DS3000: Navigating the Numbers of Future NBA Stars

Peter So, Andrew Klaey, Sahil Kudva, Toby Chan

Abstract

The goal of this project is to compare the effectiveness of using the RAPTOR metric versus the Box Score metric in determining an NBA player's value to their team for a specific season. The data was derived from popular files on Kaggle containing the RAPTOR and Box Scores of players in the NBA (between the years 2014-2022). 2023 RAPTOR data was also tested to determine how accurate our models are for the most recent NBA year. A total of 6 models will be used and assessed to determine which model (and dataset) is the most accurate in predicting the number of (All-NBA) votes a player gets at the end of the season, and whether they make the all-NBA team.

Introduction

As fans of the NBA and avid basketball watchers/enjoyers, we wanted to center our project around the analytics of the sport. Through our discussions and searches for open source NBA data, we discovered the RAPTOR dataset. RAPTOR is a newly developed, much more holistic metric used to evaluate an NBA player's impact on the court. The metric takes into account more context and advanced statistics in its calculations, returning offensive and defensive rating scores for each player on different ends of the court. This is a contrast to traditional box score data, which is a table that solely keeps track of players' counting statistical averages (such as number of points, assists, rebounds, steals and blocks gotten) every season. Both metrics have their benefits but also their limitations. While the Box Score data lets us know about the actual/visual impact an NBA player has on the court, RAPTOR may reveal more of the intangible actions (such as hustle play, or defensive intensity) that casual fans may not be aware of.

As the RAPTOR metric is still relatively new, it is currently unclear how accurate of a metric it is, and if it can accurately predict trends and patterns of NBA players' performances in the modern day. Box Score, on the other hand, has been a long-standing metric (being tracked since the beginning of the NBA in the 1940s), and has always been the go-to metric for evaluating players' impact. Therefore, we thought that it would be very interesting to use our project to compare these 2 types of metrics, to see which one is actually more insightful into watching and learning about NBA players, and evaluating impact and (individual award) success on their teams.

To determine if RAPTOR or Box Score statistics is more indicative of measuring players' impacts, we will be developing models (using both datasets) to predict the number of All-NBA votes each player gets at the end of each season. The All-NBA Team is an annual honor bestowed upon the top players in the NBA. It recognizes the top players at each position (guards, forwards, and centers) based on their performance throughout the regular season. The All-NBA Team consists of three separate teams: First Team, Second Team, and Third Team. These teams are determined through a voting process conducted by a panel of sports journalists and broadcasters. There are 100 voters and each player receives points based on their placement in the voting. A first-team vote earns a player 5 points, a second-team vote earns 3 points, and a third-team vote earns 1 point. Thus, the voting range is between 0 to 500 points, as the highest number of votes a player can achieve is 500. The players with the highest point totals at each position are selected for the respective All-NBA Teams. Each member of the panel selects two guards, two forwards, and one center for each team, resulting in a total of 15 players being selected across the three teams. However, our models don't take into account the different position groups and instead base the votes on the stats included in the box score and RAPTOR models. By testing how accurate the datasets (with different models) are, at predicting the number of all-NBA votes each player gets at the end of each season, we will discover which dataset evaluates players' impact and effect on their teams better.

Related Work

While there isn't any directly related work in terms of examples of analysis using Python, FiveThirtyEight has created an article for RAPTOR which includes visualizations and explanations of their metrics.

Methodology

To begin the project, we gathered data from two popular Kaggle datasets and used Python and Pandas to analyze the data. We then looked at the basic information about the dataframes (such as its descriptive statistics and data types), and performed preliminary EDA. We did this by checking and converting for suitable data types, and dealing with missing or invalid values. Afterwards, we performed a visual analysis to see the distributions of values and verify that the data we were using was appropriate for our goal. The visualizations (histogram) showed that there were many outliers in the dataset and, through further analysis, we realized that the outliers were due to many players playing very few minutes or games in the season, which was ruining the integrity of the data. To account for this, we decided to filter out players that played less than the mean (of games and minutes) on both the RAPTOR and Box Scores datasets. In the end, we were left with more useful datasets, which only consisted of NBA players who played notable or significant minutes throughout the season.

Once the cleaning was completed, we began looking at correlations between variables to select which ones to choose for our models. We graphed var correlation to the total votes column. Based on this analysis, we decided to use the features 'pts', 'usg_pct', 'ast', 'ast_pct', 'reb', 'dreb_pct', 'net_rating', 'ts_pct', and 'dreb_pct' for the box scores. For RAPTOR we used 'season', 'poss', 'mp', 'raptor_box_offense', 'raptor_box_defense', 'raptor_onoff_offense', 'raptor_onoff_defense', 'raptor_offense', 'raptor_defense', 'war_total', 'war_reg_season', 'war_playoffs', 'predator_offense', 'predator_defense', and 'pace_impact' as the selected features.

Our final decision in model selection was on the basis of the answer to our underlying question containing continuous values for RAPTOR and box scores. We used three different models to try and dissect which dataset had a stronger relationship to discern All-NBA votes for individual players.

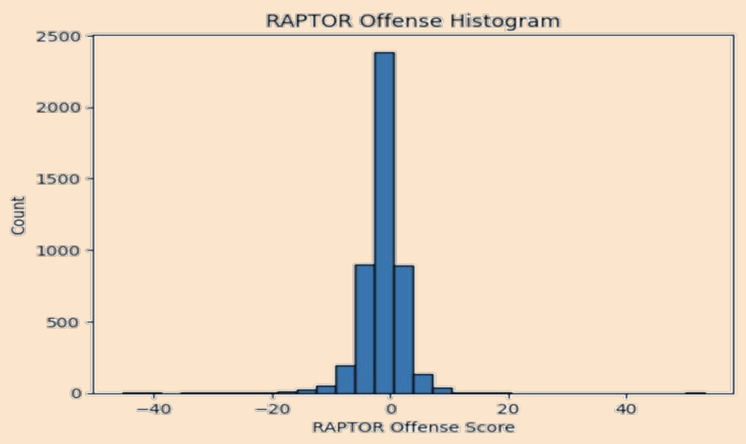
We started with a Polynomial Linear Regression model. Polynomial regression allows for more flexibility in capturing nonlinear relationships between variables. In the context of comparing impact metrics like RAPTOR and box score stats with All-NBA votes, the relationship between these metrics and player performance may not always be linear. Polynomial regression can capture these nonlinearities better than simple linear regression. It can also account for interactions between predictor variables. In the case of comparing RAPTOR and box score stats, there may be interactions between certain metrics that affect player impact differently.

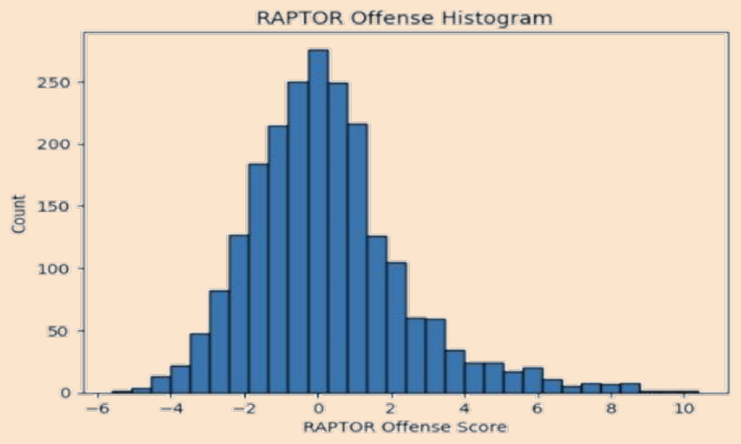
We also decided to use Support Vector Regression models. SVR is less sensitive to outliers compared to some other regression techniques. In the context of NBA player analysis, outliers in player performance or impact metrics may exist due to various reasons such as impactful injuries (torn achilles or ACL), exceptional performance, or changes in team dynamics. SVR can help mitigate the influence of these outliers on the model's predictions.

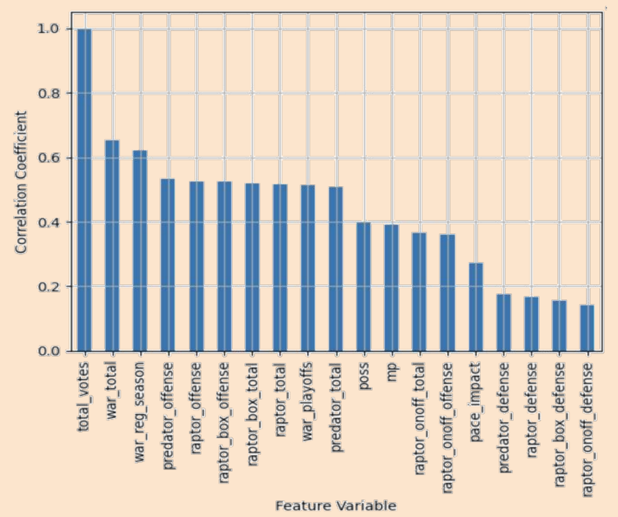
The third model that we chose is the Random Forest Regression model. This model generally has high accuracy and is able to handle large dimensional datasets while still being resistant to overfitting in comparison to a decision tree that will split data using splits that don't always make the most sense. Each tree is trained on a random subset of the data and features, which helps in capturing different aspects of the data without relying too heavily on any single predictor or subset of predictors. To train the random forest regressor model, we gave it the features discussed above and the total votes as the target.

Data was partitioned into a reproducible training and test set, where the training set is 70% of the data and the test set is 30% of the data. Once partitioned, we set up a grid search utilizing GridSearchCV to find the best hyper-parameters for each model. It is important to note that we intentionally chose not to normalize the data since the data is already standardized in different ways. For example RAPTOR is already standardized and box score has per-game as well as seasonal statistics.

For evaluating the performance of the model, we used the mean squared error (MSE) value and R-Squared value of the models so that we can get a numerical representation of the errors and correlations. The model with the best performance was the random forest regressor for the RAPTOR scores. This model provided both the highest R-Squared score and the lowest MSE value for the train and test data compared to the other two models.







Results and Evaluation

After training Polynomial Regression, Random Forest Regression, and Support Vector Regression models on the RAPTOR data and Box Score data, we can compare the models to the counterparts to determine which data set is better at predicting All-NBA votes.

For predicting All-NBA votes using RAPTOR data, the best polynomial regression model (which used a degree of 1) resulted in a mean-square error of 3483.5 and a R2 value of 0.577. However, some of the predictions contained negative values which is not possible. In the actual voting format, there is no way to give a player negative votes. For this reason, we decided to convert all negative votes to 0. This gave a slightly better MSE of 3207.7 and a R2 value of 0.610. The best Random Forest Regression model (which had a max depth of 5 and 200 estimators), resulted in a mean-square error of 2361.8 and a R2 value of 0.713. This model didn't predict any negative votes. The best SVR model (which had a C value of 1, degree of 3, and polynomial kernel), resulted in a mean-square error of 4869.8 and a R2 value of 0.408. This model also predicted negative votes. So after converting negative values to 0, the model had a mean-square error of 4864.6 and a R2 value of 0.409. Overall, when using the RAPTOR dataset, the Random Forest Regression model was the best at predicting All-NBA votes.

For predicting the All-NBA votes using the box score data, the best polynomial regression model (which used a degree of 3), resulted in a mean-square error of 3067.9 and a R2 value of 0.640. Like the model for RAPTOR, some of the predictions contained negative values, which is not possible. After converting all negative votes to 0, the model had a better MSE of 2973.7 and a R2 value of 0.651. The best Random Forest Regression model (which had a max depth of 5 and 50 estimators), resulted in a mean-square error of 2999.1 and a R2 value of 0.648. This model didn't predict any negative votes. The best SVR model (which had a C value of 1, degree of 4, and polynomial kernel), resulted in a mean-square error of 4100.4 and a R2 value of 0.518. Using Box Score as the datasource, the Random Forest Regression model was also the best in predicting All-NBA votes.

After running the 6 models, we can see that the box score data outperformed the RAPTOR data for the polynomial and support vector regression models (the Box Score models scoring an MSE of 2973.7 and 4100.4 compared to RAPTOR's MSE of 3207.7 and 4864.6 respectively), while RAPTOR outperformed in the random forest regression models (RAPTOR had a lower MSE of 2361.8 compared to Box Score's 2999.1). This indicates that both datasets are insightful in evaluating player impact, despite the difference in what their numbers indicate. Ultimately, the best model for predicting All-NBA votes was the Random Forest Regression using RAPTOR data.

To further test our best model, we decided to predict the All-NBA votes (and the top 15 players who make the All-NBA team) for the 2023 season to ultimately score the accuracy of the model. Our model scored a 95.1% accuracy in predicting whether or not players made the All-NBA team.

Impacts

Overall, we believe that the results of our project mostly impacts fans and audiences of the NBA. The metrics and scores generated by RAPTOR give us insights into an NBA player's offensive and defensive impact for their team, which might not always be directly visible or evident to us fan, as we have a tendency to only focus on wherever the ball is, and missing out on the off-ball actions or intensity/hustle plays that players commit to elsewhere on the court. Rather, it is much easier (and common) for us fans to visualize and track Box Score data, as it solely tells us about the counting statistics that players get, and is something we can see and count ourselves. However, the results of our project tell us that RAPTOR is actually a more valuable metric in evaluating an NBA player's impact on their team, suggesting that our current ways of watching and looking at basketball might be flawed, and not the most effective.

Therefore, the project impacts all fans and casual watchers of the NBA, as it sends a message to us that we as audiences should reconsider the way in which we think and watch basketball to evaluate and judge the quality of players. We should not solely look at the numbers players put up in games or at the end of the season (the act of "box score watching"), but instead look beyond that and into the intangibles, such as whether a player passes the eye-test, their abilities to read schemes and coverages, or just the overall effort and skill they play with to impact their team's winning as a whole. As this is not very straightforward, and requires the casual fan to be more educated on the game and its tactics overall, this may be a bit difficult to implement and take fans some time to adjust towards. However, in the long run, this will be beneficial to the basketball community, as we will learn to truly appreciate the game of basketball, and give unrecognized impactful and effective players their rightful flowers.

Conclusion

Through this project, we explored various NBA player statistics, combining our passion for the sport of basketball with our knowledge of data science. Overall, we were able to achieve our goal of determining which NBA dataset (RAPTOR metrics versus traditional box scores) was best to use for evaluating player impact, along with the best model to pair with it to make future predictions about the stars of the league. Ultimately, our analysis showed that the RAPTOR metric was a better evaluator of player performance and predictor of player success in end of season All-NBA voting. Specifically, the Random Forest Regression model (paired with the RAPTOR data) was most successful in handling predictions (with the lowest MSE), and yielded a 95.1% success rate in further testing.

While our model was 95.1% accurate in the test we did involving current (2023) data, we believe that there is definitely still room for improvement for the model. As mentioned in the introduction, each voter can only select a specific amount of players for each position group (for example, only two guards, two forwards, and one center for the 1st 2nd and 3rd teams). This means that while some players might get more votes than others overall, they still might not be considered for an all-NBA team, as they had less votes than others within their position group. However, as this would be a bit too complex, and the datasets we worked with did not have a section for position groups, our model does not account for this, and considers the top 15 scoring vote getters to be on all-NBA teams. This could potentially explain why we were only 95% accurate, as some players who deserved/had higher votes eventually got snubbed from the teams, whilst others who had less votes (but were in a less competitive position group) sneaked through. Therefore, if we were to develop our model in the future, one thing we could look at adjusting for would be position groups, to make our model perform much closer and realistically to how it is done in the NBA by voters.

References

1. *All-NBA Voting*, pr.nba.com/wp-content/uploads/sites/46/2023/05/2022-23-Kia-All-NBA-Team-Voting-Results.pdf. Accessed 13 Apr. 2024.
2. Ryanabest. "The Best NBA Players, According to Raptor." *FiveThirtyEight*, 14 June 2023, projects.fivethirtyeight.com/nba-player-ratings/.
3. Viswa. "Support Vector Regression: Unleashing the Power of Non-Linear Predictive Modeling." *Medium*, Medium, 28 July 2023, medium.com/@vk.viswa/support-vector-regression-unleashing-the-power-of-non-linear-predictive-modeling-d4495836884#:~:text=Robustness%20to%20Outliers%3A%20SVR%20is,more%20resilient%20to%20noisy%20data.
4. R, Sruthi E. "Understand Random Forest Algorithms with Examples (Updated 2024)." *Analytics Vidhya*, 3 Jan. 2024, www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=One%20of%20the%20most%20important,for%20classification%20and%20regression%20tasks.