**Data Science: Case Study Prompts**

Please choose one of the following prompts to research, analyze and present. It is preferred that you select a prompt for the sport that you’re most interested in working in. Each presentation should include the data that was collected, the methodologies employed and a proposed framework on how to automate or implement in real-time. Each presentation should last 30 minutes with 30 minutes for Q&A and discussion.

**NBA Player Usage Rates:** Player usage rates (% of offensive plays that end with a FGA, FTA or TO for that player) are context dependent. When a high usage rate player shares the court with another high usage rate player, usages rates often decline because there are only so many possessions to go around (see Durant/Harden/Irving experiment in Brooklyn). Similarly, when a high usage rate player is absent, other players will need to step up to fill the offensive void.

Please design a model/framework that can predict usage rates for players in a particular game. This framework should be agile enough to account for late scratches, minutes restrictions, etc. Use the last two weeks of the current NBA season for out of sample testing.

**MLB Player Clustering:** Pitchers and batters vary considerably in strengths, weaknesses, and tendencies. Some pitchers feature a dominant fastball with less effective secondary pitchers. Other pitchers lack velocity but thrive on movement and accuracy. Similarly, some batters perform well on certain types of pitchers and struggle against others.

Please conduct a player clustering analysis which groups pitchers into categories based on their pitch profiles and pitch success rates. Among other things should take into consideration metrics such as velocity of various pitches, pitch selection and pitch success rates. The framework should be designed to predict expected run values of plate appearances by a particular batter facing a particular pitcher. Use the last month of the 2021 MLB regular season for out of sample testing.

**NFL Data Weighting:** With the fewest games of every major US sport, the NFL has considerably less data than other sports. This introduces challenges of how much to weight recent data versus older data when assessing player performance. The challenge of updating rating is increasingly difficult for young players or players who’ve changed teams. For example, after three lousy seasons with the Jets (as one of the youngest QBs in the league) in 2021 Sam Darnold led the Panthers to a 3-0 start while averaging 296 passing yards per game and a 3:1 TD:INT ratio. Those first 3 games had him at an QB Rating of 98.9, which would have finished in the top 10 if he sustained that over the full season. Is this a small sample size of an outlier performance or an indicator of things to come?

Please construct a model that optimizes a weighting function NFL QBs. The dependent variable should be QB Rating or another commonly used QB rating metric. This type of analysis should follow the Bayesian approach to constructing posterior estimates of player skill levels.