A1. Background on you/your team

- Competition Name: March Machine Learning Mania 2022 - Women's
- Team Name: AI Fortune-telling
- Private Leaderboard Score: 0.39924
- Private Leaderboard Place: 4th Place
- Name: Kevin Liu
- Email: wii365@gmail.com

A2. Background on you/your team

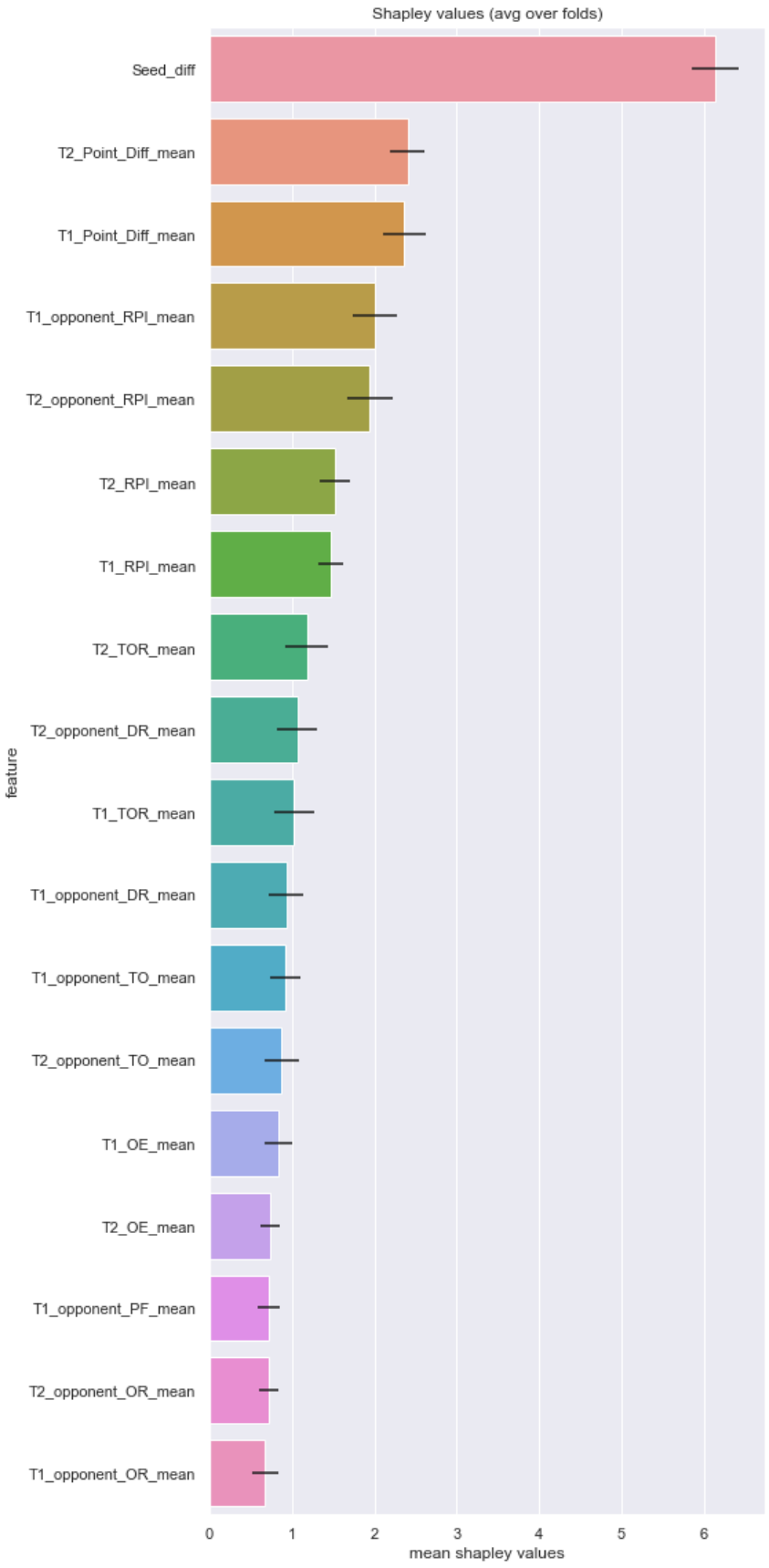
- What is your academic/professional background?
  I am a machine learning practitioner fascinated by predicting the future.
- Did you have any prior experience that helped you succeed in this competition?
  I've acquired a lot of experience from 9 other competitions that I joined on Kaggle, mostly with structured tabular data.
- What made you decide to enter this competition?
  This is a unique competition on Kaggle. The NCAA tournaments attract millions of fans each year. Watching the results unfold on TV adds another layer of excitement. I started my first NCAA competition in 2020, however, the tournaments got canceled due to COVID. My model in 2021 was defeated halfway by the upsets. I decided to make it more resilient for 2022.
- How much time did you spend on the competition?
  About a week for 2022, 3-4 weeks back in 2021 and 2020.

A3. Summary

- The training method(s) you used (Convolutional Neural Network, XGBoost)
  XGBoost to predict point spreads. Univariate spline to get win probability.
- The most important features
  Seed difference, mean point difference, etc.
- The tool(s) you used
  Anaconda, VS Code.
- How long it takes to train your model
  About 45 seconds.

A4. Features Selection / Engineering

- What were the most important features?


Shapley values (avg over folds)
Seed_diff
T2_Point_Diff_mean
T1_Point_Diff_mean
T1_opponent_RPI_mean
T2_opponent_RPI_mean
T2_RPI_mean
T1_RPI_mean
T2_TOR_mean
T2_opponent_DR_mean
T1_TOR_mean
T1_opponent_DR_mean
T1_opponent_TO_mean
T2_opponent_TO_mean
T1_OE_mean
T2_OE_mean
T1_opponent_PF_mean
T2_opponent_OR_mean
T1_opponent_OR_mean
feature
0
1
2
3
4
5
6
mean shapley values

- How did you select features?
  I used SHAP to explain the feature importance of the XGBoost model to reduce the number of features to 50, and then I manually removed another 30 by trial and error.
- Did you make any important feature transformations?

  I created many composite features using various team evaluation metrics.
  https://www.nbastuffer.com/analytics-101/team-evaluation-metrics/
- Did you find any interesting interactions between features?

  Seed difference dominated the feature importance, making the seed number of both teams virtually useless.
- Did you use external data? (if permitted)

  No.

A5. Training Method(s)

- What training methods did you use?

  5-fold cross-validation, repeated 5 times for averaging.
- Did you ensemble the models?

  No.
- If you did ensemble, how did you weigh the different models?

  N/A.

A6. Interesting findings

- What was the most important trick you used?

  Boosting UConn to #1 seed as they had 4 home games before the Final Four.
- What do you think set you apart from others in the competition?

  Feature engineering, game overriding, and a bit of luck.

A7. Simple Features and Methods

- I tried removing more features and re-ran my model against past competitions, but the LB positions only dropped.

A8. Model Execution Time

- How long does it take to train your model?

  About 45 seconds.
- How long does it take to generate predictions using your model?

0.5 second.

- How long does it take to train the simplified model (referenced in section A6)?

  About the same.

- How long does it take to generate predictions from the simplified model?

  About the same.

A9. References

- Raddar's well-known 2018 1st place solution: https://www.kaggle.com/competitions/womens-machine-learning-competition-2019/discussion/80689