607_Assignment1

Ariann Chai

2023-09-01

Overview

The article and dataset I choose was 2019 Womens' World Cup Predictions (https://projects.fivethirtyeight.com/2019-womens-world-cup-predictions/).

This article was used to predict the outcome of the 2019 Womens' World Cup and to record the results of the tournament. There was 2 datasets included in this article: wwc_matches and wwc_forecast. wwc_matches was more focused on the individual matches of the tournament and had more specific prediction variables as opposed to wwc_forecast which more showed the results of the tournament with the teams' spi rating being the main prediction factor. I choose to work on/fix up the wwc_matches file.

Findings and Recommendations

While comparing the predicts made and results of the tournament, I conclude the multiple factors used did a pretty good job of predicting the outcome of these games. SPI (or soccer power index) rating was the clear best predictor as 40 out of 52 games where won by the team with the higher spi rating (the 12 games also included tie games). SPI ratings were generated by authors using 4 metrics. If I were to recommend a change for the data, I would have minus one of these metrics which had to deal with accounting for red & yellow cards and the time of the game when a goal was scored. It's stated in the article that one of these predicted problems with the data was if there was no play by play available for a match. The solution they decided on was to only include the final scores for that match. I don't think it is necessarily important to have the score times included especially if it could cause a problem where a match does not have this data.