Predicting Pass Completion Percentage for the 2022 World Cup Final
By Alex Veroulis

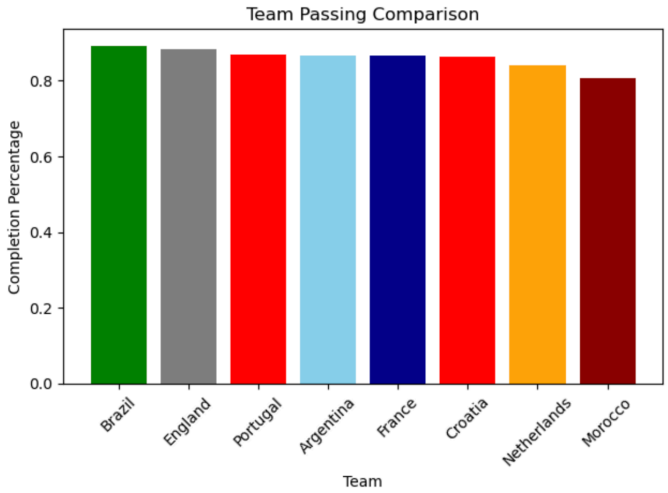

Figure 1: Pass Completion Percentage Among Final 8 Teams

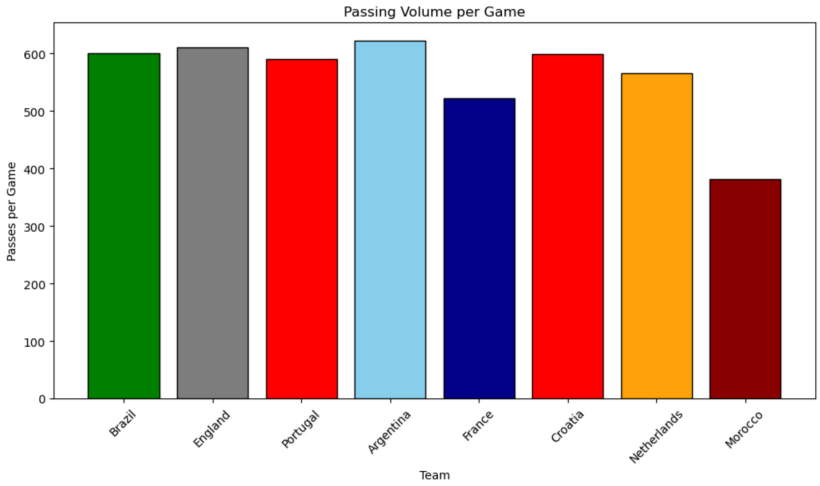

Figure 2: Average Pass Volume Among Final 8 Teams

With the rise of betting across the globe, especially within the United States, professional and recreational bettors alike have taken interest in the world's greatest sporting event, the World Cup. Many states either have legalized sports betting or daily fantasy in one form or another, and there are new niche markets that focus on player performance, mid-game stats, and even team-wide metrics. With the World Cup Final right around the corner, I'm excited to predict the passing completion percentages for France and Argentina in the big game.

The first step of my analysis focused on event data, which is a detailed account of what happened from one play to the next. Key parts of the data included the play type, the team in possession, and touches per possession, among more. Since there was a vast amount of data, my goal was to get a sense of play type frequency in each game (ex. how many clearances were there, etc.), so I counted different play types (including passes and completions) each team carried out over their individual games.

As seen above, I looked at how teams that made the quarterfinals passed the ball throughout the tournament (Figure 1). Most teams are well above 80%, with Brazil leading the pack at 89.1%, and Argentina and France are not far behind at 86.7% each! In addition, when looking at pass volume (Figure 2), Argentina leads the pack with over 600 passes attempted per game, which could easily explain their success up to this point.

Next, I went on to create several variables based off the existing data. I noticed that the pressure_type on passes had three levels, two of which were deemed "successful." So, I changed this variable to 0's and 1's, where 1 signifies a successful pressure and 0 an

unsuccessful one, and I added these instances for each team over each game. After this, I got creative with a custom rating system of opportunities created. With several positive and negative opportunity types possible, I wanted to quantify the impact of these types relative to one another, so I made a variable grading system evaluating the impact of each type. Finally, since PFF provides player grades during the game, I took the average of each team's passing grades throughout their games.

Then, I looked at tracking data, which contained player coordinates throughout the game. With so much data, I wanted to look at the most relevant parts, so I only considered games where either Argentina or France were previously involved. For consistency, I used the smoothed coordinates for analysis, which we previously done so by PFF. I went on to create one variable for each team: the average distance of players from midfield during each of the game's passes. To do this, I used the distance formula to find each player's distance from midfield, and then took the average distance for each team's 11 players for every passing play during a game; then I took the average of these averages to get a game-wide average distance.This is a good indication of how conservative or aggressive a team could be and is indicative of the style a team could play in the future.

The final step was the modeling process, where we used the predictor set from previous games to train a random forest model that outputs predicted pass completion percentages. With the small set of training data, I decided to employ a technique called leave one out cross validation (LOOCV), which treats every observation as the test data and cycles through each observation to see the model's performance. After averaging the errors, we found the average prediction (root mean squared error) was off by roughly 2.62%, a decent performance given the small training set.

After training, I modified the predictors for Argentina and France for their world cup game by taking their variable averages from the previous games in the tournament and then standardized them by comparing their averaged metrics to the game-by-game metrics of the other teams. Then, I input these predictors in the model, which gave us predicted pass completion percentages of 87.5% for France and 87.4% for Argentina. We could have a very competitive game on our hands, only time will tell if these close percentages foreshadow a nail-biting finish!