Fall 2021

Final Project Description and Requirements: Portfolio Optimization

Project Description

This problem is posted by an asset management firm as their Portfolio Optimization Horse Race for Summer 2019. The goal of the project is to construct a portfolio of U.S. stocks which tracks the expected return of a given benchmark index and yields the highest *information ratio* (as defined by the annualized monthly return divided by the standard deviation of the returns). The benchmark index is the Russell 1000 Index from Dec. 2002 to Aug. 2019.

Courtesy to the firm, a set of more than 20 variables (a.k.a., factors) on a universe of about 1000 US and ADR stocks for 16 years are provided. These monthly factor data cover the period from 12/2002 to 11/2018. The data is contained in a zip file posted on Canvas. The definitions of the factors are given in the Appendix.

Complete the following tasks:

1. Use the monthly returns of the Russell 1000 index from Jan. 2003 to Nov. 2018 contained in "Benchmark Returns.csv" and compute the information ratio (defined below) and maximum drawdown of the buy-and-hold strategy for this index.

2. Write a function which takes input arguments including the data file name "`rus1000_stocks_factors.csv`" and returns an 10-element array. The elements of the array correspond to the number of SEDOLs (namely, tickers for stocks) with "return" values falling in each of the 10 intervals (0, 10], (10, 20], (20, 30], … , (90, 100].

3. Fit a 10-factor return prediction model as given by equation (15) in Reference [1] using all stocks contained in the stock universe given in the factor data file. You may use any programming tools to fit this multivariate linear regression model across all securities using factor values in Oct. 2004, either Python or C++, with "return" of each SEDOL in Nov. 2004 being the dependent variable. Basically, you are asked to fit a multi-variate linear regression model with data in Oct. 2004 and Nov. 2004, then apply the fitted model to predict returns of all SEDOLs in Dec. 2004 with factor values of each SEDOL in Nov. 2004. Interpret the fitted results in terms of how factors influence the predicted returns and explain whether the 11 fitted coefficients of the factors (a0, a1, a2, … , a10) obtained for Nov. 2004 are statistically significant.

4. For the predicted return values in Dec. 2004, write a function to compute the Information Coefficient (IC) and the associated t-statistics in the following fashion. Suppose there are

1000 SEDOLs in Dec. 2004. Then, we will get 1000 predicted return values when applying the 10-factor return prediction model in part 3 to the 1000 stocks. Create a scaled_return of each stock which is defined as the percentile value of each individual predicted-return value within the 1000 predicted returns. For example, if the predicted return of stock ABC is 23.4 and 23.4 is the 17.6 percentile of the 1000 predicted returns in Dec. 2004, then the scaled_return for ABC is 18 (round 17.6 to integer 18). Now, we have 1000 scaled_returns for the 1000 stocks in Dec. 2004. The IC is then computed by regressing the 1000 scaled_returns onto the actual returns in Dec. 2004. The regression coefficient in front of actual returns is the IC and the t-statistic of the regression coefficient is the t-statistic of the IC. Report the IC and t-statistic for the predicted returns generated by the 10-factor model in Dec. 2004.

5. Form a portfolio $w_t \equiv (w_{1,t}, w_{2,t}, \dots, w_{N,t})$ where N is the number of stocks in `rus1000_stocks_factors.csv` in each month t (t ranges from Dec.2004 to Nov. 2018) based on the following strategy. Set two portfolio strategy parameters H = 70 and K=4. Suppose N = 2500, we need to decide on the fraction of investment to put into each of the 2500 stocks in Nov. 2004, namely, the vector $w_1 \equiv (w_{1,1}, w_{2,1}, \dots, w_{2500,1})$ where each element of $w_t$is greater than or equal to 0 and the sum of the 2500 fractions is equal to 1. We consider an equal-weight strategy over all stocks with predicted return values in Dec. 2004 being greater than H (70 in this case). For example, if in Nov. 2004, the predicted returns in Dec. 2004 of the 2500 stocks have only 3 values which are greater than 70. Say, they are the predicted returns of stock 17, stock 346, and stock 1123 in Dec. 2004. Then $w_{17,1}$ = $w_{346,16}$ = $w_{1123,1}$ = 1/3, and all other w's are set to 0 in Nov. 2004. Next, we need to determine portfolio weight in Dec. 2004. Consider the set of stocks in Dec.2004 which have predicted returns in Jan. 2005 being greater than H. Denote this set by A. Let B denote the set of stocks which have positive weights in Nov. 2004. We pick K stocks which are in B but not in A, and replace them with K stocks which are in A but not in B. After the replacement, set B becomes B*. The portfolio in Dec. 2004 is then formed by holding equal-weight in all stocks in B* and 0-weight in other stocks. Repeat this procedure to construct 179 monthly portfolios from Nov. 2004 to Oct. 2018. Compute the 179 monthly returns from Dec. 2004 to Nov. 2018. Once the portfolio weights of all selected stocks are determined, use these weights and the actual returns of the corresponding stocks contained in "cleaned_return_data_sc.csv" to compute the monthly returns of the portfolios formed every month.
   a. Compute the mean, standard deviation of this monthly return series. Report the annualized information ratio as defined by (sqrt(12)*mean(returns)/stddev(returns)).
   b. Plot the cumulative returns from Dec. 2004 to Nov. 2018.

(Remark: to get robust predicted returns in Dec.2004, it is suggested that one repeats the model fitting task described in part 3 for Nov. 2004, Oct. 2004, …, Jan. 2004, and Dec. 2003. Then use the average of the 12 sets of coefficients (a0, a1, a2, … , a10) to get the predicted returns in Dec. 2004. Similar procedure shall be carried out for the subsequent months.)

6. Consider the same portfolio strategy applied to stocks selected by the CTEF scores. This is simply done by repeating part 5 for t = Nov. 2004, Dec. 2004, Jan. 2005, Feb. 2005, Mar. 2005, … , Oct. 2018, Nov. 2018 with the predicted returns of the stocks replaced by the CTEF scores observed in current month. Compute the IC of CTEF scores in current month with respect to the realized returns in next month and the information ratio of this portfolio strategy. Remark: this strategy is based on a simple return-prediction model which uses the CTEF score of a stock in month (t-1) to be its predicted return in month t.

7. Implement TWO of your own return-prediction models which utilizes any machine learning/deep learning model to predict the stock returns in month t based on factor values observed in previous months. Namely, repeat part 3-5 with your own return-prediction model replacing the 10-factor model. Compute the IC and t-statistics of your predicted returns and the information ratio of your portfolio based on stocks selected with your predicted returns from Nov. 2004 to Oct. 2018.

Submission requirements:

1. All source codes and a README.txt file containing: a) a list of all files submitted; b) a list of ALL major dependence of supporting libraries/packages and their versions; c) instructions on how to compile and run the main code. Source codes shall be readily run, and tested for expected outputs without any issue. (These include all the C++/header files and/or Python scripts used in the project).

   The codes and implementation shall demonstrate proficiency in the following aspects:

   - Object-oriented design
   - Clean structure of the task flows and the corresponding functional modules
   - Adequate use of functions
   - Clear input/output interfaces

2. A written report describes the approaches taken, results found and analysis performed which pertain to each of the 7 parts in the project description.
   a. The construction of the input variables (i.e., features) and output variables to the price prediction models need to be clearly explained.
   b. The statistical validity and the accuracy of the price prediction models shall be provided.
   c. Clear justification of no look-forward bias in the models/strategies.
   d. The report shall contain the following sections:

        i. Problem description and data preparation.
        ii. Specification of each of the 4 return-prediction models (10-factor, CTEF, two of your own models). Discussions of the model fitting results and the validity of fitted parameters.
        iii. Performance metrics (using tables for ICs, t-statistics and Information Ratios and figures of cumulative returns) and discussions of each of the 4 portfolios constructed based on the 4 return-prediction models.
        iv. Conclusions.
        v. Appendix: explicit description of the contribution of each group member in “Contribution” section. Under the name of each member, list the following information,
            1. the work done such as performed data cleaning/feature generation, built a Support Vector Machine regression model to predict price, conducted model validation with backtesting, etc.
            2. the codes or parts of a code written by the member.

    e. Four csv output files generated by the TWO of your own models so one can directly compute the monthly returns of the two respective portfolios by multiplying the two matrices contained in these csv files. Specifically, for each model, submitting the following two csv files:
        i. One csv file has the first row being all the SEDOLs which you use to create portfolios at a monthly frequency in Dec. 2004, Jan. 2005, Feb. 2005, Mar. 2005, … , Nov. 2017. Put all SEDOLs of the union of the stocks appearing in the 168 portfolios as the labels of all columns in the first row of the csv file, then for the subsequent rows, use YYYY-MM as the index of these rows, and put the portfolio weights of the portfolio in YYYY-MM in the corresponding row with each weight entered into the corresponding cell with the correct SEDOL column label.
        ii. The other csv file has the exact row index and column headers. All the cells contains the monthly returns predicted by the model for each of the SEDOLs appearing in the column headers in all YYYY-MM rows.

References

[1] Guerard, J.G., S.T. Rachev, and B.P. Shao (2013). “Efficient global portfolios: big data and investment universes”, IBM J. RES. & DEV. Vol. 57, No. 5.

[2] Guerard Jr., J. B., Markowitz, H.M., & Xu, G. (2014). The role of effective corporate decisions in the creation of efficient portfolios, *IBM Journal of Research and Development* 58, No. 4, Paper 11.

Appendix

Explanation of data labels in the factor data file

1.DATE: MM/YYYY

2.CUSIP

3.SYMBOL

4.COMPANY NAME

5.SEDOL: stock identifier

6. FS_ID: FactSet ID

7. EP - Earnings/Price

8. BP - Book/Price

9. CP - Cash Flow/Price

10.SP - Sales/Price

11.DP - Dividend Yield

12.EP1 or FEP1 - 1 year ahead IBES Forecasted EPS to Price/Last year's forecasted earnings per share

13.EP2 or FEP2 - 2 year ahead IBES Forecasted EPS to Price/Last year's forecasted earnings per share

14.RV1 - FEP1 IBES Revisions

15.RV2 - FEP2 IBES Revisions

16.BR1 - IBES Breadth

17.BR2

18.CTEF - Consensus EPS I/B/E/S forecast, revisions and breadth

19.PM1 - price momentum as price(t-1)/price(t-12)

20.PM2 - price momentum as price(t-1)/price(t-7)

21.ES - (Eli Schwartz) Corporate Exports

22.RETURN - Returns

23.REP - Current EP/Average EP of last 5y

24.RBP - Current BP/Average BP of last 5y

25.RCP - Current CP/Average CP of last 5y

26.RSP - Current SP/Average SP of last 5y

27.RDP - Relative Dividend Yield

28.VOL - Monthly stock Volume

29.CRET - Monthly Stock Return

30.STATPERS - Date of IBES Forecast

31.USFIRM - US Firm (=1)

32.CURCODE - Currency Code

33.TOT - Total Number of FY1 Analysts

34.FGR1 - 1 year ahead forecast earnings per share monthly breadth

35.FGR2 - 2 year ahead forecast earnings per share monthly breadth

36.MRV1 - Mckinley definition of revisions in 2005

37.MRV2

38. ROIC: past 12-month return on invested capital

39. RSTDEV: standard deviation of past 12 monthly returns

40. RPM71 (see PM2): reverse price momentum of month-7 price divided by month-1 price.

41. ROA1, ROA3: 1-year, 3-year return on asset

42. ROE1, ROE3, ROE5: 1-year, 3-year, 5-year return on equity

43. 9MFR: return forecast by Mckinley 9-factor model

44. 8MFR: return forecast by Mckinley 8-factor model

45. LIT: legal insider trading index