Roni Kaakaty

8/14/2021

DSC630-T302- Predictive Analytics

Milestone#5

**Predicting NBA Wins**

**Executive Summary**

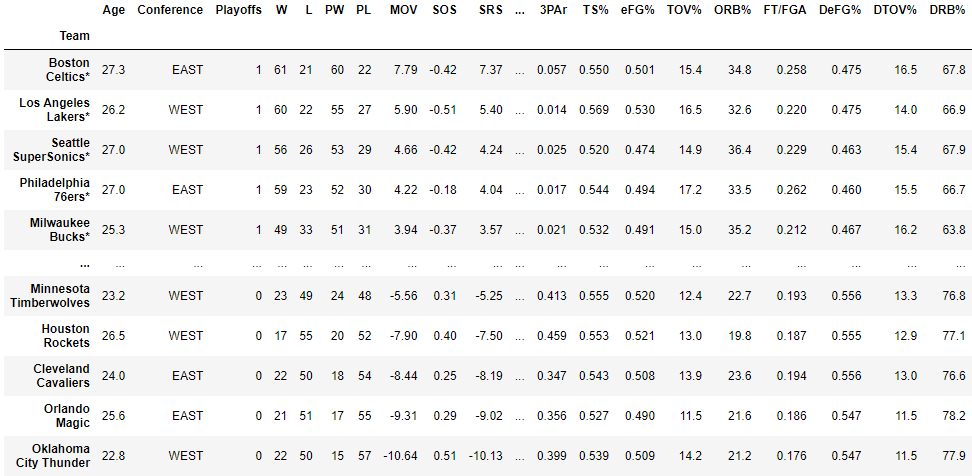
Making the NBA playoffs is a lucrative financial proposition for team owners. Other than the notoriety and publicity that goes with being on national TV for an extended period of time, teams that make the playoffs are also entered into a playoff pool that will provide incentives based on the performance of the team in the playoffs. In 2020, it was estimated that the playoff pool was around $24 million dollars. My model aims to accurately identify what variables will contribute the most to wins resulting in playoff qualification based on historical data, beginning with the 1979-1980 regular season. The model will incorporate various features that positively correlate with wins. With this information, business owners will be able to see if their team is expected to fall short of expectations and they can rectify the situation by improving aspects of their team that aren’t as strong as others. My findings found that business owners should target offensive players when trying to increase their win total. Teams that made the playoffs from each conference scored better in various offensive categories than those that didn’t qualify for the playoffs, whereas their defensive metrics didn’t differ as much. I was able to create a model that had an accuracy of 61% in being able to determine win output.

**Introduction**

The topic I chose to do my final project on is predicting whether or not a team would have enough wins to make the playoffs based on advanced metrics. The NBA is a professional basketball league in North America that consists of 30 teams. Those 30 teams are broken equally into two separate conferences, the Eastern and Western conferences. Each conference is made up of 3 divisions that include 5 teams each. Since the playoffs have already started, I will have a source of truth to see how accurate my models are. Making the playoffs is a goal of every organization as it provides the team an opportunity to advance each round and compete for an NBA championship. Teams from each conference play an 82-game regular season schedule, 41 games at home and 41 games away from home. At the end of the regular season, the top 8 teams in order of wins from each conference then qualify for the playoffs. While in the playoffs, these teams play a best of seven series in bracket format (1 seed vs. 8 seed, etc.) to determine who advances. To create the model, I used historical data from as far back as the 1979-1980 NBA regular season as that is when the 3-point line was introduced. I do not want to include data before that time since the numbers will be skewed too drastically. The data sets I am using will provide me with the team statistics broken down into different quantitative categories that I will then be able to train the model on. I will be able to see what features contribute to wins the most, which in turn will provide business owners the information they need to see what areas of focus they should place in the off-season to improve their team.

Methods

This project utilized advanced stats metrics found on basketball-reference. I have taken each individual season’s data and have compiled all the information into one master dataset. I had to create an additional column to add the conference designations as I realized the need to be create two different models, based on the team’s conference. I also included another column that notated whether a team made the playoffs or not. I will need one for the East and one for the West, due to the difference in number of wins needed to make the playoffs in each conference. After making sure all the data in the dataset was aligned in the proper format, I loaded the CSV file into Jupyter Notebook using Python. After the file loaded, I used the dropna function to get rid of unnamed columns and missing values.

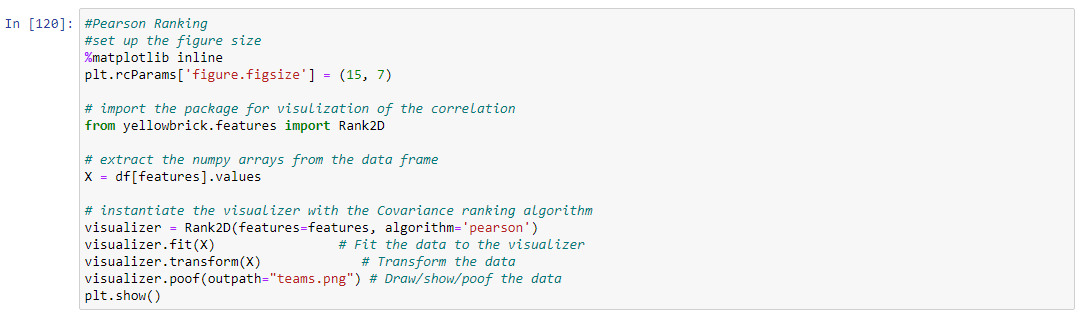


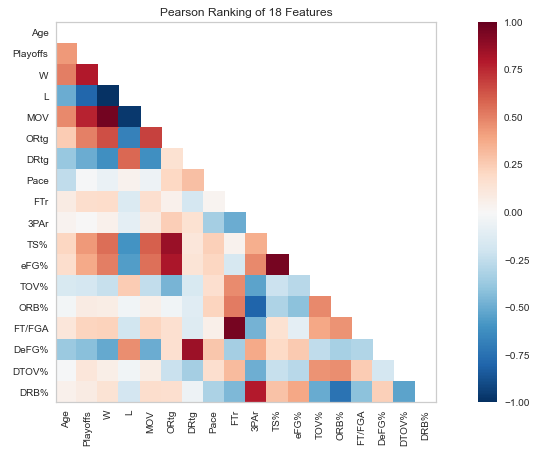
The cleaned dataset was now ready for me to use. The variables that I have included in my model are as follows:

* Age – Average age of the team.
* Playoffs- Did the team make the playoffs? (0 = No, 1 = Yes)
* W- Number of wins during the regular season.
* L – Number of losses during the regular season.
* MOV – Margin of victory.
* ORtg- Offensive team rating.
* DRtg- Defensive team rating.
* Pace- Estimate of possessions per 48 minutes.
* FTr- Free throw attempt rate.
* 3PAr- 3pt. attempt rate.
* TS%- True shooting percentage that is a measure of shooting efficiency.
* eFG%- Effective field goal percentage.
* TOV%- Offensive turnovers committed per 100 plays.
* ORB%- Offensive rebound percentage.
* FT/FGA- Free throws per field goal attempt.
* DeFG%- Defensive field goal percentage.
* DTOV%- Defensive turnover percentage.
* DRB%- Defensive rebound percentage.

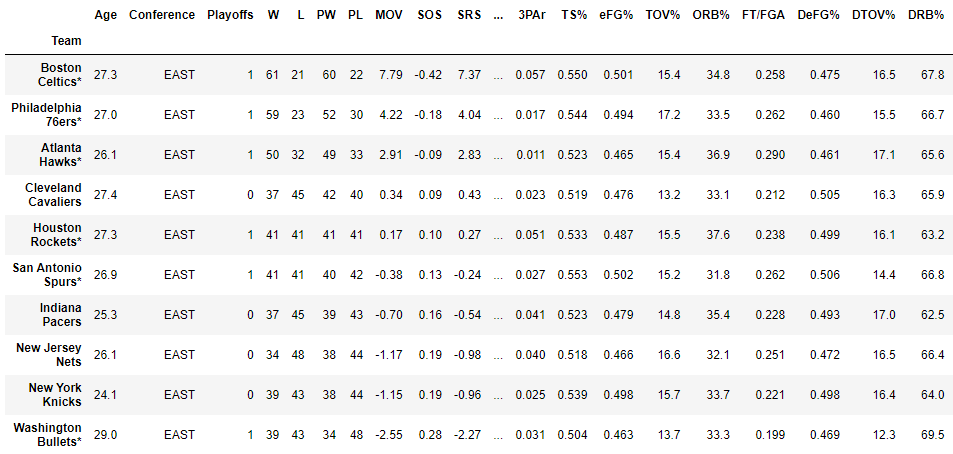
Results

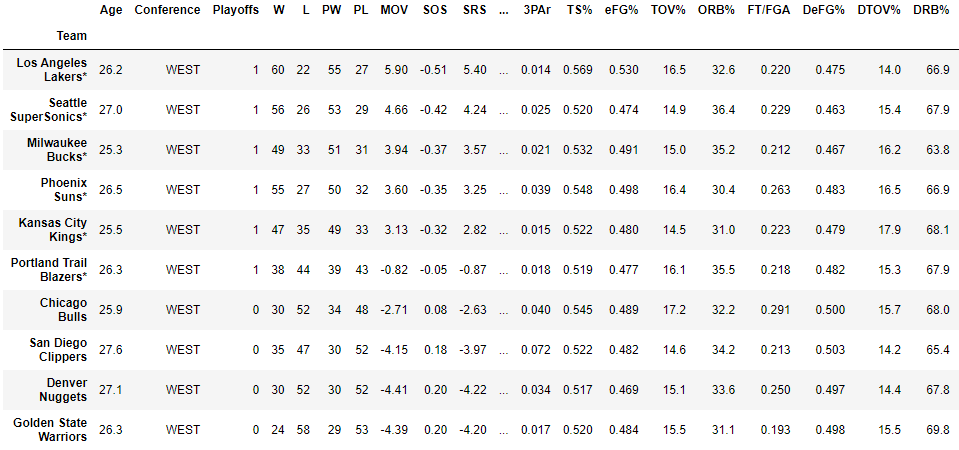
After construction my feature list of the variables I am looking to use, I proceeded to create a Pearson’s correlation visualization that provided me with a heatmap of how correlated two variables are with each other.



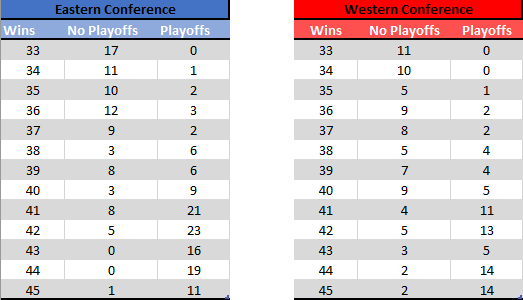


What was interesting about this visualization is that it illustrated the importance of the offensive categories relative to whether a team made the playoffs or not. The adage “defense wins championships” might be true in the playoffs, but it looks like the offense is what gets teams to the playoffs. Offensive rating, true shooting percentage, and effective shooting percentage had a positive correlation with making the playoffs. Whereas typically strong defensive metrics such as defensive rebound percentage and defensive rating, did not play as large a role. Now that I had a general idea of what advanced metric contributed to wins the most, my next step was splitting out the data into an Eastern conference dataset and Western conference dataset to see the number of wins typically required at minimum in each conference to qualify for the playoffs. I made sure to omit the 1995, 2011 and 2020 seasons due to the NBA lockout and COVID pandemic, which impacted the number of games played in the regular season.

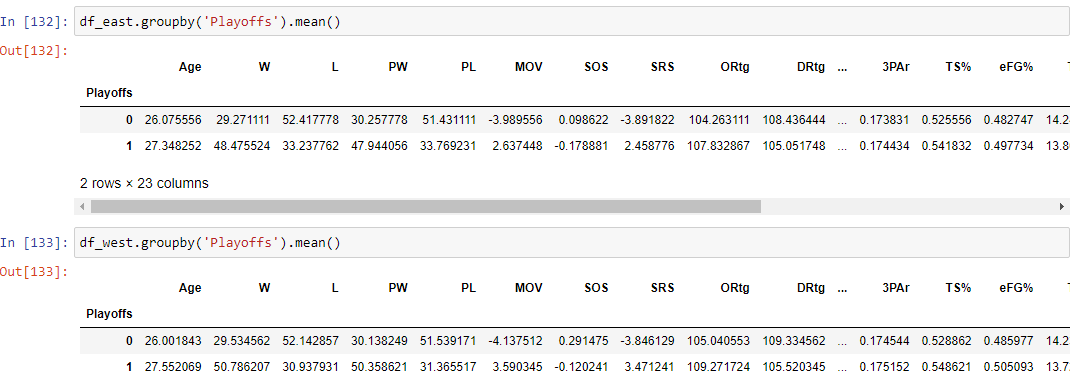




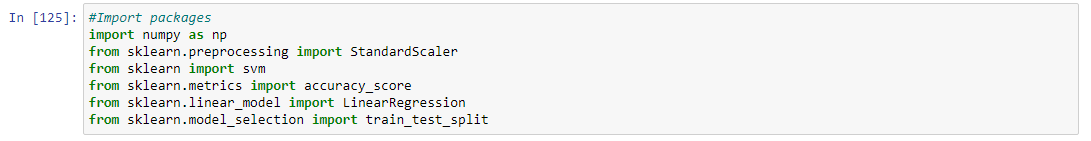
After I split the dataset based on the teams’ respective conference, I used R to create a table that populated the Playoffs vs. Wins column to get an idea of the minimum number of wins required to make the playoffs in each league. The table shows that the Eastern Conference has had more teams make the playoffs with fewer wins required than the Western Conference. Since 1979-1980, more teams have made the playoffs in the East with only 38 wins, than in the West. The likelihood of making the playoffs in the East with more than 42 wins is far greater than what the odds would be in the West.

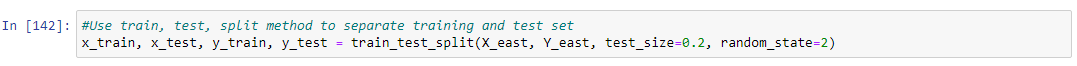


Then I incorporated the groupby function in Python to provide a side-by-side breakdown of the mean attributes in each variable between playoff vs. non-playoff teams in each conference.

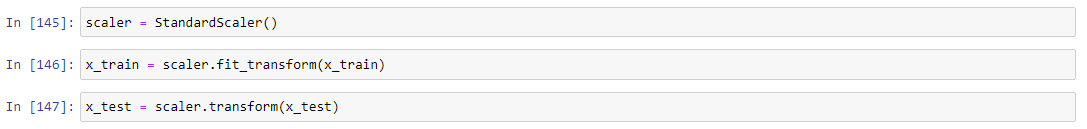


The average number of wins of a playoff team in the East was 48, whereas in the West it was 50 wins. We also see a significant difference of 2 points in offensive rating between Eastern conference playoff teams vs. Western conference playoff teams. After identifying what the minimum thresholds should be for each conference, my next step involved training the model along with scaling my data set. I used Sklearn to train, test and split the model.

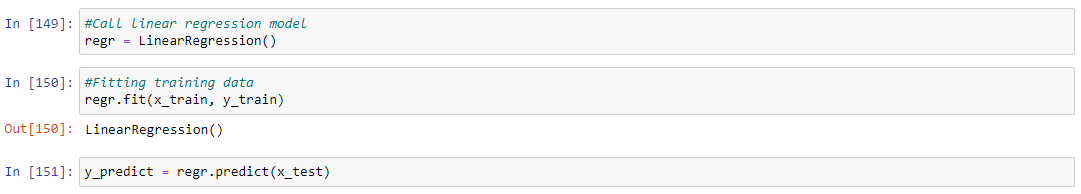




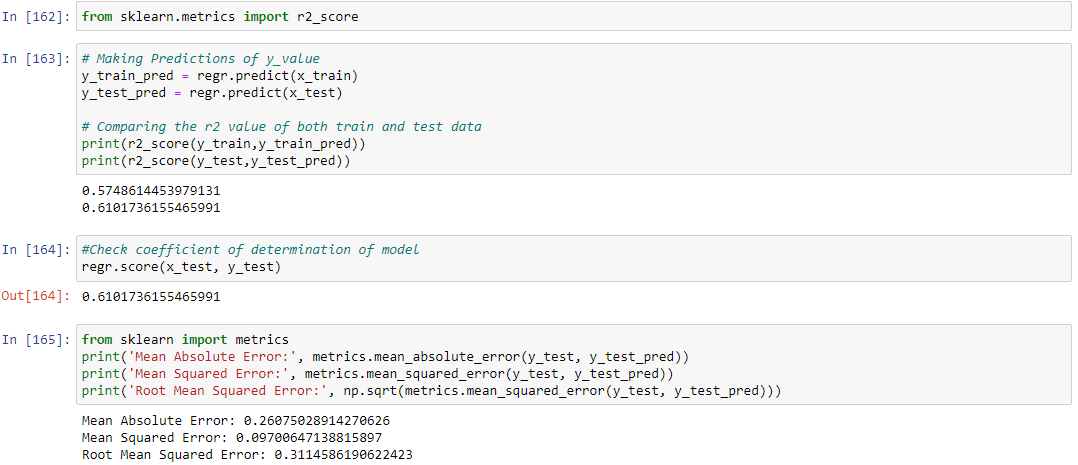
I utilized the StandardScaler package to scale my data:



I used a linear regression model to help predict the number of wins as I am targeting a continuous variable.



After fitting my linear regression model, I used Sklearn to capture the accuracy score and provided me with other metrics that validated my model.



The model had a low mean absolute error of .26, a mean squared error of .09 and a root mean squared error of .31. The training model had an accuracy score of 57% and a test score of 61%. There isn’t a wide range between the two, which validates the accuracy score of the model.

Discussion/Conclusion:

Based on what I have discovered, it can be determined that teams in the Eastern conference have a lower threshold to meet to make the playoffs. I have also found that offensive attributes have contributed to making the post season more so than defensive attributes. I suspect that is due to the various rule changes that have occurred over time that have benefited the offensive player. Rule changes such as no longer being able to put your hand on an opponent on the perimeter, defensive 3 seconds in the paint, and goaltending, have all contributed to the rise of all the offensive categories. The league is in the business to entertain, so making it easier to score more points is in their benefit. With this model, team owners will be able to put more of an emphasis on acquiring players that excel in offensive categories. For future considerations, I will analyze if the same categories correlate to wins as strongly in the playoffs as they do in the regular season. This model shouldn’t be used as the end all be all when making team construction decisions, however it has value as a tool that can help validate a team owner’s reasoning when making personnel decision.

Acknowledgements:

I would like to thank basketball-reference for containing all the historical datasets I needed for this project. Their database is impressive and easy to scrape from. I would also like to thank Bellevue University for teaching me the skills necessary to complete such a detailed real-world project.

References:

Accounting for basketball: The money behind the madness. (n.d.). Retrieved from https://www.kenan-flagler.unc.edu/news/accounting-for-basketball-the-money-behind-the-madness/#:~:text=In the NBA, there is,playoff dollars however they wish.

Burd, B. (2020, October 01). How Much Do NBA Players Make in the Playoffs? Retrieved from <https://thesportseconomist.com/how-much-do-nba-players-make-in-the-playoffs/>

How NBA Analytics is Changing Basketball: Merrimack College. (2020, June 25). Retrieved from <https://onlinedsa.merrimack.edu/nba-analytics-changing-basketball/>

Jj. (2016, March 23). MAE and RMSE - Which Metric is Better? Retrieved from <https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d>

McFarlane, G. (2021, May 19). How The NBA Makes Money: The Second-Largest Sport in the Country. Retrieved from https://www.investopedia.com/articles/personal-finance/071415/how-nba-makes-money.asp

1979-80 NBA Season Summary. (n.d.). Retrieved July 05, 2021, from <https://www.basketball-reference.com/leagues/NBA_1980.html>

Reynolds, T. (2020, October 10). 10 things to know as 2020 NBA playoffs begin. Retrieved from https://www.nba.com/news/10-things-know-playoffs-begin