Traditional vs. Advanced Statistics and the NBA Regular Season MVP Award

Alex Spence
NYU CDS
New York, New York
ajs811@nyu.edu

Mike Urciuoli
NYU Courant
New York, New York
mlu216@nyu.edu

***Abstract*—**

Traditional NBA player statistics have been collected since the onset of the NBA in 1946. The development of advanced player statistics has grown in the NBA from the 1990s through today. This project looks to answer the question which statistics (advanced or traditional) correlate most to the regular season MVP award. The seasons used for comparison were 2002-03 to 2016-17. It was determined that on average, the metrics which most correlated to the regular season MVP were estimated wins added, value added and player efficiency rating (all advanced statistics). The analytic developed correctly predicted the MVP for 10/16 seasons studied. For further study, it is recommended to analyze more seasons and apply machine learning techniques on the most informative statistics to develop regular season MVP prediction models.

I. Introduction

Advanced player stats have grown in the NBA since 2001. Also important are traditional player statistics. This project looks to answer the question which dataset (advanced statistics or traditional player statistics) correlates most to regular season MVP award. The seasons used for comparison are 2002-03 to 2017-18. This study also intends to reveal if the MVP award is given accurately based on what the statistics reveal.

II. Motivation

The users of this analytic are NBA franchises, players, the media, and NBA fans. This analytic is important for providing an objective assessment of player performance as it relates to the regular season MVP, and to determine which statistics are the most informative. Since the media votes on this award, it is important to understand upon what criteria they may be basing the voting. This analytic can be used to determine whether the regular season MVP is awarded fairly. It can also be used to determine whether advanced or traditional statistics correlate more with regular season MVP.

III. Related Work

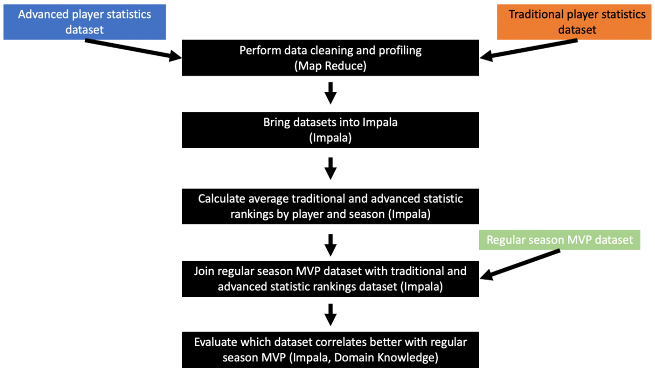
The following is a summary of “Revisiting the Correlation of Basketball Stats and Match Outcome Prediction”, which studies topics related to the analytic of this project. The author, Li, starts off by commenting that previous works have used linear regression using previous match data to predict win/loss in NBA games with 73% accuracy, but that due to the simplicity of the stats that existed at the time there was a "glass ceiling" as to how well prediction could be using this info. This has given rise to more advanced stats, RAPM and PER are named, which encapsulate more info on players and teams. Li leverages a dataset that contains NBA statistics from 2012-2018 to conduct his experimentation between the two stat groups, which totals to 7380 games. Long sliding windows looking back 15 games was important to Li's method, in an attempt to avoid overfitting. Machine learning techniques are used to train classifiers, and most importantly LASSO (Least Absolute Shrinkage and Selection Operator) is used for two purposes: regularization and feature selections. What results is 14 features most useful to Li's classifiers, which are shown in Fig. 1 as well as listed in paragraph 4.1. [1] Some of these parameters were included as part of developing the analytic for this project.

The paper, A Neural Network Model of the NBA Most Valued Player Selection, developed a neural network on NBA player statistics from 1997 to 2019 to predict NBA MVP. Three datasets were used during model development - a total set, which includes basic statistics on each player, an advanced dataset, which includes advanced analytics on each player, and a mixed dataset which included both. Some example statistics used for the "totals" section is FG%, DRB, AST, and some example statistics used for the "advanced" section was PER, TS%, VORP. The neural network was trained using mini-batch gradient descent and a validation set was used for hyperparameter tuning. A test dataset was used to test the accuracy of the model. The results were successful - the correct MVPs were predicted for both the 2009-2010 and 2016-2017 seasons. [2]

IV. Design and Implementation

A. Design Details

The following diagram describes the analytic pipeline. First, data cleaning and profiling were completed on each of the datasets using MapReduce. Then, the datasets were brought into Impala, and player rankings were calculated by season and player for each statistic. After this, the rankings were joined with the MVP dataset and various queries were performed to determine which statistics most correlated with the regular season MVP award. Other results were investigated as well through Impala SQL queries.


Advanced player statistics dataset
Traditional player statistics dataset
Perform data cleaning and profiling (Map Reduce)
Bring datasets into Impala (Impala)
Calculate average traditional and advanced statistic rankings by player and season (Impala)
Regular season MVP dataset
Join regular season MVP dataset with traditional and advanced statistic rankings dataset (Impala)
Evaluate which dataset correlates better with regular season MVP (Impala, Domain Knowledge)

V. Datasets

A. Advanced Player Statistics (2002-03 to 2017-18) [3]

This dataset is 0.6 MB and includes advanced NBA player statistics from 2003-04 to 2018-19. The dataset can be found on Kaggle. Full descriptions of each variable are shown in the Appendix. The schema is as follows (except where noted, all variables are of the data type float): index (int), rank (int), games played (gp) (int), mpg (minutes per game), true shooting percentage (ts%), assist ratio (ast), to ratio (to), usage (usg), offensive rebound rate (orr), defensive rebound rate (drr), rebound rate (rebr), player efficiency rating (per), value added (va), estimated wins added (ewa), player (string), team (string), and season (string).

B. Traditional Player Statistics (1949-50 to 2016-17) [4]

This dataset is 5.1 MB and includes traditional NBA player statistics from 1949-50 to 2016-17. The dataset can be found on Kaggle. Full descriptions of each variable are shown in the Appendix. The schema is as follows (except where noted, all variables are of the data type float): index (int), year (int), player (string), position (pos) (string), age (int), team (tm) (string), games (g) (int) , games started (gs) (int), minutes played (mp), field goals made (fg) (int), field goal attempts (fga) (int), field goal percentage (fg%), three pointers made (3p) (int), three point attempts (3pa) (int), three point percentage (3p%), two pointers made (2p) (int), two point attempts (2pa) (int), two point percentage (2p%), effective field goal percentage (efg%), free throws made (ft) (int), free throws attempted (fta) (int), free throw percentage (ft%), offensive rebounds (orb) (int), defensive rebounds (drb) (int), trb (total rebounds) (int), assists (ast) (int), steals (stl) (int), blocks (blk) (int), turnovers (tov) (int), personal fouls (pf) (int), points (pts) (int).

C. Regular Season MVPs (1949-50 to 2018-19) [5]

This dataset is 4 KB and includes the regular season MVPs from 1949-50 to 2018-19. The dataset can be found on NBA.com. The schema is as follows: (year (string), player (string), team (string), player_team (string), and mvp_index (int).

VI. Results

One challenge with the datasets was that the traditional dataset contained more players than the advanced dataset. This was solved by treating the datasets independently, then joining on the MVP dataset for the final average rank calculations. In addition, some cleaning was performed on the advanced dataset to convert years from "YYYY-yyyy" to "yyyy". All queries were performed using Imapala and the results are shown below.

1. Most informative advanced statistics by avg mvp ranks: ewa (3.8) / va (3.8) / per (4.1) / usg (10.6) / ts_perc (35.9)
2. Most informative traditional statistics by avg mvp ranks:
   pts (12.6) / fg (12.67) / tov (15.2) / ft (17.1) / ast (17.6)
3. Composite advanced statistics are the most informative overall (ewa, va, per)
4. Advanced statistics seem to be valued less in earlier years:
   2003 to 2010 avg MVP ewa (5.75 rank / 593 value)
   2009 to 2017 avg MVP ewa (1.67 rank / 810 value)
   2012 to 2018 MVP ewa rank is 1 for 5/7 seasons
5. 2005 and 2006 has the lowest ranked ewa MVPs Steve Nash (18 rank, 15 rank)
6. An analytic was developed to predict the MVP of each season. The analytic is based on the highest average player rank of ewa, va, and per for each season. This analytic correctly predicts the MVP for 10/16 seasons studied.

It was somewhat expected that the most informative statistics were from the advanced dataset. Advanced statistics were designed to reveal more about overall player performance than traditional statistics. In addition, it is intuitive that advanced statistics more closely correlate with MVP in more recent years. It takes time before new ways of evaluating players are commonly accepted and the most informative advanced statistics of ewa, va, and per were introduced into the mainstream in the early 2000s.

VII. Future Work

Given more time, it is recommended to bring the data into Tableau to create dashboards for further analysis. In addition, it is recommended to connect to the NBA API and get advanced statistics and traditional statistics for all available years. Then the analytic could be rerun on more years and the conclusions could be verified or tweaked. It would also be informative to include more analysis on a player's team performance as it relates to awarding the regular season MVP. Finally, machine learning techniques could be applied to develop a more robust prediction model.

VIII. Conclusion

MapReduce and Impala were used to develop an analytic to compare traditional and advanced statistics and evaluate the correlation to the regular season MVP award. Advanced statistics which correlated most with regular season MVP were ewa, va, and per. Traditional statistics which most correlated with regular season MVP were pts, fg, and tov. It was confirmed that the advanced statistics more closely correlated with MVP than the traditional statistics. It was also concluded that as time increases, advanced statistics more closely correlate with MVP. The goodness of the analytic was verified by crosschecking statistics/rankings on NBA.com [6] and ESPN.com [7], as well as by developing an analytic to predict the regular season MVP. It is recommended to continue verifying the analytic by including more seasons and looking at team statistics for each player.

Acknowledgment

Thank you to Prof. McIntosh for answering our analysis questions, to Kaggle for compiling the data sets, to HPC for providing the technology to make this project possible, and to the NBA for providing a medium for professional basketball.

References

1. Zovak, Sarcevic, Vranic, Pintar Game-to-Game Prediction of NBA Players' Points in Relation to their Season Average link
2. Chen, Dai, Zhang A Neural Network Model of the NBA Most Valued Player Selection link
3. https://www.kaggle.com/hultm28/nba-player-hollingers-stats
4. https://www.kaggle.com/drgilermo/nba-players-stats/home?select=Seasons_Stats.csv
5. https://www.nba.com/history/awards/mvp
6. https://stats.nba.com/leaders/
7. http://insider.espn.com/nba/hollinger/statistics

Appendix

Results of data profiling