MAST679 Project 1

Francis Fregeau 26647227

October 2022

Current metric and its issues

The proposed metric is

$$f(t,\ t+\epsilon,\ p_a,\ p_b) = \sum_{p \notin T(p_a)} \mathbf{I}_p(t,\ t+\epsilon)$$

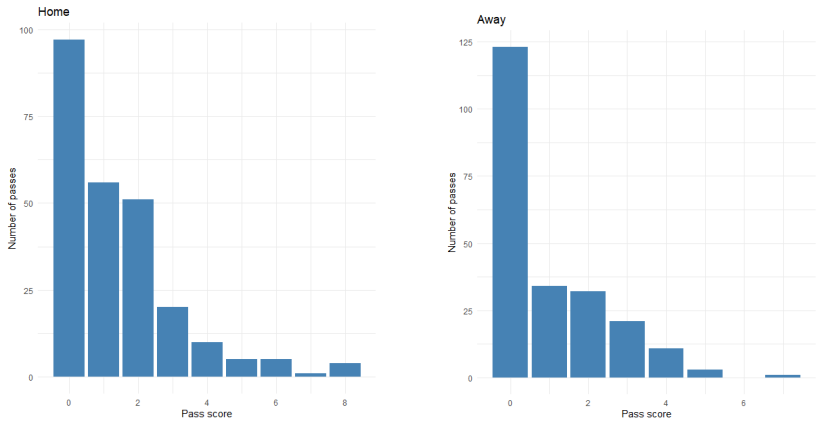
where p_a and p_b are players belonging to the same team $T(p_a)$, t and $t+\epsilon$ are the time at which p_a initiated the pass and p_b received the ball, and

$$\mathbf{I}_p(t,\ t{+}\epsilon) = \begin{cases} 1 & \text{if p was ahead of the ball at time t and behind at time t}{+}\epsilon \\ 0 & \text{otherwise} \end{cases}$$

There are a few issues with said metric, namely:

1. A lot of passes are made backward or laterally. These passes are not necessarily bad per se as they constitute a key part of building proper positioning before initiating forward passes with non-zero metric.
2. The closer we move to the opposite team's goal, the easier it is to score hence passes being all the more impactful. However, said passes are more likely to be made in a lateral fashion due to diminished forward space between the striker and the goal. As a result, the metric is bound to have a lower expected value under such scenarios.
3. A lot of possibly relevant statistics are being left out, such as subsequent events (intercepts, goals, etc.), average player distance from the ball at time $t+\epsilon$, improvement in distance between the ball and the goal, the angle between the ball and the goal at time $t+\epsilon$, etc.

As such, we deem the metric to be of rather poor quality, and intrinsically naive. The distribution of $f(t,\ t+\epsilon,\ p_a,\ p_b)$ for match 1 is shown below and most of attributed scores are 0 as expected.



(a) Barchart, home team (b) Barchart, away team

Figure 1: Distribution of pass scores according to the default metric

Decision trees based metric

We propose to use a metric based on decision trees in lieu attempting to build one which would rely on intuition alone. A pass should lead to one of the following events:

1. Another pass (enemy or friendly)
2. A challenge (won or lost)
3. An interception
4. A lost ball
5. The ball moving out of bound
6. A shot on goal

In order to evaluate the quality of a pass, one can use a multi-class Xgboost model and use the resulting probabilities $\mathbf{p}$ alongside a vector of scores $\mathbf{s}$ tied to each outcomes in order to compute the metric

$$\mathbb{R}^4 \to \mathbb{R} : \ \psi(\mathbf{p}) = \mathbf{p} \cdot \mathbf{s}$$

where $\cdot$ denotes the dot product operator.

Unfortunately, a single match will not provide enough data in order to train a reasonably accurate model, but we nonetheless computed a set of 19 key metrics to work with such as the distance between the goal and the ball at time t and $t+\epsilon$, the angle between the ball and the goal at time $t+\epsilon$, the average distance between the 3 closest players and the ball at time $t+\epsilon$, the number of enemy

and friendly player ahead of the ball at time t and $t + \epsilon$, whether the ball is on the right or left side of the goal at time $t + \epsilon$, etc. As for the values of $\mathbf{s}$, we choose the following quantities:

1. Another pass: -1 for enemy, 1 for friendly
2. A challenge: -2 for a loss, 1 for a win
3. An interception: -2
4. A lost ball: -1
5. The ball moving out of bound: -1
6. A shot on goal: 5

Result

Combining the initial proposed metric with the newly computed one and training an Xgboost model (10 folds cross-validation) yielded poor results (all predictions were friendly passes as the next event). We hypothesise that this is due to the fact that friendly passes as a subsequent event make up the vast majority of the data (high class imbalance). This issue could be somewhat alleviated by down-sampling, but we choose against modifying the distribution of the data. Scores were computed using the out-of-fold probabilities tied to the best model (based off log-loss) and their distribution is shown below.

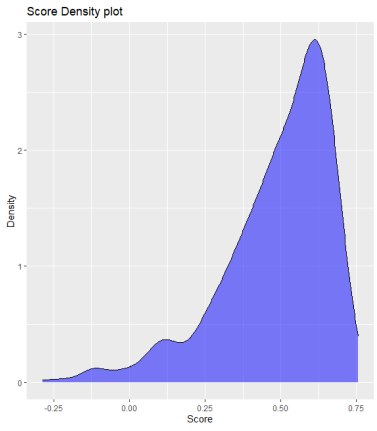


Figure 2: Density of proposed metric

Conclusion

We computed a probability-based metric using the expected score $\mathbf{E}[S|\,\mathbf{x}] = \mathbf{p}(\mathbf{x}) \cdot \mathbf{s}$ for any given pass given a vector of covariates $\mathbf{x}$ tied to said pass. Further development in terms of appropriated values for $\mathbf{s}$ alongside an increased amount of data to work with could most likely lead to substantial improvements in terms of score relevance. We believe that this framework is more desirable than merely using the initially proposed metric as more factors are taken into account, and passes with equal values of $f(t,\ t+\epsilon,\ p_a,\ p_b)$ can still be compared to one another since the probabilities that $\psi(\mathbf{p}|\mathbf{x_i}) = \psi(\mathbf{p}|\mathbf{x_j})$ given $i \neq j$ are almost certainly null.