Challenge for Job Candidates

Thank you for your interest in working with us.

First, we want to be upfront with you that we are open to all arrangements; however, given the intense and collaborative nature of the work, our preference is for folks for whom this work would be their top professional priority, and who can be based or spend a significant amount of time in Boston for the next 6-12 months.

Assuming you are still interested, please answer exactly two of 1) – 3), and all of 4) – 6).

The goal of this Challenge is to help us gauge if you'd be a good fit on our team, and also help you gauge if the type of work you'd be doing would be a good fit for you. At the least, these exercises should be helpful for you to learn more about the types of problems we face.

We want to emphasize confidentiality. We are a small group, and trust is key. Please do not post or share this Challenge anywhere, and please do it yourself; by submitting a response, you are giving us your word that you are the sole author of your response.

Please show your work – share your code, and be clear on what data you've used. Please do not provide answers that are more complicated than they need to be; please do not go out of your way to use fancy jargon or demonstrate your knowledge of advanced techniques. Use what's appropriate. Simple, effective, interpretable >> complicated, effective, uninterpretable.

Answer exactly two of 1) – 3); choose any two you'd like

1) Using data from this webpage alone (https://www.pro-football-reference.com/years/2019/) - do not use any data that can be found only through links on this webpage - build a model which you'll use to determine the probability that the home team wins, and to predict the final score of the game, for each game for week 15 (http://www.nfl.com/schedules/2019/REG15). Send us model details and prediction results. How confident are you in your predictions? If you wanted to get a more quantitative measure of your confidence in your predictions, what might you do?

2) Using the data in the attached file ('cfb_games_for_ml_task'), and no other data, you are tasked with coming up with win probability forecasts for future college football games based on the point spread (and total, if you so desire).

Be conscious of overfitting to historical biases. Also be aware that since scoring is (usually) in increments of 3 and 7, certain score differentials will occur with greater frequency than others. This should have an impact on your analysis.

For simplicity, assume that each line is 50% cover probability.

3) You need to forecast NFL quarterback (QB) performance. Using the data in the attached file ('qb_by_game' file), and no other data, forecast a number for yards_adj using previous yards_adj and any other variables in the dataset you think are relevant. Think about how to properly regress the QB to a relevant "mean" without introducing survivorship bias.

Several notes about the data:

- 'dpos': QB's draft positions

- 'start': QB's first year
- 'reg_play': number of effective plays (not every play is weighted equally; garbage time is de-weighted)
- 'yards_adj': yards per offensive play, contextualized (not every play weighted equally)
- 'yob': QB year born
- 'value': value of draft position (based on Thaler-Massey methodology)

Answer all of the below 4) – 6)

4) You are interested in building profitable gambling models in soccer, despite having no experience with soccer data and little knowledge of the sport. A friend recommends this paper to you, which you read: https://arxiv.org/pdf/1802.07127.pdf

Now it's time to create your own team and player ratings, predict game outcomes, etc., with an eye to making profitable bets. Do you use the framework used in this paper? Why or why not? If not, what framework do you use?

5) Give an example of a project you've done where you've had to clean, organize, combine multiple datasets. Think about something you’ve done that would be relevant if you were putting together a database of college, G League, international and NBA players, recognizing that some players end up spending some time in all of those four places, have many seasons of data associated with them, and also have non game playing metrics that might be of interest.

6) What experience, if any, do you have working with player location (coordinates) data?

Again, please do not provide answers that are more complicated than they need to be. Please do not go out of your way to use fancy jargon or demonstrate your knowledge of advanced techniques. Use what's appropriate. **Simple, effective, interpretable >> complicated, effective, uninterpretable**.

Finally, we realize that the above is a lot, and you have other obligations, so take your time completing the challenge. Just keep us posted on approximate timing.

Thank you, and we look forward to hearing from you soon!

Rufus Analytics Team