Comparison of Predictive Models on Soccer Matches

Karanmeet Khatra, University of British Columbia – Okanagan

ABSTRACT

Soccer is the most popular sport in the world. It is played and watched globally, and billions of dollars are spent betting on the results of these matches. In this paper, I dwell into previous methods that researchers have used to try and predict the outcome of these matches as well as designing my own predictive model using a unique dataset containing predictive parameters. This is done using the probabilities of a team winning and comparing that to the importance of the game to each team. As well as using three different projected score parameters to answer which one does the best at matching the correct result.

INTRODUCTION

The sports betting industry is worth over 3 trillion dollars, with an estimated 110 billion spent on global soccer betting (AFootballMatch 2020). Not only is it fun to watch the matches, but also having real money on the line takes it to another level. Imagine spending $5 to potentially take-home thousands of dollars if every one of your bets is correct.

The common phrase “the house never loses” is one that has proven to be correct. Sports bettors have been trying to beat these online betting websites for years and although some have been successful in doing so the vast majority are unsuccessful. In order to beat the “house”, predictive models are used to determine the result of the matches using numerous amounts of factors. Some researchers have investigated the idea of using running performance

(intensive runs, total distance covered) and the performance of players on the pitch and how that affects the win probability for a team (Soebbing *et al*. 2020, Al-Mulla *et al.* 2020). Others have used team rankings, statistical models and machine learning models to predict how a game will end (Andreou 2020, Pappalardo *et al*. 2020).

In this dataset, there are multiple predictive models given to us including: the probability for either team winning and the probability of a tie occurring, how important the game is for each team, projected scores of both team (proj_score), an estimate of how many goals a team should have scored given the number of shots they took (xg), and an estimate of how many goals a team should have scored based on actions taken around the opposing goal (nsxg). My interest in this dataset lies around how well each of these do in predicting the results of the match. This is broken into two sub questions: which of the two, probabilities and importance, do a better job at predicting the actual outcome of the game? And how accurate are the projected score (proj_score), as well as xg and nsxg at predicting the actual score? Both of these questions have been explored in some areas of prior research such as predicting goals scored to determine the outcome of UEFA champions league matches (Andreou 2020), and using a team's position in a competition's final ranking to relate it to their performance (Yi *et al.* 2020)

MATERIALS AND METHODS

The dataset initially contained over 10000 records from matches played in 2019 and played/ still to play in 2020. As I am interested in figuring out how good the predictions did in predicting the results of the match, any columns containing empty values were dropped, and any match yet to play was also dropped. After doing so just over 2500 rows were left.

Relationship between Variables

I conducted a deeper analysis into the variables using both the data frames describe function as well as an overall look using the pandas profiling function. Following this, relationships between variables were observed. The first plot I looked at was a correlation matrix. The correlation matrix showed correlation coefficients between variables. Any positive value in a cell indicated two positively correlated variables and vice versa with a negative value. The values in the cells can only be between -1 and 1, and the further away the correlation coefficient is from zero, the stronger the relationship between the two variables. Furthermore, two pair plots were constructed to take a deeper look into any relationship that existed in the correlation matrix.

Analysis on Research Questions

Lastly, I conducted the analysis on the research questions. For the first research question, I needed to look at how well each parameter (probability and importance), did at predicting the actual outcome of the game. For the probability parameter, if the probability for a team to win was higher than the other, I would say that team is predicted to win, and if the probability for a tie was greater than 0.3, I would say the outcome of the match would be a tie. Consequently, for the importance parameter, if one team had a higher importance percentage than the other team, that team would be projected to win and if the importance percentage was the same, then a tie would be projected. Afterwards, I took a deeper look into this comparing them to the actual result of the match, with different plots. For the second research question, I needed to compare the projected score parameters (projected score, xg, and nsxg) to the actual score of

the match. The first thing I did was construct hex bins between each of these parameters and the actual score to observe how correlated they are, and then afterwards make count plots to indicate whether a parameter correctly predicted the correct score. Note that a prediction was classified as correct if it is +- 0.5 from the actual score. For example, if the prediction for team 1 was 1.5 and team 1 scored 1, the prediction would be correct.

RESULTS

Relationships between Variables

The correlation matrix was the first step in observing any type of relationship between variables (Figure 1). I was mainly interested in relationships for the probability, projected scores, importance, xg, and nsxg parameters. For the probability parameters, they have a positive correlation with the actual score (score1/2), projected score (proj_score1/2) of their team and a negative with the opposing. The same correlation was also displayed between the importance parameters. And then for each of the projected score parameters, a positive correlation was seen amongst one another, as well as a positive correlation between each of these parameters and the actual score for their team in the match (negative correlation with the opposing team).

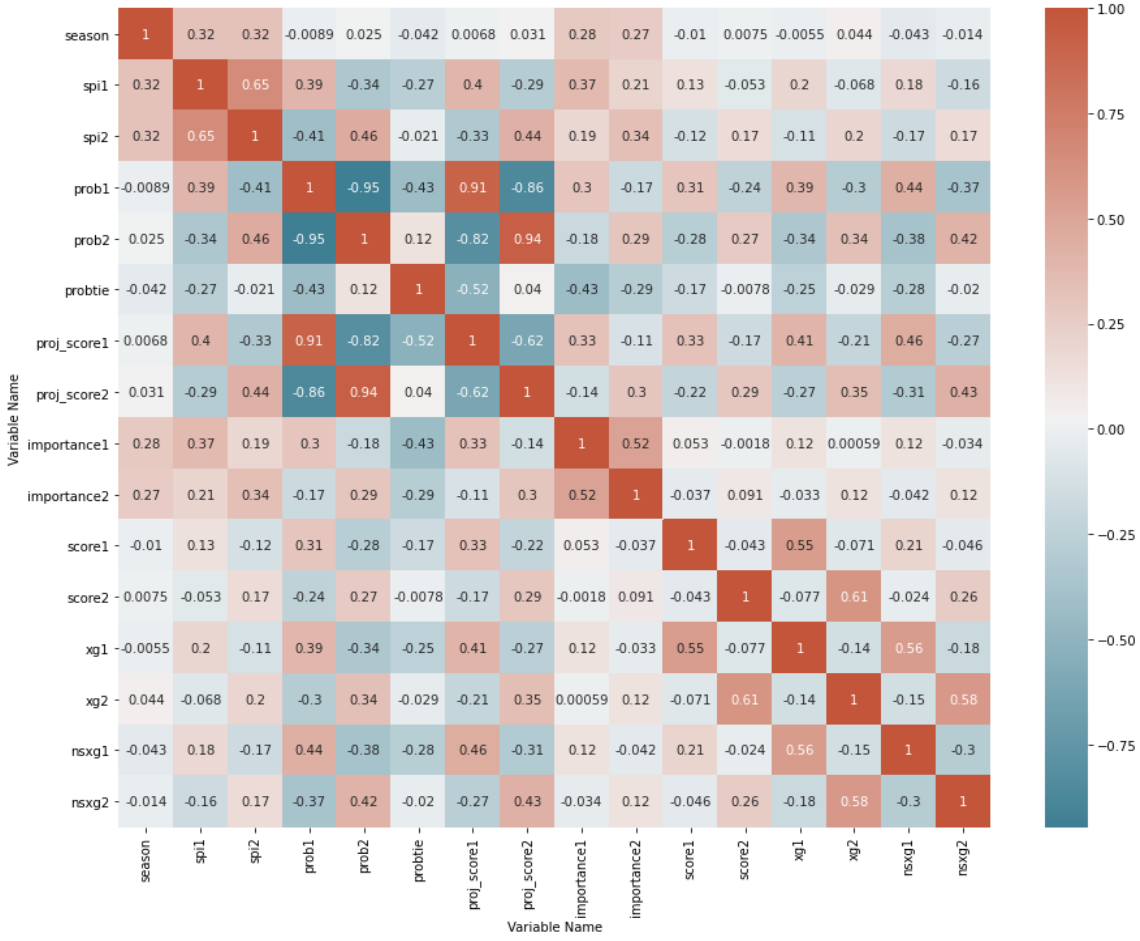


Figure 1 Correlation Matrix showing the correlation coefficients between different variables

The pair plots showed us more of the same (Figure 2). Observing this pair plot, we can see the relationship between importance and some of the other parameters. There is a small positive correlation between each team's importance and their projected score and a negative with their opposing teams. Also, worth noting that there isn't a direct correlation between the importance parameters and the actual scores of the game.

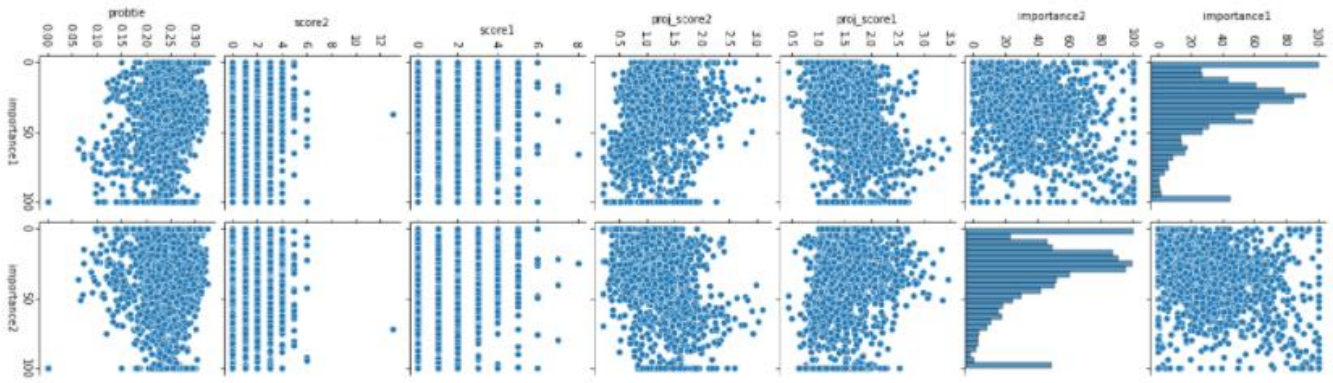


Figure 2 Correlation between importance and other parameters using a pair plot

Analysis on Research Questions

Research Question 1: which of the two, probabilities and importance, do a better job at predicting the actual outcome of the game?

A scatterplot was constructed to observe the relationship between the probability of each team winning and their respective score as well as the importance of each team and their respective score (Figure 3). For Figure 3 (a,b) we can see as the probability for a team increases, so does their goals scored. In addition, for Figure 3 (c,d) they display the same conclusion with the importance parameter.

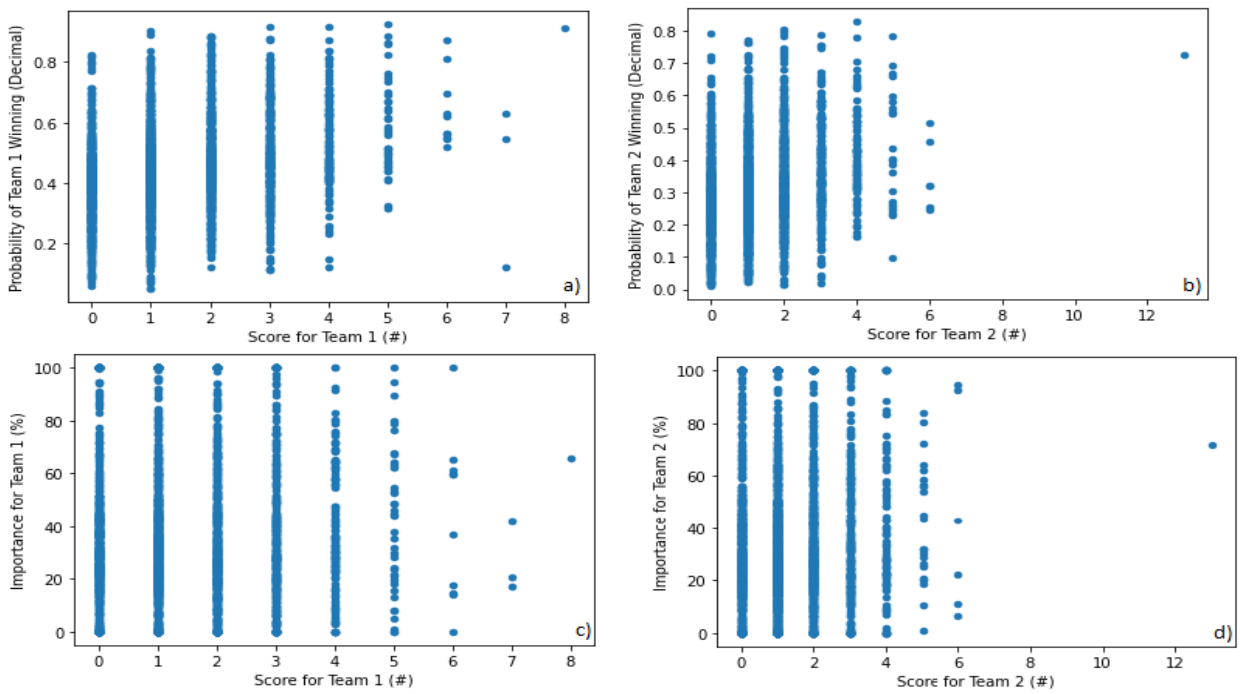
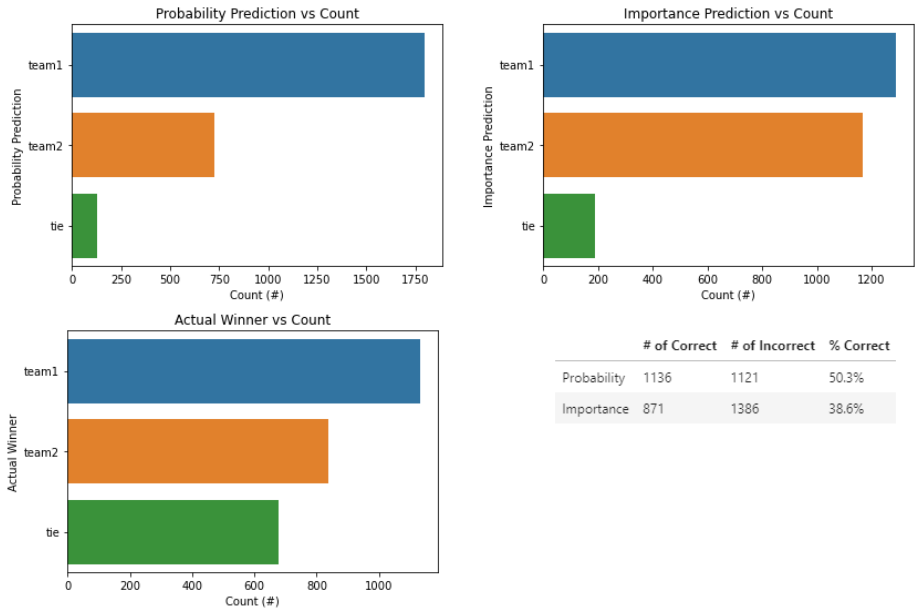


Figure 3 Scatterplots showing the relation between: a) Probability for Team 1 and Score 1 b) Probability for Team 2 and Score 2 c) Importance for Team 1 and Score 1 d) Importance for Team 2 and Score 2

Following, I plotted both the probability and importance parameter with how many times they predicted team 1/2 to win and how many times a tie was predicted. This was then compared to the actual result of the matches (Figure 4). From this I can see that the number of ties that were predicted were much fewer than the actual number of ties that happened. In the same way, the number of times that team 2 was projected to win was way higher in the importance parameter than the probability parameter. These two plots were then transformed into two tables which conveyed the percentage of times team1/2 and a tie was predicted and how many times the parameter was correct (Figure 4). We can observe from this table that the probability parameter was correct 50.3% of the time and importance 38.6%.



	# of Correct	# of Incorrect	% Correct
Probability	1136	1121	50.3%
Importance	871	1386	38.6%

Figure 4 Comparison of the probability/importance predictions to the actual result

Research Question 2: how accurate are the projected score (proj_score), as well as xg and nsxg at predicting the actual score?

The hex bins for each parameter and their respective score predicted were plotted first (Figure 5). All the hex bin plots convey the same thing with each parameter being heavily correlated and most goals falling around 0-2.

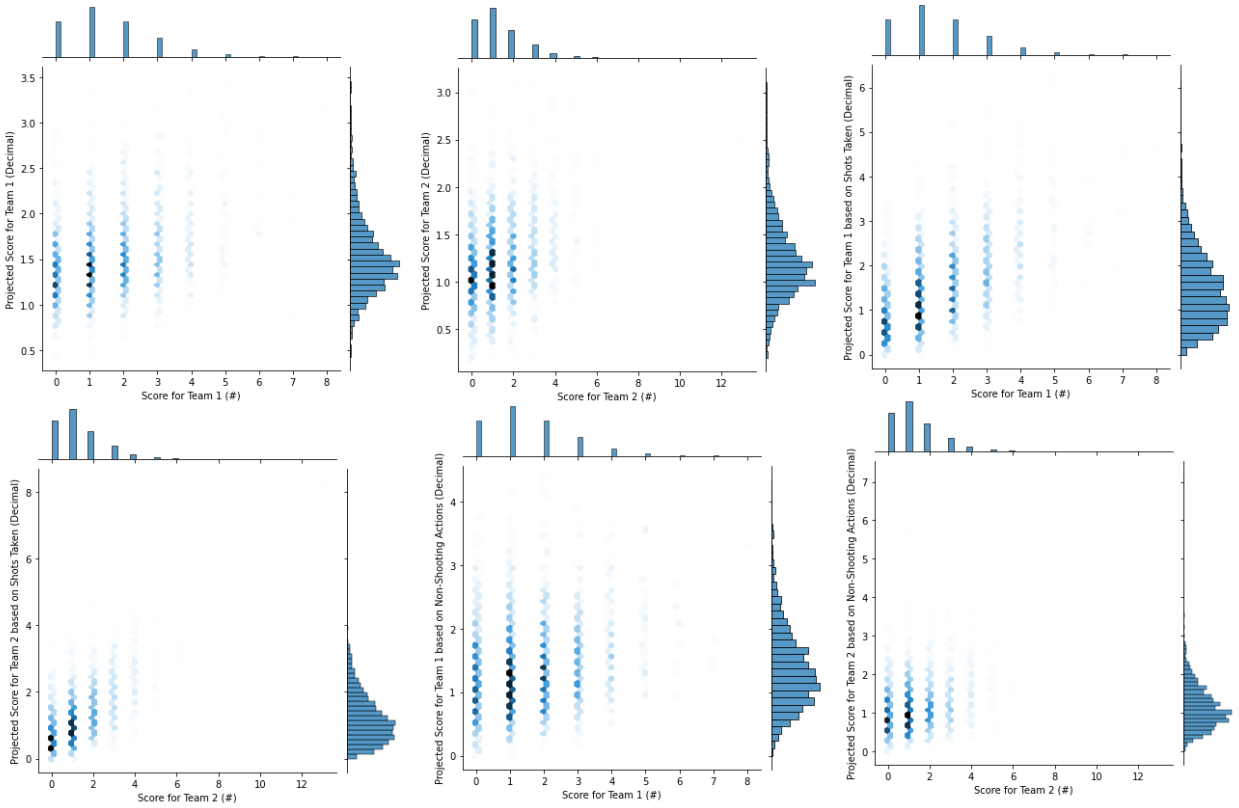


Figure 5 Hex bins for each parameter and their respective score

As was the case with the first research question a count plot for each parameter was made, and if their prediction for that teams score was within +- 0.5 of the actual score the prediction was correct (Figure 6). Table 1 conveys all this data put together and gives the percentage of times each is correct. It can be seen from this table that both xg1/2 do the best job at predicting the actual score.

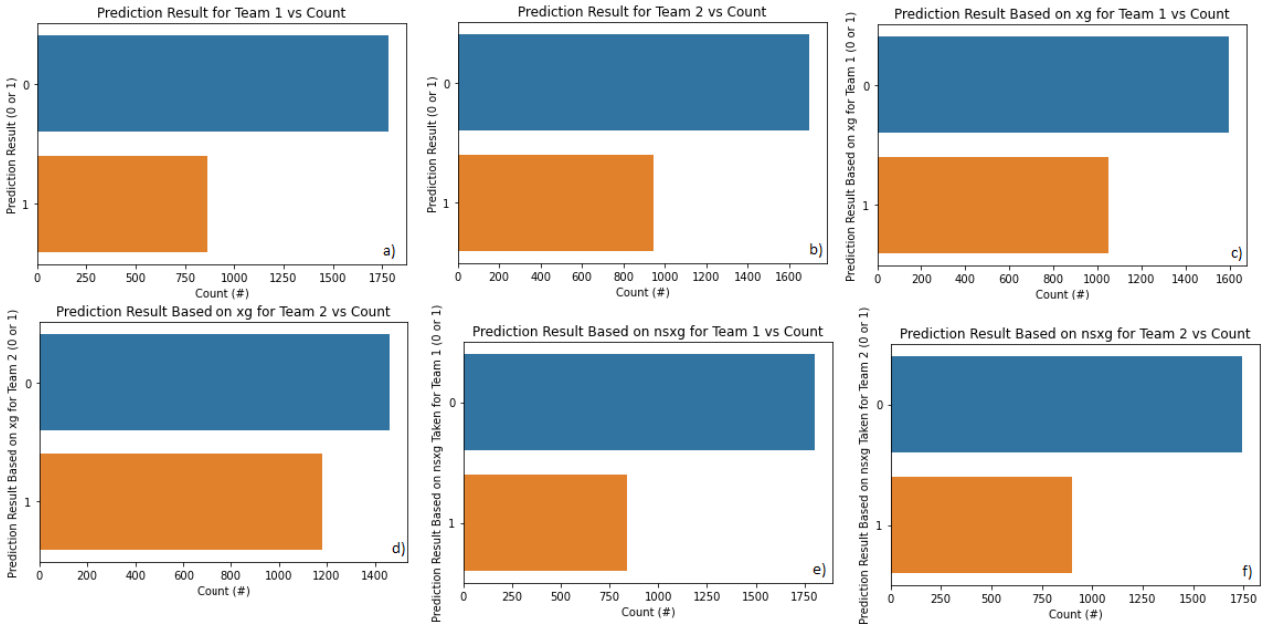


Figure 6 Count plot displaying how many times each parameter was correct (0 - incorrect, 1 - correct) a) Projected Score 1 and Count b) Projected Score 2 and Count c) xg1 and Count d) xg2 and Count e) nsxg1 and Count f) nsxg2 and Count

	# of Correct	# of Incorrect	% of Correct
proj_score1	738	1354	32.7%
proj_score2	807	1450	35.8%
xg1	903	1354	40.0%
xg2	1007	1250	44.6%
nsxg1	706	1551	31.3%
nsxg2	775	1482	34.3%

Table 1 Displaying the prediction results for each parameter

DISCUSSION

The main question I had about this dataset was how well each parameter does in predicting the correct results of the match. From both sub questions we can see that there did exist correlations between the predictive parameters and the actual results. Although I can clearly identify which ones do the best (probability and xg), neither do a very good job, and are correct

less than 50% of the time. Having said that, as it relates to sports betting, being right 50% of the time is still very hard to do. There are so many ways a game can end and thus almost impossible to get right every time. I believe the big influencer of our predictive models being less successful was in fact how many ties resulted in a match. Prior research had even shown that it was difficult to predict ties using a machine learning approach (Yi *et al.* 2020). Both of our parameters (probability and importance) predicted less than 200 ties whereas the actual number was around 600. As a whole, the way I conducted my research proved to answer both my questions in determining which parameter was best; however, there are ways that this research could be expanded to in the future to potentially generate better results.

Future Work

The first way our methods of analysis could be changed would be the way we predicted a tie for each of the probability and importance parameter. A tie was predicted if the probability for a tie was greater than 0.3. It would be interesting to see how the results would change if this probability was lowered. On the other hand, a tie was predicted for the importance parameter if the importance percentages were the same. One thing to change here would add a threshold and if the importance percentages fell within them a tie would result.

Secondly, for the score parameters instead of trying to predict each individual score correctly, the total score could be predicted. This would mean combining the results of proj_score1/2, xg1/2, and nsxg1/2 and seeing which combination does best at predicting the actual total score. This would eliminate half of the parameters that need to be correct.

Finally, I could combine different parameters to predict the result. Such as using the probability and importance parameter and saying if the probability for a team winning is greater than 0.5 and the importance percentage for that team is greater than 70%, we say that team is predicted to win. This would then combine both predictive models together and thus should produce a better result.

CONCLUSION

To conclude, I was able to confidently say which parameters were better at predicting the results of the matches. However, the prediction parameters didn't exhibit high rates of success. This could be due to the wide array of outcomes in a soccer match or just the simple fact it is extremely difficult to predict anything in the real world. It would be interesting to take these predictive models to the real world and test on live betting sites to view the success rates and observe how often the "house" is beaten. Altogether, it was a great experience conducting a full-on analysis with a dataset I was interested in and using different methods to complete this.

REFERENCES

1. Soebbing, B. P., Wicker, P., Weimar, D., & Orlowski, J. (2020). How do Bookmakers Interpret Running Performance of Teams in Previous Games? Evidence from the Football Bundesliga. Journal of Sports Economics. https://doi.org/10.1177/1527002520975827
2. Wheatcroft, E. (2020). Profiting from overreaction in soccer betting odds, Journal of Quantitative Analysis in Sports, 16(3), 193-209. doi: https://doi.org/10.1515/jqas-2019-0009
3. Al-Mulla, J. M., & Alam, T. (2020). Machine learning models reveal key performance metrics of football players to win matches in Qatar Stars League. IEEE Access, 1-1. doi:10.1109/access.2020.3038601

4. Andreou, Zacharias. (2020) Who Will Be Crowned King of Europe? A Predictive Model for the Uefa Champions League. Senior Independent Study Theses. Paper 9034. https://openworks.wooster.edu/independentstudy/9034
5. Pappalardo, L., & Cintia, P. (2018). Quantifying the Relation Between Performance and Success In Soccer. Advances in Complex Systems, 21(03n04), 1750014. doi:10.1142/s021952591750014x
6. Yi, J. H., & Lee, S. W. (2020). Prediction of English Premier League Game Using an Ensemble Technique, 9(5), 161–168. https://doi.org/10.3745/KTSDE.2020.9.5.161
7. AFootballReport. (n.d.). How much Money is being bet on Sports every Year? Retrieved December 05, 2020, from https://afootballreport.com/blog/how-much-money-is-being-bet-on-sports-every-year