Problem Set 1

Business 35137

Winter 2026

Due: January 23rd

1. Download the file `gw.csv` from canvas. This file contains monthly S&P 500 index returns along with a series of predictors used to forecast the market. The S&P returns are offset by one month from the predictors. In the first part of the problem set we will explore how well we can forecast market returns using machine learning methods.

   (a) For each of the predictors, regress the S&P 500 index returns on the predictor using the full sample of data. Report the R^2s of these regressions. Next, evaluate the out-of-sample performance of each predictor individually using an expanding sample of data starting in 1965. How do the out-of-sample R^2s compare to the in-sample R^2s? Interpret what this means for the usefulness of these predictors in forecasting the market.

   (b) Next, try the same expanding sample exercise but include all the predictors in a single regression, compare the out-of-sample R^2 here to those in part (a). Let's now incorporate a penalty term into the regression to counteract overfitting. Compute results for lasso, ridge, and elastic net and use K-fold cross-validation to select the optimal penalty term. Plot the out-of-sample R^2 for each month for each of the three methods along with the un-penalized regression. How do the methods compare? What does this tell us about the predictability of market returns?

   (c) Next, lets introduce some non-linearities into the model. Use the radial basis function kernel to generate non-linear expansions of the underlying predictor set (use the `RBFSampler`

from `sklearn`). Generate these features for a number of different feature counts. Plot the out-of-sample R^2 as a function of the number of features generated by the kernel. How do the results compare to the linear models? Interpret the importance of the number of features in the kernel expansion.

(d) To what extent do our results depend on the training window? Refit the model from part (c) using a rolling window of 12, 36, 60, and 120 months. What do you observe about the out-of-sample R^2 as the training window changes?

(e) To what extent do our results depend on the cross-validation method? Refit the model from part (c) using a range of folds for cross-validation. What do you observe about the out-of-sample R^2 as the number of folds changes?

(f) Next, download the `FREDMD.csv` file from canvas. Incorporate the macroeconomic variables from this file into the model from part (c). How do the out-of-sample R^2 change when we include these variables? What does this tell us about the virtue of complexity?

(g) Lets compare the results from part (c) to some alternative methods. Compare the results to the `KernelRidge`, principal components regression (combine `PCA` with a standard regression framework), `PLSRegression`, and `GradientBoostingRegressor` methods from `sklearn`.

(h) Using everything you've learned up to this point, construct the best possible model for forecasting the S&P 500 index returns. Explain the reasoning behind your choices.