Final Report

DevonGrace Tax, Maggie Truong, Michael Helm, John Ramirez, Smyan Jaipuriyar

Introduction

For our project, we focused on two main research questions to explore what factors contribute to team success in NCAA Division I basketball. The **first question** looks at the relationship between a team's defensive efficiency (`ADJDE`) and how far they progress in the NCAA tournament during the 2021 to 2024 seasons. Adjusted Defensive Efficiency (`ADJDE`) measures how many points a team allows per 100 possessions, which accounts for differences in the pace of play. The **second question** investigates how a team's three-point shooting accuracy (Offp_3p) and free throw rate (`FTR`) influence their win percentage (`W/G`) and postseason success.

- **Null Hypothesis 1:** Teams with better defensive efficiencies do not perform better in the tournament or make it further than teams with weaker defensive metrics.
- **Alternative Hypothesis 1:** We hypothesize that teams with better defensive efficiencies will perform better in the tournament and make it further than teams with weaker defensive metrics.
- **Null Hypothesis 2:** Offensive metrics such as three-point shooting accuracy and free throw rate are not positively associated with a team's ability to win games or perform well in the postseason.
- **Alternative Hypothesis 2:** We hypothesize that these offensive metrics are positively associated with a team's ability to win games and perform well in the postseason.

The dataset we are using includes NCAA Division 1 College Basketball statistics from 2013 to 2023, although there is no postseason data for the 2020 season due to COVID-19. Originally, the dataset had 368 rows and 24 columns, covering a wide range of stats from over 10 years. For this project, we shortened the scope to focus on the seasons from 2021 to 2024 and cleaned the data to make it more relevant to our research questions. This allows us to concentrate on recent trends and ensures the analysis stays manageable.

This project is important because it helps us better understand what factors really matter for success in NCAA basketball. These insights could be useful for coaches, players, and analysts looking to improve team strategies, as well as for fans who want a deeper understanding of how the game works. Additionally, the project ties into the growing field of sports analytics and shows how data can be used to evaluate and predict team performance.

Exploratory Data Analysis (EDA)

During this phase of our project we attempted to take a deeper look into the data itself and draw conclusions and insights from it.

1. We dealt with NA values in the dataset

- These NA values that were encountered made sense that they were left NA as they dealt with post season seeds, and round exits
  - For the `POSTSEASON` variable NA value we changed it to "Did Not Make PostSeason"
  - For the `SEED` variable we changed the NA value to "No Postseason"
- Aside from that mentioned above, the dataset was well-structured, and there were no issues with the other variables.

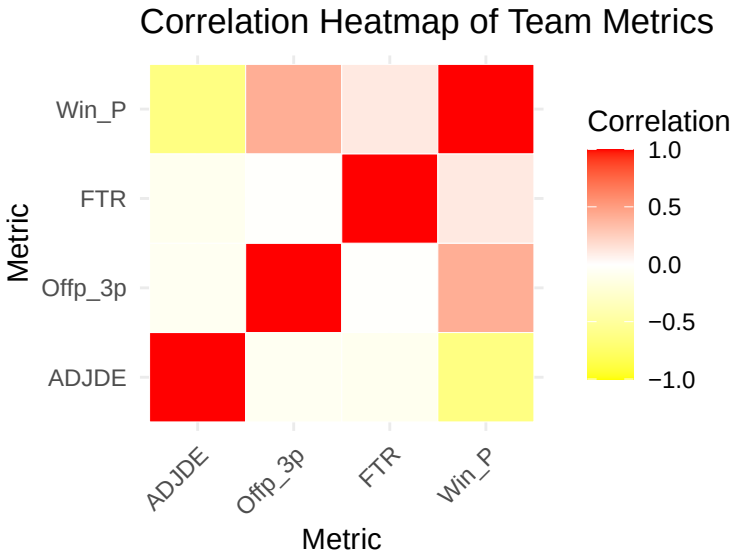
2. Create Descriptive Statistics

- We took the numerical variables to see how the summary statistics looked, it was able to give us insights on the range and other metrics like mean for the variables to better understand the data we dealt with (Not included as it is quite long). This helped guide us in our data visualization.

3. Data Visualization

From the heat map below that shows correlations between our variables of interest for the hypothesis we created, we are able to see that:

1. A strong positive correlation between Adjusted Defensive Efficiency (ADJDE) and Win Percentage (Win_P), suggesting that better defensive performance is closely tied to team success.
2. Moderate positive correlation between Free Throw Rate (FTR) and Win Percentage (Win_P). Teams that get to the free-throw line more often tend to perform better overall.
3. Offensive 3-Point Shooting Percentage (Offp_3p) and Adjusted Defensive Efficiency (ADJDE) is weak, implying that a team's offensive 3-point shooting doesn't significantly impact their defensive efficiency.



Modeling Process

To investigate our two research questions, we implemented distinct models each carefully selected to best fit the characteristics of each hypothesis and the dataset.

For **Hypothesis 1**, a **classification model** was the most appropriate choice. Postseason success is a categorical variable, represented by discrete stages such as "R64," "E8," or "Champions," rather than continuous numeric values. Therefore, we implemented a multinomial logistic regression model to classify teams into postseason categories based on their defensive metrics. This approach allowed us to predict the likelihood of a team reaching various stages in the NCAA tournament, providing insights into how defensive efficiency impacts postseason progression.

For **Hypothesis 2**, we chose to utilize a **regression model**. These offensive metrics (three-point shooting accuracy (`Offp_3p`) and free throw rate (`FTR`)) and win percentage (`WinPerc`) are continuous variables, making linear regression an ideal method for identifying and quantifying the strength of the relationship between shooting efficiency and win rates. By combining the effects of `Offp_3p` and `FTR`, we aimed to determine how these key offensive factors contribute to overall team success throughout the season. We also implemented a logistic regression model for predicting postseason outcomes (categorical variable like "Champions," "Elite Eight") based on the same predictors (Offp_3p and FTR).

Hypothesis 1: Classification Model for Predicting Postseason Success

Results

- **Model Description and Evaluation**

We applied a multinomial logistic regression model to predict the postseason success of NCAA basketball teams based on defensive efficiency (`ADJDE`). The dataset was filtered to include only teams that made the postseason, with `POSTSEASON` treated as a categorical variable. After splitting the data into a 70% training set and 30% test set, we fit the model using the `nnet` package's `multinom()` function. The model converged after 70 iterations.

- **Accuracy and Confusion Matrix**

The model achieved an accuracy of **51.94%**, indicating it correctly classified the postseason stage approximately half the time. While this suggests moderate predictive power, further improvement is possible by incorporating additional predictors. Below is a summary of the confusion matrix (Columns are from Orginal Categories, Rows are the predicted):

	2ND	Champions	E8	F4	R32	R64	R68	S16
2ND	0	0	0	0	0	0	0	0
Champions	0	0	0	0	0	0	0	0
E8	0	0	0	0	0	0	0	0
F4	0	0	0	0	0	0	0	0
R32	2	3	6	1	11	8	2	10
R64	1	0	7	2	35	96	10	10
R68	0	0	0	0	0	0	0	0
S16	1	0	0	1	0	0	0	0

The matrix reveals that the model struggles with less frequent categories as seen with "2ND," "Champions," "Elite Eight" (E8), and others, where the model fails to predict any instances accurately. However, it performs reasonably well for more common categories like "R64" and "R32," reflecting class imbalance issues where these categories have more observations in comparison to the less frequent stages like those listed above.

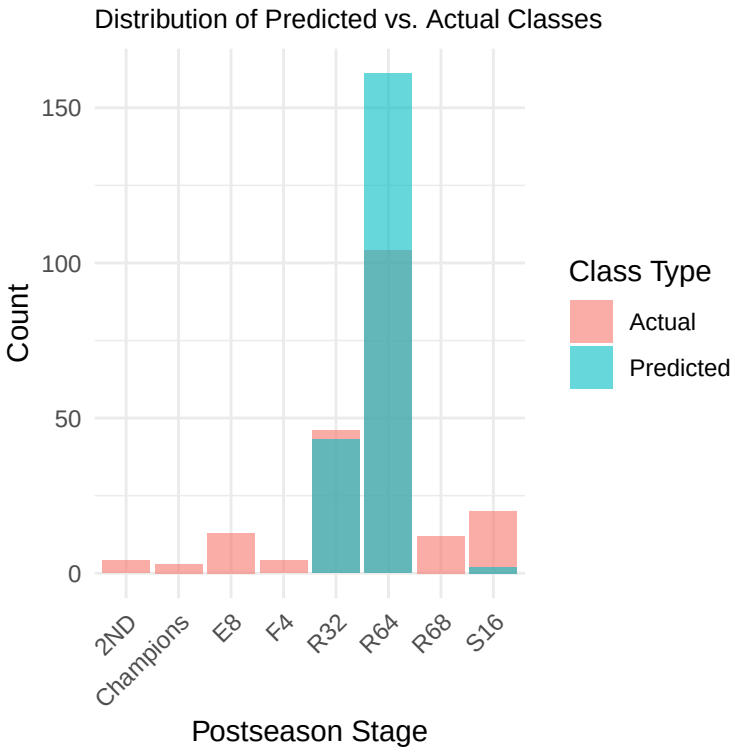
Interpretation

The model's moderate accuracy shows how it is necessary to have additional predictors to enhance its predictive capability. It performs better at predicting early postseason stages due to the larger sample sizes available (these stages have more observations in the dataset), while later stages suffer from insufficient data. The confusion matrix shows a bias toward more frequent classes, with "R64" being correctly predicted 96 times but often confused with "R32" (on 35 accounts) and "S16" (on 10 accounts)

- Key Takeaways:

1. The accuracy can be improved by including more predictors such as offensive efficiency or team experience.
2. Increasing the dataset size, especially for less frequent classes like "Champions," could help the model better generalize to all stages.
3. Applying class balancing techniques or weighted loss functions could mitigate the impact of class imbalance.

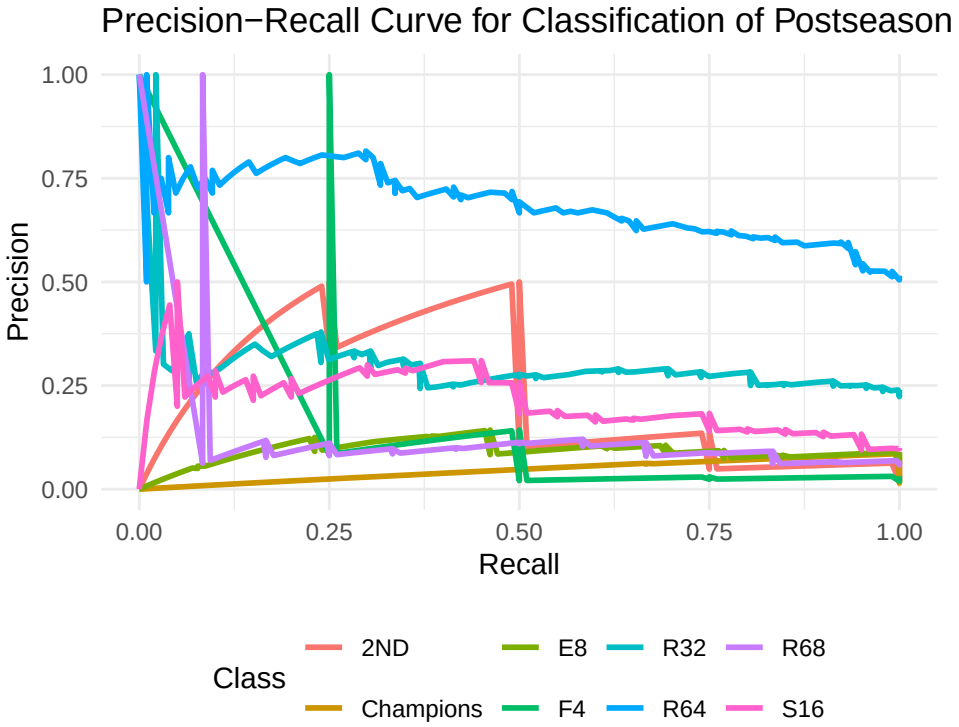
Visualization and Communication



This bar graph compares the predicted and actual postseason stages, highlighting discrepancies between them. The blue bar represents the amount or predictions for each category where the pink is the actual number per category.

- Commentary & Relevance

The bar graph shows that the predictions for "R64" exceeded actual outcomes, while predictions for "R32" closely aligned with the actual outcomes. As for the other categories, predictions for later stages like "S16" and "E8" were minimal or non-existent. We can see that amount of prediction decreases over these other categories (if any at all). This can be assumed to be because of the bias toward most frequent classes– as mentioned earlier– since there are more observations to train/test on with the early round exits (such as `R64` and `R32`), where as later round exits (such as `S16` and under) struggle and have weak predicting power. This underscores the need for more data on these stages.



Above, the visualization depicts Precision-recall (PR) curves for each class show varying model performance across postseason stages.

- Commentary & Relevance

The PR curves highlight the model's strengths and weaknesses across classes. The model performs best for "Champions," maintaining high precision at low recall levels. As for the early stages including "R68," "R64," and "S16," these classes show inconsistent precision, which indicates difficulty in predicting earlier rounds. Also, it can be seen tht there are a series of sharp spikes throughout the PR curves for "Elite 8" and "R68," suggesting potential overfitting. Objectively, we can say that the model can improve in areas like feature selection, class balancing, and data expansion to enhance its predictive accuracy for later stages. But overall, this model proves to promisingly predict early-stage postseason outcomes.

Hypothesis Conclusion We fail to reject null hypothesis ($p = 0.3638$), thus defense efficiency is not a significant factor for postseason advancement

Hypothesis 2: Regression Models for Predicting Team Success (Win Percentage and Postseason Success)

(1) Linear Regression Model: Predicting Win Percentage

- Results 1

- **Model Descriptions and Evaluations**

  We used a linear regression model to predict Win Percentage (WinPerc) based on 3-Point Efficiency (Offp_3p) and Free Throw Rate (FTR).

- **Summary of Win Percentage Model Key Takeaways:**

  – The model's summary (given below) shows insights into the significance of the predictors. The summary explains how the model produced significant coefficients for Offp_3p and FTR, indicating their influence on win percentage.

```
##
## Call:
## lm(formula = WinPerc ~ Offp_3p + FTR, data = cbb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53985 -0.11517  0.00313  0.11759  0.53848
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5796124  0.0389089 -14.897  < 2e-16 ***
## Offp_3p      0.0280415  0.0010096  27.775  < 2e-16 ***
## FTR          0.0039894  0.0004997   7.984  1.9e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1635 on 3520 degrees of freedom
## Multiple R-squared:  0.1906, Adjusted R-squared:  0.1901
## F-statistic: 414.4 on 2 and 3520 DF,  p-value: < 2.2e-16
```

- Visualizations 1 & 2 (for Linear Regression Model)

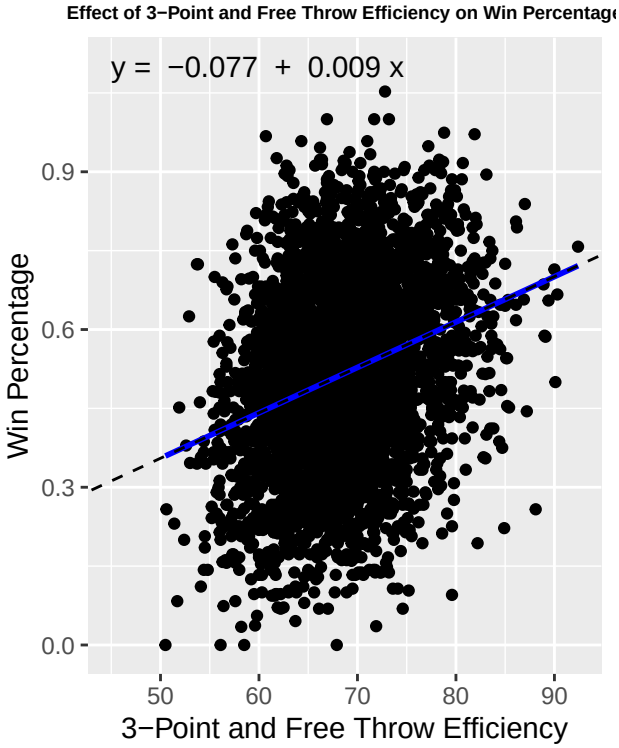
- *Interpretation for Visualization 1 (left visual pm page 7)*

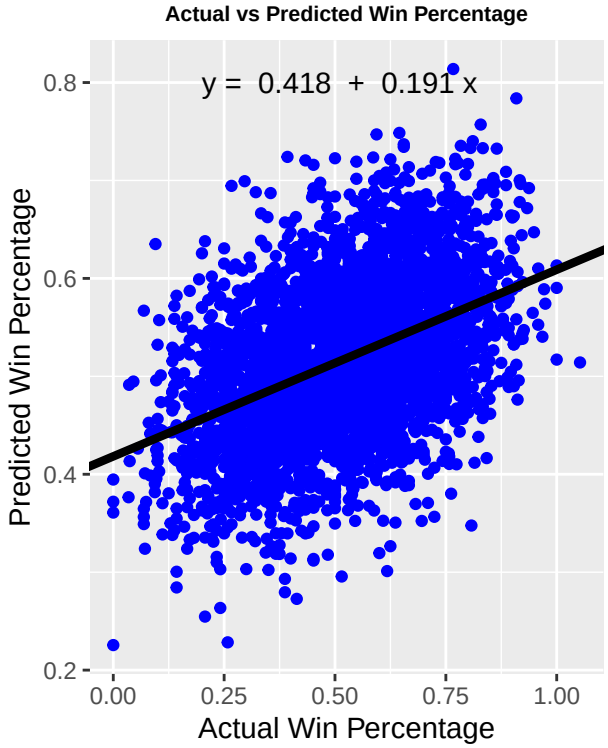
This visualization of the Effect of 3-Point and Free Throw Efficiency on Win Percentage (Visualization 1) shows the positive correlation between combined efficiency (3-point percentage + free throw rate) and win percentage. Teams with higher shooting efficiency tend to achieve higher win percentages.

- *Interpretation for Visualization 2 (right visual on page 7)*

The scatter plot (Visualization 2) compares the actual win percentages with the predicted ones for the Actual vs Predicted Win Percentage visualization. The RMSE indicates how well the model fits the data, and the equation shows the linear relationship. The smaller the RMSE, the better the fit.

```
## [1] "RMSE:  0.163433913465462"
```





(2) Logistic Regression Model: Predicting Postseason Success

- Results 2

- **Model Descriptions and Evaluations**

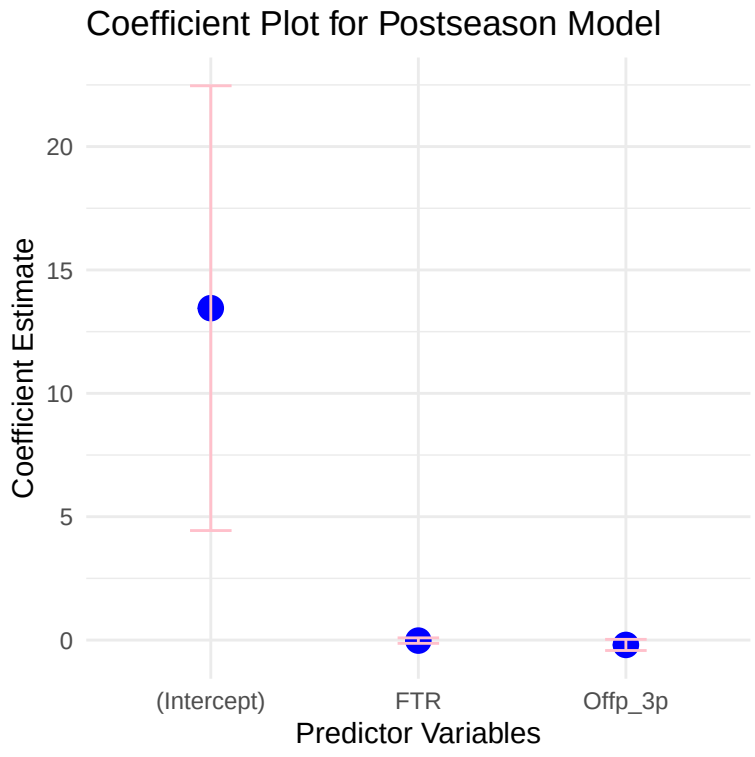
We used a logistic regression model to predict Postseason Success (POSTSEASON) using Offp_3p and FTR.

- **Summary of Post Season Success Model Key Takeaways:**
  - **Offp_3p in the Summary:** A significant coefficient suggests that higher three-point shooting efficiency reduces the likelihood of a lower postseason outcome.
  - **FTR in the Summary:** This variable has less consistent significance, showing variability in its impact on postseason success.

```
##
## Call:
## glm(formula = POSTSEASON ~ Offp_3p + FTR, family = "binomial",
##     data = cbb)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.45167    4.59824   2.925  0.00344 **
```

```
## Offp_3p     -0.19619    0.11533  -1.701  0.08891 .
## FTR         -0.02125    0.05686  -0.374  0.70858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 137.26  on 3522  degrees of freedom
## Residual deviance: 134.25  on 3520  degrees of freedom
## AIC: 140.25
##
## Number of Fisher Scoring iterations: 9
```

- **Coefficient Plot Key Takeaways:**
  - For the coefficient plot (shown below on page 9), the intercept has a large coefficient with a wide confidence interval, indicating high uncertainty in the baseline odds of postseason success.
  - Offp_3p has a negative estimate, suggesting a statistically significant negative relationship with postseason success.
  - FTR's confidence interval crosses zero, indicating it may have a weaker or non-significant impact on postseason success.



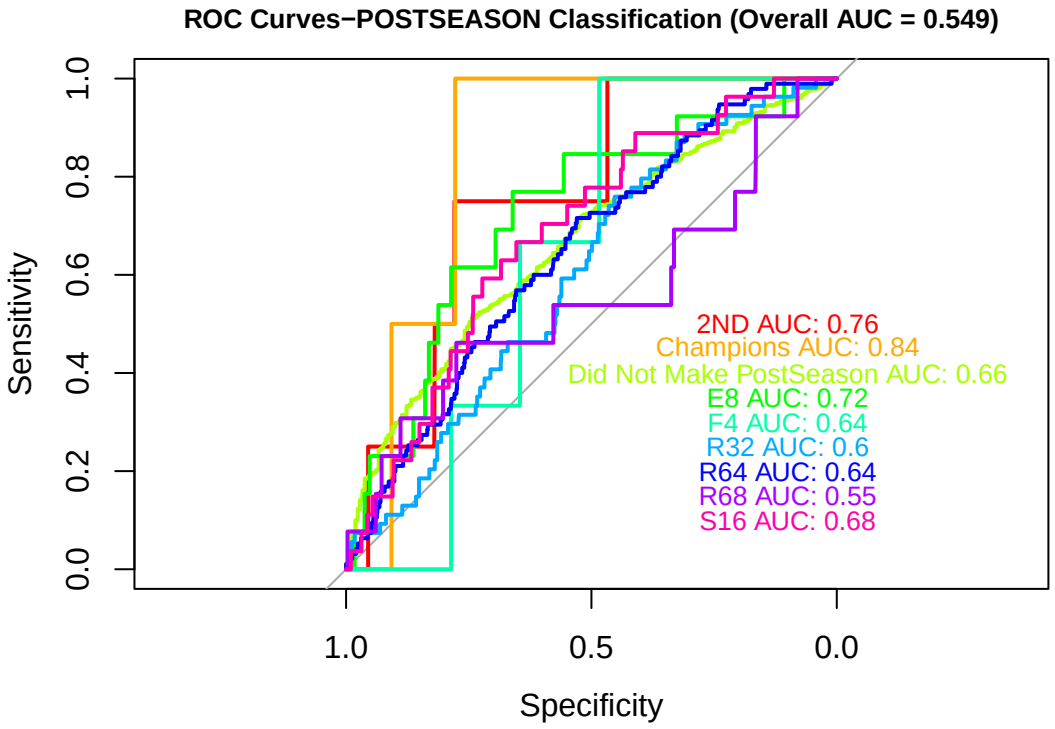
- Visualizations 3 & 4 (for Logistic Regression Model)

Visualization 3: Precision-Recall Curve for Postseason Outcomes

```
## # weights:  36 (24 variable)
## initial  value 5433.736380
## iter  10 value 3158.032005
## iter  20 value 2284.668945
## iter  30 value 1959.839298
## iter  40 value 1779.893149
## iter  50 value 1778.026494
## iter  60 value 1778.023893
## iter  70 value 1778.001875
## iter  80 value 1777.986936
## iter  90 value 1777.971948
## final  value 1777.971620
## converged
```

```
## [1] "Multiclass AUC:  0.549282046510828"
```

Visualization 4: ROC Curves for Postseason Classification



- Interpretation for Visualizations 3 & 4

The AUC values that are higher means that our model did well at identifying teams that had that post season ranking. The highest AUC was for Champions and was 0.84 which means that our model was fairly good at identifying its overall discriminatory power for each classification category. We see that the model

demonstrates it best performs at differentiating "Champions" (AUC: 0.84) and reasonably well for "2ND" (AUC: 0.76) and "Elite Eight" (AUC: 0.72). Nevertheless, it has trouble distinguishing these classes from others in previous rounds such as "R68" (AUC: 0.549), which may be because of less clear patterns or unbalanced data.

Hypothesis Conclusion We reject the null hypothesis, (p = 2.2e-16), thus 3 point percentage and free throw rate positively affects win percentage.

We reject the null hypothesis, (p = 0.003), thus 3 point percentage and free throw rate positively affects post season ranking.

Conclusion and Recommendations

Overall in our findings we can see that our regression model did a good job in predicting champions but struggled with predicting earlier rounds. We determined that there's a high correlation between 3 point percentage and free throw accuracy on win percentage but might not be the only factors. We could enhance the dataset by collecting additional data, particularly for underrepresented categories, to improve the classification model's ability to predict these stages more accurately. Expanding the dataset will reduce biases caused by imbalanced data. We could also include more diverse variables to provide a holistic view of team performance. Overall the analysis focuses on offensive metrics, which limits the scope of the insights. Other critical aspects such as turnovers or rebounds should be taken into account to gain a more comprehensive understanding of success in basketball. We also notice a lack in contexual metrics, We can say that the data fails to include contextual metrics such as home-court advantage and player injuries, which are factors that may influence a team's success. Additionally in regards to seasonal patterns, it doesn't include how the team performance changes overtime, such as improvements/declines throughout the season.