Mariners 2026 Data Science Intern Problem Set

Below is the set of problems we would like you to complete. You will be evaluated not just on your predictions, but also on your ability to communicate clearly and the quality of your process.

1. In the Google Drive folder linked below you will find four files: 'train_data.csv', 'test_data.csv', 'outfield_position.csv' and 'glossary.csv'.

   a.   'train_data.csv': data from selected "sacrifice" plays where an out is recorded by an outfielder and a runner on 2nd or 3rd base has an opportunity to advance (or score from 3rd). Successful advances are indicated in the 'runner_advance' column with a value of 1. A value of 0 can either indicate a failed attempt to advance or no attempt at all.
   b.   'test_data.csv': equivalent data from a separate selection of plays where the 'runner_advance' information has been withheld.
   c.   'outfield_position.csv': player tracking data, if available, for the outfielders involved in the sacrifice plays.
   d.   'glossary.csv': a list of the columns included in the other tables explaining what each column represents.

   Your objective is to predict and fill the 'runner_advance' column in 'test_data.csv' with an estimate from 0 to 1 of the probability that the runner will successfully advance (for example, a 'runner_advance' of 0.9 indicates a 90% probability that the runner will advance on the play). You may use whatever method or language you like, but **you must submit the code you used to generate and evaluate your predictions**. You will be evaluated on both the log loss score of your predictions and on your process in generating the predictions. Please also include a brief explanation of your process **for an R&D audience** and what steps you would take to improve on this model if you had more resources.

2. Outfielder **6HuoaRi0** has just been called up from Triple-A and the Major League coaching staff would like to know how effective he is at preventing runners from advancing on sacrifice plays, believing this could be a way for him to separate himself and earn more playing time. Using the data provided, prepare a 1-2 page report for **a coaching staff audience** to answer their query. Please supply any code used to generate this report.

3. In approximately 300 words, what is a recent mistake the Mariners organization has made, and why do you consider it a mistake?

Data for problems 1 and 2 can be found here:
https://drive.google.com/drive/folders/1T8_L1OpDDs4wEJlNEVe7M-twPytV97x-?usp=sharing

Due to the large size of the test dataset, consider either zipping the file before emailing or storing it in a public cloud storage site like Google Drive or Dropbox.

If you have any questions, or need clarification of any kind, please message Forrest Diamond at fdiamond@mariners.com.