Assignment: Signal and Noise in Batter Outcomes

This assignment is designed to simulate the type of open-ended data assessment you might face in an interview with a professional baseball team's analytics department. Your task is to build, validate, and apply models of batted ball outcomes, with a focus on distinguishing signal from noise in player performance.

Part 1: Predictive Modeling

Fit a linear regression model[1] and a random forest model (recommended package: ranger) to predict the linear weight[2] of the batted balls in event.csv (on Canvas) using launch speed, launch angle, and spray angle. For each model, show a summary or visualization[3] of the fit. Explain the rationale behind your modeling choices, including your feature selection and transformations. Describe what aspects of batted ball outcomes your models capture well, and what they might miss.

Bonus: In addition to the linear regression model and the random forest model, fit either a generalized additive model (recommended package: mgcv) or a gradient boosting model (recommended package: xgboost).

Part 2: Model Validation

Test the out-of-sample predictive performance of both the linear regression model and the random forest model and report the out-of-sample root mean square error (RMSE) of each model. For the random forest report the tuning parameters, and explain how you chose them. Discuss which model performs best for this task and why (i.e., why does that model achieve the lowest test error?).

Bonus: Include either a generalized additive model (GAM) or a gradient boosting model in this comparison.

Part 3: Player Evaluation

For the best performing model from Part 2, calculate the residual (linear weight minus prediction) on each batted ball. Fit a linear mixed-effects model to estimate how much of this residual is signal and how much of it is noise. Using regression to the mean, estimate each batter's true talent for over-performing their linear weight predicted from batted ball trajectories. Report the top five and bottom five[4] mean-regressed over-performers (including the estimated values).

Bonus: Select two players from your list and make a baseball-based argument for why they may have a true skill for over- or under-performing relative to batted ball outcome model expectations.

Deliverables

- A PDF report (max 4 pages) with labelled sections corresponding to Part 1, Part 2, and Part 3. This report should contain absolutely no R code. It should be well written and easy to follow. Summarize your findings and leverage data visualization as appropriate.
- A R script (.R file extension) with all of the code you used to generate the results for your report

[1] For the linear regression, think carefully about transformations of variables. For example, launch angle would probably be better as a quadratic term rather than a linear term, and you should consider interactions between variables as well.

[2] You'll need to join linear_weight.csv (on Canvas) into the event table to get the linear weight of each batted ball.

[3] See `ranger::importance.ranger()` for random forest variable importance plots.

[4] i.e., the top five under-performers

Rubric

You will get feedback on your work product based on several criteria. Within each of those criteria, the feedback will be: Missing (0%), Needs Improvement (70%), Good (85%) or Exceeds Expectations (100%). Your grade on the assignment will be the average of the grades across criteria. The criteria are:

1. **Predictive Modeling.** Is the implementation correct? Are your interpretations justified?
2. **Model Validation.** Is the implementation correct? Are your interpretations justified?
3. **Player Evaluation.** Is the implementation correct? Are your interpretations justified?
4. **Data Visualization.** Have you used data visualization effectively to convey your findings?
5. **Written Communication.** Is the report well written and easy to follow?

Reminders

- Please **anonymize** your submission by removing any personally identifiable information (including file paths in your R script that contain things like usernames!).
- In this class we value **creativity** and **critical thinking**! Don't just parrot what you've been taught—bring your own experiences to bear on the assignment. If you disagree with something we've covered in class, that's strongly encouraged! Be sure to explain your reasoning.