Samuel Kellum and Ethan Perello

Mentor: Prof. Mattei

CMPS 4010-01

Prof. Zheng

September 13, 2022

For the final project, we will be participating in a kaggle machine learning competition for the 2023 NCAA men’s basketball tournament (2022 competition URL: <https://www.kaggle.com/competitions/mens-march-mania-2022/data>).

Every year, there are thousands of basketball games played between Division 1 men’s teams, ending with the March Madness tournament. This tournament consists of 68 teams and is single-elimination in style. In other words, a team that loses a single game during the tournament is eliminated.

The objective of this competition is to utilize historical data in addition to any external data to build a machine learning model that predicts probabilities for every possible matchup in the past 5 NCAA® tournaments (2017-19 and 2021-22) in addition to the 2023 tournament. The predictions will be submitted before the tournament begins, and the model(s) with the most accurate predictions will win the competition. We are allowed to submit two submissions to the competition so we will keep this in mind as we work. Having two submissions that use slightly different models to predict the outcomes will maximize our chances of winning.

Fortunately, we have access to all of the submissions from previous years, which means we can study the models of previous winners as inspiration for our model.

We will need to decide which technologies we will use to build our machine learning model. As of now, we will likely use Python’s Pandas and Scikit-Learn libraries to build our model. One immediate challenge we face is our lack of machine learning knowledge. Therefore, a significant amount of our time in the fall semester will likely be spent self-learning the fundamentals of machine learning. Another challenge we face is deciding what data to use and how to build our model. Fortunately, there is a ton of available data on the website such as game-by-game stats (points, free throws attempted, defensive rebounds, turnovers, etc.) for all regular season, conference tournament, and NCAA tournament games in the last 20 years. Our plan is to focus on the data already provided by the kaggle website before looking into external data.

We know some external data that we will likely have to find because previous winners of the competition have used these outside sources. Most successful past submissions have used basketball data outside of NCAA basketball. Basketball data is very easy to find and access though and we are confident that we can find sources for this outside data as well. There are many basketball leagues around the world that record and publish basketball data.

There will also be many submissions from other people which will create a challenge because in order for our model to be effective it will have to have something about it that is different from all the other models. We do not know yet what specifically will make our model unique, but we plan to come up with potential solutions as we go through the process. Submissions for the challenge are scored on the log loss function so we plan on frequently testing our model using the log loss function. As we test our model we will be able to discover which of our new ideas are effective and which new ideas are not.

A submission that won the competition in 2017 used its competitors' submission's predictions as part of the model. This is the type of novel idea that we will need to make our model the most effective. We plan on including this in our model, but since it has already been done we will need to come up with something else in addition to this.

After discussion with our mentor, we are highly confident in the feasibility of this project, especially since it is already well defined. Although we have not assigned roles yet, we will always ensure that we are dividing the work evenly. Additionally, we plan on meeting with our mentor every two or three weeks throughout the semester.