Identifying Determining Factors for At-Large March Madness Bids Using Machine Learning Methods

An Independent Research Project

by

Shaye O'Beirne

June 2023

Identifying Determining Factors for At-Large March Madness Bids Using Machine Learning Methods

Independent Research Project

Shaye O'Beirne

Abstract:

Each year, 68 Division I Men's and Women's College Basketball Teams participate in the annual NCAA Tournament, famously known as March Madness. Although there is no perfect science to determine which teams will be selected to take part in the NCAA Tournament, aside from the guaranteed spots held by the conference championship winners, there are certain factors that play an important role in this determination. The question of which non-conference-winning teams will be selected at large serves as a wonderful application for various Machine Learning methods to identify which factors are most influential in the probability that a team is selected to participate in the tournament. In this analysis, a Logistic Model will be employed to determine which commonly (as well as not-so-commonly) used metrics and statistics affect the probability a team is selected at-large, as well as to what extent they impact the team's participation outcome.

Contents

CHAPTER 1

Introduction

Over the last few decades, College Basketball has emerged as one of the most competitive and electrifying forms of spectator-sport entertainment. As the game itself continues to evolve and teams grow in their development, college basketball games yield millions of viewers each season. To keep up with the ever-growing strength of opposing teams, coaches have begun to leverage resources available to them with the hopes of creating an edge for their team. One advantage that many collegiate and professional teams have begun to implement into their operations is analytics. The use of Sports Analytics comes in many forms, such as player evaluation in scouting, developing offensive and defensive strategies, or analyzing training regimens. While the applications may seem limitless, they all have one common goal: **to get wins**.

The use of analytics has become increasingly popular, among both coaches and fans, in predicting outcomes for the annual NCAA Division I Basketball Tournament. March Madness is a highly-followed series of single-elimination tournament games that follow regular-season play and post-season conference tournaments. Of the 68 tournament invitees, 32 are automatically granted a spot in the tournament following their winning of their conference's tournament championship, while the other 36 teams are invited to the tournament via an at-large bid. Following the conclusion of all conference tournaments, a 10-member committee convenes to select 36 teams (28 of which are guaranteed to play in the tournament, with the other 8 competing in a play-in tournament to secure their bracket spot) that will participate in the tournament on the basis of various stats, metrics, and rankings. [10] While some metrics may inherently be favored in the deciding which teams will be recipients of an at-large bid, there is no perfect science or standard formula used to determine these teams.

Within Sports Analytics is a high-demand topic known as Machine Learning, which are algorithms constructed with the ultimate goal of leveraging data and finding insights that can forecast the data's future behavior and classify observations. [20] A popular area of study in both Mathematics and Computer Science, Machine Learning has extensive capabilities to draw conclusions and highlight patterns from data sets of any size. The analysis in this paper aims to find the most efficient and effective method of predicting which non-conference-winning teams will receive an at-large bid to participate in the March Madness Tournament. Additionally, the resulting regression coefficients will be used to determine which factors hold the most weight, and to what extent, in predicting if a team will receive a bid.

CHAPTER 2

Datasets

2.1. Division I Men's Basketball Dataset

The data for this analysis has been acquired primarily from Kenpom.com, a website created by Kenneth Pomeroy that displays and archives various statistics for all Division I Men's Basketball Teams, as well as their players. Kenpom uses both box score records as well as play-by-play data in its aggregation of team and player statistics, dating from the 2022-23 season back to 2002. The uniqueness of Kenpom come from the use of proprietary algorithms to rank teams and provide ratings, based on both conference and non-conference play. [16]

The website is broken down into several categories, including General Rankings, Four Factors (Field Goal Percentage, Rebounding, Turnovers and Free Throws, for both Offensive and Defensive metrics), Individual Player Statistics, Point Distribution, Team Height/Experience, and Miscellaneous Statistics. The aggregation of these tables contains over 160 statistics, both observed and derived. [14] [13] This selection was narrowed down to 10 variables for use in analysis:

- **TourneyTeam**: A binary variable that indicates whether or not the team in question was selected to participate in the tournament (1 = Tournament Team, 0 = Non-Tournament Team) [2]
- **WonConf**: A binary variable that indicates whether or not the team in question won their conference tournament (1 = Conference Champion, 0 = Non-Conference Champion) [11]
- **AdjEM**: Adjusted Efficiency Margin, the difference in the number of points a team scores per 100 possessions and the number of points that team allows per 100 possessions (Offensive Efficiency - Defensive Efficiency) [15]
- **Exp**: Team mean of players' college play, in years
- **AdjTempo**: Average Number of Possessions per 40-Minute Game
- **eFGPct_Off***: Offensive Efficient Field Goal Percentage (Weights 3-Point Shots more than 2-Point Field Goals)
- **TOPct_Off***: Offensive Turnover Percentage (Percent of Possessions that end with a Turnover)

- **ORPct_Off***: Offensive Rebound Percentage (Percent of Possessions that end with a Rebound)
- **FTPct_Off***: Offensive Free-Throw Percentage (Percent of successful free-throws)
- **FF_Weighted**: A weighted differential score of the Offensive Four Factor variables and the Defensive Four Factor Variables [12]

Another statistic was used from TeamRankings.com, which provided the ratings for each team's Strength of Schedule [21], which is determined by the aggregation of the ratings of each of the opposing teams a given team faces in their season, both conference and non-conference. The final variable used in the analysis is:

- **StrOfSched**: Rating of each team's Strength of Schedule in both conference and non-conference play, based primarily on Win Percentage of the Team's Opponents [22]

All of the data used in the analysis was collected from 2022-23 season, pre-tournament (no data collected during the NCAA Tournament period is used). There are 331 Division I Men's Basketball Teams whose data was used in the analysis.

2.2. Significance of Metrics

With the availability of 162 different potential variable, it's important to narrow down a concise list that are all relevant to the analysis. The variables used in this analysis were carefully selected on the basis of which factors will fundamentally lead to a winning team, and thus make them a potential candidate for an at-large bid to March Madness.

TourneyTeam will serve as the target variable, as we are trying to predict whether or not the team will secure a bracket spot (on the merit of an at-large bid). Since TourneyTeam is a binary variable, the resulting value of the analysis can be interpreted as the probability a team receives an at-large bid to participate in the tournament.

AdjEM was selected due to its prominence in evaluating teams, due to its fundamental ability to quantify the scoring performance of a team. It is also the method by which Kenpom.com ranks all Division I teams; the highest Adjusted Efficiency Margin preceding the NCAA tournament was held by the University of Connecticut, with an AdjEM of +29.86, effectively ranking UConn as the #1 Team, according to Kenpom. Following the result of the 2023 NCAA Tournament, this ranking method seems to be somewhat valid, as the UConn Huskies were the National Champions of the NCAA Tournament in 2023.

The variables **Exp** and **StrOfSched** serve more as control variables. An experienced team will likely have more players that have participated in high-level play with or on tournament teams, as well as players who are more prepared and conditioned to withstand the demand from the competition within the tournament. Teams with stronger schedules are also typically candidates for a bid, as these teams are more often than not members of power conferences with high-performing teams that are nationally ranked.

The remaining variables, **eFGPct_Off, TOPct_Off , ORPct_Off,** and **FTPct_Off** are each considered to be what are known as the **Four Factors**. The Four Factors was a simple and concise evaluation system for any basketball team, as determined by statistician and professional basketball coach Dean Oliver in 2002. It aims to evaluate a team by simply looking at the team's ability to "score, protect, crash, and attack", [8] since all of those factors are highly correlated with the ability to end a possession for a team. A weighted Four Factors metric (**FF_Weighted**) has also been incorporated into the analysis, which assigns different weights to the difference in offensive measures and defensive measures for each of the Four Factors, and combines them into one comprehensive score (see Section 2.3 for further information on the derivation of **FF_Weighted**).

2.3. Data Pre-Processing

The data was first filtered to only include teams that had not won their conference tournament. Since winning the conference tournament automatically guarantees a spot in the NCAA tournament, it would be weighted too heavily in the analysis, resulting in an underestimate of the effect of other variables. The resulting data set included 331 teams.

The next step was to ensure that the variables generally adhered to the same scale, as to not over- or under-estimate the effect of each variable. There were five variables that were fairly large relative to the remaining variables in the dataset: AdjTempo (maximum value of 73.45) and the Four Factors. Since each of the Four Factors can be interpreted as percentages, they were each scaled down by a factor of 100, bounding each variable between 0 and 1. As for AdjTempo, each observation was scaled down by a factor of 40, since Adjusted Tempo is the (Average) Number of Possessions per 40-Minute Game. Therefore, the rescaled AdjTempo can be interpreted as the (Average) Number of Possessions per Minute in a game.

The final adjustment to the data set was the calculation of the **FF_Weighted** variable. Dean Oliver's derivation implies that each of the Four Factors are not weighted equally. The weighting is as follows; Shooting (40%), Turnovers (25%), Rebounds (20%), and Free Throws (15%). [8] Therefore, in creating a comprehensive metric that properly weights the difference in the Offensive and Defensive measures of each of the Four Factors, the resulting equation that was calculated and used in the data set is:

$$\begin{aligned}\textbf{FF\_Weighted} =& 0.4 \times (eFGPctOff - eFGPct\_Def) + \\ & 0.25 \times (TOPct\_Off - TOPct\_Def) + \\ & 0.2 \times (ORPct\_Off - DRPct\_Def) + \\ & 0.15 \times (FTPct\_Off - FTPct\_Def)\end{aligned} \tag{1}$$

The difference in each of the Four Factors represents the net result of the team's ability to score, protect, crash, and attack.

CHAPTER 3

Methods and Models

3.1. Logistic Regression

Logistic Regression is a commonly used Machine Learning tool that uses one or more statistics (known as **features**) to predict the probability that an event occurs (known as a **binary target variable**) using observations from a set of data. Each feature in the regression equation corresponds to a numerical coefficient, which can be interpreted as the marginal effect of an additional unit of that feature on the binary target variable. [17] The Logistic Regression equation is defined as

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{(-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n))}} \tag{2}$$

where x_1, x_2,...,x_n are the variables in the data set, and β_0, β_1, β_2,...,β_n are the coefficients on the variables in the regression. [9] The Cost Function of Logistic Regression is defined as

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\beta(\mathbf{x}^{(i)})) + (1 - y^{(i)})\log(1 - h_\beta(\mathbf{x}^{(i)}))] \tag{3}$$

where m is the number of training examples in the sample, $\mathbf{x}^{(i)}$ is the feature observation, $y^{(i)}$ is the target observation, and h_β is the prediction function. The ultimate goal of Logistic Regression is to minimize the **Cost Functions**, or the measure of how well the model has performed on the given data, with respect to the parameters of interest used in the regression.

3.2. Cross-Validation

The Logistic Regression model was implemented using the Scikit-Learn Regression Packages, as available in Python. [19] The data was split randomly (by observation); 70% of the data was randomly selected to construct each of the models, and the remaining 30% was used to test the performance of the model's ability in predicting the observation's corresponding tournament participation value.

The use of Cross-Validation was also employed in the evaluation of the Logistic Regression model. Cross-Validation is accomplished by splitting the data into a pre-specified number of "folds", or equally-sized subsets of the data, where each fold will be used once as a testing split, and the remaining folds will be used to train the model. This is an effective method of testing the model, as it iterates the training process and fine-tunes the model for different properties and nuances that may exist within the data set that aren't necessarily accounted for in the original training and testing split. Since the data set is relatively small, 5 folds will be used in Cross-Validation, which is equivalent to an 80/20 training/testing split. [1]

3.3. Prioritization of Features Using Classification

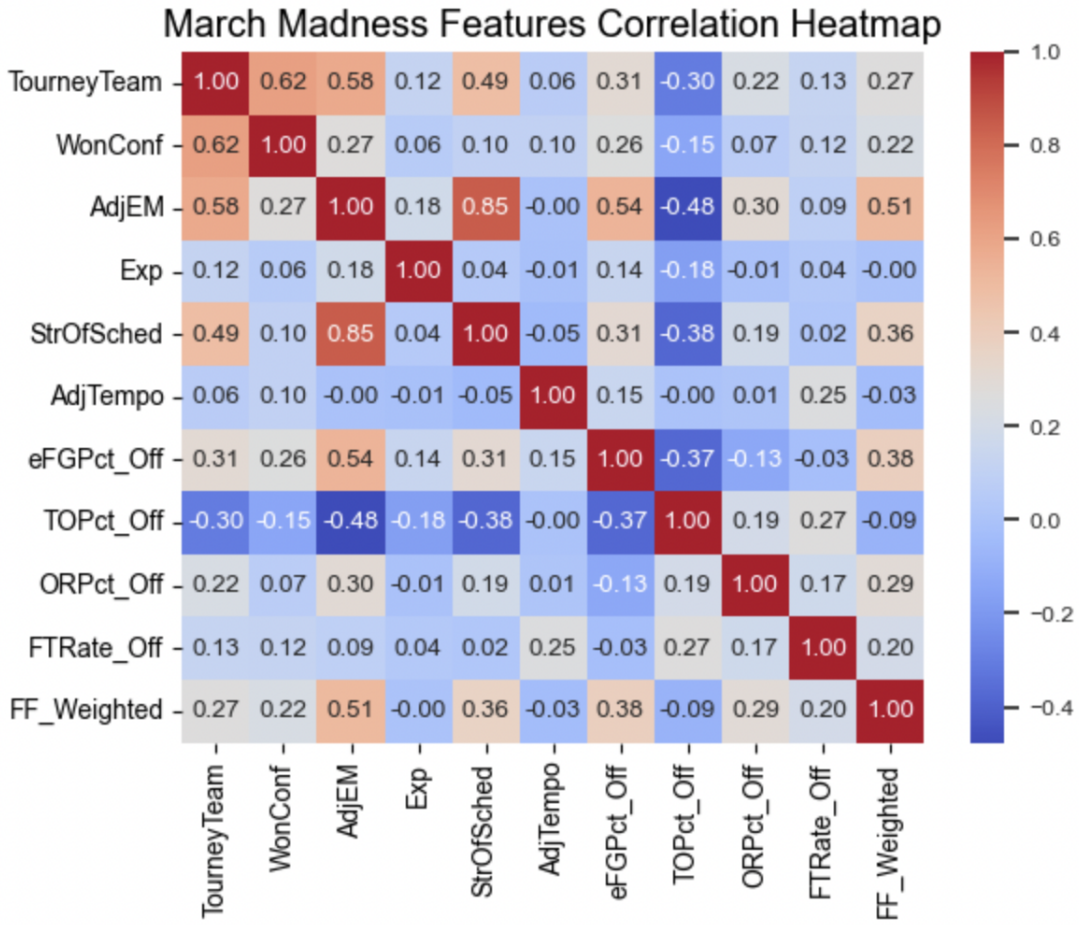
Another method that will be used to identify the most important features is from Classification Decision Trees. **ExtraTreesClassifier** and **RandomForestClassifier** are two Classification methods available in SKLearn that will be used to identify features that impact a team's probability of receiving an at-large bid most:

- **ExtraTreesClassifier**: Creates a randomly-generated decision tree from a pre-specified number of features; since the split factor of the tree is random, the resulting trees that are generated are generally diverse [3]. There are multiple trees that are constructed over the entire data set.
- **RandomForestClassifier**: Uses a greedy algorithm to construct decision trees; chooses the most optimal split for the data [4]. There are multiple trees that are constructed over bootstrapped subsets of the data.

Within these three models is the aforementioned `feature_importances_` attribute, which assigns a score to evaluate the role of each feature in the model and the effect it has on the model's ability to predict the target variable. The process of identifying the most important features is by fitting each of the Classifier models to the training data, then calling the `feature_importances_` attribute on each of the models to list out the weighted percentage of how much each feature contributes to accurately predicting the target variable. [18]

3.4. Feature Selection using RFE

In running the regressions, it is important to minimize any noise created by the multicollinearity of features. **Data multicollinearity** [23] occurs when two or more features are highly or perfectly correlated with each other. This could lead to issues within our models, since the coefficients on the features could be highly responsive to small changes within the data, resulting in a less accurate estimate of the true effect of the feature(s) on the target variable.



As visible from the heat map, most of the features do not seem to have a high correlation with each other. Some potentialy correlated features with relationships to keep note of are AdjEm ∼ StrOfSched, AdjEm ∼ eFGPct_Off, and AdjEM ∼ FF_Weighted, with R-Squared values of 0.85, 0.54, and 0.51, respectively. These relationships seem valid, as a team with a higher Adjusted Efficiency Margin will likely have a stronger schedule, better shooting, and be stronger overall in the four factors. In completing the analysis, however, these relationships will be noted.

In contrast, there are also some relationships between features that have little to no correlation. AdjEM ∼ AdjTempo, TOPct_Off ∼ AdjTempo, and Exp ∼ FF_Weighted have no relationship, each with an R-Squared values of 0.

Various dimension reduction packages available in Python, specifically those in Scikit-Learn, were considered for implementation in the analysis. Originally, Principle Component Analysis (PCA) had been the primary interest in dimension reduction, as it is typically the most effective reducing the data set to its most important components. However, the goal in this analysis was to return coefficients to interpret from the dataset, which cannot be properly interpreted following the implementation of PCA.

PCA uses linear algebra to construct a linear combination of the original data set that can efficiently and effectively be fit by a model, so the resulting coefficients cannot be interpreted as the direct effect of each feature on the respective team's participation in March Madness.

Instead, Recursive Feature Elimination (RFE) from SKLearn was employed. This is the process of selecting the most important features from a data set by recursively considering smaller and smaller subsets of the features. [5] The algorithm recursively loops through the less-significant features until a pre-specified number of the most relevant features are determined while trying to eliminate dependencies and collinearity. The function then ranks the features from most effective (1) to least effective (9, for the March Madness data set).

CHAPTER 4

Analysis

4.1. Exploratory Analysis

Prior to building the Logistic Model, an exploratory analysis was performed to gain some insight on the characteristics of the individual features. A histogram was constructed for each feature to examine their respective distributions, split into two groups (Tournament Teams and Non-Tournament Teams). A Two-Sample T-Test was then performed to determine if the difference of means for the two groups was statistically significant, indicating if there is a difference in the overall distribution of values for each feature between Tournament Teams and Non-Tournament teams (See *Figures Appendix* in Chapter 5). Table 1 Displays the difference in means between the Tournament and Non-Tournament Teams for each feature, as well as the resulting p-values for each T-Test. As the p-value crosses the three thresholds listed below in Table 1, the difference in means becomes more and more significant.

TABLE 1. Exploratory Analysis of Feature Distributions

	Mean		**T-Test Results**
Feature	Tourney Team	Non-Tourney Team	p-value
AdjEM	-3.2	13.8	<0.01***
Exp	1.88	2.02	0.025**
StrOfSched	-1.62	5.27	<0.01***
AdjTempo	1.68	1.69	0.221
eFGPct_Off	0.50	0.52	<0.01***
ORPct_Off	0.28	0.30	<0.01***
TOPct_Off	0.17	0.19	<0.01***
FTRate_Off	0.31	0.33	0.017**
FF_Weighted	-0.45	1.28	<0.01***

* = 10% Level of Significance, ** = 5% Level of Significance, *** = 1% Level of Significance

The results in Table 1 indicate the features that show the greatest differences in their means for Tournament Teams and Non-Tournament Teams are **Adjusted Efficiency Margin**, **Strength of Schedule**, and the **weighted Four Factors Score**. This analysis is simply exploratory, so it doesn't necessarily dictate the extent to which these features affect a team's probability of being selected in an at-large bid to the NCAA tournament. Therefore, the next step is to build the Logistic Model.

4.2. Naive Logistic Regression

The Naive Logistic Regression Model was originally constructed using all nine features. The resulting model, when tested on a random 30% split of the data (disjoint from the split of data that was used to train the model), scored **94%** in an evaluation of accuracy. Additionally, when tested on the entire data set, the model was able to accurately predict the outcomes of the target variable for **95%** of the observations. These accuracy scores were reflected in the resulting Cross Validation scores, as presented in Table 2.

TABLE 2. Cross Validation Scores for Naive Logistic Regression

Folds				
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
0.970	1.000	0.924	0.924	0.955

The highly scoring model provides confidence that the coefficients on each of the features will accurately reflect each feature's impact on the probability a team receives an at-large bid for the NCAA tournament. Despite the fact that the model is performing well, analysis using feature selection will continue to be implemented, not necessarily to improve the Naive Model, but more so to narrow down the selection of features with the greatest impact on the target variable.

4.3. Results Using Features Selection Methods

Two new models were constructed using the classifiers instead of Logistic Regression, in order to identify the classifier's interpretation of most important features. Table 3 presents the results from the classifier models:

TABLE 3. Testing Regressions Using Ranked Features from RFE

	Features									
Model	AdjEM	Exp	Sched	Tempo	eFG	TO	OR	FT	FF	Score
RandForest	0.05	0.02	**0.38**	**0.10**	0.04	0.03	**0.30**	0.06	0.03	93%
ExtraTrees	0.08	0.05	**0.24**	**0.11**	0.10	0.04	**0.24**	0.08	0.06	93%

Both the RandomForestClassifier and ExtraTreesClassifier scored over 90% in accuracy on the testing data split, indicating the models were mostly able to accurately predict the outcomes on the target variable. For both models, **StrOfSched**, **AdjTempo**, and **ORPct_Off** were selected as the most important features. The favoring of StrOfSched and ORPct_Off is in agreement with the results from Table 1 in Section 4.1, as these two features had a statistically significant difference-in-means between Tournament Teams and Non-Tournament Teams. AdjTempo, however, did

not have a statistically significant difference-in-means between Tournament Teams and Non-Tournament Teams, as its p-value did not cross any of the three thresholds for level of significance.

The results from RFE imply that the most relevant features to accurately predict whether or not a team receives an at-large bid to participate in the NCAA tournament are **Years of Experience (1)**, **Adjusted Efficiency Margin (2)**, **Adjusted Tempo (3)**, the **Weighted Four Factors Score (4)**, and **Free Throw Percentage (5)**. According to the Correlation Heatmap, each of the relationships between the five features show weak to no correlation, so there is little to no concern of multicollinearity.

TABLE 4. RFE Feature Importance Rankings

Features								
AdjEM	Exp	Sched	Tempo	eFG%	TO%	OR%	FT%	FF
2	1	7	3	9	8	6	5	4

The Logistic Regressions were then reconstructed using the ranked features from RFE. This was done by running the regression nine times, starting with the highest-ranked feature first (FTRate_Off), then consecutively adding the next best-ranked feature until all nine features were used. Table 5 presents the resulting scores of each model using this method.

TABLE 5. Testing Regressions Using Ranked Features from RFE

Features									
AdjEM	Exp	StrOfSched	Tempo	eFG	TO	OR	FT	FF	Model Score
	*								89%
*	*								**98%**
*	*		*						98%
*	*		*					*	98%
*	*		*				*	*	98%
*	*		*			*	*	*	98%
*	*	*	*			*	*	*	**94%**
*	*	*	*		*	*	*	*	94%
*	*	*	*	*	*	*	*	*	94%

From Table 5, there are evident jumps in the score at the separate points where **AdjEm** and **StrOfSched** are included in the regression. Incorporating AdjEM into the regression resulted in the accuracy of the model increasing by 11%. After the addition of AdjEM, the addition of the following features made little to no effect on the resulting model score. On the other hand, incorporating StrOfSched into the model the regression resulted in the accuracy of the model decreasing by 4%. Again, the addition of the following features made little to no effect on the resulting model score after the addition of StrOfSched. While StrOfSched had a negative impact on the model's predicting ability, both AdjEM and StrOfSched will be considered in continuing the analysis.

4.4. Final Logistic Regression Results

After conducting the exploratory analysis and determining the most important features from feature selection, the four features that were most prevalent showed to be **Adjusted Efficiency Margin**, **Strength of Schedule**, **Experience**, **Adjusted Tempo**, and the **Weighted Four Factors Score**. The resulting Logistic Regressions using these selected features scored **94%** when evaluating accuracy on the testing split, indicating the fit of the model is very strong. This score can be corroborated by the cross-validation scores, as presented in Table 6:

TABLE 6. Cross Validation Scores for Final Logistic Regression

Folds				
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
0.970	1.000	0.924	0.924	0.955

4.5. Interpretation of Coefficients

Using the results from the final Logistic Regression, the coefficients can be interpreted as the marginal effect of one additional unit of the feature in question on the probability the team of observation will receive an at-large bid to participate in the NCAA Tournament. The Logistic Regression equation will take the form:

$$\widehat{Pr(Bid)} = \widehat{\beta}_0 + \widehat{\beta}_1 AdjEM + \widehat{\beta}_2 StrOfSched + \widehat{\beta}_3 Exp + \widehat{\beta}_4 AdjTempo + \widehat{\beta}_5 FF \quad (4)$$

Where $\widehat{\beta}_0$ represents the estimated intercept of the regression equation, and $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, $\widehat{\beta}_4$, and $\widehat{\beta}_5$ represent the estimated coefficients on Adjusted Efficiency Margin, Strength of Schedule, Average Years of Experience, Adjusted Tempo and the Weighted Four Factors metric, respectively.

TABLE 7. Coefficients from Final Logistic Regression Model

	Features				
Intercept	AdjEM	StrOfSched	Exp	AdjTempo	FF_Weighted
-5.87	0.257	0.244	0.130	0.146	-0.228

Table 7 gives the resulting coefficients from the Logistic Regression equation. Of the five selected features, the three that most increase a team's probability of receiving an at-large bid are the Adjusted Efficiency Margin, Turnover Percentage, and Free-Throw Percentage. The value of 0.257 for the coefficient on AdjEM indicates that an additional point on a team's Adjusted Efficiency Margin is estimated to increase the probability they receive an at-large bid by **25.7%**. Similarly, the value of 0.244 for the coefficient on StrOfSched indicates that an additional percentage point on a team's Strength of Schedule score is estimated to increase the probability they receive an at-large bid by **24.4%**. Finally, the value of 0.130 for the coefficient on Exp indicates that an additional year on the team's mean years of college play experience is estimated to increase the probability they receive an at-large bid by **13.0%**, and the value of 0.146 for the coefficient on AdjTempo indicates that an additional possession on a team's average number of possessions per game is estimated to increase the probability they receive an at-large bid by **14.6%**.

The negative coefficient on the Weighted Four Factors score implies that an increase in the FF_Weighted score decreases a team's probability of receiving an at-large bid. While at first glance this seems unusual, as a team that is offensively better at shooting, foul shooting, rebounding, and preventing turnovers would likely be a team that's a better candidate for an at-large bid, there may be some confounding variables at play. For example, a team that is much better than the other teams in a lower-performing conference may be able to yield themselves a high FF_Weighted score, but may not size up to other teams in stronger conferences, and therefore would not receive a bid.

Equation (5) is the result of plugging the results from Table 6 into Equation (4):

$$\widehat{Pr(Bid)} = -2.90 + (0.231 \times AdjEM) + (-0.032 \times StrOfSched) + (-0.116 \times TOPct\_Off) + (0.243 \times FTRate\_Off) + (0.002 \times FF\_Weighted) \tag{5}$$

To validate the results from Equation (5):

Take the University of Connecticut, a team that did not win their conference tournament but was selected at-large to participate in the NCAA tournament in 2023 – and won! Their AdjEM was +25.37, StrOfSched was +11.1, Exp was 1.74 years,

AdjTempo was 66.7 possessions per game, and FF_Weighted was +5.34. The result plugging all of these measurements into Equation (5) is 2.62. The threshold for Logistic Regression in SKLearn is 0.5, so if the predicted value for the target variable is ≥ 0.5, the observation is classified as a binary variable of 1; in this analysis, that indicates that the team received an at-large bid, which is true for the University of Connecticut.

On the other hand, take Loyola Marymount University, a team that did not win their conference, and was not selected at-large to participate in the NCAA tournament in 2023. Their AdjEM was +6.32, StrOfSched was +2.5, Exp was 2.29 years, AdjTempo was 67.0, and FF_Weighted was -0.21. The result plugging all of these measurements into Equation (5) is -3.04, which is below the 0.5 threshold, so the model would predict that Loyola Marymount would not receive an at-large bid in 2023 – and they didn't.

4.6. Discussion

The data used in this analysis was exclusively from the 2022-23 season. This allowed the regression to highlight the features that most recently had an impact on a team's bid status, as opposed to building the model from previous years' data and attempting to generalize the results to this most recent season. A solution to the issues that would arise from attempting to build a model using data that spans over multiple season would be to implement a **panel regression** [6]. By doing so, the unobservable differences in variables that change from year to year would be controlled for, and the results from models built from previous years' data could more accurately predict valid results for future seasons. This would also allow for a larger sample size in the data set, which would allow for the analysis of different progressions over time among teams.

Another improvement that could be made to the analysis is by adjusting the model for effects caused by **regression discontinuity** [7]. Since winning a conference tournament guarantees a spot in the NCAA tournament, there is a discontinuity in regressing all features on the target variable. Winning a conference tournament acts as a binary cutoff in this case (WonConf = 1), where the non-conference-champions can then only secure a spot in the tournament through an at-large bid. In this analysis, the data was filtered to only include non-conference winners, due to this issue of regression discontinuity, however in future analysis the incorporation of these missing observations would strengthen the validity of the regression results, and can be done so by applying regression discontinuity remediation techniques.

CHAPTER 5

Conclusion

Of the 358 Division I Men's Basketball Programs, only 68 will participate in March Madness each year, with 32 programs being selected by virtue of winning their conference tournament. Some programs are notoriously frequent seed-holders, while other teams make a Cinderella run to have a shot at being the NCAA National Champions. Either way, there is no perfect science to predicting which teams will participate in March Madness, however analytical strategies such as Machine Learning can certainly shed some light on what factors matter most in such a determination.

The variety of methods used in this analysis led to the construction of a high-performing model that can accurately predict **94%** of target variables in the data set. Additionally, the greatest determinants of whether or not a team would receive an at-large bid to participate in the NCAA tournament, according to the Logistic Model, were **Adjusted Efficiency Margin**, **Strength of Schedule**, **Average Years of Team Experience**, **Adjusted Tempo**, and the **Weighted Four Factors Score**.

The conclusions drawn in this analysis can be used by college coaches, front office staff, and other relevant stakeholders to potentially increase the probability that their team is selected for an at-large bid, given that they are not the winners of their conference. Perhaps a coach may opt to recruit more experienced players, or speed up the tempo of their team's play, or even make schedule adjustments to increase the strength of their competition (granted, this may be difficult or not possible in conference play, as conferences are pre-determined, but it is much more feasible for pre-season or non-conference matches). Whatever avenue the stakeholders choose to pursue, it will nonetheless give their team a quantitatively-informed advantage to get closer to a chance at dancing in the postseason.

Figures Appendix

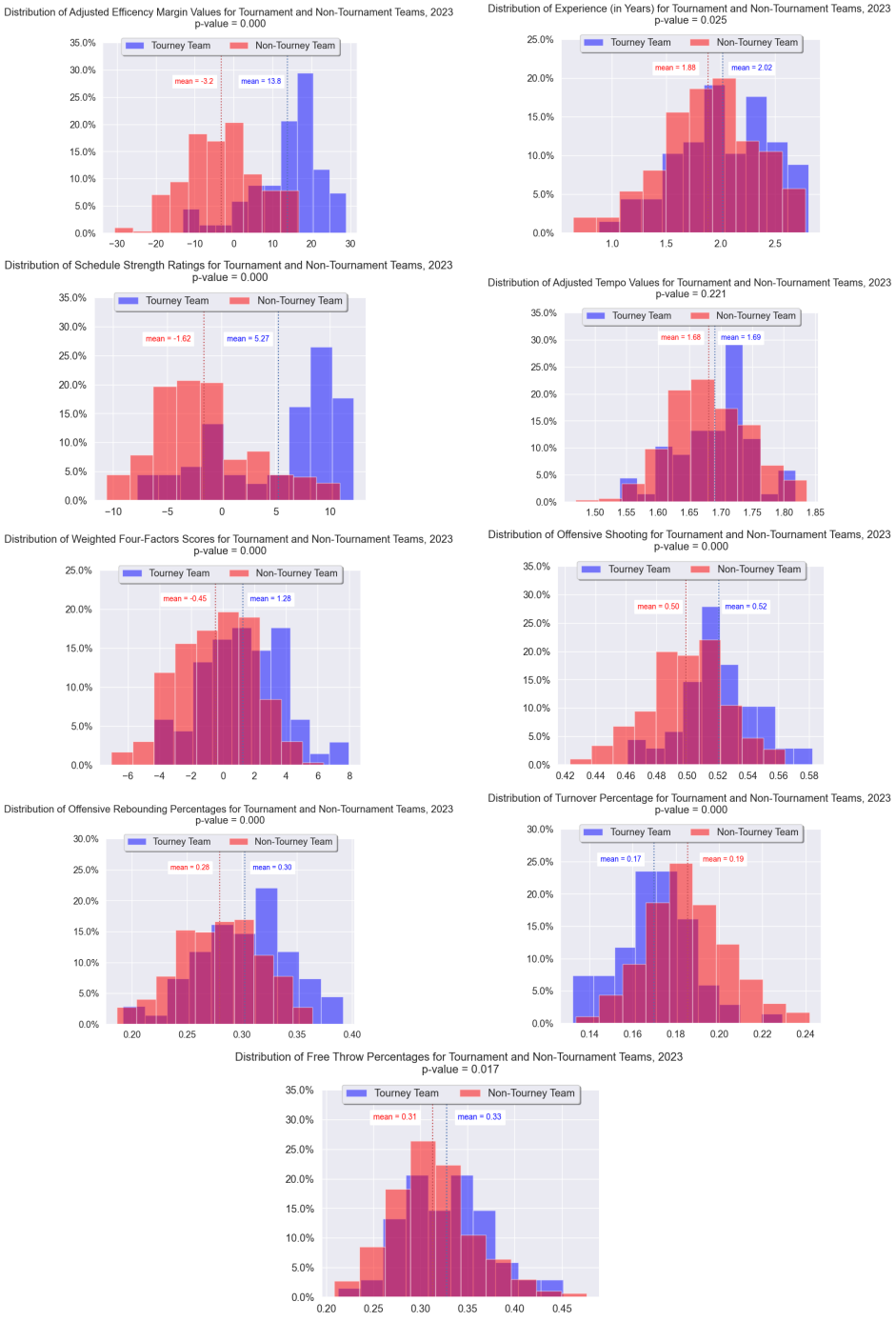


FIGURE 1. Distribution Plots of Data Features with T-Test Results

Bibliography

[1] Mohammed Alhamid. What is cross-validation? 2020.

[2] Kyle Boone. March madness 2023: Committee reveals official ncaa tournament bracket seed list from 1-68. 2023.

[3] Scikit-Learn Developers. sklearn.ensemble.extratreesclassifier.

[4] Scikit-Learn Developers. sklearn.ensemble.randomforestclassifier.

[5] Scikit-Learn Developers. sklearn.featureselection.rfe.

[6] Nick Huntington-Klein. *The Effect*. Chapman Hall, 2023.

[7] Nick Huntington-Klein. *The Effect*. Chapman Hall, 2023.

[8] Justin Jacobs. Introduction to oliver's four factors. 2017.

[9] Learn By Marketing. Logistic regression explained. 2023.

[10] NCAA. How the field of 68 division 1 men's teams is picked for march madness. 2023.

[11] NCAA. Tracking all 32 ncaa men's basketball conference tournaments, auto bids for 2023. 2023.

[12] Kenneth Pomeroy. Four factors. 2005.

[13] Kenneth Pomeroy. Ratings glossary. 2012.

[14] Kenneth Pomeroy. Stats explained. 2012.

[15] Kenneth Pomeroy. Ratings methodology update. 2016.

[16] Kenneth Pomeroy. 2023 pomeroy college basketball ratings. 2023.

[17] Abhibhav Sharma. Logistic regression explained from scratch (visually, mathematically and programmatically). 2021.

[18] SKLearn. Feature importances with a forest of trees. 2023.

[19] SKLearn. sklearn.linear_model.logisticregression. 2023.

[20] Devin Soni. Supervised vs. Unsupervised Learning: Understanding the differences between the two main types of machine learning methods. *Toward Data Science*, 2018.

[21] TeamRankings. Ncaa college basketball strength of schedule rankings ratings. 2023.

[22] Daniel Wilco. Explaining college basketball's strength of schedule. 2019.

[23] Songhao Wu. Multicollinearity in Regression: Why it is a problem? How to check and fix it. *Toward Data Science*, 2020.