Beating the Bettors – English Premier League

Jean-Marc Ruffalo-Burgat
Computer Science
Colorado State University
frenchy9@rams.colostate.edu

1. Introduction

Soccer, one of the world oldest sports, the origins of which can be traced back as far as 250 BC to the Chinese game of Tsu Chu ("Kicking Ball"), a game in which a ball was to be placed in a net and the use of hands was not permitted [1].

The game as we know it, was created in Britain along with the Football Association, which would oversee and manage all rules for the sport. This new popular sport took over the school grounds and pockets of many citizens. Yet at this point gambling on sporting events was heavily regulated and therefore not very popular. Two major events would see more than 50% of UK residents gamble at least once a week, 41% of which placing bets several times a week [2].

These events were: The Betting and Gaming Act of 1960 [3] and the formation of the Premier League in 1992, the highest level of Professional soccer in Britain and the most watched league in all the world [4].

This pairing of lessened betting restrictions and unified national soccer league led to a rise in gamblers and an increased demand in data that would assist in making informed betting decisions. However, to make more informed betting decisions on soccer, more detailed data was needed. At this point the only stats kept on games was the final score, and in some cases: who played and who scored.

This was miles behind the stats revolution taking place across the pond in the United States, spearheaded by Baseball. A sport which saw its first detailed stat (the box score) printed in a newspaper in 1859 [5]. The disparity between data collection in Baseball and data collection in soccer described well the difference in nature of the two sports. Baseball unlike soccer is very static, one player throws the ball while another attempts to strike it. There is a limited number of potential outcomes that would need to be recorded to accurately represent a game.

The play in baseball is analogous to a set piece in soccer, where the play stops, and teams can attempt a pre-planned event of play. However, set pieces only make up a small portion of play, while the rest is a continuous run of play that gives each game and play a unique feel. While the intention in soccer is to put the ball in the back of the net, no two goals are ever really the same. The distance and angle from goal, the number of players between the shooter and the goal, the position of the keeper, the speed and position of the ball under the shooter's feet (or air), all these factors contribute to a unique shot every time.

All those factors impact a shot, but a shot is only one piece of data that could be collected, the same factors are in play for every pass, the length of the pass, the height of the ball, the amount of pressure on the player. Each team will record hundreds of passes in a game. To compound on this issue, the majority of play for each player is without the ball, how can we record detailed stats for what players do without the ball?

Therein lies the issue in betting on Soccer, it is a perfectly imperfect game in which one team can dominate nearly all statistical metrics kept on a game, yet lose in the only one that matters, total goals scored. This is the problem I intend to address; can I create a soccer prediction algorithm to predict the result of a soccer match? The success of which will be tested against 3 key metrics.

To accomplish this, I will create two sperate models, one which will generate weights for each statistical point kept on a game, the values of each of these statistical points in conjunction with their weights will output a likelihood that a team will win, draw, or lose. The second model will be used to generate accurate statistical points for each future game based on past performances, to be plugged into the weights.

This will output a percentage tied to each possible outcome for a match (win, draw loss) which will be used to choose a winner. A successful prediction being one in which the model predicts the same result as which occurred in real life. The accuracy of the model will be compared on the following metrics.

1. Random prediction (33%)
2. Always choosing the home team (44%)
3. Make a profit VS. Betting odds

The thirds metric is the most important for this research paper, as it is a way of comparing how my model performs to others created by the betting industry. Additionally, it provides a better description of a model's performance than just prediction rate, a problem which will be explored further.

2. Literature Review

One of the first people to ever start collecting in depth soccer stats was Charles Reep, who would write the passes and shots down in a notebook by hand by simply watching the match in the stands

[6]. The result of his data collection led him to believe that long sequences of passing and possession did not heighten a team's chance at scoring, and that a team should simply boot the ball across the field to their strikers. An action that could occur many more times that a long string of passes and would theoretically achieve the same result. This thought process influenced the play styles of many teams of that time, and now with further detailed data has been thoroughly disproven. As keeping possession of the ball not only provides a dangerous attack but is the best defense. This is a perfect example of the danger of drawing conclusion by relying on limited stats, a common occurrence in the birth of soccer stats, and a dangerous outcome even today.

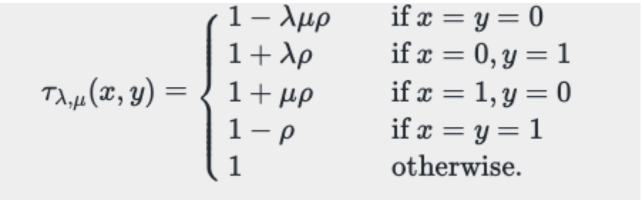
This lack of in-depth stats for each game was reflected in the limited nature of published works on predicting outcomes of soccer matches.

Two works, one by Moroney [7] and the other by Reep and Benjamin [8] both used a negative binomial distribution, however while both were successful in predicting the overall success rate of teams across an entire season, it was unable to do so successfully for individual games. The conclusion being that while skill triumphed across an entire season, individual games were too biased by chance.

This problem was rectified by using two independent Poisson distributions, a model demonstrated by Maher [9]. Now models could accurately (to a degree) predict the result of individual games. This model would also automatically be adjusted based on a team's recent performance.

This model would create new ills as it dispelled old, limitations of this model are two-fold, first it only would output the anticipated result and not a confidence interval of the result. Additionally, the usage of a Poisson distribution is faulty in that soccer matches tend to be low scoring affairs, with many games having fewer than 2 goals total.

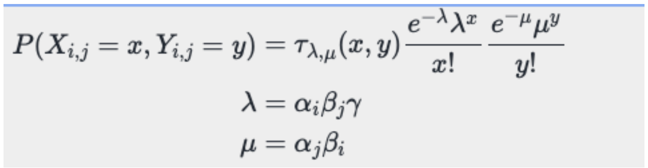
A groundbreaking new model based on Maher's earlier work was published by Dixon and Coles [10]. Using the same independent Poisson distributions this model was unique in its ability to also account for: home field advantage, a team's proclivity to score as well as not to concede, as well as recent performance relative to opposition strength (meaning a recent string of wins against poorer opposition is not worth as much as a string of wins vs the best opposition). Additionally, it solved the problem of low scoring matches by modifying the probability distribution with the below parameter (X – home team goals, Y – away team goals)



$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1-\lambda\mu\rho & \text{if } x=y=0 \\ 1+\lambda\rho & \text{if } x=0, y=1 \\ 1+\mu\rho & \text{if } x=1, y=0 \\ 1-\rho & \text{if } x=y=1 \\ 1 & \text{otherwise.} \end{cases}$$

Figure 1. Low Scoring Games [11]

The full probability distribution including this modification for low scoring games, where αi is the attacking strength, βi is the defensive strength and γ is the implicit home team advantage.



$$P(X_{i,j}=x, Y_{i,j}=y) = \tau_{\lambda,\mu}(x,y)\frac{e^{-\lambda}\lambda^{x}}{x!}\frac{e^{-\mu}\mu^{y}}{y!}$$
$$\lambda = \alpha_i\beta_j\gamma$$
$$\mu = \alpha_j\beta_i$$

Figure 2. Poisson Probability Distribution [11]

The only thing this model was missing was more detailed data, and at the same time as this paper was published mere miles away was a new startup that would do just that. In 1996 a company called Opta Sports began collecting the most in depth and accurate data on the Premier League, and in a few years' time they would expand to collect detailed soccer data on every major league across the globe [12].

Opta's strategy was simple: sit 3 people down in a room with a computer displaying the game to collect data on, and a live stream of the match. Then for 90+ minutes every pass, shot, goal, and auxiliar action would be tracked and logged. For tens of thousands of games every year. This level of detailed data allows the expansion of previous models to better predict the outcomes of games. One real world example is the company AskBettor which use a modified version of the Dixon Coles model alongside a neural network and have achieved an accuracy rate of over 75% [13].

This does bring up one of the two issues with current prediction models, this first being that of accuracy rate. Accuracy rate means very little when it is not given with auxiliary information. What is this model betting on? Match results, half-time results, total goals, both teams to score? While betting on match results has a random hit rate of 33%, both teams to score is a much higher 50%.

A model which does not predict on every match but rather a subset of all possible matches leads to a dishonesty in accuracy rating. One could create a simple model that predicts the teams: Liverpool, Manchester City, Paris Saint Germain, and Bayern Munich to always win at home. In the past 3 years this simple model would have an accuracy rate of 75%.

Therefore, for a true representation of the success of a model, the games the model predicts on must be known, or a comparison must be made, for example against betting odds. The previous model mentioned has an impressive accuracy rate of 75% but if 1 unit was bet on the result predicted by the model, it would have a negative return on investment, as the above-mentioned teams are the some of the best teams in their league, and are expected to win a home match, which is reflected in their betting odds.

This is an important factor to mention when creating a model to predict on soccer matches, as otherwise it would have been a problem solved long ago, a successful model is not one with just a high accuracy percentage on a subset of games, but one with a high accuracy rate on any game, which should in turn result in a positive return on investment if 1 unit was bet on each game.

The second issue (not seen in models using the original Dixon-Coles Model) is seen in models that train on game data, but then also test on game data, one such example can be seen in the paper by Ulmer and Fernandez [14]. In which a high accuracy rate is produced but the model is predicting on information that would be impossible to know beforehand.

This key fact is something that must be taken into consideration, if a model tested on game data produces an accuracy rating of 60% this is not the *true* accuracy rating of the model, but rather the ceiling, thus the author of said model must then attempt to accurately predict each statistical point of the match to be predicted on, such that the trained model can accurately perform.

The values of the predicted statistical points must only be generated using data known before a game is played. This constraint limits the future predictive power of the model as its prediction accuracy deceases the further in the future it must predict. This, however, is a non-issue and is acceptable as predictions should only be made on the next sequential unknown game, as predicting the result of a games more than one week in the future adds no benefit as only decreases the accuracy,

The last question that is not answered in previous models, is the intricate components that make up a team and its performance. Newer models are based on in-depth statistics, but they are statistics of the team as a whole, while a team is made up of 11 players. Players who vary in skill, and importance for a team. Whose inclusion or exclusion can completely alter the result of match. Any of the 11 players may be tired, injured or even suspended based on actions from previous games. A team may be on a hot streak and dominating all opponents, but if a star striker is injured that team's likelihood of continuing that winning streak should be decreased to reflect their decrease in attacking power, however a model trained based on a only on s team would fail to take this into account.

Additional qualitative factors that are hard to quantify can also affect match outcomes: weather, manager (hired or fired), travel distance, and luck. The effect of these factors is debatable on a match's outcome, but this could be the same fallacy that Charles Reep found himself in, and until new metrics can be created to take these factors into accounts, models may be performing sub optimally.

One new controversial metric that aims to consider a team's "luck" in a performance is Expected Goals (XG), a model which quantifies what the expected chance of a goal going in was, based on the shooter's position, the ball height, defenders' positioning, strike location (left foot, right foot, head) etc. The possibilities of this new metric could allow models to not diminish a team on a poor run of form if their expected goals would say they should have won those games, or to not rank to highly a team who appears to be outperforming their actual skill level.

The usage for this model is still contentious, as there is no consensus if a team who underperforms their XG should be considered an unlucky team and therefore likely to perform better than previous results dictate or if they are underperforming their expected goals simply because they are a poor team.

The approaches and metrics used to create prediction algorithms is still in its infancy, and models will continue to improve in performance as computers are able to train faster on large amounts of data, and new metrics to improve models are created and peer reviewed.

3. Methodology

As the founder of The Match Predictor, I have access to a huge database of soccer results spanning 100+ years, 2 million+ games, and 100 million+ in-game match events which are usable for training an algorithm to predict match results. As mentioned in in the Dixon-Coles model, past performance become less and less indicative of future results as time progresses.

Therefore, for my dataset I will only consider games from the past three years. An additional constraint will be to only train and test on games from the Premier League. The benefits of this is the model which is trained on statistical points of games allows it to fit to the unique play style of one league, as the playstyles change and skew statistics from one league to another. Additionally, by using such a popular league, this allows reproducibility as data for the Premier League is among the most accessible.

Other data used is the home win, draw and home loss betting odds for each match retrieved from Football-Data [14]. This data is used on the predicted results of my model, to analyze the model's efficiency as described in Metric 3. Lastly a subset of games from the second tier of English football (The Championship) will be included, as the model requires previous games played to predict future performance, and the nature of the Premier League is such that the worst teams at the end of each season are replaced by the best teams from the championship.

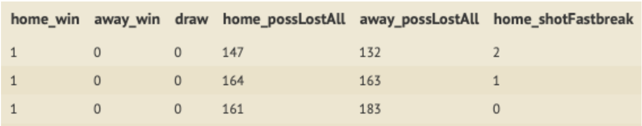
At this point I was left with over 3000 games to test and train on and over 200 statistical points for a single game. Some statistical points could be as broad as: possession, shots, and yellow cards. Or as detailed as passes left, passes in defensive half, and shots from the right on target.

In the end this data was brought down to just 56 stats per team per game, this was to reduce noise, as well as prevent the model from focusing on statistical points that directly impacted the game and were not predictable such as: goals, penalties, red cards.

This is one of many places that may be altered in the future to explore how adding or removing stats may affect the model.

After the statistical points were chosen, the data was flattened from its original JSON format to a more readable CSV format.

To generate the first of two models used to make predictions the CSV data was read into a DataFrame, and each statistical point was normalized around 1. The data was passed to one of three neural networks, each corresponding to a possible result (home win, draw, home loss).



home_win	away_win	draw	home_possLostAll	away_possLostAll	home_shotFastbreak
1	0	0	147	132	2
1	0	0	164	163	1
1	0	0	161	183	0

Figure 3. Data used to train first model

The usage of three independent neural networks was inspired by the separate independent Poisson distributions by Maher [9], as each neural network could be tuned to find which stats were most indicative of each result, as well as produce a percentage chance of that result occurring or one of the other two possibilities.

At this point three neural networks were created and trained on game data. Therefore, the accuracy rate of this model on predicting match results will be the ceiling for the eventual model as to predict future match results the game data would need to be predicted and no prediction can be better than the exact stats from the game.

Two different approaches were used for the second model which would attempt to predict the game statistics of future matches.

A. Average of past X games

This approach was the simplest, I simple aggregated the stat values from each teams' previous matches and divided by the total number of previous matches.

B. Hyperparameter Tuning

Approach two was more complex but considers the intricacies of soccer. The predicted game stats would be a weighted average based on how each team performed in the following categories,

- Recent performance (last 5 games)
- Head-to-Head
- Home vs Away

The average statistical value for each metric was calculated from the three game subsets, and the final average was calculated by adding the three averages after applying some weight to it based on the three above categories.

This approach incorporates the two important features which effect match results as outlined by Dixon-Coles [10] as well as an additional feature I deemed valid. Head-To-Head was added to incorporate how one team's style of play might skew results one way or another. As well as the oddity that is a weaker team's proclivity to overperform when playing a rival.

The weights were generated through hyperparameter tuning and nearest neighbor, in which a many different weights were tested for each category and the weights which produced match stats closest to the real values would become the default weights for predicting all future matches. Meaning once the optimal weights were found using the entire dataset, those were the weights used on all future testing data, and weights were not regenerated on each test session to prevent overfitting.

These two approaches were used to generate each teams predicted value for each stat value to then be passed to previously trained three models to predict the outcome of that match.

The goal of this model being to match the actual stats as closely from the game as to be as close to the ceiling (our max accuracy rate)

The output of this predicted stats model would be passed to each of the three neural networks which would return percentages mapping to the models predicted chance of each result, all three values were stored and used to calculate the model's success rate vs the three defined success metrics.

4. Results

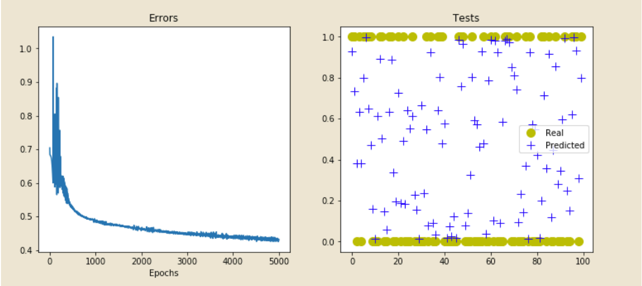


Figure 4. NN for Home Loss

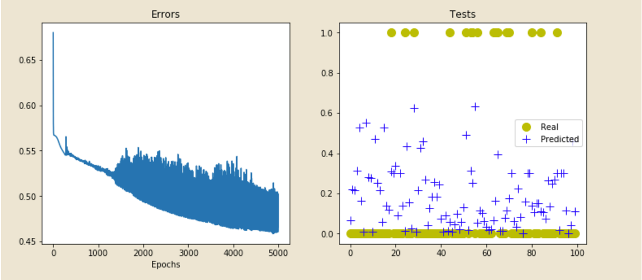


Figure 5. NN for Draw

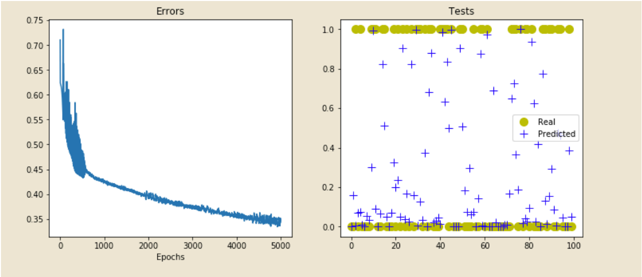


Figure 6. NN for Home Win

The left graph shows the change in error rates over time for each neural network. The right graph displays the predicted result percentage (plus) vs the actual result (circle). Unsurprisingly very few predicted values match up exactly to the actual values, as the model returns a percentage chance of the result, while the actual result is a binary 1 or 0. Most of the models converge at or below a 40% error rate.

Several interesting findings from this model, is its impressive accuracy for predicting a Home Loss, which is the least common result of the three possibilities. Additionally, the high variance in error rate for draws, which is reflected in its poor ability to predict those results.

A. Prediction results

The simplest way of having the model predict a result based on the three percentages returned is by simply predicting the result that has the highest percentage chance of occurring. This is a valid approach and the results of which will be shown below.

A new approach I introduced to my model used the betting odds for each game to optimize the return on investment but had an inverse effect on the prediction rate, though this result was anticipated. This approach made predictions based on each of the model's predicted chance of result compared to the betting chance which can be calculated by 1 / betting odds of result.

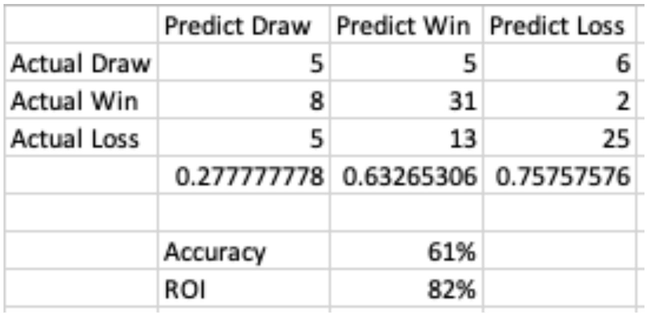
For example, my model may anticipate a team has a 60% chance of winning and only a 20% chance of drawing. Meaning a win should be predicted, however, if the betting odds predict a 55% chance of winning, and a 10% chance of drawing, this new approach will instead predict a draw, despite the lower chance of predicting correctly. The idea being despite the lower chance of the bet resulting in a win, in the long run that result will occur more often than the betting odds predict, and thus the return is greater than the lower returns achieved by betting on the agreed upon most likely result.

This approach was added to solve the previously mentioned issue of models with high accuracy rates, but low return on investments as the predictions were made on obvious results that did not yield a return high enough for the number of times it occurred.

The following three models were trained on 10% of the games ~100 games.

B. The Perfect Model

As expected, the model tested using real game data, which is impossible to know beforehand, performed optimally. With a 61% accuracy rate, this outperforms the 33% blind rate and 44% home team win rate, the first two of the 3 metric to compare the model. An 82% ROI means not only can the model predict accurately but does well across games, regardless of if the betting market considers that result a favorite.



	Predict Draw	Predict Win	Predict Loss
Actual Draw	5	5	6
Actual Win	8	31	2
Actual Loss	5	13	25
	0.277777778	0.63265306	0.75757576
	Accuracy	61%	
	ROI	82%	

Figure 7. Confusion Matrix for Perfect Model

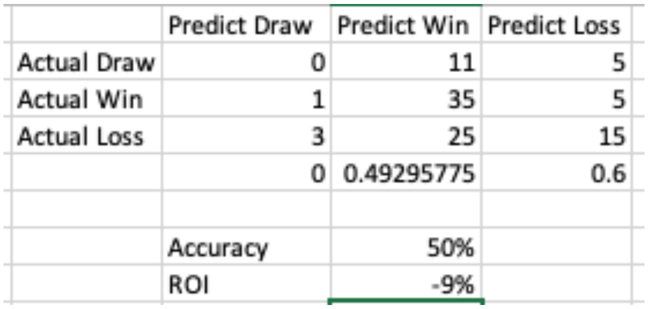
By testing this model using stats from the real game, the results of this model are to be considered optimal for this approach, and all other models which are tested on predicted game stats, should use these results as its ceiling for success.

Attempting to optimize the ROI by predicting on the relative difference between my model's predicted percentage and the betting odds percentage saw the accuracy rate decrease 5% as well as a decrease in the ROI by 20%.

This finding goes against what was predicted as the Accuracy AND Return were decreased. My hypothesis for which is, given this is the optimal model, the chance the predicted percentage for a result is wrong is the lowest as it predicts on the real game data, therefore choosing any result other than the one with the highest likelihood of happening will result in poorer results.

C. Average stats from all previous games

By simply predicting the game data by averaging the teams' previous values for each stat, resulted in an accuracy rate of 50% and a ROI of -9%.



	Predict Draw	Predict Win	Predict Loss
Actual Draw	0	11	5
Actual Win	1	35	5
Actual Loss	3	25	15
	0	0.49295775	0.6
	Accuracy	50%	
	ROI	-9%	

Figure 8. Confusion Matrix for AVG Stats

With a 50% accuracy rating, the model shows improvement over the first two metrics. The model underperforms relative to the betting odds, which is metric 3.

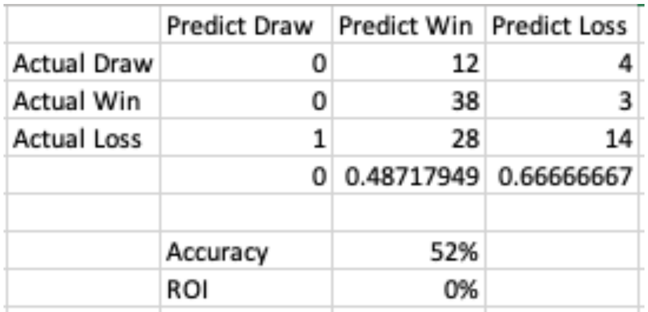
Optimizing for ROI by using the second prediction approach, gives us predicted result. The accuracy decreases significantly to 25% but the ROI increases to 7% (an increase of 18%).

D. Hyperparameter Tuning

By generating the game data used to test our model by getting a weighted average of previous games split in three categories: head-to-head, home and away, and recent performance.

This predicted game data had a lower MSE than the average of all previous games, which as expected saw it perform better.

This model predicted at a 52% accuracy rate, better than our metrics but still below the ceiling of 61%. The ROI just outperforms our metric with a return of 0.8%.



	Predict Draw	Predict Win	Predict Loss
Actual Draw	0	12	4
Actual Win	0	38	3
Actual Loss	1	28	14
	0	0.48717949	0.66666667
	Accuracy	52%	
	ROI	0%	

Figure 9. Confusion Matrix for HPT

Optimizing for ROI causes the Accuracy to drop to 33% but an increase in ROI to 31%.

5. Conclusions

The Hyperparametric tuned model achieved values for its Accuracy and ROI that exceeded the metrics defined for a successful model.

Many modifications and improvements can be made to increase these percentages, as the max possible values for each are known. One major improvement to be made is the model's ability to predict draws, as it only did so 1 time out of the entire testing dataset.

Another improvement would be to decrease the MSE of the predicted stats, the closer the predicted stats are to the real values the closer the model will be too optimal performance.

The original dataset was trimmed down significantly to only include certain games, and certain stats, future models may attempt to change which games and stats are chosen, which would result in different results (as well as new ceiling).

While the HPT model for predicting stats was optimal it is possible that the weights provided for each category represent a local maximum rather than a global, therefore comparing the MSE with additional games not included in this dataset would be wise.

Yet no model will be perfect, nor should any model be perfect, what makes sports so great is the belief that anything can happen. With the right mixture of skill, hard work and luck anything can happen. It doesn't matter how much better the opposition is, or how the badly the odds may be stacked against your favorite team, billions of people watch the sport every year because of that believe that anything is possible and that is truly ***unquantifiable.***

References

[1] Mental Itch. 2021. *The History of Tsu Chu.* Retrieved November 18, 2021 from https://mentalitch.com/the-history-of-tsu-chu/

[2] Lock. Feb 2021. *Frequency of gambling in Great Britain in 2020.* Retrieved November 18, 2021 from https://www.statista.com/statistics/543450/gambling-frequency-united-kingdom-uk/

[3] The Betting and Gaming Act, 1960. The Journal of Criminal Law. 1961;25(2):149-155. doi:10.1177/002201836102500209

[4] Premier League. 2021. *Entertaining Audiences.* Retrieved November 18, 2021 from https://www.premierleague.com/this-is-pl/the-fans/686489?articleId=686489

[5] Pesca. July 2009. *The Man Who Made Baseball's Box Score A Hit.* Retrieved November 18 2021 from https://www.npr.org/templates/story/story.php?storyId=106891539

[6] Arastey. November 2019. *History of Performance Analysis: The Controversial Pioneer Charles Reep.* Retrieved November 8 2021 from https://www.sportperformanceanalysis.com/article/history-of-performance-analysis-the-controversial-pioneer-charles-reep

[7] Moroney. (1956) *Facts from Figures*. 3rd edn. London: Penguin.

[8] Reep and Benjamin. (1968). *Skill and Chance in Association Football.* J. R. Statist. Soc A, 131, 581-585.

[9] Maher. (1982) *Modelling Association Football Scores.* Statist. Neerland. 36, 109-118.

[10] Dixon, Coles. (1997). *Modelling Association Football Scores and Inefficiencies in Football Betting Market.* Applied Statistics, 46, 265-280.

[11] Winchester. Sept 2019. *Dixon Coles Model.* Retrieved November 12, 2021 from https://philipwinchester.github.io/dixon-coles-model/

[12] Arastey. June 2018. *Opta Sports: The Leading Sports Data Provider.* Retrieved on November 18, 2021 from https://www.sportperformanceanalysis.com/article/opta-leading-sport-data-provider

[13] Betegey. 2021. *Pro Accuracy Tracker.* Retrieved on November 18, 2021 from https://askbettor.com/tools/accuracy_tracker

[13] Ulmer and Fernandez. (2014). *Predicting Soccer Match Results in the English Premier League.* https://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf

[14] Football-Data. 2021. *Historical Football Results and Betting Odds Data.* Retrieved on October 20, 2021 from https://www.football-data.co.uk/englandm.php