Dylan Buchheim
R11524739

CS 4331 Data Mining Project I

Problem Understanding Phase

Objectives for this project :

- Learn about the characteristics and player statistics associated with the winners of the NBA MVP award in order to determine the defining factor(s) for winning the award.
- Develop models that can predict who will win the award based on a few important statistics.

Translating the Objectives:

In order to learn about the characteristics of MVP winners and determine the most important factors for winning the award I will…

- Perform graphical analysis of key player statistics and overlay those graphs with the statistics of past award winners to discern patterns between those who have won the award and those who have not.

In order to predict future winners of the MVP award I will…

- Develop a series of classification models which will identify the most likely candidate for the MVP award using key player statistics.
- Relevant models include:
    - Decision Trees
    - Naïve Bayes Classification
    - Neural Networks

Data Preparation Phase

During the data preparation phase I took several steps to clean the dataset and augment it with the appropriate information so that it was ready to be modeled in a later phase.

Data Preparation Tasks:

- I removed all rows from the data set that came from a year where there was no MVP Award.
- I removed the asterisks that were appended to some of the player's names. These asterisks were throwing off my ability to subset the data by player's names.
- I removed rows in the dataset that were completely null and contained no information.
- I removed the two attributes in the dataset that were completely null and contained no information across all rows.
- I re-indexed the trimmed dataset.
- I added a new attribute to the dataset, “MVP”, and defaulted all the values to “no”. This attribute is the target variable that will keep track of whether or not a player in the dataset won the MVP award that year.

- Using another dataset that contained each season’s MVP winner, I changed the appropriate “MVP” values to “yes” in my original dataset.
- I re-expressed the “Pos” attribute, a categorical variable that tracks the player’s basketball position, as a numeric variable for ease of use within graphs and models.
- I created standardized versions of the “G” (games played) and “MP” (minutes played) attributes to help find player stat blocks that were outliers.
- I removed players from the dataset who played in very few games because many of their stats were extreme outliers due to having a small sample size for their stats.
- Finally, I trimmed out the attributes from the dataset that were not useful to my model or were redundant with another attribute.

A Note on Outliers:
After removing players with very little games played, there are still some statistical outliers amongst many of the stats. However, for my purposes of trying to identify the league MVP based on a handful of these stats, these outliers are important data. Naturally, the winner of the “Most Valuable Player” award will tend to be an outlier in a few statistical categories. Therefore, I must keep these outliers in my dataset, otherwise the model will only learn about the average player and not MVP caliber players.

Exploratory Data Analysis Phase

During the exploratory data analysis phase, I observed the relationships between each player statistic and the target variable, “MVP”. I did this by creating a distribution of each statistic with an “MVP” overlay. This allowed me to see the trend between the statistic and the player’s MVP status quite clearly.

I also wanted to observe the distribution of how many MVPs played at each basketball position. To do this, I created a bar graph of the player’s positions with an MVP overlay. Additionally, I created a contingency table between player position and MVP status, and made an alternate version of the table which displayed proportions instead of counts.

Due to the fact that many basketball stats have a wide range of values, I decided to bin players based on a few key stats. This will allow me to categorize players more effectively and take advantage of more models in the future. I created a binned version of the player’s age, games played, rebounds, asists, and points. Then I graphed each of these binned values with an “MVP” overlay to get a better understanding of the categories that MVP caliber players fall into.

Lastly, I created a new variable, IPG (Impact Per Game). This variable is a combination of the three major basketball statistics, rebounds, asists, and points, averaged on a per game basis. Then I graphed a distribution of this new variable with an “MVP” overlay to see if it had any predictive merit.