Data Science

Msc Supply Chain & Retail Management (2025-2026)

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

Context: You are working as a data scientist in a retail supply chain company. Your goal is to use machine learning to improve decision-making in purchasing, supplier selection, and inventory management.

Exercise 1: Understanding Overfitting, Underfitting, and Hyper-parameter Tuning

1. Explain the concepts of **overfitting** and **underfitting** in machine learning models. How are they related to training and test error?
2. Describe the **bias-variance tradeoff**. How does model complexity affect bias and variance?
3. Give one example each of a model that tends to underfit and one that tends to overfit in the context of supply chain forecasting.
4. What is a **hyperparameter**? How does it differ from model parameters?
5. For each of the following models, name at least one important hyperparameter that affects overfitting:
   - Decision Trees
   - Random Forests
6. Briefly describe how **cross-validation** can be used to tune hyper-parameters.
7. How can we improve generalization capacities of a simple decision tree?
8. What are the main differences between *bagging* and *boosting*.

Exercise 2: Problem Identification (Classification or Regression?)

Classify each scenario below and justify your answer:

1. Predict the total delivery time (in days) of an order.
2. Predict whether a supplier will delay delivery beyond the agreed lead time.
3. Estimate the total cost of procurement based on supplier, quantity, and market prices.
4. Classify whether a product should be reordered this week.

Exercise 3: Procurement Cost Prediction

You are predicting procurement cost from:

- `Region (A, B, C)`
- `Quantity`
- `Delivery Time`

Region	Quantity	Delivery Time	Cost
A	100	5	1200
B	200	4	2400
A	100	6	1100
C	300	7	3100
B	150	5	1900

1. Propose a split on one feature (e.g., `Quantity` < 150). Compute the **variance of cost** in each resulting subset.
2. Calculate the **reduction in variance (information gain)** after the split.
3. Which split (by region, quantity, or delivery time) gives the best regression performance?
4. Discuss the risk of overfitting in regression trees and how pruning can help.

Exercise 4: Delivery Delay Prediction

You aim to predict a potential delay in supply chain using the following data:
Exam

Supplier	Items	Delayed?
A	50	Yes
A	40	No
B	60	Yes
A	30	No
B	70	Yes

1. Calculate the **Gini impurity** of the root node.
2. Compute the Gini impurity of the children nodes after splitting on `Supplier`.
3. Calculate the **information gain** from this split.
4. Suggest an alternative split on `Items < 45` and perform the same calculations.
5. Which split is better? Explain.