CS-4641-Machine-Learning-Project

CS-4641 Machine Learning Project

Proposal | Midterm Report | Final Report

Group 9 Final Report

Introduction

The goal of the project is to develop a machine learning model to predict whether a given basketball player would become the NBA MVP based on previous NBA MVPs and player statistics. The NBA MVP is an annual award given to the best performing player in the NBA during a regular season. A player is chose and awarded based on a panel of sportswriters, broadcasters, and former players' votes based on a current player's seasonal performance such as points per game. The benefits of being chose as an NBA MVP include things like earned bonuses, and becoming seen as more desirable to other teams in the league. Being able to develop a model that can help sports analysts effectively predict the next NBA MVP would allow for benefits in decision making and scouting for players, along with that it can potentially provide the team with more opportunities for sponsorships and media attention.

Problem Definition

Sports media is split across many platforms with little variation between platforms, so it would be difficult for one company to attract and maintain a large user base over their competitors. With the rapid advances in machine learning, it is now possible for a sports media company to implement a way to predict future NBA mvp's as a way to interact with users. MVP predictions could be a novel feature that attracts new users to the platform, and an accurate prediction of the MVP would only serve to increase the platform's noteriety and reputation amongst other sport media companies. The added interaction could create a community for users to become active and debate with other fans about the predictions. Findings with machine learning could also be extending to improve the quality and safety of basketball and other sports [1]. This same data could be used by teams to increase skill, health and fitness, and player recruitment which would lead to higher quality sport entertainment for fans [2].

Methods

For our data preprocessing, we used multiple techniques that consist of: Feature Selection, One-Hot Encoding, Handling Missing Values, Standardization, and a Correlation Heatmap. First, we better contextualized the data through One-Hot Encoding of positions in the NBA to give them numerical values. We also filled missing values with 0's to account for players that did not play in any games for that NBA season. Then we used a correlation heatmap to find any relationships in the data as well as redundancy. Alongside the heatmap, we also standardized all of the data and created box and whisker plots of every stastical measure to find where past MVPs lie amongst each distribution. These visualizations allowed us to select the statistical categories that were likely to yield a MVP award. In the end, we settled with these statistical categories: 'G (Games played)', 'GS (Games Started)', 'MP (Minutes Played)', 'FG (Field Goal)', '2P (2 Pointers Made)', '2PA (2 Pointers Attempted)', 'FT (Free Throws Made)', 'FTA (Free Throws Attempted)', 'BLK (Blocks)', 'PTS (Points)', 'MVP (MVP status, binary value)'. Each of the models will use these data columns in their respective predictions. We explored linear regression to predict NBA MVP scores with the use of features like points, games played, minutes, and various shooting stats. We started by calculating an MVP score through a weighted formula, which added a small amount of random noise to simulate unpredictable elements in MVP selections. Then, we scaled the features to allow for consistency in the model's interpretation of the data, which is an important step given the range and variance in our dataset. Our model was cross-validated across five folds, yielding a mean R-squared score that indicated how well it could generalize to unseen data. We achieved an R-squared value and Mean Absolute Error (MAE) that revealed both the strengths and limitations of our approach. While we anticipated a degree of predictive power, the R-squared score emphasized some unexpected variability. This can show that certain intangible qualities of an MVP candidate might not be fully captured by statistical performance alone. In addition to linear regression, we implemented a Random Forest Classifier to predict the NBA MVP. We applied Random Over-Sampling to balance the target classes due to imbalance in the dataset caused by the scarcity of MVP candidates relative to non-MVP players. The features were scaled using StandardScaler to standardize the data, ensuring consistent model performance. We trained the Random Forest model with 100 estimators, leveraging balanced class weights to further address class imbalance. The dataset was split into training and testing sets with an 80-20 ratio, and the model's performance was evaluated using metrics such as accuracy, precision, recall, and a confusion matrix. This approach provided insight into the classification dynamics of MVP versus non-MVP candidates, complementing the linear regression analysis. We also applied logistic regression as a baseline classification model for predicting the likelihood of a player being an MVP. Prior to training, the features were standardized to ensure consistent scale and avoid bias. To improve model performance, we used balanced class weights to reduce the effects of class imbalance in the dataset. The model was trained and evaluated using a five-fold cross-validation scheme, with key metrics including accuracy, precision, recall, and the F1-score. This approach allowed us to assess the predictive power of logistic regression and its ability to distinguish MVP

candidates from non-candidates based solely on statistical performance. The results highlighted logistic regression's effectiveness as a transparent and computationally efficient method, while also emphasizing areas where more complex models might provide an edge.

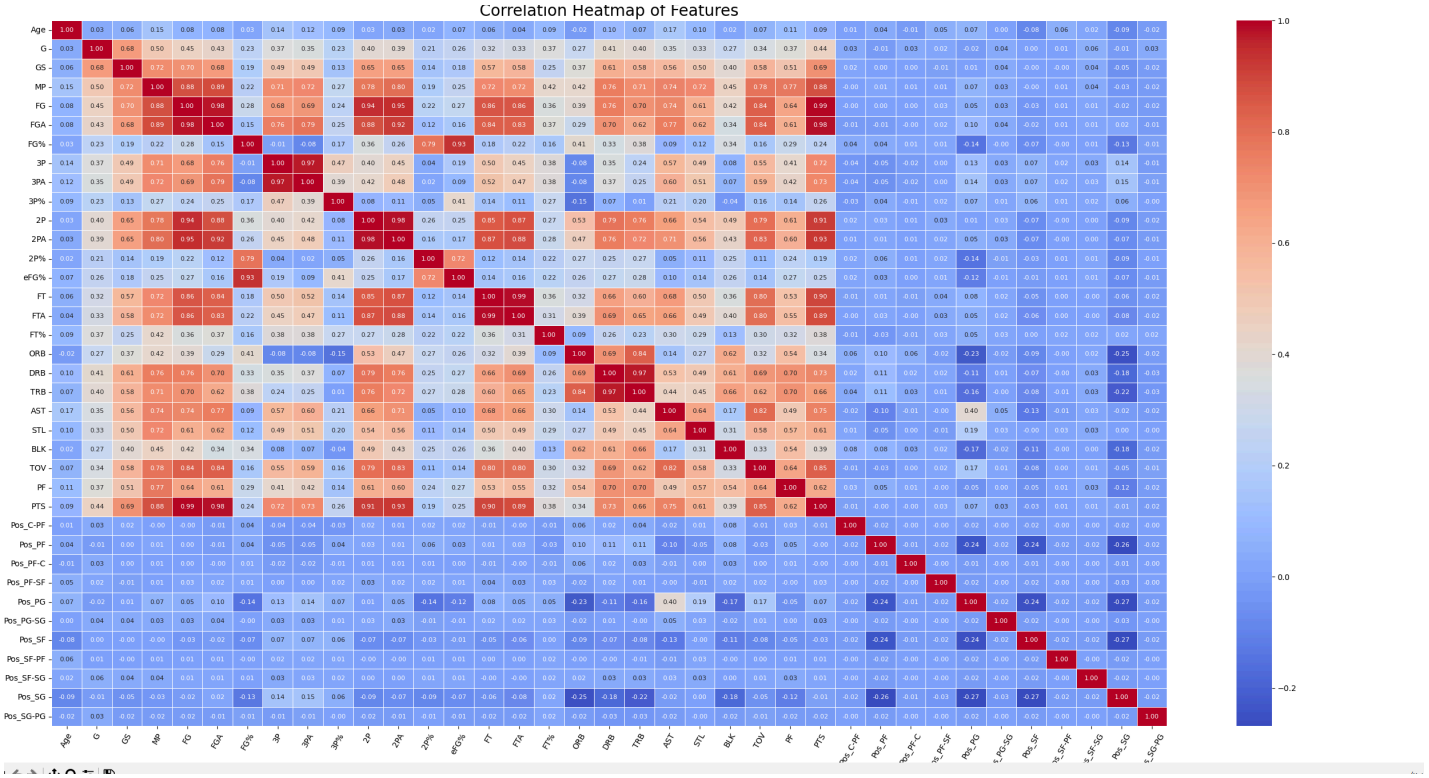
Results and Discussion

Model 1: Linear Regression

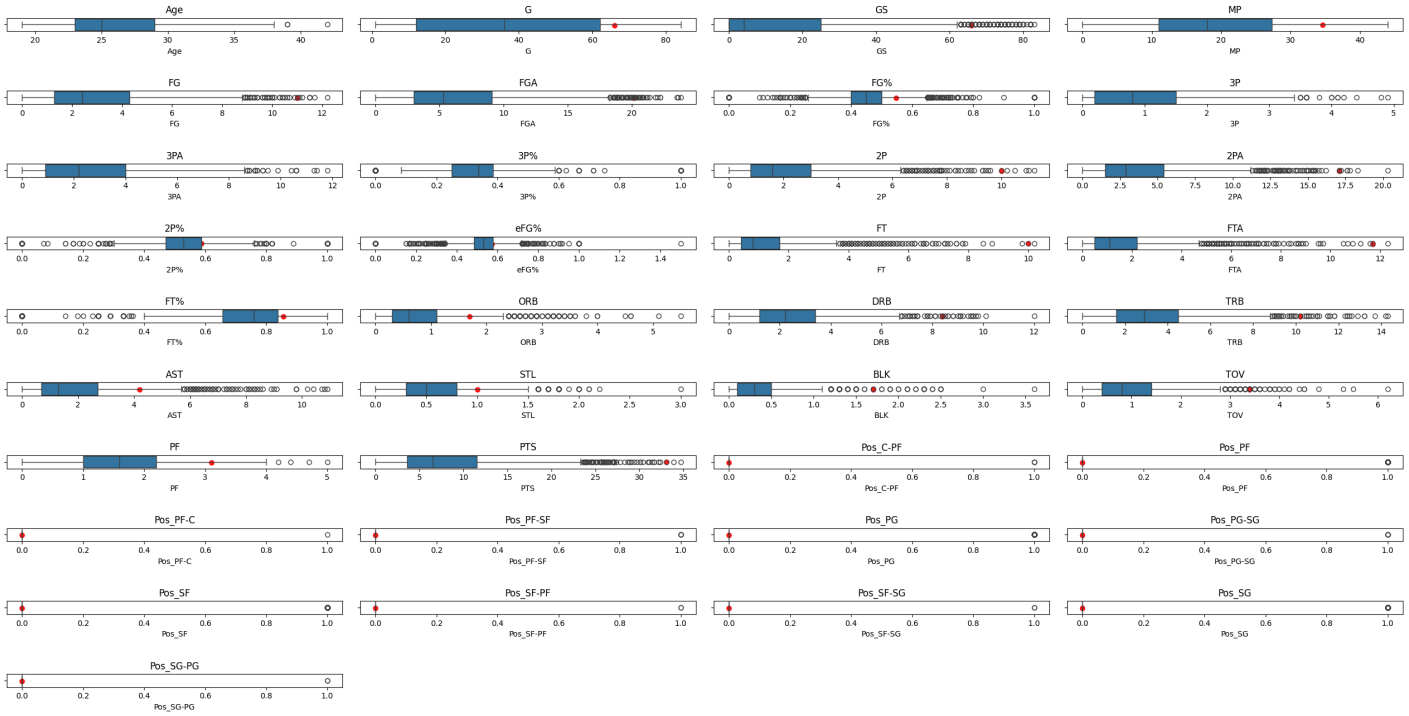
We used linear regression to predict NBA MVP scores based on features like points, games played, and shooting stats. To account for variability, we calculated MVP scores using a weighted formula with added noise and scaled all features for consistency. Cross-validation showed a high mean R-squared score, indicating the model could generalize well to unseen data. However, the R-squared also revealed variability, suggesting that some intangible MVP qualities are not fully captured by the statistics alone.

We used Matplotlib and Seaborn to create clear visualizations of our model's performance, with Seaborn adding smoother insights through KDE curves for residual patterns. Sklearn helped us evaluate the model and scale features for consistency. Using StandardScaler, we ensured each feature contributed equally to the model.

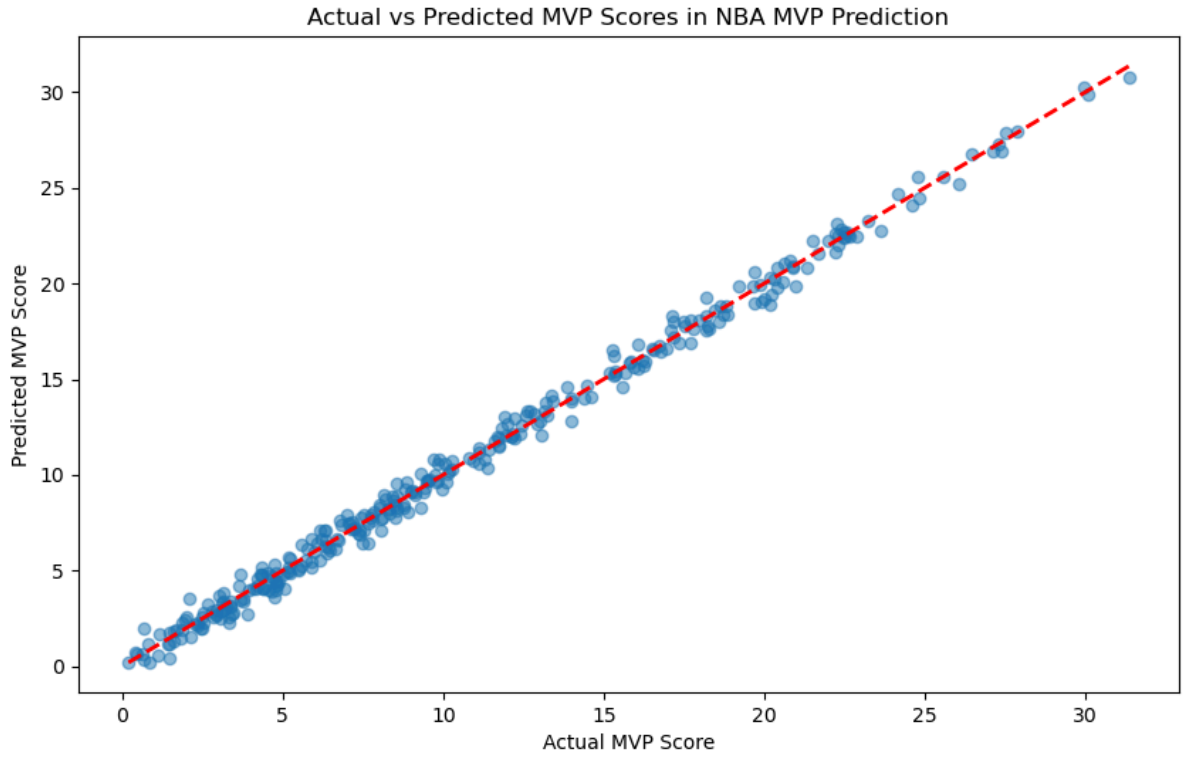
Result of the heatmaps:



Here, we found relationships such as 3 pointers and points not being as important as the average fan may think. We found that the past three MVP players are not exceptional in 3P shooting, so we did not include 3P in our features.

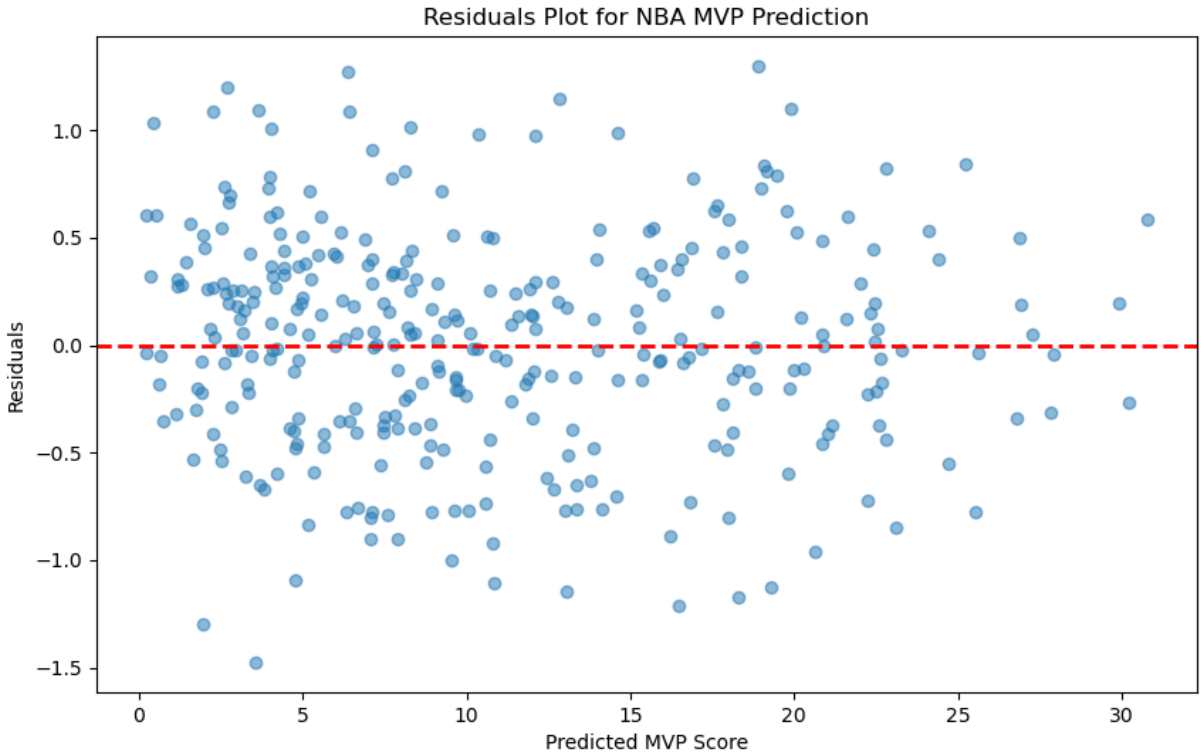


For the box and whisker plots, we highlighted the past MVP players in red to differentiate them from the other players. This displayed the statistical categories that past MVPs excel in, meaning high performance in these categories are likely to grant a MVP award.

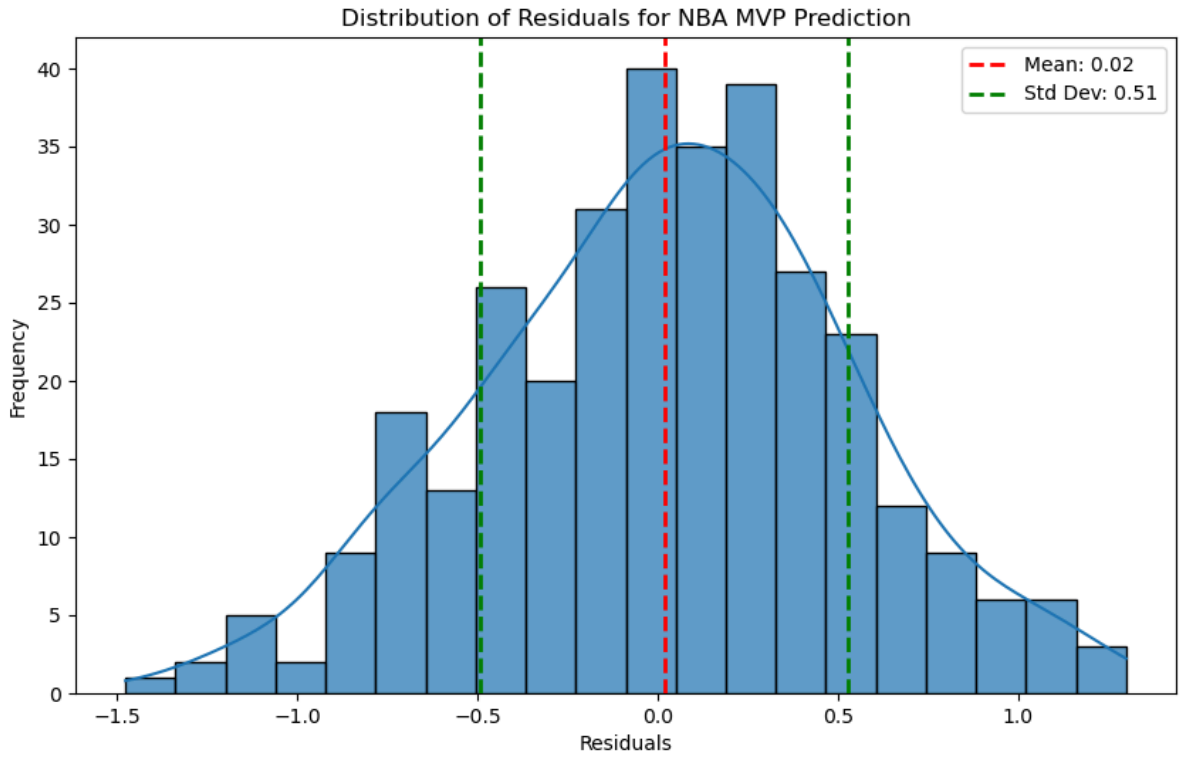


In the Actual vs. Predicted plot, we observed a linear trend, showing that the model captured a relationship between the features and MVP scores. However, some points deviated significantly,

indicating patterns of over- or underestimation. These deviations suggest possible systematic bias or unaccounted player characteristics in our model.



Our residuals plot showed that the mean and standard deviation were mostly within acceptable ranges. However, some residuals deviated noticeably, highlighting where the model struggled to fit the data accurately. This suggests limitations in capturing non-linear relationships or the stronger influence of certain features on MVP scores.



Finally, the distribution of residuals showed a near-normal pattern, which is a good indication that our model's errors were evenly distributed. However, the existence of some outliers may also indicate that there exist certain unpredictable features of MVP selection.

Additionally, the linear regression model also yields useful quantitative scoring metrics:

```
- Cross-validated R-squared scores: [0.99445718 0.99485757 0.99449042 0.99528302 0.99
- Mean Cross-validated R-squared: 0.9946708877239944
- Mean Absolute Error (MAE): 0.40619466688704325
- R-squared: 0.9948891093927336
```

Our cross-validated R-squared scores ranged from 0.994 to 0.995, showing the model explained 99.5% of the variance in MVP scores. The primary R-squared value of 0.9949 confirmed high predictive accuracy. With a Mean Absolute Error of 0.406, our model's predictions deviated by only 0.41 points on average, indicating low residual error.

Model 2: Logistic Regression

To prepare the data for the Logistic Regression model, we applied several preprocessing steps to enhance its ability to classify MVP status effectively. First, we addressed missing values by replacing them with zeros, ensuring that all players were represented, even if they had no

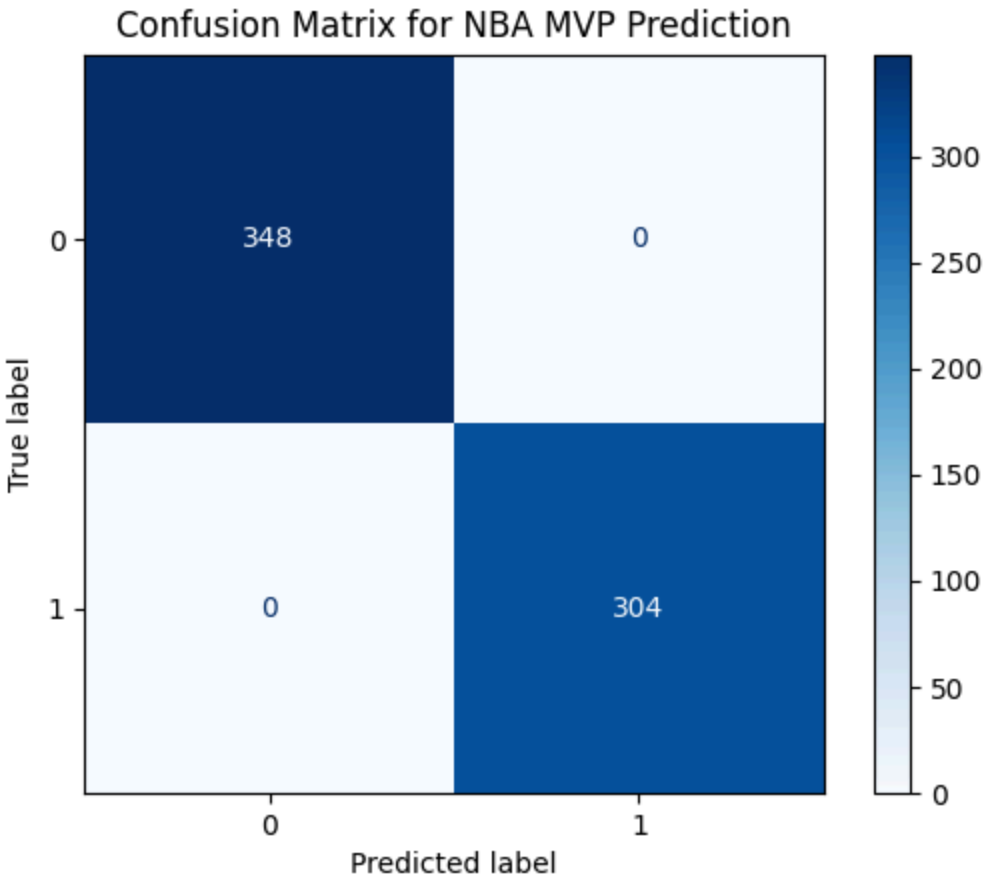
recorded statistics for a season. We also used one-hot encoding for categorical features like player position, transforming them into numerical values suitable for the model. StandardScaler was applied to normalize the dataset, allowing each feature to contribute equally to the classification process. Finally, we performed feature selection to identify the most relevant features for predicting MVP status, focusing on metrics like points, minutes played, and free throws made.

```plaintext
Top 5 Predicted MVP Players:
        Player                    Season         MVP_Prediction_Probability
401     Joel Embiid               2022-2023                      0.348372
1082    Joel Embiid               2023-2024                      0.237679
426     Shai Gilgeous-Alexander   2022-2023                      0.128840
229     Giannis Antetokounmpo     2022-2023                      0.069435
910     Giannis Antetokounmpo     2023-2024                      0.056769
```

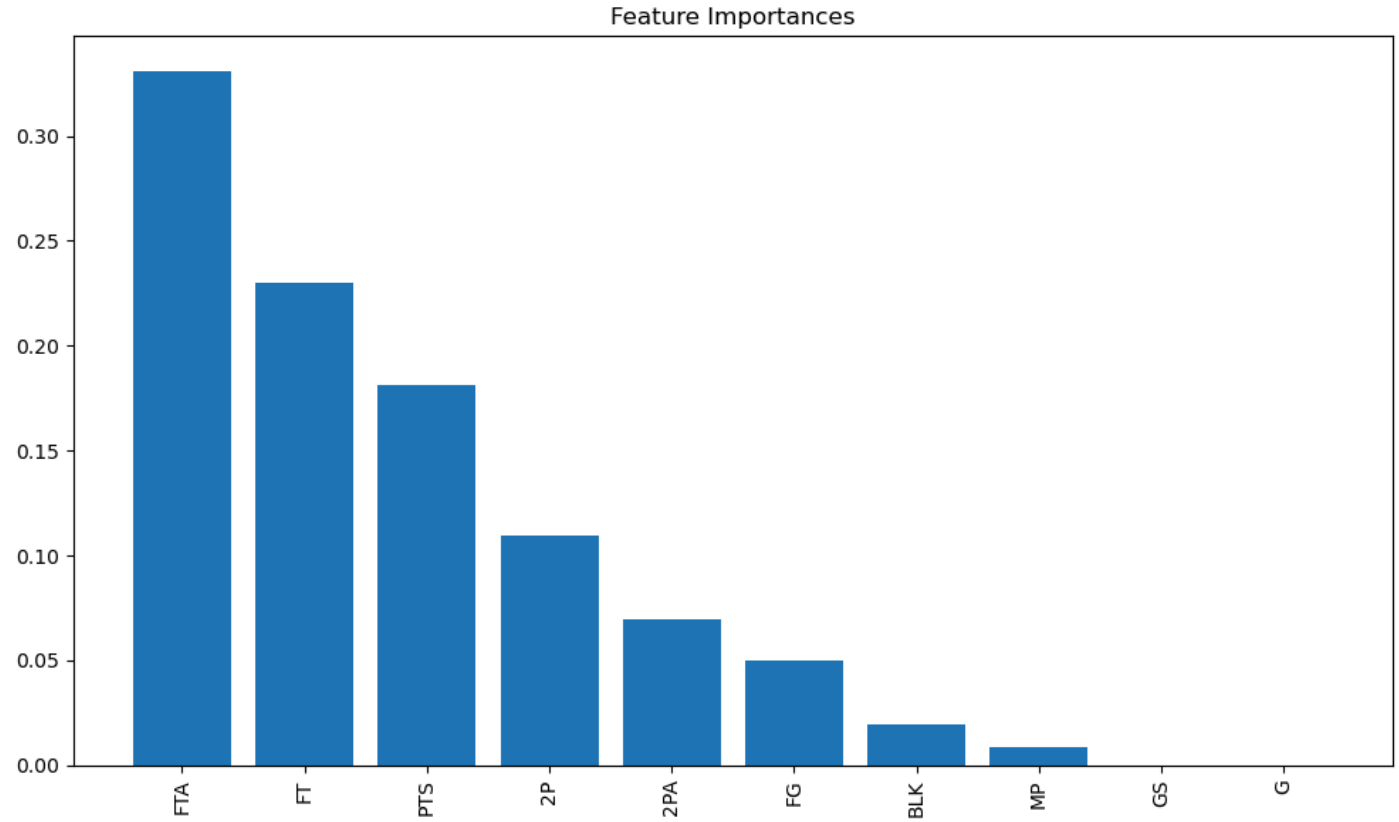
The results from our Logistic Regression model show that it can predict MVP status with reasonable probabilities. Joel Embiid was the top predicted MVP for both the 2022-2023 and 2023-2024 seasons, with probabilities of 34.8% and 23.7%, reflecting his strong performance. Other players like Shai Gilgeous-Alexander and Giannis Antetokounmpo were also recognized, but with lower probabilities, suggesting the model saw them as less likely to win. While these probabilities give us an idea of potential MVP candidates, they also show that the model gives lower confidence, likely because it can't fully account for subjective factors like leadership or the storylines that influence MVP voting.

Model 3: Random Forest

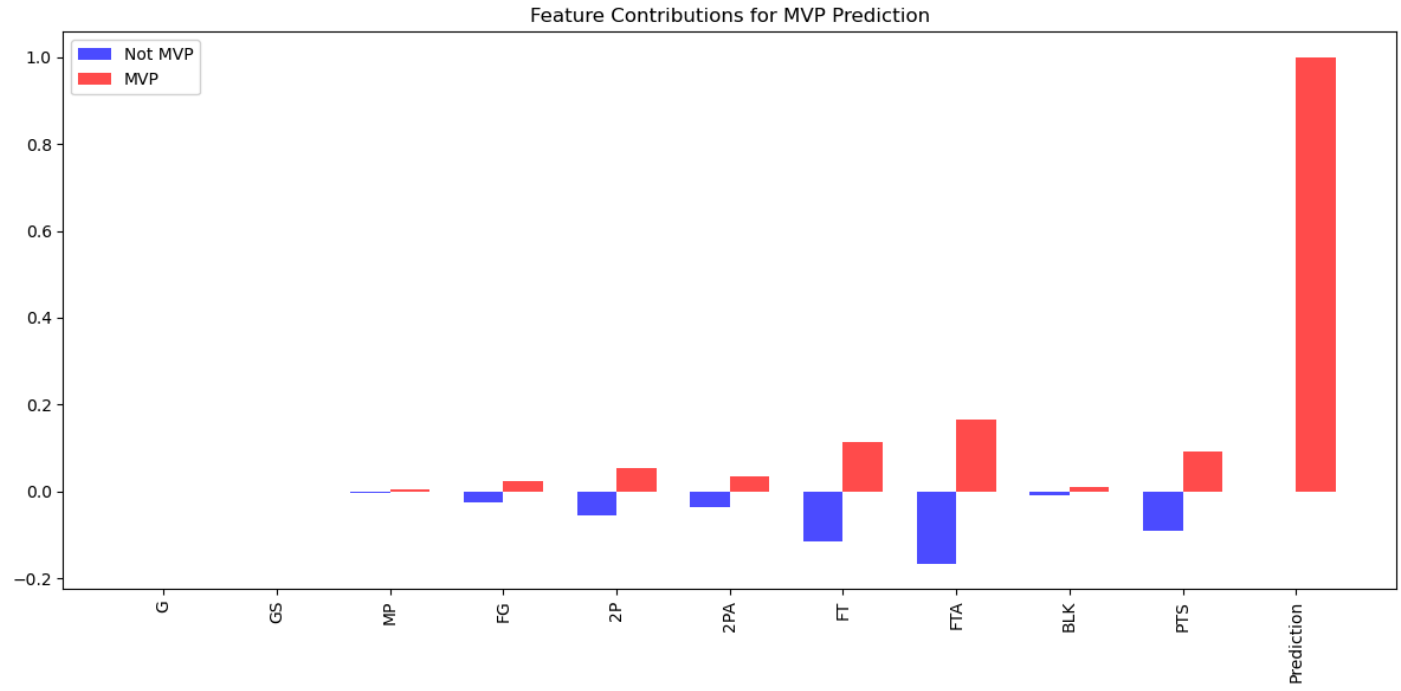
We implemented the Random Forest model to enhance classification performance for predicting MVP status. Random Forest builds multiple decision trees and combines their predictions for robust and accurate classification. We chose this model because it handles feature interactions well and is less prone to overfitting compared to single decision trees.



The confusion matrix shows that our Random Forest model perfectly classified both MVPs and non-MVPs, with no false positives or false negatives. This high accuracy is due to balancing the dataset with oversampling, ensuring equal representation of both classes. These results highlight how effective Random Forest is in distinguishing MVP players from non-MVPs.

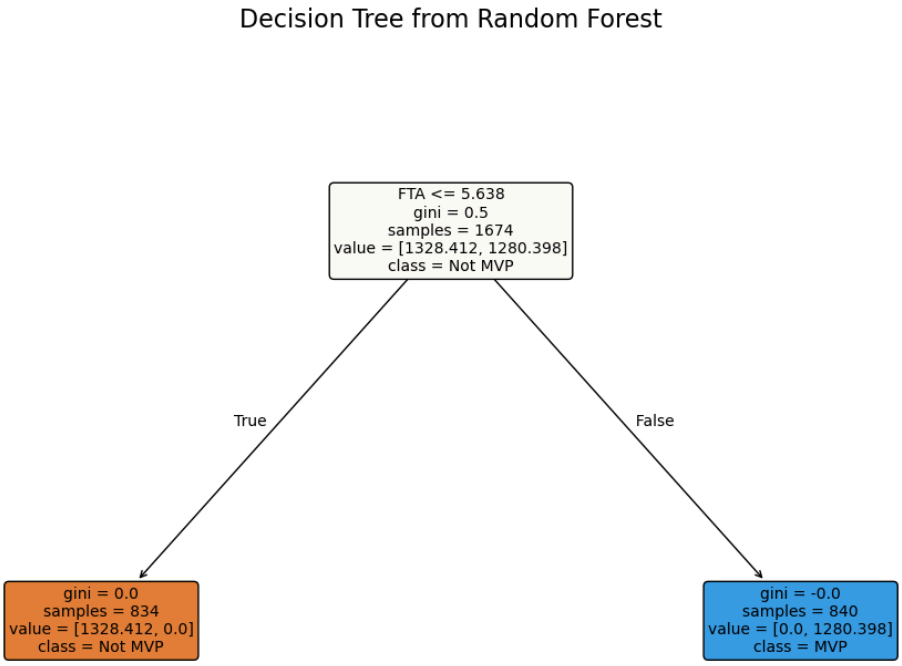


The feature importance plot shows that points scored, minutes played, and free throws made were the most influential factors in predicting MVP status. These findings align with our Linear and Logistic Regression results and confirms that these features are important for identifying MVP players.

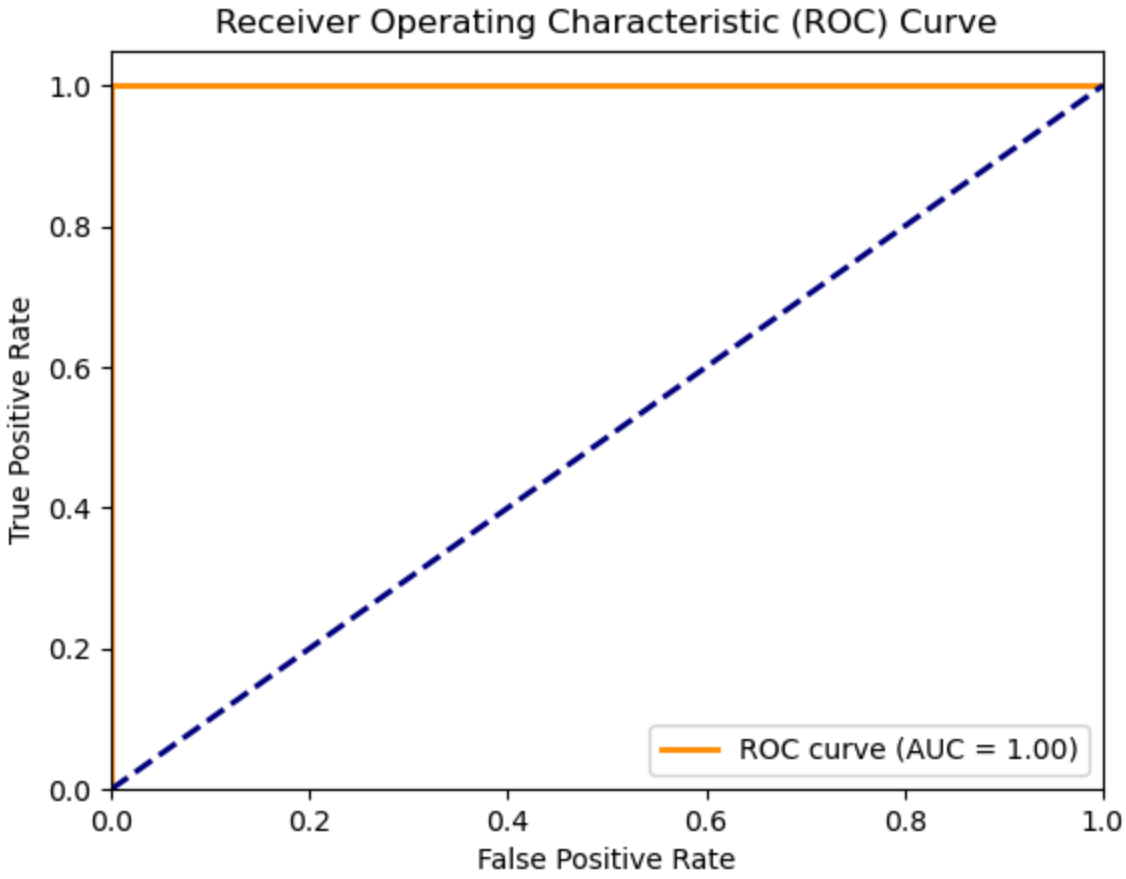


Here, the feature contribution plot explains how each feature affected the prediction for a specific player. For instance, a player with high points and minutes played sees strong positive

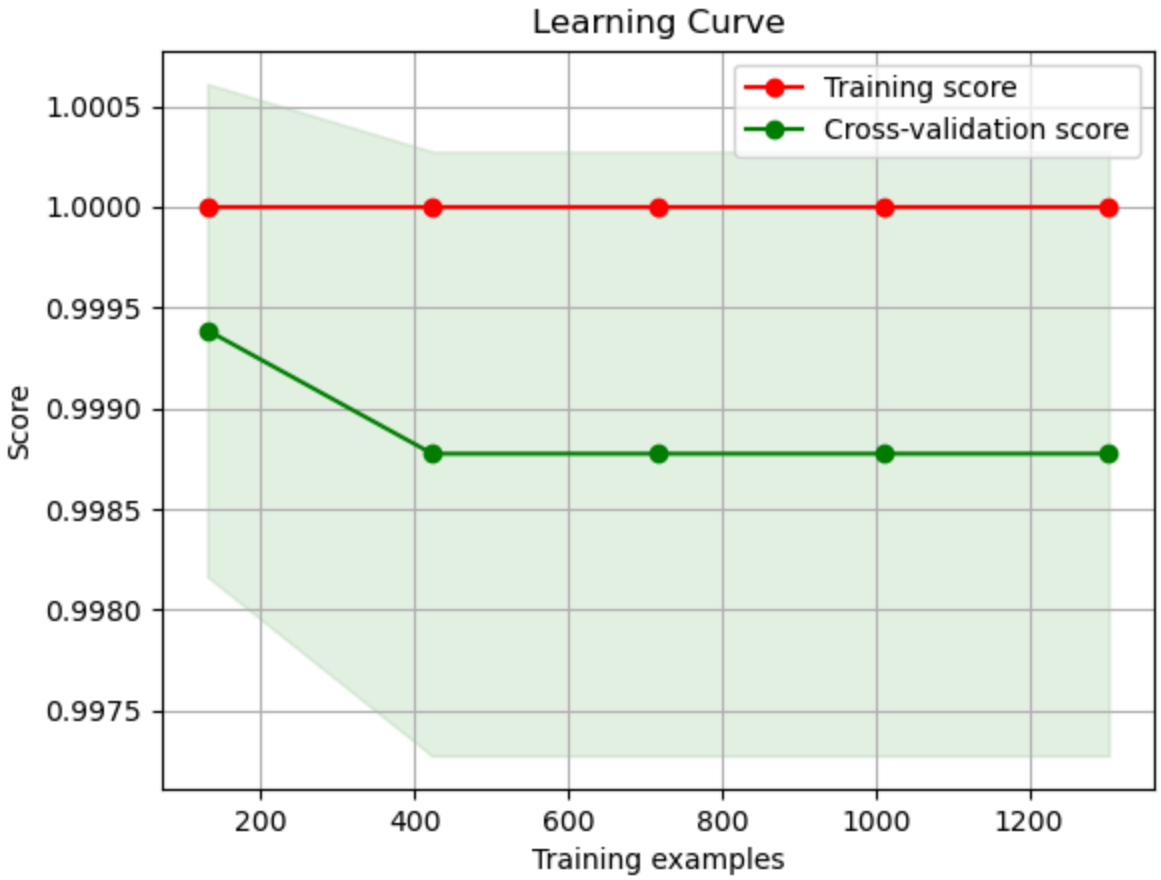
contributions from these features toward being predicted as an MVP. This allows us to understand the model's reasoning behind its decisions.



Additionally, we also created a decision path visualization to break down how one decision tree in the Random Forest made its prediction. It shows splits based on thresholds that would give us some insight into how the model evaluates players in the dataset.



The Receiver Operating Characteristic curve confirms the model's exceptional performance, with a near-perfect AUC score close to 1.0. This shows that the Random Forest model can reliably distinguish between MVPs and non-MVPs across different probability thresholds.



We also added the learning curve to show that our model generalizes well, with high training and cross-validation scores. The small gap between these scores indicates minimal overfitting, and the plateau in performance suggests that our dataset size was sufficient for training the model effectively.

To implement the Random Forest model, we started by preprocessing the data with standardization and oversampling to address class imbalance. The model was trained on this balanced dataset, and we evaluated its performance using metrics like the confusion matrix, feature importance, and ROC curve. The classification report confirmed perfect precision, recall, and F1 scores for both classes, with an overall accuracy of 1.0. These results demonstrate that our Random Forest model is highly effective and reliable for this classification task.

```
Class distribution in original data:

```plaintext
 MVP
0    1630
1       1
Name: count, dtype: int64
```
```

Class distribution before resampling:

```plaintext
 MVP
0    1630
1       1
Name: count, dtype: int64
```

Class distribution after RandomOverSampler:

```plaintext
 MVP
0    1630
1    1630
Name: count, dtype: int64
```

Classification Report:

```plaintext
                precision    recall  f1-score   support

             0       1.00      1.00      1.00       348
             1       1.00      1.00      1.00       304

      accuracy                           1.00       652
     macro avg       1.00      1.00      1.00       652
  weighted avg       1.00      1.00      1.00       652

Accuracy: 1.0
```

Model Comparison

Linear Regression provided a high R-squared value, explaining 99.5% of the variance in MVP scores. However, it struggled with capturing intangible MVP qualities, which would result in some systematic over- or underestimations. Logistic Regression, on the other hand, classified MVP candidates well and highlighted top-performing players like Joel Embiid with reasonable probability scores. Its predictions reflected a reliance on quantifiable metrics, though it lacks the addressing subjective factors like leadership and narrative, which often influence the MVP selections.

Random Forest model outperformed both Linear and Logistic Regression by achieving perfect classification accuracy after oversampling the dataset. It provided clear insights into feature

importance and individual feature contributions, showing how specific metrics like points and minutes played influenced MVP predictions significantly. Unlike the other models, Random Forest handled complex feature interactions and avoided overfitting, as shown by the consistent training and cross-validation scores that we had yielded. While its accuracy was impressive, this was partly due to the balanced dataset, which might not reflect real-world performance as we desired.

Conclusion

We found that these ML models had demonstrated how data can help predict NBA MVP outcomes. They emphasize the importance of stats like points, minutes played, and free throws but also show the limits of relying only on numbers. While the Random Forest model worked the best, our results also showed that subjective factors like leadership and player narrative play a big role in MVP decisions. This shows how difficult it is to actually predict MVPs and suggests that combining stats with expert opinions could make predictions more accurate and useful for the NBA.

Project Timeline

Timeline

Dataset link

2021-2022 Season
2022-2023 Season
2023-2024 Season

Final Contribution

Team Member	Contribution
Rayan Ahmed Shamsi	Random Forest Modeling, Data Visualization
Eric Hoang Phan	Random Forest Modeling
Ibaad Sayeed	Report
Jinseok Hwang	Report Modification, Video Presentation
Hieu Nguyen	Data Visualization, Report Creation

References

[1] “Sport analytics leverage AI and ML to improve the game,” CIO. https://www.cio.com/article/2081595/sport-analytics-leverage-ai-and-ml-to-improve-the-game.html

[2] J. Davis et al., “Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned,” Machine Learning, Jul. 2024, doi: https://doi.org/10.1007/s10994-024-06585-0

[3] G. Papageorgiou, Vangelis Sarlis, and Christos Tjortjis, “An innovative method for accurate NBA player performance forecasting and line-up optimization in daily fantasy sports,” International journal of data science and analytics, Mar. 2024, doi: https://doi.org/10.1007/s41060-024-00523-y

[4] Z. Yu, “Using machine learning to predict the NBA MVP,” Samford University, 2023. https://www.samford.edu/sports-analytics/fans/2023/Using-Machine-Learning-to-Predict-the-NBA-MVP