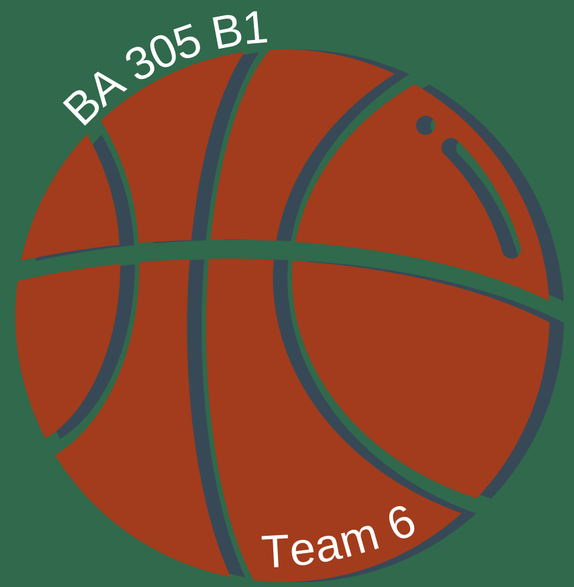
THE CELTICS:

Predicting 2023 Season Wins



MEET THE TEAM

Tessa Wu

Edwin Suh

Phillip Kim

Mary Nguyen

1 Introduction

Why Sports Analytics?

Our team wanted to combine our own interest in with NBA with our academic knowledge.

Sports analytics is very crucial in today's sports industry as it helps gain competitive advantage for teams, improves fan engagement, and ultimately plays a huge role in sports betting.

Why the Celtics? Edwin's favorite team.

2 Data Pre-Processing

COMPOSITION — CLEANING — PCA

- 2018 - 2022 game by game stats of all teams (past 5 seasons)
- 1 row = 1 game
- Selected ONLY Celtics games

TOTAL GAMES: 4599 ROWS → **CELTICS GAMES: 308 ROWS**

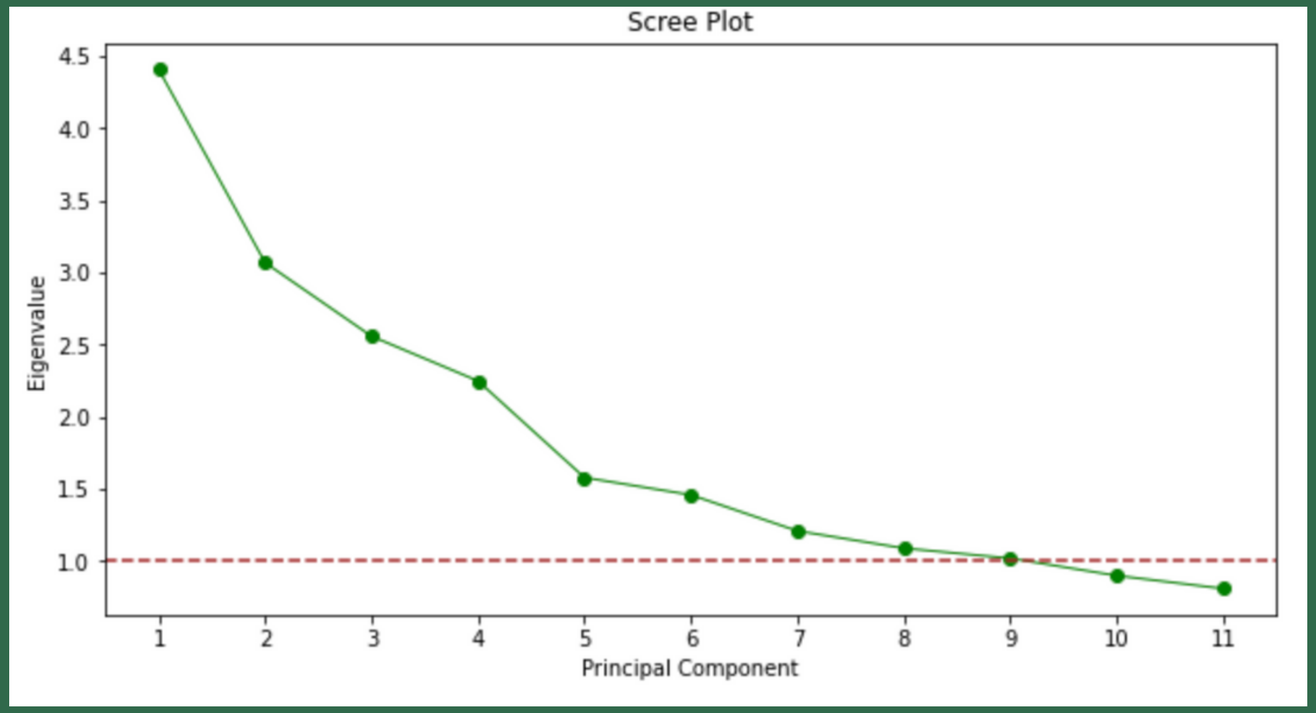
Variables: 40 columns

COMPOSITION — CLEANING — PCA

Original Columns	Dropped Columns	Reason
Unamed: 0, season_id, game_id, game_date, team_id, team_abbreviation, matchup, wl_x, fgm, fga, fg_pct, fg3m, fg3a, fg3_pct, ftm, fta, ft_pct, quality, oreb, dreb, reb, ast, stl, blk, tov, pf, pts, plus_minus, date, season, team, elo1_pre, elo1_pre, elo1_post elo2_post, raptor1_pre, raptor2_pre, score1, score2, wl_y	Unnamed: 0, season_id, game_id, game_date, team_id, team_abbreviation, matchup, wl, plus_minus, score1, score2, wl_x, elo1_post, elo2_post	• no predictive power (season_id, team_id) • affect prediction (wl, plus_minus, elo1_post, elo2_post)

COMPOSITION — CLEANING — PCA

LATENT ROOT CRITERION



Principal Component	Proportion of Variance	Cumulative Proportion
0	0.191	0.191
1	0.133	0.324
2	0.111	0.435
3	0.098	0.532
4	0.069	0.601
5	0.063	0.664
6	0.052	0.717
7	0.047	0.764
8	0.044	0.808

COMPOSITION — CLEANING — PCA

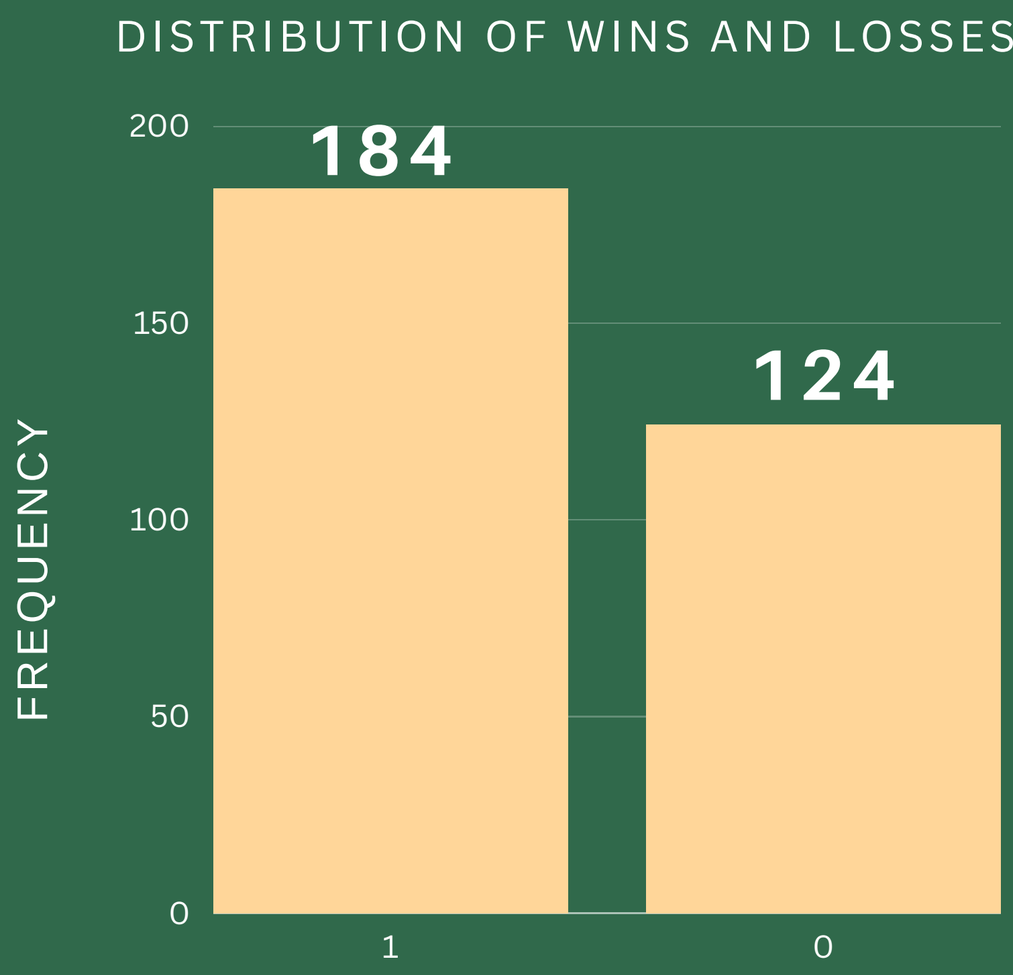
FULLY CLEANED DATASET

	Scoring	OpponentStrength	SecondChance	FreeThrows	DefensiveRebounding	CelticStrength	PerimeterShooting	Turnovers	FreeThrowEfficiency
0	-1.099	-0.603	2.975	0.452	-1.707	-0.633	0.896	-0.176	0.277
1	-0.499	-1.527	3.144	-0.083	-0.767	-0.396	1.266	1.071	1.422
2	-2.509	2.002	-2.585	1.592	-0.161	0.043	-0.409	0.128	-0.024
3	-2.145	1.182	1.946	-2.146	-1.530	-0.186	-1.075	-1.210	0.352
4	-3.124	-1.030	-0.826	2.175	0.344	0.866	0.415	-0.054	1.122
...	...	...	...	...	...	...	...	...	...
303	1.502	0.307	-2.909	0.806	0.641	-1.908	-1.427	-0.150	-0.193
304	6.686	-1.561	-0.214	-1.236	0.930	-1.063	-0.435	-1.209	2.599
305	1.702	-0.917	-0.476	-0.621	0.434	-1.579	-0.435	0.336	2.347
306	2.115	-2.108	0.898	0.113	-0.239	0.392	-1.181	0.650	1.550
307	4.401	-0.489	2.668	1.919	-0.885	-1.308	-1.039	-0.691	-0.884

Principal Component	Measures
Scoring (0)	pts, ast, fgm, fg_pct, fg3m, fg3_pct
Opponent Strength (1)	elo2_pre, raptor2_pre, quality
Second Chance (2)	fga, fg3a, oreb, reb
Free Throws (3)	ftm, fta
Rebounding (4)	dreb, reb, tov
Celtic Strength (5)	elo1_pre, raptor1_pre
Perimeter Shooting (6)	fg3a, ft_pct
Turnovers (7)	tov
Free Throw Efficiency (8)	ft_pct

3 Data Visualization

Finding 1: Naive Model

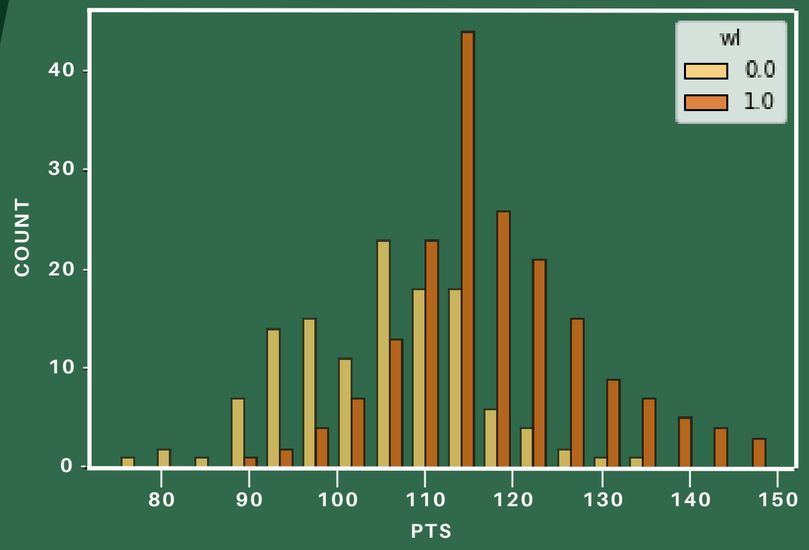


BASELINE ACCURACY:

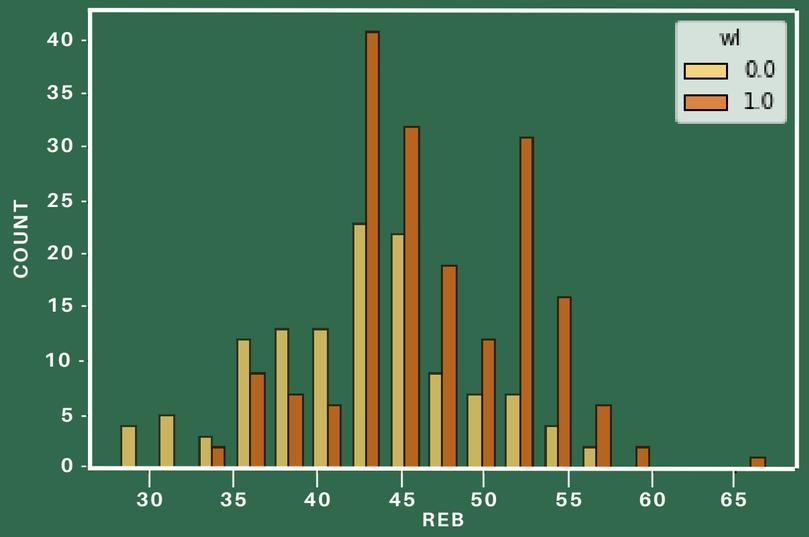
$$\frac{184}{308} = 59.74\%$$

Finding 2: Team Statistics & WL

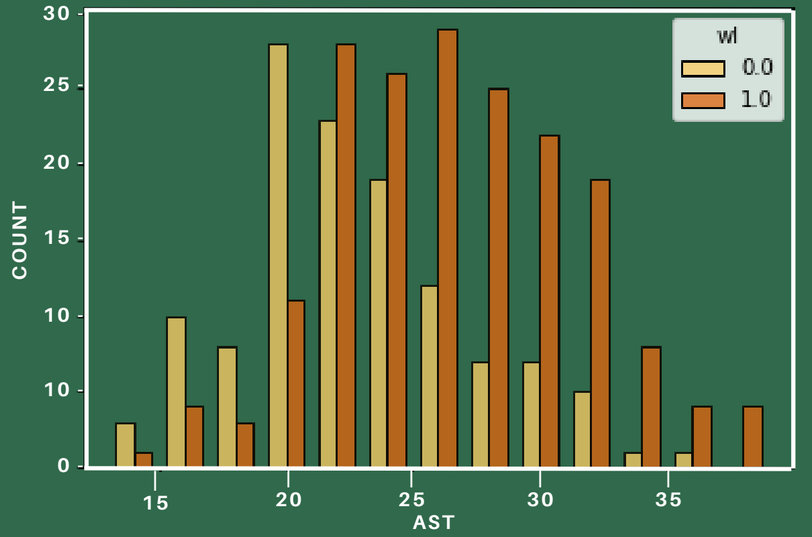
POINTS & WL



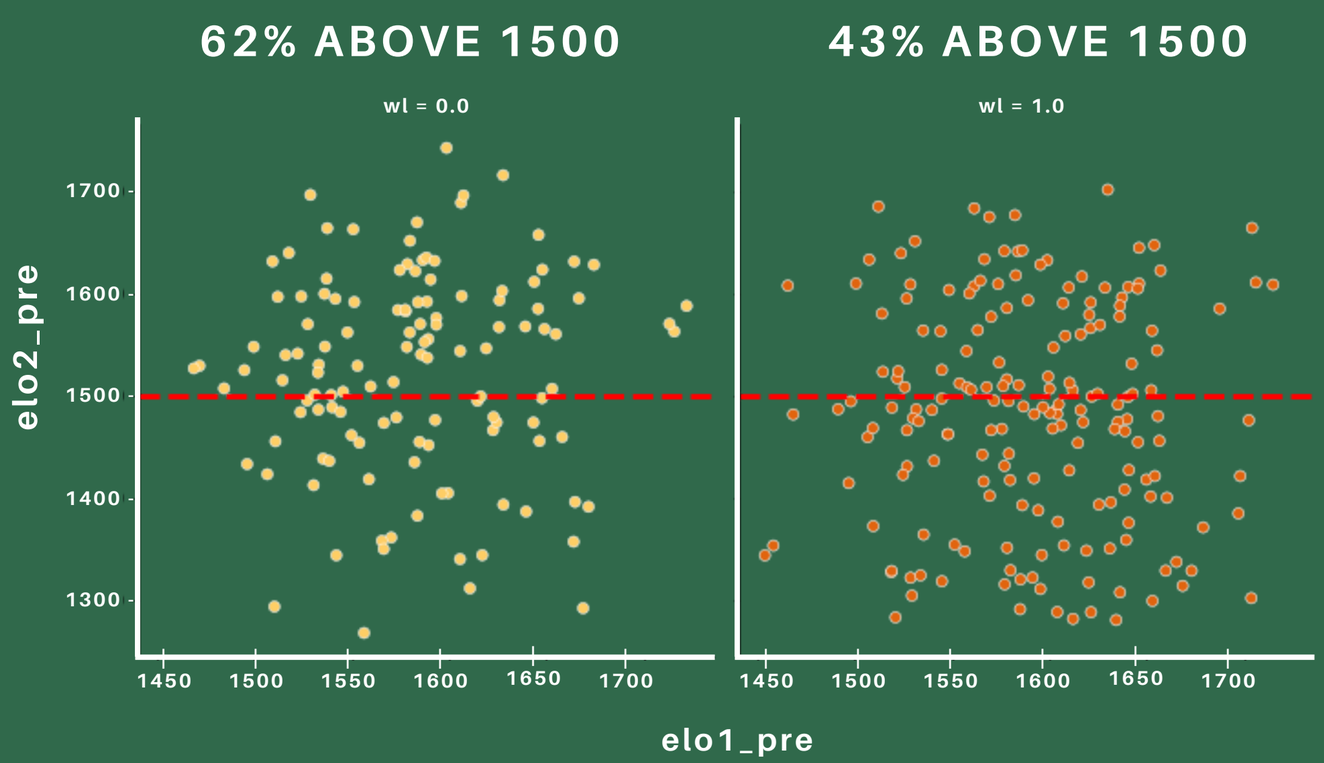
REBOUNDS & WL

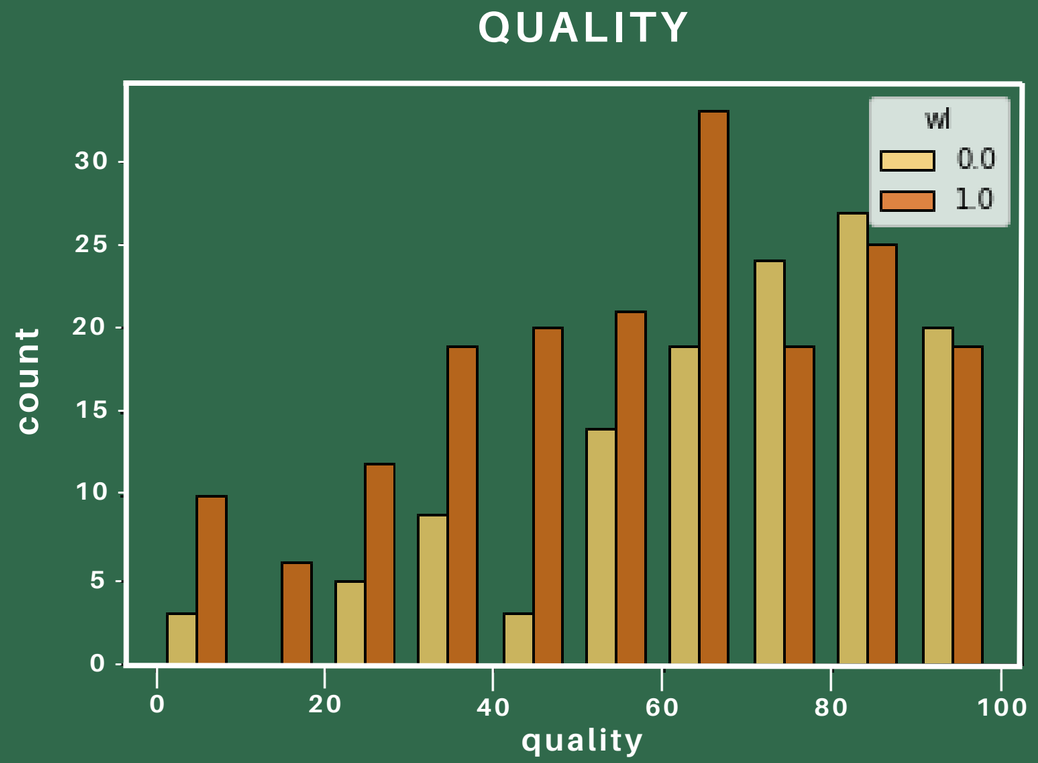


ASSISTS & WL



Finding 3: Opponent Strength





4 Modeling

Logistic Regression

ACCURACY SCORE: 84.22%

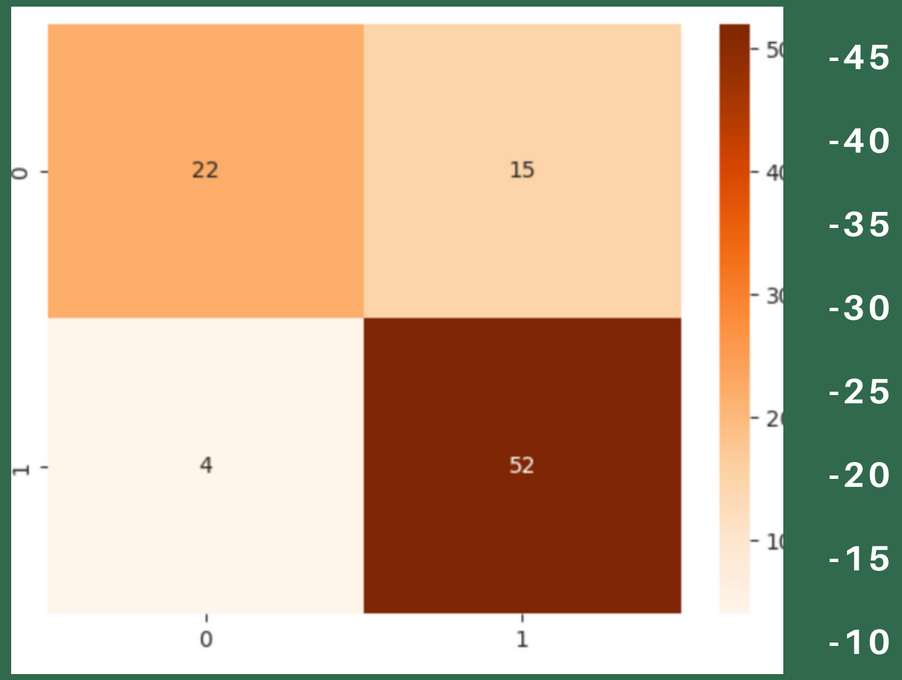
Top 3 Most Powerful Features

- **Scoring**
- **Defensive Rebounding**
- **Free Throws**

intercept	0.8829760921694616
	coeff
Scoring	0.980094
OpponentStrength	0.436194
SecondChance	-0.044902
FreeThrows	0.537107
DefensiveRebounding	-0.712681
CelticStrength	-0.093717
CelticFT	-0.032516
Turnovers	-0.480565
OppFreeThrows	-0.283187

K- Nearest Neighbors

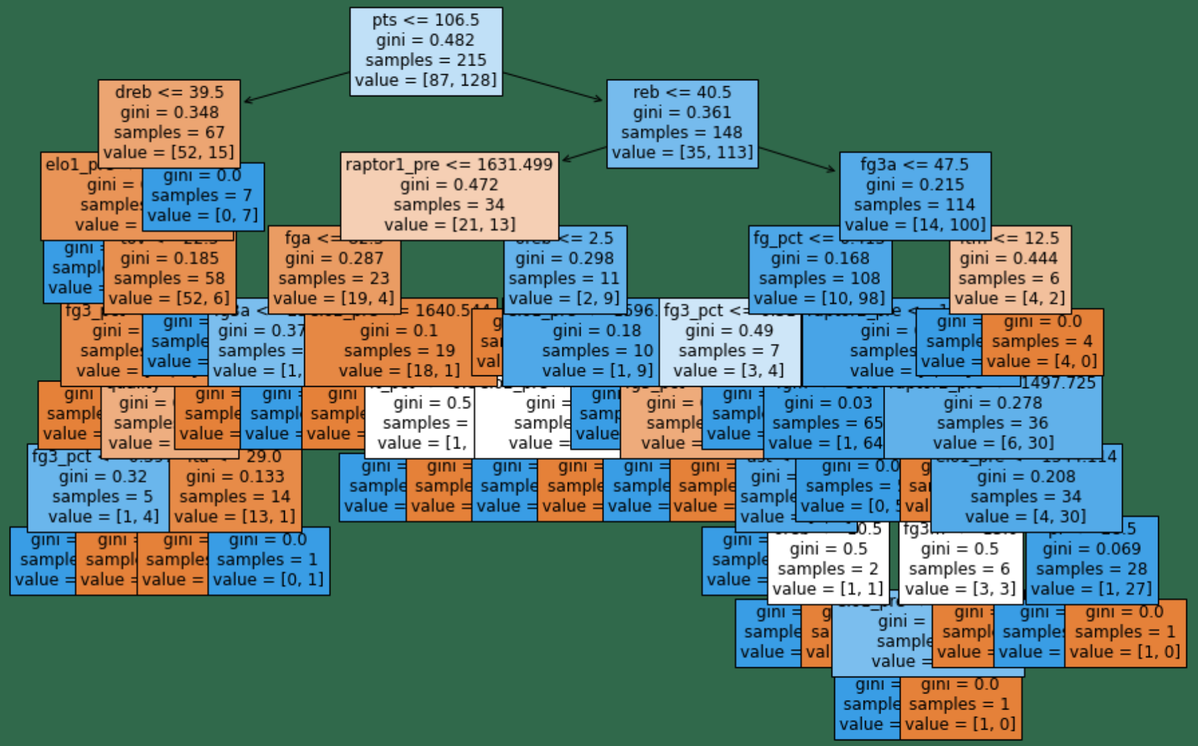
- BEST K = 9
- ACCURACY SCORE: 80.45%



Decision Trees

FULL TREE

# of Nodes	# of Leaves	# of Max Depth	Accuracy Score
63	32	9	66.95%

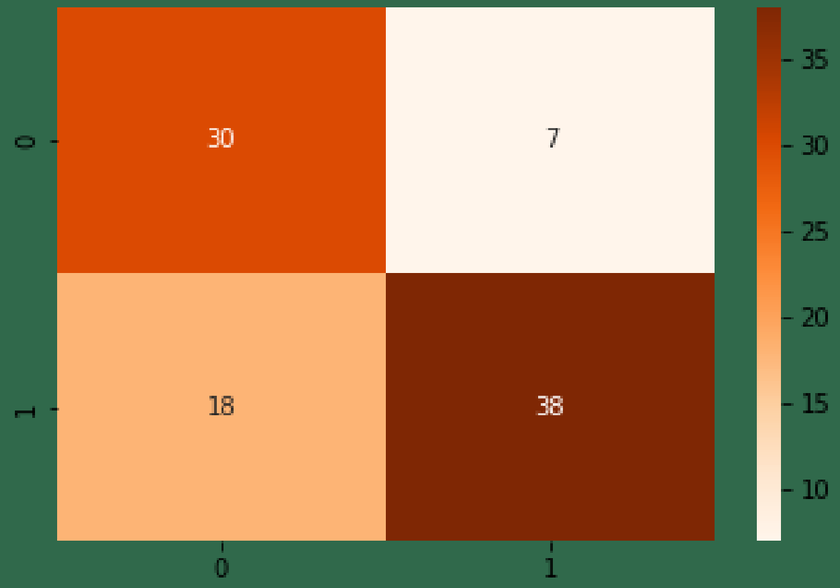


Parameter	Selection
Splitter	'Random'
Max_Depth	6
Min_Sample_Split	25
Min_Impurity_Decrease	0.001
Criterion	'Entropy'
Random_State	1

Decision Trees

PRUNED TREE

# of Nodes	# of Leaves	# of Max Depth	Accuracy Score
25	13	6	77.77%

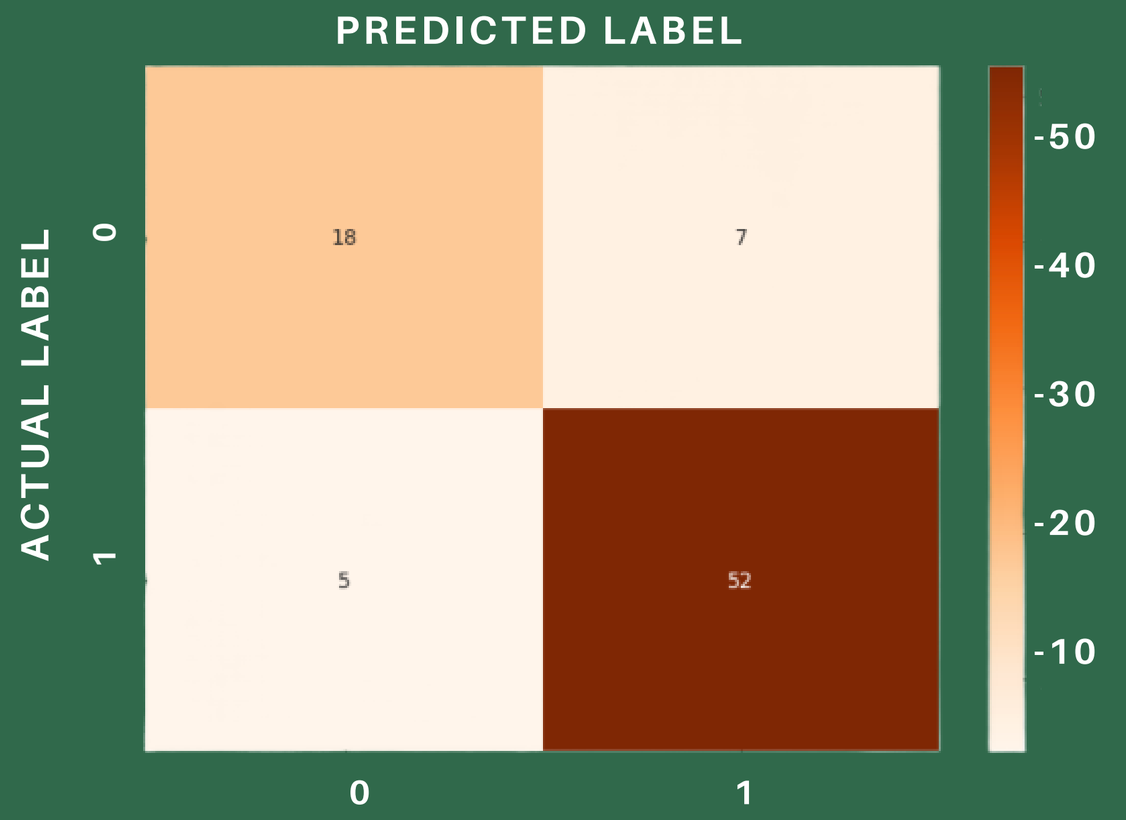


Random Forest

BEST MODEL: CRITERION "LOG_LOSS" & N_ESTIMATORS 125

ACCURACY SCORE = 83.87%

	criterion	n estimators	n jobs	random state	accuracy of model
0	gini	75	-1	1	79.03%
1	entropy	75	-1	1	79.0%
2	log_loss	75	-1	1	79.0%
3	gini	100	-1	1	80.45%
4	entropy	100	-1	1	78.55%
5	log_loss	100	-1	1	78.55%
6	gini	125	-1	1	80.91%
7	entropy	125	-1	1	79.48%
8	log_loss	125	-1	1	79.48%



Random Forest

Top 5 Features
Field Goal Percentage
3 Point Field Goal Percentage
Defensive Rebounds
Total Rebounds
Points Scored

5 Model Comparisons

MODEL COMPARISON

Model	Naive	Logistic Regression	k-NN	Decision Tree (Full)	Decision Tree (Pruned)	Random Forest
Accuracy	59.74%	**84.22%** 	80.45%	66.95%	77.77%	80.91%
Parameters		solver = liblinear penalty = 'l1' cv = 10	k = 9 cv = 10 weight = uniform metric = manhattan	random_state =1	criterion = entropy max_depth= 6 min_impurity_decrease = 0.001 min_sample_split = 25 splitter = random cv = 10	n_estimators = 125 random_state = 1 criterion='gini' n_jobs = -1 cv = 10
Important Features		Scoring, **Defensive Rebounds**, Free Throws			**Defensive Rebounds**, Field Goals Made, **Field Goal Percentage**, Free Throw Made	**Field Goal Percentage**, 3 Point Field Goal Percentage, **Defensive Rebounds**

2023 Regular Season

Model	Logistic Regression	k-NN	Decision Tree (Pruned)	Random Forest
Accuracy	48.78%	47.56%	79.27%	**82.93%**
Parameters	solver = liblinear penalty = 'l1' cv = 10	k = 9 cv = 10	criterion = entropy max_depth= 6 min_impurity_decrease = 0.001 min_sample_split = 25 splitter = random cv = 10	n_estimators = 125 random_state = 1 criterion='gini' n_jobs = -1 cv = 10

6 Conclusion

Challenges & Improvements

Challenges	Improvements
Limited Dataset • RAPTOR scores were recently introduced so NBA seasons prior to the 2018-2019 season did not have RAPTOR scores	• Waiting for more seasons in order to increase size of data
Factoring in Opponent Strength • Elo ratings, RAPTOR scores, and quality are the only variables that take into account opponent	• Add to our dataset the opponent's game by game statistics, adding more opponent features
Limited Models • There could better models to represent the data	• More research into other predictive models for better predictive analysis
The NBA is Unpredictable • The randomness of sports is an aspect that cannot be controlled	• Try to approach the problem from a different angle, so that the randomness can be accounted for in some way

THANK YOU, ANY QUESTIONS?