Abstract

Data is information and information provides useful insight to help its users make better decision. In order to make impactful decisions, data must be gathered and put through rigorous testing to see what it outputs. For this linear regression project, I chose to look at NBA Team stats spanning across 20 years, to predict win percentage.

Design

The purpose of the project was to use machine learning models to predict the winning percentage for the past 20 seasons of the NBA by team. A deeper insight into the stats could possibly point to a specific and/or group of stat columns that have the greatest influence on winning percentage. Teams can build a roster and/or obtain talent to capitalize on those "influential stat" to product a winning team.

Data/Algorithms

20 years of data was web-scrapped using BeautifulSoup. The primary source was from ESPN. I gathered team stat data and their respective winning percentage for each year. Before starting the process of predicting the winning percentage, the data had to be cleaned first. Since the data was scrapped the data type had to be changed to reflect a numerical data-type. Also had to make sure that there were no "NA" data.

Ran the data through some initial tests to make sure there were no outliers and/or influential data-points that would skew the analysis.

Models:
A simple linear regression model and a polynomial model were used before settling on a simple linear regression model. After running both models the linear regression model produced the closest score between the training set and the validation set.

Model Evaluation and Selection:
The testing data was split into a 80/20 split and the training/validation data was split 75/25. After selecting the simple linear regression model, I proceeded to fit the model and it produced a R^2 value of 0.81 and MSE of 0.005. Although, with more time and tweaks the ideal scenario would have been to output a R^2 value closer to 0.90 or above.

Tools

NumPy and Pandas for data manipulation
Matplotlib and Seaborn for plotting
BeautifulSoup for web-scraping
Scikit-learn for modeling