Isabel Osgood

Data Science Tools 2 Final Project Paper

**‘Put Me In Coach’:**

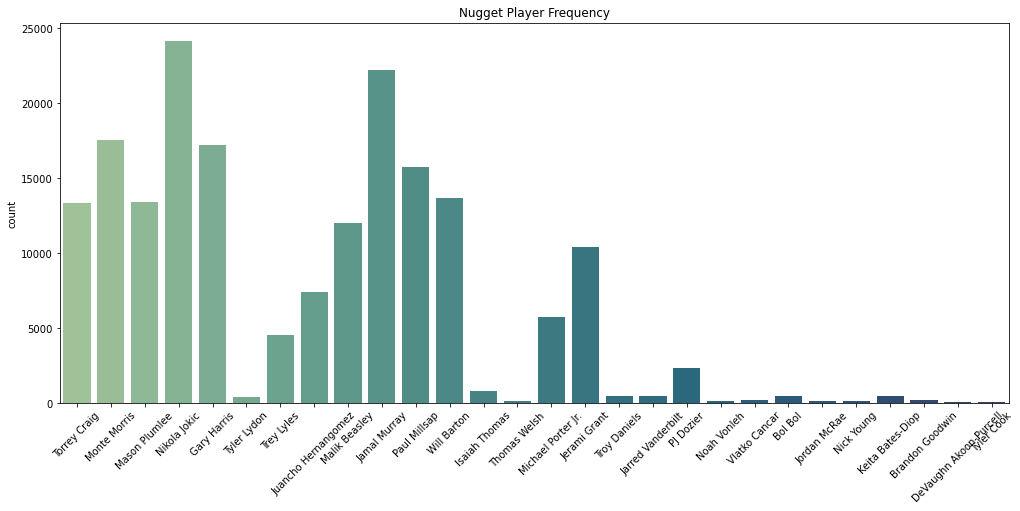
**Investigating Basketball Player Interaction’s Effect on Performance**

**MOTIVATION:**

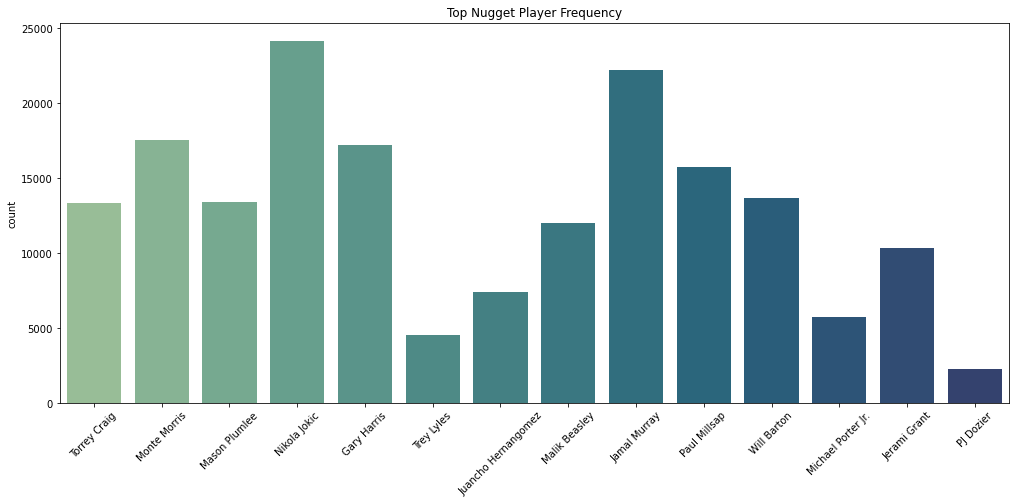
This project investigates the interaction of professional basketball team members on the court by predicting the number of points scored for each play with a given 5 players on the court. I became interested in thinking about player interaction when thinking of the most legendary basketball teams: The Dream Team, The Fab Five, and the 1995-1996 Chicago Bulls. Not only are these teams stacked with the greatest players the game has ever seen but there is a comradery and energy between players that translates to them dominating the court. The aim of this project is to attempt to measure this interaction for the Denver Nuggets from the past two seasons.

**DATA PREPROCESSING AND SPLITTING:**

The play by play data was gathered from basketball-reference.com, then parsed to contain features for time remaining in game, play length, and each of the players on the roster for the past two seasons – 2018-19 and 2019-20. Only Denver’s offensive plays were selected and free throws were also dropped because they are one man plays so interaction does not affect the outcome. The features of this dataset are points scored on play – which is this projects outcome variable, the players on the court, the type of event, the shooter, the time remaining in the game, the period, the play length, shot distance and others that aren’t included in the project. This dataset provides the following player frequency plot.



Now this does not provide an unbiased platform to work with. To limit noise, the bottom half of the players were removed from the dataset. This is the last step of data parsing and provides 36,753 plays (instances) for the eventual model building.



The player columns then went through a one-hot-encoding process to indicate which players were on the court during any given play. The data was then split using a standard 70-30 split for training and testing.

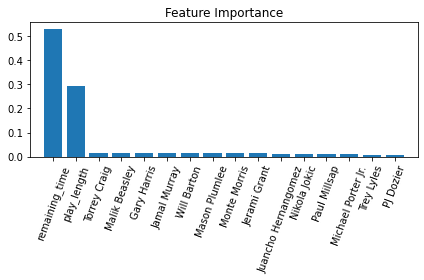
**BUILDING MODELS:**

To investigate player interaction the data was sent through the model building and evaluation cycle 3 times. The first time the input features Remaining Time, Play Length, and each player. For each round Random Forest, Logistic Regression, and K-nearest-neighbors methodology were compared. These algorithms were selected because they are three of the most popular multiclass classifier algorithms but are also different enough to provide an interesting comparison.

Round A:



There is some very obvious overfitting with the random forest model. There is not a significant difference in logistic regression or k-nearest neighbors but both models performed well and there is no overfitting or underfitting. Using sci-kit-learn’s SelectfromModel package to look at feature importance coming out of the random forest modeling.



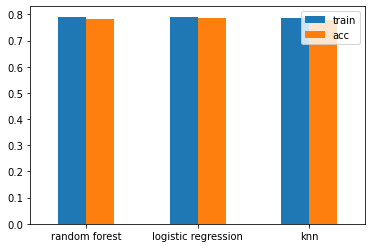
Remaining time and play length and clearly having a much larger impact on this modeling so in the next rounds remaining time and players exclusively to see if successful models can be created from those features alone.

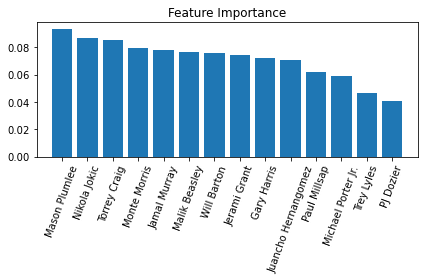
Round B: Remaining Time.

For this random forest, logistic regression, k-nearest neighbors again were the algorithms compared. But this time random forest does not produce the overfitting.

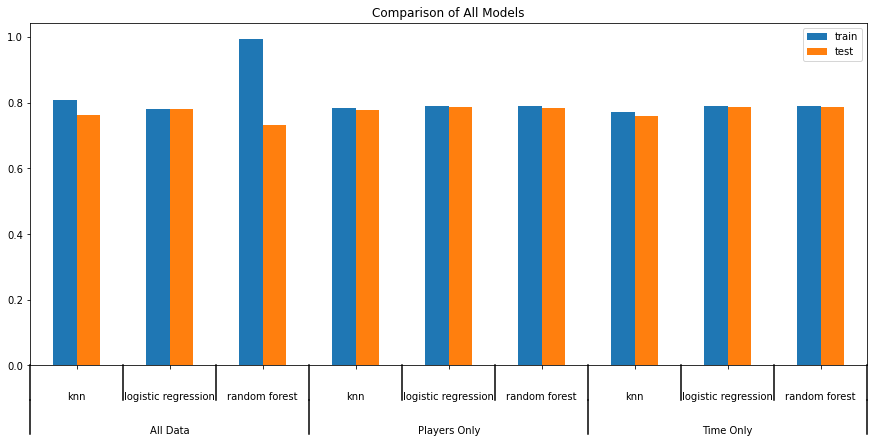


Round C: Only using players. This also eliminates any interaction between players and remaining time. Again none of the models show any overfitting or underfitting and are moderately successful. Similar to round A – feature importance analysis was conducted again.





Now all 9 models are compared with each other.



None of the models really stand out amongst the rest except for the entire dataset’s random forest is overfit. All of the models are moderately successful which is cool but not super cool.

**CONCLUSION:**

This project didn’t really give us a definitive answer to the question of how the interaction between Denver Nuggets effects the game but did provide some really interesting insights on interaction. For simplicity of discussion if interaction between players/time will not be investigated to be positive or negative although the importance is measured. Some interesting tidbits coming from this analysis include: Torrey Craig is more important when remaining time is included in the model but when time is removed and the interaction between players and time is no longer a factor Mason Plumtree has a much longer impact on the model.

FUTHER CONCLUSIONS:

The relationship between players on the court and points scored we can just look at pivot tables of the best and worst performing squads on the Nuggets’ rooster. This analysis was done on the full rooster not just the top players.

TOP PERFORMING SQUADS:

There is a bubbling effect on the top of the performance table – meaning that the top performers have very low levels of play counts and time together on the court.



LOWEST PERFORMING SQUADS:

The Nuggets should probably avoid putting these squads together on the court.

