Dylan Buchheim
R11524739

CS 4331 Data Mining Project II Model Deployment Report

Setup Phase

Data Partitioning:
I began the setup phase by partitioning the cleaned dataset, *PlayerData*, into two sets, a training set and a test set. I configured the partition so that roughly 75% of the data went into the training set and 25% of the data went into the test set.

Cross Validation:
After partitioning the dataset, I ran a couple of statistical tests to validate that the partition was truly random. First, I ran a t-test for the difference in means on three key numeric attributes. These three numeric values were *TRB* (Total Rebounds), *AST* (Assists), and *PTS* (Points). Then I ran a two sample z-test on my target variable *MVP*. None of the tests provided sufficient evidence that there was a significant difference between the means or proportions of these attributes across the partitioned datasets.

Rebalancing the Data:
Once I had validated the partition of my dataset, I set out to balance my training set to produce a better proportion of "yes" values for my target variable, *MVP*. I decided to increase the percentage of "yes" values to 20% in my training set. So I resampled the appropriate number of rows to create this proportion and appended the newly sampled rows to the end of my training set.

Establishing Baseline Performance:
Due to the fact that my model will be performing binary classification, I decided to use the all negative model to evaluate my models. Since the vast majority of the *MVP* values in my dataset are "no", the all negative model will produce a very high accuracy. Applying the all negative model to my test set yields a 99% accuracy. However, the all negative model generates a 0 for sensitivity because it does not classify any entries as positive. These are the two main metrics I will use to evaluate my models.

Modeling Phase

I decided to use the decision tree model, C5.0, in order to accurately predict my target variable, *MVP*. I built several different C5.0 models with various combinations of predictor variables being input into the algorithm (*Note: I removed many of the worse performing models from my R script in order to significantly reduce the length of my R script submission*). The three models I left in my R script produced some of the highest levels of accuracy. Ultimately, only the third model reached the level of baseline accuracy while also providing a high level of sensitivity.

In the beginning of the modeling process, I was factoring minimal player information into my models and then assessing model accuracy. Slowly, I began adding other player information that I believed to have a high predictive value into my model, all while removing data that resulted in worse performance. Eventually, after analyzing data distributions once again, I realized that the importance of some player stats have changed as the game of basketball has progressed through different “eras”. In an effort to capture the “era” that a player belonged to I added a new attribute *Year_Binned* which classified each entry into the decade it came from. After adding *Year_Binned* to the third model it reached the highest level of accuracy and sensitivity I was able to achieve.

Evaluation Phase

After fine tuning my classification models I finished with a model that can classify a player’s season as “MVP worthy” with 99% accuracy and a sensitivity level of 87%. This model met my baseline accuracy while also providing a high level of sensitivity, which the baseline model did not. These models also helped to accomplish the other objective I set for this project as they highlight the statistics that are most influential in determining the NBA league MVP.

The only shortcoming of my final model would be the precision value of 38%. I believe this is due to the fact that every season there is only one MVP award given out and many times the runner up has extremely similar stats to the winner. Consequently, my model is classifying many second place finishers as “MVP worthy”. After all, the MVP award is granted based on votes from NBA fans, so I believe that player notoriety is often what separates the winner from the runner-up. Therefore, I see no way of significantly improving model accuracy or precision beyond this point based on player stats alone.

Cost Evaluation:

The problem I set for my project was an exploration of the capabilities of data mining and was not economic in nature. Consequently, there are no costs associated with classifications my model makes.