A colleague from the Learning Team, who is responsible for digital trainings, has approached you and would like your help in understanding the completion rates for their online trainings. They would also like you to build a model that estimates the probability that an employee completes a training.

To this end, they have provided you two data files: “employee.csv” and “performance.csv”. The first file (“employee.csv”) contains HR data regarding our employees, while the second file (“performance.csv”) contains information about an employee’s performance rating from our performance management system.

Using these two files (containing synthetic data) and either R or Python:

Undertake the necessary steps to

1. Build a model that estimates the probability that an employee completes a training
2. Write a short summary (bullet points and comments in your code/notebook/markdown are perfectly fine) that gives your colleague some insights into the top 5 drivers of your estimates as well as an evaluation of the model’s performance.
   Since we do not have provided a detailed codebook, feel free to provide your own interpretation as to what these variables might mean (don’t worry, there are no wrong answers).

Please note:
This should take not more than 1 hour (at maximum 2 hours). We are interested in seeing your general workflow and how you approach data problems; while important, model performance is not the ultimate outcome of this task.