NBA Basketball Case Study:

What makes an MVP Special?

By Taher Darwesh

***Abstract* —** In this paper we look at historical NBA data from 1950 analysing basic and advance performance statistics to understand what makes the MVP special. Our data is aggregated at a yearly player level, we use this to prove that indeed the MVPs performance is higher than the rest of his peers. We are also able to quantify the effect he has on winning using Bayes Theorem. We then analyse positions to then find a style of play and player within each, we further this by using unsupervised learning to discover new playing styles and positions based on MVP stats using Kmeans clustering. Finally, we used supervised learning to build a logistic regression model to predict MVPs, we were unable to accurately predict MVPs.

I. Introduction

Since its inception in 1950, the National Basketball Association (NBA) has become the most iconic basketball league in the world. It is considered the pinnacle of basketball and attracts the best talent from across the globe. Since the 1955-56 NBA season, every year one player is crowned with the Most Valuable Player Award. This award is given to the best performing player during the regular season. Sports is performance based, every team is in competition to attract the best talent or to elevate their talent to build a high-level performing team, to win. Having the best performing player in the best league on the team should suggest glory in the horizons. So, it is quite important to identify what makes an MVP different and the effect they have on winning. NBA teams can benefit from this data for scouting and development.

II. Analytical Questions and DAta

A. Data Sources

1) Main Source:

- NBA Player data since 1950 from Kaggle.com [1].

2) Other Sources:

- Team names data sourced from [2][3].
- MVP data sourced from [4].
- Championships data was sourced from [5].

B. Analytical Questions

1) What makes the MVP different? What is the effect on becoming a champion?

- Importance - Discover the statistics behind the identity of the MVP. Quantifying the effect on being champion translating their value.

2. Do positions have playing styles? does this effect MVP award? Can categorisations of players be improved?

- Importance - Identify the MVP positions and whether any are more favored. Unsupervised learning will be used to re-categorize positions.

3. Can we predict the MVP based on their performance?

- Importance - to see if there are patterns amongst their performance statistics which our supervised learning model can detect and help predict them.

C. Key Data Charactersitcs

1) 'mvp':

- MVP by year as a binary calculated column

2) 'finals_won':

- Binary column for finals won. Champion players alongside teams can be highlighted.

3) 'Pos':

- Position of each player every year linked to the above two.

These characteristics are perfect in answering the questions as they allow us to see the yearly statistics of each MVP and whether they were champions answering question 1 & 3. Having Position with MVP and performance statistics will answer question 2.

I believe we will be able to identify and predict the MVP which would help basketball teams understand the DNA of the MVP and what it takes to get their current players to that level. I believe we will be able to identify different playing styles but potentially not improve on the categorizations of positions as they have never changed. If so, should teams adopt these new positions if found?

III. Analysis

The project is founded on four stages: Preparation, Analysis, Modelling and Validation:

A. Preparation

Includes the following:

- Importing
- Exploring data
- Merging
- Replacing values
- Cleaning
- Visualizing variables
- Dealing with null values

During the preparation stage the all the datasets were imported and explored. Some of the highlight information found were most MVPs won were by Kareem Abdul-Jabbar at six, the team to have the most players win the award was Boston Celtics at ten, the team to win the most championships were the Lakers and the Boston Celtics at 16 apiece and Michael Jordan is ranked first at being a champion alongside being an MVP on four occasions.

- *Data Preparation* – After merging Championships data, we decided to check if the number of rows made sense but after basic estimating, we were under. We decided to check if the maximum squad number 15 was always met. In doing so we used a group by function to count the number of players in

a squad for the 'Los Angels Lakers' for every year. We found out that the maximum squad number of 15 wasn't always used throughout history but the interesting thing we found was in the current era were the max is 15 some years the squad number was more than 15. After investigating we found out that some players get traded mid-season which means they have a row for each team they played for and a totally row. We used the function drop_duplicates on subset year and player to remove all the extra rows for each player each year. This was very important piece of analysis as leaving this data would have inflated players performances and potentially classifying them as better players than they really are.

- *Data Preparation* – After completing all merges it was time to deal with null values. Our data had a significant amount of missing data. We tackled it using two approaches, dropping years with significantly missing data and replacing with category averages. The first approach we decided to only keep the seasons ending greater than 1979. This immensely reduced the number of null values while still giving us 38 seasons of NBA data to work with. The second approach we look at how each variable was affected by either position, age or year. It was then decided to take the average of the variable grouped by either year, position and age or year and position. This would avoid a general average across a column which could distrust local averages.

B. Analysis

Includes the following:

- Feature engineering
- Bayes Theorem to answer MVP impact on winning
- Functions in conjunction with boxplot analysis to answer MVP statistical performance and MVP position style
-

At this stage the data is completely ready for analysis. One of the important stages at this point is where we engineer our own features that will be useful for our analysis and modelling

- *Data Derivation* – Given that our data is at player year level, this means all the statistics are aggregated to a complete season level which isn't a good indicator to how consistent a player is during the season. So, we engineered all the important basic statistics such at 'PTS' and 'AST' and we divided them by the number of games each season to give us a stat per game. This is a more of a consistent stat to compare players.

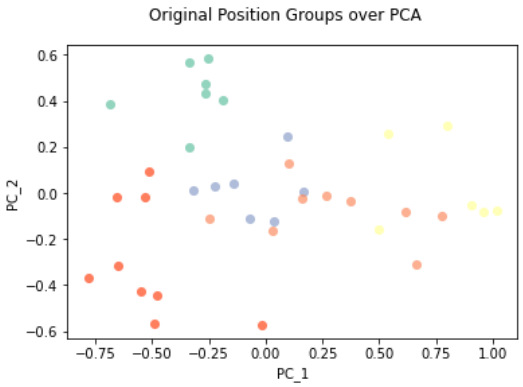
C. Modelling

Includes the following:

- Scaling
- Understanding Outliers
- Principle Component Analysis
- Correlation Analysis
- Using KMeans Clustering to improve positions based on MVP performance statistics.
- Using Logistic regression to predict the classification of a player as MVP or not

At this stage we are adapting our data to be ready to be used by our constructed model.

- *Construction of Models* – One of the models we construct is a clustering Model using a KMeans algorithm. This Algorithm is distance based so we are required to normalize our data so its within the same scale to help give more accurate results. Given that we used 9 features we then were required to run PCA on two components to help us visualize our new clusters on a scatter plot to see the difference pre and post clustering. After that we were able to use the new cluster labels and append it to the dataset to then analyse the distribution of positions in each cluster.



- *Construction of Models* – Before building the Logistic Regression model we had to implement the assumptions associated with it, such as the response variable being binary, ensuing the observations are independent from each other, there are no Multicollinearity amongst explanatory variables, there are no extreme outliers and the sample size is large enough. To ensure we met these we converted 'mvp' column to binary, removed advance stats that are calculated from basic stats and removed highly correlated variables. All the MVP points were regarded as outliers, so we were unable to remove them.

D. Validation

Includes the following:

- Computing performance metrics
- Analyzing results – pre & post
- Confusion matrix
- Using K-fold validation
- Improving the model

At this stage we have the result from our model, and we wish to understand them better and ensure they are consistent.

- *Validation of Results* – After running our Logistic regression and predicting on our testing set, we test the performance of the classification model by

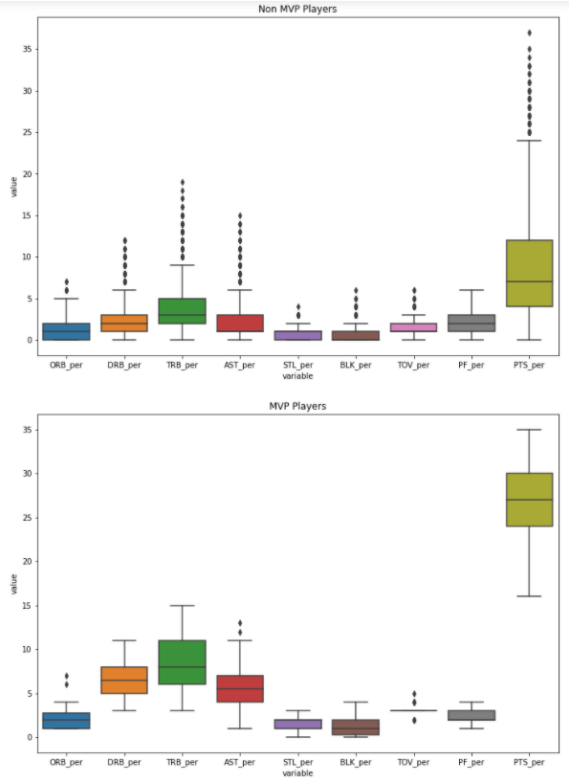
computing a confusion matrix. This method validates our model by calculating the true positives, false positives, true negatives and false negatives. We then visually express this the confusion matrix function in a heatmap to illustrate how the predicted values and true values algin. We also compute performance metrics such as Accuracy, Recall and F1 to understand the performance of our model and to compare against iterations and improvements.

IV. Findings, Reflection & Further Work

A. *Findings*

1. *What makes the MVP different? What is the effect on becoming a champion?*

As we can see from the averages plot on the right that the MVP is on average better in every important basic statistic per game especially in points scored, assists and total rebounds. From the boxplot analysis below, we can see that MVP data slice hardly has outliers meaning they are very consistent as a performing group in comparison to the rest of the population that has many outliers.



In terms of the effect on winning we used Bayes Theorem to determine probability of being a champion given that a player is awarded the MVP, which is 0.39, seen in the figure below. Each player is 6.7% of the team and if effect of winning was equally distributed each should have the same effect of 6.7% but the MVP has nearly 6 times that effect at 39%.

Bayes Theorm:

P(Champ|MVP) = P(MVP|CHAMP) * P(CHAMP) / P(MVP)

= P(MVP|CHAMP) * (0.03240188) / (0.00238619)

P(MVP|CHAMP) is unknown so we can use conditional probability to calculate it using the joint probability:

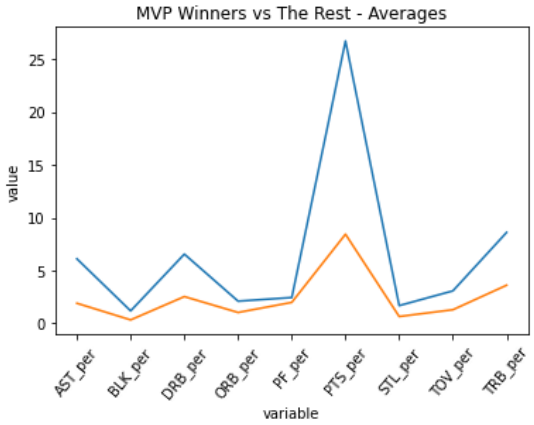
P(MVP|CHAMP) = P(MVP,CHAMP) / P(CHAMP)

= (0.00094192) / (0.03240188)

= 0.0290699

Now we can replace back into the first formula:

P(Champ|MVP) = (0.0290699) * (0.03240188) / (0.00238619) = 0.39



2. *Do positions have playing styles? does this effect MVP award? Can categorisations of players be improved?*

After using boxplots, we discovered every position has unique playing styles. Each position was coined with a tag describing its playing style after the analysis.

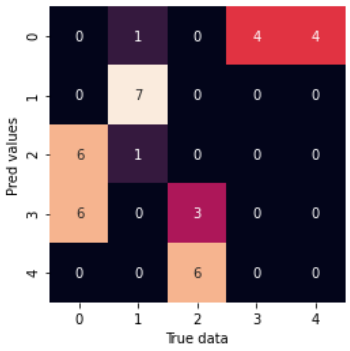
Original Positions:

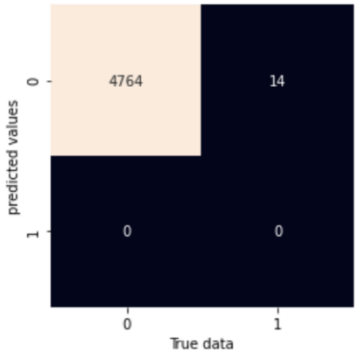
- Point Guard - “The Cool One”
- Shooting Guard - “The Talisman”
- Small Forward - “The Main Man”
- Power Forward - “The Defender”
- Center - “The Beast”

I then clustered on MVP stats, only one position and playing style remained as seen in the confusion matrix below.

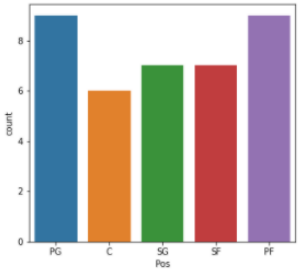
New Positions* and styles:

- Shooting Guard - “The Talisman”
- Center Forward* - “The Big Man”
- Power Guard* - “The Leader”
- Dynamic Center* - “The Machine”
- Assisting Forward* - “The Play Maker”

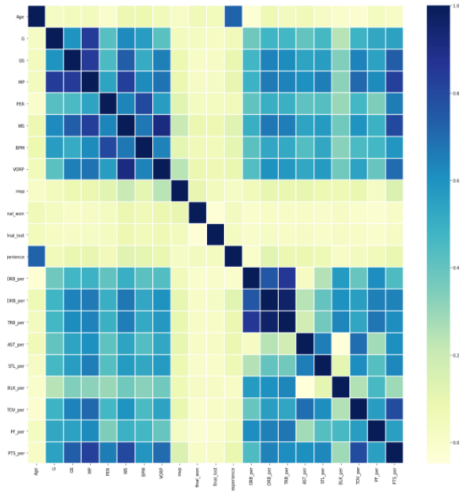




Finally, position doesn't seem to have a significant impact on the number of MVP award as seen in the fig below.



3. *Can we predict the MVP based on their performance?*
We used Logistic Regression to classify MVP based on selected basic performance statistics. We prepared the data by feature engineering, normalizing and removing correlated predictors.



Our results showed that for value"0" – Non-MVPs the model was 100% accurate and managed to classify every value correctly. As of the value "1" – MVPs the model wasn't able to predict any values and all were wrongly classified as Non – MVPs, this is shown by the recall value for "1" which is 0. We tried to improve validate the model using K-fold validation to run iteration of the model to see if we would get consistent results which we did. None of the iterations managed to highlight any MVPs correctly.

B. Reflection

I feel like I was able to answer all the analytical question I put forward successfully except creating a model that was successful in predicting MVPs, which has more to do with the dataset rather than the modelling. The dataset suffers from a common issue in classification which is class imbalance. I would think this project would be interesting to those in the professional game to see a scientific approach to the effect of the MVP and how potentially new positions and styles could be tested within the game.

C. Further Work

Try different classification techniques such as decision trees to improve the classification rate. Manipulating the data to have a more balanced spread of either class to improve the models or use an algorism that is better equipped for this kind of dataset like anomaly detection. Also, how to deal with outliers in sports when analyzing high preforming athletes as they will always outperform the population and be classified as outliers.

I. REFERENCES

[1] *https://www.kaggle.com/drgilermo/nba-players-stats*

[2] https://www.nba.com/teams

[3] http://www.apbr.org/abbreviations.html

[4] http://www.espn.com/nba/history/awards/_/id/33

[5] https://en.wikipedia.org/wiki/List_of_NBA_champions

[6] Best, C. (1987). Experience and career length in professional football: The effect of positional segregation. Sociology of Sport Journal, 4, 410–420.

[7] Altavilla, G., and Raiola, G. (2014). Global vision to understand the game situations in modern basketball. J. Phys. Educ. Sport 14, 493–496.

[8] Herzog, Brad (2003).Hoopmania:The Book of Basketball History and Trivia. Rosen Pub. Group

[9] Klein, Christopher. "How A Canadian Invented Basketball." History, 22 August 2018. Accessed September 30, 2020. https://www.history.com/news/how-a-canadian-invented-basketball

[10] Conte, D., and Lukonaitiene, I. (2018). Scoring Strategies Differentiating between Winning and Losing Teams during FIBA EuroBasket Women 2017. Sports 6:50.

II. WORD COUNT

- ✓ Abstract - 124
- ✓ Introduction - 143
- ✓ Analytical Questions & Data - 298
- ✓ Analysis - 948
- ✓ Findings, reflections & Further Work - 600