Outlier Detection for Identifying Outstanding Players in the NBA

Riya Gori
rvgori@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Cole Sanders
cgsande2@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Manideepika Reddy
mmyaka@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

ACM Reference Format:
Riya Gori, Cole Sanders, and Manideepika Reddy. 2024. Outlier Detection for Identifying Outstanding Players in the NBA. In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND PROBLEM STATEMENT

In today's era of data-driven decision-making in sports, leveraging advanced analytics to identify outstanding players through outlier detection can revolutionize team strategies. Our project plans to explore this field. This project proposes to utilize unsupervised learning techniques such as clustering and anomaly detection to identify outstanding players in NBA history. We also plan to answer the question of what factors contribute to making outstanding players. The novelty of aspects of this project are in the methods we plan to use and the hypothesis we plan to explore. Unsupervised learning is not a topic that has been covered in the homework, and we plan to implement several models for outlier detection including synchornization-based, clustering-based, distance-based, and univariate based. Then we plan to answer the non-trivial question of which model performs better for detecting outliers in sport-based contexts.

1.1 Related Work

(1) Synchronization Based Outlier Detection
https://link.springer.com/chapter/10.1007/978-3-642-15939-8_16
We have summarized the paper's understanding into the following: The paper "Synchronization Based Outlier Detection" presents a unique approach to outlier detection, which is based on the concept of synchronization. The authors propose a method where each data object is regarded as a phase oscillator, and its dynamical behavior over time is simulated according to an extensive Kuramoto model. In the context of this paper, the oscillators are the data objects, and their phase is determined by their respective data values. The interaction between these oscillators is governed by the Kuramoto model, which leads them towards a state of synchronization. During this process, the authors observed that regular objects and outliers exhibit different interaction patterns. Regular objects tend to synchronize with each other, while outliers remain out of sync. This difference in behavior allows for the natural detection of outliers. The measure used to detect these outliers is the local synchronization factor (LSF). The LSF is a measure of the degree of synchronization of a particular oscillator with its neighbors. Outliers, which do not synchronize well with their neighbors, will have a high LSF. The authors conducted an extensive experimental evaluation on both synthetic and real-world data to demonstrate the benefits of their method. They found that their synchronization-based approach was able to effectively detect outliers in various datasets. This approach to outlier detection has potential applications in a wide range of fields. For example, it could be used to detect criminal activities, analyze athlete performance, or identify rare events or exceptions.

. In conclusion, the paper presents a novel and powerful concept for outlier detection. By leveraging the principles of synchronization and the Kuramoto model, the authors were able to develop a method that can naturally flag outliers in complex datasets.

(2) Anomaly and Event Detection for Unsupervised Athlete Performance Data
https://ceur-ws.org/Vol-1458/E27_CRC63_Odonoghue.pdf
This paper proposes a new methodology for anomaly detection within a dataset. It involves combinations of algorithms to accurately identify outliers.
The dataset under consideration comes from GPS trackers monitoring the athletic performance of Gaelic football players. Due to the contact nature of the sport, there is a concern that some of the GPS units may be damaged during the game and provide incorrect data. The team of researchers plan to use anomaly detection to identify any faulty trackers in this dataset.
Using boundary detection, univariate outlier detection, principal component transformation, and principal component classification, the researchers identified anomalies and classified them to be single outliers or a marked shift in the pattern of data.
In their results they found that using a combination of boundary detection, principal component transformation, and principal component classification was the most efficient method of identifying anomalies. Univariate outlier detection did not make a significant difference in the number of anomalies detected when the above methods were already employed.

(3) Prediction-based outlier detection with explanations
https://ieeexplore.ieee.org/document/6468672
The paper proposed a domain independent unsupervised

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

approach to detect outliers in a heterogenous dataset. It presented a model that does not only consider how one instance differ from others externally but also considers the internal dependency and abnormality among it's own attributes. This model uses neighbors(Numerical/Categorical/Social based) as basis to predict a certain attribute. To implement this, LOESS Regression (Local linear regression method) is used that builds model for each data point based on it's neighboring datapoints. This method has an advantage of dealing with heterogenous data. The model also implemented an explanation framework which is capable of producing natural language explanations of each outlier. It provides explanation on the spot with constant complexity thereby aiding in advanced verification. And the evaluation of algorithm is done by considering four metrics.

(4) Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering
https://www.hrpub.org/download/20170830/WJCAT2-13708454.pdf
This paper focuses on identifying outliers in K-Means clustering clusters. The goal for the paper is to remove outliers to achieve better clusters; however, for our purposes, we will focus on their methods to identify the outliers. The first is a distance based algorithm. It takes the maximum and minimum pairwise distance for each point in the dataset calculated using Euclidean distance. A threshold value is set and if the distances are above the threshold it is classified as an outlier and if it is below the threshold it is classified as an inlier. A second clustering based approach is proposed where k-means clustering is run with a relatively large value of K and the smallest cluster is identified. This cluster is removed as an outlier and the process is repeated until accuracy improves. The paper found that distance based outlier detection and removal increased the accuracy of their K-means clustering algorithm by 1 percent and clustering based detection and removal increased it by 6 percent. This indicates that clustering based outlier detection is the stronger method.

(5) A Probabilistic Transformation of Distance-Based Outliers
https://arxiv.org/abs/2305.09446
This paper covers numerous methods for distance based outlier detection and attempts to assign a probabilistic score to each outlier prediction. This way observations are not determined to be either outliers or not, but rather each point is given a predicted probability of whether it is an outlier or not. This is done to show a clearer gap between outliers and non outliers as sometimes it can be difficult to distinguish between the two. The weighted k nearest neighbor outlier detection algorithm was picked for the experiment and probabilistic scores were generated and normalized using distance and outlier scores as a metrics. They found that doing so was effective in creating a clearer contrast between outlier and non-outlier data points and emphasized the importance of tuning normalization metrics to maximize this contrast. They also show that this type of probabilistic output and normalization can be applied on other distance based outlier detection algorithms to find similar improvements in outlier detection.

2 METHOD

2.1 Approach

In this project, we employed machine learning techniques to analyze an NBA dataset, aiming to discover insightful patterns and anomalies in player performance from the research papers analysed by us:

(1) Distance-Based Outlier Detection: Adopting the approach from the fifth paper, we utilized the concept of Euclidean distance where for each data object (in our case, player performance metrics) distance from mean is calculated and considered as an outlier if it is more than a certain threshold value
(2) Anomaly Detection using univariate: Influenced by the second paper, we incorporated a method that normalizes data based on data per minute and specifies a threshold value based on mean and standard deviation.
(3) LOESS-Based Outlier Detection: The third paper inspired us to implement a LOESS-based outlier detection method that examines not just the external attributes but also the internal dependencies within a player's statistical record. In our implementation, the LOESS algorithm was pivotal in smoothing and predicting data points based on localized subsets of the dataset. By integrating both numerical and categorical data into our outlier detection model, we enhance our ability to uncover deep, actionable insights across diverse player metrics. After applying LOESS smoothing, we calculate the residuals, which are the absolute differences between the actual points scored and the LOESS-predicted points. A threshold for identifying outliers is set at two standard deviations above the mean residual. Players whose performance metrics significantly deviate from their predicted values are marked as outliers.
(4) K means clustering: Normalize the data and reduce the dimension into 2D using PCA. Then find out the optimal number of points using elbow point and silhouette score. Cluster the points using K means and apply gaussian function to detect outliers.

2.2 Rationale

The chosen techniques for this analysis— are LOESS Regression based outlier detection, anomaly detection using PCA and clustering, Univariate based outlier detection and Distance-based outlier detection—are particularly suited for the NBA dataset due to its complex, multidimensional nature and the specific challenges it presents. Here's how these methods align well with the nature of the dataset and the analytical objectives, compared to other possible approaches

(1) Anomaly Detection with PCA and Clustering: The NBA dataset is high-dimensional, with numerous statistics recorded for players, making it challenging to discern patterns or anomalies directly. PCA reduces this dimensionality, distilling the data into principal components that capture the most variance and, thus, the most significant patterns. Following PCA, clustering (e.g., k-means) groups players with similar performance profiles, facilitating the detection of outliers

as those who do not fit well into any cluster. This method contrasts with univariate analyses that might overlook complex, multivariate relationships or with naive clustering that doesn't first reduce dimensionality, potentially getting muddled by irrelevant variance.

(2) LOESS-Based Outlier Detection: The NBA dataset, being heterogenous with numerous interconnected performance metrics, requires a nuanced approach to outlier detection that considers these interdependencies. LOESS-based detection, especially using LOESS regression, assesses how well an individual player's statistics conform to patterns inferred from their peers. This method is particularly suitable for datasets where inter-attribute relationships are critical to understanding context and performance nuances, offering a more refined analysis than simpler, rule-based systems that might miss subtleties or inter-variable influences. Compared to other possible approaches, such as straightforward thresholding or basic statistical anomaly detection, these chosen methods provide a deeper, more nuanced understanding of the data.

(3) Univariate-based outlier detection: It analyzes individual variables independently to identify extreme values, making it particularly useful for simple data structures and specific contexts where the influence of other variables is negligible. This method is advantageous due to its simplicity, as it involves basic statistical calculations like mean and standard deviation, and is computationally efficient, which is ideal for large datasets and preliminary outlier analysis to quickly spot extreme values. For instance, it can be crucial for detecting unusually high temperatures in climate data. However, its limitations include providing limited insight as it only assesses the distribution of individual variables and fails to detect outliers apparent only in a multivariate context, such as anomalous combinations of variables. Additionally, its simplicity might lead to oversimplification, potentially missing critical outliers or incorrectly flagging too many data points as outliers if variables vary in scale or distribution.

(4) Distance-based outlier detection: It identifies outliers by evaluating the distances between data points in a dataset, typically using metrics like Euclidean distance. This method is particularly effective in multivariate data where assessing relationships between multiple variables is crucial for outlier identification. It offers flexibility as the threshold for what constitutes an outlier can be adjusted according to different data types and contexts, and it is robust in identifying outliers that differ significantly in their feature space, not just as extreme values. However, this method faces challenges such as computational complexity, as calculating distances between all pairs of points can be resource-intensive, especially in large datasets. Additionally, its performance may suffer in non-uniformly distributed data, where sparse regions might wrongly appear as clusters of outliers, and it is also prone to the "curse of dimensionality," where the effectiveness decreases as the number of dimensions increases since points in high-dimensional spaces tend to become equidistant from each other.

3 EXPERIMENT

3.1 Dataset

It comes from a comprehensive NBA database, which includes various statistics and information about NBA players up until 2005. Here's a detailed breakdown of the dataset:

Players Table (3573 entries): This table provides a detailed list of players, including both physical attributes and career descriptions up to the year 2005. The physical attributes likely include height, weight, and position, while the career descriptions could encompass years active, teams played for, and perhaps college or country of origin.

Player regular season Table (19113 entries): This table contains season-by-season statistics for each player during the regular NBA season. The stats are detailed and could include points per game, assists, rebounds, steals, blocks, shooting percentages, and more, offering a comprehensive view of a player's performance in each season.

Player regular season career Table (3760 entries): This table aggregates each player's total statistics over their entire regular-season career up until 2005. It provides a cumulative view of a player's performance metrics, consolidating their regular season achievements into a single record per player.

Player playoffs Table (7544 entries): Similar to the regular season stats, this table provides detailed statistics for each player in the playoffs. These statistics are crucial for analyzing a player's performance under the heightened pressure and competitiveness of the postseason.

Player playoff career Table (2056 entries): This table summarizes each player's career statistics in the playoffs, giving an aggregate view of their postseason contributions. It's an essential resource for evaluating a player's performance and impact during the most critical part of the NBA season.

Player allstar Table (1463 entries): This table focuses on the performances of players in All-Star games, showcasing the stats of those who have participated in this prestigious annual event. It provides insights into the elite level of play and recognition achieved by the top players in the league. Each table in this dataset can be analyzed independently or combined to extract comprehensive insights, such as longitudinal performance analysis, comparisons between regular season and playoffs, career progressions, and the impact of physical attributes on performance

3.2 Hypotheses

We predict that LOESS-based outlier detection will be the most effective at finding overperforming players, followed by clustering-based, distance-based, and finally univariate-based approaches.

3.3 Experimental Design

In our experiment, we have implemented different kinds of methods and compared the effectiveness of different methods for detecting outliers among NBA players. Using our NBA dataset, we employed four different methods on the data to gather our results: a distance-based, a clustering-based, a univariate-based approach, and LOESS Regression. For evaluating our models, we assumed that outstanding players were those who were picked to go to the All-Star game.

We then compared the detected outliers to the All-Star data set to compute metrics such as accuracy, precision, recall, and F1 score.

(1) Distance-Based Outlier Detection Implementation:
Data- The player_regular_season_career and player_playoffs_career files are concatenated into one larger dataset.
Preprocessing- All numerical stats are normalized from [0,1] for each player.
Method- The means of all normalized numerical stats are calculated. The Euclidean distance from each players' normalized stats to the normalized mean vector is then calculated for all players. Players in the top 97% or above of distance from the mean are determined to be outliers.

(2) Clustering-Based Outlier Detection Implementation:
Data- The players dataset is joined to the player_regular_season dataset based on player identifiers.
Preprocessing- The applicable features in the new dataset are normalized from [0,1]. PCA is performed with two principal components.
Method- K means clustering is performed with five clusters. Additionally, a Gaussian multivariate probability density function is used on the distribution over the principal components, and any player objects mapping to an area of the distribution where the probability is less than .05 is determined to be an outlier.

(3) Univariate-Based Outlier Detection Implementation:
Data- This method uses the player_regular_season_career dataset.
Method- It calculates outliers based on three expert-chosen stats: rebounds per minute, points per minute, and turnovers per minute. If any player is over one and a half standard deviations from the mean in any of these stats across their career, they are classified as an outlier.

(4) LOESS Regression-Based Outlier Detection Implementation:
Data- The players dataset is joined to the player_regular_season dataset based on player identifiers.
Method- The model uses a domain independent unsupervised approach to detect outliers in a heterogenous dataset. It considers how one instance differ from others using neighbors as basis to predict a certain attribute. LOESS Regression (Local linear regression method) is implemented such that it builds model for each data point based on it's neighboring datapoints. The neighbors are chosen by integrating two types of neighborhood functions
   (a) Numerical based: Identifies objects with similar numerical attribute values. Examples: Points, Rebounds, Assists, Blocks and Steals
   (b) Categorical based: Identifies objects with identical categorical attribute values. Example: position, division

   We have also implemented an explanation framework which is capable of producing natural language explanations of each outlier. It provides explanation on the spot with constant complexity thereby aiding in advanced verification.
   The Prediction based approach, works as follows:
   (a) Assume we have n data points $\{P_1, P_2, \ldots, P_m\}$
   (b) The goal is to learn whether $P_i = \{P_{i1}, P_{i2}, \ldots, P_{im}\}$ is prediction-based outlier or not.

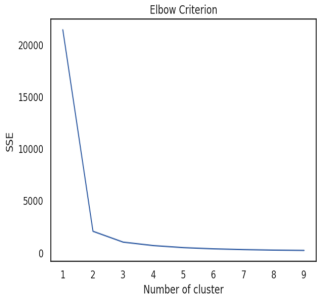

Figure 1: Elbow Criterion

   (c) Let P_{ik} denote the k^{th} attribute of P_i.
   (d) For this, we need to construct a neighborhood function $f(P_i)$ that returns a set of s neighborhood instances N.
   (e) We then use these N instances as training samples and construct m Regression models $R_1, R_2, \ldots, R_m$ each trying to predict one attribute using the rest of $m-1$ attributes.
   (f) Let R_k be used to predict P_{ik} value.
   (g) If the predicted value is not equal to the true value then it can be marked as an outlier.

   We have used Prediction based outlier factor to detct the outliers in the given data
   Prediction-Based Outlier Factor (PBOF): PBOF for the P_{ik} attribute is calculated as:

$$PBOF(P_{ik}) = \frac{|(\hat{y}_{ik} - y_{ik})|}{\sqrt{WMSE}}$$

   where: - $\hat{y}_{ik}$ is the predicted value of the k^{th} attribute of instance P_i. - y_{ik} is the true value of the k^{th} attribute of instance P_i. - $\sqrt{WMSE}$ is the square root of the Weighted Mean Squared Error (WMSE) of the training in regression, which captures the unpredictability of the neighborhood. - $|(\hat{y}_{ik} - y_{ik})|$ captures the unpredictability of point P_i with respect to attribute k.

4 RESULTS

We have implemented K means Clustering, Distance based outlier detection, Univariate based outlier detection and LOESS based outlier detection and evaluated these metrics. For K means clustering, we have used the sum of squarred distances and plotted a graph to determine the elbow criterion. The elbow point gives us the optimal number of clusters needed for our dataset to cluster them effectively. In the figure there is a steep curve at point 2 indicating that 2 clusters are needed for our dataset.

We have also implemented Silhouette score to verify that the optimal number of clusters is 2. The silhouette score for 2 clusters is higher than the others. So, we can tell that the optimal number of clusters is indeed 2.

We then used 2 clusters and implemented K means clustering method to cluster our data points. As our final step, we have applied Gaussian distribution that is fitted into 2 dimensional dataset to find the outliers in the dataset. It is done in such a way that, we have pre selected a specific threshold value epsilon=0.05. The data points which fall below this threshold value are identified as outliers.

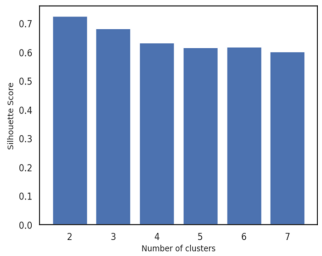

Figure 2: Silhouette score

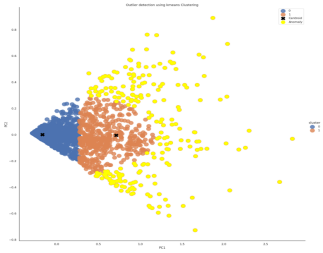

Figure 3: Kmeans clustering based outlier detection method

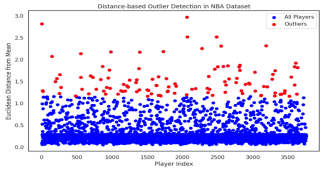

Figure 4: Distance based Outlier detection

Outliers based on distance:

	ilkid	distance
1	ABDULKA01	2.822399
11	ADAMSAL01	1.294483
17	AGUIRMA01	1.280511
142	BARKLCH01	2.079898
189	BAYLOEL01	1.502867
...	...	...
3592	WILKELE01	1.426989
3595	WILKIDO01	1.840616
3609	WILLIBU01	1.922448
3627	WILLIHE01	1.216836
3636	WILLIKE01	1.820884

Figure 5: Distance values of the points that are considered as outliers

From the figure, we can see that the data points depicted by the yellow dots represent outliers

As our second approach, we have applied Distance based outlier detection method which calculates the Euclidean distance from the mean. If the distance exceeds the pre specified threshold (which is 97th percentile of distances) then it is considered as an outlier.

As we can see from the figure, the mean is around 1.2. So, the points that are below 1.2 are considered as normal points (Blue dots) and the ones that are exceeded are considered as outliers (Red dots).

Our next approach is based on Univariate based outlier detection. In this approach the players whose normalized values (Points per minute, Rebounds per minute, Turnover per minute) is more than one and a half standard deviations away are flagged as outliers.

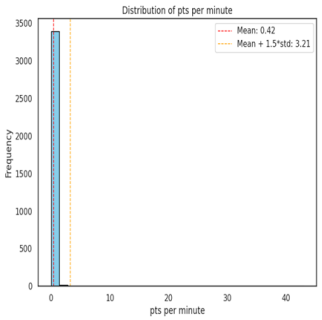

Figure 6: Outliers based on points per minute

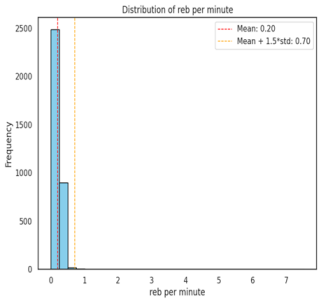

Figure 7: Outliers based on rebounds per minute

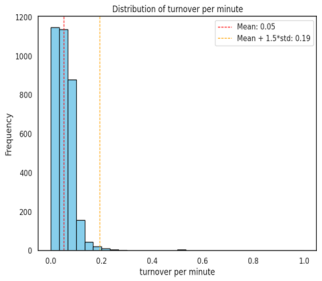

Figure 8: Outliers based on turnover per minute

Based on the figures, we can say that Univariate model failed to detect outliers efficiently.

LOESS Regression is selected as our final approach because of it's advantage of considering Categorical attributes along with Numerical attributes. LOESS calculates the smoothed values for target variables based on predicted values and categorical attributes. It calculates the residual errors between the original and smoothed values and identifies outliers based on a threshold (set at 2 times the standard deviation of the residuals). For a detailed explanation as to why a certain point is considered as an outlier, we made use of natural language explanation.

For each method, we have calculated the Confusion matrix and found Accuracy, Precision, Recall and F1 Scores for each method.

From the Results we can argue that

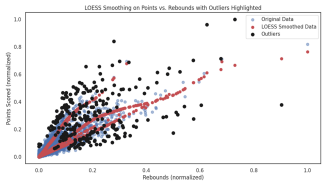

Figure 9: LOESS Regression based outlier detection method

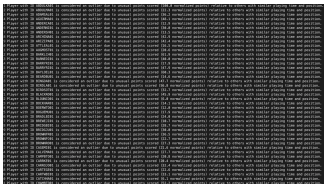

Figure 10: Enter Caption

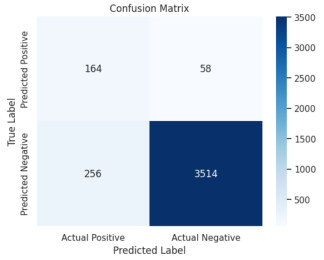

Figure 11: Confusion matrix for K means

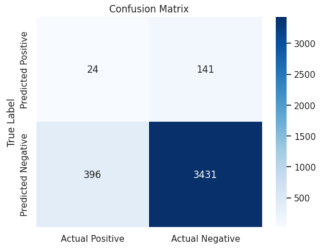

Figure 12: Confusion matrix for Univariate based method

(1) High accuracy (kMeans and Distance-based models at 0.92 and LOESS regression at 0.91): Indicates that these models are generally effective at correctly identifying both (outliers) and typical players (non-outliers).
(2) High precision (Distance-based model at 0.88): Means that when this model predicts a player to be an outlier, it is very likely to be correct. There are few false positives, i.e., regular players wrongly identified as outliers.
(3) Low precision (Univariate-based model at 0.14): Implies that many players this model identifies as outliers are actually not, leading to many "false alarms."
(4) Higher recall (kMeans model at 0.39): Indicates that this model is relatively better at detecting most of the (true outliers) from the dataset.

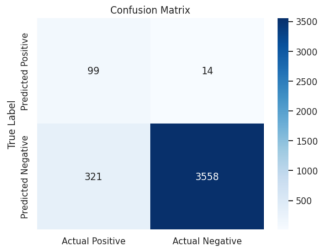

Figure 13: Confusion matrix for distance based method

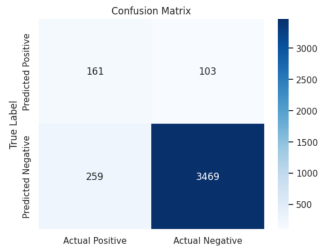

Figure 14: Confusion matrix for LOESS Regression

Model	Accuracy	Precision	Recall	F1 Score
K means	0.92	0.73	0.39	0.51
Distance based	0.92	0.88	0.24	0.37
Univariate based	0.86	0.14	0.057	0.082
LOESS Regression	0.91	0.61	0.38	0.47

Table 1: Table depicting Accuracy, Precision, Recall and F1 Score for every model

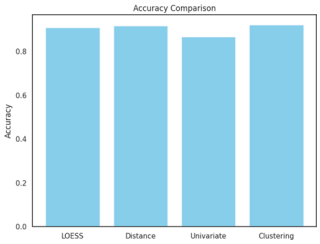

Figure 15: Graph depicting Accuracy

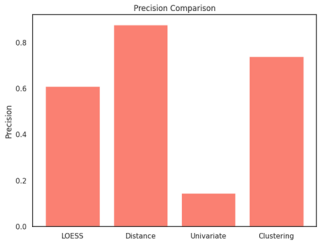

Figure 16: Graph depicting precision

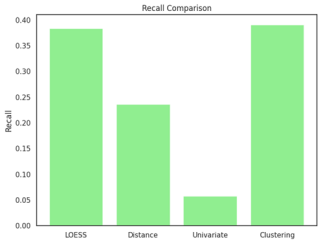

Figure 17: Graph depicting Recall

(5) Balanced F1 Score (kMeans at 0.51): Shows a good balance between not missing true outliers and not falsely identifying regular players as outliers. This is important in the NBA context, as teams would want to recognize true talent without overestimating player capabilities.
(6) Low F1 Score (Univariate-based model at 0.082): Indicates the model is neither good at identifying true outliers nor at avoiding misclassification of regular players as outliers.

5 DISCUSSION

All four models we explored had relatively high accuracy values ranging from .86 to .92; however, this does not necessarily mean they performed well for outstanding performance detection. The dataset was heavily imbalanced with many more standard or substandard players compared to outstanding players. Based on the F1 score, precision, and recall values, kMeans was the best overall model with the best balance between all three scores, followed by LOESS regression, then distance-based, and finally univariate-based.

It is important to note the high precision score of the distance-based model as it significantly outperformed the other models in this area with a score of .88. In fact, precision scores for all models was nearly double that of their recall scores. This suggests that these models would perform well in scenarios that discourage false positives. The high precision scores relative to the recall scores suggest that there are either inadequacies in the data we picked or the methods themselves which lead to a large number of outstanding players to be passed over and falsely labeled as negatives. Potential future work should evaluate aggregating the results from the methods with high precision (kMeans and distance-based) to see if their combined results detect more outstanding players, leaving less players classified as false negatives and hopefully not significantly increasing the number of false positives due to the high precision scores.

Our hypothesis was not entirely supported as LOESS regression was outperformed by the clustering-based approach. However, the ordinal effectiveness of the remaining models does follow our hypothesis. It is interesting to note that LOESS regression used both numerical and categorical attributes of the players while the remaining three algorithms were based on the numerical attributes alone. LOESS being outperformed by kMeans suggests that categorical data was not significantly relevant in the NBA dataset. It further suggests that NBA player performance is best modeled as groups of similar-performing players and outliers are determined based on their proximity to these groups rather than their proximity to other distance-based measures such as statistical averages across the full set of players.

Our research agrees with some of the previous findings in our related works section. Firstly, univariate outlier detection is a weaker predictor and should be used as a baseline model only. Secondly, the only consistently identified outlier found in all relevant papers that used an NBA dataset was Jason Kidd and he was picked up by all models except for univariate-based outlier detection.

6 CONCLUSION

Conclusion In our comprehensive study on outlier detection for identifying outstanding players in the NBA, we implemented and evaluated four distinct statistical and machine learning methods: K-means clustering, distance-based outlier detection, univariate-based outlier detection, and LOESS regression. Each of these methods was applied to a rich dataset consisting of various NBA player performance metrics up until 2005.

Our findings reveal that while each method has its strengths, K-means clustering and distance-based outlier detection proved to be the most effective in distinguishing the truly exceptional players from the rest. K-means clustering, with an F1 score of 0.51, demonstrated a superior balance in precision and recall, indicating its robustness in identifying true outliers without overwhelming false positives. On the other hand, the distance-based method excelled in precision (0.88), underscoring its ability to accurately flag outliers when they are detected, albeit at a lower recall rate.

Contrarily, the univariate-based method exhibited limitations in both precision and recall, suggesting its lesser suitability for complex datasets like those of NBA players where multiple performance indicators need to be considered simultaneously. LOESS regression, while not outperforming K-means or distance-based methods, still showed commendable accuracy and was particularly adept at integrating both numerical and categorical data, providing a nuanced view of player performance.

The success of K-means in our study aligns with the hypothesis that clustering, coupled with PCA for dimensionality reduction, effectively highlights outliers in high-dimensional data. This approach not only aids in the clear delineation of performance metrics but also enhances the detectability of exceptional cases by grouping similar performances and isolating those that are significantly different.

The lesson we learned from our work was that it is relatively difficult to accurately identify all over-performing players based on the data alone. However, due to our high precision scores, it is possible to identify a subgroup of players who are outstanding with high certainty. This would apply to scenarios where coaches are recruiting players for their team. A false positive outstanding player in this case would be much more costly than a false negative given the naturally large pool players. Thus it is useful for them to identify a true positive with high certainty. If we had more time we would experiment with aggregating the results of our better-performing approaches. As mentioned in the discussion, we believe the right aggregation method would decrease the number of false negatives significantly without a significant change in the number of false positives. The goal of this would be to accurately identify all outstanding players instead of subsets of the group.

Overall, our research contributes to the sports analytics field by enhancing the understanding of how different outlier detection methodologies can be strategically utilized to pinpoint excellence in sports data. This not only aids in team management and scouting but also enriches the tactics employed during games. Future studies may explore integrating these methods with real-time data and expanding them to other sports for broader applicability.

7 REFERENCES

(1) Shao, J., Böhm, C., Yang, Q., Plant, C. (2010). Synchronization Based Outlier Detection. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science(), vol 6323. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15939-8$_1$6

(2) Donoghue, Jim, et al. Anomaly and Event Detection for Unsupervised Athlete Performance, ceur-ws.org/Vol-1458/E27$_C$*RC*63$_O$*donoghue.pdf*. *Accessed*14*Apr*.2024.

(3) Chen, Liang-Chieh, et al. "Prediction-based outlier detection with explanations." 2012 IEEE International Conference on Granular Computing, Aug. 2012, https://doi.org/10.1109/grc.2012.6468672.

(4) Barai (Deb), Anwesha, and Lopamudra Dey. "Outlier detection and removal algorithm in K-means and hierarchical clustering." World Journal of Computer Application and Technology, vol. 5, no. 2, May 2017, pp. 24–29, https://doi.org/10.13189/wjcat.2017.050202.

(5) Muhr, David, et al. "A Probabilistic Transformation of Distance-Based Outliers." arXiv.Org, 18 July 2023, arxiv.org/abs/2305.09446.

(6) https://github.com/jason-huynh83/NBA-cluster/blob/master/NBA_cluster.ipynb

8 MEETING ATTENDANCE

Date: 2/29/2024
Attendance: Riya, Manideepika, Cole
Date: 3/22/2024
Attendance: Riya, Manideepika, Cole
Date: 3/30/2024
Attendance: Riya, Manideepika, Cole
Date: 4/5/2024
Attendance: Riya, Manideepika, Cole
Date: 4/13/2024
Attendance: Riya, Manideepika, Cole
Date: 4/14/2024
Attendance: Riya, Manideepika, Cole