Gulay Samatli-Pac

Best NBA Players, Value for Money

Web Scraping and Linear Regression Project to Predict Salaries

1. Introduction

In this project, I will explore the NBA player's performance statistics to predict a player's annual salary based on his game statistics throughout the season. There are two potential business cases for this project. The first one is to identify players who have potential to improve their team with great value. This information will help a team manager who is exploring a new team mate to his/her team. On the other hand, the result of the project will also help an athlete and his agent to identify the real value of him in the league. This information will give an athlete negotiation power during the recruiting process.

2. Data and Tools

The data is scraped from the Basketball Reference website using BeatifulSoup. The metrics that have been collected are:

- Players annual salaries,
- Players per game statistics:
    - Age -- Player's age on February 1 of the season
    - G -- Number of Games
    - GS -- Number of Games Started
    - MP -- Minutes Played Per Game
    - FG -- Field Goals Per Game
    - FGA -- Field Goal Attempts Per Game
    - 3P -- 3-Point Field Goals Per Game
    - 3PA -- 3-Point Field Goal Attempts Per Game
    - 2P -- 2-Point Field Goals Per Game
    - 2PA -- 2-Point Field Goal Attempts Per Game
    - eFG% -- Effective Field Goal Percentage (adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.)
    - FT -- Free Throws Per Game
    - FTA -- Free Throw Attempts Per Game
    - ORB -- Offensive Rebounds Per Game
    - DRB -- Defensive Rebounds Per Game
    - TRB -- Total Rebounds Per Game
    - AST -- Assists Per Game

- STL -- Steals Per Game
- BLK -- Blocks Per Game
- TOV -- Turnovers Per Game
- PF -- Personal Fouls Per Game
- PTS/G -- Points Per Game

There are a total 30 NBA teams and more than 400 players play in each season. The data covers statistics for 3 seasons, season 2019, 2020 and 2021.

The following tools are used to explore and analyze the data:

- BeautifulSoup for data scraping,
- Pandas, Numpy, statmodels and scikit-learn for EDA, data cleaning and regression modeling,
- Matplotlib and Seaborn for visualizations.

3. Approach

After collecting the data, I have simultaneously conducted the exploratory analysis and cleaned the data by removing duplicates, dealing with missing values. The purpose of the project is building a model with a significant prediction power to identify the underpaid players. I have removed about 8% of top paid players and players who played less than 20% of the season to have a broad representation of the majority of players and to decrease the bias.

The initial model is with all numerical features using Salary data without outliers. Linear Regression model using kfold = 5 gives cross_val_scores that change between 44% and 57% (13% higher the lowest one) for which implies overfitting.

More than 10 regression models including LASSO, Polynomial, Ridge, ElasticNet with GridSearchCV to the raw target variable and also log transform of the Salary features are built to find a good fit to predict the salaries. From these models, a LASSO model with Polynomial features (degree = 2) provides the best out of sample Rsquare (r^2 = 0.60). I used this model to identify 15 most undervalued NBA players (i.e. with highest negative error).

4. Next Steps

- Add more metrics to describe a player's performance: injuries, years of experience.
- Add more features to capture that might indirectly affect salaries: team market size, expected fan-based revenue, merchandising and advertising.
- Apply different models which require less assumptions about the data like Decision Tree Regressor, Random Forest Regressor, Gradient Boosted Tree Regressor.
- Apply deeper feature selection to simplify the model without losing the prediction power.