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 salary for every NBA player for the past 2 seasons, 2019 - 2021, to predict salary and which stat column has the biggest impact in salary payout.

Design

The purpose of the project was to use machine learning models to predict salary pay for every NBA player. A deeper insight into the stats could possibly point to a specific and/or group of stat columns that have the greatest influence on salary payout. With this information, GM/Owners can better provide contract to their players to optimize a winning team with the right amount of contract.

Data/Algorithms

2 seasons of data was web-scrapped using BeautifulSoup. The primary source were from Hoopshype (salary information) and Basketball-reference (for player stats). After cleaning the data, I was left with around 720 data points across 11 stat columns

Models:

The following models were used in this project to determine which model was the most effective: (all R^2 values are with the test dataset)

1. Simple Linear Regression Model
    1. MAE: $1,996,622
    2. R^2: 0.212
2. Polynomial
    1. MAE: $1,999,022
    2. R^2: 0.213
3. Lasso
    1. R^2: 0.210

Model Evaluation and Selection:

The model suffered from the fact that there was multi-collinearity present in the dataset. Removing certain variables helped with the issue, but there is a still a lot work needed to be done in order to fix the multi-collinearity issue. With that in mind, I split the data using a 80/20 split and ran three separate models. All models had a similar R^2 value indicating that one model wasn't particularly stronger than another model.

Tools

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