Assignment 2: Evaluating players using machine learning

Arin Rahim

October 2024

Contents

1 Introduction

Going into the season 2018/2019, Leicester city FC are competing in the Premier League, EFL cup, and the FA cup. These are three competitions that require high-level players with both physical and technical abilities to adapt to the tight schedule and the match intensity that the games bring. The squad we have right now looks promising, but reinforcement to our midfield is needed since we lost our two most prominent midfield players N'golo Kanté a couple of seasons ago and Danny Drinkwater last season.

For this summer transfer window, evaluation of the midfielders in the top European leagues has been done using data from multiple matches and a machine learning model to predict each players expected passing accuracy (xPA). The following document will give an explanation of the measure and assumptions of the machine learning model, the result, and a more advanced explanation about the creation of it.

2 Measures and Assumptions

A definition of what type of midfielder we are on the lookout for was done and the conclusion was a player with a high xPA. Having a player with this skill improves the team's ability to maintain possession and enhance the possibilities to break through the first line of press, initiating attack towards the goal.

While we could simply assess a player's passing accuracy by looking at their total number of passes and counting the accurate ones, we also need to consider additional factors to properly evaluate the difficulty and the outcome of each pass. Difficulty of a pass can be defined by looking at the distance the pass covered, angle of the pass, the body part used, the type of pass, and the circumstance of the pass for instance was it made under pressure. The distance of a pass reveals if it was short or long, the angle of the pass can tell us which direction the pass went, the body part used for a specific pass tells us the players technical abilities to make a pass, the type of pass can be simple, cross or another variation, and the circumstance of the pass reveals the players ability to make quick and accurate decisions in stressful situations. The outcome of the pass can be if it lead to a counter attack, if it was a through ball, or a key pass that gained significant advantage in attack. The model takes into account all the assumptions of a pass and tries to predict midfielders in the top European leagues - specifically England, Spain, Italy, France, and Germany - expected passing accuracy and compare the expected value with their actual value to assess a player's performance in terms of passing.

3 Result

The model predicted that many players in the top five European league performed better than their expected passing accuracy, while some performed less. Table 1 summarizes the top five best players in Europe, including current club, xPA, and their actual passing accuracy.

Name	Team	xPA (%)	PA (%)
Moussa Dembélé	Tottenham Hotspurs	90.0	93.2
Fabian Delph	Manchester City	89.4	94.5
Adrien Tamèze	OGC Nice	89.2	92.5
Nampalys Mendy	OGC Nice	89.1	92.5
Javi García	Real Betis	89.0	94.1

Table 1: Top five European players and their respective xPa together with their actual passing accuracy for comparison.

The surprise with the result is that Nampalys Mendy — who was out on loan from us to OGC Nice — was ranked at 4Th place according to the model's prediction. Initially, keeping him would be ideally since he is already familiar with the club. To further assess which of the players to sign, I looked at their age and current market value, see table 2.

Name	Market Value (€)	Age
Moussa Dembélé	18	30
Fabian Delph	15	29
Adrien Tamèze	4	24
Nampalys Mendy	5	25
Javi García	4	31

Table 2: Top five European players and their market value together with their age, from Transfermarket.com

Nampalys Mendes being the next youngest after Adrian Tameza, and also having a low market value indicates that keeping him would be great since he is young and there would be no need to bring in a player of similar profile, saving us from unnecessary spending of the transfer market budget.

4 Conclusion

As a conclusion, keeping Nampalys Mendy would be ideally since he has one of the highest expected passing accuracy in Europe and already is included in our squad. However, further research can be done to reach an ultimate solution to find the proper midfielder.

5 Creation of Model

This section is meant to give an in-depth explanation of the model, together with the data, the features, and target used for the evaluation.

5.1 Wyscout Data

The data was sourced from Wyscout and includes information on teams, players, and match events was from the 2017/2018 season, as well as the 2016 European Championship and 2018 World Cup. However, only events from England, Spain, Italy, France, and Germany first divisions was utilized. For each event, the passes made by midfielders were extracted along with the tags that provide additional details about each pass, see table 1. A threshold was also added to remove players with less than 500 passing attempts to only include midfielders that are heavily involved in a teams possession.

Action	Tag ID
Left Foot	401
Right Foot	402
Head Body Pass	403
Through ball	901
Under Pressure	2001
Key Pass	302
Counter Attack Pass	1901
Blocked	2101
Interception	1401
Accurate	1801
Not Accurate	1802

Table 3: Each possible action of a pass. One pass could include multiple actions.

Each pass in the data was also divided into the type of pass but they had no specific Tag ID. Hand pass was removed from the dataset but the rest can be seen in table 2.

Pass
Simple Pass
High Pass
Head Pass
Smart Pass
Launch
Cross

Table 4: Different types of passes a player could perform.

5.2 Features

After the data cleaning process, the features x_i were generated. A column for each action was created, where a value 1 indicates that the pass included the specific tag ID and the type of pass, and a value of 0 indicates that the tag ID and type of pass were not present for that specific pass. In addition to previously mentioned actions, two additional features were created. The first feature was the distance of the pass, calculated using included coordinates of the ball and applying the euclidean distance formula for each pass. The second feature was the angle of the pass, calculated using trigonometry and the coordinates. The end result for a pass angle was a value between 0 to 360, with 0 and 360 meaning that the specific pass was going straight forward.

5.3 Target

The intended target was the accuracy of each pass. It was classified in the dataset as either 0 for inaccurate or 1 for accurate.

5.4 Logistic Regression Model

In a logistic regression model, a set of variables is given to the model and the goal for it is trying to predict a binary outcome, usually represented as 0 or 1. In this case, the variables given to the model were the features and the outcome of the pass, either accurate or not, as the binary target. The model is trying to predict a probability value between 0 to 1 for each pass and does so by using the sigmoid function. if the probability of a pass is 0.5 or higher, it is classified as accurate; if the probability is lower than 0.5, it is classified as inaccurate.

Before training the model, the data was split into a training set and a test set. The purpose of the training set is to teach the model, while the test set is used to evaluate the model's prediction accuracy. In this case, the training set only included event data from the Premier League, whereas the model was tested both on Premier league data and on the Spanish, Italian, French, and German first division league data.

During the training of the model, the features are adjusted to maximize the probability of predicting the correct result. This is done through a method called maximum likelihood estimation (MLE), where the goal is to find the set of coefficients that maximize the likelihood of observing the actual outcomes given the model. MLE adjusts the model's parameters so that the predicted probabilities are as close as possible to the observed results.

6 Evaluation of Model

To evaluate the model's ability to predict the expected passing accuracy after training it, two metrics were used. First, ROC AUC score revels how well the model distinguishes between the two classes (0 or 1). Second, the log-loss value assesses how far the model's prediction are from the actual outcome. The metrics was first used to evaluate the models prediction for the players in the Premier League, and then again together with the the players from the other leagues. The score showed slight variations when the model was tested on the dataset containing players from the top five European leagues. The result for each metric can be seen in table 3.

Metric	Data	Value
ROC AUC	Premier League	0.78
Log-Loss	Premier League	0.34
ROC AUC	Top Five European Leagues	0.77
Log-Loss	Top Five European Leagues	0.35

Table 5: The two metrics with each of their score for both the Premier League and top five European Leagues, which includes the Premier League.

For the ROC AUC score, a value of 0.78 and 0.77 indicates a fairly good model performance but still room for improvement. Similarly, the Log-Loss scores 0.34 and 0.35 are considered reasonably good but are in need of improvement as well.

Further evaluation was done by analyzing the classification report, which provides metrics such as precision, recall, f1-score, accuracy, macro average, and weighted average for each class. Precision measures the accuracy of the model when it predicts a certain class, while recall measures how well the model captures all the actual instances of a certain class. The F1-score, the harmonic mean of precision and recall, balances both metrics. It provides a balance between precision and recall. Support represents the number of actual occurrences of each class in the dataset, and accuracy measures how often the model is correct overall. The macro average computes the average of precision, recall, and F1-score for both classes, giving equal weight to both classes, while the weighted average considers the support (frequency) of each class while calculating the average of precision, recall, and F1-score. Although some metric values varied depending on whether the report was from the Premier League or the top five European Leagues, the results between the classes were consistent, see table 4.

Class	Precision	Recall	F1-score	Support
Premier League				
Inaccurate	0.70	0.31	0.43	3765
Accurate	0.88	0.98	0.93	20102
Accuracy	0.87			23867
Macro avg	0.79	0.64	0.68	23867
Weighted avg	0.86	0.87	0.85	23867
Top Five European Leagues				
Inaccurate	0.71	0.30	0.43	84057
Accurate	0.88	0.98	0.93	445538
Accuracy	0.87			529595
Macro avg	0.80	0.64	0.68	529595
Weighted avg	0.85	0.87	0.85	529595

Table 6: Classification Report for Premier League and Top Five European Leagues.

The classification report revels that the model performed significantly better on predicting accurate passes than inaccurate. This can be due to the imbalance of the dataset, which is dominated by accurate passes. Different balancing methods were tried without any improvement but a possible solution might be to train the model with data containing more inaccurate passes, which may increase the model's prediction ability across different classes.

7 Conclusion

The model predicted the expected passing accuracy for all midfielders in the top five European leagues and the result showed that many players performed better than their expected value, while some of them performed under their expected value. Even though the model's prediction ability can be improved, using the results together with other performance evaluation methods can help Leicester City FC bring the optimal midfielder to reach success during the upcoming season and the years to come.