NBA DFS Capstone Milestone Report

Davy Gia Nguyen

The main goal of this capstone project is to look at two sets of NBA data and to come up with a model that will be useful in differentiating the stats that matter the most in predicting and projecting daily fantasy points for each player.

The task of wrangling and analyzing NBA data to aim at predicting daily fantasy sports is very close to my heart because it led me on a journey that opened my eyes to the extraordinary world of data science.

I was using Google Sheets as my main tool to analyze and construct DFS lineups when I first became interested in this hobby, but I soon realize after the first year that the highest-ranked DFS players were using much more sophisticated tools like R and Python. Further research led me to signing up for DataCamp and then I eventually decided to switch my entire career to Data Science and joining SpringBoard's program as I become convinced that Data Science is like a superpower. It can provide meaningful insights in countless industries, it can predict the future, and after listening to a data science podcast where one data scientist discussed how his team literally save lives using data science to predict and help patients with sepsis in emergency rooms.

Thus, when I was considering a few different ideas for my Springboard's capstone project, I ended up choosing to look at predicting NBA DFS because I can both hone my superpowers while analyzing the subject that I'm very interested in and have some domain expertise in.

I hope that the analysis and conclusions can provide some meaningful insights to the readers since I know the NBA league itself is the fastest growing sport and with on-demand services like Netflix being more popular, sports is becoming more and more attractice due to more lucrative TV contracts, and the DFS industry itself is just in its infancy with plenty of room for growth. Therefore, I hope that this phenomenon will lead to more and more daily fantasy players who might find this capstone useful.

My main hypothesis from playing DFS for two years is that performing clustering analysis to engineer a feature called DvC, defense vs clustering, will be very valuable in predicting future fantasy points performance for each player.

For my capstone project, I utilized two main sets of data and employed various data wrangling techniques to prepare them for analysis and for performing computations that will yield meaningful insights that will compose a model that can predict a player daily fantasy points. The first dataset is comprises of season stats for all active NBA player for the 2018 season acquired from stats.nba.com, while the second dataset was acquired from rotoguru.com.

After the first part of clustering analysis, we learn from our clustering operations that our hypothesis is corroborated such that simply relying on DvP or defense vs position stats for DFS is not as reliable or accurate because a player conventional position is misleading and depends much more on coaching changes, player roles, and countless other factors. DvC or defense vs cluster would help us solve this problem better, and performing regular clustering operations will keep each player up to date with his current cluster type due to the many changes. For instance, a DFS player looking at a DvP stat and see that a team is weak vs point guards so they will pick a PG player from the opposing team for that day such as Stephen Curry (PG/cluster 2) from the Warriors to take advantage of this opportunity, but the actual DvC stat will show that the particular team is actually weak against cluster 0 so the DFS player using DvC will actually roster Klay Thompson (SG/cluster 0) and perform much better. Note that clusters can have multiple positions such as for this simulation our cluster 2 comprises of SG, SF and PG with the following counts: SG 40 SF 13 PG 11.

After confirming that DvC is more useful than DvP, we will proceed to add it to one of our many features and then use regression operations to find out if it is indeed an important stat to predict actual Fanduel Fantasy Points over the course of the season so far and we also examine how does DvC compare to the other features and which features are the most important in projecting FDP.

We confirm our initial hypothesis that Defense versus Clustering, DvC, is indeed a very important predictor of Fanduel Fantasy Points with seasonal FP right behind it.

We notice that other features are ranked as expected such as pace and player impact estimate and age. A slower team will score more FP if they are playing a faster team who dictates the pace. Younger and more-impact players will score more FP.

We do notice some surprising and intriguing results such as certain opponents greatly impact FP such as Cleveland and San Antonio. One possible interpretation is that these teams draw more attention and likely is televised nationally which could incentivize players to hustle more on the court. Additionally, after DvC and FP, the rest of the features are very close with each other and that the coefficients for some stats that are generally regarded as very important and predictive of future player performance in the basketball analytics community are actually not very predictive of daily fantasy performance such as minutes, effective field goal percentage, true shooting percentage, etc. This phenomenon could be due to these stats don't change on a daily basis often enough that the changes are accounted for in Fanduel's daily pricing algorithm.