Soccer Intelligent Business using Linear and Polynomial Regression in Premier League teams

Francisco Javier Castillo Hernandez - A01208970
ITESM Campus Querétaro
Intelligent Systems - Linear and Polynomial Regression

Abstract - This paper aims to help Premier League teams managers to win or take better decisions with the use of different AI techniques in order to provide more data. With the use of Linear regression, we can predict how many points are necessary to stay in the first spots of the tournament and the necessary time per player to do their best in their respective position so the club maintains its goals. The use of polynomial regression is meant to help to buy decisions based on player statistics, predicting the type of player you need and how will perform based on previous data. All of this so the managers can have more information based on previous performance and how the league is behaving.

Keywords - Linear Regression, Polynomial Regression, Premier League, Business Intelligence

1. Introduction

Premier League is an English soccer league, where 20 teams battle for being crowned English champions. The league takes place between August and May and involves the teams playing each other home and away across the season, a total of 380 matches. Three points are awarded for a win, one point for a draw, and none for a defeat, with the team with the most points at the end of the season winning the Premier League title.

The teams that finish in the bottom three of the league table at the end of the campaign are relegated to the Championship, the second tier of English football. Those teams are replaced by three clubs promoted from the Championship; the sides that finish in first and second place and the third via the end-of-season playoffs.

If any club finishes with the same number of points, their position in the Premier League table is determined by goal difference, then the number of goals scored. If the teams still cannot be separated, they will be awarded the same position in the table. (*Premier League Competition Format & History*, 2021)

The teams that finish in the top four of the Premier League qualify for the next season's UEFA Champions League group stages. A fifth-place Premier League finish will put a team into the UEFA Europa League but the next best-placed teams who have not qualified for Europe will also enter the competition if the winners of the FA Cup and/or League Cup qualify through their league position.

In soccer one of the most important things is the team and how it is built. According to the global football market size it was valued at $ 1,883.6 million in 2019, it is estimated to reach $ 3,712.7 million by 2027.

The **Transfermarkt market** values are calculated by taking into account various pricing models. A major factor is the Transfermarkt community, whose members discuss and evaluate player market values in detail. In general, the Transfermarkt market values are not to be equated with transfer fees.

Most important factors:

- Age
- Performance at the club and national team
- Reputation/prestige
- Development potential
- League-specific features
- Marketing value
- Number & reputation of interested clubs
- Performance potential
- Experience level
- General development of transfer fees (*Transfermarkt Market Value Explained - How Is It Determined?*, 2021)

Intelligent business is described as a technology-driven process used for analyzing strategic data and delivering actionable information. Business intelligence (BI) is a solution that helps executives, managers, and workers collaborate to make informed business decisions by allowing them to discover more about trends and processes that are affecting their organizational performance.

Utilizing BI within football clubs, which hold mass amounts of important data, can optimize and maximize revenues by placing a particular focus on relevant KPIs.

Tracking and monitoring athletes also allow clubs to identify the potential talents of each player, enabling them to implement strategic training sessions that encourage maximum growth and nourishment of each individual that represents a future asset of the club.

The competitive advantage of football clubs is not in relation to the large amounts of data they have, but more the people, processes, and tools at their disposal that can make data meaningful and transform it into actionable insights.

Particularly during the time of negotiations, decision-makers within technical departments must be able to make informed choices by having instant access to all relevant information required as necessary. (Savino, 2020)

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting (ML | Linear Regression, 2018).

There are 2 types of regression:

- Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

- Multiple Linear Regression there are more than one independent variable for the model to find the relationship.

Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below: $y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+..... b_nx_1^n$ (Gallo et al., 2015)

2. Implementation

a. Linear Regression

Cleaning data and data loading: The dataset was loaded as its respective file

and data to gather in a .csv format (Defenders, Midfielders, Forwarders) using pandas (library of python).

Data pre-processing: For implementations purposes, the Learning rate was predefined as a global variable with the value of .01, with a bias mapped on the first element of the lists with a constant value of 1. (Implementation by professor Benjamin Valdés used in the project fo the bias)

In order to process the data for a bigger dataset of players, a pre-processing method was defined in order to filter by position and add into a list the attributes Age, Appearances, and depending on the dataset a different column (Fouls Committed/Defenders, Goal Assists/Midfielders & Shots on Target/Attackers).

Null hypothesis: From what was mentioned before about the attributes the list of parameters was defined as 4 lengths filled by ceros, which correspond to the bias, and the 3 attributes mentioned. (The null hypothesis states that **the slope is equal to zero**)

Scaling: Using the min-max scaling also known as x normalized where the following formula is used: x - Xmin / Xmax - Xmin, that basically means that it scales from one to zero all the data taking the max values and the min values of each row

MSE: We use MSE in Gradient Descent in order to find the optimal parameters for our linear regression by minimizing the epochs or iterations. It is used with the formula:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

where Yi is the Hypothesis value and ^Yi is the real value to find the average of the squared differences between those values.

Gradient Descent: In order to perform Linear Regression, gradient descent is an iterative optimization algorithm to find the minimum of a function that helps to find the best-fit line for a dataset.

$$\theta_j := \theta_j - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

The algorithm stops whether it reaches the same parameters as before or the max epochs or iterations are given. What will help to know if can be trusted is the coefficient of determination, at soccer BI an avg above 70 is considered as good.

Coefficient of determination: R squared score is used to evaluate the performance of a linear regression model. Is the amount of variation in the output dependent attribute which is predictable from the input independent variable.
It follows the following formula **R2= 1- SSres / SSto**t, where SSres is the sum of squares of the residual errors and SStot is the total sum of the errors.

SSres: It is the sum of the differences between the predicted real value and the mean of the dependent variable

SSTot: Are the squared differences between the observed dependent variable and its mean

b. Single Polynomial Regression

Cleaning data and data loading: The dataset was loaded as its respective file and data to gather in a .csv format (Defenders, Midfielders, Forwarders) using pandas (library of python).

Scikit-learn: With the use of the Scikit learn kit for the polynomial regression the features and degrees were added, the data was transformed by the polynomial form.

Polynomial Degree: For the following data, first a Linear Regression was made but since the shape of the data was following more of a polynomial form it was

decided to change it to polynomial regression, with a 2-degree base, for all of the models, the degrees were changed (2-10) until it gets the highest R2.

Polynomial Regression: The Linear Regression function was called with the fit method in order to train the data in it, then the function predicted with the test array was made to get the predictions set. In order to check the efficiency of the model, the R squared and MSE was evaluated in our data. Lastly, the predictions were plotted compared to the real dataset.

3. Results

Polynomial regression:

Defenders

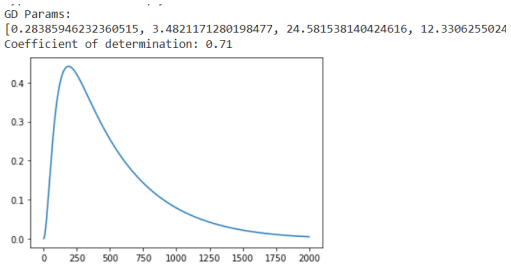


Fig. 1: Defender's MSE Plot (Goals Conceded - Age, Appearances & Fouls Committed)

Midfielders

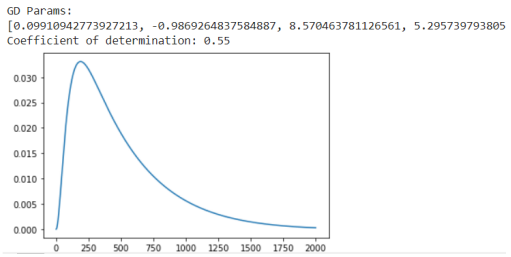


Fig. 2: Midfielder's MSE Plot (Shots on Target - Age, Appearances & Goal Assists)

Attackers

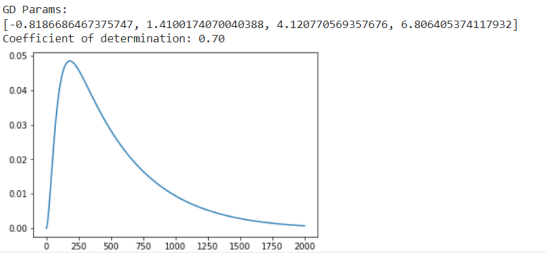


Fig. 3: Attacker's MSE Plot (Goals - Age, Appearances & Shots on Target)

Overall all the coefficients of determination are good except the midfielders coefficient that has a .55 and for soccer is not useful at all since it is not truthful at all, but defenders and attackers have a coefficient of .70 that is very useful since soccer has a lot of variables but the main one such as age, appearances and shots are the main ones to consider for a soccer player as Transfermarkt mention. For a first approach using a linear regression taking into account 3 different factors is a very and truthful coefficient that can help managers to estimate how many goals can score depending on their stats of the attackers and how many goals conceded based on their age, appearances and fouls committed can a defender can concede. At least for a highly competitive league, .70 is very good for the teams that are in the mid-lower table in order to make decisions based on these attributes to buy a player.

Defenders

```
Coefficients:
 [[ 1.48172068 -0.01052918]]
Mean squared error: 42.98
Coefficient of determination: 0.75
```

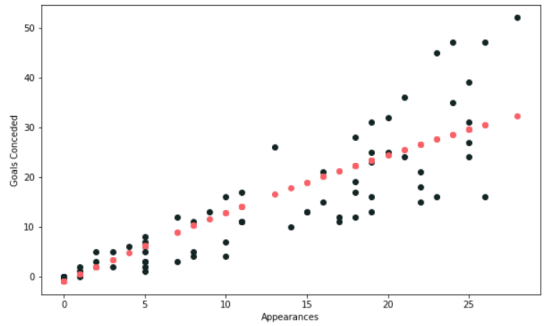


Fig. 4: Polynomial Regression Plot (Appearances/Goals Conceded [Degree - 2])

Midfielders

```
Coefficients:
 [[0.26015281 0.00589971]]
Mean squared error: 1.09
Coefficient of determination: 0.66
```

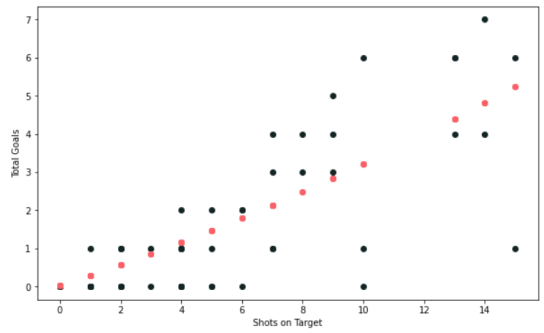


Fig. 5: Polynomial Regression Plot (Shots on Target/Total Goals [Degree - 2])

Attackers

```
Coefficients:
 [[-0.0404123   0.00138568]]
Mean squared error: 5.01
Coefficient of determination: 0.68
```

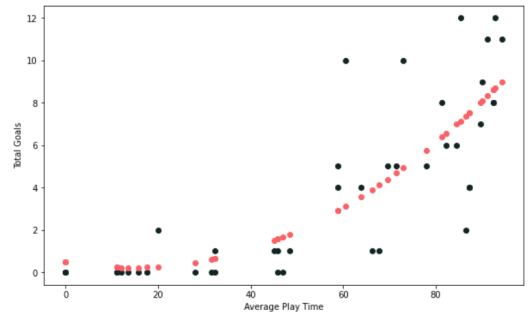


Fig. 6: Polynomial Regression Plot (Avg Time Play/Total Goals [Degree - 2])

```
Coefficients:
 [[0.28930753 0.00374511]]
Mean squared error: 2.27
Coefficient of determination: 0.86
```

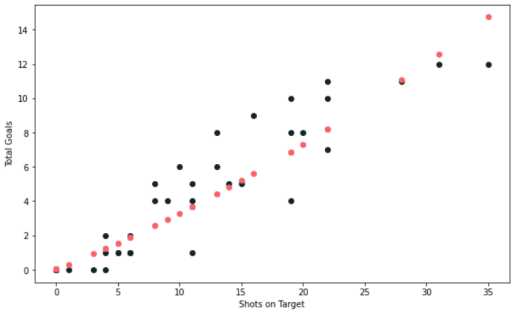
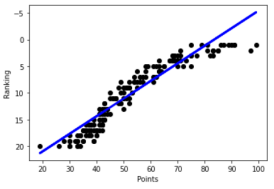


Fig. 7: Polynomial Regression Plot (Shots on Target/Total Goals [Degree - 2])

The general overview of the polynomial is that approximately the coefficients is around .70 that as mentioned before is pretty good for mid-lower league table that can be confident on how many goals conceded can a defender give for appearances, how many goals can score the midfielder depending on their shots on target, how many goals can score an attacker based on their shots on target and average time play.

Auxiliar Graphs:

Rankings



	Rk	Squad	MP	W	D	L	GF	GA	Pts	Attendance	Top Team Scorer	Goalkeeper	Year
0	1	Manchester Utd	38	25	7	6	73	35	82	41001	Eric Cantona - 14	Peter Schmeichel	1996
1	2	Newcastle Utd	38	24	6	8	66	37	78	36501	Les Ferdinand - 25	Shaka Hislop	1996
2	3	Liverpool	38	20	11	7	70	34	71	39553	Robbie Fowler - 28	David James	1996
3	4	Aston Villa	38	18	9	11	52	35	63	37492	Dwight Yorke - 17	Mark Bosnich	1996
4	5	Arsenal	38	17	12	9	49	32	63	32614	Ian Wright - 15	David Seaman	1996

Fig. 8: Linear Regression Plot (Points/Ranking)

4. Conclusions

After evaluating all of the regressions and performed all of the algorithms, we can summarize the results in 3 key points:

- The Regression tasks that had a high coefficient of determination were the defenders and attackers

with a coefficient of approximately of .70 that overall with very lower MSE, that can tell us that it could reliable but we have to consider other attributes to take a final decision, since is a very nice approach to consider to player statistics to decide for a buying decision.

- For this dataset midfielders need other attributes in order to find better relations and can have a higher coefficient so it can be a reliable source to start a buy decision, since having a coefficient lower than .5 is not acceptable.
- Overall the results were good but can be better with a complex dataset and a more complex algorithm in order to have more attributes to consider such as the ones mentioned on Transfer Market. The results from other researchers is with a single algorithm the coefficient can be between .45 and .68.

The aim of this project is to have a better overview of the Premier League data, where decisions can be made through numbers, for the teams that are in the mid-lower table sometimes they need to plan a whole project for the next tournament where they have to consider their players and their new acquisitions, and this project aims to help to them to determine a general view of how they players can perform for their next tournament, and to know the statistics of the tournament to do a forecast based on all of this information given.

Right now the project has a limited dataset perfect for a first approach but in order to get more information a more specific dataset and a complex algorithm in order to have more accurate results and not limited ones. Out there in order to decide they use at least 3 mixed algorithms to have a better forecast for their team and their players. It is expected to continue this project to do a comparison with a complex algorithm.

5. References:

Gallo, A., Davenport, T. H., & Kim, J. (2015, November 4). *A Refresher on Regression Analysis*. Harvard Business Review. Retrieved March 29, 2022, from https://hbr.org/2015/11/a-refresher-on-regression-analysis

ML | Linear Regression. (2018, September 13). GeeksforGeeks. Retrieved March 29, 2022, from https://www.geeksforgeeks.org/ml-linear-regression/

Premier League Competition Format & History. (2021). Premier League. Retrieved March 29, 2022, from https://www.premierleague.com/premier-league-explained

Savino, M. (2020). *THE IMPORTANCE OF BUSINESS INTELLIGENCE IN A FOOTBALL CLUB*. Linkedin. Retrieved 03 29, 2022, from https://www.linkedin.com/pulse/importance-business-intelligence-football-club-marco-savino/

Transfermarkt Market Value explained - How is it determined? (2021, May

13). Transfermarkt. Retrieved March 29, 2022, from https://www.transfermarkt.co.in/transfermarkt-market-value-explained-how-is-it-determined-/view/news/385100