Predicting the NBA's MVP Using Machine Learning: A Comparative Study

Warren Sucklal
Department of Engineering
Western University
London, Canada
wsucklal@uwo.ca
251102996

Stefan Ilic
Department of Computer Science
Western University
London, Canada
silic4@uwo.ca
251091076

***Abstract*—This study investigates the performance of Random Forest (RF) and Neural Network (NN) models in predicting the National Basketball Association (NBA) Most Valuable Player (MVP) in comparison to other forecasting techniques. In addition, the study aims to account for the phenomenon of voter fatigue in MVP voting. The analysis involved applying these models to a dataset of NBA MVP contenders from 1980 to the present day. The accuracy of the models was evaluated using various metrics, including MSE, MAE, R2, and most importantly recall. Experimental results demonstrate that our RF model outperformed other readily available models in terms of accuracy. This can be attributed to the RF's ability to capture and handle non-linear relationships among MVP voting. The NN did a better job at modelling voter fatigue; however, it failed to accurately predict the NBA's MVP consistently. We hypothesize that this is due to the limited data available to properly train the NN. Future work might explore methods to improve upon our models, such as implementing techniques to balance class distribution, including qualitative assessments, or extending the model to dynamically adapt to in-season changes.**

***Index Terms*—National Basketball Association, Most Valuable Player, Random Forest, Neural Network, Machine Learning**

I. Introduction

A. Context

In the high-stakes world of professional basketball, the National Basketball Association (NBA) Most Valuable Player (MVP) award stands as a pinnacle of individual achievement. The MVP is awarded to the player who has been the most valuable to their team in the regular season. The determination of the MVP is a subject of intense scrutiny and debate among fans, analysts, and stakeholders, underscoring its significance in the sport's landscape.

The NBA's MVP voting process is as follows: A 100-person panel is selected by the league to vote on an MVP. Panelists are independent media figures and are considered basketball subject matter experts. The vote follows a process similar to the Borda Count where panelists rank their first-place through fifth-place candidates. Each player receives a certain amount of points from being placed first, second, third, fourth, and fifth. In the 2022-2023 season, the rankings of 1st place, 2nd place, 3rd place, 4th place and 5th place corresponded to 10, 7, 5, 3, and 1 point/points respectively. The player with the most points wins the award. In the 2022-2023 season, as there were 100 panellists, the maximum number of points a player could have been awarded was 1000.

Player, Team	1st Place Votes	2nd Place Votes	3rd Place Votes	4th Place Votes	5th Place Votes	Total Points
Joel Embiid, Philadelphia	73	25	2	0	0	915
Nikola Jokić, Denver	15	52	32	0	0	674
Giannis Antetokounmpo, Milwaukee	12	23	65	0	0	606
Jayson Tatum, Boston	0	0	1	89	8	280
Shai Gilgeous-Alexander, Oklahoma City	0	0	0	6	28	46
Donovan Mitchell, Cleveland	0	0	0	1	27	30
Domantas Sabonis, Sacramento	0	0	0	1	24	27
Luka Dončić, Dallas	0	0	0	2	4	10
Stephen Curry, Golden State	0	0	0	1	2	5
Jimmy Butler, Miami	0	0	0	0	3	3
De'Aaron Fox, Sacramento	0	0	0	0	2	2
Jalen Brunson, New York	0	0	0	0	1	1
Ja Morant, Memphis	0	0	0	0	1	1

TABLE I
2022-23 KIA NBA Most Valuable Player Award Voting Results [1]

B. Current Situation and Problems

As these votes are subjective and analysts often differ in their opinions, we want to investigate which player is most likely to win the award and what factors analysts value most in picking their regular season MVP. The implications of the MVP selection extend far beyond the hardwood floors of NBA arenas. In the dynamic world of sports betting, the MVP award plays a crucial role. Bettors often wager money on who will win the award. Furthermore, in recent years, there has been a large increase in sports betting following its legalization in Ontario [2]. Prediction is essential across the betting industry for bettors, bookies, and the average sports fan; therefore, we hope that by democratizing and comparing complex models with non-linear capabilities, we give more insight to the MVP race across the betting industry. In addition, both of the authors of this paper are NBA fans, meaning that this report and model function as a passion project as well.

C. Research Gap and Novelty

We hope to enhance existing models by specifically addressing voter fatigue, which is an overlooked factor in previous studies. Pastorello's analysis, which utilized a random forest model, showed promising results in modeling the NBA MVP. The gap suggests the potential for a custom random forest model that includes a measure of voter fatigue, which could improve the accuracy of NBA MVP predictions [3].

The second area of interest is the exploration of neural networks for predicting the NBA MVP. Previous models have not experimented with neural networks, and there is curiosity about whether neural networks could yield better predictive accuracy compared to existing models. This represents a significant gap in the research, as neural networks have the potential to uncover new insights and enhance prediction models for the NBA MVP award.

D. Objectives

In terms of objectives, we hope to create a new MVP prediction model that accounts for voter fatigue. In addition, we hope to create a neural network to predict the NBA MVP.

E. Results and Impact

The RF model accurately predicted the NBA MVP in 79.5% of cases, outperforming others but failed to capture voter fatigue due to its rarity. The NN model, despite overall lower accuracy, successfully predicted Embiid as the 2023 MVP, indicating better handling of complex patterns like voter fatigue. Both models predict Jokic as the 2024 MVP. We hope that the higher achieved accuracy by attempting to include voter fatigue within our model will motivate others to do the same.

II. Background Related Work

A. Previous Work in the Field and Literature Analysis

Most of the previous work done on modelling NBA MVP share has been done through the use of linear prediction methods. For example, the Shen-Nagy model for NBA MVP prediction applies linear transformations to a few advanced NBA statistics to come up with an NBA MVP prediction. Shen and Nagy use statistics such as PER (Player Efficiency Rating), True Shooting Percentage, and Offensive and Defensive Rating. They use Usage Rates and Winshares as prerequisites. They achieve a claimed 0.882 accuracy for MVP predictions across all of their tested seasons; however, they predict the 2022-2023 MVP wrong as Nikola Jokic [4]. In reading through their methodology, we are concerned that this model may be overfit to the training data.

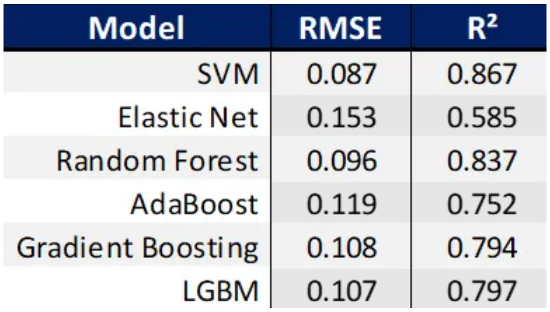
Tummala does better by implementing a random forest to capture non-linear relationships in order to predict the NBAs MVP. The random forest model is a non-parametric algorithm, which means it makes no underlying assumptions about the distribution of data or the form of the relationship between features and the target variable. This is useful because the relationship between player statistics and the MVP award is likely complex and non-linear. Random forests naturally take into account interactions between features. In basketball, player performance is not just about individual stats but how different aspects of a player's performance interact (e.g., points per game combined with assists or rebounds). There are many stories that can be woven into an MVP story beyond just simple stats, and the hope is that by using a non-linear model, some of these stories can be captured [5].

Yoo and Pastorello explore the use of many different models in both of their blogs published in Towards Data Science including Support Vector Machines (SVM), Elastic Net, Random Forest, AdaBoost, Gradient Boosting, and Light Gradient Boosting Machine. However, as highlighted by Pastorello in his article, his model predicts for Nikola Jokic to win his third MVP in the 2022-2023 season. Pastorello acknowledges that this is unlikely and underscores that at the time Joel Embiid was the heavy favourite to win NBA MVP [3][6].

This phenomenon, which is modelled poorly by the above mdoels, is known among analysts as voter fatigue. This was seen most recently with Nikola Jokic after having won two MVPs in the 2020-2021, and 2021-2022 seasons. Despite having the most impressive stats that typically lead to NBA MVPs, such as win shares, he still did not win MVP as he had won it both prior years. NBA analysts are human beings and love a good story; therefore, they get tired of voting for the same great player over and over, and become more likely to award an MVP to another player.

B. Research Gap

With the previous work in mind, we have outlined two major areas that are worth investigating. Firstly, we want to look into improving on the previous work done in modelling the NBA MVP by taking into account voter fatigue. In particular, a random forest model seems to have done particularly well in modelling the NBA MVP in Pastorello's analysis.



Model	RMSE	R^2
SVM	0.087	0.867
Elastic Net	0.153	0.585
Random Forest	0.096	0.837
AdaBoost	0.119	0.752
Gradient Boosting	0.108	0.794
LGBM	0.107	0.797

Fig. 1. Performance of Pastello's Models [6]

Therefore, we wonder if augmenting our custom random forest model with some sort of fatigue measure will help us in predicting the NBA MVP.

Secondly, as all of the models that have been published don't experiment with neural networks, we wonder if the use of a neural network for this problem will allow us to come up with any interesting findings and beat any of the other models in terms of accuracy.

III. Methods

A. Research Objectives

For both of our objectives, we will measure accuracy by using the following simple formula:

$$\frac{\text{number of correct mvp predictions}}{\text{number of seasons analyzed}}$$

Objective 1: We hope to compare the accuracy of our random forest model compared to the accuracy of other models found in the research literature.

Objective 2: We hope to compare the accuracy of our neural network model compared to the accuracy of other models found in the research literature.

B. Data Collection and Interpretation

Two essential data categories are necessary the historical data of past MVP contenders, encompassing both the recipients and the considered nominees, and the performance data of the MVP nominees for the ongoing 2023–2024 season.

For the first category, comprehensive records of MVP candidates were retrieved from Basketball-Reference.com, which hosts a dedicated page for each season's MVP competition. Though the NBA has a history dating back to the 1949–1950 season, the gathered historical data pertains to the modern era, spanning from the 1979-1980 through the 2022-2023 seasons. Data collection was done using a modified version of the Basketball Reference web scraping and data gathering tool published by Yoo [6][7].

Regarding the second category, the current 2023–2024 season's MVP hopefuls were identified via the NBA.com MVP KIA Ladder feature. Throughout the season, the NBA updates this section weekly to highlight the top 10 contenders for the MVP title. The statistics for these selected players were subsequently acquired from Basketball Reference using the same tool above [6][7].

To align individual player basic and advanced statistics along with team performance metrics to the correct player and season, a combination of data extraction techniques was employed. This included utilizing the Pandas library for HTML table extraction, the Beautiful Soup library for web scraping, and a specialized basketball reference scraping tool for acquiring the raw data [7]. In addition, a VPN and a scraping delay had to be used in conjunction with the scraping script in order to not be kicked from the website. Finally, certain players like Luka Doncic had to have their statistics extracted manually due to data extraction and formatting issues within the script. We assume that this is because Luka Doncic's last name is often spelt with an extra accent on the middle and the last C, as we experienced similar issue pulling some of Nikola Jokic's data that we had to rectify.

In order to model MVP voter fatigue, we created a script to add two new derived columns for each observation, won1yearago and won2yearsago. These columns are a one-hot encoding of whether a player has won the MVP one season ago, or two seasons ago respectively. These features are added using the AddingMVPFatigue.ipynb script. We hope that these added features will allow us to better model voter fatigue.

Another point we want to bring up is regarding class distribution of MVP contenders against non-contenders. Due to the vast majority of NBA players being non-contenders for the MVP award, we have a major class distribution problem, as most players would receive an MVP share of 0. This is especially problematic as the NBA is rife with talent, and you have many players who have impressive stats on poorly performing teams in highly correlated features such as Win Shares. Therefore, we have decided to only include the players that are considered by the media to be MVP contenders. This point is discussed more in the threats to validity portion of the report.

C. Exploratory Analysis

When performing exploratory analysis, the first order of business is to plot the distribution of the dependent variable. We will choose the "share" feature as our dependent variable. This is a derived variable calculated by taking the MVP voting points the MVP candidate scored and dividing it by the total number of MVP voting points the player could have been awarded in a season. We choose MVP Share as the dependent variable as it is a ratio and accounts for the varying number of points that can be awarded in any one season.

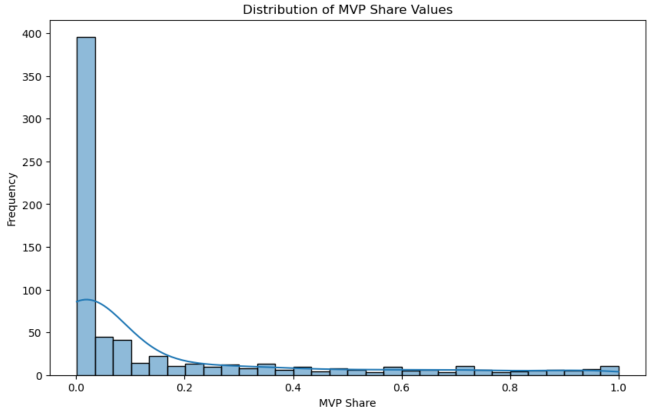


Fig. 2. Distribution of MVP Share

As we can see, the data is heavily right-skewed, and the average value of MVP share is quite low. Due to this skew, the assumption that the errors of our prediction are normally distributed may not hold, potentially leading to bias in parameter estimates and predictions. In addition, we have to be careful regarding the predictions near the mode of the distribution. Where there are many observations, the prediction quality might be better due to the model having more data to learn from. However, the prediction quality can decrease for larger values due to fewer training examples. This is of great concern to us as we are concerned with the players with the largest MVP share values. It is worth noting that models like Random Forest are non-parametric and can sometimes handle skewed data better than parametric models since they do not assume a normal distribution [8].

We can also look into which of our features are most correlated with MVP share to get a better idea for feature selection.

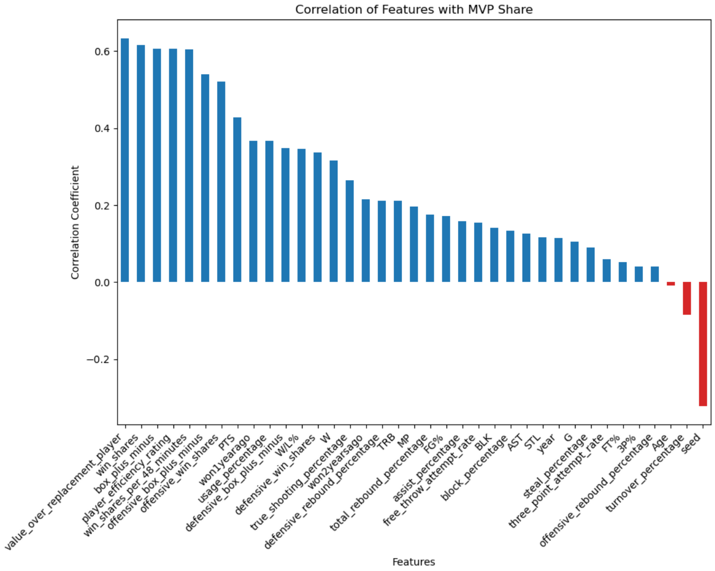


Fig. 3. Correlation of features to MVP Share

There are no major surprises among these correlations to an avid basketball fan. With a correlation coefficient of 0.6334, Value Over Replacement Player (VORP) is the most strongly positively correlated with MVP share. This indicates that players with a higher VORP tend to have a higher MVP share, suggesting that players who contribute more compared to a replacement-level player are more likely to be valued in MVP considerations [9].

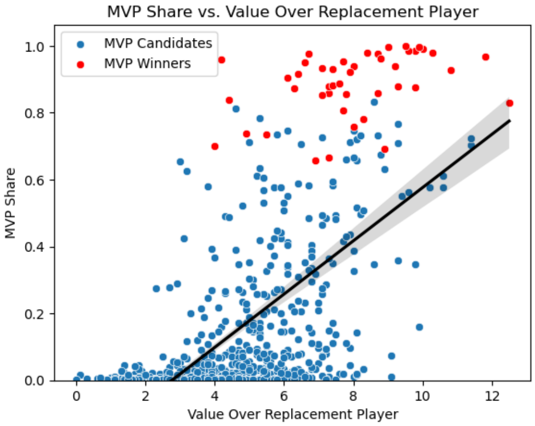


Fig. 4. Correlation of VORP to MVP Share

Win Shares (WS) has a correlation of 0.6152 with MVP share, indicating that players who contribute more to their team's victories have a higher chance of achieving a higher MVP share [10].

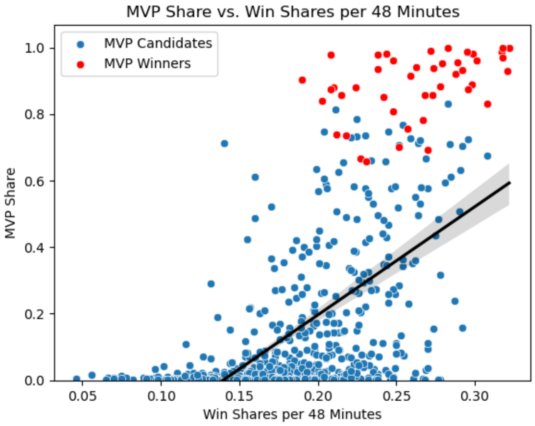


Fig. 5. Correlation of WS per 48 to MVP Share

Box Plus/Minus (BPM) with a correlation coefficient of 0.6061, BPM is another strong indicator. A higher BPM, which indicates a player's overall contribution to the team per 100 possessions, correlates with a higher MVP share.

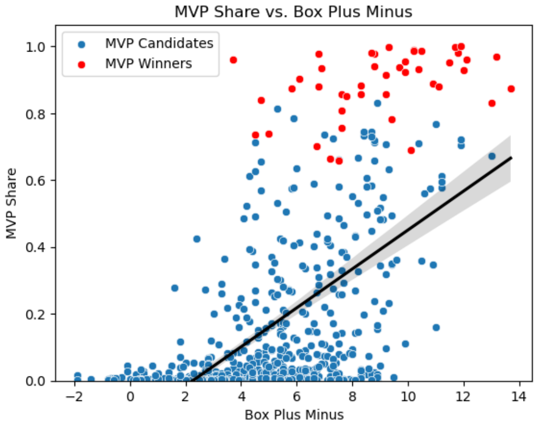


Fig. 6. Correlation of BPM to MVP Share

Player Efficiency Rating (PER) is very close to BPM in its correlation with MVP share (0.6061). However, unlike player efficiency, BPM is a better measure for a player's intangible impact on the floor. Players with a higher BPM may play better defence and may be better leaders when on the NBA floor.

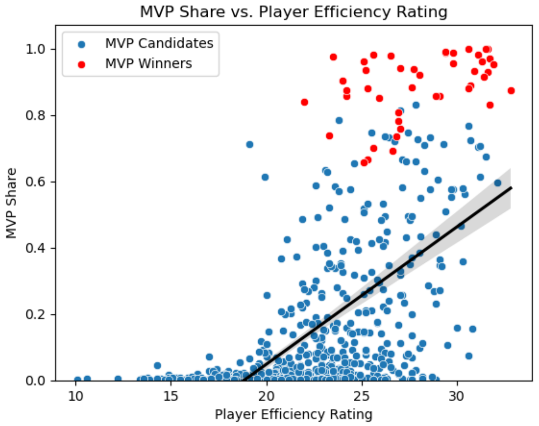


Fig. 7. Correlation of PER to MVP Share

Interestingly, at the bottom of the list, the "seed" feature has a notable negative correlation of -0.3212 with MVP share. This suggests that players from lower-seeded teams are less likely to have a high MVP share. Lower-seeded teams contributing to a lower MVP share make sense, as one of the ways that analysts evaluate MVPs is on their ability to bring their team wins. MVPs almost always come from teams that are highly seeded and perform well [11].

One important distinction that needs to be made is regarding counting and rate statistics. In our feature list, almost all features represent rate statistics, whether it be a measure like points per game, or box plus-minus. These statistics remain along a relatively similar range throughout the entire NBA season. On the other hand, counting measures such as Win Shares and VORP increase as the season goes on. This poses a problem for analysis as certain NBA seasons don't have 82 games in them, and the current NBA season has not had all 82 games for each team played. To combat this, we have decided to use rate statistics wherever we can in our model. For example, we use win Shares per 48 minutes over win shares, and we use win/lose percentage instead of the number of wins. Furthermore, we have decided to assume VORP grows linearly for each player throughout the season, allowing us to normalize each NBA season to 82 games for the VORP statistic for each player [9][10][12].

D. Random Forest Model Implementation

To predict the NBA MVP we decided to use Random Forests (RFs) for their ability to handle non-linear data. RFs operate by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees, making them robust against overfitting and capable of capturing complex patterns.

Our initial step involved rigorous cleaning of the dataset to ensure its integrity. This included handling missing values, eliminating duplicates, and rectifying inconsistencies.

We then moved onto feature engineering. Features were selected using Mutual Information, emphasizing those that historically influenced MVP voting including both advanced and basic statistics [13]. The crux of our RF model's success lies in judicious feature selection. Drawing on our deep understanding of basketball analytics, and through expert interviews, we selected features that encapsulate a player's contribution to their team, including advanced metrics like Win Shares per 48 minutes, Player Efficiency Rating (PER), and others [6][12]. This selection was informed by both domain expertise and the correlation analysis highlighted in our exploratory analysis.

In regard to categorical variable encoding and normalization, while RF models exhibit a degree of indifference to feature scale, standardization practices were nonetheless applied to facilitate interpretation and ensure a uniform data structure.

To build the model we used scikit-learn [14]. We instantiated a RF model with an initial configuration of parameters. The model was then trained on a curated set of features, with MVP share serving as the target variable. Our RF model was configured with a specific number of trees, carefully chosen to balance computational efficiency with predictive performance. We also experimented with different depths for the trees to prevent overfitting while ensuring sufficient complexity to capture the nuances of MVP voting behavior.

The performance of the RF model was evaluated using accuracy as the benchmark, along with the importance of each feature in the prediction. This not only provided insights into the model's effectiveness but also shed light on the factors most influential in determining the MVP.

Performance evaluation utilized metrics such as MAE, MSE, and R², offering a comprehensive view of the model's accuracy and predictive power. As seen in class, to ensure the robustness of our model, we employed cross-validation techniques. This approach involves partitioning the data into complementary subsets, training the model on one subset, and validating it on the other to ensure generalizability to new, unseen data. Cross-validation further validated the model's consistency across various data splits.

Hyperparameter tuning was also done through grid search and random search techniques, played a crucial role in refining the model. This iterative process aimed to enhance the model's precision and generalizability [15].

Continuous refinement and feature reevaluation were pivotal in the model's evolution. Strategies to mitigate overfitting, such as adjusting the maximum depth of trees and experimenting with the number of trees, were methodically employed.

In summary, the implementations of both the Random Forest and Neural Network models were underpinned by an iterative methodology, encompassing comprehensive data preprocessing, strategic model building, and steadfast refinement.

E. Neural Network Model Implementation

The allure of Neural Networks (NNs) lies in their ability to model complex, non-linear relationships between inputs and outputs, making them a promising tool for deciphering the nuanced dynamics of MVP voting. The endeavour to utilize NNs in predicting the NBA's MVP extends beyond mere model architecture and training. A meticulous approach to data preprocessing, model building, and iterative refinement underscores our methodology, aiming to harness the full potential of NNs in capturing the multifaceted determinants of MVP success [16].

Prior to feeding data into our neural network, we performed meticulous data preprocessing. This involved normalizing the statistics to a common scale, addressing the challenge of disparate ranges among variables such as points per game, rebounds, and assists. Normalization enhances the model's efficiency and convergence speed during training.

Echoing the RF approach, our dataset underwent cleaning and preparation. This step was pivotal in ensuring the NN model operated on clean and consistent data.

In regard to feature selection, we utilized Mutual Information and other feature importance metrics, we selected a blend of basic and advanced basketball statistics that hold significant sway in MVP considerations [13].

We also wanted to have the option to analyze categorical variables such as team names. Team times were standardized using a predefined JSON mapping to team abbreviations, enhancing the model's interpretability [7]. Further, one-hot encoding was applied to other nominal categorical variables, enriching the model's input data.

Despite NN's resilience to varied feature scales, we normalized our data to ensure uniformity, aiding in model convergence and facilitating a coherent interpretation of the learned weights.

The foundation of our NN model is a carefully structured architecture designed to capture the multifaceted nature of basketball statistics and their impact on MVP selection. We opted for a sequential model, layering densely connected neurons to facilitate the deep learning process. Each layer is equipped with a rectified linear unit (ReLU) activation function to introduce non-linearity, ensuring the model's ability to learn sophisticated patterns in the data [17].

Leveraging the flexibility of frameworks like TensorFlow and Keras, we constructed a sequential NN model. The training phase was meticulously monitored to balance learning efficiency with the risk of overfitting, utilizing techniques such as dropout and regularization [18].

The model's training was executed over multiple epochs, with each iteration presenting an opportunity for the model to refine its predictions. We employed a split of historical data for training and validation, allowing the model to learn from past MVP seasons while validating its predictions against a subset of the data not seen during the training phase.

Upon training, the model was evaluated against a held-out test set. Performance metrics such as accuracy and loss provided insights into the model's predictive capability. Iterative refinements, including adjustments to the network's depth and width, were made in pursuit of optimal performance.

To optimize the neural network, we utilized the Adam optimizer, renowned for its adaptive learning rate capabilities, facilitating faster convergence to the optimal solution. The model's performance was evaluated using accuracy as the primary metric, supplemented by loss reduction over epochs to monitor improvement [19].

Features
PTS
AST
TRB
seed
value_over_replacement_player
box_plus_minus
win_shares_per_48_minutes
won1yearago
won2yearsago
player_efficiency_rating
offensive_box_plus_minus
offensive_win_shares
W/L%
true_shooting_percentage

TABLE II
LIST OF FEATURES USED IN BOTH MODELS

IV. RESULTS

A. Random Forest Model

The results from the RF model presented an insightful evaluation of the effectiveness of machine learning in predicting the NBA's MVP. The RF model was implemented with 10 features, demonstrating its predictive capability through a series of evaluations. Specifically, for the 10-fold cross-validation approach, the model achieved a mean squared error (MSE) of 0.025, with an R-squared value of 0.629, indicating that approximately 62.9% of the variance in MVP voting could be explained by the model. The mean absolute error (MAE) was reported as 0.097, offering insights into the average deviation of predicted MVP shares from actual values.

The model's consistency and reliability were highlighted by its performance across different folds. For 10-fold cross-validation the model had an average MSE of 0.025 and a standard deviation of MSE at 0.0072. This demonstrates the model's robustness, as there are only slight variations in error rates across different subsets of the data, ensuring confidence in the model's predictive power.

The application of the model with a 5-fold cross-validation methodology yielded comparable results. The model maintained a consistent MSE, R-squared value, and MAE. The only slight variation in the average MSE (0.0265) and the standard deviation of MSE (0.0034) for the 5-fold cross-validation further illustrates the model's adaptability and stability in handling different data partitions.

A critical achievement of the RF model was its ability to correctly predict 79.5% of all MVPs. This high success rate not only exemplifies the model's efficacy in identifying potential MVP candidates but also highlights the potential of machine learning techniques in augmenting traditional MVP prediction methods.

In summary, the RF model's performance, characterized by its low error rates and high R-squared value, underscores the potential of applying advanced analytical techniques to sports analytics. By leveraging the Random Forest algorithm, we demonstrate a promising avenue for enhancing the accuracy and reliability of MVP predictions. We hope that we have uncovered valuable insights for stakeholders in the sports betting industry and the broader NBA community.

B. Neural Network

Our Neural Network model underwent rigorous testing, involving a series of predictions to ensure reliability and robustness in its forecasting ability. Across 10 separate predictions, the model exhibited an Average Mean Squared Error (MSE) of 0.0297, with a Standard Deviation of MSE standing at 0.003. This indicates a consistent performance with minimal variation in predictive accuracy across trials. Furthermore, the model achieved an average R-squared value of 0.416. This is lower than our previous RF model.

Further analysis with a reduced set of 5 predictions yielded similar results, underlining the model's stability. The Average MSE slightly improved to 0.0293, and the Standard Deviation of MSE narrowed to 0.001. The average R-squared value in this subset of predictions was 0.403, affirming the model's consistent performance.

A critical measure of the model's effectiveness is its success rate in correctly identifying potential MVP candidates. In this regard, the model demonstrated a meager success picking rate of 56.8%. This success rate is indicative of the model's current lack of utility in the complex arena of NBA MVP predictions. Numerous variables and unpredictable factors influence the final outcome; however, data quantity remains a significant hurdle. The model significantly under performed when compared to our RF model, and many of the other models analyzed in the literature.

C. Comparison Between Models and Summary

Method	Recall	Correctly Predicts 2023 MVP
Shen Nagy	88.2%	No
Yoo RF	73.8%	No
Our RF	79.5%	No
Our NN	56.8%	Yes

TABLE III
PREDICTION ACCURACY AND SUCCESS FOR THE 2023 MVP.

To summarize, our RF model manages to accurately predict the NBA MVP in 79.5% of test cases, outperforming all other machine learning models we analyzed that published their testing results. Despite our best effort, it appears that The RF model was still not able to effectively model the phenomenon of voter fatigue. We hypothesize that this is due to the phenomenon of voter fatigue not appearing too often in the training data. In general, a player has to be a generational talent to have won back to back MVPs. This has only happened a few times in modern NBA era.

In regard to the NN model, it effectively predicted that Embiid would win the MVP in 2023 ahead of Jokic, potentially indicating that it models voter fatigue better. This makes sense as NN would be much better at modeling non-linear relationships compared to other models. However, the model in general performs very poorly and has the lowest accuracy among all methods analyzed. The NN's proficiency at taking voter fatigue into account was hard to quantify due to the small number of times voter fatigue has come into play in the modern NBA.

Both our NN and RF model predict that Jokic will win MVP in the 2024 season.

D. Threats to Validity

In the process of modeling the NBA's MVP predictions using machine learning techniques, several threats to the validity of our results have been identified. Addressing these concerns is crucial for ensuring the reliability and robustness of our predictive models. These reasons include the class distribution of MVP candidates and non-candidates, incomplete season data, historical data constraints, the evolution of the NBA, and subjectivity in voting.

The disproportionate class distribution of MVP candidates versus actual MVP winners poses a significant threat to the validity of our model. The model is exposed to a large number of MVP candidates compared to a much smaller group of actual MVPs. This imbalance could lead to a tendency for the model to overfit to the characteristics of MVP candidates without discerning the nuanced differences that distinguish an MVP. Overfitting might result in excellent performance on training data, but it would likely falter in generalizing to new, unseen data.

With 2024 season not yet concluded, our models rely on incomplete data for metrics such as Win Shares and VORP. This necessitates the estimation of these statistics, introducing potential biases and inaccuracies. Since these metrics are highly correlated and contribute significantly to a player's MVP candidacy, any error in estimation could lead to faulty predictions. Furthermore, player performance can easily change, injuries happen, and team dynamics shift. Any early prediction of MVP success can be easily invalidated later on in the season. For example, in the 2024 season Joel Embiid was the favourite to win MVP until he got injured [20].

Focusing exclusively on the modern era of the NBA post-1980 may limit the scope of our dataset and the resulting analysis. With fewer MVP seasons to analyze, the model has limited examples from which to learn, potentially hindering its ability to capture the full spectrum of variability within MVP selections. Furthermore, the game of basketball has evolved significantly, with changes in strategies, player roles, and the relative importance of different statistics. This evolution could mean that the factors contributing to MVP selection have shifted over time, making it challenging to create a model that is time-agnostic. For example, in the 70s centers tended to win MVP as they were the most valued players in the league; however, in the 80s guards were seen as the most likely to win MVP. These two positions have vastly differing playstyles [21].

MVP voting is inherently subjective, with voters' preferences and biases possibly changing over time. This subjectivity is difficult to quantify and therefore to model accurately. Our machine learning models can capture trends and patterns in the data but cannot fully account for the individual voter's decision-making process. In addition, our models primarily focus on quantitative data and may overlook qualitative factors such as player leadership, team chemistry, narrative, and media influence, all of which can impact MVP voting. The quantification of such aspects is complex and not directly represented in the dataset, which could lead to a misalignment between model predictions and actual MVP outcomes. For example, the 2006 Steve Nash MVP awarding was seen as very controversial. Nash's coach, Mike D'Antoni promoted a high power fast paced offence that was innovative and people enjoyed watching, helping Nash in MVP voting. In addition, Nash was a part of the coveted 50-40-90 club which means he shot 50, 40, and 90 percent from the field, behind the 3-point line, and free-throw line respectively. Our model would struggle to capture the above subjective measures [21].

Future work might explore methods to mitigate these threats, such as implementing techniques to balance class distribution, including qualitative assessments, or extending the model to dynamically adapt to in-season changes.

V. Conclusion

Our exploration into RF and NN models has yielded notable distinctions in their predictive capabilities. In particular, there were notable distinctions concerning the elusive variable of voter fatigue, a factor often understated in MVP discussions.

The RF model's 79.5% accuracy in MVP predictions marks a significant advancement in the field of studies what be analyzed. The RF model leaves significant room for improvement in modeling voter fatigue. The NN's adeptness at identifying Joel Embiid as the 2023 MVP hints at its potential to navigate the complex terrain of non-linear variables like voter fatigue. However, its overall performance leaves much to be desired. The consistent prediction from both models that Nikola Jokic will claim the 2024 MVP title sets a curious precedent for the upcoming season.

We hope that this study encourages future scholars and analysts to dive deeper into the intricacies of MVP prediction. A promising avenue lies in expanding the dataset to encapsulate a more extensive range of variables. Variables that are less quantifiable such as an MVP's marketability, or the narrative surrounding their season are important inputs that haven't been modelled. This could be done by potentially exploring unstructured data such as player narratives and media sentiment. Moreover, refining our models to dynamically adjust to in-season developments and adopting techniques to manage class imbalance could significantly bolster predictive precision.

The project underscored the importance of data quality and variety. There were major leanings regarding the challenge of balancing feature selection, and the need for models to adapt to the non-static nature of sports analytics. Perhaps the most crucial lesson was the acknowledgment of the inherent subjectivity in MVP voting. We learned that a theoretically simple narrative factor such as MVP fatigue ended up being very hard to model. That being said, there remains an opportunity for machine learning to intertwine quantitative analysis with the qualitative essence of sport.

REFERENCES

[1] "2022-23 Kia NBA Most Valuable Player Voting Results," NBA, 2023. [Online]. Available: https://ak-static.cms.nba.com/wp-content/uploads/sites/46/2023/05/2022-23-Kia-NBA-Most-Valuable-Player-Voting-Results.pdf.

[2] N. Turner et al., "Brief Report: The Rise of Online Betting in Ontario," Journal of Gambling Studies, vol. 39, no. 4, December 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10899-023-10268-1. DOI: 10.1007/s10899-023-10268-1.

[3] G. Pastorello, "Predicting the NBA MVP with Machine Learning," Towards Data Science, 2023. [Online]. Available: https://towardsdatascience.com/predicting-the-nba-mvp-with-machine-learning-c3e5b755f42e.

[4] L. Shen and G. Nagy, "Calculating NBA MVPs Using Advanced Sports Analytics," University of Pennsylvania, Philadelphia, PA, USA, 2023. [Online]. Available: https://wsb.wharton.upenn.edu/wp-content/uploads/2023/05/Shen_2023_Basketball_MVP.pdf.

[5] "Using Machine Learning to Predict the NBA MVP," Samford University, 2023. [Online]. Available: https://www.samford.edu/sports-analytics/fans/2023/Using-Machine-Learning-to-Predict-the-NBA-MVP.

[6] D. Yoo, "Predicting the Next NBA MVP Using Machine Learning," Towards Data Science, 2023. [Online]. Available: https://towardsdatascience.com/predicting-the-next-nba-mvp-using-machine-learning-62615bfcff75.

[7] D. Yoo, "NBA MVP Project," GitHub, 2023. [Online]. Available: https://github.com/DavidYoo912/nba_mvp_project.

[8] "Skewness and the Meanings of Mean," Newcastle University, 2023. [Online]. Available: https://www.mas.ncl.ac.uk/ njnsm/medfac/docs/skewness.

[9] "Box Plus/Minus and VORP," Hackastat, 2023. [Online]. Available: https://hackastat.eu/en/learn-a-stat-box-plus-minus-and-vorp/.

[10] "Advanced NBA Stats for Dummies: Understanding the New Hoops Math," Bleacher Report, 2023. [Online]. Available: https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math.

[11] "NBA Most Valuable Player Award," Basketball Reference, 2023. [Online]. Available: https://www.basketball-reference.com/awards/mvp.html.

[12] Interview with J. Cameron, former competitive basketball player and NBA data science analyst, April 2024.

[13] "Mutual Information with Python," TrainInData, [Online]. Available: https://www.blog.trainindata.com/mutual-information-with-python/.

[14] "sklearn.ensemble.RandomForestRegressor," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

[15] W. Koehrsen, "Hyperparameter Tuning the Random Forest in Python Using Scikit-Learn," Towards Data Science, [Online]. Available: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74.

[16] "Supervised learning with neural networks," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/neural_networks_supervised.html.

[17] M. Di Pietro, "Deep Learning with Python: Neural Networks Complete Tutorial," Towards Data Science, [Online]. Available: https://towardsdatascience.com/deep-learning-with-python-neural-networks-complete-tutorial-6b53c0b06af0.

[18] "Keras," TensorFlow, [Online]. Available: https://www.tensorflow.org/guide/keras.

[19] "Adam Optimization Algorithm for Deep Learning," Machine Learning Mastery, [Online]. Available: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/.

[20] "Joel Embiid knee injury timeline, return updates and latest news on 76ers star," Sporting News, [Online]. Available: https://www.sportingnews.com/ca/nba/news/joel-embiid-knee-injury-timeline-return-updates-76ers/3ffd4fd4359e8baa94f3323b.

[21] "2005-06 MVP Race: Steve Nash Upset LeBron James and Kobe Bryant, Shocked the World and Won the Award," Fadeaway World, 2023. [Online]. Available: https://fadeawayworld.net/nba/2005-06-mvp-race-steve-nash-upset-lebron-james-and-kobe-bryant-shocked-the-world-and-won-the-award.