Instructions

Please perform the given assignment alone and submit the solution within 5 working days. The code needs to be written in R using the common numerical, statistical, machine learning and visualisation libraries, and should be clear and well annotated. Estimated time required is 3-5 hours.

Task

Data: You are given 2 data files:

1. Baseball_train.csv: a dataset containing 900 data samples
   - Team name (Team)
   - Run scored by the team (RS),
   - Run scored against the team (RA),
   - Team's on base percentage (OBP),
   - Team's Slugging percentage (SLG)
   - Team's Batting average (BA)
   - Whether team Qualifies for playoffs or not(Playoffs) -if 1 then team qualifies and if 0 it doesn't.

Your target variable is Playoffs.

2. baseball_unknown.csv: a dataset containing around 300 data samples in similar format, but where the value of target variable Playoffs is missing.

Objectives

The 3 main objectives of the task are:

1. Explore the data, understand its structure and identify the key input variables that drive the target.
2. Train and test a model that predicts, to the best extent possible, the target value from some or all the input variables.
3. Generate a prediction for the unknown dataset.

Deliverables

1. R code file
2. Predicted values of playoffs for unknown data set. The file should be in .csv format
3. A brief explanation of your approach. This should include which model you chose, what steps did you do before building the model, why you selected certain variables.