Forecasting Premier League team point totals for the next season

Ralph Bergisch
Aleksandar Stoychev
Rowald Paardekooper

April 7, 2024
1955 words

Abstract

This study explores the feasibility of predicting Premier League teams' points totals for the upcoming season based on their performances in the preceding season. Utilizing a dataset of team performances from Kaggle and Transfermarket, we applied various machine learning models, including linear regression, decision trees, random forests, and support vector machines. The mean absolute error (MAE) served as our primary performance metric. Our findings suggest that while forecasting football outcomes is inherently challenging due to the sport's unpredictability, certain models, particularly regression with lasso regularization, can offer reasonably accurate predictions with a MAE of approximately 7 points.

1 Introduction

In the realm of sports analytics, forecasting the performance of Premier League football teams stands as a compelling challenge faced by many private companies and every single fan of the sport. This study poses the research question: Can the total points for each team in the Premier League be predicted based on their performance metrics from the previous season? Embedded within this inquiry are hypotheses grounded in the assumption that historical performance indicators—such as goals scored, goals conceded, win/loss records, and transfer market activities—bear a significant correlation with future success, measurable in points accumulated throughout the season.

Addressing this question necessitates a regression task approach, wherein the continuous outcome of season points totals is predicted from a set of independent variables derived from prior season performances. The predictive models explored in this research—linear regression, decision trees, random forests, and support vector machines—are chosen for their varying capacities to handle the intricacies of the dataset and their potential to uncover underlying patterns within the historical data.

The dataset, sourced from Kaggle and Transfermarket, comprises 204 instances, encapsulating a rich selection of performance metrics alongside transfer market dynamics over several seasons. This collection features 41 variables, offering a broader view of team performance that includes not only on-field statistics but also market activities, thus painting a comprehensive picture of the factors contributing to a team's seasonal achievements. Despite its big choice of independent variables, the dataset presents challenges, such as multicollinearity among predictors and a limited number of instances relative to the complexity of the predictive task.

2 Methods

The machine learning models used to answer the research question all were programmed to give continuous output. To test our research question, multiple machine learning models are built to estimate the points total of every team. Due to the nature of our dataset, and sport statistics in general, multicollinearity is our biggest obstacle. To account for this, various models are created, each with their own unique solutions. The recent development in the machine learning area allows us to approach the problem from different angles. The models can be split into 4 categories:

1. Ordinary Least Squares regressions
2. Decision Trees
3. Random Forests
4. Support Vector Machines

The performance measurement to compare the different model will be the Mean Absolute Error, henceforth MAE, because of its nice interpretability.

Firstly, we delved into the nuances of linear regression and the examination of multicollinearity within our dataset. We start with a standard OLS for comparative reasons. Next, to enhance the robustness of our regression analysis, we employed the Variance Inflation Factor (VIF) to evaluate each predictor. This step was crucial to identify and mitigate multicollinearity, ensuring that our regression models were not influenced by highly correlated predictors.

Secondly, Ridge and Lasso regressions are used. The Ridge and Lasso models are regularization techniques used in regression analysis, particularly when dealing with multicollinearity. In standard OLS, multicollinearity can lead to unstable parameter estimates and inflated standard errors. Ridge and Lasso regressions address this issue by adding a penalty term (or alpha's) to the objective function, which penalizes large coefficients. The penalty term encourages simpler models by shrinking the coefficients towards zero. This filtering process can be particularly useful for our high dimensional dataset, where there are many predictors which are probably irrelevant. Multiple penalty terms are tested to achieve the lowest MAE.

Next, both a decision tree and a random forest were used, as these methods are prominent amongst machine learning models. For both models no extra preparation or transformation of the data is required. Since all the data is numerical an encoding is not needed. The decision trees used, have depths from 1 till 20, while the random forest was tested with a depth between 1 and 12 and with the number of trees between 10 and 100.

Lastly, a support vector machine was used. For the support vector machine, multiple parameter specifications for the penalty parameter and separation function were implemented. The penalty parameter is tested for values between 0.0001 and 1000000 with iteration of times 100 each. After finding the best penalty parameter in this rough search, numerous close penalty parameters are tested to find the optimal one. For the kernel, a linear, polynomial, sigmoid and radial basis function were used.

3 Results

3.1 OLS Regressions and Multicollinearity loop

Preliminary results from our linear regression analysis revealed insights into the predictive power of the previous season's performance metrics on a team's point total. The initial regression model yields a MAE of 9.42, and a Coefficient of Determination (R-squared) of 0.537, indicating a moderate level of prediction accuracy. Following the identification and removal of variables exhibiting high multicollinearity with 7.5 as our chosen threshold, the model improved with a reduced MAE of 8.48. These enhancements showcase the importance of addressing multicollinearity in predictive modelling.

Graph 4 shows the performance of the Ridge regression. After running the regression for different penalty terms, the optimal alpha is 400. The MAE for the training and test set, is 6.80 and 7.33 respectively.

Graph 5 shows the performance of the Lasso regression and after running for different penalty terms, the optimal alpha is 0.5. The MAE for the training and test set, is 6.98 and 6.99 respectively. The negligible difference between the training and test MAE shows that both models perform relatively well. The coefficients estimated on the training data predict the observations in the test set almost as good as in the training set.

3.2 Decision Tree, Random Forest, and Support Vector Machine

The decision tree model scored a 9.81 as MAE which is the worst of all models used. The smallest MAE was achieved with a tree depth of 5. The random forest achieved its best value with a depth of 10 and 100 as the number of trees. The random forest scored a 7.70 as MAE. The last model used is the support vector machine which performed only slightly worse than the best linear regression. The support vector machine performed best when trained with the radial basis function and a penalty parameter of 0.85. For the support vector machine, a normalized version of our dataset was given as input since this gave a lower

MAE. The MAE the support vector machine achieved was a 7.18 after retaking the original mean and variance into account.

4 Discussion

Multiple machine learning models are used to forecast the following season's total points and a regression with lasso regularization scored best on our chosen performance measure which was the mean absolute error (MAE). This regression has a MAE of 7.0, meaning that for the testing data, the model is on average 7 points off the points total. The MAE might not be the best performance measure to estimate the predictive power of this model, as it could be heavily influenced by outliers. Thus, for further research we strongly advice to calculate the standard deviation of the errors as this makes the model's performance more interpretable. We think our model predicted the points total decently, but not having the standard deviation limits the certainty we have in answering if our model predicted the points total correctly.

The methods used to answer the research question all have their strengths and weaknesses. The basic regression suffers from the highly correlated dataset, violating certain OLS conditions. These correlation problems were solved with the Lasso and Ridge regularization, but these models suffered from independent interpretation issues. A shortcoming of both ridge and lasso is its inability to point out the true explanatory variables. To give an example, since wins and goals scored are highly correlated, it is difficult to tell which is the more important variable in estimating the next seasons points total.

The decision tree was not able to forecast any acceptable prediction which probably was because of noise in our data. The random forest and support vector machine both overfitted our data which can be mostly seen in our MAE of our testing data. Both these models share the problem that they are hard to interpret and took long to train. Another problem for the models is that they assume that variable importances and effects stay the same over the years while in reality, the game of football changes.

The problems encountered were mostly connected with the properties of the data. The dataset had only 204 instances and 41 variables, which is a relatively low number of observations to train a machine learning model with. With so little data, the models soon started to overfit which is not preferable for the testing data.

We believe that the dataset itself is not a complete information set to properly forecast a team's points total in the next season. Transfer flows of teams have majorly grown in the last years and should have been indexed. In this way, a machine learning model could better distinguish between big and small transactions in a particular season. Other variables could also be included to better forecast a team's performance, for example a trainer switch, or an important player leaving the club.

5 Conclusion

This paper investigated to what extend it is possible to forecast a teams points total based on performances from its direct previous season. The forecasting was done with machine learning models and to be more specific with four different regressions, decision trees, random forests and support vector machines. All models were trained with the same data in an 80%/20% split for training and testing. The performance measure chosen to compare the machine learning models was the mean absolute error (MAE) since it is very clear to interpret. A linear regression was first run to investigate the baseline benchmark the models should attain which turned out to score a 9.4 as MAE but after correcting for multicollinearity this came down to 8.5. After training and testing the models, the regression with lasso regularization scored best with a 7.0 as MAE. All other models with exception of the decision tree, scored just slightly worse with all of them having a MAE smaller than 8 while the decision tree even failed to outscore the baseline linear regression with a MAE of 9.8.

To answer our research question, the forecasted points totals are promising in the sense that they were forecasted acceptable but not with spectacular accuracy. Not getting perfect accuracy is however to be expected since football is in general hard to predict. In conclusion, it is possible to forecast a teams next years points total but the accuracy of the forecast is restricted to an average error of 7.

For further research it could be of value to look at other factors that influence a teams points total. The data we mainly used was of on-field performances but a trainer switch or a star-player leaving the club transfer free due to an expiring contract would not be picked up by our model. To add these bits of information to our model, a future research could add dummy variables for various effects that could happen to the model such that the model becomes more advanced. Another thing to note is that the points total in our research is being forecasted by only the statistics of the previous season while it could also be that certain statistics of more distant seasons in the past could also be of importance.

6 Appendix

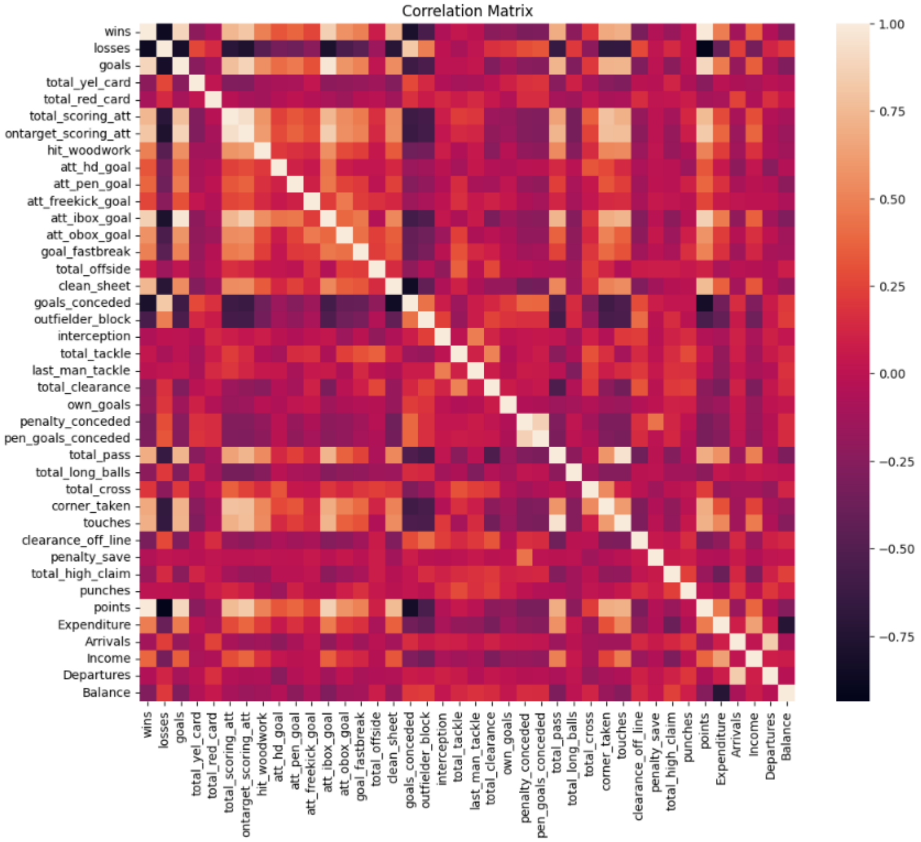

Figure 1: correlation matrix for the data used

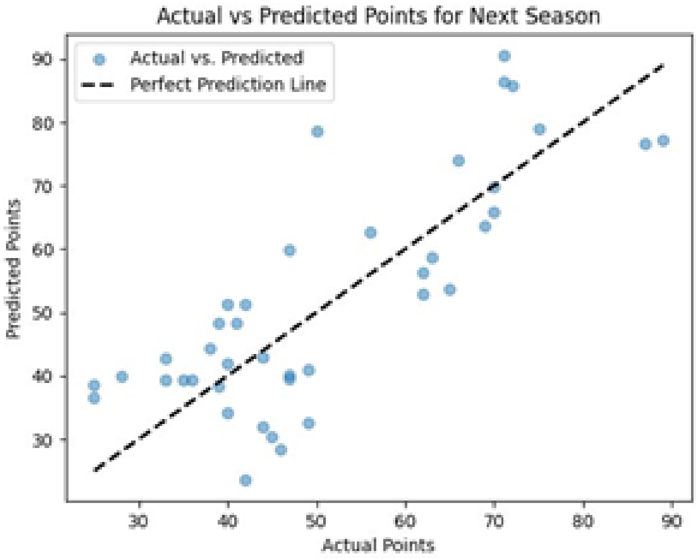

Figure 2: prediction of basic linear regression without VIF correction

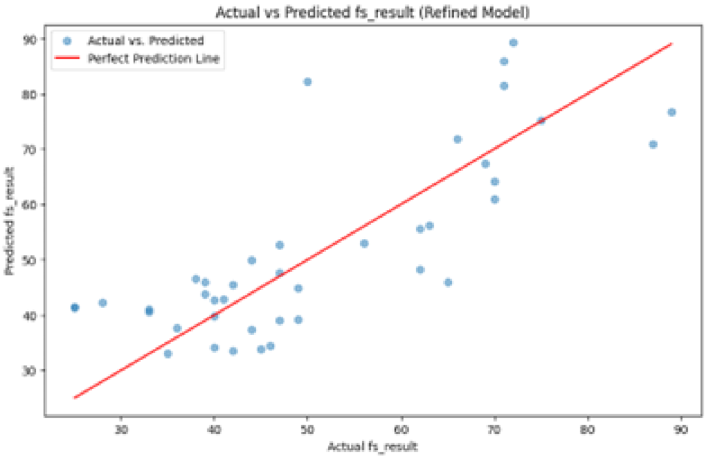

Figure 3: prediction of basic linear regression after VIF correction

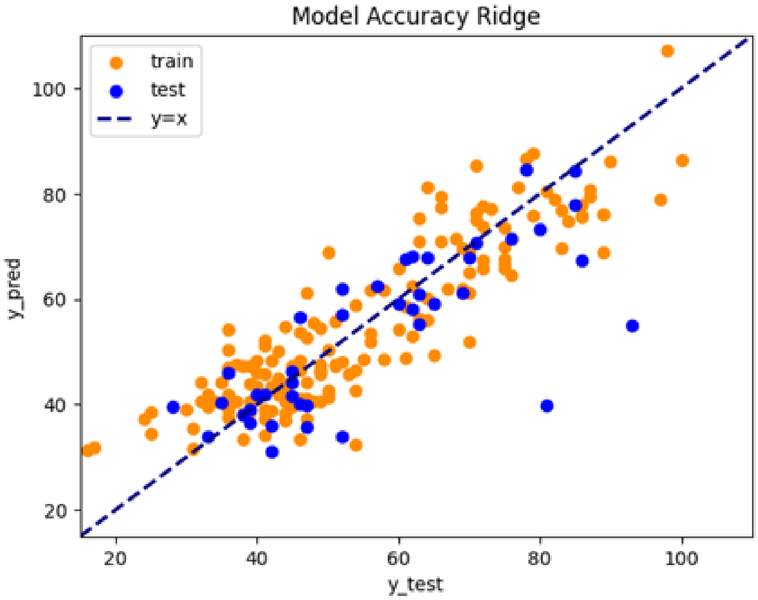

Figure 4: prediction of Ridge regression

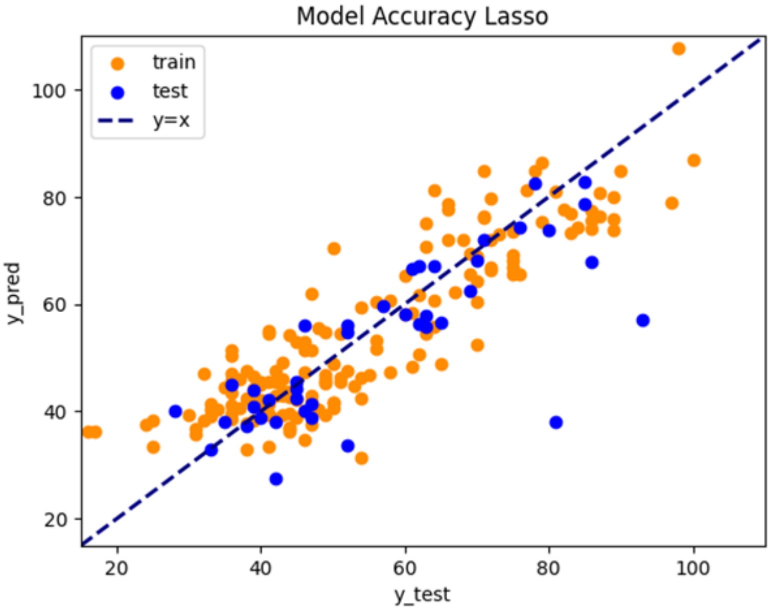

Figure 5: prediction of Lasso regression

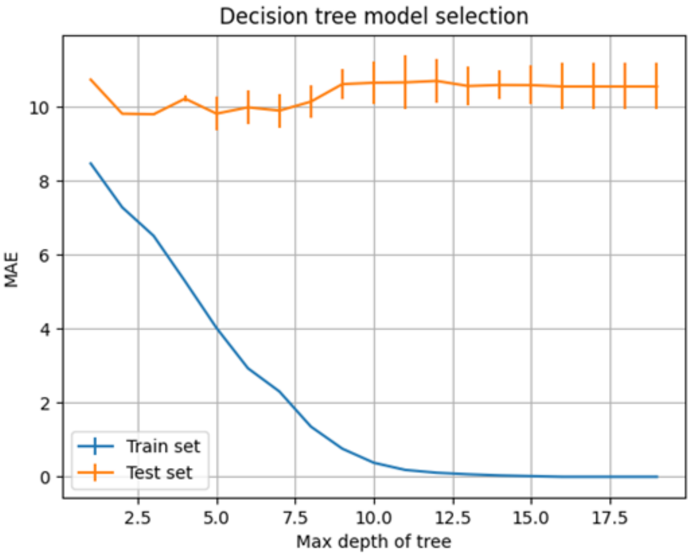

Figure 6: MAE of decision tree

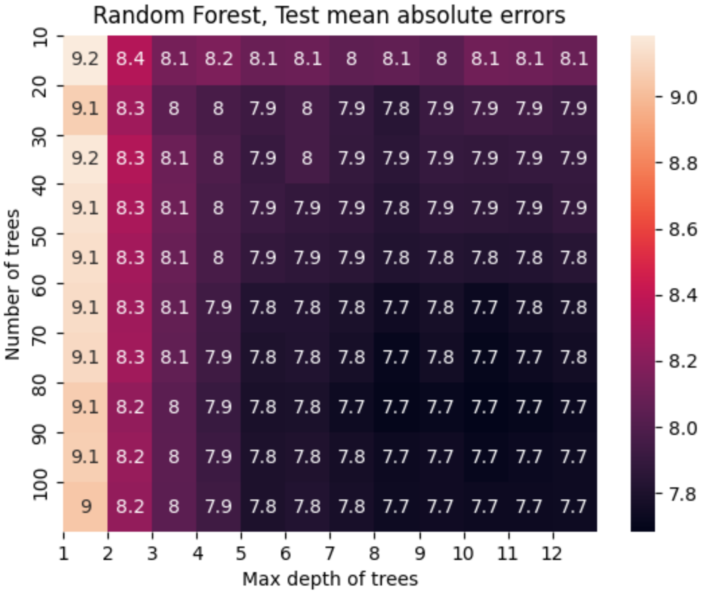

Figure 7: MAE of random forest

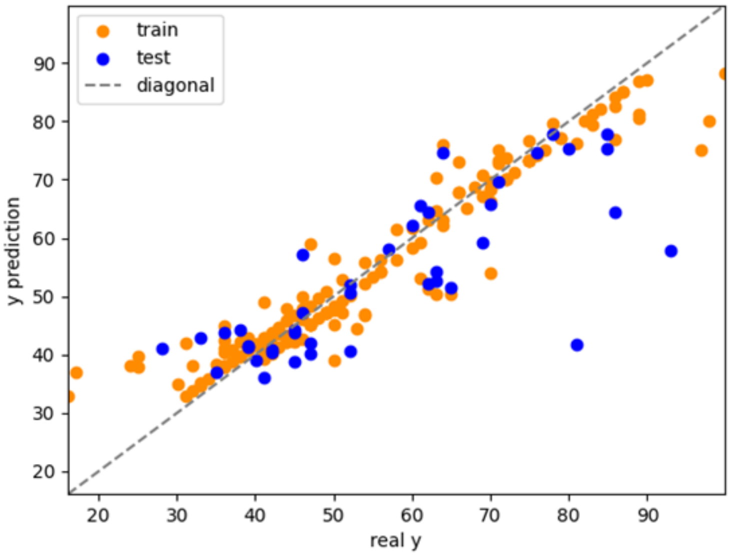

Figure 8: prediction of support vector machine