**2021 NCAA MBB Tournament Team Prediction**

Langdon Hatton, [lhatton@bellarmine.edu](mailto:lhatton@bellarmine.edu)

**ABSTRACT**

I wanted to make some kind of predictive model that involved basketball. I found a dataset that a variety of team statistics from every division one team in the year 2021 season. I performed logistic regression on this college basketball dataset using python. The dataset I chose consists of 347 entries with 21 total columns, and it contains various types of variables. I will use this dataset to test a prediction based on the independent and dependent variables we choose. Included in the project will be visual representation of the data to show any interesting findings.

1. **INTRODUCTION**

For the final project, I will be using logistic regression to predict whether 2021 D1 men’s basketball teams made the NCAA tournament. I will also be changing the test cases to see how I can make my model the most accurate.

1. **BACKGROUND**
   1. *Data Set Description*

This dataset was collected during the 2020-2021 NCAA division 1 men’s basketball season. A variety of stats are presented for each team. There were 347 D1 teams in 2021, so there are 347 entries. There were no null values in the original dataset. I removed the team column, because it did not make sense to have the team's name in my model. In the original dataset instead of the NCAA\_T column, it was a column called SEED. If a team made the NCAA tournament, there was a seed, 1-16, entered. If they did not make the tournament, then N/A would be the entry. I changed this column. I made it so a one was entered for teams that made the tournament. For teams that did not make the tournament, I entered a zero. I did this in excel. I found this dataset on Kaggle. It can be found here - <https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset?select=cbb21.csv>

* 1. *Machine Learning Model*

Logistic regression is a machine learning model used for binary classification tasks, where the goal is to predict the probability of an event occurring. In this particular case, the goal is to predict whether a 2021 men's college basketball team made the NCAA tournament or not. The model is trained on a dataset that contains 347 entries and 21 columns, where each row represents a team and each column represents a different feature. The dataset has been preprocessed by dropping some unnecessary columns, handling missing values, and creating independent and dependent variables. The independent variables are the features that will be used to make predictions, such as the team's conference, the number of games played, the team's offensive and defensive efficiency, and more. The dependent variable is a binary variable that indicates whether the team made the NCAA tournament or not.

Logistic regression works by estimating the probability of an event occurring using a logistic function, which maps any real-valued input to a value between 0 and 1. In this case, the input is a weighted sum of the independent variables, where each variable is assigned a weight or coefficient that represents its importance. The logistic function is then applied to the input, and the resulting output is interpreted as the predicted probability of a team making the NCAA tournament. To train the model, the weights are initialized randomly, and the model is fitted to the training data using an optimization algorithm that minimizes the error between the predicted probabilities and the actual outcomes. Once the model is trained, it can be used to make predictions on new, unseen data. The accuracy of the model can be evaluated using various performance metrics, such as accuracy, precision, recall, and F1-score.

1. **EXPLORATORY ANALYSIS**

Our dataset has 347 entries. It contains 21 columns. There are no missing values in the dataset.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| CONF | Object |
| G | int64 |
| W | int64 |
| ADJOE | float64 |
| ADJDE | float64 |
| BARTHAG | float64 |
| EFG\_O | float64 |
| EFG\_D | float64 |
| TOR | float64 |
| TORD | float64 |
| ORB | float64 |
| DRB | float64 |
| FTR | float64 |
| FTRD | float64 |
| 2P\_O | float64 |
| 2P\_D | float64 |
| 3P\_O | float64 |
| 3P\_D | float64 |
| ADJ\_T | float64 |
| WAB | float64 |
| NCAA\_T | int64 |

**Here is a description of what each variable means**

G: Number of games played

W: Number of games won

ADJOE: Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense)

ADJDE: Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division I offense)

BARTHAG: Power Rating (Chance of beating an average Division I team)

EFG\_O: Effective Field Goal Percentage Shot

EFG\_D: Effective Field Goal Percentage Allowed

TOR: Turnover Percentage Allowed (Turnover Rate)

TORD: Turnover Percentage Committed (Steal Rate)

ORB: Offensive Rebound Rate

DRB: Offensive Rebound Rate Allowed

FTR : Free Throw Rate (How often the given team shoots Free Throws)

FTRD: Free Throw Rate Allowed

2P\_O: Two-Point Shooting Percentage

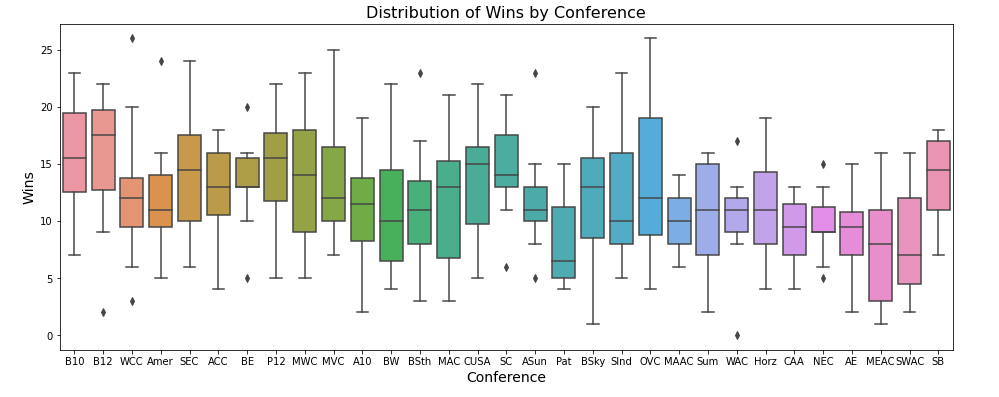
2P\_D: Two-Point Shooting Percentage Allowed

3P\_O: Three-Point Shooting Percentage

3P\_D: Three-Point Shooting Percentage Allowed

ADJ\_T: Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)

NCAA\_T: 1 indicates that the team made them NCAA tournament, 0 indicated that the team did not make the NCAA tournament.

*I figured I would make some charts to visualize some of the data.*  


*Figure 1: Box plot of the distribution of wins by conference*

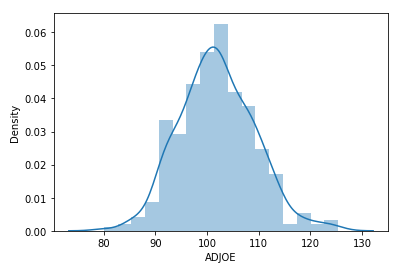


Figure 2: Dist Plot of adjusted Offensive Efficiency

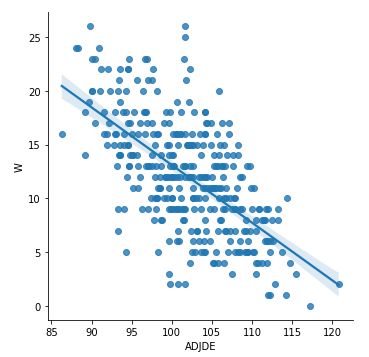


Figure 3: LM Plot of Adjusted Defensive Efficency vs. Wins

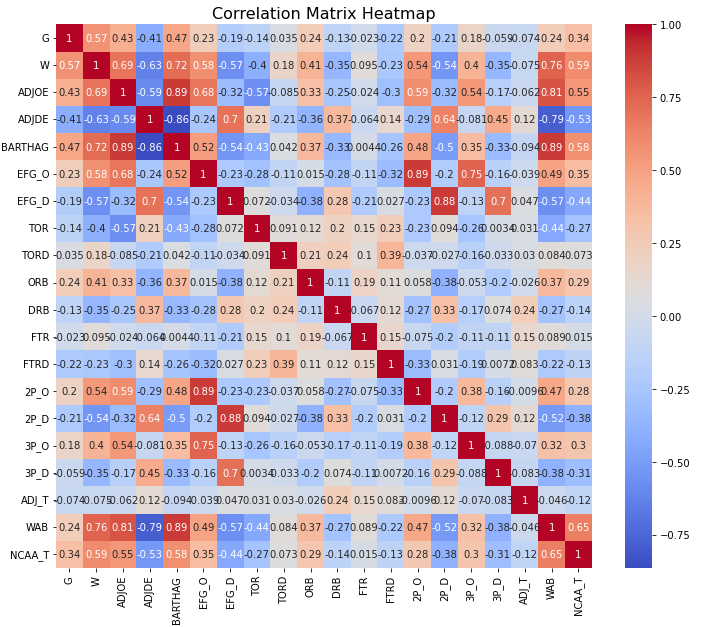


Figure 4: Heatmap for every column

1. **METHODS**
   1. *Data Preparation*

First, I explored the dataset. I saw that there were no missing values. I did go into Excel and change the SEED column to NCAA\_T, like I mentioned above. After that, I was ready to split the independent and dependent variables. Then, I got dummies for the columns that were not numerical.

* 1. *Experimental Design*

Table 2: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | Done in python with the normalized dataset. The only normalization was removing the team name column and changing seed column to just if they made the tournament or not. The test size was 30%. |
| 2 | Same as experiment 1, but the test size was 20%. |
| 3 | Same as above experiments, except I also removed the ADJ\_T column and set the test size back to 30% (I removed this column because it had basically no correlation to making the tournament) |
| 4 | Same as experiment three except I changed the test size to 20% |
| 5 | Removed the team, W, and G columns. Test size was at 30%. With this model, you could run it on teams that have not finished their seasons yet. |

* 1. *Tools Used*

The following tools were used for this analysis: Python running the Anaconda environment for windows was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas, NumPy, Matplotlib, Seaborn, and SKLearn. I chose python because I find it easier than R, and I just think it is easier to work in. I used Pandas to manipulate my dataset. SKLearn was used for logistic regression directly. The rest of the libraries were used for visualization.

1. **RESULTS**
   1. *Classification Measures*

Experiment #1

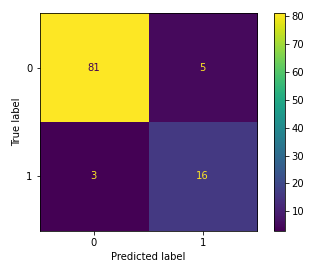


FIGURE 5: confusion matrix for experiment 1

precision recall f1-score support  
  
 0 0.96 0.94 0.95 86  
 1 0.76 0.84 0.80 19  
  
 accuracy 0.92 105

macro avg 0.86 0.89 0.88 105  
 weighted avg 0.93 0.92 0.93 105  
  
Experiment #2

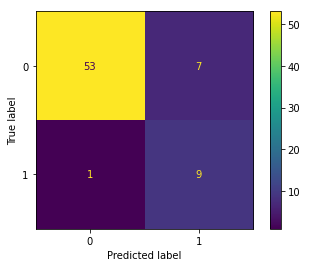


FIGURE 6: confusion matrix for experiment 2

precision recall f1-score support  
  
 0 0.98 0.88 0.93 60  
 1 0.56 0.90 0.69 10  
  
 accuracy 0.89 70  
 macro avg 0.77 0.89 0.81 70  
weighted avg 0.92 0.89 0.90 70

Experiment #3

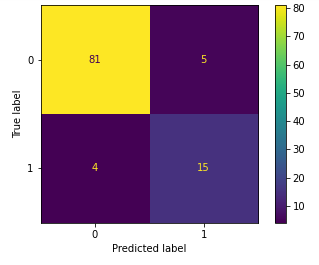


FIGURE 7: confusion matrix for experiment 3

precision recall f1-score support  
  
 0 0.95 0.94 0.95 86  
 1 0.75 0.79 0.77 19  
  
 accuracy 0.91 105  
 macro avg 0.85 0.87 0.86 105  
 weighted avg 0.92 0.91 0.92 105

Experiment #4

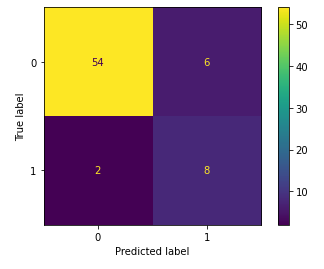


FIGURE 8: confusion matrix for experiment 4

precision recall f1-score support  
  
 0 0.96 0.90 0.93 60  
 1 0.57 0.80 0.67 10  
  
 accuracy 0.89 70  
 macro avg 0.77 0.85 0.80 70  
 weighted avg 0.91 0.89 0.89 70

Experiment #5

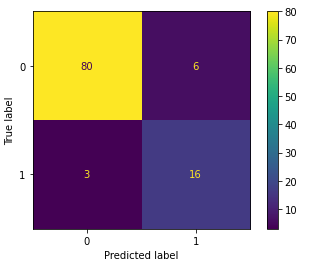


FIGURE 9: confusion matrix for experiment 5

precision recall f1-score support  
  
 0 0.96 0.93 0.95 86  
 1 0.73 0.84 0.78 19  
  
 accuracy 0.91 105  
 macro avg 0.85 0.89 0.86 105  
 weighted avg 0.92 0.91 0.92 105

* 1. *Discussion of Results*

You can see that Experiment #1, experiment #3, and Experiment #5 have relatively higher accuracy scores of 0.92 and 0.91, respectively, indicating that the model's overall performance is good for these experiments. However, we also observe that the F1-score for the minority class (class 1) is relatively lower than the majority class (class 0) for the experiments, indicating that the model is not performing as well in identifying the positive cases. The models did do very well at predicting case 0, but it did not do as well at predicting case 1. This was probably due to having data that had more case 0’s than 1’s. Like mentioned, this was unavoidable due to the nature of college basketball.

* 1. *Problems Encountered*

I had a hard time just finding a dataset. Datasets with large amounts of data about college basketball are hard to find, so that was the hardest part. After that, I was not sure what I could predict regarding basketball. Once I figured that out, the project went smoothly.

* 1. *Limitations of Implementation*

The main limitation I have is the lack of size. There are only 347 teams in D1 basketball, so there was no way for me to have more entries. If possible, a larger dataset would have been great. I do think logistic regression worked great on the dataset, as the results were pretty decent.

1. **CONCLUSION**

In this project, I used logistic regression to predict whether 2021 D1 men's basketball teams made the NCAA tournament. The dataset contained a variety of team statistics for all 347 D1 teams in the 2021 season. I preprocessed the dataset by dropping some unnecessary columns, handling missing values, and creating independent and dependent variables. I trained a logistic regression model on this preprocessed dataset and evaluated its performance using various metrics. All models achieved around 90% accuracy on the test data, which indicates that it is a reasonably accurate predictor of the NCAA tournament qualification.

In addition, I conducted exploratory data analysis to understand the distribution of variables and their relationship with the dependent variable. I found that variables such as ADJOE, BARTHAG, and WAB had a strong positive correlation with the NCAA tournament qualification, while variables such as ADJDE and TORD had a strong negative correlation.

Overall, this project demonstrates the utility of logistic regression for binary classification tasks in sports analytics. The predictive model I developed can be used to identify teams that are likely to make the NCAA tournament and inform bracket predictions. The methods of this project can also be extended to other sports and other binary classification tasks.

.