A Data Scientific Approach to Equity Backtesting Research

Masters in Data Science (STA????W) Research Proposal

University of Cape Town

Author: Riaz J Arbi
Supervisor: Associate Professor Tim Gebbie

Abstract

Keywords

1. Hypothesis

- The replicability issues surrounding academic research into the cross-sectional variation in stock returns can be largely mitigated by adhering to a data-scientific approach to data analysis.
- Robustness of results in the field of stock return research can be can be significantly improved by accounting for the risk of overfitting as the number of trials increases.

2. Literature Review

Replicating Anomalies Financial Charlatanism

3. Aims and Objectives

This project is firmly rooted in the meta of finance research. The objective is not to validate whether particular anomalies in the cross-sectional variation of stock returns exists. Rather, it is to outline and implement a system wherein researchers can investigate these questions in a statistically rigorous manner.

1. Survey of current academic backtest methods (see the github dissertation repository)

- A critique of the challenges around replicability because of lack of documentation.

- A critique of the challenges around validity because of poor statistical methods (IS/OOS).
- Discussion on how these challenges can be mitigated using standard data science tools.

2. Documentation of an working demonstration system that mitigates these challenges (see the github code repository)

- Source code along with README documentation of every phase of the project will be released to a public repository under and Apache 2.0 license.

3. An original replication case study which makes use of the demonstration system to replicate a widely cited academic paper in the field (see the github replication example)

4. Data Requirements Specification

All stages of the data collection and transformation will be transparently documented and source code will be made avaialble on a hosted repository.

All raw data wll be programmatically extracted from a Bloomberg Terminal using the Excel Add-In. Validation of this data will be conducted by randomly selecting ten financial reports from the universe and cross-checking the contents of those financials against the raw data.

All data cleaning and interpolation will be done using the dplyr and tidyr packages in the R statistical computing language.

5. Systems Requirements Specification

Hardware Requirements

```
- A computer running an x86 processor
- 50gb of hard drive space
- At least 8gb or RAM
```

Software Requirements and Packages Used

```
- Access to a Bloomberg Terminal with Excel installed (does not have ot be the development machine)
- A research machine running
   - Ubuntu 16.04 LTS
   - Jupyter Server
   - RStudio Server
   - Python 3.5
   - R 3.4.3
   - Nextcloud 12 for transferring data from the Bloomberg Terminal to the research machine
```

Software Development Framework, configuration control and version control

The primary objective of this dissertation is the creation of a script-based workflow that improves the equity backtesting research process. The complete codebase, as well as the dissertation and demonstration case will be made avaialble on a GitHub repository. The code will be released under an Apache 2.0 license.

The data will not be released due to data vendor licensing constraints. However, macro-enabled Excel workbooks that programmatically query the Bloomberg Excel Add-In for relevant data will be made avaialable in the public repository.

Version control of all project deliverables will be managed using the Git version control system. Commits will be pushed regularly to the publically avaialable GiHub repository to ensure timeous backups of the

6. Project Milestone Deliverables

7. References