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

Due date: 20 February 2026

Team Work: at most 2 students per group

The work will be submitted directly to Brightspace in the space provided for this purpose. Be sure to include the names of the group members so that you can also report the grades for your partner. Your submission will include a report explaining the methodology used, the results obtained, as well as the analyses, conclusions, and any recommendations, but also the code used to obtain the results. Only one submission per pair.

1 Linear Regression: a case study

In this part, we start by studying the data related to matches that took place the previous week in order to deduce different information for the following week's matches. Saying differently, we are dealing with a regression task.

1.1 A linear model to predict attendance

For this, we will consider the dataset named **attendance_train.csv** which contains several information about matches that took place in different stadiums. We will first study this data in order to build a model that will allow us to predict the number of spectators at a match.

Tickets and Promotions We also have the following information: the price for a place is 20 euros during the week and 25 euros during the week-end. The promotion means that all tickets are 10% cheaper Cost Stadium We have the following information regarding the size of the stadium:

- when the stadium/hall has less than 1000 seats, the match has a cost of $50,000$ euros.
- when the stadium/hall has between 1000 and 2000 seats, the match has a cost of $60,000$ euros.
- when the stadium/hall has more than 2000 seats, the match has a cost of $70,000$.

Other revenues (consumptions) We make the following assumpions regarding the consumption(food and drink) during a match :

- 40% of the seated people consume an average of 6 euros worth of drinks (2 drinks)
- 30% of the seated people consume an average of 8 euros worth of food (one sandwich)
- 80% of the people standing consume an average of 10 euros worth of drinks (3 drinks)
- 20% of those seated consume an average of 8 euros worth of food (one sandwich)

Cost employees Finally, the cost of an employee to serve the audience must be taken into account. The average cost of an employee is 500 euros per game. However, the current regulations require one employee for every 100 spectators.

Finally, we assume that people take all the seats before filling in the remaining standing room or loft seats.

1.2 Expected work on the Training Data

Your objective is to build the best regression model that can predict the number of spectators at this event. You will therefore compare the performance of a linear regression and a regression tree on this set, compare the outputs of these different models and their performance, and identify the key variables that led to the decision made by the models studied.

1.3 Prediction tasks

In the following we will try to predict the number of spectators for a set of games that will be held next week in different stadiums based on the the previous learned model. We will use the same information as before to make this prediction. All variables are known in advance except the temperature which has been estimated using a model based on time series. The information are available in the file **attendance_test.csv**.

1. Use the previous learned models to predict the attendance for these future match and compare their performances. Comment the obtained results
2. Based on predictions made using test data and previously provided information regarding customer consumption during a game. Estimate the quantity of food required to meet spectator needs, bearing in mind that an additional 10% margin should always be added to the quantities ordered to avoid stock shortages.
3. Does the strategy adopted allow for events that are profitable? Do we observe the same behavior from one model to another?

2 Classification Task

Artificial intelligence (AI) is transforming supply chain management, yet progress in predictive tasks—such as delivery delay prediction—remains constrained by the **SynDelay**, a synthetic dataset specifically designed for delivery delay prediction. Generated using an advanced generative model trained on real-world data, **SynDelay** preserves realistic delivery patterns while ensuring privacy protection. Although not entirely free of noise or inconsistencies, **SynDelay** provides a challenging and practical testbed for advancing predictive modelling in supply chain AI.

You have various pieces of information in this dataset, and the goal is to predict whether the order will arrive late (label = 1 or 2) or whether it will be delivered on time.

We want to compare the performance of a decision tree, a random forest, and a set of trees trained by boosting, and we will use decision trees to identify the key factors that determine whether or not an order will be delivered late.

Care will be taken to perform the training properly and to reserve part of the data to test the performance of the different models.