# Reproduction-based Learning and Replicating Investment Strategy Research

**Joey Bortfeld**
Georgia Institute of Technology
New York, USA
jbortfeld3@gatech.edu

**ABSTRACT**
We present a framework for using an existing research study as the centerpiece of a student project. The student's objective is to replicate the study and in doing so, actively learn and experience the research process. The project supplements the research piece with supplemental content as well as critical thinking exercises to assess mastery of the material.

**INTRODUCTION**
This paper will discuss a project to guide OMSCS Machine Learning for Trading (ML4T) students through understanding and replicating the paper "Return Predictability and Market-Timing: A One-Month Model" by Hull, Qiao and Bakosova (2017). The goals of this project are 1) facilitate student learning by reviewing the papers content and providing supplemental content to expand upon it 2) use engagement techniques so that students learn by actively collecting and analyzing the data and deciding what methodology to use and 3) demonstrate a framework for educators to use for future replication-based learning projects. The project uses jupyter notebooks to deliver text content, multiple choice quizzes, and essay questions, and to provide a workspace for code to download, analyze and chart data.

The project will attempt to address some of the current deficiencies in finance education whereby course content is poorly connected to real life financial issues and students are disengaged from actively participating in their education.

This paper will proceed with a discussion of some of the challenges in teaching about investment strategy, then provide a blueprint to address those issues, then samples and a discussion of the actual project content, and finally a discussion of weaknesses and areas for further development.

**Challenges - Lack of Classroom Content**
There is not presently a resource that guides students through the practical aspects of doing a research project in financial trading.

Investing and trading assets such as stocks and bonds is important for the financial security of everyone. It's an avenue for people to generate new income streams, grow their accumulated wealth and afford their retirement. Furthermore, investing is near the heart of capitalism and directs savings from people to the businesses or governments that need it to fund investments and expand operations.

Fortunately there are many resources that will teach very basic financial concepts such as what stocks are, the importance of investing in low fee products, etc. The target audience for these lessons are mostly individuals who want to manage their own financial assets. For individuals that want more advanced lessons, on the level of how professional money managers work, there are unfortunately far fewer options. Specifically, there is a gap in instruction on how to construct successful trading strategies. Part of this is completely understandable given the competitive context: people generally don't want to share profitable strategies with others.

One existing option for learning about trading strategies is to take massive open online courses (MOOCs) in economics or finance. For example, MIT has freely published class lecture videos from its MBA-level Financial Theory course by Andrew Lo. Many of these courses however are more concerned with the worthwhile exercises of understanding markets from an academic perspective rather than the related, but more specific, task of devising profitable strategies.

Financial trading and investment draws from several disciplines, first amongst them economics and finance. Existing teaching methods employed in economics and finance classrooms however are known to suffer from several problems that lead to a lack of student engagement and shallow understanding of the material.

- "The vast majority of the literature seems to suggest that the style and the method of teaching" lead students "to perceive economics as a difficult

subject with negative repercussions on their performance." (Becker & Watts, 2001)

- In particular, the "chalk and talk" teaching method whereby a presenter lectures for the duration of a class while writing equations on a chalkboard has been singled out for not engaging higher order thinking skills such as applying, analyzing, evaluating, and creating.
- Lecture-style classrooms do not train students in the practical everyday tasks of economists who must collect hard-to-find data and utilize computers and programming to analyze data (Becker & Watts, 2001)
- Classroom economics lectures ignore current events and therefore lack context and relationship to the real-world. For example, classroom topics may cover the mathematics of economic growth and inequality without connecting it personally relevant experiences such as job loss or indebtedness. (Becker, 1998) (Carlson, 1999)
- Student assessment (testing) in economics does not encourage critical thinking and tends to favor tests of recall and reciting information back to the instructor rather than assessments that invoke "proofs of independent judgement." (The Economist, 2017)

**Solution**

We use an adapted case-based approach to address some of the above problems in economics instruction. Case-based learning is used widely in legal and medical education but less so in social sciences like economics. Cases come in two broad varieties. First, a narrative case is at its core a story about what individuals and groups did when they were confronted with a problem and involves the information available to them, the actions they took to collect and analyze data, and the decisions they took to try to solve the problem. The case serves as an opportunity for students to learn from this past experience. Second, a simulation case puts students in the case itself and challenges them to act as decision makers. Adaptations of this method also involve withholding information about the decisions the actors actually made so that students can make their own independent decisions based on the presented data.

The purpose of cases is to engage students with hands-on, realistic experiences by noting that "the most vivid and powerful lessons from our own educational experiences are related to projects in which we were actively involved. Concepts, ideas and experiences are harnessed and clarified in our mind more easily and quickly through direct experience than through the reading of books and abstract theories and concepts." (Volpe, 2002)

Case studies are typically implemented as classroom debates or simulations but for our project we adapt the simulation case method as the inspiration for a project undertaken by a sole individual. Here, the case is the research paper and students will simulate performing the research itself, critically assess the methodology, and execute the analysis that they feel best for addressing the main hypothesis of the paper. Our project will address deficiencies in prior instructional methods by:

- Eschewing the lecture-based instructional method and instead engage students with hands-on activities.
- Making difficult material more accessible by chunking material into smaller modules.
- Tasking students to collect data, write code, and analyze data to simulate the real work of researchers.
- Focusing on a actual trading strategy to excite students and illustrate the real-world applicability of their analysis.
- Encouraging students to disagree with the paper's methodology and then investigate alternative ways to conduct the research.

From a pedagogical perspective we lean heavily on Bloom's taxonomy of learning to engage six separate cognitive processes in order to facilitate deep knowledge development.. These six cognitive dimensions are remembering, understanding, applying, analyzing, evaluating and creating. Accordingly, there will be activities for students to engage in throughout the project including quizzes, essays and coding exercises. For example, below are sample questions related to each cognitive dimension:

- Remembering: "Identify some of the claimed results in the paper abstract from the following multiple-choice list."
- Understanding: "Outline in pseudocode the process for running a dynamic, weighted-least squares regression as detailed in the paper."
- Applying: "Implement in code a process to download and combine the mentioned datasets."

- Analyzing: “Explain in an essay your regression results.”
- Evaluating: “Evaluate in an essay whether the trade’s performance is consistent with the results claimed in the abstract.”
- Create: “How would you design a new methodology for testing the hypothesis if you disagree with the author’s approach?”

**Methodology**

We design a project for students that leverages the aforementioned pedagogical techniques and principles to further and deepen student learning experiences. The project revolves around a research article and immerses students into it from the perspectives of both the student and researcher. At a high level, the student’s learning experience entails:

- Reading the research article including methodology, results, and interpretation sections.
- Understanding the design decisions the researchers took in their methodology and possible alternatives decisions they could have taken instead.
- Replicating the research itself by collecting data and writing code to process the data.
- Analyzing the data they have generated using analytical techniques from the paper and extensions thereof.
- Interpreting the results and drawing conclusions and contrasts between their data and the reported data in the research article.
- Developing further question and designing future research plans.

The selected research article for the project was “Return Predictability and Market-Timing: A One-Month Model” by Hull, Qiao and Bakosova (2017). The key purpose of the paper is to describe a trading strategy that predicts the forward return of the broad stock market based on a set of macroeconomic variables using an extension of a linear regression model. The paper then describes a methodology to invest as appropriate in an exchange traded fund (ETF). based on the model’s predictions. This article was selected for several reasons including its accessibility (open source research paper), feasibility (the research was based on public data), opacity (the methodology was open and clear on most items), appropriateness (the difficulty level was appropriate for the target audience), and relevance (the paper describes a trading strategy that is used to actually manage an almost $100 million exchange-traded fund or ETF).

The target audience for the project are masters students in Georgia Institute of Technology’s Machine Learning for Trading course. Though the material can be accessible to a broader audience, the narrower demographic allows us to assume intermediate to advanced skill levels in programming, data analysis, statistics, machine learning, economics and finance.

The project is delivered as a series of jupyter notebooks which enables content, assessment, coding, and analysis to be unified onto a single platform. Students can potentially perform coding and data analysis on a separate platform or in any programming language they want but Python is the preferred language and a natural fit for the jupyter notebook environment.

**Results**

The project is spread across five jupyter notebooks, corresponding to distinct sections of the research paper.

Notebook 1 covers the paper’s introduction which provides information on related works and an outline of the paper. The notebook provides supplemental content to explain cited research and student exercises involve downloading and charting data. An example student activity from this section is as follows and focuses on the cognitive skill of analysis:

- *Let's do some exploratory analysis now. If we're trying to predict the monthly returns of the S&P 500 as part of a trading strategy, then we should know what those returns look like. What's the mean? Are the returns bell-shaped? What are the maximum and minimum monthly returns? Examine the S&P 500 data using the SPY ticker, which is an ETF that tracks the index. At a minimum, plot the price and returns over time and plot a histogram that shows you the distribution of returns. If you want to be an "expert" on the S&P 500, what other charts would you like to see?*

Notebook 2 covers the paper’s description of the data used for the research. Supplemental content centers on best practices for properly assembling research datasets while student exercises involve writing code to download, process and combine data series into a dataset that can be used later. Below is a student activity for cleaning data that asks them to exercise creativity:

- *Now design a method to control for outliers. Often outlier control methods might entail just dropping data that “looks” like an outlier or to use winsorization (create a floor/ceiling to the data range). Sometimes people transform the data by taking a log or exponent of the data.*

*What you do will depend on your data. For example, taking a log value is useful when your data can have a wide range with a long right tail. This commonly happens when you're dealing with income distribution and you're trying to compare people who all usually make less than $100,000 but there is a small percentage of people that makes 100x that amount. Based on the data you have and the outliers you've identified above, implement a strategy to control for outliers.*

Notebook 3 addresses the paper's description of the dynamic weighted linear regression model used while supplemental content dives more deeply into the concept of unequally weighted modelling techniques. Student activities focus on estimating and comparing models that were calibrated on different historical datasets. Here is a student quiz that asks students to remember content from the reading:

*Which of the following are true?*

- *In the WLS formulation we use, older data points have lower influence on the model than more recent observations*
- *OLS is a special case of WLS but with a half life of 0.99*
- *In WLS we calibrate model coefficients by minimizing the sum of squared weights.*
- *The equity risk premium we use should be the return of the SPY ticker, to proxy the return of the S&P 500.*
- *We use WLS, rather than OLS, in order to represent more expressive, nonlinear functions*

Notebook 4 centers on the paper's method of variable selection and the supplemental content discusses why using fewer variables in a model is generally more desirable than using many. As a coding activity, students iteratively consider all possible models (based on which variables to include) and choose the optimal configuration according to different evaluation criteria. One assessment question in this notebook tests students understanding of the reading:

- *In variable selection we are picking which variables to use out of a set of all possible variables. This is akin to selecting which model to use out of a set of possible models. Hull et al. use 15 explanatory variables in their variable selection process. How many possible models (explanatory variable combinations) are they choosing from? What is the complexity of the "model space" as the number of explanatory variables grow?*

Finally, Notebook 5 concerns the paper's trading strategy stemming from the model's predictions. Additional content discusses various ways of measuring the performance of different investment strategies and the student activity section focuses on asking students to imagine ways to improve the research:

- *Did your results match the results of the paper? If it did not, what are some possible explanations? Do you think the paper's results are credible?*
- *Having gone through the process of replicating Hull et al.'s strategy, do you like it? Would you invest your money in it? Would you want to develop your own model further and invest in that?*

**Limitations**

The project is an example of using a published piece of research as the focal point for a learning experience involving reading, replicating, and critiquing. While this specific project is focused on a finance-related topic, it can potentially be applied to a variety of fields.

Future projects could improve on this demonstration project by:

- Incorporating multisensory teaching delivery mechanisms such as videos, audio, pictures, etc. This would break up the monotony of long text passages which tend to dull engagement and weaken student attention. Images and video can be integrated into the jupyter notebook platform so there are no technical limitations that prohibit this feature.
- Facilitating group collaboration. Given that scientific research and workplace organization are increasingly team-centric, the preparatory experience from a class project should reflect that. The paper chosen as the focus of the demonstration project itself has three authors. Reproduction-based projects can focus on organizational and management skills necessary to divide work, communicate and debate results, and negotiate interpersonal dynamics.
- Improving methods for communicating and sharing data and research results. One of the objectives of the project is to test existing research through reproducing the study. For future researchers that in turn want to reproduce, expand, or combine student projects, there should be an easy and efficient way to do this. Developing standards for data gathering and labeling as well as coding style and quality could help facilitate more cooperative research.

## Conclusion

In this paper we outlined the design of a student project to reproduce a piece of financial research. We surveyed some of the current weaknesses in financial and investment education including the lack of hands-on experience in the field. We proposed that our project would solve some of these challenges by immersing students in the process of testing investment ideas and having them apply financial analysis under conditions that simulate the real life difficulties and messiness of data. Finally, we suggest that future work should continue to add more additional content to enhance the material, possibly using new mediums such as video.

## REFERENCES


1. Becker, W. (1998) 'Engaging students in quantitative analysis with short case examples from the academic and popular press', *American Economic Review*, Papers and Proceedings, vol. 88, pp. 480–6.
2. Becker, W. and Watts, M. (2001) "Teaching methods in US undergraduate economics courses", *Journal of Economic Education*, vol. 32, pp. 269–80
3. Bloom's Taxonomy of Learning Domains. Knowledge Jump. Clark, Donald. June 5, 1999. Retrieved from http://www.nwlink.com/~donclark/hrd/bloom.html
4. Carlson, J. (1999) 'A case method for teaching statistics', *Journal of Economic Education*, vol. 30, pp. 52–8
5. The Economist. " The teaching of economics gets an overdue overhaul" Retrieved from https://www.economist.com/finance-and-economics/2017/09/21/the-teaching-of-economics-gets-an-overdue-overhaul
6. Volpe, G. "Case Studies" "The Handbook for Economics Lecturers". Retrieved from https://www.economicsnetwork.ac.uk/handbook/casestudies/welcome.