Parsa Zilaee

Summary Paper

Statistical/Hypothetical Question

The primary question driving this analysis was: What physical and historical factors significantly impact a boxer's success? Specifically, I aimed to understand how variables like age, height, reach, and fighting stance influence a boxer's win rate. Additionally, I explored whether past performance metrics, such as total wins, could be used to predict future outcomes.

Outcome of EDA

The exploratory data analysis (EDA) provided several key insights:

Histograms and Descriptive Statistics: Variables like age_A, height_A, and reach_A showed fairly normal distributions, although age had a few extreme outliers. Wins (won_A) were right-skewed, indicating most fighters had a lower number of wins, with a few exceptions having very high win counts.

PMF: Comparing win probabilities between Orthodox and Southpaw fighters revealed that Southpaws often had a higher probability of achieving high win counts.

CDF: The cumulative distribution function for reach_A highlighted that the majority of fighters had a reach between 170-180 cm.

Regression Analysis: A multiple linear regression model showed that reach and age positively correlated with wins, while height had a slight negative effect. The model, although statistically significant, had a low R-squared value, reflecting the complexity of predicting fight outcomes with limited data.

Missed Variables

Several factors that could enhance the analysis were missing:

Opponent Metrics: Including data on opponents' attributes (e.g., height, reach) would provide a more holistic view of match dynamics.

Fight Context: Variables like match location, referee bias, and crowd influence could offer additional predictive power.

Training and Injury Data: Fighters' training history, injury status, and fitness levels were not available, limiting the model's accuracy.

Incorrect Assumptions

One significant assumption was that all fighters' data were equally reliable. However, extreme outliers indicated potential data entry errors. The regression analysis also assumed a linear relationship between predictors and wins, which may not fully capture the interactions in boxing performance.

Challenges and Learning

Handling Missing and Erroneous Data: Cleaning and interpreting outliers required subjective decisions.

Statistical Interpretation: Understanding the implications of low R-squared values in regression was initially confusing but ultimately aligned with real-world variability in sports data.

Categorical Variables: Properly handling and interpreting the influence of stance on performance was complex, requiring deeper exploration.

Despite these challenges, this project significantly improved my ability to conduct and interpret statistical analyses in a real-world context.