Modelling Football Scores in the Italian Serie A

Kurt Darmanin (21207174)
Meghashree Madhava Rao Ramachandrahosur (21200301)

29th July 2022

Abstract

The immense economic growth experienced in the footballing world has sparked an interest in the evolution of sports modelling, both in academia and in the gambling industry. This paper demonstrates a novel goals-based model for predicting the results of a football match, with this model being fitted to data from the Italian Serie A between 2017 and 2021. The predictive performance of this model is then tested and compared against the bookmakers' odds from 2021 to 2022. The technique is based on a linear Poisson regression model, however a non-linear version of this model that makes use of neural networks is also proposed. It has been found that the predictive performance of our linear Poisson regression model is essentially comparable to that of the bookmakers', making its use suitable for the development of a profitable betting strategy.

Keywords: Poisson Regression, Neural Networks, Scores Modelling, Football (Soccer), Italian Serie A.

1 Introduction

Over the years, the footballing world has grown into a multi-billion Euro industry, with the Italian first division, the Serie A, having a total of around 520 million fans worldwide (Verri, 2022). With regards to the betting industry, the global amount wagered on this league alone has amounted to €38.4 billion in 2020, with the total number of monies being wagered on European football tallying up to around €188 billion (Complete Sports, 2021). The demand stemming from this immense interest within the sport has thus led both academia as well as industry to propose several novel ideas in order to model the outcome of the game, with the aim here being to accurately come up with the odds of a team to win, draw or lose a match, prior to the game taking place.

Most of the research within literature can be summed up into two groups, namely those putting forward results-based models (Forrest & Simmons, 2000; Baboota & Kaur, 2019) and those suggesting goals-based models (Dixon & Coles, 1997; Egidi et al., 2018; Ankomah et al., 2020; Wheatcroft, 2020). The former approach is a multi-class classification problem, where the model has to correctly classify if a team will win, draw, or lose the match. On the other hand, the latter approach focuses on the scoreline by simulating how many goals the home and away teams will score, and then subsequently inferring the outcome of the match from the simulated scoreline.

A goals-based model is put forward here based on the bivariate Poisson model by Dixon and Coles (1997), with the Poisson parameter being expressed in terms of a regression model as in Ankomah et al. (2020). The choice of variables is inspired by Baboota and Kaur (2019), whereby in addition to the teams' match statistics, the difficulty of the match, the scoring and defending forms of the teams, and their FIFA ratings are considered. The main variation here is that a number of regression models are examined, namely the linear regression model and the neural network model. Several hyperparameter settings for the latter model are finely tuned, so as to make sure that the optimal settings are chosen. Following the training and validation stages, the performance of the models are then compared to the bookmakers' odds using the test set. In order to make this comparison, the ranked probability score (RPS) is used as the metric that determines the accuracy of the models' and the bookmakers' odds.

The remainder of this paper is organised in the following way. Section 2 describes the data that is available to us, the features and transformations that are considered, and the variables that are ultimately chosen as inputs for the models. The models themselves are illustrated in detail in Section 3. In Section 4, the results of these models and how they fare against the bookmakers' probabilities is explored. A betting strategy based on the best performing model is developed in Section 5. Finally, Section 6 summarises the results and suggests improvements that would be beneficial for future research.

2 Data

2.1 Data Description

The data for the match statistics were obtained from a public source, this being Football-Data.co.uk (`https://www.football-data.co.uk/italym.php`). This website provides a plethora of regularly updated data sets for match statistics from various football leagues in Europe and beyond. Apart from the

scoreline, these data sets also include other important features such as the number of shots, corners, and shots on target attempted, as well as the number of fouls and bookings committed by each team during the match. Moreover, the betting odds for a home win, an away win, or a draw by a number of key bookmakers is also provided. We will focus on the odds produced by Bet365 when making the comparison with the models' predictions, given that the company is one of the leading players in the gambling industry.

Interestingly, aside from using the aforementioned match statistics, Baboota and Kaur (2019) also incorporate the team ratings from the FIFA video game series by Electronic Arts. The same will be attempted here, with the data being acquired from FIFA Index (`https://www.fifaindex.com`). The defending, midfield, attacking, and overall ratings for each team were thus scraped from this online database.

By combining the two data sets from Football-Data.co.uk and FIFA Index, the engineered final data set is structured in such a way that each observation row i represents the inputs for a function that predicts the number of goals that a single team will score in a particular match. Hence, there are double the number of rows i as there are matches in the data set. For instance, for any given match, there would be a row where the home team is the reference team, $RefTeam_i$, and the away team is the opposing team, $OppTeam_i$. Then there would be another row where the away team is the reference team, $RefTeam_i$, and the home team is the opposing team, $OppTeam_i$.

In total, five seasons' worth of data were collected, with these being the seasons that range between the years of 2017 and 2022. The former four seasons are used as training and validation sets with an 80%-20% random split. On the other hand, the remaining 2021-2022 season is used as the test set that compares the accuracy of the predicted probabilities with those produced by Bet365.

2.2 Feature Engineering

As is noted in Dixon and Coles (1997), for a statistical model of football matches to be representative of reality, various types of features need to be considered. The authors summarise the required considerations that need to be taken into account in the following points:

a) The model should take into account the different abilities of both teams in a match.

b) Teams playing at home are believed to have an advantage, this being the so-called *home effect*.

c) A reasonable measure of a team's ability is likely to be based on a summary measure of their most recent performances.

d) Given the way that the game is played, a team's ability is likely to be best summarised by having separate measures for their ability to attack (to score goals), and their ability to defend (not to concede goals).

e) When summarising a team's performance in terms of its recent results, the abilities of the teams that they have played against should be taken into account.

To accommodate for the last point, two weighting parameters, w_{DEF} and w_{ATT}, are employed. The former weighting parameter, w_{DEF}, adjusts the variables according to the past opponents' defending strengths. This weighting parameter can take values $w_{DEF} = \{1.950, 1.446, 1.183, 1.000\}$. The higher the value of w_{DEF}, the higher the quality of the opponents' defensive capabilities. On the other hand, the latter weighting parameter, w_{ATT}, adjusts the variables according to the past opponents' attacking strengths. This weighting parameter can take values $w_{ATT} = \{1.000, 1.335, 1.671, 2.095\}$. The higher the value of w_{ATT}, the lower the quality of the opponents' attacking capabilities. The values for these two weighting parameters are inspired by the ratios of goals scored and conceded by the Serie A teams between the years 2017 and 2021. The way that these values are assigned is mainly dependent on the teams' FIFA ratings.

Hence, with the purpose of tackling each and every point mentioned above, the following features are engineered by combining the teams' match statistics with the weighting parameters w_{DEF} and w_{ATT}:

- $RefGSF_i$: A measure for the goal scoring form of the reference team, calculated from the number of goals scored by the reference team in the past 6 games, $RefGS_{i-j}$, and adjusted for the reference team's past opponents' defending strengths, w_{DEF}.

$$RefGSF_i = \log(\frac{1}{6}\sum_{j=1}^{6} w_{DEF} RefGS_{i-j})$$

- $RefSF_i$: A measure for the shooting form of the reference team, calculated from the number of shots attempted by the reference team in the past 6 games, $RefS_i$, and adjusted for the reference team's past opponents' defending strengths, w_{DEF}.

$$RefSF_i = \log(\frac{1}{6}\sum_{j=1}^{6} w_{DEF} RefS_{i-j})$$

- $RefSTF_i$: A measure for the shooting accuracy form of the reference team, calculated from the number of shots-on-target attempted by the reference team in the past 6 games, $RefST_i$, and

adjusted for the reference team's past opponents' defending strengths, w_{DEF}.

$$RefSTF_i = \log(\frac{1}{6}\sum_{j=1}^{6} w_{DEF} RefST_{i-j})$$

- $RefCF_i$: A measure for the pressing form of the reference team, calculated from the number of corners won by the reference team in the past 6 games, $RefC_i$, and adjusted for the reference team's past opponents' defending strengths, w_{DEF}.

$$RefCF_i = \log(\frac{1}{6}\sum_{j=1}^{6} w_{DEF} RefC_{i-j})$$

- $OppGDF_i$: A measure for the goal defending form of the opposing team, calculated from the number of goals conceded by the opposing team in the past 6 games, $OppGC_{i-j}$, and adjusted for the opposing team's past opponents' attacking strengths, w_{ATT}.

$$OppGDF_i = \log(\frac{1}{6}\sum_{j=1}^{6} w_{ATT} OppGC_{i-j})$$

- $OppSTDF_i$: A measure for the defending form of the opposing team, calculated from the number of shots-on-target attempted against the opposing team in the past 6 games, $OppSTC_i$, and adjusted for the opposing team's past opponents' attacking strengths, w_{ATT}.

$$OppSTDF_i = \log(\frac{1}{6}\sum_{j=1}^{6} w_{ATT} OppSTC_{i-j})$$

- $Home_i$: A binary variable that takes a value of 1 if the reference team is playing in their home stadium, and a value of 0 if the reference team is playing away.

$$Home_i = \begin{cases} 1 & \text{if } RefTeam_i \text{ plays Home} \\ 0 & \text{if } RefTeam_i \text{ plays Away} \end{cases}$$

- Y_i: The number of goals scored by the reference team, $RefGS_i$, which is the output variable in our models.

$$Y_i = RefGS_i$$

The logarithmic transformation, $\log(x)$, is applied to several of the above features. The reasoning behind this is that since these variables are in the range of $[0, \infty]$, applying a log-transformation not only changes their range to vary in $[-\infty, \infty]$, but it also makes it easier for these input variables to have a linear relationship with the output variable. The latter statement especially comes in handy when it comes to satisfying the conditions that need to be met in order to apply a linear regression model. Furthermore, prior to applying the log-transformation $\log(x)$, in the very few cases where the values $x = 0$, these were changed to $x = 0.1$ so as to make the transformation possible.

The focus in Dixon and Coles (1997) is to capture the teams' short-term form. However, they do not place any emphasis on a team's long-term qualities. As such, the FIFA ratings are coupled with the match statistics, whereby whilst the latter are used to assess the form of a team, the former are used as a measure of the players' talent within a team. In this regard, the following features are engineered:

- $ATTDEF_i$: The difference of the FIFA defence rating of the opposing team, $OppDEF_i$, from the FIFA attack rating of the reference team, $RefATT_i$.

$$ATTDEF_i = RefATT_i - OppDEF_i$$

- $MIDMID_i$: The difference of the FIFA midfield rating of the opposing team, $OppMID_i$, from the FIFA midfield rating of the reference team, $RefMID_i$.

$$MIDMID_i = RefMID_i - OppMID_i$$

- $DEFATT_i$: The difference of the FIFA attack rating of the opposing team, $OppATT_i$, from the FIFA defence rating of the reference team, $RefDEF_i$.

$$DEFATT_i = RefDEF_i - OppATT_i$$

- $OVROVR_i$: The difference of the FIFA overall rating of the opposing team, $OppOVR_i$, from the FIFA overall rating of the reference team, $RefOVR_i$.

$$OVROVR_i = RefOVR_i - OppOVR_i$$

Other variables that are included in the data set but have not already been mentioned are the following:

- $RefBPF_i$: A measure for the discipline form of the reference team, calculated from the number of yellow cards, $RefYLW_{i-j}$, and red cards $RefRED_{i-j}$, received by the reference team in the past 6 games. Each yellow card has a weighting of 10 points, whilst each red card carries a weighting of 25 points. Lower values of this variable signify that the reference team is a well-disciplined team.

$$RefBPF_i = \log(\tfrac{1}{6}\textstyle\sum_{j=1}^{6}(10\, RefYLW_{i-j} + 25\, RefRED_{i-j}))$$

- $OppBPF_i$: A measure for the discipline form of the opposing team, calculated from the number of yellow cards, $OppYLW_{i-j}$, and red cards $OppRED_{i-j}$, received by the opposing team in the past 6 games. Each yellow card has a weighting of 10 points, whilst each red card carries a weighting

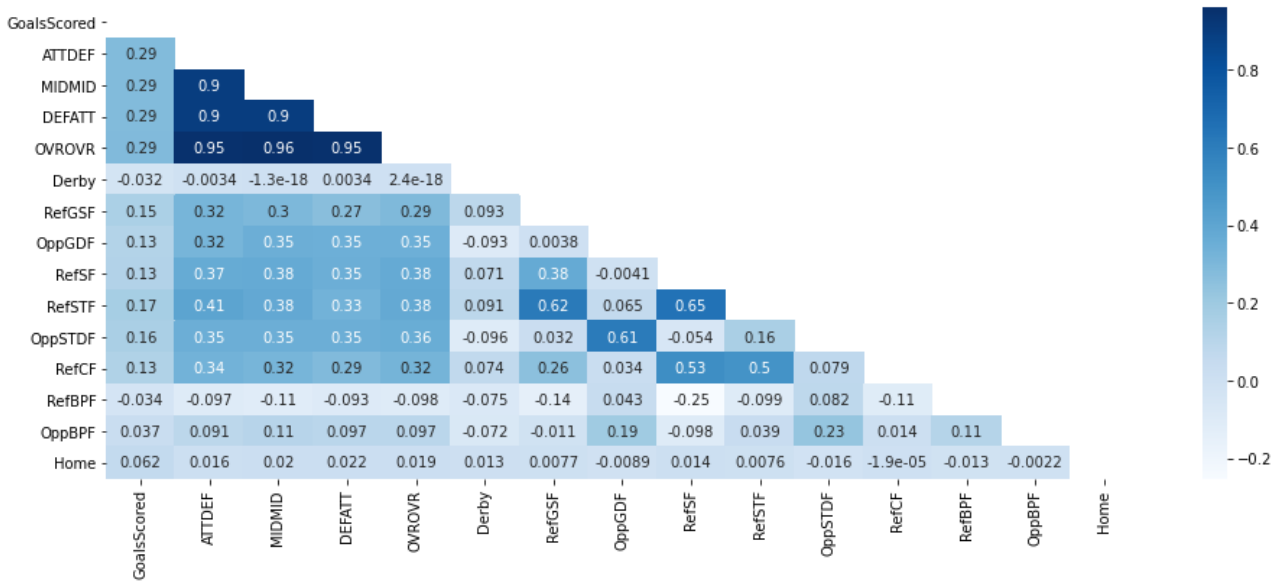

Figure 1: Correlation Heatmap for Engineered Features under Consideration

of 25 points. Lower values of this variable signify that the opposing team is a well-disciplined team.

$$OppBPF_i = \log(\tfrac{1}{6}\sum_{j=1}^{6}(10\,OppYLW_{i-j} + 25\,OppRED_{i-j}))$$

- $Derby_i$: A binary variable that takes a value of 1 if the reference team and the opposing team are rivals, and a value of 0 otherwise.

$$Derby_i = \begin{cases} 1 & \text{if Teams are Rivals} \\ 0 & \text{if Teams are not Rivals} \end{cases}$$

It is important to note that the features engineered here are computed across each season independently. Hence, there is no sharing of information across seasons. What is more is that since most of the variables use data from the previous six games, and since no data from the previous seasons are considered, then the models discussed here only work from 7th game week of the season onward.

2.3 Feature Selection

The aim of feature selection is to select only the most relevant of variables, so as to reduce the total number of features. In this context, the chosen features should be able to predict the number of goals scored, Y_i, whilst ideally, also not being highly correlated with the other explanatory variables. There are in total 14 features that were engineered, from which a select few have to be chosen as inputs for our models.

Prior to choosing which features should be included, the correlation matrix of such features along with Y_i is analysed. Only data from the training set are considered here. A correlation heatmap for the engineered features is illustrated in Figure 1. Starting off with the features stemming from match statistics, the variables that have the highest correlations with the number of goals scored are $RefSTF_i$ and $OppSTDF_i$. Moreover, $RefGSF_i$, $RefSF_i$, and $RefCF_i$ are highly correlated with $RefSTF_i$, whilst $OppGDF_i$ has a strong relationship with $OppSTDF_i$. On the other hand, both $RefBPF_i$ and $OppBPF_i$ have close to no correlation with the number of goals scored. Hence, from the subset of match statistics variables, only $RefSTF_i$ and $OppSTDF_i$ are chosen.

Moving on to the FIFA variables, it is evident that $ATTDEF_i$, $MIDMID_i$, $DEFATT_i$, and $OVROVR_i$ are highly correlated with each other. The latter feature has the highest correlation with the number of goals scored, albeit, not by much. As such, $OVROVR_i$ is selected as part of the final data set. The remaining two variables, $Home_i$ and $Derby_i$, is also selected. These two binary variables are known to change the dynamics of a football match. The $Home_i$ represents the so-called *home effect*, whereby the home team is known to have an advantage over the away team due to their fans' support. The other binary variable, $Derby_i$, is also believed to have an effect on the number of goals scored. The teams in derby matches tend to be fierce rivals, possibly adding pressure to the players on the field. As such, it makes sense how the added pressure might reduce the number of goals scored in a match.

In summary, the chosen features are $RefSTF_i$, $OppSTDF_i$, $OVROVR_i$, $Home_i$, and $Derby_i$. The first two variables, $RefSTF_i$ and $OppSTDF_i$, represent the short-term form of any two teams playing each other from their recent match statistics, with these features describing the team's attacking abilities and their opponent's defending abilities, respectively. The $OVROVR_i$ feature takes into account the quality of players on the field by taking the difference between

the International Federation of Association Football (FIFA) ratings of the two teams playing a match. Finally, the $Home_i$ and $Derby_i$ binary features describe the *home effect* and the added pressure that comes with a derby match. Therefore, the set of input variables, X_i, that are used to model the number of goals scored may be described as $X_i = \{1, RefSTF_i, OppSTDF_i, OVROVR_i, Home_i, Derby_i\}$.

3 Predictive Models

3.1 Poisson Regression

The model put forward here aims to first predict the scoreline, and then infer the outcome of the match from said scoreline. As is shown in Dixon and Coles (1997), the number of goals are assumed to follow a Poisson distribution, with the goals scored by the home team assumed to be independent of the goals scored by the away team. Inspiration is also taken from the methodology in Ankomah et al. (2020), whereby Poisson regression is used to model the scores of football matches.

The difference here is that the Poisson parameter is not constant for the whole season, but rather, it will change for every team depending on the match. The form of the teams at play, their quality of players, whether the match is a derby or not, and whether the team will be playing at home or away will thus all be taken into account. In this way, the Poisson parameter, λ_i, is expressed in terms of a regression model, $E(Y_i|X_i)$, whereby equidispersion is assumed, such that:

$$\lambda_i = E(Y_i|X_i) = Var(Y_i|X_i)$$

The regression model stated here could either take the form of a linear model or a neural network model, both of which will be delved into further detail later on in this section. Given that $\lambda_i > 0$, the output of these regression models are made to be equivalent to the logarithm of the expected number of goals scored, $\log(E(Y|X))$. As such, any $Y_i = 0$ was changed to $Y_i = 0.1$ in order to make this logarithmic transformation possible. To get the estimated Poisson parameter, we would then simply need to take the exponential of the regression output. Hence, after acquiring an estimate for the Poisson parameter, λ_i, the number of goals scored by a particular team in a given match, Y_i, could then be simulated from the following Poisson model:

$$P(Y_i = y) = \frac{e^{-\lambda_i}(\lambda_i^y)}{y!}$$

The number of goals are simulated $N = 10,000$ times for each Y_i. The goal simulations of both teams in a match are then grouped in order to get N scorelines for every single game. Suppose that the number of goals scored by the home team in a given match, Y_h, and the number of goals scored by the away team in the same match, Y_a, are independently modelled by a Poisson distribution with parameters λ_h and λ_a, respectively. Then, the simulated scorelines are equivalent to being drawn from a bivariate Poisson distribution, as follows:

$$P(Y_h = y_h, Y_a = y_a) = \frac{e^{-\lambda_h}(\lambda_h^{y_h})}{y_h!}\frac{e^{-\lambda_a}(\lambda_a^{y_a})}{y_a!}$$

Therefore, the proportion of simulations where $Y_h > Y_a$, $Y_h = Y_a$, and $Y_h < Y_a$, estimate the model probabilities of a home win, $P(H)$, a draw, $P(D)$, and an away win, $P(A)$, for a given match, respectively.

$$P(H) = \frac{n(Y_h > Y_a)}{N}$$

$$P(D) = \frac{n(Y_h = Y_a)}{N}$$

$$P(A) = \frac{n(Y_h < Y_a)}{N}$$

3.2 Linear Model

The first proposed method for estimating λ_i is to make use of a linear regression model, resulting in a linear Poisson regression model. As was previously stated, $\log(E(Y|X))$ is taken as the model's output, whilst X is taken as the model's input. The regression weights, w, are estimated using the ordinary least squares method.

$$\log(E(Y|X)) = X'w$$

After the model is trained on the training set, estimations for λ_i may be produced by taking the exponential of the linear model's output.

$$\lambda_i = E(Y_i|X_i)$$

3.3 Non-Linear Model

The second proposed method for estimating λ_i is to make use of a neural network model, resulting in a non-linear Poisson regression model. Similar to the linear model, $\log(E(Y|X))$ is taken as the model's output, whilst X is taken as the model's input. In this case, the input X is normalised prior to training the model. The number of hidden layers, G, within the network is a hyperparameter, such that $G \in \{0, 1, 2\}$. The ReLU function, $f(z) = max(0, z)$, is used as the activation function for all of the hidden layers. A different set of weights are present within this model, naturally depending on G.

When there are no hidden layers in the network, the set of weights $w^{(1)}$ connect the input layer with $V + 1$ nodes to the output layer with a single node. In this case, $V + 1 = 6$ since we have 6 variables, including the intercept. It is important to note that this model would be equivalent to the linear model. However, the weights are trained differently, as the Adam optimisation algorithm is used instead of the

ordinary least squares method.

If $G = 0$:

$$\log(E(Y|X)) = \sum_{j=0}^{V} w_j^{(1)} x_j$$

When there is a single hidden layer in the network, there are two set of weights, these being $w^{(1)}$ and $w^{(2)}$. The weights $w^{(1)}$ connect the input layer with $V + 1$ nodes to the hidden layer with $H_1 + 1$ nodes, including the bias terms. The weights $w^{(2)}$ then connect the hidden layer to the output layer with a single node. The number of nodes in the hidden layer is a hyperparameter, such that $H_1 \in \{2, 3, 4, 5\}$.

If $G = 1$:

$$\log(E(Y|X)) = \sum_{k=0}^{H_1} w_k^{(2)} f(\sum_{j=0}^{V} w_{jk}^{(1)} x_j)$$

When there are two hidden layers in the network, there are three sets of weights, these being $w^{(1)}$, $w^{(2)}$, and $w^{(3)}$. Again, the weights $w^{(1)}$ connect the input layer with $V + 1$ nodes to the first hidden layer with $H_1 + 1$ nodes, including the bias terms. The weights $w^{(2)}$ connect the first hidden layer to the second hidden layer with $H_2 + 1$ nodes, including the bias terms. Finally, the weights $w^{(3)}$ connect the second hidden layer to output layer with a single node. The number of nodes in the hidden layers are hyperparameters, such that $H_1 \in \{2, 3, 4, 5\}$ and $H_2 \in \{2, 3, 4\}$.

If $G = 2$:

$$\log(E(Y|X)) = \sum_{l=0}^{H_2} w_l^{(3)} f(\sum_{k=0}^{H_1} w_{kl}^{(2)} f(\sum_{j=0}^{V} w_{jk}^{(1)} x_j))$$

In order to avoid overfitting, a number of regularisation methods are used. L2-regularisation, a type of penalty-based regularisation method, is used in every hidden layer. Different penalty parameters which control the amount of regularisation are used in each layer, with λ_1 and λ_2 being the penalty parameters for the first and second hidden layers, respectively. These penalty parameters are treated as hyperparameters, with $\lambda_1, \lambda_2 \in \{0, 0.001, 0.0025, 0.005, 0.01\}$.

Dropout is also employed as a method of regularisation. When training the model, a percentage of the network nodes are switched off, leading to underfitting. All the network nodes are then employed at the testing stage, with the weight parameters scaled appropriately. The rates at which these nodes are switched off vary by hidden layer, with these rates being finely tuned and taking possible values of $\{0, 0.3, 0.4\}$.

The final regularisation method used here is that of early stopping. This involves stopping the training procedure when the validation error starts to increase. In this case, the training will stop if after 20 consecutive epochs, there has been no improvement in the validation loss.

We previously mentioned that the Adam optimisation algorithm is used to train the different set of weights. In general, the Adam algorithm tends to perform well in many applications, especially since the convergence and performance is usually less sensitive to the tuning of the learning rate when compared with other optimisation methods. Nonetheless, the learning rate is also made to be a hyperparameter, with the possible values being in the set of $\{0.001, 0.002, 0.005, 0.01\}$. A larger learning rate allows the model to learn faster at the cost of a potentially sub-optimal set of estimated weights. On the other hand, a smaller learning rate could lead to an optimal set of weights but at the cost of higher training time, and at the risk of the training algorithm becoming trapped in a local region.

The neural network model described here involves quite a number of different hyperparameters that need to be tuned. The optimal set of hyperparameters settings is sought for by using a tuning algorithm. For this reason, the Hyperband algorithm as described in Li et al. (2018) is used. The ideal result would be for the tuning algorithm to yield the best possible model that generalises the problem at hand, mainly by changing the architecture, the amount of regularisation, and the learning rate.

Following the hyperparameter tuning search, the model is trained on the optimal set of hyperparameters, after which estimations for λ_i may be produced by taking the exponential of the neural network model's output.

$$\lambda_i = E(Y_i|X_i)$$

4 Results and Discussion

The models are coded using the Python programming language. Several libraries are made use of, namely, *NumPy* and *pandas* for data wrangling, *Matplotlib* and *Seaborn* for plotting, *Keras*, *Tensorflow*, and *statsmodels* for modelling purposes, and *Sklearn* in order to assess these models.

Traditional metrics such as the accuracy, precision, recall, and F1 scores are utilised so as to describe the models put forward here as well as the bookmaker's predicted outcomes. These metrics are calculated on the test data, which solely consists of the 2021-2022 season. As was previously mentioned, due to first six games being used to engineer some of the features, predictions could only be made for the matches taking place from 7th game week onward. The outcome with the highest predicted probability are taken in each case in order to classify if a match will be won by the home team, drawn, or won by the away team.

A more novel metric, the Ranked Probability Score (RPS), is also used. The RPS is a metric which assesses how well predicted events fit probability distributions, with a value of 0 signifying a prediction that is entirely accurate, whilst a value of 1 signifies a prediction that is entirely inaccurate. This metric is particularly useful in football modelling, given that rather than being solely dependent on the classification of the outcome itself, the RPS is sensitive to the probabilities given to each possible outcome. In fact, a thorough analysis is undertaken in Constantinou and Fenton (2012), demonstrating that the RPS is the ideal metric to measure the accuracy of a probabilistic football forecasting model.

The RPS for a single instance is defined mathematically as follows:

$$\text{RPS} = \frac{1}{r-1}\sum_{i=1}^{r}(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j)^2$$

where r is the number of outcomes, p_j is the predicted probability of outcome j, and e_j is the actual probability of outcome j. For the sake of comparison, when references are made to the RPS metric, what we are referring to is actually the mean RPS values for that particular model.

4.1 Linear Poisson Regression Model

The results show that the Linear Poisson Regression (LPR) model has a classification accuracy of 51.56% and a mean RPS score of 0.2016. With a recall value of 0.63, the LPR model does exceptionally well when it comes to predicting home wins, albeit being somewhat biased towards the number of games that the home team wins, as evidenced by the slightly lower precision value of 0.53.

On the contrary, the precision in predicting away wins is proportionally quite high, given a value of 0.57. The relative weakness in the LPR model lies in the underprediction of away wins, as evidenced by the low recall value when compared to the other model and the bookmaker's predictions.

Unfortunately, although the model is fairly accurate when it comes to predicting home or away wins, the F1-scores show that it fails to be accurate when predicting draws. It is worth noting that the rate at which draws are predicted is actually slightly lower than the rate of observed matches that end in draws. Hence, the problem here is not one of underprediction, but one of misclassification. However, as will be seen later on in this section, both the NLPR model and the bookmaker fail to ever predict a single draw.

The strength of the LPR model mainly lies in its simplicity. This model is parsimonious enough given that it has a handful of variables and an equivalent handful of parameters that need to be estimated, whilst at the same time, it has enough features to correctly model the problem at hand. A summary of these results may be found in Table 1.

Table 1: Metrics for the LPR Model

(a)Confusion Matrix

	Pred. Home Win	Pred. Draw	Pred. Away Win
Act. Home Win	79	20	26
Act. Draw	41	23	21
Act. Away Win	30	17	63

(b)Precision-Recall Table

	Precision	Recall	F1-score
Home Win	0.53	0.63	0.57
Draws	0.38	0.27	0.32
Away Win	0.57	0.57	0.57

4.2 Non-Linear Poisson Regression Model

Prior to discussing the results of the Non-Linear Poisson Regression (NLPR) model, we first describe the model configurations for the neural network model. The Hyperband tuning algorithm found that the optimal set of hyperparameters are a neural network model with a single hidden layer, $G = 1$, which contains $H_1 = 4$ nodes, not including the bias terms. With regards to regularisation, no dropout or L2-regularisation is employed, as the dropout rate and the λ_1 penalty parameter are both set to 0. This leaves early stopping as the only method of regularisation in the neural network model. Furthermore, the optimal learning rate is set to 0.01.

The subsequent NLPR model has a classification accuracy of 48.75% and a mean RPS score of 0.2062, thus lagging behind the LPR model in both metrics. The reasoning behind these lackluster results is that the NLPR model overpredicts the number of home wins and away wins at the cost of never predicting draws. This is evidenced by the high recall values in home and away wins, but extremely low values in the corresponding precision figures. Hence, although the F1-scores for the the home and away wins are comparable to those of the LPR model, an F1-score of 0 for classifying draws highlights the main problem with the NLPR model. A summary of these results may be found in Table 2.

Table 2: Metrics for the NLPR Model

(a)Confusion Matrix

	Pred. Home Win	Pred. Draw	Pred. Away Win
Act. Home Win	81	0	44
Act. Draw	49	0	36
Act. Away Win	35	0	75

(b)Precision-Recall Table

	Precision	Recall	F1-score
Home Win	0.49	0.65	0.56
Draw	0.00	0.00	0.00
Away Win	0.48	0.68	0.57

The failures behind the NLPR model are likely due to a lack of features being available to model the λ_i

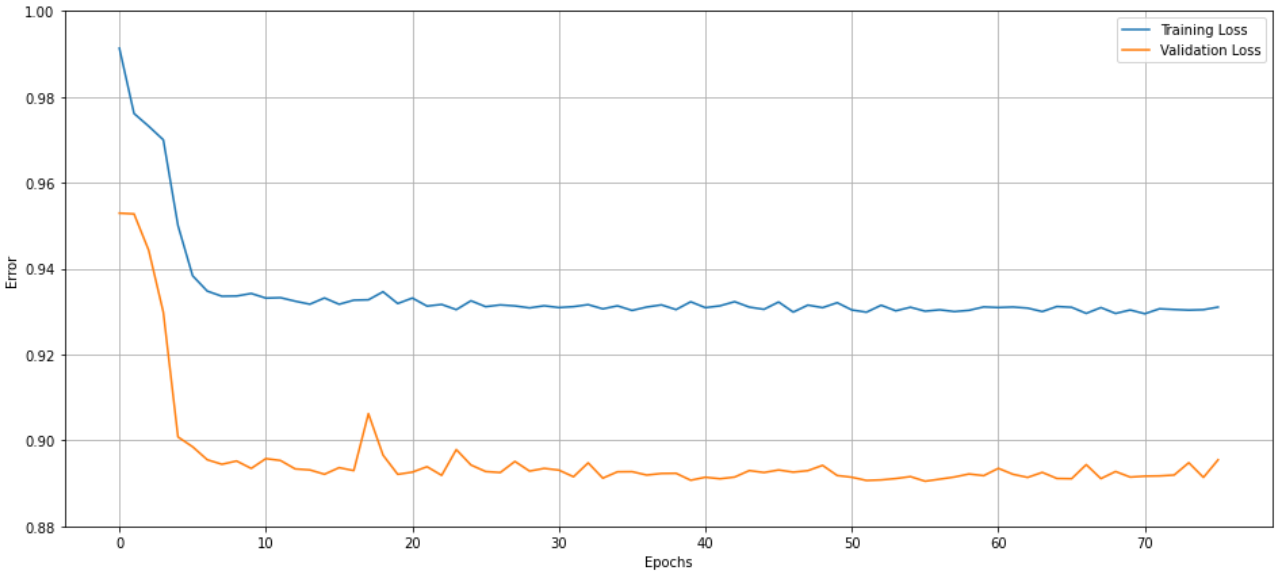

Figure 2: Learning Curve for the Neural Network Model

parameter. Neural network models typically thrive in scenarios where there are numerous features available. In this case, there are only $V = 5$ variables to model the task at hand. The learning curve illustrated in Figure 2 shows quite a difference between the training error and the validation error, which gives a further indication towards the model being underfit. Moreover, the relative success of the LPR model might suggest that following the various transformations applied to both the output and input variables, the problem of modelling the λ_i parameter might be a linear one, thus eroding the benefits of employing a non-linear model.

4.3 Bookmakers' Predicted Outcomes

The odds produced by the bookmaker, Bet365, are utilised as a standard of comparison for our models, from which the implied probabilities may be calculated. The implied probabilities reflect what the bookmakers think the likeliness of an outcome is to occur. These are converted from the betting odds, taking into account the bookmaker's margin which intentionally skews the odds in favour of the bookmaker for them to be able to turn a profit. As such, the inverse of the odds is always expected to be greater than 1. Basic normalisation is employed to alleviate for the bookmaker's margin in a similar fashion to what is discussed in Dixon and Coles (1997), Strumbelj (2014), and Baboota and Kaur (2019).

Let $O = \{O_H, O_D, O_A\}$ be the bookmaker's odds, where O_H, O_D, and O_A represent the odds of a home win, a draw, and an away win, respectively. Furthermore, consider the set of implied probabilities $\pi = \{\pi_H, \pi_D, \pi_A\}$ that are computed by taking the inverse of the odds as follows:

$$\pi_i = \frac{1}{O_i} \ \forall i \in \{H, D, A\}$$

Then, the normalised probabilities, $B = \{B(H), B(D), B(A)\}$, are computed by dividing the inverse odds, π, by the booksum, Π, such that:

$$B(i) = \frac{\pi_i}{\sum_i \pi_i} = \frac{\pi_i}{\Pi} \ \forall i \in \{H, D, A\}$$

The normalised probabilities show that the predictions from Bet365 have a classification accuracy of 51.25% and a mean RPS score of 0.1992. Although the recall values signify that the bookmaker has managed to correctly classify home and especially away wins at an astounding rate, the precision values show that this was done at the cost of overpredicting a match being won by either side. Surprisingly, not a single draw was predicted by the bookmaker. Therefore, even though the F1-scores for the home and away are relatively high, the massive underprediction of draws lowers the overall prediction accuracy of the bookmaker. A summary of these results may be found in Table 3.

Table 3: Metrics for the Bet365 Predicted Outcomes

(a)Confusion Matrix

	Pred. Home Win	Pred. Draw	Pred. Away Win
Act. Home Win	96	0	29
Act. Draw	51	0	3
Act. Away Win	42	0	68

(b)Precision-Recall Table

	Precision	Recall	F1-score
Home Win	0.52	0.62	0.56
Draws	0.00	0.00	0.00
Away Win	0.51	0.77	0.61

4.4 Comparing the Models with the Bookmaker

In summary, the LPR model has the highest accuracy value at 51.56%, closely followed by Bet365's predicted outcomes which have an accuracy of 51.25%,

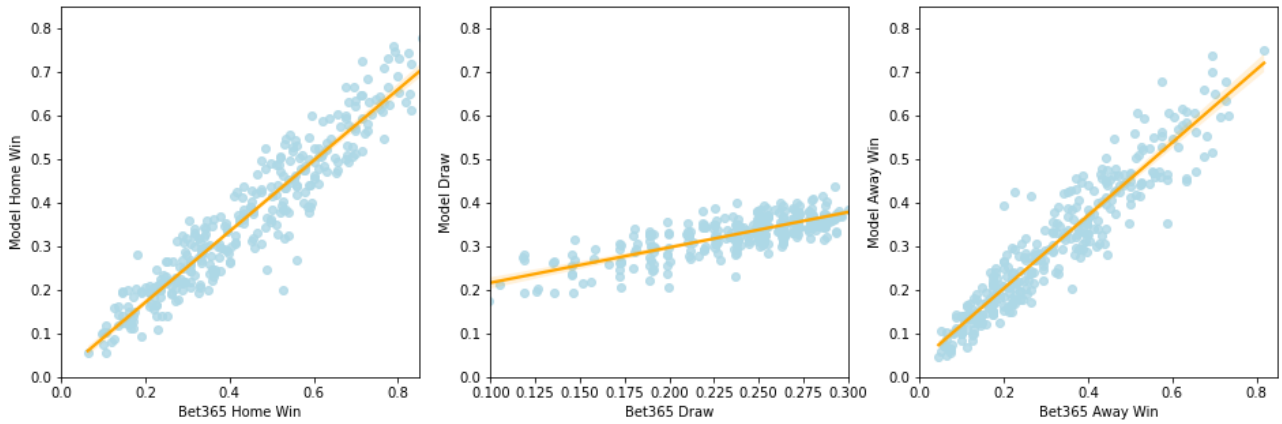

Figure 3: Linear Poisson Regression Model's Probability Estimates against Bet365's Probability Estimates

with the NLPR model lagging behind the former two after after correctly classifying 48.75% of the observations within the test set. The discrepancy in results is also quite marginal when the RPS metric is used. The bookmaker's predicted outcomes have a RPS of 0.1992, whilst the LPR model and NLPR model have a RPS of 0.2016 and 0.2062, respectively. In this regard, the best performing model is clearly the LPR, with the predictive performance of this model being practically indistinguishable from that of the bookmaker. A summary of these results may be found in Table 4 below.

Table 4: Summary Metrics for Models and Bet365

	Accuracy	RPS
Linear Poisson Regression	51.56%	0.2016
Non-Linear Poisson Regression	48.75%	0.2062
Bet365	51.25%	0.1992

The discrepancy in predictive performance between Bet365's forecasts and those of the LPR model is only 0.0024 according to the RPS metric. In fact, this difference in RPS is smaller than the best results-based model in Baboota and Kaur (2019), with their best performing model, the Gradient Boosting model, having a discrepancy of 0.0144 when compared to Bet365.

The similarity in results between the LPR model and Bet365 is illustrated in Figure 3, where the two sets of probabilities for each outcome of every match is shown. Overall there is a good agreement between the two probability assessments, with seemingly less variability being present here when compared to the model in Dixon and Coles (1997). The main differences lie with the LPR model having a tendency to give a higher probability of a match ending in a draw then the bookmaker, which in turn reduces the probability of a match going either way.

In conclusion, the LPR model is not only the simplest and most efficient model provided here, but it also has the best predictive power in estimating the probability of the match outcomes. Moving forward, the LPR model will be used as the basis for a betting strategy against the odds provided by Bet365.

5 Betting Strategy

In order to devise a strategy that beats the bookmaker's take and guarantee a profit, it is crucial to identify outcomes where the model probabilities are significantly higher than those obtained from the bookmaker's odds.

It is evident that the bookmaker significantly undervalues the probability of a draw occurring. At the same time, draws are typically not as likely to occur in football as much as either team winning the match. Hence, an ideal strategy would be one that makes use of double chance betting.

Essentially, double chance betting involves a bet on two possible outcomes occurring. In this case, a double chance bet could be placed on either a home win or a draw, a draw or an away win, or a home win or an away win. Only the former two will be considered given the bookmaker's handicap in predicting draws. Hence, although the potential profit is sacrificed, the purpose of placing a double chance bet is to minimise the risk of losing money by increasing the chance of success.

Unfortunately, the data that was available to us only consisted of the odds for a home win, a draw, or an away win. As such, the only alternative is to artificially create the odds for the double chance bets from the odds that are readily available.

Let $\hat{O}_{HD}$ and $\hat{O}_{AD}$ represent the artificial odds of a double chance bet of a home win or a draw, and an away win or a draw, respectively. Then, the double chance odds are created artificially as follows:

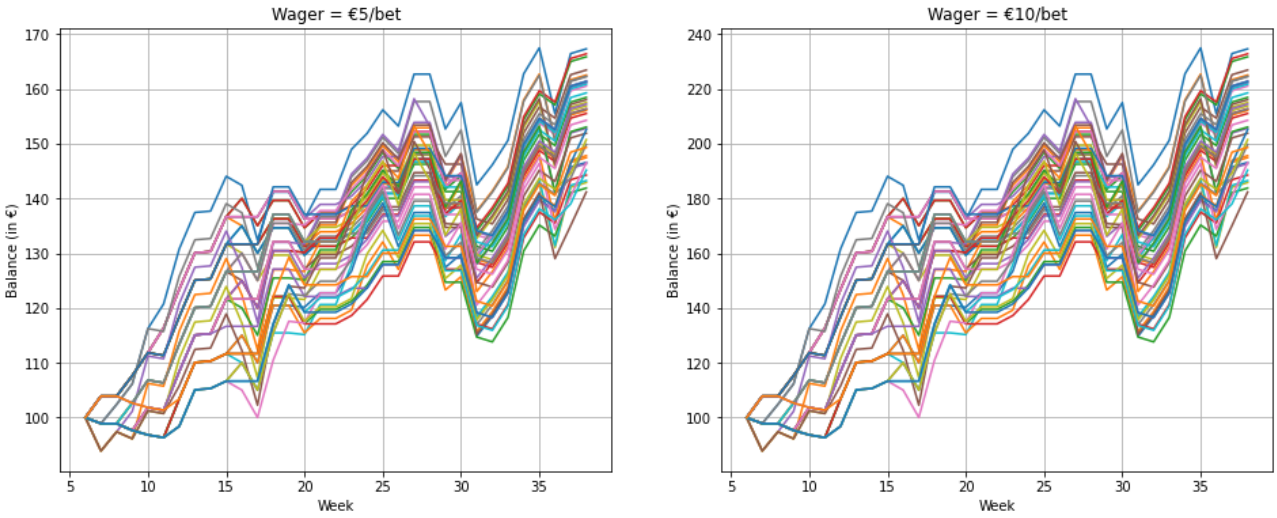

Figure 4: Betting Simulations from Week 6 to Week 38

$$\hat{O}_{iD} = \frac{1}{\frac{1}{O_i} + \frac{1}{O_D}} = \frac{1}{\pi_i + \pi_D} \ \forall i \in \{H, A\}$$

Furthermore, the normalised probabilities for the double chances, B_{HD} and B_{AD}, are simply computed in the following way:

$$B(iD) = B(i) + B(D) \ \forall i \in \{H, A\}$$

The strategy proposed here takes into consideration two main things, namely, the ratio between the model probability and the bookmaker's probability, and the risk tolerance that we are able to handle. Ideally, the ratio ought to be greater than 1 so as to represent an outcome that the model thinks is more likely to occur than the bookmaker does. The higher the ratio, the bigger of a bargain the bookmaker's odds will be, assuming that our model is the correct one. The risk tolerance assesses the probability that all of the money placed on a particular bet will be lost. Another important factor that needs to be taken into account is that in reality, only a single bet can be placed per match.

$$\text{Ratio} = \frac{P(i)}{B(i)} \ \forall i \in \{H, D, A, HD, AD\}$$

$$\text{Risk} = 1 - P(i) \ \forall i \in \{H, D, A, HD, AD\}$$

The betting strategy may be undertaken systematically by successively following the steps below on every game week. At every stage, if no bets are found, then no money is wagered.

1. Filter the bets by those that have Ratio $\geq p$.
2. Filter the bets by those that have Risk $\leq q$.
3. Since only one bet can placed on each match, filter the bets by the least risky outcome in each match.
4. Bet €x on each of the remaining bets.

An algorithm was designed in order to find the optimal values for p and q in the betting strategy. After running this algorithm a number of times, the resulting bands for these thresholds were found to be somewhere in the ranges of $p \in [1.17, 1.23]$ and $q \in [0.40, 0.45]$. Given an initial balance of €100, if €5 is wagered on every bet, the final balance would be in the range of [€143, €167]. On the other hand, given the same initial balance, if €10 is wagered on every bet, the final balance would be in the range of [€183, €235]. Figure 4 shows the progression of the balance by each game week, given wagers of €5 per bet and €10 per bet, respectively. The profits from this strategy highly depend on the thresholds p and q, as these parameter values ultimately select which bets are taken. Moreover, a caveat for the success of this betting strategy is that it was designed following the testing of the model accuracy.

In practice, the actual odds for the double chance bets would be a bit lower than the ones artificially created here, especially if the probability of that outcome is low. However, if the outcome is more likely to happen, the difference between the artificial and actual double chance odds would in fact be minimal. Hence, although the profits shown here would actually be lower in reality, since the strategy does not compound returns and the more likely outcomes are bet on, then this difference should be insignificant.

6 Conclusion

The bivariate Poisson model in Dixon and Coles (1997) is used as the foundation for the goals-based model put forward here, with the Poisson parameter represented in terms of a regression model similar to what is proposed in Ankomah et al. (2020). A number of original features are introduced, with these variables taking into account the difficulty of the match, the

scoring and defending form of the competing teams, and the quality of the players playing the game through the teams' FIFA ratings. These features helped to induce a surge of realism in our models, thus improving their accuracy in predicting the outcome of a football match, and ultimately faring better than other models suggested in literature. The LPR model, which is the simplest of the two models proposed here, had by far the best performance, mostly due its parsimony. In fact, the predictive performance of the LPR model is practically comparable to that of Bet365, as evidenced by both the prediction accuracy and the RPS. This is a remarkable feat, especially given the lack of resources available to us.

A profitable betting strategy based upon the LPR model is also proposed here. This strategy is a simple one, as it filters through both result bets and double chance bets, firstly by only considering those bets which have a greater value for the risk taken as per our model, and secondly by discarding any bets whose risk tolerance does not match our own. Nevertheless, the odds for the double chance bets were artificially created, making the actual profit from this strategy slightly inflated. Any future research should look into testing the profitability of this betting strategy by using the actual odds for double chance bets published by the bookmakers .

Unfortunately, given how the features are engineered, the model has a fatal flaw in that it is only able to predict games taking place from the 7th game week of the season onward. Furthermore, the NLPR model would have likely performed better if there were more features within the input layer. However, due to the limited availability that was experienced with regards to the data, we did not have enough resources to produce more variables. This resulted in the neural network model being underfit. It would be interesting to see how the introduction of an increased number of variables would affect the results of this non-linear model.

Another problem is that one of the main assumptions when using any type of Poisson regression is that of equidispersion, that is, the expectation and variance for the number of goals are assumed to be equal. Statistical distributions that relax this assumption ought to be implemented in the future in order to see if there is any improvement in the results. Although not shown here, we tested a linear model coupled with the Negative Binomial distribution. This distribution assumes overdispersion, wherby the variance is always greater than the expectation. Nevertheless, the resulting predictive performance actually ended up being worse than both the LPR model and the NLPR model.

A final suggestion for future research includes the implementation of the models and strategies put forward here, but instead using other time-frames or different football leagues. In this way, we would be able to know if the conclusions reached here would also hold in other scenarios.

References

1. Ankomah, R.K., Amoah, E.K. and Obeng, E.A. (2020). Predictve Modeling of Association Football Scores Using Bivariate Poisson. *American Journal of Mathematics and Statistics*, 10(3), pp.63-69. DOI: 10.5923/j.ajms.20201003.01.

2. Baboota, R. and Kaur, H. (2019). Predictive Analysis and Modelling Football Results using Machine Learning Approach for English Premier League. *International Journal of Forecasting*, 35(2), pp.741–755. DOI: 10.1016/j.ijforecast.2018.01.003.

3. Complete Sports (2021). Premier League Generated Over €68B Of Wagers, More Than Serie A And Bundesliga Combined. [online] Available at: `tinyurl.com/32vs2h2v` [Accessed 9 Jun. 2022].

4. Constantinou, A.C. and Fenton, N.E. (2012). Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models. *Journal of Quantitative Analysis in Sports*, 8(1), pp. 1-14. DOI: 10.1515/1559-0410.1418

5. Dixon, M.J. and Coles, S.G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2), pp.265–280. DOI: 10.2307/2986290

6. Egidi, L., Pauli, F. and Torelli, N. (2018). Combining Historical Data and Bookmakers' Odds in Modelling Football Scores. *Statistical Modelling*, 18(5-6), pp.436-459. DOI: 10.1177/1471082X18798414

7. Egidi, L. and Torelli, N. (2020). Comparing Goal-Based and Result-Based Approaches in Modelling Football Outcomes. *Social Indicators Research*, 156, pp.801-813. DOI: 10.1007/s11205-020-02293-z.

8. Forrest, D. and Simmons, R. (2000). Forecasting Sport: The Behaviour and Performance of Football Tipsters. *International Journal of Forecasting*, 16(3), pp.317–331. DOI: 10.1016/s0169-2070(00)00050-9.

9. Goddard, J. (2005). Regression Models for Forecasting Goals and Match Results in Association Football. *International Journal of Forecasting*, 21(2), pp.331–340. DOI: 10.1016/j.ijforecast.2004.08.002.

10. Karlis, D. and Ntzoufras, I. (2003). Analysis of Sports Data by using Bivariate Poisson Models. *Journal of the Royal Statistical Society: Series*

D (The Statistician), 52(3), pp.381-393. DOI: 10.1111/1467-9884.00366.

11. Karlis, D. and Ntzoufras, I. (2009). Bayesian Modelling of Football Outcomes: Using the Skellam's Distribution for Goal Difference. *IMA Journal of Management Mathematics*, 20(2), pp.135-145. DOI: 10.1093/imaman/dpn026.

12. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(1), pp. 1-52. DOI: 10.48550/arXiv.1603.06560

13. Schauberger, G. and Groll, A. (2018). Predicting Matches in International Football Tournaments with Random Forests. *Statistical Modelling*, 18(5-6), pp.460–482. DOI:10.1177/1471082x18799934.

14. Shin, H.S. (1991). Optimal Betting Odds Against Insider Traders. *The Economic Journal*, 101(408), pp.1179-1185. DOI: 10.2307/2234434.

15. Shin, H.S. (1993). Measuring the Incidence of Insider Trading in a Market for State-Contingent Claims. *The Economic Journal*, 103(420), pp.1141-1153. DOI: 10.2307/2234240.

16. Strumbelj, E. (2014). On Determining Probability Forecasts from Betting Odds. *International Journal of Forecasting*, 30(4), pp.934-943. DOI: 10.1016/j.ijforecast.2014.02.008.

17. Verri, D. (2022). What does football look like in the metaverse? *BBC Sport*. [online] 18 May. Available at: `https://www.bbc.com/sport/football/61473631` [Accessed 9 Jun. 2022].

18. Wheatcroft, E. (2020). A Profitable Model for Predicting Over/Under Market in Football. *International Journal of Forecasting*, 36(3), pp.916-932. DOI: 10.1016/j.ijforecast.2019.11.001

Acronyms

FIFA International Federation of Association Football

LPR Linear Poisson Regression

NLPR Non-Linear Poisson Regression

RPS Ranked Probability Score