Data Driven Methods in Finance: intro

Fall 2023: IEOR 4576

Naftali Cohen

Course Summary

The Data-Driven Methods in Finance course at Columbia University provides a hands-on, project-based learning experience focusing on quantitative finance. Drawing from multiple data sources, the course takes students through the entire data science workflow—from conceptualization to performance evaluation.

- Core topics include statistics, forecasting, machine learning, data scraping, and MLOps.
- Students will compete in a real-time financial forecasting competition as part of the course.
- Participants will make investment decisions and present their analyses using open-source financial and alternative data.
- Prerequisites for the course include a background in Python, linear algebra, statistics, and probability. Familiarity with operations research topics and optimization packages is also recommended.
- Attendance with a laptop is mandatory, as the course is computer-intensive and incorporates both lectures and lab work.

This course is ideally suited for students aiming for careers as quants or data scientists in the financial sector.

Logistics

- **Course Title**: Data-Driven Methods in Finance (IEOR 4576)
- **Department**: Industrial Engineering and Operations Research (IEOR)
- **Term**: Fall 2023 (previous offering: Spring 2023, Fall 2022)
- **Instructor**: Dr. Naftali Cohen
- **Email**: nyc2107@columbia.edu or naftali.cohen@columbia.edu
- **Level**: graduate, 3 credits
- **Time**: 7:10-9:40 pm
- **Dates**: Monday, 09/05/23-12/11/23
- **Location**: 140 URIS
- **Logistics page**: https://www.columbia.edu/~nyc2107/Teaching.html
- **Lecture notes**: https://naftalic.github.io/ddmif/
- **Slack**: click here to join
- **Courseworks**: click on me!
- **Office hours**: Tue, 2 pm, via zoom
- **Google Colab**: click on me!

Typical class structure

- **First segment**: 7:10 pm - 8:00 pm (50 minutes)
  - Lecture/ HW answers
- **First break**: 8:00 pm - 8:10 pm (10 minutes)
- **Second segment**: 8:10 pm - 9:00 pm (50 minutes)
  - Interview question/s
  - Project reports
- **Second break**: 9:00 pm - 9:10 pm (10 minutes)
- **Final segment**: 9:10 pm - 9:40 pm (30 minutes)
  - Hands-on exercise/ HW questions

More details

- **Data and Methods**: We will source data from Yahoo Finance, Factset, OpenBB, WRDS, and others. We will only discuss and use known financial strategies, textbook algorithms, open-source software, and freely available datasets during class.

- **Course Prerequisites**: This course is computer-intensive and assumes a working knowledge of Python. Students should know basic Python packages such as Numpy, Pandas, and Matplotlib. Knowledge of linear algebra, probability, and statistics is required. Familiarity with Operation Research topics, web scraping (e.g., Beautifulsoup), optimization packages (e.g., Gurobi or CVXPY), and the Google Colab environment is recommended. Financial background is optional.

- **Topics**: During class, we will focus on the main data science workflow of generating ideas, sourcing information, extracting features, combining signals, optimizing decisions, and evaluating performance. We will discuss sample statistics, forecasting, machine learning methods, data scraping, MLOps, and more.

More details

- **Class Meetings**: Attending each class and bringing a laptop to class is necessary. This course employs lectures and computer labs, and we will devote significant time to practicing the techniques presented during class.

- **Project**: Students will participate in an in-class real-time financial forecasting competition (similar to the M6 competition). They will be asked to augment open-source financial data with external open-source datasets.
  - Students will be asked to present their findings to the class by the end of the course.
  - The forecasting competition will focus on
    - the ability to estimate future returns and uncertainty,
    - the ability to combine estimates into an investment decision,
    - the importance of a consistent investment strategy,
    - the importance of alternative datasets and proper use of data, and
    - the importance of teamwork, transparency, and learning from mistakes.

- **The winning students are guaranteed an A+ and a special prize.**

Class schedule

	Date	Dues	Notes
1	Sep 11, 2023	test submission	First class
2	Sep 18, 2023	submission 1	-
3	Sep 25, 2023	submission 2	-
4	Oct 2, 2023	submission 3	-
5	Oct 9, 2023	submission 4	-
6	Oct 16, 2023	submission 5	-
7	Oct 23, 2023	submission 6	-
8	Oct 30, 2023	submission 7	-
9	Nov 6, 2023	-	Academic holiday - no class
10	Nov 13, 2023	submission 8	-
11	Nov 20, 2023	submission 9	-
12	Nov 27, 2023	submission 10	-
13	Dec 4, 2023	-	Final presentations
14	Dec 11, 2023	-	Final exam

Texts

Required Textbooks:

- Quantitative Equity Portfolio Management

Recommended Reading:

- Rational Decision-Making under Uncertainty: Observed Betting Patterns on a Biased Coin
- Seven Sins of Quantitative Investing
- Causal Factor Investing: Can Factor Investing Become Scientific?

Recommended Textbooks:

- Introduction to Mathematical Portfolio Theory

Grading

Grading:

- **HW**: 40%
- **Final exam**: 30%
- **Project presentation**: 30%

Course overview I

- Active vs. Passive: Can we outperform an index/benchmark?
- Qualitative vs. Quantitative: Stock coverage, scientific process
- Manual vs. Systematic: Rules, Algorithms, etc.
- What does it mean to outperform? Benchmark, CAPM, multifactor
- How to measure performance? Return (alpha), Risk (beta), IR (Sharpe)
- Ex-ante, Ex-post
- Market efficiency: Weak, Semi-strong, Strong
- Well-known anomalies: Neglected firm effect, January effect, Momentum, Index change effect, etc.
- Behavioural biases: Ambiguity aversion (preference to the familiar), availability bias (memory of similar recent event), confirmation bias (higher weight to confirming info)
- Fundamental vs. economic factors
- Data dimensions: cross sectional, time series, panel
- Basic model: alpha, exposure, premium, and noise

Course overview II

Pipeline components/ topics:

- Universe
- Data
- Features
- Predictiveness
- Forecasting
- Positions
- Portfolio construction
- Performance evaluation
- Rebalance

Course overview III

What are we (relatively) good at?

- Stock picking
- Correlation

What are we bad at?

- Pointwise forecasting
- Market Timing

Strengths:

- Large cross section
- Data availability
- Computing power
- (some) market inefficiencies
- Behavioural biases
- Scientific thinking

Weaknesses:

- Market efficiency
- Short history
- Non-stationarity
- Regime change
- Low signal content
- Market impact

Recommended meetings

- 2023 Future Of Finance Conference: Sep 22, 2023
- The future of forecasting & the M6 competition: Nov 6-7, 2023
- 4th ACM International Conference on AI in Finance: Nov 27-29, 2023

Disclaimer

This course is for educational purposes only and does not offer investment advice or pre-packaged trading algorithms. The views expressed herein are not representative of any affiliated organizations or agencies. The main objective is to explore the specific challenges that arise when applying Data Science and Machine Learning techniques to financial data. Such challenges include, but are not limited to, issues like short historical data, non-stationarity, regime changes, and low signal-to-noise ratios, all of which contribute to the difficulty in achieving consistently robust results. The topics covered aim to provide a framework for making more informed investment decisions through a systematic and scientifically-grounded approach.