

Our goal is to validate the paper's claim of outperforming consensus forecasts by using ridge regression to predict a target stock's earnings based on a feature space of:

1. current prices of the target stock
2. current prices of a set of other stocks
3. moving averages of prices of the target stock
4. moving averages of prices of the other stocks weighted by their correlations to the target stock

All our data came from Bloomberg, namely closing prices (PX\_Last), some EPS measure, and a comparable earnings consensus estimate. Our earnings data came from the field EARN\_ANN\_DT\_TIME\_HIST\_WITH\_EPS which required the BDS function to call.  ~~But some data cleaning is needed since the earnings consensus data could not be populated together with the price data in excel because price uses BDH function but consensus data uses the BDS function via the field.~~

link to paper: http://www.ferari-project.eu/wp-content/uploads/2014/12/earningsPrediction.pdf

The paper's method was conducted on stocks of the S&P 100 index over the period 2008-2012. Their cross validation for the algorithm's parameters was done on a different set of 100 randomly picked stocks during the period 2004 to 2006. What we did: Implement the algorithm in the paper (the algo retrains the ridge regressor after each earnings announcement), then cross validate for the two parameters as specified in paper using stocks of the S&P100 index for the period 2013-02-08 to 2015-02-09 (note that to simplify data cleaning, we plan not to randomly pick another set of stocks like in the paper), we found the parameter combination to be nu=10000 and W=11, we then applied the algorithm to stocks of the S&P100 index for the period 2015-03-09 to 2019-03-11, then compared Mean Relative Error of our predictions with the consensus forecasts over the same period. The paper claimed that the MRE was much lower using their algorithm, our result did not support the claim.

some facts:

the S&P 100 index data from Bloomberg had 101 stocks, after dropping stocks with Nan data, 88 stocks were left, our analysis uses these 88 stocks only.

our cross validation didn’t include the cases W=500 and W=750 as in the paper. More data is needed if we test these two as well. Without these two cases, our grid still has 6x6=36 points which is plenty. The cv took 6 hours to run.

The code’s largest bottle neck is fitting sklearn’s Ridge, so making the code faster doesn’t seem to be in our control. Our CV is very computationally intensive since we are running 36 runs (each run for a parameter combo) where each run involves 88 stocks over a two year period with refitting of the ridge regressor each time there is an earnings announcement and this happens for each stock.The run on the test set involves 88 stocks over a four year period