Université Paris-Saclay

# Football Match Outcomes Prediction Using Machine Learning Approaches and Bookmakers Data

Ali Golmakani, Kuijun Liu

March 3, 2022

# Abstract

Machine learning has shown to be powerful in classification and regression problems. One of the applications of machine learning is to predict outcomes of sports events with historical data. This report aims to build mathematical models with machine learning algorithms to predict different football match outcomes suck as the match result and the number of goals by each team. In this report we have used the conventional classification algorithms to predict the result of the matches and we evaluate their performances using the bookmakers data of football betting market. Using these models, we can make comparison between official odds and our calculated probabilities of each outcome, and we have analysed the possibility of gaining a positive profit from the market. Furthermore, we also used other algorithms such as artificial neural networks and random forests in this context. For predicting the exact scoreline, we have used Poisson regression methods, which could model the number goals scored by each team. We have tested the performance of the different methods by calculating their accuracy and then we find the best method among our selection.


**Keywords**: machine learning, bivariate Poisson regression, classification

# List of Contents

# 1 Introduction

Several studies have been applied to the betting markets in order to scienti cally investigate the prices (odds) and also to  nd a con dent way to invest the money (betting strategies). Considering the weak, semi-strong, and strong forms of ef ciency of the markets (Fama (1970)), one can see that due to a competitive nature of the bookmakers market, they will satisfy on the weak form ef ciency criteria. After neutralizing the bookmakers commissions from the odd prices, we can derive the probabilities of the outcomes predicted by the bookmakers. Agreeing to the weak form ef ciency, Gross and Rebeggiani (2018) have recently performed the statistical tests in their review to measure the ef ciency of football betting markets.

The ef ciency of sports betting markets have been a popular subject for the researchers to work on. Before the emission of the online bookmakers, they used to have a so called  xed-odd betting option, which starts and ends at certain times with a  xed odd. This is the more traditional way of betting despite online markets who change their odds during the process. Vlastakis, Dotsis and Markellos (2009) believe the odds of  xed-odd betting markets contain less information and more risk comparing the odds of online markets right before the match starts. They mention despite  nding most of the online bookmakers ef cient, there are still some evidences of inef ciency in the market, for example arbitrage opportunities mostly in the  xed odds. This fact is con rmed also by Angelini and De Angelis (2019) and later by Elaad, Reade and Singleton (2020). They also mention that the ef ciency of betting markets has been improved signi cantly in the recent years.

On the other hand, there are evidences supporting the existence of many inef ciencies and biases. For example, Cain, Law and Peel (2000) and Dixon and Pope (2004) discussed about the existence of favorite-longshot bias in football betting markets. This type of bias  rst introduced in horse riding betting market, and con rmed by Dowie (1976), Ali (1977), Quandt (1986) and Shin (1993). It states that the punters tend to over-invest on the longshot odds, which means investing on favorite options should have more pro t. Vlastakis, Dotsis and Markellos (2009) and Angelini and De Angelis (2019) claim to  nd a pro table strategy using favorite-longshot bias and also away-favorite bias.

Other than these works around testing of the weak (and semi-strong) ef ciency of the betting markets by analyzing some empirical frequently reported biases, we will focus on the Machine Learning (ML) based approaches to measure the ef ciency. In this paper, we de ne formal notations and indicators for ef ciency in football bet-

ting markets, and we will analyse the strengths of bookmakers probability estimation by comparing their performances (measures) with each other and also with ML predictions. The ef ciency in betting markets is related to having a good accuracy in predicting the probabilities of the outcomes, we have multiple choices to measure ef- ciency, such as cross-entropy loss, and Brier score (Brier (1950)). There are more details about this subject in section 4.

In our analysis we have used a large data of football matches and odds to reduce the effect of variance as much as possible. We have applied the fundamental ML algorithms on this data, to t a model to estimate the probability of outcomes. For each football match we create a set of features gathered from the last 5 games of both teams to use it for training of our models. We use these features to predict the probabilities of different match outcomes, such as the result (home, draw, away) or number of goals of each team. The resources of the data is presented in section 3 and the details of our training processes have been provided more precisely in section 4, 5, and 6.

## 2 Related Works

Using conventional scoring measures for empirically evaluating the quality of predicted probabilities, is a popular way to compare the ef ciencies of different bookmakers. For instance, Štrumbelj and Šikonja (2010) uses the mean quadratic error to compare the ef ciency of the most popular bookmakers. And, Graham and Stott (2008) and Forrest, Goddard and Simmons (2005) used both mean quadratic error and maximum likelihood to evaluate their forecasting models.

Among the most common measures of probability predictions, there are two general types used in literature. One type called sensitive to distance which are kind of measure that depend on all predicted probabilities, therefore even the wrong predictions that are close to the actual outcome should be rewarded more that the distant predictions (e.g. Ranked Probability Score). The other type, are the ones who are calculated only from the predicted probability of the actual outcome that happened (e.g. cross-entropy loss). Constantinou and Fenton (2012) and Wheatcroft (2019) have summarized and compared the effectiveness of different scoring rules in UK football betting market. They argue that in a case of football results prediction, the scoring rule is better to be sensitive to distance. Hence, they claim RPS (Ranked Probability Score) is a robust measure in this context.

There are also a variety of works that examine the ef ciency of betting markets by trying to conduct a pro table betting strategy using statistical models. These statistical models use the historical data to t a prediction model on the match outcome, and outperforming the bookmakers predicted probabilities means contradiction of the ef ciency of market. Depending on the type of the model to forecast the real probabilities of the outcomes, we can have different algorithms. For example, Buraimo, Peel and Simmons (2013) claimed the inef ciency of the UK xed-odd betting market by showing a pro table betting strategy using Fink Tank predictions provided by University of Warwick between 2006 to 2012.

Studies on predicting different football match outcomes have been very popular. There are many approaches to train a forecasting model for a certain match outcome using the historical data. Machine Learning suggests variety of algorithms that can be applied to this eld. On the other hand, one can evaluate the ef ciency of bookmakers by having a comparison between the score of their predicted probabilities (implied from their odds) and the score of the probabilities predicted by regression models. Therefore regression can either result in supporting the ef ciency of the market (by

being outperformed by the odds), or in other case, it can result in providing a pro table betting strategy (by outperforming bookmakers odds).

One of the most popular approaches to estimate match outcomes are ordered and multinomial logit and probit models. The logit models, follow a logistic regression criteria in their loss or pro t function. Pope and Peel (1989) uses logit model to test the ef ciency of the betting market. Despite their conclusion on not being able to nd reasonable inef ciencies in the market, Vlastakis, Dotsis and Markellos (2009) uses the multinomial logit model to implement a pro table betting strategy.

On the other hand, the probit model also can be used to capture the probabilities of the match outcomes. Forrest, Goddard and Simmons (2005) used this algorithm to test the ef ciency of odds with a complete set of features for each match. Graham and Stott (2008) assigned an objective quality measure $_i$ to each team, and then used $_i$ $_j$ as an indicator for an ordered probit model to estimate the probability of match results between teams i and j .

Kuypers (2000) was able to implement a pro table strategy in xed odds market in England using an ordered probit model predictions. Goddard and Asimakopoulos (2004) also uses an ordered probit model to predict match results and test the weak form ef ciency of bookmakers odds. They also showed that the ordered probit model has almost similar performance whether it is based on the probabilities of the match result or exact scoreline. Extending this idea for exact scoreline estimation leads us towards Poisson regression models.

Generally, most scienti c predictions of football outcomes are centered around the match result probability prediction (Home, Draw, Away). However, there are also some works that focus on the other outcomes, like the total number of goals, difference in goals, and the exact scoreline. Reade, Singleton and Williams (2020) claim to nd a favorite-longshot bias in the exact scoreline odds, and they also provide a pro table betting strategy using it. Karlis and Tsiamyrtzis (2008) tries to t a Bayesian model to predict goal differences.

In the study of exact scoreline forecasting, there is a common idea that the number of goals of each team in a match has a Poisson distribution and hence a Poisson regression method to t into the historical data can be applied. This idea is inspired by assuming that the probability of scoring a goal for each team in each individual attack is xed value p. Therefore the total goals of a team should have a binomial distribution

with parameters $(n; p)$ where n is the number of attack opportunities of the team. And by using Poisson approximation for binomial variables we can conclude our initial fact. The rst statistical approaches using this idea appeared in Moroney (1952) and Reep, Pollard and Benjamin (1971).

However in the Poisson regression model, the dependencies of the two random variables (the number of goals of each team) has been analysed. Maher (1982) parameterized the information of the match by assessing an attack and defence score to each team. They assumed if $_i$ and $_i$ are indicators of attack and defence powers of a team. Then in the match between team i and team j comes from a bivariate poisson distribution $(X_{ij}; Y_{ij})$ where $X_{ij}$ and $Y_{ij}$ are poisson with parameters $(_i\ _j;\ _j\ _i)$. Dixon and Coles (1997) further investigated this approach by strengthen their assumptions and improving their results. They found evidence for the dependence of $X_{ij}$ and $Y_{ij}$ specially in low scoring games. Therefore they add a wight to the probabilities of the match outcomes where $\max(X_{ij}; Y_{ij})$ 1. They have used maximum likelihood approach to estimate the real probabilities. This idea can be seen in the later works such as Karlis and Ntzoufras (2003), Goddard and Asimakopoulos (2004) and Herbinet (2018).

Poisson models also are used in order to obtain a rankings from the team strengths. Ley, de Wiele and Eetvelde (2019) have compared the bivariate Poisson model to the Independent Poisson model and other models from the literature, and found that the bivariate Poisson model is the best one which assigns a strength ranking to the teams. Herbinet (2018) is a more recent work on predicting the result and scoreline of football matches. They use Dixon and Coles (1997) model as a benchmark to evaluate their results. They implement a dynamical ELO ranking system Elo (1986) for the both teams to use it for predicting the outcomes. This ranking system was initially developed for assessing the strength of chess players. Herbinet (2018) have applied more modern ML algorithms to this problem, such as Decision Trees, SVM (Support Vector Machines), and ANN (Arti cial Neural Networks).

Among the more frequent machine learning models and algorithms that have been applied in this context, we can name random forests. There are some comparisions between the performance of random forest and other conventional ML algorithms in Schauberger and Groll (2018) and they claim that random forest approach outperforms other regression models in predicting match results and scorelines. Later, Groll et al. (2019) extended random forest algorithm that have been used by Schauberger and Groll (2018) in FIFA World Cup data. They introduced the hybrid random forests

which somehow combines random forest with bivariate Poisson ranking method introduced by Karlis and Ntzoufras (2003).

As a summary and review for the performances of ML algorithms in football prediction, Joseph, Fenton and Neil (2006) have compared different methods such as KNN (k-nearest neighbors), Decision tree, and different versions of Bayesian networks. Later, Hucaljuk and Rakipović (2011) did the same think but they added a neural networks model to their comparision. They conclude that ANN has the best accuracy in predicting match results.

With the growth of ANN usage in the classi cation and regression problems, many researchers have been trying to use neural networks and deep learning approaches to model football match outcome. ANN have been applied to this context in several ways. The most common way to achieve a prediction model for match result, is by using multilayer perceptron. Arabzad et al. (2014) applied this method to Iran premier league and An lets et al. (2020) have also used multilayer perceptron with both L1 and L2 normalization. They have used a similar data to ours to predict English premier league result, they claimed to have a positive return on Bet365 odds using their model.

There are also more complicated neural networks and deep learning models that have been used in this context. For example Tsampazis and Tefas (2018) tried to use a modi ed version of autoencoders to construct a pro table model on the English premier league. Also the usage of recurrent neural networks is another way to model the performance of the teams and sometimes the players through the time. Jain, Tiwari and Sardar (2021) and Rahman (2020) used Long Short Term Memory (LSTM) modules to predict the match results. Lastly Cheng et al. (2003) combined the ANNs with the Learning Vector Quantization (LQV) method to estimate the general ratings of the teams and consecutively classify the match result.

# 3 Data and Resources

## 3.1 Datasets and Pre-processing

We have used the statistical information that is provided in https://www.football-data.co.uk/. Speci cally for the UK Premier League, we have used the historical data for all of the matches from 2010 to 2020. Since there are a lot of missing values in both match statistics and odds, we have restricted our attention only to the most fundamental statistics. For match statistics we use Number of Goals (Full time and Half time), Number of the Shots (On Target and Off Target), Number of Corners and Number of Fouls, for both Home and Away teams. And for odds statistics, we only use the odds of math results (Home, Draw, Away), and the odds for having more than 2.5 goals (>2.5 and <2.5). In this section, we are going to explain our methodology to analyse each odd.

At rst, we should de ne and build our training data to be able to apply the ML algorithms on it. For each match in the data, we can de ne different types of features. The rst type, is just the features that are related to the Home and Away performances in a recent window. For each team, we have put into our training data the averages of the match statistics of that team from their previous 5 matches. Forrest, Goddard and Simmons (2005) have also used the features gathered from recent matches in their analysis. They also used the league standings information of the teams which we haven't considered in our work. Another type of features, is the features that depend on the match but not on the teams, like weather and time. For simplicity issues, we also dismissed this kind of features in our data.

For each match between home (H) and away (A) team, at the rst step we extract the recent 5 matches of both H and A, and then compute the average number of the goals (FG), goals in the rst half (HG), shots (S), shots on target (ST), corners (C) and Fouls (F). All of this 6 features that we mentioned will be also computed for the opponents of our teams in the recent matches, in this case we put an additional C and the end of the feature name. For example, HFGC means the average number of the goals (FG) that the home team (H) conceded (C) in their recent 5 games.

We expect these features to have a Poisson distribution, also we expect the number of goals to have a Poisson distribution. In (Figure 1) you can see the total frequencies of each score for home and away and in (Figure 2) you can see the total distribution of scorelines in the train data. The effect of playing in the home eld is observed. you can see the histograms, pairwise scatter plots, and correlations of these 6 features in (Figure 3).

After applying these steps, we can successfully split our data to training, validation and test. For the training data we have 3104 samples, and for the validation and test data we have 1035 samples, and the data consists of 24 features (normalization applied).

Another idea that has been used for pre processing the features, is assessing a weight for the performance features of the teams in the past matches which decays exponentially with being more distant to the current match in time. (e.g. Dixon and Coles (1997), Karlis and Ntzoufras (2010))

Figure 1: Histogram of the number of goals for home and away. Horizontal lines are their corresponding mean value

Figure 2: scoreline distribution

# 4 Football Betting Market Analysis

As mentioned in the previous sections, there seem to be too many works on analyzing the ef ciency of the betting markets. In case of sports betting, we usually deal with several types of odds by the bookmakers which provides a market for the gamblers to use various strategies for investment. In this section, we will try to analyse the simplest type of odds where you can only gamble on an event, and depending the outcome of the event, either the gambler wins or loses the bet.

Measuring the degree of ef ciency in betting markets can be analysed in different ways. The bookmakers ef ciency can be measured by comparing their ef ciency indicators estimated from their empirical performances. Since providing the odds is related to forecasting the probabilities of the outcomes, so in the context of forecast veri cation, we can evaluate the quality of the predicted probabilities by looking at their scores. There are many ways to de ne a score for probability forecasts that satisfy in the strictly proper scoring rules criteria Murphy (1969).

There are many ways for the bookmakers to provide their odds. Of ine bookmakers odds ( xed odds) use a xed value for their odd price to announce at the start time of the market. Despite them, online bookmakers have a dynamical odd price

Figure 3: Pairwise plot of team features

which can be changed during the market time. They can apply the new information to their prices at any time, and using pari-mutuel system is very common among them. Bookmakers using pari-mutuel system may provide inef cient odds to increase their expected pro t.

An of ine bookmaker xes the odd price at some point, and then the punters who think this odd has a positive expected pro t, invest on their offer. Let $o > 1$ be the odd price provided by a bookmaker, then if the desired outcome happens, the investors money will be multiplied by $o$, otherwise they lose their money.

We can de ne $p = o^{1}$ to be the bookmaker's predicted probability of an event. If the event happens with probability $p$, then the expected pro t of the bookmaker is $1 \quad \frac{p}{p}$ (considering a unit amount of investments). Since the bookmakers probably have a positive pro t in average, it helps us to get an intuition that $p$ should be slightly greater than $p$. But we should also note that, the higher values of $p$ (equally smaller odd prices) results in having less demands from the punters because of the competitiveness of the bookmakers market. On the other hand, smaller values of $p$ reduces

the pro t of the bookmakers. So $p$ can be interpreted as a forecast of $p$ computed by the bookmaker including commission.

Let be the empirical expected pro t of a bookmaker (with a unit amount of invest in each provided bet) $= 1 \quad E[\frac{y_i}{p}]$, where $y$ is our observations from the match outcomes, in this case, it is a random variable which takes the values in $f0; 1g$. If we assume that this commission is linearly distributed among probabilities, We can deduce that $(1 \quad )p$ is the true probability of the bet predicted by the bookmaker. Being less than this value means less pro t for the bookmaker and being more than this value means less demands in competitive market for the bookmaker. Angelini and De Angelis (2019) also support the fact that the true probability is a linear function of the bookmaker probability by applying Ramsey RESET tests and detecting signi cant non-linear dependencies.

Also in this context, is related to the de nition of reliability and calibration of probabilistic predictions. The reliability diagram is a function that takes a value in $[0; 1]$ as input, and lters the forecasts in which the predicted probability is near the input value, then it assigns to it the frequency of the occurrences of target values in the ltered data. Let $f(x) = P(y_i = 1 j p = x)$ be the reliability diagram of $p$ (Hartmann et al. 2002), then the expected pro t of $p$ can be computed as $= E_{q \ D}[1 \quad \frac{f(q)}{q}]$ where $D$ is our historical observations of $p$ and its corresponding relative frequency $f(p)$. By supposing that the bookmakers adjust their commissions linearly through the unit interval, then we can approximate $f(x) = (1 \quad )x$.

In our rst step, we empirically estimated the expected pro t (commission) of bookmakers to compute their true implied predicted probabilities of the outcomes by subtracting their commission from their pro t through scaling their odds $\hat{p} = (1 \quad )p$. (Figure 4) shows the reliability plots of $p$ and $\hat{p}$. Now we can compare their forecasting accuracy measures with each other and also with ML estimated probabilities.

## 4.1 Classi cation Analysis

In this section, we are going to train classi cation models to predict the real probabilities of the outcomes. One of the common ways to measure the probability forecasting quality, is using the Brier score which in our case (uni dimensional), it's the same as mean quadratic error between the observations and the predictions $BS = E[(y_i \quad \hat{p})^2]$. This score has been introduced by Brier (1950) for evaluating weather forecasting models, but it became one of the popular ways to evaluate betting market predictive

Figure 4: Reliability diagram of the probabilities implied from the bookmakers odds. BbMx: probability implied from the maximum odd of the market. BbAv: probability implied from the average odd of the market. CEP: Conditional Event Probability

performance. Murphy (1973) introduced a decomposition formula for the Brier score to be the sum of calibration (reliability) and re nement. $BS = CAL + REF$ where calibration is computed as $CAL = E_{q\ D}[(q \quad f(q))^2]$ and re nement is $REF = E_{q\ D}[f(q)(1 \quad f(q))]$. Since re nement is invariant to the linear transformations of our prior distribution of $q$, therefore converting $p$ to $\hat{p}$ by multiplying $(1 \quad)$ doesn't change this term and only affects reliability. Re nement is also related to the area below the diagram of the ROC curve.

On the other hand, we can also use Maximum Likelihood method to measure the predictions. It leads us towards a logistic regression model in this context (sometimes referred to as the ignorance score). If $\hat{p}$ is the probabilities estimated by the model and $y$ is the vector of the results coming from Bernouli variables with parameter vector $p$ then by nding $\max_{\hat{p}} P(yjp = \hat{p})$ we will have an estimation of the real parameter $p$.

The likelihood value for y and $\hat{p}$ can be computed as follows:

$$L(y;\hat{p}) = P(y|p = \hat{p}) = \prod_i (y_i \hat{p}_i + (1 - y_i)(1 - \hat{p}_i))$$

Therefore, after taking the logarithm, maximizing the log-likelihood is equal to minimizing the cross-entropy loss between $\hat{p}$ and y, which is linearly equivalent to minimizing the Kullback–Leibler divergence of the predicted and real distribution.

Other than the mentioned measures, there are other measures that have been used in evaluating the (probability) forecasting performances for example, AUC, F1 score, and skill scores .

Now we are ready to train our models. We have used linear logistic regression, random forests, and multilayer perceptron neural networks to train our probability estimation models on our training data to be able to challenge the ef ciency of betting market or nd a pro table model.

## 4.2 Fitting Models and Probabilistic Inferences

In this section we will examine the performances of different ML algorithms on predicting the match outcomes. After applying the different algorithms to achieve a prediction function, we can compare them by their quality measures. We have focused on the average and maximum odd of BetBrain for having more than 2.5 goals in total. The empirical expected pro ts of the average and maximum odds are 6:12% and 1:25% respectively. You can see their reliability plot in (Figure 4). To measure the quality, we can observe the Brier score (containing Calibration and Re nement), and the logistic regression loss (negative of likelihood).

For our models, we have used the validation set to tune our model hyperparameters. For the random forest, we found the optimum number of trees and the maximum nodes in each tree as our hyperparameters to using the validation data. Also for neural network, the size of hidden layer is estimated from the performances on the validation data.

After the training process, we can look at the posterior distributions of our models. In probability estimation by random forests, there is a common step after training the model which is changing the posterior distribution of the model by a speci c map. It is referred to as the calibration process for probability estimation. In other words, it

means transforming classi er scores into class probabilities. We have used the Isotonic Regression method which estimates this transformation ma with partially linear maps Zadrozny and Elkan (2002).

Now we are ready to compare the measures of our models with each other and also with the bookmakers, in (Figure 8) you can see the pairwise plot of our model predictions. Notice the high correlation between ANN and GLM. (Figure 7) is the plot of train and validation loss during the ANN training. And (Table 1) shows the performance of our models on predicting more than 2.5 goals in the match.

| single class classi cation | | | | |
|---|---|---|---|---|
| | negative log-likelihood | Brier | reliability | re nement |
| BbMx | 0.6833679 | 0.2453246 | 0.0005 | 0.2448382 |
| BbAv | 0.6852494 | 0.246221 | 0.0016 | 0.2446663 |
| BbAvC | 0.6833051 | 0.2452842 | 0.0003 | 0.2449532 |
| LR | 0.6894725 | 0.2481594 | 0.0004 | 0.2477197 |
| RFC | 0.6897751 | 0.2483468 | 0.0002 | 0.2481223 |
| NN | 0.6893291 | 0.2480245 | 0.0008 | 0.2472735 |

Table 1: Single Class Models Evaluation. BbMx: Probabilities implied from the maximum odd. BbAv: Probabilities implied from the average odd. BbAvC: Probabilities implied from the average odd with calibration. LR: Linear Regression. RFC: Random forest with Calibration. NN: Neural Network

## 4.3 Generalization to Multiple Classes

As we mentioned in section 2, there are two common measures of probability predictions: Ranked Probability Score(RPS) and cross-entropy loss. RPS was proposed by Epstein (1969)as a scoring system for probability forecasts of weather states, which is a useful proper scoring rule for ranked categories. The original RPS was formulated with the framework of $K$ possible states and vector $r = (r_1; :::; r_K)$ be a probability forecast of an ordered variable. The RPS when class $j$ occurs, $RPS_j(r)$ is

$$RPS_j(r) = \frac{3}{2} \quad \frac{1}{2(K \quad 1)} \sum_{i=1}^{K \quad 1} [(\sum_{n=1}^{i} r_n)^2 + (\sum_{n=i+1}^{K} r_n)^2] \quad \frac{1}{K \quad 1} \sum_{i=1}^{K} ji \quad j jr_i;$$

when $K = 2$, RPS reduces to Probability Score(PS) or Brier Score(Brier (1950) ). The RPS for overall forecasts is de ned by

$$RPS = \frac{1}{K \quad 1} \sum_{i=1}^{K} \sum_{j=1}^{i} r_j \quad \sum_{j=1}^{i} e_j \quad !^2;$$

Figure 5: Random Forest Feature Importance

Figure 6: Random Forest Hyperparameters

Figure 7: Loss of ANN during training. Left: Single Class. Right: Multi Class.

Figure 8: Pairwise plot of the models predictions. BbMx: probability implied from the maximum odd of the market. lr: Logistic Regression. rfc: Random Forest with Calibration. nn: Neural Network

where $r_j$ is the forecasted probability of outcome $j$ and $e_j$ is the actual probability of outcome $j$. According to the de nition of $RPS_j(r)$, it is not hard to see such a scoring system does not only depend on real outcome but also depend on predicted probabilities. According to Epstein (1969), RPS can be viewed as a measure of 'distance' between different categories. Murphy (1970) introduced two de nitions of distance based on RPS and PS respectively. Meteorologists have indicated the for the need

for a scoring rule based upon considerations of distance that are particularly relevant for forecasts of ordered variables such as ceiling height, temperature and football scores. RPS is sensitive to distance, which implies that it can be robust in measuring probability forecast.

Another useful measure of forecast is cross-entropy loss. It measures the performance of a classi cation model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label. Suppose we have discrete real distribution p and predicted probability distribution q with the same support X , the cross-entropy is de ned by

$$H(p; q) = \sum_{x \in X} p(x) \log q(x):$$

Compared with RPS, cross-entropy loss depends only on the true outcome, it does not concern about 'distance' between different prediction categories. In (Table 2) these scoring measures have been shown for both bookmakers probabilities and the different models that we trained on the data.

To evaluate the quality of the bookmakers predictions, we should nd the true probabilities implied from their odds. What we suggested in the previous section, was to divide the inverse odds to the empirical expected pro t of the bookmakers to have their real predicted probabilities. But here we can use other algorithms because in this case we are sure that one of the output classes is going to happen for sure. In other words let $s = \sum 1=q$ be the sum of inverse odds of the output classes. Then we can conclude that $s > 1$ and here the expected empirical pro t of the bookmakers can be derived from $s - 1$. This provides a natural way to estimate the real predictions of the bookmakers implied from their odds. This method supposes that the bookmakers apply their commissions equally among the outputs. We call this method the basic way of extracting the probabilities.

Forrest, Goddard and Simmons (2005) and Gross and Rebeggiani (2018) believe commission is uniformly distributed among different outcome probabilities. Shin (1993) on the other hand discusses about the effect of insider traders in betting markets and bring a new formula for calculating the outcome probabilities implied by odds provided by bookmakers. Štrumbelj (2014) claims that Shin's formula has the best performance to predict the probabilities by comparing the Brier scores of basic uniformly scaled probabilities and Shin's probabilities implied from the bookmakers odds.

| multiclass classi cation score | | |
|---|---|---|
| | negative log-likelihood | Ranked Probability Score |
| B365B | 1.011027 | 0.2065123 |
| B365S | 1.010823 | 0.2064822 |
| WHB | 1.010913 | 0.2065046 |
| WHS | 1.010818 | 0.206444 |
| LR | 1.038011 | 0.2149896 |
| RFC | 1.054481 | 0.2196784 |
| NN | 1.037853 | 0.214916 |

Table 2

## 4.4 Betting Strategy

One of the ways to testify the ef ciency of the market is by trying to nd a pro table betting strategy, because by achieving it, we can conclude the existence of information that are not included in the prices. The eld of providing and analysing trading strategies in nancial markets is vast. Similar to stock markets, betting markets also provide several investment options for the punters, therefore there are many different ways to implement a strategy for investment. In this section we will analyse the effectiveness of our classi cation models on nding a betting strategy.

There have been many successful and unsuccessful attempts to nd a pro table strategy in football betting market. On the other hand, like stock markets, we can expect the behavior of market to vary over time which means the pro tability of a certain strategy also depends on the different time periods. For example, we can nd evidences supporting the existence of positive pro t betting strategies based on statistical biases in the bookmakers prices, such as favorite-longshot bias and etc.

Among the highly reported biases, there is also best-odd bias which uses the strategy of investing on the best odd for each match whenever the variance of the odds of different bookmakers is too much. Forrest, Goddard and Simmons (2005) claim that the commission is virtually eliminated by using this strategy. Deschamps and Gergaud (2007) also found evidence for best-odd bias. They also report a draw-bias in the sense that betting at draw odds yields a higher return than betting at home or away odds in football betting markets.

Over the time, more complicated approaches on nding a pro table betting strategy are being tested. Using machine learning models, is a common way to nd patterns in historical data to help us have a more accurate prediction of the outcomes. The power of this method highly relies on the quality and size of the historical data that we

want to train our models on.

Note that the pro tability of a prediction model is not necessarily related to having more or less accuracy comparing to the bookmakers odds in the outcome prediction. It means different probability prediction models, can capture different information from the historical data to help them to estimate the probability of the outcomes. For example in (Figure 8) you can notice that the correlation between RF and ANN predictions are less than the correlation between GLM and ANN predictions, which means there are more samples that RF and ANN prediction do not totally agree to each other. However we saw that the accuracy of ANN is better than RF in prediction, but it doesn't mean that RF is not pro table against ANN. We have analysed the cases for having more than 2.5 goals when RF assigns 5% more probability comparing to ANN in the validation and test data. In this case if the actual odd were set by the ANN probabilities, then we could have 578 bets with average pro t of 2:57% per bet in the validation data and 724 bets with average 2:49% pro t in test data. On the other side, if we the odds were set by RF probabilities, and we invest on the bets where ANN predicts 5% more chance comparing to RF, we will have 685 bets with average 7:36% pro t in the validation data and 512 bets with 4:15% average pro t.

Using our models we can implement a betting strategy by a simple common algorithm that whenever our predicted probability divided by the bookmakers probability is more than a threshold for any bet, we invest a unit amount of money in that bet. In (Figure 11) and (Figure 10) you can see our pro t in the validation and test data respectively, using this strategy for different values of the ratio.

From these gures we can conclude that it is not possible to conduct a betting strategy with a reliable positive return using our models. It was expected to be like this since we only consider limited amount of features to estimate our probability, and in reality there are too many conditions that may have effect on the posterior probability distributions of the match outcomes.

Figure 9: Reliability plots of our models. LR: Logistic Regression. RFC: Random Forest with Calibration. NN: Neural Network

Figure 10: Average pro t of the models in the Test set

# 5 Score Prediction and Poisson Regression

In this section we are focusing on the exact scoreline prediction using our covariates (features) of each match. It is intuitive to consider that the number of goals has a Poisson distribution. As we mentioned earlier, a way to model the probabilities of

Figure 11: Average pro t of the models in the Validation set

scoring is assuming that there is a xed probability for each team to score in a time interval with xed lengths. It implies that the $Y_1$ $Poisson(\ _1)$ and consecutively $Y_2$ $Poisson(\ _2)$. The simplest way is to predict $_1$ and $_2$ by generalized linear model. Since we know they are positive, we can use the logarithm as the link function. In the following section we will further investigate this method.

On the contrary, the dependence of $Y_1$ and $Y_2$ has always been discussed in the literature. The previous setting that we introduced, supposes that they are independent. However, by observing that the marginal distributions are poisson, the natural way to model $Y_1$ and $Y_2$ is by using bivariate poisson variables. Together with $_1$ and $_2$ that we introduces, there is another parameter $_3$ in bivariate poisson model which somehow controls the correlation between the target variables. Karlis and Ntzoufras (2003) used bivariate poisson regression to model the number of goals. They used the same generalized linear method to estimate $_1$ and $_2$, but for $_3$ they used several cases to model it. In the rst and second case they analyzed $_3 = 0$ and $_3 > 0$, where in the third case they used generalized linear estimation $_3 = \log(X\ _3)$. Also a natural way to model $_3$ is by using the general match day features instead of home and away features.

Karlis and Ntzoufras (2003) also introduced another method for robust modeling of the number of goals. Originally, Dixon and Coles (1997) observed that the pure Poisson distribution may not be the most accurate way to model the number of goals. They used the Poisson regression method introduced by Maher (1982) to statistically show that the match result tend to be nished with a draw more than what the model predict. Especially in low scoring games, hey showed that the double Poisson model underpredict 0-0 and 1-1 scores and overpredict 0-1 and 1-0 scores. Therefore, they used a method called In ation to force the posterior distribution to be slightly different. Karlis and Ntzoufras (2003) extended their idea and introduced another in ation methods in this context. They used an additional Poisson and Geometric distributions on the diagonal of the joint distribution to x the underestimation of draw outcomes.

## 5.1 Generalized Linear Model

A generalized linear model is a generation of ordinary linear model. The ordinary linear regression has following form:

$$Y = X \quad + \tag{1}$$

, where

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}; X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}; \quad = \begin{bmatrix} 1 \\ \vdots \\ p \end{bmatrix}; \quad = \begin{bmatrix} 1 \\ \vdots \\ n \end{bmatrix}$$

It is assumed that $Y_i$ has normal distribution. The error terms $_i$ N (0; $^2$) and $_i$ and $_j$ are independent for i ~~6~~ j . The most common estimator for standard linear regression model is the least square estimator which is

$$\hat{} = (X^T X)^{-1} X^T Y \tag{2}$$

For the generalized linear model, the dependent variable $Y_i$ does not have to be normally distributed but allow it to be a member of exponential family. The GLM is developed by Nelder and Wedderburn (1972) as a way generalized some statistical models relating to four distributions: normal, binomial, Poisson and gamma. A GLM consists of 3 elements: an exponential family of probability distributions with probability density function f (Y ; ), a linear predictor = X and a link function g such that E(Y jX ) = = g $^1$( ), where the general form of f (Y ; ) can be expressed as

$$f(Y ; ) = h(Y ) \exp(b( ) \quad T(Y) \quad A(Y)) \tag{3}$$

where $h; b; T$ and $A$ are known functions.

In the previous section, we used the logistic version of the generalized linear model to predict the match result and the probability for having more than 2.5 goals. Using the cross entropy loss function we and setting the link function $g = \log(\frac{x}{1\ x}) = \ ^{1}$, where is the softmax function. Using the fact that $@_j =@_i x = \ _i(1_{i=j} \ _j)$ we can calculate the derivatives of the loss function.

$$\hat{p} = \ (X \ ) =) \ \frac{@\ell}{@} = \frac{@\ell}{@} \ ^{0}(X \ ) \tag{4}$$

Now we can use gradient based methods to optimize the parameters. Note that minimizing this loss value in this case is the same as maximizing the likelihood of our observed data with our parameters. For a probability regression, the logistic loss function seems to have the best performance for training models. We can alternatively use a quadratic loss function but it has worse accuracy in multiclass classi cation.

## 5.2 Poisson Regression Model

The Poisson regression model assumes that dependent variable $Y_i$ follows Poisson distribution. Recall that if a random variable $X$ follows Poisson distribution, then its probability mass function is given by

$$f(X = k; \ ) = P(X = k) = \frac{^{k}e}{k!}$$

Compared with equation 4, we know that Poisson distribution belongs to the exponential family

$$f(X = k; \ ) = \exp(k \log( \ ) \quad \log(k!))$$

and we can derive a generalized linear model for Poisson regression. The link function for Poisson regression is $g(x) = \log( x)$, thus the model can be expressed as

$$E(Y \ jX \ ) = \ = \exp(X \ )$$

## 5.3 Parameters Estimation

In this section, we discuss how to estimate in the Poisson regression model.

Consider $Y = [Y_1; \ ; Y_n]^T$, where $Y_i$ are independent random variables with $E[Y_i] = \ _i$, we can estimate by maximum likelihood. With the probability mass

function f (X = k; ), our log-likelihood function for for each $Y_i$ is

$$l_i( _i; y_i) = y_i \log( _i) \quad _i \quad \log(y_i!):$$

In a compact form,

$$l( ; y) = \sum_{i=1}^{N} y_i \log( _i) \quad \sum_{i=1}^{N} _i \quad \sum_{i=1}^{N} \log(y_i!):$$

The method of scoring provides a way to obtain the maximum likelihood estimates for $_j$, at the rst step, the score function is de ned by

$$s_j = \frac{@l}{@_j} = \sum_{i=1}^{N} (\frac{@_i l}{@_i}\frac{@_i}{@_j});$$

where

$$\frac{@_i l}{@_i} = \frac{y_i}{_i} \quad 1$$

and

$$\frac{@_i}{@_j} = \frac{@_i}{@_i}\frac{@_i}{@_j} = \frac{@_i}{@_i} x_{ij}$$

with = X .

We denote the observed information matrix w.r.t. as J ( ), which is the covariance matrix of score function s

$$\begin{aligned} J( )_{jk} &= E(s_j s_k) \\ &= \sum_{i=1}^{N} \frac{E((y_i \quad _i)^2) x_{ij} x_{jk}}{^2_i} (\frac{@_i}{@_i})^2: \end{aligned}$$

Let W be an N N diagonal matrix with elements $w_{ii} = \frac{1}{_i}(\frac{@_i}{@_i})^2$, the information matrix can be written as

$$J( ) = X^{>} W X$$

Then the m + 1-th iteration of is given by

$$^{m+1} = \quad ^m + J \quad ^1( ^m)s( ^m)$$

## 5.4 Kolmogorov-Smirnov Test

Before applying Poisson regression to our data set, we use Kolmogorov-Smirnov (KS) test (Massey Jr (1951)) to select features that are from Poisson distributions. K-S test is a non-parametric test that can be used to compare a sample with a known

probability distribution. In this case, we evaluate the goodness of t between a sample and the Poisson distribution. We de ne the empirical distribution function $F_n$ of n i.i.d. observations $X_i$ as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[1 \quad ;x]}(X_i);$$

where $1_{[1 \quad ;x]}$ is the indicator function. $F_n(x)$ gives the fraction of sample points less than x. Suppose we have some cumulative function $F_0(x)$, then the K-S statistic for $F_0(x)$ is given by

$$D_n = \sup_x |F_n(x) \quad F_0(x)|$$

By the fundamental theorem of statistics, if $X_i$ follows Poisson distribution with cumulative function $F_0$, then $D_n$ converges to 0 almost surely.

However, in practice, the parameter of cumulative function is unknown. In this case, we have to estimate the unknown parameter which is the mean for Poisson distribution from the sample. By Campbell and Oprian (1979), we have a table obtained from a Monte-Carlo calculation to verify whether a set of observation follows Poisson distribution. The rate parameter of the Poisson distribution is estimated by the mean value of the sample. If D is larger than the critical value provided in (Campbell and Oprian (1979)), we reject the hypothesis that the sample follows Poisson distribution.

## 5.5 Bivariate Poisson Models and In ations

We have seen in the previous section that the marginal distributions of number of the goals scored by the teams, can be approximated by two Poisson distributions. In other words, for a certain match, if we denote the number of goals of home and away by $Y_1$ and $Y_2$ respectively, then we know they have distributions $Poi(\ _1)$ and $Poi(\ _2)$, therefore, we can estimate $\ _1$ and $\ _2$ using linear functions from our features.

In previous setting, we assume that the target variables $Y_1$ and $Y_2$ are independent. However, the independence assumption may not hold. For example, When the difference $|Y_1 \quad Y_1|$ is large, one of the teams may have a lower willingness to attack or defense. In this case, it is natural to consider a bivariate Poisson model which drops the independence assumption. It is suggested by Karlis and Ntzoufras (2003) that a bivariate Poisson model improves the prediction of the number of goals in football games.

Figure 12: Difference of actual scoreline distribution and cumulative distribution of the predicted probabilities of the scorelines using Poisson regression

Figure 13: Difference of actual scoreline distribution and cumulative distribution of the predicted probabilities of the scorelines using Poisson regression

### 5.5.1 The bivariate Poisson distribution

There are different approaches to construct bivariate Poisson distribution discussed by Kocherlakota and Kocherlakota (2017). A common way discussed by Johnson (1997) is the trivariate reduction method. Consider random variables $X_k; k = 1; 2; 3$ which are independent Poisson random variables with non-negative parameters $_i; i = 1; 2; 3$. Then we de ne random variables $Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$, which follow a bivariate Poisson distribution with parameters $_i$ denoted by $BP( _1; _2; _3)$.

Figure 14: Difference of actual scoreline distribution and cumulative distribution of the predicted probabilities of the scorelines using Poisson regression

The joint probability mass function is

$$f_{JBP}(y_1; y_2; \ _1; \ _2; \ _3) = e^{(\ _1 + \ _2 + \ _3)} \frac{\ _1^{y_1}}{y_1!} \frac{\ _2^{y_2}}{y_2!} \sum_{r=0}^{\min(y_1; y_2)} \binom{y_1}{r} \binom{y_2}{r} r! \left(\frac{\ _3}{\ _1 \ _2}\right)^r :$$

$Y_1$ and $Y_2$ follows Poisson distribution marginally with expectations $E(Y_1) = \ _1 + \ _3$ and $E(Y_2) = \ _2 + \ _3$. The covariance of $Y_1$ and $Y_3$ is $\ _3$. If $\ _3 = 0$, then the two random variables are independent which is the case of double-Poisson distribution.

### 5.5.2 Parameters Estimation

In the joint bivariate Poisson regression model, we assume that the parameters $\ _i; i = 1; 2; 3$ depend on explanatory variables. Using the exponential link function, the joint bivariate Poisson regression can written as

$$\log \ _{k;i} = X_i \ _k; k = 1; 2; 3;$$

where $i$ denotes the observation number, $X_i$ denotes the vector of explanatory variables for the $i$ th observation and $\ _k$ is the vector of coef cients related to the $k$ th parameter.

Jung and Winkelmann (1993) and Kocherlakota and Kocherlakota (2001) use Newton-Raphson method to estimate the parameter. Karlis and Ntzoufras (2005) developed an EM algorithm to deal with convergence problem with Newton-Raphson method. Karlis and Tsiamyrtzis (2008) present a Bayesian estimation approach for the param-

eters of the bivariate Poisson model and provide the posterior distributions in closed forms.

Since for canonical exponential families the method of scoring and the method of Newton–Raphson coincide, we can still use method of scoring to maximize likelihood. In this case, for each observation $(y_1i; y_2i)$, the joint biavariate Poisson distribution

$$f(y_{1i}; y_{2i}) = e^{(\ _1;\ _2;\ _3)} g(y_{1i}; y_{2i});$$

where

$$g(y_{1i}; y_{2i}) = \sum_{r=0}^{\min(y_{1i}; y_{2i})} \frac{\ r_{3i}\ \ y_{1i}\ \ r\ \ y_{2i}\ \ r}{r!(y_{1i}\quad r)!(y_{2i}\quad r)!}:$$

The corresponding score functions are $U_{k;j} = \frac{@l}{@\ _{kj}}; k = 1; 2; 3.$ More explicitly, they can be written as

$$U_{1;j} = \sum_{i=1}^{n} \ _{1i} X_{ji} \quad \frac{g(y_{1i}\quad 1; y_{2i})}{g(y_{1i}; y_{2i})} \quad 1 \quad ;$$

$$U_{2;j} = \sum_{i=1}^{n} \ _{2i} X_{ji} \quad \frac{g(y_{1i}; y_{2i}\quad 1)}{g(y_{1i}; y_{2i})} \quad 1 \quad ;$$

$$U_{3;j} = \sum_{i=1}^{n} \ _{3i} X_{ji} \quad \frac{g(y_{1i}\quad 1; y_{2i}\quad 1)}{g(y_{1i}; y_{2i})} \quad 1 \quad :$$

The whole process is exactly the same as the case of calculation in independent Poisson regression model. A better estimation is using EM algorithm. To construct EM algorithm, we de ne 3 latent variables $Z_1; Z_2; Z_3$ which is of Poisson distribution with parameters $\ _1;\ \ _2$ and $\ _3$ respectively. We assume that $Y_1 = Z_1 + Z_3$ and $Y_2 = Z_2 + Z_3$. Now the complete data is $(Z_1; Z_2; Z_3)$ but only $(Y_1 = Z_1 + Z_3; Y_2 = Z_2 + Z_3)$ are observed. The complete data log-likelihood is $l(\ _i)$ and marginal log-likelihood $l_{Y_i}; i = 1; 2$ is based on observations only. The EM algorithm estimates $\ _3$ via the conditional expectation at E-step. At M-step, we maximize the complete-date likelihood by tting three Poisson regression models. More precisely, the E-step calculates the expected complete data log-likelihood ratio $q(\ _i j\ _i)$

$$q(\ _i j\ _i) = E_{\ _i}[l(\ _i) \quad l(\ _i) j Y_1; Y_2];$$

where $\ _i$ be and arbitrary but xed value, typically the value of $\ _i$ at the current iteration. The M-step maximizes $q(\ _i j\ _i)$ in $\ _i$ for for xed $\ _i$, i.e.

$$\ _i = \arg\max_{\ _i} q(\ _i j\ _i):$$

### 5.5.3 In ated Bivariate Poisson model

A major drawback of the bivariate Poisson model is the restriction of positive correlation only. Since its marginal distributions are Poisson, they cannot model over-dispersion or under-dispersion. It has been found that the number of draws in a game is larger than those predicted by a simple bivariate Poisson model (Karlis and Ntzoufras (2003)).

We can consider mixtures of bivariate Poisson distributions but such models suffer from complicated structure. A more applicable approach is in ating the probabilities of certain values of variable under consideration. Dixon and Coles (1997) have proposed a zero-in ated model which considered in ation only for the $(y_1 = 0; y_2 = 0)$ cell. In this report, we extend the zero-in ated model by adding an in ation component on the diagonal of joint probability function. A diagonal in ated model is speci ed by

$$f_{IBP}(y_1; y_2) = \begin{cases} (1 \quad) f_{JBP}(y_1; y_2 | \ _1; \ _2; \ _3) & y_1 \neq y_2; \\ (1 \quad) f_{JBP}(y_1; y_2 | \ _1; \ _2; \ _3) + D(y_1 | \ ); & y_1 = y_2 \end{cases};$$

where $D(y_1 | \ )$ is a discrete distribution with parameter vector . When $= 0$, we have the simple bivariate Poisson model. Diagonal in ated models can be tted using the EM algorithms provided in Karlis and Ntzoufras (2005). A proper choice of $D(y_1 | \ )$ is geometric distribution. $P(y_1 = j) = \ _j$ for j=0,1,2,...,J, where $\sum_{j=0} \ _j = 1$. If $J = 0$, then we have a zero-in ated model. One important property of in ated models is the marginal distributions are not Poisson distributions, which means they can be either under-dispersed or over-dispersed depending on the choice of $D(y_1 | \ )$. In the simplest case of zero-in ated models, the marginal distributions are also over-dispersed relative to the simple Poisson distribution. Another important characteristic is that, even if $\ _3 = 0$ (double Poisson distribution), the resulting in ated distribution introduces a degree of dependence between the two variables under consideration. Moreover, diagonal in ation can possibly correct both over or under-dispersion and correlation problems encountered in modeling count data.

In (Figure 15) you can see the best hyperparameter for Dixon-Cole in ation, and in (Figure 16) you can see the hyperparameter table for diagonal Poisson in ation. (Table 3) shows the effect of different types of in ation on validation and test data using double Poisson model.

Figure 15: Finding best  for Dixon-Cole in ation

Figure 16: Finding best p and  for diagonal Poisson in ation

| different ways for in ation for double Poisson models | | | | |
|---|---|---|---|---|
| | Negative Log-Likelihood | | | |
| | valid | test | value ( tted on train) | |
| no in ation | 2.902616 | 2.887559 | - | |
| Dixon-Cole | 2.90176 | 2.887176 | = -0.027 | |
| diagonal Poisson | 2.901885 | 2.887013 | p = 0.01325, | = 0.96 |
| diagonal geometric | 2.901915 | 2.887198 | p = 0.01129, | = 0.66 |

Table 3: Effect of Different Types of In ation

# 6 Neural Networks and Rating Systems

The work of Maher (1982) for predicting the exact scorelines using Poisson regression was a breakthrough in this context. As we mentioned earlier, they used a rating method to predict the number of goals. There are many rating algorithms in football which is used by the researches, specially Bayesian networks bene t from this ratings systems. The basic way to have a rating, is to assign a strength rating for each team. It means if we have n teams and we assign $r_i; i \ 2 \ f \ 1; :::; ng$ for each team to represent their strengths, we can use them to model match outcomes. For example in match result prediction (home win, draw, away win) between team i and team j we can use $r_i \quad r_j$ to estimate the outcome. ELO rating system Elo (1986) is one of the most common strength indicators of football teams that is used in the literature. It was originally designed for rating chess players for it became also popular in other sports like football. Herbinet (2018) used ELO ratings together with a attack and defence rating system to predict the number of goals.

Maher (1982) assigned attack ( $_i$) and defence ( $_i$) ratings for each team $i \ 2 \ 0; :::; n$ and used $_i \ _j$ and $_j \ _i$ to estimate $_i$ and $_j$ in a Poisson regression context. However, this idea has been completed by many works afterwards, but it is still one of the main ideas to build a prediction model for the number of goals based on it. In our case, since we want to testify the power of prediction of the outcomes only by using the performance features of the teams in the recent games, so we will not consider the rating of teams as a parameter in our estimation. In other words, we are not going to add the team name to our features data. Therefore, using the conventional attack and defence rating system is not applied in this cases because they are assigned to the teams.

Alternatively, there are some method that we can use to predict the attack and defence rating of home and away teams using their features. The rst method that we analyse, is to use a clustering algorithm on the features space, and then assign an attack and defence rating for each cluster instead of each team. Note that an speci c team may be in different clusters over the time depending on their previous performances. Another method is to estimate $_i$, $_j$, $_i$, and $_j$ directly through our features. Instead of using generalized linear model, to have a more effective approximation, we have used neural networks to estimate the attack and defence ratings of the teams using their features that represents their performances in the recent matches. To train our neural network, we can use the Poisson regression method and Poisson negative log likelihood as our loss function. We will discuss it more at the end of this chapter.

## 6.1 Estimating Team Attributes

Many statistical and Bayesian approaches in predicting match outcome comes from the popular algorithms to estimate the different attributes of the teams. For example, let $R_i$ represents the total strength of team $i \in \{1, \ldots, n\}$, then after a game between team $i$ and team $j$, we change the rating from $R_i$ to $R_i + \Delta R$ where $\Delta R = Y \cdot P$ and $Y = Y_i - Y_j$ is the difference of goals and $P = P - \hat{P}$ is the difference of the predicted match result and the actual match result. A generalization of this process gives us a rating system which called ELO, which is widely used in probability estimations and Bayesian inferences. This indicator which is also used in FIFA ratings, is included in the features for predicting match outcomes in many works.

Similar to using the strength rating to predict the scorelines, by using Poisson regression we can provide a way to model the strength rating. In this case, we can again estimate the number of goals of home and away by a bivariate Poisson distribution using the difference of the ratings of two teams as our indicator. Suppose team $H$ and team $A$ have ratings $R_H$ and $R_A$ respectively and $\Delta R = R_H - R_A$, then one way to estimate the average goals of home and away is

$$\begin{cases} \log(\lambda_H) = f(\Delta R) + h \\ \log(\lambda_A) = f(-\Delta R) \\ \log(\lambda_C) = c(X) \end{cases}; \tag{5}$$

where $h$ is the home effect (around $0.26$ in our data) and $c(X)$ is a function to estimate the correlation between the number of goals where $X$ is our match features. These equations imply $\log(\lambda_H / \lambda_A) - h = g(\Delta R)$ where $g(\Delta R) = f(\Delta R) - f(-\Delta R)$.

Now, we can linearly model our rating system together with $c$, $f$ and consecutively $g$. Since we have already modeled $\log(\lambda_H)$, $\log(\lambda_A)$ and therefore $\log(\lambda_H / \lambda_A)$ to have a linear distribution, so it is fair that we approximate $g$ with linear functions, we can set $g(\Delta R) = \gamma_0 \Delta R$ (the intercepts of $f$ cancel each other). We know that our data consists of the features of both home and away teams. It means, if we split our features for home and away teams $X = (X_H; X_A)$ that represent the features of teams $H$ and $A$ respectively, and if we linearly model our rating system to be $R_H = \gamma X_H$, then we can have

$$\log(\lambda_H / \lambda_A) - h = g(\Delta R) = \gamma_0 \Delta R = \gamma X_H - \gamma X_A \tag{6}$$

where $= {}_0$ . The coef cients of can be interpreted as the effect of each feature in the team strength. (Figure 17)

In the previous chapter, we used generalized linear model with Poisson regression to estimate ${}_H$ and ${}_A$. For a match between teams H and A we assumed that

$$\begin{cases} \log({}_H) = X \ {}_H + h \\ \log({}_A) = X \ {}_A \\ \log({}_C) = X \ {}_c \end{cases} ; \qquad (7)$$

where ${}_H$; ${}_A$; ${}_c$ are the corresponding linear coef cients estimated in Poisson regression model. Here we can split X to $(X_H; X_A)$, We can also split ${}_H = ({}_{HH}; {}_{HA})$ and ${}_A = ({}_{AH}; {}_{AA})$ to represent attack and defence ability of home and away teams. Then we can deduce that $\log({}_H) = X \ {}_c + {}_{HH} X_H + {}_{HA} x_A + h$ and $\log({}_A) = X \ {}_c + {}_{AH} X_H + {}_{AA} X_A$. In this case

$$\log({}_H = {}_A) \quad h = ({}_{HH} \quad {}_{AH}) X_H \quad ({}_{AH} \quad {}_{AA}) X_A \qquad (8)$$

This equation is similar to the previous equation where we used the strength ratings $R_H$ and $R_A$ to estimate $\log({}_H = {}_A)$. In the previous case, we calculated the rating of the teams by multiplying their features to . But in this case we computed the rating of the teams by multiplying their features to ${}_H$ and ${}_A$ depending that the team is playing in home or away, where ${}_H = {}_{HH} \quad {}_{AH}$ and ${}_A = {}_{AA} \quad {}_{HA}$. In (Figure 17) you can see the values of ${}_H$ and ${}_A$ that we derived in the previous chapter using bivariate Poisson regression, you can see that they are similar.

Another way to estimate the attributes of the team is by using the independent Poisson regression method introduced in Maher (1982). They assigned attacking and defending ratings to each team. So for each match between H and A we have 4 parameters, ${}_H$, , ${}_H$, and ${}_A$ representing their attack and defense ratings. Then they modeled ${}_H = {}_H \ {}_A e^h$ and ${}_A = {}_A \ {}_H$ where h is the home factor. A simple bivariate Poisson approach using their idea to assign attack and defense rating, is to linearly model and . Then we will have the following model

$$\begin{cases} \log({}_H) = X_H \ + X_A \ + h \\ \log({}_A) = X_A \ + X_H \\ \log({}_C) = X \ {}_c \end{cases} \qquad (9)$$

Figure 17: Coef cients of team rating linear mappings

It is good to note that in this situation, is actually the inverse of defense rating because here having more means having worse defense. Similar to what we did for strength rating, we can interpret our generalized bivariate Poisson regression to extract these attack and defense rating factors. We can split $X = (X_H; X_A)$ to rewrite equation (7). Recall the de nition of $_H H$, $_H A$, $_A H$, and $_A A$ from equation (8), then

$$\begin{cases} \log(\ _H) = X_H\ _{HH} + X_A\ _{HA} + h \\ \log(\ _A) = X_A\ _{AA} + X_H\ _{AH} \\ \log(\ _C) = X\ _c \end{cases} \tag{10}$$

Like what we saw in coef cients of strength rating estimation. Here we can observe that $_H H$ and $_A A$ play the role of assessing the attacking ratings, and on the other hand, $_H A$ and $_A H$ compute defense rating. In (Figure 18) we can see they are similar to each other and also to the solution of optimization (9)

Originally, Maher (1982) assigned the attack and defense parameters to each team instead of estimating it from the match features. In their work, each team could have a separate attack and defense rating whether they play in home or away, but later Dixon and Coles (1997) did not consider this separation to reduce the number of parameters as well. Instead, they applied an in ation effect which increases the likelihood and we analysed it in the previous chapter. By assigning a xed rating for each team, one could think that it's not actually very realistic because the team performances could be variant through the time. To model this inef ciency, Dixon and Coles (1997) used

a time exponential weight decay to compute the likelihood.

By clustering the data, we can think of each cluster as a team. We applied K-Means algorithms to devide our data into 20 clusters. And then using the conventional methods we can assign ratings to the clusters. In (Figure 19) you can see that the attack and defense rating can be calculated for each cluster using Dixon and Coles (1997) algorithm, which is similar to the attack and defense ratings of cluster centers computed by the weights of the generalized linear bivariate Poisson regression model.

Figure 18: Coef cients of attack and defence linear mappings

Figure 19: Attack and defense rating of the different clusters

## 6.2 Neural Networks

With the advances of technology, the usage of ANNs (Arti cial Neural Networks) has become more frequent due to their high capability to t on the high dimensional spaces. There are many gradient based optimization methods which have helped Deep Learning model to be applied to the classi cation and regression problems and perform better than the previous models specially in high dimensional spaces thanks to more calculation power that we have recently achieved.

By using neutral networks we can reconstruct the models for predicting the match result and scoreline. However we believe that the degree of complexity of our data is not high enough to have a major improvement by using neural networks, but as we saw in chapter 3, we expect a better result from neural networks comparing to the linear models.

Neural networks give us the advantage of being able to have more complicated models by changing the architecture of the model and the loss function. In this chapter we are going to analyse the performance of ANNs in scoreline prediction and we will compare it with the models that we introduced in the previous chapters.

As we already mentioned, to model a bivariate poisson distribution we should have the estimations of $_H$, $_A$, and $_C$. Similar to the generalized linear models, we can estimate these values from the match features, but here we can add hidden layers to capture the nonlinear behavior of the data. Therefore we have implemented a simple multilayer perceptron with one hidden layer and 3 outputs representing $_H$, $_A$, and $_C$. Since the natural logarithm function is an appropriate link function for estimating our outputs, we have used the exponential function as the activation for the output layer. For the hidden layer we used $\tanh$ activation, which implies that if $X$ is the input data and $h = \tanh(XW)$ is the hidden layer (here $W$ represents its weights), then we can have $_H = \exp(h\ _H)$, $_A = \exp(h\ _A)$, and $_C = \exp(h\ _C)$ where $_H$, $_A$ and $_C$ are the weights of the output nodes $_H$, $_A$, and $_C$. And then we can use $_H + _C$, and $_H + _C$ to estimate the average goals of each team by assuming that they have a bivariate Poisson distribution. In (Table 4) we can see that the loss of this multilayer ANN bivariate Poisson is slightly less than the linear bivariate Poisson model.

Another way to use ANN in this context is by modeling the rating methods introduced in the previuos section. We have seen the the general rating method and the attack defense method can provide us the indicators to predict the scoreline using them (Figure 20 and 21). With ANNs we are able to modify the estimation models for computing these ratings, we have implemented the ANN versions of these methods.

The simplest case for estimating the ratings with neural networks is to use multilayer

perceptron. But we also can generalize this idea to have a more effective model. Instead of a single value for the rating, we can extend these ratings to have more dimensions. It means we can use the match features to estimate the team ratings but we will a have vector that represents these ratings.

To model the general strength rating of each team using their features, we have used a multilayer perceptron with one hidden layer. Therefore the team ratings $R_H$ and $R_A$, can be seen as two vectors that we can get by applying the mentioned multilayer perceptron on the home team and away team features respectively. And then similar to the original case, the scoreline can be predicted from $R_H$ $R_A$ which is a vector now. Again, we have used a multilayer perceptron to derive $_H$ and $_A$ from $R_H$ $R_A$, and we have used the initial match features to estimate $_C$. In (Figure 22) you can see a visual representation of this model.

We can have a similar approach to use attack and defense ratings of the teams using ANNs. Again, we extend the ratings to be a vector instead of a single value, and we use multilayer perceptron to model them. Note that, since in the original model $_H$ depends only on the attack rating of the home team and the defense rating of the away team, here we model $_H$ from the vectors that represent the home attack and away defense. In other words, we have another multilayer perceptron to predict the average number of goals of a team using their attack vector and the defense vector of their opponent.

Therefore, after estimating the ratings, we concatenate home attack and away defense vectors and we use it to estimate $_H$. Similarly, $_A$ can be model by concatenating away attack and home defense and applied the mentiond multilayer perceptron. Lastly, $_C$ is modeled similar to the last case by uaing the match features. In (Figure 23) you can see the visual representation.

In all of our models, we had to use some speci c con gurations to help our neural networks in the training process. Since the usage of in ation is independent to the choice of model, like the other models, we can also apply different in ation methods that we introduced in the previous chapter to the neural network models. We have used tanh function as the activation function of our hidden layers and. To avoid over- tting, we have included L1 and L2 regularizations in our hidden layers and use early stopping to capture the best performance of the models on the validation data, then we test them on the test data. We also used a stochastic gradient descent algorithm and batch size equal to 128. The trainings take 5 minutes on GPU: Nvidia Geforce GTX 1060. In (Figure 24) you can see the loss plot during the training of last two mentioned neural network models respectively. In (Table 4) you can see the performances of these models comparing to the other models on the test set.

| Scoreline Prediction Evaluation | | | |
|---|---|---|---|
| | Negative Log-Likelihood | | |
| | No In ation | Dixon-Coles | Diagonal Poisson |
| GLM | 2.888148 | 2.887778 | 2.887611 |
| BVP | 2.887559 | 2.887176 | 2.887009 |
| GR | 2.891967 | 2.891552 | 2.891261 |
| AD | 2.887924 | 2.887541 | 2.887415 |
| NNBP | 2.887202 | 2.886817 | 2.885813 |
| NNGR | 2.886271 | 2.88588 | 2.886719 |
| NNAD | 2.885894 | 2.885505 | 2.885377 |

Table 4: Performances of the scoreline prediction models. GLM: Generalized Linear Model. BVP: Bivariate Poisson Model. GR: Prediction using difference of rating. AD: Prediction Using Attack and Defense Ratings. NNBP: Neural Network Bivariate Poisson Model. NNGR: Neural Network General Rating Model. NNAD: Neural Network Attack Defense Rating Model

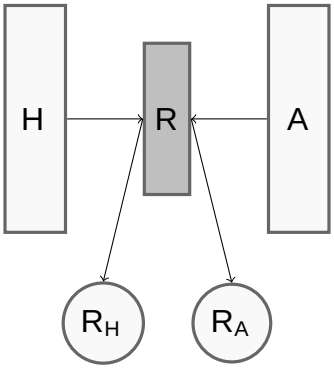


Figure 20: real valued general strength rating system

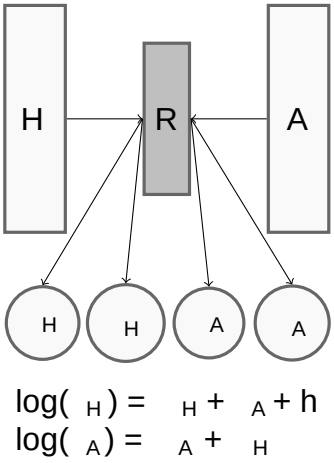


Figure 21: real valued attack defense rating system

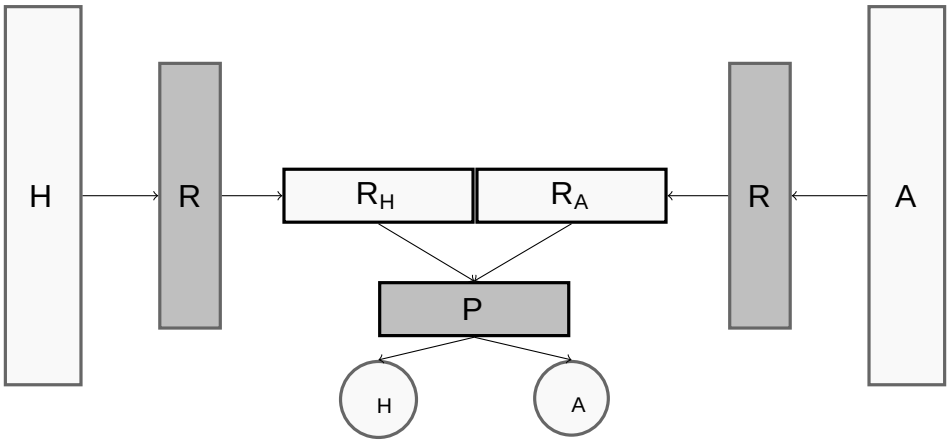


Figure 22: neural networks general strength rating system

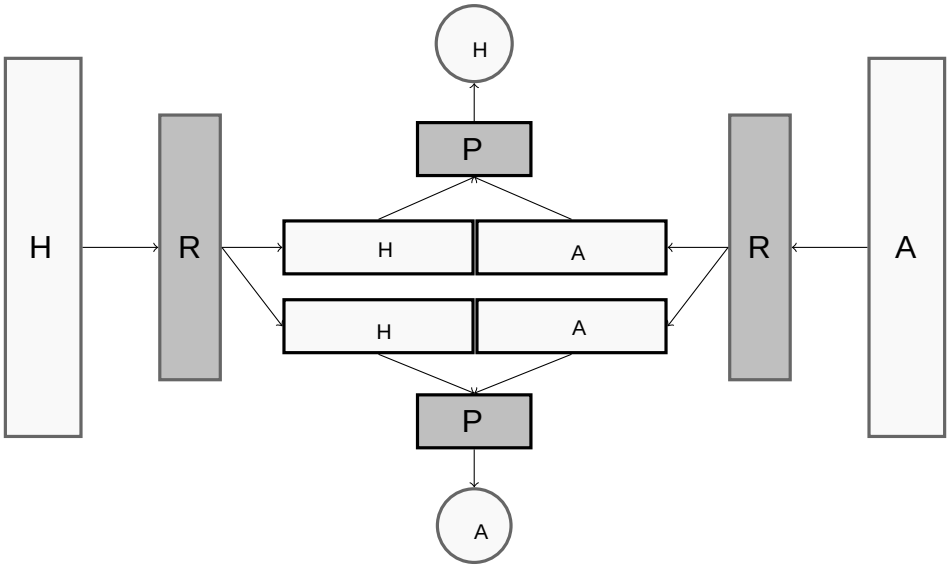


Figure 23: neural networks attack defense rating system

# 7 Conclusion and Future Works

We can split our work into two general parts. In the rst part we have analysed the quality of bookmaker odds, and the performances of the conventional classi cation models on predicting the match result. We observed that the probabilities implied from the bookmakers odds, are more accurate than the models that we trained on the historical data. Note that we haven't used a powerful data in the sense that there are a lot of other parameters and conditions that have a direct impact on the match outcomes, but we haven't considered them. Therefore it wasn't possible to have a reliable positive return in match result betting, however, we found a slightly positive return in betting on having more than 2.5 goals from the maximum odds of the bookmakers. In the future works, we are intersted in using a more complete data to improve our result.