VHB ProDok – Machine Learning

Welcome & Introduction

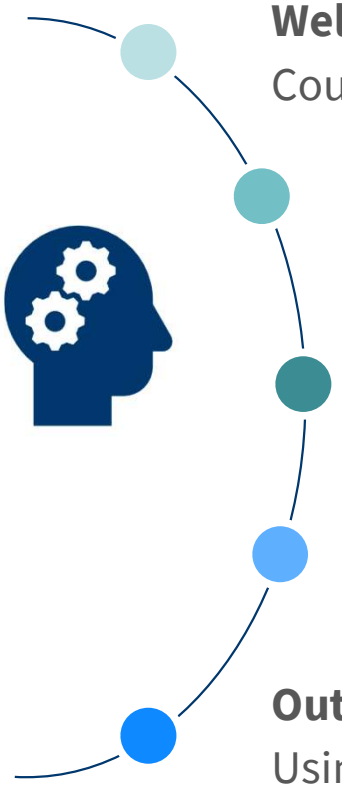
Stefan Lessmann

Lecturer

Stefan Lessmann

- **Studied at University of Hamburg (Germany)**
  - Diploma in Business Administration, 2002
  - PhD. / Habilitation in MIS, 2007 / 2013
- **PhD:** Data Mining Using Support Vector Machines
- **Habil:** Ensemble learning for decision support and analysis
- **Professor of Information Systems, Humboldt-University of Berlin, Germany**
- **Research interests within applied machine learning**
  - Marketing & risk analytics
  - Natural language processing (NLP)
  - Causal ML and policy learning
- **Contact: stefan.lessmann@hu-berlin.de**

Agenda



Welcome & Introduction
Course format, schedule, tooling and data, grading and evaluation

Block I: Fundamentals of Machine Learning
Types of ML, use cases, algorithms and practices

Block II: Artificial Neural Networks for Deep Learning and Text Analytics
Foundations of neural networks, text and sequential data, NLP transfer learning

Block III: Selected Topics in Machine Learning Research
Interpretable machine learning, potential outcomes and treatment effects

Outlook and Next Steps
Using ML in your research, course assignment

Course Format and Objectives

Ready participants for using machine learning (ML) in their research

- **Learning goals**
  - Familiarity with selected ML concepts, theories and algorithms
  - Understanding of ML uses cases in business and economics research & practice
  - Ability to apply ML algorithms to data and interpret their results
- **Blend of teaching approaches (presentation, discussion, programming)**
- **Target audience**
  - PhD student in business or economics with working knowledge of statistics / econometrics
  - No prior knowledge of machine learning needed or assumed
- **Onside format: three full days + one half day**

Course Outline

Day 1			
		Arrival of participants	
09:00	10:30	Welcome and introduction	
10:30	11:00	Coffee break	
11:00	12:30	L.I.1	Introduction to machine learning
12:30	13:30	Lunch break	
13:30	15:30	P.I.1	Data exploration & preparation using Python
15:30	16:00	Coffee break	
16:00	17:30	L.I.2	Supervised learning principles

Day 2			
09:00	09:15	Welcome and recap of day I	
09:15	10:30	L.I.3	Machine learning model validation
10:30	11:00	Coffee break	
11:00	12:30	L.I.4	Advanced supervised learning algorithms
12:30	13:30	Lunch break	
13:30	15:30	P.I.2	Supervised ML in action: Credit risk analytics case study
15:30	16:00	Coffee break	
16:00	17:30	L.II.1	Introduction to neural networks

Course Outline

Day 3			
09:00	09:15	Welcome and recap of day II	
09:15	10:30	L.II.2	NLP foundations and Word-to-Vec
10:30	11:00	Coffee break	
11:00	13:00	L.II.3	Toward modern NLP
13:00	14:00	Lunch break	
14:00	15:30	P.II.1	Text processing and sentiment analysis
15:30	16:00	Coffee break	
16:00	17:30	L.II.4	Sota models for text analytics

Day 4			
09:00	09:15	Welcome and recap of day III	
09:15	10:30	L.III.1	Interpretable machine learning
10:30	11:00	Coffee break	
11:00	12:30	L.III.2	Causal machine learning primer
12:30	13:30	Lunch break	
13:30	15:30	L.III.3	Closing session: Discussion of the course assignment & next steps

Block I: Fundamentals of Machine Learning

Learning algorithms for processing structured data

- **ML processes, algorithms, and applications in decision analysis and support**
  - L.I.1 Introduction to machine learning
  - L.I.2 Supervised learning principles
  - P.I.1 Data exploration & preparation using Python
  - L.I.3 Machine learning model validation
  - L.I.4 Advanced supervised learning algorithms
  - P.I.2: Credit risk analytics case study
- **Focus on supervised ML**
  - Prevails in business and economics research
  - High relevance for PhD candidates who use ML in their research

Churn	No. of orders	No. of returns	Days since last order	Total purchases	…
Yes	3	1	7	€150	…
Yes	1	0	13	€75	…
No	5	2	5	€33	…
No	2	0	2	€24	…
No	1	0	25	€120	…
No	3	1	17	€41	…
Yes	9	1	9	€284	…
No	2	2	14	€10	…
Yes	1	0	11	€18	…

Structured tabular data with a **dependent variable (target)** and a set of **independent variables (features)**.

Block II: Artificial Neural Networks for Deep Learning and Text Analytics

Fundamentals and approaches for natural language processing (NLP)

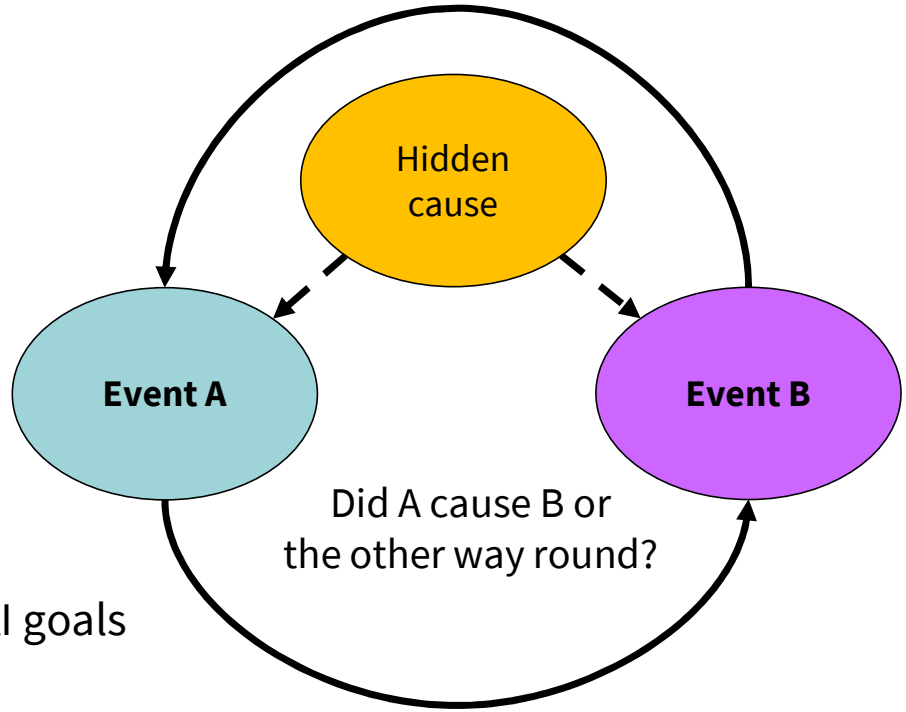
- **Neural networks, deep learning, and their applications for text data processing**
  - L.II.1 Introduction to neural networks
  - L.II.2 NLP foundations and Word2Vec
  - P.II.1 Neural networks in Python (self-study)
  - P.II.2 Processing text data in Python
  - L.II.3 State-of-the-art models for text analytics
  - P.II.3 Sentiment analysis case study
- **Focus is on deep learning and text**
  - Hot topic in academia and industry
  - Opportunity for business and econ research where NLP papers are yet scarce

This is a piece of text. There is no a priori defined structure in this text. Sentences can be long or short. The building block of text data is a word (or character). We can interpret text as a sequence of words. A learning algorithm for text data needs to understand the meaning of words. NLP is the discipline concerned with crafting such algorithms. Much corporate data is available in textual form, which makes NLP a mega-topic on managers' agenda.

Block III: Selected Topics in Machine Learning Research

Advances in the scope of causal inference and ML model interpretation

- **Selected recent developments in ML research**
  - L.III.1 Interpretable machine learning
  - L.III.2 A causal machine learning primer
  - P.III.1 Uplift modeling case study (self-study)
- **Selection motivated by**
  - Reluctance to trust opaque, black-box ML models
  - Explainability is a prerequisite for achieving higher-order AI goals (Fairness, Compliance, Robustness, Safety, etc.)
  - Paradigm shift in ML and econometrics to exploit each others' strengths
  - Strong preferences for interpretable cause-effect models in business & economics research



Tooling

- **Python 3.10 or later**
  - Probably the most popular languages for machine learning and data science
  - Open source
  - Great online resources including courses, videos, tutorials, blogs, etc.
- **Core Python Libraries**
  - Numpy, Pandas, Matplotlib, Seaborn
  - Scikit-learn for standard machine learning
  - Keras for deep learning and natural language processing (NLP)
- **Development environment**
  - Google Colab: recommended for students w/o prior experience in Python programming
  - Many other options (PyCharm, VSC, Spider, local Jupyter installation, …)

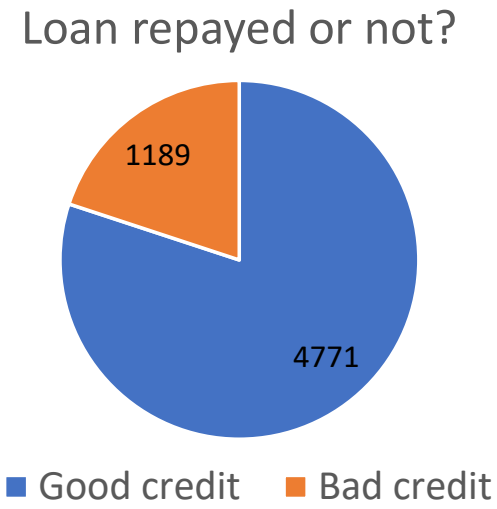
Home Equity Data Set

Coding sessions of block #1

CREDIT RISK ANALYTICS

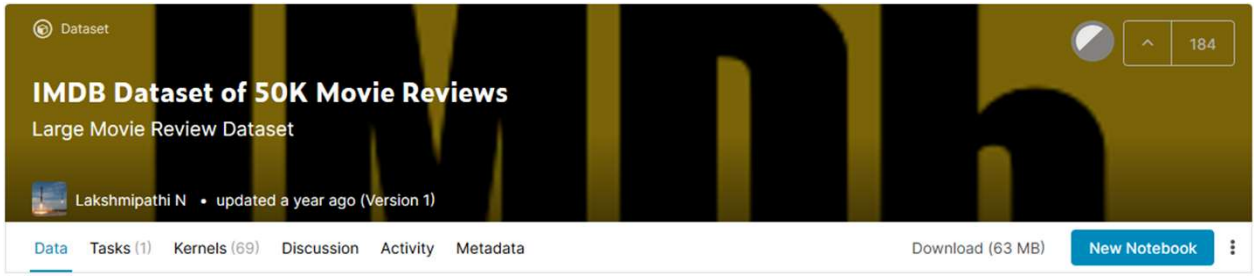
MEASUREMENT TECHNIQUES, APPLICATIONS, and EXAMPLES

- **Assessing the risk of retail credit applicants**
- **Binary classification problem: default or not?**
- **Simple, easy-to-use data set for practicing**
  - Everything available in one table
  - Basic data preparation suffices
- **Data set characteristics**
  - 5960 observations
  - 12 variables describing each loan
    - LOAN: amount of the loan request
    - MORTDUE: amount due on existing mortgage
    - VALUE: value of current property
    - REASON: DebtCon or HomeImp
    - JOB: occupational categories
    - YOJ: years at present job
    - DEROG: number of major derogatory reports
    - DELINQ: number of delinquent credit lines
    - CLAGE: age of oldest credit line in months
    - NINQ: number of recent credit inquiries
    - CLNO: number of credit lines
    - DEBTINC: debt-to-income ratio

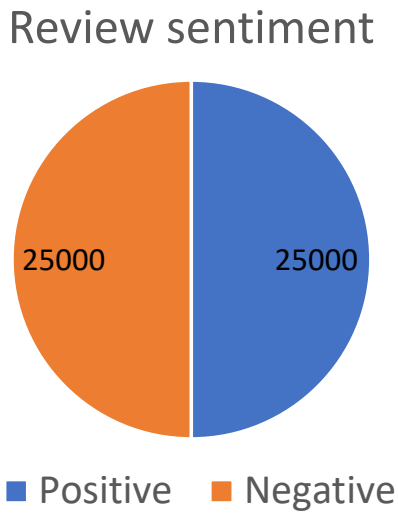


IMDB Movie Review Data Set

Coding session of block #2



- **Sentiment of movie reviews**
- **Binary classification problem**
- **Popular data set for NLP**
  - 50K reviews from IMDB
  - From very short to rather verbose
  - Two columns, review and sentiment indicator



Probably my all-time favorite movie, a story of selflessness, sacrifice and dedication to a noble cause, but it's not preachy or boring. It just never gets old, despite my having seen it some 15 or more times in the last 25 years. Paul Lukas' performance brings tears to my eyes, and Bette Davis, in one of her very few truly sympathetic roles, is a delight. The kids are, as grandma says, more like "dressed-up midgets" than children, but that only makes them more fun to watch. And the mother's slow awakening to what's happening in the world and under her own roof is believable and startling. If I had a dozen thumbs, they'd all be "up" for this movie.

Encouraged by the positive comments about this film on here I was looking forward to watching this film. Bad mistake. I've seen 950+ films and this is truly one of the worst of them - it's awful in almost every way: editing, pacing, storyline, 'acting,' soundtrack (the film's only song - a lame country tune - is played no less than four times). The film looks cheap and nasty and is boring in the extreme. Rarely have I been so happy to see the end credits of a film. The only thing that prevents me giving this a 1-score is Harvey Keitel - while this is far from his best performance he at least seems to be making a bit of an effort. One for Keitel obsessives only.

https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data

Digital Couponing Data Set

Coding session of block #3

- **Identify surfers who should get a coupon**
- **Estimate a causal *uplift* model**
  - More than a binary classification problem
  - Discounts change surfer behavior
  - Counterfactual problem
- **Data set characteristics**
  - 60K observations (sessions)
  - 26 variables describing each session
  - Two possible targets: conversion and spent
  - Treatment and control group

Private data set obtained from industry partner and used in, e.g., https://arxiv.org/abs/2003.06271

Course Logistics

- **The VHB offers two options for certifying your participation**
  - Confirmation of participation (German: “Teilnahmebestätigung”)
  - Graded Ph.D. course with 6 ECTS (German: “Leistungsschein”)
- **Second option requires completing a course assignment**
  - Computational essay in the form of a Jupyter notebook
  - Submission deadline: **30. April 2026**
- **Assignment objectives**
  - High match with your research (or research that generally interests you)
  - Hands-on practice of machine learning

Assignment Tasks

Two alternative options for deciding on an assignment task

- **Option 1 (preferred)**
  - Brief proposal of a topic that is relevant to your research and related to the course
  - Discussion and approval (during the course of by afterwards email)
  - Complexity should be comparable to the pre-defined tasks
- **Option 2:**
  - Selected set of pre-defined modeling tasks
  - Based on Kaggle or a paper that you are to '*replicate*'
- **Submission consists of an executable Jupyter notebook**
  - Let me know if you plan to use confidential data
  - We will make it possible

Evaluation and Assignment

Evaluation criteria

- **Quality of the exposition**
  - Mix of a research paper and a good ML blog
  - Proper organization, sound argumentation, language quality, etc.
  - Sensible blend of prose, notation, codes, results (this makes a good comp. essay)
- **Modeling approach**
  - Soundness of the empirical design (experimental factors, performance measures)
  - Suitability and sophistication of the employed methods
- **Empirical results**
  - Use of sensible and appealing visualizations
  - Precise interpretation and discussion of findings
- **Tutorial notebooks provide some guidance on what is expected**

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742
Fax. +49.30.2093.5741

stefan.lessmann@hu-berlin.de
http://bit.ly/hu-wi

www.hu-berlin.de

Photo: Heike Zappe