Machine Learning for Predicting Football Games' Outcome

Rıdvan Sözen
Mehmet Mücahit Sayar
Yavuz Selim Uçar

Abstract

Sports betting is one of these perfect problems for machine learning algorithms. Availability of tons of data makes it easier and attractive to study this problem with machine learning techniques. In this paper, we tried to solve only the soccer part of this problem because soccer is the most famous sport now. So, sources and data that we could get are much broader than any other sport area. Our model is based on numerous factors in the games such as results of historical matches that both teams played, players' performances in those matches and an overall performance of players. This paper analyses application of various ML methods, such as Artificial Neural Network, XGBoost and Vector Support Machine, on prediction of sport games' results. Then, after we used these methods, we compare and conclude which method gives us the best solution for our problem.

Introduction

Football industry has gained huge importance over the last decades. The money and the followers changed the perception of football from a sport to a market where both insiders, managers and players, and outsiders, fans and businessmen, play with a huge money. These developments also generate another market so called the betting market. In these two markets, answer of a question rises and plays an important role "Who is going to win this match?".

Predicting the outcome of a football match gives various advantages to a manager or a bookmaker. Managers determine his tactics and bookmaker determine his odds according to these outcomes. When this is so important, these circumstances make this question a problem that is needed to solve.

ML techniques can give a reliable solution to this problem. Nowadays, when electronically available data are so numerous, it is easy to employ a ML method. We construct

a classification model based on our training data set because we try to see that the match is going to end with win, draw or lose for a team. As classification necessitates, we use supervised learning techniques since we try to develop a predictive model based on both input and output data.

In other section of this paper, we give some statistical analysis and data exploration such as correlation between our features in our raw data. Then, we study how we reconstruct our data by selecting our features based on some feature selection methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Recursive Feature Elimination (RFE) and Least Absolute Shrinkage and Selection Operator (LASSO). Then, we examine the results we get from every ML methods and we will compare ML methods with each other.

Statistical Analysis and Data Exploration

The Source of Data

As indicated above football industry bears great importance for the ones who want to make huge profits. When this is the case finding the data of the industry is nearly impossible. Hopefully, the game industry and the supervisory agencies must keep all the data on soccer. Therefore, we took our data from European Soccer Database, which is transmitted to Kaggle, enetscores.com and EA Sports FIFA Games.

Data Clearing

After merging the datasets, we reached from the channels above, the raw data was included 13459 observations with 50 columns. However, some columns had to be extracted like DateTime, ID etc. due to their insignificant contributions to our ML methods, additionally missing values were dropped. Detailed feature list and their meanings are in the next section.

Feature List and Their Descriptions

Above all, we have 43 features related to our dependent variable “results”

- home_or_away: Which team is home team
- home_player_overall (1-11): Home players’ overall ratings
- home_player_defenders: How many defence players does home team have
- home_player_forwards: How many forward players does home team have
- Last_5_match_win_home: How many wins does home team have in its last 5 matches
- Last_5_match_goal_diff_home: Goal difference that home team has in its last 5 matches
- Last_5_match_card_home: Home team’s card score in its last 5 matches
- Last_5_match_faul_home: How many fouls does home team commit in its last 5 matches
- Last_5_match_corner_home: How many corners does home team have in its last 5 matches
- Last_5_match_shout_total_home: How many shouts does home team have in its last 5 matches
- Last_5_match_shout_rate_home: Shout accuracy of home team in its last 5 matches
- Last_5_match_shout_on_home: How many accurate shouts does home team have in its last 5 matches
- Team_variance_home: Variance of home team’s players
- Last_5_match_btw_home: How many wins does home team have against away team
- away_player_overall (1-11): Away players’ overall ratings
- away_player_forwards: How many forward players does away team have
- away_player_defenders: How many defence players does away team have
- Last_5_match_win_away: How many wins does away team have in its last 5 matches
- Last_5_match_goal_diff_away: Goal difference that away team has in its last 5 matches
- Last_5_match_card_away: Away team’s card score in its last 5 matches
- Last_5_match_faul_away: How many fouls does away team commit in its last 5 matches
- Last_5_match_corner_away: How many corners does away team have in its last 5 matches
- Last_5_match_shout_total_away: How many shouts does away team have in its last 5 matches
- Last_5_match_shout_rate_away: Shout accuracy of away team in its last 5 matches

- Last_5_match_shout_on_away: How many accurate shouts does away team have in its last 5 matches
- Team_variance_away: Variance of away team's players
- Last_5_match_btw_away: How many wins does away team have against away team

Statistical Exploration of Features

We looked at the main statistics of the data in the Table 1 and Table 2 by using the describe function and checked the following statistics for each feature: count, mean, standard deviation, min, 25%, 50%, 75% and max. On the other hand, it can be said that variance of the total scores of matches is huge and anything that comes to mind may effect results. For example, one may believe weather, players' happiness before the match, players' marital status, even the hours they sleep etc. effect the results. However, until the data included the features mentioned above is prepared, we should pay attention the features shown in features part. Therefore, after examining the data and check for the nulls and insignificant ones we conducted the statistical exploration part.

Correlation

When the number of features increases, there is a better chance for features to be affected by similar cause, basically multicollinearity problem. Therefore, we created a correlation matrix and checked the correlations of each feature in the Table 3.

We discovered that heatmap shows no crucial correlations among independent variables. On the other hand correlations among players in the same team may be seemed as critical however it is normal to get these results due to fact that the players in a team are chosen by managers as the budget of the team allows hence it can be said that players in a team have similar talents. In the ML part we will get rid of these intermediate correlations among independent variables.

Methods and Feature Selection

We employ three Machine Learning methods namely Artificial Neural Network (ANN), XGBoost and Vector Support Machine (VSM). By conducting XGBoost, we use ensemble approach which consist of supervised learning meta algorithms for predicting output target feature by aggregating individual learning algorithms to lower their variance error source or by boosting sequentially built once to simultaneously lower their squared biased error and variance error sources. By conducting ANN, we use multi-layer perceptron method which consists of supervised network based on learning algorithms for predicting output target feature by dynamically processing output target and input predictors data through multi-layer network of optimally weighted connections of nodes. Thirdly, VSM consists of supervised boundary based learning algorithm for predicting output target feature by separating output target and input predictor features data into optimal hyperplanes.

In order to have a simpler and well-working mechanism, data is needed to simple as well. To do so, we need to reconstruct our raw data and transform it to smaller in dimensions. Working with less dimensions gives us some advantages. As the number of features increases, the model becomes more complex. The more the number of features, the more the chances of overfitting. A machine learning model that is trained on many features, gets increasingly dependent on the data it was trained on and in turn overfitted, resulting in poor performance on real data, beating the purpose.

We used four dimensionality reduction methods. PCA is the first linear dimensionality reduction method we use in our data. PCA is simply based on variance of feature we have in our raw data. It takes features with maximum variances as principal components. LDA is the second linear dimensionality reduction method. LDA is based on class separability. According to this method, examples from same class are put closely together by the projection and

examples from different classes are placed far apart by the projection. PCA orients data along the direction of the component with maximum variance whereas LDA projects the data to signify the class separability. RFE is a feature selection method which fits a model and removes the weakest feature until the specified number of features is satisfied. Here, the features are ranked by the model's coefficient or feature importance attributes. LASSO is a linear model which estimates sparse coefficients and is useful in some contexts due to its tendency to prefer solutions with fewer parameter values.

Results

In this paper, we have tried analyse application of various ML methods, such as Artificial Neural Network, XGBoost and Vector Support Machine, on prediction of sport games' results. Then, after we used these methods, we compare and conclude which method gives us the best solution for our problem.. Finding the proper data and feature selection are the two most important and difficult parts of the work due to fact that there is a huge industry make profits from soccer. We have 4 dimensionality reduction methods and we use 3 ML techniques. So, we have 15 different result, including each ML with raw data (not reduced). All accuracy rates, Mean Squared Errors and Mean Absolute Errors as shown in Table 4. We get the highest accuracy rate with XGBoost by using LASSO dimensionality reduction method and XGBoost with raw data. Both accuracy rates are 56 percent. However, other methods that we use are not useless since their accuracy rates varies between 46 percent and 56 percent. So, we can conclude that any method we use in our prediction model can give us the expected results.

In addition, those methods which we get the highest accuracy rates are also with the lowest MSE scores. This supports XGBoost is the most suitable method to use.

References

https://www.sciencedirect.com/science/article/pii/S2210832717301485#b0155

https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e

https://sebastianraschka.com/faq/docs/feature_sele_categories.html

https://www.analyticsindiamag.com/what-are-feature-selection-techniques-in-machine-learning/

Appendix

Table 1

```
In [67]: data[inp].describe().transpose()
Out[67]:
```

	count	mean	std	min	25%	50%	75%	max
HomePlayer_1_Overall	13459.0	76.265844	5.451313	43.000000	73.000000	76.000000	80.000000	91.000000
HomePlayer_2_Overall	13459.0	72.901924	5.071392	51.000000	70.000000	73.000000	76.000000	89.000000
HomePlayer_3_Overall	13459.0	74.665428	5.048232	45.000000	72.000000	75.000000	78.000000	89.000000
HomePlayer_4_Overall	13459.0	74.521733	5.259335	48.000000	72.000000	75.000000	78.000000	89.000000
HomePlayer_5_Overall	13459.0	72.690393	5.099230	45.000000	70.000000	73.000000	76.000000	92.000000
HomePlayer_6_Overall	13459.0	74.515491	5.173175	46.000000	72.000000	74.000000	78.000000	92.000000
HomePlayer_7_Overall	13459.0	74.516606	5.217039	45.000000	71.000000	74.000000	78.000000	92.000000
HomePlayer_8_Overall	13459.0	74.496025	5.351867	45.000000	71.000000	74.000000	78.000000	91.000000
HomePlayer_9_Overall	13459.0	75.228843	5.518267	46.000000	72.000000	75.000000	79.000000	94.000000
HomePlayer_10_Overall	13459.0	75.653615	5.737594	46.000000	72.000000	75.000000	79.000000	94.000000
HomePlayer_11_Overall	13459.0	75.894420	5.512998	46.000000	72.000000	76.000000	79.000000	93.000000
AwayPlayer_1_Overall	13459.0	76.277732	5.474348	51.000000	73.000000	76.000000	80.000000	91.000000
AwayPlayer_2_Overall	13459.0	72.840478	5.132908	48.000000	70.000000	73.000000	76.000000	89.000000
AwayPlayer_3_Overall	13459.0	74.607177	5.121790	45.000000	72.000000	75.000000	78.000000	89.000000
AwayPlayer_4_Overall	13459.0	74.488669	5.330000	48.000000	71.000000	75.000000	78.000000	90.000000
AwayPlayer_5_Overall	13459.0	72.673378	5.099822	47.000000	70.000000	73.000000	76.000000	88.000000
AwayPlayer_6_Overall	13459.0	74.333011	5.241610	46.000000	71.000000	74.000000	78.000000	92.000000
AwayPlayer_7_Overall	13459.0	74.358942	5.309509	46.000000	71.000000	74.000000	78.000000	93.000000
AwayPlayer_8_Overall	13459.0	74.374619	5.366319	45.000000	71.000000	74.000000	78.000000	92.000000
AwayPlayer_9_Overall	13459.0	75.088640	5.598291	46.000000	72.000000	75.000000	79.000000	94.000000
AwayPlayer_10_Overall	13459.0	75.530054	5.830089	41.000000	72.000000	75.000000	79.000000	94.000000
AwayPlayer_11_Overall	13459.0	75.762092	5.568913	46.000000	72.000000	76.000000	79.000000	93.000000

Table 2

AwayPlayer_11_Overall	13459.0	75.762092	5.568913	46.000000	72.000000	76.000000	79.000000	93.000000
Last_5_match_goal_diff_home	13459.0	-0.113679	4.670938	-17.000000	-3.000000	0.000000	3.000000	29.000000
Last_5_match_win_home	13459.0	1.811650	1.207839	0.000000	1.000000	2.000000	3.000000	5.000000
Last_5_match_goal_diff_away	13459.0	-1.826956	4.481613	-18.000000	-5.000000	-2.000000	1.000000	22.000000
Last_5_match_win_away	13459.0	1.388067	1.136348	0.000000	1.000000	1.000000	2.000000	5.000000
Last_5_match_btw	13459.0	-0.006315	2.300317	-5.000000	-2.000000	0.000000	2.000000	5.000000
Team_variance_home	13459.0	14.018262	11.055234	0.247934	6.876033	10.925620	17.322314	118.809917
Team_Mean_home	13459.0	74.668211	3.922293	60.181818	72.000000	74.272727	77.181818	86.545455
Team_variance_away	13459.0	14.246703	11.269896	0.380165	6.975207	11.090909	17.603306	121.289256
Team_Mean_away	13459.0	74.575890	3.969314	61.181818	71.909091	74.090909	77.090909	87.272727
Last_5_match_shout_total_home	13459.0	62.505090	14.827743	4.000000	53.000000	62.000000	72.000000	122.000000
Last_5_match_shout_on_home	13459.0	23.461178	8.522315	1.000000	18.000000	22.000000	28.000000	73.000000
Last_5_match_card_home	13459.0	11.168363	4.256197	0.000000	8.000000	11.000000	14.000000	31.000000
Last_5_match_faul_home	13459.0	71.569805	16.377796	2.000000	60.000000	72.000000	83.000000	150.000000
Last_5_match_corner_home	13459.0	25.354930	7.356339	0.000000	20.000000	25.000000	30.000000	60.000000
Last_5_match_shout_total_Away	13459.0	55.319043	14.875643	0.000000	46.000000	55.000000	64.000000	119.000000
Last_5_match_shout_on_Away	13459.0	20.594844	7.972766	0.000000	15.000000	20.000000	25.000000	60.000000
Last_5_match_card_Away	13459.0	11.858088	4.434887	0.000000	9.000000	12.000000	15.000000	36.000000
Last_5_match_faul_Away	13459.0	71.664685	17.994058	0.000000	60.000000	73.000000	84.000000	132.000000
Last_5_match_corner_Away	13459.0	22.048072	7.094609	0.000000	17.000000	22.000000	26.000000	54.000000
home_team_last_result	13459.0	-0.130247	0.851976	-1.000000	-1.000000	0.000000	1.000000	1.000000
away_team_last_result	13459.0	0.144662	0.849385	-1.000000	-1.000000	0.000000	1.000000	1.000000

Table 3

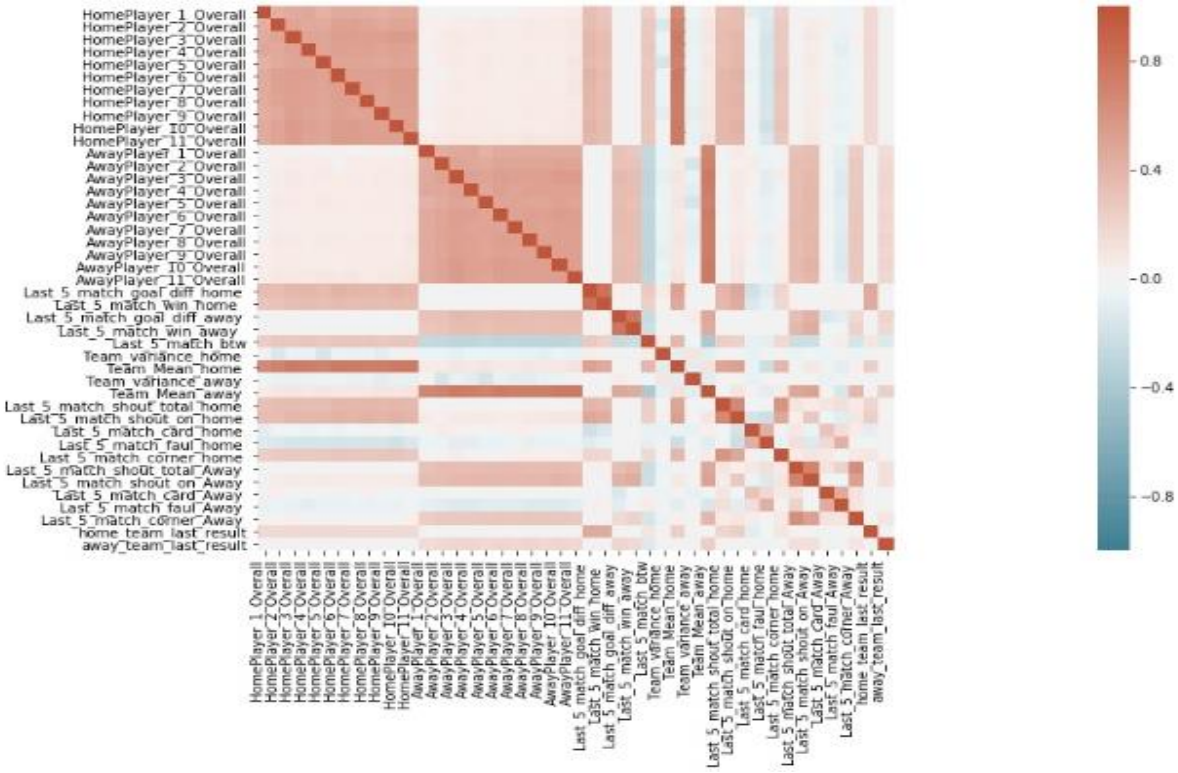

Table 4

	F1Score	AccurancyRate	MSE	MAE	PrecisionScore
XGBoostesulterWithRawDatA	0,438951599	0,567979198	0,985029095	0,611441308	0,503162482
XGBoostesulterWithLassot	0,433615272	0,565750371	0,988417319	0,615156018	0,489138812
ANNWithRawDatA	0,412165893	0,555349183	1,019133449	0,642644874	0,693483709
XGBoostesulterWithLDA	0,387465853	0,544576523	1,037377251	0,662332838	0,363029519
XGBoostesulterWithRFE	0,386066358	0,541976226	1,042912826	0,667904903	0,443170148
SVMWithLDA	0,383401089	0,541976226	1,043980837	0,668647845	0,364054027
ANNWithLDA	0,384809154	0,541604755	1,044692238	0,669390788	0,360329723
ANNWithLassot	0,41426516	0,540490342	0,988793072	0,632243685	0,48898119
ANNWithRFE	0,320435802	0,514858841	1,09469883	0,722882615	0,379093075
XGBoostesulterWithPCA	0,313996491	0,472139673	1,167762876	0,806463596	0,369558051
SVMWithRawDatA	0,213124421	0,469910847	1,173949532	0,812778603	0,156636949
SVMWithPCA	0,213124421	0,469910847	1,173949532	0,812778603	0,156636949
SVMWithLassot	0,213124421	0,469910847	1,173949532	0,812778603	0,156636949
SVMWithRFE	0,213124421	0,469910847	1,173949532	0,812778603	0,156636949
ANNWithPCA	0,329656781	0,468053492	1,159462652	0,802748886	0,393769842