PROJECT II
UNIVERSITY OF MIAMI FINTECH BOOTCAMP 2023
BY DANIEL MOLNAR

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, their nationality, and the average age of players.

FEATURES:

TEAM'S VALUE

We theroize that there is a direct correllation 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.

AVERAGE AGE OF THE TEAM

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 statiscical 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.

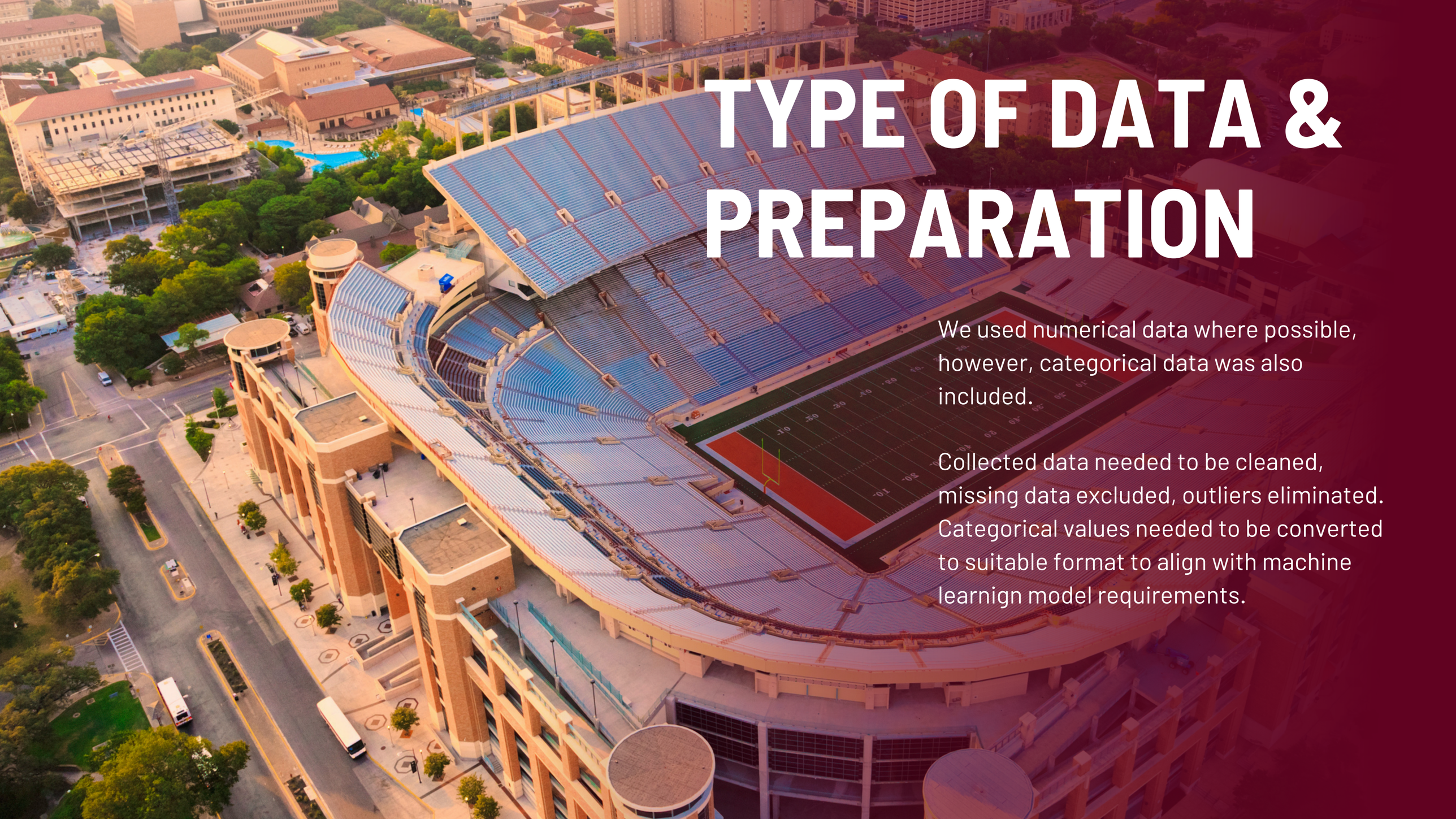
WHAT DATA DID WE COLLECT?

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

LIMITATIONS

This prediction model is not free of limitations. The available data, especially historical, cannot be verified, therefore, the result is based on the belief that the values, and numbers 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, however, categorical data was also included.

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

MACHINE LEARNING MODELS USED

24

RANDOM FOREST

12

LICERIA SUPER CUP

8

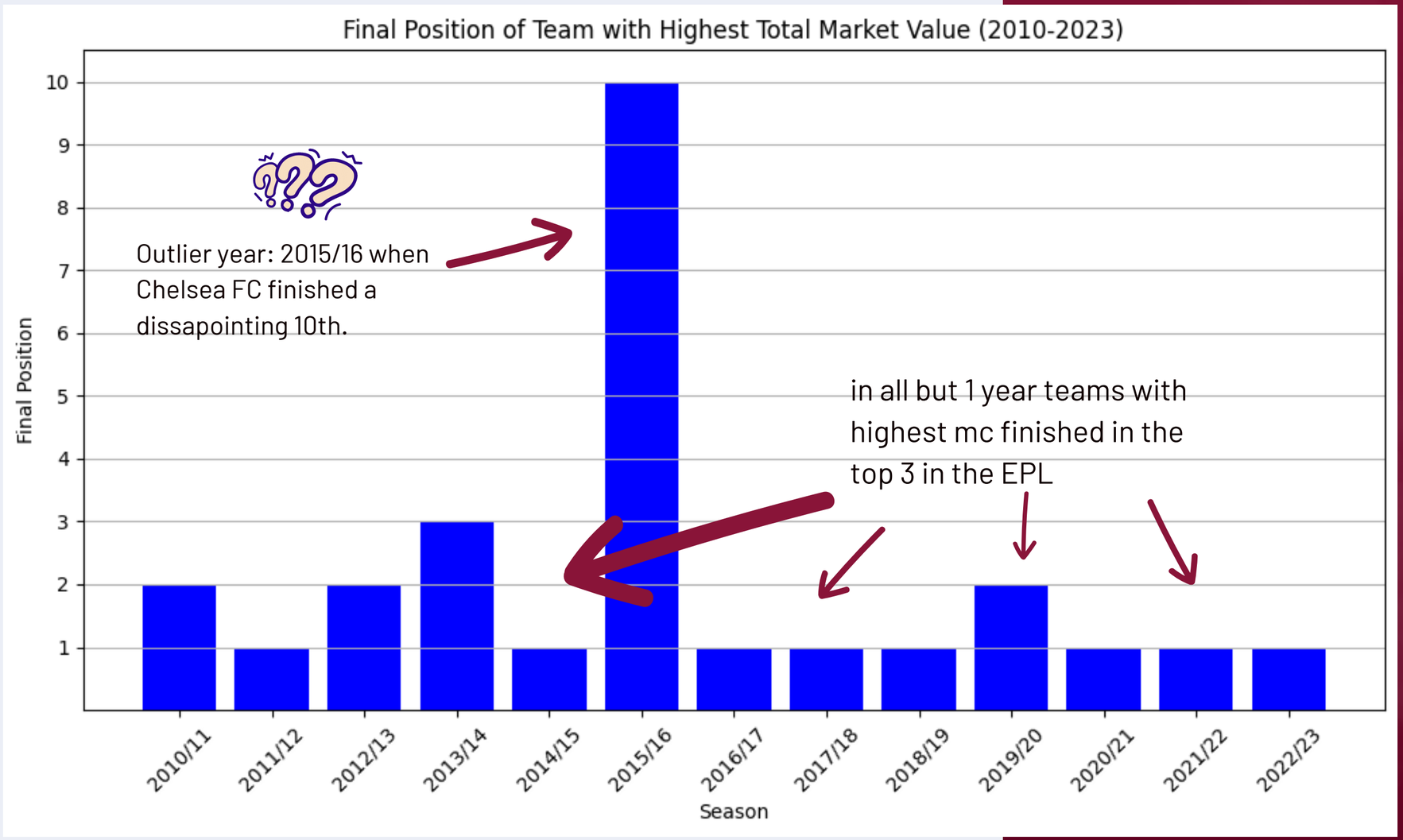
LARANA CHAMPIONS LEAGUE

16

COPA RIMBERIOFRADEL

RESULTS

Shows the final position of the team with the highest market value in a given season.



RESULTS

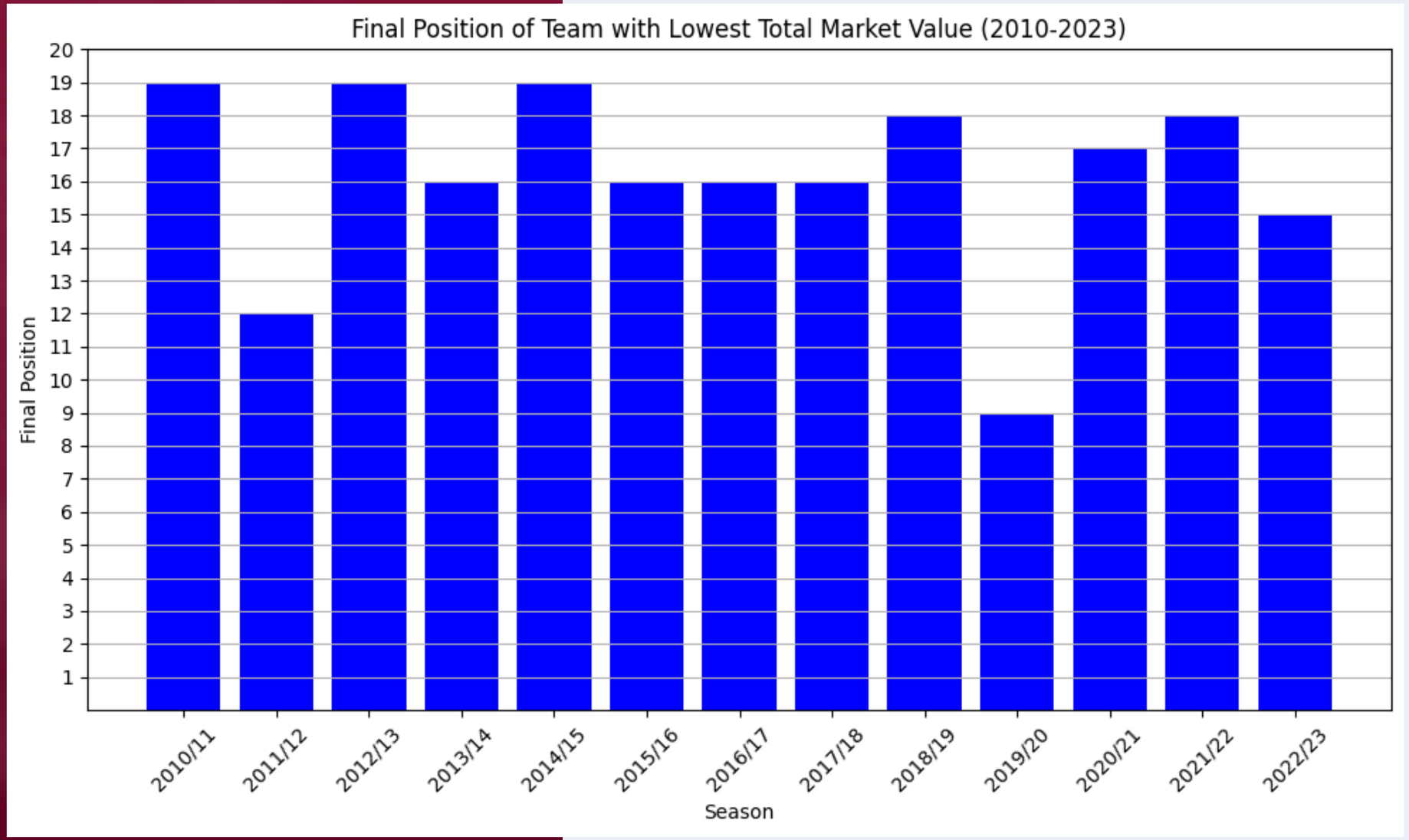
Shows the final position of the team with the lowest market value in a given season.

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%

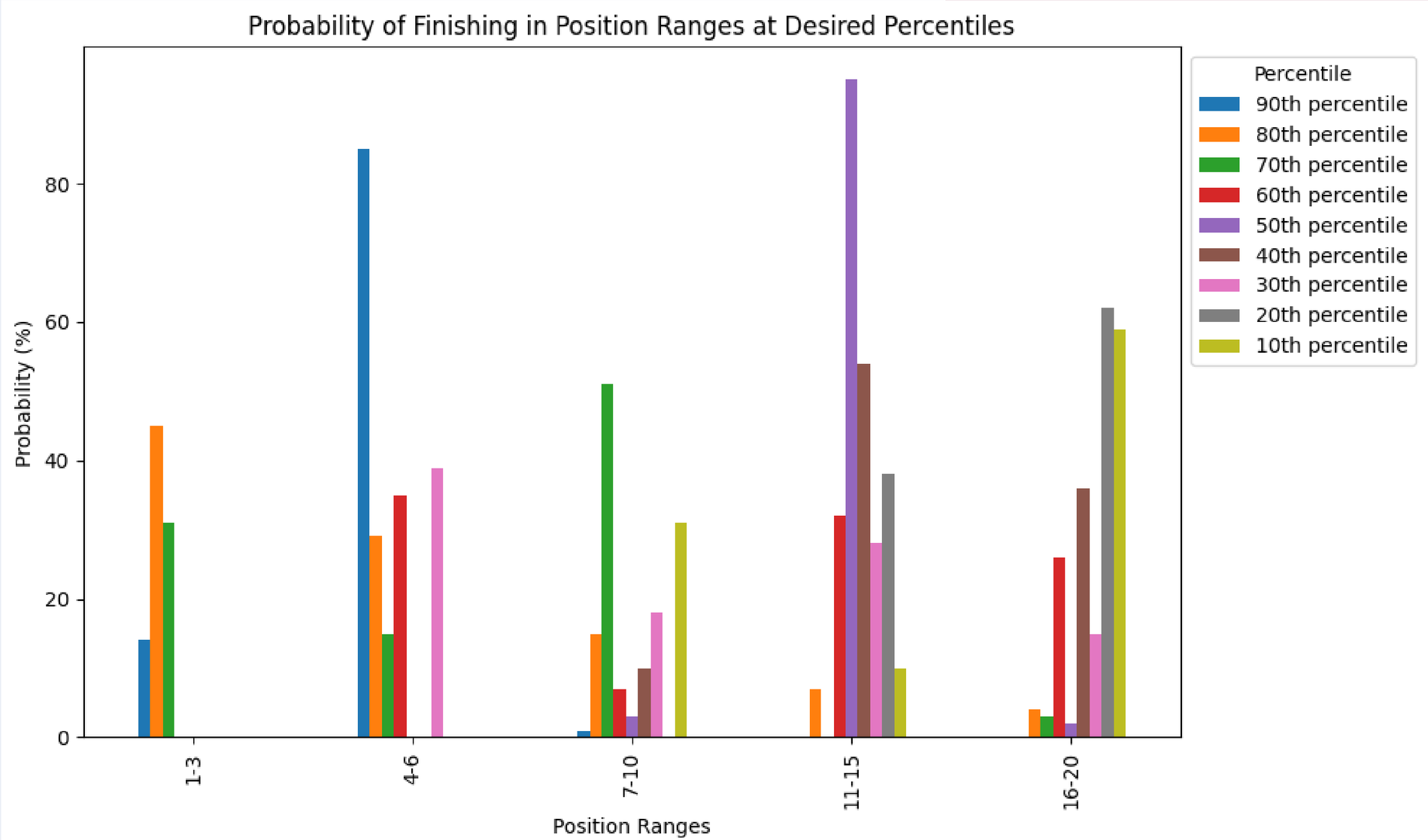
Teams in the 10th percentile have the following probabilities:
To finish in positions 1-3: 0%
To finish in positions 4-6: 0%
To finish in positions 7-10: 31%
To finish in positions 11-15: 10%
To finish in positions 16-20: 59%

RESULTS

We Received the Results in the Following Format

In plain English: Teams with higher mc are going to finish higher in the standing.

RESULTS

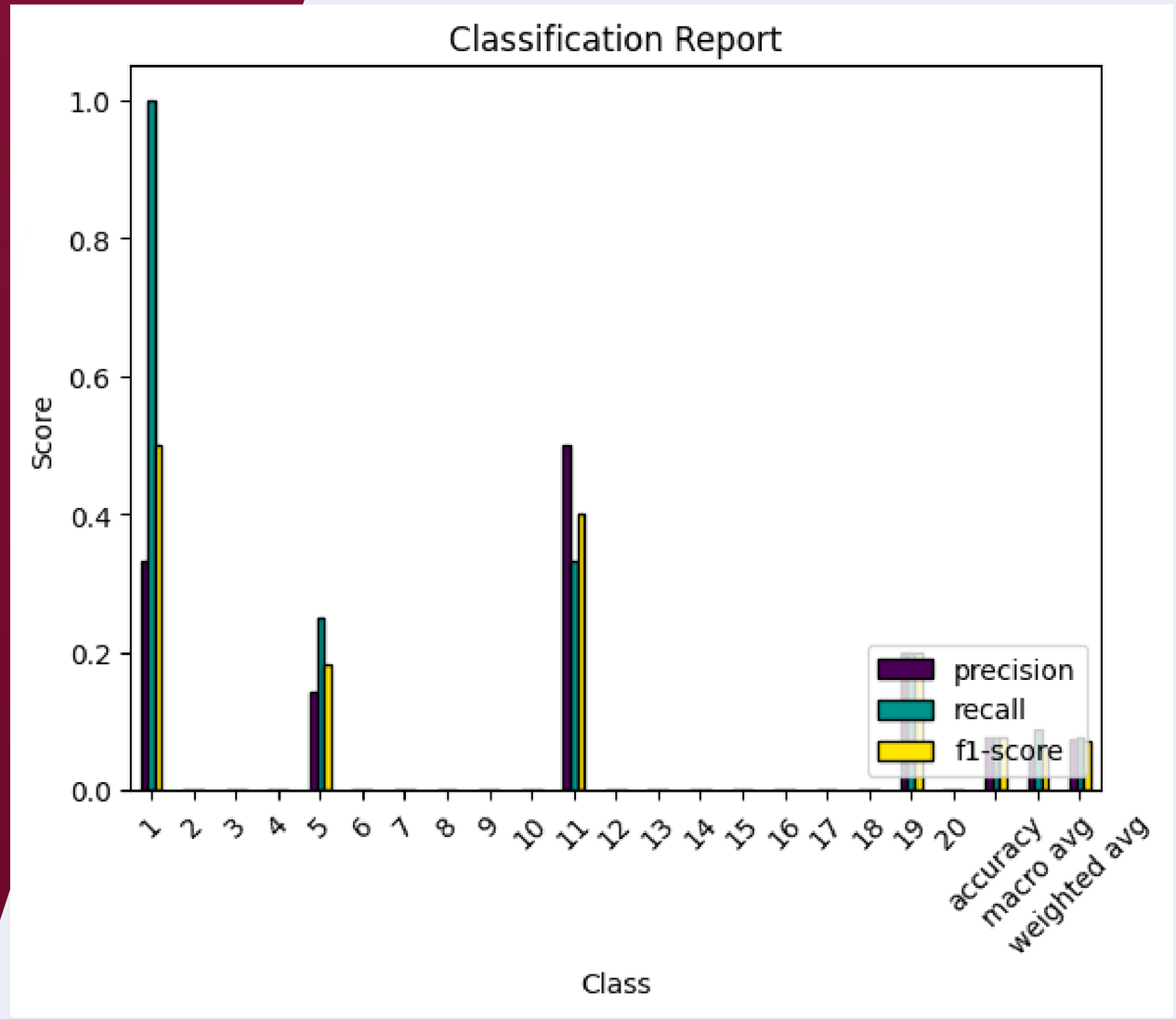


What are we seeing?

Teams in certain percentile brackets (90th, 80th, etc), color coded, and their respective probability to finish in a certain position range (1-3, 4-6, etc)

Classification Report Based on Balanced Random Forest Classifier Model's Results

	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



WHAT'S IN THE FUTURE?

1. CREATING PREDICTIONS FOR FUTURE SEASONS

2. CREATING PREDICTIONS FOR FUTURE TEAM STRATEGIES

3. USE FUTURE DATA TO FURTHER STRENGTHEN THE MODEL'S PREDICTIONS