Analyzing Premier League Data Using Cross Validated Logistic Regression

Andy Elgin

Department of Mathematics and Statistics
Coastal Carolina University

1. Introduction

THE Premier League has grown into the most prominent and lucrative football league in the world. Billions of dollars are spent annually on players and television rights. Every year, 20 clubs vie for the title of Premier League Champion. At the conclusion of the season, a club's place within the league table determines many things, such as intercontinental tournament bids (Champions League and Europa League) and monetary rewards. Last year, Manchester City received over $54 million for winning the Premier League.The bottom 3 clubs are relegated, or demoted, to a lower division for the following season. Classifying relegation as binary (yes or no), we can then investigate potential factors that influence the probability of relegation.

Some things to make note of as we progress through the study:

- The time frame of interest was from the 08/09 season to the conclusion of the 21/22 season.
- Avg. Squad Age refers to the average age of the entire lineup in years.
- Net Spend is calculated by: revenue-expenditure of a club for that season.

2. The Data

GATHERING data from the Premier League website archives, we collect information on 10 teams from each of the 14 seasons of interest, giving a total of 140 observations. From this website, detailed statistics of each club's season was gathered, and compiled into a singular data set.

For each of the 140 teams, the following variables were considered:

- The net spend of the club for that particular season
- Average age of that year's club
- The final position of the club on the league table

Additionally, we will define our response variable as 1 if the club was relegated and 0 if the club was not.

3. Cross Validation

CROSS validation is a method of evaluating how well our model performs at predicting probabilities. The entire data set is partitioned into two pieces: the training set and the testing set. The training set, which is roughly 80% of the data set, is used to create a predictive model, while the test set is "hidden" from R. This allows us to develop a model and see how well it performs on the unseen portion of data. In this instance, we utilized a 10-fold cross validation method. This means the data was split into 10 folds, and the remaining 9 folds are used to train the data. The process is repeated until all 10 folds are used as the test set once. Upon implementation of this in R, the following accuracies are obtained:

- Internal estimate of accuracy: 0.671
- Cross-validation estimate of accuracy: 0.686

Accuracy is defined as the proportion of correct predictions over the total predictions. Nearly 70% of the predictions were correct using this k-fold validation technique.

4. The Logistic Regression Model

RELEGATION is our binary response variable of interest. Using R, we can develop a regression equation that takes into account the different predictors β_1 to β_n:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Using this general format, the estimated logistic regression equation using our predictors is as follows:

$$\log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right) = -10.6053 + 0.009(NetSpend) + 0.3788(AvgAge)$$

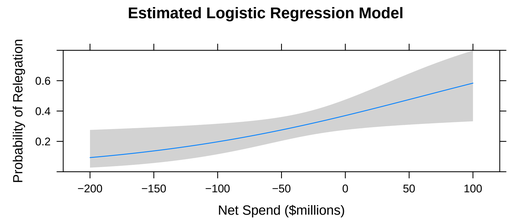
In addition to the coefficients of the model, the corresponding p values are just as important to understanding the output. Both net spend and average squad age were statistically significant, with p values of .0214 and .0382 respectively.

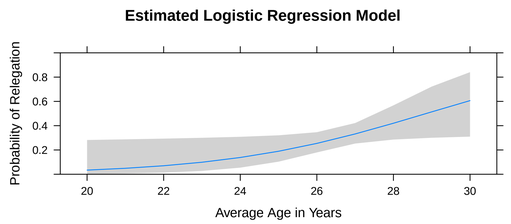
Now that we have an equation, we can use it to estimate probabilities, denoted by $\widehat{\pi}$, based on our desired explanatory variable values. To estimate probability values, the following equation is used:

$$\widehat{\pi} = \frac{e^{\beta_0+\beta_1(x_1)+\beta_2(x_2)}}{1+e^{\beta_0+\beta_1(x_1)+\beta_2(x_2)}}$$

Inserting the values of β from the regression model along with net spend values and average age will result in the probability of relegation. For example, a team with a positive net spend of $20 million and an average squad age of 26.5 years has a 40% probability of facing relegation. Odds ratio confidence intervals are another way to explain the results we are seeing:

- Average Age: As the average age of a club increases by one year, the odds of being relegated increase by a factor of between 1.0285 and 2.115.
- Net Spending: As net spend increases by $1 million, the odds of relegation increase by a factor of between 1.002 and 1.017.





The plots above depict the logistic regression models we have created, along with 95% confidence bands for the predicted probabilities corresponding to each predictor of interest.

5. Conclusion

SINCE we are estimating probabilities, it can be difficult to draw any concrete conclusions about particular outcomes. However, using logistic regression models is one of the most powerful tools when it comes to modeling binary responses. While it may seem intuitive that the more a club spends on players, the less probable they are to be relegated, there have been instances where this is not the case.

To further build on this study, data from earlier seasons could be included to determine how/if results would vary. Should this be the case, it is important to take into account currency value and how it fluctuates over time. Additionally, with a much larger dataset, a more powerful form of cross validation could be conducted. As the number of folds increases, cross validation requires stronger computational power. Although this study stops at the conclusion of the 21/22 season, further research can be conducted to see if this model holds up with current and future seasons.

References

[1] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[2] Zhang, Liu, C.-A. (2022). "Model averaging prediction by K-fold cross-validation". *Journal of Econometrics*.