A Brief Bayesian Analysis to Discover the Optimal Region to Save Penalties in Fútbol

Aditya Sriram

2024-05-22

Bayesian Analysis of Penalty Saves in Football

Introduction:
In a recent conversation with my old college roommate, we discussed the question: "What sport skill do you think you could reasonably perform ***at a professional level*** given one full month of training?" For context, if I did nothing but play basketball and work on my skills with a trainer for an entire month, I still don't think I would be able to get off a quality shot attempt against NBA-level defense. Same thing with a forward pass in the NFL, even with an elite offensive line. We went through a bunch of different sports and kept coming to these types of conclusions.

However, once we ended up on the topic of soccer and penalty kicks, we agreed that kicking a penalty would be one of the few viable sports skills we could hope to succeed in after a month of training, even against professional-level goalkeeping. But this begs the following question: *As a goalkeeper, is there a region on goal where the chances of stopping a penalty are the highest?* This quick analysis aims to determine the most likely region where penalties are saved by goalkeepers, using a Bayesian inference approach. The data is derived from a research paper, and the data used here is:

1. the number of penalty kicks per goal region,
2. % of penalties saved, and
3. penalties just outright missed (without a save attempt) in each area of the goal.

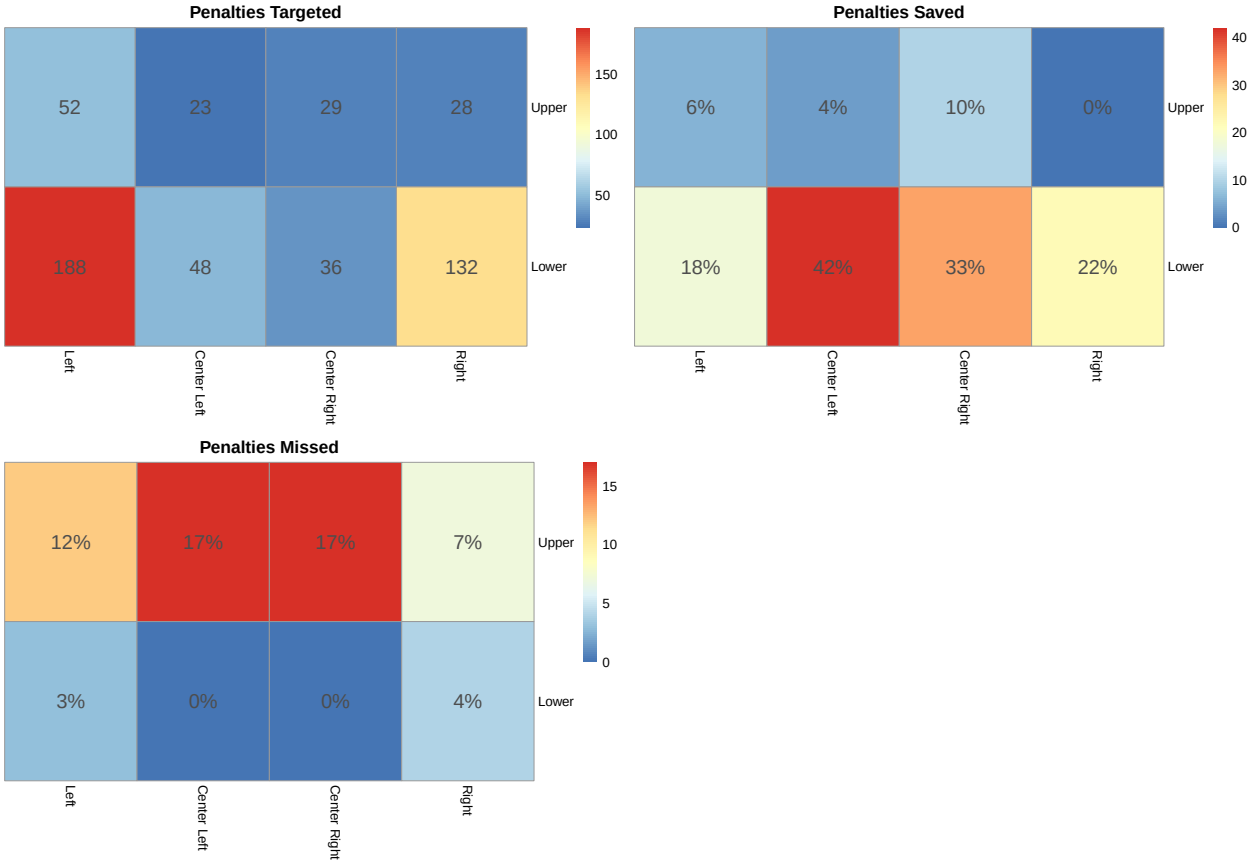
For this exercise, the soccer goal is split into 8 unique goal boundaries: upper and lower for the horizontal boundaries; and Left, Center Left, Center Right, and Right for the 4 vertical boundaries.

Data for penalty kicks is based on a sample of 536 penalty kicks between the *2010-11* and *2014-15* UEFA Champions' League and Europa League seasons. Other variables can theoretically be added to this model such as:
- footedness of player
- grounded shot or elevated shot
- likelihood of making a PK given average number of non-PK shots attempted per match

I will keep this analysis relatively simple for now and just consider the location of the penalty kick taken for this exercise. Also, why in the world am I doing this? If I know the regions where a goalkeeper is *most likely* to stop a penalty, I will put my month's training into kicking PK's at zones that are hard for goalkeepers to stop even with their world-class training and skill.

Today's Coding Time: 90 minutes max.

Data from Heatmap - Penalties Targeted and Saved per Region



Bayes' Theorem Application

Bayes' theorem helps us update our existing knowledge/belief about the probability of a penalty being saved in each region given that a save has occurred.

(1)

$$P(T_i|S \cup M) = \frac{(P(S|T_i) \cdot P(M|T_i)) \cdot P(T_i)}{P(S \cup M)}$$

Where:

- $P(T_i|S \cup M)$ is the posterior probability of targeting region i given a save or miss.
- $P(S|T_i)$ is the likelihood of a save given targeting region i.
- $P(M|T_i)$ is the likelihood of a miss given targeting region i.
- $P(T_i)$ is the prior probability of targeting region i.
- $P(S \cup M)$ is the marginal probability of a save or a miss.

Prior Probabilities $P(T_i)$:
Calculate the prior probabilities for each region.

```
library(kableExtra)
# Calculating the total number of targeted penalties
total_targeted <- sum(targeted)

# Calculating the prior probabilities of the targeted penalties
prior_probabilities <- targeted / total_targeted
prior_probabilities_df <- as.data.frame(prior_probabilities)
kable(prior_probabilities_df, caption = "Prior Probabilities of Targeted Penalties") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = FALSE, position = "center") %>%
  column_spec(1, bold = TRUE) %>%
  column_spec(2:5, background = "white")
```

Table 1: Prior Probabilities of Targeted Penalties

	Left	Center Left	Center Right	Right
Upper	0.0970149	0.0429104	0.0541045	0.0522388
Lower	0.3507463	0.0895522	0.0671642	0.2462687

Likelihoods $P(S|T_i)$ and $P(M|T_i)$:
Convert the percentages to probability values for saved and missed penalties.

Marginal Probability $P(S \cup M)$:

$$P(S \cup M) = \sum_i (P(S|T_i) \cdot P(M|T_i)) \cdot P(T_i)$$

$P(S \cup M) = 0.2422015$

Posterior Probabilities $P(T_i|S \cup M)$:

```
library(kableExtra)
# Calculating the posterior probabilities
posterior_probabilities <- ((prob_saved + prob_missed) * prior_probabilities) / P_S_M

# Reshaping the posterior probabilities into a matrix
posterior_matrix <- posterior_probabilities
rownames(posterior_matrix) <- c("Upper", "Lower")
colnames(posterior_matrix) <- c("Left", "Center Left", "Center Right", "Right")

posterior_matrix <- as.data.frame(posterior_matrix)
kable(posterior_matrix, caption = "Posterior Probabilities of Penalty Saves by Region") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = FALSE, position = "center") %>%
  column_spec(1, bold = TRUE) %>%
  column_spec(2:5, background = "white")
```

Table 2: Posterior Probabilities of Penalty Saves by Region

	Left	Center Left	Center Right	Right
Upper	0.0720998	0.0372054	0.0603143	0.0150978
Lower	0.3041134	0.1552919	0.0915113	0.2643660

Findings and Conclusion

Based on the Bayesian analysis, using a combination of save likelihoods and miss likelihoods leads to these most probabilistic regions for a penalty to be ***unsuccessful*** (not just saved, but unsuccessful since we also are accounting for non-save misses):

1. **Lower Left (30.41%)**
2. **Lower Right (26.44%)**
3. **Lower Center Left (15.53%)**

This helps predict the likelihood of a penalty being stopped based on historical data. Again, the penalty kick data comes from **UEFA Champions League and Europe League** spot-kick tracking from the **2010-11 to 2014-15** seasons.

View the reference paper from the **International Journal of Performance Analysis in Sport** with tracking data by clicking the link here.