Project Summary

Harry Zhenghan Zhang

Before the start our course, I had experience programming (in any programming language: Python, Java, R, C, JavaScript, etc.) and had defined/written my own functions or methods and also defined my own classes or objects.

The goal of this project is to use NCAA(basketball,men) historical data to analyze and predict the result of a potential matchup between any of the two teams.

The data used are separate csv files containing information of all historical matchups since 2003,each team's information and conference information.All data are gathered on Kaggle website, and can be downloaded using this API: 'kaggle competitions download -c mens-machine-learning-competition-2019'. Here, I have already downloaded the data beforehand, as they are presented in zip format and need to be unzipped first.

The project was already encapsulated, so assuming the user uses the .py file, he/she would only need to input the two teams he/she wants to predict. And the program looks for a saved model, if not found, it runs the training algorithm on the data (already preprocessed) and saves the model locally in h5 format. Then it uses the model to predict which of the two teams would win, and displays to the user.

The features used in this project: data handling and cleaning, plotting, model training, file I/O, use of pandas, numpy, keras and sklearn, and user interaction.

IMPORTANT:

1. this project requires the following packages: sklearn, numpy, pandas, keras, tensorflow, h5py, matplotlib, seaborn.
2. All paths I used are absolute paths, under the "readcsv" section and "userInterface" and "getTeamNames" function. Please adjust accordingly.
3. Check .ipynb file for detailed illustration and description, and also a more detailed explanation of the codes. Use.py file for prediction only.