Vietnam National University
John von Neumann Institute

Entropy Competition Organizer

Socialist Republic of Vietnam
Independence - Freedom - Happiness

Ho Chi Minh, August 06th 2019

ENTROPY - DATA ANALYTICS COMPETITION 2019

FINAL ROUND

PLEASE READ THESE BELOW INSTRUCTIONS CAREFULLY

- The Final Round will be composed by 2 sessions:
  - Session 1: Top-20 contestants will work in Session 1 of this subject and submit your work **before 23h59, August 21th 2019**.
  - we will announce the list of top 10 contestants based on the output results of Session 1 **before 23h59, August 24th 2019**.
  - Session 2: Top-10 contestants based on the result of the Session 1, will continue with the Session 2 and submit your work **before 11h59AM, September 3rd 2019**.
  - Top 10 contestants still have three days to prepare a well-organized presentation for the Final Day at John von Neumann Institute.
- The technical report should be presented in well-defined format and English usage.
- The technical report must have clear structure: introduction, approaches, evaluation, discussion, etc which are suitable for delivering your work efficiently.
- The contestant have the first week for investigate and explore the data and problem, even build your own models (if you have this ability). After the first week, we will open a link and publish API for you to test your results.
- The contestant can only use Python or any equivalent IDEs to complete this Final Round.
- During the competition, if the contestants have any concern, please post your question here, or email to the Organizer.
- Lastly, good luck and **please do not quit**.

Vietnam Stock Market Challenge

INFORMATION SECTION

For this section, starting from **I** to **V**, we would like to introduce to you the motivation of the topic of the fourth Entropy competition, provide you all necessary information you need to know regarding the data, the evaluation method, the way of submission.

I. Introduction

Vietnams stock market has been evaluated as a bright spot in the region in terms of growth speed and foreign capital absorption for years. It has also been added to the watch list for possible upgrade to Secondary Emerging Market by FTSE Russell, a leading global provider of financial services. And more importantly, our government's vision is to develop high technology-based resources working on financial market in the next several years. Therefore, the topic of the final challenge for the fourth Entropy competition is in financial analysis from data science perspective.

Here, the question is can we leverage the noisy/dirty/large financial data which is very popular in financial market to create a portfolio that meet client's requirements? It turns out leading to the main challenge is how to ingest/analyze/interpret the data in the most feasible sense in order to determine the list of feasible signal(s) in this pool of information.

In this final round, you are provided the stocks data in Vietnam market (in HOSE exchange) from their listed dates to the end of July 2019, the main answer we would like to know is could you propose an innovative approach and implement your approach to select a list of tickers that potentially produce a highest Sharpe Ratio on the future data.

II. Dataset

The information of the data:

- The file name: `entropy2019_dataset.csv`.
- The data has 520432 rows and four attributes of 443 tickers.
- The data can be downloaded here.
- The data has the following schema (see Table 1).

Field	Type	Description
ticker	string	an unique ticker of an asset
date	datetime	a current date
close	float	an adjusted close price (adjusted for splits or dividends)
volume	int	trading volume in shares

Table 1: List of features

III. Performance Metric

There are actually several indicators employed to measure the performance of a portfolio, but here we only focus on Sharpe Ratio. *Assuming the risk free rate is zero*, the Sharpe Ratio is the performance of an investment by adjusting for its risk which is defined as:

$$Sharpe\ Ratio = \sqrt{n} * \frac{\mathbb{E}[R - R_f]}{\sqrt{var[R - R_f]}} \tag{1}$$

where:

- n is an annualization factor of your period, e.g. $n = 252$ for 252 trading dates in one year, 12 for months or 4 for quarters,
- R is daily-returns vector of the whole portfolio or one asset,
- R_f is the risk free rate vector (assumed as zero).

See https://en.wikipedia.org/wiki/Sharpe_ratio for more details.

IV. Portfolio creation

Using the data from date t to T ($t < T$), we can evaluate your selected list of tickers such as A, B and C with their returns are R_A, R_B and R_C respectively. We always hold a portfolio uniformly on your selected tickers, in another words, the return of your portfolio is average of tickers returns. The formulas of those returns are defined as:

$$R_i = \frac{P_i^T}{P_i^t} - 1 \tag{2}$$

$$R = \frac{1}{N} \sum_{i}^{N} R_i \tag{3}$$

where:

- R_i is the return of the asset i
- P_i^T is the price of the asset i at date T
- P_i^t is the price of the asset i at date t
- N is the number of tickers in the portfolio
- R is the return of the whole portfolio

In the above example with tickers A, B and C, you can get your portfolio cumulative return as following:

$$R = \frac{1}{3}\big(R_A + R_B + R_C\big) \tag{4}$$

After that, we can compute standard deviation of the portfolio return then use formula 1 to have Sharpe Ratio (here we're assuming the risk free rate is zero).

V. Submission

The submission includes two parts:

V.1. The test result via API

The structure of test result is your selected tickers that your model predict those tickers will create a portfolio with highest Sharpe Ratio on future data. The format of this test result is as following:

Listing 1: `result_<ENTROPY-ID>_<Name>.csv`

```
ticker
AAA
BBB
CCC
```

Please take a look at our `result.csv` file to make clear the submission requirement.

V.2. The solution via a zip/rar file

The structure of the solution here (for both two sessions) includes:

- Source code
- A clear technical report

- A file readme.txt

The submission links of two sessions:

- The submission form of Session 1: form 1
- The submission form of Session 2: form 2

The submission name of the solution part MUST FOLLOW the structure below:

```
<ENTROPY-ID>_<PRIVATE-ID>_<Name>.zip/rar
(ZIP/RAR archive file format)
```

where:

- `<ENTROPY-ID> is your provided unique public code.`
- `<PRIVATE-ID> is your unique private code shared secretly to you.`
- `<Name> is the name of the contestant.`

Example: A contestant named Nguyen Van An with the ENTROPY-ID E-Q1234 and PRIVATE-ID 1as456, he needs to upload his submission with the file name: `E-Q1234_1as456_NguyenVanAn.zip/rar`

The contestant can submit multiple times and the Organizers will take the latest version of submission for the final grading.

CONTENT SECTION

For this section, starting from **VI** to **VII**, it includes all the tasks of the Final Round.

VI. Requirements of Session 1 - for all 20 contestants

The whole process includes multiple steps from preprocessing to evaluation. You are requested to do and report your work step by step:

1. **Data Exploration (20 points)**: you need to explore the data to show some statistical and analysis information of given training data. The insight information will be useful for further steps. To make sure you understand the context, here are some tasks for you:

   (a) Show a price path of one ticker and its Simple Moving Average (SMA),

(b) Show a price path of ticker VNINDEX and its Simple Moving Average (SMA),

(c) Show a price path of one ticker and its Relative Strength Index (RSI),

(d) Show a price path of one ticker and its rolling 60 days Sharpe Ratio,

(e) Does your chosen ticker has seasonal effect? Furthermore, do tickers in Vietnam market have seasonal effect?

(f) Does your chosen ticker's daily returns greater than our market index (ticker VNINDEX) statistically?

2. **Feature Engineering (20 points)**: some features are useful, some are irrelevant while others are missing from given data set. In this step, you should consider:

   (a) **Features Extraction (10 points)**: list of generated features from the given data. Provide your reasons and evidences,

   (b) **Featutes Selection (10 points)**: list of useful features from your viewpoint. Provide your reasons and evidences.

   After this step, provide list of features which you will use for modeling. In the case you have other strategies for feature engineering, provide detail of your approaches carefully.

3. **Model Selection (10 points)**: which model should be chosen for this challenge with your features? Please describe your approach and the clear pipeline you apply to model this challenge and data to fit into your approach.

4. **Tuning Parameters (10 points)**: providing only workable models is good but not enough. In this task, you are asked to improve your model performance via clear step of tuning parameters.

5. **Discussion (10 points)**: Normally, as a daily task in data science when a data scientist tackles a challenge, you are asked to provide your opinion on the quality of your current data for this challenge. Please provide your discussions and ideas for new, feasible features which should be collected to increase model performance.

6. **Model quality (30 points)**: You will be provided the train data to build your own models (in the period 01/01/2013 - 31/07/2019). Then your selected tickers (i.e., your final model) will be evaluated via API, and the final result will be realistically tested from the first of August to the day before August 24th 2019.

VII. Requirements of Session 2 - only for the top 10 contestants

1. **Innovation (20 points)**: Do you think you could leverage more indicators in market to enhance your list of features and how can you judge your proposal?
2. **Improvement (30 points)**: Could you consider to add more filtering on the stock universe to enhance your selection?
3. **Quality of deliverables (50 points)**: The remaining score will be applied to evaluate the quality of your report and presentation, including the way you defense your ideas and implementation to the Jury.

END OF PROBLEM