Sumo Stakeholders and KPIs

Dataset:
https://www.kaggle.com/datasets/thedevastator/sumo-wrestling-matches-results-1985-2019

Notes: Our dataset includes the tournament/venue, day (of the week?), wrestlers' names, wrestlers' ranks, what their record is, if they won their match, and the "deciding" move. This is presented such that for a given tournament, every athlete is mentioned as the "first" wrestler on one line and as the "second wrestler" on a different line.

Goal: We intend to predict the outcome of sumo matches. This could mean a variety of things. The most straightforward is the winner of the match, but we can also look at the final (or "deciding") move of a match.

Ideas to consider:

- Match fixing (八百長) scandals have been revealed several times. (The book *Freakonomics* by Levitt and Dubner goes into detail on this.) We ought to take this into account in our analysis.
- One potential indicator that is not explicitly in the data but is tacitly present is "length of time in professional sumo." We can compute this by listing the year that each player first appeared at a tournament.
- If we are not satisfied with this data, we can scrape data (up to the present day!) from https://sumodb.sumogames.de/.

Potential Stakeholders: This analysis would likely appeal to sumo fans, as well as sport officials and team owners. The owners/coaches specifically may be interested because they could reference this analysis to decide when and with which opponents to schedule their players at any given tournament to maximize their athletes' chances at success. I am not sure if sports betting on sumo is an active pastime, but the case that it is, this analysis may aid gamblers.

Key Performance Indicators:

- Percentage of matches correctly predicted (should be significantly higher than 50%)
- Percentages of matches whose final move was correctly predicted (should be significantly higher than (100/{# of standard end moves})%.
- Ability to estimate the rank of players moderately well without looking at the ranking data. (What is the average residual of our ranking to the actual ranking?)