European soccer match prediction

Quang Pham

Temple University

[tug47906@temple.edu](mailto:tug47906@temple.edu)

**Introduction:**

Soccer is one of the most popular sport in the world and Europe is the center of soccer, with 5 biggest soccer leagues of five countries: England, Spain, Germany, France and Italy. Nowadays, soccer is not just a sport, but also an industry, which people try to applied many fields of science to improve the quality of soccer matches. One of the famous field applied to improve matches is Data Science. For example, from a large amount of data, each team will have a data scientists team working hard in order to predict the best strategy for the upcoming match, or to find the way to improve the team’s quality through transfer and trading market. For others business, such as gambling company or sport analytics company, data science is used a lot to achieve many insights about teams and player, or to predict the future results in order to give the best odd ratio.

However, predicting the result is not an easy job. First, the dataset is very complex, because there are many kinds of events happen in a soccer match, with different times and locations have different impact to the match. For example, an attempt at first minute from middle of the field is totally different with an attempt in the last minute near the goal-line. Moreover, there are many external attributes can have impact on the result: weather, winning streaks, players’ health, etc. Secondly, players are not robots, so we cannot predict everything correctly. In Premier League 2015-2016 season, the odd for Leicester to win the title is 5000-1 against, so no one believed that Leicester will win the league, but at the end of the season Leicester became the champion. Soccer has a lot of surprise, and it is a big challenge to predict the match results.

There are many researches with the goal to get the best model to predict match result. One of the common machine learning tasks is classification, as in the research [1] of Rory P.Bunker and Fadi Thabtah, which used Neural Network model to make a classify prediction. This research not really focus on finding the best model, but to prove that Machine Learning is appropriate for predicting sport and the result turn out that after testing with different data from different sports, results from different types of sport need different approach, and ML models stills need improvement. Additionally, different approach algorithms can lead to different accuracy [2], for example: Naïve Bayes, Support Vector Machine(SVM) and Random Forest. This classifiers were tested on the dataset with 3800 games in training set and 760 games in testing sets, each game contains team, score, winner and the number of goals. The results of different classifiers is different, the highest result is formed by Baseline model(0,6 in accuracy), and the second one is Naïve Bayes(0,56) accuracy.

**Approach and Methodology:**

My project goal is to build a model and train it with a dataset of 940.000+ events from 9704 soccer games across Europe. Each event has many attributes, such as type of events, player, time, position,etc. Moreover, the dataset also has the record of each games odd and results. From the dataset of 900000+ events sorted by time, I changed into the dataset of 10000+ matches, each example is a single match with attributes are the counts of each type of. From this dataset, I will do preprocessing steps to make a new dataset with each match is an example. For an example, we will have home team and away team name, teams record (goals, games, etc.) and the past record between every 2 teams matches from all matches of them in the past. For example, for the match number 1001, I will have the records of Home team in first 1000 matches, away team in first 1000 matches, and the records of every matches that Home team met Away team in first 1000 matches. My training dataset will start from match number 1001 . If the match is number X, the row of that match will contain the records of Home and Away team in first (X-1) matches, as well as record of matches between Home Team and Away team in [0:X] matches. One other important thing is that because the record could be cumulative, so I will use average (sum of record divided by number of matches) to make the scale of data. By that way, I will always use the data from the past with a limited range as the input to predict the results of future matches.

For this project, my goal is only to predict whether home team can win or not, which is a binary classification. I will build model based on kNN, Decision Tree, Random Forest, Naïve Bayes and a Feed Forward Neural Network model trained 1000 epoch to predict the outcome. However, each model requires different type of input to work efficiently, so for each model I will use features selection to select the best number of input attributes to for the best results. The input data have over 300 features, so without features selection, the accuracy is just about 52%, not better than a random guest. But when I only use small number of features, the accuracy of kNN and Random Forest increased.

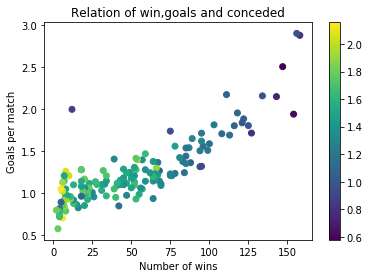
**Results:**

**EDA:**

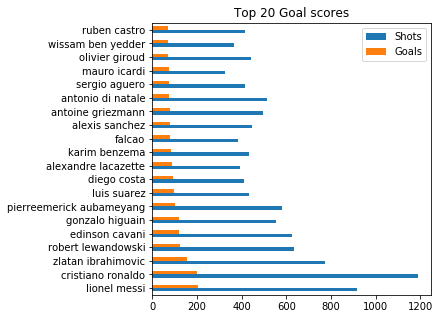
Before starting with prediction steps, I have done some EDA to get interesting insight from the dataset. There are many elements can affect the result of a soccer match, and I perform EDA to see the overall quality of leagues, players, as well as the trend in playing of teams. For example, we can see the details about matches, goals and cards of the 5 soccer leagues in the table below:

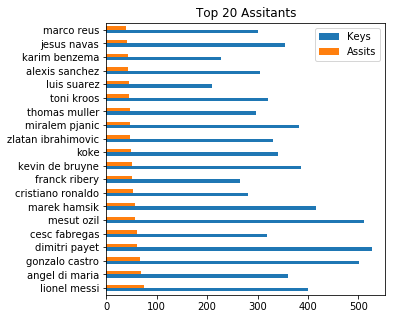
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Leagues** | **Games** | **Goals** | **Goal per game** | **Cards** | **Cards per game** |
| Bundesliga | 1690 | 4871 | 2.88224 | 6253 | 1.850 |
| Ligue 1 | 2107 | 5283 | 2.50735 | 7891 | 1.87523 |
| Serie A | 2106 | 5580 | 2.64957 | 10703 | 2.5410 |
| Premier League | 2120 | 5812 | 2.74150 | 7891 | 1.8610 |
| La Liga | 2089 | 5787 | 2.77022 | 11634 | 2.7845 |

From the table, we can conclude that Bundesliga has highest rate of goal per games, so we can expect that match in Bundesliga will have more goals than other leagues. Premier Leagues and La Liga have a quite similar rate of goal per game, but the number of cards per game of La Liga is higher, so matches in La Liga will have high number of goals and cards. Based on the fact that La Liga has many famous players played, we can conclude that La Liga is very competitive and famous, which lead to high number of goals and cards.

****

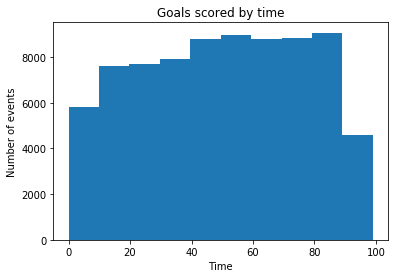
In the plot above, we can see a strong relation among number of wins, goals per matches and conceded goals per match. The higher goals per match and lower conceded goals, the higher number of wins team can have. For example, two teams with highest numbers of wins have very high number of goals per-match( near 3.0) and low number of conceded goals per match. We can see the correlation in the plot is near to 1 because we can see the increase rate. However, there are some outlier in the plot. For example, we have a team has a very low number of conceded goals, but because that team also has low number of goals per match, the team also has low number of wins also.

Like many other sports, one important path of soccer is scoring. So to evaluate the value of players, one important aspect is number of goals players have. 

In the plot, we can see that Ronaldo and Messi are two most valuable players if we focus only in number of shots and goals. Ronaldo and Messi dominate soccer world in the past 10 years, and in statics field they also dominate with very high number of goals. 

If we use number of assists and key pass as a scale to evaluate players, we can also see that Messi is still the most valuable players in making chances for teammate, while Di Maria and Castro finish in second and third place.

One more thing I notice that most of the goals happened from 40 minutes and 80 minutes, while there is less goal at the beginning and the end of matches. As we can see in the plot below, the time between 50 and 60 and between 80 and 90 can be very interesting to audience to watch because there a lot of goals at that time.



**Predicting Model**

After having the preferred dataset, I started building model, trained on first 6000 variables and tested on the others. My first approach is directly used the dataset with more than 300 features.

When using all features, my accuracy is quite low(57% for kNN with 250 neighbors, 65% for FNN with 400 epoch, 51% for Decision Tree and 55% for Naïve Bayes). I realize that among many features, there are many features can make noise to the data, beside the fact that for kNN or Decision Tree, high number of features can lead to the low accuracy. So I tried SVD and features selection, but only features selection can increase the accuracy. After training and testing with most suitable hyper parameter, here is the best accuracy I have:

Number of features used is an important element lead to the accuracy of model. Choosing very small number of features in kNN and Decision Tree (k=5) can increase the accuracy dramatically. If I change number of features too much, the accuracy can change around 50 to 63. For Naïve Bayes and Random Forest model, changing number of features from 5 to 100 can only change the accuracy lower by 1 to 2 percent.

For FNN model, my approach is quite different, I tried different number of features and trained on a model with 8 layers for 40 times, with ReLu is the main activations, and “softmax” is the last activation for classification model and compile with ‘sgd’. The results when choosing different number of features is not so different, so I decided to go ahead with 200 features and train the model for 1000 epoch. Among 1000 times, my highest accuracy is around 66,5%. My result for FNN is not very satisfy myself because I expect the accuracy higher than 70%.

**Conclusion:**

Predicting sport results is a very difficult task[3]. My project is to predict whether home team can win or not, based on a dataset of more than 950.000 events. My models are trying to make a binary classification, and the accuracy is mostly around 65%. For lazy model like kNN or model with a low requirement of computational costs like Naïve Bayes, the accuracy is quite good. However, for FNN model which cost a lot of time to train, 65% may not very good. There are many reasons lead to this problem, first is predict sport itself never an easy problem since there are many elements can lead to the results, and even our human cannot predict the results correctly because there are many times weak team can defeat strong team, and dominant team cannot win the match. Secondly, I believe my FNN model is not build strongly enough, because using ReLU and sigmoid may lead to the loss of data as well as I don’t have enough time and resources to train the model for the higher accuracy. However, 66% is not a very bad number and I believe with more work, I can increase the accuracy of my model.

**References:**

1:” A machine learning framework for sport result prediction”, Rory P.Buner and Fadi Thabtah. <https://www.sciencedirect.com/science/article/pii/S2210832717301485>

2: “Predicting Soccer Match Results in the English Premier League”, Ben Ulmer and Mathew Fernadez.<http://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf>

3.” Lucky is Hard to Beat: The Difficulty of Sports Prediction”, Raque Y.S.Aoki, Renato Martins Assuncao, Perdo O.S..Vaz de Melo.