Dissertation Write Up V1

Idris Hedayat

Planning:

Description: Earlier in 2021, a group of major European football club have attempted a coup to establish a new international competition in which they would have participated permanently. The European football association (UEFA) threatened to exclude them from the national championships, had they gone on with their plan. The aim of this project is to estimate the performance of European teams in their respective national championships and then to estimate the results in a hypothetical scenario in which the 12 clubs have been excluded from participation".

structure of report? planning below

Intro: - Bakgroound - Research Quesion - Introduction to statistical methods: A Literature Review - Hierarchical Modelling - introducing r inla packages - report structure

Introduction to data: - Data description: Domestic League Data - Data preperation - Covariates

Exploratory Data Analysis: EDA - - - -

Model Building: Bayesian Hierarchical Modelling - - - -

Model Checking: - Model validation

Discussion: - Conclusion - Limitation - Future Work

Introduction

Background and Motivatoin

In 2021 many of Europe's most prominent and successful Association Football teams engaged in discussions that led to the development of the 'European Super League' (ESL). The Super League was spearheaded by it's 12 founding members from 4 of Europe's top 5 leagues including the biggest names in football worth billions of dollars on their own such as Manchester United, Real Madrid, Barcelona. The competition proposed to create a new international standard of Football with aim of providing a more sustainable and lucrative business model for the associated clubs as a direct competitor to the existing UEFA Champions League. The Super League was conceptualised as a closed competition; implying that the founding members would be guaranteed permanent membership and exemption from relegation regardless of their performance in a season. With 20 teams in total, this would mean only the 12 founding members were guaranteed to enjoy the expected revenues from the competition, while the 8 others faced a constant threat of relegation. Another aim of the league was to maintain a more competitive balance, in the sense that the league would only contain games of the highest quality from Europe's Best. The founders believed this would retain a larger global audience, thus increasing the value of broadcasting rights, sponsorship and merchandise sales. The organisers also proposed the reduction of financial disparities with the rest of European football by proposing to share a proportion of the revenue generated with other clubs and investing in grassroots football development.

However, following the announcement April 18th 2021, the organizers and teams faced vast backlash among the various stakeholders of the football community, including fans, football governing bodies, domestic leagues, smaller clubs, and even governments.

The proposed Super League marked a significant departure from the traditional format of European Football, causing concern from European and Global football's governing bodies; UEFA and FIFA, as well as governments including the UK government, with prime minister at the time Boris Johnson threatening to consider what he labelled as a "legislative bomb" . Critics argued the closed format would lead to the erosion of sporting merit and integrity, many fearing the superior ESL would damage the structure of Europe's domestic leagues, given that the resources of the founding clubs would shift to the new competition, devaluing the domestic leagues as a result and thus widening the financial gap. This concentration of wealth and power was argued to making it more difficult for smaller clubs to compete hence threatening their long term survival. While the reader may not believe this to be of vital concern as other matters, it must be considered that the competition alone was expected to generate Four billion euros , and Europe's top 5 leagues already generating billions of euros alone, notably the Premier League with €5.492 billion. Thus there are billions of euros at stake with the proposition of this new competition, with many subsidiary industries such as sports betting, analytics, and sponsorships that could be affected. So, it is of our interest to investigate what may potentially have occurred as a result of this possible intervention in the world of football, with Juventus, Barcelona, and Real Madrid still yet to withdraw from the competition's establishment, there is still a possibility for the European Super League's fruition into reality. With the 12 founding members leaving their respective leagues, and other clubs such as Paris Saint Germain (Ligue 1, France) potentially leaving theirs, one might consider how this would effect the state of the domestic league should they not participate, who would take crown in each of the top 5 leagues?

Research Question

This project aims to explore and analyse the predicted performance of European teams in their respective domestic leagues by modelling and analyzing a range of factors yet to be discussed. In addition, we will use the resulting methodology and analysis to estimate and discuss the hypothetical outcomes in a scenario where 12 clubs have been excluded from participating in their leagues. This exclusion will have a significant impact on the outcome of the leagues, and in reality would have various potential ripple effects on the teams, competitions, and industries as a whole. By conducting a thorough analysis and exploring the potential outcomes, we hope to provide valuable insights into the current state of European football and the potential impact of excluding certain clubs from participation. In doing this we hope to explore statistical techniques that can be used to model and predict the outcome of match results. We will look at the recent scope of literature on statistical techniques that have been used in the prediction of match results.

Introduction to statistical methods: A Literature Review

A Brief History

The prediction of football match results has been of interest to researchers and statisticians for decades. The use of binomial and negative binomial models were explored as early as 1968 by Reep and Benjamin in their paper "Skill and Chance in Association Football", before the Poisson distribution became the prominent method of modelling these relevant quantities. Many cite and credit Maher (1982) who successfully demonstrated the uses of Poisson Modelling,later addressing the relation between teams in a given match through Bivariate-Poisson modelling. One of the most influential studies in the field came from the work of Dixon and Coles (1997) who proposed a far more effective Poisson Model framework. In their study they introduced further improvements with time-weighting factors and correction terms for low-scoring matches. Their method demonstrated improved accuracy over previous models and has been widely reference in subsequent research.

Bayesian framework:

Bayesian frameworks have gained popularity over recent years due to their ability to incorporate prior knowledge and ability to update probabilities based on newer data, providing promising results with their

predictive capabilities. Extensions of the Bayesian framework include the Bayesian generalised linera model of Rue and Salvesen (2000), which has been widely referred to in subsequent research. The framework was later improved by the promising Hierarchical Bayesian Models of Gianluca Baio in his 2010 and 2018 papers (Baio et al.,2010) (Baio et al., 2018). Baio's contributions have been significant to the field of statistical methods in football match modelling. The 2010 paper introduced a multilevel structure in their modelling to estimate team-specific parameters nested within a mixed effects model, applied in the context of the Italian Serie A league. They also later specified a more complex mixture model that aimed to overcome the issue of over shrinkage produced by the Bayesian multilevel model, in order to provide a better fit to the data of previous season. Baio later made further improvements to this area of the Bayesian framework in his 2018 paper incorporating even more complex and effective mixture models accounting for the data available at the time. His works have been widely cited in later literature, including comprehensive reviews of Bayesian statistical methods in football (Santos-Fernandez et al., 2019) and studies of more modern techniques (Hubáček et al., 2019). The Bayesian Hierarchical Model demonstrates advantages by naturally accounting for relations between variables through the assumption that they come from a common distribution. Overall this approach enables the effective and comprehensive account of team performance while accounting for the hierarchical structure of football league data. More recent research using Bayesian frameworks include that of Razali et al (2017) exploring the use of machine learning methods, in this case Bayesian Networks (BNs) to model, predict, and validate match results for the English premier League. Other recent applications have included Naive Bayes and Tree Augmented Naive Bayes Models in Rahmanet et al. (2018),

Other methods

Although Bayesian methods have become more prominent in the field of predictig football results in recent years, other methods have also been observed. In the past Bradley-Terry models (Bradey and Terry 1952) with comparison modes pairing teams to determine the outcome of a game (Kuk 1995) have been used to estimate the probabiliies of winning, drawing, or losing a match. Related studies in the field include Godin et al's (2014) leveraging of contextual information via "Twitter Microposts" and machine learning techniques in order to comprehensively beat expert and bookmaker predictions, a dynamic approach wit a different end goal to the purpose of our study. Furthermore, other more relevant and modern research cover Machine Learning Methods such as Random Forests (Groll et al. , 2018), Gradient Boosting and Linear Support Vector Machines, notably by Baboota et al. (2018) who's work has been well cited as developments in the use of Artificial Intelligence and Machine Learning methods continue to develop in this field.

As we can observe, the aims of the papers discussed in the literature review vary. Some aim to simply model the outcome of the came like in Fahrmeir and Tutz (1994) using models of paired data with time varying features. Others investigate outcomes by predicting goals scored as seen in Dixon and Coles 1997 and Baio et al. (2010, 2018) while others address other characteristics used to predict outcomes, such as passing movements and shots per game see Reep et al (1968).

Introduction to Bayesian Hierarchical modelling

For the purpose of this study regarding statistical modelling in the prediction of football outcomes, we will be predicting outcomes of the top five European domestic leagues utilising the Bayesian Hierarchical Model (BHM).

(Need to expand on why?) By adding many degrees of hierarchy, BHM enables the representation of complicated relationships and structures within the data. This strategy is based on the Bayesian probability theory discussed above, which offers a methodical manner to revise beliefs or probabilities in light of new information by applying the Bayes' theorem.

The framework used in BHM adopts a

From Bayesian Inference theory we know:

$$P(\theta, y) = P(\theta)P(y \mid \theta)$$

is the joint probability distributions for θ and y, written as the product of the distributions: - the prior distribution $P(\theta)$, which estimates the parameter θ before data is observed - the sampling distribution $P(y \mid \theta)$, referred to as the Likelihood; the distribution of observed data conditional on θ .

Using Bayes Theorem:

$$P(\theta \mid y) = \frac{P(\theta, y)}{P(y)} = \frac{P(y \mid \theta)P(\theta)}{P(y)}$$

Where $P(\theta \mid y)$ represents the posterior distribution of the parameter θ given observed data .$P(y)$ reflects the marginal likelihood, or model evidence, that is derived as the integral of the joint probability distribution of θ and y:

$$P(y) = \int P(\theta)P(y \mid \theta)d\theta$$

In scenarios where the marginal likelihood is not easily obtained, one may express the postierior distribution as:

$$P(\theta \mid y) \propto P(y \mid \theta)P(\theta)$$

Hierarchical Models:

The extension of the Bayesian framework into hierarchical modelling incorporates two added features used in deriving the posterior distribution: - Hyperparameters: set of parameters that are used to determine prior distributions of other parameters uesd in the model often referred to as lower level parameters. The hierarchical nature of BHM allows for multiple levels, each level having its own distribution. Parameters at the higher levels are used to determine properties of the priors for lower-levels, which are the hyperparameters. - Hyperpriors: these are the prior distributions on the hyperparameters, used to express uncertainty in a hyperparameter.

The framework of the BHM: We consider the structure of a two level Bayesian Hierarchical Model let y be a set of obesrvations $y_1, ..., y_n$ from random variables $Y_1, ..., Y_n$ and θ be the set of parameters from each Y_i; $\theta_1, ..., \theta_n$ from a common population with distribution determined by hyperparameter ϕ

The Likelihood above is $P(y \mid \theta, \phi)$ with $P(\theta, \phi)$ as its prior distribution.

$$\text{Stage 1: } y \mid \theta, \phi \sim P(y \mid \theta, \phi)$$

The prior can be expressed as $P(\theta, \phi) = P(\theta \mid \phi)P(\phi)$ using the definition of conditional probability

$$\text{Stage 2: } \theta \mid \phi \sim P(\theta \mid \phi)$$

The next stage being the hyperparameter: ϕ with prior distribution $P(\phi)$, referred to as the hyperprior.

$$\text{Stage 3: } \phi \sim P(\phi)$$

Using this structure we obtain the posterior distribution using bayes theorem, expressed as:

$$P(\phi, \theta \mid y) \propto P(y \mid \theta, \phi)P(\theta, \phi) = P(y \mid \theta)P(\theta \mid \phi)P(\phi)$$

Using this we can obtain probabilities from the posterior distribution.

The R-INLA package:

In surrounding literature regarding BHM theory, various languages are used for modelling including R, Jags, Python, Stan (Hilbe et al., 2017), with popular related studies such as Baio et al.'s paper (2010) addressing the use of WinBugs software. Here they used standard Markov Chain Monte Carlo (MCMC) methods that were used to generate samples from the posterior distribution. However this particular method requires a large number of iterations in order to converge, which computationally is incredibly intensive and time consuming as the number of samples needed increases. Although MCMC methods are able to handle very complex and dynamic models that include a large amount of parameters making it flexible in the modelling process, for the purpose of this particular study it raises questions as to whether it is optimal or necessary.

Instead we will consider the R-INLA package, as used in Baio et al.'s 2018 paper for football prediction modelling. The package refers to the Integrated Nested Laplace Approximation (INLA). This method is far more computationally efficient compared to MCMC techniques used in fitting Bayesian Models, particularly those with latent Gaussian Structures such as Gaussian processes and grouped random effect models. Using a combination of analytic approximations and numerical integration with posterior densities, the obtained posteriors can be use to get posterior expectations and quantiles. Thus, in hope of being able to efficiently and effectively generate many models along with their predictions for the different leagues we will use the INLA package going forward.

Report Structure

In the next chapter of the report we will introduce and inspect the data that will be used to determine the outcome of the domestic leagues. We then go onto introduce main concepts of the model building process, before discussing the results obtained along with any limitations and future outlook on the research.

Introducing the Data

Fbref Datasets

Data of the top 5 leagues are available from a vast array of sources, and we select Fbref because of its comprehensive and well structure format of each league, While providing insightful information on characteristics such as individual player statistics, and complex measures of performance including pass progression types and expected goals, which may be considered for more complex models. Fbref provides the data for each league in the same format, and unlike other sources splits the fixtures for each team into rounds, making it very useful when predicting and updating posterior probabilities for each round as the leagues progress in time. Fbref gives access to multiple seasons in time for each league, which also gives rise to the possibility of using these past seasons in the modelling process.

Data Description

The initial Fbref fixture dataframes include fixture lists of 380 games for leagues with 20 teams; Premier Leage, La liga, Ligue 1, and Serie A, while Bundesliga has 18 teams and thus 306 games in a season. They provide information on "Wk"; the round that a particular fixture belongs to, Date, Home Team, Away Team, Score, Venue, and other variables that were not included in the final dataframe, namely; xG, Attendance, Referee, as these would not be able to be reproduced every round of prediction.

Data preperation

We then clean the data handling any missing values and correcting any inconsistencies, particularly in the case of Premier league data. There are 38 rounds of fixtures for each team in all leagues apart from Bundesliga

where there are 34. However in some cases there are rounds that are rescheduled thus creating double game weeks and other conflicts, and so in the case of the Premier League the number of rounds has been adjusted so there are 44 rounds to avoid any conflicts making predictions more effective in the long run. We do this and other data preparation steps using the tidyverse set of packages in r, particularly dplyr and tibble. We first order the data in order of data, adding ID_game indicating the number of the specific fixture out of the 380 that are to be played. The input data is converted into long format by duplicating the rows of each fixture for a home and away team row, adding a binary variable Home, 1 if the team in the row of a given fixture is Home and 0 if away. The Opponent variable was created grouping data by ID_game and assigning each team's opponent as necessary. Venue variable is adjustedinstead of being the nstadium name, to the name of the team it belongs to; "Old Trafford" becomes "Manchester Utd". We then create the following variables that are to be used in the Bayesian Hierachical modelling processing:

Variable	Description
Goal	Number of goals scored by the specific team in a given fixture (this will be our response)
Points Won	Points Gained after each fixture, 3 for Win, 1 for Draw, 0 for Loss
Days Since Last Game	Number of days since the last fixture played by the specific team
Total Points	Cumulative points acquired by each team after accounting for the points won in the specific game
Points Difference	Difference in points between the 2 teams for each fixture
Relative Strength	Weighted value of total points between 2 teams of each fixture
Form	proportion of points won in the last 5 games, i.e x points out of the possible 15
Goals Conceded (GC)	Amount of goals the opponent scored against a specific team
Goal Difference (GD)	Difference between the goals scored and concede in a given game
Total Goals Scored (TG)	Cumulative total of goals scored in all games played after given game
Total Goal Difference (TGD)	Total Goal Difference (Total Goals Scored - Total Goals Conceded)
Goal Difference-Difference (TGDDiff)	Difference in the totalgoal differnece between the teams of a fixture
Rank	The league position of a team at any given game week, (1 to 20), decided by goal difference if tied
Rank Difference	Difference in league position of teams for each fixture

Exploratory Data Analysis

Model Building

References:

https://www.theguardian.com/football/2021/apr/20/uk-government-may-legislate-to-stop-european-super-league-says-minister

[https://www.statista.com/statistics/1230111/european-super-league-sponsorship-revenue/#:~:text=Early%20reports%20su

Maher, M. J. (1982). Modelling association football scores. Statistica Neerlandica, 36(3), 109-118.

Rue, H., & Salvesen, Ø. (2000). Prediction and Retrospective Analysis of Soccer Matches in a League. Journal of the Royal Statistical Society. Series D (The Statistician), 49(3), 399–418. http://www.jstor.org/stable/2681065

Godin, F., Zuallaert, J., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2014, June). Beating the bookmakers: leveraging statistics and Twitter microposts for predicting soccer results. In KDD Workshop on large-scale sports analytics (pp. 2-14). New York, NY, USA: ACM.

Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning, 108, 29-47.

Santos-Fernandez, E., Wu, P., & Mengersen, K. L. (2019). Bayesian statistics meets sports: a comprehensive review. Journal of Quantitative Analysis in Sports, 15(4), 289-312.

Schauberger, G., & Groll, A. (2018). Predicting matches in international football tournaments with random forests. Statistical Modelling, 18(5-6), 460-

Hilbe, J. M., De Souza, R. S., & Ishida, E. E. (2017). Bayesian models for astrophysical data: using R, JAGS, Python, and Stan. Cambridge University Press.