Metis Project 2: Written Description

By Will Carnevale

Abstract:
The goal of this project was to apply Dean Oliver's "Four Factors of Basketball Success" on a player-to-player basis. In 2002, Dean Oliver identified what he considered the most important factors in predicting a basketball team's success:

1. Shooting (40%) → positive
2. Turnovers (25%) → negative
3. Rebounding (20%) → positive
4. Free Throws (15%) → positive

The percentages were the approximate weight Dean Oliver assigned to each factor. His model would predict the number of wins a team might get. I wished to use similar model based on individual stats which reflected the same key aspects as Dean Oliver's model in order to predict a player's win shares.

Design: *(Hypothetical context)*
This project originates from the Golden State Warriors Organization's player development department. In light of more recent trends in NBA playstyles, the Golden State Warriors wished to re-assess how they view a player's impact on a game. To this end, they have hired me to scrape the web for data from the most recent NBA season (2020-2021) and perform an exploratory data analysis. In particular, the GS Warriors wish to have a regression model on how certain key player features might impact win shares.

Data:
For this project, I drew my data from two different sources. The first source was nba.com, which I used to obtain data on the height and weight of all NBA players. The second source was Basketball Reference. I utilized multiple tables (Totals, Per Game, Advanced) from this source, and drew upon multiple columns in order to obtain a variety of features for my regression model.

Algorithms:
1. Web-scraping the datasets using Selenium and pandas
2. Cleaning individual dataframes in preparation for multiple merges on "Player" column
3. Re-scaling desired columns as necessary
4. Performing ordinary least squares linear regressions on the cleaned data to begin determining how well fit the model is, or which variable(s) express enough multicollinearity to justify removing (according to their p-values)
5. Introducing new engineered features to attempt to improve the model's fit
6. Performing Lasso to reduce the model to its most important features

Tools:
1. Selenium for primary webscraping
2. Pandas for data manipulation and web-scraping
3. Matplotlib and Seaborn for plotting data
4. Sklearn and Statsmodels.api for regression operations
5. Unidecode to deal with odd characters in strings

Communication: All slides and visuals will be available on my github as a public repo called metis-project-work.