STATS 201

Machine Learning for Social Science

Session 3, 2025-26

Course meeting time: Tuesday & Thursday 8:30am – 11:00am
Course meeting location: LIB 2121
Course format: Lecture & seminar
Academic credit: 4

Instructor's information

Markus Neumann
Assistant Professor of Political Science and Computational Social Science
Email: markus.neumann@dukekunshan.edu.cn
Office: WDR 1104
Office hours: Friday 8:30am-12:30pm, or by appointment
Professional website: https://markusneumann.github.io/

I am an Assistant Professor of Political Science and Computational Social Science at DKU. My research revolves around the application of statistics and machine learning methods to social science data, especially text, images and audio. The substantive focus of my research is political advertising.

Getting in touch with me
Feel free to send me an email about any questions you may have. Make sure the subject line contains 'STATS201'. During the week, I will try to respond within 24 hours. If you don't receive a response in that time, feel free to email me again. You can also come to my office hours, or make an appointment if the office hours times don't work for you.

What is this course about?

In almost every field, there is a need to draw inferences from or make data-based decisions. This course aims to provide an introduction to machine learning that is approachable to diverse disciplines, empowering students to become proficient in the foundational concepts and tools while working with interdisciplinary, real-world data at the intersection of machine learning and social science. You will learn to:

- Structure a machine learning problem,
- Determine which algorithmic tools apply to a given problem,
- Apply those tools to diverse, interdisciplinary data,
- Evaluate the performance of your solution, and
- Interpret and communicate your results accurately.

This applied introduction to machine learning will arm you with the essential skills to conduct analyses and communicate results effectively. Additionally, the course content will always be updated to the latest academic advancements and industry practices, ensuring that you are working with cutting-edge data, algorithms, and social science issues.

What background knowledge do I need before taking this course?

Prerequisite: MATH 101 or 105 and STATS 101; or MATH 205 or 206.

No prior Python experience is assumed. Students are expected to have some experience working with data and basic statistical concepts using a programming language, most commonly R (as taught in STATS 101). Python will be introduced and used throughout the course.

What will I learn in this course?

By the end of this course, you will be able to:

- Recognize pertinent research questions in social science that can benefit from machine learning techniques and understand interdisciplinary methodologies.
- Describe the foundational principles of machine learning algorithms.
- Construct machine learning challenges based on social science topics.
- Identify appropriate algorithmic methodologies for specific problems.
- Evaluate the effectiveness of machine learning solutions.

What will I do in this course?

You will spend this course actively building, evaluating, and explaining machine learning models in the context of real social science problems. Most class sessions combine short lectures with hands-on coding, structured in-class activities, and discussion of empirical research that uses machine learning in practice. You will work directly with data, implement models from scratch or with standard libraries, assess their performance, and interpret their outputs with attention to substantive meaning and limitations.

A central component of the course is a multi-week group project in which you apply machine learning methods from the course to a social science research question of your choosing. Through weekly progress reports, GitHub-based workflows, in-class oral check-ins, peer feedback, and a final public-facing project website, you will practice framing research questions, making methodological choices, and communicating technical results clearly and responsibly. The course places particular emphasis on understanding why methods work, when they are appropriate, and how to explain them to a social science audience.

How can I prepare for the class sessions to be successful?

The point of college is to learn how to work independently—to take responsibility for your own learning rather than simply following instructions. In a technical, applied course like this one, that means managing your time carefully and developing habits that help you keep up with both conceptual material and hands-on work. To make the most of this class:

- Read actively and come to class having attempted the readings, even if parts are unclear. Focus on understanding the problem being addressed, the assumptions behind the method, and what the results do and do not show.
- Plan ahead so you have time to think about the material, not just finish tasks.
- Keep your work organized and reproducible. Maintain clean notebooks, clear variable names, and a consistent GitHub workflow so that you can return to earlier work without confusion.
- Ask questions early if something isn't clear.

What required texts, materials, and equipment will I need?

All readings will be available through DKU/Duke Libraries, uploaded to Canvas, or freely available on the web.

- There is no course textbook, but "*An Introduction to Statistical Learning with Applications in Python*" will provide useful background reading. The book is freely available on its website.
- For the course project, you will need a GitHub account.
- For day 11, you will need the free software Audacity.
- Python will be the primary programming language in this course. We will mainly be using Python through Google Colab, which you can access in your web browser.
- Please bring your laptop to class. If you don't have a laptop, you may be able to borrow one from the library.

How will my grade be determined?

Participation (15%):
Your participation grade is based on both attendance and active engagement during class. Class sessions will include lectures, discussions on assigned readings, coding exercises, and in-class activities—all offering opportunities for you to contribute. Participation grades reflect general engagement during class and are distinct from graded oral progress updates. You will however have the chance to ask thoughtful questions during other students' oral progress updates.

Course Project (85%):
You will undertake a group project (1–3 students per group) that applies machine learning methods from this course to a social science research problem. Your project should engage seriously with both the technical ML components and the substantive social science question.

The project unfolds through a sequence of staged assignments that build toward a final deliverable. From weeks 2 through 6, each group submits a weekly progress update consisting of a short written report and a corresponding GitHub commit history that documents ongoing work, such as data preparation, modeling, experiments, figures, or drafting. Each week, students will also give a brief in-class oral progress update and answer questions about their work. In week 7, groups will give a formal presentation of their project. The project concludes with the creation of a public-facing website, hosted on GitHub Pages, accompanied by a complete and well-organized GitHub repository.

The final group project factors into the course grade as following:

Week 2 progress 10% (5% report, 5% oral check (during week 3))

Week 3 progress 10% (5% report, 5% oral check (during week 4))
Week 4 progress 10% (5% report, 5% oral check (during week 5))
Week 5 progress 10% (5% report, 5% oral check (during week 6))
Week 6 progress 5% (5% report)
Presentation 20%
Final deliverable 20%

The oral checks and part of the presentation are graded individually and more strictly to ensure that everyone does their fair share. The rest are group grades.

The work you submit should be clear, professional, and well structured, with careful attention to both technical correctness and substantive interpretation. You will be evaluated on the quality of your problem formulation, the appropriateness and understanding of the machine learning methods you use, the soundness of your reasoning, and the clarity with which you communicate your results in written, visual, and oral form. Additional guidance and detailed criteria will be provided with each assignment.

Grade scale:
A+= 98% - 100%; A = 93% - 97.9%; A- = 90% - 92.9%; B+ = 87% - 89.9%; B = 83% - 86.9%; B- = 80% - 82.9%; C+ = 77% - 79.9%; C = 73% - 76.9%; C- = 70% - 72.9%; D+ = 67% - 69.9%; D = 63% - 66.9%; D- = 60% - 62.9%; F = 59.9% and below.

For final grades, .05 is rounded up. For example, a 92.94999 is an A-, a 92.95 is an A. Grades are non-negotiable and can only be changed due to an error in calculation or transcription.

Don't trust Canvas' grade calculation – if you want to know what your current grade is, calculate it yourself.

Grade definitions:

Grade	Description
A+ (Exceptional)	Reserved for *extraordinary* work—original, insightful, and executed with exceptional depth or creativity. Awarded only in rare cases when the work clearly stands out from even the best A-level submissions.
A / A- (Excellent)	Exceeds course expectations through clear, polished, and thoughtful work showing full command of the material. Demonstrates strong analytical skill and careful attention to detail.
B+ / B / B- (Good)	Represents solid, competent work that meets course expectations. Shows clear understanding and consistent effort. A "B" indicates the expected level for students who engage seriously with the material; "B+" reflects above-average insight or execution.
C+ / C / C-	Meets the basic requirements of the assignment and shows adequate

Grade	Description
(Satisfactory)	understanding, but with notable gaps in clarity, depth, or accuracy.
D (Marginal)	Below expectations. Demonstrates limited understanding or incomplete engagement with the material but meets the minimum standard for passing.
F (Failing)	Does not meet minimum course standards or requirements. Work is incomplete, incorrect, or missing.

What are the course policies?

Generative AI Guidelines and Policy

Generative AI (such as ChatGPT) is a novel technology to which higher education is learning to adapt. In this course, **the use of generative AI is explicitly allowed**, but still governed by DKU's rules:

- Use of these tools is governed by DKU's Academic Integrity Policy, and students must employ this technology consistent with expectations of the instructor, course, or assessment.
- **Fake citations** made up by AI will be treated as violations of DKU's Academic Integrity Policy.
- In any situations in which such tools are used (with or without permission) in the process of completing assignments, students are obliged to cite fully any use of generative AI tools in the formulation of their work, including by preserving a record of the use of the tool as original source material.
- Students should be encouraged to save all rough drafts and notes for papers, in case any concerns arise.

DKU also has a licensed version of ChatGPT that you can (but don't have to) use.

Late Penalties

This course will move quickly so therefore it is imperative that you do not fall behind by submitting late assignments. Assignments will be subject to a 30% late penalty for each portion of a day they are overdue past the deadline. For example, if an assignment is submitted 6 hours late, which is equivalent to 1/4th of a day, it will incur a penalty of 7.5% (0.25 * 30%). Rescheduling of assignment due dates will only be permitted for serious medical and personal matters, **and requires advance notice**. Unless stated otherwise, all assignments are due at 11:59pm China time.

Class Attendance

As outlined in the University Bulletin: "Responsibility for class attendance rests with individual students, and since regular and punctual class attendance is expected, students must accept the consequences of failure to attend. […] A student who has failed to attend a course for the equivalent of 2 or more weeks **may be assigned a grade of F** in that course." Absences will only be excused for serious medical and personal matters. If you will be absent from a class for a university-sponsored activity, please make arrangements with me – **beforehand** – regarding any work you might miss.

Discussion Guidelines:

Civility is an essential ingredient for academic discourse. All communications for this course should be conducted constructively, civilly, and respectfully. Differences in beliefs, opinions, and approaches are to be expected. Please bring any communications you believe to be in violation of this class policy to the attention of your instructor. Active interaction with peers and your instructor is essential to success in this course, paying particular attention to the following:

- Be respectful of others and their opinions, valuing diversity in backgrounds, abilities, and experiences.
- Challenging the ideas held by others is an integral aspect of critical thinking and the academic process. Please word your responses carefully and recognize that others are expected to challenge your ideas. A positive atmosphere of healthy debate is encouraged.
- Read your online discussion posts carefully before submitting them.

Academic Integrity:
As a student, you should abide by the academic honesty standard of the Duke Kunshan University. Its Community Standard states: “Duke Kunshan University is a community of individuals from diverse cultures and backgrounds. We are dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Members of this community commit to reflecting upon and upholding these principles in all academic and non-academic endeavors, and to protecting and promoting a culture of integrity and trust.” For all graded work, students should pledge that they have neither given nor received any unacknowledged aid.

Academic Policy & Procedures:
You are responsible for knowing and adhering to academic policy and procedures as published in University Bulletin and Student Handbook. Please note, an incident of behavioral infraction or academic dishonesty (cheating on a test, plagiarizing, etc.) will result in immediate action from us, in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising). Please visit the Undergraduate Studies website for additional guidance related to academic policy and procedures. Academic integrity is everyone’s responsibility.

Academic Disruptive Behavior and Community Standard:
Please avoid all forms of disruptive behavior, including but not limited to: verbal or physical threats, repeated obscenities, unreasonable interference with class discussion, making/receiving personal phone calls, text messages or pages during class, excessive tardiness, leaving and entering class frequently without notice of illness or other extenuating circumstances, and persisting in disruptive personal conversations with other class members. Please turn off phones, pagers, etc. during class unless instructed otherwise. Laptop computers may be used for class activities allowed by the instructor during synchronous sessions. If you choose not to adhere to these standards, I will take action in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising).

Academic Accommodations:
If you need to request accommodation for a disability, you need a signed accommodation plan from Campus Health Services, and you need to provide a copy of that plan to me. Visit the Office of Student Affairs website for additional information and instruction related to accommodations.

What campus resources can help me during this course?

Academic Advising and Student Support

Please consult with the instructor about appropriate course preparation and readiness strategies, as needed. Consult your academic advisors on course performance (i.e., poor grades) and academic decisions (e.g., course changes, incompletes, withdrawals) to ensure you stay on track with degree and graduation requirements. In addition to advisors, staff in the Academic Resource Center can provide recommendations on academic success strategies (e.g., tutoring, coaching, student learning preferences). All ARC services will continue to be provided online. Please visit the Office of Undergraduate Advising website for additional information related to academic advising and student support services.

Writing and Language Studio

For additional help with academic writing—and more generally with language learning—you are welcome to make an appointment with the Writing and Language Studio (WLS). To accommodate students who are learning remotely as well as those who are on campus, writing and language coaching appointments are available in person and online. You can register for an account, make an appointment, and learn more about WLS services, policies, and events on the WLS website.

IT Support

If you are experiencing technical difficulties, please contact IT:

- China-based faculty/staff/students 400-816-7100, (+86) 0512- 3665-7100
- US-based faculty/staff/students (+1) 919-660-1810
- International-based faculty/staff/students can use either telephone option (recommend using tools like Skype calling)
- Live Chat: https://oit.duke.edu/help
- Email: service-desk@dukekunshan.edu.cn

It is recommended that you familiarize yourself with DKU's VPN and proxy, since these may enable you to access resources you might be unable to reach otherwise. When using the VPN, you will usually want to select Duke VPN (portal.duke.edu), then INTL-DUKE.

What is the expected course schedule?

(1)	2026-01-06	Course Overview & Python Intro
(2)	2026-01-08	Python Intro, GitHub Intro **Reading** *Guide for Reproducible Research* (The Turing Way)
(3)	2026-01-13	Logistic Regression; Supervised vs. Unsupervised ML **Reading** *Machine Learning for Social Science: An Agnostic Approach*
(4)	2026-01-15	Metrics & Evaluation; additional models (Naive Bayes, Trees, SVM) **Reading** *An Empirical Evaluation of Explanations for State Repression*
(5)	2026-01-20	Optimization & Gradient Descent; Parameter Tuning **Reading** *The role of hyperparameters in machine learning models and how to tune them*
(6)	2026-01-22	Overfitting, Regularization, Bias–Variance Tradeoff **Reading** *To Explain or to Predict?*
(7)	2026-01-27	Text 1 (Fundamentals) **Reading** *Human Rights Texts: Converting Human Rights Primary Source Documents into Data*
(8)	2026-01-29	Text 2 (Modern Approaches) **Reading** *Word embeddings quantify 100 years of gender and ethnic stereotypes*
(9)	2026-02-03	Vision 1 (Fundamentals) **Reading** *A Framework for the Unsupervised and Semi-Supervised Analysis of Visual Frames*
(10)	2026-02-05	Vision 2 (Modern Approaches) **Reading**

		Body Language and Gender Stereotypes in Campaign Video
(11)	2026-02-10	Audio 1 (Fundamentals) **Reading** *Pitch perfect: Vocal pitch and the emotional intensity of congressional speech*
(12)	2026-02-12	Audio 2 (Modern Approaches) **Reading** *A Dynamic Model of Speech for the Social Sciences*
	2026-02-17 & 2026-02-19	Chinese New Year – No class
(13)	2026-02-24	Time Series **Reading** *ViEWS: A political violence early-warning system*
(14)	2026-02-26	Course Project Presentations