Data Science Methods in Finance

Group Assignment

Jens Kvaerner, Ole Wilms

Submission deadline: March 21, 2021 at 23.59

Important Instructions

- Submit a report with your results and a zipped folder with the script.
- The code must run without errors and generate the same solutions as you have reported
- The data you will use is "2020_CRPS_DataScience.RDS", and it is available on Canvas
- We care about the layout, and we want a "consultant" professional report. That means you should prepare a set of slides, and not a long document (more on this below).
- We will not upload a detailed solution manual. However, we will ask 1-2 groups to share their solutions with the rest of the class.

General Instructions

This exam is based on WRDS data and consist of a large cross-section of US stocks. You have data on monthly returns, volume, etc. and a set of accounting data. A description of all variables present in the dataset is available in Gu et al. (2020) and additional information is available at Shihao Gu's website. The ultimate goal of the assignment is to identify a good model for predicting stock returns, and to use those predictions to form investable portfolios that outperform the market (more on this below).

Please create a report in power point or with Beamer (we recommend overleaf) where you summarize the results. Each item should be answered in a different slide. Start all headers with the question you are answering. For example, when presenting summary statistics for the first question, write "Q1: Summary statistics", etc.

There are 14 questions that should be answered with 14 slides (put additional calculations, if needed, in your appendix). Each question is given equal weight (i.e., 1/14) and they add up to 100 points. 10 extra points will be given to solutions with extraordinary good layouts.[1]

The group assignment are quite long and demanding but you have all resources available. We encourage discussions on ment, but sharing code is not allowed.

[1]The template "example_template.pdf" on Canvas is a good starting point. We strongly encourage you to follow this structure as close as possible.

Task

1. **Report summary statistics** for the data. Include in this statistics, at least 1) the number of stocks with missing and non-missing values by year (e.g., graphically), 2) mean returns (reported monthly in percentage points), 3) standard deviation of returns, 4) mean log market cap, and 5) the standard deviation of log market cap. For 2-5), report the time-series averages of the cross-sectional means in a table.

2. Create two portfolios using all stocks. The first portfolio is an equally weighted portfolio of all the stocks and the second portfolio is a value weighted portfolio. **Plot the cumulative return** of these two portfolios together with the market factor available at Kenneth French's website (to make the market factor comparable with your factors you need to add "Mkt.RF" and "RF". Download all the factors (we will use them later).

3. Create 100 portfolios based on the stocks market cap that are monthly rebalanced. Do not include the last five years of the data in this exercise. Portfolio 1 contains the smallest stocks, etc. **Create a figure where you report the full period annualized Sharpe ratio of these 100 portfolios**. Before you calculate the Sharpe Ratios, convert the monthly returns into excess return by subtracting off the risk free rate, "RF" (hint: make sure to get the units correct).

4. Use those 100 portfolios to find a linear combination of these portfolios that have the highest in-sample Sharpe ratio (make sure the weights add up to one). **Create a table that compares** this Sharpe ratio with the Sharpe ratio of the market and the equally weighted portfolio you have created earlier for two periods. First the period you used to calculate the weights in the tagency portfolios. Then use the same weights for the remaining five years of data that you have not used yet. In total you will report six Sharpe ratios. Comment (maximum two sentences)

5. Perform a PCA on those 100 factors for the two periods (in-sample and out-of-sample). **Create two figures that show** much of the variation is captured with the first 3 PCs. Provide one figure for each period.

6. **Create a table with the Sharpe Ratios of each of the 3 PCs in both periods**, and interpret them in light of the portfolios they are made up from (maximum three sentences)

7. **Create a figure that illustrates a 6-by-6 correlation matrix** of the PCs from the two sample periods. Comment (maximum two sentences)

8. Missing values are prevalent in the earlier years of the data set. Impute these with the cross-sectional median of the respective characteristic.[2] Before you fill in missing values, drop all variables (characteristics) that have more missing values in a given year than a reasonable cutoff. **Explain on one slide** how you decided the cutoff, how many variables you lost, and potential issues with this selection method. Hint: You should have more than 50 predictors. As a final pre-processing step, scale all the features to the [-1, 1] interval (and if necessary include only stocks with say at least 12-24 months of return history. One way to do this is to use dplyr and "group_by(id)" and then count the number of months. You can then filter data accordingly with "filter()".

9. **Report the same summary statistics as in question 1**. Include also the summary statistics from question 1, and make it easy to compare the two samples. Comment on the "representativeness" of the sample (maximum two sentences).

10. Organize your data so it is ready to be used to forecast stock return (Table 1 in the Appendix provides an example). Continue to leave the last five years for testing. Split the remaining data into a 70/30 split for training and validation. Keep the ordering of the data. Chose 3 methods of which one should be (lasso, or ridge), one should be tree-based, and one should be either PLS or PCR. For each model fit as many versions you want on the training set, and use the validation set to pick the best model of each type. **Create one slide where you state (and explain) the metric you use to chose between models**. One such metric is the cross-sectional average mean-squared-error give by

$$\frac{1}{N}\sum_{i}^{N}\sqrt{\frac{1}{T_i}\sum_{t_i}^{T_i}(r_{t_i}-f(r_{t_i})^2}, \tag{1}$$

where $f(r_{t_i})$ is the predicted stock return for stock i in period t, and r_{t_i} is the realized one. Another alternative is to use

$$R^2 = 1 - \frac{\sum_{(i,t)}(r_{i,t}-f(r_{i,t}))^2}{\sum_{(i,t)} r_{i,t}^2}, \tag{2}$$

which sums over all stocks and months in the relevant period.

11. **Make one slide with your three best models**. Write down the model, its hyperparameters, etc. Report all the information we need to use your model. Include the performance

[2] In other words, if we are missing characteristic $c_{i,t}$, where i is the individual stock and t the month, we impute it with the median of c of all stocks at month t.

in the training and validation set.

12. Choose one model and **create one slide with a figure** that shows which features are most important for predicting stock returns.

13. Use your model to create a portfolio for the last five years of the sample (i.e., the test data you have not used it). Make sure the portfolio weights add up to one each month. One simple way to do this is to predict returns for every month and then rank all stocks from high to low return predictions. To ensure the weights add up to one you can use the softmax function. **Create a figure that shows the cumulative performance of this portfolio** for the last five years of the data (i.e., the test set) and compare it with 1) the market portfolio and 2) the tangency portfolio you identified earlier. We want one slide, but feel free to add similar charts for other two models in your appendix.

14. **Create one table with Fama-French regressions**. That is, regress the two portfolios you plotted in the previous question on the Fama-French 5 factors. Comment (maximum three sentences). Extra: Calculate the statistical power you have in rejecting the null of a zero intercept against the alternative of a positive alpha.

1 Appendix

Table 1: Example of data structure

Year	Month$_{t+1}$	Stock ID	Return$_{t+1}$	Feature 1$_t$	Feature 2$_t$	Feature 3$_t$
1995	31/01	10006	0.069	0.632	-0.297	0.511
1995	31/01	10014	0.000	-0.434	-0.485	0.137
1995	31/01	10030	-0.029	0.476	-0.664	-0.334
...	...	...	...	...	...	...
1995	31/12		0.072	0.712	-0.785	0.333
1995	31/12	93434	-0.041	-0.612	-0.515	-0.167
1995	31/12	93436	0.128	0.926	0.625	-0.540

References

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.