A Data Scientific Approach to Equity Backtesting Research

Riaz J Arbi[a]

[a] *University of Cape Town, Cape Town, South Africa*

Abstract

Contemporary research into the cross-sectional variation in stock returns is fraught with replication challenges. This makes it difficult to validate important results and hampers the advance of knowledge in the field. This project addresses some of these challenges by fitting the backtesting research process to a data scientific workflow. By making extensive use of standard open source statistical programming languages the authors provide an environment in which total replication is trivial and common statistical errors are avoided by default.

Keywords: backtest, data workflow, overfitting, historical simulation, replication

JEL classification G12

This is a working paper. It is extremely messy and not intended to be read by anybody. It has been made publically avaialable to facilitate discussion among collaborators on the project.

1. Introduction

The objective of this thesis is to rebuild the Equity Asset Management research and trading pipeline in a Data Scientific way. The final body of work should provide a set of best practices guidelines that Asset Managers can use in production to improve their processes. This project has two objectives.

- The first is the rigorous documentation and construction of an equity research environment.
- The second is to engage in a process of discovery in the equity research space with a view to uncovering interesting topics to drill down to for future research.

It is expected that this body of work can sustain several thesis topics. I have structured it in a modular way so that I can start at the beginning of the data science stack and move towards actual

*Corresponding author: Riaz J Arbi
Email address: riazarbi@gmail.com (Riaz J Arbi)

analysis. But if I don't get that far then I think the building and documentation of an equity research environment would be of considerable use to academics and professionals in the field.

2. Modular Approach

A guiding principle in the construction of my thesis topic is "don't reinvent the wheel." My topic is structured as a series of layers. Each successive layer is enabled by the layers below it. If a solution exists for any particular layer, the existing in-the-wild implementation can be documented and implemented and I can move on to the next layer. If a layer has not been implemented, or if no readily accessible open-source solution exists, I will endeavour to document a solution and compile a set of best practices.

My thesis will serve several purposes. It will be -

1. A collection of methods and recommendations for conducting equity research in conjunction with common statistical software packages.
2. An instruction manual for building an equity research environment to current statistical best practices.
3. A working open source system which can be cloned by equity researchers to enable furthering of the discipline of equity research.
4. A demonstration set for what can currently be achieved in the space using existing open source tools

The layers, ordered according to the Data Science stack, are as follows -

1. Documentation and construction of an updatable dataset that is free from survivorship and look ahead bias, accurately models slippage and transaction costs.
2. Implementation of an event-based backtester that can compute optimal trades at each rebalancing period using both market and fundamental data.
3. Evaluation of various machine learning prediction methods for their alpha-generation capabilities.

3. Proposal

Although each of the layers mentioned above appear simple, they each contain deep, difficult to surmount challenges. I expect that each of these layers are sufficient for a full thesis. I would like to spend the first four months of the 2018 calendar year building a data pipeline and research environment.

The objective of this phase is to construct a working research environment and to search for interesting areas of further research. In the fifth month, I will review the potential topics I have uncovered and select one for further research. If no topics are suitable, I will refactor the code generates in the platform building phase, thoroughly document it, and submit it as my thesis. In this fallback scenario my thesis will be an open source software package and manual for the construction of a full equity research environment, form data collection through to trading. I intend to release this as open source software to serve as a guide for future researchers and enthusiasts. I have already identified the following topics that could be interesting subjects for deeper study:

3.1. Dataset Construction

Clean, accurate datasets are important for both the backtesting and management of investment portfolios. The absence of such a dataset severely limits the management capabilities of an asset management firm and renders most academic research into systematic investment management useless. By building and documenting best practices for equity dataset construction, I hope to enable quality, real-world applicable research in my academic peers and to provide the tools that asset managers need to rigorously apply portfolio management theory to practice.

Equity asset managers typically measure their investments against a benchmark. However, most asset managers obtain the benchmark timeseries from data vendors, and if they do have the ability to create it in-house, it is a manual human-driven process.

Data from vendors is often contradictory.

Problems to be solved in this section would include -

3.1.1. Selection and construction

Selection and construction of an appropriate database solution weighing speed and simplicity against consistency and atomicity. Review of existing technologies and schema. Research into the appropriate ways of dealing with validation time versus transactional time and define a method for dealing with contradictory data from multiple data sources.

3.1.2. Automate

Automate the consumption of underlying constituent data and the building of the benchmarks from first principles. Consume multiple sources and use selection or averaging rules to reconcile conflicting datas sources.

3.2. Building a Backtester

Backtester construction is well documented, and there exist some excellent open source solutions. However, these solutions tend to be focused on US markets, and they lack certain capabilities. The primary shortcoming of these packages is that they tend to be security-centric, not benchmark-centric. For instance, zipine, the popular python algorithmic trading package, currently provides no functionality for simply executing a passive market-cap-weighted investment strategy. Baskets of stocks can only be specified by manual lists. This renders the zipline package useless to professional asset managers, since modern portfolio management is centered around active-weight deviations from a benchmark. The starting point of any portfolio should be a benchmark portfolio, and a trading algorithm should specify deviations from the benchmark.

Problems in this section would be -

3.2.1. Survey

A survey of existing implementations, along with a summary of capabilities and shortcomings

3.2.2. Assessment

Assessment of these systems from a modern portfolio theory perspective

3.2.3. Building

Leverage the work done in dataset construction to build an event-driven backtester that allows an asset manager to rigorously backtest investment ideas.

3.2.4. Evaluation of various Machine Learning prediction methods

This topic can only be tackled once the above two topics have been dealt with. There is a fair amount of research on the implementation of machine learning algorithms for the prediction of stock prices. Research in this area appears to be fraught with poor data input and inaccurate modeling of transaction and slippage costs. This section would involve selecting several heavily-cited papers in the space and attempting to replicate them using our dataset and trading models. I am especially interested in whether results are replicable, and what the impact of real-world transaction and spread costs would be.

3.2.5. Retooling

Retoolthe portfolio implementation and research process along the lines of the Good Judgment Project. This would be more of a theoretical topic. There has been a lot of good work done at the Good Judgment Project on the process of forecasting, and on the management of professional forecasters. See https://www.edge.org/conversation/philip_tetlock-a-short-course-in-superforecasting.

Equity analysts and portfolio managers are professional forecasters, but there is very little quantifiable measurement of the forecasting abilities of these experts. Typically, analysts are simply measured on the excess return of their stock picks relative to the universe. The term for this is attribution, and it is the answer to the question, "what percentage of our fund's return is attributable to each analyst?".

The Good Judgment Project has developed a much more fine-grained approach to measuring forecasting, and they recognize that a forecast can be broken into separable components. In the equity price forecasting space, these component would be, for example, timing, direction of move, magnitude of move, underlying driver of move. A poor stock picker may be a very good predictor of underlying drivers of stock returns, but a poor predictor of timing responses of stocks to drivers. This topic would constitute a literature review of the new work done at the Good Judgment Project and translate that work into a model framework for recording the motivations for stock picks. Since the thesis has to be completed by the end of 2018 I do not think I could accumulate enough data for an evaluation of the model.

3.2.6. Extension into a multi-year research project

I use this research as a springboard for getting actual asset managers to pilot the frame- work and create an ongoing research project on the impact of this paradigm for analyst accuracy.

I could integrate the retooled research process into an automated trading process. Trades can only happen through the bet-logging tool; no trade occurs if any fields are null; entry and exit form the stock are determined by the bet-logging information. If an analyst wants to extend / reduce / cancel a bet, a new bet must be created. For instance, a long bet in gold can only be cancelled by a short bet in gold. This basically makes deviation from rational investing an 'opt-in' option: if no action is taken, no behavioural biases are introduced.

The historical bets can be cubed by portfolio, analyst, time period and stock industry. and can be plotted against the benchmark. We can use statistics to determine who the good, well calibrated analysts are. We can mine this data for all sorts of interesting in- sights. Does a particular analyst generally get their initial bets correct but stymie their bets with re-bets? Then they don't have a handle on their behavioural biases. Does an analyst generally get their pharmaceutical bets right, but get their iron-ore mining bets wrong? Move them away from mining.

4. Concluding Remarks

The topics touched upon above are by no means a comprehensive account of what I think can be done in the space. But it should serve to convince the reader that there is massive scope for data science in the asset management and equity research space, both in the portfolio management workflow and in research. I would argue, in fact, that there is a significant overlap in the work that equity analysts do and that which data scientists do.

Too often a finance graduate glosses past the data collection, cleaning and modeling phase of their project in order to get to their problem of interest. This invariably results in results which are worthless in the wild and is primary contributor to the maxim that "every academic backtest yields alpha, but none actually makes money". My singular conviction is that equity researchers need to be familiarized with data science tools; they need data science-grade environments and data sources. I believe that it would be a tremendous contribution to this field if I rigorously document how a researcher can make that transition. A secondary goal is to actually implement the environment and then start using the tools of data science to solve problems in the areas of asset management and equity research.