Big Data in Finance: In-Class Presentation Assignment

Prof. Tarun Ramadorai

Due: 3rd/4th March, 2021 (Core/Elective)

As we have seen in class, we can build a machine-learning algorithm with good out-of sample predictive power for the aggregate equity premium. In your in-class presentation, we'd like you to try to build such a model to predict *daily, stock-level* returns out of sample, based on lagged own and cross-stock returns, and lagged institutional trading flows of a given stock and other stocks. While existing research has found some predictability in stock-level returns, new machine-learning tools may be better suited to exploit the rich cross-sectional information in cross-stock returns and trading flows, and the possibly short-lived nature of useful trading signals in high-frequency data.

There are three datasets on the Hub: "Returns_Clean", which gives you daily stock returns for 100 (randomly sampled) New York Stock Exchange (NYSE) listed stocks, and "Flows_Clean" which gives you the matched (i.e. each column corresponds to the stock / column in the stock return dataset) institutional net trading flows, from 1993-04-01 to 2000-12-29. And finally, "PERMNO_Tickers" which gives you the link between the PERMNOs contained in the two previous files to stock ticker symbols (this is so that you can merge additional data in for further analysis—more details on this below).

This document can help guide you towards building your own model to predict stock-level equity returns. You should feel free to deviate from or indeed, completely disregard any of the suggestions here, and experiment with any methods you like – the presentation that you finally come up will gain points for originality and differentiation from your peers (conditional on being correct and accurate, of course!).

The final presentation should be in the form of powerpoint or other presentation slides. Presentations will be quite short, at maximum 10 minutes, so **please make sure that you have no more than 8-9 slides in total including title and conclusion**. Also, please practice your presentation in advance—**note that you will need to designate ONE member of your group to present**.

Variable Construction

We have constructed daily returns and standardized net institutional trading flows for you.

Prediction

You can now start to build *predictive* models using machine learning techniques. Let us call $r_{i,t}$ the return for stock i at date t, which is our targeted variable to predict. Our possible predictor variables are then $r_{i,t-1}$, i.e. the stock's lagged return, and the set of variables $\{r_{j,t-1}\}$, $\forall\ i \neq j$, i.e. the lagged cross-stock returns; together with $flow_{i,t-1}$, i.e. the stock's lagged trading flow, and the set of variables $\{flow_{j,t-1}\}$, $\forall\ i \neq j$, i.e. the lagged trading flows for all other stocks except for i. Feel free to use other lags as well if you find it useful (e.g. $t-2$, $t-3$, etc), the key is to use *past* information to predict current stock returns.

Additionally, the top rows of the data are PERMNOs, a unique stock identification code assigned by CRSP. You can use it to add other variables that you may find relevant. We also give you a link from PERMNOs to tickers in the third file ("PERMNO_Tickers"). Note that these links are in the form of a panel dataset, because tickers can change through time. For example, the company with PERMNO 10145 (AlliedSignal Inc.) merged with Honeywell Inc. in 1992 and its ticker changed on 1999-12-02 from ALD to HON (the merger agreement is mentioned in this SEC filing: `https://www.sec.gov/Archives/edgar/data/48305/0001047469-99-032694.txt`).

To warm up, we will briefly discuss the linear regression model (as we did in class). We will start with a small set of predictors: stock i's own returns up to three lags ($r_{i,t-1}$, $r_{i,t-2}$, $r_{i,t-3}$), and stock i's net trading flows up to three lags ($flow_{i,t-1}$, $flow_{i,t-2}$, $flow_{i,t-3}$). Stack these in predictor matrix X. We will refer to our target to predict as y, which is $r_{i,t}$ (current return of stock i).

To make sure you understand what is going on, we have provided a little table below to show you how to compare the linear regression model to the historical mean return model on the data set for each stock using rolling windows, with fixed window size w. We have written some simple equations in each step in the following table, but have omitted the equations for the predictions, errors, and cumulative errors for the linear regression model so you can check your understanding from the notes on out of sample forecasting. The comparison of cumulative out-of-sample squared forecast errors here is between the historical mean return model, and the linear regression model.

Steps	**Process**	**Historical Mean**	**Linear Regression**
Step 0	Window size $= w$		
Step 1	Training in each window	$\overline{y} = \frac{1}{w}\sum y_s$	$\widehat{\beta} = (X'X)^{-1}(X'y)$
Step 2	Prediction	$y\_pred_t = \overline{y}_{t-1}$	$y\_pred_t = ?$
Step 3	Sqd. Forecast Error	$error^2 = (y\_pred - y\_true)^2$	$error^2 = ?$
Step 4	Cumulative RMSE	Cum. $RMSE = ?$	Cum. $RMSE = ?$

Now you have the basic steps to start predicting equity returns using a linear model. You are at liberty to select the window size w when you are performing the regression, but I would caution against "snooping" this size outside of cross-validation.[1] Using the

[1] Hint: Choose w such that you have enough observations to estimate the model (minimum of 30), but

errors of the two models, you can compute the out of sample R^2 and plot the cumulative RMSE differential between the linear regression and historical mean return models for all stocks as a warm-up exercise.

Next, we will extend our set of predictor variables from above (let us rename these to X^{small}) and add stock i's lagged cross-stock returns $\{r_{j,t-1}\}$ and net flows $\{flow_{j,t-1}\}$ $\forall\ i \neq j$ (and again add lags of your own choice) to get the full set of predictor variables, X^{full}. Machine-learning models may be much better suited to use the cross-sectional variation across all of these predictor variables (why? make sure you understand the dimensions of X^{full} and remember that the sample is fixed with time-series dimension T). Now you can replace the linear regression with predictions from other models, including the LASSO, Tree, Random Forest, or whatever models you like best!

Note that in order to do so, you will need to tune the parameters for these models using cross-validation techniques (in each training window). Convince yourself that you understand how to do this. To make sure you follow what needs to be done here, you can check your understanding using the parameters for the LASSO and the Random Forest approaches in the table below.

MODELS	Parameters	To be tuned
LASSO	?	alphas
Random Forest	?	min samples leaf

Regression or Classification?

One possibility is that predicting the noisy variation in daily equity returns is hard, but predicting general outperformance or underperformance is easy. To test this, one option you might try is to predict which percentile of the expanding window distribution the future stock return lands in.

Consider a training dataset that is 100 observations in length. For this time period, you can compute the 33 and 66 percentiles of the distribution of stock-level equity returns. For example, the 33 percentile might be a -2% return, with all observations below this percentile being lower than this number, and the 66 might be +2% if the distribution is roughly symmetric.

Your goal would then be to predict whether the out of sample equity return return falls a) below this 33 percentile b) between 33 and 66, or c) above the 66 percentile, i.e., below -2%, between -2% and +2%, or above +2%. A simple benchmark to beat would be a basic prediction that the previous day's equity return percentile position is the best prediction of the next day's percentile position. That is, if the previous day's return was +3%, then the next day's return is predicted to land above the 66 percentile.

Does this approach seem to improve the performance of the ML models?

small enough to have a meaningful period over which you can evaluate the performance of the model.

Understanding performance

If you have successfully implemented these steps, you should be in a position to analyze the performance of various machine learning techniques for daily stock market returns! How would you compare the predictive performance of the various models to one another?

Since this is a finance course, rather than simply a course in prediction methods, you should be able to understand and explain, wherever possible, the underlying *economic* forces at work which generate the predictability of these variables. Do you expect any challenges to implement your strategies in the present day? What else might you want to consider?

1 Useful references

John Y. Campbell, Tarun Ramadorai and Allie Schwartz, 2009, "Caught on tape: Institutional trading, stock returns, and earnings announcements", *Journal of Financial Economics*. `http://www.tarunramadorai.com/TarunPapers/caughtontape.pdf`

Alex Chinco, Adam D. Clark-Joseph, and Mao Ye, 2018, "Sparse Signals in the Cross-Section of Returns", *Journal of Finance, forthcoming* `http://www.alexchinco.com/sparse-signals-in-cross-section.pdf`