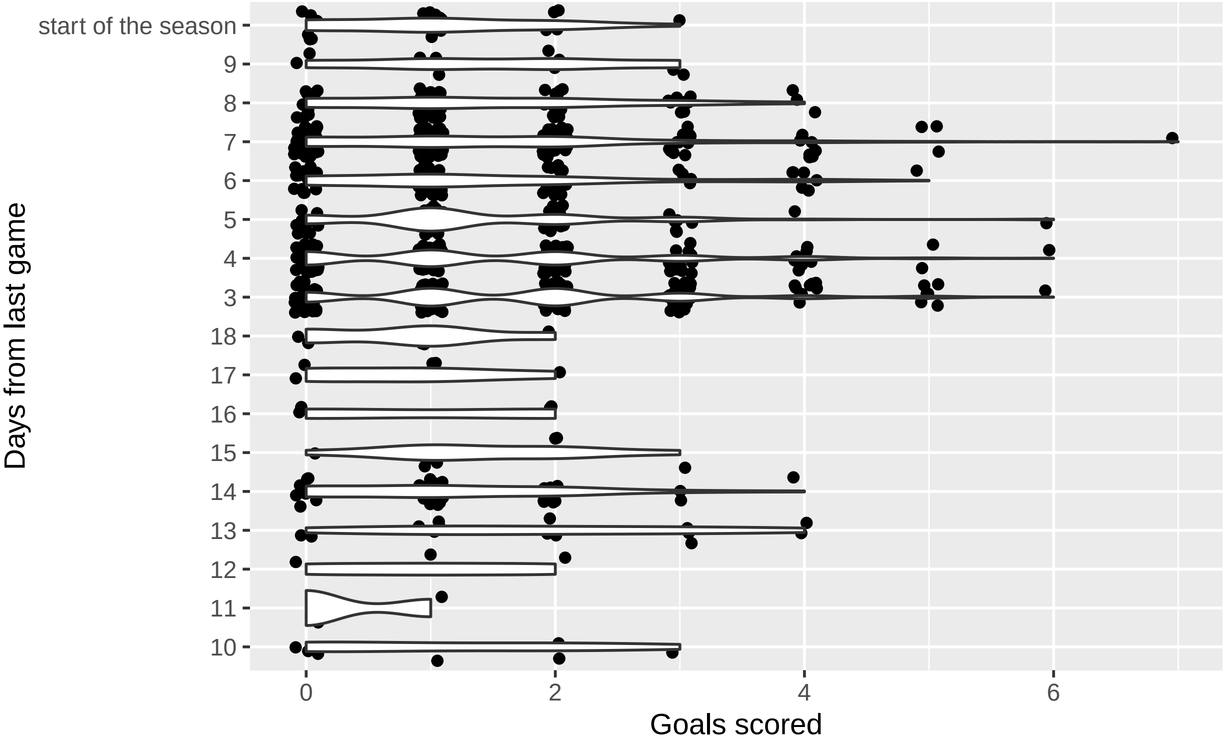
Premier League regression models

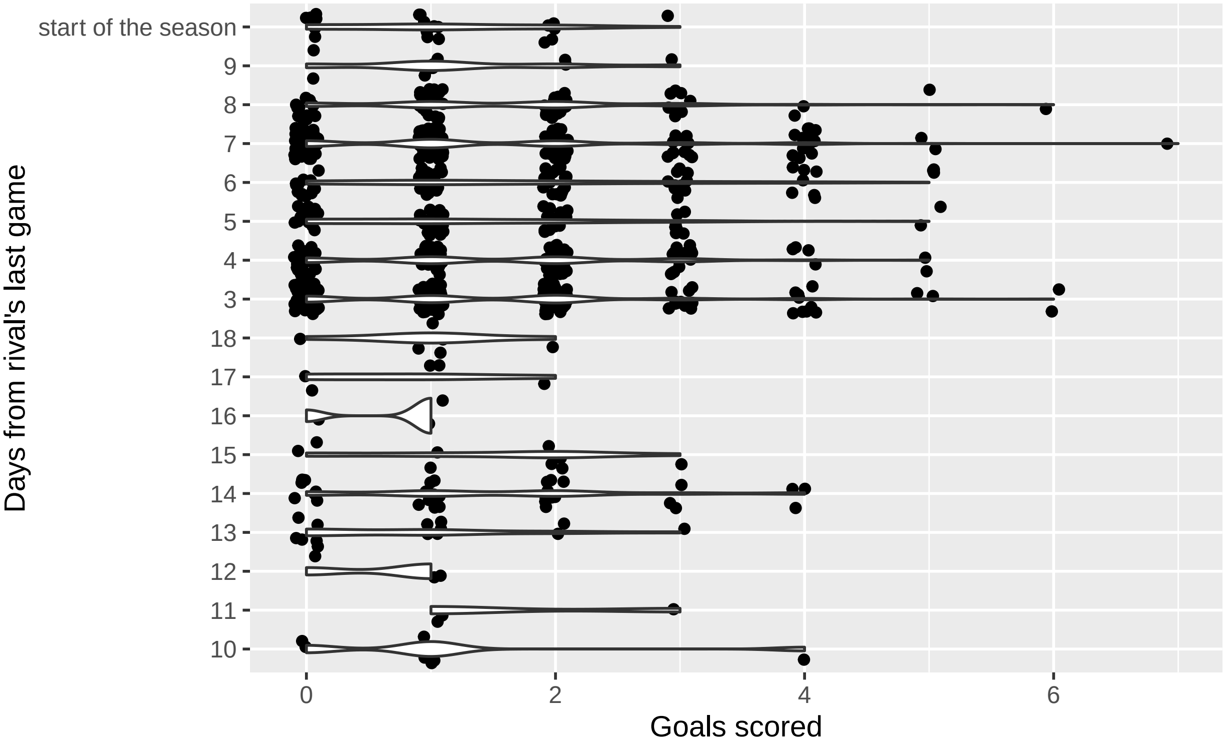
Michał Porczyński

2025-08-28

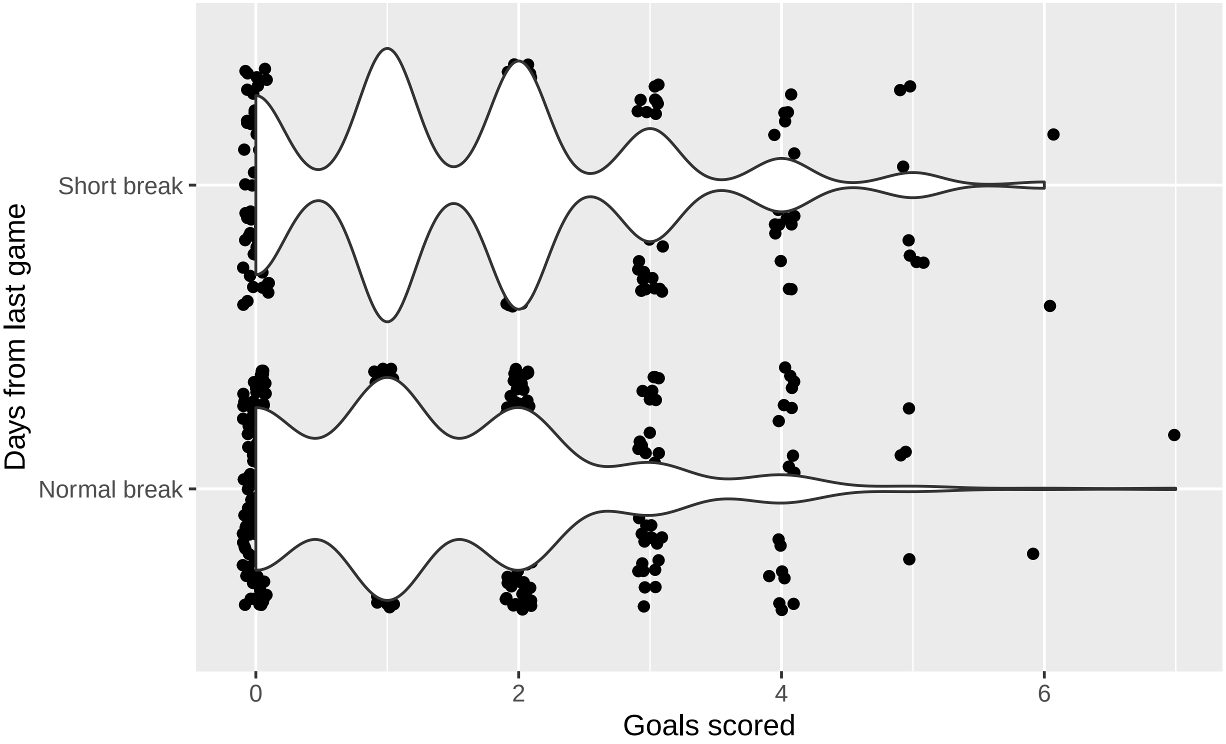
The number of goals scored is one of the most important outcomes in football matches. Predicting goals can provide valuable insights for team analysis, player evaluation, and match forecasting. The aim of this report is to build a regression model that estimates the number of goals based on available match data. We are going to base it on Elo ranking, which provides information about a team's strength, formation used by a team in a match, days since the last game, and goals scored and conceded by a team and its opponent in the last three Premier League games. We will use results from the Premier League 2024/2025 season and then test the best possible model on data from the 2025/2026 season up to Gameweek 2.

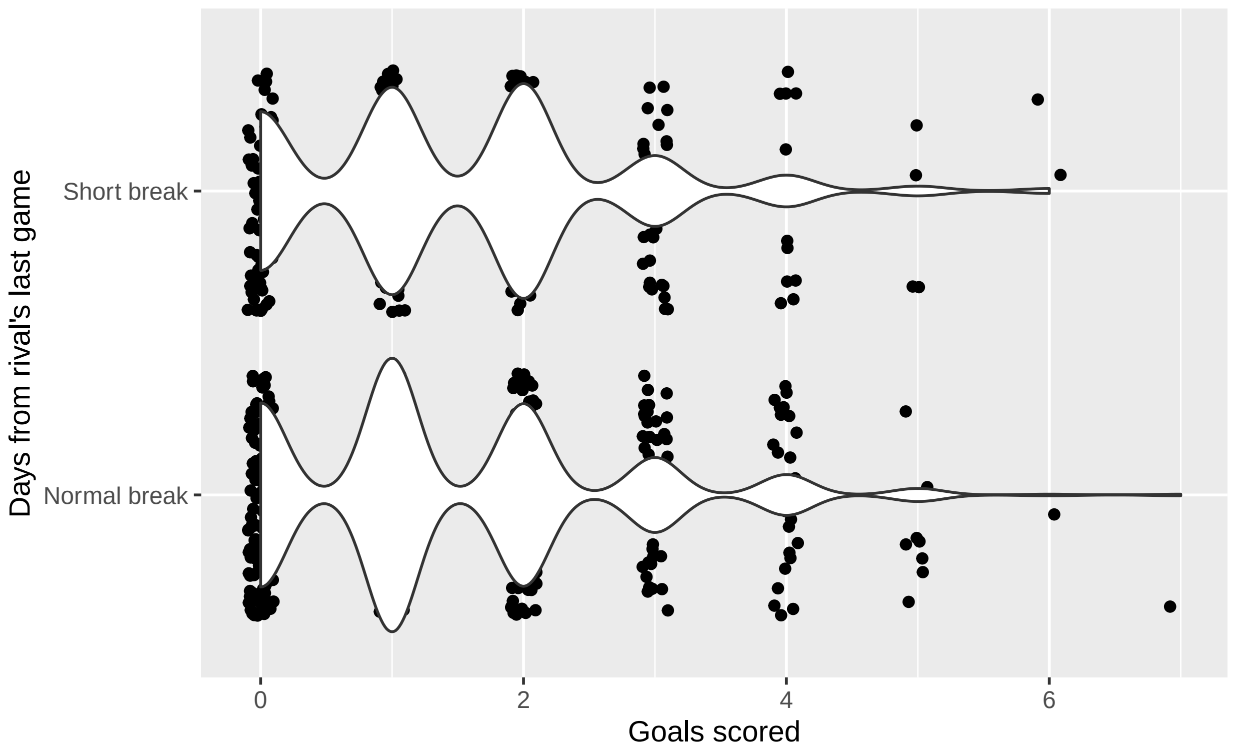
Before building a model, we would like to clean our data and group it properly. I am going to use decision trees to form buckets and then group specific data into categories. I will do this with a few variables — Team, Formation, Days from last game, and Goals scored in the last three games. The same applies to data for the team's opponent. For instance, let's look at the goal chart over days since the last game.





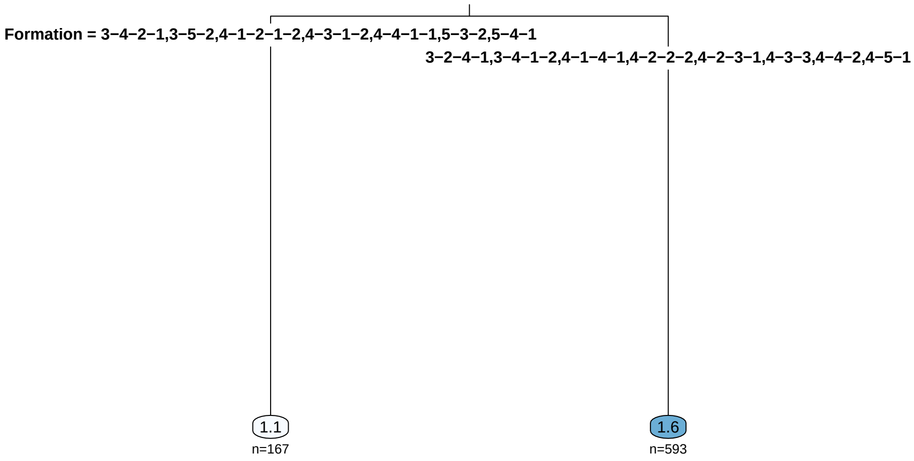
As you can see, there are many possibilities, and the distributions among specific values are quite similar. Some of them also do not consist of many observations, which is why we are going to form them into two groups, referring to whether the time from the last game was a short break (3 or 4 days) or a normal break (5 days or more). Below you can see how the charts look after transformation.

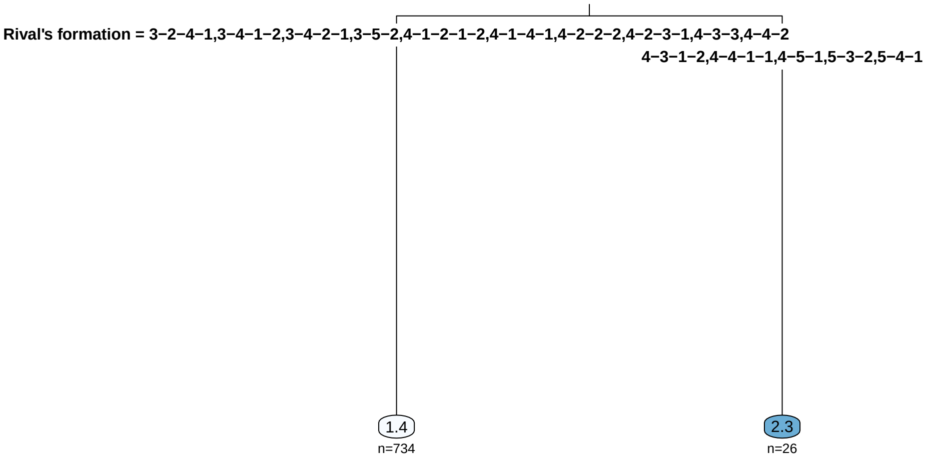




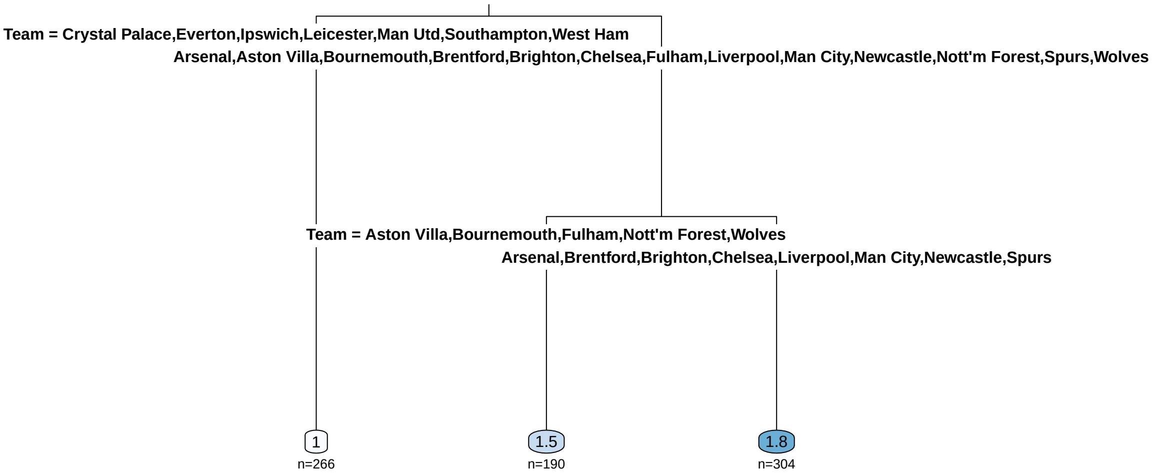
As you can see, they seem quite similar, but in general, when we look at the 'Days from last game' variable, I would say that when a team has a short break before a game, it is more likely to score more goals than when there is only a normal break. This may be because, in general, better teams that play in cup competitions like the Champions League have less time to rest before a game.

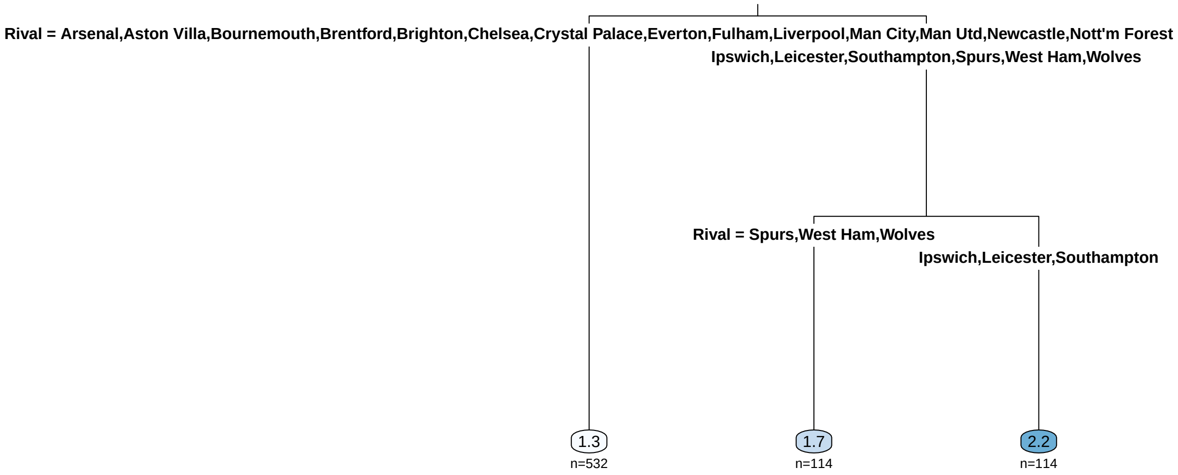
The same transformation will be applied to the variable 'Formation', but in this case, we will use the decision tree. As you can see below, the tree gives the best possible partition for the variable 'Formation', and that's how we are going to divide observations from the variables 'Formation' and 'Rival's formation' into two groups.



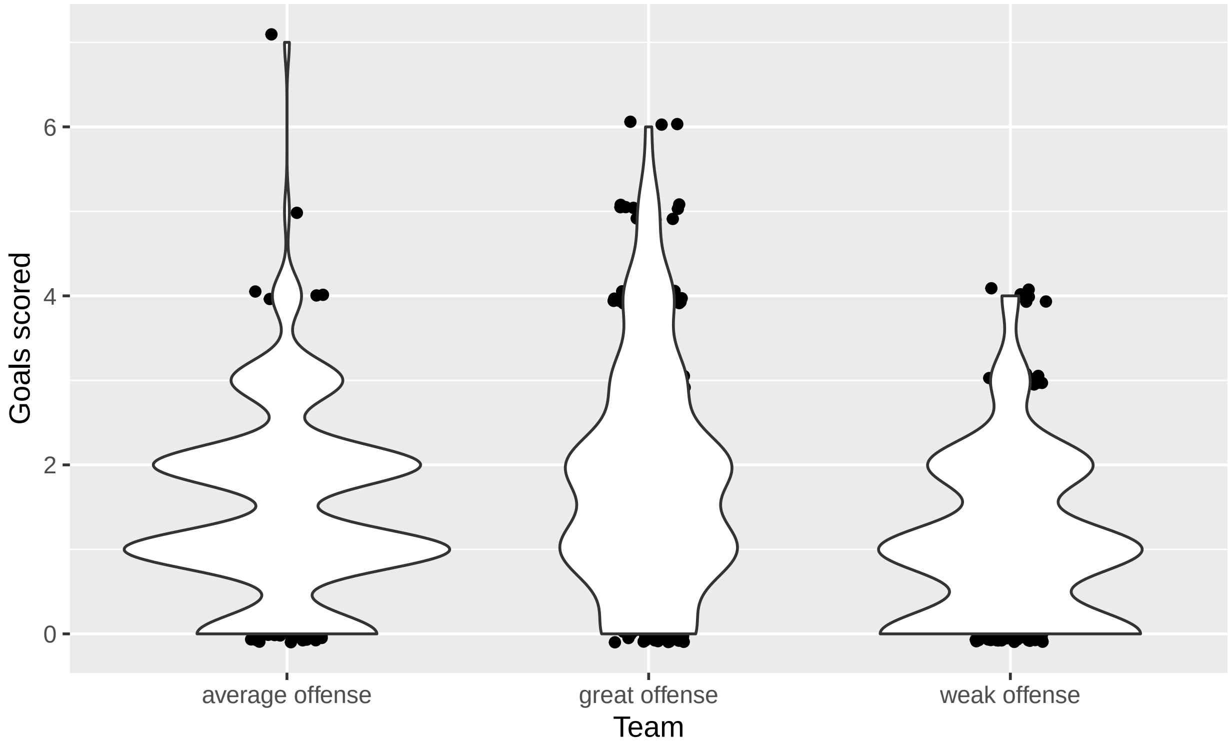

Rival's formation = 3−2−4−1,3−4−1−2,3−4−2−1,3−5−2,4−1−2−1−2,4−1−4−1,4−2−2−2,4−2−3−1,4−3−3,4−4−2
4−3−1−2,4−4−1−1,4−5−1,5−3−2,5−4−1
1.4
n=734
2.3
n=26

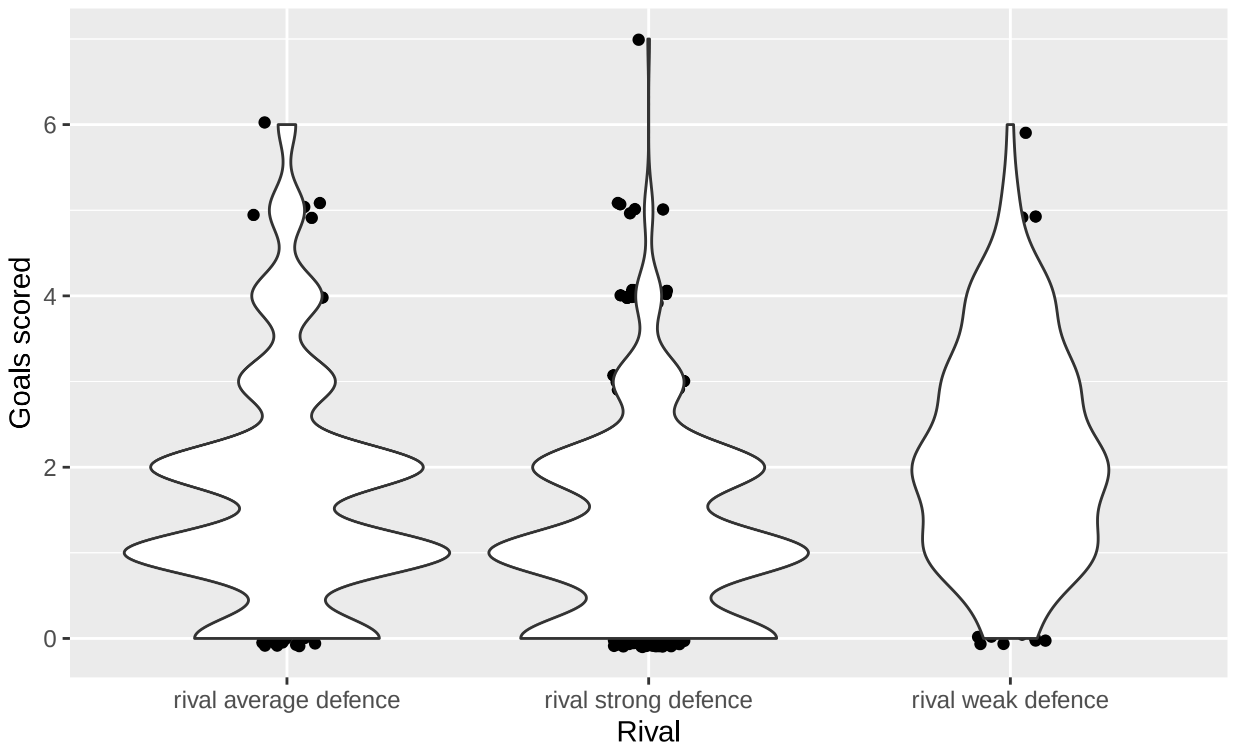
The same method with a decision tree was used to divide observations for the variables ‘Goals scored in the last 3 games’ and ‘Goals conceded in the last 3 games by the opponent’. The first group is for less than four goals scored and more than three goals scored, and the second group is for observations with ‘less than seven goals conceded’ and ‘more than six goals conceded’.


Team = Crystal Palace,Everton,Ipswich,Leicester,Man Utd,Southampton,West Ham
Arsenal,Aston Villa,Bournemouth,Brentford,Brighton,Chelsea,Fulham,Liverpool,Man City,Newcastle,Nott'm Forest,Spurs,Wolves
Team = Aston Villa,Bournemouth,Fulham,Nott'm Forest,Wolves
Arsenal,Brentford,Brighton,Chelsea,Liverpool,Man City,Newcastle,Spurs
1
n=266
1.5
n=190
1.8
n=304


Rival = Arsenal,Aston Villa,Bournemouth,Brentford,Brighton,Chelsea,Crystal Palace,Everton,Fulham,Liverpool,Man City,Man Utd,Newcastle,Nott'm Forest
Ipswich,Leicester,Southampton,Spurs,West Ham,Wolves
Rival = Spurs,West Ham,Wolves
Ipswich,Leicester,Southampton
1.3
n=532
1.7
n=114
2.2
n=114

That partition is quite reasonable because, in general, teams with weaker offense are in the same group as newly promoted teams or Wolves, while teams like Arsenal, Man City, or Liverpool are in the group with stronger offense. The same applies to the team's defensive strength.

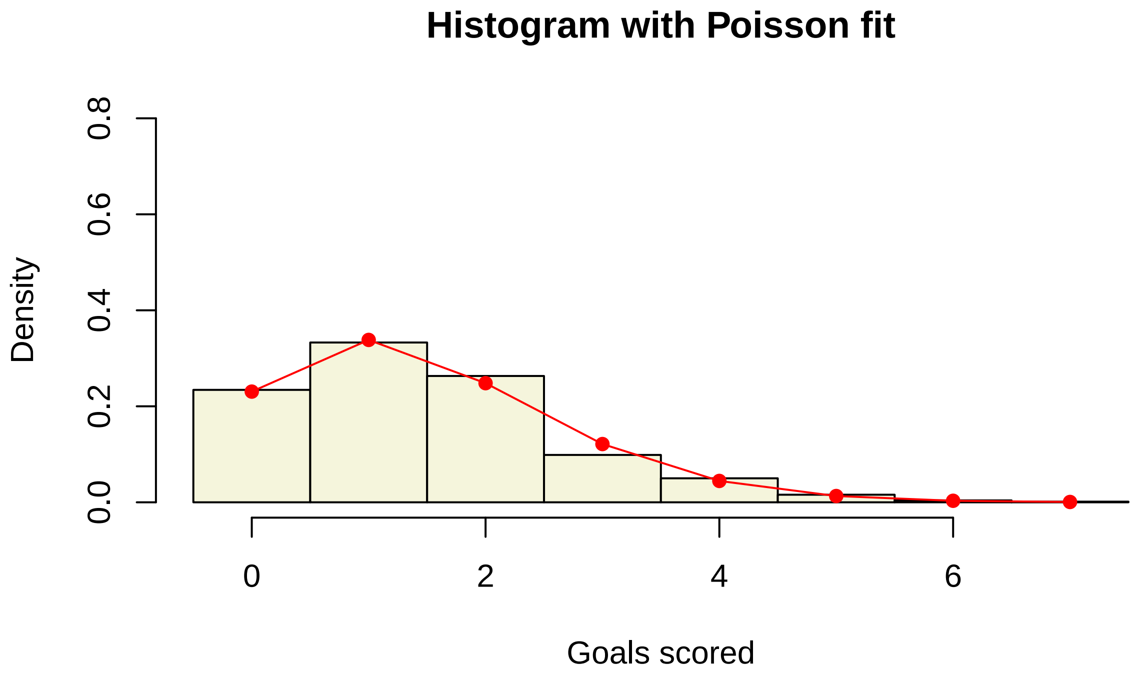




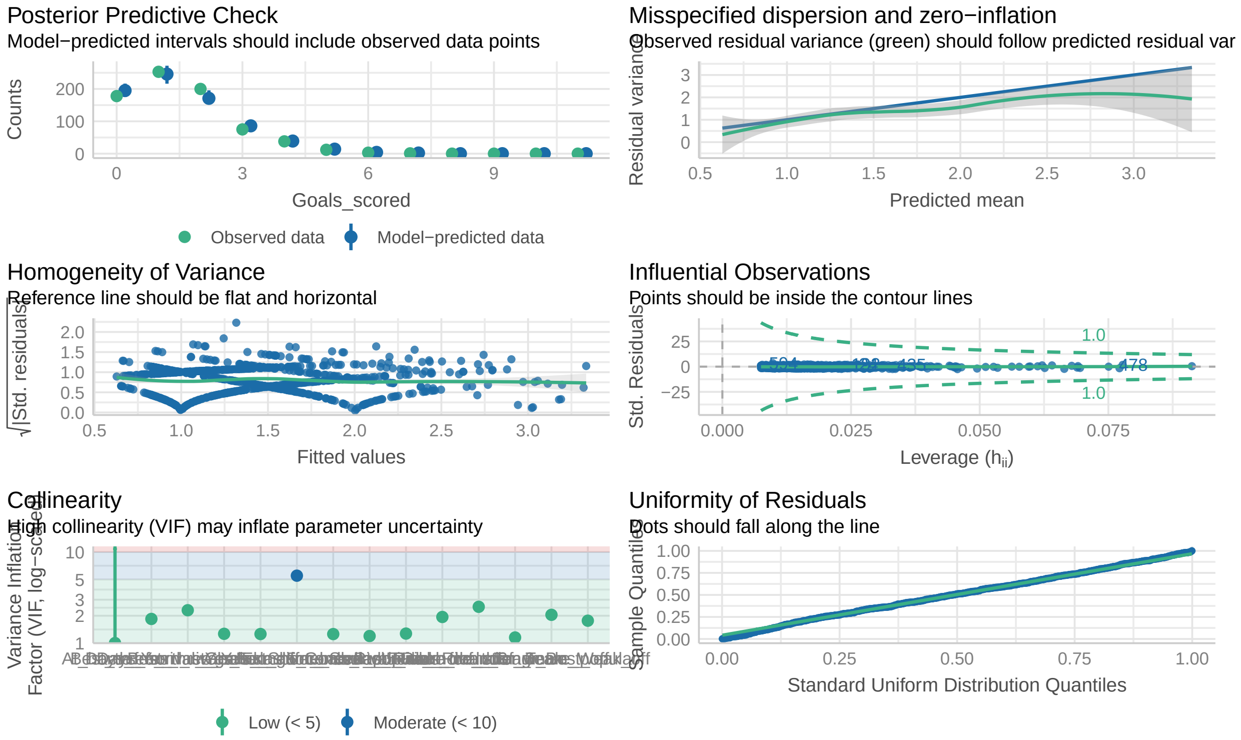
There you can see how it looks. In general, teams with weaker offense score fewer goals, while teams with stronger offense are more likely to score more goals. The same applies to the opponent's defensive strength. If a team's opponent has a weak defense, the team is more likely to score more goals.

Finally, we are going to build a model. We are going to use the Poisson Regression Model with or without interactions. We will also consider the Negative Binomial Regression Model in case of overdispersion. In our data, there might be issues with that because the mean (≈ 1.46) is quite similar to the variance (≈ 1.52). If we encounter zero inflation, we will also have to consider models that handle zero inflation, like the ZIP model, but it is probably unnecessary.

I should also mention why I decided to use the Poisson Regression model. In general, we use it when we deal with countable data that represent the number of times a specific event happens in a given period. Here, the event is goals scored by a team in a game, and the period is one game (ninety minutes). Below, you can also see a histogram of goals scored by teams, which is quite similar to a Poisson distribution, apart from values two and three where the theoretical distribution does not match the empirical probability.

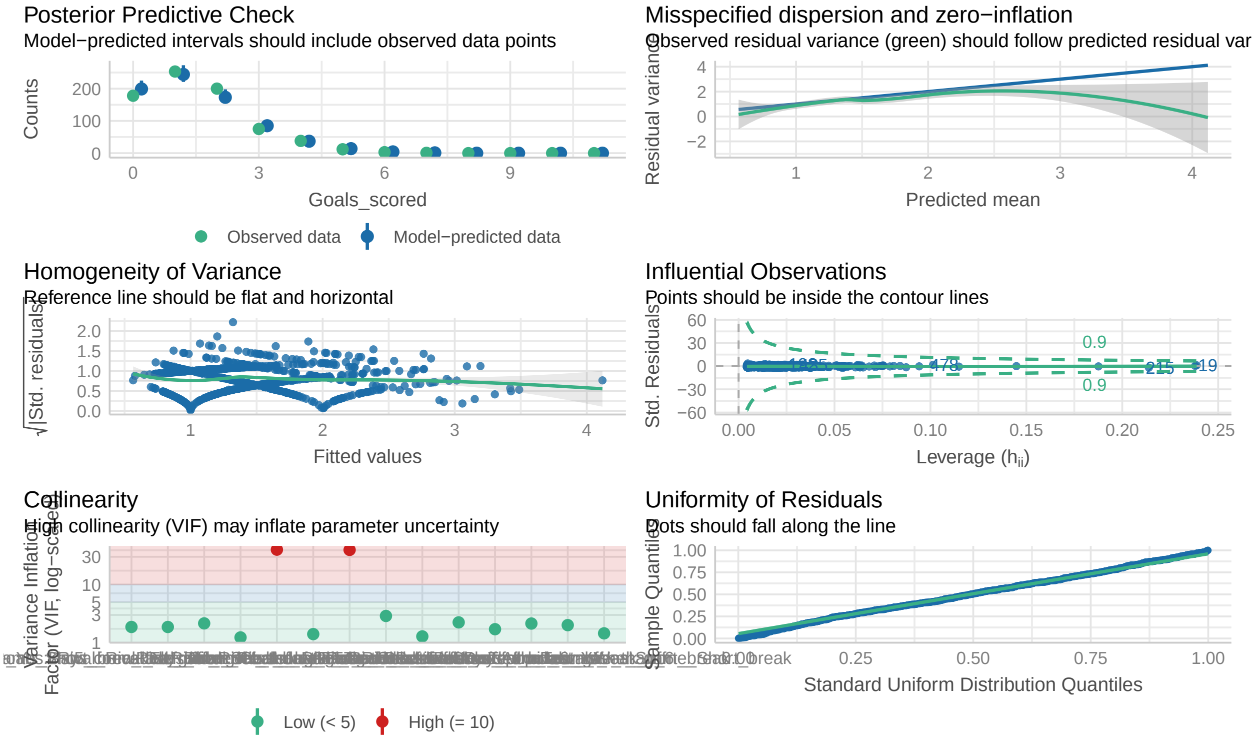


Now we can move on to the process of model bulding. Im a going to present a couple of Poisson Regression models. The first one will be the easiest one, where we will use all possible variables in model building. Then we are going to check collinearity between variables, overdipsersion and zero inflation.

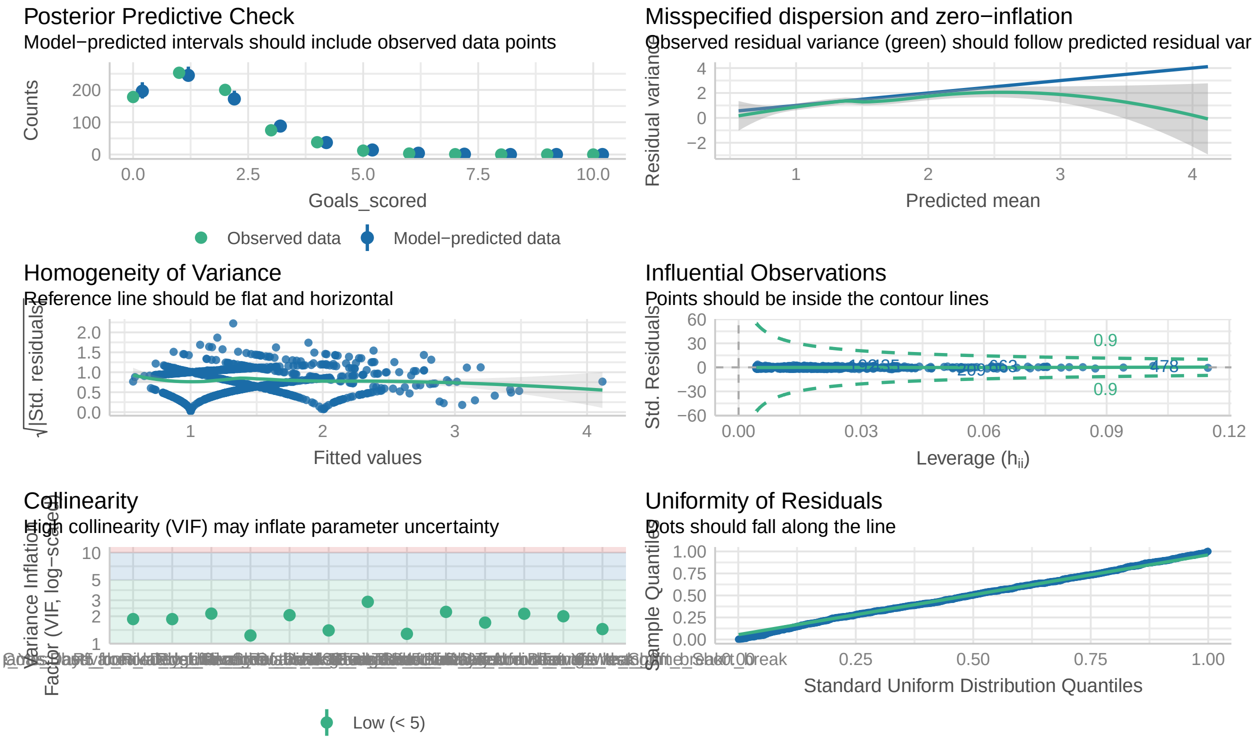


This model does not look bad. The main drawback is that it underestimates the number of two goals scored. When we look at other plots, for instance the one showing collinearity, we see that one variable has a VIF value above 5 — Elo difference. Its VIF value is 5.5, but it is a crucial variable for our model, and apart from it, the VIFs are not high, so I am not going to remove it. Thanks to the performance library, we can also check that we do not have any problems with overdispersion or zero inflation (probably a Negative Binomial model would not improve things).

Now the second model, where I am going to introduce interactions to the model. But I am also going to build that model wisely because of the amount of variables in my model so I am going to use Lasso Regression which will get rid of unnecessary variables in my model. Then I am going to check model's performance.

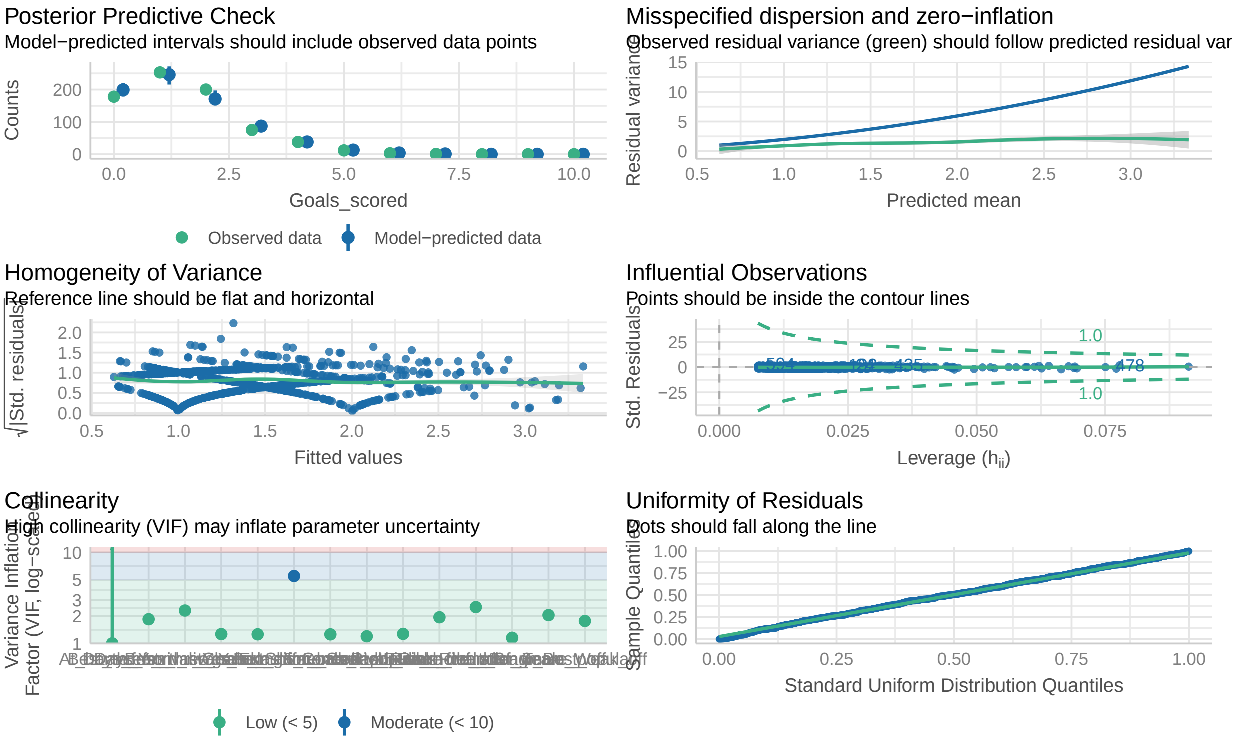


At first glance, there are predictors that are highly correlated. I had to remove one of these variables (in my case, the interaction between two dummy variables) and see how the model looks after this reduction.



It is much better now because the model does not suffer from multicollinearity. Unfortunately, there is still a problem with underestimating two goals scored, and the model slightly overestimates zero goals scored. This shows that zero inflation would not be a good choice because we do not have problems with zero inflation. Overdispersion is also not an issue.

The third model is going to be negative binomial regression model without any interactions. I would like to see if this model will help to improve predictions.



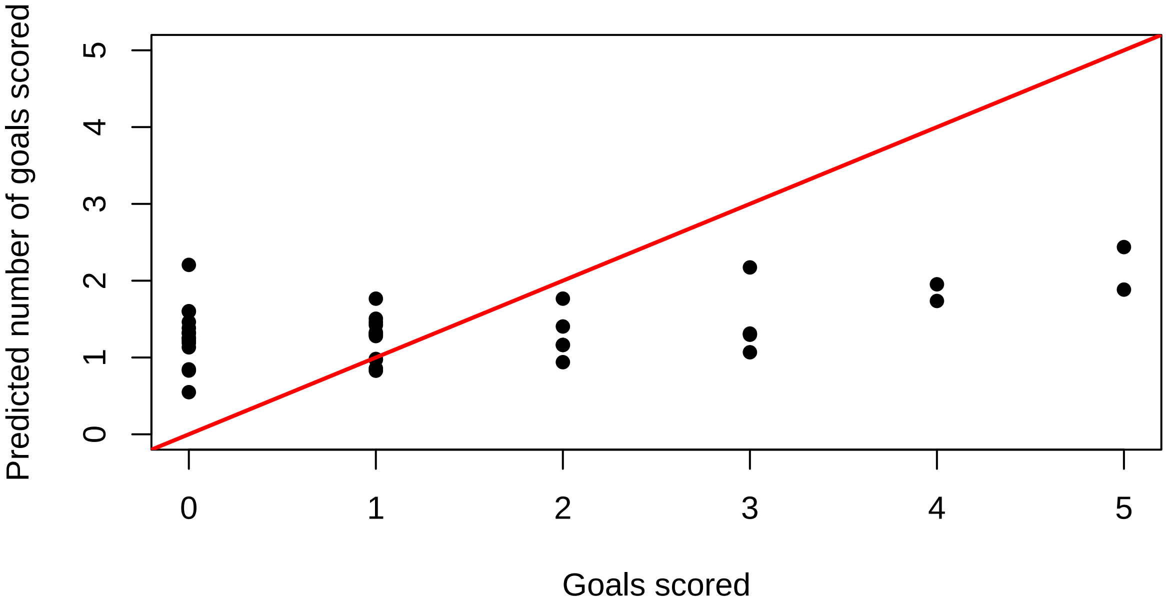
For me, the Negative Binomial model would not be a good choice, especially when we look at the plot with misspecified dispersion where residual variance does not follow the predicted mean. This probably indicates that we did not have problems with overdispersion.

Know it is time to compare models that we have built. We are going to look at the number of features in the model, AIC and BIC values and also how well it predicts zero, one, two and three goals scored because they are in general the most probable numbers of goals scored in a football game.

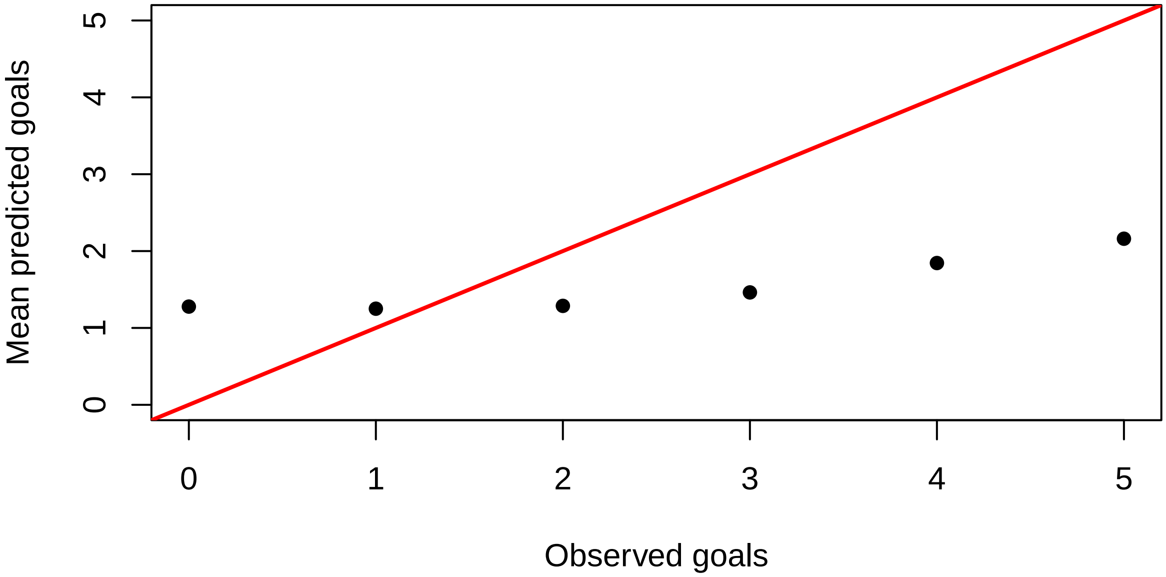
Generally, as I mentioned before, Negative binomial model has not improved anything and it is unnecessary to use if we look closer at AIC and BIC values. We can also test a null hypothesis that these models are statistically equivalent vs alternative hypothesis that negativ binomial is much better. We just have to calculate deviance statistics which is $D = \text{deviance for poisson model} - \text{deviance for negative binomial model} = 770.20 - 770.15 = 0.05$ and then compare it with quantile from chi squre distribution with one degree of freedom. As we see $D = 0.05 < F^{-1}_{\chi^2_1}(0.9) = 2.71$ so we cannot reject null hypothesis.

Looking at AIC and BIC, the LASSO Poisson model is better and has fewer coefficients. It has slightly worse predictions compared to the default Poisson model, but the difference is not large, so I am going to choose the LASSO model to see how it performs on the test data.

I have prepared test data, which I collected from two first gameweeks from Premier League season 2025/2026. We are going to see how good the predictions are generally looking at the plot Predicted goals scored over Goals scored.



As you can see in the plot, generally the chosen model does not give very accurate predictions, or at least there is room for improvement. In general, the model rarely predicts more than two goals when a team scored zero or one goal. In other cases, it depends on the situation, but I would say that when more than three goals were scored, the model predicts that around two goals are expected. Below you can see a chart with the mean predicted goals for each observed goals group.



There is no significant difference between groups with 0, 1, or 2 goals scored. There is a slight difference when we look at groups with 3, 4, or 5 goals scored.

To sum up, I would say that we can use our model to get an idea of expected outcomes for each game, such as which team is likely to win or lose, but I would not recommend using it to predict the exact number of goals scored. Probably, some machine learning models should be used for that, or I should consider adding other features to the model.