Pre-Class Discovery: Solutions and Discussion Notes

Instructor copy – do not distribute before the lecture.

Data Science with Python – BSc Course

Task 1: Sort the Companies by Outcome

Model answer

(a) Predicted defaulters: companies 2, 4, 6, 8, 10. All five share three warning signs:

- Negative revenue growth (business shrinking)
- High debt ratio (≥ 0.78)
- Large negative past-year return (the market already priced in distress)

Borderline case: company 10 has the mildest signals of the five. Some students will keep it out of the "default" list; that is defensible.

Predicted non-defaulters: companies 1, 3, 5, 7, 9, 11, 12. All show positive revenue growth, low-to-moderate debt (≤ 0.55), and positive past-year returns.

(b) Most useful features: Debt ratio and revenue growth. Past 1Y return is highly correlated with both — it mostly confirms what those two already say. Sector is weakly informative (airlines, energy, retail appear more often among defaulters than tech or pharma). Size is useful but secondary: small companies dominate the defaulters, large companies dominate the survivors.

(c) Key difference from unsupervised learning: Each row has a *label* (defaulted / did not default). We are not just finding structure in the data; we are learning a rule that maps features to a known outcome. This is the defining feature of **supervised learning**: a teacher (the labels) tells us the right answer for each training example.

Common misconceptions

- Students sometimes rely on sector alone ("airlines always fail"). Push back: sector is a feature, but not a reliable predictor on its own — large airlines with low debt can be stable.
- Some students memorize the 5 IDs. The point is not which rows are defaulters, but which *features* predict default.
- A few students will ignore past return because "the stock price does not cause the default." Correct — but in practice, past return is often the most predictive feature because it aggregates all public information. This is a good conversation about causation vs. prediction.

Discussion prompt

This task sets up the entire module. Reveal during the lecture:

- The feature "past 1Y return" duplicates information from the other features — a Lasso penalty (L22) would drop it.
- Sorting companies into defaulters / non-defaulters is exactly binary classification (L25–L28).
- The students' intuition of "which features matter" becomes formal via coefficient magnitudes and feature importance.

Task 2: Fit a Line Through Points

Model answer

(a) Best-fit line by eye: Students should draw a line that passes roughly through the cloud of points, rising from lower-left to upper-right.

(b) Estimated slope and intercept: The points span roughly from (1, 2.2) to (10, 9.4). Slope $\approx (9.4 - 2.2)/(10 - 1) = 7.2/9 \approx 0.80$. Intercept $\approx 2.2 - 0.80 \times 1 = 1.4$. A reasonable eyeball line is $y \approx 0.80x + 1.4$. (The actual ordinary-least-squares fit is very close to this.)

(c) Objective rule: Yes, classmates will draw different lines. The objective rule used in ordinary least squares (OLS) is: *pick the line that minimizes the sum of squared vertical distances between the line and the points.* The lecture will show that this choice yields a unique "best" line.

(d) Quadratic through every point: A polynomial of degree 9 can pass through all 10 points exactly. But between and outside those points it will wiggle wildly. On the *next* quarter's data the predictions will be much worse than the simple line. This is **overfitting**: fitting the noise, not the signal.

Discussion note

This is exactly the bias–variance trade-off preview for L21–L23. High-variance models match the training data perfectly but generalize badly. The line is biased (it misses some points) but stable (generalizes well).

Common misconceptions

- Students often read the intercept as "where the line crosses the drawing," which may be above $y = 0$. Remind them that the intercept is y when $x = 0$, which may require extrapolating.
- Some confuse "best fit" with "goes through the most points." The OLS line may pass through none of the points exactly.

Task 3: The Exam Score Debate

Model answer

(a) The trivial 95% model: `predict_pass()`, i.e. always predict "student will pass." Since 95% of students do pass, this model is right 95% of the time — with zero learning effort and zero usefulness.

(b) Does 95% mean the model is good? No. With a 5% minority class, any trivial rule reaches 95%. To judge the model we must know what it does on the *minority class* (the struggling students). A 95% accuracy tells us almost nothing.

(c) Better questions:

- Of the students my model flagged as "will fail," what fraction actually failed? → **Precision**.
- Of the students who actually failed, what fraction did my model catch? → **Recall** (sensitivity).
- A single number that balances both: **F1 score**.

This task directly motivates L27 (classification metrics beyond accuracy) and L28 (class imbalance techniques).

Common misconceptions

- Students sometimes insist that accuracy is "the" metric. Emphasize: accuracy is one metric; it is often misleading on imbalanced problems.
- Some write "the model should be better than random." True, but with 5% minority, "random guessing" is 95% of the time "pass" — the bar is much lower than students expect.

Discussion prompt

Ask the class: if the school uses this model to decide which students get extra tutoring, what happens when the model has high accuracy but low recall? (Answer: struggling students are missed; the policy fails precisely where it should help most.)

Task 4: Which Features Matter?

Model answer

(a) Top 3 features (typical good answers):

1. **Momentum** — the most robust predictor of short-horizon returns in empirical finance (Jegadeesh & Titman, 1993).
2. **Volatility** — high-vol stocks have different expected returns than low-vol ones (low-vol anomaly); volatility clusters.
3. **Beta** — mechanically relates stock returns to market returns; if you have a market view, beta converts it to a stock-level prediction.

Other defensible top-3 picks include P/E ratio (value anomaly) and market cap (size factor). These show up in the Fama-French factor models covered in L24.

(b) Bottom 2 features:

- **R&D spending** — reported annually, slow to change, not very predictive at the monthly frequency.
- **Dividend yield** — relevant for long-horizon returns, but monthly movements are dominated by price noise.

The point is not the specific answer but the reasoning: match the feature frequency to the prediction horizon.

(c) 500 features: Manual selection is impractical. Algorithms can help via:

- **Regularization**: Lasso shrinks many coefficients to exactly zero, performing feature selection automatically (L22).
- **Feature importance**: decision trees and random forests rank features by how much they reduce impurity (L26).
- **Univariate screening**: discard features with near-zero correlation to the target.

Discussion note

Highlight during the lecture: students' intuition (top-3 choices) is often right — but data-driven methods scale to 500 features or more, and they discover surprising interactions intuition misses.

Task 5: Draw the Decision Boundary

Model answer

(a) Straight line: A diagonal from roughly (0.3, 9) down to (0.8, 0) separates the two groups well. It will likely misclassify **1–2 points**: the ambiguous middle pair (5, 5) and (6, 5) are hard to separate with any line.

(b) Wiggly curve: Yes, a sufficiently flexible curve can wrap around every point to achieve 100% accuracy on this training set — but only by carving out tiny pockets to isolate the middle ambiguous points.

(c) New company at (0.55, 5):

- The straight line predicts based on which side of the line (0.55, 5) falls on. Depending on where the student drew the line, this will be either default or non-default — but the classification is stable.
- The wiggly curve may put (0.55, 5) in a tiny pocket created to accommodate noise in the training data, giving an arbitrary prediction.
- **Trust the straight line**, because the wiggly curve overfits the training data and does not generalize.

Discussion prompt

This directly sets up three lecture topics:

1. Logistic regression (L25) draws a linear decision boundary like the student's straight line.
2. Decision trees (L26) can draw axis-aligned wiggles, useful but prone to overfitting without pruning.
3. The bias–variance trade-off: simpler models have more bias but lower variance; more flexible models fit training data better but generalize worse.

Common misconceptions

- "A model that gets 100% on training data is the best model." No — training accuracy is not the goal; test accuracy is.
- "More flexibility is always better." No — unconstrained flexibility leads directly to overfitting.

Task 6: Your 3 Questions

Typical student questions and where they are answered

- "How does a computer actually find the best line?" — L21 normal equations / gradient descent.
- "What if I have more features than data points?" — L22 Ridge / Lasso / ElasticNet regularization.
- "How do I measure how good a regression model is?" — L23 MSE, RMSE, MAE, R^2.
- "What is a factor model? How does it explain returns?" — L24 CAPM and Fama-French.
- "How do I convert a linear output into a probability?" — L25 logistic regression and the sigmoid.
- "How does a decision tree decide which feature to split on?" — L26 Gini impurity and information gain.
- "What metric should I use for fraud detection?" — L27 precision, recall, F1, ROC/AUC.
- "What do I do when one class is 1% of the data?" — L28 SMOTE, class weighting, threshold tuning.
- "How do I know if my model is overfitting?" — L23 cross-validation, train vs. test gap.
- "How is supervised learning different from unsupervised?" — Opening slides: labels vs. no labels.

Instructor tip

Collect student questions at the start of the lecture on sticky notes or via a shared doc. Revisit them at the end to check whether the lecture answered each. Unanswered questions make good exam review topics and flag which sections may need more depth next time.