4 - 13/7/2021

INVESTING BOT

Introduction

Given more than 400 stocks on HOSE market, the goal is to find a handful of stocks that performs the best in a long-term run (consistent growth)

Metrics used in assessment are:

- PE
- EPS
- ROA
- ROE
- ROS

from 2009 to 2020 (12 years)

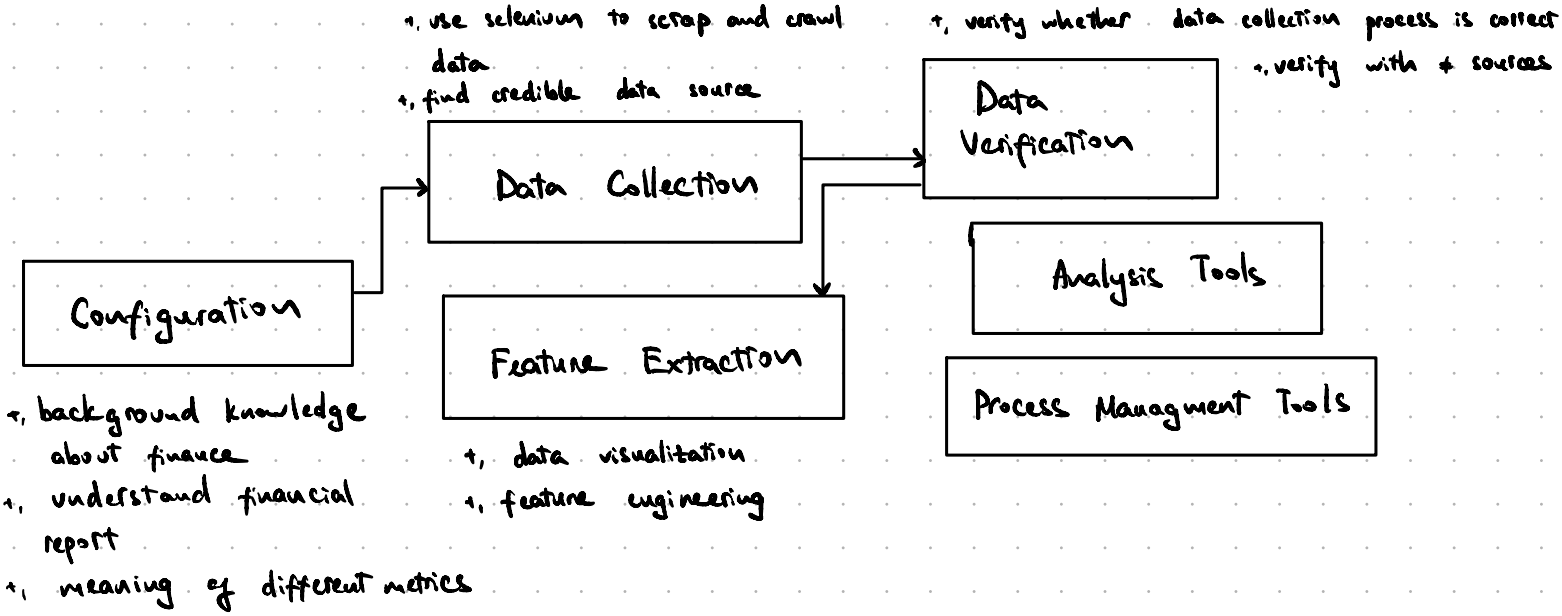
Another point that should be taken into account is that the goal of the project is to **INVEST** not **TRADE**. Therefore, long-term growth is taken into account, which means daily trading data can be neglected.

Additionally, assume that all 'great' companies have some common trend in statistics (eg: low PE, high EPS, sustainable growth in EPS,...) => the project's aim is to look for that common trend.

Also, it is unreasonable label each stock whether it is a good investment or not (it could be, but in that case, the algorithm would be subjective and only stocks with familiar trend will be chosen). Plus, labelling 400 stocks is a tedious job.

With all being said, this is an unsupervised learning problem.

Pipeline



Data Collection

I have always been using 'FireAnt' and 'vietstock finance' to look for stock data, so when it comes to data collection, those were the first two pages that I thought of. Vietstock has a 'get data' feature which export data into Excel file. However, a subscription is needed to use this feature, or else only data from last year can be downloaded. For the project, we want to use data from quite a number of years → vietstock not feasible.
Alternatively, FireAnt does not provide all metric information from finacial report.
Also, vietstock has data available to be seen, but not be able to be downloaded.
Therefore, the next thing that I think of is to use web scraping and crawling to obtain these information.

While looking at how I can scrap info from 'vietstock', I have realised that 'vietstock''s web is actually pretty hard to interact with. I tried another website called 'CafeF', which has the same feature as vietstock but its interface was more easily interacable. Therefore, I would use this website to scrap data.

Another significant point I realised during reading metrics' meanings was that when assessing how a stock is doing, it should be compared to those who are in the same industry. Hence, data of stocks and their corresponding field are collected and stocks will be compared within a field only.

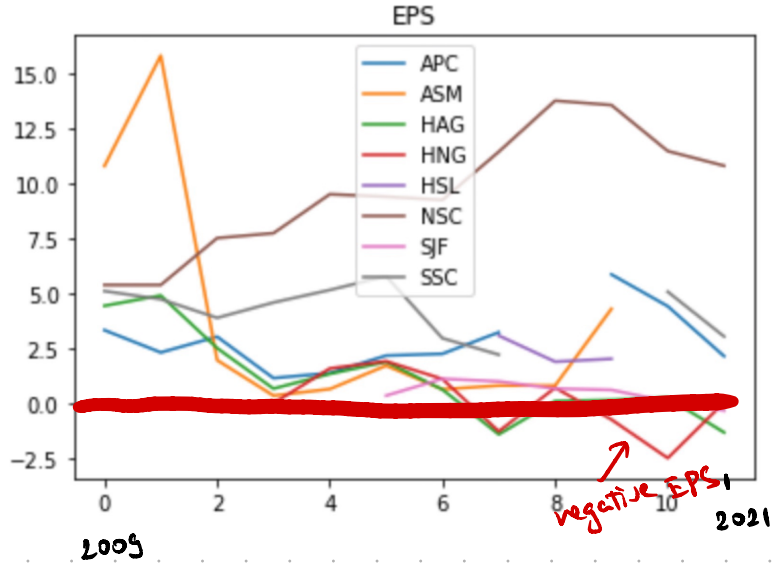
Scraping and crawling for over 400 stocks takes about 12.000s ≈ 200 mins = 3hours 20 mins to complete. Notice that stocks from finance industry have a very different interface from the others. Hence could not fetch using this method. In fact, this companies' metric is not visible on CafeF in the web interface (no idea about the reason behind this, or maybe it was hidden somewhere else, idk)

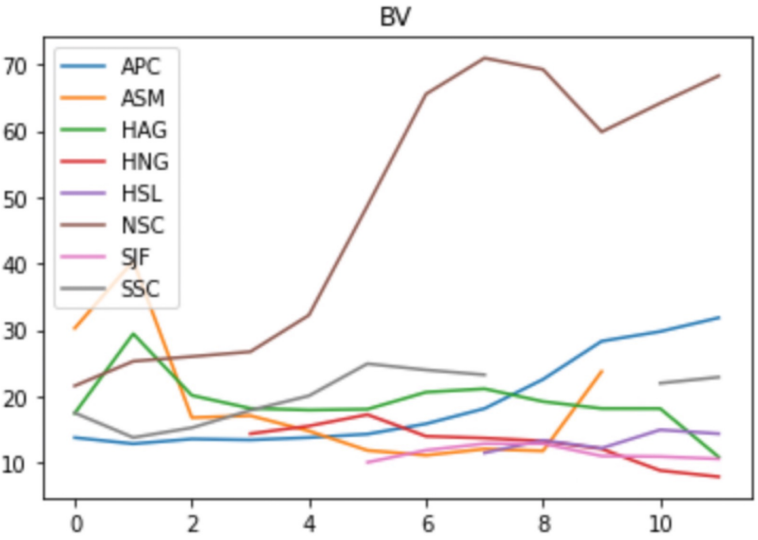
Results:
- all stocks in HOSE market was successfully collected (except for those in finance industry)
- metrics including PE, EPS, ROA, ROE, ROS, GOS, DAR are collected for years from 2009 - 2021
  for the years where these metrics unavailable, they will be left Nan

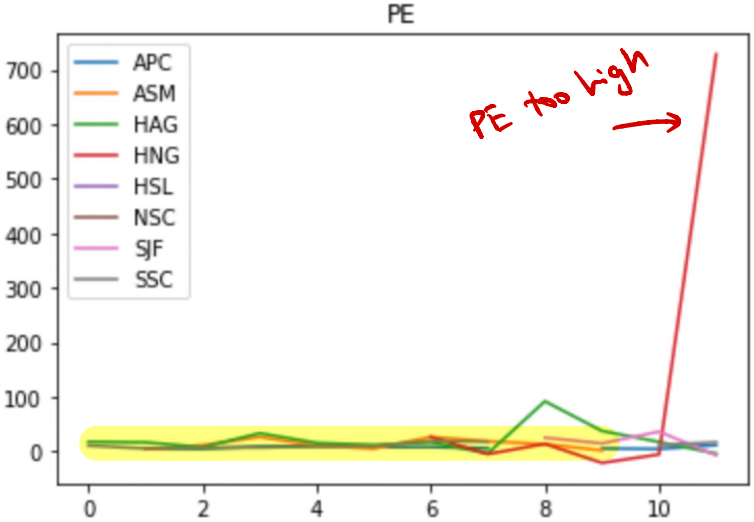
Data Verification

- Collected data has been randomly verified for a small number of stocks.
- Collected data has been verified with different web pages as well (vietstock, FireAnt)

Data Visualisation & Feature Extraction







Process of Feature Extraction

- By plotting and visualizing, data trends can be revealed

- Also, it is assumed that similiar data trends repeat and there is a reason behind the change (standing from a ML/AI knowledge it can be true, but from a finance background, this can be false. More reasonings why the accumption does not work will be investigated later)

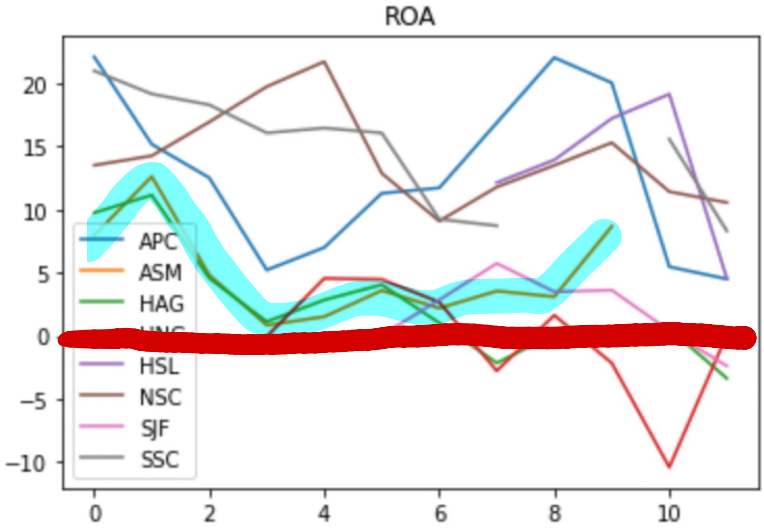
- Also assume that trends in a field can be applied to another field

- Feature is extracted by compairing data trends and stock performance. A good performing stock is considered to be the one who has consistently upward trend.

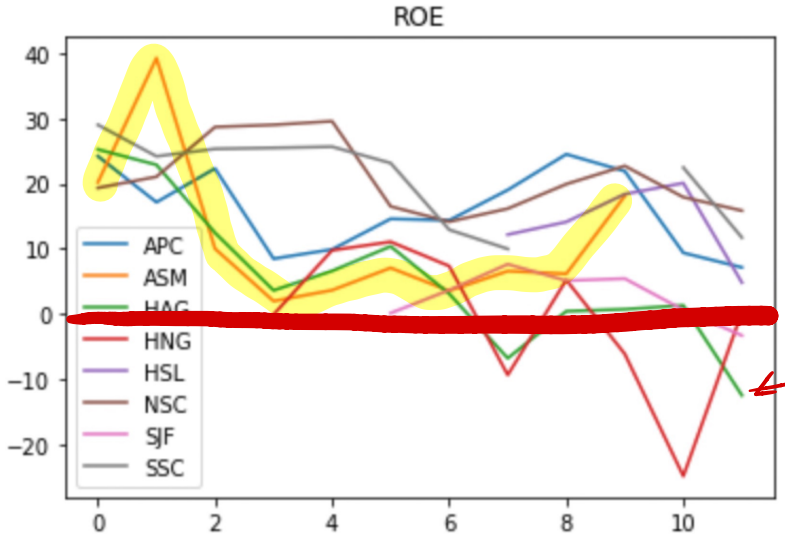
- Among the following stocks from 'Agriculture' industry, ASM is the only one that merely satisfy the condition. NSC is also okay, while the rest performs really bad.

By observations, the following features are engineered:

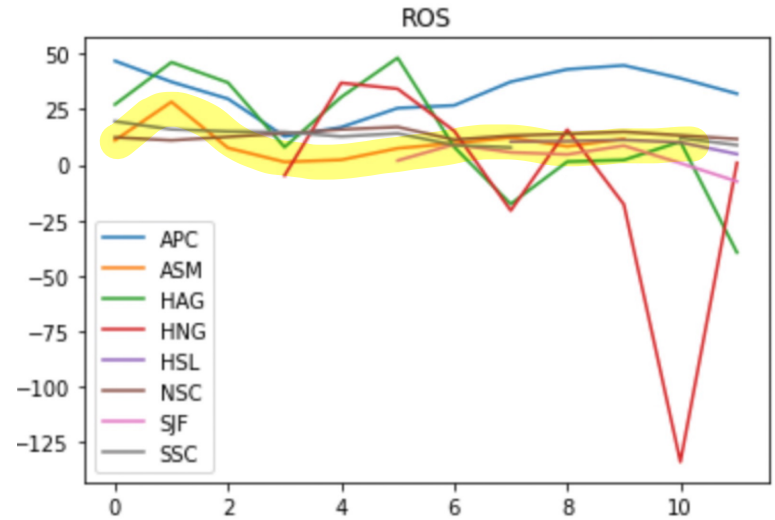
+, Negative ROE?
+, Negative ROA?
+, Negative EPS?
+, Downward ROA?
+, Downward ROE?
+, Sudden change in BV?
+, PE too high to its peers?

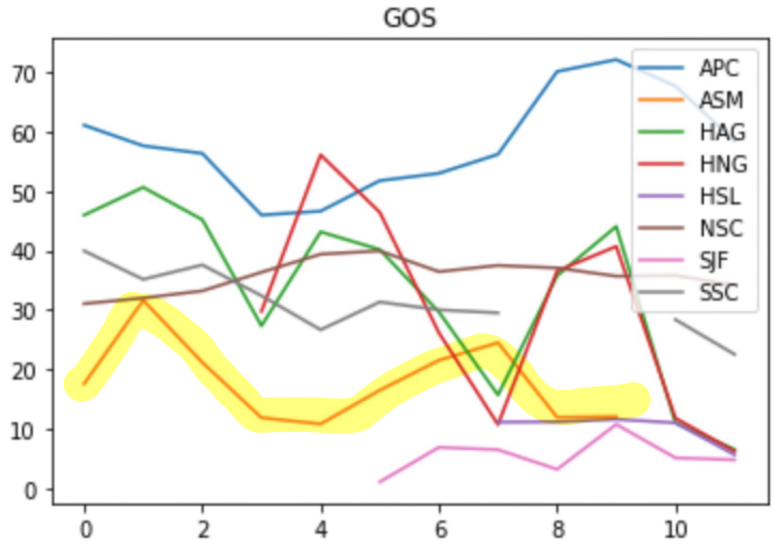


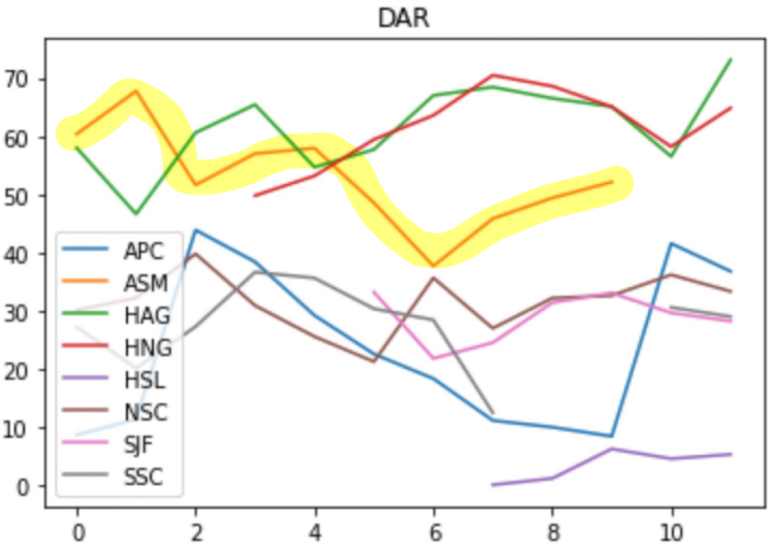
negative ROA → bad



negative ROE? → bad







Machine Learning Model

After successfully engineered new features, a machine learning model is applied to categorise different stocks. An unsupervised learning model is used to accomplish the task.

(Clustering - the number of clusters chosen is 2
Ideally, 0 - not worth investing
1 - worth-while investing)

Results

Clustering does give out some result, but apart from the results, there were many other concerns realised during the process (most of which are related to naive assumptions:

+, **not any industry has the same trend.**
There exists industry where mostly all companies perform well altogether, while there are industries where very little to none companies do well.
Therefore, the assumption that characteristic / data trend of a field can be applied to another is false and misleading. Therefore, feature engineering can lead to misleading result.

+, **difficult to assess result labelling**
Results from clustering does separate and categorize stocks into different bins. However, assessing whether the clustering works is a subjective and labour intensive task. There were over than 400 stocks on HOSE market, it is possible to label all of those by looking at the stock price chart simultaneously; but it is a hard, ineffective work to be done.

+, **false repeated trend assumption**
The ML model works well when there exist a repeated trend in the data; the computer is able to reveal what humans hard to detect. However, it's probably not the case in financial model. For example, sudden crisis can come, like the COVID-19, which is unexpected and changes stock market dramatically. In fact, many crisises are out there and there is absolutely no similiar trends in those crisis. They may come and be predicted for economists who understand and predict firm's behaviour. But those signs may take a long time to be reflected on numbers on stock market.

+, **ineffective use of resources**
If the investing bot were successfully built, it will be used to be run for three or four times per year, because the data taken was from financial report and it would be updated quarterly or annually. This will waste computing resource. Additionally, a quarter update could miss a very significant change. For example, the company is expected to have report on Dec, Mar, Jun, Sep. A company could have been bankrupted in the middle months unexpectedly. In that case, our investment could be vanished without any notice.
On the other hand, if trading is used, the program will be run at anytime, computing resource will be used effectively. Also, by continuously updating stocks' share price, the scenario mentioned earlier when a company suddenly exited the market can be somewhat be avoided.

Reflection :

- The project was not successful achieving what was initially set up as goals, due to lack of financial knowledge that lead to misleading assumption.
- Things that have been achieved during the project:
  + web scraping and crawling stock metrics for over 400 companies
  + categorise companies by industry
  + data visualisation
  + feature engineering
  + applying ML model
- Problems: unable to assess or evaluate model's performance due to subjective viewpoint and labour-intensive work.
- Even though this project did not achieve what it was set to do, during research time, the door to another gate is opened: QUANTITATIVE TRADING

  The project will be closed temporarily so that I can do research about quantitative trading and will get back to it later.