AI and Data Analytics in Finance

Elective · Undergraduate, with an online MBA block variant · Sample syllabus

Course number	TBD — 3 credit hours
Term	TBD
Meeting time	TBD
Location	TBD (computer classroom or bring-your-own-laptop)
Instructor	Yichen Wang
Office	TBD
Office hours	TBD, and by appointment
Email	TBD — replies within one business day
Course site	TBD (learning management system)

COURSE DESCRIPTION

Financial analysis increasingly begins with data that is messy, unstructured, or too large to read. This elective teaches students to work with that data end to end: assembling market and accounting panels in Python, extracting information from filings, transcripts, and news, and using large language models to classify and retrieve content at a scale no analyst could read manually.

The organizing commitment of the course is validation. Every technique students learn is measured against something: a dictionary method against hand-coded text, an LLM against a dictionary baseline, a model output against a benchmark the students built themselves. Students finish able to say not only what a model produced but how often it was wrong and in which direction. That habit, not familiarity with any particular tool, is what the course leaves behind.

PREREQUISITES

Principles of Finance or Corporate Finance, introductory statistics, and one introductory programming course in any language. Prior Python experience is not required; the first three weeks establish a common environment and the syntax the course relies on. Students without the programming prerequisite should speak with the instructor before enrolling.

SOFTWARE AND MATERIALS

- No textbook is required. Readings are journal articles, practitioner notes, and technical documentation, all posted to the course site.
- Python 3 with pandas, matplotlib, and scikit-learn, installed through a distribution the course will specify. Jupyter notebooks are the submission format for all coding work.
- Access to a large language model through an application programming interface. Course credits are provided; no student is required to purchase a subscription.

- A private on-premise environment running open-weight models, provided by the University, for work involving licensed or restricted data that may not be sent to a commercial provider. Assignments run in either environment.
- Filings and market data from public sources and databases the institution licenses. Students may use only data they are licensed to access, and licensed data stays in the on-premise environment.

LEARNING OUTCOMES

On successful completion of this course, students will be able to:

LO1. Assemble and clean financial data from market, accounting, and filings sources in Python, and document the pipeline so that a third party can reproduce the result.

LO2. Produce standard structured analyses, including returns, risk measures, and valuation inputs, from raw data rather than from prepared extracts.

LO3. Measure sentiment and topical content in financial text using dictionary methods, and state the assumptions those methods impose on the result.

LO4. Apply large language models to classification, extraction, and retrieval over a corpus of financial documents using structured, reproducible prompts.

LO5. Validate model output against a hand-coded benchmark, quantify error rates, and identify look-ahead bias, hallucination, and other sources of model risk.

LO6. Communicate an evidence-based financial recommendation to a non-technical audience, disclosing methods, prompts, and limitations.

COURSE SCHEDULE

The schedule is tentative. Sessions pair a short lecture with a hands-on lab, and the Outcomes column maps each week to the learning outcomes above.

Week	Topic	Outcomes	Lab / deliverable
1	What AI does and does not do in finance	—	*Environment setup*
2	Financial data sources and Python foundations	LO1	*Lab 1*
3	Cleaning, merging, and reproducible pipelines	LO1	*Assignment 1*
4	Returns, risk, and portfolio statistics in code	LO2	*Lab 2*
5	Accounting data and valuation inputs from filings	LO2	*Assignment 2*
6	Visualizing and communicating quantitative results	LO2, LO6	*Lab 3*
7	Unstructured financial text: filings and transcripts	LO3	*Lab 4*
8	Dictionary methods and their assumptions	LO3	*Assignment 3*
9	Benchmark study: hand-coding and agreement rates	LO3, LO5	*Benchmark study*
10	Large language models: prompting and cost	LO4	*Lab 5*
11	LLM classification and extraction over a filings corpus	LO4	*Assignment 4*
12	Retrieval over a document corpus	LO4	*Project proposal*
13	Model risk: look-ahead bias, hallucination, drift	LO5	*Lab 6*
14	Project presentations and course synthesis	LO6	*Memo, presentation*

INSTRUCTIONAL APPROACH

Each session opens with a short conceptual segment and spends the remainder on a live lab. Students write code during class, and the instructor works the same problem at the front so that errors surface and get debugged in the room rather than at midnight. Labs are graded on completion and effort; the applied assignments that follow them are graded on correctness and on the quality of the write-up.

Two design choices shape the sequence. Students build a dictionary-based measure of text sentiment before they may use a language model for the same task, so they have a baseline rather than an impression. And in Week 9 they hand-code a sample themselves and measure how often the automated method agrees. Students who have counted the disagreements treat model output differently afterward.

ASSESSMENT

Component	Weight	Outcomes assessed
Participation and lab engagement	10%	LO6
Applied assignments (four)	32%	LO1–LO4
Benchmark validation study	18%	LO3, LO5
Project proposal	5%	LO6
Final project: memo and presentation	35%	LO4–LO6
Total	**100%**	

Applied assignments (32%)

Four assignments, each building on the lab that precedes it. Submissions are Jupyter notebooks that run from top to bottom without manual intervention, accompanied by a short written interpretation. Code that produces the right number but cannot be re-run does not receive full credit.

Benchmark validation study (18%)

Students hand-code a sample of financial documents, run an automated method over the same sample, and report agreement rates, disagreement patterns, and what the disagreements have in common. The deliverable is a short report with the coding protocol attached. This is the central assignment of the course.

Final project (40%)

Working individually or in pairs, students pose a question that requires unstructured financial data, build a pipeline to answer it, validate the output, and present a recommendation. Deliverables are a proposal, a memo of at most eight pages for a non-technical reader, and a fifteen-minute presentation. The memo is assessed on the defensibility of the method and the honesty of the limitations section.

GRADING SCALE

A	93–100	**B**	83–86.9	**C**	73–76.9
A–	90–92.9	**B–**	80–82.9	**C–**	70–72.9
B+	87–89.9	**C+**	77–79.9	**F**	Below 70

Final grades are not curved. Grade appeals must be submitted in writing within one week of the grade being posted and must state the specific basis for the appeal.

COURSE POLICIES

Use of generative artificial intelligence

This course inverts the usual policy. Using AI tools is not merely permitted, it is required, because learning to use them well is the subject matter. What is graded is transparency about that use.

Every submission that involved an AI tool must include two attachments: a prompt log recording the prompts, model, version, and date; and a validation appendix stating what was checked, how, and what was found, including anything the model got wrong. A submission without both is returned ungraded.

Two limits apply. Students remain responsible for every number and claim submitted; "the model produced it" is not a defense. And where an assignment specifies a step be done by hand, such as the Week 9 hand-coding, automating it is treated as an integrity violation.

Attendance and participation

Labs are conducted in class and cannot be reproduced from notes. Regular attendance is expected, and students who anticipate missing more than two sessions should speak with the instructor early in the term.

Late work

Assignments are due at the start of class. Late submissions receive a 10 percent deduction per calendar day up to three days, after which they are not accepted. Commit work in progress rather than waiting until it is finished.

Academic integrity

Students may discuss approaches and may use published code with attribution. Submitting another student's notebook, misrepresenting results, or fabricating validation figures violates the University's academic integrity policy and will be reported.

Accessibility and accommodations

The University provides reasonable accommodations to students with documented disabilities. Register with the disability services office and provide the accommodation letter early in the term.

Communication

Email is the primary channel outside of class, with replies within one business day. Debugging questions are better brought to office hours or the course discussion board, where other students benefit from the answer.

ONLINE MBA BLOCK VARIANT

For an accelerated seven-week online block, the fourteen weeks fold in pairs, with recorded lectures and one synchronous lab per module. Assignments drop from four to two; the benchmark study and final project are retained in full. Weights become participation 10 percent, assignments 30 percent, benchmark study 20 percent, project 40 percent, with a recorded presentation.

Module	Covers weeks	Focus and deliverable
1	1–2	Orientation and Python foundations; environment setup
2	3–4	Pipelines and portfolio statistics; Assignment 1
3	5–6	Filings data, valuation inputs, and visualization; Assignment 2
4	7–8	Financial text and dictionary methods
5	9–10	Benchmark validation study; introduction to LLMs

Module	Covers weeks	Focus and deliverable
6	11–12	LLM classification, extraction, and retrieval; project proposal
7	13–14	Model risk and disclosure; final memo and recorded presentation