1RT001 Assignment 1: Evaluating Passes

While attempting to identify creative players, we are tasked with creating a statistical model that predicts what constitutes a difficult pass, where the first run should only look at pass success (xP) as a function of where on the pitch the pass was taken. We use the WyScout open data from the 2017/2018 season for the Premier League, Primera Liga, Ligue 1, Serie A and Bundesliga as a basis for our work. Due to most passes in the team's own half being completed, we will only be looking at passes that end in the opposition half. Headers and set pieces are also excluded from our analysis, as we are only interested in creativity from open play. Initially, when looking at the pass completion ratios for the actual data, it is tempting to replicate the work done on Expected Goals models, as we observe some of the same features for pass success vs location as for goals scored vs location (Figure 1).

However, when we train a logistic regression model on the data set, we find that the start location of a pass has extremely poor prediction strength – in fact, this model (Model 1 in Figure 2) predicts all passes to succeed! Adding end locations (Model 3) improves our results, but we also see from the calibration curve that the model performs erratically for passes with a low probability of success. If we are to use a statistical model to identify players that are good at completing difficult passes, it is crucial that the model performs well for these cases.

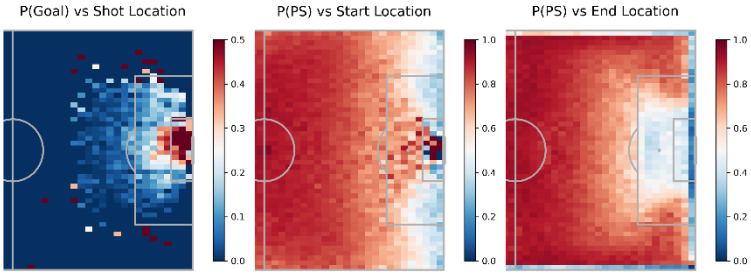

Figure 1: Empirically derived probabilities of goal (left) and pass success as a function of start location (centre) and end location (right) for all passes in the opposition half of the WyScout open data set.

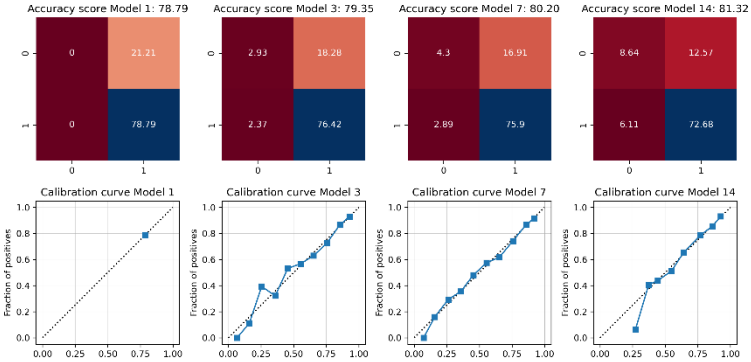

Figure 2: Results from four different Logistic Regression models. M1 is the intial attempt with pass start locations only, Model 3 also includes the end locations. Model 7 uses only the end locations of the pass, plus the derived parameters pass length and xGA. Model 14 is Model 7 plus the dervied parameters pass angle, delta goal angle, xG and end xG in addition to the SMART & SIMPLE classicification tags from the original data set.

We therefore introduce more variables, both to help the model and to help us identify creative players. In addition to physical parameters like pass length, pass angle, distance to goal and goal angle, we also introduce expected goal values (xG) for both the start and end locations, as well as defining expected goals added (xGA) as the difference in xG between start and end locations. For these calculations we use a simple xG model base on goal distance and goal angle. The rationale for introducing xGA is that while goals are the most rare and valuable events in football, good chances come a close second. A pass that significantly increases the xG between the start and the end locations should thus be a good measure of both creativity and pass difficulty. We also investigate if the classification tags for the passes from the data set (CROSS, SIMPLE, SMART, HIGH) will improve the prediction strength of our model.

Tests show that we get the best prediction strength for low probability passes when including just the end locations as well as pass length and xGA (see Figure 2). Adding classification tags helps the overall accuracy, but at the expense of high false positive rate for the low probability passes we want the model to identify. We thus settle on Model 7 to help us predict the probability that a given pass in the data set will be completed.

When ranking the players in the data set, we look at several variables (all normalized to per 90 minutes): Overall pass volume and pass completion rates, net pass progression, high-risk pass volume and high-risk pass success, high xGA pass volume and high xGA pass success, xGA total, xP overperforance and xP final 3rd overperformance. A high-risk pass is defined as a pass that has less than 20% chance of completion by our model, while a high xGA pass is defined as a pass that increases the xG by 0.1. We assume that there is little correlation between high-risk passes and high xGA passes, as a high-risk pass does not necessarily create a good chance, and we want to identify high-risk passers that are not creating a high volume of xGA.

It should be noted that there are several weaknesses to our approach. For instance, we are only looking at the events related to the players themselves and not the wider context. For instance, we would like to take players between the passer and goal into account and to assess passes that create space for other team members in this model. The latter is a crucial skill, as space is the most important commodity in football. Both features would however require tracking data, or at least freeze-frames for each event. Furthermore, as WyScout does not track pressure events, we cannot assess the player's press resistance.

However, the strength of the model is that it is simple, require only sparse event data and emphasizes player creativity. As a face value test, we bin players into defenders, midfielders and forwards and then filter on all players above the 95th percentile for at least four of the above selection criteria. This results in a selection of 14 defenders, 18 midfielders and 11 forwards. All players with less than 900 minutes game time are excluded.

- The 14 top ranked defenders include established stars Kimmich, Alves, Marcelo, Carvajal as well as future world-beater Alexander-Arnold. More affordable[1] and young options found are D. Juncà (24/€1.50M) and M. Navarro (22/€2M), with Ben Davies of Tottenham a surprise inclusion.
- The top 18 ranked midfielders include de Bruyne, Fàbregas, di María, Isco and Perišić along veterans Albrighton and Milner. Young and affordable inclusions are Alejo (23/€4M) and Pereira (22/€6M).
- The top 11 ranked forwards include both Neymar and Messi, as well as Sánchez, Suso, and Isigne. Break-out stars Coman & Thauvin are also present, as is a surprise inclusion of Max Gradel.

Finally, we use the model to look closer at two right-sided full-backs: established international Kyle Walker (27) vs 19-year old starlet Trent Alexander-Arnold. From the radar plot we see that Walker edges it on pass volume and overall success, while TAA is more progressive with his passing and has an expected output much higher than his sole recorded assist. In contrast, Walker seems to have overperformed with his 6 assists. The heatmaps indicate that TAA has a more direct style, with more passes behind the defensive line. Walker is more likely to cut back into the 16-yard box from very advanced positions – possibly a passing style underrated by the model. Walker also seems to recycle the ball more in the opposition half where TAA's passing is more progressive, fitting with City's more elaborate playing style versus Liverpool's more direct approach.

TAA also outperforms both the model and Walker in the opposition half and the final third, on the flip side his high-risk pass success is non-existent. Given his development since this season, we can argue that the model has a point about young Alexander-Arnold.

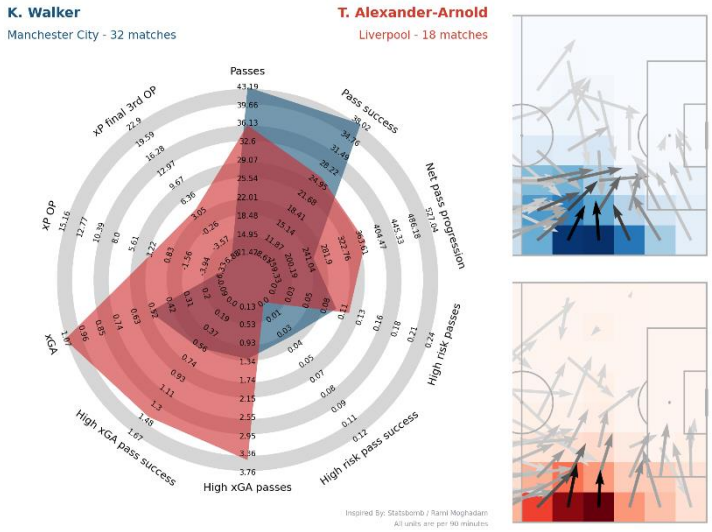

Figure 1: Style comparison between Kyle Walker and Trent Alexander-Arnold. Radar outer bounds at 99th percentile for all players in group. Pass heat maps indicate where passes are taken from, length of arrows indicate average pass length. Darker means more passes.

[1] All prices according to transfermarkt.com