Predicting College Basketball Success – Does Conference Really Matter?

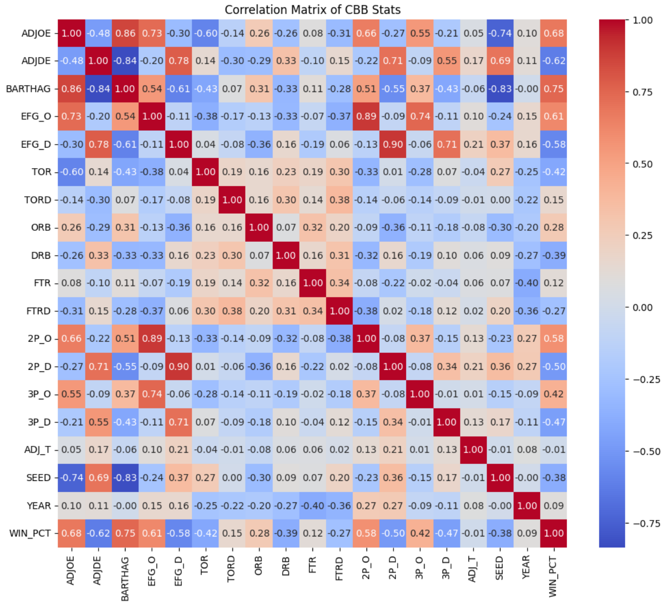
Andre Archer, Mary Meyer, and Ashley Foster

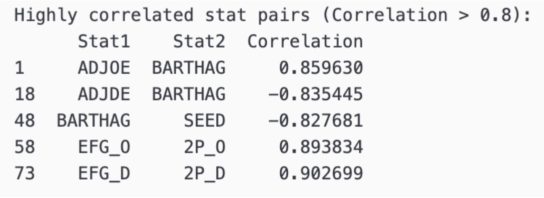
Introduction

The goal of this project was to create a series of notebooks and models to understand the most important statistics influencing a college basketball team's seasonal win percentage and their chances of qualifying for the NCAA tournament. The chosen college basketball dataset included performance data for 361 teams for the 2013 to 2025 seasons. Data prior to 2013 contains inaccuracies in the chosen statistics and is commonly seen in the basketball community as unrepresentative of a team's performance prior to the tournament. The college basketball dataset contains statistics including the team name, conference, number of games and wins, adjusted offensive efficiency (ADJOE), adjusted defensive efficiency (ADJDE), power rating (BARTHAG), effective field goal percentage shot and allowed for offense (EFG_O) and defense (EFG_D), turnover rate, and additional statistics. The chosen models for the seasonal win percentage prediction were a linear regression (LR) and a gradient boosted regressor (GBR), and a random forest classifier (RF) was used for predicting if a team qualified for the NCAA tournament. As the conference identifier for each team was represented by a string, a few methods were utilized to convert this string into a value suitable for the chosen models. First, the conferences were split into two groups (power conference or non-power conference), and a binary assignment of values was applied based on the group (1 for power conference and 0 for non-power conference). An additional method of one-hot encoding for each individual conference (e.g. assigning an integer for each unique conference) was applied to determine which encoding method performed better. The results of these methods were then compared based on the model performance.

Data Processing

The dataset first needed several cleaning steps prior to model input. There were 2 datasets available for Division I college basketball teams: a dataset from 2013 – 2024 and a partial dataset from 2025. First, the dataset columns were explored to determine any column differences between the two datasets. Any discrepancies (e.g. a column named 'TEAM' for the 2013 – 2024 dataset and named 'Team' for the 2025 dataset) were either fixed or the columns were dropped to concatenate the datasets. To analyze a team's win rate, a column was added to the combined dataset as the percentage of games a team won out of played games per season. Any string or object columns (e.g. the conference name) were changed to integer values through one of the following methods: binary assignment (column has a 1 if the conference is a part of a "power conference" and a 0 if not) or one-hot encoding (each conference assigned a different integer value). Additionally, any empty cells or NaN values were replaced with a 0 or the row/column was removed. Some columns were also removed since they were not assumed to be important to the model or were highly correlated: team name, season year, wins above average, BARTHAG (power rating), and offensive and defensive 2-point rates (2P_O and 2P_D, respectively). Some visualizations were also created to determine most important and highly correlated features. The correlation matrix of the dataset columns can be seen in Figure 1. From this matrix and the variables with correlations above 0.8, the BARTHAG and the 2P_O and 2P_D columns were dropped. This ensures better efficacy of the model by eliminating weights on highly correlated variables. The following columns were kept after dataset cleaning and removing highly correlated columns: efficiency on offense/defense, turnover rates, rebound rates, free throw statistics, conference name (encoded), tournament qualification, and adjusted tempo.





```
Highly correlated stat pairs (Correlation > 0.8):
      Stat1    Stat2  Correlation
1     ADJOE  BARTHAG     0.859630
18    ADJDE  BARTHAG    -0.835445
48  BARTHAG     SEED    -0.827681
58    EFG_O     2P_O     0.893834
73    EFG_D     2P_D     0.902699
```

Figure 1. Correlation matrix (left) and the highly correlated pairs of variables (right) with correlations above 0.8.

Visualizations of the trends between variables were used to confirm model training over the correct trends. First, the win percentage versus ADJOE and ADJDE were plotted, as shown in Figure 2. The ADJOE and ADJDE are statistical measures that estimate a team's points scored or allowed, respectively, per 100 possessions against an average offense/defense. Therefore, it makes sense that a higher ADJOE corresponds with a higher win percentage and vice versa for ADJDE.

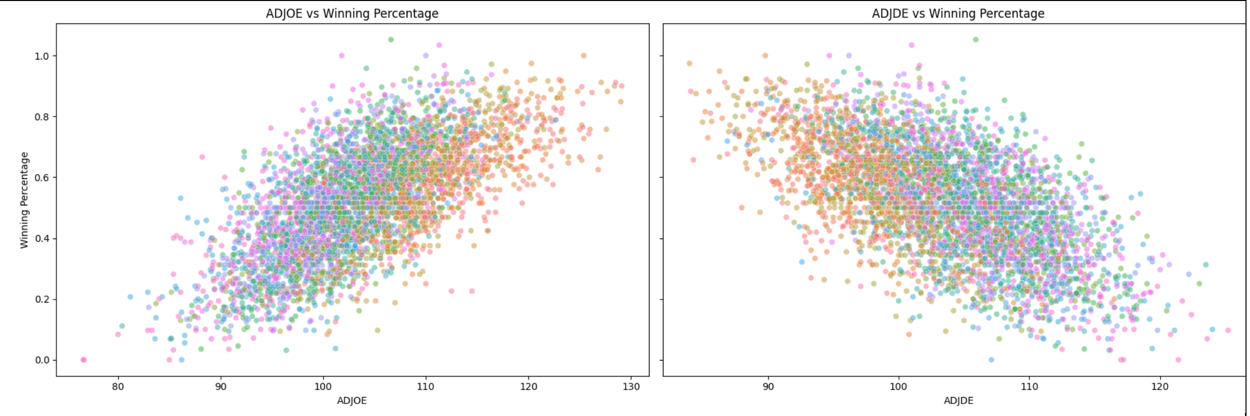
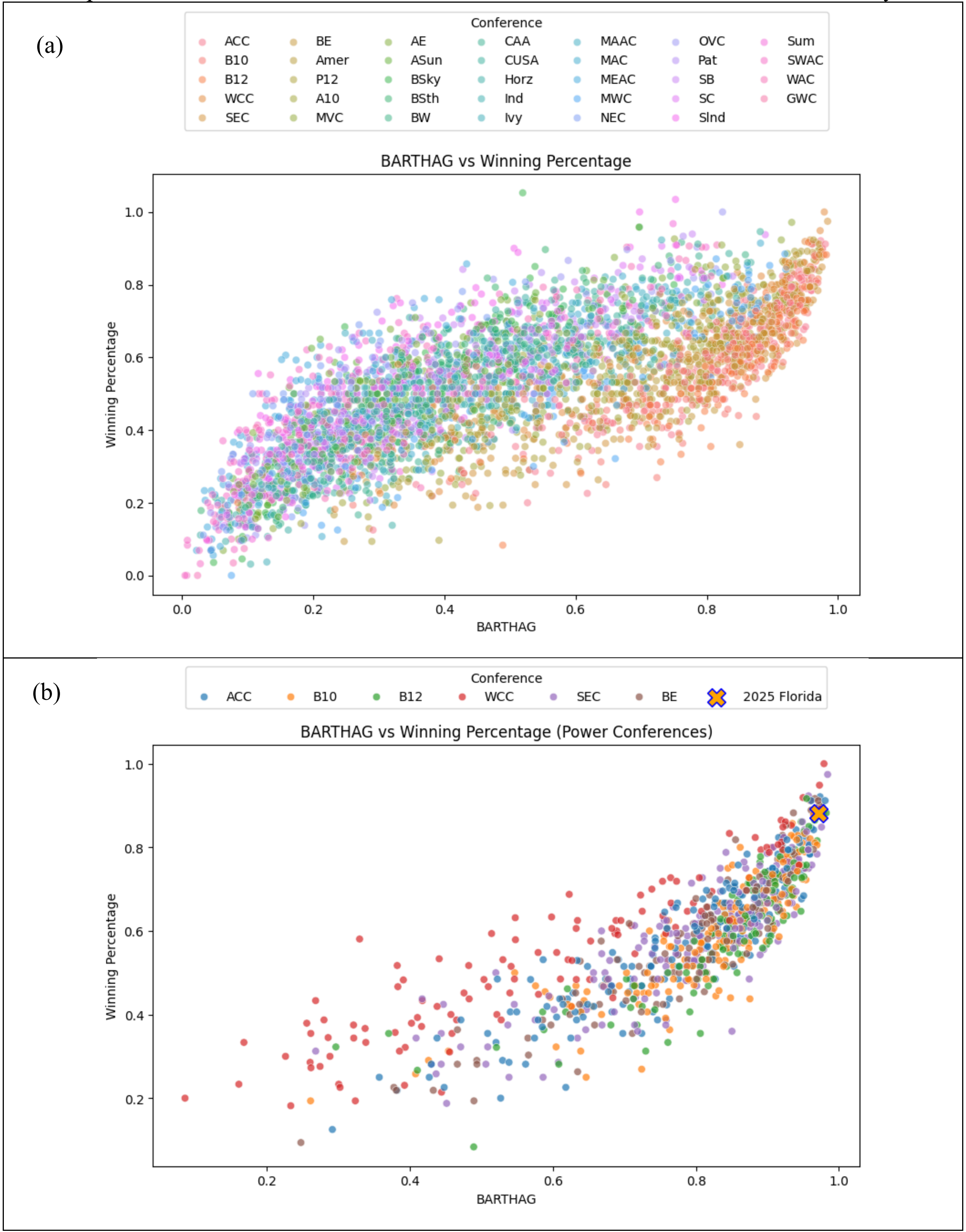


Figure 2. Winning percentage versus ADJOE (left) and ADJDE (right).

Figure 3(a) illustrates the win percentage versus BARTHAG for all conferences and figure 3(b) for just the power conferences to see how important inclusion in a power conference is for a team's overall winning percentage and BARTHAG rating.


(a)
Conference
ACC
B10
B12
WCC
SEC
BE
Amer
P12
A10
MVC
AE
ASun
BSky
BSth
BW
CAA
CUSA
Horz
Ind
Ivy
MAAC
MAC
MEAC
MWC
NEC
OVC
Pat
SB
SC
SInd
Sum
SWAC
WAC
GWC
BARTHAG vs Winning Percentage
Winning Percentage
0.0
0.2
0.4
0.6
0.8
1.0
BARTHAG
0.0
0.2
0.4
0.6
0.8
1.0
(b)
Conference
ACC
B10
B12
WCC
SEC
BE
2025 Florida
BARTHAG vs Winning Percentage (Power Conferences)
Winning Percentage
0.2
0.4
0.6
0.8
1.0
BARTHAG
0.2
0.4
0.6
0.8
1.0

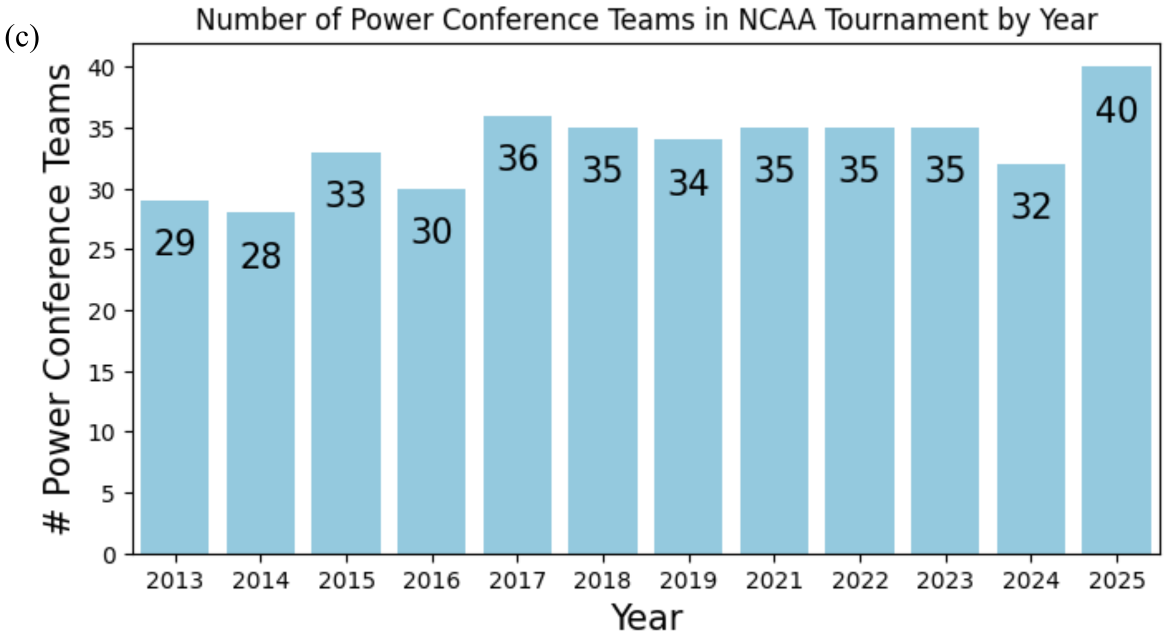


Figure 3. Win percentage versus BARTHAG rating for (a) all conferences and (b) only power conferences. (c) Number of power conference teams in NCAA tournament by year (out of 64 teams that enter the conference).

As can be seen from figure 3, the power conferences tend to make up most of the teams with both high BARTHAG scores and high winning percentages. These teams also make up over half of the teams each year in the NCAA tournament. This likely has to do with the number of resources these teams have (e.g. good coaches and funding) as compared with other conferences. Therefore, perhaps the different encoding mechanisms (binary vs. one-hot encoding) might train differently since the conference appears to have a strong impact on winning percentage. Note, Florida 2025 is highlighted on the plot as they won the national championship this year, illustrating where a successful team should be on this plot.

After these cleaning and visualization steps, the data was split into 80% training and 20% testing with a random seed of 16. After initial fitting of the models, the hyperparameters were then optimized for each model to improve overall performance.

Predicting Win Percentage

Two models were used for predicting win percentage: a Linear Regression model and a Gradient Boosting Regression model.

Linear Regression

Linear regression (LR) assumes that there is a linear relationship between the input and output, represented by a straight line on the data. It is a simple and easy to understand supervised machine learning algorithm. It works by minimizing the error between predicted in actual values for a dataset. If the dataset, however, is not linear or is too complex, LR may perform poorly as compared to more robust and complex models.

Before modeling the data, the data required scaling as some of the numerical values were presented with different orders of magnitude. The StandardScaler function from Scikit-Learn was used to scale the data prior to model fitting. After scaling, all values for variables were between -10 and +10 and would therefore impact the model with similar weights.

The model was then fit with the default model parameters for the binary encoding of the conference. The model achieved a mean squared error (MSE) of ~0.005 and an R^2 score of 0.844. This R^2 value is fairly good, but a higher value (above 0.9) would provide a better fit. After this, the hyperparameters were fit using the RandomizedSearchCV function from Scikit-Learn with the following parameters tested: 'fit_intercept': [True, False], 'copy_X': [True, False], 'n_jobs': [None, 1,5,10,15]. The following parameters were chosen as the best fit from the RandomizedSearchCV function: 'n_jobs': 1, 'fit_intercept': True, 'copy_X': True. The model was fit again using these updated parameters, and the new MSE was ~0.005 and the R^2 value was still 0.844. The results from each of these models can be seen in figure 4. The highest coefficients for each of the dataset variables were also analyzed, and the highest coefficients were for the offensive and defensive effective field goal (EFG_O = 0.077 and EFG_D = -0.070). This makes sense since, intuitively, scoring more points would give more wins and being scored on more would give less wins. Interestingly, inclusion in a power conference gave a coefficient of -0.038, which was unexpected since it would be assumed that a team would be more likely to win with more resources. However, these teams generally have harder schedules than teams with less resources, and thus power conference teams would be expected to lose more.

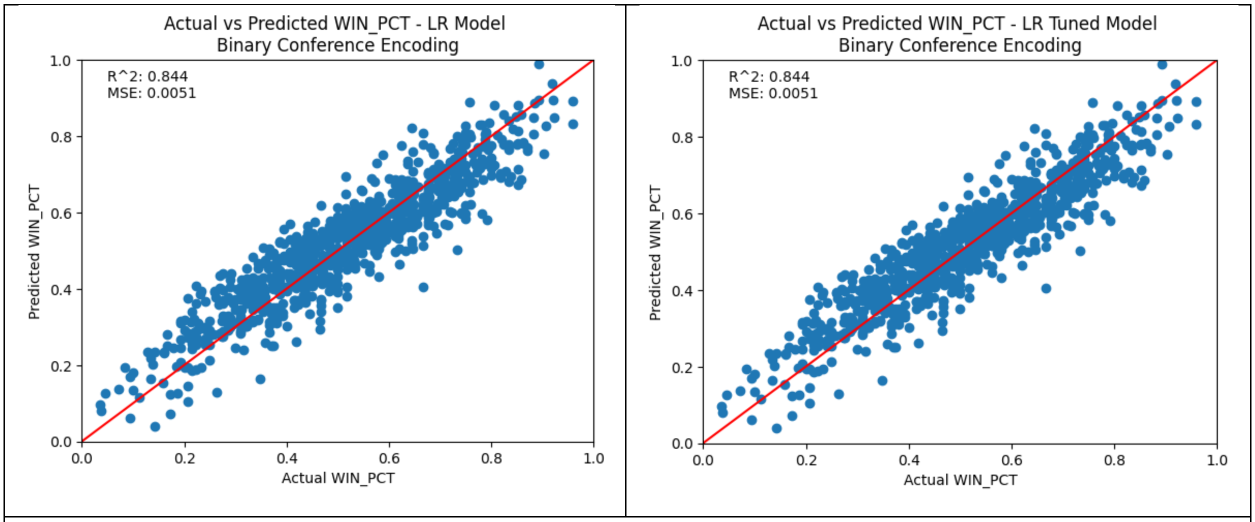


Figure 4 The actual versus predicted win percentage for the binary conference LR model for the initial model (left) and the model with refined hyperparameters (right).

Next, the same strategy was applied to the one-hot encoded data to see the most significant conferences and if those conferences were the power conferences. Once again, the model was fit both with and without hyperparameter optimization. The same hyperparameters were tested for this model, and the following were chosen from the optimization: 'n_jobs': 10, 'fit_intercept': True, 'copy_X': True. The results from this can be seen in Figure 5. Again, tuning hyperparameters did not help with the model performance. The most important features from this model were also extracted to determine which conferences were most significant. These conferences were: SWAC (0.135), MEAC (0.131), Slnd (0.128), Ind (0.102), and Asun (0.102). Interestingly, none of these conferences are considered power conferences and are actually usually considered as some of the weaker ones. As discussed earlier, however, the power conferences have more difficult schedules than non-power conferences, often contributing to lower win percentages. Additionally, the model performed better using the one-hot encoded conference than the binary power conference inclusion encoding. This suggests that inclusion of a team in a power conference does not necessarily indicate that they will have a higher seasonal win rate than a team not in a power conference. Inclusion of game-by-game data would help strengthen this claim, especially if a ranking were applied to the strengths of the teams in the game to weight game wins by matchup.

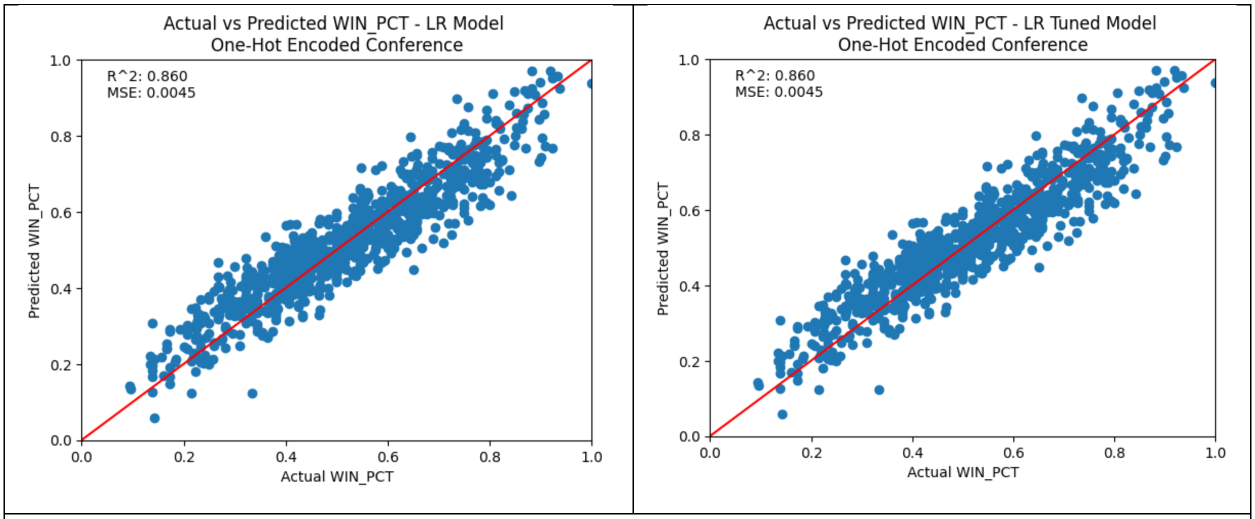


Figure 5 The actual versus predicted win percentage for the one-hot encoded LR model for the initial model (left) and the model with refined hyperparameters (right).

Gradient Boosting Regression

Gradient Boosting Regression (GBR) is a tree-based, ensemble technique that relies on the reduction of model residuals. GBR uses many decision trees throughout its iterative fitting process.

After data processing and scaling, both of the binary encoded and one-hot encoded datasets were assessed by the GBR model. Initially, both dataset fittings were completed without hyperparameter tuning, and after these preliminary results, the models underwent hyperparameter tuning using the RandomSearchCV modules from the SciKit-Learn library. For the binary encoded case, the MSE was ~0.0064 and the R^2-score was 0.803 prior to parameter tuning. The parameter grid space for tuning was: {'loss': ['squared error', 'absolute error', 'huber', 'quantile'], 'n_estimators': [100,200,300,400,500], 'learning_rate': [0.01,0.05,0.1,0.2], 'max_depth': [1,2,3,4,5]}. The best hyperparameters were: {'loss': 'squared_error', 'n_estimators': 400, 'learning_rate': 0.05, 'max_depth': 3}. After tuning, the MSE was ~0.0056 and the R^2-score was 0.823. The results for the binary encoded data for the GBR model can be seen in figure 6.

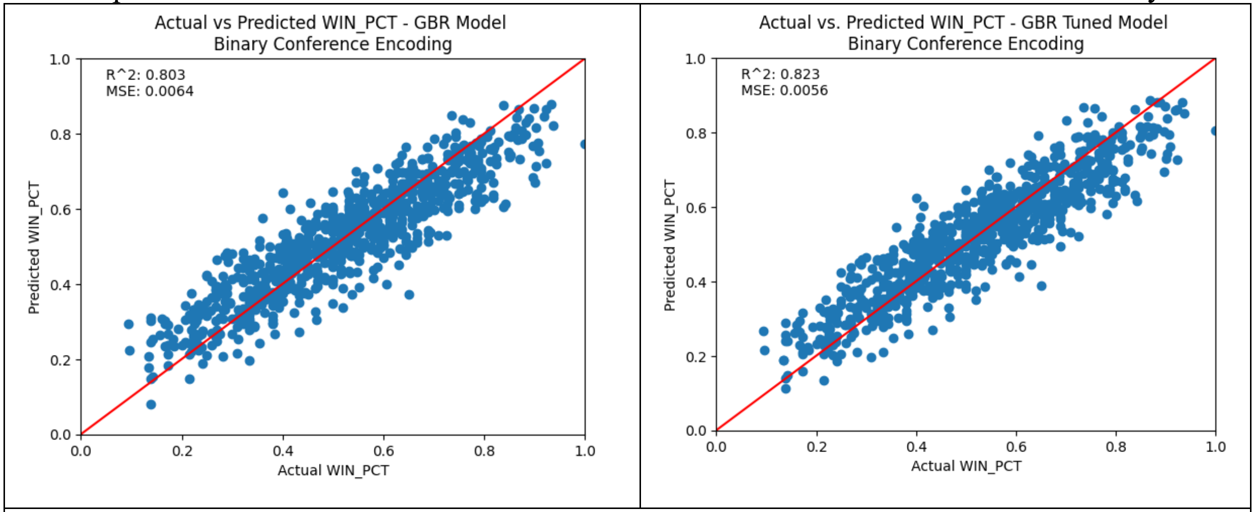


Figure 6 The actual versus predicted win percentage for the binary encoded GBR model for the initial model (left) and the model with refined hyperparameters (right).

Regarding the GBR model with the binary encoded dataset, the highest features of importance are ADJOE (0.374), EFG_D (0.156), EFG_O (0.139), and ADJDE (0.124). As previously discussed, all four of these performance metrics are sensical in predicting win percentage.

For the one-hot encoded dataset, the GBR model performance, prior to tuning had an MSE of ~0.0064 and an R^2-score of 0.798. The parameter grid space for hyperparameter tuning was the same as for the binary encoded dataset. The tuned parameters were: {'loss': 'squared_error', 'n_estimators': 200, 'learning rate': 0.1, 'max_depth': 3}. After tuning, the MSE was ~ 0.0055 and the R^2-score was 0.828. The GBR model performed better after hyperparameter tuning for the one-hot encoded dataset. These results can be seen in figure 7.

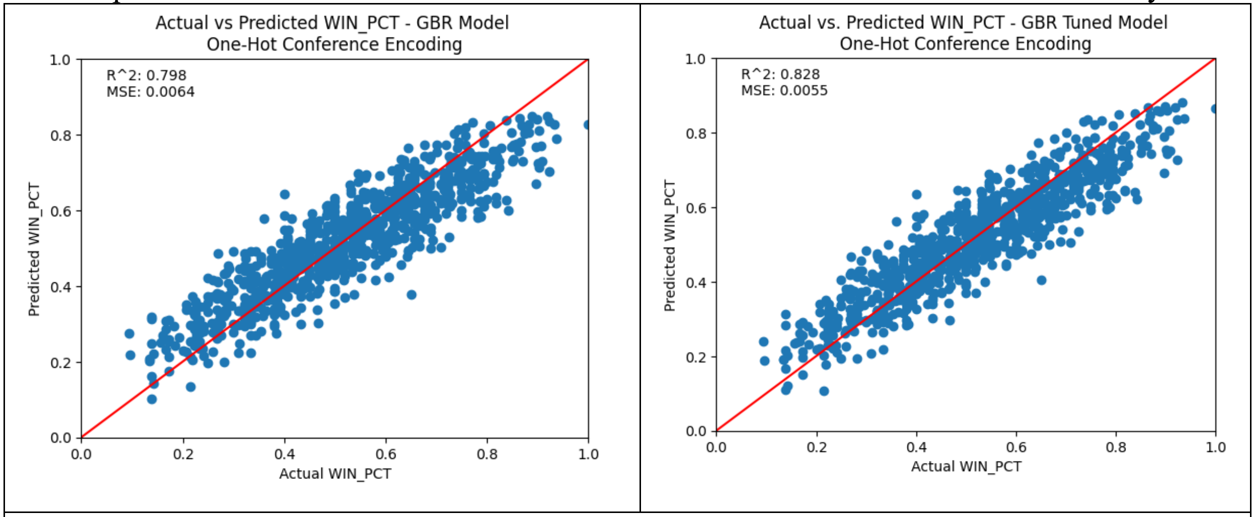


Figure 7 The actual versus predicted win percentage for the one-hot encoded GBR model for the initial model (left) and the model with refined hyperparameters (right).

Regarding the one-hot encoded dataset with the GBR model, the main features of importance are: ADJOE (0.372), EFG_D (0.163), EFG_O (0.140), and ADJDE (0.114). These features show similar importance between the binary and the one-hot encoded datasets.

Overall, the performance of the GBR over both data classification sets was worse than the LR model. This was not an expected outcome. If the data is truly linear with respect to the factors assessed, then that would explain the better LR model performance. Additionally, GBR is prone to overfitting, which may have caused a performance issue. A more likely possibility is that there are confounding factors in the dataset of limited size that change over time. For example, data from 2014 and 2025 are both in used in training, but the landscape of college basketball has changed significantly between these two points in time. Therefore, it is reasonable to expect a maximum performance that is able to be achieved by these models with the dataset used. This can likely be improved on with more data or properly considering changes in the rules and payment system in college sports, but that would require resources and time outside the scope of this course and project.

Predicting Tournament Qualification

NCAA tournament qualification is a main goal for, assumingly, all NCAA teams every year. Therefore, it is important to understand what goes into tournament placement and if there are any performance predictors that contribute to the end-of-season competition. We used a

Random Forest Classification (RFC) to predict tournament qualification. RFC is a tree-based approach that takes the most common output of many decision trees as the final predictor of a dataset. Here, we assessed tournament data for the 64 teams that qualified for the NCAA tournament for each year 2013 – 2025. Using a 80% training and 20% testing split and a random state of 12, both the binary encoded and one-hot encoded versions of this data were fitted with the RFC model.

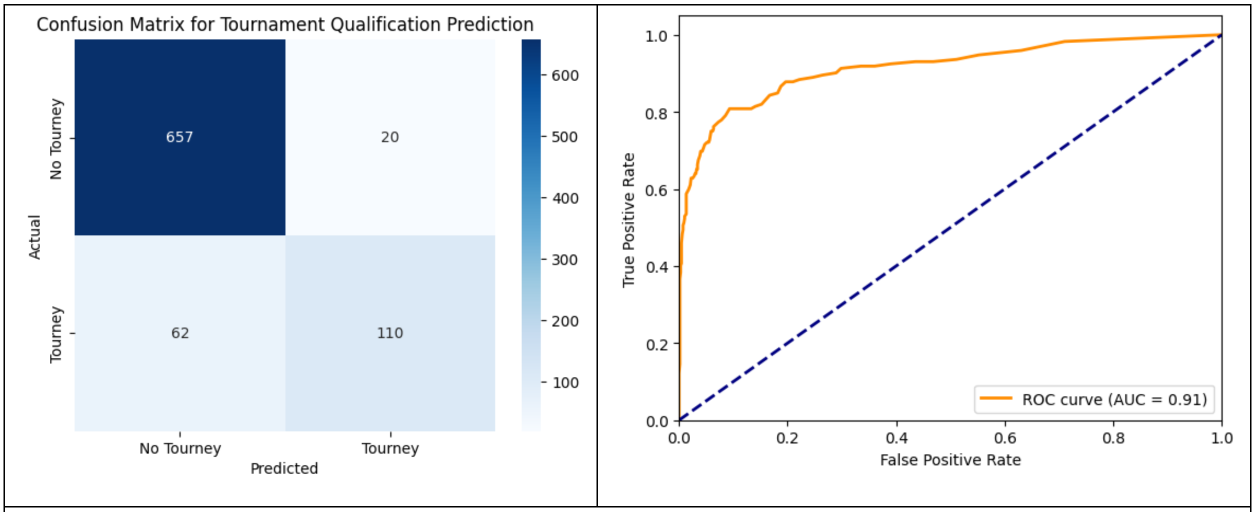


Figure 8. Confusion matrix (left) and ROC-AUC curve (right) for the RFC model fitting of the binary encoded tournament qualification dataset.

Figure 8 illustrates the confusion matrix and ROC-AUC curve for the RFC model prediction of the binary encoded dataset. The top 4 most important features are ADJOE (0.230), ADJDE (0.161), EFG_O (0.081), and EFG_D (0.079). None of these features are conferences, but rather performance statistics for how each team plays.

Figure 9 illustrates the confusion matrix and ROC-AUC curve for the RFC model prediction of the one-hot encoded dataset. The top 4 most important features are ADJOE (0.197), ADJDE (0.150), EFG_D (0.088), and EFG_O (0.078). There is a slight switch in the importance of the EFG_O and EFG_D statistics for the one-hot encoded dataset. Still, none of these 4 most important features are conferences. This is to say that the conference a team is in does not matter when it comes to tournament qualification, but rather how the teams perform throughout the season.

There are marginal differences between the results presented in figures 8 and 9, with the RFC model of the one-hot encoded dataset appearing slightly better in performance. Table 1 confirms this result with the comparison of performance metrics, such as, accuracy, precision, recall, and F1-score.

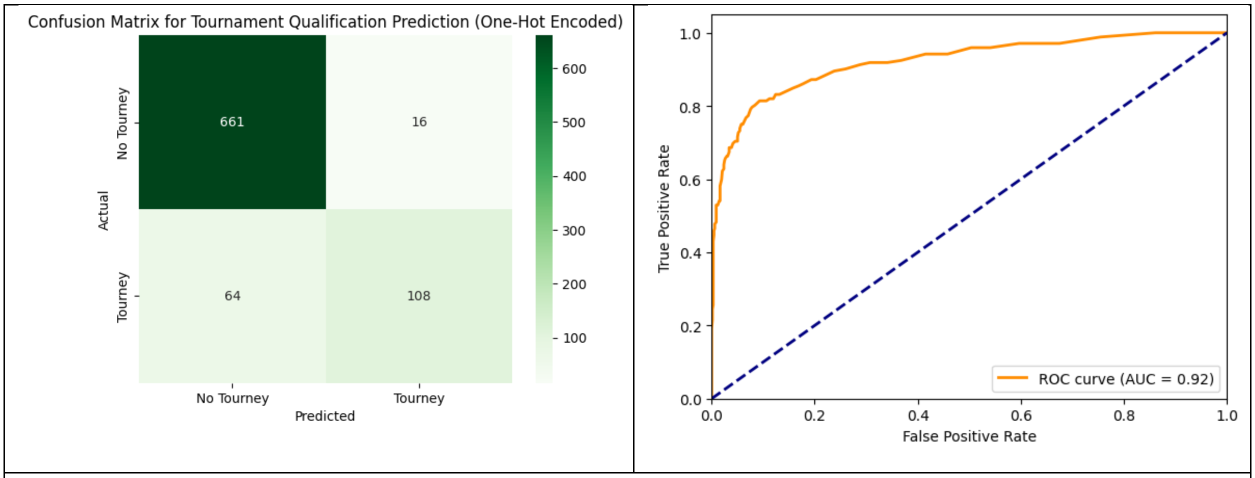


Figure 9. Confusion matrix (left) and ROC-AUC curve (right) for the RFC model fitting of the one-hot encoded tournament qualification dataset.

Table 1. RFC model performance for binary and one-hot encoded tournament datasets.

	Accuracy	Precision	Recall	F1-Score
Binary	0.90	0.88	0.80	0.83
One-Hot	0.91	0.89	0.80	0.84

Conclusion

After processing and fitting machine learning algorithms to NCAA college basketball statistics for the 2013-2025 seasons, there are firm conclusions that in-game performance is more of a predictor for qualifying for the end-of-season tournament. The conference a team is placed in does not contribute to the tournament qualification, but rather metrics such as points scored on opponent teams and points allowed by opponent teams matter more. Overall, an efficient offense is the most important in predicting team performance. Efficient defense is the second most important statistic. These results were confirmed with both Linear Regression and Gradient Boosting Regression models. To increase the win percentage of a team, make an easier schedule. To make the end-of-season tournament, perform better throughout the season – regardless of the conference or difficulty of season schedule. This makes sense when considering why power conferences also produce more tournament teams, although it is not just being in the power conference that causes this result. Typically, power conference teams are larger schools with more funds, pull factor, and name recognition, so they are then able to gather more talent and pay them in today's game than in the past. However, there are schools with high recognition and talent that make the tournament regularly being outside of a power conference, such as Wofford, Loyola-Chicago and Xavier. These schools rely on focusing on basketball talent and historical name recognition.

Improvements in the dataset to include individual player statistics, coaching metrics, and amount of money teams spend per player and staff could help improved model predictions if these were deemed features of importance. However, many of these statistics are either difficult to acquire or guarded behind API walls for use by gambling platforms or sports firms.