Moneyball - EPL

Charles Peters

2019

I, Charles Peters, have read and understood the School's Academic Integrity Policy, as well as guidance relating to this module, and confirm that this submission complies with the policy. The content of this file is my own original work, with any significant material copied or adapted from other sources clearly indicated and attributed.

1 Introduction

In 2002, Major League Baseball team Oakland A's won their division and set a record breaking 20-game win streak. However, what makes this feat even more remarkable is that despite the A's success, they had one of the lowest budgets in the league. Much of the A's success was attributed to an algorithm that the A's used to evaluate players in a way that other teams at the time were not. In modern day sport, the fans' opinions of a player's worth is based of several factors, such as recent results, style of play, with elements of bias. Fans often overlook the underlying statistics, and the A's capitalised on this by evaluating players' on overlooked statistics, allowing them to pick up good players with much lower wages than the rest of their competition.

Now days many teams, not just in baseball but all sports, are starting to use these statistical analysis techniques. My aim for this project was to perform my own statistical analysis on English Premier League football players, and attempt to identify which players are good value for the wages they demand.

2 Data

Using appropriate models in python, data was collected from two websites and then sorted into tables. Tables were created for the four separate positions being considered : Goalkeepers, Defenders, Midfielders, Forwards. In-game statistics were collected from the official premier league website(1), and data relating to wages and payrolls were collected from spotrac(2).

3 Model and Predictions

Once the data had been collected, a method for evaluating each player had to be determined. The method used was to have a model predict a win rate for each player based on their stats. Leave-one-out cross-validation was used to leave one player out of the model each time and fit the model to the rest of the data points. The expected win rate for the left out player was then calculated. The idea behind this is that players with a higher expected win rate than their true win rate would be players that perform well, based on their in-game statistics, yet this is not reflected by their true win rate. This may indicate that the player's true win rate is a reflection of the whole teams performance, rather than individual performance.

Seeing as the task at hand was to predict quantitative values, a regression model is an appropriate choice of model here. Two models were tested: linear regression and ridge regression. The linear regression model generally gave quite similar results to the ridge regression model, with comparable mean-square errors (MSE), but also lead to some predictions for win rate being negative. As a result, the linear regression model was not used in further analysis, and the ridge regression model was the model to be taken forward. A separate model was fitted for each position, and a function was defined to determine the appropriate α value for each model.

Once all the models had been fitted several functions were defined to return different information – *best_players()* returns the best players in each position for a given budget; *build_team()* builds a team of eleven players for a given budget and formation, attempting to maximise the average expected win rate of the team; *build_squad()* builds a 25-man squad for a given budget, allowing the user to adjust the number of players they want for each position, again with the goal of maximising overall expected win rate.

4 Results

It is no lie that the players who demand the highest wages tend to be some of the best players in the league, and this is reflected in our models too – the players with the highest expected win rates tend to have the highest wages. However, there are more deviations from this trend than one might expect. In fact for all positions, the highest valued player was not the player with the highest wages. As an example see Table 1 for data on Goalkeepers. Sergio Romero, who our model predicts to be the best goalkeeper, does not have low wages by any means, but there are goalkeepers with much higher wages yet valued less by our model (e.g. Kepa Arrizabalaga). Perhaps who is more interesting in this table is Emiliano Martinez - who seems like a steal when compared to other players with similar expected win rates.

The same logic we applied for goalkeepers can be applied to the other positions – there are many players with expected win rates far greater than their wages suggest and vice versa.

Name	True win rate	Expected win rate	Annual Salary
Sergio Romero	0.57	0.92	£3,640,000
Alisson	0.80	0.66	£4,680,000
Ederson	0.81	0.59	£3,380,000
Kepa Arrizabalaga	0.56	0.53	£7,800,000
Emiliano Martinez	0.50	0.51	£416,000
Vicente Guaita	0.47	0.50	£3,120,000

Table 1: The top goalkeepers in the league as evaluated by our model, alongside their annual salaries.

As far as constructing an optimal team, a club with an unlimited budget could play the following team, in a 3-4-3 formation, to maximise their expected win rate:

Sergio Romero
Oleksandr Zinchenko — Aymeric Laporte — Nicolas Otamendi
Paul Pogba — Youri Tielemans — David Silva — Mesut Ozil
Alex Iwobi — Mohamed Salah — Roberto Firmino

This team would cost you £85,958,000 a year in wages - meaning you have already spent more than 14 of the premier league clubs' full budgets just on a starting 11!

5 Limitations

In order to reduce the complexity of this project several assumptions had to be made, which as a result reduce the accuracy of the predictions.

Firstly, the validation of the statistic used to evaluate each player is hard to quantify. The model was chosen to reduce MSE, which indicates the difference between the predicted and true win rates. However, as stated earlier, the point of the project is to identify players which have a significant difference between expected win rate and true win rate, so a higher MSE doesn't necessarily correspond to a poor choice of model. A different statistic could be chosen which better indicates a player's individual performance, and also allows us to quantify the uncertainty more reliably.

Other assumptions made were that we completely neglected the effect of non-statistical factors on a players value. These include age, particular play style, recent form, and in general are the factors that humans generally consider more when rating players. Rather than only considering qualitative or quantitative factors separately, it would be more appropriate for a football manager to combine their knowledge of football with the statistical analysis of this project to best determine which players might be suitable for their squad.

References

[1] https://www.premierleague.com

[2] https://www.spotrac.com