Calculated Fields MLB Win Predictions

1. My calculated field was run differential, and it was created by subtracting first half runs allowed from first half runs scored. The result was a superior R-squared for a simple linear regression model. The model between first and second half wins produced an R-squared of 15.54%, while the model between run differential and second half wins produced an R-squared of 20.52%.
2. I think the improvement exists because run differential provides more context to the strength of a team compared to just wins or losses. Obviously, wins and losses are the most important thing in the mind of the front office. However, when thinking about just 81 games, run differential paints a more detailed picture of the true performance of a team. For example, a team with a lot of close wins and blowout losses may not be better than a team with a lot of close losses, even if they have a few more first half wins.
3. I'm not entirely sure if run differential would improve year-over-year performance predictions when compared to wins and losses. In reality, both of these predictors might not be very good at projecting performance for the following year due to year-over-year changes in the 40-man roster. I think that general managers should instead use runs allowed and runs scored on their own, as it would provide more context on the strength of a team's scoring and run prevention capabilities.
4. If it were discovered that run differential had a strong correlation with successful year-over-year projections, then my decision-making as a GM would change. Instead of wins, run differential would become my priority when considering the strength (or weakness) of the upcoming year's team. If the past season's run differential was poor but wins were high, I would still be somewhat aggressive when it comes to bringing in more talent. If wins were low but run differential was high, I would feel less pressure to make franchise-changing moves in the offseason.

Golfing Score Predictions

1. Formula:

$$\widehat{Avg} = 92.89 - 0.025 AvgDriveDistance - 0.134 GreensinRegulation\% \\ - 0.614 AvgStrokesGainedPutting - 0.102 Scrambling\%$$

2. The highest variable for impact on a score is average strokes gained putting. Because of this high impact, an improvement by one standard deviation would have the biggest impact on the predicted score. The least impact on their score would come from a one standard deviation increase on average drive distance, which had the lowest impact in the model.

NFL 3rd and 4th Down Entropy Revisited

1. From the simple linear regression model, I do not believe that there is a meaningful correlation between standardized entropy and first down rate. The p-value was somewhat high (0.1439) and the R-squared was very low (3.88%). Additionally, this regression does not control for offensive and defensive grades, which makes it difficult to say that standardized entropy is significant.
2. If we control for team offensive grades, I expect there to be an increase in the strength of the correlation between these two variables. If offensive talent is controlled for, it is most likely better to have variability in your team's play-calling.
3. In the multiple linear regression model, I believe that pass grade, run block grade, and standardized entropy are all significant in the model. Therefore, I would leave just these three variables in the multiple linear regression model.
4. Standardized entropy's p-value dropped to 0.02, which suggests it is statistically significant in this model. I believe that this drop happened because we accounted for team offensive grades.

By accounting for pass grade, pass block grade, receiving grade, rush grade, and rush block grade, the omitted variable bias from the previous regression was removed. Therefore, the estimate of standardized entropy's relationship with first down rate became more accurate.