Gulay Samatli-Pac

NBA Salary Prediction

Web Scraping and Linear Regression Project

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. The purpose is to provide fair salary expectations during the recruiting process to each side - to players and teams. The model hopefully helps the team to evaluate a player's previous performance and offer a baseline salary. Meanwhile, a player can use the model to understand what statistics will make more impact on his salary and he would have more power during his salary negotiation. Thus, this project will focus on building a model which will have both interpretative and prediction powers.

2. Approach

The data is scraped from the Basketball Reference website. The metrics that will be 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

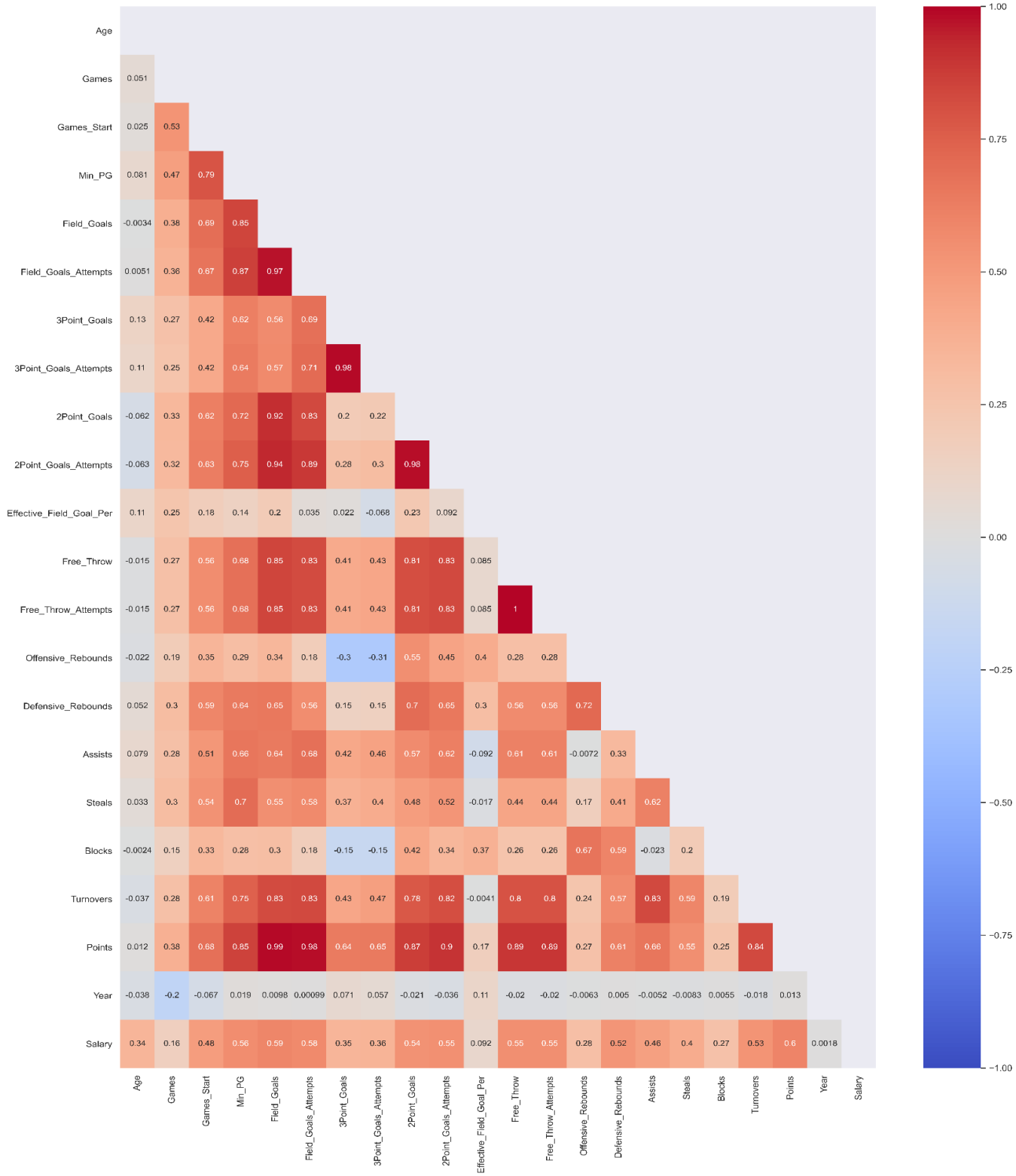
- AST -- Assists Per Game
- STL -- Steals Per Game
- BLK -- Blocks Per Game
- TOV -- Turnovers Per Game
- PF -- Personal Fouls Per Game
- PTS/G -- Points Per Game

There are a total 30 NBA teams and more than 400 players play in each season. To reach the desired amount of data asked for this project, I have collected statistics for 3 seasons, season 2019, 2020 and 2021. Next, I have simultaneously conducted the exploratory analysis and cleaned the data by removing duplicates, dealing with missing values and outliers.

The following visualization shows the correlation between the target variable, Salary, and all numerical features after data cleaning. As seen, the Salary variable has significant positive correlation with most of the features. It is also seen that there exists multicollinearity between independent features.

Feature Name	**Correlation**	**Feature Name**	**Correlation**
Points	59.68%	*Assists*	46.21%
Field_Goals	59.08%	*Steals*	40.33%
Field_Goals_Attempts	57.73%	*3Point_Goals_Attempts*	35.53%
Min_PG	56.19%	*3Point_Goals*	35.00%
Free_Throw_Attempts	54.82%	*Age*	34.50%
Free_Throw	54.82%	*Offensive_Rebounds*	28.50%
2Point_Goals_Attempts	54.58%	*Blocks*	27.30%
2Point_Goals	53.65%	*Games*	16.37%
Turnovers	52.74%	*Effective_Field_Goal_Per*	9.18%
Defensive_Rebounds	52.02%	*Year*	0.18%
Games_Start	48.07%		

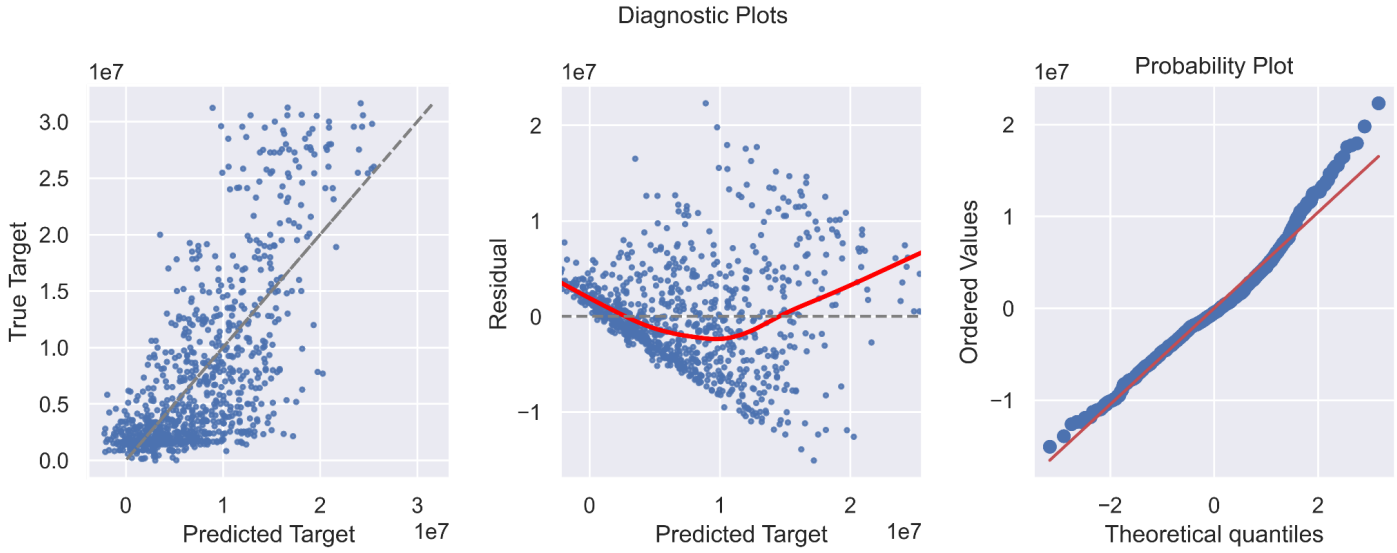
Gulay Samatli-Pac



Gulay Samatli-Pac

2.1. Baseline Model

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.
The diagnostic plots indicate that a nonlinear model might be a better fit to the data.
Transforming the target variable might also be helpful to increase the model efficiency since the Salary variable is skewed since one of the assumptions of Linear Regression is the normal distribution of the predictor variable.



3. Next Steps

- Apply the log transform to the Salary feature,

- Apply feature selection and engineering with various regression models, e.g. LASSO, Ridge, ElasticNet to improve the prediction on the test data.

- Report the final result!