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>

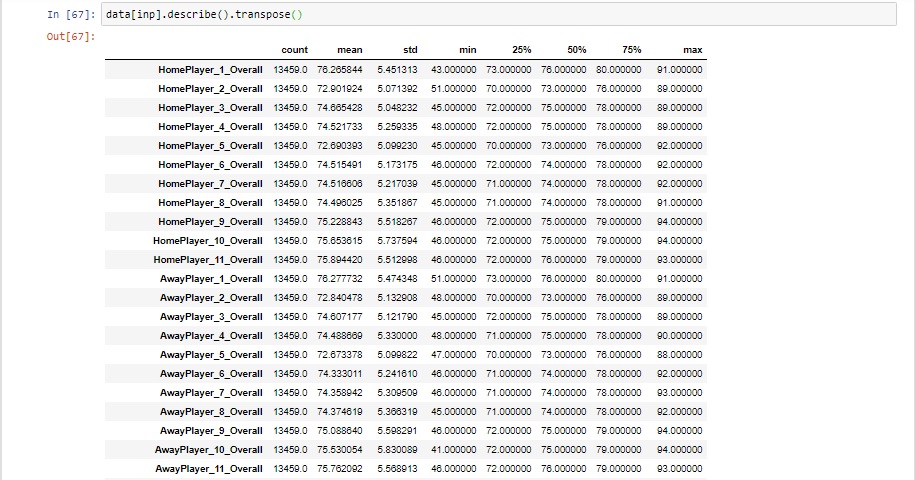
<https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e>

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

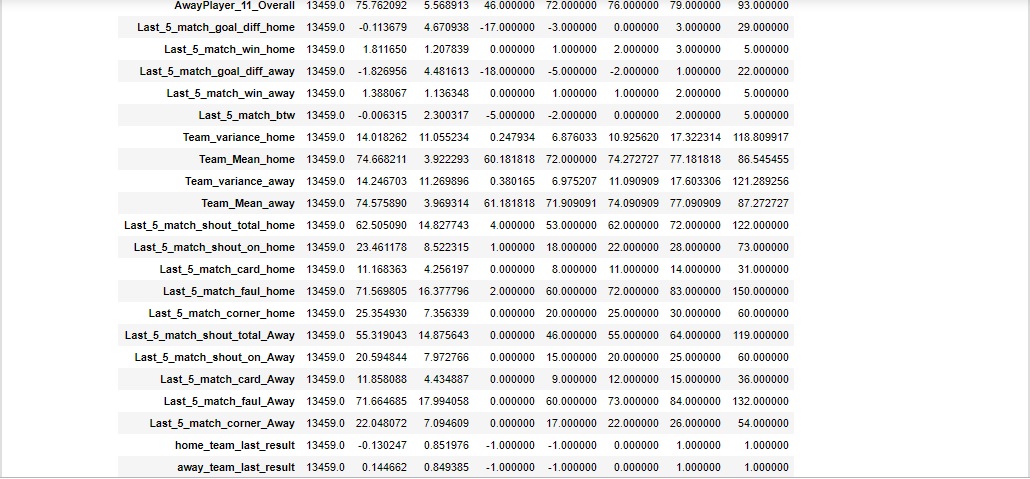
<https://www.analyticsindiamag.com/what-are-feature-selection-techniques-in-machine-learning/>

***Appendix***

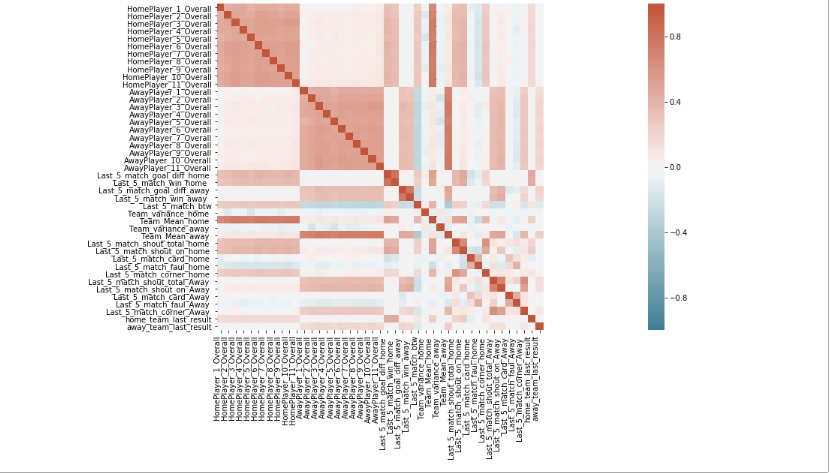
**Table 1**



**Table 2**



**Table 3**



**Table 4**

