☆ UNIVERSITY OF MIAMI ☆
FINTECH BOOTCAMP
2023
PROJECT II

THE ENGLISH PREMIER LEAGUE
FINAL STANDING PREDICTOR

BY DANIEL MOLNAR

SUMMARY OF THE PROJECT

This project aims to develop a machine learning model that predicts the final standings of the Premier League football season. The prediction will be based on several factors including the total sum of all players value, their nationality, and the average age of players.

FEATURES:

TEAM'S VALUE

We theroize that there is a strong correlation between the total value of each team and it's performance in the league.

PLAYERS' NATIONALITY

We are curious to see if players' nationality play a role in how the team performs in the season.

PLAYERS' AGE

Does age matter in the success of a team? Do younger players improve team's performance of perhaps higher average age contributes to better performance?

SCOPE AND PURPOSE

WE ASSUME THAT THE FOLLOWING GROUPS MAY BENEFIT FROM OUR MODEL.

CLUBS

Clubs can make better decisions in light of this knowledge; they may be able to tailor their signing strategies based on various statistical models and machine learning predictions.

THIRD PARTIES

Fans, analysts, and bookies can all benefit from the innate knowledge of probability. While there are heaps of historical data to assist people in decision-making, only a few focus on predicting the future based on raw data.

PLAYERS

Finally, players may be able to make better decisions as to better gauge their own net worth, future transfers, contract negotiations.

DATA COLLECTION METHODS

The data for this project was sourced from Transfermarkt, and the official Premier League website. The available APIs are very limited in this niche field, and those that exist are meant for profit-making enterprises. Therefore data collection was limited to manually scraping the data from websites.

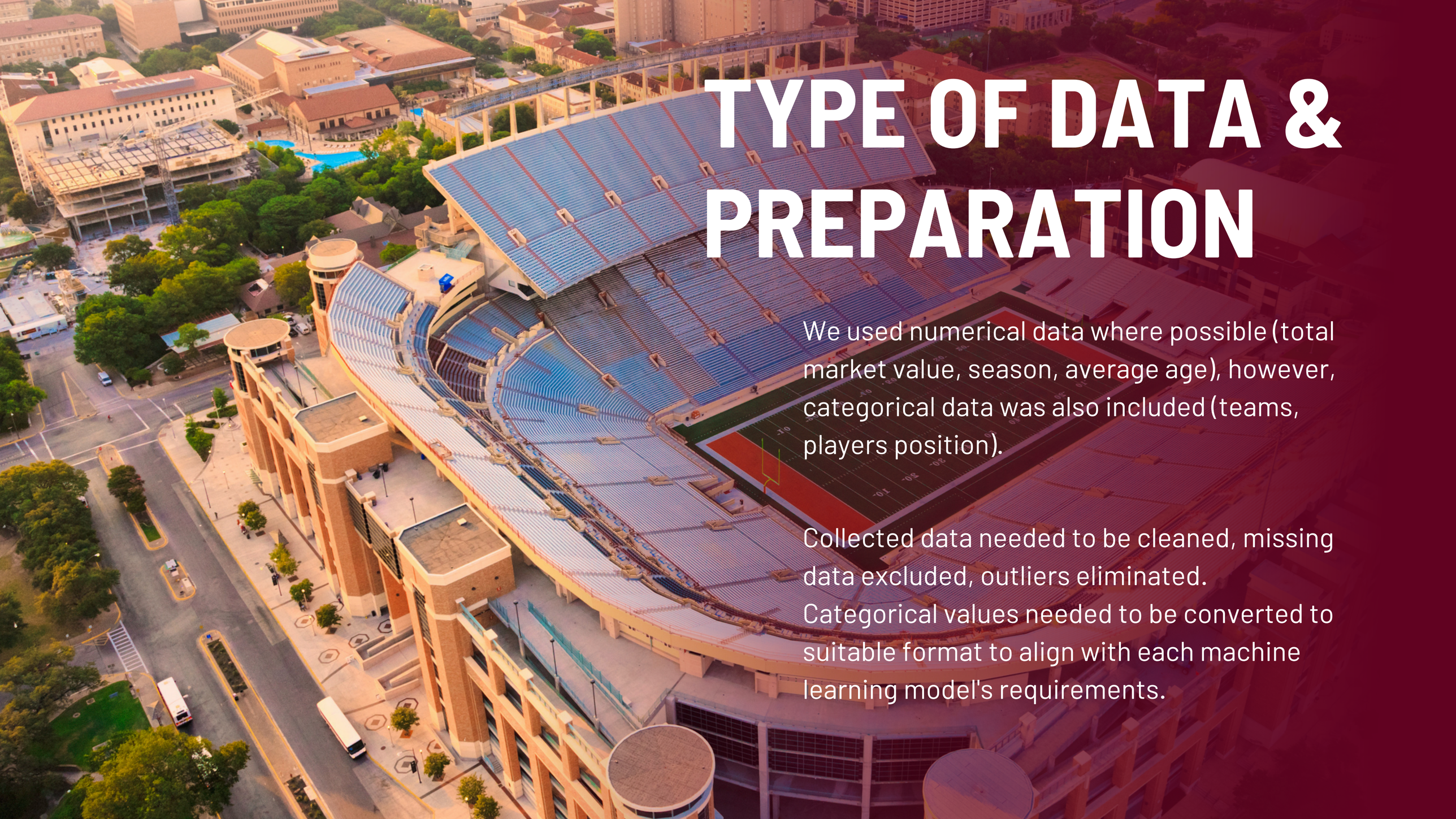
WHAT DATA DID WE COLLECT?

- Premier League standing between 2010-2013
- Players and teams market values
- Players position
- Players age
- Players nationality

LIMITATIONS

This prediction model is not free of limitations. The available data, especially historical, is difficult to verify, therefore, the results are based on the faith that the values, numbers, and information are true.
Other factors that may change the outcome of the season are listed here:

- players' market value
  - source/ estimates
- health and fitness
  - injured players
- coaching/ support staff
  - coaches, availability of resources
- team chemistry
  - nationality?
- luck



TYPE OF DATA & PREPARATION

We used numerical data where possible (total market value, season, average age), however, categorical data was also included (teams, players position).

Collected data needed to be cleaned, missing data excluded, outliers eliminated.
Categorical values needed to be converted to suitable format to align with each machine learning model's requirements.

MACHINE LEARNING MODELS USED

AND THEIR UNIQUE CHARACTERISTICS

RANDOM FOREST

1. HANDLES NON-LINEARITY
2. FEATURE IMPORTANCE
3. HANDLES OVERFITTING

GRADIENT BOOSTING

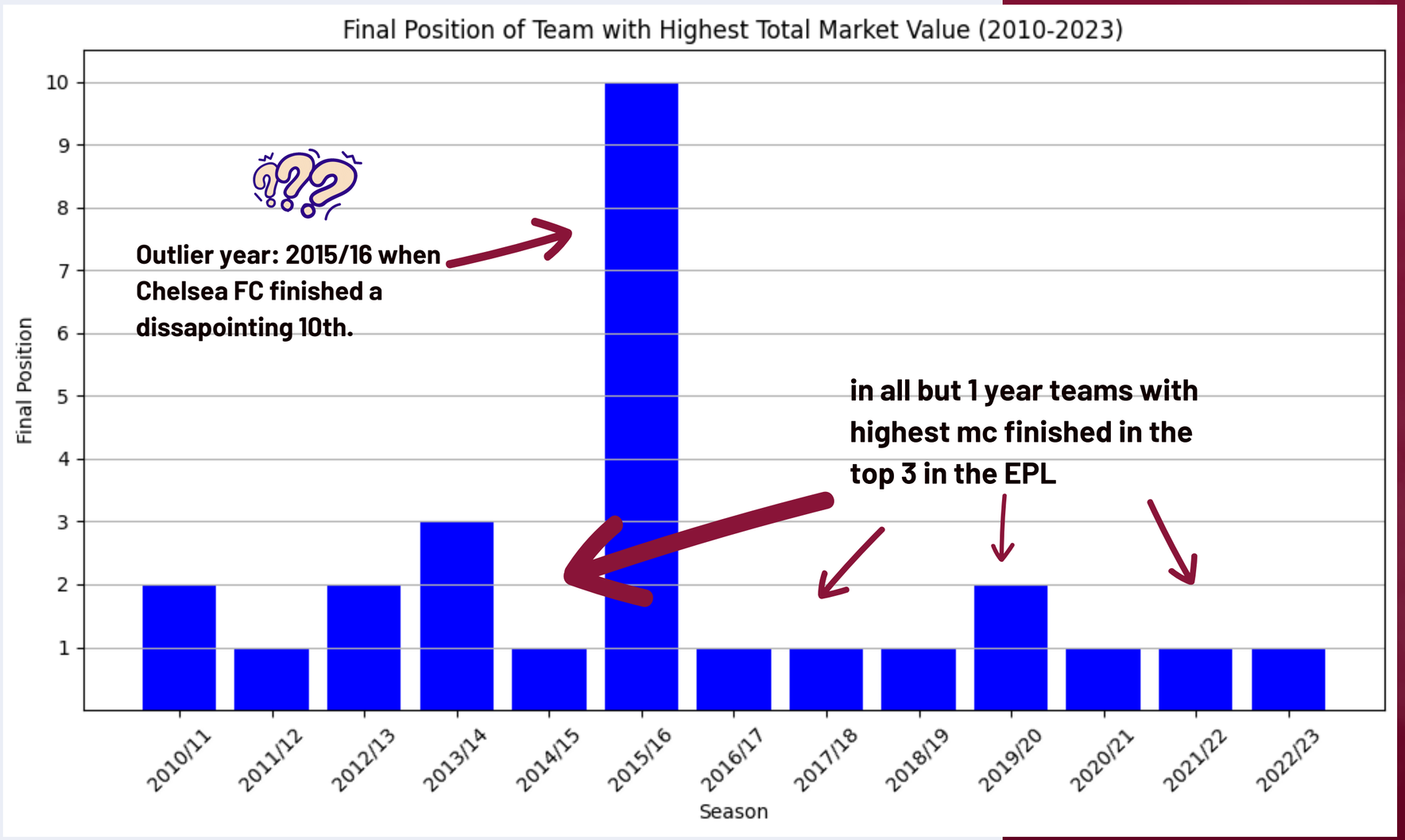
1. ENSENBLE METHOD
2. HANDLES NON-LINEARITY
3. REGULARIZATION AND AVOIDS OVERFITTING

LINEAR REGRESSION

1. INTERPRETABILITY
2. EASE OF IMPLEMENTATION
3. BASELINE MODEL

RESULTS

The graph shows the final position of the team with the highest market value in a given season.



RESULTS

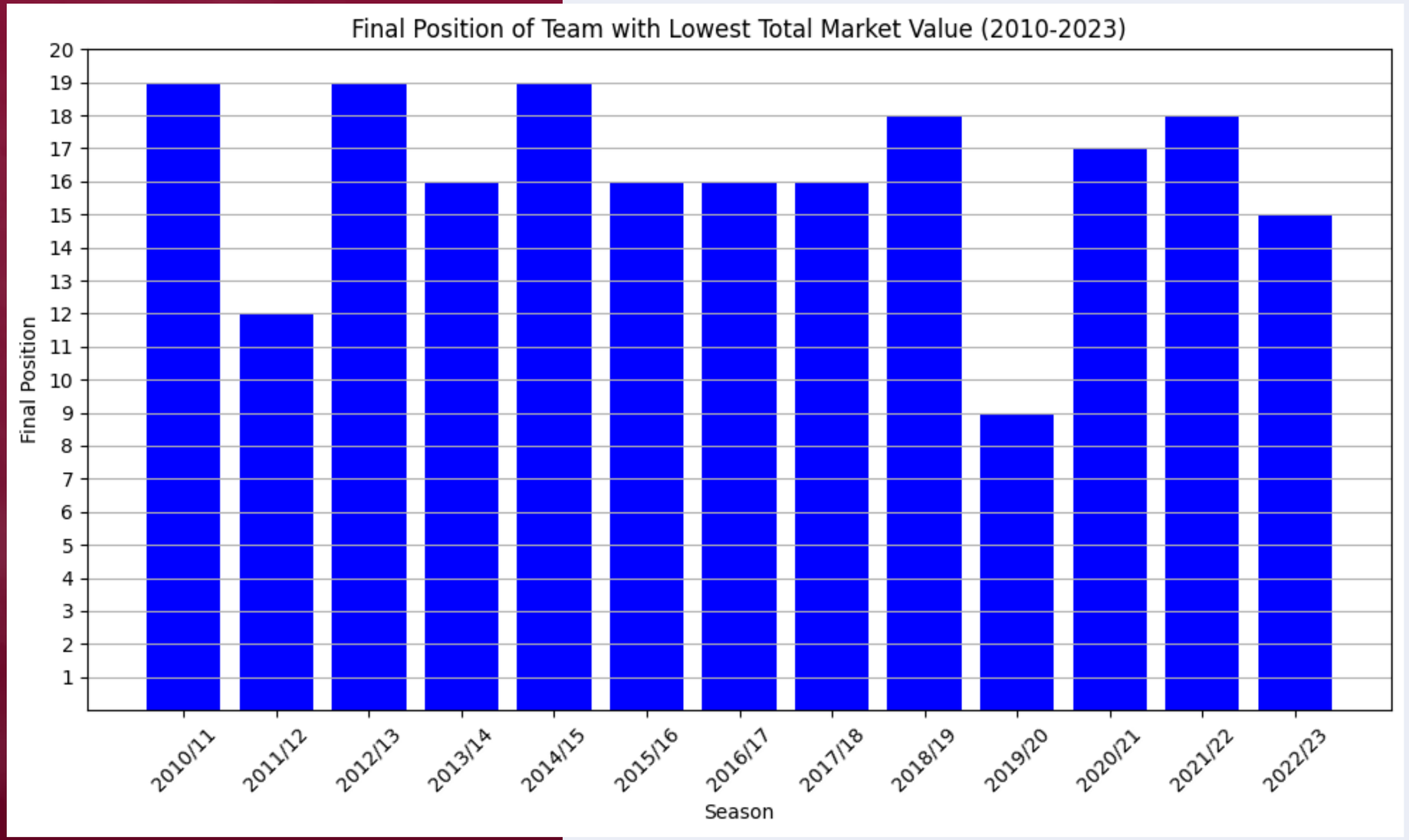
The graph shows the final position of the team with the lowest market value in a given season.

Interpretation:
The lowest mc teams have never finished higher than 9th place.

DID YOU KNOW?

FUN FACT

The lowest mc team has never finished last either!



Teams in the 90th percentile have the following probabilities:
To finish in positions 1-3: 14%
To finish in positions 4-6: 85%
To finish in positions 7-10: 1%
To finish in positions 11-15: 0%
To finish in positions 16-20: 0%

Teams in the 80th percentile have the following probabilities:
To finish in positions 1-3: 45%
To finish in positions 4-6: 29%
To finish in positions 7-10: 15%
To finish in positions 11-15: 7%
To finish in positions 16-20: 4%

Teams in the 10th percentile have the following probabilities:
To finish in positions 1-3: 0%
To finish in positions 4-6: 0%
To finish in positions 7-10: 31%
To finish in positions 11-15: 10%
To finish in positions 16-20: 59%

RESULTS

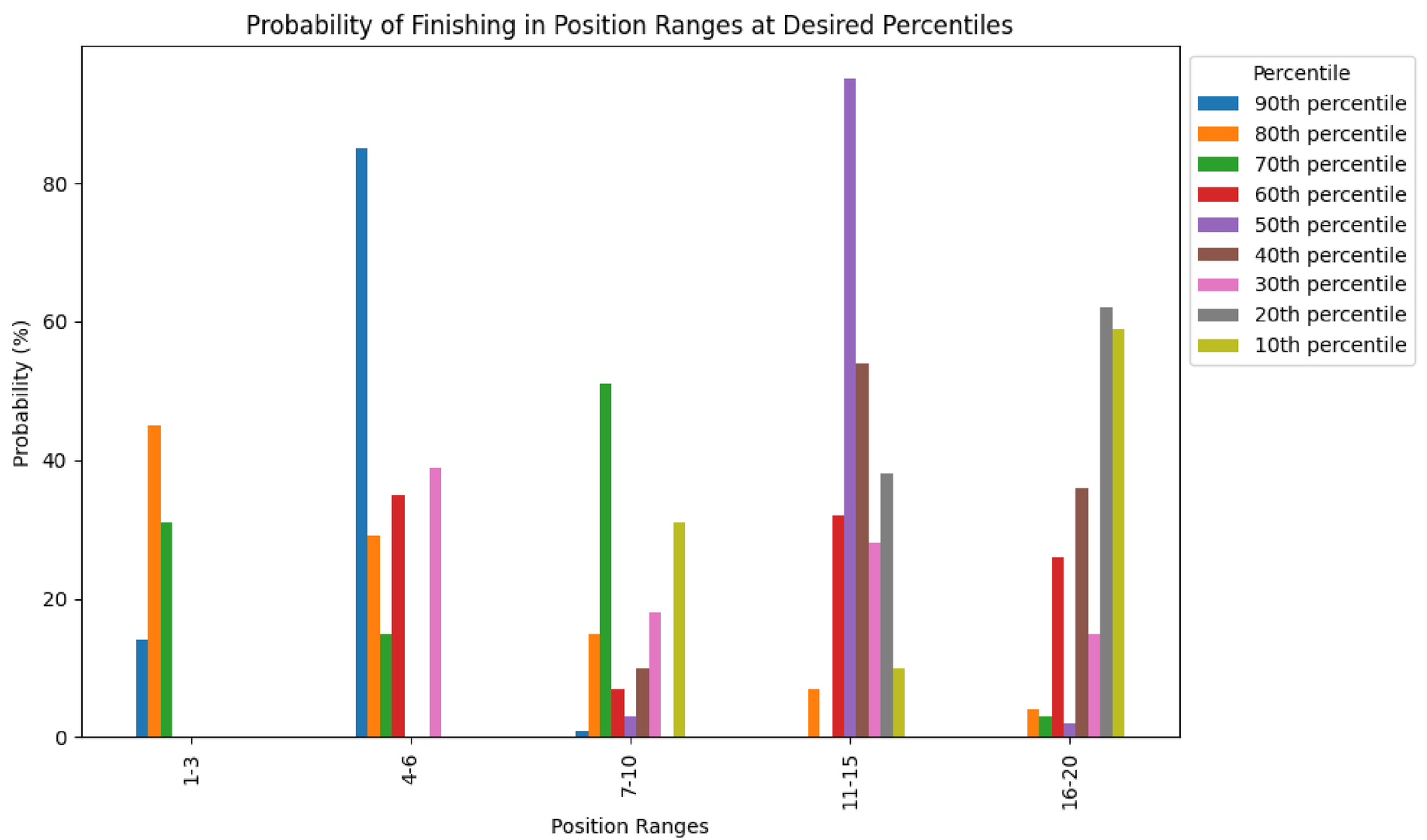
What are we seeing here?

Teams were classified in increments of 10% from 10th percentile (lowest 10% of the teams based on their Total Market Value) to 90th percentile.

The probability in % gives us the indicator about their probability how they will finish at the end of the season.

RESULTS

In plain English: Teams with the higher Total Market Values are going to finish higher in the standing.



Classification Report Based on Balanced Random Forest Classifier Model

	precision	recall	f1-score	support
1	0.33	1.00	0.50	1
2	0.00	0.00	0.00	2
3	0.00	0.00	0.00	1
4	0.00	0.00	0.00	2
5	0.14	0.25	0.18	4
6	0.00	0.00	0.00	2
7	0.00	0.00	0.00	4
8	0.00	0.00	0.00	4
9	0.00	0.00	0.00	2
10	0.00	0.00	0.00	4
11	0.50	0.33	0.40	6
12	0.00	0.00	0.00	4
13	0.00	0.00	0.00	1
14	0.00	0.00	0.00	3
15	0.00	0.00	0.00	6
16	0.00	0.00	0.00	4
17	0.00	0.00	0.00	5
18	0.00	0.00	0.00	3
19	0.20	0.20	0.20	5
20	0.00	0.00	0.00	2
accuracy			0.08	65
macro avg	0.06	0.09	0.06	65
weighted avg	0.08	0.08	0.07	65

Precision: proportion of true positives predictions among all instances.

Recall: proportion of true positives predictions among all actual instances. in that class.

F1-score: the harmonic mean of precision and recall.

Unfortunately, the overall performance of the model is generally low.
Low precision and recall values.

Side by Side Comparison

Balanced Random Forest

BRF Classification Report

	precision	recall	f1-score	support
accuracy			0.08	65
macro avg	0.06	0.09	0.06	65
weighted avg	0.08	0.08	0.07	65

Gradient Boosting Regressor

GBR Classification Report

	precision	recall	f1-score	support
accuracy			0.06	52
macro avg	0.04	0.04	0.04	52
weighted avg	0.05	0.06	0.05	52

Linear Regression

LR Evaluation Metrics

```
Mean Squared Error: 18.862298883490 49
R-squared: 0.4333680081962633
```

WHAT'S IN THE FUTURE?

1. USE DIFFERENT DATA TO FURTHER STRENGTHEN THE MODEL'S PREDICTIONS

2. CREATE PREDICTIONS FOR FUTURE SEASONS

3. USE THE PREDICTIONS FOR FUTURE TEAM AND PLAYER STRATEGIES