Machine Learning for NBA All-Star Game selection

Final project - Introduction to Machine Learning

Tiago Carvalho
up201906163

Henrique Castro
up202007917

Daniel Lopes
up202006301

***Abstract*—This project was conducted as part of the "Aprendizagem Computacional" course in the first year of the Master's degree in Electrical and Computer Engineering at FEUP. The main objective is to develop a machine learning model to predict NBA All-Star selections based on players' performance statistics. The dataset comprises various player metrics, including points, rebounds, and assists. Initial data exploration and preprocessing encompass handling missing values, converting formats, and grouping statistics by year. Exploratory data analysis provides valuable insights into trends and patterns in player performance over the years. A correlation analysis is performed to identify significant features for predicting All-Star selections. The ultimate goal of the project is to deliver a robust model capable of identifying potential All-Star players based on their performance metrics.**

I. Introduction

In the ever-evolving landscape of professional sports, leveraging data and cutting-edge technologies has become paramount for gaining a competitive edge. This project, conducted within the scope of the "Aprendizagem Computacional" course during the first year of the Master's degree in Electrical and Computer Engineering at FEUP, delves into the intersection of basketball, data science, and machine learning.

The focal point of this endeavor is the prestigious National Basketball Association (NBA) All-Star Game, where the league's top talents converge to showcase their skills annually. The objective is ambitious yet compelling: to develop a machine learning model capable of predicting NBA All-Star selections based on players' performance statistics.

Our dataset is a comprehensive collection of player metrics, encompassing points scored, rebounds, assists, three-pointers made, and more, laying the foundation for our analytical journey. Through varied data exploration and preprocessing techniques, we meticulously examine the data, addressing issues such as missing values, format conversions, and grouping statistics by year. One of the key points of our dataset is that we only process positive data, i.e. we only have the stats of the players that have been selected to the All-Star Game and we don't have any information about the ones that weren't selected.

Beyond the technical intricacies, this project embarks on an insightful journey of exploratory data analysis, unraveling the intricate trends and patterns woven into player performance over the years. The correlation analysis not only identifies key features influencing All-Star selections but also sheds light on the evolving dynamics of the sport.

In a broader context, this project emphasizes the importance of incorporating data-driven methodologies in molding contemporary sports landscapes. The fusion of basketball performance and data science not only deepens our insights into All-Star selections but also illustrates the transformative influence of technology in redefining excellence in professional sports. As we work towards constructing a reliable predictive model, our objective extends beyond forecasting All-Star selections. We seek to actively contribute to the ongoing dialogue surrounding the evolving role of data in shaping the trajectory of sports in the future.

II. Dataset Analysis

In this chapter, we dive deep into the NBA All-Star dataset, aiming to uncover valuable insights into the factors shaping player selections. Our exploration begins with a close look at raw metrics, spanning essential performance indicators such as points, rebounds, and assists. The dataset acts as a rich source of information, reflecting the nuanced journey of player statistics over the years. We navigate through the data with a focus on thorough treatment, addressing missing values, converting formats, and organizing statistics by year. Throughout this analytical journey, we trace the evolution of critical metrics, examine the team's impact on All-Star selections, and highlight players with notable selections.

The dataset used within the scope of this project can be found in Kaggle and contains the information about all the NBA players select to the All-Star Game since 1980 to 2022. The stats presented to each players are:

- First and last name;
- Team in the year of selection
- Year of selection
- Games and minutes played
- Field gold made, attempted and percentage
- Three points shots made, attempted and percentage
- Free throws made, attempted and percentage
- Rebounds, offensive and defensive
- Assists
- Turnovers
- Steals
- Blocks
- Personal fouls

A. Game Evolution

If we now dive into the evolution of NBA gameplay over the years, specifically focusing on points per game (PPG) and three-point shooting, we can draw significant conclusions. The landscape of basketball has witnessed a remarkable transformation, marked by substantial changes in playing styles, strategies, and the overarching significance of offense in the game.

In its historical context, the NBA was characterized by players who relied heavily on mid-range jumpers and close-range shots, with a substantial portion of scoring concentrated in the paint. However, as the game unfolded, a paradigm shift took place. The incorporation of three-point shooting emerged as a pivotal aspect, reshaping the dynamics of offensive strategies.

Points per game act as a metric reflecting the league's offensive output, demonstrating not only individual players' scoring contributions but also the overall pace and intensity of the game. As illustrated in Figure 1, there is a clear increase in PPG, signaling a transition toward a more fast-paced and high-scoring style of basketball.

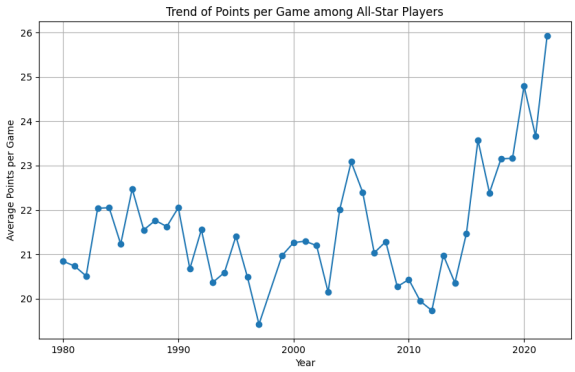

Fig. 1. Evolution of points per game.

The introduction and growing usage of three-point shooting further contributed to this evolution. Teams recognized the strategic advantage of the three-point line, and players began developing and refining their long-range shooting skills. This shift not only elevated individual scoring performances but also led to a more dynamic and unpredictable style of play.

Analyzing the evolution of three-point shooting, figure 2, provides additional insights. Initially, three-point attempts were sporadic, and the focus was on high-percentage shots closer to the basket. As the years progressed, the frequency and accuracy of three-point attempts increased, with players becoming adept at shooting from various positions on the court, and shooting beyond the arc plays a central role in offensive strategies.

This journey through the evolution of points per game and three-point shooting reflects not only statistical trends but also the dynamic nature of the NBA. It showcases the adaptability

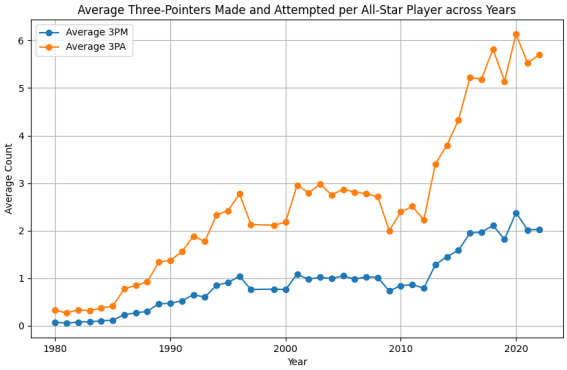

Fig. 2. Evolution of the use of the 3 points line.

of players and teams to changing trends, contributing to the league's constant innovation that defines modern basketball.

B. Game Influence

The influence of the teams that players are a part of plays a crucial role in their selection to the All-Star Game. The team's performance, overall record, and the player's contribution to the team's success are significant factors considered by voters. A player on a successful and high-profile team is more likely to garner attention and be selected for the All-Star Game.

As we delve into our analysis, it's intriguing to note how team dynamics and success impact individual player selections. The All-Star selection process not only recognizes individual success but also acknowledges the players' role in contributing to their team's achievements. Is not a coincidence that the most successfully teams in the NBA are the ones with the most selections.

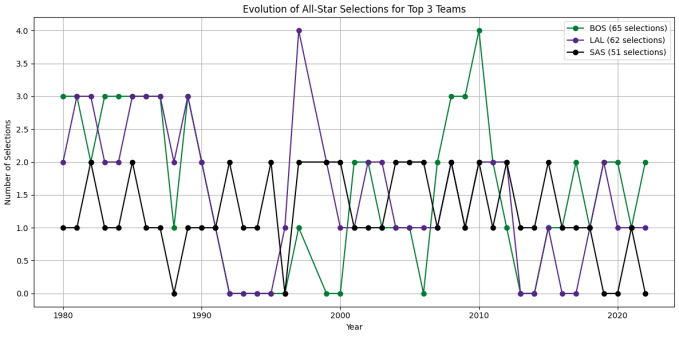

Fig. 3. Teams with most selections to the All-Star Game.

In Figure 3, it is evident that the top three teams in terms of All-Star Game selections between 1980 and 2022 are the Boston Celtics, Los Angeles Lakers, and San Antonio Spurs. This trend aligns with the historical success of Boston and the Lakers, both holding the record for the most championships, and highlights the Spurs as a team with significant historical prominence in the NBA.

An illustrative example of this correlation is LeBron James, who emerges as one of the most decorated players in terms

of All-Star selections. LeBron's remarkable career has witnessed significant contributions to multiple teams, showcasing his consistent excellence and leadership on the court. His numerous All-Star selections not only attest to his individual brilliance but also mirror the success of the teams he has been a part of throughout his illustrious career. Moreover, his enduring popularity and widespread admiration among fans contribute to his status as one of the most beloved players in the league.

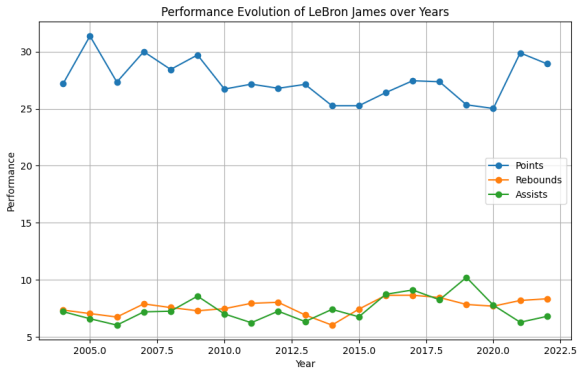

Fig. 4. Evolution of the Game of LeBron James.

III. Machine Learning Methods

A. Standardization

Standardization plays a pivotal role in our machine learning pipeline, contributing to the robustness and effectiveness of our predictive model. In the realm of predictive modeling, the StandardScaler function emerges as a valuable tool for transforming and standardizing the features within our dataset.

The StandardScaler operates by centering the data around its mean and scaling it to have a standard deviation of 1. This process ensures that all the features share a common scale, preventing any particular feature from dominating the learning process due to its larger magnitude. This is particularly crucial in our context, where player statistics encompass a diverse range of metrics, such as points, rebounds, and assists, each measured in different units.

The StandardScaler function calculates the standard score of a sample x as the following:

$$z = \frac{(x - \mu_{year})}{s_{year}}$$

Where z is the stardardized value, the μ is the mean of the training samples and the s is the standard deviation of the training samples.

By standardizing the features, we create a level playing field for the One-Class SVM model. This enables the algorithm to treat each feature with equal importance during the training process, fostering a more nuanced understanding of the patterns that distinguish All-Star players. The standardized values contribute to the model's ability to discern subtle variations and identify the essential factors that elevate a player to All-Star status. In essence, standardization enhances the model's interpretability and its capacity to make accurate predictions based on a unified metric.

B. One-Class SVM

In our search to build a robust predictive model for NBA All-Star selections, we select the One-Class SVM (Support Vector Machine) algorithm. This choice is driven by the unique nature of our dataset, which exclusively have positive instances—players who have been selected to the All-Star Game. The One-Class SVM is well-suited for such scenarios where the dataset primarily contains information about the target class, as it focuses on identifying deviations from the norm.

The One-Class SVM operates on the principle of mapping the input features into a higher-dimensional space and constructing a hyperplane that encapsulates the majority of the data points. In our case, this hyperplane delineates the characteristics common to All-Star players. By doing so, the algorithm effectively learns the intrinsic patterns associated with All-Star selections, making it adept at identifying players who exhibit similar statistical profiles.

This model choice is particularly advantageous when dealing with imbalanced datasets, where the positive class (All-Star selections) is significantly outnumbered by the negative class (non in this case). Given that not all players make it to the All-Star Game in a given year, our dataset inherently reflects this class imbalance. The One-Class SVM's ability to operate effectively in such scenarios positions it as a suitable candidate for our predictive modeling task.

Furthermore, the interpretability of SVMs allows us to gain insights into the decision boundaries that contribute to a player's likelihood of being selected as an All-Star. As we aim to develop a model that not only predicts but also sheds light on the factors influencing All-Star selections, the One-Class SVM aligns well with our objectives.

IV. Results

Applying the One-Class SVM to our dataset, which exclusively consists of positive instances (players selected to the All-Star Game), provides valuable insights. However, it's important to acknowledge the limitations of working only with positive data. The model primarily focuses on recognizing patterns associated with All-Star selections, making it less robust in predicting instances where players deviate from this norm.

Despite these inherent challenges, our model demonstrates a notable capacity to identify players with statistical profiles resembling those of All-Stars. The decision boundaries learned by the One-Class SVM provide a glimpse into the distinguishing factors that contribute to a player's likelihood of being selected to the All-Star Game.

Next, the One-Class SVM function is used with the following parameters:

- γ = 'scale' - Kernell coefficient, calculated as the following: $\frac{1}{(n_{year_{features}} \times \sigma^2_{X_{year}})}$
- μ = 0.10 - An upper bound on the fraction of training errors and a lower bound of the fraction of support vectors

The One-Class SVM results hinge on the distribution of the dataset, distinguishing between normal and abnormal patterns. This approach involves considering all player statistics, establishing a hyperplane to identify outliers, and leveraging these insights for predicting All-Star selections. Notably, we conducted two distinct training sessions—one incorporating the year of selection and the other disregarding it.

Surprisingly, the inclusion or exclusion of the selection year exerted a more substantial influence on the results than initially anticipated. The visual representation of these outcomes is depicted in the following images, providing a tangible view of the model's performance under these distinct training scenarios.

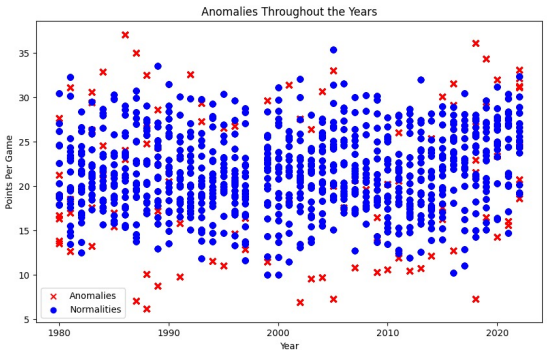

Fig. 5. Anomalies without year consideration.

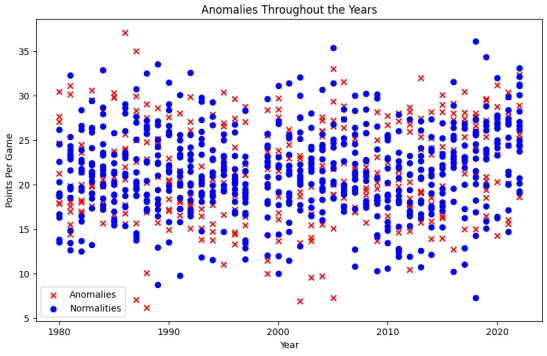

Fig. 6. Anomalies with year consideration.

As we can see in figures 5 and 6 if we consider the year of selection there is gonna be a lot more outliers in opposition to the result without considering the year.

A similar result can be seen in the images 7, 8, 9, 10. Various combinations of data can be selected for plotting based on specific aspects of interest.

When interpreting these results, it's crucial to consider that our model is trained with 20 or 21 different parameters. Drawing conclusions based on 2D and 3D representations of only a small subset of these parameters can be risky and may lead to misinterpretations. Additionally, upon observing these results, we can readily identify instances where certain players were incorrectly classified as outliers. A notable example is the miss classification of players with higher points per game as anomalies, whereas being the best scorers in the league would typically secure an easy entry into the All-Star Game.

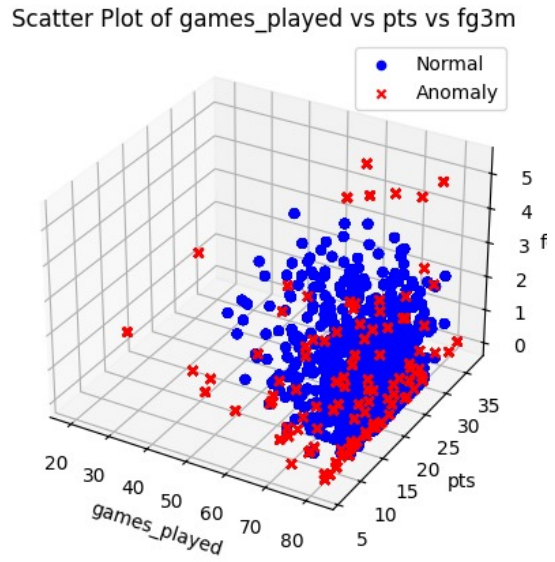

Fig. 7. Influence of games, points and 3 points without considering the year.

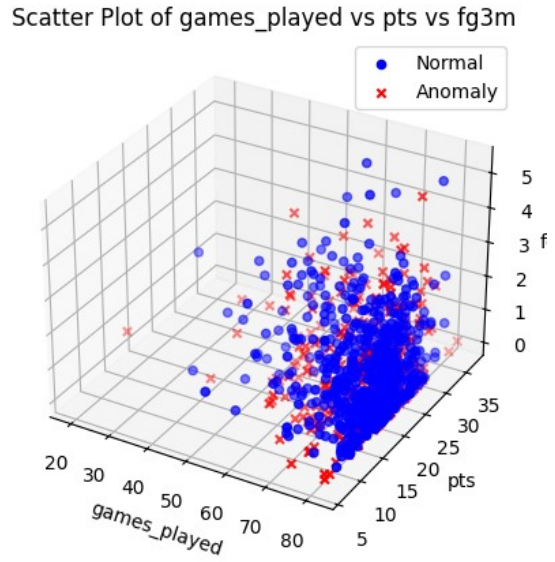

Fig. 8. Influence of games, points and 3 points considering the year.

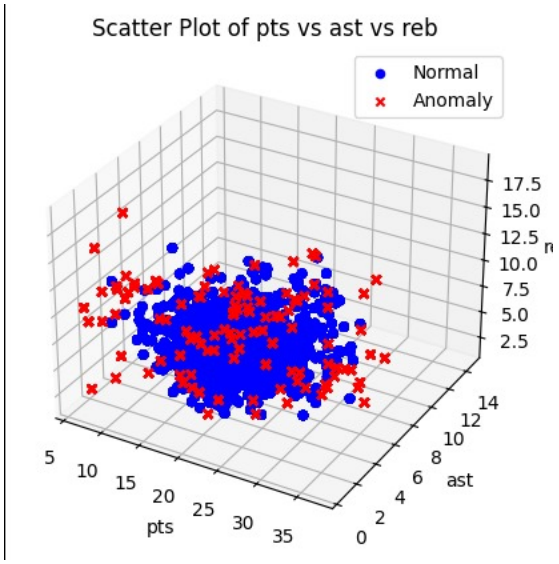

Fig. 9. Influence of points assists and rebounds without considering the year.

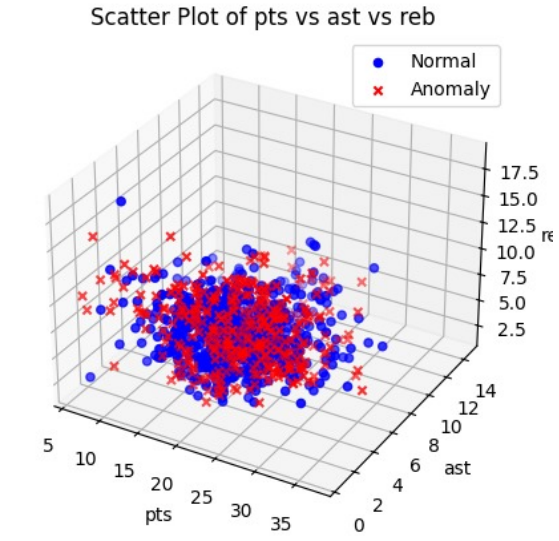

Fig. 10. Influence of points assists and rebounds without considering the year.

A final interesting result to analyze involves expanding the classification beyond normalities and anormalities. By introducing two additional classifications, namely ”very unlikely” and ”very likely,” we aim to provide a more nuanced perspective on the likelihood of a player being selected to the All-Star Game.

Examining the results depicted in image 11, we observe that only a single observation, representing the points per game from selected players, is truly an anomaly compared to the rest of the player population. This nuanced classification approach allows us to identify not only outliers but also instances that are highly unlikely or highly likely to be selected based on specific performance metrics.

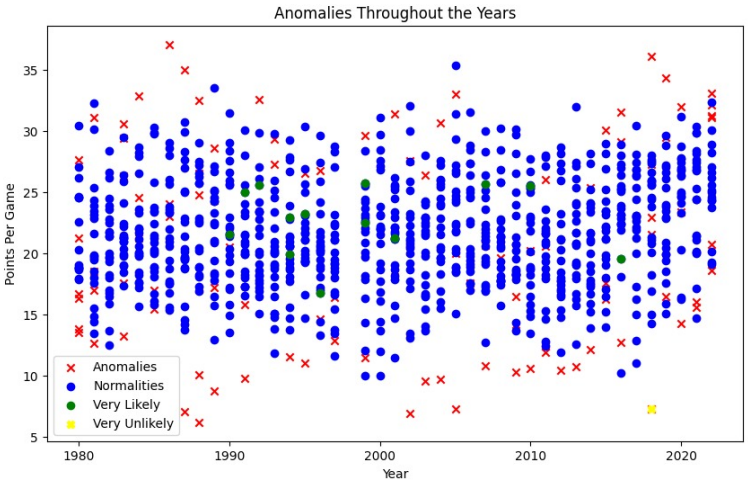

Fig. 11. Anomalies in Points per Game.

V. Conclusions

In conclusion, the absence of negative data in our dataset poses a significant challenge for the application of the One-Class SVM in predicting NBA All-Star selections. The model relies on distinguishing between normalities and abnormalities within the dataset, and without instances of non-selected players, it struggles to define a clear boundary between the two classes. Consequently, the predictive accuracy of our model may be compromised, as it lacks a comprehensive understanding of the factors influencing non-selection.

Moreover, the choice of the One-Class SVM for this specific task may not be the most optimal. While the model is adept at identifying outliers in datasets with clear distinctions between normal and abnormal instances, its efficacy diminishes when applied to a dataset primarily composed of positive instances. In essence, the One-Class SVM might not be the most suitable algorithm for predicting All-Star selections, given the nature of the dataset.

To address these limitations, alternative machine learning methods could be explored. Ensemble methods like Random Forests or Gradient Boosting could offer better performance, as they are capable of handling imbalanced datasets and capturing complex relationships within the data. Additionally, the incorporation of negative data, representing players who were not selected to the All-Star Game, would provide a more balanced training set, enhancing the model’s ability to discriminate between selected and non-selected players. While the current exploration employed the One-Class SVM, future iterations of this project may benefit from considering alternative models to improve predictive accuracy.