Sports Prediction Using Machine Learning

Ragini Singla , Dr.Amardeep Singh
[1]CSE, Punjabi University, Patiala, India,
[2]CSE, Punjabi University ,Patiala , India.

Abstract –Sports Prediction is one of the most used applications in betting industry. Machine learning is used extensively in this application. We have used one of the largest soccer datasets and used it to predict the outcome of the most popular football leagues in Europe and World. We have used Naïve Bayes, Support Vector Machine and Logistic Regression for classification and compared them before and after applying normalization. We have used various parameters while performing prediction which are related with offensive and defensive properties of the team like Goals scored, conceded, corner kicks, red, yellow cards etc. every season. Classification using machine learning algorithm brings displays the most accuracy with and without normalization.

Keywords – SVM, Sports-Prediction, Naïve-Bayes, Logistic-Regression, Football, Soccer.

1. INTRODUCTION

Machine learning refers to the study of different computer algorithms and dataset of artificial intelligence (AI)[10]. It enables us to automatically learn and ameliorate itself from experience without any human intervention. In the study of Machine Learning, we access the data and compute it further to achieve goals. The learning of data means observing the data or datasets, for instance, experiences or set of instructions to check the pattern inside the given data. The system should be highly reliable. Machine Learning algorithms[2][7] have been widely used in number of applications such as email filtering, Number Plate recognition, computer vision and many more. It mainly focuses on computation predictional statistics, which makes predictions using algorithms of computers. Machine Learning also sometimes refers as predictive analytics[9] as it predicts the data.

There are several Machine Learning algorithms which are worth to look at:

1.1 Support Vector Machine: Support Vector Machine or SVM[17] is one of the most commonly used algorithm of machine learning. It has the ability to be utilized for regression or classification-based challenges or problems. However, the most important feature of this algorithm is that it is used for solving the classification problems. By using this algorithm, we can plot every single data item in n-Dimensional form of points, where n is number of features, we are using along with the value of feature used of a value for specific coordinate. After this, we apply classification method to find out the hyper-plane, which classifies two classes vividly.

The main advantage of this algorithm is that it is quite efficient when dealing with high dimensional spaces. Also, SVM offers better results when number of process are much more as compared to number of given samples. Apart from this, in the decision function, it uses subset of training points. Consequently, it requires less memory resources when compared with other algorithms. Nonetheless, there are some downsides as well. When the features are greatly outnumbered than number of samples, to avoid over-fitting of kernel functions, it is imperative to regularize the term. Furthermore, SVMs usually unable to provide the probability estimate because it can be measured using five-fold cross-validation, which is expensive again.

1.2 Multinomial Naive Bayes

Another type of popular algorithm is Multinomial Naïve Bayes. This algorithm is impeccable for classification[14] along with some discrete features, for example, word count of text data classification. Although it requires integer feature counts, tf-idf vectors i.e fractional count can be used as well. Multinomial Naïve Bayes algorithm is a special type of Naïve Bayes[16] algorithm. It estimates probability P(c|x) where c represents the class of the result and x refers to the instances.

$$P(c|x) = P(x|c) * P(c) / P(x)$$

This algorithm has numerous benefits. Firstly, it is easily implementable. Besides, it requires small of training datasets to know about the test data, which makes it more efficient, accurate and highly scalable. On the flip side, this algorithm assumes automatically that each and every attribute belongs to independent vector, which is problematic in real-life as this type of situation rarely happens. Moreover, when categorical variable encounters category in the test dataset, model sets the value 0 of probability and halts the prediction[13]. This situation often refers to as Zero Frequency. However, this can be resolved through smoothing technique.

1.3 Logistic Regression

Logistic Regression is regression analysis[20] algorithm, which is used to analyze dependent variable. In other words, it is used to describe the relation between dependent binary and nominal vectors. It is taken from the concept of statistic. This is the reason statistical model. In Logistic Regression or logit regression, it offers only two values i.e. pass/fail which can be noticed through indicator variable. By using this algorithm, we have two values either 0 or 1. The main advantage of Logistic Regression algorithm is that it is one of the simplest algorithms of machine learning[12]. Apart from this, it is easy to implement and train the data. It has very less computation power. Moreover, it has high tolerance to over-fitting values. Having said that, it makes assumption of linearity data which lies in between the dependent and independent variable, which is not feasible in actual world as there is hardly linear separable. Furthermore, it is mainly used for discrete functions. Also, when the observations are less as compared to features, Logistic Regression algorithm

does not perform well because it enables the problem of overfit. And, sometimes it is quite difficult to get the relation of complex datasets. In such case scenarios, Neural Network[1][11] performs extremely well.

2. Sports Prediction

We are making prediction on who will have the best chance to win European Football Championship[15], which is one the favourite sport and league in the world. Prediction in football is used widely in betting[6][24-25]. Large number of bookies have their own prediction model and their prediction is success [5] around 53 percent of the time. One of the best Football pundits Mark Lawrenson has a prediction accuracy of around 52.6 [19]. We have used machine learning algorithms like Support Vector machines (SVM), Multinomial Naive Bayes, and Logistic Regression[3][4] for prediction of Football Match outcomes.
We have used two datasets, first one is taken from Kaggle [5], which is used to get exploratory data analysis. But there are some more data needed for prediction analysis [21-23] and for that we have used dataset from footballs-data.co.uk website [18] with large number of data including seasons of leagues.
We have started with the Kaggle dataset which is in sqlite format and has tables related with the country, league, player, team etc[8]. With the information of over 25000 matches and 10000 players, it is one of the largest datasets of eleven football leagues in Europe.
We have used python libraries like numpy for numerical calculations, matplotlib for plotting graphs, pandas for data analysis and manipulation, sqlite3 for database connectivity, scipy for mathematical computations and sklearn for machine learning algorithms.
We have connected the database in the initialisation phase of the code after importing the libraries using following commands:

```
with sqlite3.connect('database.sqlite') as con:
    countries = pd.read_sql_query("SELECT * from Country", con)
    matches = pd.read_sql_query("SELECT * from Match", con)
    leagues = pd.read_sql_query("SELECT * from League", con)
    teams = pd.read_sql_query("SELECT * from Team", con)
    tempmatch = pd.read_sql_query("SELECT * from Match", con)
    matches3 = matches2 = matches
```

Figure 2.1 - Connecting Sqlite with python

After connecting the database, one will be able to call the values from countries, matches, leagues, teams, temp match, and matches3. For example, if we call leagues, we will get the following output:

```
leagues
```

	id	country_id	name
0	1	1	Belgium Jupiler League
1	1729	1729	England Premier League
2	4769	4769	France Ligue 1
3	7809	7809	Germany 1. Bundesliga
4	10257	10257	Italy Serie A
5	13274	13274	Netherlands Eredivisie
6	15722	15722	Poland Ekstraklasa
7	17642	17642	Portugal Liga ZON Sagres
8	19694	19694	Scotland Premier League
9	21518	21518	Spain LIGA BBVA
10	24558	24558	Switzerland Super League

Figure 2.2 - Calling leagues from database

After verifying the database connectivity, we are now able to call the attributes using the column name as displayed in figure above.

We have selected three countries for our analysis. As per the popularity of the league is concerned, i have selected English, German and Spanish Leagues for analysis. These are the most watched and competitive soccer leagues in world. After selecting the leagues, we have merged the league matches. Dataset that we have also does not have the results data of the matches, so we added the results of the matches by comparing the home and away team goals. So, if one team has more goals than the other, then that team won the match, and if the number of goals are same for both the teams, then the match is ended as a draw. Function used for fetching results is shown below where x is home team goals and y is away team goals:

```
def results(row):

if row['x'] == row ['y']:
        val = 0

elif row['x'] > row['y']:
val = 1

else:
        val = -1
returnval
```

Leagues are separated after fetching the results by creating a new dataframe for storage. And after separating, grouping is performed along with the division on the basis of seasons using group by functions. For example:

eng = englishpl.groupby('season')

After getting the data ready, we started to plot different data using ggplot like total goals scored every season as shown below:

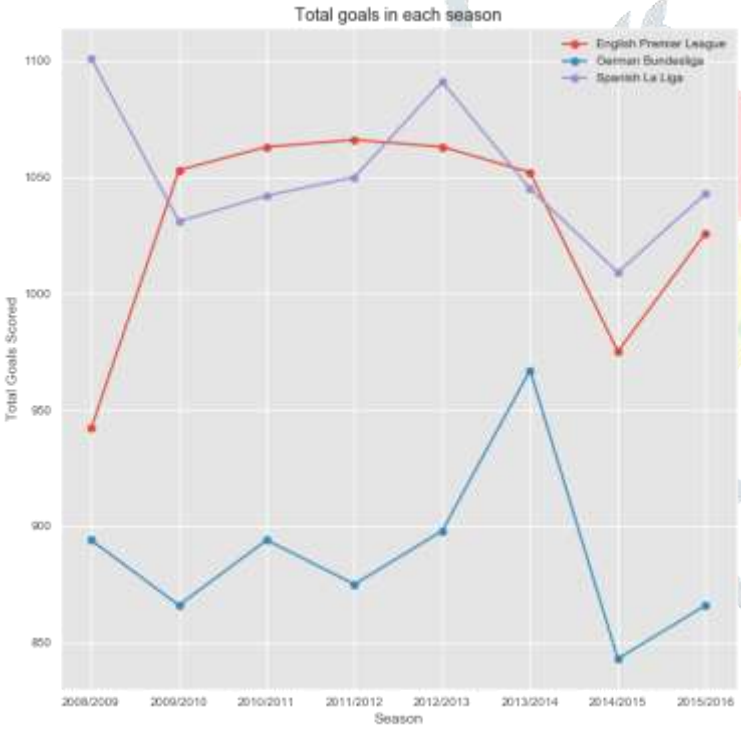

Figure 2.3 - Plotting total goals scored every season.

As above graph shows, there are more goals scored in Spanish La Liga than any other leagues we are analysing , but German Bundesliga has two teams less than other two leagues, so total games played are 306 in Bundesliga, as compared with 380 in other two leagues. Next calculation we made is the average number of goals per game that shows that Bundesliga teams score more than 3 goals per game on average as their totally tally is around 970 goals in 306 matches, whereas the average is 2.7 in English Premier League and 2.8 in Spanish La Liga. In our analysis phase another important calculation we made displays which team scores most at their home ground. For that we have merged home team and their matches and then plot them as shown below:

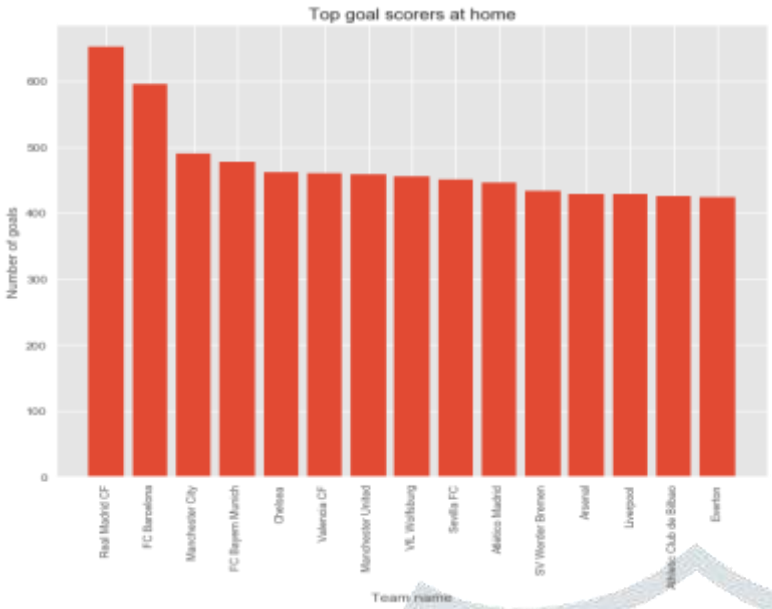

Figure 2.4 - Top Goal Scorers at home turf.

The above figure displays the top 15 highest home goal scorers and along with this, we have also calculated and figured out the top 15 teams with best gal scoring record in away matches as shown below:

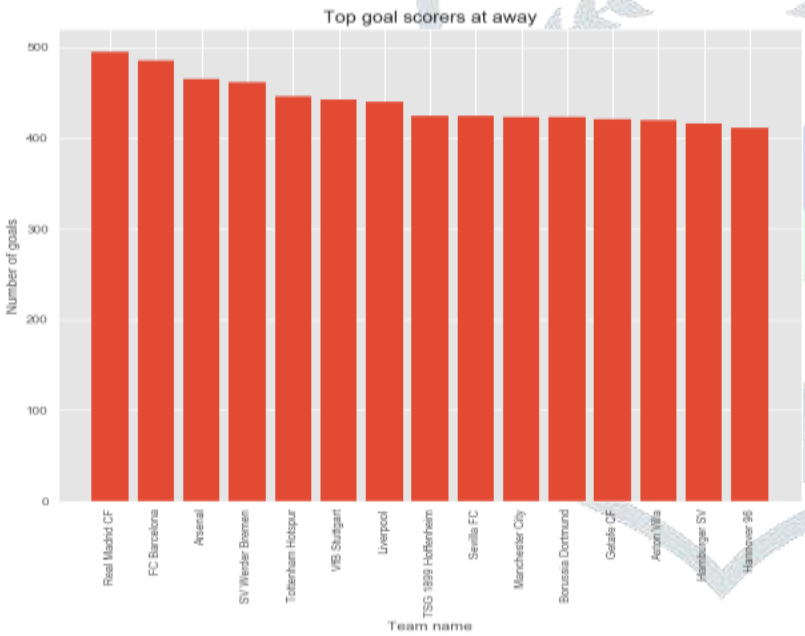

Figure 2.5 - Top 15 goal scorers in away matches.

Next, we found the probability using entropy. We calculate entropy to find the predictability as higher the value of entropy more the unpredictability in the results. We have used scipy library to import entropy and created the function to calculate match entropy. Entropy is then plotted as below:

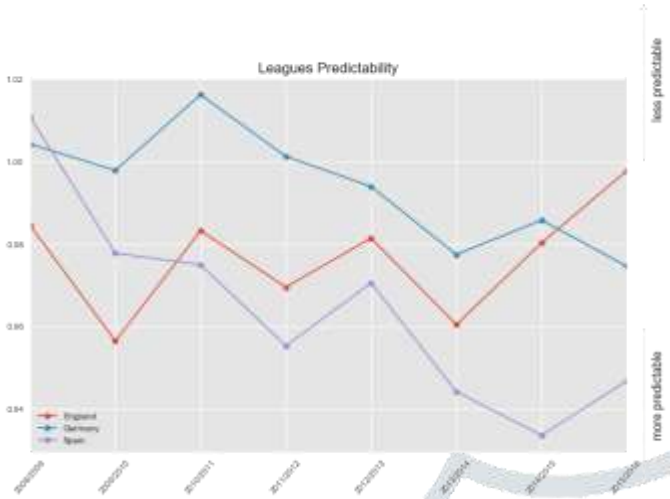

Figure 2.6 - Entropy Values displaying probability

Spanish La Liga has same top three teams in around half of seasons. Bayern Munich won Bundesliga 6 times since 2008, while in English Premier League, Manchester United and Chelsea won three times since 2008 and Manchester City won twice. Therefore EPL is more unpredictable than other leagues.

EPL also seeing decrease in home advantage in every season. We also downloaded the EPL data from year 2000[18]. The dataset includes large number of game statistics like Full-time Goals, Half-time Goals, number of corners, free-kicks, red and yellow cards, total number of shots on goal, successful number of shots etc.

After reading the dataset, first add it to the dataframe, then various features are selected which are used for prediction. We put all the data into a single dataframe. Various features used are:

Table 2.1 – Attributes used for Prediction

Label	Full Form
FTHG	Full-time Home Goal
FTAG	Full-time Away Goal
FTR	Full-time Result
HTHG	Half-time Home Goal
HTAG	Half-time Away Goal
HTR	Half-time Result
HS	Home Shots
AS	Away Shots
HST	Home Shots on Target
AST	Away Shots on Target
HF	Home Team Foul
AF	Away Team Foul
HC	Home Team Corner
AC	Away Team Corner
HY	Home Team Yellow Card

Label	Full Form
AY	Away Team Yellow Card
HR	Home Team Red Card
AR	Away Team Red Card

After creating a merged data frame, we created a csv file which can be used for future analysis. After the data merged in a csv file, we calculate the average home and a ways goals scored and conceded. We calculated offensive or attacking strength of every team by looking at the number of goals they have scored at home and away, which can be a part of prediction strategy. These are calculated using tables we put all the team home and away goals into four different variables and then divide it with 304 home and away matches and the output to further divide by average home and average away goals scored and conceded. We have used following procedure for calculation:

table.team = results.home.team.all().values
table.x = results.home.FTGH.sum().values
table.y = (table.x/304)/z

Where x==Home Goals Scored; y==Home Attacking Strength; z==average-home-scored
Using the above methods, we calculated Attacking and Defensive strengths of teams.
After getting all the values, we created two training sets where one has attacking and defensive strength related data and other has attacking and defensive along with shots and corners. These two training sets are then classified with the three machine learning algorithms i.e. Support Vector Machines, Logistic Regression and Multinomial Naïve Bayes. Sklearn library is used for calling machine learning algorithms. When applied with the training datasets, we found the following accuracies:

Table 2.2 – Algorithm comparison before normalization

Algorithm	Accuracy (Score 1)	Accuracy (Score 2)
SVM	0.539	0.612
Naïve Bayes	0.484	0.573
Logistic Regression	0.554	0.591

The above table shows the accuracy without normalization, and we have also performed normalization on the trained datasets, for which normalize function is imported from sklearn library and accuracy after normalization is shown below:

Table 2.3 –Algorithm comparison after normalization

Algorithm	Accuracy with Normalized Train data (Score 1)	Accuracy with Normalized Train data (Score 2)
SVM	0.556	0.594
Naïve Bayes	0.484	0.484
Logistic Regression	0.559	0.610

This is clear from the above table that performance of Naïve Bayes and SVM reduced after normalizing train dataset, while the performance of Logistic Regression improves little after normalizing train data.

3. Conclusion

Prediction is one of the most popular applications in the mathematical and machine learning. There are large number of machine learning algorithms available with python libraries, we have used Support Vector Machine (SVM), Multinomial Naive Bayes and Logistic Regression. This is seen that when we add corners and shots in our work, then we get better results. Naive Bayes and SVM provides better accuracy without adding normalization. And if we add normalization, then in that case, Logistic Regression provides best results. The betting accuracy in our work is around 53 percent, and the SVM model brings around 61 percent accuracy. Accuracy becomes better with the addition of new data attributes like player or team form. To better the accuracy of the prediction, we can also use twitter analysis and XG Boost model.

References

[1]McCullagh, J., 2010. Data mining in sports: A Neural Network Approach. Intl. J. of Sciences and Eng, 3, pp.131–138.
[2] Tan, P.N., Steinbach, M. & Kumar, V., 2006. Introduction to data mining, Pearson Addison Wesley Boston.
[3]Landwehr, N., Hall, M. & Frank, E., (2005). Logistic model trees. Machine Learning, 59(1), pp.161–205.
[4] Blundell, J., (2009). Numerical Algorithms for Predicting Sports Results. University of Leeds, School of Computer Studies.

[5]The ultimate Soccer database for data analysis and machine learning, "https://www.kaggle.com/hugomathien/soccer"
[6]Buursma, D. (2011). Predicting sports events from past results: Towards effective betting on football matches. In Proceedings of the 14th Twente Student Conference on IT. Twente, Holland.
[7]Alpydin, E. (2010). Introduction to machine learning (2nd ed.). Cambridge, MA, London, England: MIT Press.
[8] Deshpande, S. K., & Jensen, S. T. (2016). Estimating an NBA player's impact on his team's chances of winning. Journal of Quantitative Analysis in Sports, 12(2), 51–72.
[9] Ruiz, H., Power, P., Wei, X., Lucey, P. (2017). "The Leicester City Fairytale?": Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons. Proceedings of KDD Conference. El Halifax, Nova Scotia Canada.
[10] McCabe, A., &Travathan, J. (2008). Artificial Intelligence in Sports Prediction. Fifth International Conference on Information Technology: New Generations.
[11] Kahn, J. (2003). Neural Network Prediction of NFL Football Games (pp. 1194–1197). Washington, DC: IEEE Computer Society
[12]Horvat, T., & Job, J. (2020). *The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1380.* doi:10.1002/widm.1380
[13] Cao, C.: Sports data mining technology used in basketball outcome prediction. Masters Dissertation. Technological University Dublin, 2012.
[14] Zhang, S., Pan, Q., Zhang, H. *et al*. Prediction of protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and Naive Bayes Feature Fusion. *Amino Acids* **30,** 461–468 (2006). https://doi.org/10.1007/s00726-006-0263-8
[15] Football betting – the global gambling industry worth billions, "http://www.bbc.com/sport/football/24354124"
[16] Naïve Bayes, http://scikit-learn.org/stable/modules/naive_bayes.html.
[17] Support Vector Machines, "http://scikit-learn.org/stable/modules/svm.html"
[18] Historical Football Results and Betting Odds Data, "http://football-data.co.uk/data.php".
[19] Pinnacle vs Mark Lawrenson, "https://www.pinnacle.com/en/betting-articles/Soccer/Mark-Lawrenson-vs-Pinnacle-Sports/VGJ296E4BSYNURUB".
[20] Logistic Regression, "https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html".
[21] KeshtkarLangaroudi, M., Yamaghani, M. (2019). Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches: A Survey. *Journal of Advances in Computer Engineering and Technology*, 5(1), 27-36.
[22] K. Apostolou and C. Tjortjis (2019) Sports Analytics algorithms for performance prediction, *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, PATRAS, Greece, 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900754.
[23] Thabtah, F., Zhang, L. &Abdelhamid, N. NBA Game Result Prediction Using Feature Analysis andMachine Learning. *Ann. Data. Sci.* **6,** 103–116 (2019). https://doi.org/10.1007/s40745-018-00189-x
[24] Zhu P., Sun F. (2020) Sports Athletes' Performance Prediction Model Based on Machine Learning Algorithm. In: Abawajy J., Choo KK., Islam R., Xu Z., Atiquzzaman M. (eds) International Conference on Applications and Techniques in Cyber Intelligence ATCI 2019. ATCI 2019. Advances in Intelligent Systems and Computing, vol 1017. Springer, Cham. https://doi.org/10.1007/978-3-030-25128-4_62
[25] Sangwon Na, Yiran Su &Thilo Kunkel (2019) Do not bet on your favorite football team: the influence of fan identity-based biases and sport context knowledge on game prediction accuracy, EuropeanSportManagementQuarterly, 19:3, 396418, DOI: 10.1080/16184742.2018.1530689