NFL and Data Science

Mengyu Jackson

I'm using Kaggle competitions to keep my skills fresh while I search for my first Data Science job. The NFL Big Data Bowl 2022 seemed like a problem with a good data set that should generate a lot of other submissions to compare myself against. I downloaded the data, opened it up, and quickly realized there was a big problem: I don't know anything about football!

I didn't know what a "Snap" was, or why it could be "ok", "bad", or "N/A" on different plays. I didn't know what any of the positions were (QB, RB, WR, K, HB, etc.), and which ones were involved in special teams, which is the focus of this competition. Luckily my father-in-law is very knowledgeable about football, and he was able to answer most of my questions. I know that kickoffs don't have snaps now, but all other plays start with a snap, which can be good or bad. I know K means Kicker, which is a special teams position, and QBs aren't important at all (for special teams).

It took me a long time to learn enough about football to begin to understand the data, and I'm only just now starting to analyze it properly. I hoped that this blog post would be about a key insight I discovered or a novel metric I was working on, but instead it's another reminder about the importance of domain knowledge. Effective data scientists need to understand statistics, communicate effectively, and have a strong technical programming foundation, but that's not enough.