Linear Regression Project by Med Billian

Predicting an NBA Player's Salary using NBA Statistics

Abstract

The goals of this project are to identify which game statistics are the best in predicting the Salary of NBA players; and to create a model to help predict future NBA Salaries. I scraped data from https://www.basketball-reference.com and https://hoopshype.com to consolidate NBA player's game statistics and salaries from 2014 to 2021.

Design

Predicting an NBA player's future salary accurately via machine learning models would greatly help the employers decide the salary payout to their players. Also, this project was inspired by the question "Is Steph Curry Overpaid?". Thus, this would help employers determine which players in their team are overpaid or underpaid.

Data

I scraped the Player Statistics data and the Salary data from the sites mentioned above. I then matched the salary data of an NBA player from a specific year with his game statistics from the previous year.

15 Features include: Year, Position, Age, Games Played, Points per Game, Total Rebound per Game, Assists per Game, Field Goal per Game, Field Goal Percent, Effective Field Goal, Block per Game, Turnovers per Game, Personal Foul per Game, Minutes Played, Salary Cap.

Initial Target: Salary

Final Target: Salary over Salary Cap (for that year).

1. Training Data

    - Salary from 2015 - 2018;
    - Game Statistics from 2014-2017.
2. Validation Data
    - Salary from 2019 - 2020;
    - Game Statistics from 2018-2019.
3. Test Data
    - Salary from 2021;
    - Game Statistics from 2020.

Algorithms

Feature Engineering

1. Feature Standardization
2. Categorical features to binary dummy variables
3. The final target was derived by dividing a player's salary by the yearly salary cap, and then getting the square root of the quotient. This was done to account for yearly salary increases and the variance of error in the initial modeling.

Models

Multiple regression, RidgeCV, LassoCV, and Elastic CV models were used to train 6 different models. Both the RidgeCV and LassoCV models displayed the highest performance. But during validation the RidgeCV model showed a slightly higher R^2, hence was decided to be the final model for predicting future salary data.

Features Selection, Model Evaluation and Selection

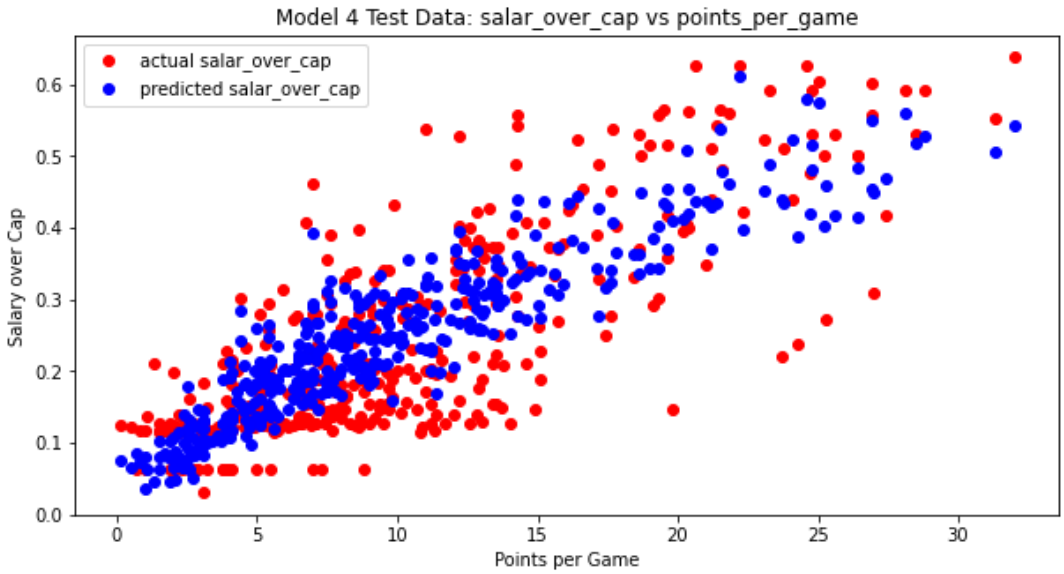
1. *Residual Error plotting for Error Variance*
2. *Cook's D to ensure there are no influential points used to train the model.*
3. *Seaborn PairPlot & Variance Inflation Factor to ensure there are no strong collinearity in the model*
4. *QQ Plot for Error Independence*

1480 data records were used to train models using data; 740 data records were used to validate models, and 392 data records were used to test the final model.

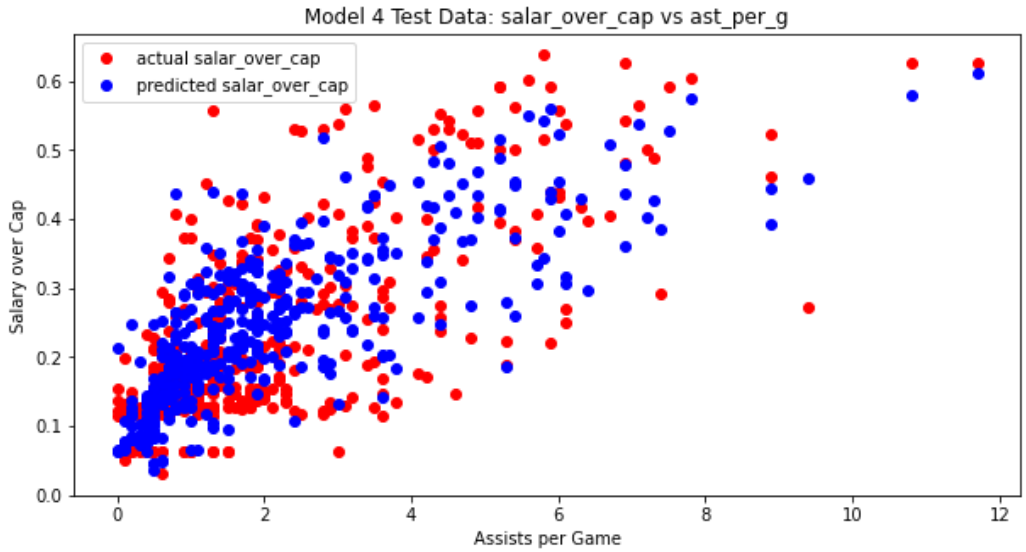
The official metric used to evaluate the final model was R^2, while the coefficient values of the model were used to determine that the ‘Points per game’, ‘Total Rebounds per game’ and ‘Assists per game’ are the best features in predicting a player’s salary.

Ridge 5-fold CV score on combined Training and Validation data: R^2: 0.61

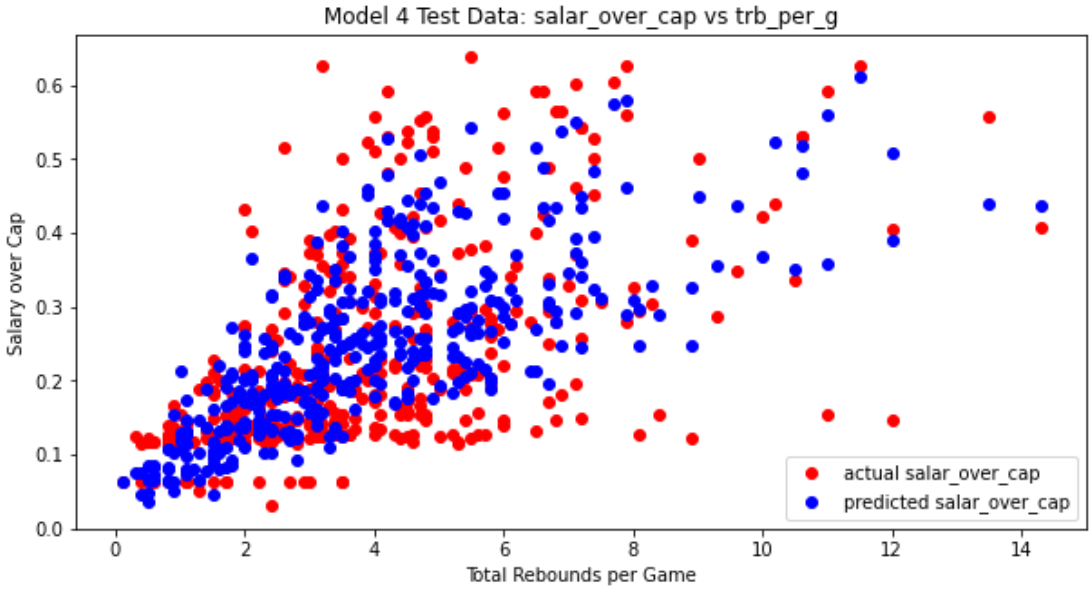
Final Ridge 5-fold CV scores on Test Data: R^2 = 0.69



Final Model Salary Prediction vs Point Per Game



Final Model Salary Prediction vs Assists Per Game



Final Model Salary Prediction vs Total Rebounds Per Game

Tools

- Numpy and Pandas for data manipulation
- Sqlalchemy, and SQLite for data storing and matching
- BeautifulSoup and requests for web scraping
- Statsmodels and Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communication

Slides and Visual graphs are created for presentation purposes.