**University of Southern California**

**“Which lineup combinations are most effective at driving team success”: An Analysis of Los Angeles Lakers 2023 Season**

**Eli Tills & Mihir Sabnis**

**DSO 579 – Fundamentals of Sports Performance Analytics**

**Lorena Martin, PhD.**

**15 Dec 2024**

## **Introduction**

*“The strength of the team is each member. The Strength of each member is the team”*. That profound statement was spoken by Phil Jackson, one of the most renowned coaches in basketball history. Coach Jackson’s statement may have numerous interpretations yet it captures the essence of teamwork and synergy in basketball. Still, the sole belief is that behind every great basketball team is the collection of players and how the specific skill set of each player works with one another.

In basketball, success hinges on finding the optimal combinations of players in a lineup to maximize performance during critical game moments. For this project, we turned out attention to the Los Angeles Laker’s lackluster 2023 season, as we are both fans of the Lakers and were interested in analyzing their in-season gameplay strategy. We want to see if we could create a model that could analyze the different combinations of their roster, and identify which lineups made the most significant contribution to victories in 2023. By understanding these contributions, we sought to determine the most effective lineups, strategically combining players whose strengths and statistics complement each other to enhance consistent team success.

Through this analysis, we sought to answer a critical question: Which lineup combinations are most effective at driving team success? By identifying these combinations, we hoped to offer actionable insights into player synergy, highlighting how specific skill sets complement one another to maximize both offensive and defensive performance. This investigation is not just about finding the best players but about finding the best way to utilize them as a cohesive unit.

## **Abstract**

This study focuses on analyzing the impact of various player combinations on game outcomes for the Los Angeles Lakers during the 2023 season. Specifically, we sought to answer the question: How do different 5-man lineup combinations contribute to game outcomes (win/loss)? Additionally, we explored whether smaller group dynamics, such as 3-man combinations, have a measurable effect on game outcomes.

To address these questions, we employed supervised machine learning models, including Logistic Regression, Decision Trees, Random Forest, and XGBoost. These models allowed us to evaluate the relationship between player combinations and game outcomes. By leveraging advanced analytics, we sought to determine the lineup configurations that maximize team success, offering insights that could inform coaching decisions and roster management.

We hypothesized that offensive combinations provide the best offensive stats, defensive combinations provide the best defensive stats, and playmaking combinations provide the best playmaking stats. The null hypothesis we came up with was that there is no relation between the player combinations and win/loss.

Another question we wanted to address during our research was whether or not a 3-man combination would affect the outcome of a game, hypothesizing that certain smaller groupings of players might amplify individual strengths or create unique synergies that influence game outcomes. This line of inquiry stemmed from the observation that, in sports, a small group of players often forms the backbone of key moments, driving offensive or defensive efficiency. Along with testing the 5-man lineups, we decided to run 3-man testing for the same features.

Through our investigation, we aimed to bridge the gap between traditional basketball strategies and data-driven decision-making, ultimately shedding light on how lineup optimization can serve as a critical tool in achieving team success.

## **[Data Collection](https://www.nba.com/stats/lineups/advanced?slug=advanced&Season=2023-24&OpponentTeamID=0&GroupQuantity=5&Period=0&TeamID=1610612747&LastNGames=0&DateFrom=1%2F17%2F2024&DateTo=1%2F17%2F2024)**

We initiated our data collection process by searching for NBA datasets relevant to building our model. The NBA.com statistics website has the most comprehensive collection of NBA data and metrics. During our research, we discovered the possibility of setting up the NBA an application programming interface (API) to access a wide range of NBA data. We found an open-source library designed to interact with the NBA stats API and attempted to use it. However, we encountered challenges with API requests, particularly when querying the “Lineups” endpoint. Due to frequent request rejections caused by endpoint rate limiting, we ultimately decided to scrape the necessary data directly from the website.

We first scraped the website for the Lakers 24-24 regular season game data. This comprised of the opposing team, win/loss, and the date of each game. We then scrapped the advanced lineup data for all the 5-man and 3-man combinations that were ever played in any game in the 23-24 regular season. The metrics collected for each lineup seem to be the sum of the players in that lineup for that game during the time they were in the lineup. This gives us a much more granular view of their individual player performance as 1 player could play in 5 different lineups and accrue different stats for each lineup which is a better outlook of theri performance compared to looking at their overall game stats. It allows us to understand how well they play with different teammates.

## **Solutions Approach**

Once we had the data we had to decide as to how we wanted to approach dummying these lineup combinations for machine learning. We could either dummy each unique lineup as its dummy variable but we had to select the most frequent lineups as there were a total of 414 unique 5-man lineups and 560+ unique 3-man combinations. For 5-man we choose any lineup that appeared in more than 4 appearances and for 3-man we choose any lineup that appeared in more than 24 appearances. This reduced the number of lineups to 64 for 5-man and 70 for 3-man lineups.

The other method was to make dummy variables for all 21 players and give each player for each row a 1 or a 0 if they appeared in the lineup or not. This means each row had a guaranteed 5/3 players that would have a 1. This method allows us to asses every single lineup without having to remove some of them due to their lack of appearance. This means that we have a larger dataset with more possible patterns that are not eliminated.

We decided to try both methods and use the methodology that produces the better-performing models in evaluation.

## **Data Manipulation**

The data we wanted to collect and build our model off of focusing on the 2023-2024 Season of the Los Angeles Lakers. The dataset itself came with numerous excess data columns, which we had to drop since the aggregates of other columns would have led to collinearity.

The dataset, which included player combinations and lineup statistics, was first loaded and explored to understand its structure and identify key features like offensive rating (OFFRTG), defensive rating (DEFRTG), net rating (NETRTG), assist percentages (AST%), rebounding metrics (OREB%, DREB%, REB%), and the target variable, WIN. Binary indicators were created for each player to represent their participation in a lineup, enabling an analysis of individual and collective player contributions. We then made the dummy variable transformations as needed.

After separating the columns into three categories, we normalized and transformed the data. The final dataset had over 20,349 possible combinations (all stemming from 21 players), 414 unique combinations, and 64 5-man combinations that had played more than 4 times.

After ensuring no missing values or inconsistencies in the data, feature scaling was applied to continuous variables to standardize their ranges, facilitating compatibility with machine learning algorithms sensitive to scale. To reduce redundancy and multicollinearity, careful checks were performed to exclude highly correlated or low-variance features.

The resulting dataset was a balanced mix of statistical features and binary player indicators, enabling both team and individual analysis. This structured and preprocessed dataset provided a robust foundation for training machine learning models to evaluate the impact of player combinations on game outcomes. This high-level manipulation process ensured that the data was well-prepared for insightful and actionable analysis.

After this was implemented, we separated the data into three categories. The three categories consisted of offensive, defensive, and playmaking stats. The categories were broken down as follows:

***Offensive Stats***

The offensive stats consisted of an offensive rating, offensive rebound percentage, efficiency goal percentage, true shooting percentage, and the number of possessions played per 48 minutes per game.

***Defensive* *Stats***

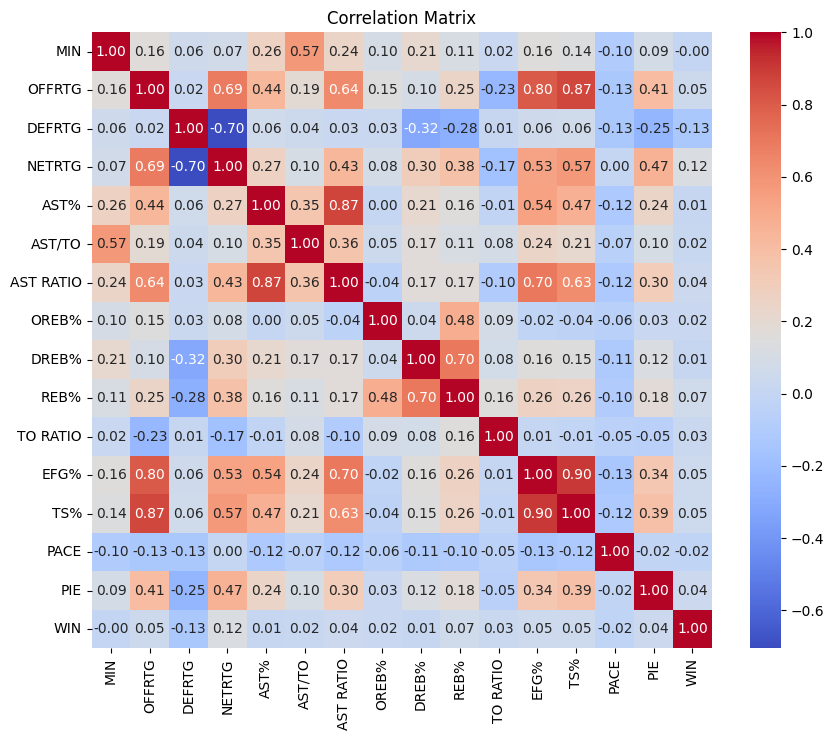
The defensive stats consisted of Defensive rating, Defensive rebound percentage, total rebound percentage, Net rating percentage, which is the difference between offensive and defensive ratings, and total minutes played by the lineup in the game.

***Playmaking stats***

The playmaking stats consisted of assists, assist-to-turnover ratio, assist ratio (which is the number of assists per 100 possessions for the lineup), Turnover ratio, and total minutes per player per game.

The dataset was then split into training and test sets using an 80-20 stratified split, ensuring the distribution of the target variable (WIN) remained consistent across both sets.

## **Model Evaluation and Analysis**

We first began our analysis by evaluating the different variables and their correlation to a win. Utilizing a correlation matrix, we determined that all predictors had a very low correlation with wins. This, of course, looks at all of the predictors without stratification of the different five-man combinations. Now that we know that none of the variables are significant for predicting a win, we wanted to bring the five-man combination data to help the model understand which rows of data contributed by one particular five-man combination.

From here, we decided that we wanted to investigate the patterns of change that were stratified by the 5-man combination. To achieve this, we applied and evaluated models, including Logistic Regression, Decision Trees, Random Forest, and XGBoost, leveraging GridSearch and Cross-Validation for optimal performance tuning. For each of the models selected we ran through 4 different player combinations: Offensive, Defensive, Playmaking, and All Stats.

When comparing the performance of models between our two dummy methods, dummying the unique lineups and selecting the most frequent lineups produced models with a higher f1 score. What we realized is that by dummying the players we will be able to identify key players but it's harder to attribute it to key lineups.

### **Logistic Regression**

The Logistic regression model provides a baseline by predicting the probability of winning based on linear combinations of features. It is useful for understanding the influence of each statistical category on the win/loss outcome through its coefficients. This will be useful in interpreting how specific combinations of stats affect game success, offering clear decision paths.

### **Decision Tree Classifier**

The Decision Tree Classifier will help identify important features and capture complex, non-linear relationships between player combinations and game outcomes. It also visualizes decision-making paths, making the model more interpretable.

### **Random Forest Classifier**

The Random Forest Classifier builds multiple decision trees and averages their predictions. This ensemble model improves accuracy, reduces overfitting, and handles large datasets efficiently. This will fit our dataset, due to the high dimensionality and different player combinations in regards to the different lineups.

### **XGBoost Classifier**

The XGBoost Classifier is known for its high predictive power and efficiency. It employs gradient boosting and optimizing performance. Given the complexity of the dataset we were using and its potential interactions, this model aligns model would align well with our multifaceted model.

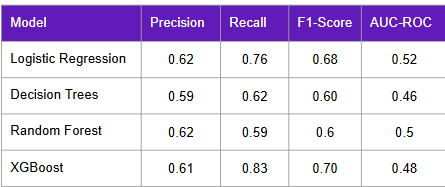
### **Chosen Metrics for Evaluation**

To evaluate the performance of our classification models, we utilized precision, recall, F1-score, and ROC-AUC as key metrics. Precision ensures the model identifies winning lineups without falsely labeling ineffective ones, which is crucial for actionable insights. Recall measures the model's ability to correctly identify all successful lineups, minimizing the risk of overlooking effective combinations. The F1 score balances precision and recall, offering a holistic view of performance, especially when both false positives and false negatives matter. Finally, ROC-AUC evaluates the model's ability to distinguish between wins and losses across thresholds, ensuring robust and reliable predictions for lineup effectiveness.

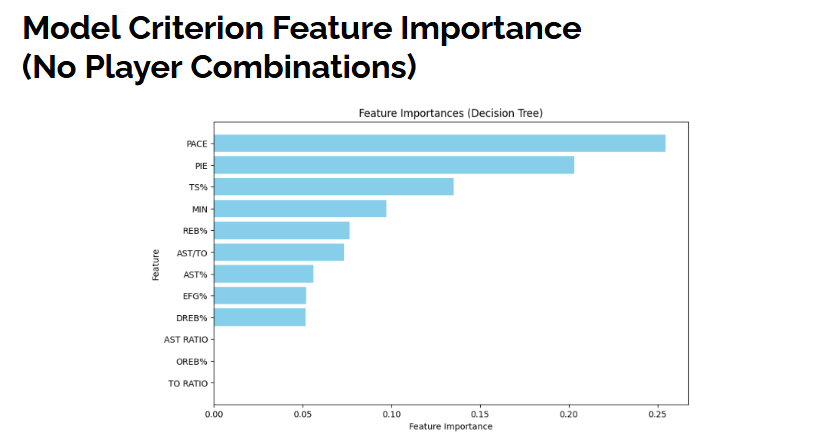
### **Model Performance**

#### **5-Man Lineups**

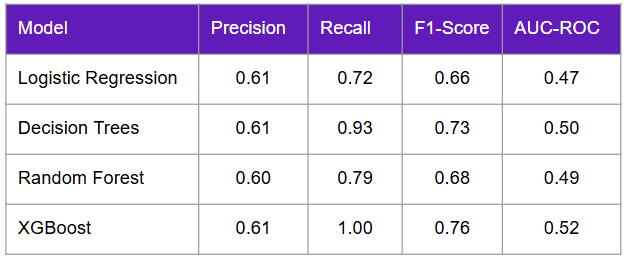
Starting with All Stats Combinations, we found no significant results regarding the four models. The All Stats Combination we are looking at the dataset that had no dummy variables, so we are only looking at the undergirded statistics across all the games in the 2023 season.



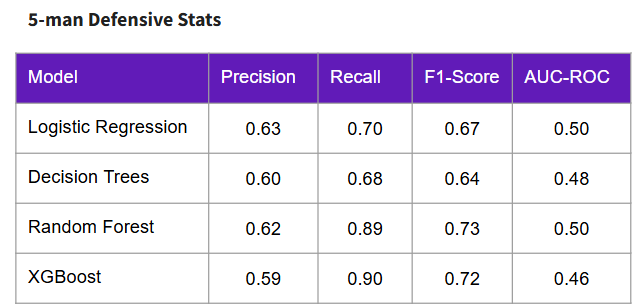
From the results, you can see the decision trees are the best model, however, it is still not a useful model. The AUC and RSA accuracy is only 0.5, which is a result that is very close to a random guess. Based on the features of importance for the No Player Combinations data, we identified as Pace being the most contributing variable.



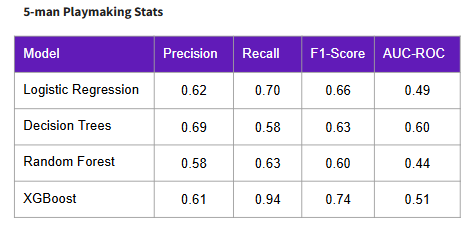
For the 5-man offensive stats, we found to have similar results, with the AUC-ROC being around 0.5. Once again this shows that the result of all the outcomes is similar to that of a random guess.



The Defensive 5-man once again had similar results to that as the offensive and all stats model results.

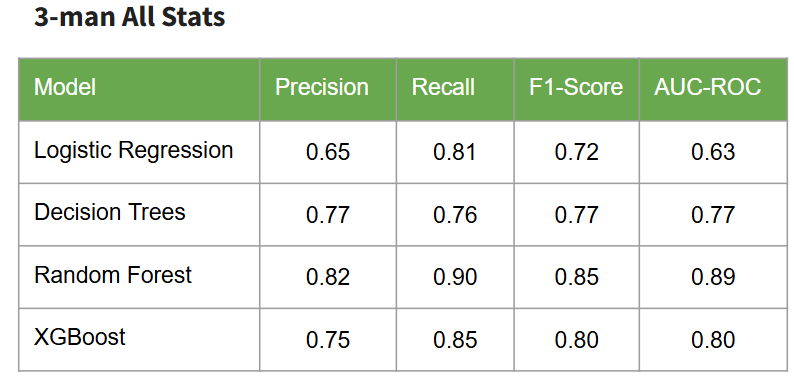


The 5-man Playmaking stats were able to deviate from the pack, having a 0.6 score for its Decision Tree model. Even though this broke away from the average of the other models, since AUC-ROC score was still low.



#### **3-Man Lineups**

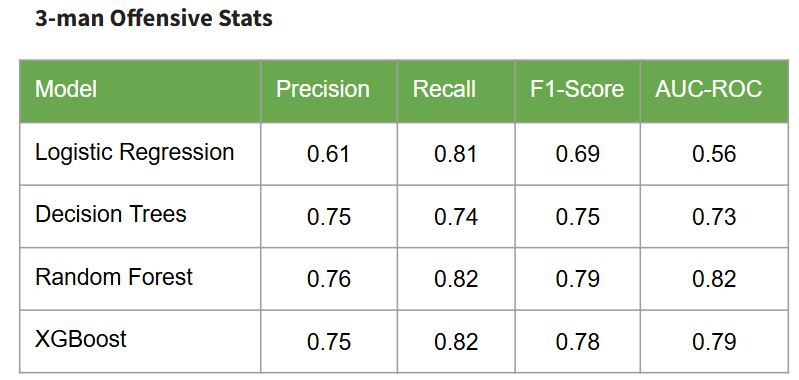
So after the undecisive results of the 5-man lineups, we decided to look into what the results would be if we focused on the 3-man lineups. Once again we ran through the same categories of Offensive, Defensive, Playmaking, and All Stats.



Starting with the 3-man All stats, we see that the Random Forest is shown to have the best fit with their AUC-ROC score a 0.89. With an F1 score of 0.85 and a Precision score of 0.82, this model was significantly more accurate for the 3-man combination than the 5-man combinations.



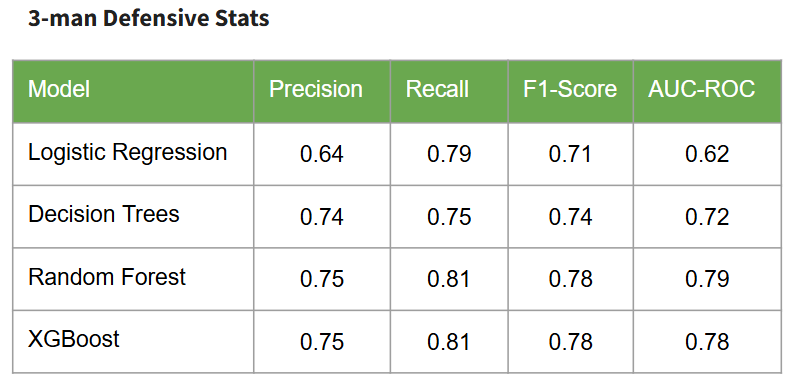
Looking at the feature importance for all stats, it clear that the player that contributed the most across all the stats that were evaluated was Taurean Prince and Austin Reaves.



The 3-man offensive stats also favored the Random Forest model, with an AUC-ROC score of about 0.82. With a F1 score of 0.79 and a Precision Score of 0.76, the Random Forest Model is significant.

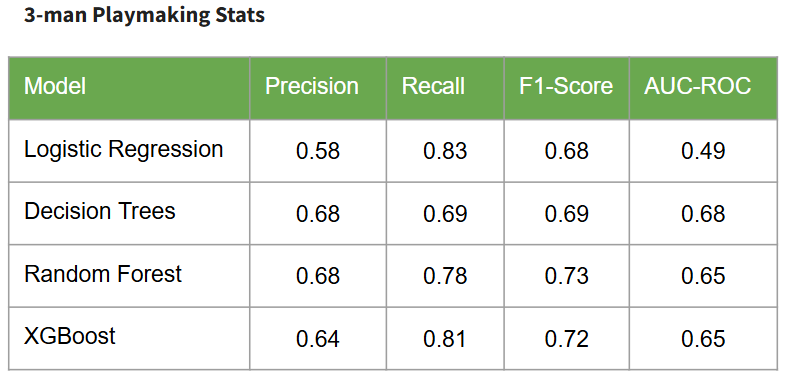


Looking at the feature importance of the offensive stats, we can identify that there is a range of dynamic combinations that can contribute on offense. The top 3 combinations here shows the which players who are on the same line up can contribute to playing better offensively.



The 3-man defensive stats also favored the Random Forest model, with an AUC-ROC score of 0.79.





The 3-man playmaking stats favored the decision trees model with a 0.68 model. From the model performance, we can determine that the model is slightly better than a random guess, but we would not apply this model to make a strategic position. This could be because the playmaking stats collected are not causality-related with win probability and therefore we need to collect alternate playmaking metrics that have a stronger correlation with win probability.



### **Model Implications: 5-man vs 3-man Lineups**

Based on the overall results of both of our models for both the 5-man and 3-man lineups, it is clear the 3-man lineups are better to evaluate then the 5-man lineups. With the 5-man lineups most of the AUC-ROC demonstrate that a random guess would be more likely to be accurate than to run data through the model. 5-man lineup is also likely to be highly skewed, with the different combinations being a trio and a duo performing differently. The 3-man lineup has a better showing of how that player performed within the team.

## **Conclusions**

### **Team Recommendations and Model Deployment**

For this Model to be implemented by an NBA team’s operations, the model must have data integration through a model training pipeline. The pipeline is a cloud-based training pipeline where the model is retrained periodically using updated season data to maintain accuracy and relevance. The key to this pipeline is to have it updated on a real-time basis, allowing for coaches and analysts to see the change in results as they occur. This could lead for both pre-game and in-game adjustments to the team's lineup, as they evaluate performance and create a better lineup efficiency. This model could also be used to evaluate the opposing team, to see the areas where a team is at their strongest, and areas where they are at their weakest.

### **Data and Model Shortcomings**

When looking at the data itself. Certain players like Austin Reeves and Anthony Davis have the highest participation rate in these lineups. Since they have more data points charted to their names, it is a lot harder to identify players who get less time on the court or appear in fewer lineups. Their impact is harder to isolate and would require another method of evaluation such as causal inference to better understand how much they contribute to the lineup’s success as an individual player on the court. Plus/Mius might give a good idea of this but doesn't take into consideration the time and duration the lineup was on the court.

There is a lot of scope with the research we have done through this project. We were able to prove that 5-man combinations are less useful than 3-man combinations in predicting a win but we must keep in mind that this conclusion was drawn from models only trained on Lakers 23-24 season. There is still room to explore how 5-man and 3-man lineups are in predicting wins for other teams in other years. Since we only focused on the Lakers, our conclusion is directly tied to the lakers performance. If their performance is inconsistent, our conclusions might not be well-founded. We might see alternate relationships for other teams that perform consistently better or worse.

Our data collection was from the only source that had collected 5-man, 4-man, 3-man, and 2-man lineup data. The metrics we used are the only ones available. We have a scarcity of relevant metrics available. We tried to scrape individual player performance data that comprised many more metrics for every game using the NBA API. We then tried to make new metrics available for lineups by grouping and summing each player's respective game metrics by their respective lineups for each game but this was tedious and time-consuming. Since this was a short project we used what was readily available. If we can pursue this method of pooling metrics using individual player data, we would have more features available to test and might give us more accurate predictions with different combinations of features.

The nature of the way the NBA is played only allows us to have 82 games of data with one team before end-season transfers are made. While we can assess lineup strategy performance within a season, its harder to assess how these lineups perform over time and might evolve. This is especially important when we consider that star players are needle movers in any lineup they are a part of, which makes it harder to assess rookies or players in the early stages of their careers.

If we had more time and processing resources, we would work with more complex models that require a greater resource load. We would require more time to explore each avenue of hyperparameter testing and methods to optimize the data in preprocessing.

### **Future Improvements**

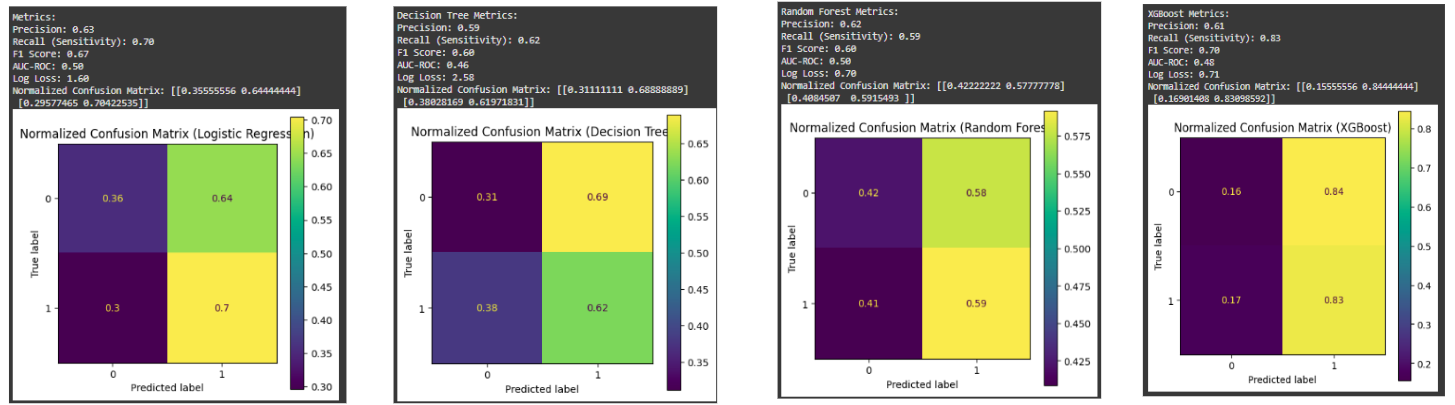
One of the things we are looking to evaluate is the two-man combinations, if they will be more efficient at predicting outcomes for games a win. With moving from the 5-man lineups to the 3-man lineups, we noticed that by making smaller combinations, we started to get a better AUC-ROC score and F1 Score. This would be a worthwhile experiment for a model and may help teams identify the strategy of having players find the best counterpart on their team to try and bring about positive outcomes. This would further the idea of helping teams develop their own dynamic duo.

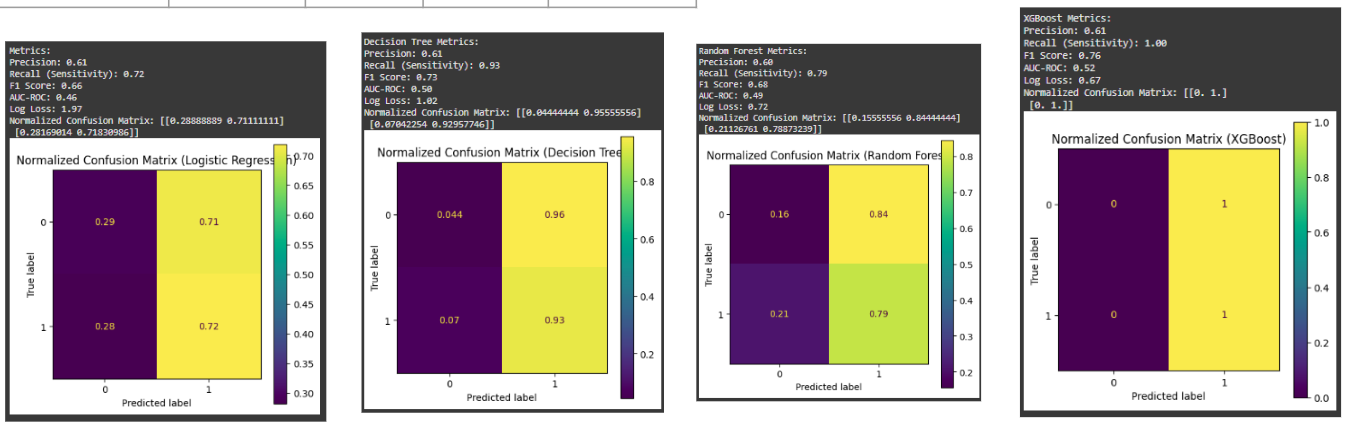
We would like to run the same research with the Lakers over a period of 5 years and also assess lineups for every other team in the NBA for the 23-24 season. We can also assess lineups for playoffs, postseason and the NBA cup, these can also be used as validation sets.

We also would like to evaluate each lineup by the quarter they are played in. That we are able to understand which quarter these lineups could possibly make an impact. That way they could strategically plan what lineup performs best in which quarter.

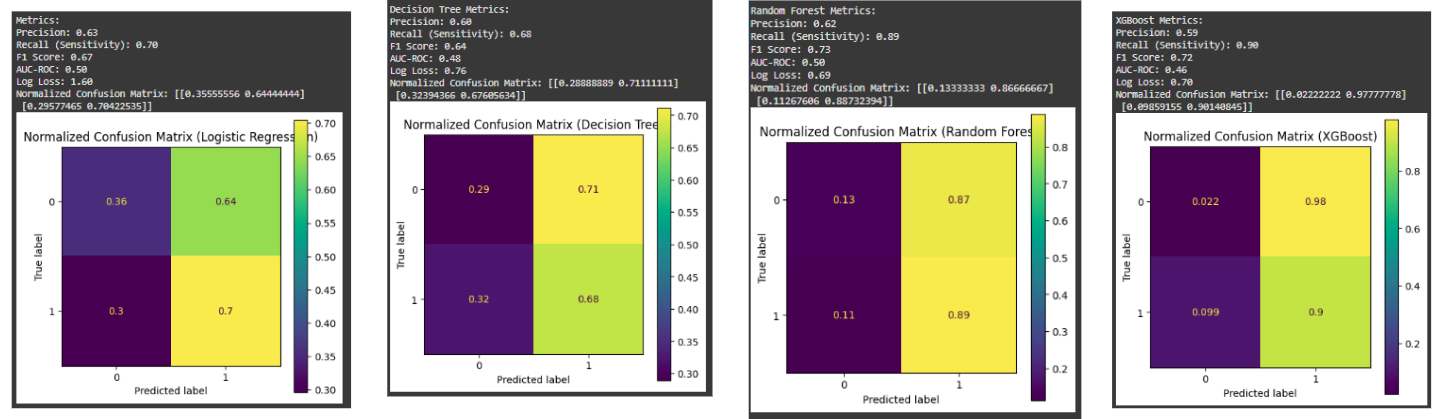
A large contributor to uncertainty in this project was the aspect of not having a comparative model. Since we only focused on the Lakers performance, this model was based on the idea that a win or loss was determined solely by the team itself, and not based on the performance of their opponent.

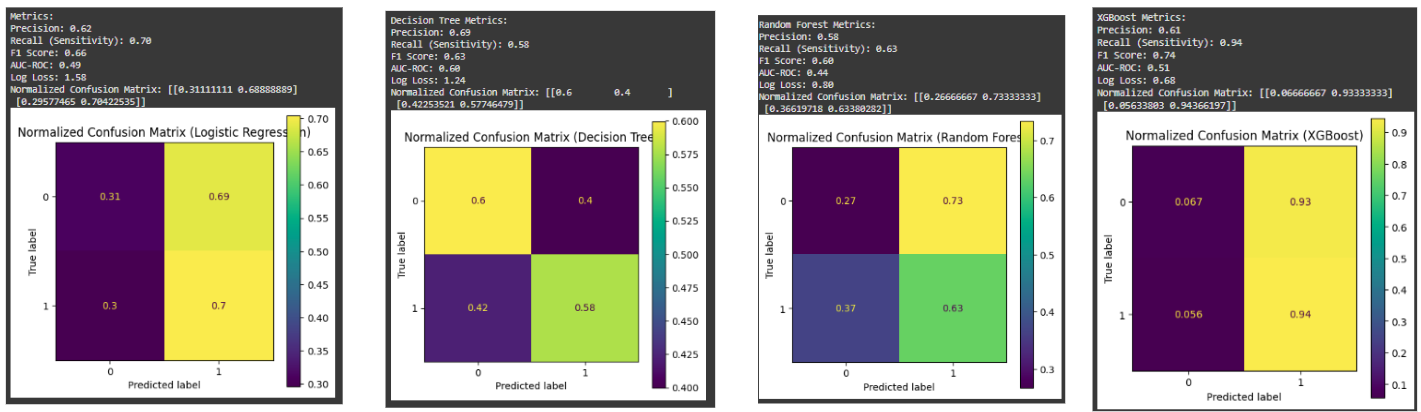
**Appendix**

5-Man All Stats Models Confusion Matrix

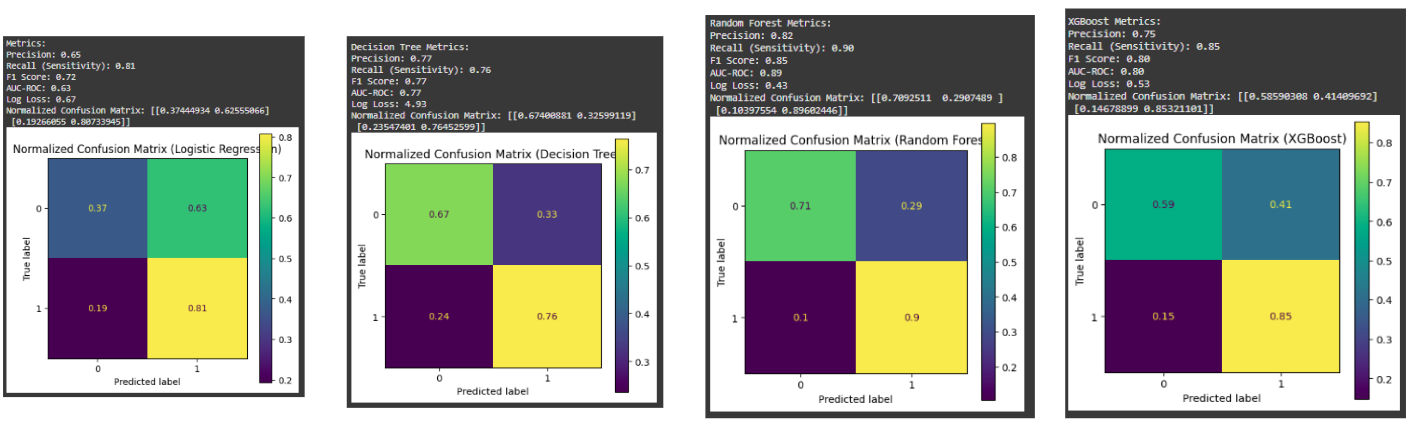
5-Man Offensive Models Confusion Matrix

5-Man Defensive Models Confusion Matrix

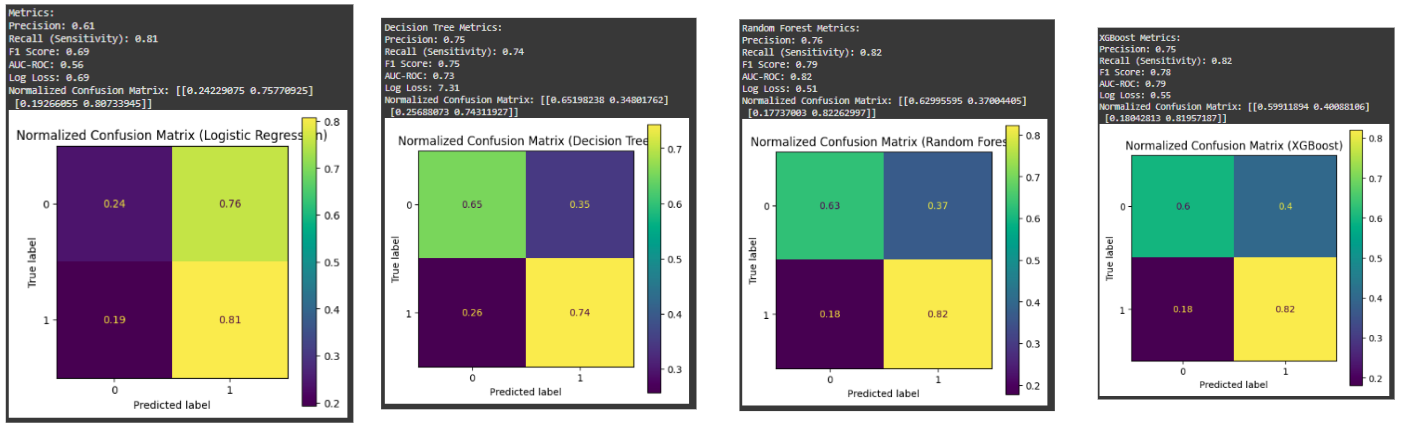


5-Man Playmaking Models Confusion Matrix

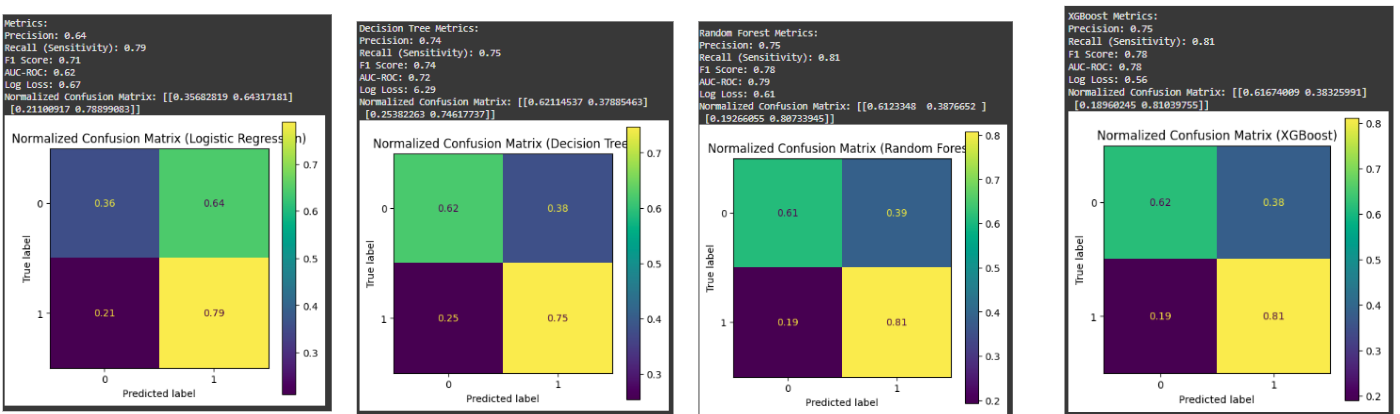
3-Man All Stats Models Confusion Matrix



3-Man Offensive Models Confusion Matrix



3-Man Defensive Models Confusion Matrix



3-Man Playmaking Models Confusion Matrix

