Football Prediction

[The Premier League]

—

Michael Causon / Liam Critcher / Jia LIN
Bo HUANG / Dailin LI / Jiayi ZHANG

CONTENTS

I Introduction

Home Advantage:

- ✓ Teams will perform better on a pitch that they constantly play on with the encouragement of the majority of the stadium.
- ✓ The home advantage will give an underlying boost to the home team.

There is some sort of effect of playing home compared to playing away.

Years 2006-2019

46.34%	25.00%	28.66%
Home Win	Draw	Away Win

Mean square error :

- ✓ Small errors means the model is good.
- ✓ Need to assume that the bookmakers are doing a good job (limitation).

PLL measure (better) :

$$\boldsymbol{PLL} = \sum_{K=1}^{N} \{\delta_K^H \log(P_K^H) + \delta_K^D \log(P_K^D) + \delta_K^A \log P_K^A\}$$

- ✓ $\delta_K^H = 1$ and $\delta_K^A = 0$ if home team wins; $\delta_K^H = 0$ and $\delta_K^A = 1$ if away team wins; $\delta_K^H = 0$ and $\delta_K^A = 0$ if teams draw.
- ✓ P_K^H, P_K^A and P_K^D represent the probabilities of home team wins, away team wins and team draw respectively.

PLL measure :

- ✓ The result is always non-positive.
- ✓ Maximise this via making it least negative as possible.
- ✓ PLL $\rightarrow$ 0 indicates the model has predicted the event with high probability.

For data of varied lengths one can take **mean PLL** as a measure instead:

Company	Bet365	Betway	Stan James	Interwetten	Ladbrokes	Sportingbet
Mean PLL	-0.9554	-0.9574	-0.9611	-0.9585	-0.9630	-0.9644

The results are **very similar** between the bookmakers (they use similar models).

1. A model with mean **PLL of -1** would be of similar quality to the bookmakers'.
2. Models that approach values **greater than -0.95** over large testing datasets will actually predict better than bookmakers.

II Methodology

Elo Ratings in Football :

Developed by Physics Professor Arpad Elo as a measure of skill level between two opponents.

01
Mean 1500

Elo ratings in the league will always have mean 1500.

02
Minimum 1200

Bad teams have Elo ratings dropping as low as 1200.

03
Maximum 1800

Historically good teams have Elo ratings that peak at 1800.

04
Difference

Difference in Elo rating is the main focus in predicting the outcome of the game.

Elo Ratings Theory :

An expected outcome :

$$E = \frac{1}{1 + 10^{-\frac{dr}{400}}}$$

Implications :

- ✓ Method is zero sum.
- ✓ Mean rating always lies at 1500.
- ✓ Large margins of victory affect Elo rating more than small ones.
- ✓ Upsets have bigger affect on Elo rating.

STEP I

The difference :

$$dr = Elo_H - Elo_A + 62$$

- ✓ $Elo_H =$ *home Elo rating*
- ✓ $Elo_A =$ *away Elo rating*

STEP II

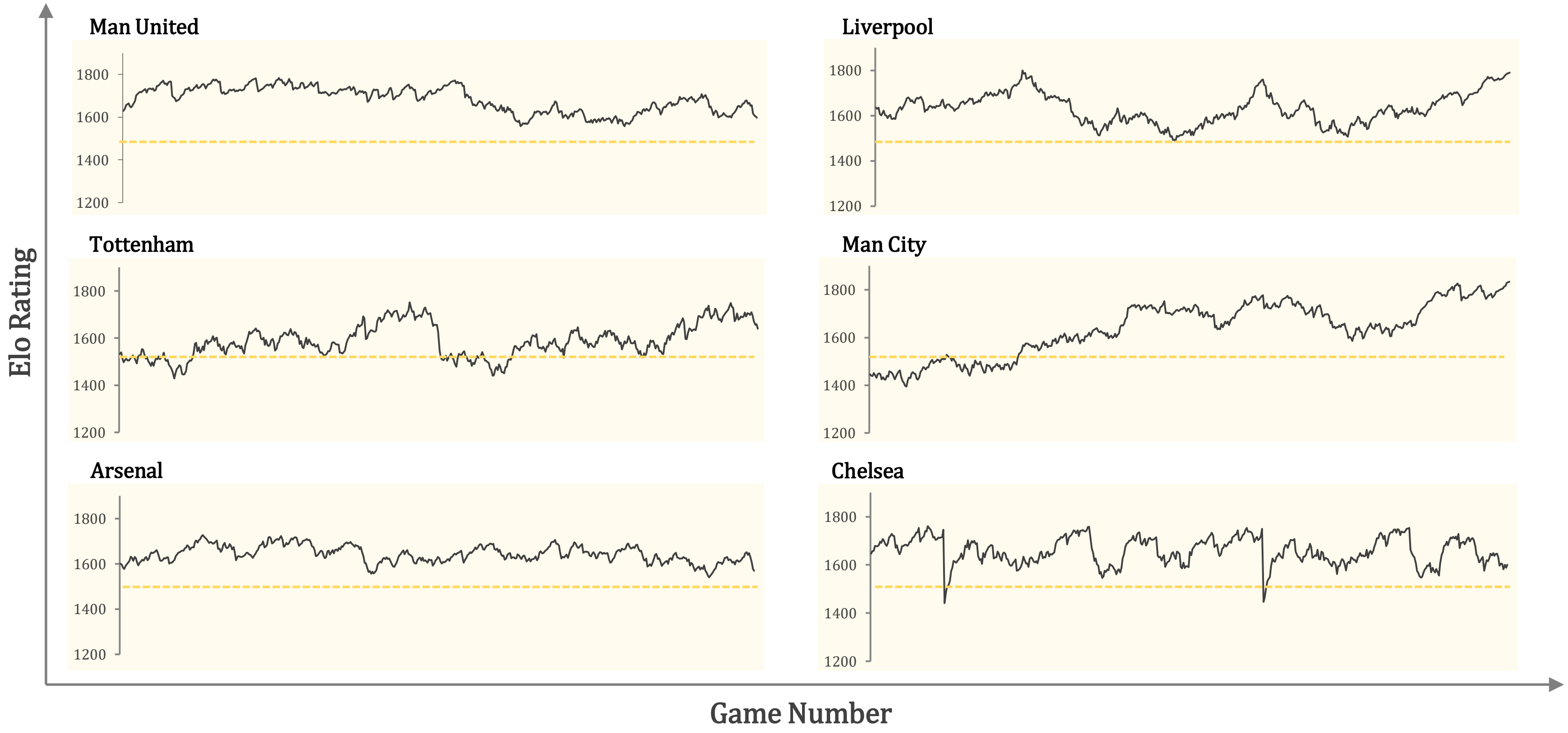
An expected outcome :

$$Elo'_H = Elo_H + KG(O - E)$$

- ✓ $O = \begin{cases} 1, \text{Home win} \\ 1/2, \text{Draw} \\ 0, \text{Away win} \end{cases}$ & K=20
- ✓ G *depends on margin of victory*

STEP III

The Big Six :


Man United
Liverpool
Tottenham
Man City
Arsenal
Chelsea
1800
1600
1400
1200
Elo Rating
Game Number

01

Two Inputs

Home Elo Rating
Away Elo Rating

02

The Model

✓ $log\left(\frac{P(Home\ win)}{P(Away\ win)}\right) = 0.3732401+$
$0.00674583Elo_H - 0.00659926Elo_A$

✓ $log\left(\frac{P(Draw)}{P(Away\ win)}\right) = 1.1073225+$
$0.00280756Elo_H - 0.00349202Elo_A$

✓ $P(Home\ win) + P(Draw) + P(Away\ win) = 1$

03

Multinomial Output

P(Home Win)
P(Draw)
P(Away Win)

Pros and Cons of the Elo Model :

Positives

- ✓ Mean PLL of within 1.5% of the best bookmakers’ for 1500 Premier League fixtures.
- ✓ Highly generalisable to other leagues.
- ✓ Can predict uncertain games at the very start of the season.
- ✓ Simple to use once the model is trained.

Negatives

- ✓ Not quite as accurate as bookmakers or other models we found.
- ✓ Performs slightly worse on other leagues (3% worse PLL than Bet365 on the 2018/19 Serie A season).

The MLE

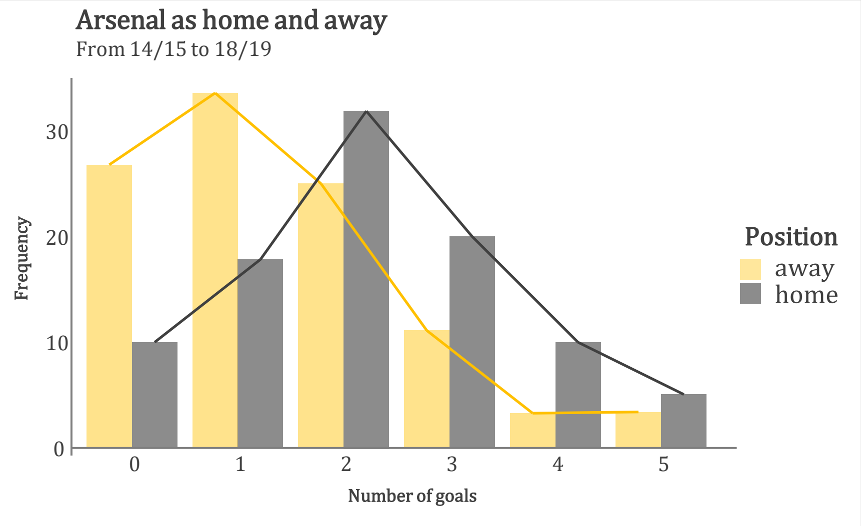
- ✓ Evaluate the parameter of each team.
- ✓ Predict the probability of the game result.

The Possession

During the possession, the team would try large number of times to attack with small probability P to shoot successfully under the defence of the opponent.

Poisson Limit Theorem

$$\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}=e^{-\lambda}\frac{\lambda^k}{k!}$$



Poisson Regression Model :

The goals generally follow the Poisson distribution :

- ✓ The goal of each team follows the Poisson distribution.
- ✓ The goal between each team and its opponent is independent.
- ✓ Each score-line is independent from match to match.

The Model :

$$P\left(X_{i,j} = x, Y_{i,j} = y \middle| \lambda, \mu\right) = \frac{\lambda^x exp(-\lambda)}{x!} \frac{\mu^y exp(-\mu)}{y!}$$

- ✓ X & Y are the goals scored in the game where team i plays against team j.
- ✓ These two random variables are independent so that the joint probability is the product of the two independent probabilities.

Parameter Estimation :

$$\lambda = exp(\alpha_i - \beta_j + \gamma)$$

$$\mu = exp(\alpha_j - \beta_i)$$

- ✓ α_i is the attack ability of the team *i*.
- ✓ β_i is the defense ability of the team *i*.
- ✓ γ is the home advantage.

Substitute the expected value into pmf and take the logarithm, if k games are observed:

$$L = \sum_{i=1}^{k} logp(x_i, y_i | \lambda_i, \mu_i)$$

To avoid multiple combination of parameters which may produce the same model:

$$\sum \alpha_i = 0 \ \& \ \sum \beta_i = 0$$

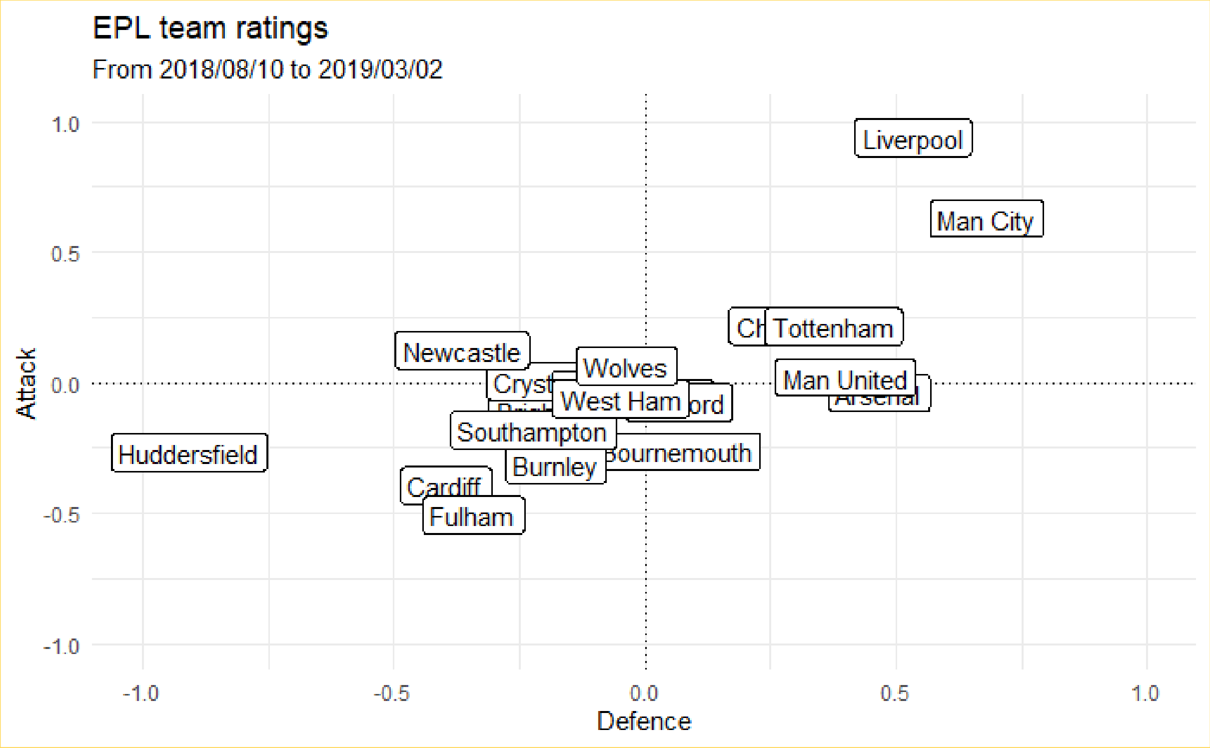
The probability of each result (win/draw/lose) be forecasted.

$$P(home\ win) \sum_{x=1}^{\infty} \sum_{y=0}^{x-1} p(x, y | \lambda, \mu)$$

$$P(away\ win) \sum_{y=1}^{\infty} \sum_{x=0}^{y-1} p(x, y | \lambda, \mu)$$

$$P(home\ draw) \sum_{x=1}^{\infty} p(x, x | \lambda, \mu)$$

Poisson Regression Model



"Attack ability vs. Defense ability in 18/19 season"

- ✓ The Dixon-Coles model think that the goal condition among 0:0, 1:0, 0:1 and 1:1 are not independent.
- ✓ Therefore, one probability modified function would be added in the Poisson regression model.

$$\boldsymbol{\tau(x, y, \lambda, \rho)} =$$

$$\begin{cases} 1 - \lambda\mu\rho, & x = 0 \,\&\, y = 0 \\ 1 + \lambda\rho, & x = 0 \,\&\, y = 1 \\ 1 + \mu\rho, & x = 1 \,\&\, y = 1 \\ 1 - \rho, & x = 1 \,\&\, y = 1 \\ 1, & otherwise \end{cases}$$

The Dixon-Coles Model :

$$\boldsymbol{P\left(X_{i,j} = x, Y_{i,j} = y \middle| \lambda, \mu, \rho\right) = \tau(x, y, \lambda, \rho) \frac{\lambda^x exp(-\lambda)}{x!} \frac{\mu^y exp(-\mu)}{y!}}$$

Time Weighting :

Time weight function:

$$\emptyset(t) = \begin{cases} exp(-\xi t) & t \leq t_0 \\ 0 & t > t_0 \end{cases}$$

Where ξ is a constant and $\boldsymbol{t_0}$ is a time at which our prediction is being made.

- ✓ Older data has less effect than more recent data.
- ✓ A game long time ago has no effect on the result of prediction.

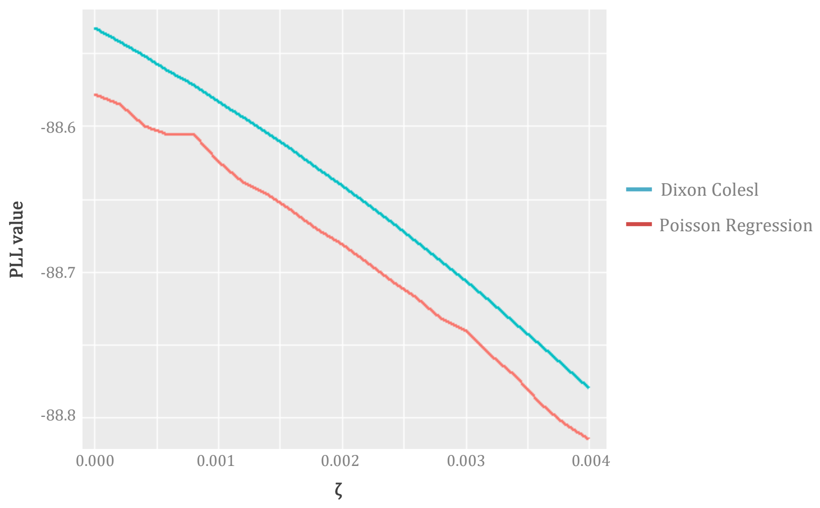
Insert the time weight function to the log-likelihood:

$$l = \sum_{i=1}^{k} \emptyset(t_i) log P(x_i, y_i | \lambda_i, \mu_i)$$

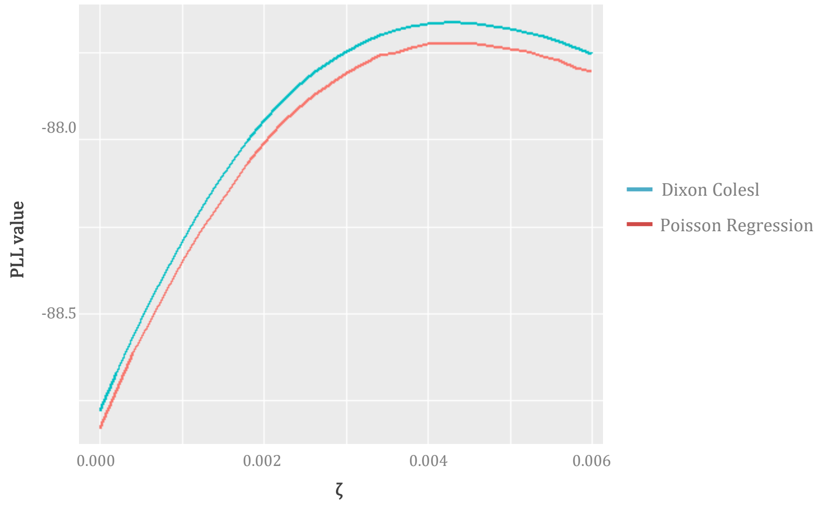
- ✓ The ξ with the highest mean PLL may represent that such model could predict the results best.

Plots of the Mean PLL :

Prediction for 16/17 season using 1 season data:



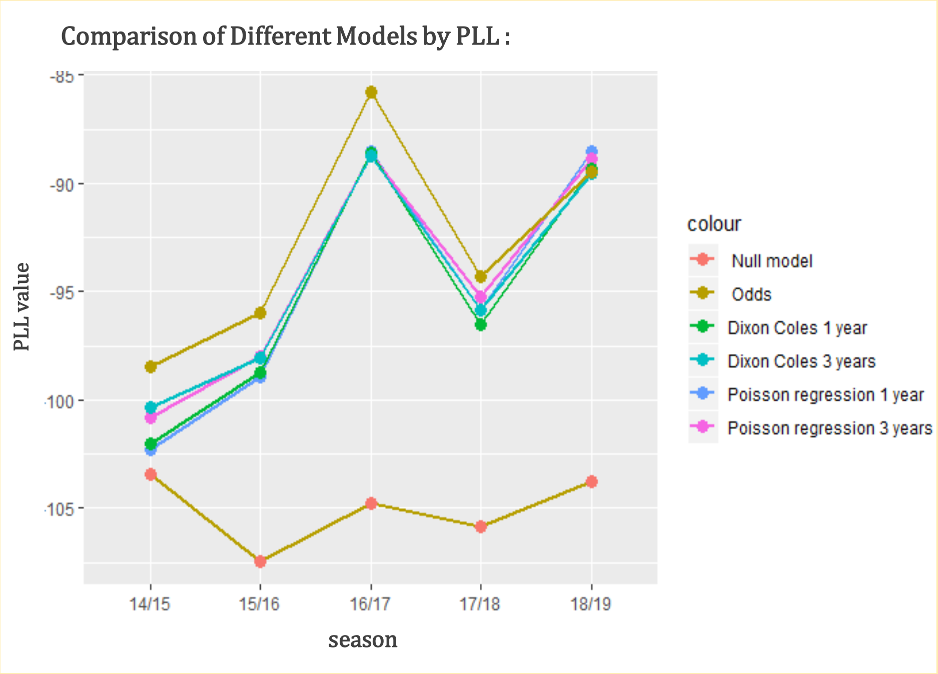
Prediction for 16/17 season using 3 seasons data:



Both Dixon Coles model and Poisson regression model could perform best without adding time weight .

The ξ of weight function should be set as 0.0044 for the best predictions.

Predicting the Results :



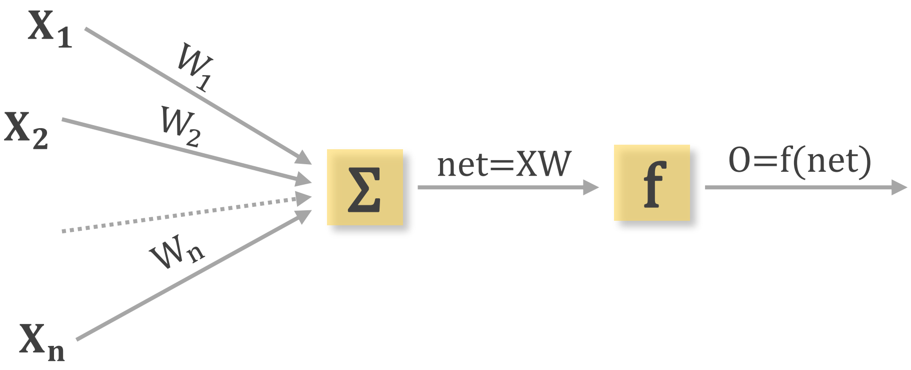
- ✓ All the models are able to predict much better than the null model.
- ✓ The predicting ability between each model and bookmaker model is very similar.

The Model	Dixon Coles	Poisson Regression
1 year data	-0.9504066	-0.9481608
3 years data	-0.9439942	-0.9424600

Artificial neural networks(ANN) :

A simulation of human neural networks, more directly, it is a mathematical model, which can be achieved by computer or electric circuit.

Artificial neuron is a basic unit of ANN, which is a simulation of biological neuron.



Input:

$$X = (x_1, x_2, x_3, \dots, x_n)$$

Weight:

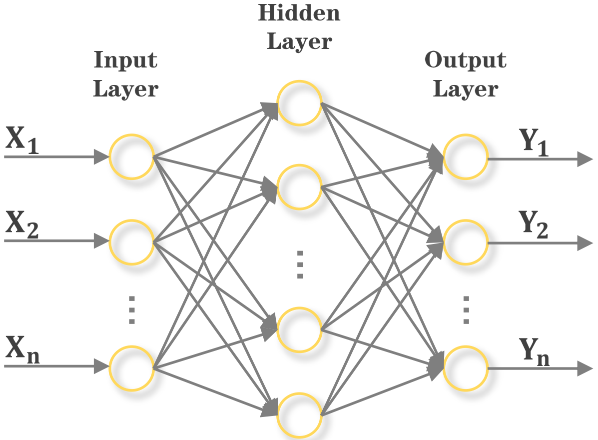
$$W = (w_1, w_2, w_3, \dots, w_n)^T$$

Net Input:

$$net = \sum x_i w_i$$

Activation Function:

$$o = f(net)$$



Basic assumptions :

- ✓ Network contains L layer.
- ✓ Connection matrix : $w^1, w^2, w^3, \ldots, w^L$.
- ✓ Sample set : $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s)\}$.
- ✓ Neurons in layer k: H_k.

Home	Away	Vector			Odds.Win	Odds.Draw	Odds.Lose
Brighton	Huddersfield	0.84	0.07	0.09	0.5168	0.2393	0.2440
Burnley	Crystal Palace	0.19	0.30	0.51	0.2865	0.3189	0.3946
Man United	Southampton	0.93	0.01	0.06	0.5512	0.2191	0.2297
Tottenham	Arsenal	0.95	0.00	0.05	0.5590	0.2146	0.2264
West Ham	Newcastle	0.72	0.15	0.13	0.4701	0.2677	0.2623
Wolves	Cardiff	0.93	0.01	0.06	0.5507	0.2194	0.2299

The Softmax Function :

$$\sigma(Z)_i = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}} \quad for\ i = 1, \dots, K\ \&\ Z = (Z_1, \dots, Z_K) \in R^K$$

57%

Prediction Accuracy

18/19

Last 100 Matches

-1.012

The Mean PLL

III Results & Discussion

Three Models :

01
The Elo Model
Long-term prediction or dealing with general situations.

02
The Poisson Model
Short-term prediction.

03
BP ANN
Potential.

Limitations :

Teams in transition :
- ✓ Player suspensions and injuries.
- ✓ A team playing particularly badly will look to change their manager.

Quality of data :
- ✓ One should also consider the data available.
- ✓ A more detailed dataset would be also surely beneficial.

THANKS

END