Fantasy Premier League With Linear Regression

Bernard Opoku

Abstract

The goal of this project was to develop a Linear model to predict the points a player will obtain in a given game week, given data from previous weeks and previous seasons/years in the Fantasy Premier league (FPL).

FPL is an online fantasy game for the English Premier League, played by approximately 8 million people across the world. The goal is to score as many points as possible from a select total of 15 players (11 starters + 4 substitutions) from the given pool of 687 players. The points scored by each player depend on upon various attributes of the player in that particular game and each carries different weightage. For example, a forward player gets rewarded 4 points for every goal he scores and 3 points for every assist. The total points scored in each game by each player are a cumulative sum of all this attributes.

My hypothesis is that given the available player data every week, I can predict points scored for the coming week with enough accuracy, that I can also personally use my model as a tool for player selection.

Design

This project was a ground up built as required by the module's project specifications and with data scraped from a website of choice. I used selenium webscraper with the help of chrome web driver and BeautifulSoup to scrape my data from the official English Premier League website

Data

The data contains approximately 20,000 observations and 33 features with majority of them being in their supposed numerical data types. Most of these feature describes what the player did in the previous game week being positive or negative.

Algorithms

I created a *for loop* to extract data to allow selenium click through different buttons to make the data accessible by the scraper. Once accessible, I use another for loop to go through the various tables to extract the data for all the 669 players and saved it an empty dictionary. I used a random sleep parameter, to avoid being blocked by the website. I Then wrote the contents into a csv file for each week.

I loaded these CVSs back by concatenating them after joining previous gameweek stats with the next weeks results which will be our target variable.

I then selected only midfield players. to start my modeling with because of same scoring metrics.

Algorithms

Models

Linear regression and ridge regression were used before settling for a simple linear regression as the model had a slightly better R square performance. With is less complexity.

Model Evaluation and Selection

I started and used only midfield players for this project. In this set, there are about 8, 000 observations that were plit into 60/20/20 train, validation and test sets. We also tried modeling with a 5-fold cross validation on the training set.

I used Error metrics Mean absolute error and mean squared error to make an ultimate selection and satisfaction to a selected model.

Final Simple Linear regression: 258 features with categorcals features which were later dummyfied.

Final Metrics

Chosen model= Log10(target+5) transformed Simple Linear regression

- train R2 = 0.3484306541172081
- test R2 = 0.3333804121855254
- Model Error
- MSE: 1.2573146034424538
- MAE: 0.8525005826756173

Tools

- Selenium and Beautiful for web scrapping
- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communication

The scatter plot below shows the linear relationship between our predicted points scored by players against the actual points scored given the previous week's

statististics(Features)