Stats 191

Project

The following project description walks you through a hypothetical scenario akin to a draft for athletic teams. Resorting to large databases of player-dependent statistics aggregated over their athletic career has become a standard tool in assembling team rosters, with groups of statisticians driving much of the decision-making. You will take on the role of these statisticians for this project, and use what you have learned to suggest team rosters for various purposes.

Statistical work in these settings is often a collaborative effort, and so you are very much invited to work in groups of up to five students—each of you will need to submit identical reports, clearly indicating whom you joined forces with. These reports are meant to communicate to the reader the exact methodology you pursued and why you chose to pursue it; please provide sufficiently detailed explanations that would both convince the hypothetical head-coach of the team you are assisting to follow your guidance, and allow any fellow statistician to exactly reproduce your work. Grading will be based both on the accuracy of your predictions as well as the clarity of your report.

Please submit your reports to Gradescope by March 20, 23:00.

College soccer season is about to commence, and so your university's soccer coach is looking to assemble a number of soccer teams to enter the competition with. To do so, a (somewhat excessive) try-out is organised, consisting of ten consecutive days and games, during which each candidate is closely monitored and evaluated on a range of metrics. More concretely, data is available on

- the overall *game score*: a score between 0 and 100 that evaluates a candidate's performance in a given try-out game,
- which *year in university* a given candidate is enrolled in: frosh, sophomores, juniors, and seniors all applied for positions on the team,
- a student's *athletic history*: the type of sports a student has engaged in prior to trying out,
- an *overall fitness score*: a score summarising a student's athletic state (as measured the morning before each game),
- what percent of *training sessions* were attended on each day before the game: the university offered soccer training sessions, which nominally are mandatory, yet attendance was not enforced,

- how many optional *strategy sessions* a candidate attended each day before the game: though optional, candidates were highly encouraged to attend these meetings,
- how many *hours of a sleep* a candidate racked up the night before each try-out game: rounded to the nearest integer,
- whether a student identifies as an *early bird* or *night owl*: early birds tend to sleep and wake early, while night owls prefer staying up late,
- the *number of meals* a student consumed before each try-out game,
- whether every try-out game itself was scheduled for a *morning, noon* or *evening* slot,
- whether the student performed the try-outs at the *west coast* or *east coast*: even though your university is located on the west coast, a few students needed to run through parts of their try-outs at the east coast due to traveling arrangements.

$n = 1000$ students underwent the try-outs and their data can be found on Canvas under the *files* folder. The head-coach would like you to analyse this data with respect to three goals:

- Two teams of 10 students each are to be assembled to enter the West-Coast-Cup and East-Coast-Cup, respectively (goalkeepers have already been picked in previous try-outs, and selecting appropriate candidates for substitutes is moved to a separate decision process).
- The regular college soccer season consists of twenty games, and the head-coach has obtained information on the expected player-score for each of these twenty opposing teams. They'd like you to suggest two candidate rosters: one that *minimises* the chances of yielding a single game, and one that *maximises* the number of expected wins.
- How to best coach a player is a complicated question, with strategies so far being based on intuition more than systematic exploration. The head-coach would like you to use available data to inform them on which player-covariates their coaching program should pay most attention to, once a student has been recruited to the team.

To do so, the following additional information may be useful:

- The expected player-scores of the twenty opposing teams, too, are uploaded on Canvas.
- A data-set recording team-scores and outcomes across 200 games is uploaded as well.
- To avoid overloading students, the three proposed rosters (for the West-Coast-Cup, East-Coast-Cup and regular season) are to be disjoint of one another; that is, no individual student is supposed to serve on more than one team.
- Games in the West-Coast-Cup are exclusively scheduled for morning slots, while East-Coast-Cup games happen during evenings.
- Regular season games are equally likely to be scheduled during any of three slots.