Caglar Altunel
caglar.altunel@student.nhh.no

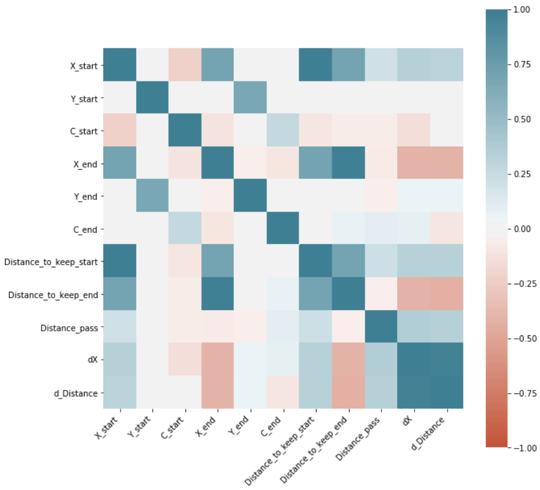
The xP Index

Short intro

In this paper, I derived an xP Index by using a logistic regression model, which has distance to the keep at pass start location and the difference of distances to the keep between pass start and pass end locations as continuous features. I enriched the model by making use of head pass, high pass and simple pass as categorical (binary) variables. By using this model, I classified good passers by filtering some of the players with respect to certain metrics. Afterwards, I put them on a scatter plot, where x-axis is pass accuracy and y-axis is xP Index, to point out a number of good passer candidates. I used Wyscout's Premier League data and used a subsample of 50,000 passing events due to computational reasons.

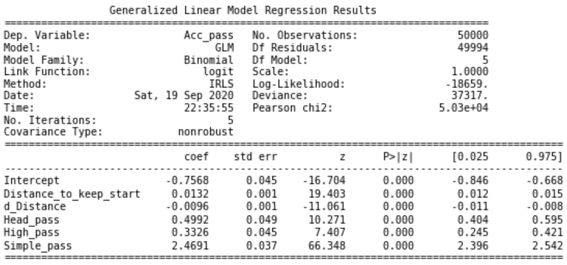
Building the model

After creating the target variable as pass accuracy (1 if the pass is successful, 0 otherwise.), I built an initial dataset with columns containing several features such as X and Y coordinates of the pass start and end locations, distance to keep at the pass start and end location, the difference of distances to the keep between pass start and pass end locations, pass distance. After data building, I reached a poor outcome by running the model where X and Y coordinates of the pass start location as features. Its quadratic form did not lead to any change in terms of model quality. However, I reached a remarkably higher log-likelihood of the model when I added all features that the dataset covers. But, due to bias variance trade off, we know that models with high number of features are good at capturing all sort of variations in the data, while they usually provide poor results when being tested on another dataset. This model indeed could be reduced to a reasonable number of features, as the correlation heatmap shows that some continuous features are strongly correlated with each other and therefore, there is no need to use all of them.



Correlation Heatmap

Due to this, I simplified the model by using only distance to the keep at the start point of the pass and the difference of distances to the keep between pass start and pass end locations as continuous features, while head pass, high pass and simple pass are used as categorical (binary) variables. The summary table of the model can be seen in the below:



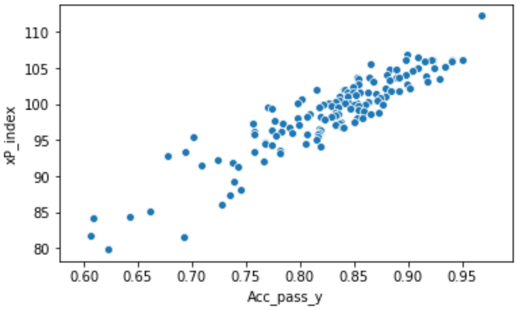
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:               Acc_pass   No. Observations:                50000
Model:                            GLM   Df Residuals:                    49994
Model Family:                Binomial   Df Model:                            5
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -18659.
Date:                Sat, 19 Sep 2020   Deviance:                       37317.
Time:                        22:35:55   Pearson chi2:                 5.03e+04
No. Iterations:                     5
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                 -0.7568      0.045    -16.704      0.000      -0.846      -0.668
Distance_to_keep_start     0.0132      0.001     19.403      0.000       0.012       0.015
d_Distance                -0.0096      0.001    -11.061      0.000      -0.011      -0.008
Head_pass                  0.4992      0.049     10.271      0.000       0.404       0.595
High_pass                  0.3326      0.045      7.407      0.000       0.245       0.421
Simple_pass                2.4691      0.037     66.348      0.000       2.396       2.542
==========================================================================================
```

I tested the xP model by rerunning it in a train data and it provided a high accuracy in classifying passes on the test data. Which is why, I concluded that the model is reasonable to build an xP Index with.

How xP Index is derived & How it works

After model building, I named the probabilities calculated for each observation as xP, and derived the xP index for each player simply by sum of the accurate passes of the given player divided by the sum of xP of the same player, and multiplying it by 100. Then we can simply interpret that players who have an xP Index of higher than 100 exceeds the expectations (completes more successful passes than the model estimates.).

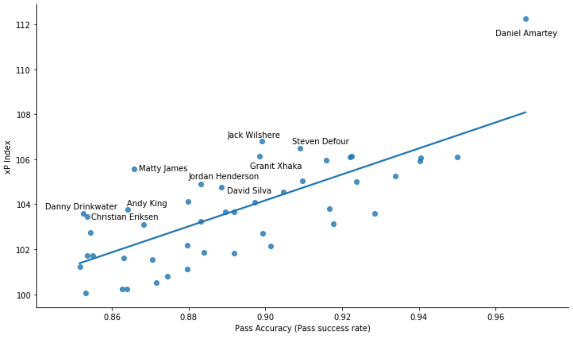
After that, I filtered the outcome only for midfielders and put it onto a scatter plot where x-axis is the pass accuracy (mean accurate pass) and y-axis is the xP Index. Not surprisingly, they are highly correlated (the graph is in the below.) but I conclude that it has enough variation to derive some interesting analysis out of it.



Findings about Midfielders

First, I focused on the top right corner of the above graph, where I believe the elite passers are at. To that end, I take only players who has higher than 200 accurate passes (within the subsample of 50,000), pass accuracy of higher than 85% and xP Index of higher than 100 into account and I ended up with a list of 15 midfielders. These players are: Fernandinho, Fabian Delph, Granit Xhaka, Ilkay Gundogan, Nemanja Matic, Jordan Henderson, David Silva, Paul Pogba, N'Golo Kante, Mesut Ozil, Christian Eriksen, Moussa Dembele, Dale Stephens, Eric Dier and Idrissa Gueye. Oriol Romeu is listed out as it could not exceed 100 xP Index threshold, even though the player achieved quite high pass accuracy.

Afterwards, I decided to relax the restrictions and re-filtered the players without a minimum number of successful passes, to see the players who had not reached enough number of successful pass, but possess a potential to be a very good passer in the Premier League. The scatter plot that contains these players can be seen in the below.



I focused on the players being located noticeably above the regression line, as these are the players having much higher-than-estimated xP Index with respect to their pass accuracies. In other words, these players tend to complete harder pass attempts successfully, as they managed to score a high xP Index than the model estimates.

I intentionally put the names of three world class midfielders, Wilshere, Henderson and David Silva, to make them as a benchmark to the rest (Wilshere is obviously a great passer.). According to my findings, Steven Defour, Granit Xhaka, Matty James, Andy King, Danny Drinkwater and Christian Eriksen are effective passers, as I conclude they tend to complete passes in tougher situations as they achieved to reach a high xP Index with respect to their pass accuracies. Some of them are potentially out of the range of the scouting departments of several clubs due to their relatively low reputation compared to globally known passing expert midfielders. Which is why, I conclude that if I were to report effective passers who are underdogs at the same time, along with the representation of the model and xP Index, I would give some these players' names, such as Matty James, Andy King and Steven Defour, to the chief scout of team (Daniel Amartey constitutes an exception as it is in the list by covering only 30 accurate passes. I would continue to follow his development before reporting his name to the chief scout.).

Strengths and weaknesses of the model

Thanks to the interpretable nature of logistic regression (linear in the log of the odds.), the model can serve the needs of both classification and causal analysis, unlike many other machine learning algorithms. Also, logistic regression enables us to assign a probability to each observation, which constitutes the base of the xP Index.

On the other hand, the model does not take into account whether the pass action occurs under pressure or not and the number of opponent players close to the pass location, between the keep and pass location etc. Which is why, I can conclude without hesitation that there is a lot of room for improvement in this model. Anyhow, it provides a fair generalization in terms of how to identify effective passers among all other players having high pass accuracy rate.