The quest for the right pass: quantify soccer player's decision making

Ali Bakkoury MSE, Hugo Artigas MSBA

http://statsbomb.com/wp-content/uploads/2021/11/Javier-M-Buldu.pdf

Introduction

The quest for the right pass: Quantifying players' decision making - Javier Buldu & Borja Burriel (Statsbomb conference 2021)

- **Objective:** Building a mathematical model to quantify the quality of a pass in a football game, Measure each pass' risk and estimated gain
- **Our approach:** Build risk and gain metrics using a model to estimate the probability a pass is completed

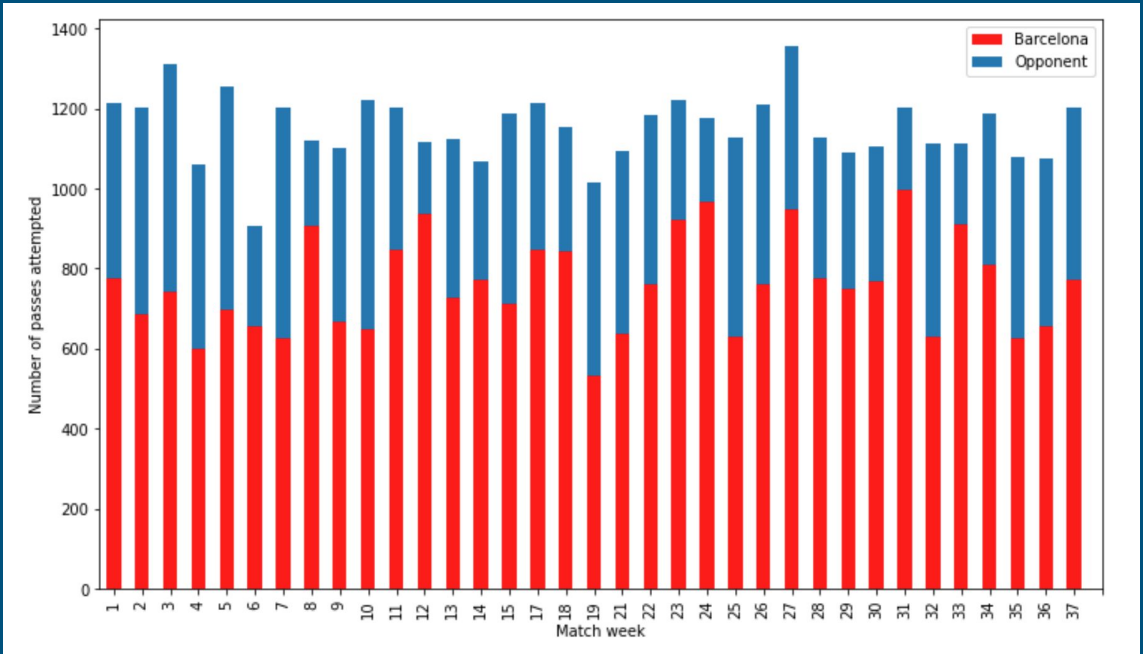
Statsbomb dataset

- 40,337 passes from Barcelona's games in the spanish league (La Liga) 2020/2021 season
- 35 features
  - Match features: date, period, minute, opponent, home/away
  - Player features: position, best foot
  - Pass features: location, length, height, angle, player situation (under pressure…)

Target

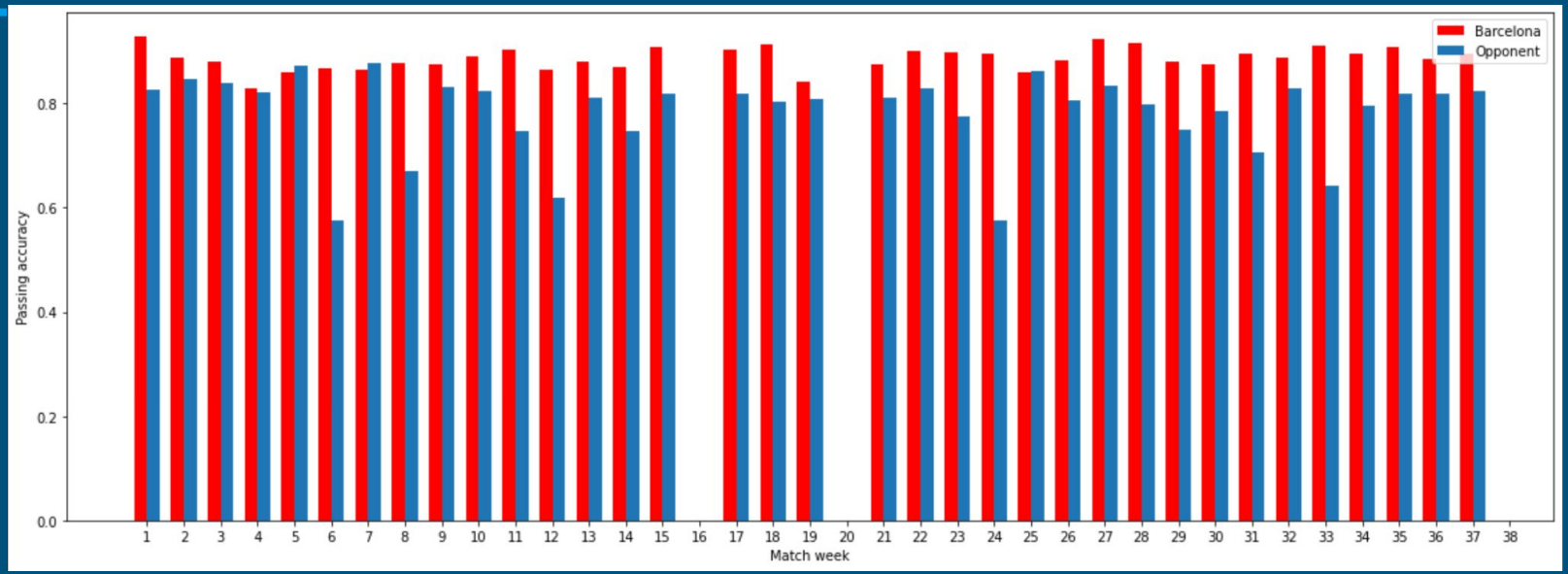
id	period	minute	possession_team	location	under_pressure	angle	…	outcome
1	1	5	Barcelona	[60,40]	True	-2.8	…	Completed
2	1	8	Real Madrid	[43,37]	False	3.0	…	Failed

Number of passes is biased towards Barça



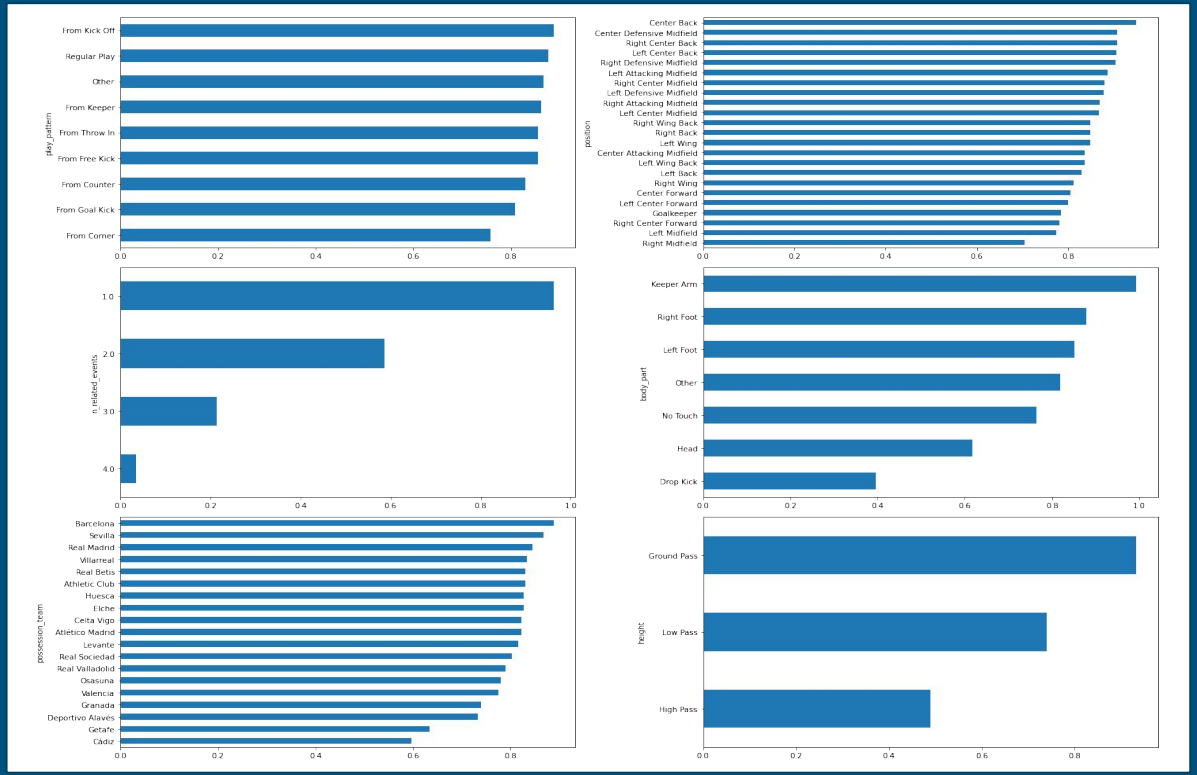
- Our model will be trained on a high proportion of Barcelona passes
- We expect it to be specifically good at predicting La Liga (specifically Barcelona) players' passes outcome

Failed passes represent ~15% of the data

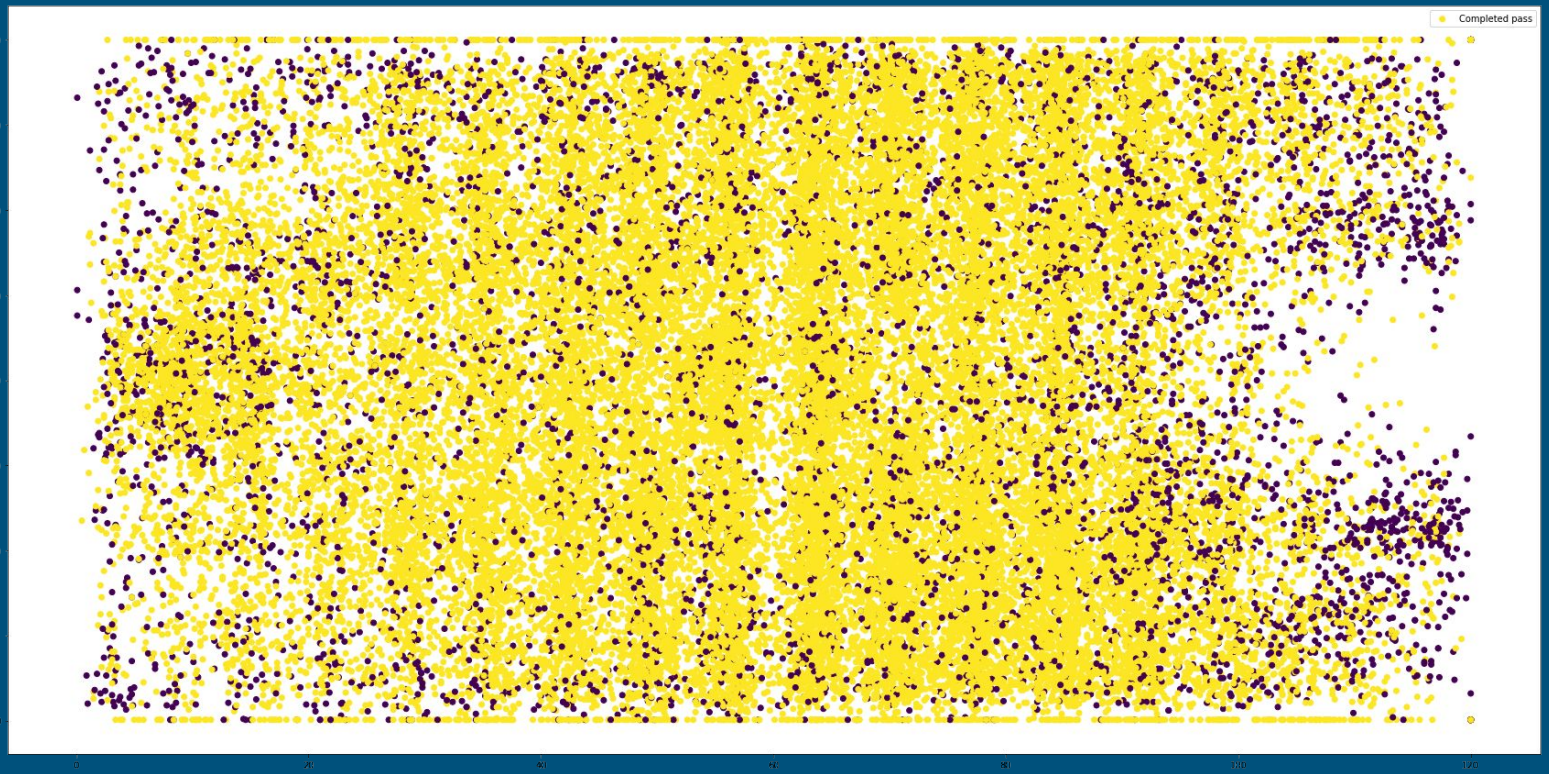


- Thankfully for football fans, players complete most of their passes
- Our dataset is imbalanced towards completed passes!

Categorical variables influence pass outcome



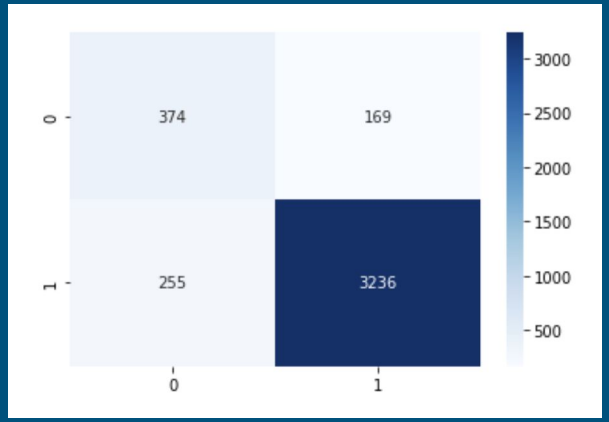
The closer to the goal, the harder the passes



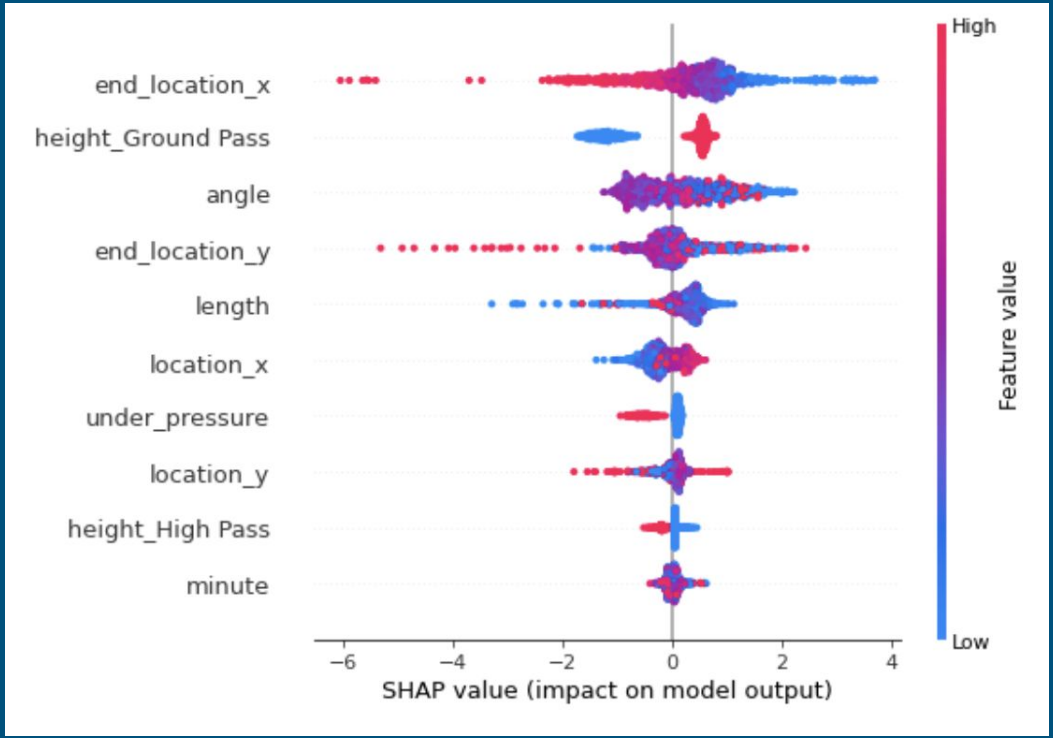
Building the passes' outcome classifier

- Gradient boosting classifier (XGBoost)
- 15 modeling features (12 numerical and 3 categorical)
  - Engineering new features on players historical passing accuracy
  - One hot encoding categorical data
- Model evaluation on holdout test set

  91.5% AUC score



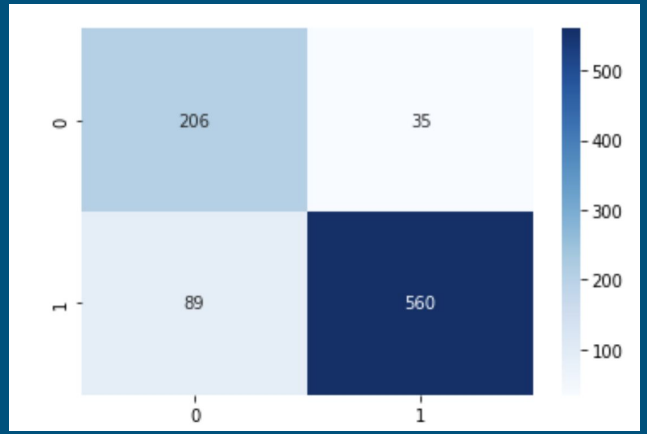
It is hard to complete a pass when pressed in the opponent's area



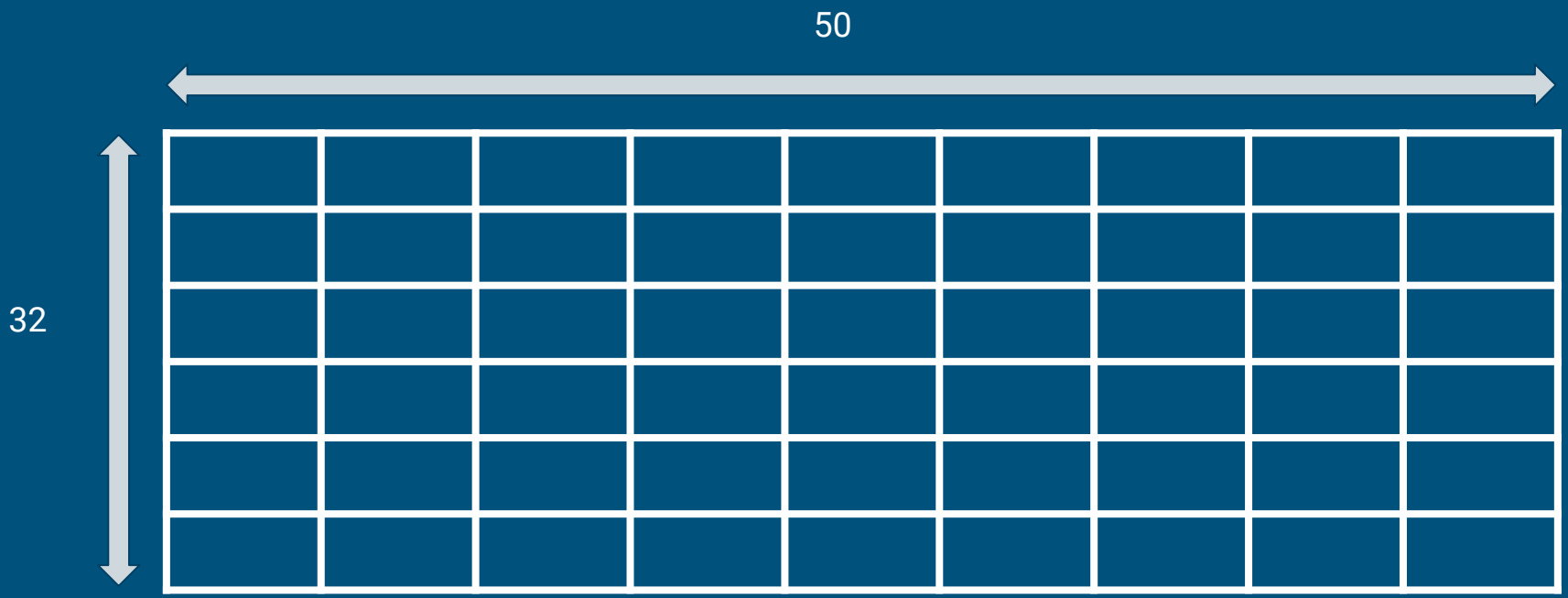
Can the model be accurate in any league?

- The model is trained on limited data from 1 league and 1 team games. Can it be accurate in other leagues?
- We gathered and processed passes data from the 2019 Champions League final (890 data points)

92.4% AUC



Divide the pitch into zones



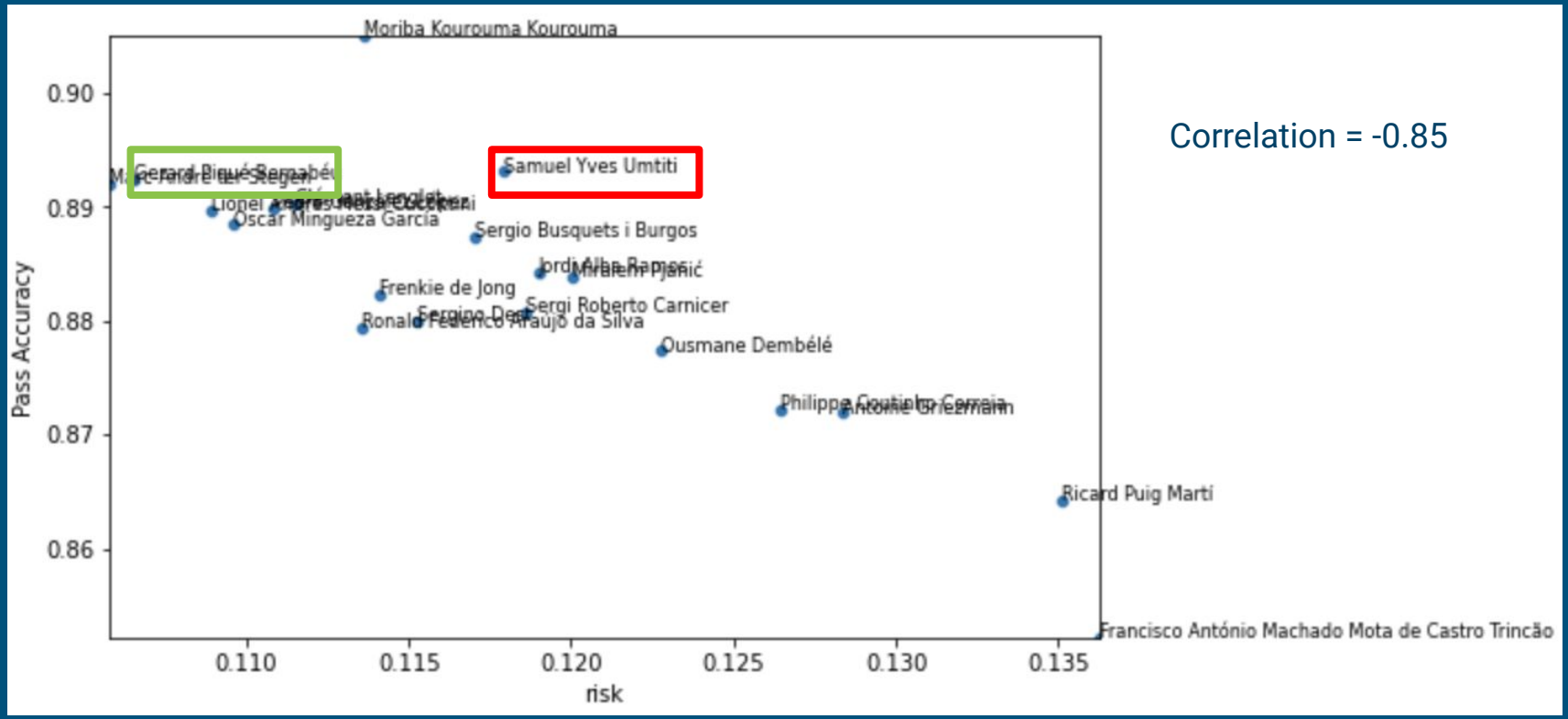
Use of the model outputs to define the average pass success rate: **the risk**

- Take a pair of zone (b_i,b_f).
- Compute the average probability (output of the model) of all the passes from b_i to b_f. Let's call this probability $p_{i,f}$.
- The risk is defined as:

$$r_{i,f} = 1-p_{i,f}$$

- Interpretation of the risk: for a pass from b_i to b_f, there is a probability $r_{i,f}$ that in average the pass will not be complete.

Pass accuracy vs. risk



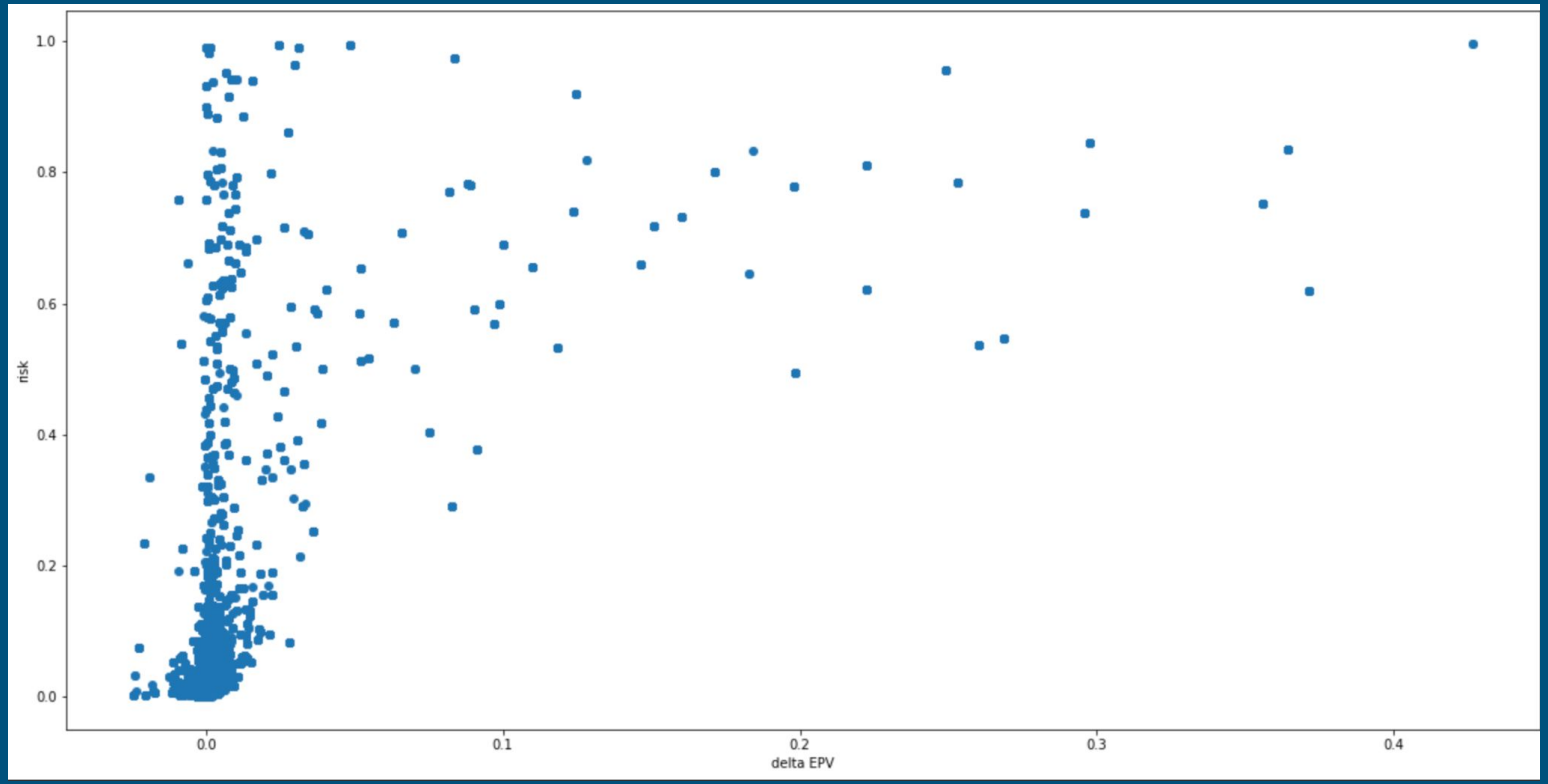
Correlation = -0.85

Define the Expected Possession Value (EPV)

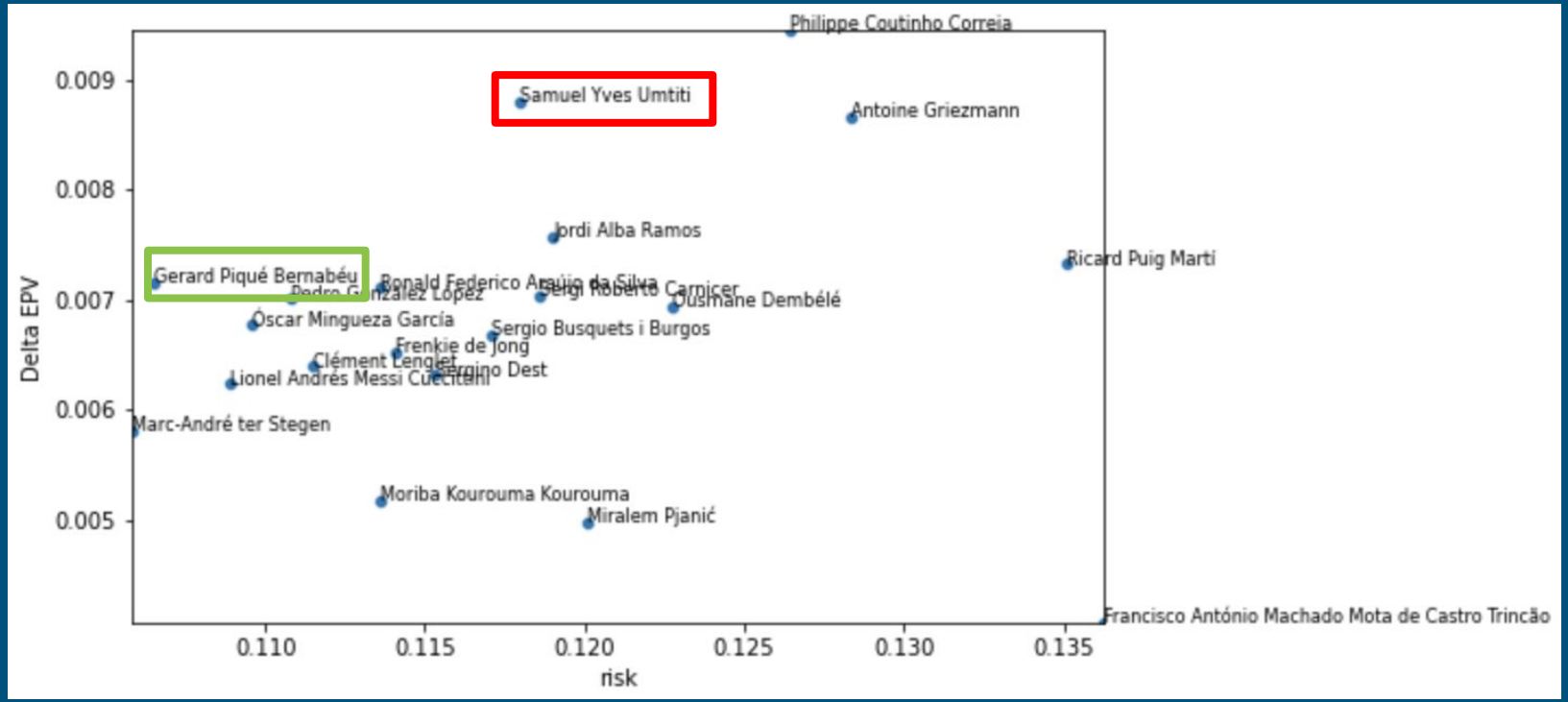
- The EPV of a zone *b* is the probability that an action starting from *b* finishes in a goal.
- To quantify the incremental probability generated by a pass from b_i to b_f, we define:

$$\Delta EPV_{i,f} = EPV_f - EPV_i$$

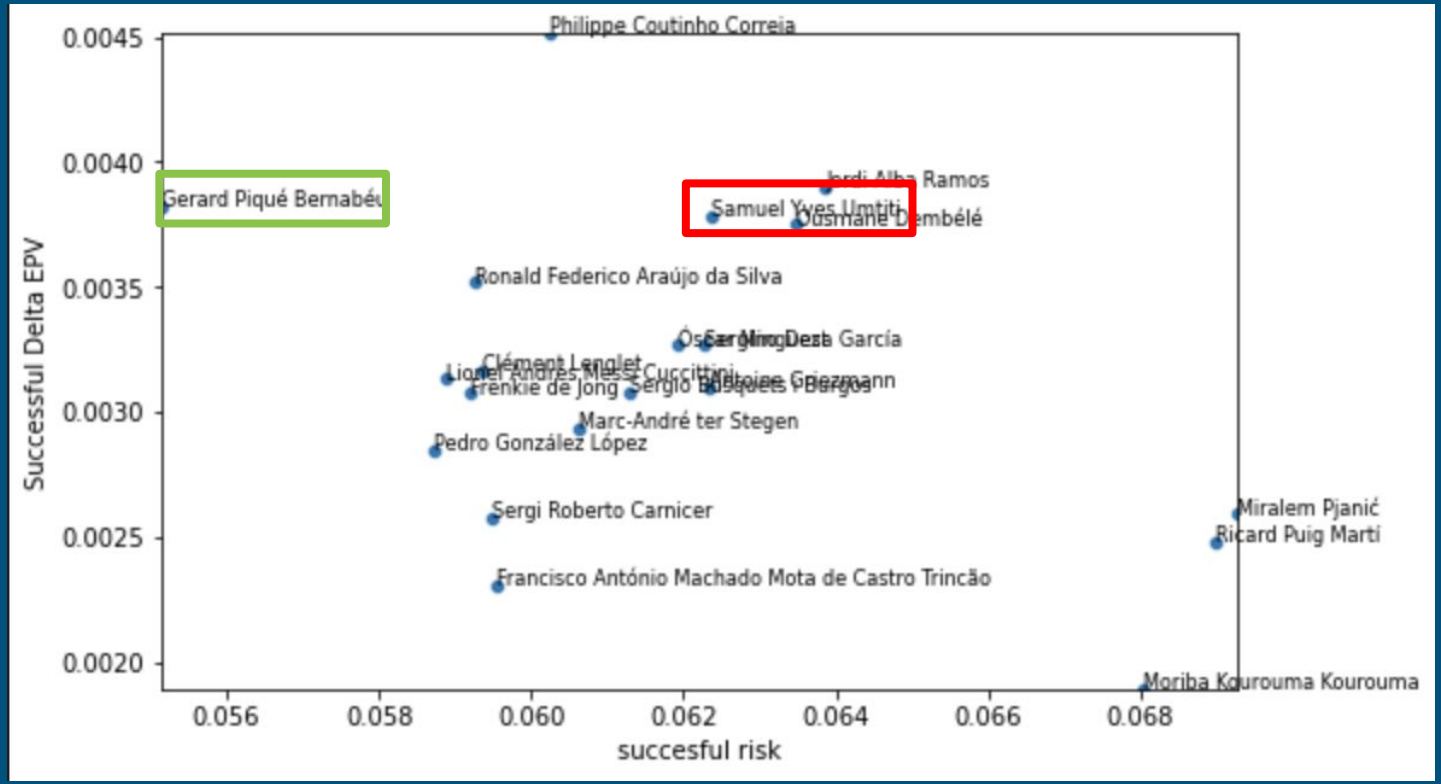
Distribution of the passes



Player pass choice evaluation



Player pass choice evaluation for the successful passes



Next Steps

- Have access to more Statsbomb data (i.e. position of other players on the pitch) to improve in the model.
- It will allow to determine for any pass the best teammate to give the ball according to his position on the pitch.