Data mining

Basketball playoffs qualification

Developed by:

Bárbara Carvalho	up202004695
Lucas Sousa	up202004682
Luís Cabral	up202006464

Introduction

- Basketball tournaments are usually split in two parts.
- **First**, **all teams play each other** aiming to achieve the greatest number of wins possible.
- Then, **a predetermined number of teams which were able to win the most games are qualified to the playoff season**, where they play series of knockout matches for the trophy.
- For 10 years, data from players, teams, coaches, games and several other metrics were gathered and arranged on this dataset.
- Goal: use this data to **predict which teams will qualify for the playoffs in the next season.**

Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was followed, in particular the following stages:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation

Business Understanding - Playoff Qualifying

4 WC Teams

Meet In the Middle

4 EC Teams

Western Conference

TEAM		W	L
1 SEA	✕	28	6
2 PHO	✕	15	19
3 LVA	✕	14	20
4 LAS	✕	13	21
5 MIN	○	13	21
6 DAL	○	6	28

Conference Semi-Finals Best-of-3			Conference Finals Best-of-3			WNBA Finals Best-of-5		
E1	*Washington*	0						
E4	**Atlanta**	**2**	E2	*New York*	0			
	Eastern Conference		E4	**Atlanta**	**2**			
E2	***New York***	**2**						
E3	Indiana	1				E4	Atlanta	0
W1	***Seattle***	**2**				W1	***Seattle***	**3**
W4	Los Angeles	0	W1	***Seattle***	**2**			
	Western Conference		W2	Phoenix	0			
W2	***Phoenix***	**2**						
W3	San Antonio	0						

Eastern Conference

TEAM		W	L
1 WAS	✕	22	12
2 NYL	✕	22	12
3 IND	✕	21	13
4 ATL	✕	19	15
5 CON	○	17	17
6 CHI	○	14	20

Data Understanding

- Figure out the Statistical Categories
  - PTS - Points
  - AST - Assists
  - DFGM - Field goals made by the opponent while the player or team was defending the rim

Data Cleaning

- Fixed missing or erroneous attribute values

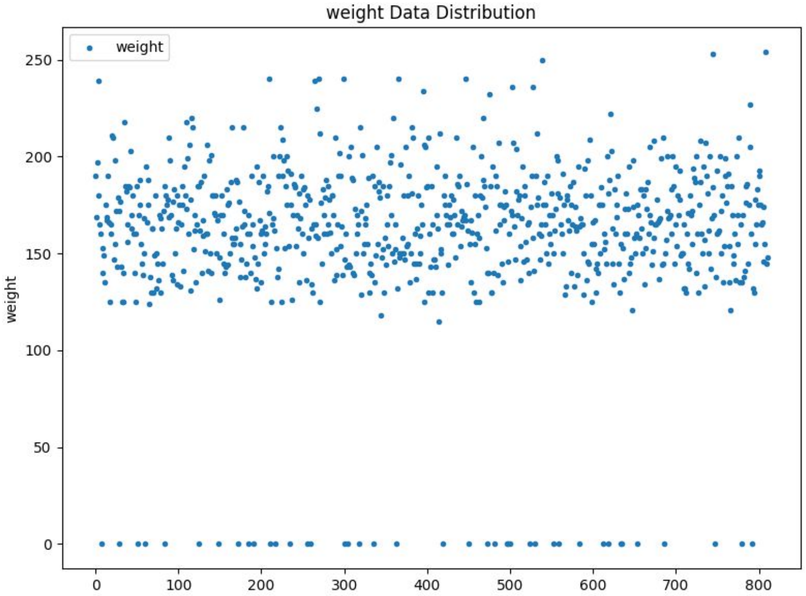
    - Kim Perrot Sportsmanship -> Kim Perrot Sportsmanship Award

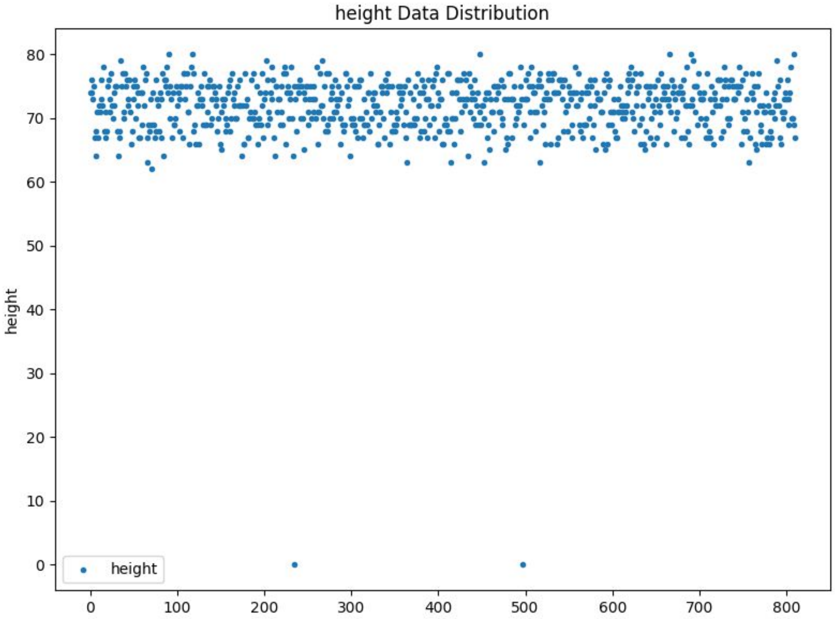
        - Changed misspelled Award

    - Use BMI to fill player's missing height or weight

- Outliers removed
- Records removed when lacking attribute values

Data Cleaning

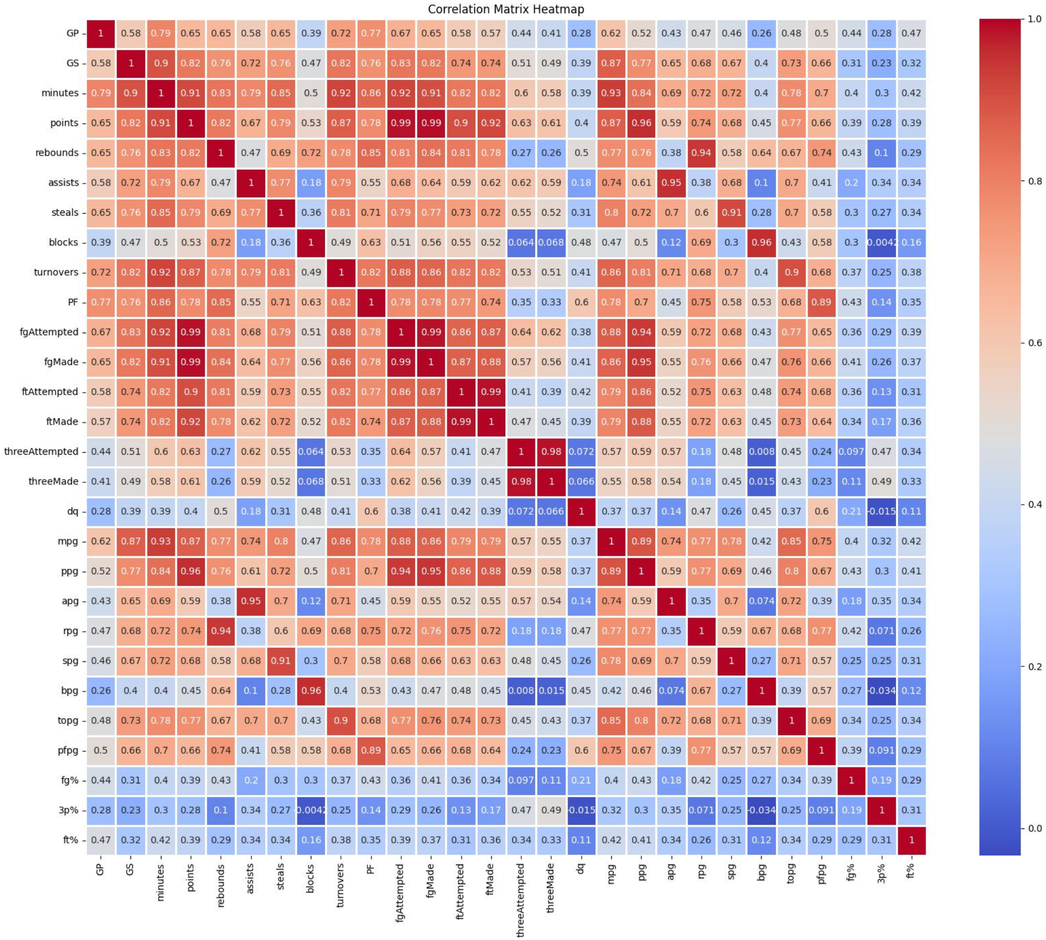




Data Preparation - Integration

- Correlation Matrix for players_teams.csv

- Since many values were heavily correlated, we decided to simplify the data
  - By rationing
  - By removing



Data Preparation - Integration

Firstly we got rid of the total stats and just kept the per-games, since there was no point in having both.
We then created new columns to reduce some abnormal correlation values.

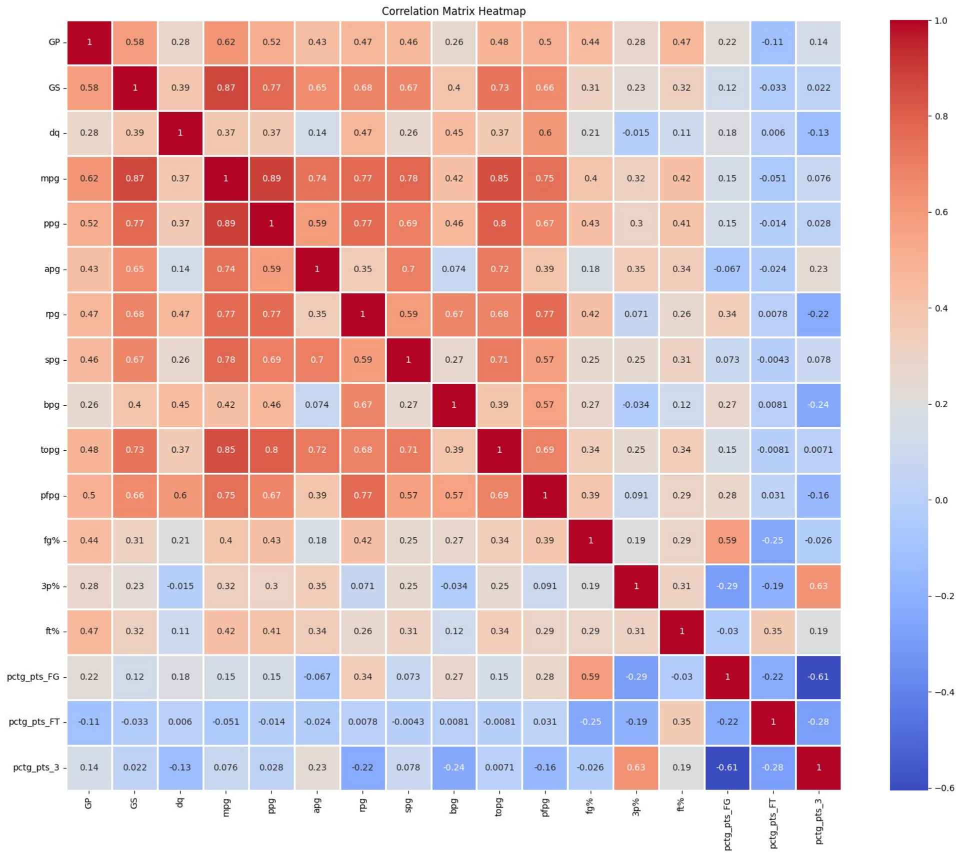
Simple ones, where we just calculated percentages:

FG% = FGM / FGA
3P% = 3PM / 3PA
FT% = FTM / FTA

And harder ones:

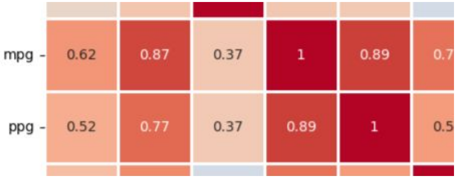
pctg_pts_FG
pctg_pts_FT
pctg_pts_3

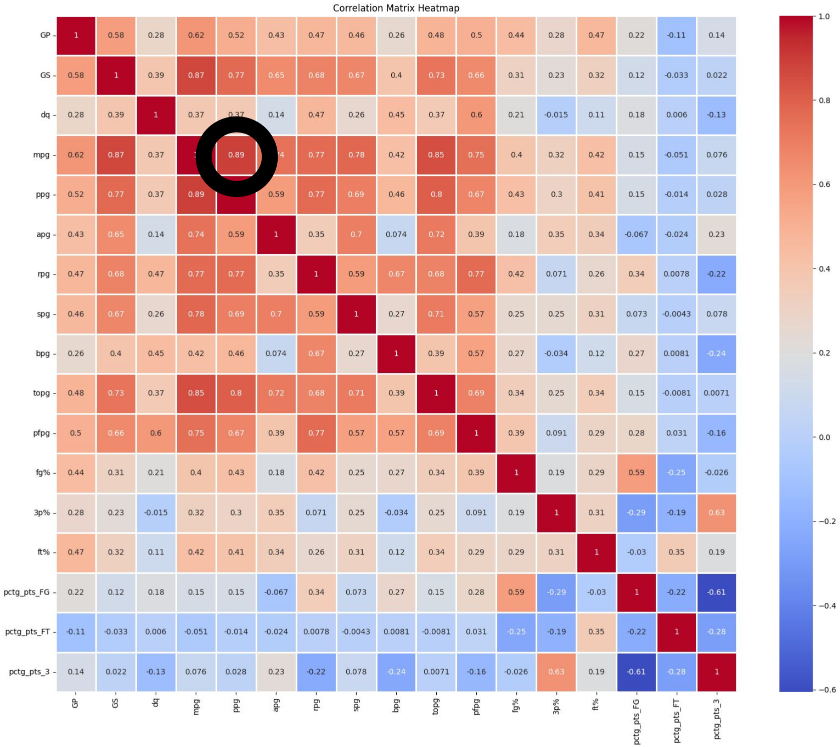
(percentage of entire points that resulted from these types of shot)



Data Preparation - Integration

We realized that the minutes per game column still showed a heavy correlation with points per game, which is expected.
So, we tried to devise a plan to get rid of the ppg stat, by replacing it with a per minute variation of it.
To do this, we needed to test whether or not the number of points a player scores in a game has anything to do with the position they are playing.

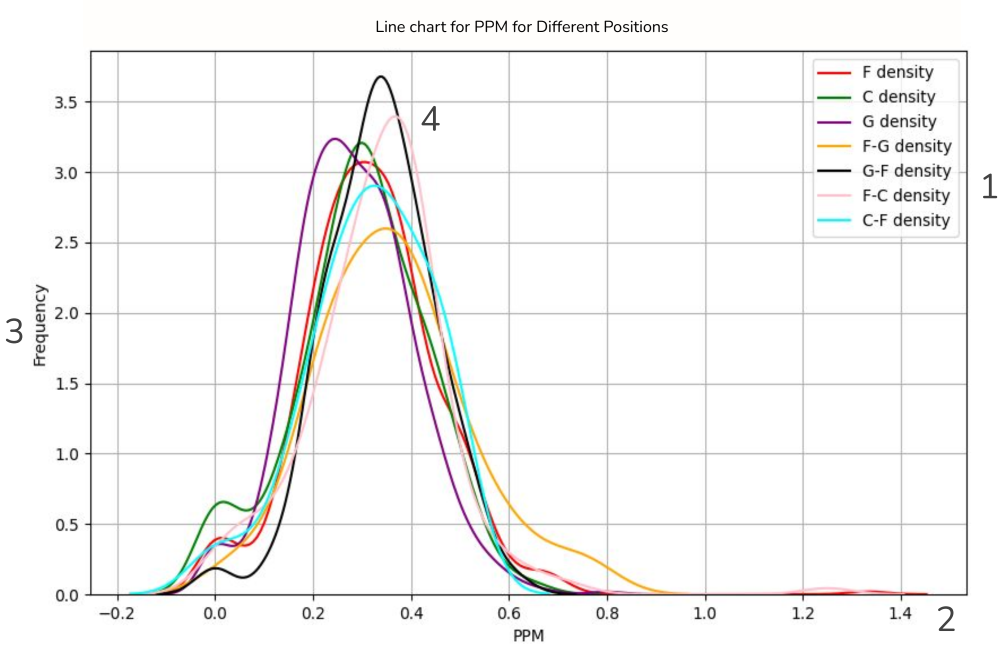




Simplified Correlation Matrix for players_teams.csv

Data Preparation - Integration

To correctly figure out whether or not a player's position has any impact on the number of points they score in a game, we decided to build line charts.



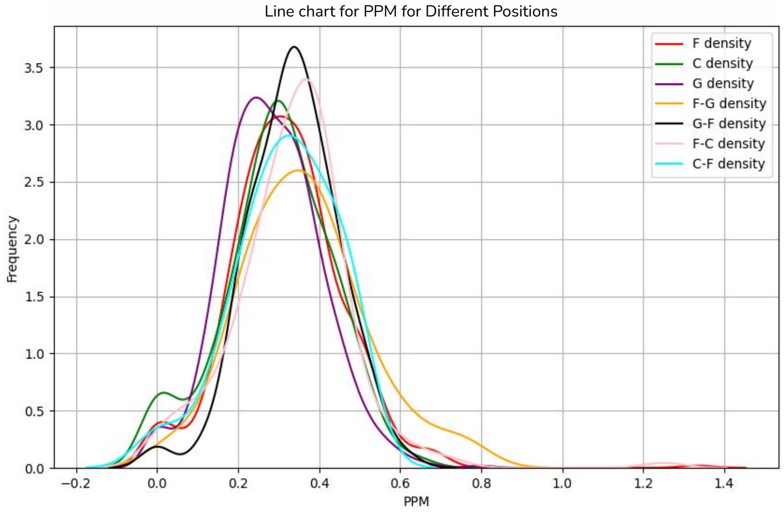
1. Players positions
2. PPM - This axis emcompasses a range of points per minute values
3. Frequency - This axis emcompasses a range of frequency values
4. Each colored line represents the **frequency** with which players from a specific **position** average a certain number of **PPM**.

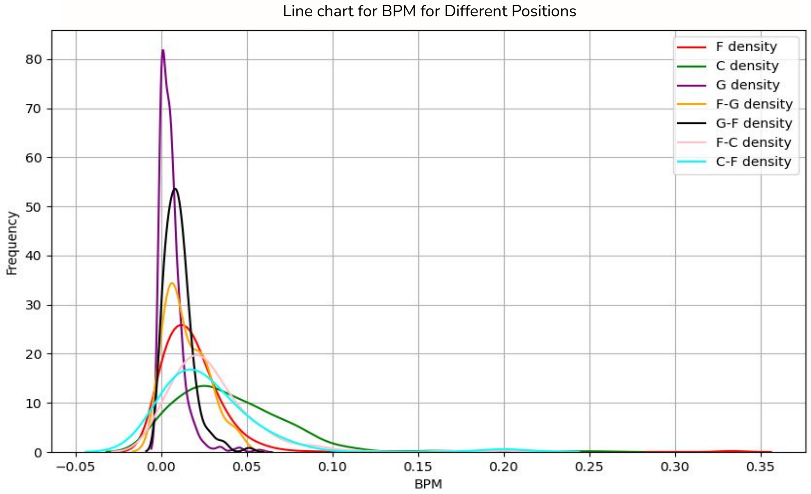
Data Preparation - Integration

How can we interpret the values?

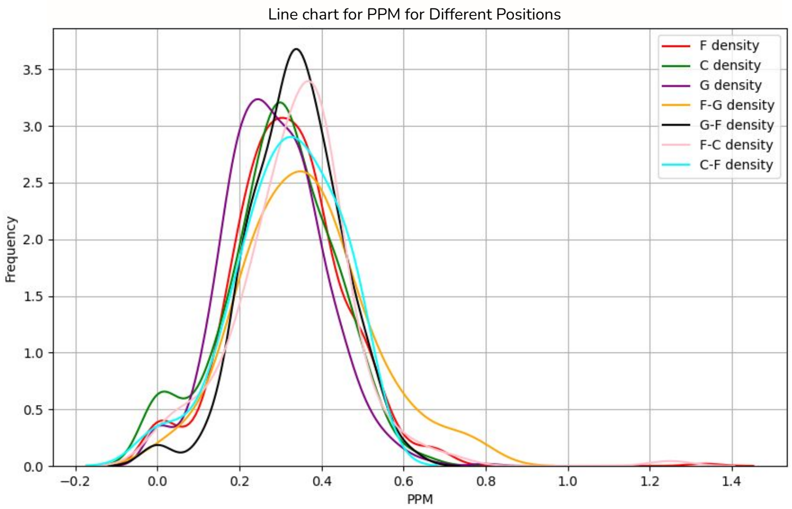
If the lines in the chart float around the same values, then the frequency with which a player achieves a certain per minute stat is not affected by its position

If there is a big discrepancy between some lines, then we can assume the position plays a factor that benefits or detriments the outcome of a player on that stat line.



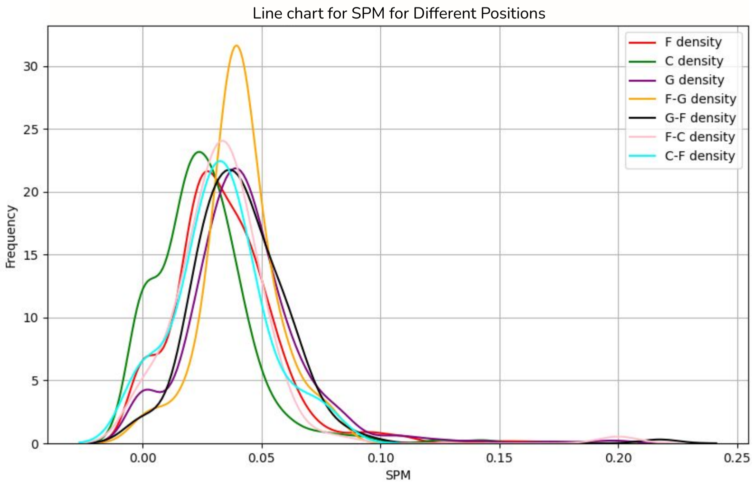


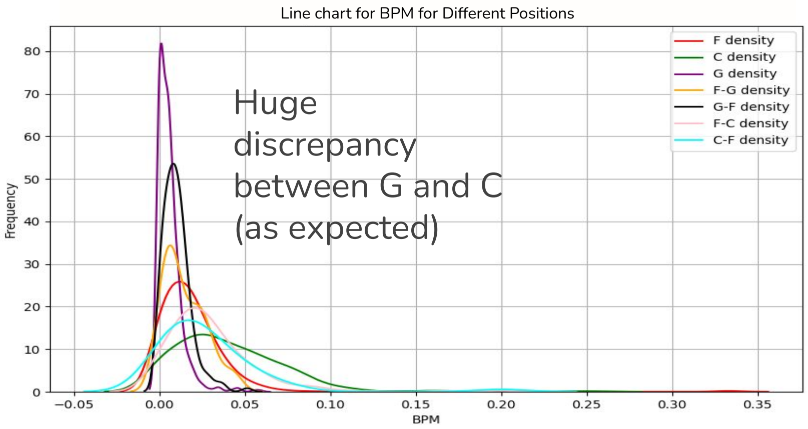
Data Preparation - Integration

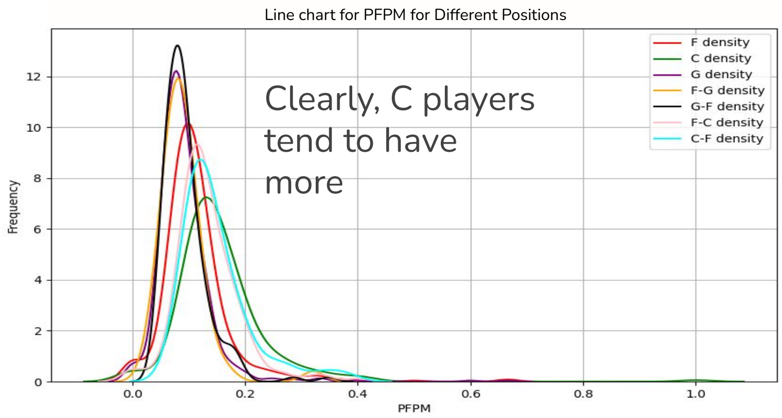


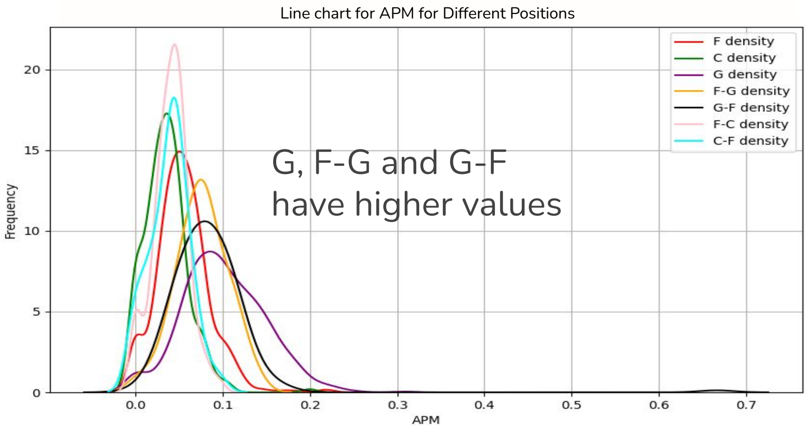
- Values are mostly independent from Position
  - (Values are considered independent if the curves for each position tend to be similar)

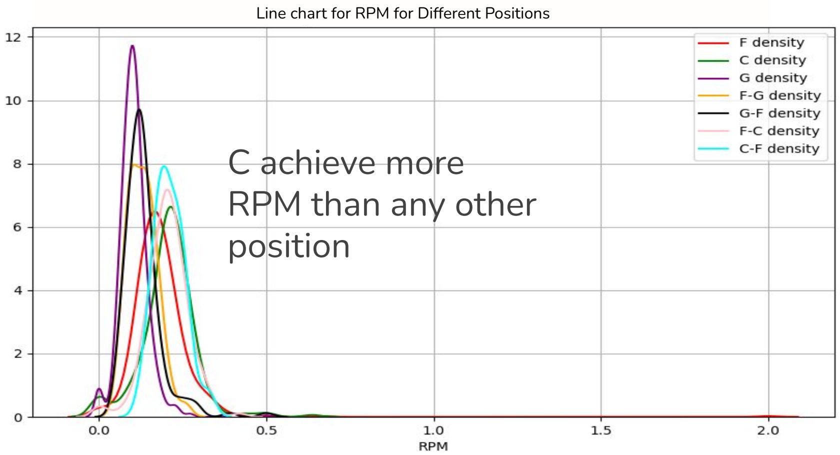
- Why not Ratio everything to Per Minute instead of Per Game?
  - Correlations aren't as high
  - Values depend more on Position
    - Here's proof




Line chart for BPM for Different Positions
Huge discrepancy between G and C (as expected)
F density
C density
G density
F-G density
G-F density
F-C density
C-F density
Frequency
BPM

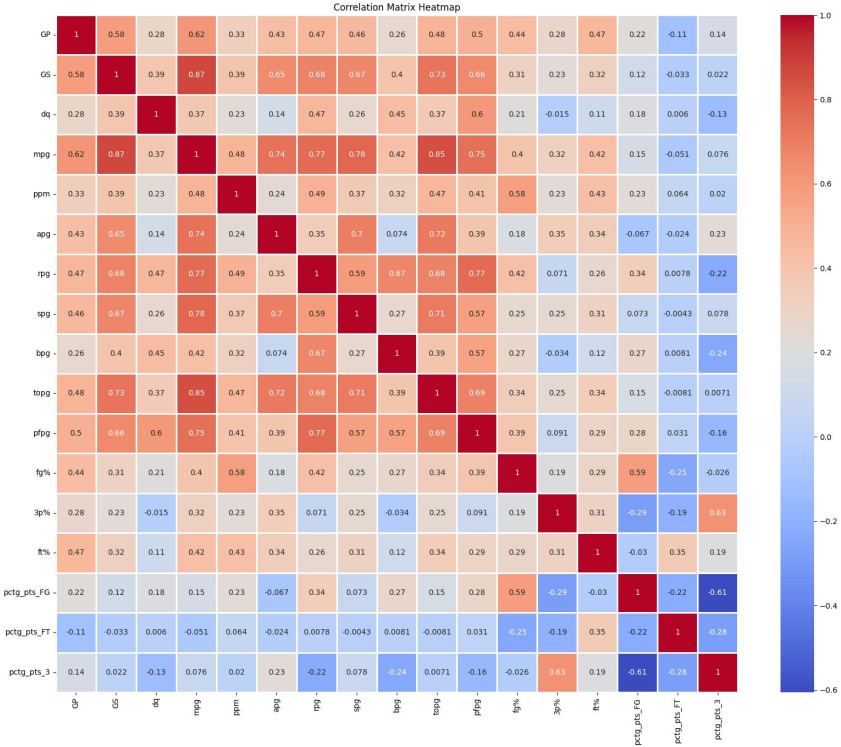

Line chart for PFPM for Different Positions
Clearly, C players tend to have more
F density
C density
G density
F-G density
G-F density
F-C density
C-F density
Frequency
PFPM


Line chart for APM for Different Positions
G, F-G and G-F have higher values
F density
C density
G density
F-G density
G-F density
F-C density
C-F density
Frequency
APM


Line chart for RPM for Different Positions
C achieve more RPM than any other position
F density
C density
G density
F-G density
G-F density
F-C density
C-F density
RPM

Data Preparation -Integration

- Final Simplified Correlation Matrix for players_teams.csv
- No extremely abnormal values that we should get rid of



Data Preparation - Reduction

- Removed irrelevant attributes
  - League ID (awards_players)
  - Death Date (players)
  - First season and last season (players)
  - Division ID (teams)
  - Arena (teams)

- Ignored post datasets, since we considered they were not relevant for the initial predictive model
  - series_post
  - teams_post

Data Preparation - Transformation

- Attribute construction
  - “results” constructed by aggregation of “first round”, “semis” and “finals”
    - Values are labels:
      - 0 - no playoffs
      - 1 - lost on first round
      - 2 - lost on semis
      - 3 - lost on finals
      - 4 - champions
  - “win_ratio”, “homeW_ratio” and “awayW_ratio” (teams)

- Transformations
  - Turn “age” to a float (players)
  - Abbreviation of award names to a standard format, capitalized and first letters only (awards_players)

Predictive Models - String Values

Dealing with string values:

- Encoded categorical features into dummy/indicator features.
- Encoded Ids into numerical values

award_WADTHM	award_WFMVP	confID_EA	confID_WE	results_Unknown	results_label0	results_label1
False	False	True	False	False	True	False
False	False	True	False	False	True	False
False	False	True	False	False	True	False
False	False	True	False	False	True	False
False	False	True	False	False	True	False
...	...	...	...	...	...	...
False	False	True	False	False	False	True
False	False	True	False	False	False	True
False	False	True	False	False	False	True
False	False	True	False	False	False	True
False	False	True	False	False	False	True

Predictive Models - Training Split

Checked for data imbalance: found that 53.82% of rows contained Y in playoffs, while 46.18% didn't.

Used time series splitting with expanding window:

- Train with the first 5 years, test with the 6th.
- Train with the first 6 years, test with the 7th.
- And so on ...

Avoiding Data Leakage

- Replaced features only knowable at the end of the season with last available years data, where possible.

Predictive Models -
Recursive Feature Selection

Process:

- Starting from the entire feature set, we tested our models.
- Iteratively removed features, one by one, and test again.
- Everytime our results got worse, we added the feature back in.

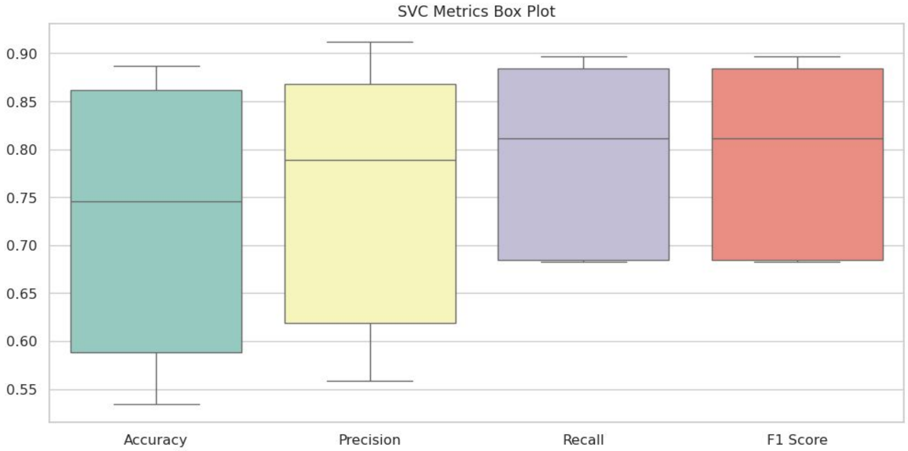
Removed features:

results, seed, college, collegeOther, birthDate, win_ratio, stint, GS, GP, rebounds, fgAttempted, ftAttempted, threeAttempted, minutes, pointsFromFieldGoal, pos, age, award, coachAward, ftMade, topg, percentage_pointsFromThree

Predictive Models
- First Approach Evaluation

- Decision Tree
- Random Forest
- SVM
- Nearest Neighbor
- Naive Bayes

Best Model Metrics - SVM:
- Average Accuracy: 72.3%
- Average Precision: 74.9%
- Average Recall: 79.2%
- Average F1: 79.2%



Problems that arose

During the dataset several teams were relocated, maintaining a large part of its roster.

Solution:

- Match the new teams with the values of the pre-relocated team.

Hyper Parameter Tuning

The following parameters were tested:

- Decision Tree:
  - max_depth: 1 -> 20
  - max_features: auto, sqrt, log2
- Random Forest:
  - max_depth: 1 -> 20
  - max_features: auto, sqrt, log2
- Support Vector Machine:
  - C: 0.1, 1, 10, 100, 1000
  - gamma: scale, auto
  - kernel: linear, rbf, poly, sigmoid
- Nearest Neighbor:
  - n_neighbors: 1 -> 20
  - weights: uniform, distance
  - metric: euclidean, manhattan, minkowski
- Naive Bayes:
  - var_smoothing: 1e-09, 1e-08, 1e-07, 1e-06, 1e-05
- Neural Network:
  - hidden_layer_sizes:
    - 1 layer with 25 -> 200 neurons, in increments of 25
    - 2 layers with 25 -> 200 neurons, in increments of 25, each
  - activation: relu, logistic, tanh
  - alpha: 0.0001, 0.001, 0.01

Predictive Models - Training Improvements

New model used: Neural Network

Through the mentioned teams matching method, as well as through hyperparameter tuning, the performance of our models was improved.

In general, their evaluation metrics were higher and more consistent, as can be seen in the box plots in the following slides.

Decision Tree - Second Approach Evaluation

Average Accuracy: 73.2%

Average Precision: 76.2%

Average Recall: 80.6%

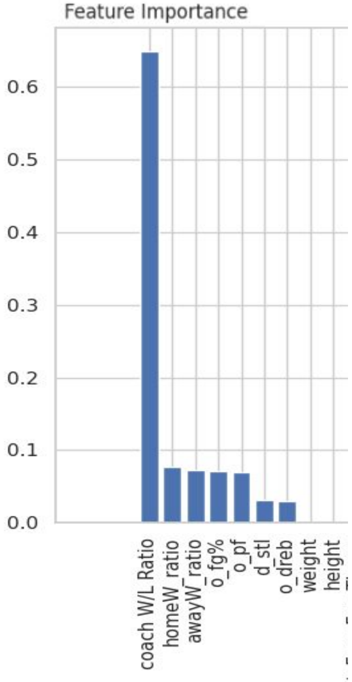
Average F1: 80.6%

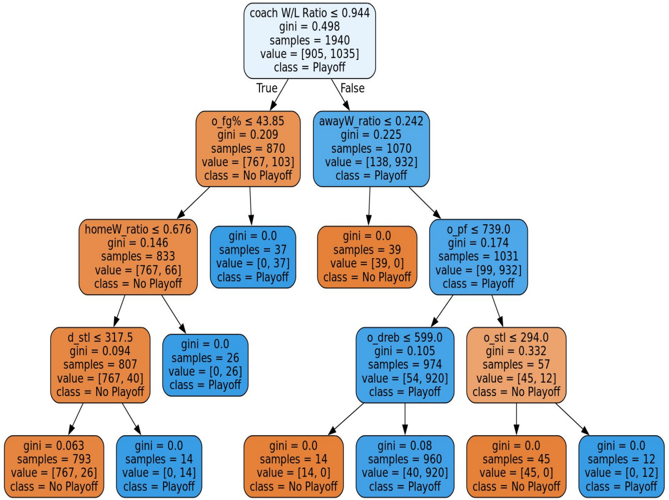
Training time: 187 seconds

Very consistent metrics during training and good average performance, with one of the lowest training times.

However, it's not consistent in predicting the number of qualifying teams per conference, even if it predicts them with high certainty.

Decision Tree -
Second Approach Evaluation





Year 11 Predictions with Conference and Certainty

Tm IDnbsp;	Conf ID_EA	Playoff	Certainty
ATL	1	0	0.41
CHI	1	1	100
IND	1	1	95.83
LAS	0	1	95.83
MIN	0	0	3.28
NYL	1	0	3.28
ORL	1	1	96.88
PHO	0	1	100
SEA	0	1	95.83
TUL	0	1	95.83
UTA	0	0	3.28
WAS	1	0	3.28

Random Forest-
Second Approach Evaluation

Accuracy Average: 71.4%

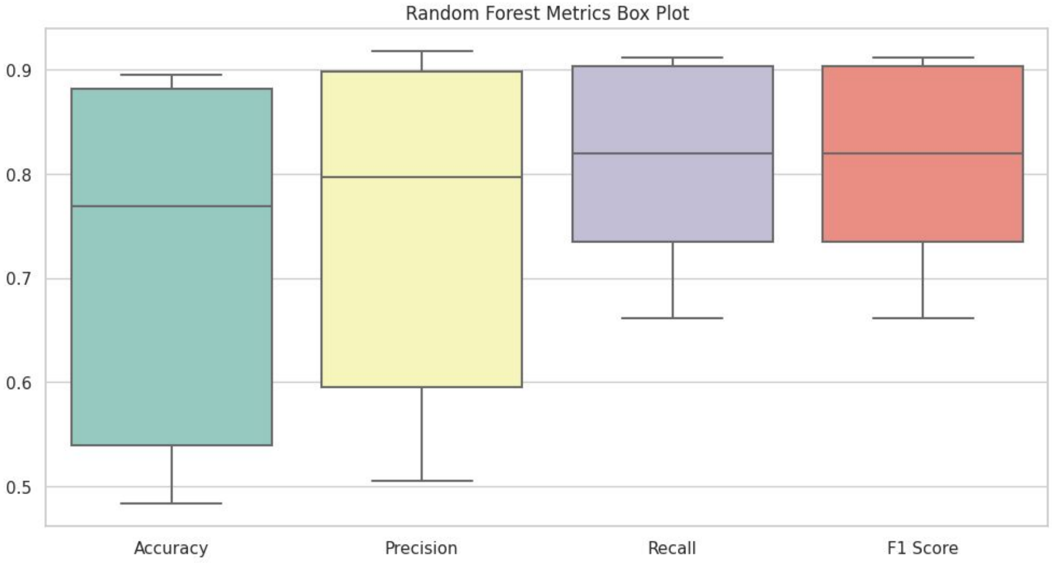
Precision Average: 74.3%

Recall Average: 80.6%

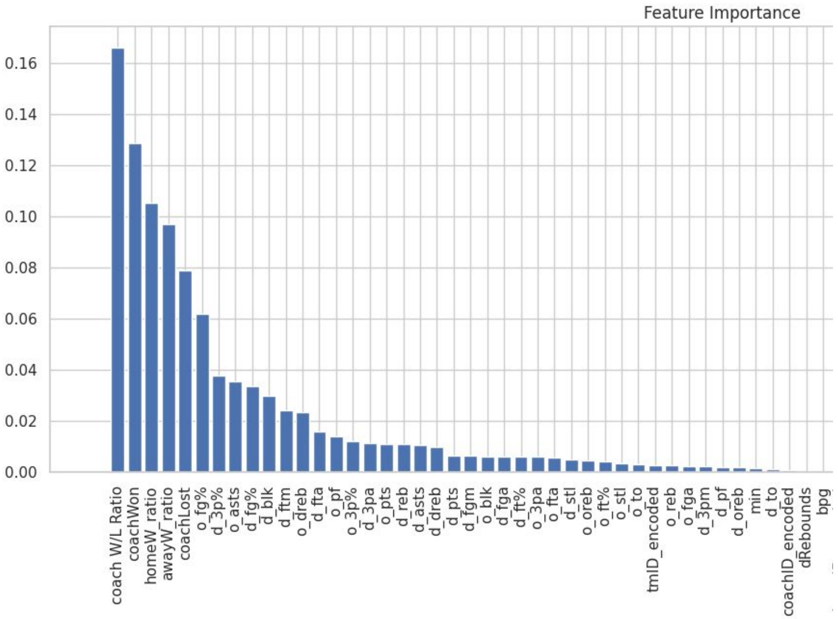
F1 Average: 80.6%

Training Time: 330 seconds

Much like the decision tree, it doesn't always predict 4 qualifying teams per conference, with lower certainties than the decision tree.



Random Forest-
Second Approach Evaluation



Year 11 Predictions with Conference and Certainty

Tm IDnbsp;	Conf ID_EA	Playoff	Certainty
ATL	1	0	47.38
CHI	1	0	33.46
IND	1	1	81.92
LAS	0	1	88.03
MIN	0	0	15.12
NYL	1	0	33.24
ORL	1	1	73.56
PHO	0	1	51.02
SEA	0	1	84.31
TUL	0	1	86.66
UTA	0	0	24.64
WAS	1	0	33.77

Support Vector Machine - Second Approach Evaluation

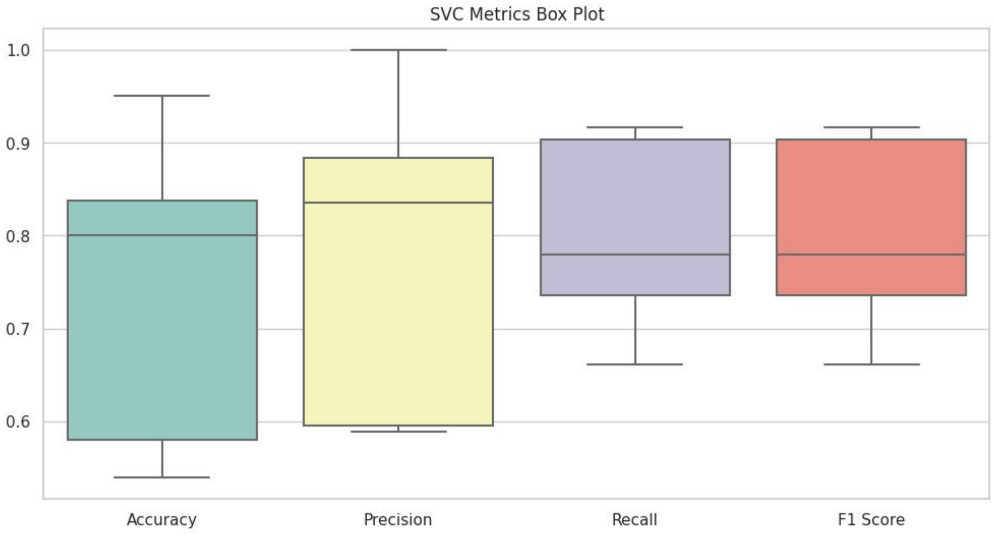
Accuracy Average: 74.2%

Precision Average: 78.1%

Recall Average: 79.9%

F1 Average: 79.9%

Training Time: 753 seconds



Only model to consistently predict 4 qualifying teams from each conference. Considering this and the fact that it scored among the highest metrics of our models, we consider the predictions it produces to be our best.

Support Vector Machine - Second Approach Evaluation

Weight	Feature
0.0532 ± 0.0299	o_dreb
0.0468 ± 0.0313	d_pts
0.0355 ± 0.0403	o_pts
0.0210 ± 0.0219	o_asts
0.0194 ± 0.0241	o_reb
0.0129 ± 0.0164	d_asts
0.0129 ± 0.0079	d_to
0.0097 ± 0.0121	o_fta
0.0081 ± 0.0102	d_blk
0.0048 ± 0.0079	dRebounds
0.0048 ± 0.0079	o_to
0.0048 ± 0.0079	o_stl
0.0048 ± 0.0079	coachID_encoded
0.0032 ± 0.0079	d_oreb
0.0032 ± 0.0079	points
0.0032 ± 0.0219	playerID_encoded
0.0032 ± 0.0079	d_fga
0.0016 ± 0.0065	fgMade
0.0016 ± 0.0158	d_3pa
0.0016 ± 0.0065	coachLost
… 53 more …	

Year 11 Predictions with Conference and Certainty

Tm IDnbsp;	Conf ID_EA	Playoff	Certainty
ATL	1	1	57.92
CHI	1	0	5.84
IND	1	1	97.8
LAS	0	1	100
MIN	0	0	0.55
NYL	1	1	53.91
ORL	1	1	78.62
PHO	0	1	56
SEA	0	1	99.42
TUL	0	1	100
UTA	0	0	37.04
WAS	1	0	17.21

Nearest Neighbor - Second Approach Evaluation

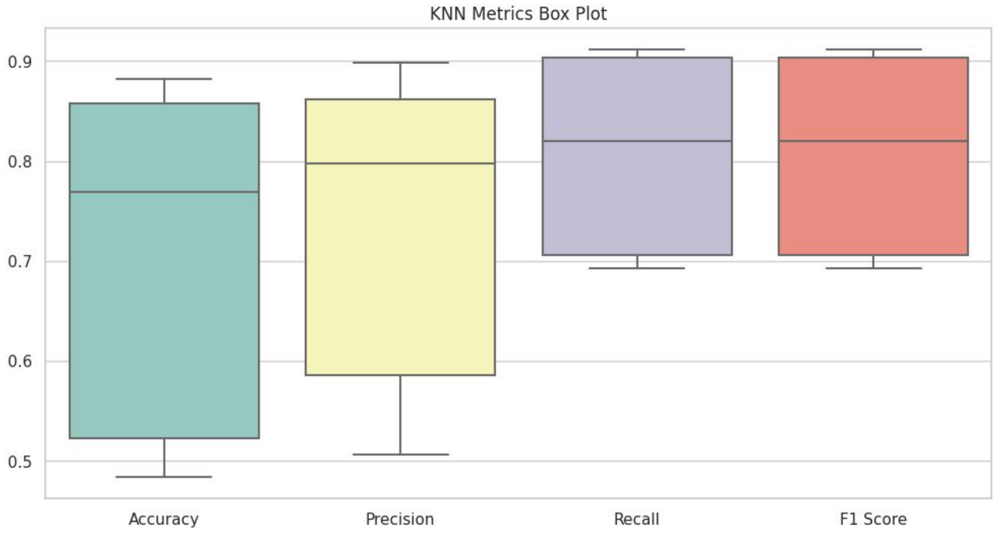
Accuracy Average: 70.3%

Precision Average: 73.0%

Recall Average: 80.7%

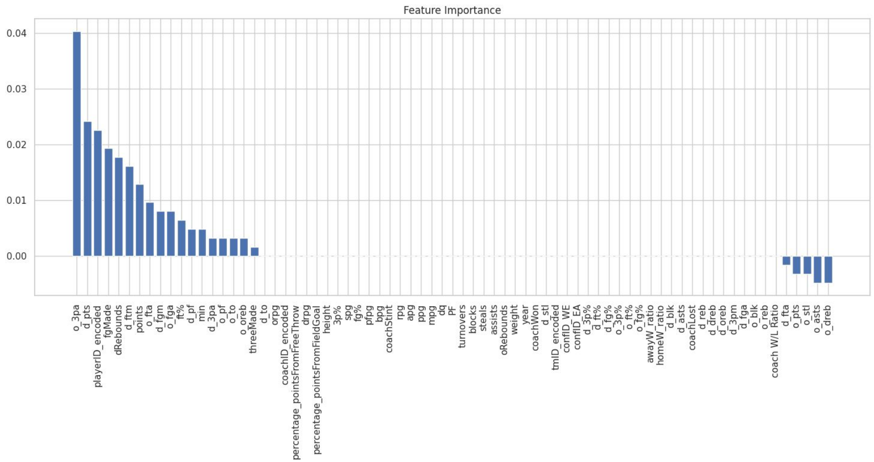
F1 Average: 80.7%

Training Time: 369 seconds



Predicts 7 qualifying teams instead of 8, but has high certainty of those 7. It's among the models with the lowest average and least consistent metrics during training.

Nearest Neighbor - Second Approach Evaluation



Year 11 Predictions with Conference and Certainty

Tm IDnbsp;	Conf ID_EA	Playoff	Certainty
ATL	1	0	2.51
CHI	1	0	15.46
IND	1	1	90.08
LAS	0	1	100
MIN	0	0	20.99
NYL	1	1	100
ORL	1	1	84.78
PHO	0	0	12.65
SEA	0	1	98.36
TUL	0	1	100
UTA	0	1	89.24
WAS	1	0	7.49

Naive Bayes - Second Approach Evaluation

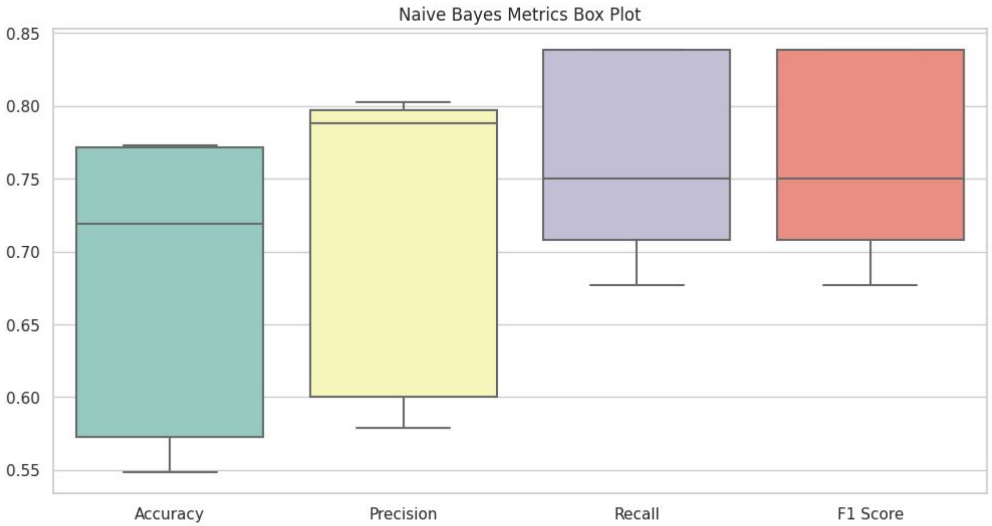
Accuracy Average: 67.7%

Precision Average: 71.3%

Recall Average: 76.2%

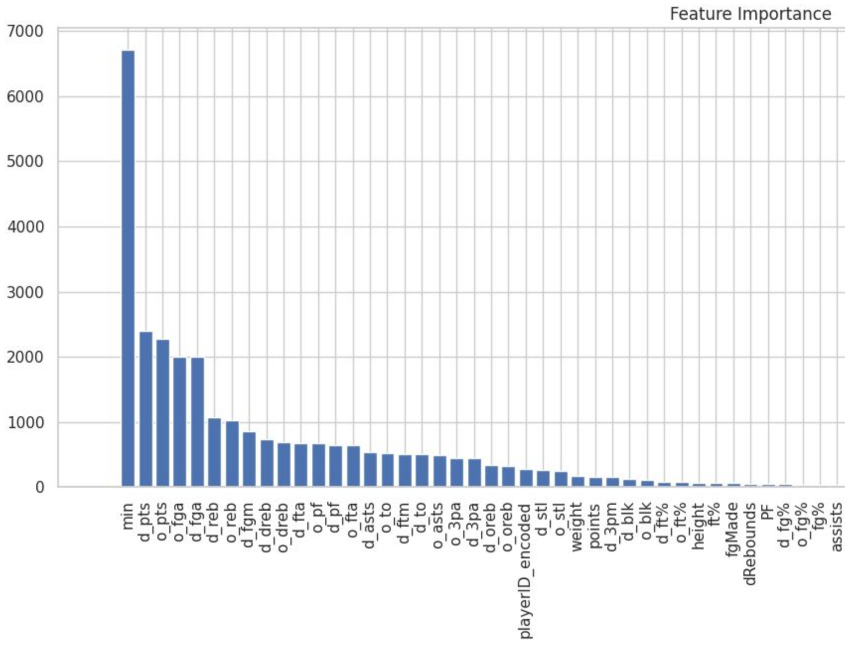
F1 Average: 76.2%

Training Time: 15 seconds



Fastest model to train, with the lowest metrics. Doesn't predict 4 qualifying teams per conference, but predicts them with high certainty.

Naive Bayes - Second Approach Evaluation



Year 11 Predictions with Conference and Certainty

Tm IDnbsp;	Conf ID_EA	Playoff	Certainty
ATL	1	0	36.64
CHI	1	0	26.83
IND	1	1	99.98
LAS	0	1	100
MIN	0	0	0.01
NYL	1	0	0.37
ORL	1	1	75.99
PHO	0	1	98.7
SEA	0	1	99.91
TUL	0	1	100
UTA	0	1	88.71
WAS	1	0	23.07

Neural Network

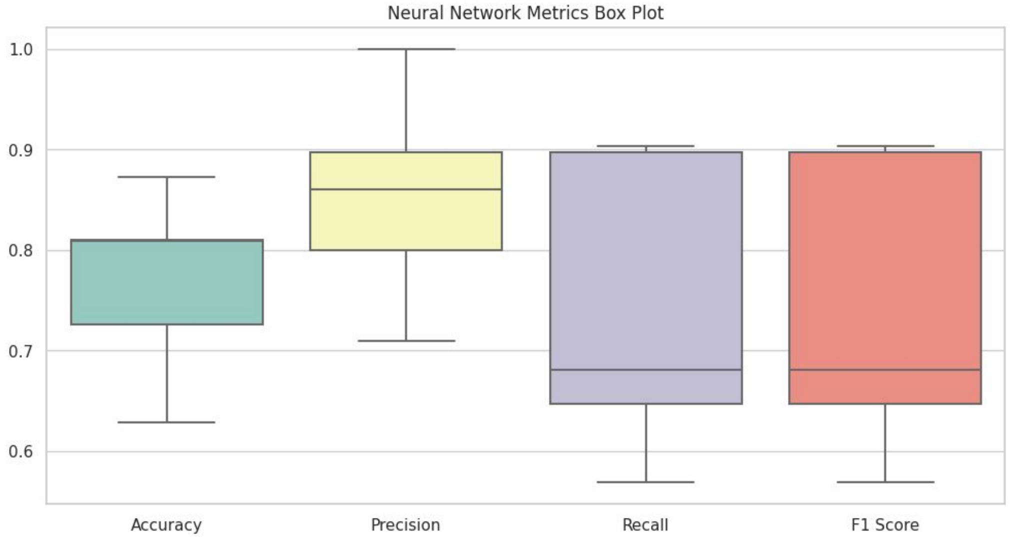
Accuracy Average: 76.9%

Precision Average: 85.3%

Recall Average: 73.9%

F1 Average: 73.9%

Training Time: 4074 seconds



By far the most time consuming to train, with the best training evaluation metrics. However, the year 11 predictions only have 5 qualifying teams, with low certainty.

Neural Network

Weight	Feature
0.0677 ± 0.0219	d_reb
0.0661 ± 0.0534	o_dreb
0.0468 ± 0.0296	d_dreb
0.0403 ± 0.0540	d_3pa
0.0387 ± 0.0724	d_pts
0.0339 ± 0.0277	o_oreb
0.0242 ± 0.0289	d_to
0.0226 ± 0.0065	d_stl
0.0210 ± 0.0262	points
0.0177 ± 0.0258	o_pts
0.0081 ± 0.0177	PF
0.0065 ± 0.0121	d_fga
0.0048 ± 0.0129	d_3pm
0.0016 ± 0.0065	d_fgm
0.0016 ± 0.0065	fgMade
0 ± 0.0000	3p%
0 ± 0.0000	height
0 ± 0.0000	percentage_pointsFromFreeThrow
0 ± 0.0000	fg%
0 ± 0.0000	ft%
	… 53 more …

Year 11 Predictions with Conference and Certainty

Tm IDnbsp;	Conf ID_EA	Playoff	Certainty
ATL	1	0	0.85
CHI	1	0	0
IND	1	1	88.91
LAS	0	1	99.41
MIN	0	0	0
NYL	1	0	0.01
ORL	1	0	28.58
PHO	0	0	0.22
SEA	0	1	97.83
TUL	0	1	78.08
UTA	0	0	0
WAS	1	0	0

References

https://www.basketball-reference.com/

https://www.wnba.com/