



Sports Analytics: NBA Simulation and Prediction

Business Analytics

Columbia University in the City of New York



1. Introduction

Basketball is one of the best examples of how analytics has changed the way sports are played and player performance is measured. In this project, we aim to analyze the NBA league as a whole and its different facets by making several predictions about the current season of NBA. In basketball, NBA teams have started using a technology called “Player Tracking” to evaluate the efficiency of a team by analyzing player movements. Similarly, we have tried to explore different use cases of business analytics in basketball. Our models and analysis will provide the coaches and managers with valuable insights and assist them in making strategic game decisions. Moreover, betting agencies can place data-driven bets to maximize profits and minimize risk. The popularity of data-driven decision-making in basketball has trickled down to the fans who are consuming more analytical content than ever before. We addressed 3 important questions that almost every fan is curious about:

- *Who will be the Most Valuable Player this season?*
- *What should be each team's best strategy for each game? Is there a potential trade that can elevate the team's performance?*
- *Lastly, which team is going to win the NBA finals this season?*

2. Data Collection and Manipulation

The data needed for our analysis was obtained using a Python API called SportsReference [1]. [SportsReference](#) is a free Python API that provides data for various sports and leagues such as the MLB, NBA, College Football and Basketball, NFL and NHL. We collected NBA data for the last 10 years to do predictive modeling. The data we included belongs to these 3 categories:

- Historical data of all NBA matches played in the last 10 years
- Player statistics and general information
- NBA schedule for the current 2019-20 season

2.1. Historical Match Data [2]:

This dataset contains the match statistics for all the matches played from NBA 2010-11 season up until today. Each data point in this dataset is a match with exhaustive details regarding which teams were playing, where the match was played, the players involved in both the team, the points scored by both teams, the outcome of the match, attendance and so on.

2.2. Players statistics [3]:

This dataset comprises information and statistics for all players who have played from the NBA 2010-11 season up until today. Also, the player statistics are provided for individual seasons allowing us to measure how a player's performance has changed over the years. The statistics for each player include their physical attributes such as height, weight, preferred playing the position as well as individual match stats like the number of games played, three-pointers scored, defensive-fouls committed, overall efficiency rating and so on.

2.3. Schedule for NBA 2019-20 season [4]:

This dataset includes all the matches scheduled for this season. Each data point in this dataset represents a match that will be played and includes variables like which teams are playing, the home and away team, match date and time, etc. After pulling the above data; preliminary EDA and data manipulation was done to get some insights on the data.

3. K-Nearest Neighbors for MVP prediction

We have used K-Nearest Neighbors to predict the MVP for the 2019-20 season. For this analysis, we used the player statistics data from 2013 to 2017 which we pulled from the sportsreference API. The data for each player can be thought of as a point in on a multi-dimensional plane. We have demonstrated the idea in 2D below for illustration. Now each point in the plane can be thought of a players' representation. The steps are as follows to predict the MVP:

1. Identify the players who have been the MVP in their respective seasons from 2013 to 2017 (4 MVPs)
2. Calculate the centroid of these four points to get a *mean MVP* representing all the key performance statistics of an MVP
3. Using the scaled performance statistics for each player in the next years (using the same scaler as we used in the training data) we identify the nearest neighbor to our *mean MVP*

Methodology:

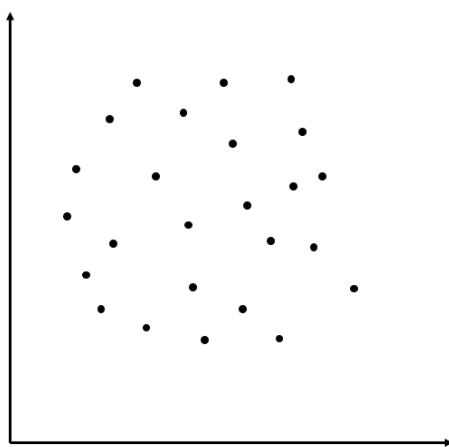


Fig.1: Data points for the seasons 2013-2017

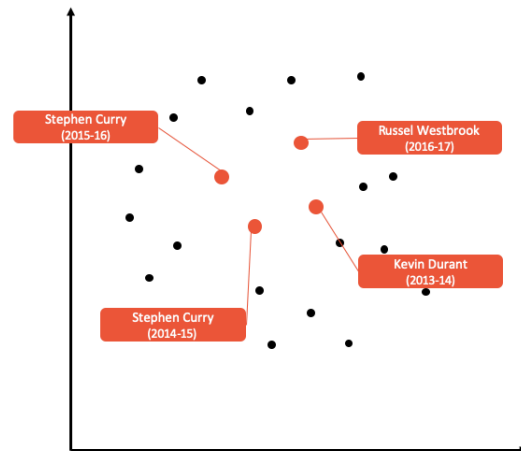


Fig.2: MVPs for each year in the training data

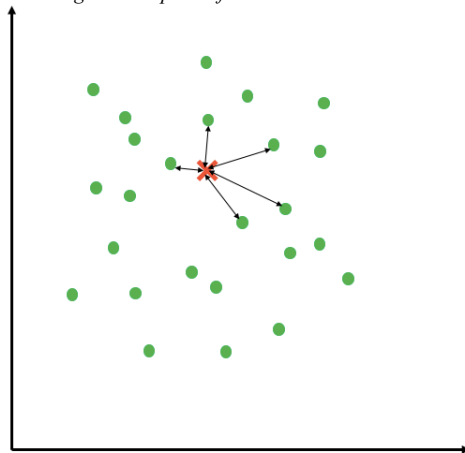


Fig.4: Nearest neighbors of mean MVP

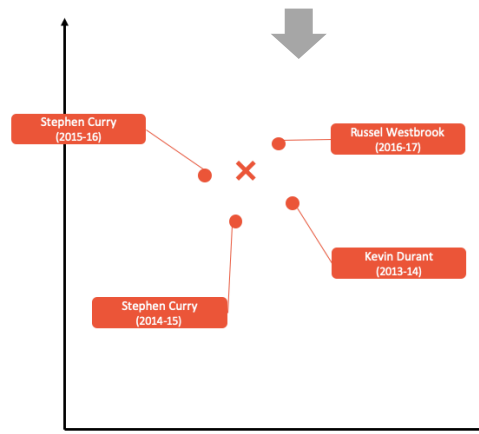
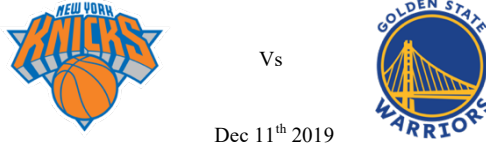


Fig.3: Mean MVP in the training data

We validated the data by predicting the MVPs for 2016-17, 2017-18 season and the model correctly predicted *James Harden* and *Giannis Antetokounmpo* validating the performance of the model. In order to predict the MVP for the current season based on the performances of players so far. The model predicted it to be *Giannis Antetokounmpo* again. We then validated this prediction with news article and NBA podcasts and observed that Giannis is outperforming himself from last season. So, it makes sense that he is a viable candidate to win the MVP award this season.

4. Logistic regression for Match prediction:



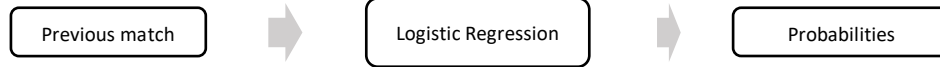
The goal of building a logistic model was to predict the probability of the home team winning based on certain parameters. Now let us consider an upcoming match between Golden State Warrior (GSW) and New York Knicks (NYK). If we look at past encounters between these teams, we would probably say that GSW has a better chance of winning the upcoming fixture. But is this really true? There are several factors to be considered:

1. Is there any key player injured in any of the teams which can affect our predictions?
2. What are the latest trades in both teams? Is there any key player who has been traded and has made the team stronger/weaker?

These questions indicate that players are an integral part in determining the outcome of a match. So due to the factors mentioned above, the chances cannot be determined only on the basis of the previous match ups? This is where a Logistic model can help by taking into account the player contributions and providing a probability of a home team winning. An analysis of the coefficients would also provide some key insight about the team strategies and each player's importance. We modelled each team based on its players as follows:



We then fed the data into a logistic regression model to get the coefficients:



Our final logistic model's equation, considering only significant variables is as follows:

4.1 Model equation

$$\begin{aligned}
 \ln\left(\frac{p}{1-p}\right) = & \beta_0 + \beta_1 PG_{rating|Home} + \beta_2 SG_{rating|Home} + \beta_3 PF_{rating|Home} \\
 & + \beta_4 SF_{rating|Home} + \beta_5 C_{rating|Home} + \beta_6 Bench_{rating|Home} \\
 & + \beta_7 PG_{rating|Away} + \beta_8 SG_{rating|Away} + \beta_9 PF_{rating|Away} \\
 & + \beta_{10} SF_{rating|Away} + \beta_{11} C_{rating|Away} + \beta_{12} Bench_{rating|Away} \\
 & + \beta_{13} PG_{rating|Home} \times PF_{rating|Home} + \beta_{14} PG_{rating|Away} \times PF_{rating|Away}
 \end{aligned}$$

4.2 Model summary













	Variables	Coefficients	p-Value	Impact	Importance
β_0	Intercept	-0.507	0.362	↑	Not in scale
β_1	PG_home_rating	0.118	0.000	↑	
β_2	SG_home_rating	0.039	0.000	↑	
β_3	SF_home_rating	0.061	0.000	↑	
β_4	PF_home_rating	0.106	0.000	↑	
β_5	C_home_rating	0.044	0.000	↑	
β_6	bench_home_rating	0.035	0.021	↑	
β_7	PG_home_rating : PF_home_rating	-0.003	0.002	↓	Not in scale
β_8	PG_away_rating	-0.100	0.000	↓	
β_9	SG_away_rating	-0.034	0.000	↓	
β_{10}	SF_away_rating	-0.040	0.000	↓	
β_{11}	PF_away_rating	-0.085	0.000	↓	
β_{12}	C_away_rating	-0.029	0.000	↓	
β_{13}	bench_away_rating	-0.046	0.003	↓	
β_{14}	PG_away_rating : PF_away_rating	0.002	0.021	↑	Not in scale

Table1: Model summary

From the above model summary, there are some important insights as follows:

4.3 Insight from logistic model

1. The coefficients are consistent with the actual match intuition. Higher the rating of the home team, more is the probability of winning. Also, higher the rating of the players of the opposite team, lower is the probability of the home team winning.
2. The Point Guard is the most important player in any team as it has the highest coefficient. This makes sense as the Point Guard is the player which is mainly responsible for making the plays in the match and the team's performance depends highly on this player. This is also validated by the highest negative coefficient of the opponent team
3. The Center position is the least important among all the players. This is the reason why players which the Center position are not paid highly. This is because the Center position is mainly responsible for rebounds and do not usually contribute in making three's an field goals. They also usually have a lower field goal percentage
4. There is an interaction term $PG_home_rating:PF_home_rating$ and $PG_away_rating:PF_away_rating$ which is significant (Point Guard and Power Forward). This has a coefficient of negative as compared to the other coefficients in a particular team. This means that if both of these players have a higher rating than the performance of the team would go down and lesser would be the probability of winning. This makes a lot of sense because both of these players are the ones which keep possession of the ball and if both the players are highly rated than both of them would have a competition for the ball possession and would result in a decrease in performance. This explains why teams which have all 'Star' players are usually not so successful
5. Finally, if a coach or a manager needs to improve the probability of winning, then investing in a Point Guard would be most profitable. Also, the coach must make sure that there is good coordination and understanding between the Point Guard and the Power Forward

5. Tournament Simulation

We carried out the simulation of the entire NBA 2019-20 tournament based on the results derived from the logistic regression model the probabilities of teams winning against other teams. This is a small snippet of the probability matrix. The rows are the home teams and the columns indicate the away teams. The cells are the probability of a home team winning against the corresponding away team. The original probability matrix is of dimensions 30 x 30.

	OKC	ORL	PHI	PHO	POR
CLE	0.658	0.659	0.611	0.601	0.563
DAL	0.733	0.734	0.692	0.683	0.649
DEN	0.705	0.706	0.662	0.652	0.616
DET	0.642	0.643	0.594	0.584	0.546
GSW	0.724	0.725	0.681	0.672	0.637

Table2: Probability matrix snippet

The steps to carry out the tournament simulation is as follows:

1. Fetch the NBA schedule from Sportsreference API
2. Store the already played matches and tournament score in a dataframe
3. Carry out a Bernoulli trial for every matchup with a probability P which is obtained from the probability matrix for the corresponding matchup, taking into consideration which team is home and away
4. Update the tournament scores based on the result of the match obtained above
5. Carry out steps 1 to 4 for all the matches in the tournament

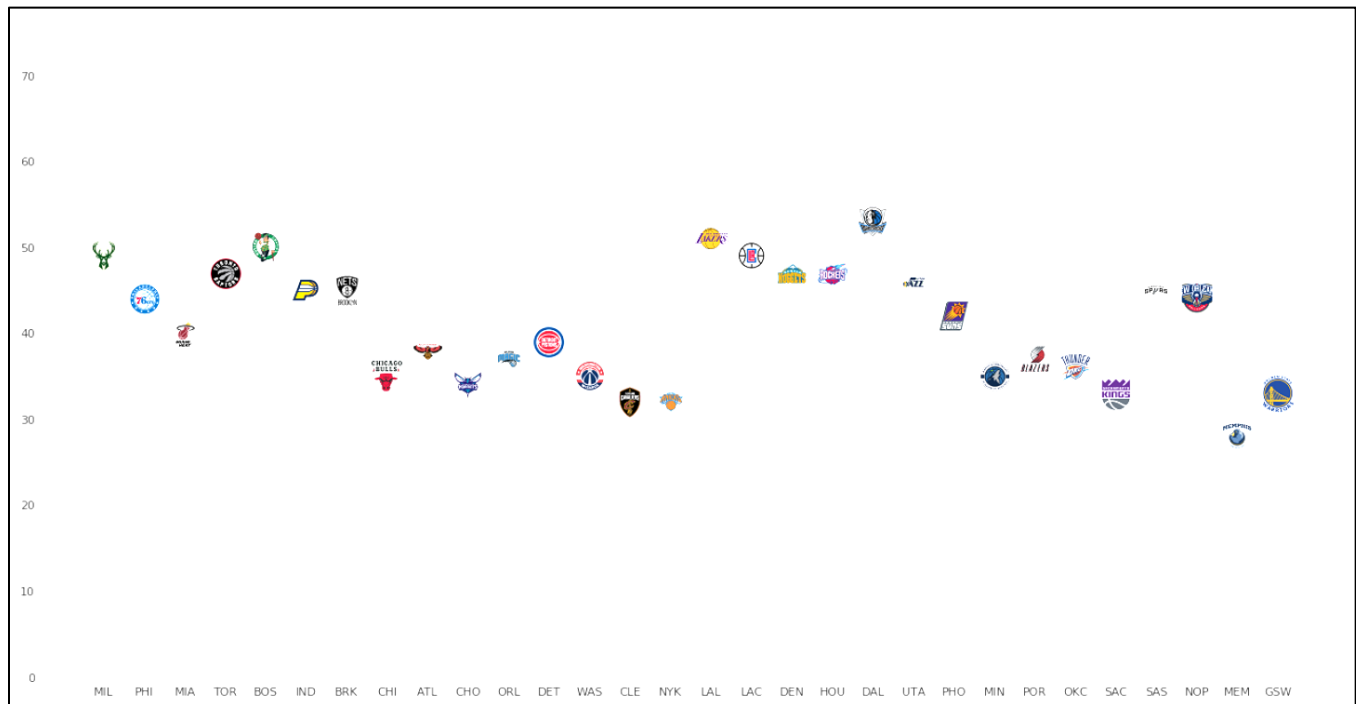


Fig.5: Tournament Simulation

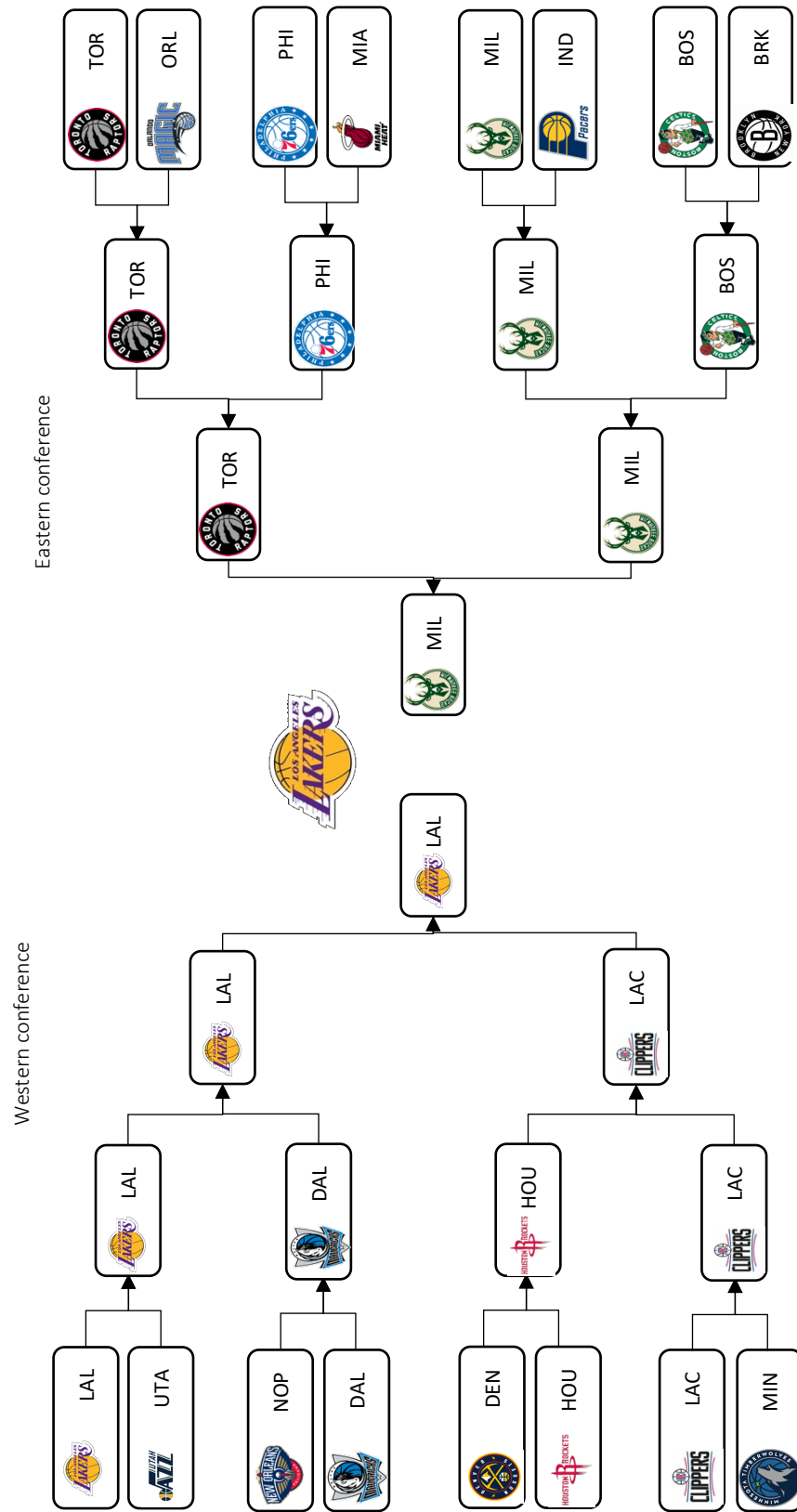


Fig.6. Simulation results

From an analytical perspective, it is imperative to observe the variation in this outcome. So, we carried out 500 such simulations to find the expectation of every team to win the championship title. From these simulations, we observed that Los Angeles Lakers have the highest odds of winning but Milwaukee Bucks and Boston Celtics are definitely not far behind.

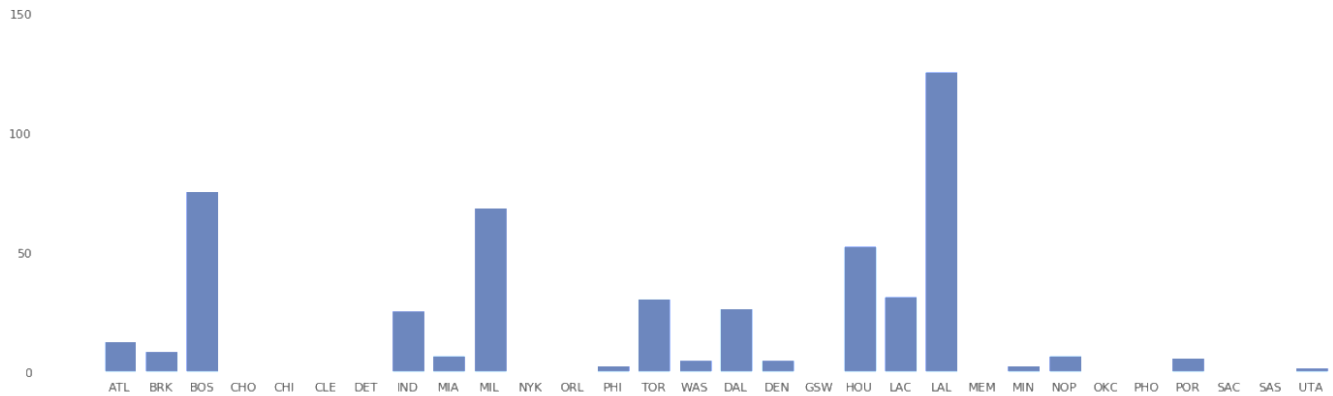


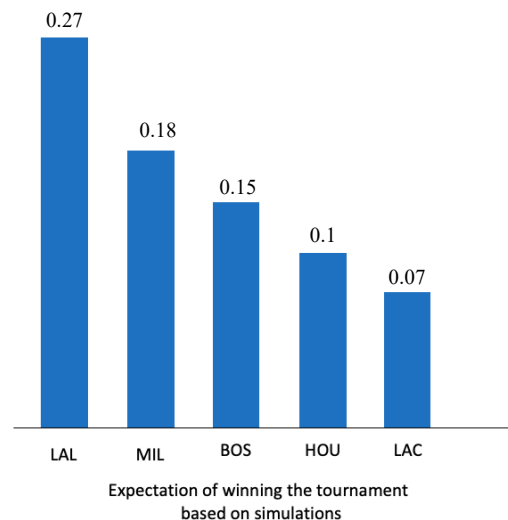
Fig. 5: Simulation results for 500 simulations

4.3 Insights from simulations

(A) Conference structure effect on team's tournament winning expectation

An interesting insight generated off the simulation aspect was that even though Houston Rockets and LA Clippers have the 2nd and 3rd highest mean probability of winning, their odds of winning the entire tournament are lesser as compared to Milwaukee Bucks and Boston Celtics. This is due to the conference structure and the way the tournament is designed. The competition amongst western conference teams is higher, making their road to the finals tougher as compared to the eastern conference teams.

Team	Mean winning probability
LAL	0.74
HOU	0.70
LAC	0.69
MIL	0.68
BOS	0.68
NOP	0.65
TOR	0.65
ATL	0.65
DAL	0.65
IND	0.64
GSW	0.64
WAS	0.62



This insight could be uncovered only by simulating the tournament 500 times, reinforcing the importance of simulation in sports analytics.

6. Conclusion and Future scope

Several valuable insights having a business value attached were generated through our project.

- The logistic regression model can be effectively used by team coaches and managers to develop customizable match strategies. The coach can simulate the starting line-up and bench squad to predict how the team will perform in the upcoming fixture. Additionally, strategies can be built for the point guard and the power forward positions, the two most important positions in the team.
- Teams can make strategic trades (Substitutes) by observing the impact of a particular player entering or leaving the team.
- Betting agencies now have access to the team winning probabilities of the next match and to construct a quantitative bet.
- Simulations are an important method to observe the odds of a particular team winning the entire tournament.

Future Scope

Analytics in basketball is a huge domain and consists of massive amounts of untapped data. So far, we have analyzed historical data to make predictions for the current season. Now, we are well-equipped to understand the different data attributes of each match. Next, we can dig deeper into every match and analyze a player's performance in a granular fashion. For instance, we can measure how fast a player moves, distance covered during the game, percentage of ball possession of each player, number of times ball is touched, number of passes and rebounds in which a particular player is involved. This will help to dive deeper into the analysis and generate accurate, personalized strategies for every game.

Another extension to our model is to predict the best starting line-up of home team given an opponent to maximize the chance of winning. This can be modelled as an optimization problem with the objective of maximizing the winning probability of the team. The available players, their ratings, their positions etc. will be the constraints of the optimization problem.

As depicted, there are many of use-cases of such data analysis project in sports. We have only scratched the surface yet. Sports analytics is a growing field with new developments happening every day and it is a domain worth looking out for.

7. Links and references

- Video Link: https://drive.google.com/open?id=1PYNzV3-7BUHznwkF0VBc2_6s0XBnMkBK
- Dataset (CSV files): <https://drive.google.com/open?id=1CCxOZ2VNdQX2aBTqK-erw8ElhxgroM8b>
- SportsReference Python API Package: <https://pypi.org/project/sportsreference/>