Data 102 Final Report

Angel Lee, Ansh Gupta, Antoine Lavoie, Nishka Govil

December 11 2023

1 Data Overview

The NBA games dataset on Kaggle was generated by being sourced directly from the Official NBA Stats website. It is a census. The only group that was systematically excluded from the data was bench players who did not directly participate in the NBA season's games. Participants are aware of this collection, as all NBA players are aware that the game stats and their own player stats will be visible on the Official NBA Stats website.

We have four different files: games.csv, games_details.csv, teams.csv, and rankings.csv, with data from the 2004 season to the 2022 season. The granularity of the data is quite high. For each of the files: Games.csv: each row represents a particular game and the game stats for the home and away team for that game (think field goal percentage, rebounds, etc.) Games_details.csv: each row represents the game stats for each player on the team for each game. Teams.csv: each row represents one of the thirty NBA teams. Stats include when the team is founded, what their arena capacity is, where they are based, etc. Rankings.csv: each row represents the thirty NBA teams. Stats include what the team's win%, road/home records are etc. in 2019.

Selection bias, measurement error, and convenience sampling are all irrelevant in the context of the data. This is because selection bias and convenience sampling do not apply because every individual in the population (players in NBA games) are included in this data set, as this is a census. Measurement error does not apply because the types of variables being measured in the dataset are not likely to be observed inaccurately. For instance, variables like the number of assists or rebounds in a game are not prone to measurement error since these are discrete variables that anyone watching the game can simply count. In addition, the data was collected by the Official NBA Stats website, which has to uphold a high level of accuracy in reporting game statistics. Furthermore, the dataset was not modified for differential privacy, as it is important to be able to see the names of each player and to analyze their true statistics.

There were some features/columns that we wish we could've had but were unavailable in the dataset. For instance, the number of years that a player has been playing, as well as the trading value of a player would have been helpful to answer our research question of what factors can predict whether the home team will win in a particular game.

In the games_details.csv file, all of the game identifier columns (e.g. GAME_ID, TEAM_ID, PLAYER_ID, etc) had values, but many of the game stats columns had NaN. In particular, we discovered that if there is NaN in one of the game stats columns, then all the other game stats columns would have NaN. By processing the COMMENTS column, we were able to conclude that these missing entries are a result of players being registered on the team but did not play in that game due to reasons such as injuries, personal reasons, blocked from participating by the NBA, etc. Since these players didn't play, they would have no impact on the state of the game, and so in this case it is fair to impute these missing values with 0s. In another file, games.csv, there were 99 rows (0.014% of the data) that had NaN values. They all happened during consecutive days of one season, so the missing values might be a result of those days' data not collected. As it is a very small sample size and were also earlier days of the season (i.e. aren't playoffs so aren't high stake games), we decided to drop these rows as they have little impact on the overall game outcome prediction.

Besides the imputation and removal of some data as mentioned above, for the GLM section we also log-transformed many of the game stats columns. This is because there are some significant outliers in our data, resulting in most of our data being heavily right-skewed. We thus chose log-transform rather than standard scaling because of several reasons:
1. We are dealing with a lot of percentages in our data, so we have a lot of multiplicative relationships. Log-transform helps to linearize our data more.
2. Our game stats are not really on different scales, so a standard scaling proved to make not much of an impact on our model outcomes.
3. The heavy skewness due to some significant outliers means that log-transform helped with the variance in our data and made it more symmetric.

Beside log-transformation, for the GLM section we also split our data into training and test sets prior to any feature engineering/ modeling so as to prevent data leakage and ensure the integrity of our results.

2 Research Questions

2.1 Bayesian Hierarchical Modeling

Research Question: Based on whether the game is during the COVID "Bubble" season of 2019-2020, do teams play more aggressively (i.e. have higher average foul count)?
Based on the fact that aggressiveness score is completely unobserved and defined by us, Bayesian Hierarchical Modeling proves to be the best available method to

gain insight into the research question. The flexibility of defining the priors for a team's aggressiveness score and the likelihood of our observed variables presents us the opportunity to control the model based on our idea of aggressiveness. However, this can also be a limitation since our choices can introduce subjectivity. Another limitation of this method is the computational intensity of Bayesian models. Given that we are working with a large set of data, we had to narrow our scope to focus only on team level granularity rather than player level.

2.2 GLM/Nonparametric Methods

Research Question: Given a particular game, will the home team win based on the team's average player statistics and team statistics? By answering this question, we can apply these findings toward several real-world decisions. For example, coaches can use these findings to determine which players should be selected to play in a game or which game strategies to employ to maximize the team's chance at winning. Furthermore, people that engage in fantasy sports can use the results of this question to help predict whether a team will win.

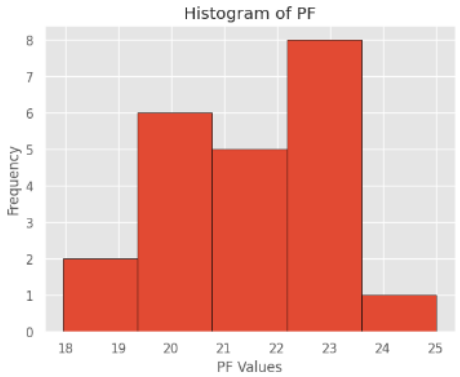
GLM and nonparametric methods are good methods to employ for this question because this is a prediction question where we are looking to predict whether a team will win based on a variety of factors. GLM's model the relationship between a number of different explanatory variables against the variable we are trying to predict and thus can be used to predict the response variable for unknown data. Nonparametric methods are also good methods to employ for this question because they are able to handle complex relationships between the game/player statistics and the variable of whether a team will win. These methods also do not make any assumptions about the underlying distribution of the data, which is helpful in this context.

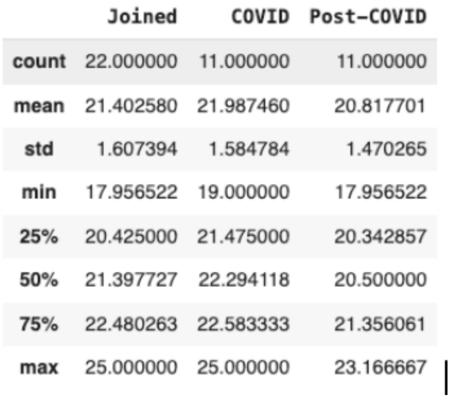
The limitations of using GLM's are that they depend on the assumptions of independence of errors, where each observation is independent, as well as the assumption of no multicollinearity. If these assumptions are violated, then the coefficients derived from the GLM may be inaccurate. The nonparametric method we chose was K-nearest-neighbors. One limitation is that when there is a large sample size, it would be computationally expensive to be searching through all rows in the dataset in order to determine which observation in the training set that the test observation is closest to. Furthermore, it is overly sensitive to outliers, which may result in a lower accuracy in those circumstances.

3 EDA

3.1 Bayesian Hierarchical Modeling

Based on the graphical model paired with our research question, we need to change the granularity of our data set. It is important to note that the current merged dataset is a combination of games_csv and games_details_csv sourced from the original Kaggle resource. Each row represents a NBA player who participated in a game, where there can be multiple rows for the same player if they participated in multiple games during the timeframe. In order to achieve a dataset where each row represents a team, we chose to group by the "GAME_ID" and "TEAM_ID" to ensure that each game has two rows (one for the home team, one for the away team), using sum as the aggregation method to total the relevant statistics for each respective team ("PF": personal fouls). Since each team played a different number of games based on their playoff performance, we decided to average the personal fouls total per game by the number of games played in the playoffs.





	Joined	COVID	Post-COVID
count	22.000000	11.000000	11.000000
mean	21.402580	21.987460	20.817701
std	1.607394	1.584784	1.470265
min	17.956522	19.000000	17.956522
25%	20.425000	21.475000	20.342857
50%	21.397727	22.294118	20.500000
75%	22.480263	22.583333	21.356061
max	25.000000	25.000000	23.166667

The histogram visualization references the "Joined" data frame, consisting of 22 teams (11 for COVID and 11 for Post-COVID) to get a general idea of the spread and distribution of personal fouls in the NBA playoffs. This is important to understand as we will be deriving the Likelihood for our X1 through Xn which are the observed average number of total fouls per game per team. We can note that the largest bar has a frequency of 8 for the bin between 22 and 23. That said the average personal fouls during both time frames sits at 21.40 where the COVID average is slightly higher and the post-COVID average is slightly lower. We can also note that the variance for COVID personal fouls is larger than that of Post-COVID indicating the tighter cluster around the mean during COVID.

Both of these visualizations provide a foundation for our further analysis and application of personal fouls data, a quantifiable attribute, to assess the aggressiveness of a team, which is a variable that is not directly observed or previously examined. We can see a consistent trend across all of the statistical measures in that there is a general decrease in the number of personal fouls post-COVID. Specifically through the violin plot, which is a hybrid between a box plot and KDE plot, we can see the larger variability in the number of personal fouls during the COVID period with more outliers or extreme values. This warrants some interesting future research regarding why this might be. Does

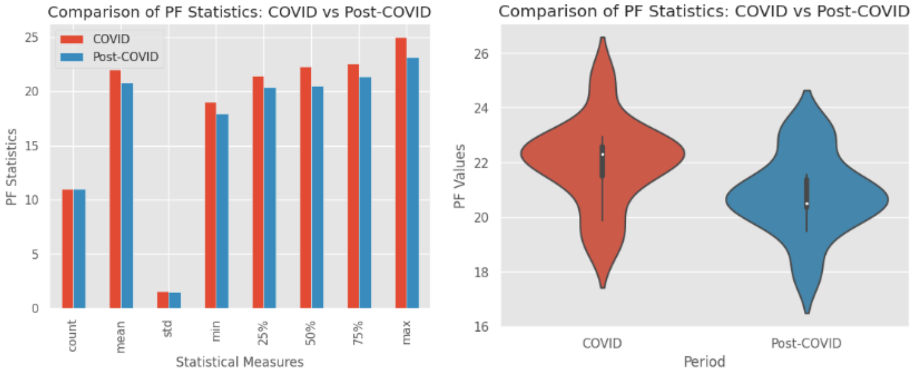
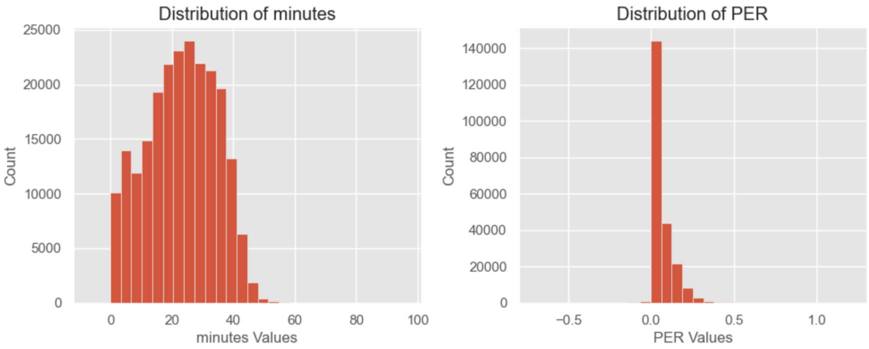


Figure 1: 12345

this change in distribution reflect adaptations in gameplay, rule enforcement, and/or a more aggressive team?

3.2 GLM/Non-Parametric



These four visualizations show the distribution of some game stats columns that were feature engineered and didn't have a visualization of their distributions on Kaggle. As we can see, TeamChemistry and PER are heavily right-skewed. This informed us of the need to log-transform our data and to observe how the distribution changes post-transformation.

For the GLM section, this correlation heatmap helped us see how much each variable correlated with the response variable, HOME_TEAM_WINS. Evidently, none of the variables had a high linear correlation (highest is PER, 0.28), which suggested to us that we need to use a link function (logit) to transform our response value so as to linearize the relationship and allow more accurate predictions.

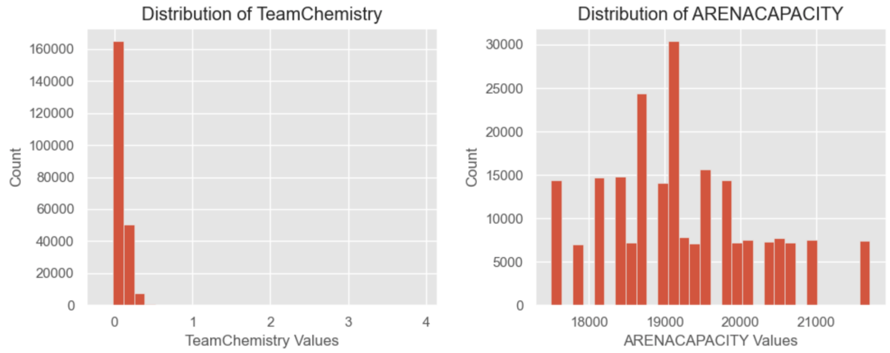


Figure 2: Distribution of engineered and select features

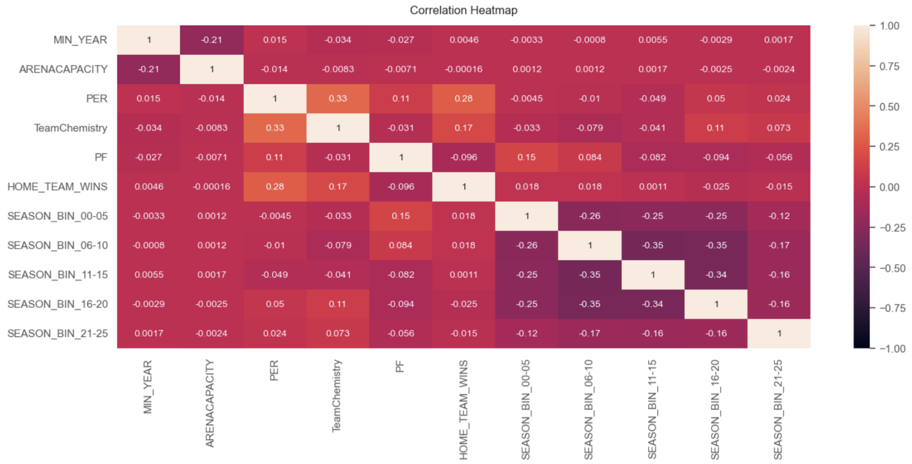


Figure 3: Correlation heatmap between features

4 Option B: Bayesian Hierarchical Modeling

4.1 Methods

We used a beta(1,1) prior for both during and after covid. We chose this non-informative prior which corresponds to a uniform distribution on the interval [0, 1] because we didn't know much about the aggressiveness of teams during & after covid. We can consider trying to implement an empirical Bayes model with a more informative (weak) traditional prior for aggressiveness based on our intuition/data analysis as opposed to an empirical prior that is data driven. Given that we are using averages, our data falls under continuous data. Along with that the possible ranges for personal fouls is between 0 and infinity. The aggressiveness score should fall between 0 and 1. This eliminates the use of

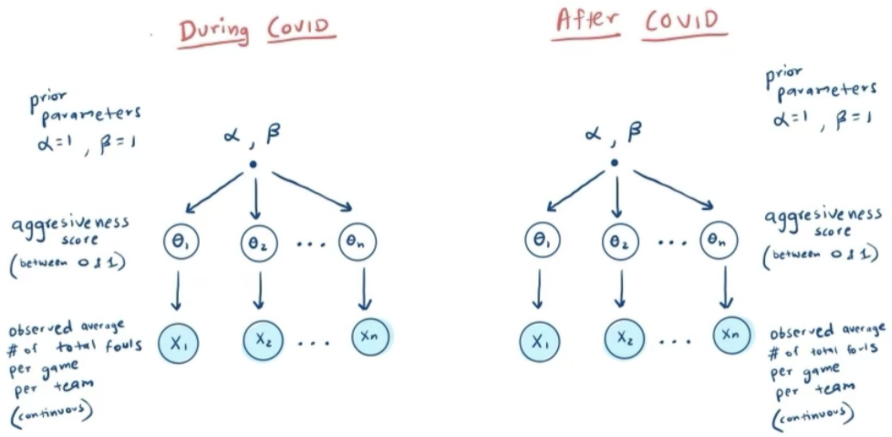


Figure 4: Bayesian Graphical Model: During COVID and Post-COVID

likelihoods/distributions like normal, binomial, and beta. Therefore based on the support of the gamma distribution being continuous and positive we know that this is our best bet. As for the parameters of the distribution, they must be a function of theta but can be interpreted by solving for the observed mean and variances calculated from our two datasets respectively.

4.2 Results

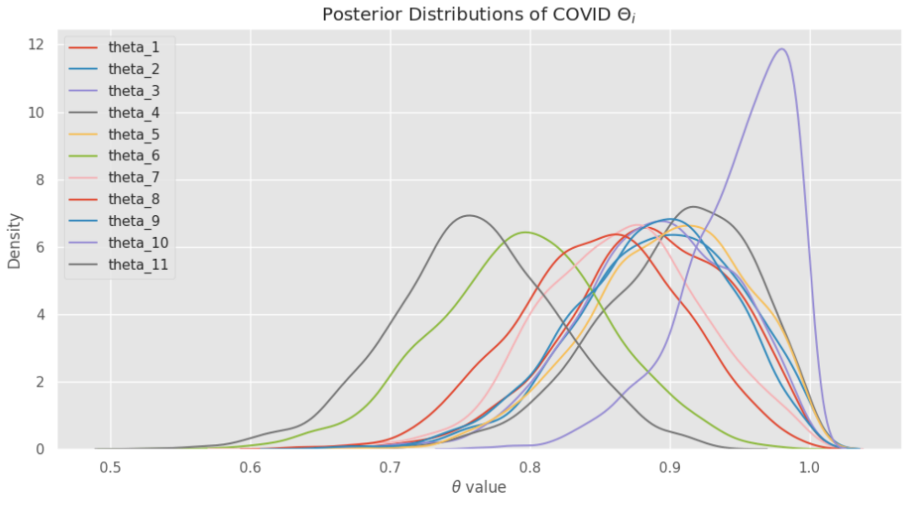


Figure 5: Visualized Posterior Distributions: COVID Time frame

In the during COVID posterior distributions, we can see that team 10 has the highest aggressiveness score, with the center of its distribution being somewhere around 0.97. In contrast, team 4 and 6 had the lowest, with the center being around 0.75 and 0.79 respectively.

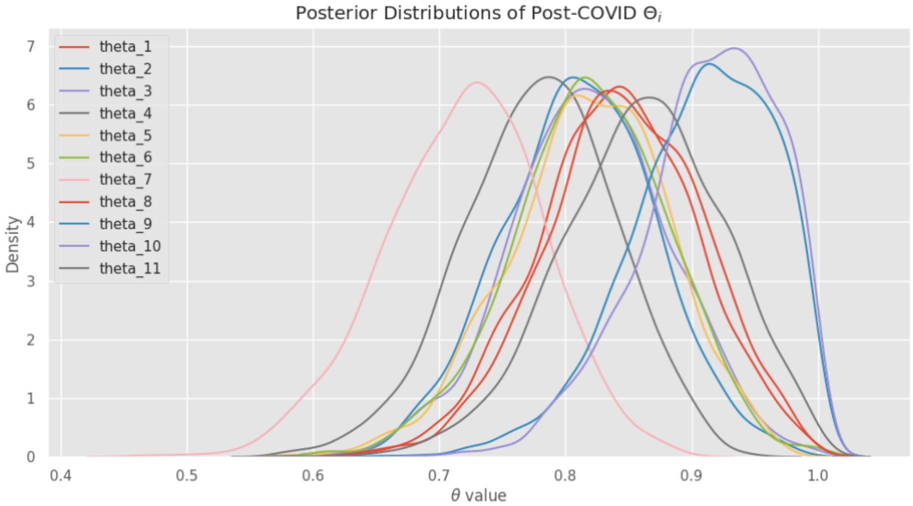


Figure 6: Visualized Posterior Distributions: Post-COVID Time frame

In the Post-COVID Posterior Distributions, while the spread seems relatively similar, the center of the distributions varies for each team. Specifically, we can see that teams 9 and 10 have posterior estimates that are closer to 1 than the other teams. The center for these two teams' distribution lies somewhere around 0.92. This indicates that they are on the more aggressive side. In contrast, team 7 has the lowest aggressiveness score with the center of the distribution lying around 0.73. The other teams lie somewhere in the middle between team 7 and team 10.

We plotted the 2000 posterior samples that we created for both during COVID and Post-COVID to visualize the distribution for each. This histogram helps us isolate each team's distribution and get a good idea of whether the distribution for their aggressiveness score has shifted between each time period of COVID and Post-COVID. For example, we can see that Teams 1, 2, 3, 4, 5, 7, and 8 all seem to have the center of their distribution shifted to the left for Post-COVID. This means that these teams were more likely to have higher aggressiveness during COVID than Post-COVID. Team 6, 9 and 11 followed a different distribution, where the mean of the distribution was shifted to the right for Post-COVID, meaning that they seemed to display higher aggressiveness scores after COVID. It is important to note that Team 6 and 9 were only slightly shifted to the right in comparison to some of the other distributions we noted earlier. Team 8 followed relatively similar distributions for both during COVID and Post-COVID, indicating that there didn't appear to be a signif-

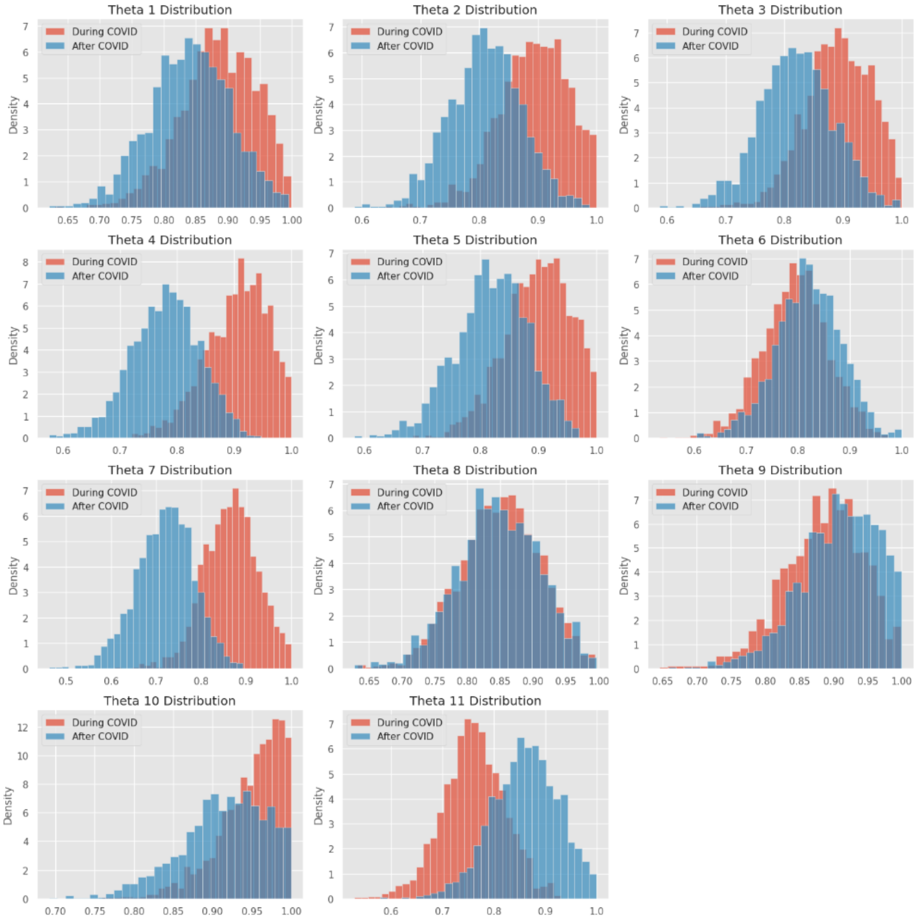
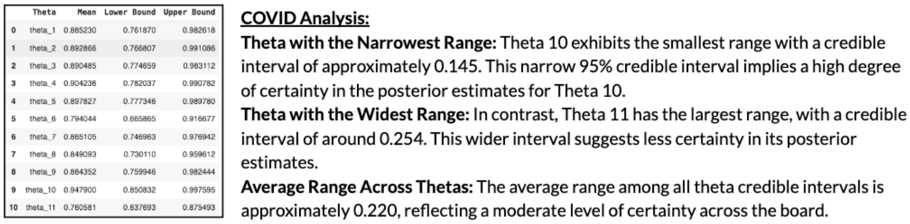


Figure 7: Overlaid Posterior Distributions: COVID vs. Post-COVID

icant change in aggressiveness for that team. We were surprised to see 7 out of the 11 teams actually were more aggressive during COVID. We had initially assumed that the lack of fans and area atmosphere would contribute to less aggressiveness amongst teams; however, the results show otherwise. Moreover, Team 10 stood out from the rest in terms of aggressiveness. The distribution was left skewed, indicating that the aggressiveness score was closer to 1 than other teams.

Quantifying uncertainty: Our model, configured with specific parameters (chains=1, tune=1000, target_accept=0.95, return_inferencedata=True), generates a nested array of 2000 samples for each of the eleven thetas, representing the posterior distribution of each team's aggressiveness score. By calculating 95% credible intervals for these theta samples, we establish lower and upper bounds within which we expect the true aggressiveness score of each team to fall with 95% probability. These Bayesian credible intervals, unlike Frequentist confidence intervals, directly relate to the probability of the parameter values themselves. The empirical range between these bounds, coupled with our understanding of aggressiveness scores constrained between 0 and 1, helps us assess the certainty of our model's predictions.



	Theta	Mean	Lower Bound	Upper Bound
0	theta_1	0.885230	0.761870	0.982618
1	theta_2	0.892866	0.766807	0.991086
2	theta_3	0.890485	0.774659	0.983112
3	theta_4	0.904236	0.782037	0.990782
4	theta_5	0.897827	0.777346	0.989780
5	theta_6	0.794044	0.665865	0.916677
6	theta_7	0.865105	0.746963	0.976942
7	theta_8	0.849093	0.730110	0.959612
8	theta_9	0.884352	0.759946	0.982444
9	theta_10	0.947900	0.850832	0.997595
10	theta_11	0.760581	0.637693	0.875493

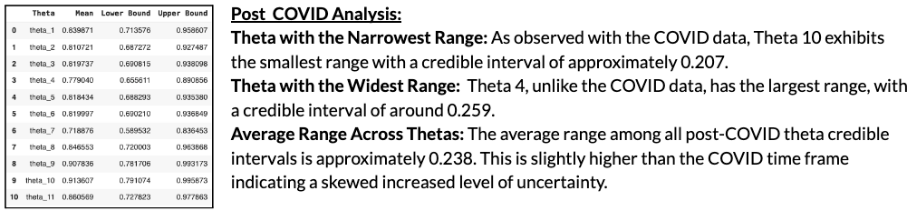
COVID Analysis:

Theta with the Narrowest Range: Theta 10 exhibits the smallest range with a credible interval of approximately 0.145. This narrow 95% credible interval implies a high degree of certainty in the posterior estimates for Theta 10.

Theta with the Widest Range: In contrast, Theta 11 has the largest range, with a credible interval of around 0.254. This wider interval suggests less certainty in its posterior estimates.

Average Range Across Thetas: The average range among all theta credible intervals is approximately 0.220, reflecting a moderate level of certainty across the board.

Theta 10 having the smallest range and therefore the narrowest 95% interval of samples indicates that our model is most certain regarding the posterior estimates for Theta 10. Throughout both timeframes, this team obtained the largest number of average fouls per game. This maximum served as a key metric in our Gamma Likelihood parameters, setting E[Xi] = maximum personal fouls * theta_i. Therefore this results aligns with our intuition as this team had a prominent role in our definition of the aggressiveness score. This understanding is confirmed with the tighter clustering of this team's posterior samples near the value 1 – evidenced in the figures from the results section – as well as the Team with the minimum number of fouls during the COVID timeframe being the least certain/widest interval.



	Theta	Mean	Lower Bound	Upper Bound
0	theta_1	0.839871	0.713576	0.958607
1	theta_2	0.810721	0.687272	0.927487
2	theta_3	0.819737	0.690815	0.938098
3	theta_4	0.779040	0.655611	0.890856
4	theta_5	0.818434	0.688293	0.935380
5	theta_6	0.819997	0.690210	0.936849
6	theta_7	0.718876	0.589532	0.836453
7	theta_8	0.846553	0.720003	0.963868
8	theta_9	0.907836	0.781706	0.993173
9	theta_10	0.913607	0.791074	0.995873
10	theta_11	0.860569	0.727823	0.977863

Post COVID Analysis:

Theta with the Narrowest Range: As observed with the COVID data, Theta 10 exhibits the smallest range with a credible interval of approximately 0.207.

Theta with the Widest Range: Theta 4, unlike the COVID data, has the largest range, with a credible interval of around 0.259.

Average Range Across Thetas: The average range among all post-COVID theta credible intervals is approximately 0.238. This is slightly higher than the COVID time frame indicating a skewed increased level of uncertainty.

Based on the larger average range across thetas, this model displays moderately increased uncertainty. Notably Theta 10, associated with Team 10 is still the most credible. Within the post_COVID data, this team has the max number of personal fouls, albeit still lower than the COVID data. It is interesting to note that Team 4 is now the most uncertain, somewhat interesting considering their personal values fall in the middle range near the average personal fouls level during the post_COVID time frame.

Overall, we can note that the size of the credible intervals remain fairly consistent between both timeframes. Given that our aggressiveness score has

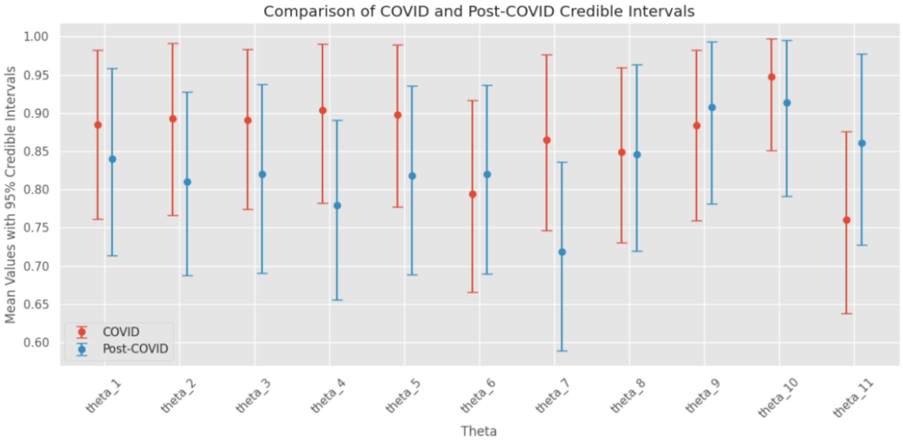


Figure 8: Posterior Estimates and Credible Interval Comparison

possible values between 0 and 1, we can note there is a moderate level of uncertainty across the teams posterior estimate samples.

4.3 Discussion

Elaborating on the limitations: Since the choice of the priors can greatly affect the results, our weak prior of Beta(1,1) is not ideal. This means that we are assuming equal prior knowledge of the successes and failures. Moreover, by choosing the Gamma distribution for the likelihood, we are assuming a specific shape for the data distribution. There is a chance that this assumption might not hold, making it difficult to capture the true patterns. Moreover, there is omitted variable bias since we aren't able to account for all relevant covariates or unobserved variables that affect aggressive behavior for NBA players. Since the model is based on observed data, unobserved confounding variables or missing data can greatly affect the underlying patterns. Lastly, a big limitation is that we need a way to connect our two models together. This is because there should be a relationship between the X_i variables (where i represents a specific team from the 11 teams) during COVID and Post-COVID. Since the teams are the same, they need to be connected in some way in both models. This means that we would need to create a separate model where the Gamma distribution for X_i in Post-COVID not only depends on the theta values but also the X_i from during COVID as well.

Convergence: Our inference procedure did not have trouble converging.

Other formulations/graphical models: In our initial modeling approach (seen in Checkpoint 2) with defining the appropriate Likelihood/Distribution for our observed variables X_1 through X_n, particularly challenged by the uneven data distribution, number of teams, players, and games during each respective playoff period. We first considered a Poisson likelihood for several reasons: its ability to model the probability of events occurring within a specific timeframe (akin to fouls in a game), its support for a count range from 0 to infinity, and the assumed independence of events, which aligned with our understanding of basketball fouls. However, the Poisson distribution, typically suited for discrete count data, was not sufficient as we were dealing with averages – continuous data. We briefly contemplated and tested the use of the Normal distribution given that averages are always normally distributed according to the Central Limit Theorem. We ruled it out due to the possibility of obtaining negative values which didn't fit the context of a score from 0 to 1. Ultimately we decided to look into more complex distributions we were less familiar with, settling with Gamma(α, β).

Useful Additional Data: As mentioned in regards to limitations, we recognize that there is some omitted variable bias given that we are defining a direct relationship between aggressiveness and personal fouls. Intuitively, there are many more variables in our accessible dataframe that could be used to create a more nuanced and representative score for aggressiveness.

In addition to our project scope we decided to take it a step further and test a new Bayesian model based on the following observed variables: personal fouls (X_i), steals (Y_i), and offensive rebounds (Z_i).

These are variables that undeniably can be mapped to a team's aggressiveness. Below are the results:

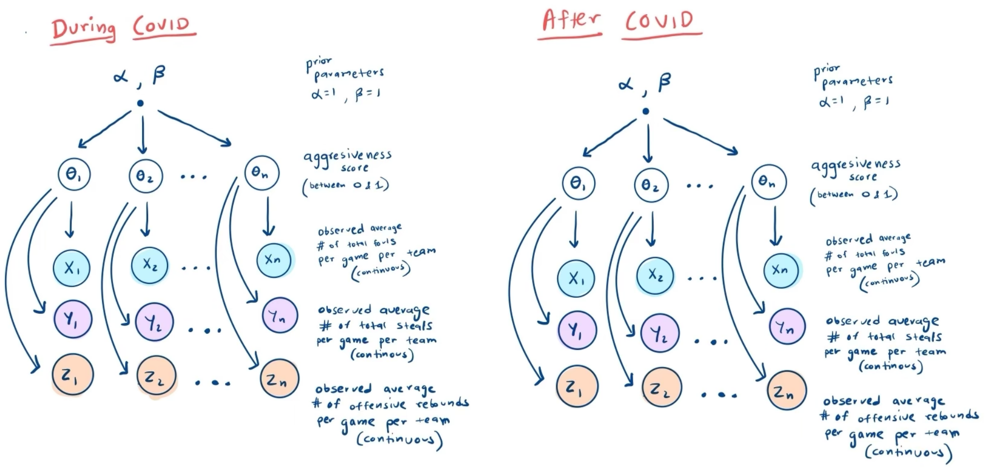

During COVID
α, β
prior parameters α=1, β=1
θ₁ θ₂ ... θn
aggresiveness score (between 0 & 1)
X₁ X₂ ... Xn
observed average # of total fouls per game per team (continuous)
Y₁ Y₂ ... Yn
observed average # of total steals per game per team (continous)
Z₁ Z₂ ... Zn
observed average # of offensive rebounds per game per team (continuous)
After COVID
α, β
prior parameters α=1, β=1
θ₁ θ₂ ... θn
aggresiveness score (between 0 & 1)
X₁ X₂ ... Xn
observed average # of total fouls per game per team (continuous)
Y₁ Y₂ ... Yn
observed average # of total steals per game per team (continous)
Z₁ Z₂ ... Zn
observed average # of offensive rebounds per game per team (continuous)

Figure 9: Nuanced Bayesian Graphical Model: COVID and Post-COVID

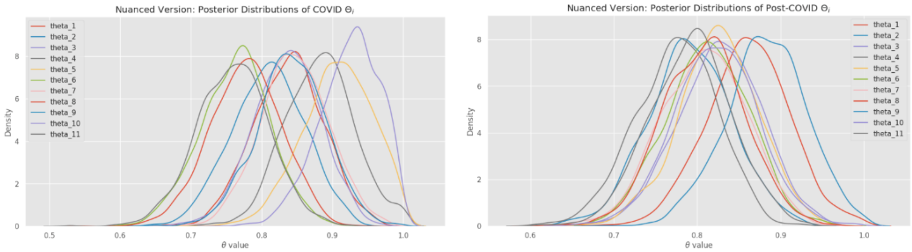


Figure 10: Visualized Posterior Distributions COVID and Post-COVID (Nuanced Model)

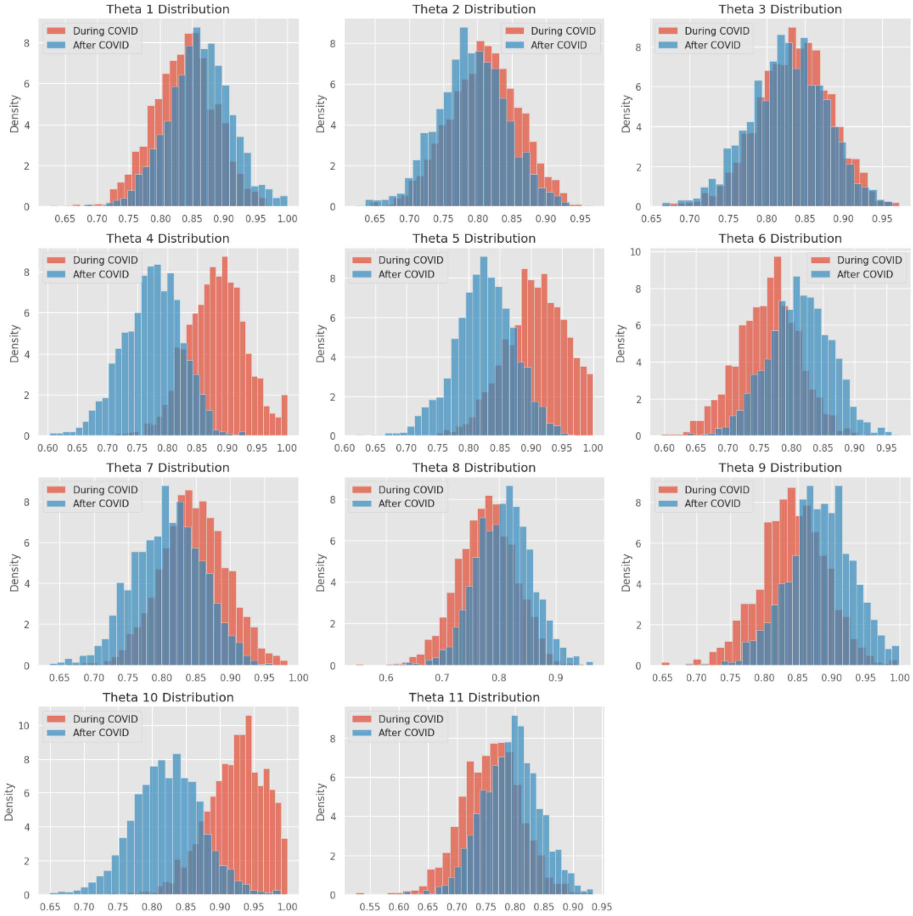


Figure 11: Overlaid Posterior Distributions: COVID and Post-COVID (Nuanced Model)

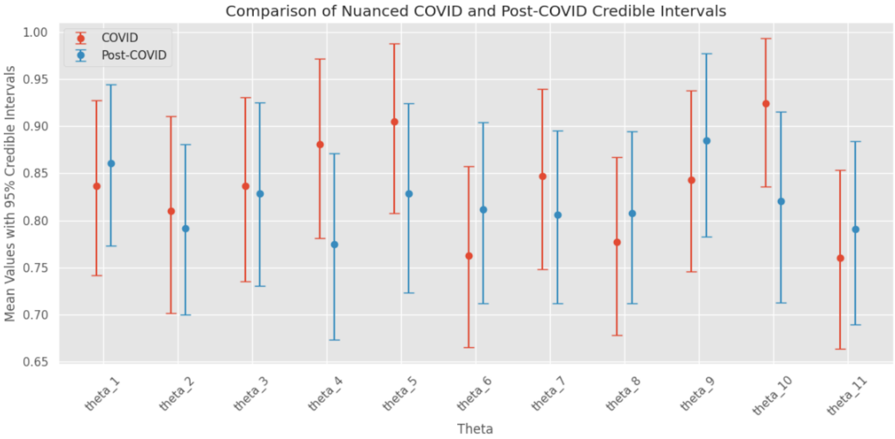


Figure 12: Posterior Estimates and Credible Interval Comparison

5 Option C: Prediction with GLMs and non-parametric methods

5.1 Methods

What we're predicting: We are trying to predict whether given a particular game, will the home team win based on the team's average player statistics and team statistics.
Features selection and justification: We selected the features in our model primarily through domain knowledge. Specifically, we identified game id, home team id, arena capacity, minimum year of the team into the NBA championship (min year), number of personal fouls, NBA season, and our two feature-engineered variables, player efficiency rating and team chemistry. We believed that teams that win in one year may be more likely to win in the subsequent year, which justified using game id, home team id, and NBA season. And we believed that the length of time a team has been in the NBA may represent that team's level of experience. We picked arena capacity because of a potential audience effect where an individual's performance is affected by the number of people watching them. We also picked the number of personal fouls to include in our dataset because it has a higher linear correlation to whether the home team wins than other variables, even if the correlation is low overall. We feature engineered a player efficiency rating modified from the formula listed on a Watts Basketball article: https://wattsbasketball.com/blog/how-to-calculate-player-efficiency-rating. And we felt that team chemistry was an important factor contributing to a team's success, which we quantified as the number of assists per minute of the game. For our GLM, we ran code that calculated the lowest AIC score for all possible combinations of features of size k where k ranges from 1 to 12, since there were 12 total features we looked at after doing one-hot-encoding for the NBA season variable. For our GLM, we then selected the subset of features corresponding to the lowest AIC. For the GLM, these features were the number of personal fouls, team chemistry, player efficiency rating, home team id, game id, and 3 bins for NBA seasons from 2000-2005, 2006-2010, and 2011-2015.

5.1.1 GLM

Chosen GLM model: logistic regression GLM
Justification: The variable we are trying to predict is "HOME_TEAM_WINS," which is a binary variable that equals 1 when the home team wins and 0 when the home team doesn't win. Because this is a binary variable, we would use logistic regression to predict it.
Assumptions: By using logistic regression, we are assuming independence of errors, where each observation is independent. We would expect that a residual plot would show no clear pattern. Another crucial assumption we are making is that there is no multicollinearity and that the variables are not correlated.

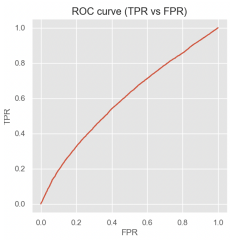


Figure 13: ROC Curve for Logistic Regression

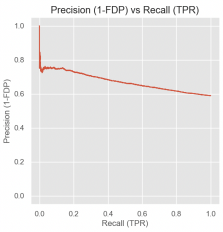


Figure 14: Precision-Recall Curve for Logistic Regression

Evaluation of model performance: We evaluated the GLM's performance by comparing the fitted values to the true values in the test set and determining the accuracy, which was 55.5% with a 0.6 threshold that we determined was optimal. We also calculated the TPR and FPR at different thresholds to plot an ROC curve, which we compared to the optimal ROC curve. We also examined the precision-recall graph at different thresholds for the model.

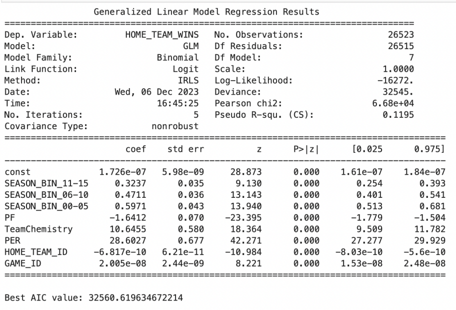
5.1.2 Non-parametric methods

Chosen non-parametric model: KNN
Justification: We tried out logistic regression, KNN, decision tree, and random forest. Since our research question is a binary classification one, we used accuracy as an indicator of model performance. We also plotted the ROC curve and Precision vs. Recall curves. For random forest and decision tree, we limited the depth and max features to 3 and 4, respectively, so as to prevent overfitting. After comparing these metrics, KNN proved to be the best model, with an accuracy of 68.5% of predicting the home team winning.
Assumptions: One of the key assumptions of KNN is that neighbors that are close to each other belong in the same class. There are many factors that contribute to whether or not a home team wins a game, and we have only selected a subset of them in our model so as to prevent a high dimension dataset. Since we have a binary classification, it means that there might be some data points that are close to one another given our features, but in reality belong to a different

class than what KNN classified it as given other features. This thus contributed to our accuracy performing decently but not the best (although a quick Google search yielded that the best NBA models out there only perform at roughly 70%!)

5.2 Results

5.2.1 GLM



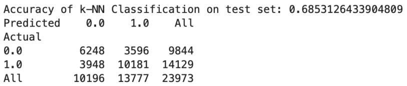
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:         HOME_TEAM_WINS   No. Observations:                26523
Model:                            GLM   Df Residuals:                    26515
Model Family:                Binomial   Df Model:                            7
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -16272.
Date:                Wed, 06 Dec 2023   Deviance:                       32545.
Time:                        16:45:25   Pearson chi2:                 6.68e+04
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1195
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const              1.726e-07   5.98e-09     28.873      0.000    1.61e-07    1.84e-07
SEASON_BIN_11-15      0.3237      0.035      9.130      0.000       0.254       0.393
SEASON_BIN_06-10      0.4711      0.036     13.143      0.000       0.401       0.541
SEASON_BIN_00-05      0.5971      0.043     13.940      0.000       0.513       0.681
PF                   -1.6412      0.070    -23.395      0.000      -1.779      -1.504
TeamChemistry        10.6455      0.580     18.364      0.000       9.509      11.782
PER                  28.6027      0.677     42.271      0.000      27.277      29.929
HOME_TEAM_ID      -6.817e-10   6.21e-11    -10.984      0.000   -8.03e-10    -5.6e-10
GAME_ID            2.005e-08   2.44e-09      8.221      0.000    1.53e-08    2.48e-08
=====================================================================================

Best AIC value: 32560.619634672214
```

Figure 15: Logistic Regression Output for GLM

The result from the model we selected shows that a higher average number of personal fouls across players in a team results in a decrease in the probability of a home team winning. Specifically, a one unit increase in the log-transformed average number of personal fouls in a team results in a lower log-odds ratio of a home team winning by 1.6412. In addition, a one unit increase in the log-transformed average team chemistry results in a higher log-odds ratio of a home team winning by 10.6455. A one unit increase in the log-transformed average number of personal fouls results in a higher log-odds ratio of a home team winning by 28.6027. Furthermore, for the non-log-transformed features, a one unit increase in the home team id results in a lower log-odds ratio of a home team winning by -6.817e-10. A one unit increase in the game id results in a higher log-odds ratio of a home team winning by 2.005e-08. And for the one-hot-encoded season features, the odds of a home team winning is 2.01 times higher when a season is between 2000-2005, compared to when it is outside this time frame; 1.6 times higher when a season is between 2006-2010, compared to when it is outside this time frame; and 1.38 times higher when a season is between 2011-2015, compared to when it is outside this time frame.
Uncertainty: We are 95% confident that the true coefficient for 'SEASON_BIN_11-15' lies somewhere between 0.254 and 0.393. We are 95% confident that the true coefficient for 'SEASON_BIN_06-10' lies somewhere between 0.401 and 0.541. We are 95% confident that the true coefficient for 'SEASON_BIN_00-05' lies somewhere between 0.513 and 0.681. We are 95% confident that the true coefficient for 'PF' lies somewhere between -1.779 and -1.504. We are 95% confident

that the true coefficient for 'TeamChemistry' lies somewhere between 9.509 and 11.782. We are 95% confident that the true coefficient for 'PER' lies somewhere between 27.227 and 29.929. We are 95% confident that the true coefficient for 'HOME_TEAM_ID' lies somewhere between -8.03e-10 and -5.6e-10. We are 95% confident that the true coefficient for 'GAME_ID' lies somewhere between 1.53e-08 and 2.48e-08.

5.2.2 Nonparametric



```
Accuracy of k-NN Classification on test set: 0.6853126433904809
Predicted    0.0    1.0    All
Actual
0.0         6248   3596   9844
1.0         3948  10181  14129
All        10196  13777  23973
```

Figure 16: Confusion matrix for KNN

We made a confusion matrix for each of the nonparametric models we tested out. The accuracy of each of are as follows:

Logistic regression: 55.5%
KNN: 68.5%
Decision tree: 56.0%
Random Forest: 57.7%

This means that KNN gave us the best classification accuracy, followed by random forest, decision tree, and then logistic regression.

Taking a look at each of our model's ROC and Precision-Recall curves:

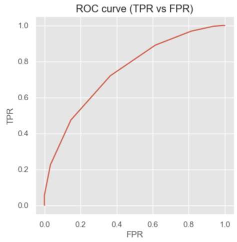


Figure 17: ROC curve for KNN

KNN: its ROC curve is above the line of no discrimination, indicating that the model has a good measure of separability. The curve also rises quickly towards a high TPR, low FPR, but plateaus, so there are room for improvement. On the precision-recall side: as recall increases, the precision rapidly decreases. All of these combined indicate that there might potentially be a class imbalance.

Decision tree, random forest, and logistic regression: the ROC curves for these

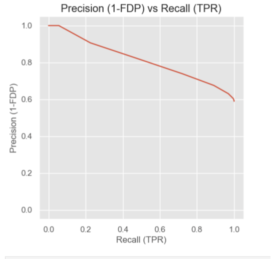


Figure 18: Precision-recall curve for KNN

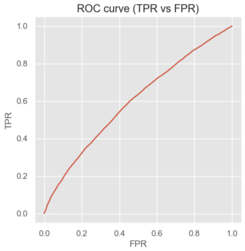


Figure 19: ROC curve for Decision Tree

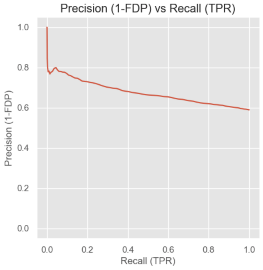


Figure 20: Precision-recall curve for Decision Tree

three models are all quite similar. They barely made it above the line of no discrimination, and does not rise quickly enough towards a high TPR while having a lower FPR as the KNN model does. On the precision-recall side: as recall increases, the precision decreases much faster than the KNN model, indicating a worse performance.

5.3 Discussion

The k-NN model performed better than the GLM because it had a higher accuracy (68.5% compared to 55.5%). The ROC curve of the k-NN model is also closer in shape to the perfect ROC curve that has an AUC of 1, compared to the ROC curve of the GLM. We are confident in applying this model to future datasets because in future NBA seasons, many of the factors contributing to whether a team wins will likely remain constant. In other words, factors like the rules of the game and the teams that tend to be strong are likely to stay the same even in future years.

The logistic regression model fits the data only moderately well. When we run the model on the test set, we generate an accuracy of 55.5%.

KNN fits the model relatively well, although there is definitely room for improvement. As aforementioned, there might be class imbalances, and indeed there are about 14,000 home team wins, but only roughly 9,900 home team losses. To address this, we might have considered oversampling the loss class, undersampling the win class, or as shown in our TPR-FPR threshold table, adjust the threshold to be 0.6 rather than 0.5 to achieve a better TPR and FPR overall. From the confusion matrix, we can see that the KNN model classifies true negative classes 63.5% of the time, and classifies true positives 72.1% of the time. There is thus a better fit for true positives than true negatives, again hinting at a potential class imbalance that we should've addressed and see if that improves the model performance as well as its integrity.

The frequentist implementation of the GLM had a higher accuracy than the Bayesian implementation. The L2 regularized logistic regression, which was the Bayesian implementation, gave a poor accuracy of 45% because it relies on Gaussian priors, which are not good priors. They assume that our data is Normally distributed but our data is not.

To interpret the results of the GLM, a higher home team id or log-transformed average number of personal fouls across players in a team results in a decrease in the probability of a home team winning. In contrast, a season bin between 2000-2005, 2006-2010, or 2011-2015 results in a higher probability of a home team winning. Furthermore, a higher log-transformed average team chemistry score, log-transformed average player efficiency rating, or game id in a team results in an increase in the probability of a home team winning.

To interpret the KNN model, we look at how well it performs on the accuracy percentage, as well as the ROC and precision-recall curve. A higher accuracy, a closer to 90 degrees angled ROC curve, and a slow decreasing precision-recall curve indicates optimal model performance. Our KNN model had a relatively decent accuracy score, 68.5%, and a higher than no discrimination ROC curve and a relatively fast downward sloping precision-recall curve. Our model is therefore satisfactory, but could be improved.

A limitation of using GLMs is that it assumes a linear relationship between the features (which can be transformed) to the transformed response variable (i.e. whether a home team wins). This prevents modeling complex relationships the way nonparametric methods can. Furthermore, logistic regression requires two key assumptions: independence of errors, where each observation is independent, and a lack of multicollinearity.

There are several limitations to KNN.

Sensitivity to feature scaling: since KNN relies upon the distance between data points to classify, it assumes that there is no unfair scaling of certain features that would dominate the calculation. We performed log-transform so as to ensure this.

"Curse of dimensionality": the higher the dimension, the less meaningful "closeness" is, and therefore reduces the usefulness of KNN. We thus did dimension reduction during our feature selection stage.

No noisy data/ outliers: KNN is sensitive to outliers and noise, so our log transform helped with this and assumed that there is no significant noise/ outlier in our data.

Access to additional features would be useful in creating more accurate models. If we had data on, for example, the number of years that a player has been playing, or the trading value of a player, we would have incorporated these into our GLM and non-parametric models. Furthermore, if we had had additional information on the coaches and their prior success rate, or the injury status of a player, we could have added these features, as well.

Our dataset size is quite large (559498 usable rows) and spans NBA games from 2004 to 2022, hence uncertainty on this front is quite low if tried to apply it onto future game stats (unless NBA rules change significantly). The data, however, was quite noisy; there were some significant outliers. This is as expected, since NBA players can have highly variable performances from game to game. Some players may have exceptional games where they score significantly more points, make more rebounds, or have more assists than their average performance. There are also star players, and if they are rested/ injured during a game it can significantly impact team statistics, leading to the team's performance deviating significantly from the norm.

6 Conclusion

6.1 Bayesian Hierarchical Modeling

Contrary to our initial hypothesis, we noticed that the majority of teams were actually more aggressive during COVID than Post-COVID. Specifically, 7 out of the 11 teams displayed a higher aggressiveness score in the posterior estimates.

This indicates that playing in "The Bubble" during the pandemic might have actually contributed to the rise in aggressiveness amongst players. Multiple external factors such as gameplay strategies and team dynamics might have shifted as a result of COVID, impacting the personal fouls rates during the time period. Through this, we found that there is a need to further investigate this trend and build a more comprehensive model. Currently, we are limited by our weak prior choice and simplified approach of defining aggressiveness. By including a multitude of factors such as steals and offensive rebounds, we can take on a more holistic approach to aggressiveness and see if there is a noticeable difference during the COVID and Post-COVID season.

Our initial modeling approach using only personal fouls to define an aggressiveness score is a relatively specific and hyper focused approach towards answering our research question. In reality our findings only give insight into one aspect of aggressiveness when looking at NBA games. This said, our proof of concept of modeling aggressiveness through fouls, steals, and offensive rebounds is more generalizable in terms of defining aggressiveness and actually confirmed our findings using only personal fouls. Through another lens, the COVID Bubble was the first time anything like this has happened during an NBA season. This rare event in itself results in narrow findings that cannot be generalized towards any nba statistics but can be used to derive insights into player, arena, and psychological dynamics. Along with this, looking only at the season after COVID narrows the scope of our project given that COVID could have had a lasting impact on NBA players just as it has had for the majority of people around the world. Applying our modeling strategies to many seasons before and after COVID would help set context and create more generalizable findings.

Merging Different Sources: While both our datasets were pulled from the NBA Stats Website, we combined different team levels with player level statistics. This allowed us to feature engineer team-level statistics like total personal fouls, steals, and rebounds.

We urge The NBA Board to further investigate these trends. There were a lot of factors that were different during the COVID Bubble which evidently resulted in more aggressive players. They should take one step further and find specific factors that most impact aggressiveness, stimulating them to revitalize future seasons. Through this analysis, the Board can better monitor their players' behaviors while being more cognizant of how factors such as audience and location impact a team. Whether it is Sports Betting Organizations (i.e Draft Kings) or Sports Analysts (ESPN), we believe that they should employ our aggressiveness scores as a feature for predicting a team's performance in a playoff game.

6.2 GLM/Nonparametric Methods

A key finding from our GLM was that variables that we engineered, such as PER and team chemistry, were the most performant features in terms of assisting us in predicting game outcomes. These have the largest coefficients in the GLM.

A key finding from nonparametric methods is that KNN's the best model at predicting home team winning outcomes. Furthermore, how strong a team is historically (which, as previously mentioned, is pseudo-indicated via team ID) is generally a good predictor of whether or not a team would win, regardless of their opponent.

The findings from both the GLM and nonparametric models are generalizable. This is because the NBA dataset we studied represented a census, where the population being studied is all NBA games from 2004 to 2022. So our findings can likely be applied to studying future NBA games, as well, since the factors that are important now to winning a game will likely remain constant. The only changes we'd expect between the dataset we studied and future datasets might be the presence of outliers for certain variables (for example if you have an NBA player that breaks current records), in which case k-NN may do more poorly.

Based on our findings, team chemistry is a particularly important factor in predicting game outcomes. A call to action we'd suggest would be for teams to place a greater focus on team-bonding and to run more drills that encourage collaboration, in light of this finding.

We did not merge data sources; all of our files came from the Kaggle dataset.

There are, however, some limitations to the dataset. For one, it doesn't capture the impact of player trades, transfers, or changes in team dynamics over time. A player's performance may be influenced by team changes, and the dataset might not reflect the nuances of these transitions. Furthermore, some players transition between positions over their careers, impacting their playing style and statistical contributions, yet the dataset only writes the position of the five starting players, hence we could not analyze this factor. Finally, the NBA has evolved its statistical tracking methods over the years. Changes in scoring rules, statistical categories, or recording techniques could thus introduce inconsistencies in the dataset, affecting the comparability of stats across seasons, yet we did not have a way of factoring these into our analysis.

Future studies could try to predict whether a team wins a given game based on additional features that we did not have access to in our dataset, such as the coaches' characteristics, the average player's age, or the presence of the highest-paid players. Furthermore, other studies can examine if there is a causal relationship between certain player characteristics and a high average team chemistry score, which can help inform team composition strategies.

We learnt several things:

1. The importance of scaling the data, especially when there are significant

outliers present. By log-transforming the data, it significantly helped improve the accuracy of our KNN model, which is sensitive to outliers if some features have outsized impact on the classification. After seeing the impact, we learnt how important it is to consider what are the fundamental principles of each modeling technique and think critically about how the data might influence/ not satisfy some assumptions needed.

2. Separating the data out into training and test sets prior to any feature engineering of any kind to prevent data leakage. While this decreased the performance of all of our models, it ensured that the results are more robust and reliable.

3. Setting max features and depth to random forest and decision trees. Without setting these, the accuracy appears to be a lot better, yet is with the risk of overfitting, overly complex model, and in the case of random forest, removes the randomization that random forest provides (since if we didn't have max features set, it would be equivalent to using all our features).

7 References

Yvette. "Beginner's Guide on How to Calculate Player Efficiency Rating." *Watts Basketball*, wattsbasketball.com/blog/how-to-calculate-player-efficiency-rating. Accessed 11 Dec. 2023.