Classification Project Report

The goal of this project was to analyze NCAA basketball team performance data and develop models to predict postseason seed groupings. By exploring key efficiency metrics such as Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, and Effective Field Goal Percentage, we aimed to uncover patterns that differentiate higher-seeded teams from lower-seeded ones. Using both statistical and machine learning approaches—including multinomial logistic regression and decision trees—we evaluated how well these models could classify teams into seed categories. Along the way, we assessed model performance using accuracy, confusion matrices, and ROC curves, and visualized results to better understand the strengths and limitations of each method.

We implemented decision trees to classify post-season seed groups. Using rpart, the classification tree was trained on a 70/30 training-test split. We used Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Effective Field Goal Percentage, Turnover Percentage, Offensive Rebounding Percentage, Free Throw Rate, and Seed Class in our trees to create splits. NA values were also removed so that this process would work properly. The top eight seeds were ranked into a “Low” category for a low number, and the last eight seeds were placed into a “High” category that corresponds to the number not ranking. After examining the complexity parameter plot, we pruned the tree ot the optimal cp to prevent overfitting. Our pruned tree identified Adjusted Offensive Efficiency as the root split, followed by splits in Adjusted Defensive Efficiency, Effective Field Goal Percentage, and Free Throw Rate. The pruned tree achieved an overall accuracy of approximately 68% on the test set.

We also used a Multiple Logistic Model to predict groups again. The original seed values were divided into four even groups: 1-4,5-8,9-12, and 13-16. The predictors used were Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, and RankeFGPct, each standardized to have zero mean and unit variance. A multinomial logistic regression model was built to predict which seed tier a team would fall into. This approach allowed us to estimate separate equations for each seed group. Looking at the results, we found that a one-standard-deviation increase in Adjusted Offensive Efficiency increased the chances of a team landing in the top-four seed group by about 1.2 units, assuming all other variables stayed the same. Our MLR model was had better predicting potential by about 2% than a Simple Logistic Model.

The KNN model that we made classified teams based on their seed group. We tested a number of different K values from 1-15, and we found that our model did the best when K was 13 or 14. We also made a plot of error rates by K size, and it showed us how much the difference varied, with it pretty much decreasing error until a plateau. At the same time, a simple logistic regression was completed with only Adjusted Offensive Efficiency as the predictor to forecast whether a team was given a Low seed. The logistic curve, plotted against a scatter of data points, had a clear trend: as offensive efficiency increased, so did the chance of receiving a

Low seed. The regression slope coefficient was exponentiated to yield an odds ratio, indicating the effect of a one-unit change in offensive efficiency on the chance of belonging to the Low seed category. Model predictions were made on a test dataset both in log-odds and probability space. KNN model offered a framework to predict real seed groupings from multivariate information and logistic regression, providing more understandable analysis focused around the individual keyvariable.Cleaning actions such as the removal of null values and input standardization were an important part of effective modeling.

Bagging was used to improve the accuracy for the classification SeedClass, a categorical variable indicating tournament seed status. The two kinds of bagging implementation we used was the ipred package and caret package. The steps were involved with the bagging process. Ipred bagging was used to train 150 decision trees using samples from the train_data. Minimal constraints were used to allow each tree to fit its subset of data. After, predictions were made and results were seen as a confusion matrix and accuracy score. Based on the results, 383 were truly classified for “low” seeds and 100 was misclassified. For the true “high” seed, 122 was classified correctly and 161 was misclassified. Overall, the accuracy was around 64.5% showing a moderate performance for predicting. Next we used the caret package with “treebag” and cross validation of 10 fold. It helped improve the model in terms of validity and see which variable is important. From the results, AJOE and ADOE were the most influential predictors, showing scores closest to 100. Following that; ORPct, eFGPct, and TOPct had moderate importance, and FTRate with the least contribution. This ranking provides key insights: models rely most heavily on overall efficiency metrics when predicting seed class, while secondary factors like rebounding and shooting efficiency add marginal predictive value. In summary, the bagging approach improved model stability and accuracy over a single decision tree. The use of both out-of-bag error and cross-validation helped validate the model effectively, and the analysis of variable importance gave valuable interpretability to the results.

In this project, we used different models to figure out what helps NCAA basketball teams get better seeds in the postseason. We focused on stats like Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, and Effective Field Goal Percentage. These stats helped us see which teams were more likely to get top-four seeds.

Each method gave us something useful. The decision tree showed that offensive efficiency was the most important factor and had an accuracy of 68%. The multinomial logistic regression gave better results than the simple logistic model and confirmed that better offense increases the chance of a higher seed. The K-Nearest Neighbors model helped us group similar teams, and the best results came when K was 13 or 14. Bagging improved accuracy and confirmed that offense and defense were the most important stats.

Overall, using several models helped us better understand how teams are ranked. While none of the models were perfect, they all pointed to the same idea: teams with strong offense and defense usually get better seeds. These findings could help with future predictions in college basketball.