Syllabus for Advanced Machine Learning in Finance

Department:	New York University Department of Finance and Risk Engineering
Semester:	Spring 2021
Course Number:	FRE-GY 7871
Course Title:	Advanced Machine Learning in Finance
Professor:	Prof. Tore Opsahl[1]
Teaching Assistant:	Sachin Labhishetty
Time:	Tuesdays at 6pm to 8:41pm (February 2nd to March 16th)
Location:	Online

I. Prerequisites

Machine Learning in Financial Engineering

II. Course Description

This course brings together advanced methods of machine learning (ML) and practical implementation understanding. The aim is to better prepare students to not only become experts in ML, but also experts in executing ML projects. From the prerequisite machine learning course with a method oriented approach, the focus in this course will shift to problem solving approach. Each lecture will tackle a particular financial problem faced by modelers, and showcase a ML solution to it. The solutions focus on the end-to-end process, including data handling and feature generation as well as techniques for gaining executive support.

Tore Opsahl[2] is a Director on BofA Securities' Scientific Implementation team (SI) and the senior member in the Americas. The team provides a wide range of customized solutions in the areas of trading optimization, risk monitoring and management, and systematic investing. The goal is to incorporate the latest advances in academic and practitioners' research in financial economics, machine learning, and data science into practical solutions that can help our clients in making informed decisions. Prior to joining the bank, he built a fintech company using real-time data from the DNS backbone of the Internet to predict website visitors and ultimately company quarterly earnings. He was also the Chief Scientist for DeepMile Networks, a Government contractor, and a Research Associate in the Innovation and Entrepreneurship group at Imperial College London's business school. He received his PhD in Complex Systems and Organization Theory, and published a numerous papers on methodological advances in network science.

III. Textbook and Papers

This course is built on the Machine Learning in Financial Engineering-course, which has used Bishop's text book Pattern Recognition and Machine Learning (2006) as a reference book. While this course will refer to material in this textbook, the lectures will be centered on various case studies. *These are*

[1] tore.opsahl@nyu.edu
[2] http://toreopsahl.com

required reading ahead of each lecture. The case studies will be provided to students through NYU Classes.

Additionally, Lopez de Prado's Advances in Machine Learning provides an up-to-date view of the challenges faced by practitioners.

IV. Course Requirements

There is a final exam in Week 7 of the course that accounts for 70% of the course grade. The remaining part of the course grade is made up of attendance and participation in class. Specifically, each class will start with a small quiz on the assigned case studies.

V. Lectures

Week 1: Improving forecasts: Feature and observation generation

Date: February 2, 2021, at 6pm

The first lecture will focus on level set students' knowledge of machine learning as well as general approaches to improving forecasts. This will include a review of methods and a discussion of data. The data discussion will touch both on the amount of data (including techniques for generating more data) and the importance of feature generation (see case study).

Case study: Opsahl, T., Newton, W., 2016. Credit risk and companies' inter-organizational networks: Assessing impact of suppliers and buyers on CDS spreads[3]

Week 2: Anti-money laundering: Binary Classification Algorithm

Date: February 9, 2021, at 6pm

Money is the prime reason for engaging in almost any type of criminal activity. Money-laundering is the method by which criminals disguise the illegal origins of their wealth. Financial institutions deal with people's money, and thus, rely on a reputation for probity and integrity. It is key to detect and report attempts to launder money to maintain this reputation. Machine learning can play an important role in the identification of suspicious activity. This lecture will examine the anti-money laundering (AML) pipeline. Specifically, we will use the a machine learning platform called DataRobot and a binary classification algorithm to reduce the number of suspicious accounts selected for human review, and potentially Suspicious Activity Report (SAR) fillings.

Case study: Haselkorn, D., et al., 2017. Finding a needle in a haystack[4]

Case study: Truong, A., et al., 2019. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools[5]

---

[3] https://arxiv.org/abs/1602.06585

[4] http://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2017/jul/AML%20Transaction%20Monitoring.pdf; See also: https://www.fincen.gov/what-we-do; https://philadelphiafed.org/-/media/consumer-finance-institute/payment-cards-center/publications/discussion-papers/2007/D2007FebPrepaidCardsandMoneyLaundering.pdf; http://www.unodc.org/unodc/en/money-laundering/introduction.html

[5] https://arxiv.org/abs/1908.05557

Week 3: Macro-economic variables: Data reduction algorithm

Date: February 16, 2021, at 6pm

Default of companies is one of the main sources of risk for financial institutions. Estimating the likelihood of companies defaulting is key for institutions when, for example, issuing loans or pricing a credit default swap. The underlying models often use a binary classification framework. It is vital that these models are sensitive to the macro-economic environment to avoid assuming that the economic environment is constant, and allow for various scenarios to impact the forecast. In the main loss forecasting process for banks, the Comprehensive Capital Analysis and Review (CCAR), the Federal Reserve provide three scenarios (baseline, adverse, and severely adverse) that are defined by 28 macro-economic variables. To enable loss forecasts to be sensitive to the various scenarios, the variables need to be included in the estimation process; however, blindly including all 28 variables leads to multicollinearity and unreliable estimates. This lecture will explore techniques for including highly correlated exogenous features in panel and time-series-based models.

Case study: Opsahl, T., 2017. Multicollinearity macro-economic variables: Forecasting using CCAR scenarios. Teaching note.

Week 4: Liquidity and utilization forecasting: Multinomial classification algorithm

Date: February 23, 2021, at 6pm

There are many classification problems in finance, and we will consider the use case of forecasting commercial loan utilization, and in particular, the classification problem of whether an obligor will have zero, partial, or full utilization at the end of the next time period. This lecture will start with a human-guided model building process for classifying the obligors (i.e., a nested multinomial regression model with linear terms). Subsequently, we will explore more advanced machine learning tools to understand the value of these tools in improving the forecast precision.

Case study: Opsahl, T., 2018. Forecasting loan utilization using neural networks: Quantifying the improvement of hidden layers. Teaching note.

Week 5: Equity trading: Empirical Asset Pricing via Machine Learning

Date: March 2, 2021, at 6pm

Equity trading is seeing a shift towards quantitative strategies from more traditional human guided approaches (e.g., hedge fund "gurus") as well as a shift towards lower cost strategies (e.g., index ETFs). These two aspects are now merging with new products based on trading strategies. This lecture will give an introduction to equity trading strategies and consider the case of calibrating a momentum strategy.

Case study: Asness, C.S., et al., 2015. Investing with style: The case of style investing. Journal of Investment Management, 13(1), 27-63[6]

Case study: Gu, S., et al., 2018. Empirical Asset Pricing via Machine Learning. SSRN.

[6] https://www.aqr.com/library/journal-articles/investing-with-style

Week 6: Loss forecasting for portfolios with CCPs: Complex systems

Date: March 9, 2021, at 6pm

As we move toward a society where no person or firm acts in isolation, it is vital to understand the systems in which people and firms interact. Nonetheless, most machine learning frameworks assume a closed system with known inputs and outputs. This assumption does not hold in most real-world contexts. This week's case study will revisit the problem of default prediction and loss forecasting for portfolios clearing with central counter-parties (CCPs). Additionally, we will explore other contexts where correlations of risk factors drive extreme tail losses.

Case study: Opsahl, T., 2015. Systemic risk of clearinghouses: Utilizing CCPs' own stress-tests to gauge losses. Teaching note.

Week 7: Final Examination

Date: March 16, 2021, at 6pm

Final exam (75 minutes; answer three out of five questions) followed by a discussion of roles within a financial and technology companies.

VI. Inclusion Statement

The NYU Tandon School values an inclusive and equitable environment for all our students. I hope to foster a sense of community in this class and consider it a place where individuals of all backgrounds, beliefs, ethnicities, national origins, gender identities, sexual orientations, religious and political affiliations, and abilities will be treated with respect. It is my intent that all students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. If this standard is not being upheld, please feel free to speak with me.

VII. Machine Learning Tools

Python / Anaconda

A useful programming framework with many machine learning techniques and tools is Python. Anaconda is one of the leading Python data science/machine learning platforms.[7]

R

Another useful programming framework for machine learning is R. Many top researchers develop methods and make them available as packages.[8]

TensorFlow

TensorFlow is a library for machine learning using data flow graphs from Google. It can be accessed both through Python and R.[9]

---

[7] http://www.anaconda.com
[8] http://www.r-project.org
[9] http://www.tensorflow.org

Keras

Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation.[10]

DataRobot

DataRobot offers an automated machine learning platform for data scientists of all skill levels to build and deploy accurate predictive models in a fraction of the time.[11]

[10] https://keras.io
[11] https://www.datarobot.com