Regression Project Write-Up

*Predicting the Win% of a NBA team based on season average stats.*

**Abstract**

The goal of this project was to build on the regression tools learned and put them into tangible use. Specifically, this project hopes to predict the win % (total wins / total losses) of a NBA team using that team’s per game stats for the season. The project begins with a webscraping of the official NBA website. The data was then scrubbed using best exploratory data analysis practices and finally used to create regression models.

**Design**

The data was pulled from [NBA.com](https://www.nba.com/stats/teams/traditional/?sort=W_PCT&dir=-1&Season=2020-21&SeasonType=Regular%20Season)’s website and begins with the 1996-97 season. The website tracks 26 features for each team on a per game and total basis. Using regression models based on the available features, teams could predict what possible categories their players and coaches should focus on to achieve success – as measured by Win %.

**Data**

The data set contains 742 rows and 26 features. Majority of the features are traditional counting stats, such as Field Goal Attempt / Made, Assists, Points, Rebounds, Blocks, and Steals. After running a multicollinearity and p-values analysis, a final selection of 12 features were used for the modeling.

**Algorithms**

*Feature Selection*

The data set was fit and scored into a regression to calculate the initial p-values and identify statistically insignificant values. Additionally, a multicollinearity assessment was made to rule out any data features that had high correlation to each other.

The remaining features were then split into an 80/20 train vs. test dataset and standardized. The training data was partitioned into a 5-fold cross validation model and put through linear, LASSO, and ridge regression models. The resulting R-Squared result showed that there was no material difference between the linear and ridge regression models. The linear regression model was chosen to keep the analysis simple. The resulting R-Squared and MAE were 0.7196 and 0.0063 (0.63% of Win %), respectively.

**Tools**

* Selenium and BeautifulSoup was used to webscrape the data
* Pandas stored the data into a dataframe
* Numpy, sklearn, and statsmodel were used to clean the data and build regression models
* Seaborn was used to create illustrations

**Communication**

Results were presented via a PowerPoint presentation.