U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Data Mining Project

Aprendizagem Computacional (AC) - 2024

G45
Miguel Lima - up202108659@edu.fe.up.pt
Pedro Romão - up202108660@edu.fe.up.pt

Contents

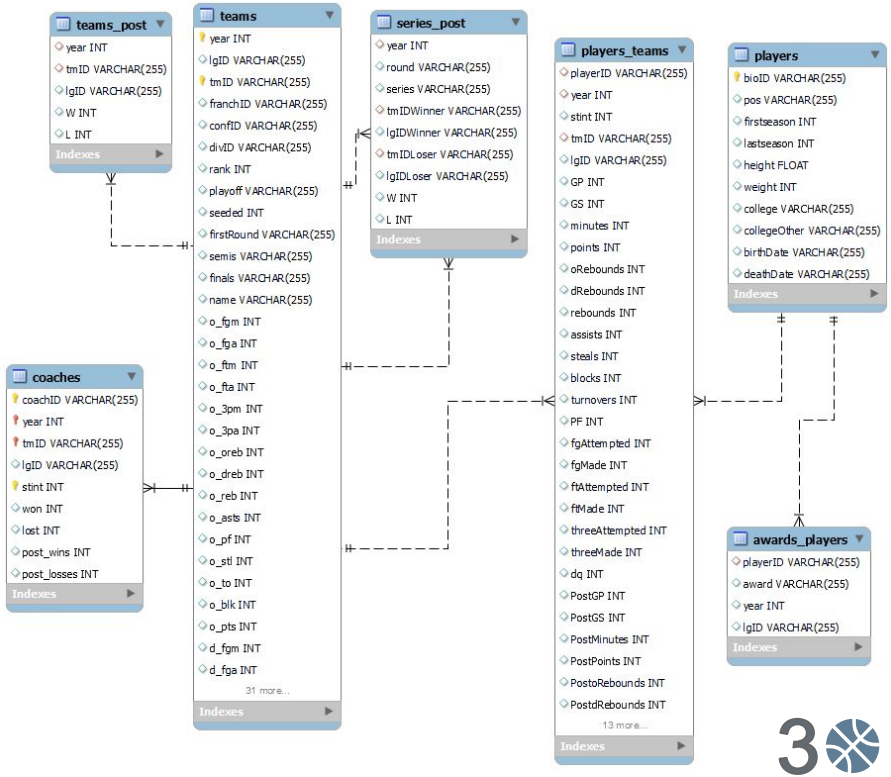
Domain Description

WNBA Competition Structure:
The WNBA season has two parts:

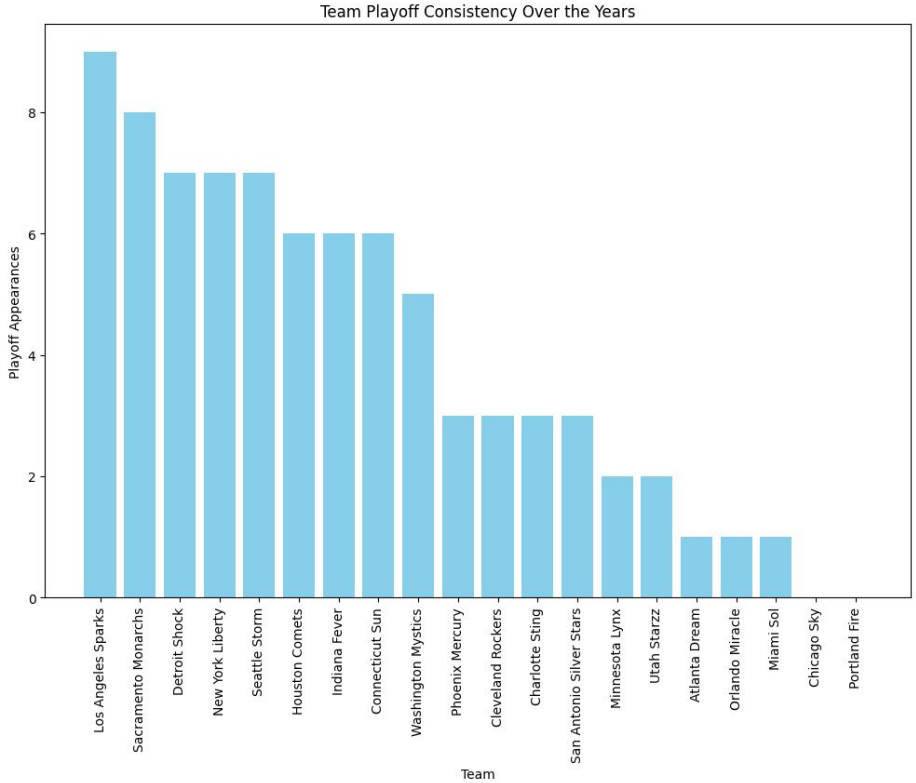
- **The Regular Season**: Teams compete to secure a playoff spot, with the best-performing teams advancing.
- **The Playoffs**: A tree-style tournament where teams compete for the championship. Regular season performance determines who gets to this stage.

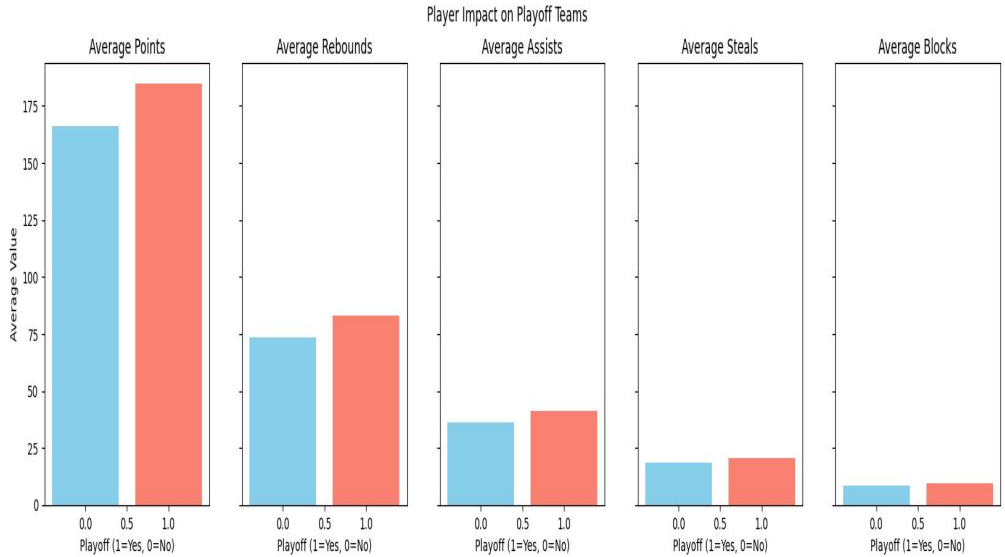
Dataset:
The dataset spans **10 seasons** of WNBA data, covering detailed information on **players**, **teams**, **coaches**, and **game metrics**. It includes records of **57** coaches, **893** players, and **20** teams.

The goal is to use this data to **train machine learning models** to **predict** which teams will **qualify** for the playoffs.

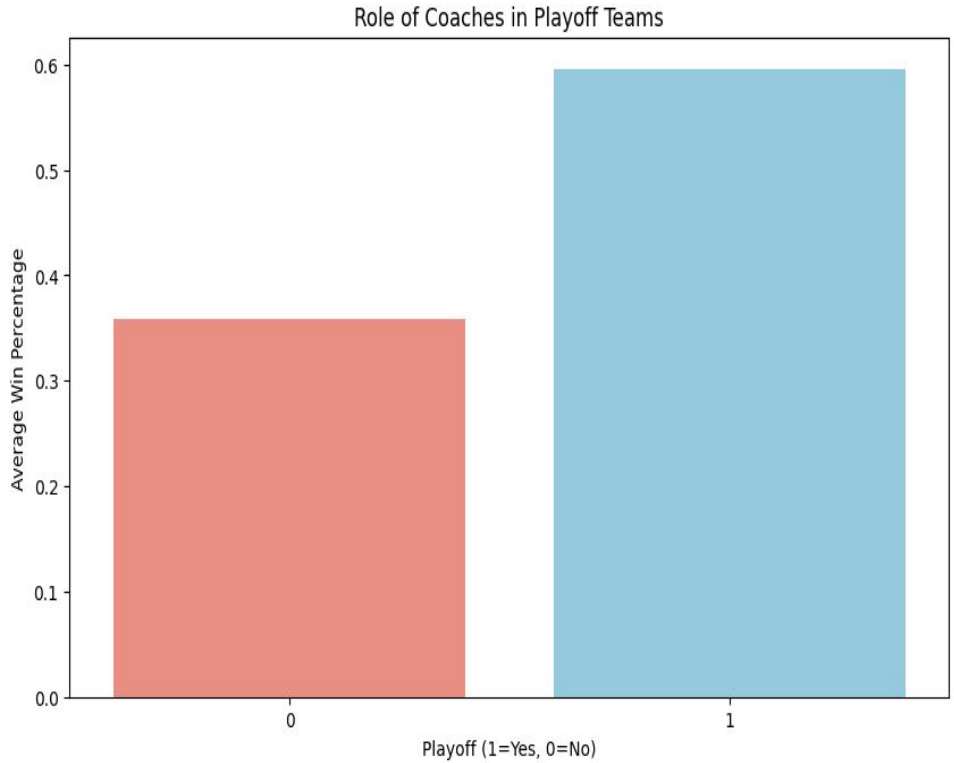


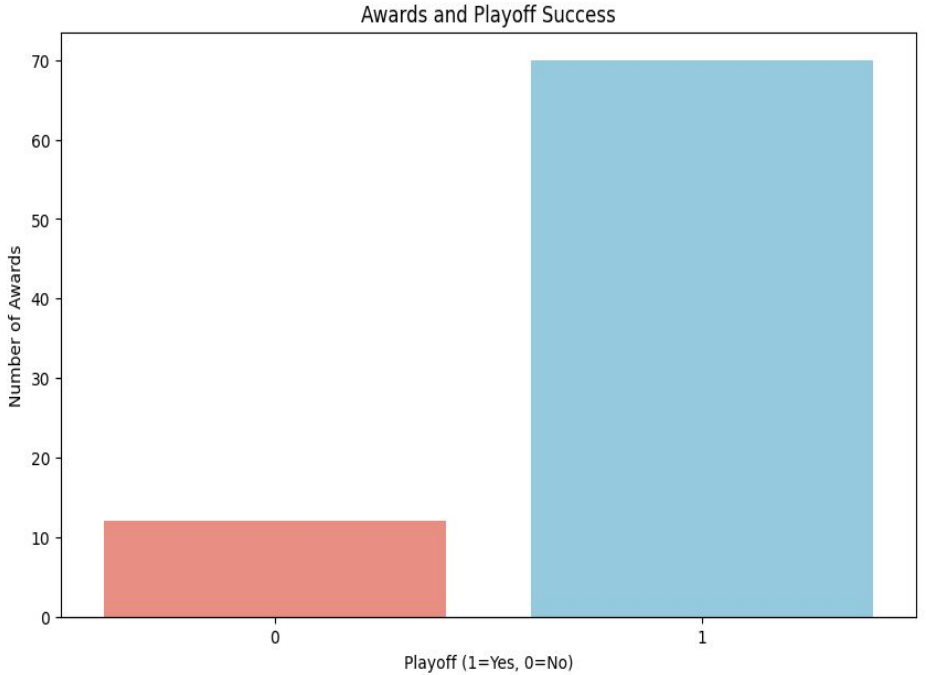
Exploratory Data Analysis (1/3)



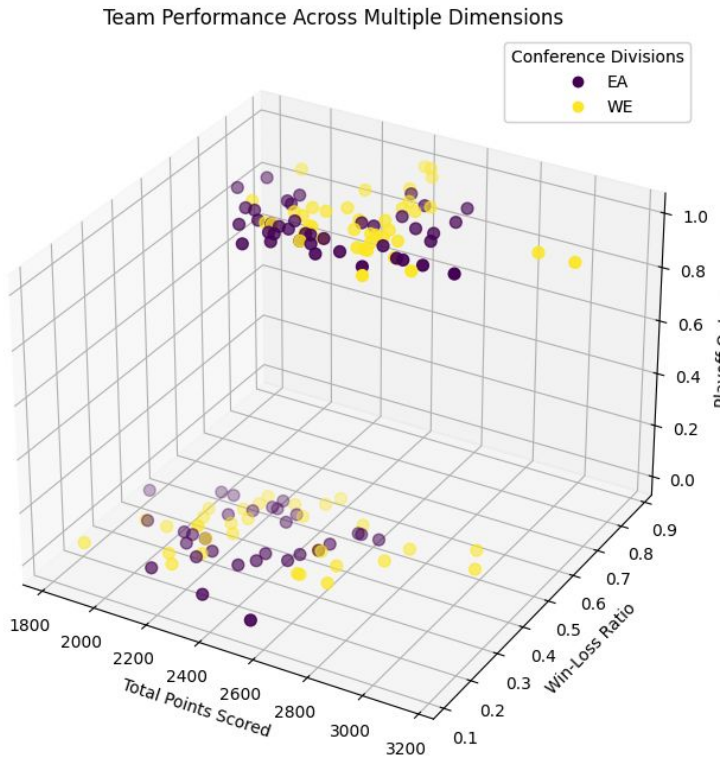


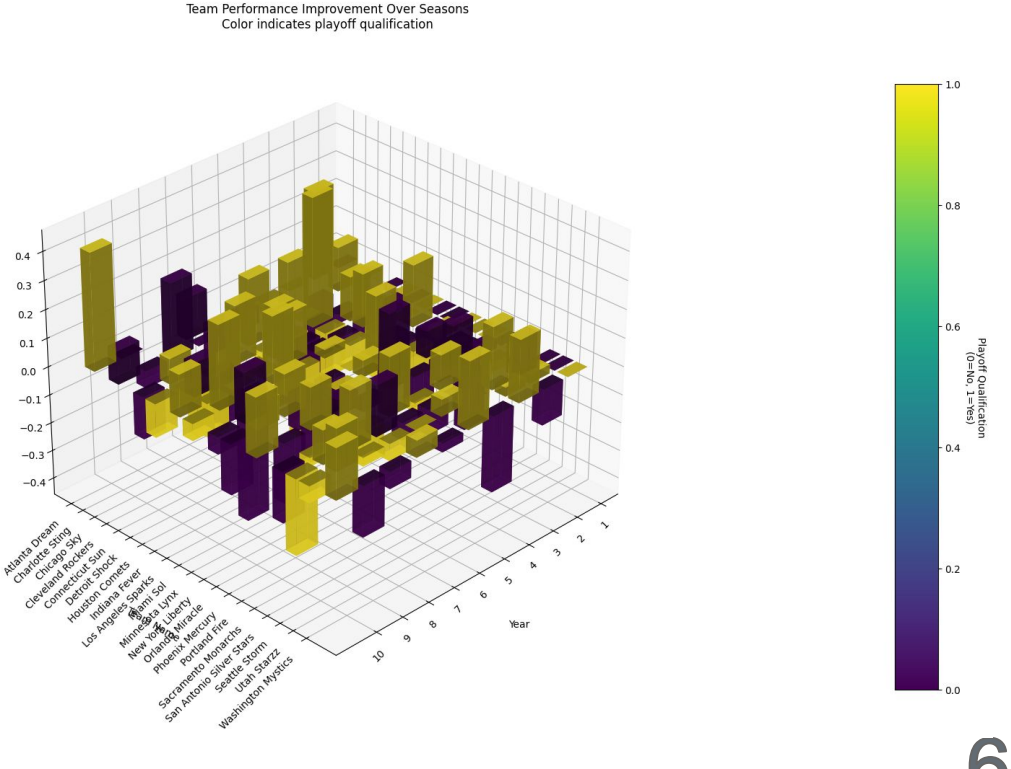
Exploratory Data Analysis (2/3)





Exploratory Data Analysis (3/3)





Problem Definition

In each season the competition consists of two phases.

First, all teams play each other aiming to achieve the greatest number of wins possible.

Following that, a predetermined number of teams with the highest number of wins advance to the playoffs

- Objective

Predict which teams will qualify for the playoffs next season.

- Data

Teams, players, coaches, awards, game statistics and other metrics

- Success Criteria

Evaluate model performance using metrics like accuracy, recall, f1-score, etc

.

Data preparation (1/4)

Remove redundant features:

- remove ‘lgID’ from all tables because its the same WNBA league
- teams: drop ‘divID’, ‘seeded’ attributes because they are not meaningful for the project
- players: erase ‘colegeOther’ and ‘deathDate’ because it had a lot of missing values

Join tables on IDs:

- join players, awards_players and players_teams on playerID and year
- join coaches and awards_players on coachID and year
- join teams and teams_post on year and teamID

3 new organized tables:

- information relative to the players and awards
- information relative to the coaches and awards
- information relative to the teams statistics

Data preparation (2/4)

Outlier Detection:
We used Z-Score Algorithm to detect outliers, although we saw a few, all values make sense in the context of our project.

Feature Engineering
Create new features to improve the performance of the machine learning model.

- Extract age from player birthdate (and drop the birthdate feature)
- Generate a column with the count of coach and player awards for each team and year
- Generate columns with the difference between the offensive and defensive statistics for each team in each year (eg. field goals scored - field goals scored by opponents, etc..)

Data preparation (3/4)

New columns with the difference between offensive and defensive statistics for each team in each year:

- difference between field goals scored and field goals scored by opponents
- difference between team offensive rebounds and opponent team offensive rebounds
- etc..

```
teams['diff_fgm'] = teams['o_fgm'] - teams['d_fgm']
teams['diff_fga'] = teams['o_fga'] - teams['d_fga']

teams['diff_ftm'] = teams['o_ftm'] - teams['d_ftm']
teams['diff_fta'] = teams['o_fta'] - teams['d_fta']

teams['diff_3pm'] = teams['o_3pm'] - teams['d_3pm']
teams['diff_3pa'] = teams['o_3pa'] - teams['d_3pa']
```

Data preparation (4/4)

Add coach and players awards to the metrics:

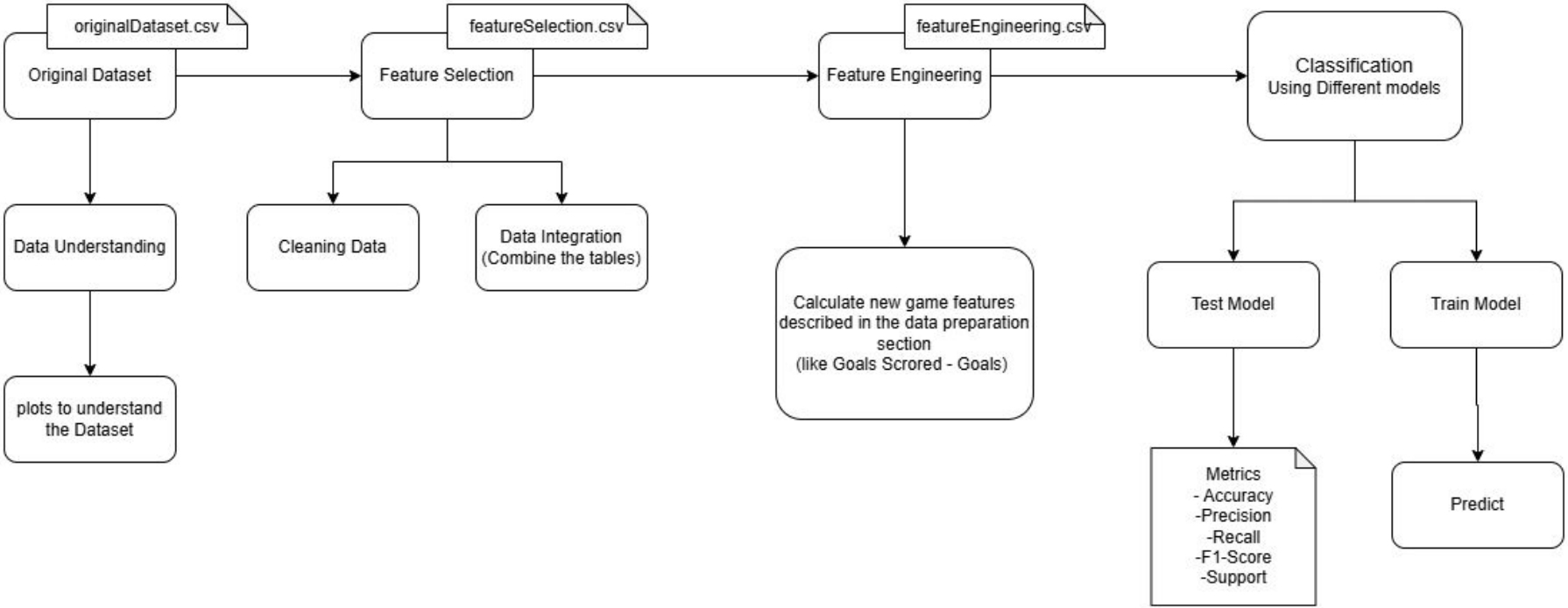
Create a new column called awards for each team each year

- If the coach was awarded coach of the year increment one in the awards teams column
- For each year, count the number of player awards earned by a team and update the awards column with incrementing that count.

Notes on Inconsistency in the dataset
Comparing some statistics don't exactly match.
Eg. Comparing team points and the points scored by every player of a team in a year don't exactly match. Same applies to the rebounds.

```
For team 'HOU' in year 3
Sum of players points: 2664
Team points: 2072
Sum of players rebounds: 1159
Team rebounds: 1001
```

Experimental Setup (1/4) - Pipeline



Experimental Setup (2/4)
Use of Sliding Window

Designed to analyze team performance across different years, enabling the capture of temporal patterns and accounting for changes over time.

Helps the model leverage historical data to adapt to evolving team dynamics as seasons progress.

Maintains prediction relevance by focusing on recent team trends rather than outdated data.

Ensures predictions remain accurate and reliable across multiple years.

We used the mean difference in teams statistics from **the actual and the previous year** to predict the next year playoffs.

Notes: We tried different numbers of years from 2 to 5 and 2 (the actual and the previous) got the highest correlation with the next year playoffs

Experimental Setup (3/4)

As we can see, the average of the current and previous year statistics features have the higher values of correlation with the Next Year Playoffs

```
Features Correlation with Next Year Playoff (Ordered by Correlation Value)
mean_diff_reb      0.363968
mean_diff_awayW    0.353762
mean_diff_dreb     0.343081
mean_diff_won      0.338469
diff_reb           0.337612
```

Experimental Setup (4/4)

After having the features we explored 5 different models:
Random Forest, **Logistic Regression**, **Support Vector Machine**, **K-Nearest Neighbors** and **Decision Tree**

Best Features Selection
For each model we used **SelectKBest** to select the best features for each model for better accuracy.

Hyperparameter Tuning
After selecting the best features we also used **GridSearch** to select the best parameters for each model.

Results (1/2) - Chosen models

For each model with the best features and best parameters using 10-fold cross validation:

We got Support Vector Machine with a mean accuracy of 0.77
We got K-Nearest Neighbors with an accuracy of 0.54
We got Random Forest with a mean accuracy of 0.54
We got Decision Tree with an accuracy of 0.54
We got Logistic Regression with an accuracy of 0.46

Classification Report for Support Vector Machine:
Accuracy: 0.77

	precision	recall	f1-score	support
0.0	0.67	0.80	0.73	5
1.0	0.86	0.75	0.80	8
accuracy			0.77	13
macro avg	0.76	0.78	0.76	13
weighted avg	0.78	0.77	0.77	13

(The metrics for the other models are in the annexes)

Results (2/2) - Prediction for year 10 playoffs

Team	Playoff	Probability	Difference
ATL	1	0.729755	0.270245
CHI	0	0.490509	0.490509
CON	0	0.67072	0.67072
DET	1	0.632019	0.367981
IND	1	0.615232	0.384768
LAS	1	0.703917	0.296083
MIN	0	0.568825	0.568825
NYL	0	0.541911	0.541911
PHO	1	0.540328	0.459672
SAC	0	0.575478	0.575478
SAS	1	0.692577	0.307423
SEA	1	0.695371	0.304629
WAS	1	0.543358	0.456642

Theoretical Max error: 13
Theoretical Min error: 0
Error: 5.694

Predictions – Prediction for year 11 playoffs

Team	Probability
ATL	0.615385
CHI	0.615385
CON	0.716084
IND	0.671329
LAS	0.794406
MIN	0.604196
NYL	0.682517
PHO	0.682517
SAS	0.682517
SEA	0.727273
TUL	0.559441
WAS	0.648951

Theoretical Max error: 12
Theoretical Min error: 0
Error: ?

Notes on year 11:

- We also explored another approach in the Prediction11.ipynb notebook, attributing the mean statistics for each player/coach from all seasons. And for the year 11, group them by team (summing performance of each individual of a team)

- This approach was logical, but the model results were worse than those from the approach described earlier in the ProjectNotebook.ipynb. Since the approach was still reasonable, we included it in the code submission and deemed it important to mention it in the notes for this report.

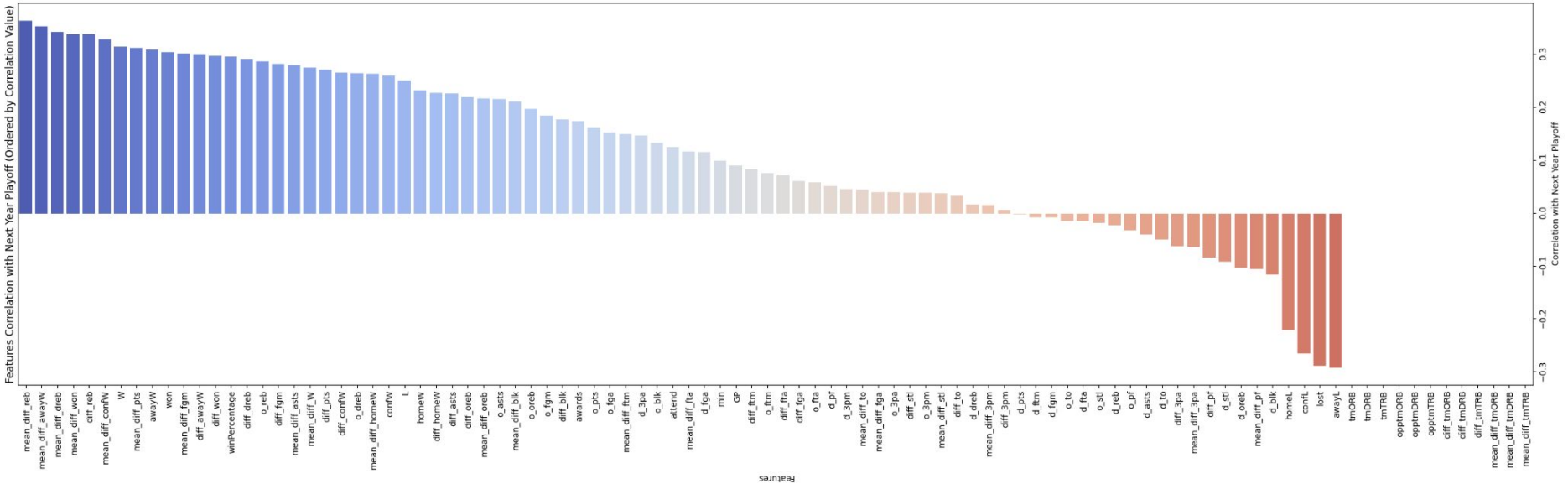
- Concluding the prediction for year 11 are from the main notebook, the ProjectNotebook.ipynb.

- The predictions for year 11 are on the playoff_normalized_ano11.csv and the playoff_predictions_ano11_binary.csv files.

Conclusions, limitations and future work

- Effective Exploratory Data Analysis
- Strong Feature Engineering
- Solid performance on the Tuned Model
- Room for improvement in model tuning for some cases
- Post-performance analysis suggests that a larger dataset could provide deeper insights into model behavior

Annexes



Correlation between the features and the next year playoffs

Classification Report for K-Nearest Neighbors:
Accuracy: 0.54

	precision	recall	f1-score	support
0.0	0.33	0.20	0.25	5
1.0	0.60	0.75	0.67	8
accuracy			0.54	13
macro avg	0.47	0.47	0.46	13
weighted avg	0.50	0.54	0.51	13

Classification Report for Decision Tree:
Accuracy: 0.54

	precision	recall	f1-score	support
0.0	0.33	0.20	0.25	5
1.0	0.60	0.75	0.67	8
accuracy			0.54	13
macro avg	0.47	0.47	0.46	13
weighted avg	0.50	0.54	0.51	13

Evaluation of K-Nearest Neighbors and Decision Tree

Classification Report for Random Forest:
Accuracy: 0.54

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	5
1.0	0.58	0.88	0.70	8
accuracy			0.54	13
macro avg	0.29	0.44	0.35	13
weighted avg	0.36	0.54	0.43	13

Classification Report for Logistic Regression:
Accuracy: 0.46

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	5
1.0	0.55	0.75	0.63	8
accuracy			0.46	13
macro avg	0.27	0.38	0.32	13
weighted avg	0.34	0.46	0.39	13

Evaluation of Random Forest and Logistic Regression