# NBA MVP Prediction Using Classification Methods

Veeti Tuppurainen, Tise Mobuse, Nicholas Selesi, Ayomide Gbadamosi

School of Information, The University of Texas at Austin

I-310D: Introduction to Human-Centered Data Science

Professor: Abhijit Mishra

May 6, 2024

**Introduction:**

Basketball is one of the most popular sports worldwide, captivating millions of fans with its fast-paced action and thrilling moments on the court. The NBA stands as the premier league in professional basketball, showcasing talent from all around the world. The Most Valuable Player (MVP) award is the pinnacle of achievement in the NBA, symbolizing excellence, leadership, and contribution to a team's success.

Throughout the extensive history of the NBA, the criteria for selecting the MVP has been the subject of ongoing debate among analysts, fans, and players alike. While individual statistics such as points per game, rebounds and assists, and defensive metrics play a significant role in MVP consideration, we decided to find out the best indicators. Our research aims to delve into the data of the NBA player statistics and honors to discern the most influential factors in determining the MVP candidacy. By analyzing a comprehensive dataset spanning decades of basketball history, we seek to offer valuable insights that not only address the MVP awards but also shed light on broader aspects of NBA performance and player evaluation.

Through examination and statistical analysis, we attempted to find out the selection of the NBA's Most Valuable Player, providing a deeper understanding of what defines success.

**Data Description:**

The comprehensive biostatistics and achievements of NBA and ABA players are included in the dataset we used, which we obtained from Kaggle. The data has a CC0: Public Domain license which allows for copying, modifying, and distributing the work. His dataset provides a rare chance to investigate many facets of player success, career longevity, and noteworthy accomplishments throughout the history of basketball. We performed extensive data transformation, data cleansing, and feature engineering to make sure the dataset was suitable for classification before analyzing it. The dataset originally consisted of 39 columns, which included, Player, From, To, Years, Pos, Ht, Height, Wt, G, PTS, TRB, AST, FG%, FG3%, FT%, eFG%, PER, WS, All Star ,All NBA, All ABA, All Rookie, All Defensive, BLK Champ, STL, Champ, TRB Champ, AST Champ, Scoring Champ, Most Improved, Sixth Man, DPOY, ROY, AS MVP, CF MVP, Finals MVP, MVP, Championships, NBA 75 Team, and ABA All-Time Team.

**Data Preprocessing**

To preprocess our data, we first utilized a process called “binary encoding”. In this process, we represent parts of our data using 0 and 1, in contrast to values like no and yes. This allows us to better interpret the specific data attribute for scalability, efficiency, and most importantly, simplicity. We did this for the “MVPTF” attribute of our dataset, as it was the basis of our project objective. The MVPTF variable consisted of two classes with class 0 having no MVP status and class 1 having MVP status. Subsequently, we checked for null values within the dataset but found none. Finally, we checked for duplicates, dropping duplicates within the “Player” column.

**Methodology:**

Although we initially thought our project objective required regression analysis, after the presentation, we soon realized that we were solving a classification problem. Classification requires predicting categorical outcomes, in contrast to regression modeling, which predicts continuous numerical outcomes and understands relationships between attributes. As a result, Machine Learning Classifier is our main technique for determining the correlations between playstyle data, physical characteristics, and the chance of winning the NBA MVP award. MLP classifier was our method of choice since it can measure these associations and offer players, teams, and analysts useful insights. We do recognize the drawbacks of this strategy, though, such as the assumptions that come with MLP Classifiers and the possibility that confounding variables could skew our findings.

**Results:**

Our model consists of 3 outputs, one for logistic regression, random forest, and k nearest neighbor. Before discussing the results, we must discuss the basis of each classifier to understand their meaning. Logistic regression is a classification model used to predict the possibility of an observation regarding a dependent variable and one or more independent variables. For our data, we tested a possible observation using the MVP award as our dependent variable, and our player,

position, and years active as our independent variables. In doing this, we found out that of all instances predicted as positive, only 71% actually is. The recall for our logistic regression is 0.62, meaning 62% of positives are correctly detected by the model. The next classification model we used was random forest. Random forest operates by constructing “decision trees” in training, outputting the average regression of the individual trees. Using the same variables to test, we found that random forest classification yielded an 86% precision, deeming it more effective than that of logistic regression. Its recall was also higher, observed to be 0.75, or 75%. The last classification model we used was K Nearest Neighbor. K-nearest neighbor is the simplest algorithm from the three, as it stores our training data and forms predictions by grouping. This model is usually used for pattern recognition, anomaly detection, and image recognition. After implementing our data, we found the precision of the model to be 0.75, or 75%. An interesting observation would be the recall, as it performed at an extremely low 0.38, or 38%. Several intriguing themes and insights have emerged from our preliminary investigation. The distribution of important measures among NBA and ABA players has been better understood thanks to visualizations like scatter plots and histograms. Even while we're still fine-tuning our classification models, preliminary findings imply that performance indicators like assists, rebounds, and points per game might be important factors in deciding who should be the MVP.

**Limitations:**

There were a myriad of limitations that arose while conducting this project. They all contributed towards the results of this project and conflict with one another. Initially, a big problem that we found was the dataset that we used. It contained all the statistics of NBA and ABA players, making it TOO BROAD for our classification model. Due to this, we had a small number of positive samples (players who were MVP’s) and a large number of negative samples, introducing bias to our results and skewing our model. Additionally, another issue that arose was with the way we chose to implement the model, using a single split. Single split means that the data is split into a single training set and a single testing set. Although this method is often used in machine learning models, this causes problems with low datasets, as there is not enough variability to strengthen the model. Another limitation that we observed after the process would be with our usage of fixed random state. Similar to the single split method, fixed random state is

a method often utilized within machine learning models, and helps with reinforcement of the model by introducing reproducibility. Unfortunately, with a skewed dataset due to our small positive cases, the model continually trained itself on skewed data, harming the accuracy of the model overall. With a combination of all of these aspects, our model has a heavy risk of potential overfitting or underfitting. Overfitting occurs when the model learns all of the noise within the training data, which leads to poor generalization on unseen data. While underfitting occurs when the model is too simplistic to capture the underlying patterns in the data. Moving forward, we'll have to instill heavier attention to detail and research skills to ensure that these limitations are not long term issues.

**Conclusion:**

In conclusion, our endeavor is an extensive attempt to identify the key performance indicators that determine the NBA MVP. Even though our investigation is still in progress, we think that the information we uncover will help the basketball community and advance our knowledge of what it takes to succeed in the NBA. We look forward to further improving our models and sharing our final results with the NBA players, fans, and pundits who are eager to learn more about this interesting issue.

**Appendix:**

Question: What was the most challenging aspect of this project?

Answer: The most challenging aspect of the project would most definitely be the sheer amount of limitations we endured during the process, as well as the lack of knowledge that we had to absolve these issues. Not only did we face plenty of challenges along the way, but we originally had a completely different project objective and dataset. Around halfway through the methodology process, we found that the dataset didn't have sufficient information for the objective we sought to accomplish. As a result, we decided to try and base a new objective around a detailed dataset. However, this would also reveal a problem, as the wide variety of attributes our dataset ultimately introduced skewed results to our classification model. I think we did the best that we could with the materials and resources that we had. Moving forward, we will

have to console the professor in frequency, as well as conduct better research to implement patches to the observed limitations.

Question: What was your favorite part of this project?
Answer:  One of the most exciting parts of this project for me was delving into the world of machine learning and linear regression. It provided me with a fantastic opportunity to dive deep into these advanced analytical techniques and gain a deeper understanding of their applications in the context of NBA player evaluation. Exploring machine learning methodologies and linear regression analysis not only expanded my skill set but also allowed me to uncover valuable insights into the intricate relationships within NBA player performance. Seeing how these advanced statistical models can be used to predict NBA MVP awards and determine the specifics of player evaluation was intense but incredibly fascinating. This practical experience has increased my understanding of the specifics of data analysis in the context of professional sports while also broadening my knowledge in that area. Overall, this project's most memorable and rewarding experience was delving deeply into linear regression and machine learning. I can't wait to use these methods more and more in the future.

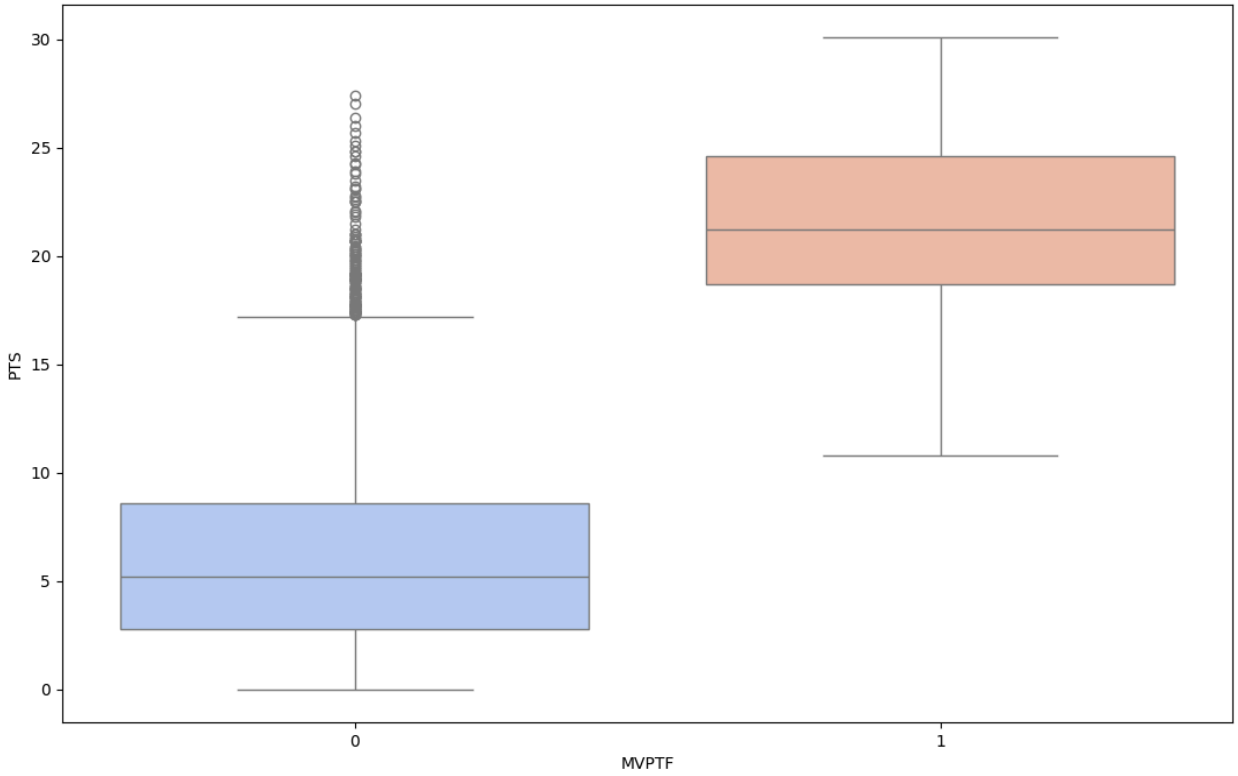
**References:**

https://www.kaggle.com/datasets/ryanschubertds/all-nba-aba-players-bio-stats-accolades/data
Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

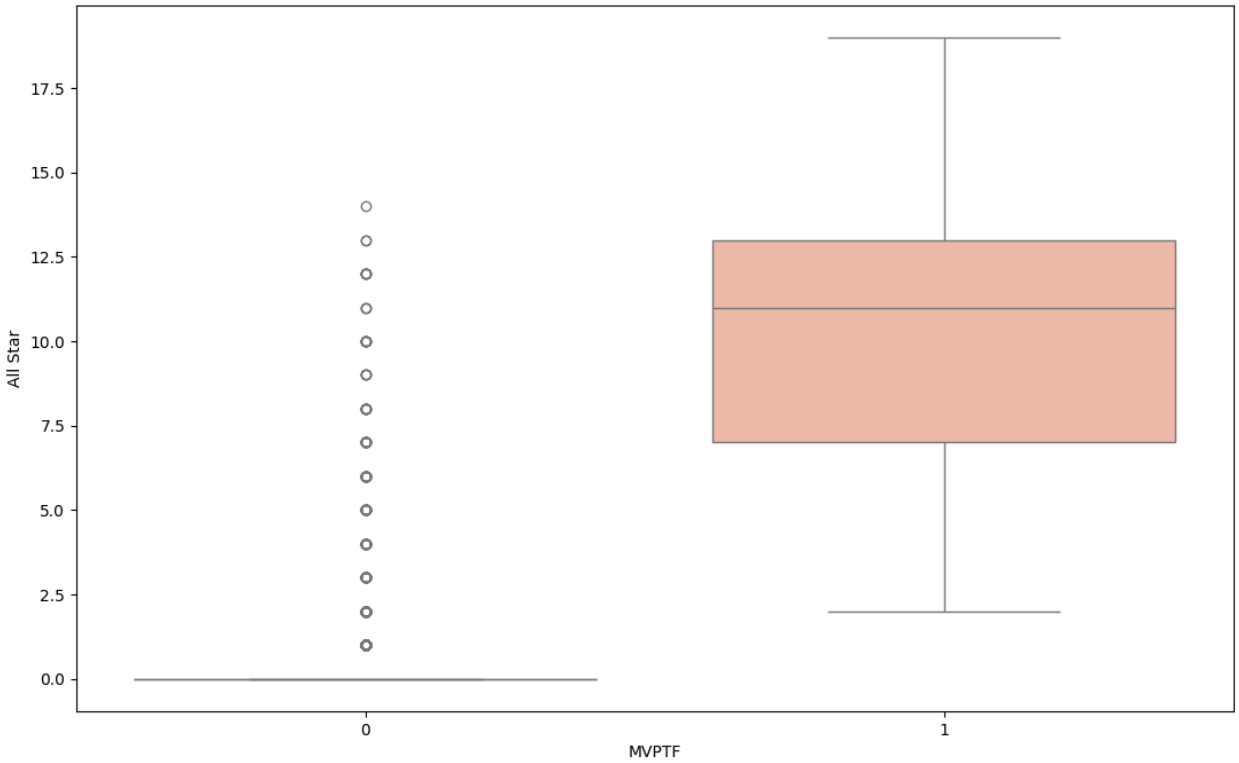
**Code:**

https://github.com/VeeJeyTee/I310D-Final-Project

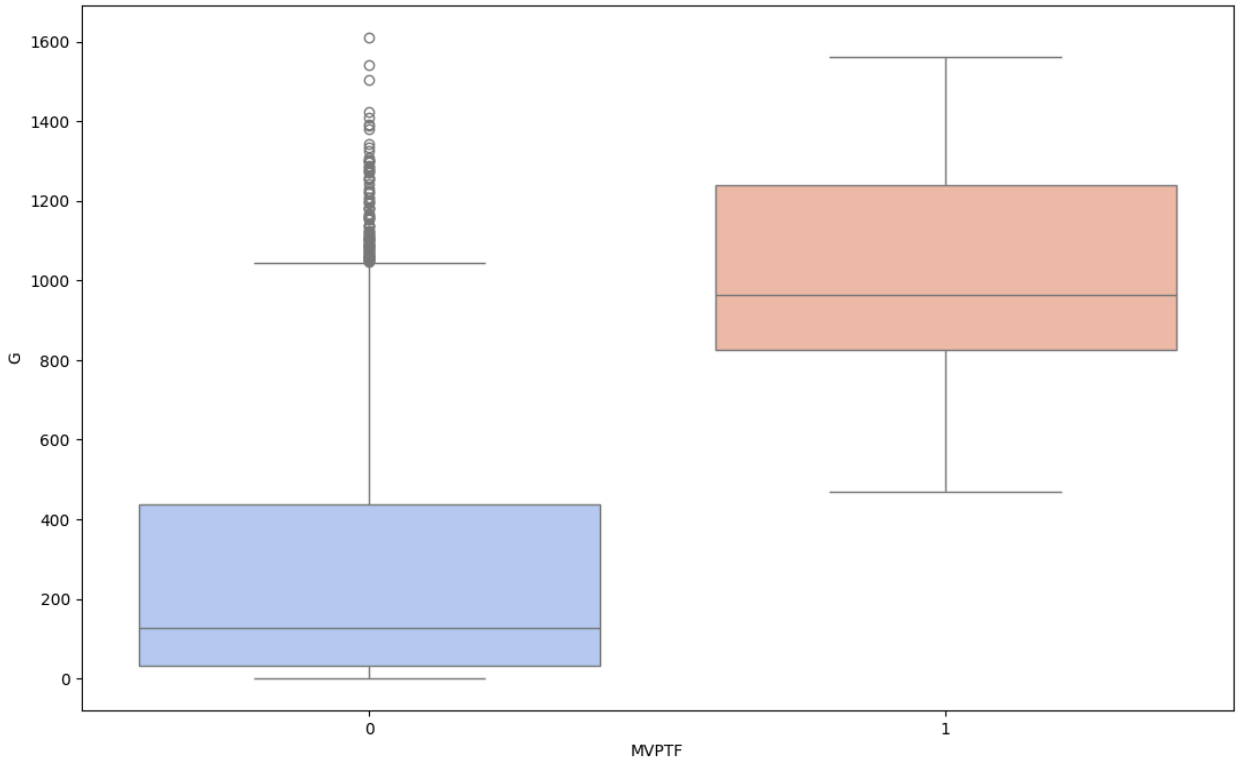
## Visual Aids:



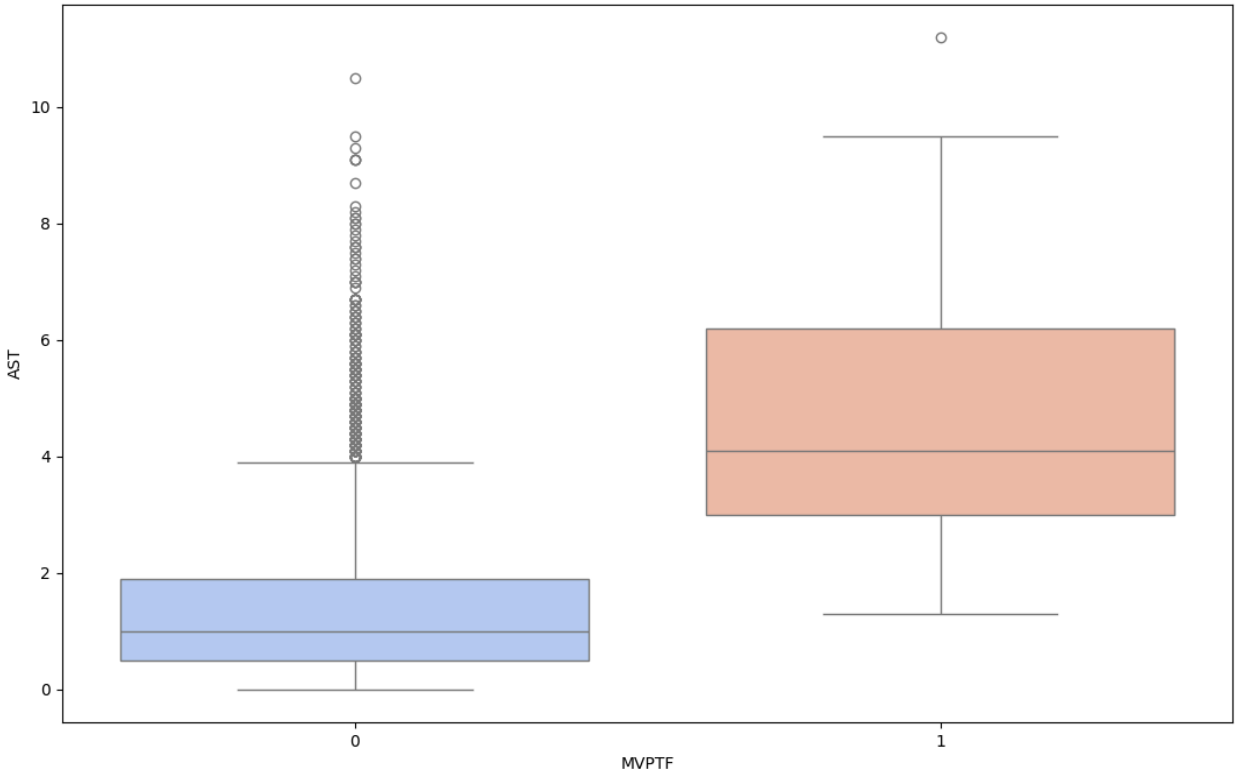

**Visual depicting relationship between career average points and NBA MVP recognition**



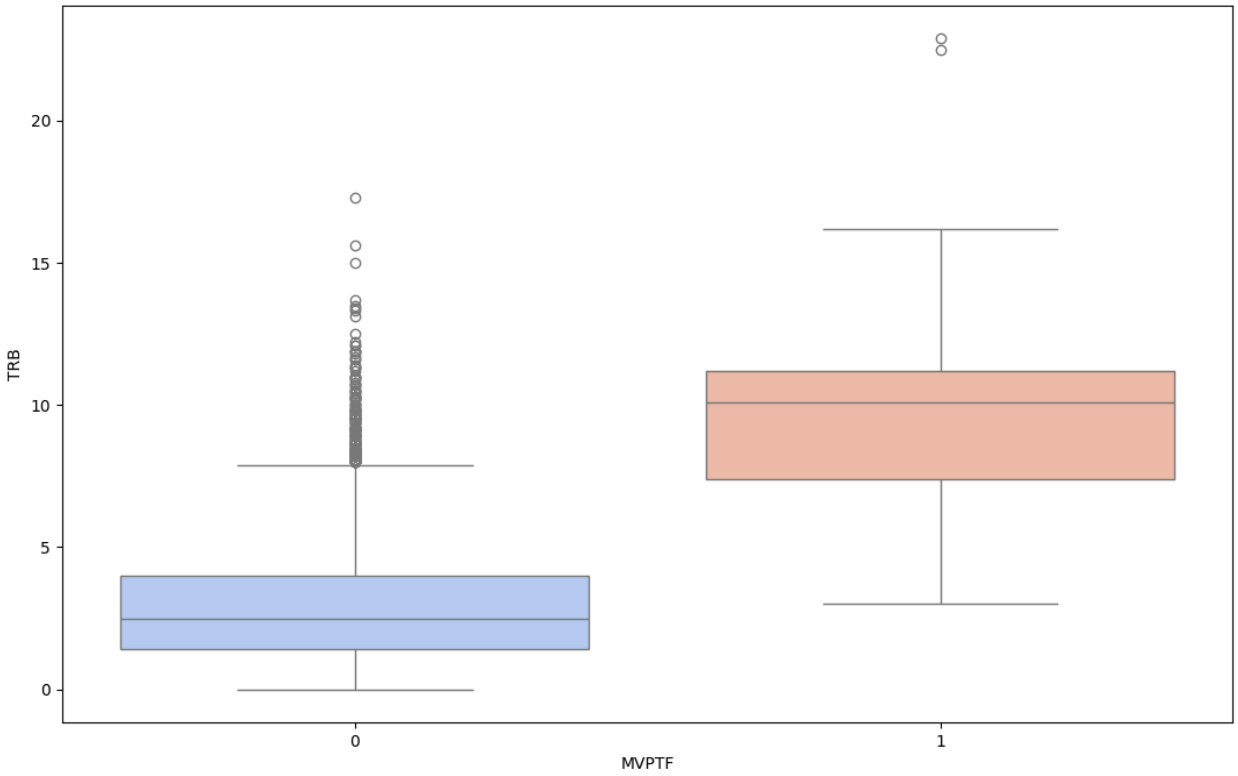

**Visual depicting relationship between NBA All star recognition and NBA MVP recognition**



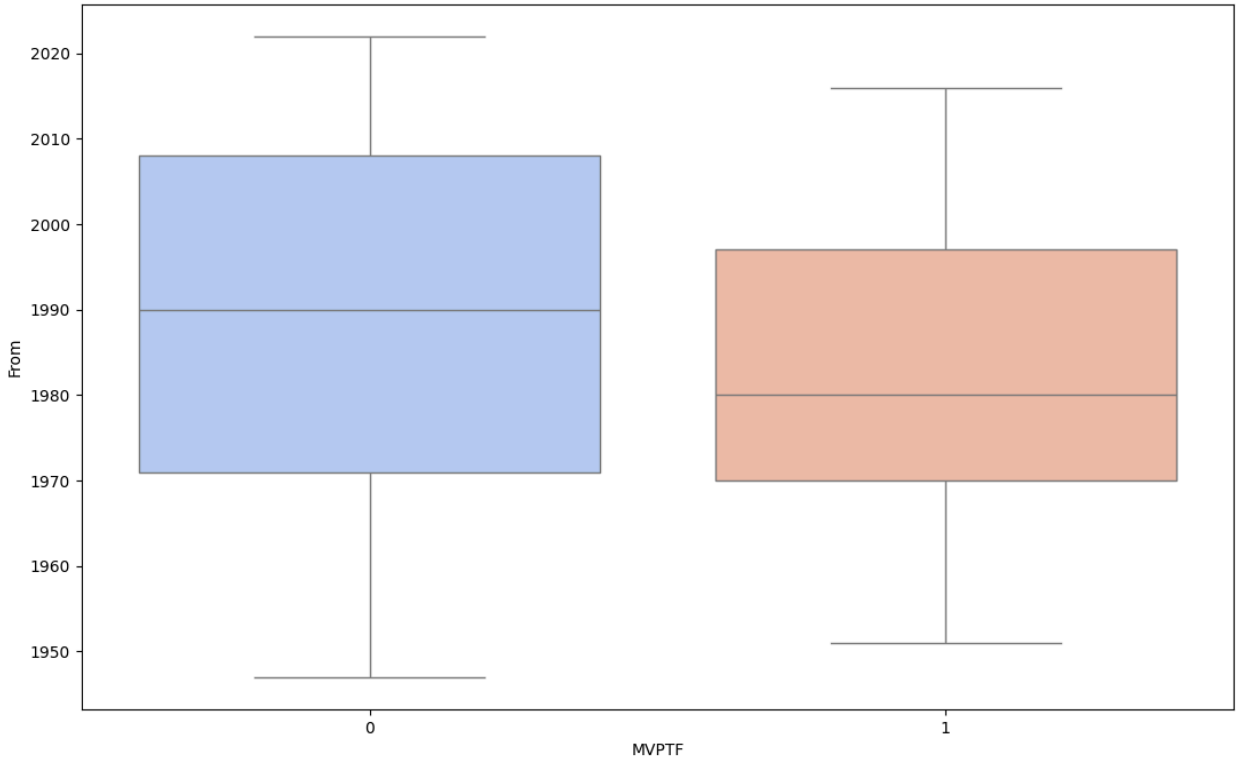
**Visual depicting relationship between total number of games played and NBA MVP recognition**




**Visual depicting relationship between career average assists and NBA MVP recognition**




**Visual depicting relationship between career average total rebounds and NBA MVP recognition**




**Visual depicting relationship between active NBA season years and NBA MVP recognition**

```
Logistic Regression
[[820   2]
 [  3   5]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       822
           1       0.71      0.62      0.67         8

    accuracy                           0.99       830
   macro avg       0.86      0.81      0.83       830
weighted avg       0.99      0.99      0.99       830

Random Forest
[[822   0]
 [  2   6]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       822
           1       1.00      0.75      0.86         8

    accuracy                           1.00       830
   macro avg       1.00      0.88      0.93       830
weighted avg       1.00      1.00      1.00       830

K Nearest Neighbor
[[821   1]
 [  5   3]]
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       822
           1       0.75      0.38      0.50         8

    accuracy                           0.99       830
   macro avg       0.87      0.69      0.75       830
weighted avg       0.99      0.99      0.99       830
```

**Classification Model results**