Problem Statement

How can an NBA team increase their win percentage by X% focusing their recruitment during the off season, training during the off season and regular season, and strategy during regular season and playoffs by focussing their efforts on a few key stats?

Approach

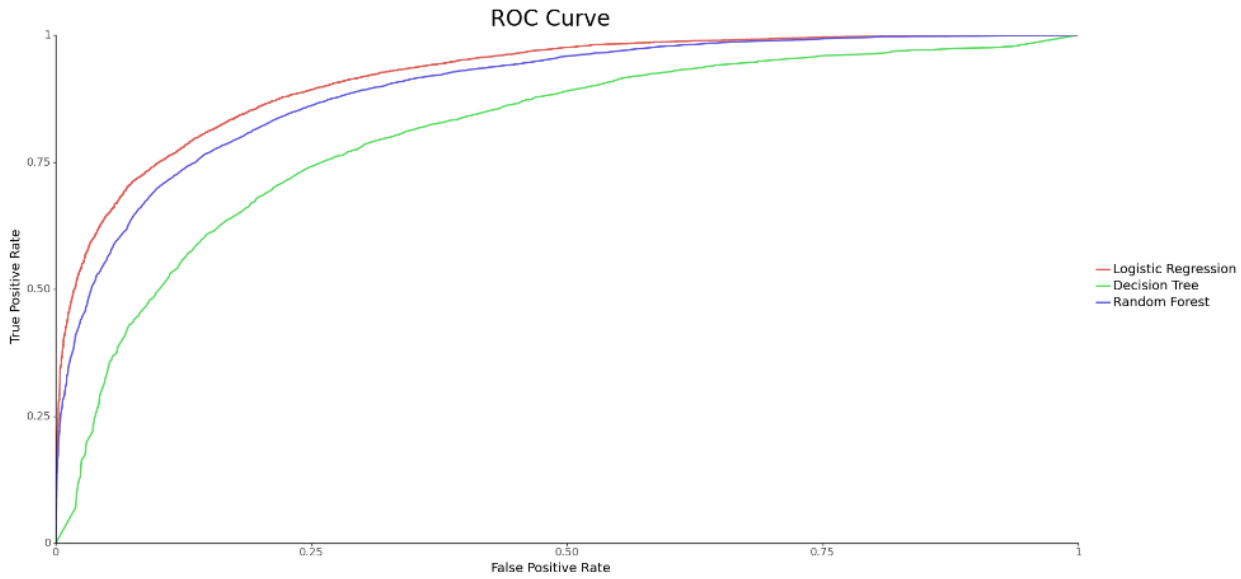
Using a dataset of Box Scores of NBA games from Kaggle, I used the stats of each team as inputs to predict the winner of the game. From there, I used the feature importance of the model to determine which stats were most important in predicting whether a team wins or loses. With this information, the General Manager of an NBA team can have a better idea of which players to sign and the coach will have a better idea of what to have the team focus on during practices.

The process in solving this problem was fairly standard. The dataset was acquired as a CSV file. Box scores are split by player so I had to do some aggregation in the data wrangling portion to get the team totals. From there it was the normal pre-processing steps such as creating dummy variables, standardizing inputs, and splitting the train and test set. Finally the model evaluation phase generated all of the insights and applications.

Model Results

Logistic Regression outperformed Random Forest and Decision Tree in every evaluation metric during both training and test phases. It also has the benefit of taking noticeably less time to run in comparison to the other models. The evaluation metrics made Logistic Regression an Easy choice

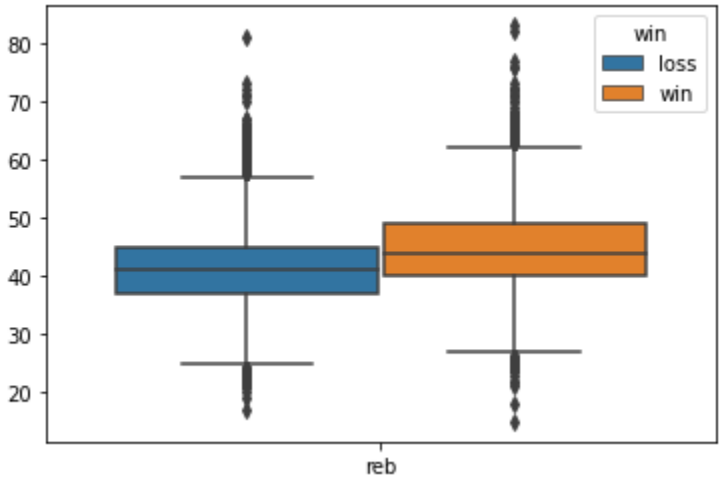
model	top_features	train_cv_mean	train_cv_std	accuracy	precision	recall	f1
Logistic Regression	[reb, fg3a, fg2a]	0.8333	0.0061	0.8315	0.8332	0.8306	0.8319
Random Forest	[fg2_pct, fg3_pct, reb]	0.8088	0.0058	0.8101	0.8129	0.8086	0.8107
Decision Tree	[fg2_pct, fg3_pct, reb]	0.7422	0.0065	0.7455	0.7429	0.7471	0.745



Recommendations

Recommendation 1: Rebounding

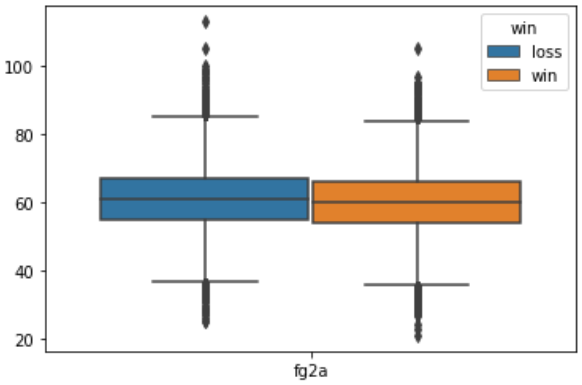
Rebounding was the number one most important feature of the model. You will oftentimes hear NBA analysts talk about “winning the glass” (“glass” = rebounding) and the model proved that to be true. Shots will be missed and it’s important your team gets possession of the ball when that happens as it gives your team more opportunities.

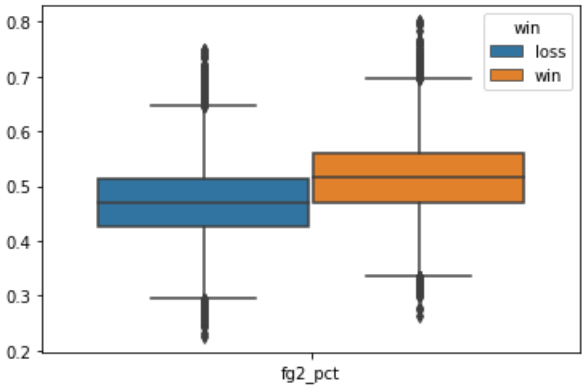


Recommendation 2: Shooting

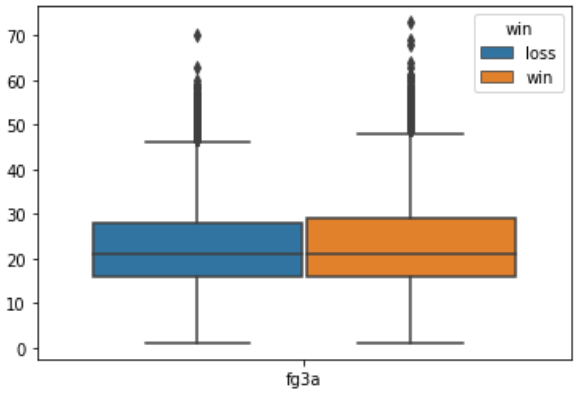
Shot percentage created the biggest divide between wins/losses
Shot attempts were most useful in predicting win/loss
Shot attempts are much more closely related to rebounds which is most important

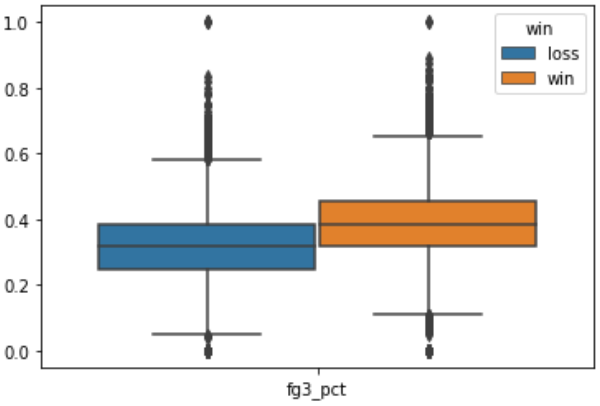
I found this result to be most interesting. The data showed the biggest divide between winning/losing teams was in 2-point and 3-point field goal percentage whereas field goal attempts were about even between winning and losing teams. However, when creating the model, it was field goal attempts that proved to have higher importance. Regardless it makes sense that shooting would be the next most important feature seeing as that is how you score.





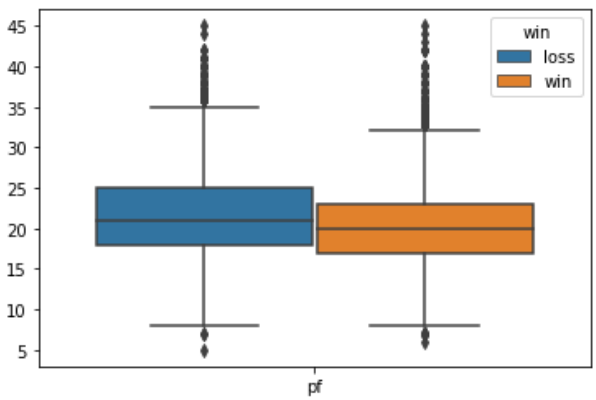
The same applies to 3-point shot attempts and percentage as it does for 2-point shot attempts and percentage

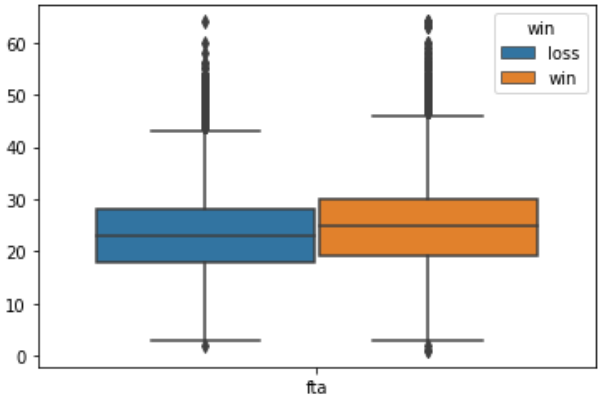




Recommendation 3: Fouls

At the bottom of the feature importance list was personal fouls and free-throw attempts. What this suggests is fouling or receiving free-throws from the other team fouling, are the least important factors in predicting wins/losses. This is likely due to free-throws being worth less (1 point instead of 2-3). Fouling is still an important part of basketball as it affects players' minutes, the team's rotation, and defensive aggressiveness. It can also be critical at the end of close games. However, it isn't as important as any of the other stats that were used in the model.





Further Research

I believe I made the best model with the data that was available, but real NBA teams have access to more advanced stats and I would have loved to build a model using those instead of the basic stuff. More defensive stats beyond just blocks and steals would have been useful and would have opened up the potential for the model to not be so reliant on offense. Advanced stats like player efficiency rating, and true shooting percentage would have been preferred over the basic shot attempts and percentage because just saying "take more shots and make more shots" is a rudimentary analysis of good basketball.