Springboard Data Science Career Track

Capstone 1: Predicting Future NBA Success

Satvik Reddy

Introduction

As an NBA Coach or General Manager, the decisions you need to make on player personnel can make or break a franchise's success for the short to long term future. The difference between making the playoffs or hitting the lottery can come down to one player, and the difference between a championship and a first round exit can also come down to one player. How do you decide who that player is? When you look at the All-NBA, All-Star, and All-Defense rosters year by year, it is clear that those players can lead franchises to success. However, every team knows who those players are. To truly gain an edge on opposing organizations, teams need to evaluate their young talent, and then decide who has the potential to reach the heights of becoming an All-Star, All-NBA, or All-Defense level player.

Using data from Basketball Reference and Kaggle, I compiled NBA rookie season statistics dating back to 1985. I used this data to create a machine learning model aimed at predicting whether a player would receive a "Future Honor" (defined as All-NBA, All-Star, or All-Defense) later in their career based off of their rookie season statistics.

Client Profile

My clients would primarily be NBA coaches and general managers. After a newly drafted player's rookie season, the clients could use my model to predict if a player would achieve a "Future Honor" based on how that player performed in the rookie season. They could then make more accurate decisions on which players to use more on the court, so that they can get closer to reaching that potential, and on which players to hold onto when they reach the end of their rookie contracts.

Data Information

The dataset was derived from a Kaggle dataset (*Nba Player Stats since 1950*) created by Omri Goldstein. It contained player statistics for every player dating back to 1950. The features include basic statistics from points, assists, and rebounds, to more advanced statistics like WS (win shares) and VORP (value over replacement player). Using RealGm.com to obtain All-Star, All-NBA, and All-Defense data, I created a "Future Honor" column in my rookie season statistics dataset, and used values of 1 to denote that a player achieved a "Future Honor" and 0 if they did not.

Data Wrangling

To get the dataset ready for analysis and machine learning, I went through a few steps. First, I took a subset of the original dataset that only included the rookie season statistics. I then took a further subset of the data that included the rookie seasons of players dating back to 1985, because this was the furthest back point at which data was available for whether or not a player achieved a "Future Honor" during their career.

There were some players that had missing statistics, so I used Basketball Reference to fill in the missing values. I deleted the following non-numeric features, such as “Tm”(Team), ”Team_abbreviation”, “College”, and “Country”. I deleted the numeric features of “draft_round” and “draft_number”. I did not want my model to be biased based on a player’s draft status, as players drafted early are already given the minutes and time to prove themselves, while my model treats all players as if they were drafted at the same time, so no one has a head start. I converted the categorical variable of “Pos” (Position) to dummy variables, because my exploratory data analysis convinced me it would be an important feature to include in my model. This will be explained in Figure 2 below. I used a RobustScaler to scale the numerical features for analysis.

Exploratory Data Analysis

I first wanted to investigate my “Future Honor” column to gain a better understanding of the prevalence of Future Honorees among all of the NBA rookies dating back to 1985.

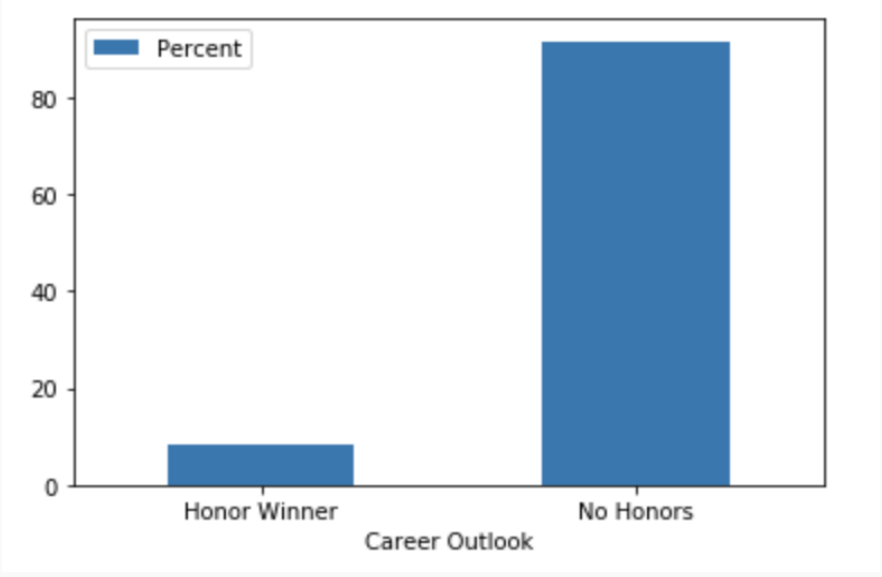

Figure 1: This figure displays the percentage of NBA rookies that have gone on to become a Future Honoree (“Honor Winner”), and the percentage that go on to win “No Honors” during their careers.

8.42% of rookies win a future honor, and this shows how rare it is, and thus how important an algorithm could be that can predict which rookies will be a part of that small percentage.

To investigate the relationship between position and honor won, I made the following chart.

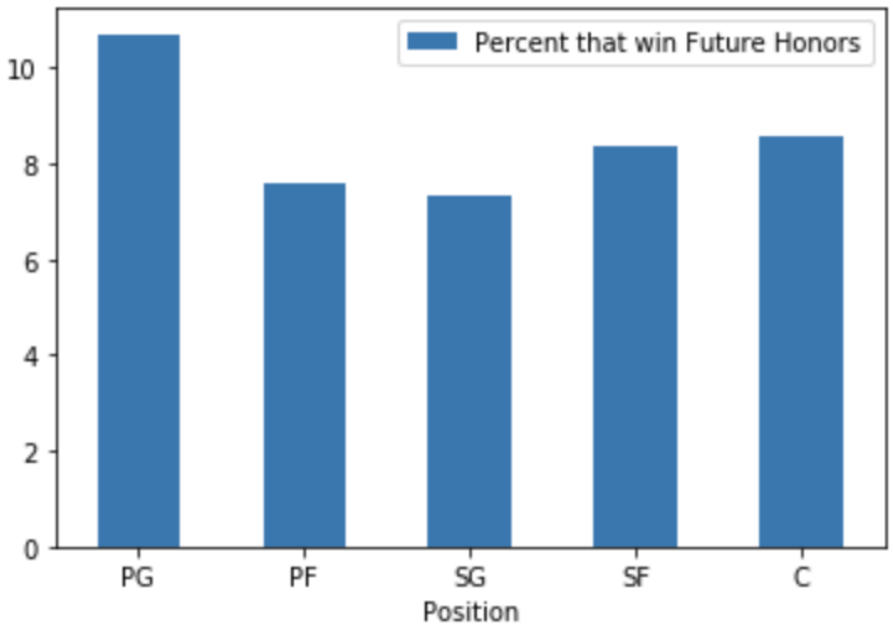

Figure 2: The percentage of players at each position that have gone on to win a Future Honor during their careers

Point guards and centers, with 10.71% and 8.55% of players winning an honor respectively, are more likely to win honors than the average across all positions (8.42%). Power forwards, shooting guards, and shooting forwards, with 7.58%, 7.31%, and 8.38% of players winning a honor respectively, are less likely than the average player to win a honor. Point Guards are the primary ball handlers, so they are involved in the highest number of offensive possessions for a team. Centers spend their time closest to the basket, and they have a general height advantage. This allows them to collect rebounds and make a higher percentage of their shots as they are closer to the rim. Overall, I was able to determine that usage, rebounds, and FG% could be important features to evaluate based off of this chart. I also determined that positionality

should be a categorical variable that I include in my model, because a player's position can have an influence on the likelihood of winning a future honor.

I took the Point Biserial Correlations between the features and the "Future Honor" column. This allowed me to see the effect of each rookie season statistic on the likelihood of winning a Future Honor during the players' careers.

Feature	Pointbiserial Correlation with Future Honor
Age	-0.23
Games	0.26
Games Started	0.38
Minutes Played	0.38
PER	0.16
TS%	0.1
DRB%	0.02
AST%	0.06
BLK%	0.09
USG%	0.04
OWS	-0.02

Feature	Pointbiserial Correlation with Future Honor
DWS	0.06
WS	0.4
WS/48	0.43
OBPM	0.15
DBPM	0.17
BPM	0.21
VORP	0.42
FG	0.4
FGA	0.39
FG%	0.09
3P	0.19

Feature	Pointbiserial Correlation with Future Honor
3PA	0.2
2P	0.03
2PA	0.4
2P%	0.39
eFG%	0.08
FT	0.08
FTA	0.41
FT%	0.05
ORB	0.36
DRB	0.4
TRB	0.39

Feature	Pointbiserial Correlation with Future Honor
AST	0.31
STL	0.36
BLK	0.35
TOV	0.39
PF	0.32
PTS	0.41

Figure 3: The table above shows all of the features that had significant Point Biserial Correlations (p-value<.05).

From this table, we can see that there are a mix of basic and advanced statistics that have the largest correlations. WS (Win Shares), WS/48 (Win Shares/48 minutes), VORP (Value Over Replacement Player) are the advanced statistics with the highest correlations. Win Shares measure a player's contribution to their teams' wins over the course of the rookie season, including both offensive and defensive contributions, and win shares per 48 minutes measures the contribution toward a team's wins per game. VORP is a measure of a player's value compared to an average quality player. The statistics of FG (Field Goals) and PTS (Points) have the highest correlations among the basic statistics.

From the correlations, I honed in on Win Shares, because this is among the few "advanced statistic" features that had strong correlations with winning a future honor. There is a lot of debate throughout the NBA and sports community about the importance of advanced statistics, so I wanted to make sure to further investigate any that I would include in my model. I made a pairplot with this feature to investigate the relationships between the variable and achieving a future honor.

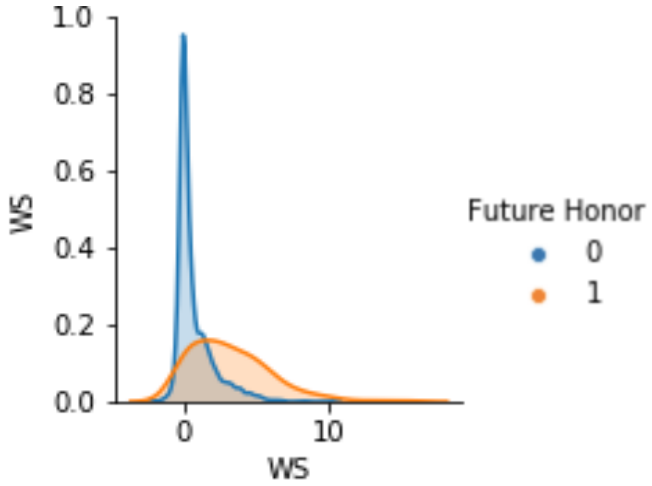

Figure 4: This figure shows the pairplot between WS(Win Shares), and the Future Honor feature.

This allowed me to see the distribution of the feature for players that did (Orange) or did not (Blue) win a Future Honor. Win shares is a statistic that measures an individual player's contribution toward a win in totatility. We can see how the bell curve for win shares of future honorees is shorter, wider, and right shifted compared to the bell curve for both features of rookie players who don't win a future honor. This tells me that this feature will be very important to investigate.

I made a correlation matrix with all of my numerical features to investigate the relationships between the features and their effects on winning a future honor.

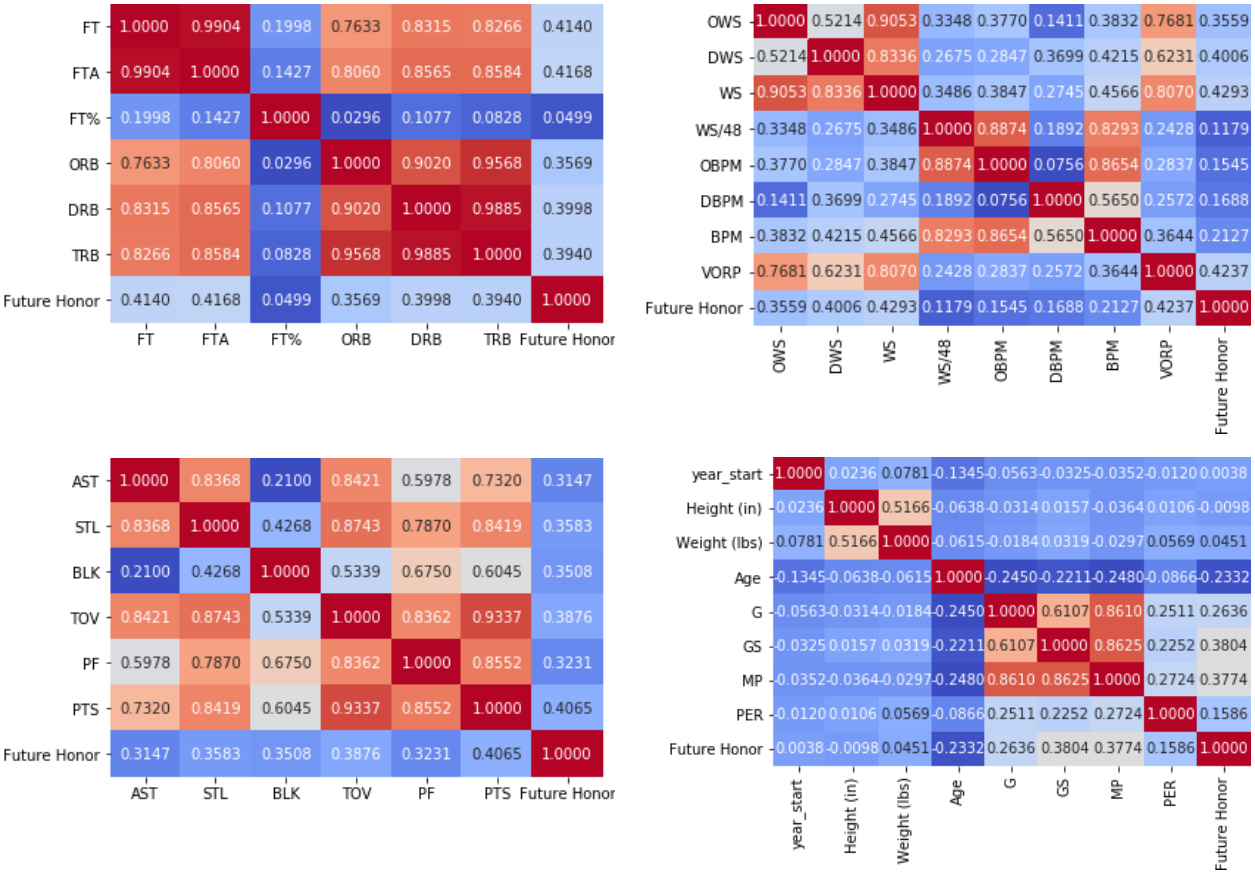

Figure 5: Correlation matrices of player statistics

I used Figure 5 to identify features with collinearity, so that I could remove what was unnecessary. Collinear features have high correlation values with each other. For example, the correlation between OWS (Offensive Win Shares) and WS (Win Shares) is .9053, and the correlation between DWS (Defensive Win Shares) and WS is .8336. This logically makes sense, because WS is calculated as a combination of OWS and DWS, but it is by using this chart that I was able to identify the features for which this collinearity was the case, so that I could remove statistics like OWS and DWS from the data before training.

Machine Learning

Feature Importance and Selection

In order to determine feature importance, we need to be able to trust the coefficients obtained through linear regression analysis, which can be skewed by multicollinearity. To

investigate multicollinearity, I obtained VIF scores for each predictor. VIF (Variance Inflation Factors) is a quantification of the extent of a predictor's correlation with other predictors. Thus, a higher value implies multicollinearity. I used a value of 10 as the VIF threshold, so I got rid of features that had VIF scores above 10. After cutting down the features, I ran logistic regression with the remaining features and the "Future Honor" column. I then only kept features that had coefficients with p-values under .05, so that I could conclude that they are statistically significant. I used a Robust Scaler to scale all my numeric features before running the logistic regression. Next, I calculated the Log-Odds of the scaled features to directly look at the change in likelihood of winning an honor, based on a one unit increase in a feature.

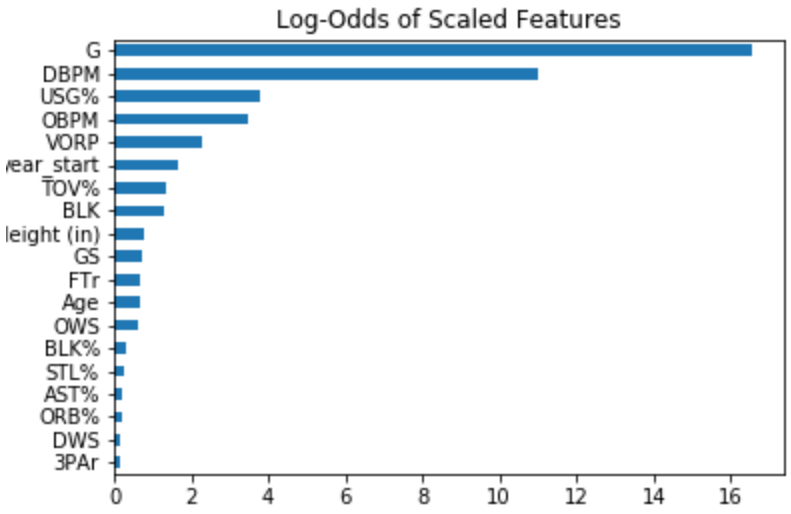

Figure 6: Log-Odds of the scaled features

Log-Odds is calculated by simply taking the logarithm of the odds of winning an honor. The odds are calculated by dividing the probability of winning an honor by the probability of failure. The Log-Odds are calculated because it creates a normal distribution of the data around 0. Here, it appears that games played, defensive box plus minus, offensive box plus minus, and usage percentage, are the features for which a one unit increase in those statistics, will lead to the highest increase in likelihood for winning a future honor. Defensive box plus minus and games played were explained in figure six. Usage percentage is the percentage of a team's possessions

that a player uses, and offensive box plus minus is a measure of a player's contribution to a team's total offensive success.

To identify possible important features, I used the StatsModel Logit function to obtain coefficients for each feature in terms of their influence on winning a future honor. Important features had p-values < .05. I also used the charts from Figures 6 and 7 to determine the important features.

I also obtained possible feature importances using a random forest for comparison. I used the scikit-learn package to obtain feature importance scores for each feature. These scores represented the level of information gained from those features. The scikit-learn package determines the information gain from each feature by determining the decrease in impurity of each decision tree due to the feature, and then averages this impurity decrease to come up with a score.

My next step was to determine what type of model would be best for predicting future honors. I chose to look at Logistic Regression and Random Forest as potential options. lI used GridSearchCV for both classifiers for hyperparameter tuning, which produced the results below:

Classifier	ROC-AUC Score	Hyper-Parameters
Random Forest	0.887	n-estimators=200, max_features=auto, max_depth = 70
Logistic Regression	0.824	Cs = 10, penalty = 12

Figure 7: Grid search results for model hyperparameter tuning

As you can see, Random Forest had a higher ROC-AUC score, and therefore I decided to use it as my predictive model going forward into thresholding.

Thresholding The Model

The Random Forest Classifier predicted probability values for each player for their likelihood of winning a "Future Honor". There were 116 distinct probabilities in the range between (0,1). However, I need my model to make binary classifications, either it predicts a player wins an honor (1) or doesn't (0). Thus, I had to pick a value, or a threshold, that I could then classify all players with probabilities over that threshold to be Future Honorees, and vice

versa for those players with probabilities less than that threshold. This threshold varies based on the business situation. I plotted precision and recall curves across the entire distribution of thresholds, and picked the threshold that had the best balance between precision and recall for the specific business case. For example, in business case 2, where recall was more important than precision (explained below), I made the following thresholding curve and picked the threshold value of 0.18.

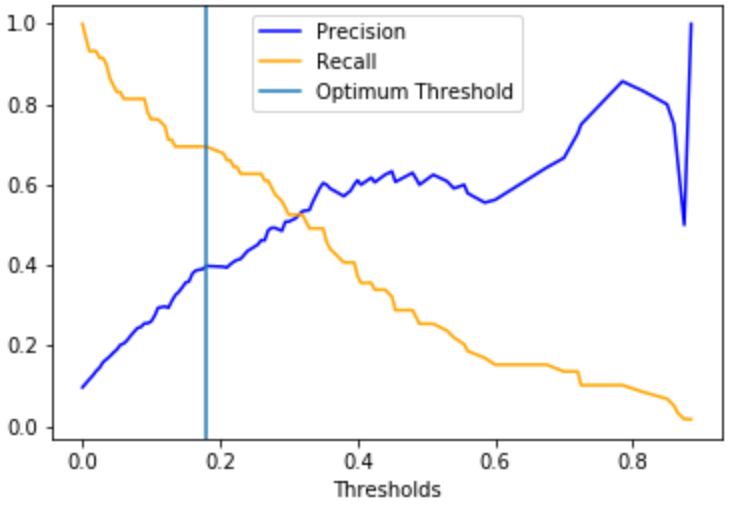

Figure 8: Precision-Recall curve as a function of varying thresholds for the Random Forest Classifier.

The vertical line denotes the optimal threshold for Business Case 2.

Business Case 1: Coaching Decisions (Short Term Sacrifices for Player Growth)

Suppose you are the coach of an NBA team at the start of a new season. You must decide who will be the starting five for your team in the coming season. This decision holds vital importance, as these five players will see the court for the most minutes, and play against the opposition's best players. Thus, you must make sure you are putting the players with the highest potential in that starting five. Last season, the team had two rookies who both played off of the

bench. They both had solid seasons, but now you must pick between the ex-rookies to decide who will be in the starting five and will play the most minutes for this upcoming season.

For this scenario, you would use my classification model to predict, based on their players' rookie year statistics, which ones have the potential to achieve "Future Honors", so that you can insert them into the starting five to give them the most experience and development possible, so that they can reach that potential. Here, precision is extremely important. Precision here is the proportion of true future honorees out of the future honorees predicted by the model. This must be high, because the players that you pick for the starting five are the ones that will largely determine your short term success on the court, and they take away development minutes from other young players, so it is vital that the players you pick truly have that exciting potential.

For this scenario I chose a threshold of 0.78, resulting in a precision of 0.83 and a recall of 0.15. This means that if the model predicts a player to receive a future honor, they will achieve this 83% of the time. While the recall is 15%, meaning that 85% of future honorees will be missed, it is more important that you choose a player for the starting five that will be a future honoree, than it is that you choose all of the players that have that potential, because you will have to also include veteran players in the starting five that have already reached that potential and will win you games.

A visual representation of this situation is below.

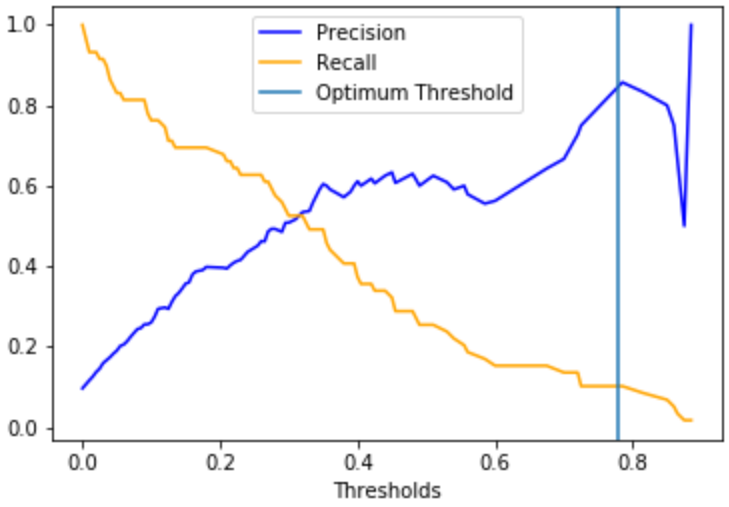

Figure 9: Precision-Recall curve as a function of varying thresholds for the Random Forest Classifier.

The vertical line denotes the optimal threshold for Business Case 1.

Business Case 2: General Manager Decisions

Suppose you are the general manager of an NBA team. Rookie contracts are guaranteed for the first two years, and you have the option of extending the contracts for the third and fourth years. You have a few players entering their fourth year with the team, and you have a few players entering their third year. You must decide whose contracts are worth extending among this group. You want to make sure you are holding onto the players that have the potential of becoming a future honoree, because it will elevate your own team, and it will prevent other teams from receiving the benefits from having those players on their teams. Recall is vital because as a general manager, you want to make sure you are keeping the highest proportion of your future honorees as possible, and precision is less important because keeping some players on the team that won't ascend to that level will hurt the team less as long as you are keeping all of the ones that will ascend. Still, there is a balance here, because the salary cap punishes teams who spend money on players that are not worth it, so precision can not be ignored.

The best threshold here would be 0.18, resulting in a precision of 0.40 and a recall of 0.66. This means that you should select players for which the model predicts a probability >= 0.18.

A visual representation of this situation is below.

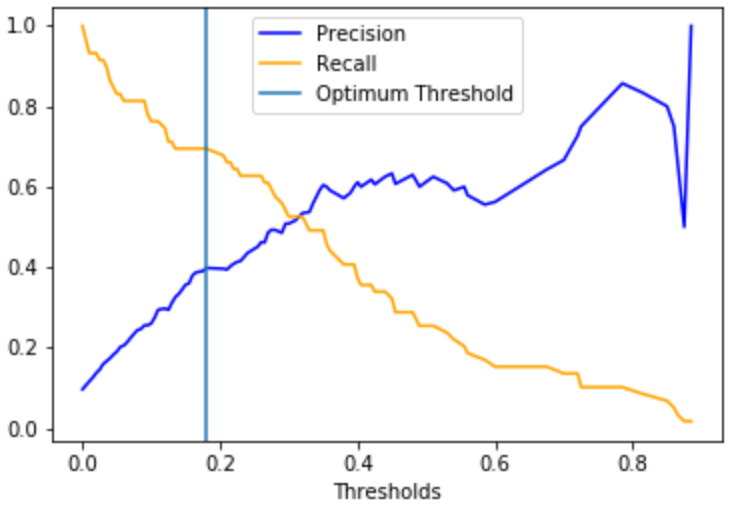

Figure 10: Precision-Recall curve as a function of varying thresholds for the Random Forest Classifier.

The vertical line denotes the optimal threshold for Business Case 2.

Conclusion

Given the small percentage of rookies that will eventually ascend to future honoree status, a team can gain a large competitive advantage by investing its time and resources in the right ones. My data exploration made it clear that both advanced statistics and basic statistics are crucial to predicting future honorees. It is high time that statistics such as WS, VORP, and BPM are evaluated by not only NBA stat junkies, but by coaches and general managers as well. At the same time, many of the important statistics were revealed to be basic ones such as PTS, STL, AST, BLK, TOV, TRB. This shows the importance of coaches giving their rookies enough time on the court to develop. My data exploration also showed that Point Guards and Centers are more likely than other positions to get Future Honors, and this is something for users of my model to keep in mind. If you are a coach deciding which young player to start, you should use my model to emphasize precision as described above. If you are a general manager deciding which player to offer an extension to, you should use my model to emphasize recall as described above.

A weakness of my model is that it does not account for external variables that could still influence a player's success. For example, a player can have very few minutes during a rookie season due to their team already having more talented players. In contrast, players on bad teams might have worse efficiency statistics because they are forced to take a lot of bad shots without a lot of talent around them. Perhaps creating new features related to team success could help account for this, or otherwise normalizing statistics to account for this, this simply would require more time.

Some other next steps for my model could be to reevaluate its success with each passing season. The current and previous year's rookie future stars will take some time to achieve distinctions, but by using my model on their rookie season statistics right now, you can find out who those future stars are right now. Another possible next step would be to look at a player's first 2 or 3 seasons rather than simply their rookie year, as many players make big leaps in their abilities between their 1st and 2nd years. It could also be interesting to make models strictly looking at a player's potential of making All-Defense, All-NBA, or All-Star, because each honor favors the skillsets of different kinds of players.