Predicting The Madness of the College Basketball National Tournament

By: Daniel Frankini

**1. Introduction-**

The goal of my project was to predict the winner of the NCAA 2022 Tournament through the use of deep learning techniques and neural network modeling. The champion of college basketball is decided by a massive 68-team tournament known as March Madness. Though the tournament is viewed by millions of people each year no one has ever been able to correctly predict every game of the tournament. I intended to use the untapped potential of artificial intelligence (AI) to improve my comprehension and predictions of the competition.

**2. Data –**

Predicting College basketball involves using a massive amount of data and multiple datasets. All of these were taken from the Kaggle competition Machine Learning Mania 2022 March Contest. The goal of these datasets is to include all possible information about college basketball regardless of how obscure. The statistics in the datasets range from record to game location to the head coaches of the teams. While preparing the data for training I organized the dataset to show important features such as win ratio, win gap average, seed number, loss ratio, and loss gap average; an example of this can be seen below. After preparing the dataset the model can provide insights into the correlation between past team characteristics and their tournament performance

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **index** | **Year** | **Day**  **Num** | **TeamIdA** | **ScoreA** | **TeamIdB** | **ScoreB** | **Round** | **SeedA** | **SeedB** | **WinRatioA** | **GapAvgA** | **WinRatioB** | **GapAvgB** | **SeedDiff** | **WinRatioDiff** | **GapAvgDiff** |
| **1176** | 2021 | 148 | 1425 | 66 | 1211 | 85 | 4 | 6 | 1 | 0.7586206897 | 9.655172414 | 1 | 23 | 5 | -0.2413793103 | -13.34482759 |
| **1177** | 2021 | 148 | 1276 | 49 | 1417 | 51 | 4 | 1 | 11 | 0.8333333333 | 10.875 | 0.6538461538 | 4.346153846 | -10 | 0.1794871795 | 6.528846154 |
| **1178** | 2021 | 152 | 1222 | 59 | 1124 | 78 | 0 | 2 | 1 | 0.8846153846 | 18 | 0.9166666667 | 17.95833333 | 1 | -0.03205128205 | 0.04166666667 |
| **1179** | 2021 | 152 | 1417 | 90 | 1211 | 93 | 0 | 11 | 1 | 0.6538461538 | 4.346153846 | 1 | 23 | 10 | -0.3461538462 | -18.65384615 |

Figure 1: Model of Organized Dataset

**3. Modeling-**

The model approach I decided to use was a Recurrent Neural Network (RNN) with a Long Short-Term Memory. This approach can identify patterns and temporal dependencies in the data. To create the model I used several important features from the structured dataset; taking into account if each feature might affect a team’s performance. Three main features used were a team win-loss ratio, the gap averaging in scoring difference, and a team seeding. These three features provide a large amount of insight into how good a team is but they weren’t the only statistics used in the model. Along with these statistics, all the teams’ previous performances were incorporated into the model.

**4. Results & Conclusion-**

A graph of blue and orange lines

Description automatically generatedFigure 2- Model Accuracy and Loss

The model was able to predict a respectable number of games correctly. The average test accuracy score was 53%, which out of context seems to be a relatively low amount. However, because of the volatile nature of college basketball and its unpredictability, it is hard to achieve a higher prediction accuracy. Overall, over multiple seasons, the model was able to predict more than half of the games correctly, which is impressive because each season of college basketball involves numerous new factors each year, which makes each year unique when compared to others. When compared to other types of models the RNN/LTSM architecture proved to be just as efficient with other model’s accuracy on Kaggle ranging from 40-57%. Showing that neural networks and deep learning techniques can provide insight into the volatile nature of college basketball. Despite the benefits, it must be noted these models have limitations because of the impossibility of fully predicting college basketball’s nature since never-before-seen events could happen at any moment. Finally, even though these models provide valuable insights, the dynamic nature of the sport makes predicting a challenging problem to conquer.