SPA DRP Writeup: Machine Learning w/ NBA

Using machine learning this quarter, I explored different models to predict the MVP of the NBA, which usually is awarded to the most valuable player in the league. They are determined by a set of voters who vote on players based on their opinions, giving players first, second, third, fourth, or fifth place vote. To do this, I used player performance variables and player perception to see how well I could predict the MVP. Overall, I wanted to see if player perception (social sentiment) would be a good predictor of the MVP. So, using machine learning I carried out the task of predicting the MVP. Machine learning is a hot topic in the sports industry and is widely used, so I believe it is important to have a basic foundation of it. It is a powerful tool that can be utilized to come up with predictions about many different topics. All in all, it's a great advantage to know how they work and how to utilize it. I was able to learn this during this project using R, and using different packages to carry out different tasks. Since I had some experience applying machine learning in Python, it was great to actually see the differences in how we apply machine learning in R compared to Python. Anyways, reading different textbooks and documentations recommended by my mentors, I was able to learn different models in R like Bayesian and XGBoost. However, I also applied models that I was already familiar like a few regression models (LASSO, Ridge, Linear, Logistic). Although I didn't become a master at utilizing these models, I definitely gained some confidence using these models and interpreting the results.

Outside of learning more about machine learning in R, I was able to learn more about the whole data analysis (with machine learning) pipeline of collecting data, cleaning data, modeling, and evaluating results. In the beginning, I had to shut down a couple of project ideas due to the fact that the data was hard to obtain, so being able to come up with a question where the data was accessible was also another step in the long process. However, once I knew the data was accessible, it was a matter of using the right APIs and libraries to scrape the data. In my case, player performance and value were the easiest to obtain, both being available in CSV formats and easily scrapeable.

Putting it all together, I finalized the results into a presentation in R (ioslides), which made some tasks easier whilst making some other tasks a bit hard. Being able to use R objects in the presentation was nice as I was able to include HTML tables and other kinds of interactive objects, especially because I had lots of ggplot files on hand. Thus, using ioslides made it a seamless process of embedding plots and visualizations. However, the content of the presentation focused on the general process of the project, how I obtained the data, and explaining how I went about text sentiment analysis, and applied it to our models. Then, going through picking different models and then finally explaining the evaluation of the results.

Overall, this whole project was a great learning experience. Being able to set aside time for actual application of machine learning to real world data with the help of the mentors really helped me stay motivated to keep learning more about the material and go through with the project. This process truly built the foundations of my machine learning knowledge and I am confident that I'll be able to carry it onto future projects and to the workplace.