MODELLING STADIUM UTILIZATION

Utilizing Modelling techniques in R

ABSTRACT

This research focuses on analyzing stadium utilization of all 20 clubs in the English Premier League 2018/19 season of all games played. The independent variables identified and utilized are home points per game, away points per game, and total goals scored to capture the stadium utilization throughout the season in terms of points fluctuation and forming a cause-effect relationship. In particular, this project tests the potential of different statistical models to predict stadium utilization based on home and away points per game, and total goals scored. Models including linear regression, logistic regression, and several general additive models were used and produced interesting results, yet the capabilities of these results are limited due to the lack of larger observations. That being said, the utilization of these statistical models can expand existing and future studies of larger data sets to make extensive predictions of stadium utilization all over the world and would be useful for club CEO's, stadium managers, and other club facilities managing staff.

Aayush Sethi

DS 612

1. Research Question

The objective of this paper is to statistically analyze the stadium attendances for the 2018/19 Premier League season and attempt to capture the attendance trends by various modeling techniques and studying the points per game for home and away teams as well as the relative strength levels of the teams.

However, the purpose of this research is not ignoring the financial aspects of stadium attendance. It is a fact that football is not like a usual business where the ultimate goal is profit, for club's ultimate measure of success is scoring goals, acquiring as many points as possible and winning trophies.

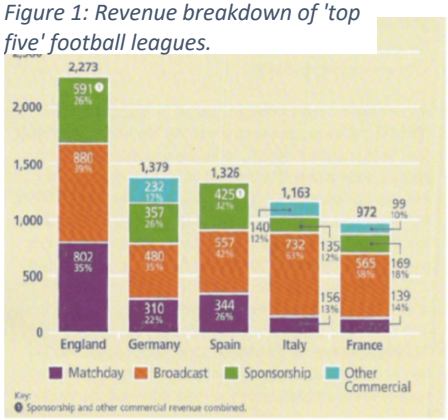
So, for the sake of performance research, we will dwell in questions that can be phrased as follows:

- Do points per game of the home and away teams affect attendance?
- Do total goals scored have a significant effect on attendance?
- What models and methods are most efficient for the given research?

2. Introduction

The English premier league (EPL) is by far the most popular and entertaining league for football fans around the world, with a potential audience of 4.7 billion worldwide and 3 billion dollars in revenue per annum. Compared to the Superbowl which generated 419 million, the premier league is therefore considered the golden class of sports league. The EPL generated a record-breaking 13.8 billion pounds in 2017/2018 season, with the clubs reporting an astonishing 4.8 billion in total revenue. Deloitte has previously reported that 70% of the clubs generated revenue stream is the matchday income and broadcasting in England. Nonetheless, there are many clubs who have been successful in building a strong brand and an international fanbase resulting in high stadium utilization throughout the season. These clubs are now in need to focus on innovative ways to deliver quality stadium experience for the value spent by fans, but most importantly capture this value.

Figure 1: Revenue breakdown of 'top five' football leagues.



As Figure 1 suggests, approximately 35 per cent of a club's revenue in England is the matchday earnings. That is larger than the share of sponsorships and on par with the broadcasting rights, so no club would say that increasing shareholder value is the driving force of a club's strategy. Ultimately, success is measured when a club scores goals, win matches and provide entertainment and excitement to the fans. Deloitte's Annual Report of Football Finance (Deloitte, 2019) provides strong data analysis in support of this argument. Ofcom (Ofcom, 2005) reported that overall match attendances have increased over the decade. In the 1999/00 season,

the average attendance of Premier League was 33,899. The 2018/19 average was 38,162 and saw a significant increase in stadium utilization as well. Stadium attendance in the Premier League is relatively higher than other major European Leagues. In the past decade, it has also had the highest average match attendances compared to Italy, Germany, France, and Spain. For the first time in history, the Premier League was overtaken by the German Bundesliga as the most attended league in 2004/5 season and has since experienced a significant upsurge in attendances. The same report shows that few fans regard live broadcast matches as the main reason for not attending the matches in stadiums. The fans' workshop indicated that televised and attended matches offer completely different experiences to Premier League fans, and for many match-going fans, broadcast matches will never be an acceptable substitute.

3. Framework

The following presents a framework to evaluate the factors affecting Premier League attendances, looking at the chosen variables for this research. Assume that home points per game, away points per game, total goals, and team strength have a role in fluctuating the rate of attendance. The framework can indicate what variables might be considered important to understand the causation of matchday attendances. The framework has three components: 1) analysis of total attendance for the 2018/19 season by using statistical modelling tools in R, 2) predictions of outcomes based on these models using ANOVA, and 3) stating the conclusions from the research.

3.1 Models for analyzing attendances

Many methods are available for classification and regression analysis. This study focuses on combining and utilizing various models to answer the research questions according to the pith of each model. The four models used for our analysis are Linear Regression, Logistic Regression, and General Additive Model.

Figure 2: Results from linear regression

	utilisation		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	28.74	25.41 – 32.07	**<0.001**
home_ppg	-1.47	-2.89 – -0.05	**0.043**
away_ppg	2.33	0.77 – 3.89	**0.004**
total_goal_count	0.02	-0.52 – 0.56	0.942
Observations	258		
R^2 / R^2 adjusted	0.050 / 0.039		

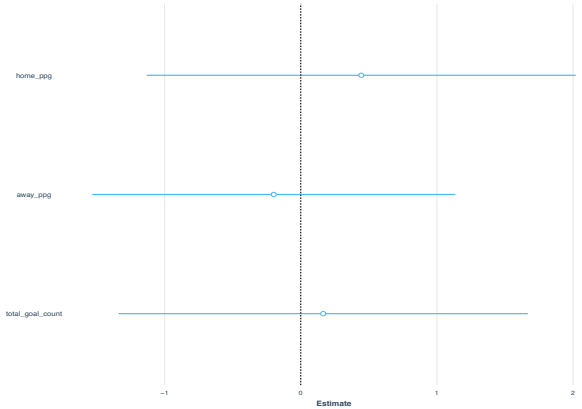
I. Linear Regression

The most basic model in the list is the linear regression model. In this model, independent variables are utilized for the approximation of the linear function to minimize the errors between response variables and predictions of the dependent variables. For this model, stadium utilization is our dependent variable and home points per game, away points per game, and total goal count are the independent variables. The objective of this model is to produce optimal regression coeffects. Due to the coefficients having an infinite number of values, the results are not always optimal to make predictions. That being said, this model still does a great job in helping us to study the relationship between the variables. Obtaining results from figure 2, we have our partial slope coefficients that provide an estimate of the change in utilization for a 1-unit change in the independent variables, holding

every other variable constant. The model has an r – squared of .50 percent, meaning that 50 percent of the variance in utilization is explained by home points per game, away points per game, and total goals scored.

II. Logistic Regression

Figure 3: Results from logistic regression

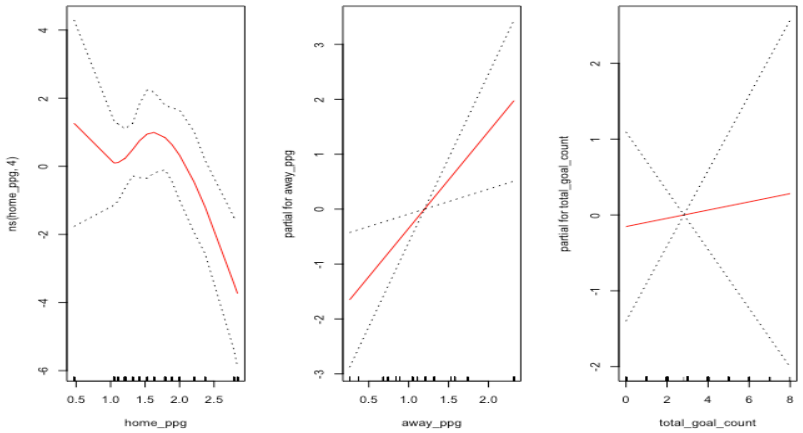


Keep in mind that the coefficients of the independent variables are now in units called logits. From the results obtained from the logistic regression, we see that home points per game and total goals scored influence stadium utilization positively, while away points per game have a slightly negative effect. Surprisingly the coefficients of all three variables are non-significant ($p > 0.05$). So, an increase in utilization by 1 unit increases the odds of home points per game by 0.75 and the total goals scored by 0.1. Whereas, away points per game decrease by 0.37 percent, this is an acceptable interpretation and falls in line with the research. Even though the p-values is non-significant and lower than the linear regression, the history of football falls in line with this model.

III. General Additive Model

The General Additive Model (GAM) is a strong statistical tool which incorporates the best aspects of all models to attain the lowest error rate. This method flexible nonlinearities for our independent variables, while not giving up the additive structure of linear models. For this method, three different GAM models are utilized to regress the independent variables (home points per game, away points per game, and total goals) based on:

Figure 4: results from GAM



a) Independent variables in the Smoothing spline (5 knots for home points per game) and utilization = linear.

b) Away points per game in smoothing spline with 5 knots.

c) Independent variables in a natural spline, with a home point per game (4 knots) and away points per game (5 knots).

3.2 ANOVA

Finally, using analysis of variance (ANOVA), showed which one of the three GAM models were best. Ideally, the model with the lowest F – value is preferred to make predictions, as it represents the lowest rate of error in the model. From the given three models, the second model with smoothing spline and away points per game with 5 knots had the F – value close to zero performed

best. In essence, this model shows that home points per game have a greater and a positive effect on stadium utilization, whereas away points per game are also significant but have a much lower effect on stadium utilization, and total goal count practically has no effect of utilization.

4. Conclusion

Many studies have been done on the causation and analysis of stadium attendances throughout the world, yet not much has been done to use different statistical models apart from the linear regressions. This research helps to further the ideas of using different models to understand the stadium utilization rates with respect to points per game. Moreover, there is evidence to prove that home points per game are the most important factor in-stadium utilization, holding every other factor constant. The methods utilized in the study have the potential to improve the predictive nature of football stadium utilization rates for future projects. Furthermore, due to the reliance on statistical predictions that depend heavily on the number of observations used in a model, that is one area where this research could improve upon. Acquiring data prior to 2005 from Germany and comparing it to England would also be useful to study the reasons behind Bundesliga over taking Premier League as the most attended league even though Premier League is still the leader in every other revenue aspect of football.

Works Cited

Deloitte. (2019). *Annual Report of Football Finances.* London: Deloitte.

Deloitte. (2019, May 1). *Annual Review of Football Finance 2019*. Retrieved from deloitte.com: https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance.html

Ofcom. (2005). *ec.europa.eu*. Retrieved from https://ec.europa.eu/competition/antitrust/cases/dec_docs/38173/38173_104_7.pdf

Data Sets

- https://www.worldfootball.net/attendance/eng-premier-league-2018-2019/1/
- https://www.whoscored.com/Regions/252/Tournaments/2/England-Premier-League