NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "HSE and University of London Double Degree Program in Data Science and Business Analytics"

Programming Project Report on the Topic: Predicting outcomes of sport events using machine learning methods

Fulfilled by:

Student of the group БПАД223
Pankova Anzhelika Sergeevna ____________ ____________
(signature) (date)

Assessed by the Project Supervisor:

Rudakov Kirill Aleksandrovich
Teacher
Faculty of Computer Science, HSE University

____________ ____________
(signature) (date)

Moscow 2024

Contents

Annotation

Футбол - один из самых популрных видов спорта, привлекающий как и заядлых фанатов самого вида спорта, так и тех, кому нравится интрига, зрелищность и неповторимость каждой игры. Ставки на спорт с каждым годом становятся все более распространенным развлечением. Ежедневно во всем мире в среднем проходит минимум 2-3 матча топ-5 лиг Европы, а также есть еще и другиче лиги и чемпионаты, то есть каждый день можно делать свои прогнозы на множество футбольных матчей. Именно это делает ставки на футбол такими популярными. Данная работа исследует уже существующие методы, служащие для прогонозирования исходов спортивных событий, а также, создает свой метод, онованный на уже известных способах предсказания и алгоритмах вычисления. Конечно же, можно проделывать всю эту работу и вручную, но гораздо удобнее и быстрее дать компьютеру возможность выполнить данную задачу. Данные, использующиеся для каждого матча должны постоянно обновляться, чтобы не утерять актуальность и помогать прогнозировать наиболее вероятный исход из всех возможных.

Football is one of the most popular sports, attracting both avid fans of the sport itself and those who like the intrigue, entertainment and uniqueness of each game. Sports betting is becoming more and more common entertainment every year. Every day, on average, at least 2-3 matches of the top 5 European leagues take place all over the world, and there are also other leagues and championships, that is, every day you can make your predictions for many football matches. This is what makes football betting so popular. This work explores existing methods used to predict the outcomes of sports events, and also creates its own method capable of predicting the outcomes of football matches based on certain data. The same algorithms will be used to determine the result of each match. Of course, you can do all this work manually, but it is much more convenient and faster to give the computer the opportunity to complete this task. The data used for each match should be constantly updated so as not to lose relevance and help predict the most likely outcome of all possible

Key words : prediction, outcomes, algorithms, betting

1 Introduction

Nowadays, betting is quite popular among football fans as bets make watching games more attractive and ticklish. Forecasting of results is hard, thus making profit from betting is a controversial question because of many factors that can influence on the outcome of exact match. However, there are some existing methods that helpful in predicting results of sport events.

Important definitions

xG - is the sum of the probabilities of scoring chances showing a mathematical value calculated taking into account the parameters of potentially scoring chances, which include shots at goal and dangerous moments without finishing shots.

Beautiful Soup - Python library for extracting data from HTML and XML files. It provides a simple and convenient way to extract data from web pages and makes it easier to work with that data.

Requests - Python module that is used to make working with HTTP requests easier. Helps in passing parameters to URLs.

Standart Scaler - a method for preprocessing data in front of machine learning models, allowing the standardization of data that has very different ranges. It is used to change the size of the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1.

Logistic regression - a machine learning algorithm that accomplishes binary classification tasks by predicting the probability of some outcome.

Multiclass logistic regression - a categorical dependent variable exhibits multiple discrete outcomes, indicating that the model accommodates more than two potential results.

Random Forest - algorithm that combines multiple decision trees to improve the accuracy and robustness of predictions.

LSTM (Long Short-Term Memory) - special kind of recurrent neural network capable of learning long-term data dependencies.

ROC (Receiver Operating Characteristic curve) - a graph that illustrates the performance of a classification model at all possible classification thresholds.

AUC (Area Under the Curve) - a measure that summarizes a model's performance in a single number, measuring the area under the ROC curve. AUC ranges from 0 to 1, where a higher AUC value indicates better model performance.

My strategy and plan

Strategy

At the start of my project, I wanted to build a model that will help to predict results of matches. I am going to use Python and its' libraries to allow myself cover my goal. It should be a model that will be able to update appropriate data automatically and "compare"any two teams from the chosen table. I was going to work with Russian Premier League, so my model will predict outcomes in this league. It is planned to obtain something similar with FIFA rating and design my own one, but for RPL and based on factors that I choose. With the help of websites containing information of each match, it is possible to make such rating. The following criteria will be used : how many goals does the team score, how many concedes, whether the team plays at home or away, possession, fouls, shots, saves, corner kicks and many other factors that will be found and valuable.

Details arisen during the process

During the work I've learned many web sites to find an appropriate one. Some sites were unsuitable because there was a few information. A large amount of data and parameters will allow the model to have higher accuracy. There were very few statistics on some sites, for example, only the number of matches played, goals scored/conceded, yellow and red cards received by each team. In addition, most sites summarized data for the entire season, without providing extended statistics for each game for a deeper analysis. At the end of research of suitable web sites, I have found the one that contains a lot of necessary data. There are many parameteres, such as : xG, the percentage of possession of the ball, the percentage of accuracy of passes, the number of shots and shots on target, saves, face-offs, number of selections. In addition, the site includes information about the stadium where the match was held, attendance, weather conditions and temperature. Most importantly, all these data were presented for each individual match of the season. A detailed review of each game made it possible to make the prediction more accurate.

2 Already known methods

Monte-Carlo method

Main idea of this method is multiple stimulation of random processes to calculate the probability of occurrence of the desired event. It is based on the principle of the possible repetition of the situation an unlimited number of times in a row. Certain conditions are created for the experiment, which do not change during the analysis. However, this technique can fail when there is need to evaluate risks, because it will be hard to collect a lot of models of an unlikely event as it will not have many repetitions in the past. As in any other field, the method works the same way in betting – you need to calculate the interaction of many outcomes that depend on incoming factors. It is enough to create different models of situations with a different set of factors affecting the result – the output will be an estimate of the probability of outcomes. There are some criteria that can be used to predict the result using Monte-Carlo method, such as : recent match statistics, the level of players (overall transfer value of the team), current place in this competition or number of points, the average performance of the team, injuries and disqualifications of players, the venue of the match (which team is playing at home, which is away)

To give an illustrative example of modeling according to these criteria we can consider match between FC Barcelona and FC Napoli that will take place on February 21 in the Champions League. It will take place in Spain, so FC Barcelona will be the host team. Here is the table that tries to predict the outcome of this event:

Table 1: Preducting outcome of the event

factor	lvl of importance	add.coeff	Team A	Team A with add.coeff	Team B	Team B with add.coeff
number of wins	3	1	4	$3 \cdot 4 = 12$	3	$3 \cdot 3 = 9$
avg scored goals	3	1	$12 \div 6 = 2$	$3 \cdot 2 = 6$	$10 \div 6 = 1.67$	$3 \cdot 1.67 = 5$
avg lost goals	2	2	$-6 \div 6 = -1$	$2 \cdot -1 = -2$	$-9 \div 6 = -1.5$	$-1.5 \cdot 2 = -3$
home/away	1	3	2	6	0	0
injuries	1	3	-6	-18	-1	-3
transfer price	2	2	861	$2 \cdot 861 = 1722$	535.15	$2 \cdot 535.15 = 1070.3$
Finally			862	1726	537.32	1078.3

In this table statistics was taken from matches in the Champions League in this season. This particular match is 1/8 of the final, so ahead of this game there were 6 games of Group Stage. To make it more clear, lost goals are taken with negative value because it worse influence on the possible success of the team. Injuries are also taken with minus as they do not help to win.

Thus, on order to trust this methodology we can conclude that FC Barcelona will be the winner of this match.

Of course, this tehnique of predicting cannot give a 100-percent confidence that the result will suit the prediction. Monte-Carlo method has also disadvantages because all factors cannot be taken into consideration. Besides, this evaluation can be subjective.

This table with prediction was created before the match between FC Barcelona and Napoli. Now, the result is known. This match ended with a result 1-1, which shows us that this method doe not give a 100-percent confidence in correct result

Fork Betting

This strategy allows to benefit from any outcome of the event. A fork is an opportunity to place bets on opposite outcomes of the same sporting event with guaranteed profit. Noticing a fork from bookmakers is not such an easy task. They are fickle and do not appear for long, so in order to catch them, you need to carefully monitor the behavior of event coefficients at different offices. Forks appear when the bookmaker's odds diverge from the market, that is, the rest of the betting shops. There are some reasons why do they appear, for example, the bookmaker overestimates the odds to attract customers or the coefficients change with different delays at different bookmakers. Here is an example of how does this strategy work : suppose we want to make a bet for match between Bayern Munchen and Bayer Leverkusen. Bookmaker A gives 2.10 for Bayern to win. Bookmaker B offers to bet on a draw for 4.10, and on the success of Bayer for 4.60. It's a fork. If you bet 1000 rubles on Bayern Munich, 500 rubles on a draw and 450 rubles on Bayer, the total amount of bets will be 1950 rubles. Potential payout in case of success of Bayern Munich: 1000 x 2.10 = 2100 rubles. In case of a draw: 500 x 4.10 = 2050 rubles. If Bayer wins: 450 x 4.60 = 2070 rubles. Accordingly, the profit will be received at any outcome of the match.

Some recent works

The research [1] showcased the use of interesting features such as the weather conditions, psychological state of players, and whether or not any of the main players were injured. The work reported very high accuracy results, with a mean accuracy of 92 percent, but the modelled system was for only one team and a single season, involving just 20 matches for the team under observation.

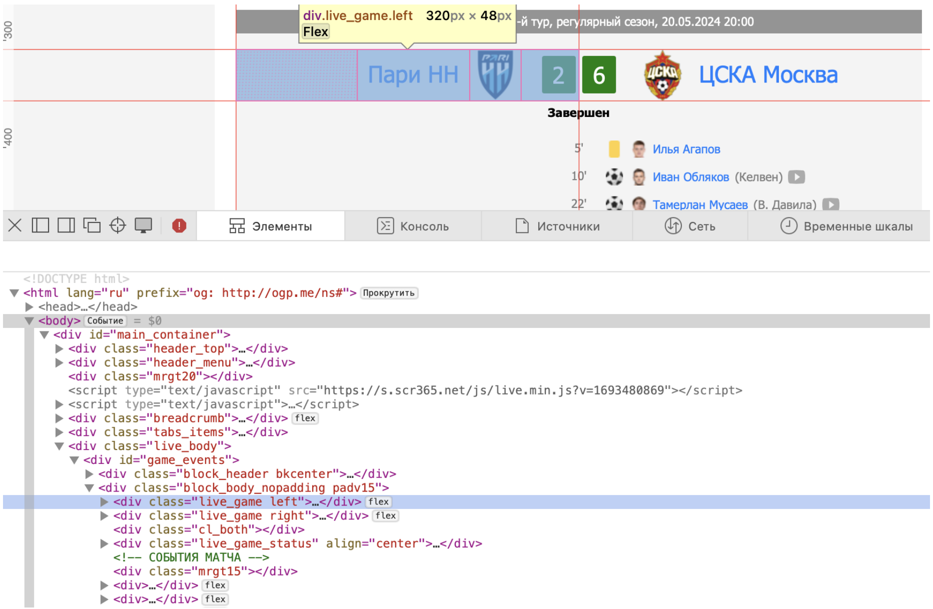
Relative importance of performance features and Correlation between performance features and success that are illustrated in Quantifying the relation between performance and success in soccer. The work [2] explored different features that can determine the success of the team, building illustrative graphs and showing the connections between results and factors.

Predicting Football Match Outcomes With Machine Learning Approaches where the following was assessed [3] : the performance of random forest, logistic regression, linear support vector classifier and extreme gradient boosting models for binary and multiclass classification. These models are trained with datasets obtained using different sampling techniques. Besides, method of random forest is also deeply researched in [4], where forecasting work was also carried out with the creation of a model based on a training sample and the implementation of a prediction model. One of the metrcis that can be used for prediction in different classificators is ROC AUC that is carefully described in [5]. An example of website citing РПЛ [6] The site with which the direct work was carried out, from where all the data for research was taken[7].

3 Data preparation and processing

Description of my work

The first step after choosing the website was parsing it. I needed to parse the site in order to pull out data on every game of the 23/24 season in the Russian Premier League. By studying the code of the page of one match, I managed to understand how the site works. This helped to write the functions needed to work with this site and create a data frame



```
<!DOCTYPE html>
▼ <html lang="ru" prefix="og: http://ogp.me/ns#"> Прокрутить
  ▶ <head>…</head>
  ▼ <body> Событие  = $0
    ▼ <div id="main_container">
      ▶ <div class="header_top">…</div>
      ▶ <div class="header_menu">…</div>
        <div class="mrgt20"></div>
        <script type="text/javascript" src="https://s.scr365.net/js/live.min.js?v=1693480869"></script>
      ▶ <script type="text/javascript">…</script>
      ▶ <div class="breadcrumb">…</div> flex
      ▶ <div class="tabs_items">…</div>
      ▼ <div class="live_body">
        ▼ <div id="game_events">
          ▶ <div class="block_header bkcenter">…</div>
          ▼ <div class="block_body_nopadding padv15">
            ▶ <div class="live_game left">…</div> flex
            ▶ <div class="live_game right">…</div> flex
              <div class="cl_both"></div>
            ▶ <div class="live_game_status" align="center">…</div>
              <!-- СОБЫТИЯ МАТЧА -->
              <div class="mrgt15"></div>
            ▶ <div>…</div> flex
            ▶ <div>…</div> flex
```

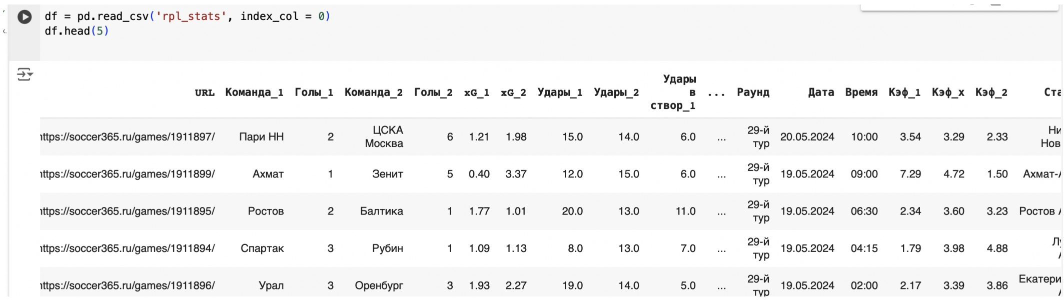
Figure 1: Work with web site

This investigation helps to write the function def parse match stats(url) to get data I've got team names, all stats, bookmakers' odds, stadium names, weather conditions.

After that I have a function to extract links on the matches. We need them to create a data frame.

Now I am able to get a data frame for future work. I have all links on each particular game with all stats and necessary data in array 'links'.

Iterating over each item from the links array, we call the parse_match_stats function and save it to df_dict. Now we have a ready-made Data Frame to work with.



```
df = pd.read_csv('rpl_stats', index_col = 0)
df.head(5)
```

URL	Команда_1	Голы_1	Команда_2	Голы_2	xG_1	xG_2	Удары_1	Удары_2	Удары в створ_1	...	Раунд	Дата	Время	Кэф_1	Кэф_x	Кэф_2	Ста
ttps://soccer365.ru/games/1911897/	Пари НН	2	ЦСКА Москва	6	1.21	1.98	15.0	14.0	6.0	...	29-й тур	20.05.2024	10:00	3.54	3.29	2.33	Ни Нов
ttps://soccer365.ru/games/1911899/	Ахмат	1	Зенит	5	0.40	3.37	12.0	15.0	6.0	...	29-й тур	19.05.2024	09:00	7.29	4.72	1.50	Ахмат-
ttps://soccer365.ru/games/1911895/	Ростов	2	Балтика	1	1.77	1.01	20.0	13.0	11.0	...	29-й тур	19.05.2024	06:30	2.34	3.60	3.23	Ростов
ttps://soccer365.ru/games/1911894/	Спартак	3	Рубин	1	1.09	1.13	8.0	13.0	7.0	...	29-й тур	19.05.2024	04:15	1.79	3.98	4.88	Л
ttps://soccer365.ru/games/1911896/	Урал	3	Оренбург	3	1.93	2.27	19.0	14.0	5.0	...	29-й тур	19.05.2024	02:00	2.17	3.39	3.86	Екатери

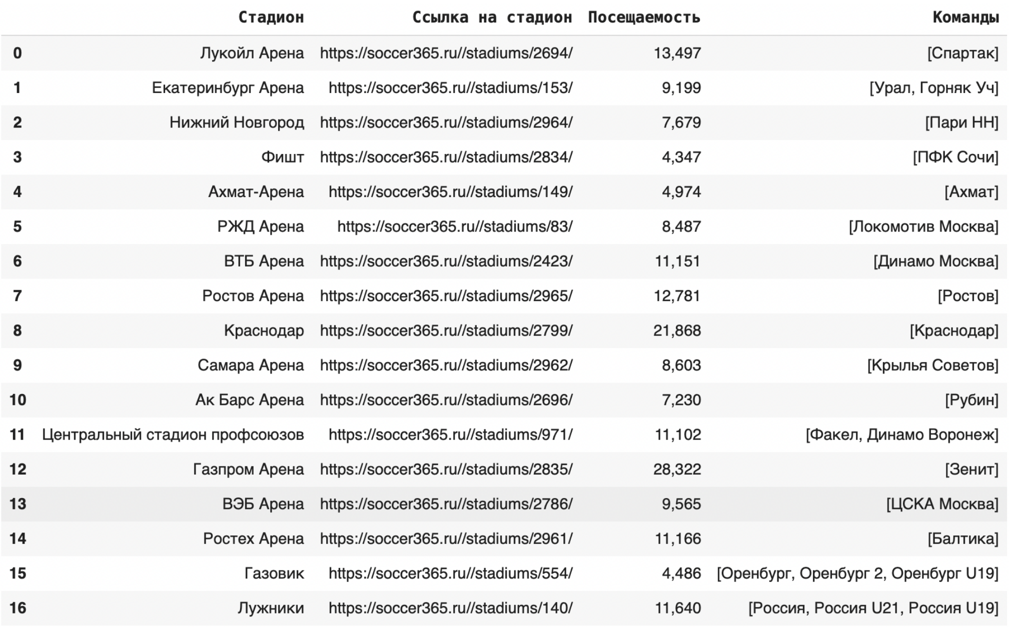
Figure 2: Data frame

To make further work with the resulting table easier and more convenient, it is necessary to analyze the data obtained. This will help us understand what kind of data we have, and whether we can work with it.
First of all, the column with URL was necessary only for debugging, so we can delete it. Secondly, the format of the 'Раунд' column entry is also not very convenient, since it is recorded not just by the number of the round, but also in each match with the word 'тур'. It will be inconvenient for us to work with such a data format, so we will leave only the numeric value.

Besides, the function info() helped to find out that column 'Дата' is a string. For convinient work, I needed to tranform in into pandas datetime. Now, it contains hours and minutes. The temperature is also in an inconvenient format. By printing out the array, it is easy to see that all the degree data is positive . Therefore, I decided to remove the Celsius and plus icon for convenience.

It seems that the information about which stadium we have is not very useful in the form in which it is. It is clear from intuitive assumptions that if a team plays at home, then it is on its own territory, therefore it plays better. Therefore, from our stadium attribute, we need to learn to understand whether it is playing at home or away. At the same time, there are non-native stadiums of the teams, we will leave them for now as guest stadiums for both teams. I need to parse the stadiums page, go to each stadium and get two dictionaries, for example, with average attendance and a list of teams for which the stadium is native.

Now it is possible to clearly understand which stadiums for which teams are home.Besides, I have decided to add column with values 1/0 (True/False) to understand which of the teams played at home.
Unfortunately, not all matches had registered attendance. Therefore, I decided to calculate the base attendance and increase it by a randomly generated percentage in the range of 10-15 per cent so that this column would be filled in front of each match.



	Стадион	Ссылка на стадион	Посещаемость	Команды
0	Лукойл Арена	https://soccer365.ru//stadiums/2694/	13,497	[Спартак]
1	Екатеринбург Арена	https://soccer365.ru//stadiums/153/	9,199	[Урал, Горняк Уч]
2	Нижний Новгород	https://soccer365.ru//stadiums/2964/	7,679	[Пари НН]
3	Фишт	https://soccer365.ru//stadiums/2834/	4,347	[ПФК Сочи]
4	Ахмат-Арена	https://soccer365.ru//stadiums/149/	4,974	[Ахмат]
5	РЖД Арена	https://soccer365.ru//stadiums/83/	8,487	[Локомотив Москва]
6	ВТБ Арена	https://soccer365.ru//stadiums/2423/	11,151	[Динамо Москва]
7	Ростов Арена	https://soccer365.ru//stadiums/2965/	12,781	[Ростов]
8	Краснодар	https://soccer365.ru//stadiums/2799/	21,868	[Краснодар]
9	Самара Арена	https://soccer365.ru//stadiums/2962/	8,603	[Крылья Советов]
10	Ак Барс Арена	https://soccer365.ru//stadiums/2696/	7,230	[Рубин]
11	Центральный стадион профсоюзов	https://soccer365.ru//stadiums/971/	11,102	[Факел, Динамо Воронеж]
12	Газпром Арена	https://soccer365.ru//stadiums/2835/	28,322	[Зенит]
13	ВЭБ Арена	https://soccer365.ru//stadiums/2786/	9,565	[ЦСКА Москва]
14	Ростех Арена	https://soccer365.ru//stadiums/2961/	11,166	[Балтика]
15	Газовик	https://soccer365.ru//stadiums/554/	4,486	[Оренбург, Оренбург 2, Оренбург U19]
16	Лужники	https://soccer365.ru//stadiums/140/	11,640	[Россия, Россия U21, Россия U19]

Figure 3: Updated stadiums' data frame

The final step in data preparation is to fill in the missing fields 'Навесы', 'Отборы', 'Передачи', 'xG', 'Точность передач'. To fill in the missing fields, I decided to create a tricky function.Since I was dealing with quantitative signs, it is possible to calculate for each team the average of the missing parameter for the matches that have already passed, where there was data. That is, for example, suppose Zenit has no data on xG in the match with Dynamo. Then, our function will take the average xG for all matches already played by Zenit and generate the missing number. For future work, I have added a column 'Таргет' which shows the result of the match. It outputs '0' if the result is draw, '1' if team 1 won and '2' if team 2 won. Moreover, I have applied one-hot encoding to to convert categorical variables in the 'Погода' column into numerical dummy variables. Also, I extracted the year, month, and day components from the 'Дата' column using the .dt.year, .dt.month, and .dt.day methods, respectively. Preparation of the data for future work now has been finished.

4 Detailed description of chosen methods

Logistic Regression

At first, I decided to try logistic regression. To do this, I encoded the signs by commands, divided them into test and training samples and mixed everything up a little. Also, to improve the convergence and quality of training, it is best to normalize the data with different range, so I have used a standard scaler.

Random Forest

After logistic regression I have decided to implement Random Forest Classifier. A random forest combines many decision trees, training each on a separate data sample, dividing nodes in each tree using a limited set of parameters. The final forecast is made by averaging the forecasts from all the trees.The approach in which each learning element receives its own set of training data (using bootstrapping), after which the result is averaged, is called bagging. Another interesting point, to speed up the work of a random forest, can be used in the parameters n_jobs = -1, in which case the construction takes place on the maximum possible number of processors. In addition, the number of trees on which the learning will take place also can be chosen.

LSTM
The main idea of LSTM is the ability to remember information over long periods of time, which allows efficient processing of sequential data such as time series, for example. This model is recurrent, so it is based on what it has learned in the past. This makes it suitable for my project as it is aimed to forecasting the results based on previous games and past statistics.

ROC (Receiver Operating Characteristic) for Multi-class We have multiclass data. Extending ROC curves in the case of classification problems with more than two classes has always been difficult, since the number of degrees of freedom increases quadratically with the number of classes, and the ROC space has c(c-1)dimensions, where c is the number of classes. I have decided construct ROC AUC for all of my three classes from 'Таргет'.

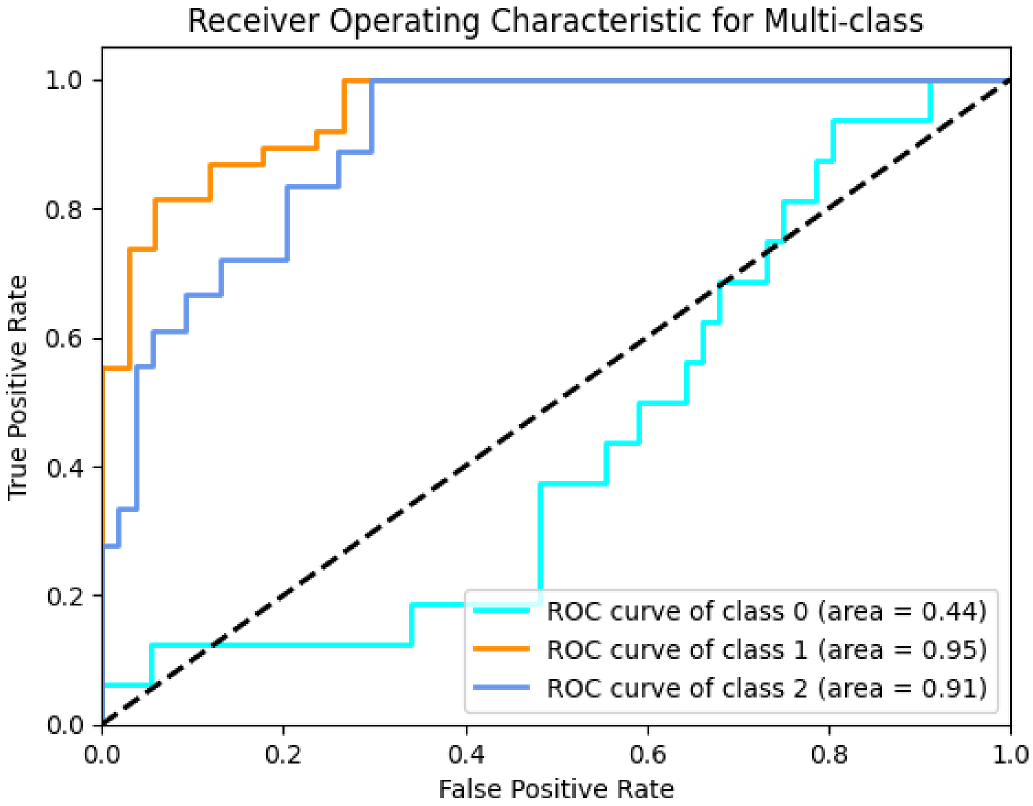
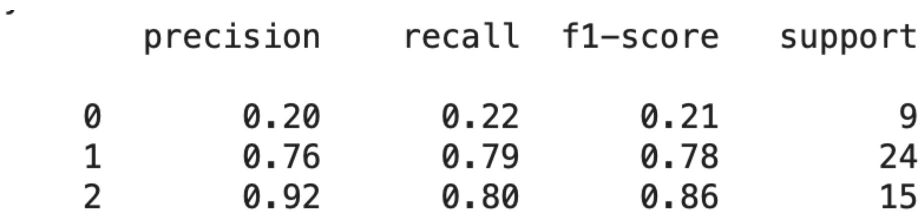


Figure 4: ROC AUC

This graph helps to get one intresting derivation. Obviously, the area under ROC curve of class 0 is much smaller than under other 2 curves. It says that draws in our data were predicted in the worst way. In addition, considering the bookmakers' quotes for matches, we can also say that the odds for a draw are usually higher. than for the victory of any of the teams. This indicates that a draw is the least likely outcome of the match.

To make sure that the draw is really predicted the worst, we can build a confusion matrix. This matrix will be able to visually show the accuracy for each of our three classes.



```
              precision    recall  f1-score   support

           0       0.20      0.22      0.21         9
           1       0.76      0.79      0.78        24
           2       0.92      0.80      0.86        15
```

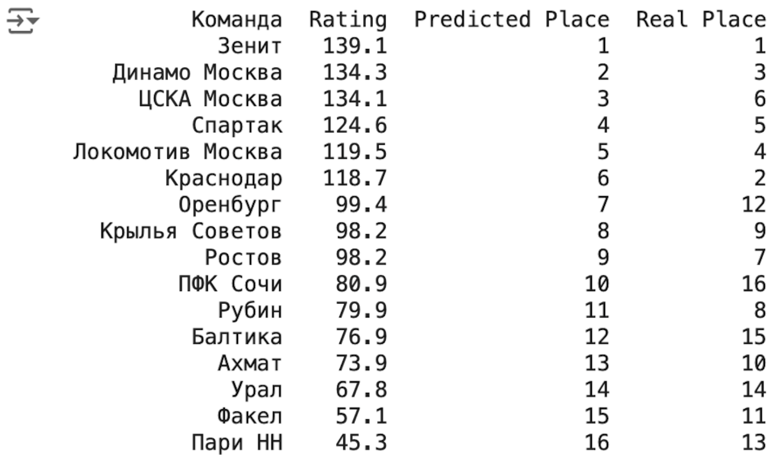
Figure 5: Confusion matrix

This matrix also illustrates that prediction of draws does not very well. Perhaps this can happen because there are few teams that will be equal in all respects. Almost always, one of the teams has some advantage, which makes predicting draws difficult.

5 Creating of a rating

From the start of working with project I was thinking about creating the rating. Here, in my data frames, I have a lot of data to build scores. At first, I create a table with each team from table with matches. Now I have summarized data within the season. One important point is that some data had to be presented in a suitable format, as they were expressed as a percentage. I took the average character for all the matches of the team by column 'Точность передач', 'Владение мячом'.

Now, using the selected coefficients based on their importance to the team, the score for each team was compiled. Creating of this rating was inspired by Monte-Carlo method that was useful for prediction by many parameters.



```
          Команда  Rating  Predicted Place  Real Place
            Зенит   139.1                1           1
    Динамо Москва   134.3                2           3
      ЦСКА Москва   134.1                3           6
          Спартак   124.6                4           5
Локомотив Москва    119.5                5           4
        Краснодар   118.7                6           2
         Оренбург    99.4                7          12
  Крылья Советов     98.2                8           9
           Ростов    98.2                9           7
         ПФК Сочи    80.9               10          16
            Рубин    79.9               11           8
          Балтика    76.9               12          15
            Ахмат    73.9               13          10
             Урал    67.8               14          14
            Факел    57.1               15          11
          Пари НН    45.3               16          13
```

Figure 6: Obtained scoring

As it can be seen from the picture, the rating made by yourself correctly determined the winner of the tournament. However, further there are discrepancies between the real places of the teams and those that were calculated by the rating. Some discrepancies are quite small, for example, Dynamo, Spartak, Lokomotiv and Krylia Sovetov actually have almost the same place as predicted by the rating.

However, there are also teams whose places in real life are very different from those determined by the rating. For example, Orenburg took 12th place in the championship, but the rating showed that it is possible to evaluate the performance of Orenburg according to the parameters that we have and with the coefficients we selected, the team should have taken 7th place.

Moreover, judging by the rating, the PFC Sochi team should have been in 10th place in the final table, but in real life the team took the last 16th place and was eliminated from the top division of Russia.

Thus, if the team performed better than the model predicted, then most likely this is due to its over perfomances. For example, there were several very strong players in this team, helping out the team despite its poor performances. In contrast, if the team performed worse than the rating model predicted, this may be due to the fact that despite the fact that it was good in a number of parameters, its rivals were much stronger, so it was not possible to achieve success.

6 Results

After conducting a study of two metrics: accuracy and f1-score, I got the following results:

	accuracy	f1-score
Logistic Regression	0.71	0.70
Random Forest	0.67	0.66
LSTM	0.62	0.71

Table 2: Results

So, now it can be seen, that the most accurate way to predict is logistic regression with metrics 'accuracy'. The values for the two metrics of both logistic regression and random forest are almost equal to each other. Besides, it seems that LTSM has got the largest variation in the accuracy of predictions across two metrics.

7 Future plans

There are many more plans that can be implemented for this project in order to expand it and make predictions more accurate. To begin with, we can see how our parameters correlate directly with the number of points scored in order to identify the most influential factor. Then, depending on the results obtained, you can change the coefficients in the rating model to make it more accurate. In addition, an important factor in the success of each team is the personal performances of its players. Therefore, for more accurate predictions, it would be good to look at the statistics of the players. For instance, it is possible to count the total statistics of each team, looking at the statistics on goals + passes for attacking players, on dry matches for defensive players.

As for the players, it's also important to look at injuries as the season progresses. Injuries to valuable players usually significantly affect the team's results, because it is difficult to replace any of the leaders.

In addition, it would be interesting to consider not one season, but several at once, for example, the last 3 seasons, in order to identify the dynamics of the teams' performances, to identify the most stable / unstable.

8 Conclusions

Summarizing the whole process of work, at first a convenient website with a lot of information was found. After that, the site is parsed with all the necessary data. Then there was a long process of data processing, filling in empty cells, bringing all the data to a convenient format for work. Then the process of working with training models began: logistic regression, random forest and LSTM. Next, we created our own rating, somewhat reminiscent of the Monte Carlo method, to put the teams in place relative to the parameters found on the site.

References

[1] P. E. Farzin Owramipur and F. S. Mozneb, "Football result prediction with bayesian network in spanish league-barcelona team," *nternational Journal of Computer Theory and Engineering*, vol. 5, no. 5, pp. 812–815, 2013.

[2] L. Pappalardo and P. Cintia, "Quantifying the relation between performance and success in soccer," *Advances in Complex Systems*, vol. 21, no. 03n04, p. 1750014, 2018.

[3] L. K. F. Bing Shen Choi and S.-L. Chua, "Predicting football match outcomes with machine learning approache," *Mendel Soft Computing Journal*, vol. 29, no. 2, pp. 229–236, 2023.

[4] K. S.B. and K. V.M., "A classification algorithm based on the principles of a random forest to solve the forecasting problem," *Software products and systems*, vol. 2, pp. 11–15, 2016.

[5] K. A.G., S. A.I., and B. D.A., "Study of the most efficient models and attribution algorithms usnig the roc auc indicator," *Modern High-Tech Technologies*, vol. 7, pp. 63–68, 2022.

[6] RPL, "Official website of the russian premier league." https://premierliga.ru, 2024. Last accessed 3 February 2024.

[7] RPL, "Website with game results." https://soccer365.ru/competitions/13/results/, 2024. Last accessed 27 May 2024.