

Problem Set 2

Business 35137

Winter 2026

Due: February 16

1. Download the file `largeml.pq` from canvas. This file contains monthly returns for 500 large-cap stocks along with a series of firm characteristics from [openassetpricing](#). The returns are offset by one month from the characteristics. This is a parquet file, which can be read into `pandas` using the `pd.read_parquet` function.
 - (a) First, rank-sort each characteristic cross-sectionally each month and use the resulting ranks to form a portfolio for each characteristic. Compute the annualized sharpe ratio for each of these portfolios, plot these results and examine the characteristics of the best and worst performing portfolios.
 - (b) Next, let's form an alternative set of signals using machine learning using the following methods:
 - i. OLS over the linear characteristics.
 - ii. lasso/ridge/elastic net over the linear characteristics.
 - iii. lasso/ridge/elastic net over the non-linear expansion formed by `RBFSampler`.
 - iv. `PLSRegression` over the linear characteristics and the non-linear expansion.
 - v. `GradientBoostingRegressor`

Use the first 20 years of data to train the model, the next 12 years to select the tuning parameters, and the remainder of the data to evaluate the models out-of-sample. Compute an out-of-sample R^2 for each method and interpret your results.

- (c) Using the forecasts from each model in part (b), form a portfolio corresponding to each method. Over the same time period, compare the annualized Sharpe ratios for these ML portfolios to the portfolios formed in part (a).
- (d) Download the file `smallml.pq` from canvas. This file contains 1000 small cap stocks, but is otherwise formatted identically to `largeml.pq`. Repeat (a), (b), and (c) and compare and interpret the results.
- (e) Finally, using everything you've learned up to this point, attempt to form a portfolio that earns the highest possible Sharpe ratio out-of-sample. Explain the reasoning behind your choices.

2. Download the `lsret.csv` file from canvas. This contains long-short portfolio returns for each of the characteristics used in [openassetpricing](#).
- (a) Using the returns from `lsret.csv`, along with the portfolios you formed in 2.a and 2.d, run PCA to estimate a set of latent factors from each set of portfolios. How many factors are necessary to explain most the variation in the portfolios? Interpret the factors associated with each set of portfolios and compare the Sharpe ratios across portfolio sets and factors.
 - (b) Using the portfolio returns from `lsret.csv` introduce an additional column corresponding to an indicator (1) for all rows. Using all data prior to 2004 as your training/validation sample, estimate lasso and ridge to “predict” the indicator. Use the estimated coefficients to form a portfolio out-of-sample and report the annualized Sharpe ratio.
 - (c) Then for a range of latent factor counts, compute PCA on the portfolio returns and repeat the exercise from part (b) for each factor count. Plot the out-of-sample Sharpe ratio as a function of the number of factors. Interpret these results.
 - (d) Repeat (b) and (c) for the large and small-cap portfolios. Compare the results across the three sets of portfolios and interpret the differences.
 - (e) Finally, using everything you’ve learned up to this point and the portfolios from `lsret.csv` attempt to form a portfolio that earns the highest possible Sharpe ratio out-of-sample. Explain the reasoning behind your choices.