MIT SLOAN SCHOOL OF MANAGEMENT

Adv. Analytics of Finance 15.457
Hui Chen Fall 2025

# Problem Set 6

Due: 2:30 PM, Thursday, December 04

1. **Interview questions**:

   (a) True or False (and explain): Since the OLS estimator is BLUE, it is the best way to train a linear prediction model.

   (b) If the true data-generating process follows an autoregressive process and the errors are IID, will it ok to apply standard K-fold cross validation for model selection?

2. **Predicting stock returns.** You are looking for signals that can predict returns for individual stocks in order to design a quantitative trading strategy.

   The file Data_HW6.zip contains monthly returns and various characteristics for a selected list of stocks for the period of 1989 to 2016. The characteristics include stock price (Price), market equity value (MV), market-to-book equity ratio (M2B), sales-to-asset ratio (S2A), short-term-debt-to-asset ratio (SD2A), long-term-debt-to-asset ratio (LD2A), price-to-earnings ratio (PE), and quarterly sales (Sales). Notice that some of these variables are only updated once a quarter.

   (a) Construct a list of features using the information provided above. You can choose which variables to include; you can also combine different variables. Briefly explain why you decide to include these features.

   Caution: You should not directly use non-stationary variables as predictors (why?). Find a way to make these variables stationary when applicable.

   (b) Next, we will turn the realized return $r_t^i$ into a binary variable. Define $y_t^i = 1\{r_t^i > 0\}$ (i.e., simply the sign of return). Predict $y_{t+1}^i$ using a logistic regression. Estimate the model using data from 1989 to 2011 (this is your **training sample**).

   (c) Redo part 2b with Ridge and LASSO (by imposing $\ell_2$- and $\ell_1$-penalty in the logistic regression). Use time-series cross validation to tune the hyperparameter $\lambda$.

   Hint: The idea is the same as when applying shrinkage to a linear regression. In sklearn.linear_model.LogisticRegression, you can specify the norm of the penalty term.

   (d) Construct the confusion matrix in the **test sample** (post 2011) with the cutoff $\bar{p} = 0.5$. Compute the Type I/II error rates and overall error rates for the three versions of logit models developed in 2b and 2c.

(e) Redo 2b and 2d using Random Forest. Compare the performance in the test sample against that of logit.

(f) (**Optional:**) Redo 2b and 2d using a feed-forward neural network (keep the architecture simple). Compare its OOS performance with the other models.