Final Write-up

Abstract

In recent years, the integration of machine learning techniques into sports analytics has revolutionized the way teams evaluate player performance and make strategic decisions. This report delves into the application of machine learning algorithms to predict NBA player performance and identify potential All-Star players. Leveraging a comprehensive dataset comprising player statistics from multiple seasons, we employ advanced modeling techniques and feature engineering to develop accurate predictive models. Through rigorous experimentation and evaluation, we aim to uncover patterns and insights that can inform talent evaluation and player selection strategies in professional basketball.

Introduction

In the dynamic realm of sports analytics, the integration of machine learning and data-driven methodologies has sparked a profound transformation. Particularly in basketball, where the quest for a competitive edge is relentless, the accurate prediction of player performance and the identification of emerging talent stand as pivotal challenges .

Problem Significance and Rationale

Predicting player performance and spotting emerging talent are essential for success in the highly competitive NBA. The teams are constantly refining their rosters and strategies, so that the insights we derive from the statistics become decisive in the field. This topic is pivotal in the evolving field of sports analytics, bridging innovation with tradition at a critical juncture.

Against this backdrop, our objectives are twofold: firstly, to unravel the potential of machine learning in augmenting the predictive capabilities essential for player performance assessment, and secondly, to elucidate its role in unearthing latent talent within the NBA.

By achieving these objectives, we anticipate triggering a change in how basketball organizations undertake player evaluation and talent scouting. Ultimately, our work aspires to not only elevate the competitive prowess of NBA teams but also to redefine the contours of sports analytics in the broader context.

The primary objectives of this study are as follows:

Developing predictive models to forecast NBA player performance based on historical data, identifying potential All-Star players using predictive modeling and feature engineering techniques, evaluating the effectiveness of different machine learning algorithms in predicting player performance, and investigating the impact of feature engineering and clustering on predictive accuracy and model performance.

Dataset and Features

Details about the API:

In the initial phase, we used an API library that contains statistics about basketball players from different NBA seasons. We used custom code to extract all relevant data for all NBA players starting in 1951.

In the second step, we used web scraping to extract a list of players selected as AllStars players.

In the third step, we merged the two data we created into one data frame by the player's name and the season. This allowed us to get stats on each player for each season and whether they were selected to be an All-Star that season.

Data description:

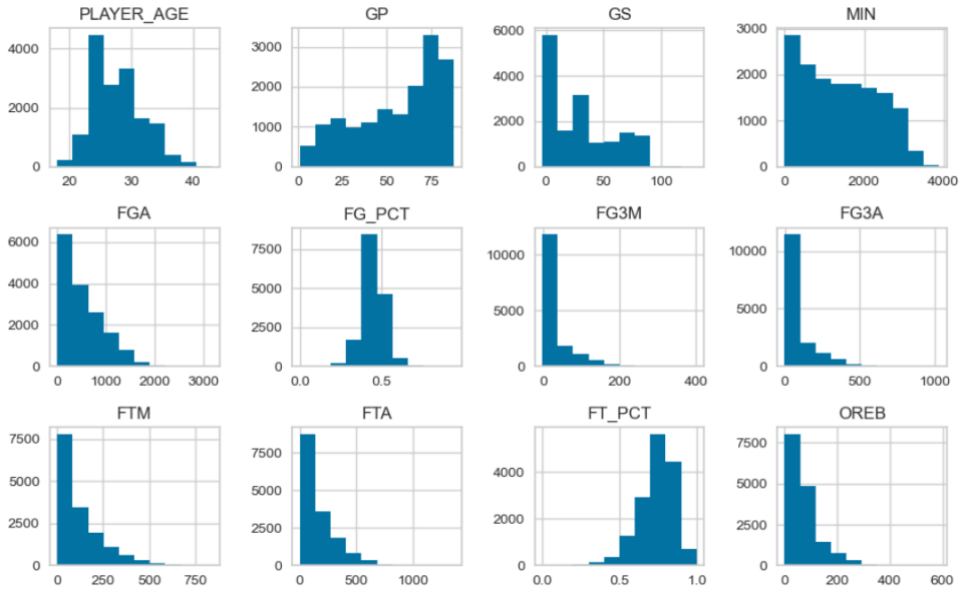
The dataset used for this study includes a comprehensive collection of player statistics spanning multiple NBA seasons. It includes various performance metrics, including points per game , assists per game , rebounds per game , field goal percentage and more.

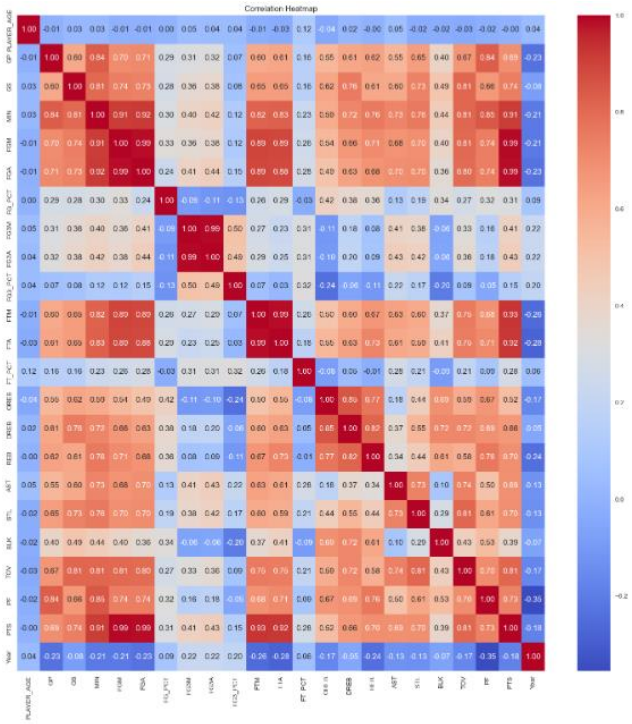
Preprocessing steps and feature selection:

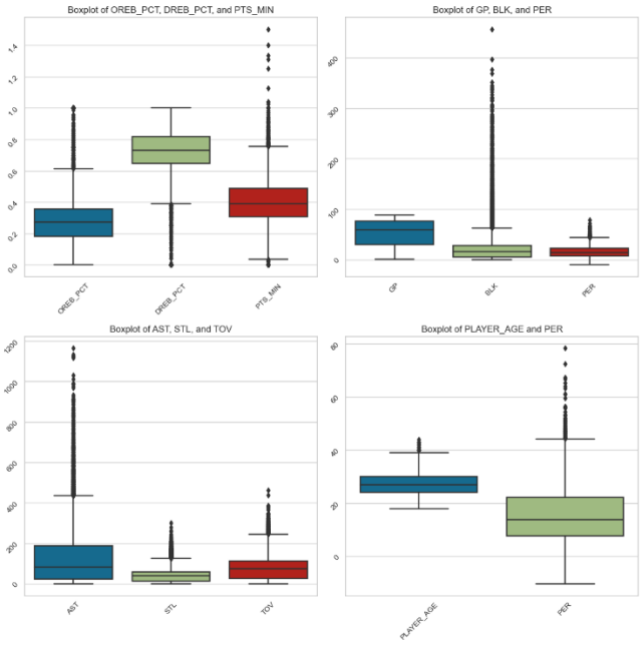
Several pre-processing steps were performed to ensure the consistency. This included data cleaning to address missing values and data normalization through rescaling. The target predictor variable represents whether a player was selected as an All-Star in a particular season (0 or 1). In this project, we assessed feature coefficients to identify the most influential variables and utilized PCA algorithms to reduce dimensionality. Additionally, we eliminated highly correlated features to mitigate multicollinearity issues and enhance model interpretability.

exploratory data analysis (EDA):

EDA serves as a fundamental step in understanding the structure of the data set and identifying potential patterns or trends. We used various statistical techniques in order to identify special patterns and correlations in the data, for example Histograms, Heatmap and Boxplot.







In continuation to the box plot, it can be observed that there are numerous outliers, therefore we will investigate them. This part will be detailed in the methodology section.

Feature engineering:

Feature engineering is essential for improving predicted performance by creating new features or modifying existing ones.

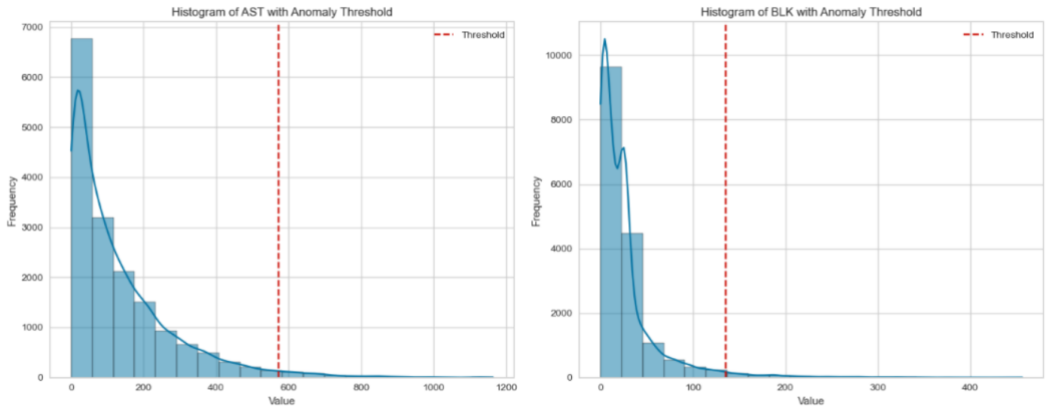
The techniques we used in this study include:

1. Polynomial Features: Generating interaction terms and polynomial features to capture nonlinear relationships.
2. Derived Metrics: Calculate metrics like Player Efficiency Rating (PER) or Points Per Minute (PTS_MIN) to enhance player performance evaluation.
3. These engineered traits are basic inputs for predictive models, contributing significantly to predictive accuracy.

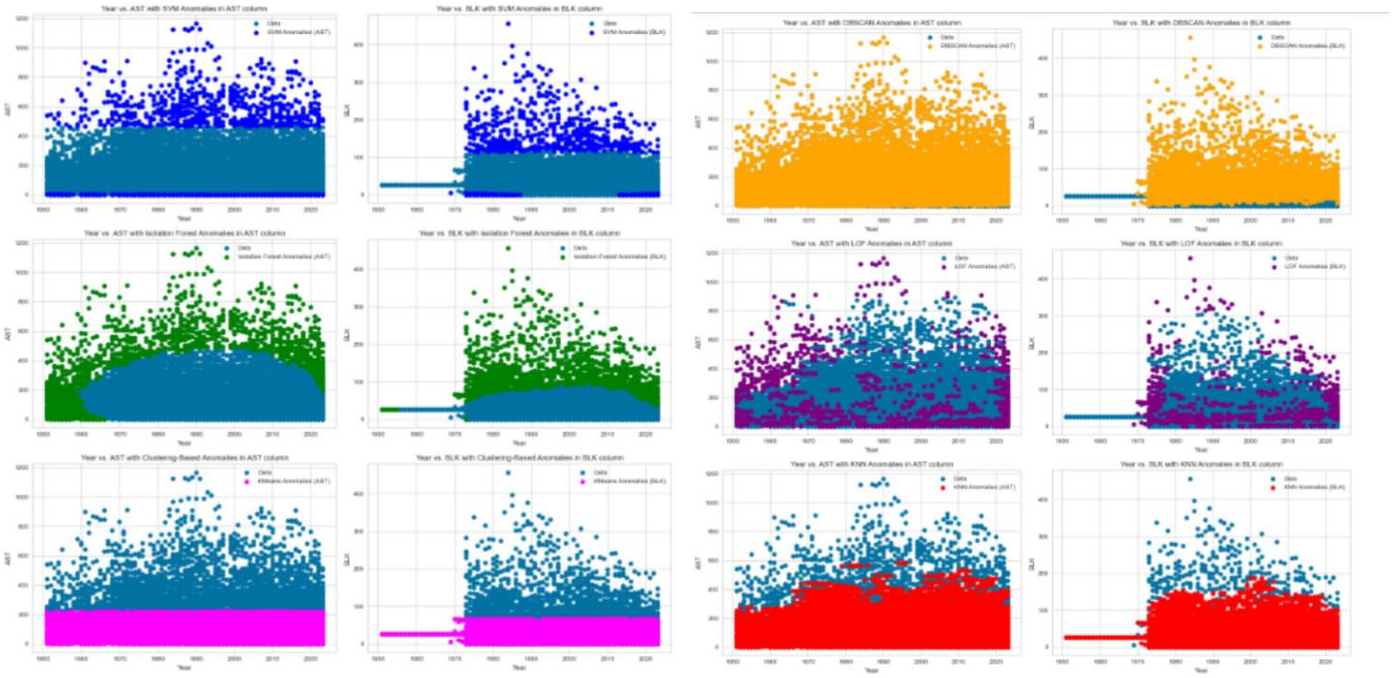
Methodology

Anomaly detection:

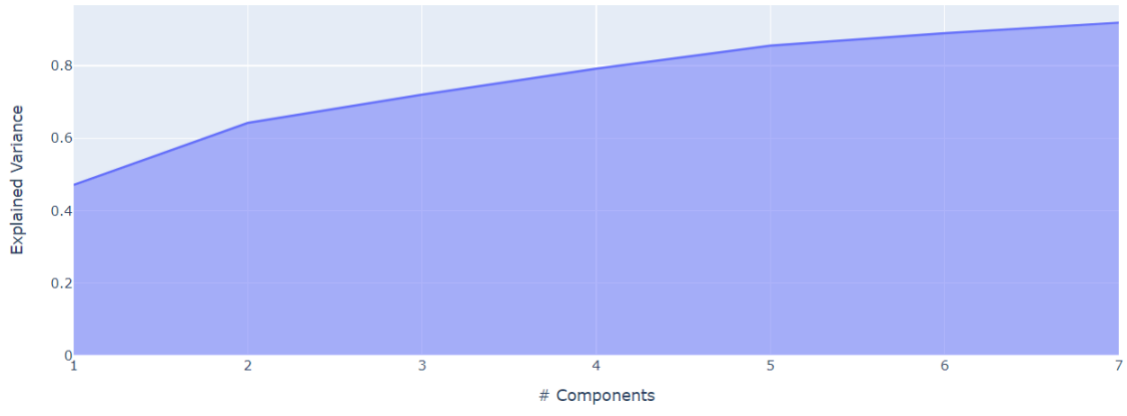
Initially, we explored the data through visualization to identify which columns have highly significant outliers. Then, we selected a threshold of 3 standard deviations and represented it on the data distribution histogram.



In the next stage, we examined the outliers using distance-based and model-based outlier detection methods. We investigated the extreme values to learn about these points and understood that these values represent the reality and not resulting from error. Therefore, they should not be disregarded and removed.



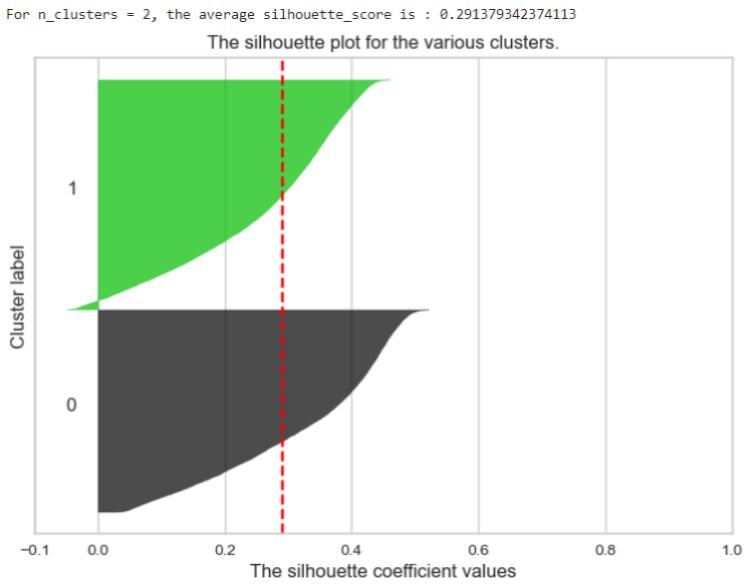
Since our data is time-dependent and constantly changing, we decided to **use time series data** to sort the data sent for training so that the training set would be past data and the test set would be future data (and not vice versa). Next, we conducted **prediction without clustering** as a reference to examine at the end of the process whether clustering and the various algorithms we tried did indeed contribute to the model's prediction. We implemented various **models: logistic regression, random forest, XGBoost, gradient boosting, AdaBoost, LGBM, including evaluation for each**. For each model, we selected the **optimal parameters** based on our data (After adjusting the hyperparameters, the model exhibits an overall enhancement in accuracy. Nonetheless, due to the unbalanced nature of the data, this improvement leads to an increase in incorrect predictions of All-Star players, except within the logistic regression model. Consequently, this adjustment does not contribute positively to our prediction task. Therefore, we have opted to retain the default values recommended by the model, with the exception of logistic regression, where we will apply the optimal parameter we have identified.). We conducted a PCA process to reduce dimensions, aiming to utilize only the relevant



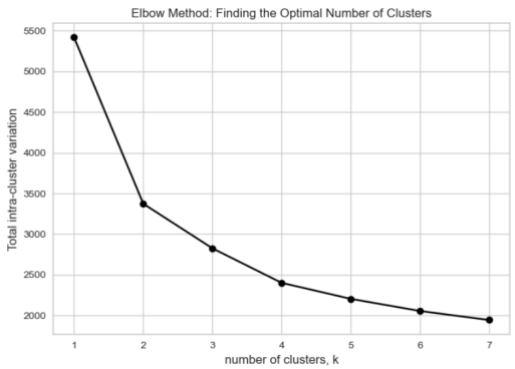
dimensions when performing clustering with KMEANS. After testing, we found that 91% of the data could be explained using just 7 features! With these seven features, we proceeded with the clustering process for our models.

We proceeded with the clustering process, aiming to choose the optimal N. We utilized two algorithms taught in class:

1. Silhouette - measuring the distance between data points within clusters and the separation between different clusters.

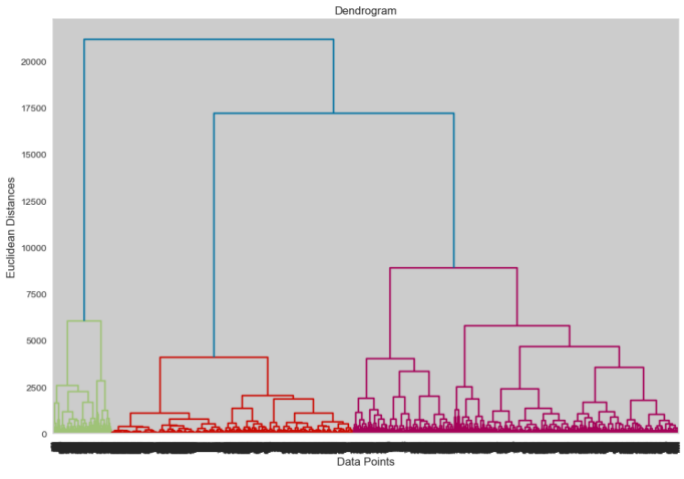


2. Elbow method - identifying the point where the rate of decrease in distortion sharply changes.



In both of these methods, it was learned that the optimal N according to our data is 2. Hence, we proceeded with clustering using two different methods:

1. Kmeans
2. Hierarchical cluster



The next step was to apply the models we had previously run with the clusters and verify the evaluation metrics and feature weights to determine if clustering had influenced the model predictions.

Experiments/Results/Discussion

During the features selection process, we evaluated various normalization methods for the data and opted for the min-max approach. Initially, we removed features with high correlation to avoid multicollinearity. Subsequently, we applied PCA (Principal Component Analysis) to reduce dimensionality, resulting in a dataset with 7 columns, termed as PCA_df.

For evaluation metrics, we employed a range of measures including accuracy, precision, recall, and F1-score. These metrics were chosen because they provide a comprehensive assessment of model performance, particularly in a binary classification problem like predicting All-Star selections. Accuracy gives an overall view of correct predictions, while precision and recall focus on the model's ability to correctly identify positive cases (All-Stars) and avoid false positives. The F1-score, which combines precision and recall, provides a balanced measure of a model's performance.

In presenting our results, we provided both quantitative and qualitative analyses to offer a comprehensive understanding. Quantitatively, we presented classification reports for each model, detailing metrics such as precision, recall, and F1-score for both classes (All-Star and non-All-Star players). Additionally, we included feature importance analyses to identify the most influential features in each model's predictions. Qualitatively, we obtained results that were unexpected. It appears that using clustering for prediction did not contribute to the model.

For Kmeans: Including the clustering feature did not significantly affect the performance of the AdaBoost, Gradient Boosting, XGBoost, and LightGBM models. In the logistic regression model, the cluster feature had a minor effect, slightly improving the model's ability to identify All-Star players. However, in the Random Forest model, there was a decrease in the accuracy of the All-Star predictions.

For Hierarchical cluster: Even after hierarchical clustering was performed, the same trends were evident: adding clustering features didn't significantly affect the performance of the AdaBoost, Gradient Boosting, XGBoost, and LightGBM models. In the logistic regression model, the cluster feature had a minor impact, slightly improving the model's ability to identify All-Star players. However, in the Random Forest model, there was a decrease in the accuracy of All-Star predictions.

The performance of the algorithms we executed was not satisfactory. Additionally, the use of the Cluster column was not efficient, and in many cases, the model decided to zero out the coefficient of this column and not give it importance. Therefore, we do not recommend using this algorithm. We estimate that the data is distributed in a way that is not suitable for analysis using clustering, and overall, clustering did not contribute to prediction.

Conclusion and Future Work

Our conclusion is that using a clustering algorithm to predict All-Star players does not help considering the data we have collected. Our dataset consists of many outlier data points that cannot be removed as they represent genuine talent outliers among players. We couldn't

improve the model to achieve satisfactory accuracy results, so it's not advisable to rely heavily on this model. In retrospect, we realized that there's no effective way to analyze All-Star players with the existing dataset. Therefore, we recommend exploring other directions such as selecting players from colleges for the NBA (draft).

Contributions

Both of us collaborated fully, so even when working separately, there was constant communication and consultation at every stage regarding the results and the continuation of work on the project (this allowed us to pick up from the same point where the other left off, with full awareness at every point in the project). Liron was responsible for extracting statistical data using API+CRAWLING, investigating the data, including outlier analysis. Maayan, after obtaining the data, merged the statistical data with another table obtained through crawling, which held the prediction variable (whether a player was an All-Star in a specific season), thus obtaining unified data for further work. After Liron's outlier analysis, she ensured to find the optimal N using learned algorithms, and to conduct clusters to improve the model. The execution of the models throughout the project (before and after the changes) was done in full collaboration, as well as drawing conclusions for each part of the project.

Appendices

Columns Legend:

1. PLAYER_ID: This could be an identifier for each player.
2. SEASON_ID: Indicates the season of the data.
3. LEAGUE_ID: Represents the league in which the player is playing (e.g., NBA, WNBA).
4. TEAM_ID: Identifier for the team the player is associated with.
5. TEAM_ABBREVIATION: Abbreviation for the team name.
6. PLAYER_AGE: Age of the player.
7. GP: Games played.
8. GS: Games started.
9. MIN: Minutes played.
10. FGM: Field goals made.
11. FGA: Field goals attempted.
12. FG_PCT: Field goal percentage.
13. FG3M: Three-point field goals made.
14. FG3A: Three-point field goals attempted.
15. FG3_PCT: Three-point field goal percentage.
16. FTM: Free throws made.
17. FTA: Free throws attempted.
18. FT_PCT: Free throw percentage.
19. OREB: Offensive rebounds.
20. DREB: Defensive rebounds.
21. REB: Total rebounds.
22. AST: Assists.
23. STL: Steals.
24. BLK: Blocks.
25. TOV: Turnovers.
26. PF: Personal fouls.
27. PTS: Points scored.
28. full_name: Full name of the player.
29. Year: Year of the data.
30. All_Star: Indicates whether the player was an All-Star that season.