Ben Oberg & Michael De Simone

Professor Aguiar

CSC542

5th May 2024

College Basketball Predictions

Introduction:

The purpose of our project was to create models that can predict regular season win totals and what variables most affect regular season wins in men's division I college basketball. We decided to choose this subject because of the rise of sports gambling in the past couple of years and especially with the recent legalization in the state of Florida. Within our project we had two main goals. The first was obviously to create successful models that could predict winning. Secondly, we wanted to understand what statistics most affected winning within college basketball. Initially, we were gonna look at regular season and postseason success. This would mean we would have classification and regression problems. However, we decided to focus on regular season success so we only had a regression problem because we were trying to predict the number of wins. There has been a lot of previous work in this field so we had a lot of previous research papers to look at.

Previous Works:

We looked at a few other research papers before beginning our project. [1] This group of researchers found that linear regression had the best success when it comes to predicting winners. The best model they had finished with an RMSE of .6. They built their model in multiple increments. First just the team, then the conference, then all of college basketball. [2]

Analyzed head to head matchups between teams from 2008 to 2013 and tried to determine who would win. . They took away two key points, with the first being that the variables selected are far more important than the models used. They also realized there is a limit to each model they created just due to the randomness of college basketball. [3] Utilized 5 different models to predict winning: Neural Networks(67%), SVM(65%), KNN(63%), Logistic Regression(63%), and RFs(61%).

Dataset:

We used a college basketball dataset from Kaggle that contains both regular season and postseason statistics from the 2013 season through the 2023 season of men's division I college basketball. 2020 was not included because the season was cut short. While the number of teams in division I has varied throughout the years there has been around 350 for the last 10 years. So we have around 3500 instances of data to work with. Within the dataset there are 23 variables varying from team name and conference name to advanced stats such as adjusted offensive efficiency to BARTHAG(the chance of beating the average division I team) to simple stats such as 2 point percentage and 2 point percentage allowed. Clearly this dataset gave us a lot to work with.

Methodology:

We began our project by preprocessing our dataset. We removed simple columns to start such as the team and conference names. So right away we removed 2 identifying variables as they were not needed for our project. Next we looked at variables that would be too direct for predicting the winner. Initially we removed 4 variables: wins, tournament seed, tournament performance, and WAB(wins above tournament bubble) as they were too strong. This then left us with 17 variables to determine what most affects winning from the original 23 in our dataset.

Within this we had our target variable (wins) and 16 other predictor variables ranging all over the place.

After this we looked at the correlation of the predictor variables with the target variable. The four most correlated variables with winning were adjusted offensive efficiency, BARTHAG, effective field goal percentage, and 2 point percentage. After this we then continued with feature selection before the initial training of our models. We used three methods: Random Forest, Correlation with a cutoff of 70% and low variance. Using Random Forest the five most important variables were adjusted offensive and defensive efficiency, BARTHAG, and offensive and defensive efficiency. For correlation the top 4 adjusted offensive and defensive efficiency, BARTHAG, and offensive efficiency. There were no variables with low enough variance to be removed. This is not surprising because of the randomness of college basketball. After this we moved on to trying to create our models. We used Linear Regression, XGBoost, and Random Forests as our models. However when using XGBoost we found little success so we decided to scrap using XGBoost and focus on the other 2 models. We also decided to use cross validation afterwards in order to attempt to try and improve our models. We used a fairly simple 10 fold cross validation in order to achieve this.

Results:

For Random Forest we ran it under multiple circumstances. We ran it using feature selection and cross validation and without using them. When using feature selection and cross validation we received an RMSE of 3.42. Looking at the importance scores the most important variables were Adjusted Offensive & Defensive Efficiency, BARTHAG, Effective Offensive & Defensive FG %. Without using feature selection and cross validation our models actually improved. We received an RMSE of 3.06, .36 better than before. We then used Linear Regression

too. We ran Linear Regression with the five variables from our correlation feature selection. With this set up we achieved an RMSE of 3.575 We then used cross validation which worsened our results to 3.7. With no feature selection or cross validation we received an RMSE of 2.83.

We were not surprised about acquiring better results with all the features present for a few reasons. 16 features and 3500 instances were not enough to create overfitting within these models that you might see with much larger datasets. Also in college basketball every statistic is important so we were not surprised that our models worsened when some were removed.

We were still pleased that we received a model with an RMSE of 2.83. We also were glad about the variables that consistently showed up as most important. We had BARTHAG, Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Effective FG percentage for, and Effective FG percentage against. The 2 point percentage also showed up sometimes too. Seeing these variables confirmed to us that our models were working correctly as it makes sense that these are the most important variables. We were surprised to see 2 point percentage as college basketball and basketball in general has shifted to shooting more threes in recent years. In general we were pleased with both the models we created and the variables we found to be most important.

Future Works:

We plan on testing our models on the 2024 season dataset once it is released. We will see how our models perform as the game changes and evolves each year. We also plan to look at our models again at the halfway point of next season in order to try and predict who will be the best teams for the rest of the season. We will also analyze the variables again and see if the most important ones have changed.

Works Cited

[1] Mente, S. N. (2022). Accuracy of machine learning algorithms in predicting college basketball games. Bachelor's thesis, Department of Computer Science, University of Virginia. https://libraetd.lib.virginia.edu/downloads/sn009z84g?filename=Mente_Sindhura_Accuracy_of_Machine_Learning_Algorithms_in_Predicting_College_Basketball_Games.pdf

[2]Zimmermann, A., Shi, Z., & Moorthy, S. (2013). Predicting college basketball match outcomes using machine learning techniques: Some results and lessons learned. Research Gate https://arxiv.org/abs/[arXiv_ID]

[3]Kim, J. W., Magnusen, M., & Jeong, S. (2023). March Madness prediction: Different machine learning approaches with non-box score statistics. *Journal of Multivariate Analysis*, https://doi.org/10.1002/mde.3814