PERFORMANCE PREDICTION IN SPORTS ANALYTICS

Dr. V. Santhi,[a] Kamya Rachel*[b] , Bharat R*[c], Rudrajit Das*[d]
[a]Associate Professor
School of Computer Science and Engineering
[b] Principal Scientist,
Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India.
E-mail: vsanthi@vit.ac.in , kamya.rachel2022@vitstudent.ac.in, bharat.r2022@vitstudent.ac.in,
rudrajit.das2022@vitstudent.ac.in

ABSTRACT

The National Basketball Association(NBA) is a professional basketball league in America composed of 30 teams and is one of the major professional sports leagues in the world. Every year 60 players are drafted into NBA, and for the sixty players to get selected, a thousand enroll each year and undergo rigorous training for six months. But a question always arises on what criteria the players are selected in the draft that year or how a billion-dollar organization overwhelmed by performance-driven athletes goes to war to lure players from opposing teams. This paper presents a systematic and analytical approach to minimize the time taken to select the best player for a particular position on the basketball court. A deep analysis strategy is adopted using the Exploratory Data Analysis technique to analyze the necessary attributes that add to the player's strengths for a particular position. The results show that it leads to better performance gains through a systematic improvement of the basketball team. Further, for prediction, an optimized model is constructed using Logistic Regression, Support Vector Machine, and Random Forest, leading to higher accuracy. This strategy will enable coaches and organizations investing in basketball teams to formulate the best team.

Keywords: Exploratory Data Analysis, Logistic Regression, Support Vector Machine, and Random Forest.

I. INTRODUCTION

Sports analytics, multidimensional data analysis, Big data, and Prophetic business analytics got significant attention from nearly all diligence to give better services to their stakeholders. At the same time, sports companies, websites, broadcasters, and online platforms considerably make use of statistical and prophetic analytics to identify players' perceptivity, scoring patterns, and relatively grounded professional players' selection using thing-asked characteristics rising far and wide. These kinds of approaches are nonstop in real-time dynamic operations, complex in nature to prognosticate the analytics, and fascinating numerous experimenters for players' performance evaluations, prognosticating optimal results, and Decision timber in a timely manner.

Utmost approaches and styles concentrate on dynamic analytics and represent epitomized scoreboards during game time grounded on situations, environment, opponent analysis, thing generation and post analytics, and performance issues of different players. In the proposed approach we substantially concentrate on previous analysis of a player's selection grounded on performance, skill set, forming platoon, and minimizing time selection by reducing the cost originally to avoid further consequences and threat factors of unproductive platoon terrain.

Vaticination analysis requires accurate queries and interpreted by sense representation in multi-dimensional data to find applicable attributes of age, count, and country. It requires two parameters, one is the age and another is the number of players in basketball having that age.

Data booby-trapping gives information on periods and count of players using the group-by system and visualizes the information in the form of a bar map.

It helps to prize the knowledge about the players with respect to periods and grounded on that club proprietor can fluently pick players for their platoon by observing that players between the age range of 21 to 30 are more active and after 30 utmost of them retiring or unfit to play. The player's support to a team is determined by the proprietor according to their contracts, the team owner and coach or proprietor can make a decision to drop or increase the contract period accordingly according to their performance. For case, 24- time-old player, a 2-year contract is acceptable, and for a 21- time-old player a 5- year contract is profitable.

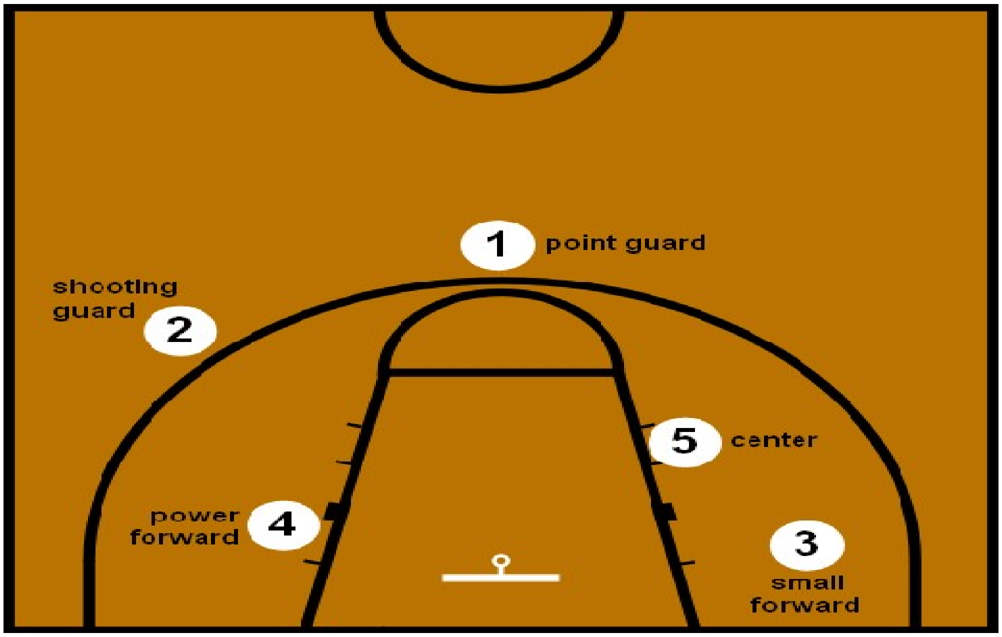


Fig 1: Positions in the Basketball Court

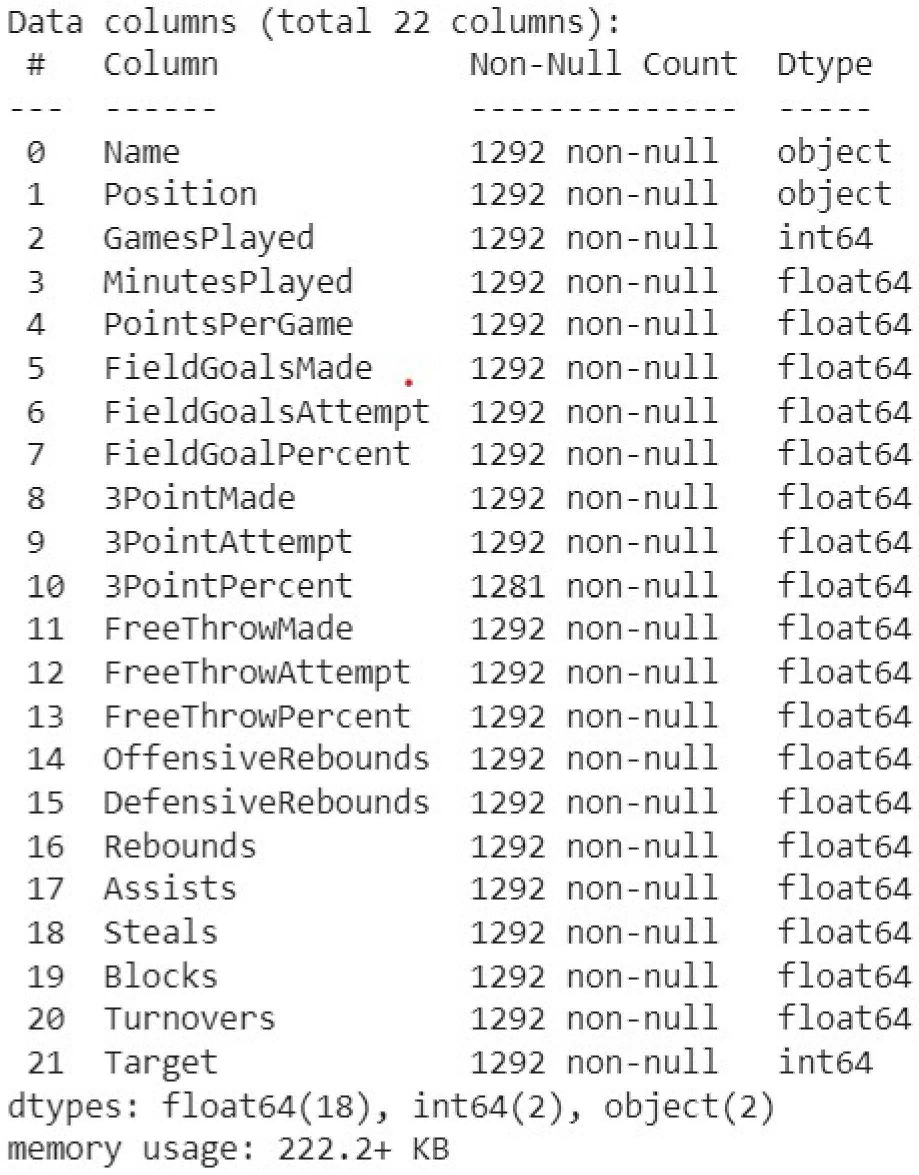
ABBREVIATION	POSITION FULL FORM
PG	POINT GUARD
SG	SHOOTING GUARD
PF	POWER FORWARD
SF	SHOOTING FORWARD
C	CENTRE

Table 1: Player Positions with Abbreviation

II. DATASET

The dataset contains all the features related to the performance of the player. There are around 1340 rows in this file and 21 feature columns indicating features like the player's Position, Games Played, Minutes Played, Points Per Game, Field Goals Made, Field Goals Attempt, Field Goal Percent, Field Goal Percent, 3 Point Made, 3 Point Attempt, 3 Point Percent, Free Throw Made, Free Throw Attempt, Free Throw Percent, Offensive Rebounds, Defensive Rebounds, Rebounds, Assists, Steals, Blocks, and Turnovers.

Data quality is the main issue in quality information management. Data quality problems occur anywhere in information systems. These problems are solved by data cleaning. the separate column of position has been manually cleaned and added to the data set for players' performance prediction with respect to position. Data scrapping was performed through Python packages. All data were retrieved from various NBA sports sources and aggregated in an Excel file followed by data cleansing actions. In addition, a normalization process of the final data was performed with the purpose to use them in the suggested formulas.



```
Data columns (total 22 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Name                1292 non-null   object 
 1   Position            1292 non-null   object 
 2   GamesPlayed         1292 non-null   int64  
 3   MinutesPlayed       1292 non-null   float64
 4   PointsPerGame       1292 non-null   float64
 5   FieldGoalsMade      1292 non-null   float64
 6   FieldGoalsAttempt   1292 non-null   float64
 7   FieldGoalPercent    1292 non-null   float64
 8   3PointMade          1292 non-null   float64
 9   3PointAttempt       1292 non-null   float64
 10  3PointPercent       1281 non-null   float64
 11  FreeThrowMade       1292 non-null   float64
 12  FreeThrowAttempt    1292 non-null   float64
 13  FreeThrowPercent    1292 non-null   float64
 14  OffensiveRebounds   1292 non-null   float64
 15  DefensiveRebounds   1292 non-null   float64
 16  Rebounds            1292 non-null   float64
 17  Assists             1292 non-null   float64
 18  Steals              1292 non-null   float64
 19  Blocks              1292 non-null   float64
 20  Turnovers           1292 non-null   float64
 21  Target              1292 non-null   int64  
dtypes: float64(18), int64(2), object(2)
memory usage: 222.2+ KB
```

III. PROBLEM DEFINITION

Long–term performance prediction for teams or individual players are fields requiring exploration. Not only coaches, but also sports agents and bookmakers are interested in how teams or players perform during a season compared to previous ones. What is discussed in this section is the context of this problem. Also, the objectives of the research are set. The unique components of basketball matches make long–term predictions very difficult; only few baskets are scored per match. Our research focuses on analyzing various attributes that improve that improves the performance of a player in a particular position.

IV. APPROACH FOLLOWED

This section showcases the flow of events taking place before we can get any meaningful experimental results, as well as the way the data were acquired and their preprocessing. The steps taken are:

Step 1 - Unstructure Data:

Step 2: Data Scraping

Step 3: Data Cleaning

Step 4: Encode the Dependent Attributes

Step 5: Data Pre-processing

Step 6: EDA

Step 7: Data Splitting

Step 8: Data Visualization

Step 9: Model (Logistic Regression + Random Forest + Support Vector Machine)

Step 10: Evaluate the model

Step 11: Process and Validate the Algorithm

Step 12: Hyper Parameter Tuning

The block diagram which summarizes the process is depicted in Fig. 1:

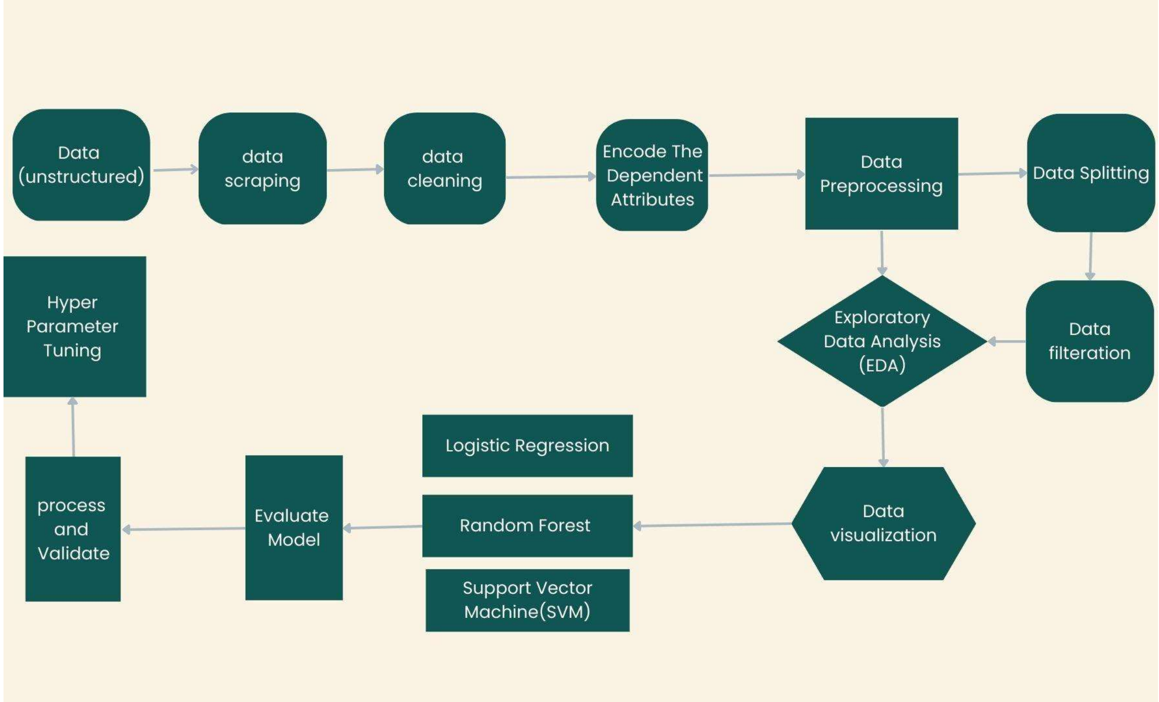


Fig 2: Data Flow Diagram

At first, the appropriate data had to be found. There are a lot of web pages that contain information and statistics regarding basketball matches and events. The data refer to both teams and players. Some of the data were accessed and collected manually, especially when that was easy. However, some of them were scraped from the internet using various scraping tools. Finally, a free database from an expired online competition had been downloaded and used for the experiments. The database contains data from thousands of players and is extracted from a famous manager simulation game. It demonstrates player ratings for several football skills. Players are rated by domain experts. After the process of data acquisition, there was a large database which needed to be organized. The database was split into different csv files, according to what data were essential for each experiment. Then, the csv files were uploaded to google colab. Naturally, the data firstly needed to be preprocessed. They were checked for null values, duplicates, noise etc. Python was used to clean the data and build the models. Then, data transformation and data reduction took place to keep only the appropriate features for each classification or regression. Finally, results were evaluated in terms of accuracy. Further, Hyperparameter Tuning was used to fine tune and improve the accuracy of the model constructed.

V. EXPERIMENTS AND RESULTS

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is one of the techniques used for extracting vital features and trends used by machine learning and deep learning models in Data Science. Thus, EDA has become an important milestone for anyone working in data science.

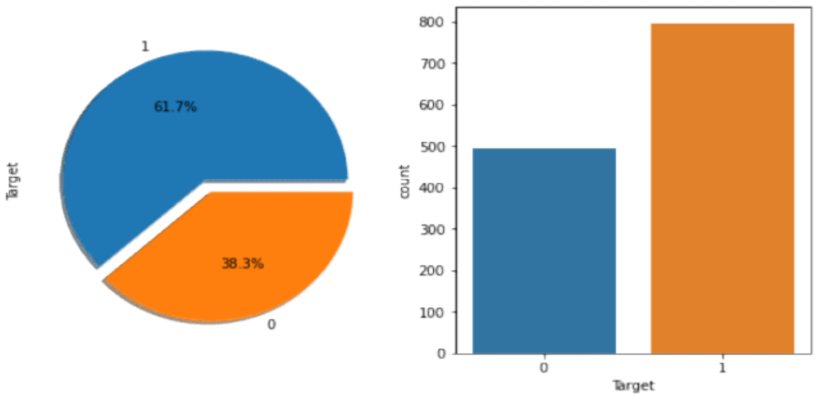


Fig 3: Desired Target variable 0 and 1

Kernel Density Estimation

The Kernel Density Estimation is a mathematic process of finding an estimate probability density function of a random variable. The estimation attempts to infer characteristics of a population,based on a finite data set. The data smoothing problem often is used in signal processing and data science, as it is a powerful way to estimate probability density. In short, the technique allows one to create a smooth curve given a set of random data.

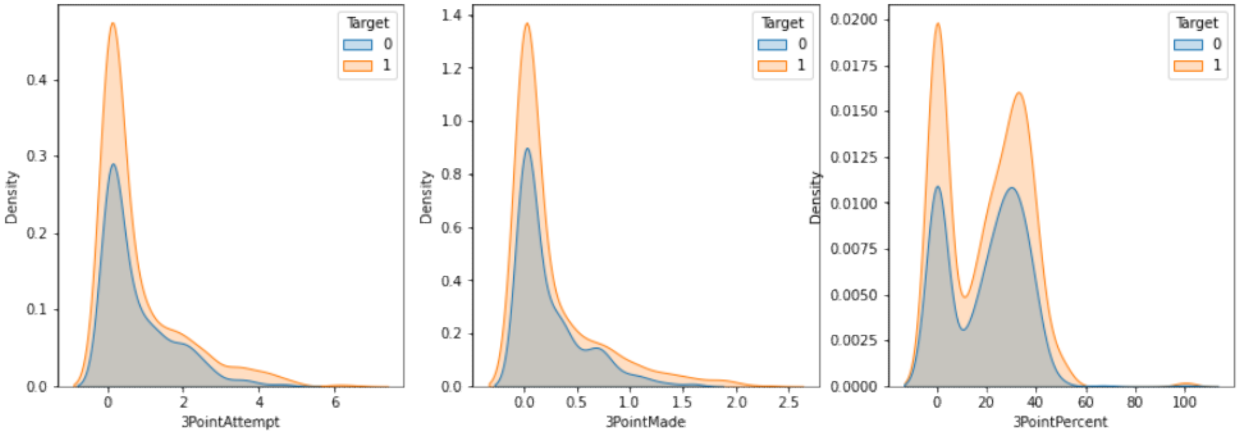


Fig 4: Density estimation of all the players for 3PointAttempt, 3PointMade and 3PointPercent

Those with more than 5 years of experience showed a significant difference in 3-point Attempts, Made, Percent

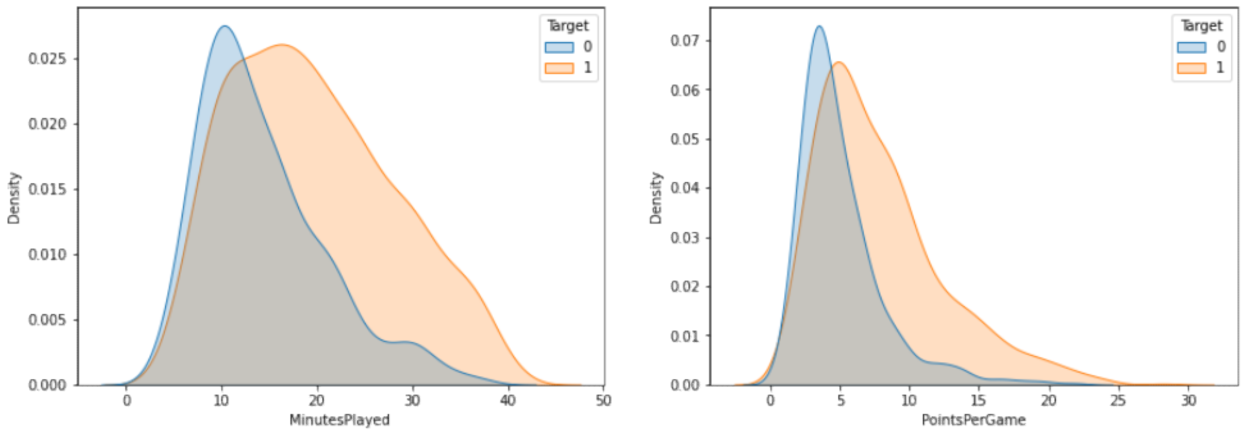


Fig 5: Density estimation of all the players for MinutesPlayed and PointsPerGame

Each position in basketball is different, and you may not know whether the positions are evenly distributed in the given data. For example, if there are 100 strikers and 80 of them are veterans, of course, the veterans have no choice but to score more.

Thus,Veterans have more play time than amateurs and score more goals per game. But we should note that this analysis is not very helpful.

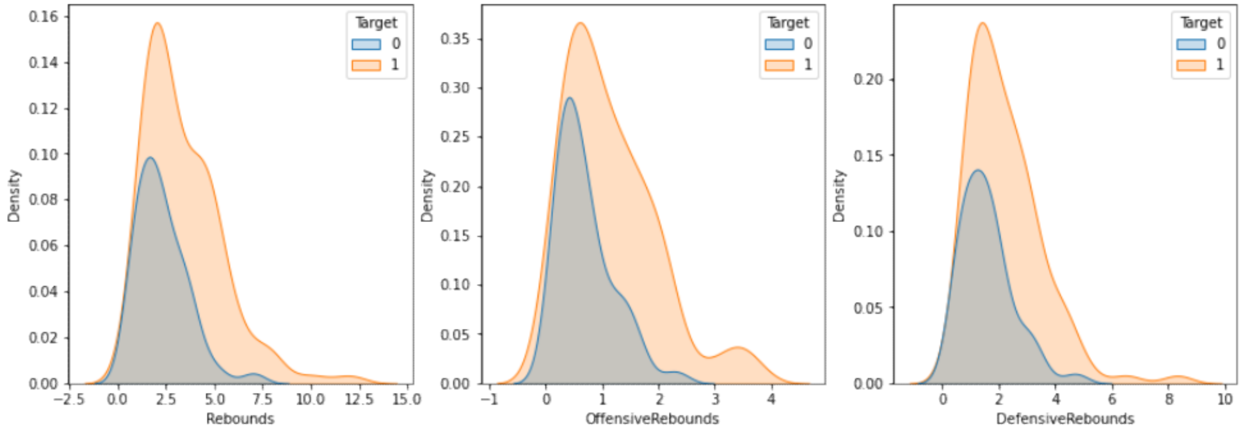


Fig 6: Density estimation of all the players for Rebounds, OffensiveRebounds and DefensiveRebounds

The most important attribute for a player in the center position in basketball is rebounding. Centers need to be great rebounders in order to help their team get extra possessions and keep the other team from scoring points. Rebounding is a key skill for centers and is often a major factor in deciding who starts and who gets more playing time. Hence, we'll analyse rebounds center players

Heat Map

Heat Maps are graphical representations of data that utilize color-coded systems. The primary purpose of Heat Maps is to better visualize the volume of locations/events within a dataset and assist in directing viewers towards areas on data visualizations that matter most.

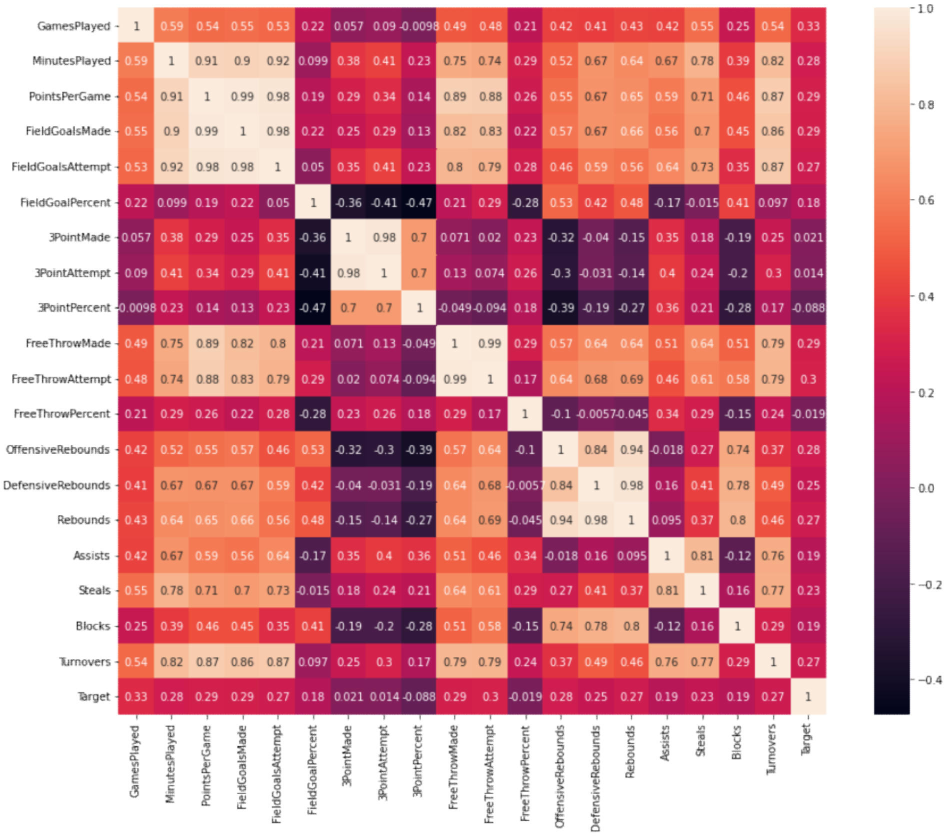


Fig 7: Heat Map for players of all positions to find the correlation between various attirubes

Extracting the dataset for players of that paticular position:

Center

The most important attribute for a player in the center position in basketball is rebounding. Centers need to be great rebounders in order to help their team get extra possessions and keep the other team from scoring points. Rebounding is a key skill for centers and is often a major factor in deciding who starts and who gets more playing time. Hence, we'll analysethe rebounds for center players.

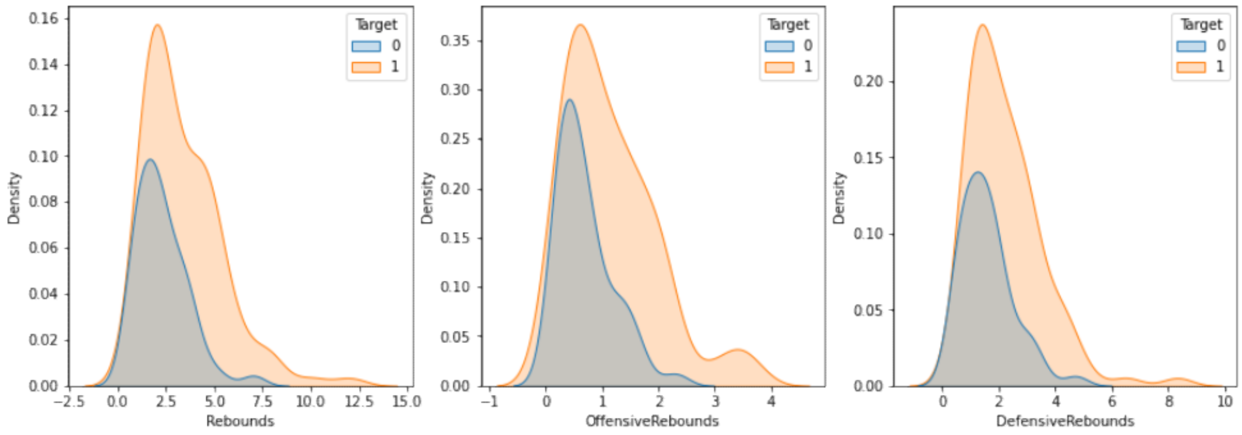


Fig 8: Density estimation of players in center position for Rebounds, OffensiveRebounds and DefensiveRebounds

Small Forward

The main attribute of a basketball player in the small forward position is athleticism. This position requires a player who is agile and has good speed, as well as the ability to make quick decisions and read the court. The small forward should also be able to shoot, pass, and dribble well, as well as have the ability to defend against opponents.

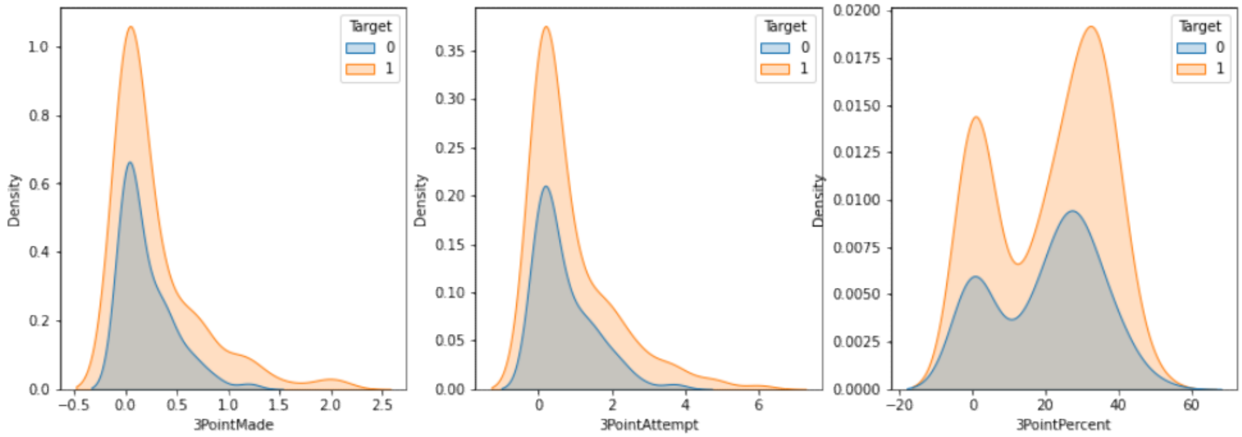


Fig 9: Density estimation of players in small foward position 3PointAttempt, 3PointMade and 3PointPercent

Point Guard

The main attribute of a basketball player in the point guard position is their ability to facilitate the team's offense by controlling the ball and creating scoring opportunities for their teammates. This includes making decisions on when to pass, when to shoot, and when to drive to the basket. Point guards must also be able to defend against their opponents' players and anticipate their opponents' moves so that they can be ready for quick counterattacks.

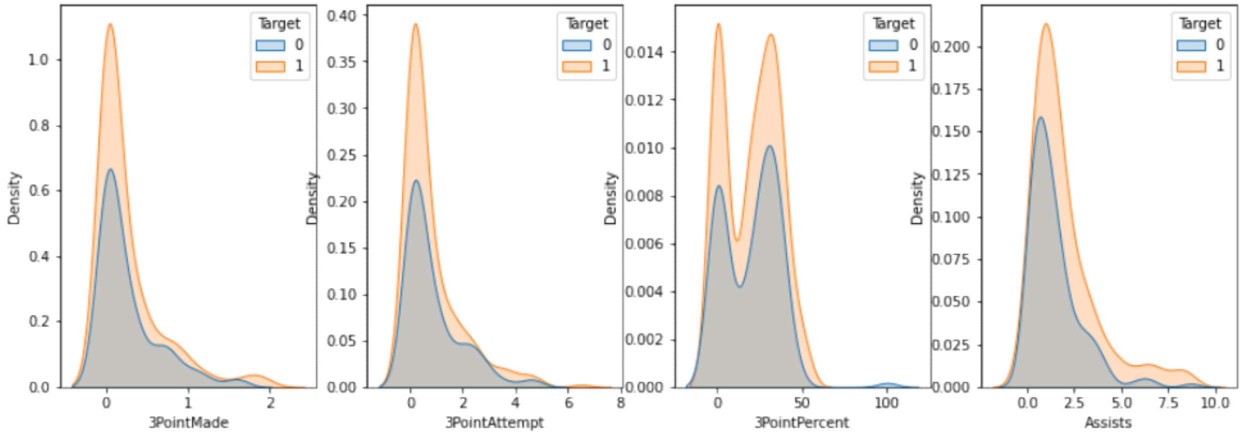


Fig 10: Density estimation of players in Point Gaurd position for 3PointAttempt, 3PointMade, 3PointPercent, and Assists

Power Forward

The main attribute of a basketball player in power forward position is strength. Power forwards need to be strong and have the ability to battle for rebounds, defend the paint, and score in the post. They also need to have good agility and court awareness.

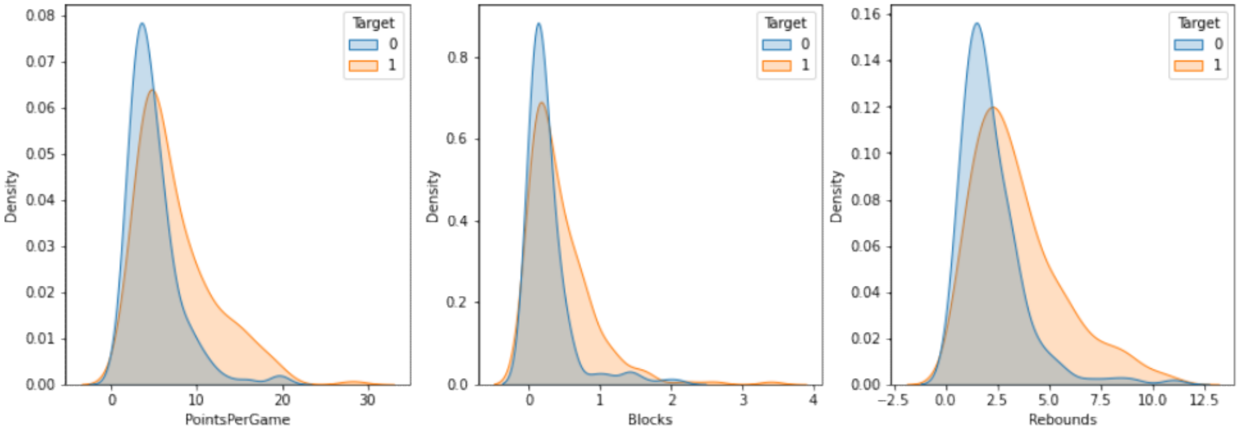


Fig 11: Density estimation of players in Power Forward position for PointsPerGame, Blocks, and Rebounds

Shooting Guard

The main attribute of a Shooting Guard in basketball is shooting accuracy. Shooting Guards are typically the most prolific scorers on a team and must possess a great shooting touch from both mid-range and long-range distances. Shooting Guards must also be able to create their own shots, as well as drive to the basket for layups and dunks. The ability to penetrate and make plays for teammates is also essential for this position.

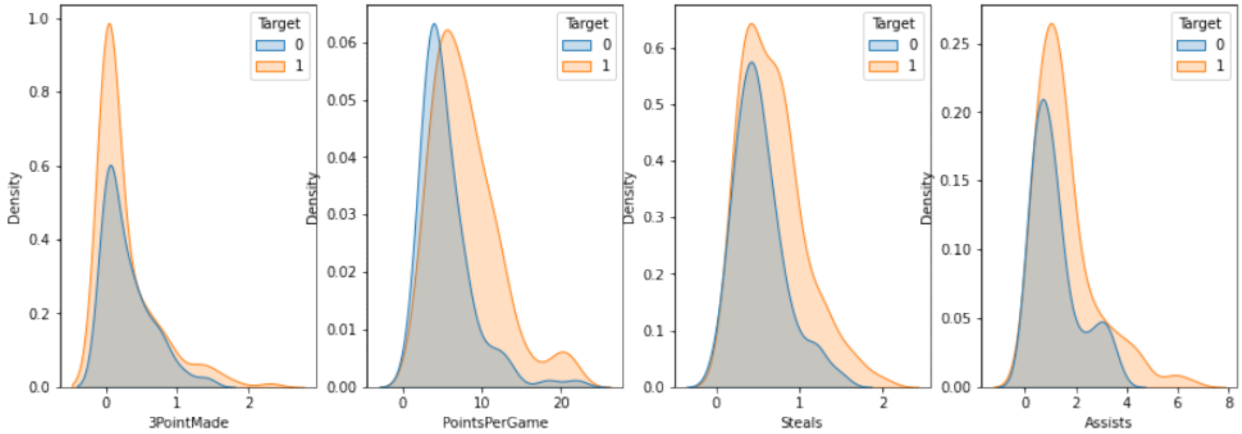


Fig 12: Density estimation of players in center position for 3PointMade, PointsPerGame, Steals and Assits

VI. CONCLUSION:

Our aim was to identify players who are bringing abilities which enhance the present players in the team and thus help in creating an identity as a single unity and bring trophies to the organization. Every player brings their own abilities to the game of basketball. With different skill sets, it becomes very difficult to choose from a pool of players. Our analysis helps in narrowing down the player abilities according to their position and their respective attributes.

1.Players who play as ***POINT GUARD*** are given the responsibility of providing ***ASSISTS,SCORING POINTS*** and mostly controlling the tempo of the game.

2.Players who play as ***SHOOTING GUARD*** are given the responsibility of providing ***STEALS,SCORING POINTS*** and acting as the second field general after PG.

3.Players who play as ***POWER FORWARD*** are given the responsibility of scoring ***POINTS*** and ***PROVIDING BLOCKS*** for the team.

4.Players who play as ***SMALL FORWARD*** are given the responsibility of ***SCORING POINTS*** and mostly defending purposes.

5.Players who play as ***CENTERS*** are given the responsibility of providing ***REBOUNDS BOTH OFFENSIVE AND DEFENSIVE REBOUNDS*** and also scoring points for the teams

As seen in the above results, RANDOM FOREST is providing the best accuracy results in terms of player performance and provides a detailed look into the abilities which the players bring to the table. HEAT MAP usage provides us with the cumulative results of how a player's attributes are mixing well with the team statistics. On seeing the heat map a team owner can decide which player they want to bring into the already present pool of players in order to create a championship caliber team.

REFERENCE:

[1] An Q, Wen Y, Chu J, Chen X. Profit inefficiency decomposition in a serial[1]structure system with resource sharing. J Oper Res Soc 2019. doi:10.1080/ 01605682.2018.1510810.

[2] An Q, Wen Y, Ding T, Li Y. Resource sharing and payoff allocation in a three-stage system: integrating network DEA with the shapley value method. Omega 2019;85:16–25. [5] An Q, Wang Z, Emrouznejad A, Zhu Q, Chen X. Efficiency evaluation of par[1]allel interdependent processes systems: an application to Chinese 985 project universities. Int J Prod Res 2019;57(17):5387–99. doi:10.1080/00207543.2018. 1521531.

[3] Andrade A, Bevilacqua GG, Coimbra DR, Pereira FS, Brandt R. Sleep quality, mood and performance: a study of elite Brazilian volleyball athletes. J Sports Sci Med 2016;15(4):601–5. [7] Ang S, Yang C, Zhao F, Yang F. Ranking of DMUs with interval cross-efficiencies based on absolute dominance. Int J Inform Decis Sci 2016;8(4):325–40.

[4] Andrade A, Bevilacqua GG, Coimbra DR, Pereira FS, Brandt R. Sleep quality, mood and performance: a study of elite Brazilian volleyball athletes. J Sports Sci Med 2016;15(4):601–5.

[5] Ang S, Yang C, Zhao F, Yang F. Ranking of DMUs with interval cross-efficiencies based on absolute dominance. Int J Inform Decis Sci 2016;8(4):325–40.] Zuccolotto, P., Manisera, M., & Sandri (2018), M. Big data analytics for modeling scoringprobability in basketball: The eσect of

shooting under high-pressure conditions. International Journalof Sports Science & Coaching, vol. 13(4), pp. 569-589.

[6] Naman Gupta, "Off the Ball: A Data Science Approach to Real-Time Football Fan Engagement" , Thesis report of University of Michigan, 2019.

[7] Tom Decroos, Jan Van Haaren, and Jesse Davis. 2018. Automatic Discovery of Tactics in Spatio-Temporal Soccer Match Data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/3219819.3219832.

https://doi.org/10.1145/3219819.3219832.

[8] Paolo Cintia, Fosca Giannotti, Luca Pappalardo, Dino Pedreschi, and Marco Malvaldi. 2015. The harsh rule of the goals: data-driven performance indicators for football teams. In Procs of the 2015 IEEE International conference on Data Science and Advanced Analytics.

[9] Nsolo, E. - Lambrix, Pa. - Niklas, C. Player Valuation in European Football. 2018. 5th Workshop on Machine Learning and Data Mining for Sports Analytics co-located with ECML PKDD 2018.

[10] Apostolou, K. and Tjortjis, C. Sports Analytics algorithms for performance prediction. IEEE 10th Int'l Conf. on Information, Intelligence, Systems and Applications (IISA 2019), pp. 469-472, 2019.

[11] Sarlis V. and Tjortjis C., Sports Analytics – Evaluation of Basketball Players and Team Performance, Information Systems, Vol. 93, November 2020, doi: 10.1016/j.is.2020.101562.