FNCE 737
Data Science for Finance

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

Contact
Office 2423 SH-DH
Email skogan@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 (see list of data partners). These projects will introduce students to alternative data coming from asset management firms 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.

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.

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:

- "Data Science: Theories, Models, Algorithms, and Analytics": https://srdas.github.io/MLBook/
- "Finance with Python": http://www.janschneider.website/teaching/financewithpython.html
- "An Introduction to Statistical Learning" http://faculty.marshall.usc.edu/gareth-james/ISL/
- "Whirlwind Tour of Python" https://jakevdp.github.io/WhirlwindTourOfPython/
- "Python Data Science Handbook" https://jakevdp.github.io/PythonDataScienceHandbook/

Grades

Grades will be determined based on:

(I) **Class Participation — 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.

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 take place in class immediately after spring break. Please note the scheduling of the exam. You are responsible for ensuring that you are available and on campus to take the exam as no make-up exam will be offered. The exam will include a series of short questions focused on the applications and analysis discussed in class.

(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 questions asked, the level of analysis, the results obtained, and the presentations made. In addition, I will conduct a 360 evaluation in which group members will anonymously evaluate each others’ contribution, and incorporate this into the individual grade.

Preliminary Meetings' Outline

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

2. **How did retail investors respond to the coronavirus crisis?** (3 sessions)
    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 markets?** (3 sessions)
    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. **Which stocks to buy and which to sell?** From traditional quant factors to ML based stock selection (3 sessions)
    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. **Can we use SEC filings to predict firm risk?** (3 sessions)
    1. extracting, cleaning, and parsing sec filings
    2. bag-of-words representation
    3. measuring text similarity to predict volatility
    4. optional: applying text regression to volatility

6. Capstone project

1. Milestone 1: describing the data and application ("what is the question?")
2. Milestone 2: identifying additionally required data sets and laying out the empirical approach ("how will we answer the question?")
3. Milestone 3: preliminary findings ("generating a minimum viable product")
4. Milestone 4: final group presentation

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

- **Point72** (https://www.point72.com): a hedge fund managing $17.2B with over 1,500 employees, focusing on discretionary long/short, macro and systematic strategies.
- **ClimaCell** (www.climacell.co/): a Boston-based technology company providing an all-in-one weather intelligence platform that predicts and automates weather challenges.
- **Estimize** (www.estimize.com): Estimize is an open financial estimates platform designed to collect forward looking financial estimates from independent, buy-side, and sell-side analysts, along with those of private investors and academics.
- **Facetrom** (www.facetrom.com): a startup developed a breakthrough, cross-platform technology that, from a single facial photograph, can analyze a person's biometric facial features to build accurate and detailed profiles about them.

During the first half of the course, senior team members from each of the teams will call in to present the company and the data that they share. These members will also be present for the final group presentations.