Machine Learning in Asset Pricing
Module 1, 2025

Course Information

Instructor: Lingxiao Zhao, PhD, Assistant Professor of Finance,

Peking University HSBC Business School

Office: PHBS Building, Room 613

Phone: 0755-2603-8442

Email: lingxiao@phbs.pku.edu.cn

Office Hour: Tuesday & Friday 8:30-10:00am;
Wednesday 9:30-10:30am

Teaching Assistant: Yutao Yuan and Xufei Lu

Classes:
Lectures: Tuesday & Friday 1:30-3:20pm
Venue: PHBS Building, Room

Course Website:

1. Course Description

1.1 Context

This course introduces recent advancements in machine learning (ML) as applied to financial markets, with a focus on asset pricing. Key topics include foundational finance concepts such as arbitrage pricing theory, asset pricing factor models, portfolio analysis, and return prediction, and investor learning. Building on this foundation, students will explore the expanding range of ML techniques used in financial research and quantitative investment.

A preliminary understanding of modern portfolio theory and econometrics is recommended. By the end of the course, students will understand how ML methods can solve empirical problems in finance and will be able to use R or Python packages to address these questions.

1.2 Textbooks and Reading Materials

1) Stefan Nagel; Princeton Lectures in Finance; 2021
Published by: Princeton University Press
Machine Learning in Asset Pricing

2) Bryan Kelly and Dacheng Xiu; Copyright © 2024 now publishers inc.; 2024
Financial Machine Learning

3) Tidy Finance
https://www.tidy-finance.org/r/
https://www.tidy-finance.org/python/

4) Related research papers

1.3 Data Sources (an account with WRDS is required)

- **French data library:** https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
- **Goyal–Welch dataset:** https://sites.google.com/view/agoyal145/?redirpath=/
- **CRSP**: stock returns, market equity, delisting returns.
- **Compustat**: firm fundamentals (book value, earnings, etc.).
- **Fama-French & Momentum Factors**: via WRDS or `frenchdata` R package.
- **Macro predictors**: FRED (interest rates, inflation, term spread, dividend yield etc)

2. Learning Outcomes

2.1 Intended Learning Outcomes

2.2 Course specific objectives

2.3 Assessment/Grading Details

- Quizzes (20%): Random assignments in class
- Assignments (10%): Hands-on coding tasks
- Midterm Exam (35%): Conceptual and analytical questions
- Final Project (35%): Empirical analysis tasks

2.4 Academic Honesty and Plagiarism

It is important for a student's effort and credit to be recognized through class assessment. Credits earned for a student work due to efforts done by others are clearly unfair. Deliberate dishonesty is considered academic misconducts, which include plagiarism; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic achievement; acting alone or in cooperation with another to falsify records or to obtain dishonestly grades, honors, awards, or professional endorsement; or altering, forging, or misusing a University academic record; or fabricating or falsifying of data, research procedures, or data analysis.

All assessments are subject to academic misconduct check. Misconduct check may include reproducing the assessment, providing a copy to another member of faculty, and/or communicate a copy of this assignment to the PHBS Discipline Committee. A suspected plagiarized document/assignment submitted to a plagiarism checking service may be kept in its database for future reference purpose.

Where violation is suspected, penalties will be implemented. The penalties for academic misconduct may include: deduction of honour points, a mark of zero on the assessment, a fail grade for the whole course, and reference of the matter to the Peking University Registrar.

AI tools requirements：
Using AI tools to complete assignments or assessments without the approval of the course instructor will be regarded as an act of academic dishonesty. Depending on the severity of the situation, penalties will be implemented in accordance with the provisions of the Peking University Graduate Student Handbook.

For more information of plagiarism, please refer to *PHBS Student Handbook*.

3. Topics, Teaching and Assessment Schedule (Tentative)

Week 1-4: Empirical Asset Pricing and Introduction to Machine Learning

Session 1: Introduction & Motivation

Theory: What is ML in finance? Expected returns, factor models, anomalies, factor zoo.

Empirical: Set up R/Python and WRDS. Simple demo: stock returns data handling.

Session 2: Financial Data & Tools

Theory: WRDS databases (CRSP, Compustat, Fama-French), CCM linking, tidy data principles.

Empirical: Build SQLite database; query monthly returns & characteristics with RSQLite/dbplyr.

Session 3: Linear Models & CAPM

Theory: OLS refresher, CAPM equation, beta vs. alpha, bias–variance tradeoff.

Empirical: Estimate CAPM betas for stocks; plot Security Market Line.

Session 4: Portfolio Sorts I (Univariate)

Theory: Univariate sorts to test return predictors; anomalies (size, value, momentum).

Empirical: Sort stocks by size into portfolios; compute average returns & spreads (Small–Big).

Session 5: Portfolio Sorts II (Bivariate & Factor Construction)

Theory: Bivariate sorts (size–value, size–momentum); factor construction (SMB, HML).

Empirical: Build 2×3 size–value portfolios; compute SMB & HML factors; compare to Fama-French.

Session 6: Evaluating Factor Performance

Theory: Factor metrics (mean return, t-stat, Sharpe, alpha). Anomalies & robustness issues.

Empirical: Test SMB & HML premiums; regress factors on market; plot cumulative returns.

Session 7: Asset Pricing Tests

Theory: Time-series (BJS) vs cross-sectional (Fama–MacBeth) regressions.

Empirical: Run Fama–MacBeth regressions on size & value; estimate factor risk premia.

Session 8: Factor Zoo & Early ML Motivation

Theory: Proliferation of anomalies, multiple testing bias. Why ML (regularization, prediction focus).

Empirical: Demo of false positives with random predictors; teaser of lasso selection.

Week 5 Midterm Week

Weeks 6–9: Other ML Algorithms & Finance Applications

Supervised Learning Algorithms

- **Linear Regression** – Predicts a continuous target by modeling a linear relationship between features and outcomes.
- **Logistic Regression** – For binary classification; uses a sigmoid function to estimate probabilities.
- **Decision Trees** – Splits data using feature-based rules; handles classification & regression.
- **Support Vector Machines (SVM)** – Finds the best hyperplane separating classes in high dimensions.
- **Naive Bayes** – Probabilistic classifier assuming independence; effective in text/spam filtering.
- **K-Nearest Neighbors (KNN)** – Classifies (or regresses) based on the labels/values of nearby points.
- **Decision Tree and Random Forests**
- **Neural Networks** – Multi-layer models capturing complex non-linear relationships.

Unsupervised Learning Algorithms

- **K-Means Clustering** – Groups data into *K* clusters based on distance to centroids.
- **Principal Component Analysis (PCA)** – Dimensionality reduction by projecting data onto principal components, retaining most variance.

Asset Pricing Applications

- **Shrinking the Cross-Section**
- **Predicting the market premium**