NBA Analytics & Machine Learning
Name:- Pranav Natarajan
Mentors:- Andrea Boskovic, Harshil Desai

Through participating in the DRP this spring, I had a chance to learn a lot about the NBA. Rookie Contracts and the draft process excited me, and the various statistics surrounding the players and the game caught my eye. I always harboured an interest in the player acquisition side of sport, and wanted to explore the use of data and analytics in that realm. Since value estimation of the players is of importance in player acquisition, I chose to work on the prediction of yearly NBA salaries, given a player's biodata and seasonal statistics.

My mentors and I then charted a plan for the project, having accepted my request to learn how to implement machine learning models in R. The first few discussions were about selecting sources of data, and what features to focus on. Further discussions resulted in the use of the `nbastatR` package (maintained by Abe Resler) as a credible source. It queried from the official NBA website, and from reputed basketball statistics website, BasketballReference. We decided to use data of players from 1985-86 to 2020-21, with the earliest players being from the draft preceding the 1985-86 season. This draft was chosen to subset players as it was the formal method to participate in the NBA — even for foreign nationals and players from foreign basketball leagues. 1985-86 was chosen as the earliest season as that was when the NCAA decided to use the 3 point line in collegiate basketball. This allowed for rule uniformity between the NBA and the NCAA, where most of the draft players have played their basketball. We also decided to normalise salaries by the yearly salary cap to negate effects of inflation and other yearly salary fluctuations in the NBA. Finally, we discussed relevant features, and ended up choosing 23. They included player age at the start of the season, the teams they played for, and more advanced metrics like the VORP, PER, and Usage Percentage.
I wished to explore ensemble models in predicting NBA salaries. So, we chose the Random Forest and Gradient Boosted Regressors. We also chose the Elastic Net as a baseline model to compare the ensemble methods' out of sample performances. For feature selection, I chose to use the Boruta Algorithm. Boruta uses the variable importance metric — a metric both ensemble methods can be evaluated by. I then learnt about these algorithms, performed relevant hyper-parameter tuning, and obtained out of sample estimates.

The final presentation was worked on in the final two weeks, with helpful feedback regarding image size and amount of text on images. Two rounds of trial presentations really helped streamline the information that I could present in about 10 minutes.

To summarise, I owe a lot to the discussions I have had with Andrea and Harshil. They really helped solidify concepts regarding the NBA contract process, rookie contract scaling, player value estimation, ensemble modelling, and hyper parameter tuning methods in R. Their feedback on my trial presentations helped me condense information that was unique and important to highlight, as well as giving me an insight to academic presentation. The resources that were shared (beyond the scope of just ensemble modelling) is something that I am really grateful for.