BA305 Section B1
Spring 2023

Boston University
Professor Tsoukalas

CELTICS

2023 Season Predictions

Data Analysis Report

Team 6

Phillip Kim

Mary Ann Nguyen

Edwin Suh

Tessa Wu

TABLE OF CONTENTS

INTRODUCTION

The Sports Analytics Market is a growing industry, projecting a 27% growth from the previous few years. Every individual from sports bettors to fans, seek their best effort to anticipate the outcome of each game, and the NBA is without exception. In the midst of the 2022-2023 playoff season, finding factors that best capture the teams' outcomes is at the height of the moment as every team undeniably wants to gain competitive advantage.

Given the Boston Celtics' recent success over the past couple of years, this report aims to explore historical game data of the Boston Celtics to help identify key determinants that could potentially explain their recent success. By providing a comprehensive and objective analysis of how each of those factors play into their outcomes for each game, the team can leverage that information to further optimize their success. Such information can also extend to sports betting where individuals are able to make more informed decisions.

Through this report, we hope to provide a better understanding of how predictive analysis is transforming the sports industry, and how teams like the Celtics leverage data to gain a competitive edge and improve their overall performance.

DATA PREPROCESSING

Data Composition

Our dataset is a combination of two separate datasets one consisting of every NBA game's box score (i.e. points, steals, fouls, etc.) with 62,367 records and the other containing 73,335 records of all historical NBA team's Elo ratings and RAPTOR ratings. Since the RAPTOR ratings have only been introduced starting from the 2018-2019 NBA season, our data acquisition is bounded by games starting from the 2018-2019 season and after that. Thus, from both datasets, we only obtained records of games played by the Boston Celtics from the 2018-2019 season to the 2021-2022 season.

With these constraints, we were able to create a singular merged dataset containing 308 records each representing a single game played in the regular season by the Boston Celtics starting from the 2018-2019 NBA season to the 2021-2022 NBA season. The dataset also comprised 40 columns relating to game performance of the Boston Celtics (points, rebounds, assists, etc) and statistics regarding the opponent's strength (elo ratings, RAPTOR ratings, quality).

Data Cleaning & Feature Selection

For detailed action on pre-processing, refer to table below:

Column_Name	**Sample**	**Decision**	**Reason**
Unnamed 0:	236	Dropped	Categorical data with no predictive power
game_id	21800001	Dropped	Categorical data with no predictive power
game_date	2018-10-16	Dropped	Categorical data with no predictive power
team_id	1610612738	Dropped	Categorical data with no predictive power
team_abbreviation	BOS	Dropped	Categorical data with no predictive power (all BOS)
matchup	BOS @ PHI	Dropped	Categorical data with no predictive power
wl_x	W	Dropped	Duplicate
fgm	42.0	NA	NA
fga	97.0	NA	NA
fg_pct	0.433	NA	NA
fg3m	11.0	NA	NA
fg3a	37.0	NA	NA
fg3_pct	0.297	NA	NA
ftm	10.0	NA	NA
fta	14.0	NA	NA
ft_pct	0.714	NA	NA
oreb	12.0	NA	NA

dreb	43.0	NA	NA
reb	55.0	NA	NA
ast	21.0	NA	NA
stl	7.0	NA	NA
blk	5.0	NA	NA
tov	15.0	NA	NA
pf	20.0	NA	NA
pts	105	NA	NA
plus_minus	18	NA	NA
date	2018-10-16	Dropped	Categorical data with no predictive power
season	2019	Dropped	Categorical data with no predictive power
team	BOS	Dropped	Categorical data with no predictive power
elo1_pre	1561.52419307645	NA	NA
elo2_pre	1607.05768800629	NA	NA
elo1_post	1573.84937213462	Dropped	Indicative of prediction
elo2_post	1594.73250894812	Dropped	Indicative of prediction
raptor1_pre	1633.0	NA	NA
raptor2_pre	1617.0	NA	NA
score1	105.0	Dropped	Indicative of prediction
score2	87.0	Dropped	Indicative of prediction
quality	90	NA	NA

wl_y	W	Converted to dummy variable	Make usable for analysis

PCA

From our correlation matrix (*see Figure 1.)*, there were many variables related to scoring that are highly correlated to each other. For example, the correlation between points and field goals made was 0.85. This is reasonable because the amount of points a team scores is based on how many field goals they can make. We saw this trend with other aspects of basketball, such as rebounding and opponent's strength. Ultimately, this led us to perform Principal Component Analysis (PCA) on our dataset to reduce multicollinearity, which can also help with over-fitting. After executing PCA, we combined like variables, reducing 40 total features to 9 principal components.

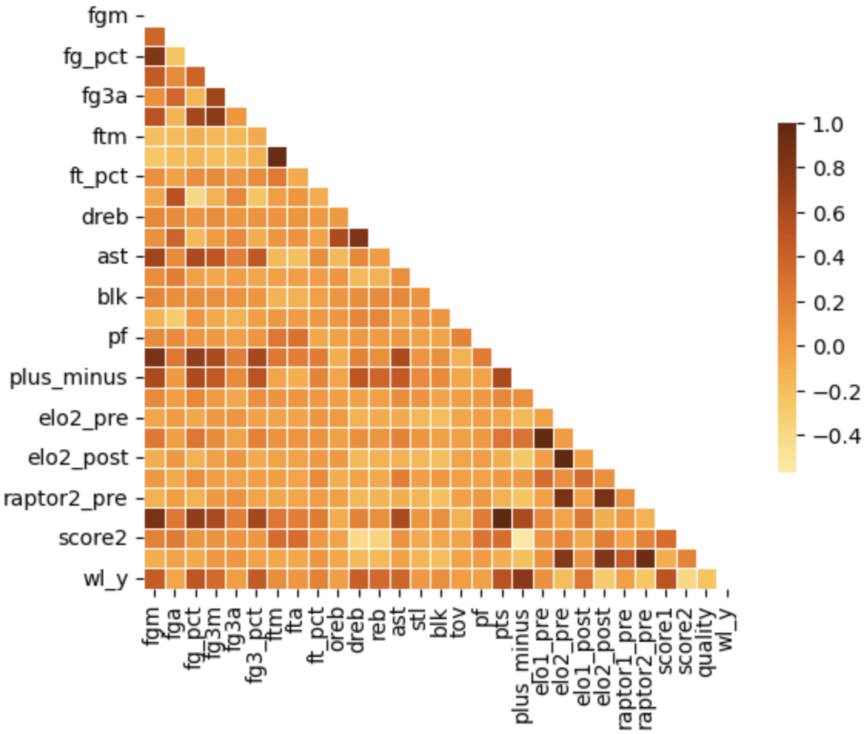


Figure 1. Correlation Matrix

DATA VISUALIZATION

We used several graphs to visualize the different variables, noting any findings that could help better understand our predictive analysis.

Finding 1: Naïve Model

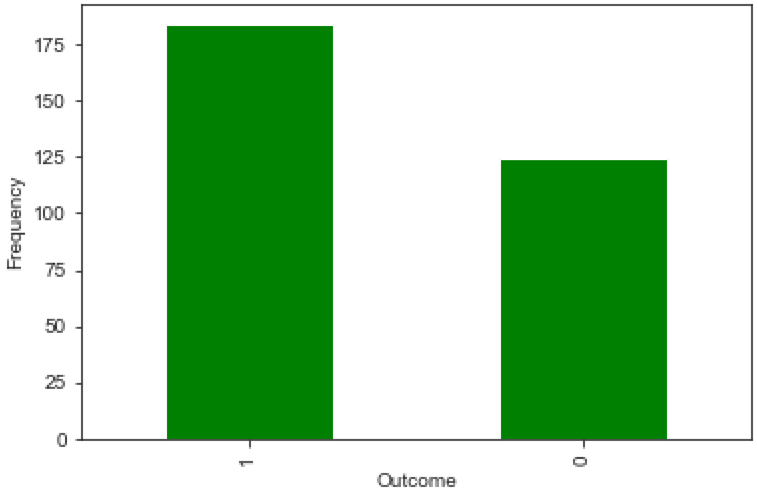


Figure 2: Distribution of Wins and Losses

When looking at the distribution of our data based on our outcome variable (Wins/Losses), we found the distribution of Boston Celtics' wins versus losses to not be significantly imbalanced (59.74%:40.26%). From this distribution, concluded that this minor imbalance should not affect our models too much and thus continued to obtain our Naïve model. Our baseline accuracy is calculated from using the majority rule by assuming the Celtics won every single game from the past 5 seasons, yielding an accuracy score of 59.74%.

Finding 2: Team Statistics and Wins/Losses

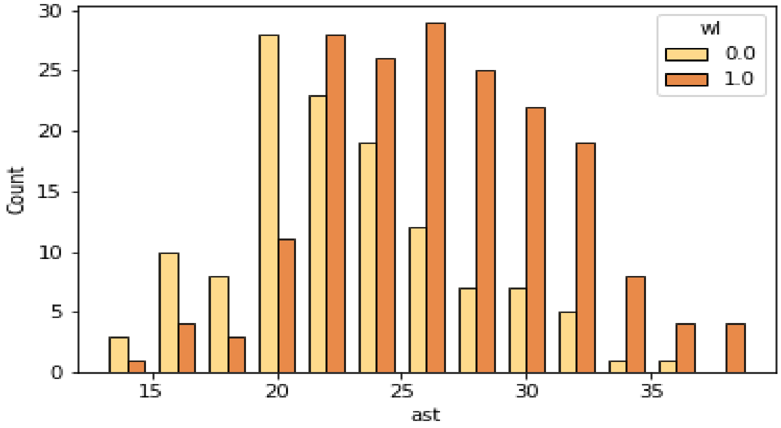


Figure 3. Distribution of Assists for Wins and Losses

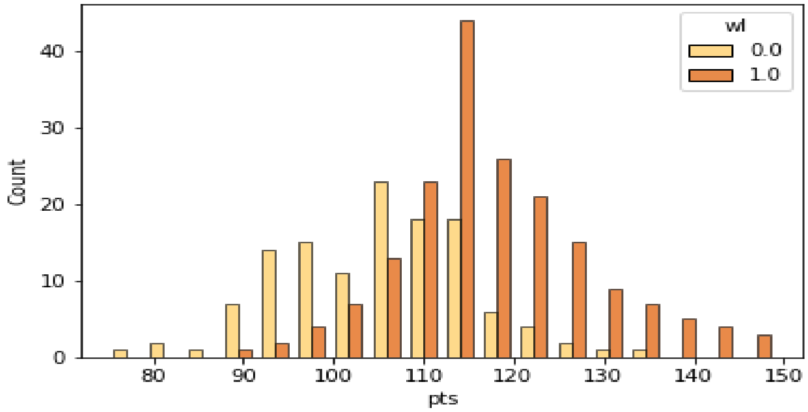


Figure 4. Distribution of Points for Wins and Losses

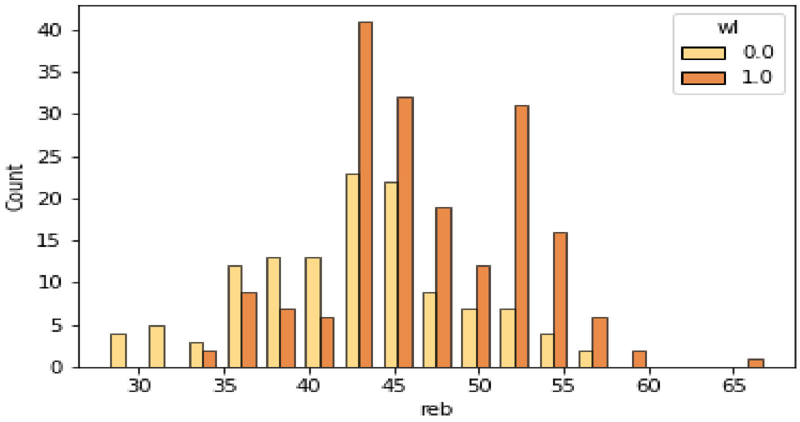


Figure 5. Distribution of Rebounds for Wins and Losses

We plotted bar charts to explore the relationship between the 3 most commonly used metrics of basketball (assists, points, and rebounds) and the number of wins/losses. From all three charts, we see a consistent trend where an increase in these metrics resulted in higher frequency of wins. That is, for instance, there is a correlation between higher number of assists and higher number of wins. Similarly, more points and rebounds showed a higher number of wins. This makes sense as all 3 of these metrics are indicators of a teams' performance; the better the performance, the higher the chance of winning.

Finding 3: Opponent Strength and Wins/Losses

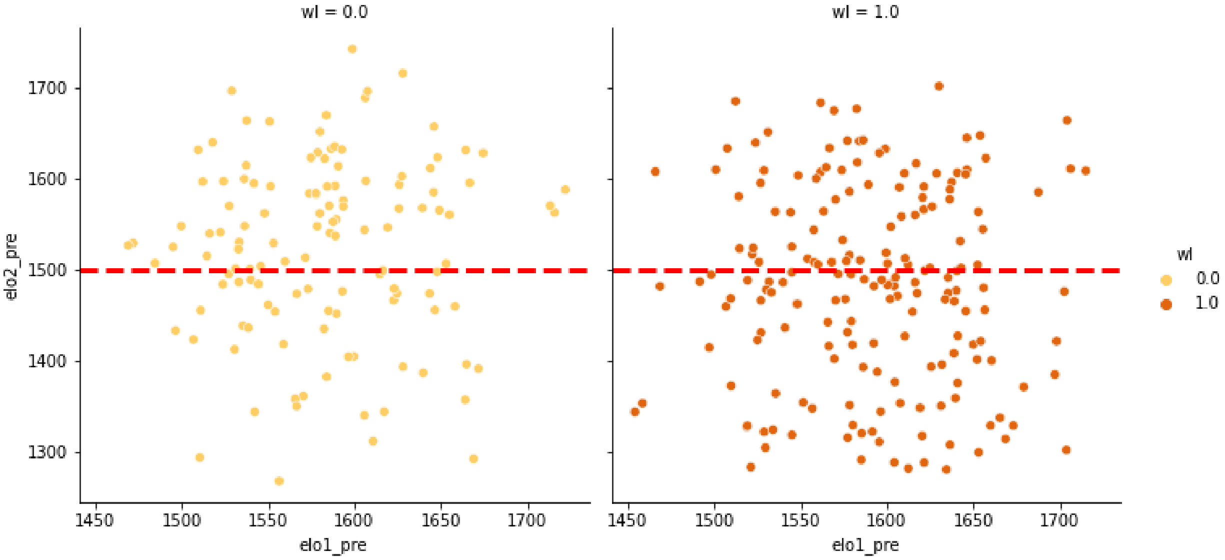


Figure 6: Distribution of Opponent Elo Score for Wins and Losses

In figure 6, we looked at the spread of opponent elo scores for games the Celtics won and lost, with elo2_pre being the opponent and elo1_pre being the Celtics. We created a threshold at 1500 elo and considered this average as every team starts with 1500 elo in the beginning of the season.

Of the games the Celtics lost, 62% of the teams had an elo score of above 1500, while of the games the Celtics won, only 43% of the teams had an elo score of above 1500. From this, we determined that elo scores were a good indication of opponent strength as the Celtics struggled to win games against better opponents.

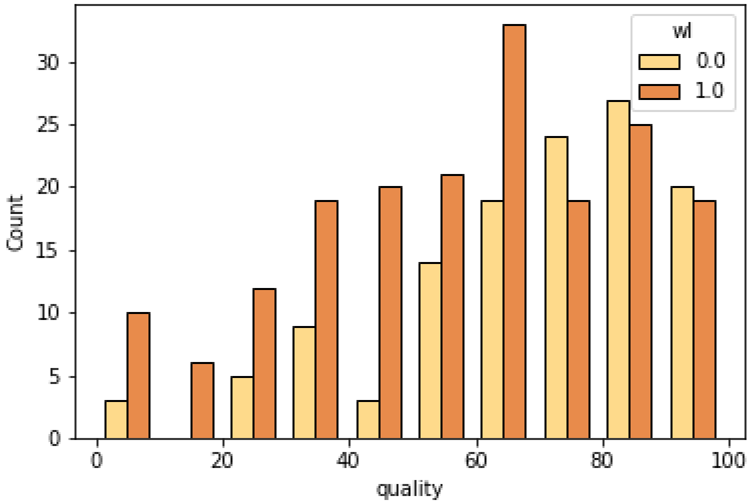


Figure 7: Distribution of Quality for Wins and Losses

We also took quality into account for assessing opponent strength. Quality measures the strength of the Celtics's matchup, ranging from 0 to 100, with a higher score meaning a harder matchup. This distribution further showed how the Celtics struggled with tougher opponents, as after a certain threshold, Celtics lost more games with a higher matchup quality score.

DATA MODELING

Logistic Regression

Logistic Regression is commonly used for binary classification that predicts the odds of a binary outcome such as predicting wins and losses. After hypertuning the parameters and performing 10-fold cross validation, the logistic regression yielded an accuracy score of 84.22% with l1 regularization and liblinear solver being the best parameters.

When observing the summary statistics for the model, we could see that the three most important features were scoring, defensive rebounding, and free throws as these features had the highest coefficient values with respective values of 0.98, 0.54, and -0.71 *(Refer to Appendix B)*.

K-Nearest Neighbors (k-NN)

K-Nearest Neighbors (k-NN), is a supervised model with one parameter which is the number of k's. It is most commonly used to classify data based on its distance from neighboring data points. Because K-Nearest Neighbors is sensitive to anomalies and distance magnitudes, we scaled the data and applied PCA before running the model.

We also performed 10-fold cross validation and applied grid search to hypertune the parameter k. After hypertuning, we obtained the optimal k of 9 with an accuracy score of 80.45%. With our confusion matrix (*Refer to Appendix C)* , we had more false positives of 22 than false negative which we think is worse because in sports, particularly in sports betting, we wouldn't want to encourage betting on a team that is inherently going to lose. This goes to show that in sports as a whole, being the favorite to win a game but actually losing is a bigger let down than expecting to lose and losing.

Decision Trees

- **Full Tree:**

Tree Summary:

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

The full decision tree was still able to achieve a higher accuracy score than the baseline model, however, the biggest concern for the model was its susceptibility to overfitting along poor visualization of the tree (*Refer To Appendix D*). Therefore, we prepruned the tree using Grid Search to find the best parameters.

- **Pruned Tree:**

Parameter Selection (Grid Search):

Max Depth	**Minimum Sample Split**	**Minimum Impurity Decrease**
6	25	0.001

Tree Summary:

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

By specifying the specific parameters, along with a cross validation of 10 folds, we were about to produce a prepruned tree that achieved a higher accuracy model and better visualization (*Refer To Appendix E*)

Random Forest

Random Forests are a commonly-used supervised model based on grouping multiple decision trees to create the most accurate prediction. Unlike decision trees, a random forest model will use feature bagging to build a forest of decision trees which is optimal for datasets with many variables that may be correlated with each other. Because of this, the nature of Random Forest models reduces overfitting and produces accurate predictions.

We hypertuned the model by performing grid search with a 10-fold cross validation which resulted in the optimal number of n_estimators of 125 and criterion of ‘log_loss’. Our final model yielded an accuracy of 81.45% which was higher than our K-Nearest Neighbors model and Decision Tree model.

Table 1. Grid Search on of Random Forest Classification

criterion	n_estimators	n_jobs	random _state	Accuracy of Model
gini	75	-1	1	79.03%
entropy	75	-1	1	79.0%
log_loss	75	-1	1	79.0%
gini	100	-1	1	80.45%
entropy	100	-1	1	78.55%
log_loss	100	-1	1	78.55%
gini	**125**	**-1**	**1**	**80.91%**
entropy	125	-1	1	79.48%
log_loss	125	-1	1	79.48%

The Random Forest Classifier predicted more false positives than false negatives which shows that the wins are more wrongly classified tha losses (*see Figure 9).*

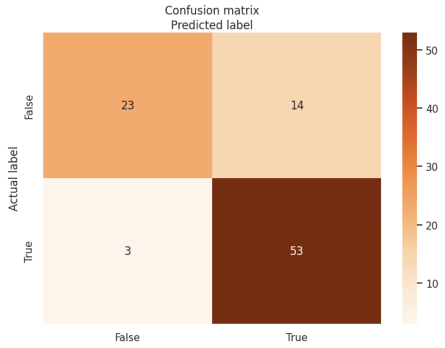


Figure 9. Random Forest Confusion Matrix

MODEL EVALUATION

We ran a total of 6 models on our dataset, including baseline, logistic regression, KNN, full decision tree, pruned decision tree, and random forest.

Table 2. Comparison of Model by Accuracy Score

Model	Baseline	**Logistic Regression**	KNN	Decision Tree (Full)	Decision Tree (Pruned)	Random Forest
Accuracy	59.74%	**84.22%**	80.45%	66.95%	77.77%	80.91%
Parameters	Majority Rule	**Penalty** = **'l1'** **Solver** = **'liblinear'**	k = 9 cv = 10 weight = uniform metric = manhattan	random_state = 1	criterion = 'entropy' max_depth = 6 min_impurity_decrease = 0.001 min_sample_spliy = 25 splitter = 'random' cv = 10	n_estimators = 125 random_state = 1 criterion = 'gini' n_jobs = -1 cv = 10

We focused on the accuracy of our models because we assigned an equal cost to false positives and false negatives. Every model we produced achieved a higher accuracy score than our naive model, proving that our models outperformed the baseline model. Logistic Regression yielded the highest accuracy score while Full Decision Trees yielded the lowest accuracy score *(see Figure 10)*.

However, when validating our models using 2022-2023 Celtics game by game statistics, we found that Logistic Regression and K-Nearest Neighbors resulted in a lower accuracy score than our baseline model (*refer to Figure 11)*. Pruned DecisionTrees and Random Forests outperformed the naive model like before with Random Forests being the most accurate.

Table 3. Comparison of Model Deployment on 2022-2023 NBA Season by Accuracy Score

Model	Logistic Regression	KNN	Decision Tree (Full)	Decision Tree (Pruned)	**Random Forest**
Accuracy	48.78%	48.78%	69.51%	79.27%	**82.93%**
Parameters	Penalty = ‘l1’ Solver = ‘liblinear’	k = 9 cv = 10 weight = uniform metric = manhattan	random_state = 1	criterion = ‘entropy’ max_depth = 6 min_impurity_decrease = 0.001 min_sample_spliy = 25 splitter = ‘random’ cv = 10	**n_estimators = 125** **random_state = 1** **criterion = ‘gini’** **n_jobs = -1** **cv = 10**

CONCLUSIONS

Challenges

We ran into four main challenges when creating our dataset and developing our models. The first challenge we ran into was how much data cleansing and data processing was required for separating game statistics and having limited data. Because RAPTOR scores were a recently introduced metric in the NBA, we only had data from the 2018-2019 season. This took extensive data processing and made it difficult to increase the size of our dataset which would have improved our issues of over-fitting and creating accurate predictions. The third challenge we faced was factoring in opponent strength since we had only three statistics such as Elo Ratings, RAPTOR Scores, and quality that accounted for opposing teams. The fourth challenge we ran into was that Sports data is extremely nuanced which made it difficult to build the best models to represent the data and accounting for the randomness of sports which cannot be controlled.

Improvements

In order to improve our issues, we decided that waiting for more seasons in order would increase the size of our data. We could also add the opponent's game by game statistics to account for the opposing team's strength. In addition to altering our dataset, increasing our knowledge of data sets and NBA statistics would improve our predictive models and sports analysis.

APPENDIX

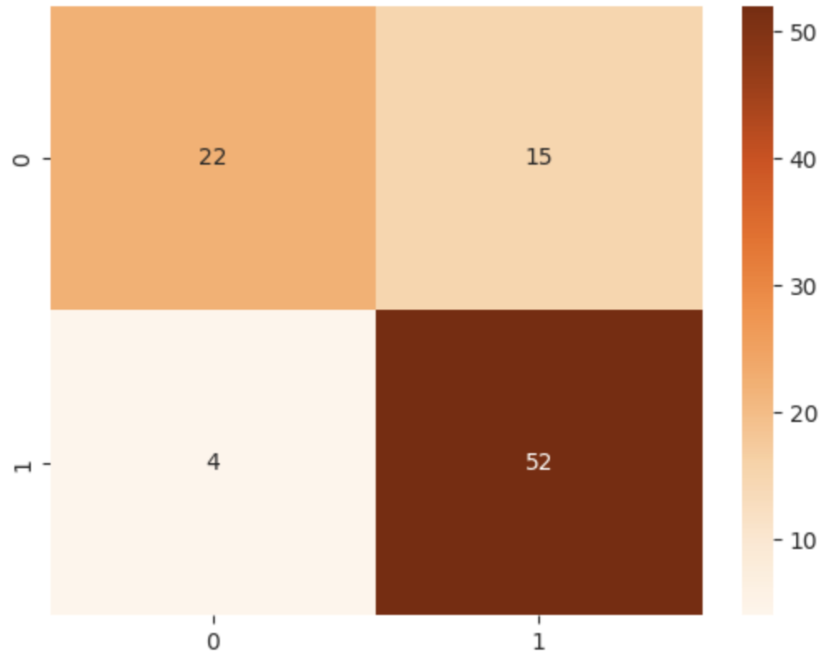
Appendix A: PCA

Principal Components	**Proportion of Variance**	**Cumulative Proportion**
Scoring	0.191	0.191
Opponent Strength	0.133	0.324
Second Chance	0.111	0.435
Free Throws	0.098	0.532
Defensive Rebounding	0.069	0.601
Celtic Strength	0.063	0.664
Perimeter Shooting	0.052	0.717
Turnovers	0.047	0.764
Free Throw Efficiency	0.044	0.808

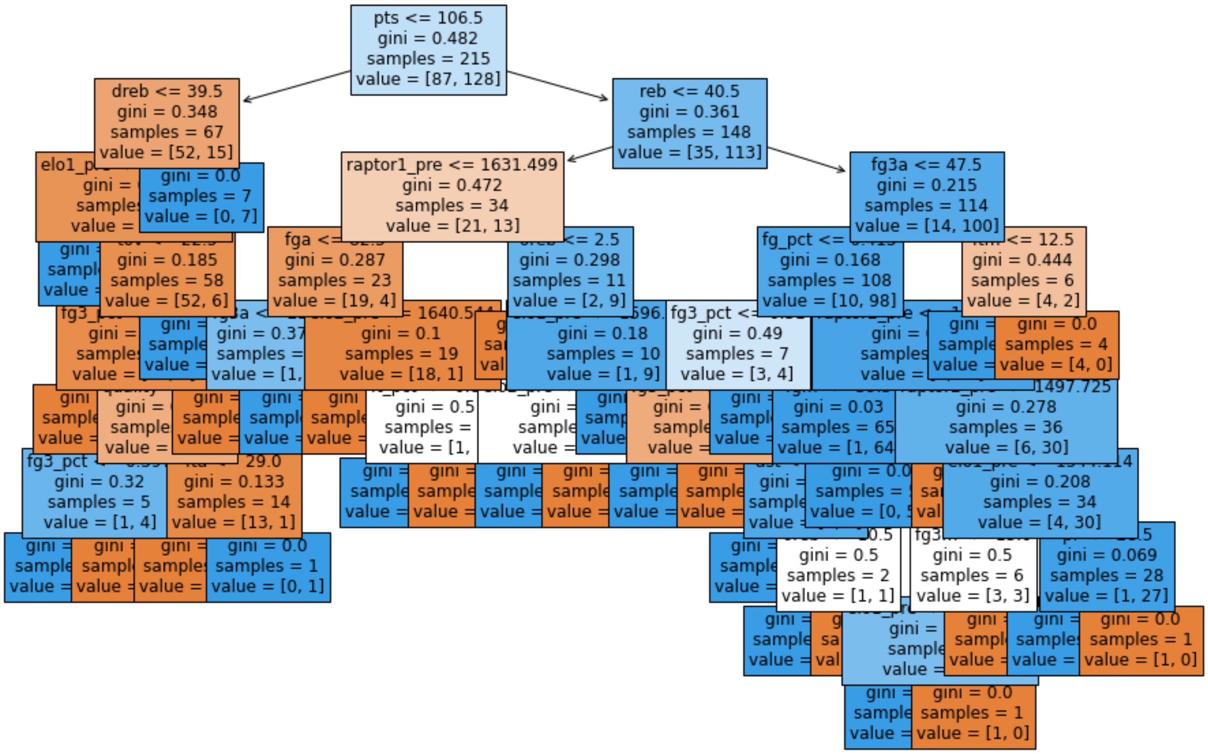
Appendix B: Logistic Regression Summary Statistic

Features	**Coefficients**
Scoring	0.98
Defensive Rebounding	-0.71
Free Throws	0.54
Turnovers	-0.48
Opponent Strength	0.44
Free Throw Efficiency	-0.28
Celtic Strength	-0.09
Second Chance	-0.04
Perimeter Shooting	-0.03
Intercept	0.88

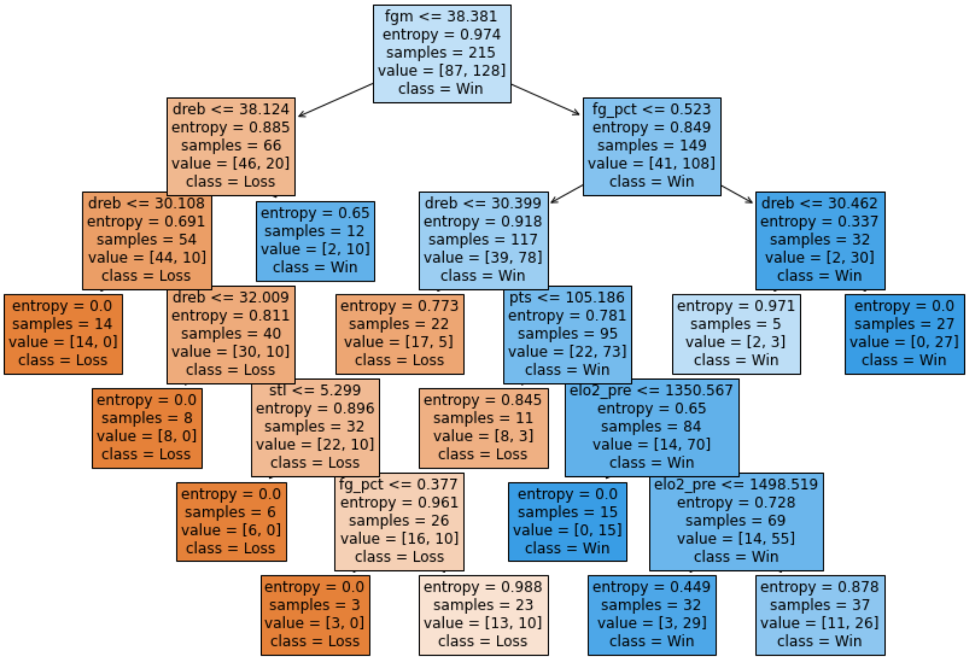
Appendix C: KNN Confusion Matrix



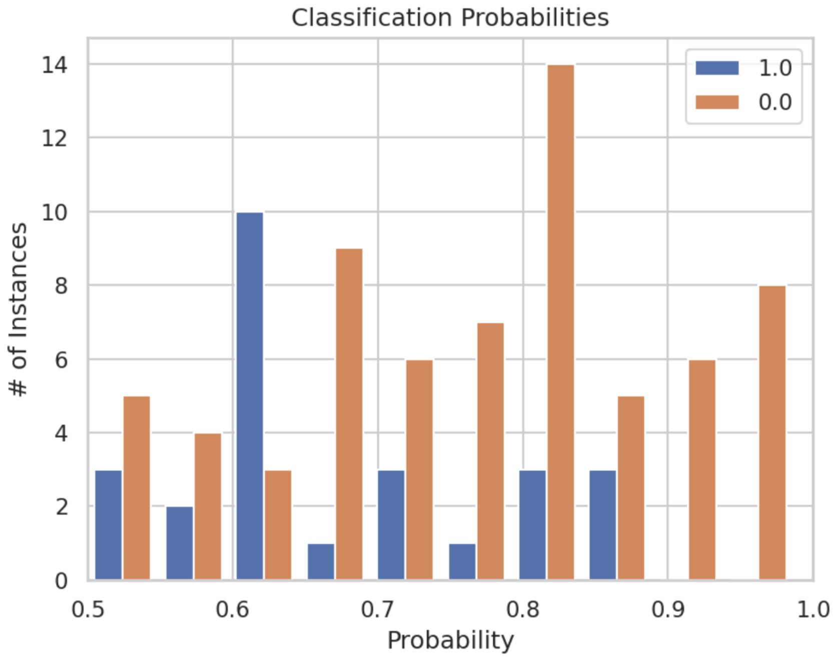
Appendix D: Full Tree Visualization:



Appendix E: Pruned Decision Tree:



Appendix F: Random Forest Classification Probability


Classification Probabilities
# of Instances
Probability