Equity Strategies

Market Timing Strategy, Analyst Earning Forecasts

## Objective

This is a two-part project on equity strategies. In the first part you will investigate a market timing strategy. In the second part you will investigate the effects of analyst forecasts on earnings.

* **Part 1:** In the first part of this this project your goal is to verify the results of Blair Hull and Xiao Qiao in their [2015 paper on Market Timing](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2609814) via a regularized regression on a set of timing indicators. You will gather as much data as you can in order to start with as close as possible dataset as they did, and you will see if you can replicate their results. You will also experiment with different ways to regularize your model. Finally, you will see what the model predicts for today’s market.
* **Part 2:** In the second part of this project you will investigate the behavior of stocks in the S&P 500 around earnings announcements conditional on analyst recommendations beforehand. You will analyze the pre- and post- earnings drift around earning announcements.

Submit all your code, reports and presentation in the directory datasets/equity/LT\_equity\_projects/ on the shared Google folder.

## Data Description

**Part 1: Market timing indicators**

You will need to do download the following historical variables in the datasets/equity/market\_timing\_sp500/team\_(teamnum) folder (where teamnum=A,B depending on which team you are from):

1. Dividend-Price Ratio.
2. Price-to-Earnings Ratio.
3. Book-to-Market.
4. Cyclically Adjusted Price to Earnings (CAPE)  
   all 4 above can be found here: <http://www.multpl.com/>
5. Bond Yields:  
   - For annual data use Shiller’s long term stock, bond, interest rate and consumption data going back to 1871 <http://www.econ.yale.edu/~shiller/data/chapt26.xlsx>  
   - For monthly data since 1971 look at [FRED data, also found on Quandl](https://www.quandl.com/data/FRED/DGS10-10-Year-Treasury-Constant-Maturity-Rate)
6. Default spread Aaa-Baa. You should be able to figure it out from the following data (or similar) from Quandl:  
   - AAA [from Quandl](https://www.quandl.com/data/FRED/AAA10YM-Moody-s-Seasoned-Aaa-Corporate-Bond-Yield-Relative-to-Yield-on-10-Year-Treasury-Constant-Maturity-DISCONTINUED)  
   - Baa spread [from Quandl](https://www.quandl.com/data/FRED/BAA10YM-Moody-s-Seasoned-Baa-Corporate-Bond-Yield-Relative-to-Yield-on-10-Year-Treasury-Constant-Maturity-DISCONTINUED)  
   - Recent Baa [data from Quandl](https://www.quandl.com/data/FRBP/SPR_BAA_TBOND_MN-Yield-Spread-of-Moody-s-BAA-Corporate-Bond-over-10-Year-Treasury-Bond-Mean-Values)
7. Term spread. You can find the term spread or calculate it from the quandl, in the same databases that the data in #5 and #6 is stored.
8. Cointegration Residual of Consumption, Assets and Wealth. Can you find this indicator anywhere in the public domain?
9. You will need to find VIX and OHLC daily SPX prices in order to construct VIX-vol forecast spread, as outlined in the paper. Both VIX and SPX prices you can find on Quandl.
10. I[mplied correlation on quandl](https://www.quandl.com/data/CBOE/SP500IMPCOR-S-P-500-Implied-Correlation-Index-Fixed-Maturity).
11. [Baltic dry index](https://www.quandl.com/data/LLOYDS/BDI-Baltic-Dry-Index).
12. Can you find New Orders / New Shipments data on public domain which is free of revision biases? Please include this if you can.
13. Principal Component of Technical Indicators (PCA-tech) indicator. Please do this last if you have time as it may take some time to replicate.
14. Consumer Price Index [from quandl](https://www.quandl.com/data/FRED/CPIAUCSL-Consumer-Price-Index-for-All-Urban-Consumers-All-Items).
15. [GSCI index](https://tradingeconomics.com/commodity/gsci) in order to compute the stock-to-comodity price ratio PCR
16. Oil futures data. This should be available from the futures dataset project in their folder. Please use their data. Make sure you copy whatever you need into your folder instead of moving it or else you will have to re-generate any missing data.
17. Short interest data [from quandl](https://www.quandl.com/data/FINRA/FNYX_MNKD-FINRA-NYSE-TRF-Short-Interest-MNKD)

**Part 2: Analyst forecasts and earnings**

The data for this part is in datasets/equity. In this folder you will find the CRSP S&P500 price time series from 1970 to 2017 dataset, the IBES analyst forecast and earnings dataset, and the Compustat S&P500 constituent dataset. Make sure to read the README file for description of each dataset and how to link them using the Link Tables dataset

## Methodology and Deliverables

**Part 1: Market Timing Strategy**

### Read the Hull Paper

Blair Hull is a practitioner and the founder of Hull trading so this was a popular paper when it first came out. Read it carefully as in this project you will attempt to replicate its results. The data indicators are all described in the Hull and Qiao paper, so go through the description of which and understand how it is obtained.

### Gather the Data

Download all the data you can. If you find more data than in the above list, please add to the list and to your data\_gathering.ipynb notebook. Make sure you describe the datasource of each dataset. Save the data in the datasets/equity/market\_timing\_sp500/team\_(teamnum) folder.

### Pre-process Your Data

In a data\_analysis.ipynb notebook create a sklearn pipeline (similar to the one shown in class) that preprocesses your data as described in the paper. If needed add any imputers or scalers in your pipeline. Split the data in Training set (until Jan 2015) and test set (anything after that). Don’t look at the test set until the very end when you have trained all your models on the training set and you are ready to give predictions for today’s market.

Re-create Table 1, Table 2 and Table 3 with your data. Are your results the same as in the paper? Where do they differ? Comment in your notebook.

### Create Predictor Classes for Each Model

As shown in class you can extend the Predictor classes by implementing, fit(), transform() and predict methods. The linear regression is already implemented in sklearn, but the Correlation Screened Regression, and Real-Time Correlation Screening Regression are not. Implement them in a class. Finally, add an elastic net linear model.

### Backtest Your Models

Run a backtest on all your models. For the regression screening models vary the screening threshold. For the elastic net, vary your regularization ratio parameter. Obtain pnl for each model which you can present in class.

### Present your methodology and results

In the first half of your talk, you will present methodology, data, and results for Part 1 of your project.

**Part 2: Analyst Earnings Forecast Study**

Separate all the data after 2014- as an out of sample test set.

### Pre- and Post- Earnings Drift

It has been known in the academic literature for quite sometime (e.g. [see Bernard and Thomas](http://sites.fas.harvard.edu/~ec970lt/Readings/Post_Earnings_Announcement_Drift/Bernard%20and%20Thomas%201989.pdf) ‘89 paper for example which cites evidence dating back to 1968) that stock prices with big earnings surprises exhibit both pre-earnings drift and post-earnings drift. Pre-earnings drift is not as surprising, especially in the presence of hidden information. However, post-earnings drift is puzzling because the drift goes in both directions. If the post-earnings drift were only going upwards, then one could explain it as a consequence of the risk premium that the market is pricing into the asset. However, the post-earnings drift seems to go both ways: Stocks with positive surprise earnings tend to continue to drift up, while stocks with highly disappointing earnings tend to continue to drift down.

Your goal in this section is not to have an explanation for the effect (although make sure you read Bernard and Thomas), but to:

* Verify the results in Bernard and Thomas. For this you will need data before 1989. Use the IBES and CRSP data since 1970.
* Analyze the magnitude of the effect since the paper was written.

Since the S&P 500 composition was different in 1989 than it is today, you will need point-of-time data. Speak to the instructors about how to get that data. One simple way, for example, would be to look at the top 500 stocks by average daily value traded.

From the daily price time series and the earnings announcement dates, construct a table of the following format:

* The multi-index is [ticker, ern\_date, qnum], e.g. the ticker of the stock, the earning announcement date and the quarter number of the announcement, e.g. 1,...,4
* There are 61 columns ranging from -30,...,-1,0,1,....,30 corresponding to the number of days after the earnings announcement. 0 corresponds to the day of the earnings announcement, e.g. the ern\_date

Select earning dates before 1989 and normalize the stock prices to be 1 at each earning date. Select the stocks with the highest 10% of 1-day price jumps between -1th and 1th day after earnings. Note that because earnings announcements can happen pre-market, a stock can jump between the -1th day close and 0’th day open. A post-market announcement can cause the stock to jump between 0th day close and 1th day open. This is why we sort by the price move between days -1 and 1.

* For each year between 1975 and 1989, plot the full 61-day time series of those top- and bottom- 10% quantile move stocks. Do you get an agreement with Bernard and Thomas?
* Do the same plots for every year since 1989. How has the pattern changed recently?

### Conditioning on Analyst Surprise

Most of the post-announcement move happens right after the announcement so in the previous section you would be missing out on alpha if you waited that long. In this section you will condition on the analyst distribution and its surprise during the announcement.

* For every earnings announcement create the analyst earnings forecast distribution before the announcement. E.g. the histogram of all the analysts’ recommendations for a given stock. Do you need to use the median or the mean analyst forecast as a predictor of the actual earnings? Explain.

Your tasks will become much simpler if you create the following two dataframes

* df\_aforcast:
  + With index [ticker, ern\_date, quarternum, analyst\_id]
  + And a single column, analyst\_ern\_forecast
* df\_actual\_ern:
  + With index [ticker, ern\_date, quarternum]
  + And a single column, actual\_ern\_value

After you create these dataframes

* Merge them and obtain different metrics of the analyst distribution and its
  + For example, you can try median forecast - actual, mean forecast - actual, but these two numbers are not dimensionless so you can’t compare them between stocks. Instead, you will have to normalize by some measure of the spread of the forecast distribution, such as the standard deviation, etc. For each choice of your surprise measure, make sure you document it and justify it.
  + Also obtain measures of the skewness of the distribution, such as median-mean (normalized by a measure of the spread of the distribution).
  + Finally, obtain measures of the kurtosis of the distribution, properly normalized so that they are dimensionless
* How do your pre-post earnings price time series from the previous section look as a function of conditioning on different values of the surprise metrics (and distribution metrics) above?

### Train Predictive Models

Train several models to predict the 10-day post-earnings time series as a function of the actual announcement and the analyst forecast distribution. Use two types of training datasets:

* Stylized surprise metrics dataset. You will use as input the different metrics of the analyst distribution and its surprise metrics that you calculated for every earnings event in the previous section. As output variable you will use a categorical variable y which is 1 every time the 10-day post-earnings stock return is in the top 10% or bottom 10% for the given quarter.
* Full distribution dataset.
  + Your input features are:
    - The relative proportions of number analysts in each of the 10 quantiles of the forecast distribution. The sum of these proportions should be 1.
    - The total number of analysts in this distribution.
    - The standard deviation of this distribution.
    - The full range of the distribution, e.g the max\_forecast-min\_forecast.
    - The actual earning announcement.
  + The output feature y is again 1 every time the 10-day post-earnings stock return is in the top 10% or bottom 10% of the given quarter.

Clearly you can’t possibly know what the actual top and bottom 10% of price moves will be before you observe all the price moves for the quarter. However, if your model is good, you will be able to *predict* how likely it is that a given earning was in the top or bottom 10%, so you will have a tradeable signal.

Create a pipeline which lets you select which training dataset you will use. For each of the training datasets backtest an SGDClassifier with both hinge loss (SVM) and log loss (logistic regression) on all the data *before* the given quarter and test your model on the current quarter. This means you will have to retrain your model on every new quarter. Create a pipeline which will let you find the optimal parameters of your model for each training quarter. You will end up with a time series of model hyperparameters. Are the optimal hyperparameters stable?

Construct the precision-recall curves for your optimal SVM and logistic regression models trained on each of the two training sets above (4 P-R curves total). What model and which training dataset work best?

What would be the PnL of the strategy which trades based on the recommendations of the optimal model? Find the PnL in-sample. Finally, what is the PnL out of sample, e.g. on data after 2014?

### Present your methodology and results

In the second half of your talk, you will present methodology, data, and results for Part 2 of your project

README file

Create a README\_db.txt file containing:

* The names and emails of all the teammates
* Description of the file naming convention, fields for each dataset
* Any comments on data features other users should be aware of when they use your data.
* Instructions about how to run the different notebooks and look at the results.