Predicting an NBA Player's Salary

Using Regression Modeling

Abstract: Is it possible to predict an NBA player's salary based on their game stats? Here I will be utilizing regression modeling to do just that. I'll be iterating through different models with the use of regularization and feature engineering tools to see which will give us the best performance based on R-square values, the accepted industry metric, as an indication of accuracy.

Design: The objective of this project is to use regression modeling to predict an NBA player's salary based on their game statistics, age, and season.

Data: Data was gathered on every player during a regular season from the years 1996 to 2022 from the official NBA website (www.nba.com). Each row of data represented a player during that given season (year) and their game stats, such as: age, games played, wins, losses, minutes played, total points, field goals made, 3 points made, free throws made, turnovers, rebounds, assists, etc. Data on salary was gathered from HoopsHype (www.hoopshype.com). Each row of data represented a player during a given season (year) and their associated salary for that year. Ways in which data were cleaned included: deleting rows with null values, updating a player's name from one table to match the other table, cleaning up column names, cleaning up values that had symbols (i.e. '$' or ','). Once data were cleaned, the two tables were joined together using an inner join on the player and season. In total, there were over 11,000 rows of data analysis.

Tools: Data was acquired through web scraping with Selenium and parsed using BeautifulSoup. All analysis was done in Jupyter Notebook with python and python libraries: Matplotlib, Seaborn, Pandas, Numpy, Sci-kit Learn, and statsmodels.

Algorithms: I started with 27 features, some of which appeared to have multicollinearity issues. However, I decided to leave them in and allow the modeling to eventually eliminate them for me. A base model (simple linear) was ran with all the features and "SALARY" was used as the target. The initial results were R squared values of 0.504 (train) and 0.525 (test). I then tried regularization, standardized the scale, then ran the model with Ridge Regression and Lasso Regression, the latter to help simplify my model. Both produced the exact scores as the base model, indicating that my model may be underfit and could benefit from adding more features.

I then applied polynomial transformation to the features and ran it in conjunction with the initial 3 models (simple, ridge, and LASSO). The R square value indeed improved with the best performative model being a LASSO with polynomial transformation. When analyzing the residual plots, there were normality issues that appeared, and so I went back and decided to try a log transformation of the target ('SALARY') to see if it would fix this issue. Unfortunately, the log transformation model produced a lower R squared value, so I decided to remain with the LASSO regression and Polynomial Transformation model. My final score came out to be 0.601 (train) and 0.582 (test). Ideally, the gap could be smaller between the two, but at least we saw a

progressive improvement through the iterations. Examining the absolute values of the coefficients revealed that the most important variables were the product of ‘GP’ (Games Played) and ‘PTS’ (Points), although this product showed a negative correlation to ‘SALARY’. ‘AGE’ was the most important variable affecting ‘SALARY’ in a positive way. My final diagnostic plot revealed that my model would perform poorly at the low and high ends. The plot had outliers at both ends.

Future Works: The R square value for this model was pretty low, just below 60%. I could definitely add more features if I had more time, such as creating dummy variables for the different teams that the players were on, or the number of times injuries excluded them from games. My predictive model did not produce any useable results. It produced a predicted value that is the median of all the salaries.