Machine Learning Final Project

Predicting Premier League Soccer Scores

MATH4050 Machine Learning

Brennan Reamer

April 19, 2023

Introduction

The aim of this project is to predict the outcome of Premier League soccer matches using data from the 2019-2023 seasons. To achieve this goal, several machine learning techniques are employed, including Poisson regression and random forest regression. The data is sourced from [1] and includes various statistics from past Premier League fixtures, such as team names, goals scored, and betting odds. After preprocessing the data, models are created that output predictions for the number of goals scored by the home and away teams. These predictions can then be combined to provide a final score prediction for each match.

While Poisson regression is used to model count data, random forest regression is a model that can be used to handle complex relationships between the features and the output. Both models have their own advantages and disadvantages, but both may be well-suited to this particular task. After fitting the models and making predictions, their performance was evaluated using metrics such as mean squared error (MSE) and R^2 value in order to determine which model performed better.

Assumptions

- The data is accurate as it has been created by an official statistics source for the Premier League.
- All observations are assumed to be independent of each other.
- For the Poisson regression model, it is assumed that the output is a count, a positive integer, and follows a Poisson distribution.

Background

The data used in this project is sourced from [1]. It consists of various statistics from Premier League soccer fixtures from the 2019-2023 seasons, such as the home and away team names, the number of goals scored by each team in each half, the number of yellow and red cards, and the betting odds from various sources.

The data was preprocessed to include only certain columns of interest. These columns included home and away team names, total goals scored, assigned referee, and betting odds from the following sites: [2], [3], [4], and [5]. The inputs for the data were chosen based on their statistic being available prior to gametime, allowing a user of the program to predict the outcome

of the Premier League match before it began. Two models were created for each machine learning technique used, one to output predicted number of goals scored by the home team, and one to output predicted number of goals scored by the away team. Combined together, these two models are then able to provide a prediction of the final score of the match.

Poisson regression and random forest regression were chosen as the machine learning techniques for this project as they are well-suited to the task. Poisson regression is a type of generalized linear model that is commonly used to model count data, such as goals scored in this case. The number of goals scored in a Premier League match can be assumed to be a count variable following a Poisson distribution.

Random forest regression is a machine learning method that uses multiple decision trees to make predictions. This technique was chosen to model the number of goals scored because it is a flexible method capable of handling complex relationships between the predictor variables and the response variable. As soccer matches in the Premier League tend to be unpredictable, a random forest regression model may be well-suited to the task.

Results

When testing the trained models, scatter plots were created to visualize the prediction results. These plots can help to visualize the accuracy of the predictions made by the models. The scatter plots for both Full-Time Home Goals (FTHG) and Full-Time Away Goals (FTAG) showed similar patterns for both the Poisson and random forest regression models, but the random forest regression model can be seen to have more accurate predictions.

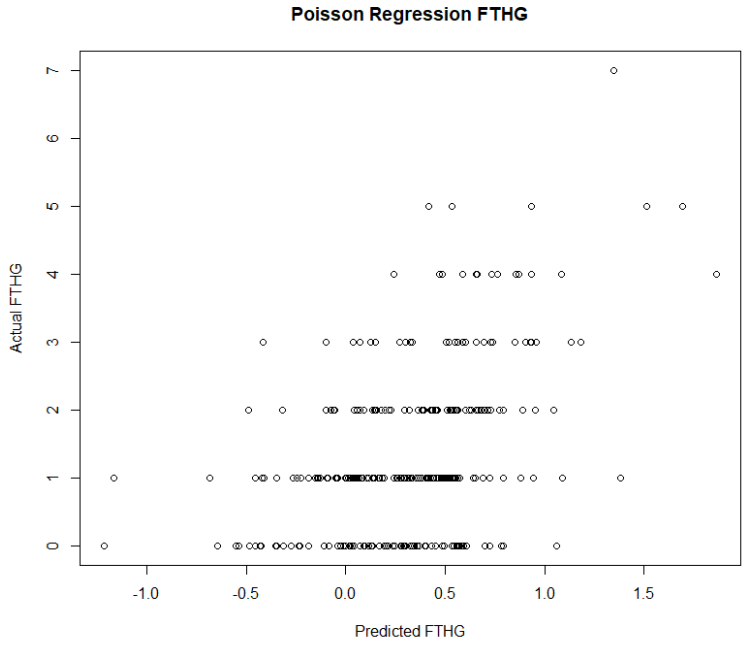

Figure 1: Poisson Regression Full-Time Home Goals

According to *Figure 1*, the predicted values for FTHG did not closely align to the actual FTHG. This may be due to the Poisson model struggling with predicting complex relationships between the features and the output. Poisson regression models typically work best when the relationship between the predictor variables and the response variable is simple and linear. However, in the case of predicting the number of goals scored in a soccer match, the relationship may be more complex and nonlinear, making it more difficult for the Poisson regression model to accurately predict the outcome.

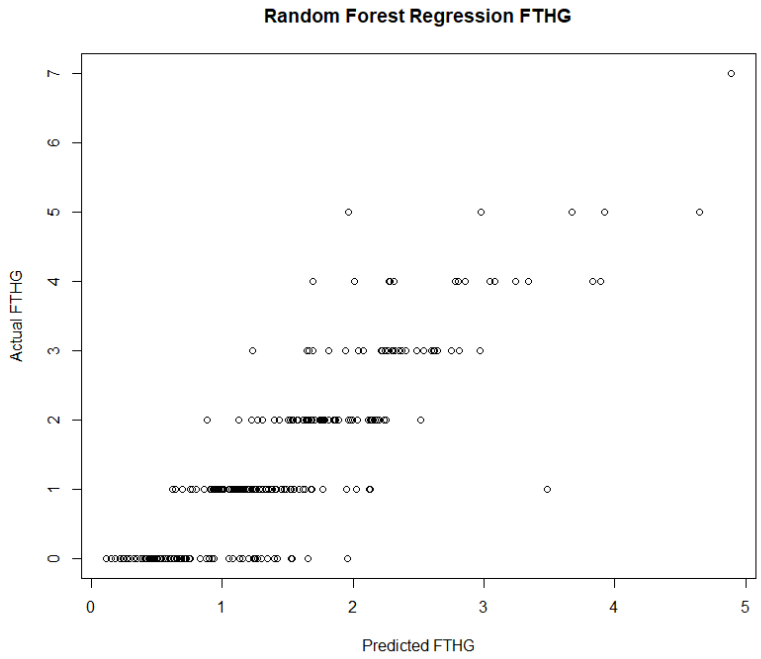

Figure 2: Random Forest Regression Full-Time Home Goals

According to *Figure 2*, the random forest regression model performed well as it was able to accurately predict FTHG values within a range of approximately one to two goals. Random forest regression may have performed better in predicting the number of goals scored because it is a more flexible model capable of capturing nonlinear relationships between the features and the output.

Discussion

After fitting the models and making predictions, their performance was evaluated using several metrics. The mean squared error (MSE) for the Poisson regression model was 4.21 for the home team and 2.25 for the away team, while the MSE for the random forest regression model was 0.68 for the home team and 0.54 for the away team. The random forest regression model has a much lower MSE, meaning it is a more accurate model for the dataset than the Poisson regression model.

The R^2 value for each model was also calculated, which measures how well the model fits the data. The R^2 value for the Poisson regression model was 0.023 for the home team and 0.152 for the away team, while the R^2 value for the random forest regression model was 0.599 for the home team and 0.621 for the away team. The random forest model has a significantly larger R^2 value than the Poisson model, meaning it is much more correlated with the data.

Based on these results, the random forest regression model performed better than the Poisson regression model in terms of both MSE and R^2. This suggests that random forest regression is a more effective technique for predicting the outcome of a Premier League soccer match using this data.

Conclusions

In conclusion, this project aimed to predict the outcome of Premier League soccer matches using machine learning techniques, specifically Poisson regression and random forest regression. After preprocessing the data and fitting the models, their performance was evaluated using metrics such as mean squared error and R^2 value. The results showed that the random forest regression model performed better than the Poisson regression model in terms of accuracy and correlation with the data. Therefore, the random forest regression model is recommended for predicting the final score of Premier League soccer matches using the data from the 2019-2023 seasons. Overall, this project highlights the potential of machine learning in sports analytics and its ability to provide valuable insights for both fans and professionals in the sports industry.

While the Poisson regression and random forest regression models were used in this project, there are several other machine learning techniques that could be applied to this task. One such technique is the Support Vector Machine (SVM) algorithm, which is often used for both classification and regression. The SVM algorithm has been shown to be effective in predicting soccer match outcomes, especially when combined with feature selection techniques to identify the most important variables.

Another technique that could be used is the neural network model, which has been applied to soccer match prediction with promising results. Neural networks are able to handle non-linear relationships between the input and output variables, making them a suitable choice for predicting complex relationships like soccer match outcomes.

The reason why these models may perform better is because they can handle more complex relationships between the features and the output. For example, neural networks are able to learn and represent complex patterns in the data. In this project, Poisson regression and random forest regression were chosen as they were well-suited to the task at hand, and their performance was evaluated using several metrics, such as MSE and R^2 value. It's possible that other models may have performed better, but given the results of this project, random forest regression is recommended for predicting the outcome of Premier League soccer matches using this data.

References

[1] “England football results betting odds: Premiership Results & Betting odds,” *Football Betting - Football Results - Free Bets*. [Online]. Available: https://www.football-data.co.uk/englandm.php. [Accessed: 15-Apr-2023].

[2] “Premier League betting,” *Odds Checker*. [Online]. Available: https://www.oddschecker.com/football/english/premier-league. [Accessed: 15-Apr-2023].

[3] “England Premier League Betting,” *Interwetten*. [Online]. Available: https://www.interwetten.com/en/sportsbook/l/1021/england-premier-league. [Accessed: 15-Apr-2023].

[4] “Premier League Betting,” *Bwin*. [Online]. Available: https://sports.bwin.com/en/sports/football-4/betting/england-14/premier-league-102841. [Accessed: 15-Apr-2023].

[5] “Premier League Betting,” *Pinnacle*. [Online]. Available: https://www.pinnacle.com/en/soccer/england-premier-league/. [Accessed: 15-Apr-2023].