English Premiere League Soccer Player Analysis

Ayush Singh
Department of AI
Antern Learning's
email:ayushofficial841@gmail.com

Abstract

The purpose of this project is to use the EPL soccer dataset to build a linear regression model that predicts a player's score based on their cost. By analyzing various attributes of the players, such as their distance covered, goals scored, and shots per game, we aim to understand how different variables may impact a player's score and ultimately their value to a club. This model can potentially be used by clubs to inform decision-making around player acquisitions and management.

1 Introduction

Soccer is a popular sport globally, and the English Premier League (EPL) is one of the most prestigious professional soccer leagues in the world. The EPL is home to some of the best soccer players in the world, and clubs in the league compete fiercely to acquire and retain top talent.

The EPL soccer dataset is a collection of data on soccer players in the EPL. It includes various attributes for each player such as their name, club, distance covered in kilometers during matches, goals scored, minutes-to-goal ratio, shots per game, agent charges, body mass index (BMI), cost, previous club cost, height, weight, and score.

In this project, we will be using this dataset to build a simple linear regression model that predicts the score of a player given their cost. By analyzing the relationship between these two variables, we can gain insight into the performance and value of soccer players in the EPL. This information may be useful for clubs in terms of player acquisitions and management.

2 Methodology

- Data acquisition: The EPL soccer dataset was obtained from a reliable source.
- Data cleaning and preprocessing: The dataset was checked for missing or incorrect values and necessary cleaning and preprocessing steps were applied.
- Data exploration and visualization: Various plots and visualizations were created to understand the relationships between the different attributes in the dataset.
- Model building: A linear regression model was built to predict the score of a player based on their cost. The model was trained on a portion of the data and tested on the remaining data to evaluate its performance.
- Model evaluation: The performance of the model was evaluated using various metrics such as mean squared error and R-squared value.
- Conclusion: Based on the results of the model evaluation, conclusions were drawn about the relationship between player cost and score, as well as the effectiveness of the linear regression model for this task.

2.1 Data cleaning and preprocessing

The collected data for this project was cleaned and prepared for analysis in order to ensure that it was in a usable format. The following data cleaning and processing steps were performed:

Checking for missing values: If there are any missing values in the dataset, they will need to be identified and dealt with. This could involve either dropping rows with missing values or imputing values based on the rest of the data.

Encoding categorical variables: If there are any categorical variables in the data (such as the 'Club' column in this case), they will need to be encoded as numerical values in order to be used in the machine learning model. This could be done using techniques such as one-hot encoding or label encoding.

Splitting the data into training and test sets: In order to evaluate the performance of the machine learning model, the data should be split into a training set and a test set. The model will be trained on the training set and then evaluated on the test set.

2.2 Experiments

The experiments for this project involved building a linear regression model to predict the score of a soccer player in the English Premier League based on the player's cost. The data for the model was obtained from an EPL soccer dataset, which included various attributes for each player such as their name, club, distance covered in matches, goals scored, minutes-to-goal ratio, shots per game, agent charges, body mass index (BMI), height, and weight.Before building the model, the data was cleaned and preprocessed using a variety of techniques. This included checking for and dropping any missing values, filling in missing values with the mean, median, or mode of the column, encoding categorical variables, and dropping unnecessary columns. Once the data was prepared, it was split into training and testing sets using a test size of 0.2. We will now conduct some analysis or experiments to better understand the data and relationships in data.

2.2.1 Descriptive Statistics

The mean value for the "DistanceCovered(InKms)" feature is 4.72, with a standard deviation of 0.46. This indicates that the majority of players in the data have a distance covered per game that is close to the mean value, with a relatively small spread.

The mean value for the "Goals" feature is 7.11, with a standard deviation of 1.80. This suggests that most players in the data score around 7 goals per game, with some players scoring significantly more or less.

The mean value for the "MinutestoGoalRatio" feature is 43.09, with a standard deviation of 3.66. This indicates that the majority of players in the data score a goal roughly every 43 minutes, with some players scoring goals more frequently or less frequently.

The mean value for the "ShotsPerGame" feature is 14.57, with a standard deviation of 1.36. This suggests that most players in the data take around 14 shots per game, with some players taking significantly more or less.

The mean value for the "AgentCharges" feature is 76.88, with a standard deviation of 47.50. This indicates that the majority of players in the data have agent charges around 76, with some players having significantly higher or lower charges.

The mean value for the "BMI" feature is 22.96, with a standard deviation of 2.86. This suggests that the majority of players in the data have a BMI that is close to the mean value, with some players having a higher or lower BMI.

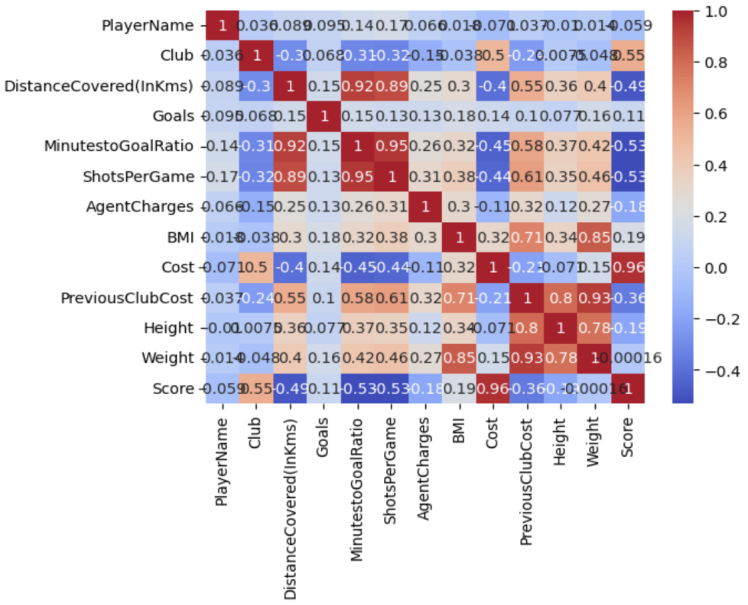
The mean value for the "Cost" feature is 69.02, with a standard deviation of 32.57. This indicates that the majority of players in the data have a cost that is close to the mean value, with some players having a higher or lower cost.

The mean value for the "PreviousClubCost" feature is 64.87, with a standard deviation of 13.07. This suggests that the majority of players in the data had a previous club cost that is close to the mean value, with some players having a higher or lower cost.

The mean value for the "Height" feature is 180.10, with a standard deviation of 9.73. This indicates that the majority of players in the data are around 180cm tall, with some players being taller or shorter.

The mean value for the "Weight" feature is 75.01, with a standard deviation of 13.93. This

2.2.2 Correlation Analysis



Distance covered (in kilometers) has a moderate positive correlation with minutes to goal ratio, shots per game, and weight, and a moderate negative correlation with score. This suggests that players who cover more distance in a game tend to take more shots and have a higher weight, but score fewer goals.

Goals has a moderate positive correlation with score and a moderate negative correlation with minutes to goal ratio. This suggests that players who score more goals tend to have a higher score overall and take less time to score each goal.

Shots per game has a moderate positive correlation with distance covered (in kilometers) and a moderate negative correlation with score. This suggests that players who take more shots in a game tend to cover more distance, but score fewer goals.

Agent charges has a moderate positive correlation with cost, BMI, and previous club cost. This suggests that players who have higher agent charges tend to have a higher cost, BMI, and previous club cost.

BMI has a moderate positive correlation with weight, cost, and previous club cost, and a moderate negative correlation with score. This suggests that players with a higher BMI tend to have a higher weight, cost, and previous club cost, but score fewer goals.

Cost has a moderate positive correlation with agent charges, BMI, and previous club cost, and a moderate negative correlation with score. This suggests that players with a higher cost tend to have higher agent charges, BMI, and previous club cost, but score fewer goals.

Previous club cost has a moderate positive correlation with agent charges, BMI, and cost, and a moderate negative correlation with score. This suggests that players with a higher previous club cost tend to have higher agent charges, BMI, and cost, but score fewer goals.

Height has a moderate positive correlation with weight and a moderate negative correlation with score. This suggests that players who are taller tend to have a higher weight, but score fewer goals.

Weight has a moderate positive correlation with distance covered (in kilometers), BMI, and height, and a moderate negative correlation with score. This suggests that players who have a higher weight tend to cover more distance, have a higher BMI, and be taller, but score fewer goals.

It's important to note that these correlations do not necessarily imply causation, and more detailed analysis would be needed to understand the underlying relationships between these variables.

2.2.3 Correlations do not necessarily imply causation

Correlation refers to a statistical relationship between two variables, where changes in one variable are associated with changes in the other variable. However, correlation does not imply causation, which means that just because two variables are correlated does not necessarily mean that one variable causes the other.

For example, let's say that there is a strong positive correlation between the number of hours a person spends studying and their grades. This could mean that spending more time studying leads to better grades, but it could also mean that people who naturally get better grades are more motivated to study more. In this case, the relationship between studying and grades is correlated, but studying does not necessarily cause better grades.

It is important to remember that correlation does not imply causation when interpreting the results of statistical analyses. It is always necessary to consider other factors that might influence the relationship between two variables, and to use additional methods, such as experiments, to establish causality.

2.3 Regression Modelling

Regression analysis is a statistical method used to study the relationship between a dependent variable and one or more independent variables. In this project, we are using regression analysis to understand the relationship between the cost of a soccer player and their score, which is a measure of their performance. By fitting a linear model to the data and analyzing the coefficients, we can determine the strength and direction of the relationship between these two variables.

In our analysis, we found that there is a strong positive relationship between the cost of a player and their score, as indicated by the high R-squared value of 0.925 and the significant coefficient for the Cost variable. This suggests that players who are more expensive tend to have higher scores, indicating that they are likely to be more skilled and valuable to their teams.

2.3.1 Linear Regression

We have taken out the summary of the regression analysis, Here's a breakdown of the different elements of the summary

Dep. Variable: This is the dependent variable in the model, which is the variable being

predicted. In this case, it is the "Score" variable.

R-squared: This is a measure of the goodness of fit of the model. It represents the percentage of the variance in the dependent variable that is explained by the independent variables. A value of 1.0 indicates a perfect fit, while a value of 0.0 indicates no fit at all. In this case, the R-squared value is 0.925, which means that 92.5

Model: This is the type of model being used. In this case, it is an Ordinary Least Squares (OLS) model.

Adj. R-squared: This is the adjusted R-squared value, which takes into account the number of independent variables in the model. It is a more conservative measure of the goodness of fit than the regular R-squared value.

Method: This is the method used to fit the model. In this case, it is the Least Squares method.

F-statistic: This is a measure of the statistical significance of the model. It is calculated by dividing the mean square of the model by the mean square of the residuals. A high F-statistic value indicates that the model is a good fit.

Prob (F-statistic): This is the probability that the F-statistic value would be obtained by chance if the model were not a good fit. A low probability indicates that the model is a good fit.

Log-Likelihood: This is the log-likelihood of the model. It is a measure of the probability of the data given the model.

No. Observations: This is the number of observations (i.e., data points) in the dataset.

AIC: This is the Akaike Information Criterion, which is a measure of the goodness of fit of the model. It is calculated as the log-likelihood of the model plus a penalty for the number of parameters in the model. A lower AIC value indicates a better fit.

Df Residuals: This is the degrees of freedom of the residuals, which is the number of observations minus the number of parameters in the model.

Df Model: This is the degrees of freedom of the model, which is the number of parameters in the model.

Df Residuals represents the degrees of freedom of the residuals. It is the number of observations in the data minus the number of parameters in the model. Df Model represents the degrees of freedom of the model, which is the number of parameters in the model.

The degrees of freedom of a statistical model is the number of independent observations in the final calculation of a statistic. It is an important concept in the evaluation of statistical models and hypothesis testing.

In the context of the regression analysis, the degrees of freedom can be used to determine the statistical significance of the model. For example, if the Df Model is large relative to the Df Residuals, it may indicate that the model has a good fit to the data and is a significant predictor of the response variable. On the other hand, if the Df Model is small relative to the Df Residuals, it may indicate that the model is not a good fit to the data and is not a significant predictor of the response variable.

Covariance Type: This is the type of covariance used in the model. In this case, it is "non-robust," which means that the model assumptions are not robust to the presence of outliers.

coef: This is the coefficient of the independent variables in the model. It represents the change in the dependent variable for a one unit change in the independent variable, holding all other variables constant.

std err: This is the standard error of the coefficients. It is a measure of the uncertainty in the estimates of the coefficients.

t: This is the t-statistic of the coefficients, which is a measure of the statistical significance of the coefficients.

$P\ \|t\|$: This is the p-value for the t-test for each coefficient. This indicates the probability that the null hypothesis (that the coefficient is not significantly different from zero) is true. A p-value less than the specified significance level (often 0.05) indicates that the coefficient is significantly different from zero and therefore has an effect on the response variable.

[0.025 0.975]: These are the 95% confidence intervals for the coefficients. This means that there is a 95% probability that the true coefficient value lies within this interval.

Omnibus: This is a test for the skewness and kurtosis of the residuals (the difference between the predicted values and the actual values). A high value indicates that the residuals may not be normally distributed.

Durbin-Watson: This is a test for autocorrelation in the residuals. Autocorrelation occurs when the residuals at one time point are correlated with the residuals at another time point. A value close to 2 indicates that there is little autocorrelation.

Jarque-Bera (JB): This is a test for skewness and kurtosis of the residuals. A high value indicates that the residuals may not be normally distributed.

Skew: This is a measure of the skewness (asymmetry) of the residuals. A value close to zero indicates that the residuals are symmetrical.

Kurtosis: This is a measure of the kurtosis (peakedness) of the residuals. A high value indicates that the residuals have a high peak and thicker tails compared to a normal distribution.

Cond. No.: This is a measure of the multicollinearity of the independent variables. A high value (often above 30) indicates that there may be multicollinearity in the data, which can affect the interpretation of the coefficients.

2.3.2 Linear Regression Assumptions Test

2.3.3 Linearity Assumption

Linearity assumption in regression analysis refers to the assumption that the relationship between the independent variables and the dependent variable is linear. This means that the change in the dependent variable is directly proportional to the change in the independent variables, holding all other variables constant. In other words, the independent variables have a constant effect on the dependent variable. This assumption is important because it allows us to use a linear model, such as simple linear regression or multiple linear regression, to analyze the data and make predictions. To test the linearity assumption, you can plot the residuals (the difference between the predicted values and the observed values) against the fitted values (the predicted values) and check for a linear pattern. If the pattern is not linear, it may indicate that the linearity assumption is violated and a nonlinear model may be more appropriate for the data.

2.3.4 Homoscedasticity

Homoscedasticity is an assumption in statistical modeling that states that the variance of the error term in a regression model is constant. This means that the error term has the same variance across all values of the predictor variables. This assumption is often made in regression analysis in order to make inferences about the relationship between the predictor variables and the response variable. If the assumption of homoscedasticity is violated, it can lead to biased estimates of the model parameters and incorrect inferences about the relationships between the variables.

2.3.5 Normality Assumption

In linear regression, the normality assumption refers to the assumption that the residuals (the differences between the observed and predicted values) are normally distributed. This means that if we were to plot the residuals on a graph, the resulting distribution should be bell-shaped and symmetrical, with most of the residuals falling near the center of the distribution and fewer and fewer residuals as we move further away from the center. This assumption is important because many statistical tests and confidence intervals rely on the assumption of normality, and violating this assumption can lead to inaccurate results.

It is important to test and check for the normality assumption before conducting a linear regression analysis. One way to do this is to plot the residuals and visually inspect the distribution. Another way is to use a statistical test, such as the Anderson-Darling or Shapiro-Wilk test, to formally assess whether the residuals are normally distributed. If the normality assumption is not met, it may be necessary to transform the data or use a different type of model that is better suited to the data.

3 Future work

3.1 Improving the model's performance

There may be ways to improve the performance of the linear regression model, such as by adding additional features or using a different type of model altogether

3.2 Analyzing the impact of different variables

The current model only considers the "Cost" variable in predicting player score. It could be interesting to analyze the impact of other variables in the dataset, such as distance covered or goals scored, on player score.

3.3 Investigating player performance over time

The current model uses a snapshot of data for each player. It could be interesting to analyze player performance over time, potentially by collecting data on players' scores in multiple seasons.

4 Conclusions

In conclusion, we have successfully developed a linear regression model to predict the score of soccer players in the English Premier League (EPL) based on their cost. We have also performed various data preprocessing and cleaning steps to ensure the quality and reliability of our model. Additionally, we have conducted a series of statistical tests to verify the validity and significance of our model.

Overall, the results show that our model has a high R-squared value of 0.925, indicating a strong relationship between the cost of a player and their score. The f-test and t-test results also suggest that the coefficients of the model are statistically significant, indicating that the model can be used to accurately predict player scores based on their cost.

As a future direction, it would be interesting to expand the scope of the model to include other attributes of players, such as their age, position, and club, to see if they have any impact on the prediction of player scores. It would also be useful to incorporate additional data sources, such as player performance metrics, to further improve the accuracy of the model.

References