Discipline:

Methods & Finance

1. Language

English

2. Title

Machine Learning in Finance

3. Lecturer

Prof. Dr. Martin Hibbeln, Mercator School of Management, Universität Duisburg-Essen
(martin.hibbeln@uni-due.de)

4. Date and Location

13.04.2027 – 25.05.2027

The course will be offered online via Zoom. After the kickoff session in week 1, weeks 2 – 5 each include a 4-hour lecture and a 4-hour computer lab session. Afterwards, there will be a small Kaggle coding competition with group presentations of the results.

5. Course Description

5.1 Abstract and Learning Objectives

Recent developments in analytical techniques, massive increases in computing power, and growing amounts of data have led to an unprecedented rise in machine learning applications across many areas of finance, such as asset pricing (Gu et al. 2020; Bianchi et al. 2021), earnings forecasts (van Binsbergen et al. 2022), and credit scoring (Fuster et al. 2021). Clearly, machine learning has become an important skill for any aspiring financial researcher.

The goal of this course is to provide participants with a solid methodological foundation in machine learning models, demonstrate how these methods can be applied to address financial research questions, and enable the participants to analyze both structured and unstructured data using machine learning in R.

At the end of the course, participants will apply their knowledge in a Kaggle coding competition, where they team up with fellow participants to develop the most effective prediction model for a prediction task in a financial context.

5.2 Content

Topic 1 – ML Fundamentals & Model Evaluation: Empirical asset pricing and credit risk

Participants will gain an understanding of the fundamental elements of machine learning, including simple linear models for regression and classification. They will explore how prediction tasks differ from statistical inference. A major focus will be on model evaluation, which will be explored from several angles, including resampling techniques, OOS vs. OOT distinctions, and evaluation metrics.

In the finance context, these tools will be applied to empirical asset pricing, where machine learning is used to predict stock returns (Gu et al. 2020), and to credit risk, where Exposure at Default is modeled for banking applications (Gürtler et al. 2018).

Topic 2 – Subset Selection, Shrinkage & Dimension Reduction: Fintech loan default and return forecasting

Participants will be introduced to various types of linear models designed to mitigate overfitting in training data. While subset selection methods such as Forward, Backward, and Stepwise Selection will be briefly covered, the main focus will be on state-of-the-art regularization techniques such as LASSO and Ridge, and dimension reduction methods such as PCA and PLS.

These techniques are motivated by finance problems with many potential predictors. Covered applications include identifying key determinants of fintech loan defaults (Croux et al. 2020) and predicting stock returns (Gu et al. 2020).

Topic 3 – Moving Beyond Linearity: Nonlinear asset pricing

This session extends beyond strictly linear relationships to nonlinear methods that define the current state of the art. The course will cover tree-based approaches, including Decision Trees, Bagging, Boosting, and Random Forests, as well as Neural Networks. In addition, participants will be introduced to feature importance techniques, which have gained prominence in addressing the black-box problem of ML models.

We discuss finance applications in empirical asset pricing, where nonlinear models capture complex interactions among predictors to improve return forecasts (Gu et al. 2020).

Topic 4 – Unstructured Data: Financial disclosures, manager & investor sentiment

The final session turns to unstructured data, which accounts for more than 80% of the data available today. Participants will learn methods for analyzing both text and image data. For text analysis, topics will include sentiment analysis and readability assessment. For image analysis, participants will build on their prior knowledge of neural networks by exploring convolutional neural networks, the state-of-the-art approach to image processing and analysis.

In a finance context, we will discuss recent research that applies these methods to the textual analysis of financial disclosures (Lang et al. 2015), to measuring manager sentiment for stock return prediction (Jiang et al. 2019), and to gauging investor sentiment from news photos (Obaid et al. 2022).

5.3 Schedule (including start and end time)

Before each session, participants are required to watch the accompanying videos from James et al. (2023), which we will provide. These videos give participants a methodological understanding of how the machine learning models work. This allows us to focus on their application in a financial context during the lecture and their implementation in R in the computer lab.

Date	Time	Session
13.04.2027	14:00 – 15:30	Kickoff Session
20.04.2027	14:00 – 17:00	Topic 1: Lecture
21.04.2027	14:00 – 17:00	Topic 1: Lab Session
27.04.2027	13:00 – 16:00	Topic 2: Lecture
28.04.2027	14:00 – 17:00	Topic 2: Lab Session
04.05.2027	14:00 – 17:00	Topic 3: Lecture
05.05.2027	14:00 – 17:00	Topic 3: Lab Session
11.05.2027	13:00 – 16:00	Topic 4: Lecture
12.05.2027	14:00 – 17:00	Topic 4: Lab Session + Introduction to Kaggle Competition
25.05.2027	10:00 – 16:00	Presentations Coding Challenge

5.4 Course format

Lectures and computer lab sessions

6. Preparation and Literature

6.1 Prerequisites

- Foundations in statistics (descriptive statistics, linear regression)
- Basic knowledge of programming in R

6.2 Essential Reading Material

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning. 2nd Edition. New York: Springer.

6.3 Additional Reading Material

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. 2nd Edition. New York: Springer.

Topic 1 – ML Fundamentals & Model Evaluation: empirical asset pricing and credit risk

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. Review of Financial Studies, 33, 2223–2273.

Gürtler, M., Hibbeln, M. T., & Usselmann, P. (2018). Exposure at Default Modeling–A Theoretical and Empirical Assessment of Estimation Approaches and Parameter Choice. Journal of Banking & Finance, 91, 176–188.

Topic 2 – Subset Selection, Shrinkage & Dimension Reduction: fintech loan default and return forecasting

Croux, C., Jagtiani, J., Korivi, T., & Vulanovic, M. (2020). Important Factors Determining Fintech Loan Default: Evidence from a Lendingclub Consumer Platform. Journal of Economic Behavior & Organization, 173, 270–296.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. Review of Financial Studies, 33, 2223–2273.

Topic 3 – Moving Beyond Linearity: nonlinear asset pricing

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. Review of Financial Studies, 33, 2223–2273.

Topic 4 – Unstructured Data: financial disclosures, manager & investor sentiment

Lang, M., & Stice-Lawrence, L. (2015). Textual Analysis and International Financial Reporting: Large Sample Evidence. Journal of Accounting and Economics, 60, 110–135.

Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager Sentiment and Stock Returns. Journal of Financial Economics, 132, 126–149.

Obaid, K., & Pukthuanthong, K. (2022). A Picture Is Worth a Thousand Words: Measuring Investor Sentiment by Combining Machine Learning and Photos From News. Journal of Financial Economics, 144, 273–297.

6.4 To prepare

Participants are expected to watch selected videos accompanying James et al. (2023), which will be provided by us, before each session. Moreover, we expect the participants to have knowledge of basic statistical concepts and have some experience with the programming language R.

7. Administration

7.1 Max. number of participants

20 participants

7.2 Assignments

None

7.3 Exam

After the course, we will host a coding competition via Kaggle where participants will team up in groups and develop predictive models for a prediction task in a financial context. The participants submit and present their results via Zoom around 2 weeks after the end of the course period.

7.4 Credits

The course corresponds to a scope of 6 ECTS.

8. Working Hours

	Hours
Preparation of programming in R	30h
Course preparation / reading material	20h
Watching videos in preparation for each session	10h
Active participation in class during lectures and lab sessions	40h
Follow-up work (read up on topics of the day to deepen understanding)	40h
Kaggle coding competition and presentation	40h
Sum	**180 h**