Pre-Class Discovery: Supervised Learning

Complete before the lecture. Bring your answers.

Data Science with Python – BSc Course

Instructions: Work through all six tasks before the lecture. There are no formulas here — just your intuition and pen-and-paper reasoning. Write directly on this sheet or bring a separate page with your answers. Total time: approximately 45 minutes.

Task 1: Sort the Companies by Outcome

10 min

Below are 12 companies described by four features. A year later, some of them defaulted on their debt and some did not. Your job: predict which companies defaulted **by eye**, based on the feature values.

ID	Sector	Rev. Growth (%)	Debt Ratio	Past 1Y Return (%)	Size (B)
1	Tech	28.0	0.15	35.0	42
2	Retail	-8.0	0.82	-28.0	3
3	Banking	5.0	0.40	8.0	120
4	Energy	-12.0	0.88	-35.0	5
5	Healthcare	12.0	0.25	15.0	80
6	Manufacturing	-15.0	0.90	-42.0	4
7	Utilities	3.0	0.55	4.0	65
8	Airlines	-22.0	0.95	-55.0	2
9	Pharma	18.0	0.30	22.0	95
10	Construction	-5.0	0.78	-18.0	6
11	Tech	32.0	0.20	40.0	55
12	Consumer	8.0	0.45	10.0	48

(a) Which companies do you predict *defaulted* in the following year? List their IDs and briefly explain your reasoning.

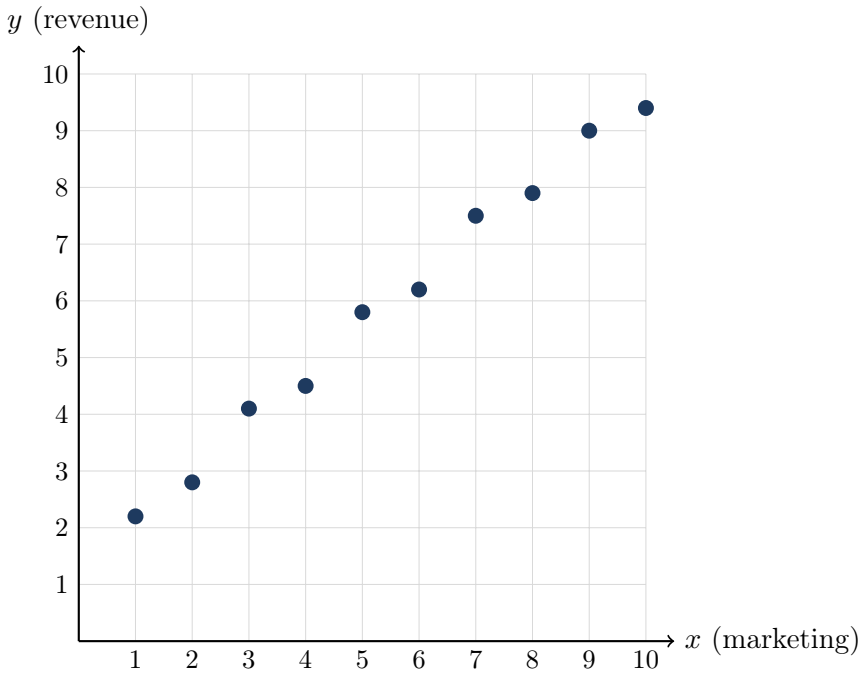
(b) Which two features did you rely on most? Are any features redundant (give similar information)?

(c) Unlike Task 1 of the unsupervised week, here each company has a *correct answer* (actually defaulted or not). What is the key difference between this kind of learning and the grouping task you did earlier?

Task 2: Fit a Line Through Points

8 min

Ten data points are plotted on the grid below. Each point shows (x, y), where x is marketing spend (in millions) and y is next-quarter revenue (in millions).



(a) Draw the single straight line that you think best summarizes the trend. Try to minimize the total distance between your line and the points.

(b) Read off your line: estimate the **slope** (how much does y change when x goes up by 1?) and the **intercept** (what is y when $x = 0$?).

(c) Suppose a classmate draws a different line. Could they still be "correct"? What objective rule could you use to decide whose line is better?

(d) If you built a quadratic curve that passed through *every* point exactly, would it predict the next point better? Why or why not?

Task 3: The Exam Score Debate

5 min

A classmate tells you:

> "I built a model that predicts which students will fail the final exam. My model achieves **95% accuracy** on the class data. This is a great model."

You later find out that in this class, only **5% of students** actually fail the final exam (the other 95% pass).

(a) Without any modelling skill, how could you write a one-line "model" that also reaches 95% accuracy on this data?

(b) Does the classmate's 95% accuracy tell you anything useful about their model's ability to actually identify struggling students?

(c) What question would you ask instead? Think about: *of the students my model flagged as likely to fail, how many really did fail?* Or: *of the students who really failed, how many did my model catch?*

Task 4: Which Features Matter? 7 min

You want to predict **next-month stock return** for a set of large US companies. Below are 8 candidate features you could feed into a model:

#	Feature	What it measures
1	P/E ratio	Price divided by earnings
2	Volatility	Recent return standard deviation
3	Momentum	Return over the past 12 months
4	Market cap	Total company size
5	Beta	Sensitivity to the overall market
6	Dividend yield	Annual dividend per dollar invested
7	Debt ratio	Total debt divided by total assets
8	R&D spending	Research and development expense

(a) Pick the **top 3 features** you believe are most predictive of *next-month* return. Rank them 1 (most useful), 2, 3 and justify each choice in one sentence.

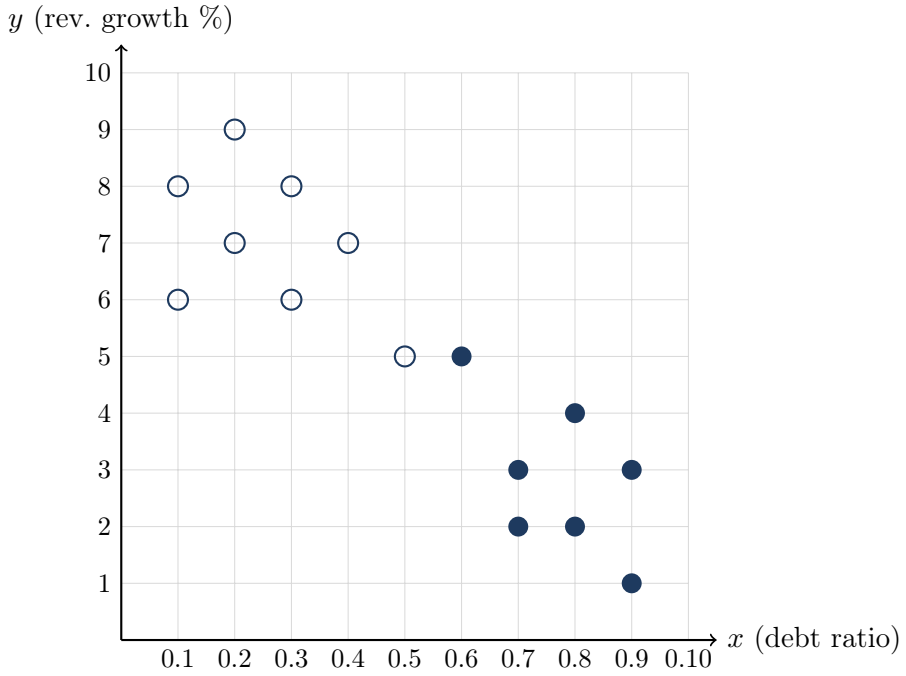
(b) Pick the **bottom 2 features** you would probably drop. Why?

(c) Suppose you have 500 features to choose from instead of 8. Would it be reasonable to decide by hand which features matter? What could an algorithm do for you that you cannot do manually?

Task 5: Draw the Decision Boundary

8 min

Fifteen companies are plotted below. Each point has two features: x = debt ratio and y = revenue growth (%). **Filled navy dots** are companies that defaulted. **Open circles** are companies that did not default.



(a) Draw a single **straight line** that separates defaulters from non-defaulters as well as you can. How many points does your line misclassify?

(b) Now try again with a **wiggly curve** that twists around to classify *every single point* correctly. Did you manage it?

(c) A new company arrives with debt ratio 0.55 and revenue growth 5%. What does your straight line predict? What does your wiggly curve predict? Which prediction do you trust more, and why?

Task 6: Your 3 Questions

5 min

After working through Tasks 1–5, you have seen several ideas: learning from labeled examples, fitting lines and curves through data, the danger of high accuracy on imbalanced problems, choosing features, and drawing decision boundaries.

Write **three questions** you want answered during the lecture. These can be about anything — how computers actually learn from data, which method to use when, what can go wrong, or how any of this applies to finance.

Question 1:

Question 2:

Question 3: