CHARTA LEARNING · LESSON NOTES

Machine learning and AI in investing

CFA Level I · Quantitative Methods

You will learn: Tell supervised, unsupervised and deep learning apart, explain overfitting, describe how a large language model is trained and writes, and say how managers use these tools and where they go wrong.

Running example: Examples use made-up figures, such as three models scored on training data and on fresh data, and an analyst who must cover 400 company filings.

1. Supervised and unsupervised learning

- Machine learning finds patterns without the analyst fixing the shape of the model in advance: the data helps choose the shape as well as the numbers. It suits messy relationships, but the model can chase pure noise.
- Supervised learning trains on inputs (features) labelled with the correct output. A category label gives classification; a number label gives regression. If any labels are missing, it is not a supervised problem.
- Unsupervised learning has no labels and looks for structure. Clustering sorts similar items into groups. Principal components analysis (PCA) squeezes many linked series into a few underlying factors. A person judges whether the result means anything.
- The deciding question: does the data come with the answer attached?

Inputs + labels → trained model → prediction on new cases

WHICH KIND OF LEARNING, AND WHICH TASK?

Will each bond default?	label is a category	**Supervised, classification**
Next quarter's earnings	label is a number	**Supervised, regression**
Group 500 stocks by trading	no labels	**Unsupervised, clustering**
Shrink 30 rate series to a few	no labels	**Unsupervised, PCA**
Tone of 10,000 labelled calls	labelled; deep network	**Supervised**

2. Deep learning

- Deep learning uses neural networks with many layers. The layers learn their own features (useful clues) instead of a person designing them. It suits huge amounts of messy material: text, images and audio.
- It is a type of model, not a third set-up. A deep network can be trained with labels (supervised) or without (unsupervised).

Watch out: Thinking deep learning is a third set-up alongside supervised and unsupervised.

3. Overfitting

- Overfitting is when a model fits its training data very closely, including random noise that will not repeat, but performs poorly on new data. Like a student who memorises last year's paper.
- Judge a model only on data it has never seen (out of sample, or held back data). A near perfect training score with a poor new-data score means overfitting, and more complexity is not better.

Gap = training score − new-data score

SHARE OF STOCKS CALLED CORRECTLY (A COIN TOSS SCORES 50%)

Model A (simple)	training 70%, new 68%: gap 70 − 68	**2 points**
Model B (very complex)	training 98%, new 54%: gap 98 − 54	**44 points**
Model C (middle)	training 85%, new 80%: gap 85 − 80	**5 points**
Best on new data	pick by unseen score	**Model C**

Watch out: Judging a model by how well it fits its training data.

4. Large language models

- A large language model (LLM) is a deep network trained on text. Text is chopped into tokens: words or pieces of words.
- Pretraining: it reads vast text and learns one trick, predicting the next token. The text supplies its own answers, so no human labelling is needed, and almost all its knowledge comes from this stage.
- Fine tuning makes it a usable assistant. Instruction tuning uses curated requests paired with good responses. Human preference feedback has people mark which of two answers they prefer. This changes behaviour, not most knowledge.
- It writes by looping: look at the text so far (the context), estimate a probability for each next token, pick one by a weighted random draw, append it and repeat. There is no fact checking step.

> **Text so far → probabilities → pick one → append, repeat**

TRAINING LESSONS AND ONE STEP OF THE LOOP

Sentence: Revenue rose 8 percent in the third quarter	8 tokens, first never a target	**7 lessons**
Shown: Revenue	right answer	**rose**
Shown: Revenue rose	right answer	**8**
After: the central bank raised interest	rates / rate / costs / every other token	**82% / 9% / 3% / 6%**
Total	82 + 9 + 3 + 6	**100%**

Watch out: Believing a language model checks facts. It picks likely tokens, so fluent is not verified, and the same prompt can give different wording.

5. How managers use these tools

- A language model's strength is unstructured text (ordinary sentences, not tidy rows) at scale: filings, research notes, transcripts and news. It can summarise, extract figures and themes, and score sentiment as an input to a model. It also helps draft reports and client messages and speed up coding.
- Retrieval-augmented generation (RAG) makes it open book: the system first searches a chosen set of documents, such as the firm's research library, and adds relevant passages to the prompt. Answers stay current without retraining and can cite sources. Made up answers become less likely, not impossible.
- In every use the output is a draft that a professional checks. A sentiment score is one input among many, never a conclusion.

TIME SAVED AND SENTIMENT

Read 400 filings by hand	400 × 3 hours	**1,200 hours**
Check an AI draft of each	400 × 0.5 hours	**200 hours**
Time saved	1,200 − 200	**1,000 hours**
Five headline scores	+1, +1, 0, −1, +1	**sum 2**
Average sentiment	2 ÷ 5	**0.4 (mildly positive)**

Watch out: Treating the output as a finished product.

6. Where it goes wrong, and the response

- Hallucination: fluent output that is false or unsupported, as sure-sounding as a correct one. Bias: training text reflects historical prejudice. Privacy: confidential data pasted into an external tool may leak. Explainability: deep models are hard to justify to a client or regulator. Instability: small wording changes alter the answer.
- Look-ahead bias: a backtest replays a strategy on past data. If the model was trained on text written after the test period, it has in effect seen the answers and results look too good. Use only what was available on each test date.
- Together these are model risk. The response has three layers: validation (test for bias, look-ahead leakage and sensitivity to wording), documented controls (who may use the tool, for what, with what data) and human oversight of every consequential output. The user stays responsible.

> **Model risk response = validation + documented controls + human oversight**

LOOK-AHEAD BIAS IN A BACKTEST, 2015 TO 2020 (ILLUSTRATIVE)

Model had read text to 2024	it knows the ending	**Look-ahead**
Backtest, per year	looks great	**14%**
Honest rerun, only what was knowable		**6%**
Gap from leaked knowledge	14 − 6	**8 points**

Key takeaways

- Supervised means labels: a category label is classification, a number label is regression.
- Unsupervised finds structure: clustering and principal components analysis.
- Deep learning is a kind of model, trainable either way, not a third set-up.
- Overfitting: great on training data, poor on new data, so judge on unseen data.
- An LLM predicts the next token from probabilities, picks, appends and repeats; fluent is not verified.
- Use the output as a draft and manage model risk with validation, controls and oversight.

Quick check

1. A model trained on loans labelled defaulted or repaid predicts new defaults. Which kind and task?
   Answer: Supervised learning: classification.
2. A model scores 95% on training data but 52% on new data. What happened?
   Answer: Overfitting: it memorised noise.
3. Why can one prompt give different wording, and why is fluent not verified?
   Answer: It samples likely tokens from probabilities; it does not check facts.