A data-driven approach to classifying difficult passes and ranking players proficient in making them

Christian Gilson
Mathematical Modelling of Football

A RECRUITMENT CHALLENGE

In 2002 the Oakland A's baseball team famously applied data science to player action data to produce player KPIs in a way that would forever change sports team recruitment. One particular metric, wins above replacement, allowed the A's to quantitatively arbitrage low cost, traditionally unconventional talent to unearth position-specific diamonds in the rough.

Applying Moneyball strategies to football has proven to be a much harder challenge. Since 2014, after the stellar purchases of Luis Suarez, Marc-Andre ter Stegen and Ivan Rakitic, Barcelona have spent over €800m on new players, without a single new recruit truly emerging as a first team leader — and Barcelona possess one of the best sports data science functions on the planet. What makes football a different animal is that it's an invasion game featuring complex inter-player dynamics and valuable game-changing contributions happening off-the-ball. It's tricky to tease out an individual's impact on team performance, and even trickier to know how a new recruit will gel with their new team.

Rather than trying to come up with one-model-to-rule-them-all that'd solve all of Barcelona's recruitment woes, here we look to adopt a practical blend of data science and domain expertise to find a particular breed of footballer that fits a pre-specified profile. In this instance, we're looking for a forward player from the English Premier League with a talent for passing, being especially adept at pulling off a difficult pass.

There are two key steps to approaching this problem: firstly to prepare features that we think impact the outcome of a pass, and secondly to fit those explanatory features with a statistical model that expresses the probability that the pass will be successful.

FEATURE ENGINEERING

Using Wyscout on-the-ball event data, it was straightforward to extract positional x,y coordinates for the start of each action, as well as filtering the full set of event actions by open play passes and free kicks (with an intent to pass). In this work, these x,y coordinates are referred to as *basic features*.

We then produced derived geometric features: the squares and product of the initial coordinates x^2, y^2, xy, the initial goal angle, as well as the x,y coordinates of where the pass ends up. The type of pass was also treated as a feature. We refer to the combination of derived geometric features, pass type, and basic features as *added features*.

We then started to add more situational features. In this recruitment challenge, we wanted to be able to differentiate (and thus have the model predict different success probabilities) between situations that make passing easier and harder. If it's more difficult for a player to successfully complete a cross during a counter attack when their team is 4-0 down, away from home, and down to 10 men, rather than a similar crossing opportunity whilst cruising to victory at home in a game where the opposition suffered an early red card, then those features should be considered. We however did not include features that we wished to summarise with or normalise over in the resulting analysis, such as player role and whether or not the pass was played with the player's weaker foot.

Whilst we know the position of the ball before and after the pass, we're blind to the off-the-ball context. A team's star trequartista may struggle to pass through a team effectively if man-marked or quickly double-teamed. To try to capture off-the-ball context representing defensive pressure, we engineered the on-the-ball data to produce the duration of the current passing sequence, the duration the passer has been in possesion for before attempting to pass, and the passing index within the sequence (i.e. the n^{th} pass within the possession).

Finally, we encoded the passing angle and change in shooting angle before and after the pass, as well as the passing distance and the distance to goal. Passing angle is defined such that it's positive for forward passes, negative for backward passes, and symmetric as to whether a passer passes left or right, thus allowing the model to reward successful forward passing players looking to penetrate opposition lines. The list of situational and contextual features (not including interaction terms or additional powers of features), beyond the *added features* to produce our *advanced feature* set are as follows:

- Binary home flag;
- Binary counter attack flag;
- Passing distance;
- Distance to goal;
- Change in shooting angle;
- Game state (the point-in-time difference in goals scored between the two teams);
- Headcount difference (e.g. is equal to 1 if 11 Vs 10);
- Third transition delta (e.g. is equal to 2 if ball passed from defensive third to attacking third);
- Cumulative team possession sequence duration;
- Player possession duration;
- Passing index within possession sequence.

MODELLING

As we applied our three sets of features to build statistical models to predict pass success, we wanted to balance the principle of parsimony — trying to explain phenomena with the

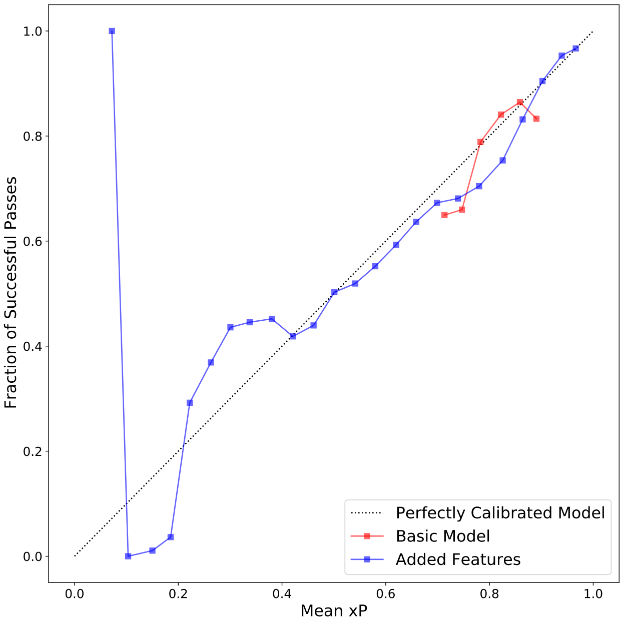

FIG. 1: Calibration plots for the simple and added feature sets using logistic regression.

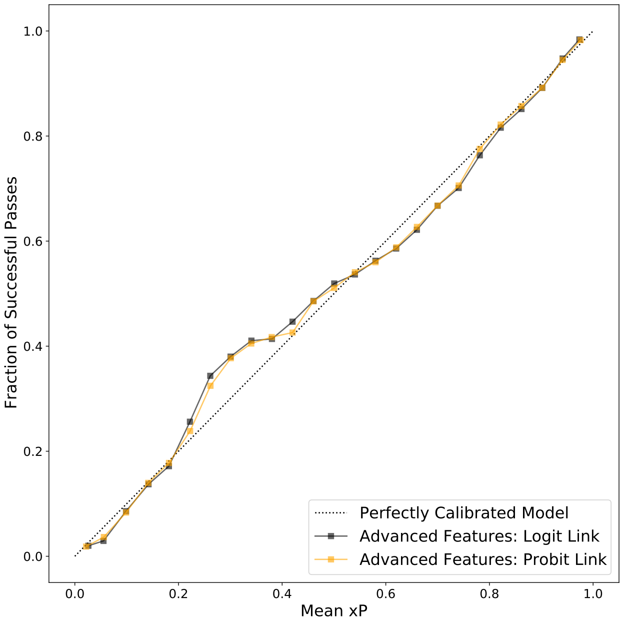

FIG. 2: Calibration plots for the advanced feature set using logistic and probit regression.

fewest features possible — with the objective of producing a model with the best calibration curve, a plot of the fraction of actual successful passes calculated within bins, against expected pass success. We used a data-driven approach to classify difficult passes, with a difficult pass defined as one which the best model predicts is more likely to fail than it is to succeed, i.e. the expected pass success is less than 50%.

Passes are Bernoulli events — they can either succeed or fail — so we needed mathematical machinery to transform our feature sets that enter the model linearly, into an expectation for pass success between 0 and 1. This expectation was defined as xP, and the machinery we used was a generalised linear model (GLM) configured to output an outcome from the binomial distribution (of which Bernouilli is a special case). The tools within this machinery that add the neccessary dash of nonlinearity between our linearly combined inputs and output xP are called link functions. We experimented with two slightly different S-shaped link functions that have the useful propertly of ranging between 0 and 1: the logit and probit link functions. All of this means we attacked the modelling problem with logistic and probit regression.

Logistic and probit regression models were trained on a statified sample of 70% of the entire Wyscout dataset containing nearly 2 million passes (spanning all domestic leagues and international competitions), leaving 30% left for testing purposes, such as calibration.

For this data-driven approach of defining and analysing difficult passes to be effective, it’s essential that the model is well calibrated at the more difficult end of the spectrum, where a closer lie to the ”perfectly calibrated” curve can be directly interpreted as a confidence level in those predictions. This means that the percentage of passes that were successful within a bin of passes attributed with xP values of, say, around 0.2, should be close to 20%.

The calibration plot in FIG. 1 shows that the *simple* logistic regression model completely fails to span the difficult range of xP, instead narrowly centering around the overall average pass success rate of 83%. The *added* logistic regression model fares better, but is poorly calibrated for xP values below 0.5, thus providing little confidence that that model could effectively classify a difficult pass, nevermind distinguish between a 10% Hollywood ball, and a pass far closer to 50-50.

The curves in FIG. 2 show that *advanced features* really made a big difference. Both models are well calibrated barring a noticable ≈10% underestimate at xP values of around 0.3. The Brier score for a model’s performance on test data provides the mean squared error of the xP forecast, effectively summarising the calibration curves into a single statistic. Brier scores for the four fitted models are summarised in TABLE I, alongside AIC scores, a measure of model parsimony that trades model deviance — the GLM analogue of the sum of squares — off against the number of parameters, where a lower score, as with Brier, is better.

Model	Brier Score	AIC Score
Basic Logit	0.139	1.14M
Added Logit	0.112	927k
Advanced Logit	0.107	878k
Advanced Probit	0.106	875k

TABLE I: Brier and AIC scores for the four models fit.

Changing link function from logit to probit has a minor but noticable impact both visually with a tighter calibration curve and statistically with superior Brier and AIC scores, however, the real stars of the show are the *advanced features*, highlighting the importance of our situational and contextual features.

A simple goodness-of-fit hypothesis test asking whether or not our probit model is missing significant features (when compared against a perfectly fitting theoretical saturated model with many more features) emphatically suggests that we are, which is no surprise. Our on-the-ball data has limitations, such as us not knowing the intended receiver and their position for unsuccessful passes, and our derived off-the-ball context is far from comprehensive.

For the purpose of this recruitment challenge, our probit model was fit for purpose for classifying difficult passes, which allowed us to rank players with proficiency in this specific skill.

RANKING PASSERS

Our model was applied to the full Wyscout dataset to produce an xP value for every pass. We then targeted Premier League forwards, and filtered out any pass with an xP of 0.5 or greater, leaving us with just the difficult passes as defined by our data-driven methodology.

We then attributed the *Over xP* statistic to each pass, subtracting xP from the binary pass outcome, rewarding the successful completion of highly difficult passes hansomely, and penalising missed opportunities to pull off a 50-50 ball more than a failed attempt at a Hail Mary. Finally, we aggregated the Over xP statistic per player by summing over all difficult pass attempts, and normalised per 100 passes, producing a ranking with the top ten passers shown in TABLE II.

Player	Over xP / 100A	Attempts	% Difficult Pass
E. Hazard	13.9	123	8.8
W. Rooney	8.4	108	9.0
S. Aguero	8.1	52	8.5
S. Mane	6.5	104	11.0
J. King	6.1	104	16.2
A. Sanchez	6.0	236	15.7
T. Walcott	4.5	69	19.4
A. Lacazette	3.6	57	8.7
A. Barnes	3.4	55	13.4
A. Martial	3.2	99	12.9

TABLE II: Top ten ranking of Premier League forwards by Over xP per 100 pass attempts.

Chelsea's Eden Hazard is the clear standout player, with a normalised Over xP score that's nearly double that of Manchester United's Wayne Rooney in second place. It's not a surprise to see Rooney so high on the list, as him dropping deep to receive the ball and attempt a killer pass became a familiar sight in his later years at United, but Hazard's ranking *is a little unexpected.* A Hazard YouTube montage will consist almost entirely of sublime dribbling and solo goals, but this analysis shows that moments of his individual brilliance extend to tricky passes that create opportunities for others.

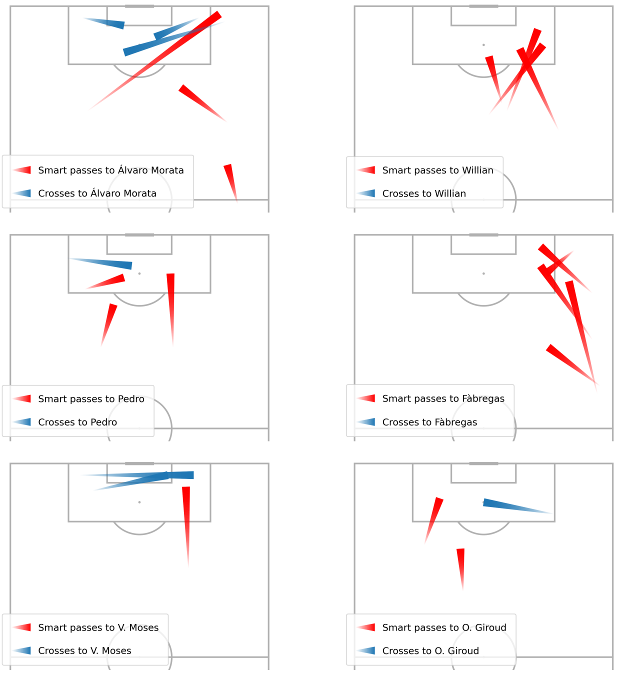

FIG. 3: High difficulty smart passes and crosses by Hazard to his six main targets in the 2017/18 Premier League season.

Chelsea often played a 3-5-1-1 formation in the 2017/18 season with Hazard given a free role behind either Morata or Giroud, with Willian, Fabregas, Pedro, and Moses often moving into advanced positions beyond Hazard during the transition from defence to attack. FIG. 3 displays these six main targets on the receiving end of Hazard's smart passes and crosses, highlighting him being adept at playing in midfield runners with penetrating smart passes into the opponents box, as well as cutting the ball back accross the face of goal or towards the penalty spot to set up shooting opportunities — it's remarkable that he only ended up with 4 assists all season!

WRAPPING UP

It's fair to say Hazard is no diamond in the rough, and it's no marvel of data science to find that a player Real Madrid paid over €100m for tops our list. Bournemouth's Joshua King however, ranking fifth, looks to be undervalued with a point-in-time Transfermarkt value of £13.5m, a steal compared to most of the other players around him in the top ten.

A natural extension of this work is to look at pass *threat*, using an approach like Karun Singh's *xT* to complement our xP, enabling us to investigate whether difficulty translates to goal threat.