Scope of Work for Data Science Final Project

Prepared by Group #28
- Mark Dinneen
- Mohammad Karim
- Ramandeep Harjai

Project Statement and Background

The World Cup is a world wide football tournament that occurs every four years. National teams from countries around the world compete for the title. While football is widely popular around the world only 32 teams qualify to compete in the world cup.

Goals: Using the players, team, and past games data, optimize the learning model which can predict which team will win the world cup.

Literature Review

- Soccernomics (2018 World Cup Edition)
- The Mammoth Book of The World Cup: The Definitive Guide, 1930-2018
- Economic Impacts of the FIFA World Cup in Developing Countries ([link](#))

Available Resources/Data

- Github Repository: https://github.com/rdharjai/csci_s109a_final_project
- Sofifa.com Player information including basic features (e.g. ID, age, height, weight, etc.). Other features include attacking, skill, movement, power each with sub-features. This data is thorough for player analysis.
- Squawka.com Stats for teams, players, and goalkeepers across different regional leagues with options for viewing previous seasons.
- Soccer club data from: https://www.soccerassociation.com
- Soccer related datasets from: https://data.world/datasets/soccer
- Player and game statistics from: https://www.whoscored.com/Statistics

Approach

- Scrape data from multiple web sources, listed above. Clean-up scraped data, prepare a dataset
- Import publicly available datasets (sources, listed above)
- Merge scraped, and imported dataset

- Perform visual inspection, and preliminary statistics analysis
- Perform model, and predictor selections
- Build and test models
- Obtain prediction results - infer model, data, and results obtained.

Preliminary EDA
- Visualizations and description of preliminary findings.