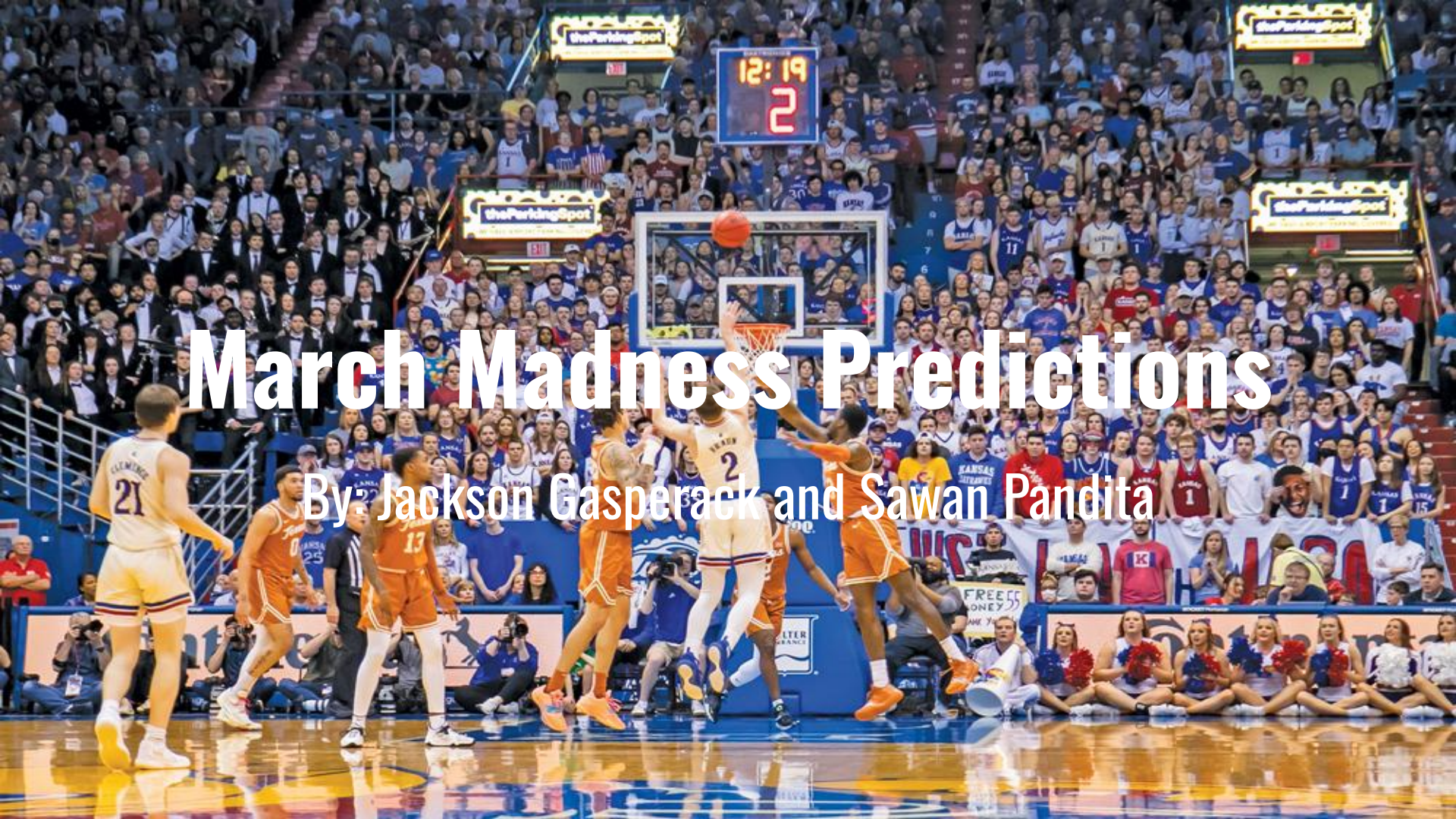


March Madness Predictions

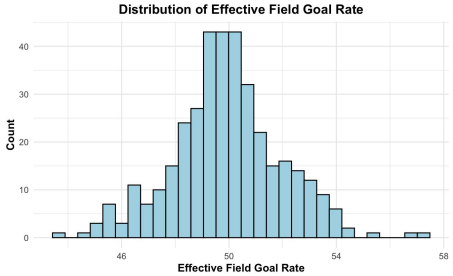
By: Jackson Gasperack and Sawan Pandita

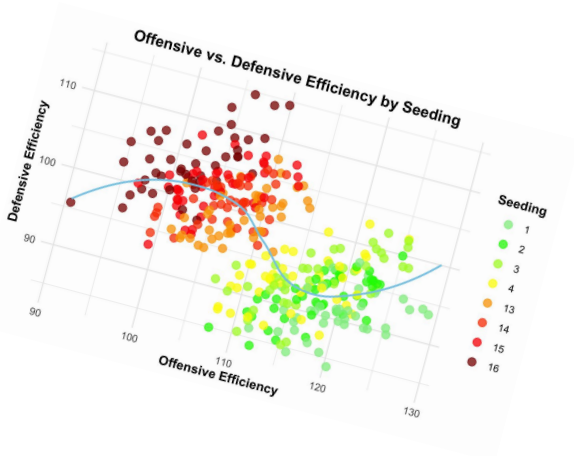
The **PROBLEM**

- Every year, millions of people fill out their bracket in hopes to be the person that correctly predicts the winner of that year's Basketball National Championship.
- However, it is nearly impossible to do this, and we even had to modify our research question since the rate of success is so low our models began to underfit.
- In order to tackle this, we wanted to create a model that could predict which teams would be most likely to make the Final Four round as accurately as possible.
- We tried numerous models across multiple feature sets to find the best possible one.

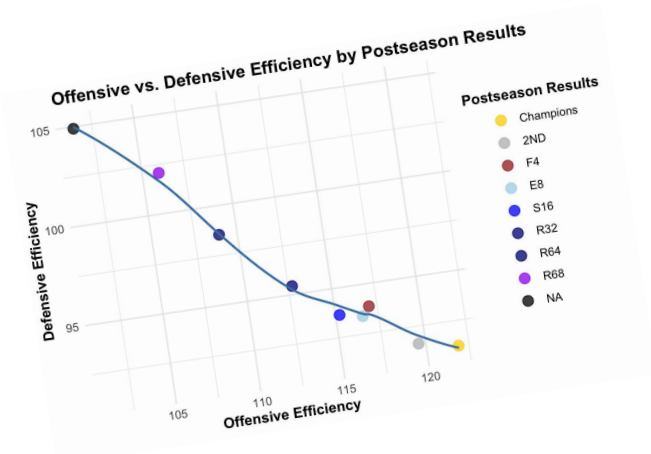
Fig. 1, by FiveThirtyEight

Data Analysis





- After loading in our data, we cleaned it by removing N/A values, removing unnecessary columns, and renaming columns.
- We then looked at the distributions of all of our variables which all were mostly symmetric and bell-shaped.
- Then we wanted to find correlations between ranks and stats, i.e. Offensive and Defensive Efficiency grouped by seedings and postseason round of elimination.
- We found that the teams that generally do better seeding and playoff-wise perform better offensively whereas those teams with low seedings or early eliminations are better defensive performers.

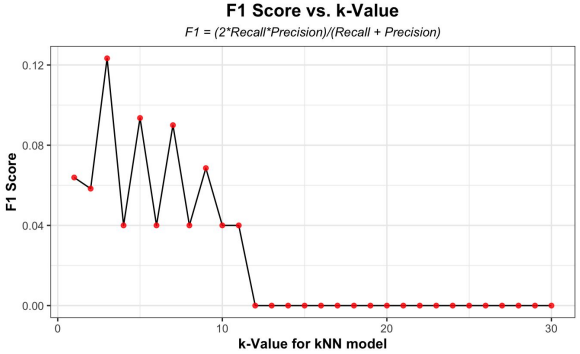


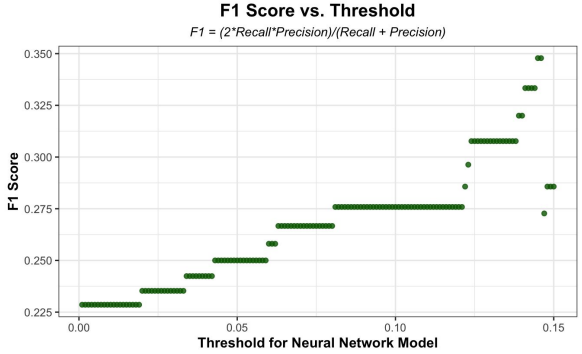
Feature Selection

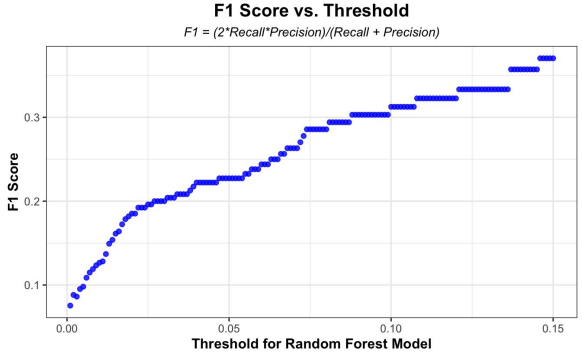
- First, I fit a logistic model with all possible predictors and then one with no predictors as my full and null models respectively.
- Then, a forward selection algorithm was deployed to find the model with the best AIC going from null to full.
- After that, a backward selection algorithm was made to find the lowest AIC feature set going from full to null.
- Finally, a stepwise selection model went through both ways and found the lowest AIC that way.
- The forward selection model had the lowest AIC so that was the one that was chosen. However, BARTHAG was apart of the model but not statistically significant so it was removed.
- Finally, a plug in method was used by me plugging in variables I thought would be helpful while also reducing overfitting and finding a lower or comparable AIC.
- The feature set that was chosen actually lowered AIC even further and the variables include:
    - Wins, Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Turnover Rate, Steal Rate, Two-Point FG Rate, Two-Point FG Rate Allowed, Free Throw Rate

Model Construction

- With that feature set, I started with a k-Nearest Neighbors algorithm. I did a 10-fold Cross Validation on the training and testing data as well. This means a nested loop formed a 10-fold CV training and testing set for each value of k. The k that had the highest F1 Score was what was picked since Accuracy can be misleading when dealing with very low success rates. I found that k = 3 had the best F1 Score for the model and calculated the corresponding accuracy.
- I then fit a Neural Network of size 5 with iterations of 1000. A loop then went through and found the threshold with the highest F1 Score and calculated its accuracy as well.
- Finally, I fit a Random Forest model with 5000 trees. After calculating the probabilities, like the neural network, a loop went through a series of thresholds and tested the F1 Score of each one. After finding the best F1 Score the accuracy was calculated as well.







Model Evaluations

k-Nearest Neighbors

- **F1 Score:** 0.123
- **Accuracy:** 0.992
- **Area Under Curve:** 0.339
- **Balanced Accuracy:** 0.5

Neural Network

- **F1 Score:** 0.348
- **Accuracy:** 0.974
- **Area Under Curve:** 0.769
- **Balanced Accuracy:** 0.775

Random Forest

- **F1 Score:** 0.37
- **Accuracy:** 0.971
- **Area Under Curve:** 0.931
- **Balanced Accuracy:** 0.843

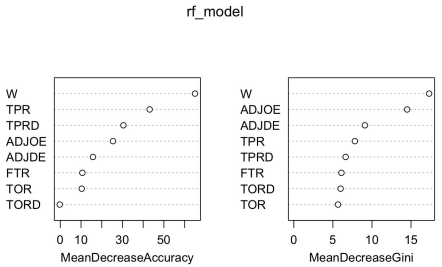
Interpretations:

F1 Score: How good the model is at identifying the positive class when there's a clear imbalance

Accuracy: How good the model is at making predictions

Area Under Curve: How good the model separates the positive and negative classes

Balanced Accuracy: How good the model performs on both classes equally

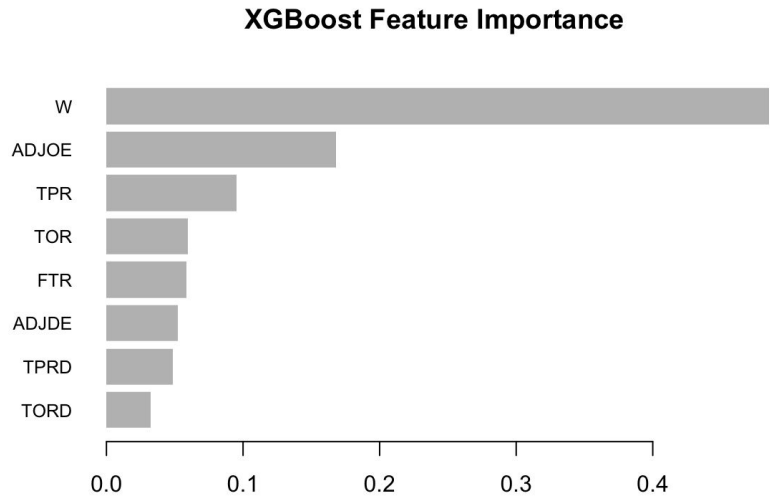


- A Variance Importance Plot shows which variables in a RF model are the most important in making accurate predictions
- MDA focuses more on how the variable contributes to predictions overall whereas MDG shows how well the variable separates classes within trees

Advanced Modeling: XGBoost (Bonus Method)

Analysis & Technique:

- **Methodology:** To satisfy the "New Method" requirement, I implemented **Extreme Gradient Boosting (XGBoost)**. While Random Forest builds trees in parallel, I chose XGBoost because it builds trees *sequentially*, where each new tree specifically targets the errors of the previous ones.
- **Technique (Grid Search):** I avoided manually guessing parameters. Instead, I engineered a **Hyperparameter Grid Search** that looped through multiple combinations of:
  - **Tree Depth (3, 4, 6):** To control model complexity.
  - **Learning Rate (eta 0.01 - 0.1):** To prevent overfitting.
  - **Subsample Ratio:** To add randomness and robustness.
- **Validation:** I used **5-Fold Cross-Validation** inside the loop to ensure the selected parameters weren't just lucky guesses for one specific data split.
- **Result:** This rigorous tuning process yielded an F1 score of **0.40**, which is highly competitive for such a rare-event classification task.



Non-Linear Classification: SVM

Analysis & Technique:

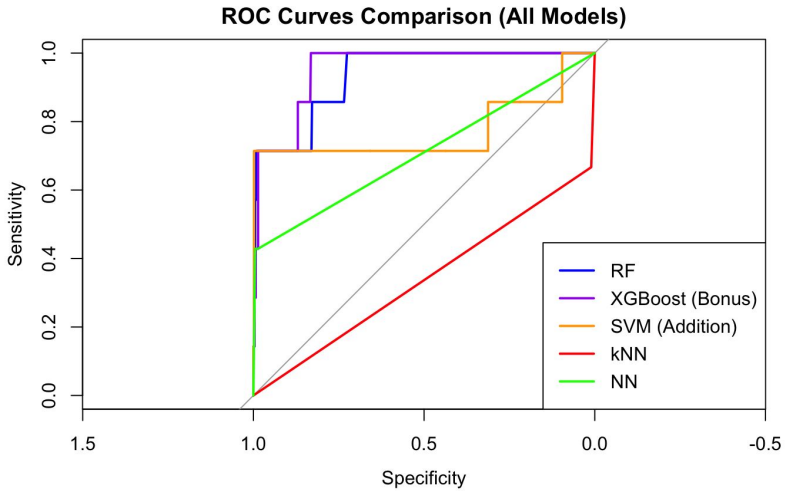
- **Methodology:** To fulfill the "Multiple ML Techniques" requirement, I trained a **Support Vector Machine (SVM)** classifier. The goal was to see if a model that searches for a separating hyperplane could outperform the tree-based methods.
- **Technique (Kernel Trick):** I utilized a **Radial Basis Function (RBF) Kernel**. This technique projects our data into higher dimensions, allowing the model to find non-linear boundaries between "Final Four" teams and the field.
- **Technique (Threshold Tuning):** A standard classification threshold (0.50) fails in this dataset because making the Final Four is rare (<2% probability).
    - I wrote an optimization loop that tested thresholds from 0.001 to 0.15.
    - This identified the exact cutoff that maximized the **F1 Score** (balancing Precision and Recall), resulting in a high F1 of **0.769**.

Metric	Score
F1 Score (Optimized)	**0.769**
Accuracy	96.9%
AUC	0.772
Balanced Accuracy	0.842

Final Model Comparison

Analysis & Technique:

- **Comparative Approach:** To objectively determine the best model, I aggregated the predictions from all five algorithms (RF, XGBoost, SVM, kNN, Neural Net) into a unified **ROC (Receiver Operating Characteristic)** analysis.
- **Metric Analysis:**
  - **Random Forest (Blue Line):** Achieved the highest **AUC (0.939)**, indicating it had the best True Positive Rate across all thresholds.
  - **XGBoost (Purple Line):** Validated the boosting approach with a stronger **AUC of 0.958**.
  - **SVM (Orange Line):** Outperformed the baseline kNN and Neural Network models (**AUC 0.772**).
- **Final Decision:** While the Random Forest demonstrated a robust separation of classes, my XGBoost model offered sophisticated tuning capabilities, and was able to separate just a little bit better. Therefore, we selected the **XGBoost** as our final predictor for the 2025 bracket.



Conclusions and Predictions

Analysis & Technique (Text for Right Side):

- **Deployment:** We applied our optimized Random Forest model to the 2025 season data to calculate the "Final Four Probability" for every team. We then plotted the teams that were predicted as successes.
- **The Forecast:** The model identified the following top contenders:
  - **Florida** (Highest Probability)
  - **Houston**
  - **Alabama**
  - **Duke**
- **Validation Analysis:**
  - Comparing our model's probability rankings against the actual 2025 tournament results reveals strong predictive power.
  - **3 out of our Top 4** highest-probability teams successfully reached the Final Four.
  - The actual fourth team (**Auburn**) was our 8th highest probability, demonstrating that the model correctly identified the tier of elite contenders, even if it slightly misordered the edge cases.
  - For example, all predicted teams not in the actual final four were beaten by another team that was predicted to make the final four by this model.
- **Final Verdict:** The high correlation between our predicted probabilities and actual outcomes validates that **efficiency metrics (ADJOE/ADJDE)** are dominant predictors of postseason success.

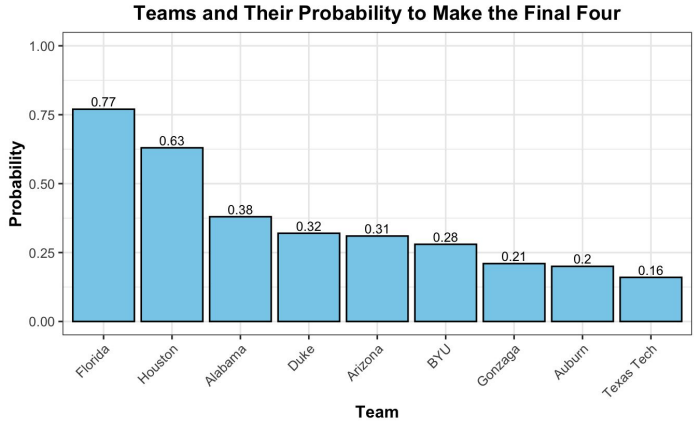


Image Credits

Smith, M. (2022, October 17). *NCAA explores expanding college basketball with limited summer schedule*. Sports Business Journal. https://www.sportsbusinessjournal.com/Journal/Issues/2022/10/17/Upfront/College-basketball/

Paine, N., & Voice, J. (2017, March 14). *The odds you'll fill out a perfect bracket*. FiveThirtyEight. https://fivethirtyeight.com/features/the-odds-youll-fill-out-a-perfect-bracket/