整整一百年的时间学习并从封建集权制过渡到资本主义议会制度 (House of Stuart, 1603 年至 1714 年，詹姆斯一世，查理一世，护国主克伦威尔...)
3. 高考刷题：五年高考三年模拟，高中三年不断尝试与改进的过程

# 人类学习的问题

- 用机器学习算法的角度总结黑格尔的话，就是人类社会的学习效率太低太慢。也就是 Learning rate ( $\eta$ ) is too low。
- 注：Learning Rate 是机器学习模型中的一个重要的超参数 (Hyperparameters)，控制着机器学习算法的学习速度

## 机器学习相对人类学习的优势

1. 机器学习用算法高度模拟人类学习的过程
2. 和人类一样，机器学习算法的秘诀也是：尝试 + 改进 + 不断重复
3. 人类和机器在学习上的差别就是在不断重复与不断优化的效率上。人类重复一次学习的过程需要几个月甚至几年。但是机器只需要不到一秒的时间。
4. 人类资深金融分析师通过十几年的工作经验与几千个工作的案例不断重复总结经验，从而对金融资产进行准确定价。可机器学习算法只需要几个小时就可以轻松分析几十万个甚至上百上千万个资产。机器学习算法通过对大数据的分析，优化自身的评估预测能力。所以机器学习算法远远超出了人类分析师日积月累的经验，其精准度与效率也远远超出同一领域的人工分析师。

# How machine learning from the data

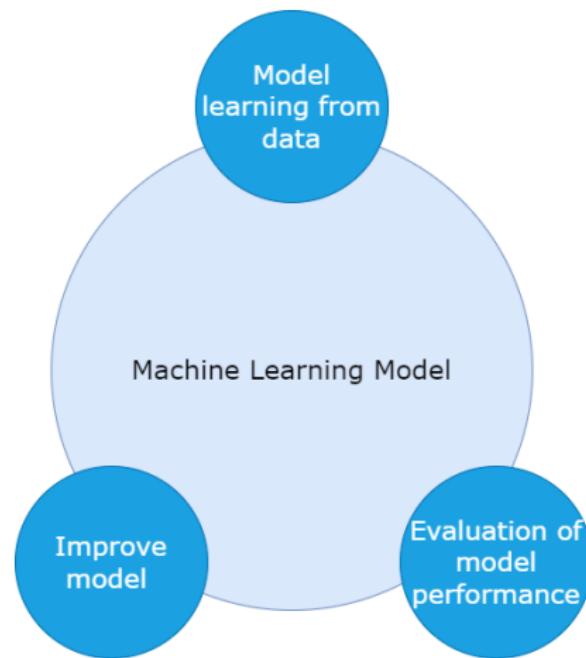


图: How machine learning from the data

# 机器学习模型在经济学的应用

1. 机器学习模型在经济金融学的应用又量化建模 (quantitative modeling) 简称 Quant, 广泛的应用于包括股市投资, 高频交易, 资产管理, 风险管理, 资产估值, 保险, 反欺诈等各个行业。
2. 很多我们经济学的同学, 都有志利用自己多年所学的经济学与金融学知识成为一个量化建模师 (Quant Researcher, Quant Economist, Data Scientist and etc.)
3. 作为一个经济学人才, 利用 Python 建立量化模型的基本功非常重要

课程安排

oooooooooooo

什么是金融人工智能模型

oooooooooooooooooooo

机器学习模型的优势

ooooooo

经济与金融行业就业需求

●oooooooooooo

学习与实战资源

ooooooo

## 课程安排

## 什么是金融人工智能模型

## 机器学习模型的优势

## 经济与金融行业就业需求

## 学习与实战资源

# 为什么说量化模型对于经济学同学是基本功？

1. 作为经济学从业者，简单的分析已经不能满足行业的需求（例如：客户不但要知道价格为什么会上升，还要知道为什么？上涨多少？）
2. 经济学量化模型（包括其中的机器学习，深度学习模型）其速度远远高于人工分析（例子：房地产价格分析、股票基本面分析、价格预测、风险预测、分析报告生成）
3. 优秀的量化模型（包括其中的机器学习，深度学习模型）的预测精度不逊于，甚至超过人工分析

⇒ 量化模型是未来经济学从业者的基本功

# 经济与金融就业市场的需求

最能反映市场趋势的就是看市场需要什么的人才。经济金融行业对量化分析师与研究员有着很高的需求。我们在 Glassdoor 上面去搜索一下量化建模所需要的人才与所具有的要求。



## Get notified of new jobs

Tired of searching for jobs? Create a job alert to see the freshest jobs daily

[Create Job Alert](#)

## Suggested Searches

quant  
Townshipin Cities Amf, MN (US)



## Get notified of new jobs

Tired of searching for jobs? Create a job alert to see the freshest jobs daily

[Create Job Alert](#)

## Suggested Searches

quant  
Townshipin Cities Amf, MN (US)

# 量化人才需求案例 1, 2



Job Company Rating Why V

## QUALIFICATIONS

7-10 years of experience in the financial services industry in a quantitative field, preferably with experience in model development/review, risk modeling and portfolio optimizatio

At least Master's Degree in a technical field such as Mathematics, Statistics, Econometrics or Operations Research

Existing experience in leading conversations in Firm Risk Committees as well as with Model risk management function is preferable

Programming skills in statistical packages such as R, python or Matlab and familiarity with database systems such as Sybase, MS SQL

Familiarity with vendor risk systems such as RiskMetrics, BlackRock Aladdin, MSCI/



Job Company Rating

## What we're looking for:

- Bachelor's Degree in quantitative discipline e.g. Finance, Mathematics/ Statistics or Economics
- Experience in statistical model development
- Knowledge of data analysis, theory and statistical techniques
- Proficiency with analytical software Python, R, SQL tools e.g., Oracle, Unix platforms, and Microsoft Office

## Skills that will help you in the role:

- Graduate studies, Masters or PhD in quantitative discipline
- Analytical work experience in a financial services company and strong technical and / or statistical skills with proven ability to process large datasets into meaningful

# 量化人才需求案例 3,4



Job

Company

Rating

## Required Skills

- Strong mathematical aptitude
- Fluency in statistical methods and modeling
- Coding skills in Python (C++ familiarity is a plus)
- Exposure and familiarity with machine learning techniques
- Values teamwork but capable of thinking independently
- Works on own Initiative / hustle – takes a pragmatic approach
- Effective communication with all levels of professional experience
- Ability to retain information and then teach others what you have learned

## Required Experience

• Commitment and interest in options and



Job

Company

Rating

## Required:

- A bachelors or higher degree in computer science, machine learning, statistics, math, economics, business or other scientific or quant-focused field

- Programming skills (esp. related to data technologies like Python, Java, C#, etc.)
- 2 or more years of experience using data/ML/AI to impact critical product or business decisions

## Preferred:

- Experience with hypothesis testing, graph theory and experiment design
- A proven track record of collaborating across organizational boundaries and

# 量化人才需求案例 5,6

1:19 4 \* • 73%  
 ← Quant Research and Model ...  
 Morgan Stanley • New York, NY

Job Company Rating Why V

## QUALIFICATIONS

7-10 years of experience in the financial services industry in a quantitative field, preferably with experience in model development/review, risk modeling and portfolio optimizatio

At least Master's Degree in a technical field such as Mathematics, Statistics, Econometrics or Operations Research

Existing experience in leading conversations in Firm Risk Committees as well as with Model risk management function is preferable

Programming skills in statistical packages such as R, python or Matlab and familiarity with database systems such as Sybase, MS SQL

Familiarity with vendor risk systems such as RiskMetrics, BlackRock Aladdin, MSCI/  
 Barron Yield Book, Barclay's POINT, SunGard

1:16 4 \* • 73%  
 ← Quant Analytics Associate  
 KeyCorp • Cleveland, OH

Job Company Rating Why V

## REQUIRED QUALIFICATIONS

- Bachelor's degree (or its equivalent) in statistics, mathematics, economics, financial engineering, data sciences, predictive modeling, or other quantitative disciplines and at least 1 year of relevant experience; 0 with Master's or PhD

## DATA LITERACY

- Understanding of:  
 Data wrangling including information documentation and importing data from different formats
- Data wrangling including information documentation and importing data from different formats

- Descriptive statistics, random variables, common distributions, outliers

## 量化建模所需人才的关键字

我们通过上述量化岗位的人才需求信息，可以得出两个跟我们密切相关的关键词：1) 经济学与 2) python。

- 经济学 (Economics): 在现如今的金融市场中，经济学常常被归类为量化学科 (STEM)，要求毕业生有很好的数量统计与分析能力。并且具有很好的编程能力。就像上述大公司的招聘广告要求的一样，经济学学生的数理统计能力需要和统计学与数学的学生比肩。并且拥有很好的编程能力。
- Python: python 是当今量化建模所使用的最热门编程语言。

# 经济学量化建模所需的知识库

1. **经济学 Economics:** economics is a social science concerned with the production, distribution, and consumption of goods and services
2. **量化建模 Quantitative:** using econometrics & machine learning skills to build model to explain and solve the economic problems
3. **Python:** trending computer language in building the quantitative models

# 经济学人工智能建模：三种知识的有机结合

这门课的内容不仅仅是 Python:

- 会用 Python  $\neq$  会应用经济学预测
- 会应用经济学预测  $\neq$  会应用量化建模
- 经济学人工智能建模 = Python + 经济学 + 机器学习模型

# 业界量化模型师的具体分类与工作流程

量化模型的构建并不简单，有着严格的模型精准度与模型质量管  
理流程。模型的质量管理流程比算法本身更重要。

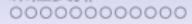
1. 一线建模型研究员 (1st line modeler)：沟通 Business 获取需求，沟通数据组获得数据资源，通过数据与 Business 需求量化建模，沟通 developer，将模型应用于 business
2. 二线建模型研究员 (Model validating)：负责检查一线模型的问题，提出模型的不足与提升意见
3. 三线建模型研究员 (Model auditing)：负责检查一线模型与二线模型的问题，检查模型合规与法律问题

## 学习经济学量化建模的人群

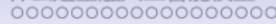
针对人群：

1. 有志于在业界从事量化研究的同学（未来的量化经济学家，数据科学家，etc.）
2. 有志于在学界从事研究的同学（经济学，金融学 PhD, 经济学，金融学，金融工程 Master）
3. 有志于研究 Policy Making 的同学们
4. 甚至是有志于创业当老板的同学们

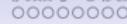
课程安排



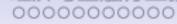
什么是金融人工智能模型



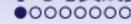
机器学习模型的优势



经济与金融行业就业需求



学习与实战资源



## 课程安排

## 什么是金融人工智能模型

## 机器学习模型的优势

## 经济与金融行业就业需求

## 学习与实战资源

# 提升能力：读模型说明书（White Paper）与论文（Paper）

- 模型的说明文件可以帮助同学们快速熟练使用模型（如：<https://xgboost.readthedocs.io/en/stable/tutorials/model.html offline>）。渠道：模型的官网 or github。
- 论文可以去更深层次的去了解模型的内部结构（如：<https://arxiv.org/abs/1603.02754>）。渠道：google scholar, Arxiv

# 遇到不会的问题怎么办 (Coding Questions)

- Search Engines such as Bing and StackOverflow
- help function in Python
- AI such as ChatGPT, Gemini, Deepseek, Kimi and etc

# 遇到不会的问题怎么办 (For Example)

Step1: Use **keywords** to search your coding questions in KIMI (or ChatGPT)

In the screenshot, a user has asked the AI "how to use python pandas to do the summation by groups". The AI, identified as "Kimi", has provided a detailed step-by-step guide:

In Python, the Pandas library provides a powerful way to perform group-wise summation using the `groupby` method. Here's a step-by-step guide on how to use Pandas to sum values by groups:

Step-by-Step Guide

1. Import Pandas

First, you need to import the Pandas library. If you haven't installed Pandas yet, you can install it using `pip install pandas`.

# 遇到不会的问题怎么办 (For Example)

Step2: study line by line the codes from AI

## 3. Group by a Column and Sum

Use the `groupby` method to group the data by a specific column (e.g., 'Category') and then use the `sum` method to calculate the sum for each group.

python

```
grouped_sum = df.groupby('Category')['Value'].sum()  
print(grouped_sum)
```

Output:

Category

# 数据资源 1: Scikit-learn 自带数据

优点：方便，一行代码读取 缺点：数据量小，不利于真实建模，

仅仅用于简单演示

## sklearn.datasets: Datasets

The `sklearn.datasets` module includes utilities to load datasets, including methods to load and fetch popular reference datasets. It also features some artificial data generators.

**User guide:** See the [Dataset loading utilities](#) section for further details.

### Loaders

<code>datasets.clear_data_home([data_home])</code>	Delete all the content of the data home cache.
<code>datasets.dump_svmlight_file(X, y, f, *, [ ...])</code>	Dump the dataset in svmlight / libsvm file format.
<code>datasets.fetch_20newsgroups(*[ ..., data_home, ...])</code>	Load the filenames and data from the 20 newsgroups dataset (classification).
<code>datasets.fetch_20newsgroups_vectorized(*[ ...])</code>	Load and vectorize the 20 newsgroups dataset (classification).
<code>datasets.fetch_california_housing(*[ ...])</code>	Load the California housing dataset (regression).
<code>datasets.fetch_covtype(*[ ..., data_home, ...])</code>	Load the covertype dataset (classification).
<code>datasets.fetch_kddcup99(*[ ..., subset, ...])</code>	Load the kddcup99 dataset (classification).
<code>datasets.fetch_lfw_pairs(*[ ..., subset, ...])</code>	Load the Labeled Faces in the Wild (LFW) pairs dataset (classification).
<code>datasets.fetch_lfw_people(*[ ..., data_home, ...])</code>	Load the Labeled Faces in the Wild (LFW) people dataset (classification).
<code>datasets.fetch_olivetti_faces(*[ ...])</code>	Load the Olivetti faces data-set from AT&T (classification).
<code>datasets.fetch_openml([name, version, ...])</code>	Fetch dataset from openml by name or dataset id.
<code>datasets.fetch_rcv1(*[ ..., data_home, subset, ...])</code>	Load the RCV1 multilabel dataset (classification).
<code>datasets.fetch_species_distributions(*[ ...])</code>	Loader for species distribution dataset from Phillips et.
<code>datasets.get_data_home([data_home])</code>	Return the path of the scikit-learn data directory.
<code>datasets.load_boston(*[ ..., return_X_y])</code>	DEPRECATED: <code>load_boston</code> is deprecated in 1.0 and will be removed in 1.2.
<code>datasets.load_breast_cancer(*[ ...])</code>	Load and return the breast cancer wisconsin dataset (classification).

## 数据资源 2: Kaggle Datasets

- Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners.
- 网址: <https://www.kaggle.com/datasets offline>
- 优点: 真实大数据、数据比较新、与经济与金融学相关数据比较多

The screenshot shows the Kaggle Datasets homepage. On the left, there's a sidebar with navigation links: Create, Home, Competitions, Datasets, Code, Discussions, Courses, More, and Your Work. The main content area has a search bar at the top. Below it, a section titled "Datasets" is described as "Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating." It features a "New Dataset" button and a "Your Work" link. To the right is a "Trending Datasets" section with a search bar and a list of four datasets:

- New & Ancient Temples - India
- maroc avito car dataset
- Heart Attack Treatment Payments By Hospital
- Mushrooms

Each dataset card includes a thumbnail, the name, the creator, the last update time, file count, size, and a CSV icon. Below this is a "Popular Datasets" section showing thumbnails of various datasets.

## 数据资源 3: Huggingface

- Hugging Face is an AI platform and community that provides a vast collection of pre-trained models, datasets, and tools for natural language processing (NLP) and other machine learning tasks.
- 网址: <https://huggingface.co/>
- 优点: 最新人工智能模型、数据 (开眼看世界)

The screenshot shows the Hugging Face website homepage. At the top, there is a navigation bar with a search bar labeled "Search models, dataset", and links for "Models", "Datasets", "Spaces", "Docs", "Enterprise", and "Pricing". On the far right, there is a user profile icon. Below the navigation bar, there are three main sections: "Following" (3 items), "Trending" (last 7 days), and a sidebar for the user "jedibear". The "Following" section shows a "Welcome" message from "Hyperbolic, Nebius AI Studio, and N...". The "Trending" section shows three items: "deepseek-ai/DeepSeek-R1", "perplexity-ai/r1-1776", and "microsoft/OmniParser-v2.0". The sidebar for "jedibear" includes links for "Profile", "Inbox (0)", "Settings", "Billing", "Get Pro", and "Create New".

## Why we use Jupyter lab for Python

- 通过 Jupyter 安装 Python:

<https://www.anaconda.com/products/individual>

<https://jupyter.org/try>

- Jupyter 是最流行的 Python 操作环境，大的金融，科技企业，政府机构都在用这个操作环境
- 方便大的团队与项目合作（一线 Quant 团队，二线 Quant 团队，EE 团队与 Business 团队）
- 方便介入云计算（企业级：AWS，MS Azure and Data Brick，个人使用：Google Colab  
<https://colab.research.google.com/>）

# Jupyter

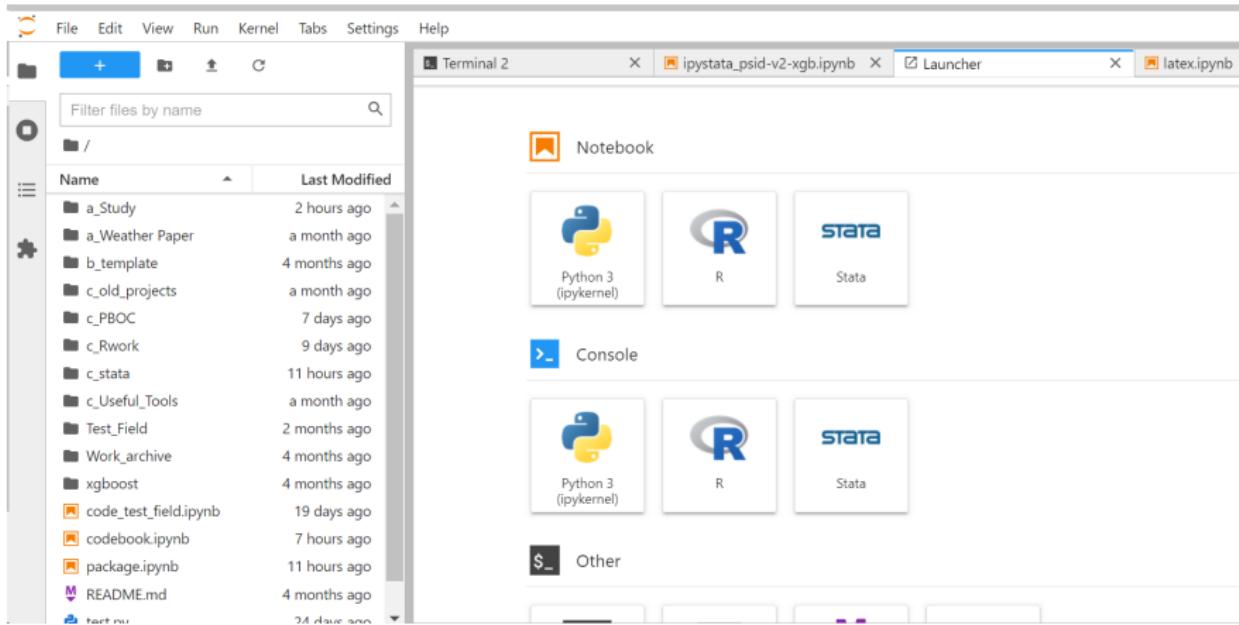


图: Jupyter 操作平台

# 安装机器学习建模所需的 python packages

安装 xgboost、scikit-learn、pandas、Tensorflow

The screenshot shows the PyPI project page for xgboost 1.6.1. The top navigation bar includes links for Help, Sponsors, Log in, and Register. A search bar is at the top left. The main title is "xgboost 1.6.1". Below it is a command line interface (CLI) snippet: "pip install xgboost". To the right, there's a green button labeled "Latest version" with a checkmark. Below the button, the text "Released: May 9, 2022" is visible. The page is titled "XGBoost Python Package". On the left, a sidebar titled "Navigation" lists "Project description" (which is currently selected and highlighted in blue), "Release history", and "Download files". On the right, the "Project description" section includes a "Project description" heading, a "pip package 1.6.1" badge, and an "Installation" section with a "From PyPI" link. At the bottom, there's a "Project links" section with a "Homepage" link.

# 不负平生所学

各位同学过去十几年勤学苦练的技能得到发挥。比如：英文能力、计量经济学能力、经济学能力、编程能力、更重要的是创造力与智慧。

1. **Math skills**
2. **English reading skills**
3. **Coding skills**
4. **Economics skills**
5. **Young talents**

# 相关文献

- Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273.
- Leippold, M., Wang, Q., & Zhou, W. (2021). Machine learning in the Chinese stock market. *Journal of Financial Economics*.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261-65.
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

# 经济量化模型研究团队（分组）

- Here we assign the groups
- 3-4 人一组
- 每组 project lead 一名
- 作业与项目以组为单位

# 团队作业 A: Question 1: 安装 Anaconda 与 Python

- 如果遇到问题团队协助解决
- 难点: windows 与苹果系统不同, 安装略有不同。遇到问题正常。
- 记录汇总问题一份报告。报告包括遇到的问题与解决方案。下节课讨论与评分。
- 不计费但是课堂提问

## Question 2: 计算 $1+1$

每个人计算  $1+1$ , and save the ipynb file

The screenshot shows a Microsoft Edge browser window with a JupyterLab interface. The left sidebar displays a file tree with the following contents:

Name	Last Modified
a_Weather_Paper	4 months ago
c_old_projects	10 months ago
c_PBOC	7 months ago
c_Rwork	9 months ago
c_Stata	3 months ago
c_Useful_Tools	4 months ago
old	4 months ago
Test_Field	4 months ago
xgboost	10 months ago
check_packages.ipynb	3 months ago
codebook.doc	5 months ago
codebook.ipynb	3 months ago
CookBook_Metrics.ipynb	4 months ago
package.ipynb	4 months ago
README.md	10 months ago
test.ipynb	3 months ago

The main area of the browser shows a Jupyter notebook titled "Untitled.ipynb". The notebook contains one cell with the code:

```
[1]: print(1+1)
```

The output of the cell is:

```
2
```

The status bar at the bottom right of the browser window indicates "Python 3 (ipykernel)".

## Question 3: test on numpy random seed , and save the ipynb file

```
import numpy as np  
  
np.random.seed(123)  
np.random.normal()
```

## Question 4: 了解 pandas, numpy, sklearn, matplotlib 是干什么的

- 预习：  
[https://python-programming.quantecon.org/about\\_py.html](https://python-programming.quantecon.org/about_py.html)
- 了解：pandas, numpy, sklearn, matplotlib 的用途。下节课 cold call。

## Question 5: 注册 Github 账号

- 了解 Git 与 Github, 注册 git 账号 (<https://github.com/>)
- 我们将使用 Git 来提交作业、考试、与小组协作  
<https://github.com/>