FNCE 737
Data Science for Finance

SHIMON KOGAN, PHD
ASSOCIATE PROFESSOR OF FINANCE
THE WHARTON SCHOOL & IDC HERZLIYA

Office 2423 SH-DH
skogan@upenn.edu

TAs

Sajad Ghorbani

sajadgh@sas.upenn.edu

Joseph Federico

jfid@wharton.upenn.edu

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 using Python and its ecosystem of packages (e.g., Numpy, Pandas, Scikit-Learn). 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). A separate module will be dedicated to the use of unstructured text data.

In the second half of the course, students will work on a capstone project in groups with real-world data provided by a number of data partners, see list at the end of the syllabus. These projects will introduce students to alternative data coming from asset management and technology companies. The goal of these projects is to identify a use case for these data and build a validation for the use case. While working separately, groups will present around a set of milestones and at the end of the course such that you are exposed to the questions, methods, and results of other teams.

Requirements

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 for that. To make the best of out of the course, students are advised to acquire some basic skills from online resources, supplied materials, and optional review sessions at the beginning of the course. More advanced application will be articulated through examples in class.

All students will be required to pass a course entry quiz covering basic Python knowledge and finance concepts during the first week of class. Only students successfully passing the quiz will be admitted to the course.

The course entry quiz will cover basic mathematical operators in Python, set/list/tuple, arrays, dictionary, if/else, for/while, writing/calling functions in Python, print, and formatting, covered in:

- "A Whirlwind Tour of Python", chapters 1 - 11
- "Python Data Science Handbook", chapters 1 - 2

The finance questions will cover basic concepts in portfolio formation, risk-return tradeoff, and the CAPM.

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/

Grades

Grades will be determined based on:

(I) **Class Participation and Assignments— 30%**
Class attendance is mandatory and you are expected to show up on the first class. The course is heavily predicated on in-class discussion. Thus, you are expected to

attend all sessions and take an active role in class. 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, assignments.
You may miss up to two classes for any reason. A third absence will lower your course grade one full level (e.g., "A" to "B", "B+" to "C+", etc.). A fourth absence will result in being drop-failed from the course. Engaging in non-course related activities in class can result in a drop-failed from the course.

(II) **Mid-course exam — 30%**

The exam will be in class on the last class prior to Spring break, on Feb 24, 2022. 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 and you will be expected to map a question into formal analysis that you will cary out.

(III) **Capstone Project — 40%**

The group project will have you apply the tools covered in the course to real-world data. Groups will be matched with the companies providing the data early on and will start exploring the data and formulating their application early on. The evaluation will be based on the milestone presentations, the questions asked, the level of analysis, the results obtained, and the presentations made.

Preliminary Meetings' Outline

1. **Motivation** // the increase role of data science in finance

2. **How did retail investors respond to the coronavirus crisis?**
    1. introduction to Python
    2. loading data and calculating returns — dividends, splits, adjusted closes
    3. merging data sets
    4. 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. Milestone 0: initial proposal and a desired matching rank

   Due 3/3/2022

2. Milestone 1: describing the data and application ("what is the question?")

   Due 3/22/2022

3. Milestone 2: identifying additionally required data sets and laying out the empirical approach ("how will we answer the question?")

   Due 3/31/2021

4. Milestone 3: preliminary findings ("generating a minimum viable product")

   Due: 4/7/2021

5. Milestone 4: final group presentation

   Due: 4/19/2022 & 4/21/2022

   Final presentations will be judged by managers at prominent quant funds

The final report, due 5/2/2022, should be up to 10 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

Milestone 0 explained

During the first half of the course, senior team members from each of the data partners will call in to present the company and the data that they share. These members will also be present for the final group presentations. The presentation, along with the data description, will provide you with an initial idea of the data type and nature.

Each data partner will be matched with 2-3 teams. It is best if your team works on an application that excites you. To that end, I will conduct a two-way match process between teams and data partners. Each team can submit one proposal. These proposals will be ranked by the data partners and this ranking will be the basis for the match. Groups should also submit a ranking of preferences should their top choice is not available.

Note: given the sensitive nature of proprietary data, some of the data partners will require the teams working with them to sign an NDA.

IMPORTANT: If you wish to work with a different alternative data set, not supplied by one of the data partners, you can do so by submitting your initial proposal for the data during this milestone.

Milestones 1-4 explained

During the second half of the semester, we will hold regular classes only the dates corresponding to the deadlines. On those dates, your team should prepare a short presentation (around 5 mins with 2 mins of Q&A) and share your progress with the rest of the class. This process is designed to ensure that you stay on track, provide you with feedback and ideas, and allow you to identify synergies with other teams.

These milestones are not the only opportunity to receive help and feedback. The course TAs and I are available throughout the course.

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

- **Tomorrow.io** (www.tomorrow.io): a Boston-based technology company providing an all-in-one weather intelligence platform that predicts and automates weather challenges.
- **Spire** (www.spire.com): a space-to-cloud analytics company that owns and operates the largest multi-purpose constellation of satellites with strong focus on maritime and weather data.
- **YipidData** (www.yipitdata.com): the company identifies, screen, license, clean, and analyze alternative data to help investors answer their key questions by generating firm/cluster-specific reports.
- **90 West Data** (www.90westdata.com): a data platform that sells data, reports, analysis and services based on our exclusive panel of US consumer transaction data. .

Tentative Schedule

Class	Date	Topic	Other
1	1/18	Motivation	
2	1/20	Retail investors trading - Part I	
3	1/25	Retail investors trading - Part II	
4	1/27	Retail investors trading - Part III	tomorrow.io
5	2/1	Predicting international markets - Part I	
6	2/3	Predicting international markets - Part II	Spire
7	2/8	Predicting international markets - Part III	
8	2/10	Which stocks to buy? - Part I	YipidData
9	2/15	Which stocks to buy? - Part II	
10	2/17	Which stocks to buy? - Part III	90West
11	2/22	Which stocks to buy? - Part IV	
12	2/24	Mid-course exam	
13	3/15		
14	3/17	Guest lecture	TBD
15	3/22	Milestone 1 presentations	
16	3/24		
17	3/29		
18	3/31	Milestone 2 presentations	
19	4/5		
20	4/7	Milestone 3 presentations	
21	4/12		
22	4/14		
23	4/19	Milestone 4 — capstone project presentations	Attended by outside judges
24	4/21	Milestone 4 — capstone project presentations	Attended by outside judges