Predicting 2023 Men’s NCAA March Madness Game Outcomes After the Fact Using Prospective Modeling & Machine Learning

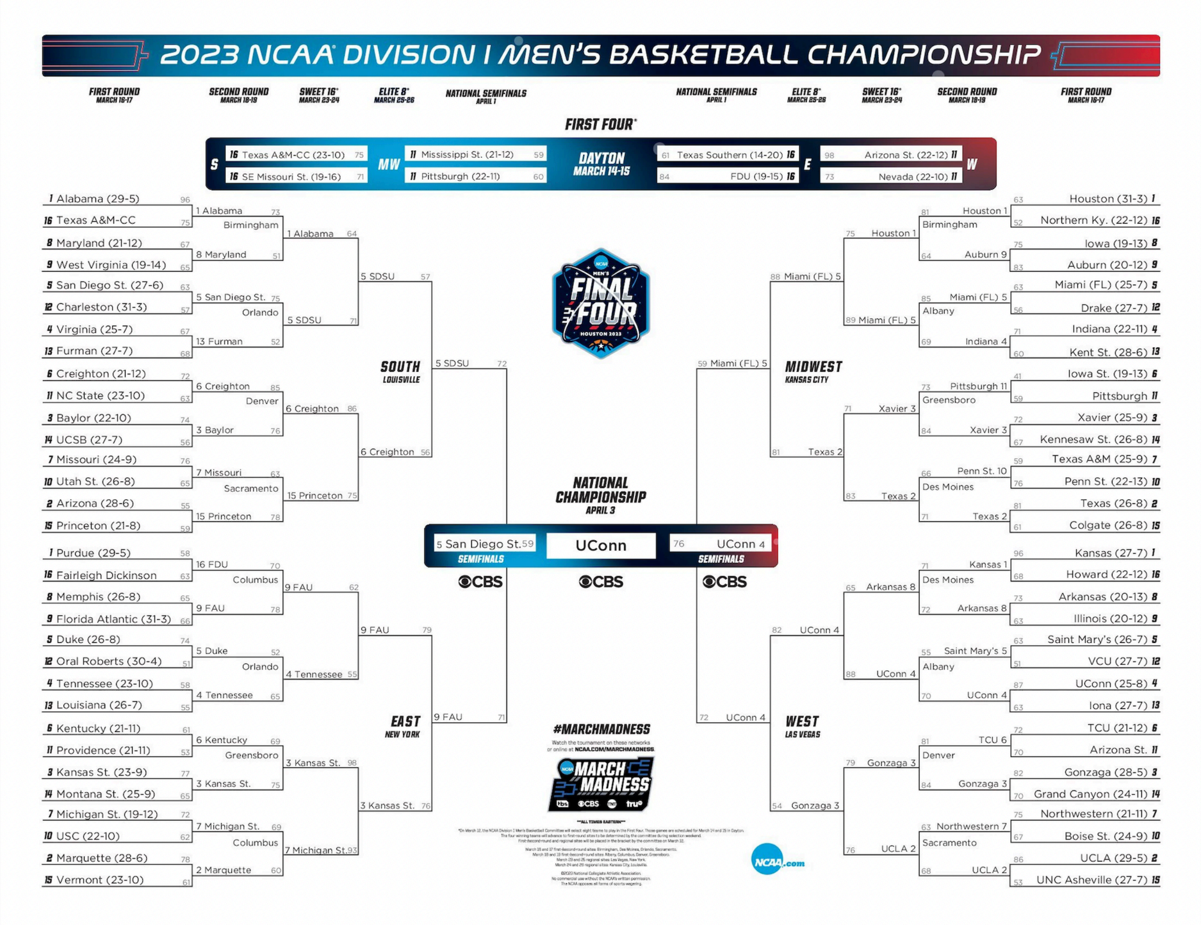
Reece Iriye, Word Count: 2,121

Abstract

This study investigates the performance of Logistic Regression, Penalized Logistic Regression, and Decision Tree models for predicting the outcomes of NCAA Division I Men’s Basketball games. Utilizing a dataset obtained from the `ncaahoopR` package, my analysis incorporates feature-engineered 10-game moving averages and whether or not a team has a home-court advantage. My findings indicate that the Logistic Regression model marginally outperforms the other models, with an accuracy of 63.2%, as well as other stronger error metrics. Game location (home or away) vastly outperforms all other proscriptive 10-game moving average variables, but this variable is rendered useless during the NCAA March Madness tournament when neither team has a home-court advantage. Some proscriptive variables, however, like 10-game moving averages of 3-point field goal attempts and win streaks play a somewhat strong predictive role in the model. In a setting like sports analytics where testing accuracy is an especially important error metric, this model has some flaws but still shines, given it is a proscriptive model and not a retrospective model.

Introduction

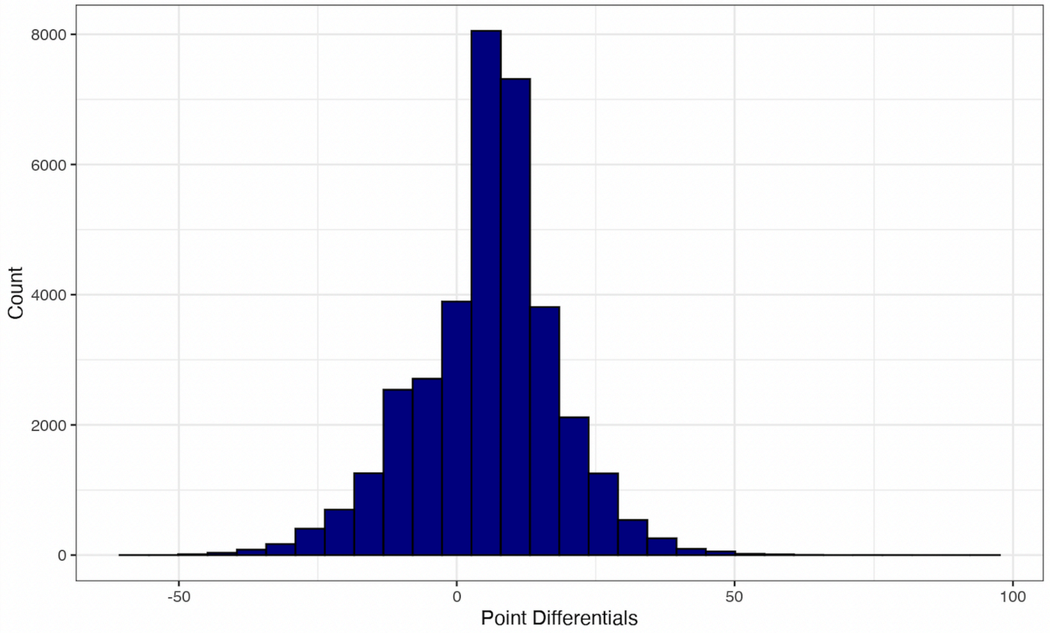
March Madness is a nickname for the annual National Collegiate Athletic Association (NCAA) Division I Men’s Basketball Tournament in the United States, which has a long history becoming the novel tournament that it is today. The first NCAA Men’s Basketball Tournament was held in 1939, with just eight teams competing in a single-elimination tournament format. The tournament expanded to 16 teams in 1951, and to 32 teams in 1975. In 1985, the tournament expanded to its current format of 64 teams, with four play-in games added in 2001 to determine the final 64 teams (Wilco, 2023).



The tournament has become a major cultural event in the United States, with millions of people filling out brackets and tuning in to watch and even bet money on the games. March Madness is known for its unpredictability and upsets, with lower-seeded teams often knocking off higher-seeded favorites. With the growth in hype emerging around March Madness in recent years, creating a prospective machine learning model that can accurately predict game outcomes could be interesting and practical for fans invested in the tournament.

Within our dataset originating from the `ncaahoopR` package, we observe 35343 total unique NCAA Men's basketball games played from the 2012-13 season up to the 2022-23 season. 364 total teams played in the league throughout the seasons, which showcases how the 64 teams competing in the March Madness tournament every year are some of the best teams in the league for that season. In **Figure 1** below, we can see the distribution of point differences depicting final game score comparisons between a team being evaluated against their opposing team. A positive score indicates that the team being observed won the game, and a negative score indicates that they lost.

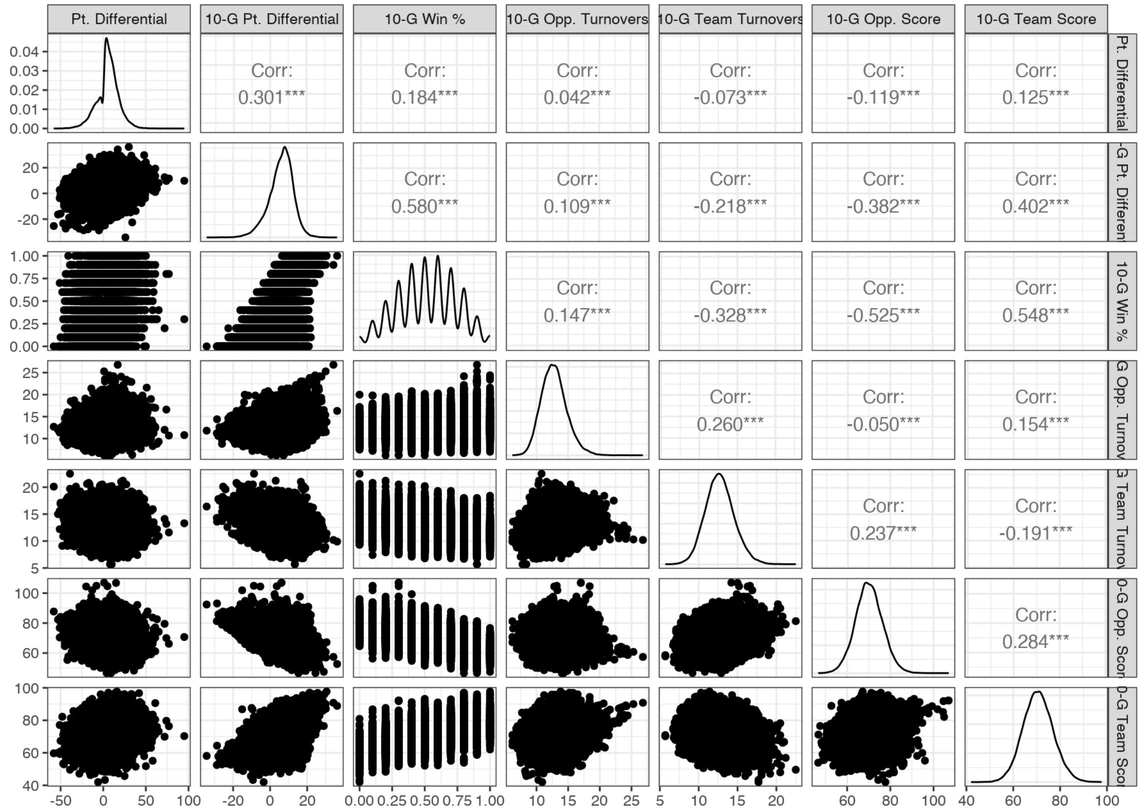
Figure 1: Histogram of March Madness Game Point Differentials



The data is approximately normal with a small number of point differentials exceeding a magnitude of 25 points, indicating that teams' final scores are often closer to one another and blowouts do not often occur in the NCAA Division I Men's Basketball league. This observation from **Figure 1** falls in line with March Madness's tendency to frequently have games with unexpected upsets, because point differentials do not have to exist several standard deviations from their mean for many game projections to result in a different team winning.

Creating an ideal prospective classification model for predicting future wins and losses for upcoming basketball match-ups should incorporate data from the recent past to incorporate teams' recent performance into the model. The dataset includes feature engineered 10-game moving averages to capture a team's momentum and utilize a team's performance in real time to suggest exactly how they will fare against their opponent in an upcoming match-up. Because the data focuses almost exclusively on 10-game moving averages as predictors of wins and losses, rows where either team had not played 10 games yet have been removed from the dataset. To understand exactly how point differentials at the end of games are correlated with some of our 10-game moving average predictors, and to evaluate whether or not multicollinearity is a concern in our model, I plotted a scatterplot matrix in **Figure 2** below to visualize some of these trends amongst our variables.

Figure 2: Scatterplot Matrix of March Madness Game Point Differentials



Multicollinearity is not a major concern in this model (at least with the variables being observed here), so we can move forward knowing that to some extent, complex relationships between involved variables in the model are not extremely prevalent. Assuming this trend persists with the rest of our data, machine learning models that tend to fit towards more complex shapes may not perform very well in this scenario (Varian, 2014). Models that are more flexible and less biased like Logistic Regression, for example, may comparatively perform better here because of the relationships seen in the data, but we will move forward testing several different models to ensure that analysis is correct.

Methods

Data Preprocessing

To preprocess the data, all rows signifying the first 10 basketball games played by either team in the game will be removed from the dataset. I mentioned this point earlier, but it is to ensure that our model is truly prospective in nature by not including incomplete data into the training and testing of our models. Additionally, all data will be normalized to ensure that stability errors do not occur, and heavily correlated predictors will be removed from the dataset to address collinearity if it exists.

Resampling Scheme

Because the dataset contains a large amount of data, I performed a train-test split where the training data is all the data from the 2012-13 season up to the 2021-22 season. 5-fold cross-validation with no repeats will be performed on the training data to mitigate overfitting and to ensure that establishing ideal tuning parameters for ML models requiring tuning are not too computationally expensive.

Logistic Regression

I plan to implement a logistic regression model in my analysis, because it is a baseline binary classification algorithm that sometimes tends to perform better than other more complex models in various settings (Wu et al., 2023). Logistic regression typically brings in less bias when fitting data, which is important especially because of our observations in **Figure 2** which displayed an unclear relationship between our predictors and overall point differences at the end of games and also our predictors alongside one another.

Penalized Logistic Regression

I will also examine a penalized logistic regression model using the NCAA Men's Basketball data to account for possible overfitting of variables. Penalized logistic regression resembles a similar shape as standard logistic regression, but it reduces the likelihood of overfitting by incorporating shrinkage parameters that "avoids extreme values of regression coefficients during model development" (Yan et al., 2022). If all predictors exhibit behavior similar to those with weak correlations, a penalty term might not be required, as there would be no overfitting to address.

Decision Tree

Decision Trees are formed using a hierarchy of branches representing conditions for our parameters (Song et al., 2015). Especially when trying to predict the probability of a team winning with 32 predictors, variable selection plays a vital role in ensuring our model does not overfit to data in case multicollinearity is an issue among predictors or predictors simply have no effect on teams winning or losing. I chose to use a decision tree instead of a random forest or KNN model, because I do not want the model I am testing to fit to an extremely unusual shape.

Results

Tuning Parameter Reporting

For penalized logistic regression using 10 hyperparameter combinations, the optimal tuning parameters are $\alpha = 8.086 \times 10^{-6}$ and $\lambda = 0.206$.

For the resulting decision tree using 10 hyperparameter combinations, the optimal tuning parameters to optimize the model incorporate a cost complexity of 3.031×10^{-6}, a tree depth of 5, and a minimum count of 2.

Error Metrics

To demonstrate the performance of each model on the testing data, I presented a range of metrics that provide a comprehensive evaluation of their overall effectiveness:

- *AUC*: The Area Under the Receiver Operating Characteristic (ROC) Curve measures the ability of the model to distinguish between classes, with higher values indicating better performance.
- *Accuracy*: The proportion of correctly classified data points out of the entire test set evaluates the overall correctness of the model.
- *Sensitivity*: Sensitivity measures the proportion of actual positives that are correctly identified by the model. It is the proportion of actual positives that are correctly identified by the model.
- *Specificity*: Specificity reflects the model's ability to avoid false alarms. It is the proportion of actual negatives that are correctly identified by the model.
- *Kappa*: Kappa is a metric that evaluates the agreement between the model's predictions and the actual outcomes, taking into account the possibility of agreement by chance. A higher Kappa value indicates better agreement.
- *Brier Score*: Brier Score measures the mean squared error between the predicted probabilities and the actual outcomes. Lower Brier scores indicate better calibration and overall performance.

Table 1 showcases the calculation of these error metrics for the Logistic Regression, Penalized Logistic Regression, and Decision Tree models.

While the scores for each category for all variables are extremely similar, the logistic regression model outperforms both the penalized regression and decision tree model in every error metric.

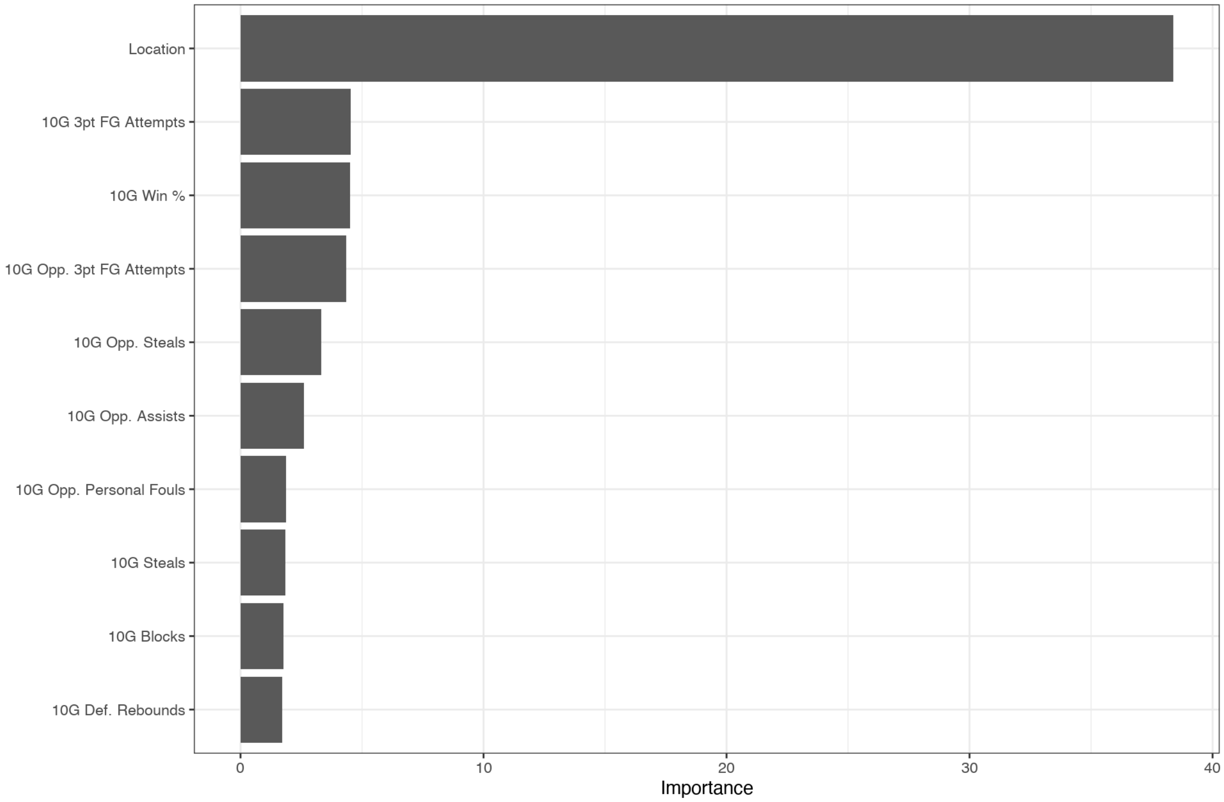
Table 1: Performance Metrics

Metrics	Logisitc Regression	Penalized Logistic	Decision Tree
AUC	0.688	0.688	0.663
Accuracy	0.632	0.632	0.613
Sensitivity	0.627	0.627	0.611
Specificity	0.638	0.638	0.614
Kappa	0.264	0.264	0.225
Brier Score	0.223	0.223	0.229

Variable Importance

Figure 3 presents a Variable Importance Plot for the Logistic Regression model identifying which variables played a key role in the prediction of whether or not a team would win a basketball match-up.

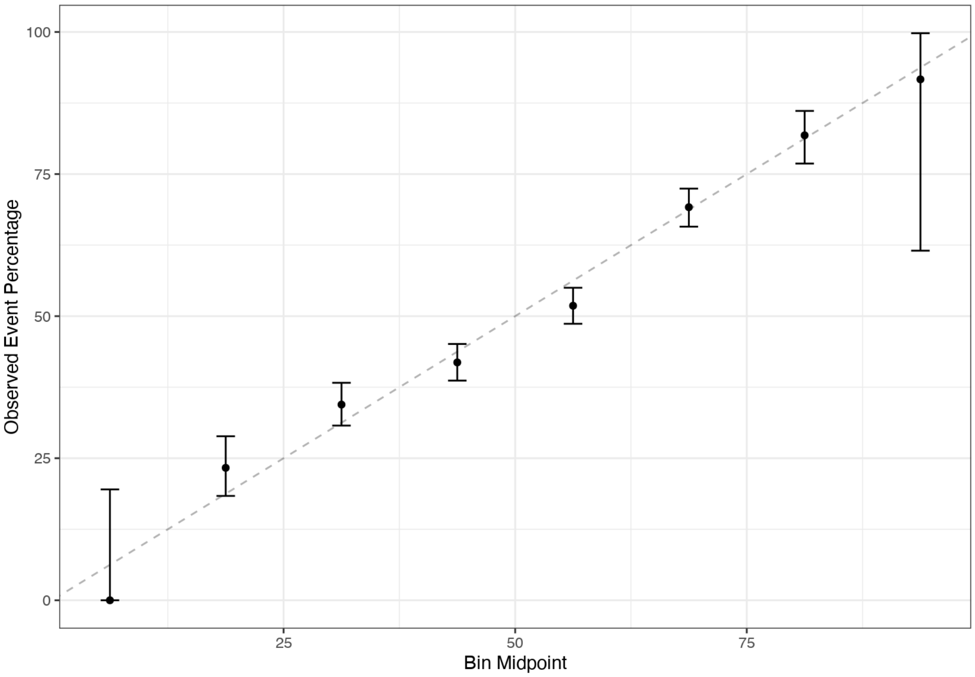
Figure 3: Logistic Regression Variable Importance Plot



Calibration Analysis

Figure 4 also analyzes the Logistic Regression model predicting wins and losses, but it showcases how the model is well-calibrated across all points of data.

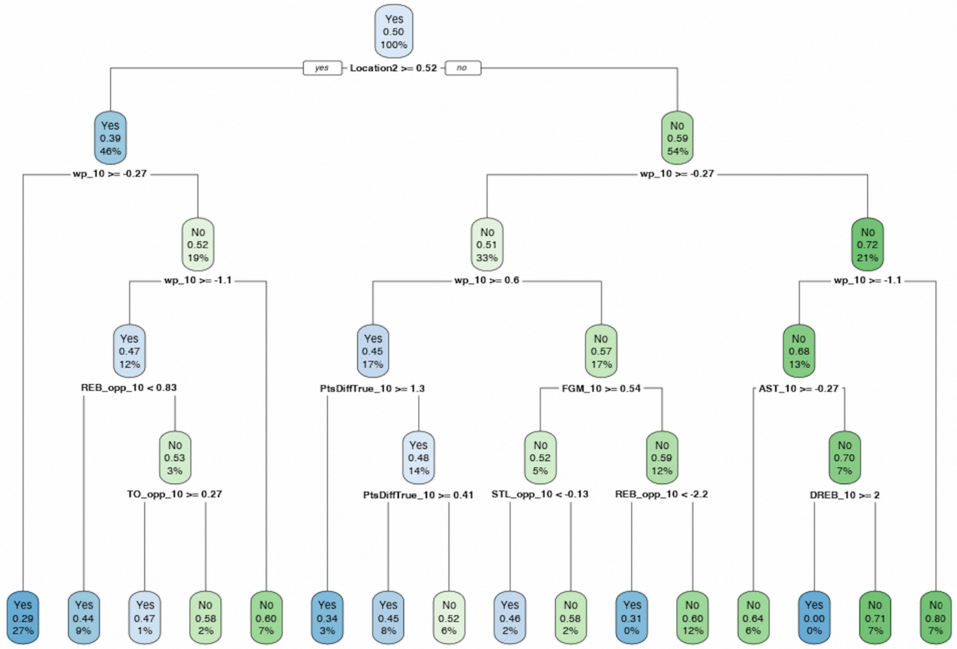
Figure 4: Logistic Regression Callibration Plot



Breakdown of Variables Analyzed

Figure 5 observes the Decision Tree model and identifies how prospective variables had been broken down to predict whether or not a basketball team would win or not in the 2022-2023 season.

Figure 5: Decision Tree Breakdown



Conclusion

In this analysis, we explored the performance of Logistic Regression, Penalized Logistic Regression, and Decision Tree models for predicting the outcomes of NCAA Men's Basketball games. Based on the error metrics presented in **Table 1**, the Logistic Regression model barely outperformed both Penalized Logistic Regression and Decision Tree models across all metrics. The Decision Tree model performed the worst in comparison to the other models, and the Penalized Regression model performed similarly to Logistic Regression.

Penalized Logistic Regression performed similarly to regular Logistic Regression because of the α is approximately equal to 0, making the model extremely similar to Logistic Regression in the first place. This result aligns with my initial assumption that a more flexible and less biased model like Logistic Regression would be better suited for this dataset, given the relationships observed in **Figure 2**.

The Variable Importance Plot in **Figure 3** provides valuable insights into the key variables that contributed to the prediction of game outcomes in the Logistic Regression model. Whether a team is playing at home or away is by far the biggest determinant of the result of the game. As seen through **Figure 5**, if a team played an away game, they were much more likely to lose a game unless their 10-game streak was strong and their opponent's performance had been lagging behind for their past 10 games. The rest of the 10-game moving average variables were not nearly as strong, but 3-point field goal attempts, winning streak, and opponent 3-point field goal attempts were the next strongest variabes in **Figure 3** and do play an important role in the decision tree in **Figure 5**. By identifying location and these 10-game moving averages, we can better understand the factors that have a significant impact on the success of teams in NCAA Men's Basketball games.

Additionally, the Calibration Plot in **Figure 4** demonstrates that the Logistic Regression model is well-calibrated across all data points, suggesting its reliability in predicting the outcomes of basketball games. It is important to note, however, that an accuracy of 63.2% is not very high, but it is somewhat strong for a prospective model in sports analytics. As mentioned earlier, NCAA Basketball games and especially March Madness games are known to have volatile scores where underdogs often come out as victorious. Especially in the NCAA March Madness tournament where both teams are playing away games in a location unfamiliar to them, our model will not be nearly as strong because our strongest predictor of whether or not a team will win disapears.

References

Song, Y. Y., & Lu, Y. (2015). "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry*, 27(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Varian, H. R. (2014). "Big data: New tricks for econometrics." *Journal of Economic Perspectives*, 28(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Wilco, D. (2023, March 23). "March Madness History: A Comprehensive Guide to the Men's Tournament." *NCAA.com*. https://www.ncaa.com/news/basketball-men/article/2023-03-08/march-madness-history-comprehensive-guide-mens-tournament.

Wu, T., Wei, Y., Wu, J., et al. (2023). "Logistic regression technique is comparable to complex machine learning algorithms in predicting cognitive impairment related to post intensive care syndrome." *Scientific Reports*, 13, 2485. https://doi.org/10.1038/s41598-023-28421-6

Yan, Y., Yang, Z., Semenkovich, T. R., et al. (2022). "Comparison of standard and penalized logistic regression in risk model development." *JTCVS Open*, 9, 303-316. https://doi.org/10.1016/j.xjon.2022.01.016