Champions League Team Analysis from 2017-2022 Using Linear Discriminant Based Classifier

Independent Study Module: Data Science In Sports

Aryan Yadav, Eryn Wali, Pranav Iyengar

Abstract

In this paper, we propose a method for predicting the outcomes of football matches. Our approach is based on a combination of statistical analysis and machine learning techniques. We apply a Linear Discriminant Based Classifier to gauge which statistics affect winning the most in the UEFA Champions League. This process is done by running the statistics of all teams in the Round of 16 through a classifier, and understanding what metrics the classifier uses to predict teams.

In order to find data for teams in the Champions League, we used every single metric from FBRef's Champion League page from 2017 onwards since this data did not have any inconsistencies. After enhancing our test-train split on the above collected data, we ran it through multiple classifiers and optimized the same for KNN and SVM. KNN was optimized using hyperparameter tuning, by finding the 'K' that gave the highest accuracy. The optimized value of K found was 3, and the optimized kernel found was 'Linear'.

We faced a few limitations as well:

1. We were limited to only a few years of Champions League data as complete statistics were not available for years prior to the 2017-18 seasons.
2. There was not enough training data per team.
3. Some teams did not have complete data.
4. Finally, the varying amount of training data per team.

In order to gauge which statistics are most significant,we divided the analysis into 3 parts. Firstly, we analyzed the statistics of the teams that got classified correctly along with their year. Then, we compared the statistics of the teams that got classified correctly regardless of their year being inaccurate. Finally, we checked to see which teams got further in the competition - whether they got classified correctly, and why/why not.

From the above, we concluded that SOT%, xnpG, Short Cmp% and Medium Cmp% are the most significant statistics for the correctly classified teams. Furthermore, Dist, PSxG, and Short CMP% are also important for the same.

Introduction

Football is one of the most watched sports worldwide. FIFA estimated that at the turn of the 21st century, there were approximately 250 million football players and over 1.3 billion people “interested” in it. It generates a large revenue and has live broadcasting - more than 26 billion people watched the World Cup finals in 2010.

While there are several tournaments that occur on a global scale, one of the most popular football competitions is the Champions League. The UEFA Champions League is an annual club football competition organized by the Union of European Football Associations (UEFA). It is the most prestigious tournament in European club football and one of the most prestigious in the world. The competition involves the top teams from each of UEFA's member countries, and it culminates in a final between two teams at a neutral venue.

The competition was first held in 1955, and it has evolved over the years to its current format, which involves a group stage followed by a knockout stage. In the group stage, 32 teams are divided into eight groups of four, and each team plays the other three teams in its group twice, home and away. The top two teams in each group advance to the knockout stage, which is a single-elimination tournament featuring two-legged ties. The final is a one-off match, with the winner being crowned the champion of Europe. The Champions League also serves as a qualifying competition for the FIFA Club World Cup, which is an international club competition featuring the best teams from each of FIFA's six confederations.

Football prediction is the process of forecasting the outcome of a football match based on statistical analysis and past performance of the teams involved. It is a popular activity among sports fans, as it allows them to show off their knowledge of the game and compete with others in predicting the results of matches. Football prediction can be based on a variety of factors, including the strength of the teams, their recent form, injuries to key players, and the head-to-head record between the two sides. Professional football prediction services use advanced algorithms and data analysis to make more accurate forecasts, but even the best predictions are not always correct due to the unpredictable nature of the sport.

Results of the UEFA Champions League are based on many independent variables (predictors). Variables such as shooting, passing, defensive actions and possession are related to the result of the match which are considered as dependent variables. In addition to these variables, the methods used to predict the match result are important. When literature is examined, it is seen that machine learning algorithms are frequently used in predicting match results.

Through this paper, we aim to use machine learning to understand which statistics affect winning the most in the UEFA Champions League. This process is done by running the statistics of all teams in the Round of 16 through a classifier, and understanding what metrics the classifier uses to predict teams.

Methodology

Finding, Compiling, Cleaning and Labeling the Data

In order to find data for teams in the Champions League, we visited FBRef's Champions League page. Here, we were able to find data for all teams in the UCL Round of 16. However, only teams after 2017-18 had complete datasets, so we had to refine our data compilation to this time period.

To compile the data, we made use of FBRef's inbuilt features - modifying and exporting of tables. We compiled 5 different statistical groups; shooting, goalkeeping, passing, defensive actions, possession and miscellaneous. We went to each individual team that made the Round of 16 (by year), and then removed columns of stats that we didn't need. We then exported the data as a CSV and added it to a spreadsheet with data from each team.

After adding the data of all teams, we had to remove a few rows as there were many empty cells. Furthermore, many columns of data had empty cells due to the nature of the statistic itself - for example, Save % is the number of saves / number of shot attempts on target. However, if there were no shots on target, Save % would be empty, as 0/0 is indefinite. Hence, we had to remove these statistics, as well as other redundant statistics.

Finally, we labeled the data by team. We created a mapping of teams to numbers - for example, Bayern Munich 2021-22 would be 0 - and then labeled the games of each team accordingly.

Formatting and Splitting the Data

Once we compiled the dataset, it had to be formatted and split in the correct way for the classifiers. The data was split into training and test data by taking 'n' games from each team as test data, and the rest for training data. The data was then split into labels and statistics, and then converted into numpy arrays and lists, from dataframes. The complete dataset can be found here.

Optimising K-Nearest Neighbours and Support Vector Classifier

After preparing the data, the next step was to find the classifier that most accurately classifies the games. There were 4 steps to doing this:

1. Gather all the classifiers needed
2. Optimize all the classifiers that need optimisation
3. Find the most accurate classifier for the current test-train split
4. Rerun code till most accurate classifier with best test-train split is found

The classifiers that we decided to test were all classifiers that were part of scikit-learn. With the help of this article, the classifiers being tested were:

1. KNeighborsClassifier
2. SVC
3. NuSVC
4. DecisionTreeClassifier
5. RandomForestClassifier
6. AdaBoostClassifier
7. GradientBoostingClassifier
8. GaussianNB
9. LinearDiscriminantAnalysis
10. QuadraticDiscriminantAnalysis

Out of these classifiers, the ones that needed optimization were KNN, and SVC. The code we ran to optimize the classifiers can be found here.

KNN was optimized using hyperparameter tuning, by finding the 'K' that gave the highest accuracy. This was done by running the model multiple times with all viable values of 'K'. SVC was optimized by running the model using all the different kernels possible. **The optimized value of K found was 3, and the optimized kernel found was 'Linear'.**

Finding Most Accurate Classifier and Best Test-Train Split

The next step was to find the most accurate classifier, with the best test-train split. In order to do so, we ran the set of classifiers 4 times for each split. In each split, there were 4 ways the data was given to the model -

1. No Change
2. Standardized using StandardScaler()
3. Principal Component Analysis performed
4. PCA and StandardScaler()

The accuracy of the model was highest for the splits where training data was more. When there were 2 games taken for each team as test data, the accuracy we found was:

	KNN	SVC	NuSVC	DecisionTree	RandomForest	AdaBoost	GradientBoosting	GaussianNB	LinearDiscriminant	QuadraticDiscriminant
No Change	4.375	0.0	3.75	3.75	6.25	3.75	2.5	7.5	16.875	0.625
Scaled	4.375	0.0	3.75	5.625	5.625	3.75	2.5	7.5	16.875	0.625
PCA	4.375	1.25	3.75	6.25	5.0	1.875	5.0	8.75	7.5	7.5
Scaled and PCA	4.375	1.25	3.75	4.375	5.0	1.875	4.375	8.75	7.5	7.5

Figure 1.

As can be seen, the model with the highest accuracy was LinearDiscriminantAnalysis, and the same trend continued for the next split.

When there was 1 game taken for each team as test data, the accuracy found was:

	KNN	SVC	NuSVC	DecisionTree	RandomForest	AdaBoost	GradientBoosting	GaussianNB	LinearDiscriminant	QuadraticDiscriminant
No Change	6.25	1.25	5.0	7.5	13.75	2.5	10.0	7.5	22.5	0.0
Scaled	6.25	1.25	5.0	11.25	8.75	2.5	7.5	7.5	22.5	0.0
PCA	6.25	1.25	6.25	7.5	5.0	2.5	3.75	10.0	8.75	7.5
Scaled and PCA	6.25	1.25	6.25	8.75	6.25	2.5	1.25	10.0	8.75	7.5

Figure 2.

As can be seen here, the accuracy overall improved when the split was 1 game. Hence, we found that the best model for our use case is Linear Discriminant Analysis, and the best split is 1 game per team, with 22.5% accuracy.

The code used to find the best split and classifier can be found here. Note: The splits were changed manually and the code was rerun - there was no loop.

Getting Predictions

After finding the best classifier, the next step was to get the predictions. The classifier was run, and once the predictions were given, they were mapped back to the team names using the original mapping structure. Finally, the data was labeled and formatted to be viewed easier. The predictions can be found here, and the accurately predicted teams are shown in our analysis.

The headings of the table are:

1. **Predicted** - What the model predicted the team to be
2. **Actual** - The actual team
3. **Same or Not** - If predicted matches actual team
4. **Same Team** - If the model predicted the right club, but the wrong year. This could happen as the core of a team stays the same across years and hence have a similar playstyle.

Limitations

The limitations we faced while carrying out our analysis was based mainly on the amount of data available.

1. We were limited to only a few years of Champions League data as complete statistics were not available for years prior to the 2017-18 seasons. Hence, we can only compare the teams in the competitions that season and seasons after.

2. There was not enough training data per team. We took the teams that reached the Round of 16 in each iteration of the competition. Hence, if they were knocked out in the round of 16 itself, and half the teams were, they only had 8 games of data available, and this had to be split into training and test data. This point also ties into point 1, as if there was data for more seasons, we could have taken a different level of the competition, such as semis or quarters, and used data from more seasons to have extra training data per team.

3. Some teams did not have complete data. For example, RB Salzburg or Besiktas, teams that did reach the Round of 16, did not have complete data as they are teams that are not part of the top 5 leagues (Spanish La Liga, English Premier League, Italian Serie A, German Bundesliga and French Ligue 1). If they did play a team from the top 5 leagues, then that specific game had statistics recorded. However, if they played another team from a non-top 5 league, such as games in the group stages, then many of the advanced statistics were missing, and hence we had to remove those rows from our dataset.

4. Finally, the varying amount of training data per team. Depending on how far they reached in the competition and the league they are from, different teams had different amounts of data, and hence were able to have a better chance of classification. Hence, our analysis might have a chance of being skewed towards statistics that are stronger for teams that reached further in the competition, even though they might not reflect as strongly on winning.

Results, Analysis and Visualisations

After training our Linear Discriminant model with the data we compiled, we tested it with 80 games (1 for each team) to see whether it could accurately predict the teams that qualified along with the year they played in. The following dataset shows us the teams our model predicted accurately:

Sr No	Predicted	Actual	Same or Not	Same Team
1	Bayern Munich 2021-22	Bayern Munich 2021-22	Same	Same Team
8	Atlético Madrid 2020-21	Atlético Madrid 2021-22	Different	Same Team
12	Benfica 2021-22	Benfica 2021-22	Same	Same Team
15	Ajax 2021-22	Ajax 2021-22	Same	Same Team
16	Villarreal 2021-22	Villarreal 2021-22	Same	Same Team
20	Real Madrid 2021-22	Real Madrid 2021-22	Same	Same Team
21	Real Madrid 2021-22	Real Madrid 2021-22	Same	Same Team
25	Liverpool 2021-22	Liverpool 2021-22	Same	Same Team
28	Chelsea 2021-22	Chelsea 2021-22	Same	Same Team
30	Lille 2021-22	Lille 2021-22	Same	Same Team
39	Barcelona 2020-21	Barcelona 2020-21	Same	Same Team
41	Manchester City 2020-21	Manchester City 2020-21	Same	Same Team
44	Real Madrid 2020-21	Real Madrid 2020-21	Same	Same Team
59	Sevilla 2020-21	Sevilla 2020-21	Same	Same Team
61	Chelsea 2020-21	Chelsea 2020-21	Same	Same Team
63	Atlético Madrid 2021-22	Atlético Madrid 2020-21	Different	Same Team
64	Bayern Munich 2019-20	Bayern Munich 2019-20	Same	Same Team
65	Bayern Munich 2020-21	Bayern Munich 2019-20	Different	Same Team
72	RB Leipzig 2020-21	RB Leipzig 2019-20	Different	Same Team
76	Manchester City 2020-21	Manchester City 2019-20	Different	Same Team
77	Manchester City 2020-21	Manchester City 2019-20	Different	Same Team
79	Real Madrid 2021-22	Real Madrid 2019-20	Different	Same Team
83	Liverpool 2018-19	Liverpool 2019-20	Different	Same Team
88	Paris S-G 2019-20	Paris S-G 2019-20	Same	Same Team
94	Juventus 2019-20	Juventus 2019-20	Same	Same Team
96	Liverpool 2020-21	Liverpool 2018-19	Different	Same Team
97	Liverpool 2018-19	Liverpool 2018-19	Same	Same Team
98	Bayern Munich 2018-19	Bayern Munich 2018-19	Same	Same Team
102	Paris S-G 2019-20	Paris S-G 2018-19	Different	Same Team

103	Paris S-G 2018-19	Paris S-G 2018-19	Same	Same Team
104	Ajax 2018-19	Ajax 2018-19	Same	Same Team
116	Porto 2017-18	Porto 2018-19	Different	Same Team
120	Manchester City 2020-21	Manchester City 2018-19	Different	Same Team
126	Atlético Madrid 2019-20	Atlético Madrid 2018-19	Different	Same Team
127	Atlético Madrid 2019-20	Atlético Madrid 2018-19	Different	Same Team
128	Bayern Munich 2017-18	Bayern Munich 2017-18	Same	Same Team
129	Bayern Munich 2019-20	Bayern Munich 2017-18	Different	Same Team
134	Manchester Utd 2018-19	Manchester Utd 2017-18	Different	Same Team
136	Barcelona 2017-18	Barcelona 2017-18	Same	Same Team
139	Chelsea 2017-18	Chelsea 2017-18	Same	Same Team
144	Real Madrid 2017-18	Real Madrid 2017-18	Same	Same Team
145	Real Madrid 2017-18	Real Madrid 2017-18	Same	Same Team
146	Paris S-G 2020-21	Paris S-G 2017-18	Different	Same Team
147	Paris S-G 2020-21	Paris S-G 2017-18	Different	Same Team
148	Liverpool 2017-18	Liverpool 2017-18	Same	Same Team
149	Liverpool 2018-19	Liverpool 2017-18	Different	Same Team
157	Roma 2018-19	Roma 2017-18	Different	Same Team

Figure 3.

Through the above results, we can infer a few main insights:

1. The classifier accurately predicted Bayern Munich, Real Madrid, and Liverpool's games OR teams the maximum number of times (6, 6, 6).
2. The classifier accurately predicted Benfica, Villarreal, Lille, Sevilla, RB Leipzig, Juventus, Manchester United, Roma, and Porto's games OR teams the least amount of times (1, 1, 1, 1, 1, 1, 1, 1, 1).
3. Some teams were predicted accurately, but their years were predicted incorrectly.

In order to gauge which statistics are most significant, we will be dividing the analysis into 3 parts. Firstly, we will analyze the statistics of the teams that got classified correctly along with their year. Then, we will go on to compare the statistics of the teams that got classified correctly regardless of their year being inaccurate. Finally, we will be checking to see which teams got further in the competition - whether they got classified correctly, and why/why not.

Analyzing Correctly Predicted Teams + Years

With reference to the above dataset, we pinpointed the rows that had "Same" and "Same Team" in the "Same or Not" and "Same Team" columns respectively. Each of these rows' compiled data was analyzed thoroughly to find out whether there were distinct differences within the statistics of each team which allowed the ML algorithm to classify each team with their individual seasons accurately. To do so, an aggregate of each statistic was taken and compared with the total list of the statistics of each team. Furthermore, the maximum and minimum of each compiled statistic was gauged in order to identify the most significant identifying statistics. An example of the above is depicted in the following screenshot:

Team	Poss	Touches	Def Pen	Def 3rd	Mid 3rd	Att 3rd	Att Pen	Live	Succ	Att	Mis	Dis	Rec
Bayern 2021-22	62.9	766.4	55.8	199.9	345	228	34.7	766	16.2	23.6	15.8	8.6	
Benfica 2021-22	38.9	515.4	74.9	199.7	217.2	104.2	15.5	515.3	10.8	20.1	14.6	10.1	
Ajax 2021-22	61.375	720.625	51.875	175.75	323.5	227.25	37.5	720.5	10.25	19.625	14.875	8.25	
Villarreal 2021-22	43.16666667	542.6666667	78.08333333	209.6666667	213.6666667	124.75	17	542.5	7.833333333	17.58333333	15.91666667	8.666666667	32
Real Madrid 2021-22	51.07692308	712	78.30769231	239.6153846	285.7692308	193.3846154	27.07692308	711.7692308	10.46153846	18.38461538	11.84615385	8.846153846	51
Liverpool 2021-22	62.5	776.8333333	50.41666667	193.3333333	384.1666667	207.25	30	776.6666667	12.5	23.16666667	18.75	9.5	55
Chelsea 2021-22	61	782.4	44.1	165.8	371.8	252.5	30.7	782.2	10.8	19.5	15.2	9.3	
Lille 2021-22	53.00309829	674.9875	62.94711538	197.3108974	299.3504274	184.8891026	26.29615385	674.8226496	10.44081197	19.72660256	15.19797009	9.110470085	46
Sevilla 20-21	60.25	725.125	55.25	186.375	355.875	189.125	24	725	9.5	15.875	15.75	10.875	
Barcelona 2020-21	62.125	862.75	49.5	158.625	412.875	296.125	36.125	861.875	14.375	23.375	11.125	12	
Man City 2020-21	60.15384615	808.7692308	43.76923077	147.0769231	457.3076923	210.7692308	28.76923077	808.6153846	12.46153846	20.76923077	11.76923077	9	62
Real Madrid 2020-21	57.58333333	784.25	65.33333333	209.5	390	192	24.66666667	784.0833333	12.25	18.41666667	12.58333333	8.416666667	58
Chelsea 2020-21	51.84615385	681.5384615	64.53846154	203.5384615	325.6153846	158.6153846	20	681.0769231	10.38461538	16.76923077	16	8	47
Bayern 19-20	63.81818182	780.4545455	54.36363636	191.5454545	378.9090909	215.9090909	37.27272727	780.2727273	11.63636364	19.09090909	12.45454545	6.363636364	58
PSG 19-20	53.18181818	708.2727273	65.45454545	220	359.5454545	135.8181818	20.18181818	708.1818182	12.45454545	20.72727273	15	8.636363636	50
Juventus 19-20	59.375	760	61.25	188.875	384.5	193	18.625	759.875	11.625	17.375	12.5	7.125	
Liverpool 18-19	51.61538462	651.8461538	55.76923077	170.3846154	321.3076923	166.5384615	28.07692308	651.4615385	9.076923077	14	14.92307692	11.53846154	
Ajax 18-19	58.03765135	749.7689127	56.05768502	186.0303904	370.1043674	200.211621	27.05945998	749.5109201	11.45873316	19.06596494	14.27109638	9.155466524	54
PSG 18-19	56.95142116	740.6125039	58.31701541	189.6188556	373.4112103	184.1077463	25.58147824	740.3847056	11.4184649	18.27678437	13.68766036	8.529449341	53
Bayern Munich 18-19	56.875	737.375	64.375	245.5	352.875	143.5	26.375	737.125	8.625	15.625	12.75	12.375	
Bayern 17-18	57.65106587	749.2302113	57.83151155	191.4224751	373.5271577	190.4766431	26.39444202	748.9551959	11.27218201	18.28050494	13.5678286	9.334587006	54
Barcelona 17-18	64.2	818.2	61.5	199.8	428	195.5	23.6	818.2	11.3	17.7	12.3	10.5	
Chelsea 17-18	45.875	624.75	52.25	171.5	297.625	163.125	24	624.5	13.875	20.75	12.875	13	
Madrid 17-18	57.38461538	745.3076923	56.61538462	188.7692308	365.4615385	198	29.23076923	745	11.92307692	17.15384615	10.53846154	9.153846154	
Liverpool 17-18	56.49189427	733.2004776	59.51198493	196.6526526	363.1447612	179.7540099	25.46648344	732.9713201	11.33076189	17.94085997	13.34238481	9.394498248	53
MEDIAN	57.58333333	740.6125039	57.83151155	191.5454545	363.1447612	192	26.375	740.3847056	11.33076189	18.41666667	13.68766036	9.153846154	53
MAX	64.2	862.75	78.30769231	245.5	457.3076923	296.125	37.5	861.875	16.2	23.6	18.75	13	
MIN	38.9	515.4	43.76923077	147.0769231	213.6666667	104.2	15.5	515.3	7.833333333	14	10.53846154	6.363636364	

The excel sheet with the entirety of the above data can be found here.

Through this methodology, the most distinct statistics for each team are as follows:

1. **Bayern Munich 2021-22**
    a. Highest - Succ, Att, Prog, PrgDist, ⅓, Prg
    b. Lowest - NULL
    c. Median - TkIW, np:G-xG, Thr,

2. **Benfica 2021-22**
    a. Highest - Def 3rd, Mid 3rd, Att 3rd, GA, PSxG, PKsV, Cmp,
    b. Lowest - Poss, Touches, Att 3rd, Att Pen, Live, Rec, GF, Sh, SoT, xG, npxG, Cmp, Att, TotDist, PrgDist, Ast, xAG, xA, KP, PPA
    c. Median - Off, Int, G-xG, Thr

3. **Ajax 2021-22**
    a. Highest - Att Pen, Fls, OG, Mid 3rd, Pass, PPA
    b. Lowest - Int, Dist, FK, Thr, Stp,
    c. Median - Won, Att, Past, GA

4. **Villarreal 2021-22**
    a. Highest - Clr, Err, AvgLen, Att, Opp, CrdR, 2CrdY
    b. Lowest - Prog, ⅓, Crs, Prog, Succ, Mid 3rd
    c. Median - Tkl

5. **Real Madrid 2021-22**
    a. Highest - Def Pen, Sh, SoTA, Saves
    b. Lowest - Fls
    c. Median - PKwon

6. **Liverpool 2021-22**
    a. Highest - Mis, Recov, Att 3rd,
    b. Lowest - CrdY, Sh, Saves,
    c. Median - G/Sh

7. **Chelsea 2021-22**
    a. Highest - NULL
    b. Lowest - Off, Past, Cmp,
    c. Median - GA, SoT, PK, PKatt, npxG, GA,

8. **Lille 2021-22**
    a. Highest - NULL
    b. Lowest - NULL
    c. Median - GF, PKA, #OPA,

9. **Barcelona 2020-21**
    a. Highest - Touches, Att 3rd, Live, Rec, PKwon, PKcon, PK, PKatt, PKA, Cmp, Att
    b. Lowest - TkIW, Won, Lost, Clr, G/Sh, Att, AvgLen, Stp
    c. Median - NULL

10. **Manchester City 2020-21**
    a. Highest - Mid 3rd
    b. Lowest - Def Pen, Def 3rd, Fld, SoTA, PSxG, Att, Opp,
    c. Median - npxG/Sh

11. **Real Madrid 2020-21**
    a. Highest - NULL
    b. Lowest - NULL
    c. Median - Poss, Att 3rd, Att, Tkl, Stp,

12. **Sevilla 2020-21**
    a. Highest - Crs, PKcon, Won, PKatt, PKA, CrsPA
    b. Lowest - Tkl, Tkl + Int, npxG/Sh, PSxG+/-
    c. Median - Opp, Stp

13. Chelsea 2020-21

a. Highest - CS
b. Lowest - GA
c. Median - Tkl + Int, GA, xG

14. Bayern Munich 2019-20

a. Highest - Tkl, Att, Past, GF, Sh, SoT, xG, npxG, #OPA, Cmp, Att, Ast, xAG, xA, KP
b. Lowest - Dis
c. Median - Def 3rd, PKcon, Saves, Pkatt

15. Paris S-G 2019-20

a. Highest - CrdY, Fld, G/sh, npxG/Sh
b. Lowest - Err, #OPA, Cmp
c. Median - Pkcon, PKatt

16. Juventus 2019-20

a. Highest - FK, Thr,
b. Lowest - Recov, Tkl, Att, Blocks, Pass
c. Median - NULL

17. Liverpool 2018-19

a. Highest - Lost
b. Lowest - Att, np:G-xG
c. Median - Def 3rd, Sh

18. Bayern Munich 2018-19

a. Highest - Def 3rd, Int, Tkl, Int, Tkl+Int, G-xG, np:G-xG, PSxG+/-
b. Lowest - NULL
c. Median - Att Pen, GA, PSxG, Att, Opp

19. Paris S-G 2018-19

a. Highest - NULL
b. Lowest - NULL
c. Median - Touches, Live, Mis, Rec, Clr, Err, FK, AvgLen, Cmp, Att, PPA

20. Ajax 2018-19

a. Highest - NULL
b. Lowest - NULL
c. Median - Fld, Lost, Dist, CS, Cmp, Att, xAG

21. Bayern Munich 2017-18

a. Highest - NULL
b. Lowest - NULL
c. Median - Def Pen, Recov, Mid 3rd, Att 3rd, Sh, Pass, SoTA, TotDist, KP

22. Barcelona 2017-18

a. Highest - Poss, Dist, TotDist,
b. Lowest - CrsPA
c. Median - NULL

23. Chelsea 2017-18

a. Highest - Dis, TkIW
b. Lowest - G-xG
c. Median - Cmp, CrsPA

24. Real Madrid 2017-18

a. Highest - Off, Blocks, Cmp
b. Lowest - Mis, CS
c. Median - Dis

25. Liverpool 2017-18

a. Highest - NULL
b. Lowest - NULL
c. Median - Mid 3rd, Succ, Prog, CrdY, Fls, Crs, Blocks, PSxG+/-, Att, PrgDist, Ast, xA, ⅓, Prog

Through the above list, we can conclude the most significant statistics for each individual team in every year of the Champions League.

A few observations that we can see are the fact that the most number of correct predictions have occurred in the season of 2021-22. This could be due to the fact that the data collected in the later years of the Champions League would be more accurate and consistent throughout the season due to improvements in technology. Furthermore, manager retention is also a factor that can be taken into account since teams in recent years have stopped switching managers as often as they used to - hence their play style would have been more consistent.

Analyzing Correctly Predicted Teams (without Years)

Sr No	Predicted	Actual	Same or Not	Same Team
8	Atlético Madrid 2020-21	Atlético Madrid 2021-22	Different	Same Team
63	Atlético Madrid 2021-22	Atlético Madrid 2020-21	Different	Same Team
65	Bayern Munich 2020-21	Bayern Munich 2019-20	Different	Same Team
72	RB Leipzig 2020-21	RB Leipzig 2019-20	Different	Same Team
76	Manchester City 2020-21	Manchester City 2019-20	Different	Same Team
77	Manchester City 2020-21	Manchester City 2019-20	Different	Same Team
79	Real Madrid 2021-22	Real Madrid 2019-20	Different	Same Team
83	Liverpool 2018-19	Liverpool 2019-20	Different	Same Team

96	Liverpool 2020-21	Liverpool 2018-19	Different	Same Team
102	Paris S-G 2019-20	Paris S-G 2018-19	Different	Same Team
116	Porto 2017-18	Porto 2018-19	Different	Same Team
120	Manchester City 2020-21	Manchester City 2018-19	Different	Same Team
126	Atlético Madrid 2019-20	Atlético Madrid 2018-19	Different	Same Team
127	Atlético Madrid 2019-20	Atlético Madrid 2018-19	Different	Same Team
129	Bayern Munich 2019-20	Bayern Munich 2017-18	Different	Same Team
134	Manchester Utd 2018-19	Manchester Utd 2017-18	Different	Same Team
146	Paris S-G 2020-21	Paris S-G 2017-18	Different	Same Team
147	Paris S-G 2020-21	Paris S-G 2017-18	Different	Same Team
149	Liverpool 2018-19	Liverpool 2017-18	Different	Same Team
157	Roma 2018-19	Roma 2017-18	Different	Same Team

Figure 4.

These are results in which the model predicted correctly for the given teams i.e. prediction is the same as the actual team but from a different season. We first calculated the mean of each statistic of the actual team and then the mean for each statistic of the predicted team. This was done to compare each statistic as means to reach a result which we did on the basis of % match - how close both means were of each statistic based on a certain threshold which we took to be higher than 94% since the range of match % of statistics was approximately from 40% to 99% with 4-6 statistics across all actual-predicted team pairs crossed the 93%.

After this analysis - of all statistics across each result - we were able to conclude some statistics that proved extremely significant in making the model classify the team from different seasons as the same. The aforementioned statistics include: Sh (Shots Total - does not include penalty kicks), SOT% (Percentage of shots that are on target), Dist (Average distance from goal of shots taken), xnpG (non-penalty expected goals), PSxG (Post shot expected goals - goalkeeping), Short Cmp% (pass completion percentage), Medium Cmp% (pass completion percentage).

SOT%, xnpG, Short Cmp% and Medium Cmp% are 4 main statistics that proved to be significant in these teams' analysis as their match % are (all rounded to closest integer) :

SOT% → 97%
xnpG% → 97%
Short Cmp% → 95%
Medium Cmp% → 94%

Higher the SOT%, higher chance of scoring as it represents the percentage of shots taken by the team on goal. A high xnpG represents a good overall attack of your team since your **non-penalty** expected goals are high. Finally, having a good short pass completion and medium completion

rate shows dominance over ball possession and displays good midfield strength making them significant statistics to consider.

Based on this, we conclude that these are some of the most important statistics used to compare teams against each other at the professional level.

Analyzing Highly Ranked Incorrectly Predicted Teams

Across UCL seasons 2017-18, 2018-19, 2019-2020, 2020-21, 2021-22 and the teams in our database, we wanted to find out why teams that were highly ranked i.e. they reached semi-finals or better in any of the seasons (given in figure 5), were classified incorrectly.

The table below shows the top four teams in the last five years.

Season	**Top 4 teams**
2021-22	Real Madrid, Liverpool, Manchester City, Villarreal
2020-21	Chelsea, Manchester City, Real Madrid, Paris Saint Germain
2019-20	Bayern Munich, Paris Saint Germain, RB Leipzig, Lyon
2018-19	Liverpool, Tottenham Hotspur, Ajax, Barcelona
2017-18	Real Madrid, Liverpool, Roma, Bayern Munich

Figure 5.

Sr No	Predicted	Actual	Same or Not	Same Team
52	Atalanta 2019-20	Ajax 2018-19	Different	
46	Tottenham 2017-18	Lyon 2019-20	Different	
54	Real Madrid 2017-18	Barcelona 2018-19	Different	
44	Benfica 2021-22	Paris S-G 2019-20	Different	
2	Atletico Madrid 2019-20	Manchester City 2021-22	Different	

Figure 6.

After analysis of these teams using the aforementioned method (on page 14) we find that they do not get classified correctly because the statistics of each of the teams are very varied - match % of means of each statistic between the predicted-actual pairs had a much wider range of 30 to 90% with only 0-1 team(s) crossing our earlier threshold.

Hence, the model was not able to predict them correctly. Since these are a mere 5 teams out of our database which has 79 teams and given the above results (match%, number of teams crossing threshold), we believe this could be classified as the ML model error.

Analyzing Low Ranked Correctly Predicted Teams

Analysis of teams that were lowly ranked, i.e. did not get through the Quarter Finals stage, across the 5 UCL seasons but were correctly predicted by our models shows us a correspondence between the important statistics for *"Correctly Predicted Teams (Without Years)"*.

Sr No	Predicted	Actual	Same or Not	Same Team
73	Paris S-G 2020-21	Paris S-G 2017-18	Different	Same Team
69	Chelsea 2017-18	Chelsea 2017-18	Same	Same Team
70	Juventus 2018-19	Juventus 2017-18	Different	Same Team
63	Atletico Madrid 2019-20	Atletico Madrid 2017-18	Different	Same Team
58	Porto 2018-19	Porto 2018-19	Same	Same Team
51	Paris S-G 2018-19	Paris S-G 2018-19	Same	Same Team
41	Liverpool 2018-19	Liverpool 2019-20	Different	Same Team
39	Real Madrid 2021-22	Real Madrid 2019-20	Different	Same Team
38	Manchester City 2020-21	Manchester City 2019-20	Different	Same Team
34	Barcelona 2017-2018	Barcelona 2019-20	Different	Same Team
31	Atletico Madrid 2021-22	Atletico Madrid 2020-21	Different	Same Team
29	Sevilla 2020-21	Sevilla 2020-21	Same	Same Team
26	Liverpool 2020-21	Liverpool 2020-21	Same	Same Team
21	M'Gladbach 2020-21	M'gladbach 20201-21	Same	Same Team
19	Barcelona 2020-21	Barcelona 2020-21	Same	Same Team
11	Paris S-G 2021-22	Paris S-G 2021-22	Same	Same Team
8	Villareal 2021-22	Villareal 2021-22	Same	Same Team

The statistics found to be almost exactly same for the predicted and actual teams in this category are: SOT% (Percentage of shots on target), Dist (Distance of goal when shot is taken at goal), PSxG (Post-Shot expected Goal), Short CMP% (Percentage of short passes completion).

Between all predicted-actual pairs from the above table they matched on these aforementioned stats at an impressively high percentage.

- SOT% average match% = 96.51%
- Dist average match% = 97.76%
- PSxG average match% = 90.11%
- Short CMP% average match % = 99.96%

Conclusion

After training our Linear Discriminant model with the data we compiled, we tested it with 80 games (1 for each team) to see whether it could accurately predict the teams that qualified along with the year they played in. Each of these rows' compiled data was analyzed thoroughly to find out whether there were distinct differences within the statistics of each team which allowed the ML algorithm to classify each team with their individual seasons accurately.

Firstly, statistics that end up being significant in teams being classified as another are SOT%, Dist, PSxG, Short CMP% which tells us that shooting and passing are by far two of the most important factors statistically for differentiating between teams, which makes sense because if a team is able to pass the ball around better than the other then it directly results in them having higher possession of the ball and more chances at goal. Based on only these two statistics a machine learning model can tell two clubs or two teams of different seasons of the same club alike or apart.

Secondly, we can see a growing trend that a team has been predicted correctly, from the same year as well, more as we go towards the 2021-22 seasons from the 2017-18 season and a big reason for this trend is big clubs opting to retain their team managers over longer periods of time, even if little to no success achieved, which would in turn make the team play a certain style of play over years giving us a similarity/match in statistics under the umbrella of attack (SOT%, PSxG, Dist) and midfield (Short CMP%, Medium CMP%).

Thus, considering the limitations in the data, we were able to pinpoint a few significant statistics that allowed for the correct prediction of the teams along with their seasons.