Data Analytics and Machine Learning | Prof. Lochstoer

Problem Set 5

Use the data in the spreadsheet "French_Portfolio_Returns.xlsx" on CCLE. This spreadsheet contains monthly returns on 138 portfolios on the Ken French webpage – industry and characteristics sorted decile portfolios. Use the "ExcessPortRet" sheet and data from 196307 (July 1963) to the end of the sample (Dec 2016).

In this homework, you will estimate decision trees and practice using clustering methods.

A: Set up data for analysis amenable to cross-validation routines.

1. Upload this data to a data.table in R, delete the "date" column and instead create a column called "Month" which counts the observations over time. I.e., 196307 is 1, 196308 is 2, etc.
2. Reshape the data so that each each observation (row) has Month, Portfolio, ExRet in the columns. "Portfolio" is a number from 1 to 138 – corresponding to the previous columns in the data set.
3. For each portfolio, add the three following columns: one month lagged return, two months lagged return, sum of months 3-12 lagged return. Then, add the following additional three columns: Do the same for squared return – one month lagged squared return, two months lagged squared return, squared sum of lags 3-12 of returns.

   Thus, for each portfolio-month you have six characteristics (functions of lagged and squared returns) as well as the current month excess return (the variable we will try to predict). Since all observations that have different timings are put on the same row, the already coded standard cross-validation routines will work well.

B: Decision Trees

1. Using data from the beginning of your sample (now, first row should correspond to 196407) until 200912, estimate RandomForest as in the class notes.
   a. Report the inputs you chose for the routine (number of trees, max nodes, mtry) and report a linear panel regression of excess returns on the predicted value, the latter obtained from the Random Forest predict function as in the notes. Cluster the standard errors by time to account for contemporaneous correlation across firms each month.

   b. Now, in the true out-of-sample period 201001-201612, get predicted values using the tree you estimated in a. Run a linear panel regression of realized returns on the predicted values, again clustering standard errors by time.

   c. Run Fama-MacBeth regressions in the true out of sample periods to get the Sharpe ratio and *t*-statistic of the implied trading strategy that uses the predicted value as the signal.

2. Repeat question B1, but instead use XGBoost. Try and report results for the following four parameter configurations in your exercise: {eta = 0.1, maxdepth = 1}, {eta = 0.1, maxdepth = 6}, eta = 0.3, maxdepth = 1}, {eta = 0.3, maxdepth = 6}. For each run, set nrounds using the cross-validation exercise in the notes.

C: Asymptotic PCA

1. Using the same data, run the APCA routine over rolling 5-year windows through the sample, setting K = 5 (i.e., get 5 biggest PCs). I.e., your sample is the excess returns from 196307 to 201612 on the 138 portfolios. At each time *t,* have the APCA routine estimate the 5 factors using the last 60 months of data on the 128 portfolios. Record at each time *t* the loadings of each of the 138 portfolios on the 5 factors.

   Now, obtain the next month's out of sample return to these factors by for each factor taking the 138 loadings and multiplying them with next month's return to the 138 portfolios. Roll through the sample so you have a time series of "out-of-sample" returns for each of the 5 factors.

   a. What is the average annualized excess return, standard deviation, and Sharpe ratio of each of the factors. Also report the monthly first autocorrelation of the factors.

   b. What are the t-statistics of the average return estimates. Estimate the standard errors controlling for heteroscedasticity and one autocorrelation of the factors.

   c. Given your results in a and b, what do you conclude about the use of these portfolios as hedges? E.g., do they affect average excess returns? Could anything go wrong in the rolling APCA you did in terms of interpreting the factors as risk factors? I have in mind a discussion of conditional versus unconditional expected returns here.

   d. You want to see if lagged return is a useful signal in a Fama-MacBeth regression (for predicting next month's return). However, you also want control for the loadings on the five factors you estimated in the APCA. Make sure you use the loadings in a rolling out-of-sample fashion – i.e.., so the Fama-MacBeth regressions indeed correspond to real-time tradeable portfolios. Explain your procedure (ie, what regressions you run and what their purpose is) and report the results.