Data mining of large Datasets

Prediction of the 2022 African cup winner

Authors:
HARRAD ABDELGHANI
MERNISSI AHMED

Content Table:

Introduction:

The African cup of football (AFCON2022) is approaching and it will be hosted by Cameroon in the period from 9 January to 6 February. The previous title was won by Algeria against Senegal in a very tough final, the final score was 1-0 for Algeria and the single goal scored was a fluke. This time, the circumstances are different and many changes have occured between-times, whether on team members and coach, FIFA ranking of teams of countries qualified for this edition of the cup. Fortunately for us, Morocco, which is our home country, is also qualified this time. Therefore, we are hoping that the title will be ours or that we will be in the top 4 teams at least at the end of the tournament. To reduce uncertainty and make our hope more realistic, we decided to calculate by ourselves the chances of our home team winning the cup by analysing all the road up to the final. Starting by predicting the results of group stage matches and then the results of elimination rounds till the final with the use of a good model, we intend to predict the identity of the AFCON2022 winner.

In order to predict the winner, we decided to work on three datasets. The first dataset contains the ranking of every team in the world in different dates from 1992 until 2021-05-27, the second one contains all the statistics of matches between different international teams in the world and finally the third dataset that we have created specially for this project and that contains the 24 teams that will participate in the African cup and the groups in which they will be playing as well as their first, second and third opponent from the group stage round in the tournament.
We found the first and second datasets in Kaggle but the third one was unfortunately not pre implemented. For that, we decided to create our own dataset since it doesn't require too much time because it contains only 24 rows. The model we choose is based on probabilities, and the victory of a team against another is decided by the winning probability that has to exceed a given value. A victory gives 3 points to the winning team and 0 points for the losing one whereas a draw gives 1 point to both of them. The AFCON consists of 6 groups of 4 teams each, from which the top 2 teams move on to the next round with the best 4 teams ranked 3rd in their groups. The 16 countries qualified play against each other in the elimination round step by step till the final that determines the winner of the cup.

First of all, we will begin by importing all the needed classes. After that we will start cleaning the data in order to keep the data that we will need in our prediction of the winner of the African cup 2022.

1-Exploring and Cleaning of data :

1.1-Data cleaning:

We will start by importing the dataset in the format excel that contains the rankings of all the international teams named fifa_ranking-2021-05-27-3.xlsx, we will keep only the columns that we will need in our study ('rank', 'country_full', 'country_abrv', 'rank_date', 'confederation') after that we will transform rank_date to the format date and finally since we just need the teams from the continent Africa we will delete all the teams that are not from Africa by dropping the teams which their confederation is different from CAF.

	rank	country_full	country_abrv	rank_date	confederation
1	107	Mozambique	MOZ	1992-12-31	CAF
5	111	Sudan	SDN	1992-12-31	CAF
6	112	Mauritius	MRI	1992-12-31	CAF
11	117	Guinea-Bissau	GNB	1992-12-31	CAF
19	104	Swaziland	SWZ	1992-12-31	CAF

We will move to the second dataset ‘AFCON-2021-Dataset-excel.xlsx’ by importing it and we kept the columns that we will need in our prediction ('Team', 'Group', 'First match \n against', 'Second match\n against', 'Third match\n against') and by precaution we decided to drop all the rows that have just null values in every case and finally we have chosen Team as the index of our dataset.

Team	Group	First match \nagainst	Second match\n against	Third match\n against
Cameroon	A	Burkina Faso	Ethiopia	Cabo Verde
Burkina Faso	A	Cameroon	Cabo Verde	Burkina Faso
Ethiopia	A	Cabo Verde	Cameroon	Ethiopia
Cabo Verde	A	Ethiopia	Burkina Faso	Cameroon
Senegal	B	Zimbabwe	Guinea	Malawi

Finally, we will clean the final dataset results.csv by importing it first then we will transform the column date to the format date so we could do some operations on it and finally we have dropped all the matches before 2011-01-01 so we could reduce the amount of data.

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
0	1872-11-30	Scotland	England	0.0	0.0	Friendly	Glasgow	Scotland	False
1	1873-03-08	England	Scotland	4.0	2.0	Friendly	London	England	False
2	1874-03-07	Scotland	England	2.0	1.0	Friendly	Glasgow	Scotland	False
3	1875-03-06	England	Scotland	2.0	2.0	Friendly	London	England	False
4	1876-03-04	Scotland	England	3.0	0.0	Friendly	Glasgow	Scotland	False

In order to extract some features such as the point and the rank difference between two teams from our datasets, we will join the dataset of matches and dataset of the rankings and since we got some duplicate columns in matches, we will eliminate them to finally get the dataset of matches with all the attributes that we need in our prediction.

Now that the dataset matches is cleaned and merged with rankings, we will add to it some relevant parameters such as 'rank_difference' to give the difference between the rank of the home team and the away team, 'average_rank' , 'score_difference' which gives the difference between the goals scored by the home and the away team, (is_won,is_loss) to express whether the home team won or lost and finally 'is_stake' that takes False if the match between the teams is Friendly, else it will take True.

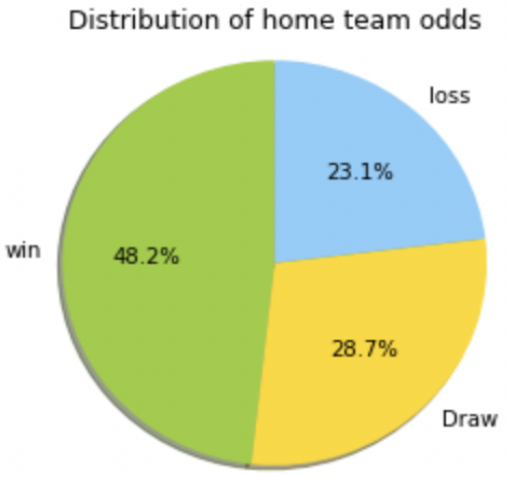
rank_difference	average_rank	score_difference	is_won	is_loss	is_stake
-30.0	54.0	0.0	0	1	False
-42.0	76.0	0.0	0	1	False
28.0	83.0	1.0	1	0	False
16.0	47.0	-1.0	0	1	False
12.0	86.0	2.0	1	0	False

To make our prediction more precise, we will add the history between the teams by calculating the difference between the wins and the losses. To do so we will add 3 columns to matches 'Previouswins_home' for the wins, 'Previouswins_away' for the losses and finally 'history_teams' which is deduced from the wins and the losses of the team.

1.2-Exploratory of data:

There are few questions in order to understand data better

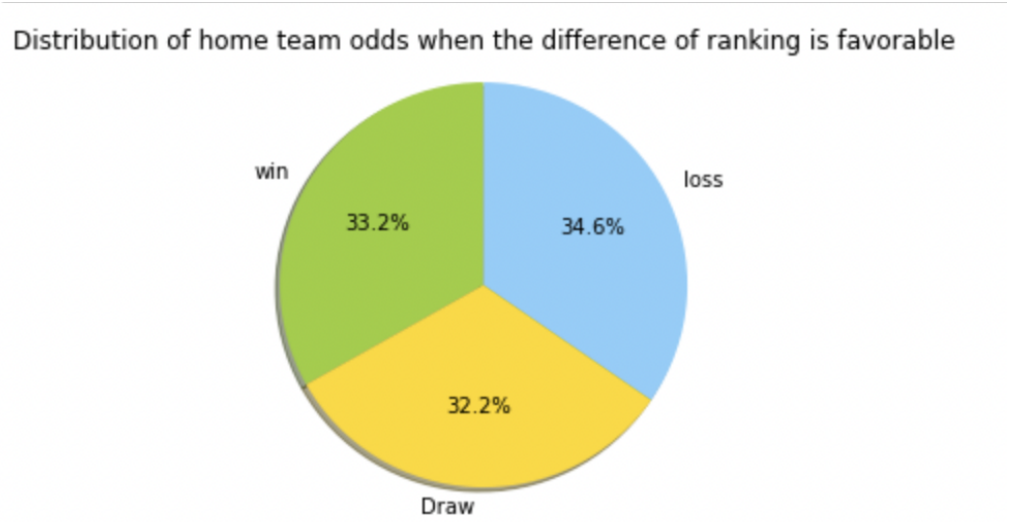
Here, we decided to draw a pie that shows the distribution of wins, losses and draws for the home team. With a rate of approximately 50%, it is clear that the home team is always more likely to win the match. For the host team, a draw is also considered as a loss and it is never a satisfying result. Hence, we took the idea of giving the hosted team a little advantage by considering it as a winner in case of equal probabilities or probability in the limit of the interval of draw, espicially in elimination rounds.



1.3-Analysis of the new features:

The rate of winning for the home team when it's better ranked is 33%, which is not that far from the previous rate, this shows that the FIFA rankings could give a first idea about the winning team knowing that there are some exceptions represented by the rest 15%.

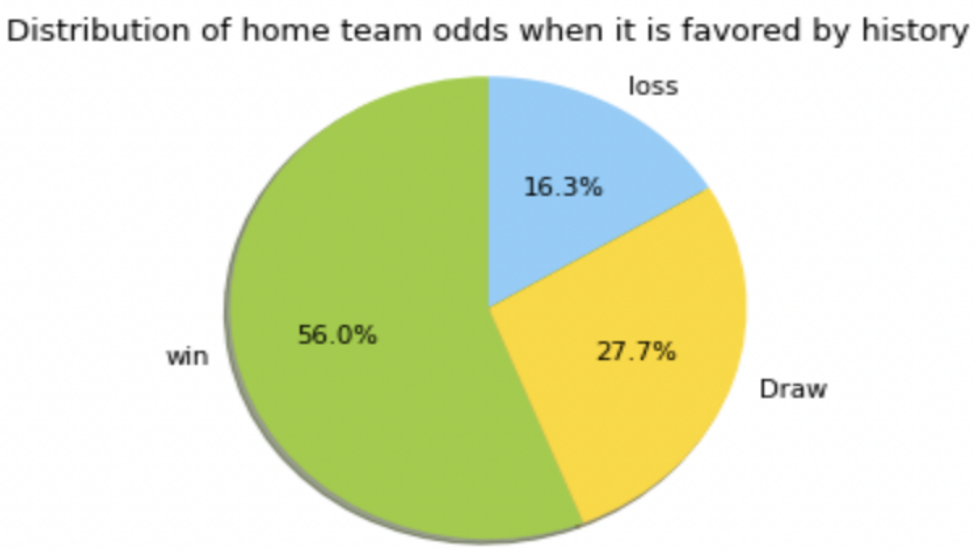
This also strengthens the fact that the home team have more chances to win even if it's less ranked or less favorite before.



1.4-Analysis of history_team feature:

The rate of winning for the home team when it's better favored by history is 56%, which shows that the reality is often loyal to the history in terms of team powers when facing each other.

Head-to-head history results have shown that they are a good factor to be taken into account when predicting the final winner of a given match.



2-Preprocessing and Classification :

2.1-Comparison between classifiers:

We will compare the three algorithms of classification (logistic regression, random forest, Decision tree) to see which one of them is the most accurate in the classification for our model. After training our model with the three algorithms,we notice in the figures that the logistic regression is the most accurate algorithm of classification for our model.In fact, logistic regression predicts correctly with an accuracy of 73% more than RandomForest with an accuracy of 66% and the Decision Tree with an accuracy of 55%.

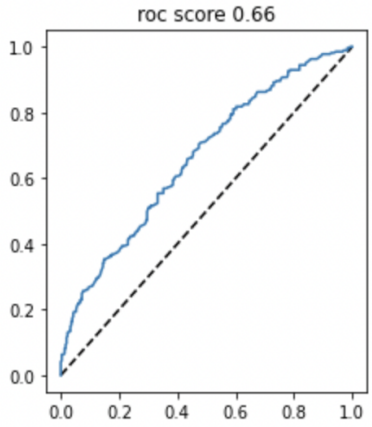


Figure 1 : Accuracy of RandomForest

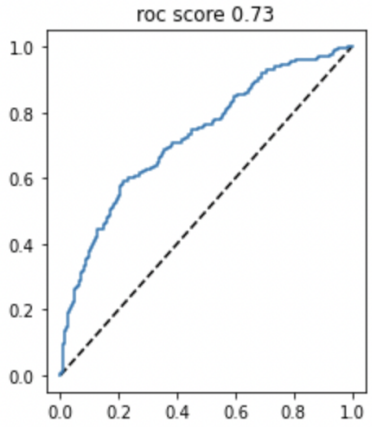


Figure 2 : Accuracy of Logistic Regression

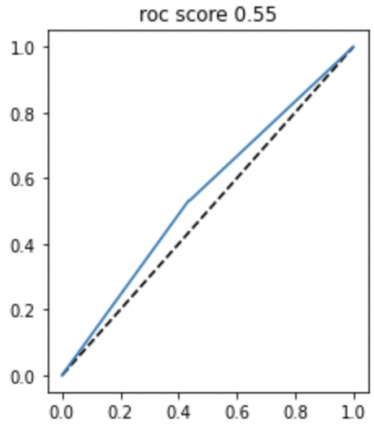


Figure 3 : Accuracy of Decision Tree

2.2-Modeling and Classification with logistic regression:

To determine who will be more likely to win a match, based on our knowledge, we come up with 4 main groups of features as follows:
rank difference: Every national team in the world has a FIFA ranking, it's a metric that qualifies the strength of the team based on their recent match results. It doesn't take into consideration friendly matches and only official ones are taken into account. The average rank for a given confrontation is the difference between ranks of the home team and the away team.
Average rank: The average of the 2 team FIFA ranks.
head-to-head match history between 2 teams aka 'history_teams' : It summarizes the historic results between the same 2 teams and it's calculated as the difference between number of victories and losses of the home team, it is an algebraic metric and its opposite gives the victories mines the losses for the hosted team.
is_stake: False if the match is friendly, True otherwise.

Logistic regression is the more powerful algorithm of classification. Hence, we chose to apply it to these features in the rest of our analysis.

2.3-Preprocessing of data and Training of the model:

Here, the preprocessing consists of a polynomial feature method, the result is a new feature matrix containing a 2-degree polynomial combination of input features. Then, we will train our model with a training set with a size of 80% out of the original dataset.

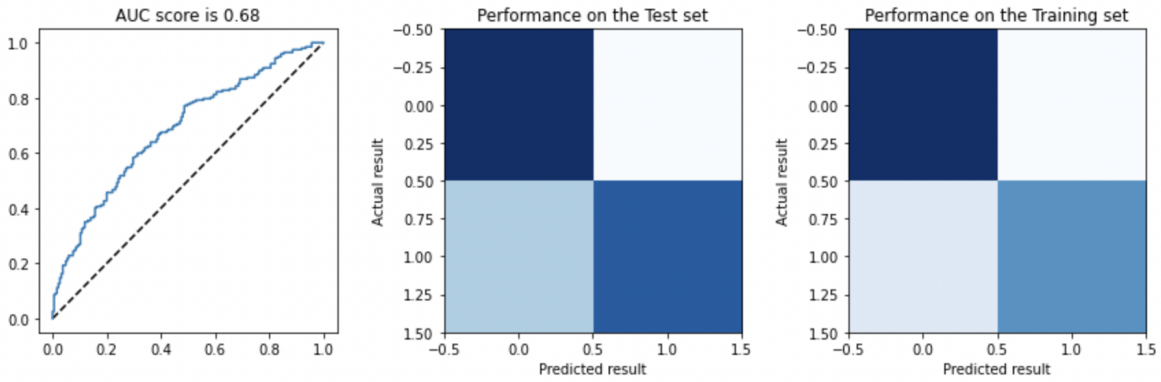
2.4-Accuracy Metrics:

2.5.1-ROC score:

The ROC score of logistic regression is not constant. In fact, it changes every time we train the model. This time, it's equal to 68% which is good knowing that soccer matches can not be perfectly predictable since there are only a few goals scored in a single match.

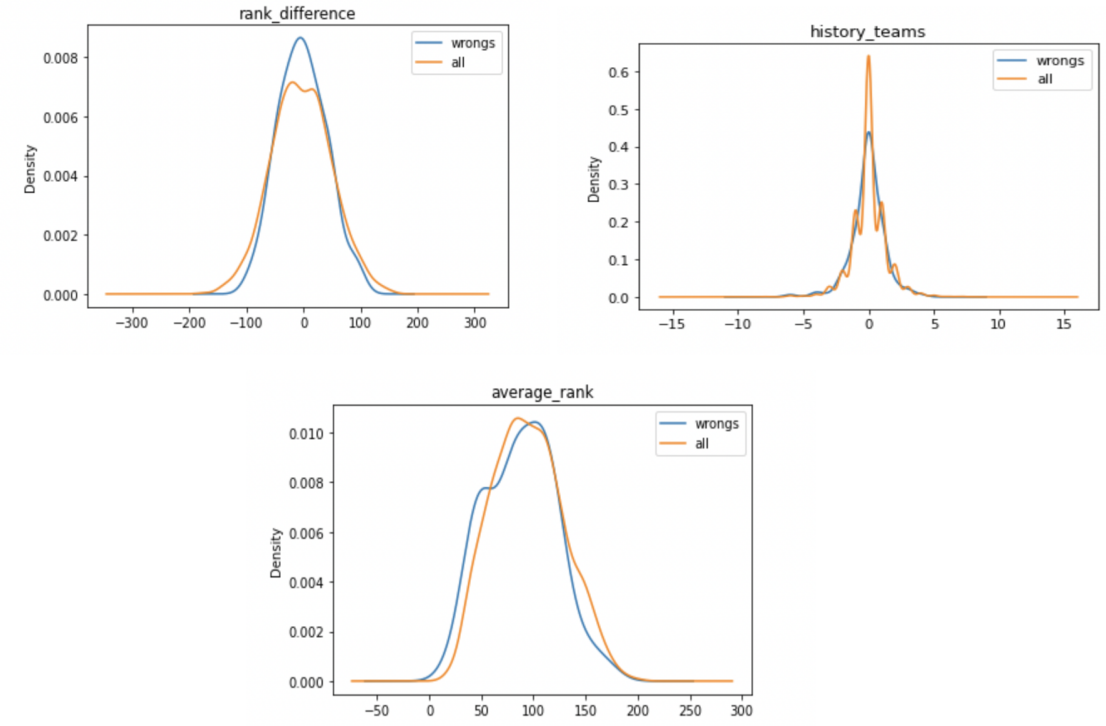
2.5.2-Confusion Matrix on training and testing sets:

The confusion matrix is a way of comparing predictions with reality. In our model, 1 means a win for the home team whereas a 0 means a loss. The colors in the confusion matrix are darker when the rate is bigger. In our case, colors are darker in the matching areas between prediction and reality with a little darkness in the testing set due to its smaller size compared to the training set. Hence, our model is doing its job perfectly.



2.5.3-Bad predictions:

To know the source of our mistakes, we tried to look at the bad predictions, to see the distribution of not friendly matches in those bad predictions and after overall. We noticed that the official matches are more likely to lead to a wrong prediction, this shows that teams play more roughly in official games. Particularly, this can be explained by the fact that official competitions are often organised in one country, so there is no home or away team besides the host one.



3-Simulation of African Cup 2022 matches Results :

3.1-Group Stage:

First of all, we will start by predicting the results of the matches of the group stage since they are already planned and we know the first, second and third opponent of each team.
To do so we will start by dropping all the rows of the rankings dataset with a rank_date different than 2021-05-27 and we will keep in this dataset just the african teams that will participate in the 2022 edition of this african cup.After creating this new dataset african_cup_rankings we will set 'country_full' as the index.

Let’s begin our prediction by adding 2 columns in the african_cup dataset the first one is ‘points’ which contains the total points that the teams will get after each game and the second column 'total_prob' which is the total probability of winning a game, Finally we will set the points and total_prob to 0.

By using model.predict_proba we will calculate the probability of winning a game(home_win_prob) or losing a game in each game in the group stage, to do so we will need 'average_rank', 'rank_difference' as well as 'history_teams', after each game played by a team we add the probability of winning or losing in the total_prob column, and concerning the column points we have defined a margin=0.05 that will help us to precise whether the game is a win or a loss or a draw.

If home_win_prob is lower or equal to 0.45 then the home_team gets 0 points and the away_team gets 3 points, if now home_win_prob is higher or equal to 0.55 then the home_team gets 3 points and the away_team gets 0 point and if we're not in either the first two situations then both the home_team and the away_team get 1 point.

Here is an example of the prediction of the games in the group C:

```
___Starting group C:___
Morocco vs. Ghana: 0.5194584428313324
Draw
Morocco vs. Comoros: 0.7964103769413908
Morocco wins with 0.80
Morocco vs. Gabon: 0.645447781293196
Morocco wins with 0.65
Ghana vs. Comoros: 0.7579560577303461
Ghana wins with 0.76
Ghana vs. Gabon: 0.5981598351696429
Ghana wins with 0.60
Comoros vs. Gabon: 0.3166887010510877
Gabon wins with 0.68
```

3.2-Analysis of group stage results:

As expected, teams that are said to be great are effectively well ranked in their groups like Algeria, Tunisia and Senegal. For other groups, there was tough competition between teams since they contained 2 or 3 big teams or teams with similar potential. Therefore in confrontations between these rival teams, the match is often more likely to end up with a draw as was the case between Morocco and Ghana, Nigeria and Egypt and 3 times between Cameroon, Burkina Faso and Cabo Verde. Even if these draws would end with another result, it would not affect the qualified teams at the end since there are 3 teams that can be qualified from each group.

At the end of the group stage round, we can affirm that the model is realistic as we are only left with the great 16 teams in the continent. These 16 teams will be divided in a predefined way to pairings playing against each other, the winner goes up to the next round step by step till the final.

3.3-Elimination round results:

Now that we got all the results of the group stage round, we will start by just letting the following columns ('Group', 'points', 'total_prob') in african_cup dataset then we will add another column in the dataset that contains their rank in their following groups.

After getting all the results of the group stage round, we will drop all the teams that got in fourth place in their group, as well as dropping the worst two teams that got in the third place from the 6 groups that we have.

The african_cup dataset will contain the teams that qualified to round_of_16 now we will match each team with their opponent by the help of the

pairing=[0,4,11,10,14,9,5,2,1,12,13,7,8,15,3,6] and after that we will calculate the probabilities of the qualified team to win or lose and with this method we will who will qualify to the quarter final and then we will do the same to find out who will qualify to the semis and then we will get who will be in the final. After running this method we got the winner of the african cup and the winner is Senegal who did beat Algeria with a probability of 0.51.

3.4-Analysis of elimination round results:

This time, there are no more weak teams, this is well explained by the algorithm predictions that were closer to 0.50. The high predictions that are still in the elimination round were between leader teams against teams that were ranked 3rd in their groups, like Nigeria vs Mauritania, Senegal vs Gabon or Cameroon vs Guinea-Bisseau. In the quarter final, the predictions became more and more tight, the winner being determined with an advantage between 2 and 7% which is not too decisive.
Rigorously, we should predict a draw in those matches but the draw is not an acceptable result and it leads to penalty shoot-out, we decided then to give victory to the team with higher probability for lack of a penalty shoot-out prediction model.

Unfortunately, Morocco played against Algeria in the quarter final and was disqualified before the semi final, but we hope Algeria will go further in the competition instead of Morocco. The semi final and the final were also tough games and the final score was determined with small details.

Congratulations in advance to Senegal for winning this edition of AFCON and their 1st one in the history.

Conclusion:

To summarize, we took the 2 datasets from kaggle and created the third one and filtered out the non desirable data.On the remaining data we did some preprocessing to clean the data. After that, we have trained our data with three different algorithms and found out that the logistic regression is the most accurate out of the three. Finally we have done a simulation to find out who is going to qualify from the group stage round and then decide a winner of the African Cup 2022.
After finishing our simulation, our model seemed to fit perfectly in the case of group stage games since a probability between 0.45 and 0.55 which is acceptable in the group round. Nevertheless, a draw isn’t acceptable in the elimination round and the two teams have to go to the penalty shoot-out to decide the winner. In our analyses, we lack a model of penalty shoot-out, so a good question for next projects would be to work out a penalty shoot-out prediction model.

References :

International football results from 1872 to 2021 dataset : Mart Jürisoo

FIFA rankings from 1992 to 2021 dataset: Alex