Data Science for Finance

SHIMON KOGAN, PHD
ASSOCIATE PROFESSOR OF FINANCE
THE WHARTON SCHOOL & REICHMAN UNIVERSITY

skogan@upenn.edu
Office 2429 SH-DH

Overview

Information is the bloodline of capital markets. The proliferation of data sources ("alternative data") along with the rise of new tools that help extract meaning out of such data ("machine learning") are therefore shifting industries that interface with capital markets. The course aims to prepare students for a wide range of careers in the financial industry and consulting, including asset management, and it places a strong emphasis on financial economics and data analysis.

The course is designed around a set of modules, each centered around a practical application relevant for capital markets. Students investigate these questions by thinking about the research design, implement it using Python and its ecosystem of packages (e.g., Numpy, Pandas, Scikit-Learn), and learn to map the analysis into measurable results. Each module goes through the data acquisition, data cleaning, visualization, and analysis process. Within each application, we develop a different machine learning approach and apply it. These include both supervised regression methods (e.g., Lasso, Ridge, Elastic Net), supervised classification methods (e.g., Decision Trees and Random Forest), and unsupervised methods (e.g., PCA).

In the second half of the course, student-teams work on a capstone project while using real-world alternative data provided by a number of data partners, see list at the end of the syllabus. These projects introduce students to alternative data relevant for asset management companies and have them identify creative and meaningful use cases for that data, and design and test these use cases. During that half of the course, there are no regular classrooms. Instead, teams setup weekly check-ins to plan ahead and share their progress and groups present around two milestones, such that you are exposed to the questions, methods, and results of other teams. At the end of the course, final project presentation is judged by a set of industry experts.

Requirements

This is NOT a coding course and the focus would be on financial research and how to design and conduct it properly. Programming is merely the last-mile tool that allows us to implement ideas and the approach to this part will be very pragmatic, focusing on the subset of most useful packages, procedures, and commands.

Programming knowledge is not a prerequisite but a desire to acquire that skill is. We will be using Python, a robust, open-source programming language often used for data science. To make the best of out of the course, students are advised to supplement the course with some basic skills from online resources, supplied materials, and optional review sessions at the beginning of the course.

Course Structure

The course mixes standard lecture, examples, cases, and guest lectures. Student are expected to work in teams and demonstrate a high level of independent learning and initiative. The course' goal is to provide students with in-depth understanding of how to integrate these technologies/analytics into new business ideas and help them be effective managers in an environment where these technologies are strategic to the organization.

Materials

During the course, I will share a large number of articles and academic papers. In addition, you may find the following free sources helpful:

- "An Introduction to Statistical Learning"
  https://link.springer.com/book/10.1007/978-1-4614-7138-7
- "Whirlwind Tour of Python"
  https://jakevdp.github.io/WhirlwindTourOfPython/
- "Python Data Science Handbook"
  https://jakevdp.github.io/PythonDataScienceHandbook/
- "Data Science: Theories, Models, Algorithms, and Analytics":
  https://srdas.github.io/MLBook/

Use of Generate AI

You may use generative AI programs (e.g., tools like ChatGPT) to help with coding and to generate ideas. However, you should note that the material generated by these programs may be inaccurate, incomplete, or otherwise problematic. Beware that blind use may also stifle your own independent thinking and creativity. In particular, please do not submit code that you do not understand in details.

You may not submit any work generated by an AI program as your own. If you include material generated by an AI program, it should be cited like any other reference material (with due consideration for the quality of the reference, which may be poor). Any plagiarism or other form of cheating will be dealt with severely under relevant Penn policies.

Grades

Ground rules:

- Class attendance is mandatory. Up to two absences are excused. More than four absences/late attendances will automatically result in a failure grade.
- On time arrival. Please be at your pre-assigned seat on time. Late arrival will be marked as an absence.

Grades will be determined based on:

(I) **Class Participation and assignments — 30%**
To obtain maximal class participation grade, you are expected to (1) participate in a way that promotes collective learning, and (2) be prepared to discuss and share your analysis/insights about the assigned readings, (3) complete the short, weekly, assignment.

(II) **Mid-course exam — 30%**
The exam will in-class, on Feb 22, 2024. Please note the scheduling of the exam. You are responsible for ensuring that you are available to take the exam as no make-up exam will be offered. The exam will be code and data driven.

(III) **Capstone Project — 40%**
The group project will have you apply the tools covered in the course to real-world

data. The evaluation will be based on the milestone presentations, the questions asked, the level of analysis, the results obtained, and the presentations made. Specifically, the capstone project grade will based on the following:

- Presentations and judges assessment (20%)
- Originality and importance of the main research questions (20%)
- Quality of the empirical design and analysis (40%)
- Quality of the interpretation and the conclusions drawn (20%)

To group-level survey, conducted at the end of the semester, will determine the allocation of project grade across group members.

Preliminary Meetings' Outline

1. **Motivation**
    1. the increase role of data science in finance and the rise of quantitative investing
    2. introduction to Python

2. **How did retail investors respond to the coronavirus crisis?**
    1. loading data and calculating returns — dividends, splits, adjusted closes
    2. merging data sets
    3. visualizing data

3. **Can we time international equity markets?**
    1. using APIs to obtain international stock returns
    2. ols and overfitting - bias-variance tradeoff, problem identification and potential solutions
    3. lasso, ridge and elastic nets
    4. trading on ML predictions

4. **Which stocks to buy and which to sell?**
   From traditional quant factors to ML based stock selection
    1. portfolio sorts — single, double and conditional sorts
    2. combining factors (z-score, rank based)
    3. dimension reduction using PCA regressions
    4. Random Forest applied to characteristics
    5. implementation issues — turnover, liquidity, capacity, etc.

5. Capstone project

1. *Initial proposal, Due 3/2/2024*
   During the first half of the course, you will be introduced to a set of publicly available datasets that have been curated into a library that you can explore. Your group will need to choose one of these datasets as the primary source of application, or identify a different data set and vet it with us ahead of time. It is best if your team works on an application that excites you.

2. *Mid-project presentation, Due 3/28/2024*
   Present to the entire class your capstone project: describing the data and application ("what is the question?"), laying out the empirical approach ("how will we answer the question?"), and sharing preliminary findings ("generating a minimum viable product").

3. *Final course presentation complete report due, Due: 4/25/2024*

The final report should be up to eight pages long exclusive of tables, figures, references, and code, which are to be included as an appendix to the report. The report should include the following section:

- *executive summary* with the main question / objective, short description of the data and the empirical approach, and the main results.

- *introduction* that motivates the question while linking to existing findings in academic / practitioner literature

- *data and analysis section* that describes the different data sources and how the empirical analysis was set up

- *results section* describing the main findings and pointing to figures/tables that support the results; be sure to discuss what you had hoped / thought that you'd find but you did not

- *future directions* discussion of what you would have added based on the current results if you had another two months to work on the project

In addition, each group will have a regular weekly check-in to share its progress, discuss open questions, and plan for the following week. These meetings will take place over zoom.

Below is a preliminary list of data partners — companies who agreed to share unique dataset to be used for this course only:

- **Revelio Labs** (https://www.reveliolabs.com/): Revelio Labs absorbs and standardizes hundreds of millions of public employment records to create the world's first universal HR database, allowing us to understand the workforce dynamics and trends of any company.
- **Thinknum** (https://www.thinknum.com/): As companies move their business operations to the Internet, new data trails are being created that can provide unique insights on these companies. Thinknum Alternative Data indexes all of these data trails in one platform, providing investors with critical data points that others miss.
- **Similarweb** (https://www.similarweb.com/): A publicly traded company (SMWB) whose mission it is to deliver the most trusted, comprehensive, and detailed view of the digital world, so our customers can outperform their competition and win their markets.
- **Earnest Analytics** (https://www.earnestanalytics.com/): Partners with companies to transform the data from their core business into a source of actionable insights for institutional investors, consumer brands, management consultants, and government agencies, while focusing on consumer spending, retail product pricing, healthcare claims, and point-of-sale transactions.

Tentative Schedule

Session	Date	Topic	Other
	Overview		
1	1/16/24	Motivation	
2	1/18/24	Introduction to programming	Data partner presentation: Earnest Analytics
	How did retail investors respond to the coronavirus crisis?		
3	1/23/24	Retail investors trading - Part I	
4	1/25/24	Retail investors trading - Part II	Data partner presentation: Revelio Labs
5	1/30/24	Retail investors trading - Part III	Assignment 1 due
	Can we time international equity markets?		
6	2/1/24	Predicting international markets - Part I	
7	2/6/24	Predicting international markets - Part II	Data partner presentation: Thinknum Assignment 2 due
8	2/8/24	Predicting international markets - Part III	
	Which stocks to buy and which to sell?		
9	2/13/24	Which stocks to buy? - Part I	Assignment 3 due
10	2/15/24	Which stocks to buy? - Part II	Data partner presentation: Similarweb
11	2/20/24	Which stocks to buy? - Part III	Assignment 4 due
12	2/22/24	Mid-course exam	
	Capstone Project		
13	*TBD*	*Guest lecture*	*This lecture is scheduled outside of normal class time*
14	3/28/24	Mid-project presentation	

Session	Date	Topic	Other
15	4/18/24	Final project presentation	Attended by outside judges