Master Engineering ELITE Programme **Sabatino Costanzo & Loren Trigo**
University of Toronto

ARTIFICIAL INTELLIGENCE IN FINANCE:

FROM TREES TO DEEP LEARNING

(APS 1052)

--COURSE SYLLABUS--

COURSE DESCRIPTION:

In this course we'll give an overview of several applications of machine learning to capital market forecasting and credit modeling, beginning with regressions, "shallow" layered machine learning models (e.g. Support Vector Machines, Random Forests), and ending with "deep" layered machine learning models (e.g. Long Short Term Memory Networks). Each model is discussed in detail as to what input variables and what architecture is used (rationale), how the model's learning progress is evaluated and how machine learning scientists and capital market traders evaluate the model's final performance so that by the end of the course, the students should be able to identify the main features of a machine learning model for capital market forecasting and to evaluate if it is likely to be useful and if it is structured efficiently in terms of inputs and outputs.

COURSE CONTENT SUMMARY:

The course covers (but it is not limited to) the following subjects: Training and testing workflow: scaling, cross-validation pipelines. Gradient descent: mini-batch, stochastic. Financial metrics: profitability and risk. Financial feature engineering. Models: multivariate regression, logistic regression, support vector machines, principal component analysis, decision trees, random forests, k-means, and hierarchical clustering, Gaussian mixtures, MLPs, LSTMs, and auto-encoder neural networks. Applications: credit modeling, financial time-series forecasting, investment portfolio design, and spread trading, credit cycle regime identification. For more details, please go to the Course Layout section below.

COURSE PREREQUISITES:

In terms of prerequisites, the participant should be familiar with the foundations of statistics, the basics of logistic regressions (desirable), and basic linear algebra (desirable); however, since our course intends to be self-contained, we will provide a review of these concepts as needed. As the examples of our course come from finance, some familiarity with capital markets and the basic financial concepts is recommended. Basic knowledge of Python or some other programming language is recommended, even though the objective of the course is not to learn how to program (shallow & deep) machine learning models from scratch, but rather, to understand how they work and to learn how to adapt them to the particular needs of the user and to optimize their application to market forecasting. The mathematical foundations of the basic machine learning models (regressions, neural networks, support vector machines, trees etc.) will be discussed and followed by a panoramic view of the inputs that are most likely to provide valuable information for market forecasting. Standard benchmarking methods used in the industry will be also covered. Subsequently, a number of basic -- already programmed-- models will be discussed in detail and their performance evaluated.

LEARNING OBJECTIVES:

The goal of this course is to expose the participant through lectures, readings and hands-on homework to the following topics:

- The machine learning workflow.
- The types of panel data encountered in finance: cross-sectional and sequential (time or location indexed). Brownian processes (random walks) and mean-reverting processes.
- The main types of models: supervised vs unsupervised, linear vs non-linear, regression vs classification, cross-sectional vs sequential. Models include: multivariate regression, logistic regression, principal component analysis, support vector machines, decision trees, random forests, k-means, hierarchical clustering, Gaussian mixtures, multi-layer perceptron, recurrent neural networks, LSTMs, and auto-encoder neural networks.
- The mathematical and algorithmic structure of the models, their assumptions and their purpose, their strengths and their weaknesses.
- Applications of the models to credit modelling, time-series and financial time-series forecasting, investment portfolio design, spread trading, credit cycle regime identification.
- Financial metrics of model adequacy: profit or risk evaluation metrics associated with financial predictive models: information coefficient, Sharpe ratio, CAGR, annualized volatility, White Reality Check (a version of Superior Predictive Ability).

LEARNING OUTCOMES:

After completing this course, participants will be able to use Python and out-of-the-box statistical learning libraries (e.g. Scikit-Learn, Keras/TensorFlow) to program a basic machine learning workflow applied to panel data and involving the following steps:

- Obtain panel data from Wharton Research Data Service using web-queries or obtain free data from data providers such as Yahoo, Quandl etc. using Pandas-Datareader.
- Prepare the data into indexed dataframes using Pandas functions for date indexing, for hierarchical indexing, and for table management: pivot, join and merge.
- Engineer alpha-factors and risk-factors with specialized libraries including Ta-lib, FINTA, the Fama-Macbeth linear factor model, Panda's date processing functions.
- Engineer time-series decomposition features such as trend, seasonality, lookback window etc. using Pmdarima, HoltWinter. ExponentialSmoothing, simple and partial autocorrelation functions.
- Engineer categorical features with one-hot-encoding.
- Extract features using principal component analysis and autoencoders.
- Select the best features using Scikit-learn feature selection functions and Quantopian's Alphalens module.

- Scale the features using Scikit-learn functions
- Construct machine learning workflows using Scikitlearn pipelines and Keras out-of-the-box functions including: splitting of data with train_test_split, model evaluation with cross-validation and model parameter tuning with grid-or-randomized-search-cross-validation.
- Apply these workflows to cross-sectional and time-series panel data using various types of models.
- Evaluate a model using simple statistical criteria (e.g. mean squared error, precision-recall), more sophisticated statistical criteria (e.g. bootstrap based), and financial criteria (Information coefficient, Sharpe ratio, CAGR, annualized volatility etc.)
- Display a model's feature importance and predictive adequacy using Scikit-Learn's and Keras out-of-the-box functions and matplotlib.

--COURSE LAYOUT—

CLASS 1:

- Why machine learning and deep learning are relevant to finance
- Modelling for inference and modelling for prediction
- Traditional statistical modelling vs machine and deep learning modelling
- Types of models (by data): cross-sectional vs time-series
- Types of machine learning models: supervised, unsupervised, semi-supervised, reinforcement
- Random walks vs stationary processes
- Classic asset price model

Homework reading review of:

- Linear regression and mean squared error optimization
- Logistic regression and maximum likelihood optimization

CLASS 2:

Basic machine learning workflow part 1

Using cross-sectional linear regression and logistic regression to illustrate:

- Train-Test workflow
- Train-Cross-Validate-Test workflow
- Scaling, cross-validation and pipelines
- Validation and test metrics
- Parameter optimization: gridsearch
- Metrics: MAE, MedAE, MSE, MSLE, MAD/MEAN-Ratio, MAPE, RSquared

Time series and feature engineering:

- TimeSeriesSplit utility
- Autocorrelation features: ACF and PACF
- Lags, date and time features
- Visualizing feature importance

Regularization of regression:

- Underfitting, overfitting, good fit, unknown fit
- Use of regularization to correct overfitting: L1 and L2 regularization
- L2 regularization to correct "jumping coefficients" caused by multicollinearity
- Optimizing regularization parameters: scaling, pipelines and gridsearch
- Scaling: Standardization, Min-Max, Mean normalization, Unit length scaling
- Scaling in finance: linear detrending (deterministic trends), differencing (stochastic trends)

Homework:

- Pandas exercise: calculating financial metrics of a trading system
- Financial metrics: Annual return, CAGR, Sharpe Ratio, Maximum Draw Down, Calmar Ratio, White's Reality Check
- Percent returns vs log returns and associated formulas

CLASS 3:

Basic machine learning workflow part 2:

- Feature reduction by selection or extraction: RFE, PCA, LDA
- Use of pipelines for feature selection or extraction

More feature engineering for time series modelling:

- Simple autoregressive models: AR(1), MA(1), ARMA(1,1)
- Time-series decomposition example using pmdarima, lags and differencing

Homework:

- Coca cola: time-series regression model
- Reading review of: F-statistic, correlation, multicollinearity, statistical models for time-series

CLASS 4:

- White's Reality Check
- Time series split utility
- Principal component analysis in depth

Examples: PCA and the yield curve, factor modelling of stock portfolios (regression, with and without PCA)

Homework:

- WRDS data wrangling exercise: downloading portfolio data, unstacking data, calculating portfolio metrics
- Regression mechanics exercise
- Readings on: PCA, cross validation types, time series cross validation
- Readings: On Factors. Fama-French risk factors, Custom MultipleTimeSeriesCV, Information coefficient, Evaluating Alpha Factors With Alpha-lens.

CLASS 5:

- Support vector machines
- Using features or kernels to model non-linearity with support vector machines
- Support vector machines parameter optimization, with scaling and pipelines
- Support vector machine and feature engineering
- Example: Support vector classifier for prediction of price movement

Homework:

- Support vector regression model of credit default swaps
- Regression mechanics exercise
- Reading on metrics for binary and multi-class classification, dealing with unbalanced data.

CLASS 6:

Trees:

- Relation between trees and binned features
- Trees and extrapolation
- Gini score and entropy score
- Tree regularization
- Tree visualization and feature importance
- 2 ways of trading a tree: extreme leaf trading and whole leaf trading
- Tree parameter optimization
- Example: factor modelling of stock portfolio (tree regressor)

Homework:

- Reading on: why entropy works to measure information complexity.
- Support vector classifier that predicts asset price return directionality: evaluation metrics in depth.

CLASS 7:

Tree Ensembles:

- Ensembles in general, the law of large numbers and the binomial distribution
- Bagging and random forests
- Gradient boosting
- Modelling correlated multiple outputs with trees or daisy chaining
- The beta of a stock, length of beta lookback window, idiosyncratic volatility
- Empirical asset pricing via machine learning: best predictors and models
- Example: factor modelling of stock portfolio (random forest regressor, with PCA), Piotroski factor model (random forest classifier)

Homework:

- Questions on trees
- Personal project (40% of grade): Absorption ratio

CLASS 8:

- Popular frameworks
- The unreasonable effectiveness of data
- Keras building blocks:
- Sequential model
- Layers
- Activation functions: sigmoid, (Leaky) ReLU, hyperbolic tangent,
- Dense layer in detail
- Dropout layer in detail
- Loss functions: MSE, MAE, MAP, binary, categorical cross-entropy, custom
- Optimization terms: pass, batch, iteration, epoch
- Optimizers: Adam, SGD, others
- Types of gradient descent: stochastic, mini-batch, batch
- Batch size
- Metrics: accuracy, etc. custom
- Compile, fit, evaluate functions, validation parameters
- Predict function
- Visualizing the training and validation errors via history object
- Callbacks
- Regularization
- Use of KerasClassifier and KerasRegressor wrappers for cross-validation and parameter optimization
- Neural networks and scaling issues
- Rules of thumb re. neural net architecture
- Multivariate processing for time-series
- Walk-forward validation with gridsearch
- Benchmarking a gridserach

Example: MLP classifier for price prediction with class weights, callback

Example: MLP regressor and classifier ensembles to predict bitcoin price

Homework:

- Optional exercise using a Python technical indicator library called Finta
- 9ETF random forest model of price prediction.
- Reading: Gradient Descent and Back Propagation

CLASS 9:

- Outlier identification
- Autoencoders
- Autoencoder and PCA comparison

Example: Credit card fraud identification

- Scaling, oversampling
- Supervised: Logistic Regression, Random Forest, SGBoost, Keras MLP
- Unsupervised: PCA, autoencoder

Homework:

- Exercise on bank failure (trees, tree-ensembles, PCA and autoencoders)

CLASS 10:

- Recurrent neural networks and LSTMS
- Relation to ARMA models
- Recurrent neural network logic: looped and unrolled
- Recurrent neural network layer in Keras
- The exploding/vanishing gradient problem
- LSTM logic
- LSTM layer in Keras
- LSTM 3D inputs
- Data generators

Example: LSTM applied to stock price prediction (with and without window normalization)
Example: RNN, LSTM and ARIMA applied to massive data (web page views)

Homework:

- Reading: VanishingGradientProblemHowToFixIt.docx, UNDERSTANDING LSTMS INTRODUCTION W ANIMATIONS & CONCLUSION

CLASS 11:

- Clustering
- Gaussian Mixtures
- The credit cycle
- Hierarchical-Risk-Parity

Example: Gaussian mixtures for price regime identification, credit cycle phase identification
Example: PCA and clustering for co-integrated pairs identification, PCA for eigen-portfolios
Example: Hierarchical clustering for portfolio construction

Homework:

- Exercise on gradient descent
- Readings: ProbabilityDensityEstimation.pdf, ExpectationMaximizationAlgorithm.pdf, MaximumLikelihoodEstimation.pdf,
- KMeansAndGaussianMixturesClustering.pdf

CLASS 12:

- An in-depth review of Technical and Fundamental Financial Indicators

Homework:

- Final Team Project

SELECT REFERENCES

- Aronson(2007) Evidence based Technical Analysis
- Chollet (2018) Deep Learning with Python
- Cleff (2019) Applied Statistics and Multivariate Data Analysis for Business and Economics
- Dixon (2020) Machine Learning in Finance
- Gueron (2019) Hands-on Learning with Scikit-Learn and TensorFlow
- Grus (2019) Data Science from Scratch
- Klaas (2019) Machine Learning for Finance
- Muller (2017) Introduction to Machine Learning with Python
- Nielsen (2020) Practical Time Series Analysis
- Patel (2019) Hands-on Unsupervised Learning Using Python
- VanderPlas (2019) Python Data Science Handbook
- Wonnacott (1977) Introductory Statistics for Business and Economics

--COURSE INSTRUCTORS—

Sabatino Costanzo-Alvarez:

He holds a Masters in Economics and Finance from Brandeis University as well as a Magister Scientiarum, a Magister Philosopharum and a Ph.D. in Mathematics from Yale University, where in 1990 achieved a significant breakthrough by solving a mathematical conjecture which had remained unsolved for more than 3 decades. Taught Mathematics of Finance at Boston University as an Associated Professor for 5 years and later co-founded the Boston Trading Group LLC, designed the trading systems used in the firm's daily Futures Trading Operations and acted as head trader of the team. Holds the licenses "Registered Representative NYSE/NASDAQ" (Series 7), "Registered Financial Advisor", "Registered Uniform State Law Securities Agent", "Registered Managed Futures Fund Representative" in the U.S. and "Canadian Securities Course" & "Conduct and Practices" in Canada, as well as products training at Morgan Stanley in Boston, and later at Merrill Lynch in New York. Chaired the Advanced Management Program for Senior Executives (PAG), an Executive MBA at the IESA Institute, where he taught Financial Engineering and Investment Management as an Associate Professor, and tutored over 70 MBA dissertations. Acted as Head of Research at Econo Invest C.A., one of the largest Investment Firms in Latin America, leading the Investment Strategy Team in charge of generating and executing the U.S. & E.U. investment strategies for Commodities, Fixed Income Instruments and Equities for the firm (published weekly in Bloomberg), as well as generating and maintaining the Sovereign Fixed Income Indexes of Brazil, Colombia, Mexico, Peru, Chile, Uruguay and Venezuela to be used in the design of international financial products. Acted as an Investment Advisor for the International Wealth Management Groups at Morgan Stanley (Boston), Merrill Lynch (NY) and the Royal Bank of Canada(Toronto), and is now a Senior Partner at the Toronto boutique Investment Firm Inter Alea, where he provides state-of-the-art mathematical modeling solutions to portfolio and risk management problems for a select group of corporate and high net worth private clients, designing and managing their investment portfolios based on their specific risk & return requirements. He teaches Portfolio Management, Statistics & Mathematical Modelling and Business Mathematics Courses at the Pilon School of Business, and is the founder and advisor of the Sheridan Students Trading and Investment Association. He is a Lecturer at the U of T Graduate School, where he is teaching Portfolio Management, Blockchain Technology, Cryptocurrencies and Artificial Intelligence applied to Finance.

Rosario Lorenza Trigo-Ferre:

Holder of a B. A. in Philosophy (Magna Cum Laude) from Yale University -where she also received training in Math & Physics-, a Ph.D. in Generative Linguistics from Massachusetts Institute of Technology (MIT) and a M. Sc. in Management of Information Systems from Boston University ("Beta Gamma Sigma Honors" award), she was a Professor at Boston University for 8 years. While a Programmer Analyst at Boston University, she designed and developed an application for the management of accounts trading stock and currency futures and co-designed financial applications under the direction of Professor Zvie Bodie at B.U. Co-founder and Trader at the Boston Trading Group and Certified Programmer Analyst in e-commerce by the University Computer Careers Program, she generated the trading signals for currencies and metals futures used in the BTG's market operations; developed an application maximizing the efficiency of trading system for currency and metal futures, and designed a client-server application for the management and operation of trading accounts. Has designed and developed many multi- tiered e-commerce applications dynamically generated from databases. Project

leader and senior programmer analyst at IngeDigit, designed and developed internet applications for banking accounts management & operation, and for international transactions between banking accounts and credit cards. She was a Professor at the Department of Production and Technical Innovation of the IESA Institute, the top -only US accredited- Venezuelan Business School, where taught courses in Information Systems, Simulation in Finance, Operations and Database Marketing. She is the author of many scientific papers in refereed journals and a Permanent Consultant for an international development bank (C.A.F, The Andean Region Development Bank), where she has designed the financial models used to evaluate the profitability, coverage and socio-economic impact of projects like the inclusion of fiber-optic cable in highways in Colombia and Peru. These models led to the enactment of new laws making such inclusion mandatory in the Andean region. Also designed the financial models used to evaluate the profitability of projects in satellite technology in Argentina (specifically the ARSAT program) by estimating the future regional demand for transponders and the impact of the project in the input-output matrix of the country, and is now a Partner at the boutique Investment Firm InterAlea, where she designs, develops, tests and implements trading and risk management strategies based on the entropy analysis of price signals, executed on stock quote-data processed through SQL-Server. She is a Lecturer at the U of T Graduate School, where she is teaching Portfolio Management, Blockchain Technology, Cryptocurrencies and Artificial Intelligence applied to Finance.