Data Analytics and Machine Learning | Prof. Lochstoer

Problem Set 4

Using the DT_StockRetAcct.csv dataset available at BruinLearn (Week 1), we will in this exercise code up a "machine learning" version of valuation using comparables, and we will use this valuation tool to form trading strategies. That is, we will program an automated learning routine that finds the model specification -- which variables to include and the functional form these variables take -- and evaluate the out-of-sample performance of this model.

As described in the lecture notes for Topic 4, we will use the Elastic Net procedure, which includes Ridge Regressions and LASSO as special cases.

Question 1: Valuation with Machine Learning, version 1

a. We will predict current log equity market value, lnME. Next, we will construct the features we will use to do so. First, create log book value by setting lnBE = lnBM + lnME. That is, add the log market value to the log book-to-market ratio. Next, in addition to lnBE, we will also consider the characteristics lnIssue, lnProf, lnInv, lnLever, lnMOM, lnROE, and rv. Thus, you have eight characteristics in total. For each of these seven characteristics, create new, additional characteristics as the squared value of the original characteristic. Name the new characteristics the same as the original, but with a "2" at the end. For instance, for lnProf, the squared value should be lnProf2. Also, create interaction terms where you multiply each of these characteristics with lnBE (except lnBE itself, as this would simply give lnBE^2, which you already have). Finally, create 11 dummy variables using the ff_ind industry variable. There are twelve possible industries, but skip the 12th one when creating dummies as the regressions we will run have intercepts. You should now have 8+8+7+11 features that you will use to predict lnME.

(i) For each year in the sample, run a cross-sectional regression of lnME on these features. Get the predicted values lnME_hat from this regression each year. Plot the R2 from these regressions across the years in the sample. That is, R2 on the y axis and year on the x axis. Comment on any interesting patterns you see in terms of this model's ability to explain equity market values across firms.

(ii) Create the variable z_OLS = lnME – lnME_hat. That is, for each firm each year create a measure of mispricing as the actual market value minus the predicted market value.

(ii) Next, you are to use the Elastic Net procedure (with alpha (l1_ratio) =0.5) to estimate lnME-hat. Each year, run a cross-validation exercise with 10 folds. Find the optimal regularization parameter, and then run the Elastic Net procedure using all the firms for that year. The sklearn procedure ElasticNet could be useful here, as well as ElasticNetCV. Plot the chosen regularization parameter for each year.

(iii) Collect the predicted market values for the Elastic Net procedure, lnME_hat_EN. Then create the mispricing variable z_EN = lnME – lnME_hat_EN for each firm and year.

(iv) Create firm excess returns as ExRet = lnAnnRet – lnRf. Given how I constructed the data, this is next year's return. Each year, run a cross-sectional regression of ExRet on an intercept and the mispricing variables z_OLS and z_EN (that is, run the Fama-MacBeth regression to get the portfolio returns based on sorts on these variables). Report the slope in the Fama-MacBeth regression (the average excess portfolio return for each of z_OLS and z_EN), as well of their t-statistics ((average excess return / stdev of returns) * sqrt(T)). Are any of these signals, z_OLS or z_EN, useful for predicting returns? Which one seems best?

(v) Choosing either z_OLS or z_EN based on which gives the highest portfolio Sharpe ratio, now run the Fama-MacBeth regressions including lnBM, lnProf, lnInv, lnMom, as well as industry dummies on the right hand side. Is the mispricing signal z that you chose marginally useful now?