LITERATURE REVIEW

Use of Data Science and its sub domains like Machine Learning, Predictive Analytics, and Artificial Intelligence is evident in each field. Some fields may have started using data science more recently as compared to other fields. One such application of data science can be found in sports analytics. Although statistics was used earlier to explain the performance of a team or individual, using predictive analytics is something that has recently came up. This project aims at collaboratively studying and presenting the results before a soccer game starts and tries to explain the same post-match. One way to approach this problem is to quantify the value of the goal scoring chance created by a team or quantifying the value of a shot while given the scoring opportunity (Eggels, van Elk & Pechenizkiy, 2016). Another way to formulate a solution for this proble is to do a comparative study and determine the best algorithm available. Baboota, Kaur (2006) did a similar comparative study where they compared various algorithms like Gaussian Naïve Bayes, Support Vector Machine, Random Forest, and Gradient Boost.

Oberstone (2009) organized the football actions into 5 broad categories; Attempts, Passing, Defending, Crossing and discipline and differentiated the teams based on the same. There were three separate classes for the teams: the top four, the bottom four and the rest 12 middle teams. He used Analysis of Variance (ANOVA) and multiple regression model to analyze the data.

However, Croucher used Poisson distribution and multiple regression and used only the goals scored by the teams to test whether any team finished significantly higher or lower. This paper also had a separate section dedicated to the drawn matches and used conditional probability and expected Poisson distribution. Similarly, Brillinger (2006) also used Poisson distribution as a measure to calculate the home and away team’s effect on the match in which he used the data of Norwegian Football League.

Based on the research done by various papers, we plan to take some concepts from these papers, and we will be doing our own research. We will be proving that the home advantage still is an important factor in Soccer and will use regression models and hypothesis testing to analyze the data.

We will be using various factors to determine the difference between what the table should have looked like and what the table finally was at the end of the season. This includes ratio of shots to shots on targets, number of goals and the number of fouls committed by the team. At the end, we also will determine whether playing at home is an added advantage to the team.

From this analysis, we would be implementing a detailed analysis that would contain regression models and descriptive statistics and would be creating visualizations for the same. Our predictions are as follows:

* Playing in the home stadium provides a substantial advantage while playing a soccer match
* The Premier League table of 2020-21 season should have been different from the final table at the end of the season.

APA Citation:

Oberstone, Joel. *Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success*. 2009, <https://core.ac.uk/download/pdf/216977938.pdf>.

Croucher J.S. (2004). *Using Statistics to Predict Scores in English Premier League Soccer*. 2004, In: Butenko S., Gil-Lafuente J., Pardalos P.M. (eds) Economics, Management and Optimization in Sports. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-540-24734-0_4>.

Brillinger, David. (2006). *Modelling some Norwegian soccer data*. 10.1142/9789812708298\_0001.

Eggels, H., Elk, R. van, & Pechenizkiy, M. (2019, May 13). *Explaining soccer match outcomes with Goal Scoring Opportunities Predictive Analytics*. Eindhoven University of Technology research portal. Retrieved October 25, 2021, from <https://research.tue.nl/en/publications/explaining-soccer-match-outcomes-with-goal-scoring-opportunities->.

Baboota, R., & Kaur, H. (2018, March 28). *Predictive analysis and modelling football results using Machine Learning Approach for English Premier League*. International Journal of Forecasting. Retrieved October 25, 2021, from https://www.sciencedirect.com/science/article/pii/S0169207018300116.