AI & ML for Data Scientists

Class 4: **Real** Big Financial Data

Lei Ge (葛雷)

Quant RUC (人大量化)

October 5, 2025

Real Financial Data

- What is the most important elements for Machine Learning? **Data**
- What makes the ML in finance unique? (we financial data)
- Why real data?

First look a the fake data

- sklearn.datasets is a good source for TOY data
- Good source for practice
- Only issue is that fake data is fake
- Lets check out why (Please follow to blank Ipynb)

CSMAR

CSMAR, short for China Stock Market & Accounting Research Database, is a comprehensive research-oriented database focusing on China Finance and Economy. CSMAR was developed by Shenzhen CSMAR Data Technology Co., Ltd based on academic research needs, meeting with the international professional standards while adapting to China’ s features.

CSMAR

- professional level financial data for stock & company study
- used by both financial companies and financial researchers

CSMAR: easy to use

- Easy to use especially for Python users
- We can use both UI and API (what is UA and API?)
- its check it with me step by step and login from lib

WRDS

USA counterpart of CSMAR [1]

[1] CSMAR followed WRDS's business model

Kaggle

- Kaggle, a subsidiary of Google LLC
- Heavely platform for Quant Research (us)
- Codes, data, competition and more
- Let check it out! (Kaggle)

Kaggle

- Kaggle is most important data source for now
- You can search and find your interested research topics
- Let check it out! (Kaggle)

Data from the internet

1. Internet has valuable data for the financial predictions
2. Internet data low quality? No
3. Selenium is a powerful and popular tool

But how to use?

- I will guide you to study this package
- but next time you should know how to learn any package by yourself

But how to use?

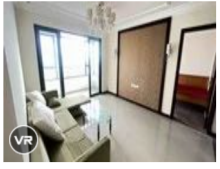
- Template + Documentation + CHATGPT + BING
- Template (from search bing and from CSDN, StackOverFlow, CHATGPT)
- Unknown knowledge $\rightarrow$ Bing + Documentation + ChatGPT

Please follow me to the selenium codes

CH1_Class4_Quant_Stock_Information.ipynb

Homework3-1: Data Mining

- Housing Price Data from https://esf.fang.com/
- Housing Rent Data from https://zu.fang.com/
- Data needed: listed below



A区双卫户型,业主急用钱!捡漏的来!!!

3室2厅 | 118㎡ | 顶层（共26层） | 南北向 | 2017年建 | 杨婷玉

京北恒大国际文化城 怀来 八达岭高速沙城东出口东约3公里

满五

45万
3813元/㎡

Homework3-2: Data Mining

默认排序 按发布时间排序 价格 面积 合并相似

国际村性价比正规三居室,南北通透,高层采光视野好,价

整租 | 3室2厅 | 139㎡ | 朝南北

朝阳-西坝河-UHN国际村

距10号线太阳宫站约585米。

交通便利 南北通透 采光好

14500 元/月

Homework3-2: Data Mining (Group)

- Team 1 北京-海淀 I：苏州桥、万柳、北太平庄、世纪城
- Team 2 北京-海淀 II：西三旗、清河、西二旗、上地
- Team 3 河北（京北）：怀来、下花园、张北、桥西
- Team 4 河北-廊坊+北京-通州：大厂、燕郊、马驹桥、亦庄
- Team 5 北京-昌平: 沙河、霍营、回龙观、天通苑
- Team 6 天津：中新生态城（滨海新区）、武清、劝业场（和平）、八里台（南开）
- Team 7 重庆-渝北：（Please choose your own blocks）

- Each person only in charge of **one block** and only get first 20 pages if too many for you

Homework3-3: Data Research (Your Own)

- Collect Data from your teammates and merge the data (please feedback to TA if someone no response, so we can help both team and other student)
- Data description of your data and whether data has outliers
- Then get housing price per m2 and housing rent per m2 ($price/m2$ and $rent/m2$) for each block
- 1) data description for each block 2) Calculate median price to rent ratio for each block
- Figure A: Bar Plot the median price to rent ratio for each block (The global fair value should around 200)

Homework3-4: Data Science Modeling

- Model 1 $price/m2_i = \beta_0 m2_i + \beta_2 location_i + \epsilon_i$
- Model 2 $rent/m2_i = \beta_0 m2_i + \beta_2 location_i + \epsilon_i$
- Use model 1 and model 2 to predict price and rent for the $m2 = 50$, $m2 = 100$
- Figure B and C: Bar Plot the price to rent ratio for each block for the $m2 = 50$, $m2 = 100$

Homework3-5: Data Science Modeling Pro Max

- Add features non-linearity and interaction to Model 1 and Model 2, then get Model 1+ and Model 2+. Compare with R2 of Model 1, Model 2 vs Model 1+ , Model 2+. Which one has higher R2 and why?
- Use model 1+ and model 2+ to predict price and rent for the $m2 = 50$, $m2 = 100$
- Figure E and F: Bar Plot the price to rent ratio for each block for the $m2 = 50$, $m2 = 100$. Compare the bar plots from these three methods. Which one should you trust?
- Submission: only Ipynb codes to your personal folder (NO DATA PLEASE, Git is for codes not for data)