Machine Learning March Mania (NCAA) Reflections

This is a classic ML problem that is not difficult or complex in any way when compared to other CV and LLP problems. Yet I was not able to create a valid submission before the deadline hit, wasting a lot of time, effort, and some money. (Kenpom and CSV Scraper subscription)

**The timeline:**

-10th of Feb: Competition started with the first submissions being available a few days after.

-8th of March: I did not do anything meaningful that contributes to the progress until that time when I began to collect data from Kenpom, Masseys and messed around with HoopR. I also tried to do some data matching (making sure that the data collected from the various sources describe the same team at the same timeframe). I also discovered that the team names were not standardized across platforms, and this is an issue that is not solved until much later into the competition. This is also one of the tasks that I assigned to my teammates and was still done by me in the end.

- For the next two weeks: We didn’t really do much besides some data matching. This was extremely inefficient and looking something like this: I send my teammates two files and ask them if they describe the same timeframe in the tournament. They give me a result 2/3 days after (maybe even longer). Repeat this process for four times.

-Final 10 days: I can say that this is the period of time in which 80% of the entire competition was completed. I standardized all team names, concatenated them into a trainable dataset, then built models that worked. Eventually I was not able to complete the submission file in time but was really close to it. This tells me that I was able to complete this in time, even without any help from the teammates, had I just done better time management.

**What I did wrong:**

-Beginning late is not even problematic if I broke down the task into smaller, more manageable pieces and distribute it across the remaining time. I did not do that, which is a major mistake. Time management has always been an issue in all the competitions that I have participated.

-I also put too much emphasize on the data collection stage. I should have built a simple baseline model as a primary task when a first began and not worry about external data until I have a working pipeline.

-Wanting to do everything the best, ended up finishing nothing. Again, this is a mistake that I have committed myself to repeatedly. I always want to go all out when I am doing something. For example, I put too much focus into the data collection stage in this competition, wanting to find all possible metrics and data available on the internet when I didn’t even have a basic model. Instead of doing all that useless feature-engineering, I should have built a pipeline that will predict the result with okay accuracy but will produce a viable submission file.

-I did put too much trust into my teammates. They do not understand anything about practical ML before this competition. I got them in because they were my friends and had some theoretical knowledge, but that is probably not the best idea if I wanted to do produce something working and productive. Most tasks I assigned to them were done by me eventually anyways.