Introduction to Statistic for Data Science

Group Mini-Project Presentation: Happiness Ladder

ISDS Group 10

Christopher Barrow, Zehao Qian, Mengyuan Zhu, Hithein Augustine

Department of Natural Sciences
Durham University, England, UK

November 28, 2023

1. Introduction
    - Background
2. Mathematical Modeling
    - Method 1: Entropy Weight Method in Football Team Evaluation
    - Method 2: Linear Transformation and Gradient Ascend
3. Result Analysis
    - Grabbing Data from Web
    - Evaluating the Teams's Tier
    - Prediction of 2023/24's winner point
4. Future Work
    - Our plans, our limitations and how these models could improve

Outline

1. Introduction
   - Background
2. Mathematical Modeling
   - Method 1: Entropy Weight Method in Football Team Evaluation
   - Method 2: Linear Transformation and Gradient Ascend
3. Result Analysis
   - Grabbing Data from Web
   - Evaluating the Teams's Tier
   - Prediction of 2023/24's winner point
4. Future Work
   - Our plans, our limitations and how these models could improve

The Premier League

- Premier League: Top tier of English Football League System.
- 20 teams play 38 home and away matches.
- Globally renowned and challenging to predict outcomes.

Background

1. Outcome predictions involve expert analysis.
2. Factors include team performance, player form, and tactics.
3. Growing data, e.g., player touches, team running stats, manager experience.

Outline

1. Introduction
   - Background
2. Mathematical Modeling
   - Method 1: Entropy Weight Method in Football Team Evaluation
   - Method 2: Linear Transformation and Gradient Ascend
3. Result Analysis
   - Grabbing Data from Web
   - Evaluating the Teams's Tier
   - Prediction of 2023/24's winner point
4. Future Work
   - Our plans, our limitations and how these models could improve

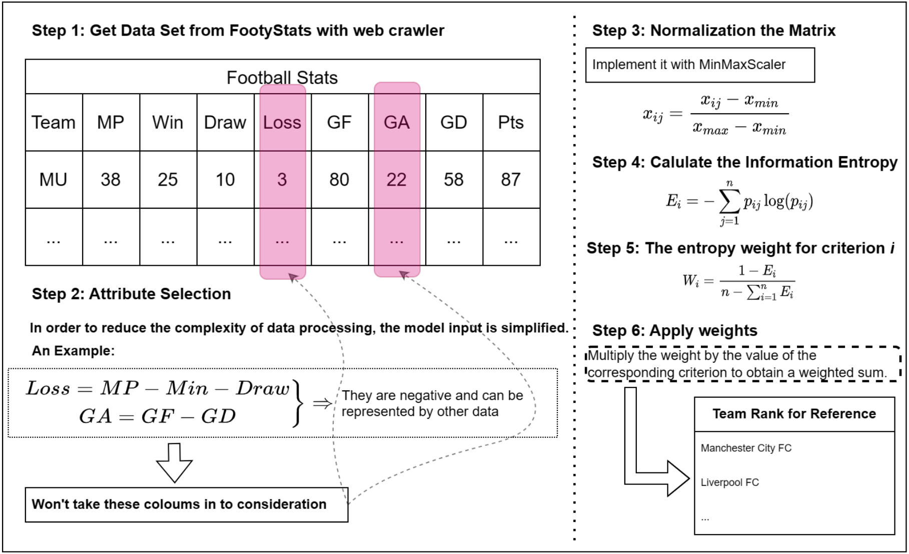
Overview of Entropy Weight Method in Football

1. Introduction
   - The Entropy Weight Method is a powerful analytical technique used in football team evaluation. It goes beyond traditional methods by considering the inherent information entropy within various performance attributes.
2. Key Characteristics
   - Entropy: Reflects the degree of uncertainty or randomness within a dataset.
   - Weight Assignment: Assigns weights to attributes based on their information entropy.
3. Objective
   - The method aims to provide a nuanced evaluation, giving higher importance to attributes that contribute more to understanding a team's performance.

Key Steps in Entropy Method

1. **Data Collection and Attribute Selection**
2. **Entropy Calculation:**
   - Utilize mathematical formulas to calculate the entropy of each selected attribute.
   - Entropy $= -\sum(p_i \cdot \log_2(p_i))$, where p_i is the probability of each attribute value.
3. **Weight Assignment:**
   - Assign weights to attributes based on their calculated entropy.
   - Attributes with higher entropy receive lower weights, and vice versa.
   - The sum of weights equals 1 for normalization.
4. **Outcome:**
   - The result is a set of weights that reflect the relative importance of each attribute in evaluating a football team's performance.

Entropy Method in Our Model



Step 1: Get Data Set from FootyStats with web crawler

Football Stats

Team	MP	Win	Draw	Loss	GF	GA	GD	Pts
MU	38	25	10	3	80	22	58	87
...	...	...	...	...	...	...	...	...

Step 2: Attribute Selection

In order to reduce the complexity of data processing, the model input is simplified.

An Example:

$$\left.\begin{array}{c} Loss = MP - Min - Draw \\ GA = GF - GD \end{array}\right\} \Rightarrow$$

They are negative and can be represented by other data

Won't take these coloums in to consideration

Step 3: Normalization the Matrix

Implement it with MinMaxScaler

$$x_{ij} = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}}$$

Step 4: Calulate the Information Entropy

$$E_i = -\sum_{j=1}^{n} p_{ij} \log(p_{ij})$$

Step 5: The entropy weight for criterion *i*

$$W_i = \frac{1 - E_i}{n - \sum_{i=1}^{n} E_i}$$

Step 6: Apply weights

Multiply the weight by the value of the corresponding criterion to obtain a weighted sum.

Team Rank for Reference

Manchester City FC

Liverpool FC

...

Outline

1 Introduction

- Background

2 Mathematical Modeling

- Method 1: Entropy Weight Method in Football Team Evaluation
- Method 2: Linear Transformation and Gradient Ascend

3 Result Analysis

- Grabbing Data from Web
- Evaluating the Teams's Tier
- Prediction of 2023/24's winner point

4 Future Work

- Our plans, our limitations and how these models could improve

Linear Transformation and Gradient Ascend

1. Linear transformation involves transforming input variables into a new space where they are more linear and easier to model
2. Our model will be taking in 5-dimension vector and transforming it to a 1-dimension vector

$$F_v : \mathbb{R}^5 \Rightarrow \mathbb{R}$$

3. The vector v basically consists of the following variables:
   - The number of games won by a team in n^{th} week, v_1
   - The number of games lost by a team in n^{th} week, v_2
   - The number of games drawn by a team in n^{th} week, v_3
   - The goal difference (goals scored - goals conceded) of a team in n^{th} week, v_4
   - The number of points scored by a team in n^{th} week, v_5

Linear Transformation and Gradient Ascend

- The linear transformation will transform the 5-dimensional vector into a one-dimensional vector which is the total number of points in the final week $F(v)$
- To calculate we will use the following formula:

$$F(v) = wv + b$$

- w is a 5-dimensional vector, weights of each variable in vector v
- b is the scalar bias - systematic error

Using Gradient Ascend

- To calculate w and b, we will use gradient ascend on historical data.
- Normalize v and $F(v)$ by centering and standardizing it.
- Use normalized data and take out predicted points using any initial value of weight.
- See the error by finding the difference between normalized $F(v)$ and predicted $F(v)$
- Find the gradient of the weight using the formula - transposed-v.error
- Find the gradient of scalar bias using the formula - sum of all elements in error vector
- Then to find the updated gradient the following formula is:

$$w = w + a * w_{gradient}, \; b = b + a * b_{gradient}$$

- Then denormalize w and b to find the actual values

Outline

1. Introduction
   - Background

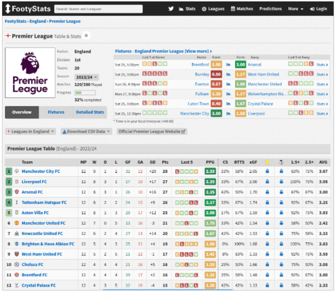
2. Mathematical Modeling
   - Method 1: Entropy Weight Method in Football Team Evaluation
   - Method 2: Linear Transformation and Gradient Ascend

3. Result Analysis
   - Grabbing Data from Web
   - Evaluating the Teams's Tier
   - Prediction of 2023/24's winner point

4. Future Work
   - Our plans, our limitations and how these models could improve

Grabbing Data from Web

Data Source: Footystats Website



FootyStats | Search Teams and Leagues | Stats | Leagues | Matches | Predictions | More | Login

FootyStats › England › Premier League

Premier League Table & Stats

Nation	England
Division	1st
Teams	20
Season	2023/24
Matches	120/380 Played
Progress	32% completed

Overview | Fixtures | Detailed Stats

Fixtures - England Premier League (View more) ›

	Last 5 at Home	Home	Form		Form	Away	Last 5 at Away	
Sat 25, 5:30pm	D D L W W	Brentford	1.50		2.00	Arsenal	W W W D L	Stats ›
Sat 25, 3:00pm	L L L L L	Burnley	0.00		1.17	West Ham United	W W L L L	Stats ›
Sun 26, 4:30pm	L L L W D	Everton	0.67		1.80	Manchester United	L L W W W	Stats ›
Mon 27, 8:00pm	L W L W L	Fulham	1.20		1.17	Wolverhampton Wa...	W L D W L	Stats ›
Sat 25, 3:00pm	L D L L D	Luton Town	0.40		1.67	Crystal Palace	D L W L W	Stats ›
Sat 25, 12:30pm	W W W W W	Manchester City	3.00		1.50	Liverpool	W W L D D	Stats ›

* Time is in your local timezone (+00:00)

Leagues in England | Download CSV Data | Official Premier League Website

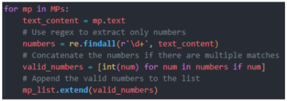
Premier League Table (England) - 2023/24

	Team	MP	W	D	L	GF	GA	GD	Pts	Last 5	PPG	CS	BTTS	xGF			1.5+	2.5+	AVG
1	Manchester City FC	12	9	1	2	32	12	+20	28	L W W W D	2.33	33%	58%	2.05	🔒	🔒	83%	75%	3.67
2	Liverpool FC	12	8	3	1	27	10	+17	27	D W W D W	2.25	33%	67%	2.00	🔒	🔒	100%	75%	3.08
3	Arsenal FC	12	8	3	1	26	10	+16	27	W D W L W	2.25	42%	50%	1.70	🔒	🔒	67%	67%	3.00
4	Tottenham Hotspur FC	12	8	2	2	24	15	+9	26	W W W L L	2.17	33%	67%	1.74	🔒	🔒	92%	67%	3.25
5	Aston Villa FC	12	8	1	3	29	17	+12	25	D W W L W	2.08	17%	67%	1.59	🔒	🔒	92%	75%	3.83
6	Manchester United FC	12	7	0	5	13	16	-3	21	W W L W W	1.75	33%	42%	1.54	🔒	🔒	58%	50%	2.42
7	Newcastle United FC	12	6	2	4	27	13	+14	20	D W D W L	1.67	42%	42%	1.53	🔒	🔒	75%	58%	3.33
8	Brighton & Hove Albion FC	12	5	4	3	25	21	+4	19	D L D D D	1.58	0%	100%	1.68	🔒	🔒	100%	75%	3.83
9	West Ham United FC	12	5	2	5	21	22	-1	17	D L L L W	1.42	8%	83%	1.33	🔒	🔒	92%	75%	3.58
10	Chelsea FC	12	4	4	4	21	16	+5	16	W D L W D	1.33	25%	50%	1.66	🔒	🔒	75%	50%	3.08
11	Brentford FC	12	4	4	4	19	17	+2	16	L W W W L	1.33	25%	58%	1.48	🔒	🔒	92%	67%	3.00
12	Crystal Palace FC	12	4	3	5	12	16	-4	15	D L L W L	1.25	42%	42%	1.13	🔒	🔒	58%	42%	2.33

Using Python Selenium Framework to automatically get data

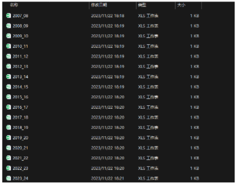
Selenium

Parse the data from HTML



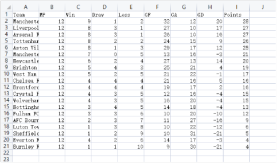
```python
for mp in MPs:
    text_content = mp.text
    # Use regex to extract only numbers
    numbers = re.findall(r'\d+', text_content)
    # Concatenate the numbers if there are multiple matches
    valid_numbers = [int(num) for num in numbers if num]
    # Append the valid numbers to the list
    mp_list.extend(valid_numbers)
```

Write Them in CSV files



名称	修改日期	类型	大小
2007_08	2023/11/22 18:18	XLS 工作表	1 KB
2008_09	2023/11/22 18:19	XLS 工作表	1 KB
2009_10	2023/11/22 18:19	XLS 工作表	1 KB
2010_11	2023/11/22 18:19	XLS 工作表	1 KB
2011_12	2023/11/22 18:19	XLS 工作表	1 KB
2012_13	2023/11/22 18:19	XLS 工作表	1 KB
2013_14	2023/11/22 18:19	XLS 工作表	1 KB
2014_15	2023/11/22 18:19	XLS 工作表	1 KB
2015_16	2023/11/22 18:20	XLS 工作表	1 KB
2016_17	2023/11/22 18:20	XLS 工作表	1 KB
2017_18	2023/11/22 18:20	XLS 工作表	1 KB
2018_19	2023/11/22 18:20	XLS 工作表	1 KB
2019_20	2023/11/22 18:20	XLS 工作表	1 KB
2020_21	2023/11/22 18:20	XLS 工作表	1 KB
2021_22	2023/11/22 18:20	XLS 工作表	1 KB
2022_23	2023/11/22 18:20	XLS 工作表	1 KB
2023_24	2023/11/22 18:21	XLS 工作表	1 KB

Read and Analyse it!



	A	B	C	D	E	F	G	H	I	J
1	Team	MP	Win	Draw	Loss	GF	GA	GD	Points	
2	Mancheste	12	9	1	2	32	12	20	28	
3	Liverpool	12	8	3	1	27	10	17	27	
4	Arsenal F	12	8	3	1	26	10	16	27	
5	Tottenhar	12	8	2	2	24	15	9	26	
6	Aston Vil	12	8	1	3	29	17	12	25	
7	Mancheste	12	7	0	5	13	16	-3	21	
8	Newcastle	12	6	2	4	27	13	14	20	
9	Brighton	12	5	4	3	25	21	4	19	
10	West Ham	12	5	2	5	21	22	-1	17	
11	Chelsea F	12	4	4	4	21	16	5	16	
12	Brentforc	12	4	4	4	19	17	2	16	
13	Crystal F	12	4	3	5	12	16	-4	15	
14	Wolverhar	12	4	3	5	16	20	-4	15	
15	Nottinghe	12	3	4	5	14	18	-4	13	
16	Fulham FC	12	3	3	6	10	20	-10	12	
17	AFC Bourr	12	2	3	7	11	27	-16	9	
18	Luton Tov	12	1	3	8	10	22	-12	6	
19	Sheffielc	12	1	2	9	10	31	-21	5	
20	Everton F	12	4	2	6	14	17	-3	4	
21	Burnley F	12	1	1	10	9	30	-21	4	
22										
23										

Outline

1 Introduction
- Background

2 Mathematical Modeling
- Method 1: Entropy Weight Method in Football Team Evaluation
- Method 2: Linear Transformation and Gradient Ascend

3 Result Analysis
- Grabbing Data from Web
- Evaluating the Teams's Tier
- Prediction of 2023/24's winner point

4 Future Work
- Our plans, our limitations and how these models could improve

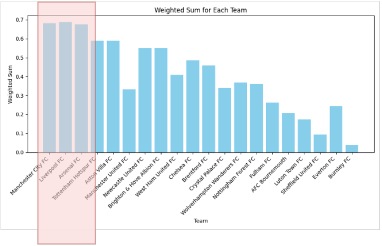
Evaluating the Teams’s Tier with Entropy Method

Attribute Weights:
MP: 0.2
Win: 0.16056986450138694
Draw: 0.15832737217198417
GF: 0.1613777266748812
GD: 0.1594934964983485
Points: 0.16023154015339913

Table of Weight Sum

	Team	Weighted Sum
0	Manchester City FC	0.681254
1	Liverpool FC	0.686918
2	Arsenal FC	0.676012
3	Tottenham Hotspur FC	0.588490
4	Aston Villa FC	0.588984
5	Manchester United FC	0.332012
...	...	...
...	...	...
...	...	...

BarChart of weight sum



After the entropy weight method calculation, the three teams in the first echelon after summing are also the current top scorers.

Outline

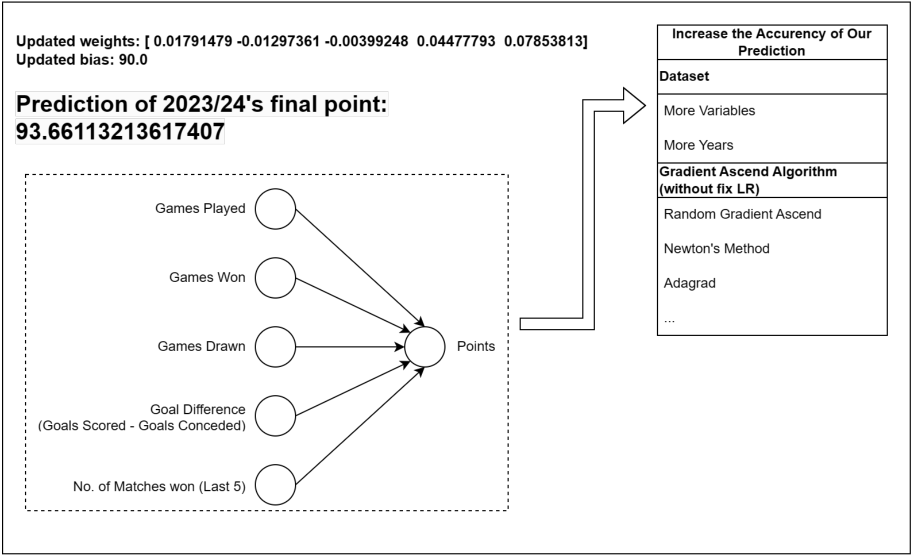
1. Introduction
   - Background

2. Mathematical Modeling
   - Method 1: Entropy Weight Method in Football Team Evaluation
   - Method 2: Linear Transformation and Gradient Ascend

3. Result Analysis
   - Grabbing Data from Web
   - Evaluating the Teams's Tier
   - Prediction of 2023/24's winner point

4. Future Work
   - Our plans, our limitations and how these models could improve

Prediction of 2023/24’s winner point



```
Updated weights: [ 0.01791479 -0.01297361 -0.00399248  0.04477793  0.07853813]
Updated bias: 90.0
```

Prediction of 2023/24's final point: 93.66113213617407

Games Played
Games Won
Games Drawn
Goal Difference (Goals Scored - Goals Conceded)
No. of Matches won (Last 5)
Points

Increase the Accurency of Our Prediction

Dataset

- More Variables
- More Years

Gradient Ascend Algorithm (without fix LR)

- Random Gradient Ascend
- Newton's Method
- Adagrad
- ...

Outline

1. Introduction
   - Background
2. Mathematical Modeling
   - Method 1: Entropy Weight Method in Football Team Evaluation
   - Method 2: Linear Transformation and Gradient Ascend
3. Result Analysis
   - Grabbing Data from Web
   - Evaluating the Teams's Tier
   - Prediction of 2023/24's winner point
4. Future Work
   - Our plans, our limitations and how these models could improve

Future plans

- Once our models are all complete, we will compare them with respect to efficiency and accuracy in predicting the winning team with this year's data as input. We could make use of Mallows' Cp and BIC statistic to establish which model has the best predictive performance.
- We will then critically analyse the most successful model and input a test dataset from earlier years to ensure we have not overfitted to our training dataset.
- To improve this model, it would be wise to consider variables other than the teams' performance and establish whether they are confounding and extraneous; for example, the about of money invested in each team.

What is this model useful for

1. Once the most significant variables have been established, a researcher could potentially use the model to advise a team's manager of where best to focus their efforts, or indeed they could consult with a bookie to help them predict the odds of the game.
2. Of course, this model overlooks the human element of the game. Factors that a manager might be aware of such as player motivation, injuries, and team dynamics can have a significant impact on the outcome of a game, but they are challenging to factor into our model.
3. Ultimately though, it is impossible to control for every variable and, even if we did the result could surprise us. But this model is beneficial in finding which variables give the teams the best chance of success.