# Predicting the Matches Entertainment on the Example of the Russian Premier League Using Machine Learning Methods and Predictions Application for Sports Broadcasts Organization

Sergey Gorshkov [1,2][0000-0001-5958-5224], Anastasia Chernysheva [1,3][0000-0003-0812-8941], and Ilya Ivanov [1,3][0000-0003-4205-590X]

[1] Lomonosov Moscow State University, Leninskie gory, 1, GSP-1, Moscow, 119991, Russia
nastya.ch.9797@mail.ru
[2] National Research University Higher School of Economics,
20 Myasnitskaya str, Moscow, 101000, Russia
serggorsar@yandex.ru
[3] Skolkovo Institute of Science and Technology,
Bolshoy Boulevard 30, bld. 1, Skolkovo, 121205, Russia
ilya01r@gmail.com

**Abstract.** The calendar of the football season of the Russian Premier League, as well as other leading European championships, is designed in such a way that several matches can take place in one period of time. This situation has become more common after the pause associated with the COVID-19 pandemic, due to the match schedule densification. The broadcaster needs to determine which match will be live shown on the main channel. Also, he can provide access to all broadcasts of the championship round for paid channels and recommend the most spectacular matches for viewers to watch. The start time of the matches in the modern world is chosen by agreement of the League and television, so it would be really convenient to put potentially the most spectacular matches on prime time. This paper introduces the concept of the entertainment index, which takes into account goals and other important events of the match. The value of the entertainment index for upcoming matches is predicted using a machine learning model based on historical data. As the result, we have a model that can predict the entertainment of a match and help you to choose the most interesting game from the viewer's side.

**Keywords:** football; Premier League; machine learning; CatBoost; matches entertainment; mathematical modeling; linear regression.

## 1 Introduction

Match entertainment is a very subjective concept. Different people like different football. Someone prefers the abundance of goals and dangerous moments, someone looks at tactical coaching steps and football "chess", someone savors a double-edged

Proceedings of the 10th International Scientific and Practical Conference named after A. I. Kitov "Information Technologies and Mathematical Methods in Economics and Management (IT&MM-2020)", October 15-16, 2020, Moscow, Russia

game with a lot of action in the center third of the field. Modern TV and Internet broadcasters are ready to offer viewers a wide range of football matches both the national championship and international competitions to watch. In many European Championships, the matches start time is chosen from certain time slots agreed with the main partner broadcasters and is based on regional specifics, participating teams' games schedule, and the broadcasters' interests. Various metrics are used to measure viewers' interest: the number of views, ratings, retention, and so on. However, these are indirect indicators that are not directly related to football, but only evaluate the reaction of viewers. Usually, are selected to be shown at the most convenient time and in the best format such as top matches, which have high-status opponents with a large fan audience is playing, and the match will be watched by the largest number of people. However, there are not many such matches, and among the others, you must also select the most and the least priority matches to display. In addition, many online cinemas and Internet TV have recommendation systems that advise you to watch the most interesting matches that the audience should like.

The second problem is related to the fact that often several matches fall on the same time slot or on overlapping time slots, and TV broadcasters need to solve the problem of showing these matches to TV viewers. Different strategies can be chosen for showing matches that take place simultaneously, for example, live broadcasting of different matches on different media holding channels, live broadcasting of individual matches on a Federal channel etc. In Russia, during the period after the pause associated with the coronavirus pandemic, matches of the Russian football championship are live shown in large numbers on the public TV channel MATCH TV. Due to the high density of the schedule, many games are held simultaneously, and therefore the TV channel must choose which match to show live on the Federal channel.

Of course, this choice is based on a lot of factors. Firstly, it is the current ranking of the teams, the competitive value of the match, principled opposition to rivals, the current form of opponents, and so on. Secondly, it is the size of the estimated match audience, which correlates with the total number of fans of a particular Premier League team. Thirdly, these are considerations related to the fact that there are no teams whose matches are not shown at all or are rarely shown on a public channel. Of course, there are matches of the leading teams of the championship, derbies, which are likely to cause the greatest audience interest, including for neutral fans. Match selection among the remaining ones becomes a difficult task.

In this study, we make a reasonable assumption that viewers are more interested in watching spectacular football, all other things being equal. Therefore, to recommend a match on the Internet or show in live time, it is suggested to learn how to predict the entertainment of a match based on the team statistics and related match's factors. Currently, there are no solutions for similar problems in the scientific literature, so in section 2 we provide an overview of related areas, where we will look in detail at existing approaches to predicting various factors, both related to the game and to the match organization, including evaluating entertainment based on other factors. Section 3 introduces the match entertainment index and describes the developed algorithm that allows predicting the entertainment of a Premier League match with acceptable accuracy based on historical data for opposing teams. Section 4 will describe the testing of the algorithm and a description of the data sets. Section 5 discusses the results of the algorithm and its correctness. Section 6 provides a summary of the done work.

## 2 Literature Review

Most sports-related research aims to predict the outcome of the matches or some match statistics. This is easily explained from the commercial side of the issue – large incomes in the field of sports for various individuals and legal entities are associated with betting. Of course, only a small part of the research is public. The most common thing that researchers try to predict in their work is the outcome of a match, based on many different characteristics [1 - 7]. In [4] are used the ELO rating (it is also used in FIDE chess rankings and other sports kinds [8]). It can be concluded from the article that statistical loss functions are more effective than economic measures when using various forecasting methods. ELO ratings are useful for encoding past results information. Applicable to football the difference in ranking is very significant in predicting the match result. The authors concluded that using the ELO rating to assess the strength of the team is a justified step. In [6], a Bayesian network is considered for predicting the results of football matches involving FC Barcelona. Many factors that influence the outcome of a football match were identified. Thus, the process of selecting the model and features to explore sets the boundaries which can be discovered. The authors of the article group factors into two types: non-psychological and psychological. Table 1 shows the main factors for predicting football matches that the authors used in Bayesian network.

**Table 1.** The most important factors for predicting the result of football matches in [6].

| Psychological | Non Psychological |
|---|---|
| Weather | Average of players age |
| History of 5 last games | Injured main players |
| Result against for teams | Average match in week |
| Home game | Performance of main players |
| Ability front team | Performance of all players |
| Psychological state | Average goal in all home |
| | Average goal for home |

The accuracy of predicting the outcome of the match was 92%. This is a very high result for such type of research, however, it can be explained by the fact that only one team is considered, which is much more likely to win than lose points. A wide range of approaches and machine learning algorithms are considered in [5, 7, 9, 10]. For example, the authors of [10] conduct experimental studies with the following models: Naïve Bayes, Bayesian networks, LogitBoost, The k-nearest neighbors algorithm, Random forest. However, none of the above models usually achieves more than 60% accuracy. Thus, based on this and other articles, we can conclude that the accuracy of close to 60% is quite acceptable since the football game is a rather non-deterministic process with a lot of random events, such as red cards and penalties, so it is not possible to predict the result with high accuracy.

There are several articles in the scientific literature regarding the evaluation of the entertainment of a match. In [11], a linear regression model with GLM encoding was chosen to predict the number of viewers based on the social network activity (Facebook) during the match, as well as activity in the last two weeks before the match. Besides that, according to the authors, club or the national team photos on the social page cause the greatest activity of the audience and fuel interest in the match. The authors of [12] come to non-trivial conclusions about the difference between the entertainment of the match for the stadium and TV audience. Using empirical methods, it is argued that the viewers attendant the stadium prefer to look at a crushing score instead of competitive confrontation. While viewers watching the broadcast on TV, on the contrary, prefer matches with an equal score and a close fight. On the Internet, various sports sites [13-16] offer methods for evaluating the entertainment of a match, but no attempts are made to its prediction.

Despite different approaches, we can identify criteria that allow us to identify a match as more attractive to the viewer. These criteria allow us to form an opinion about the game entertainment, based on a system for evaluating interconnected indications. So, on the popular Russian sports portal Sports.ru, the index is calculated by analyzing the following parameters: shots, goals, hot passes, strokes, fast attacks, and constructive passes. Each parameter is assigned a coefficient that depends on the rarity of the event. Federal Russian sports channel Match TV calculates the entertainment of the match takes into account several other factors: the accuracy of passes, the number of accurate passes over a certain period, the complexity of the combination that led to the goal, the number of shots on target relatively to all shots on goal, and of course, the average number of goals.

In addition, there are many studies aimed at predicting match attendance [17-19]. As well as articles analyzing TV views and thus suggesting the number of views of the upcoming match and the possibilities of their commercialization [20, 21]. In [17], the authors study the audience's interest in the Italian top division and conclude that the greatest interest for viewers is not even the fight for the title, since for objective reasons it is claimed by a small number of teams, but the fight for a place in European competitions. In [20], the authors investigate the relationship between the popularity of the world Cup and the national team of the country hosting the championship participation. So, in their observations, they describe the relationship between the views reducing the championship and of the host country’s national team outflow from the championship. In addition, they come to the fairly obvious conclusion that holding a match on the weekend greatly increases its popularity among fans. In the field of the organization of football matches in Russia involves leading universities, for example, HSE has helped the Russian Football Union with the establishment of the match calendar depending on team participation in the UEFA Champions League games and peculiarities of the Russian climate. It is also planned to create a calendar for referees [22] in the near future.

## 3 Proposed Method

In this article, we present *EntertainmentIndex* – an index that will be calculated based on various statistical data, and its prediction method. In section 3.1, we describe the principles of feature formation and the feature-object matrix, as well as the principles of calculating the entertainment index. Section 3.2. describes the methodology for testing the hypothesis that there is a relationship between the entertainment indicator and the obtained features. Section 3.3 covers the process of machine learning method choice for solving the regression problem, and finally, in section 3.4, we describe the process of training the model directly.

### 3.1 Feature preparation

We will use the basic statistical characteristics of the match as model features: the number of scored and conceded goals, the number of shots on goal and on target, the number of beats off goalkeeper's goals (saves), number of corners, offsides, violations, warnings, penalties, percentage of ball's possession and correct passes. In addition, we will take into account some of the indicators that are responsible for the course of the game: a missed penalty, the goal after 85 minutes, the missed decisive penalty (which could change the score), the decisive goal in added time, strong-willed victory (victory of the team that had a lower score during the match), strong-willed draw (during which one of the teams brought back two goals). We also take into account the match attendance and the team's place in the standings.

The entertainment index will be a weighted sum of some groups of match events. Let us describe the main components that will be included in this linear combination with the corresponding weights:

1. The number of scored goals. More goals lead to more interest in the match. However, if the game is on "one goal", then the match is less interesting than if the game did not have one-team domination. This component will take into account the total number of goals scored by both teams, minus the number of goals scored by the winning team with a goal difference of more than four. This is the basic component, the total number of goals scored is included in a linear combination with a coefficient of 1, i.e. 1 point for each goal.
2. The number of shots on goal. Not all spectacular matches are replete with a large number of goals, the interest adds a large number of shots on goal and on target, in particular. These events are less important than goals, so we believe that every 10 shots over 5 shots for 1 point, an additional bonus – 1 point for every 10 shots on target. Since goals almost always follow after shots on goal, it turns out that we take goals into account in the metric three times, which increases their significance for the metric.
3. The number of saves by the goalkeepers. Spectacular saves increase entertainment, so we will count every 7 saves over 5 for 1 point in addition to entertainment.
4. We add 1 point to the metric for each unrealized penalty, a goal after 85 minutes, a goal after 90 minutes that affected the outcome of the match, and

2 points for an unrealized penalty after 90 minutes that would have changed the outcome of the match.

5. Add 1 point for each ball played in a strong-willed victory (the victory of a team that was behind in the score for a certain period of the match) by one of the teams and 1 point for each ball played over one in a draw if the team lost two goals or more during the match.

Note that all the indicators used to build the entertainment index can be calculated for a large number of tournaments since it uses basic statistics, which makes the solution extensible. We can also calculate these indicators for several past rounds to further use them to predict the index value for individual matches in the future.

The entertainment index must be calculated for the match, so, we calculate it as the sum of the indices for two opponents, based on the description above. From the feature description of the match for each team, we will create a new matrix of feature objects as follows. Let's select the three rounds preceding the match and average all the indicators for the three rounds. Thus, we get for each match: historical data for features for each of the teams and the entertainment of the match, which we need to predict. We use the assumption that during the last rounds the team plays more or less evenly. It makes no sense to take less than three matches because football has a high uncertainty in all indicators and averaging smooths out the variance. If we take a longer period, the team's game may already change. We will add additional information to the match description that is known along with historical statistics before the game starts – the name of the chief referee, the stadium, and the start time of the match.

Just note, why information about the players, the various advanced statistics, etc. is not added. The fact is that we consider predictions within a single season, because teams change too much over a year, and therefore the sample consists of a small number of objects. In this regard, it is necessary to control the number of features so there weren't too many of them.

### 3.2 Testing the hypothesis about the relationship between the target variable (entertainment) and features

Up to this point, we have conducted research based on the assumption that there is some relationship between the previously entered *EntertainmentIndex* and the feature description based on historical data. Let's check this assumption using the statistical test engine. Because we want to test the dependence of a real-valued target variable on a features-objects matrix (and there are more than two of these features), we cannot use standard approaches, such as calculating the Pearson or Spearman correlation, as well as the Fisher exact test and other frequently encountered parametric and non-parametric statistical tests.

The main problem is that we have many features, so it is impossible to evaluate their contribution separately. This is because when the feature set is expanded, the significance of individual features may significantly change. Besides, the target variable may not depend on some features individually, but on a specific group of features. Linear regression without regularization is used to solve a problem with similar constraints. Let’s find the solution by the least-squares method. Define the ordered number of the feature as $i$. Thus, our target variable is the entertainment coefficient $e$, and $i$-th feature $x_i$. Then the linear regression problem will be presented as follows: it is necessary to find the vector $\alpha$, such as $e \approx \alpha x$. Zeroing coefficients for all variables except one $x_i$., we can consider the effect of the target variables on the $i$-th feature. At the same time, we can consider various features combinations. Of course, the machine learning model that will be used to predict the target variable is non-linear, but a linear model is sufficient for testing.

To test the hypothesis about $\alpha_i = 0$ (null hypothesis), we will use the Student's T-test. If the null hypothesis holds, then the null distribution will be the Student's distribution. In Python, the necessary functionality is implemented as the ols (ordinary least squares) function in the statsmodels.formula.api [23] module.

### 3.3 About CatBoost

There are many machine learning methods available for solving the regression problem, ranging from classical methods such as linear regression to deep learning. When choosing a method for training, it is necessary to start, first of all, from the nature of the data and their characteristics. The main point to consider in our task is the small number of observations in the sample. Therefore, it is quite problematic to use deep neural networks due to the high probability of their overfitting. Gradient boosting algorithms demonstrate good quality for solving such problems. There are many different implementations of this idea, such as XGBoost [24], LightGBM [25], and H2O [26], which have their strengths. In the current study, we will use CatBoost (for "categorical boosting") – an open-sourced gradient boosting library [27, 28]. The advantage of this method is the processing of categorical features that are present in our feature space (names of teams, stadium, name of the main referee of the meeting). Standard approaches for processing categorical features – label encoding, one hot encoding [29, 30] – in this task will either reduce the information content of features or create many times more features than it was originally. CatBoost internally processes categorical features in some way, and in addition to this, it handles data gaps well, which are quite a lot in our task. Among the advantages of CatBoost, it is also worth noting that in addition to the features themselves, some combinations of features are also considered. When constructing a split for the current tree, CatBoost greedily examines various combinations, including combinations of numeric and categorical features. In [27, 31], the CatBoost algorithm is compared with XGBoost and LightGBM. In recent years, this algorithm is often used in various studies, for example, [32, 33].

### 3.4 Model training process

To train the model, we will use CatBoostRegressor from the catboost python package. Using the cross-validation method, we will select important hyperparameters – the maximum size of the tree and the number of iterations. The optimal number of iterations will be small (less than a hundred), which can easily be explained by the small number of objects in the sample. As the minimized quality functional, we choose the metric MAE (mean absolute error):

$$MAE = \frac{\sum_{i=1}^{n} |a(x_i) - y_i|}{n},$$

where $a(x_i)$ is the entertainment value predicted by the algorithm and $y_i$ is the original entertainment? The choice of this metric is based on the fact that it is important for us to take into account the absolute values of the deviation of the algorithm's response from the correct answer. Further, we will interpret as the result of the MAE value too.

# 4 Testing *EntertainmentIndex* Model and Evaluating Algorithm Accuracy

The approach described in section 3 is suitable for any football championship for which can be collected enough basic statistics. Let's test the quality of the algorithm on Russian Premier League matches of the 2019/2020 season.

## 4.1 Dataset creation

To get statistics, we used the data from the sports website Sportbox.ru [34]. Using methods for working with HTML pages, we extract the following information from the various site's pages:

1. Information about the match known before the start of the round – team names, date, number of the round, name of the referee, stadium, start time, day of the week.
2. Information about the position of teams before the round.
3. Statistics of the last match – the number of goals, shots, percentage of possession, etc.
4. Information about the course of the match – goals, penalties with time stamps.

Based on the collected information, we will prepare a table with all features described in section 3 and calculate the entertainment coefficients for all games. We will also calculate historical indicators for features and form the features-objects matrix. The resulting distribution of the entertainment index for the 2019/2020 season is shown in the Fig. 1.

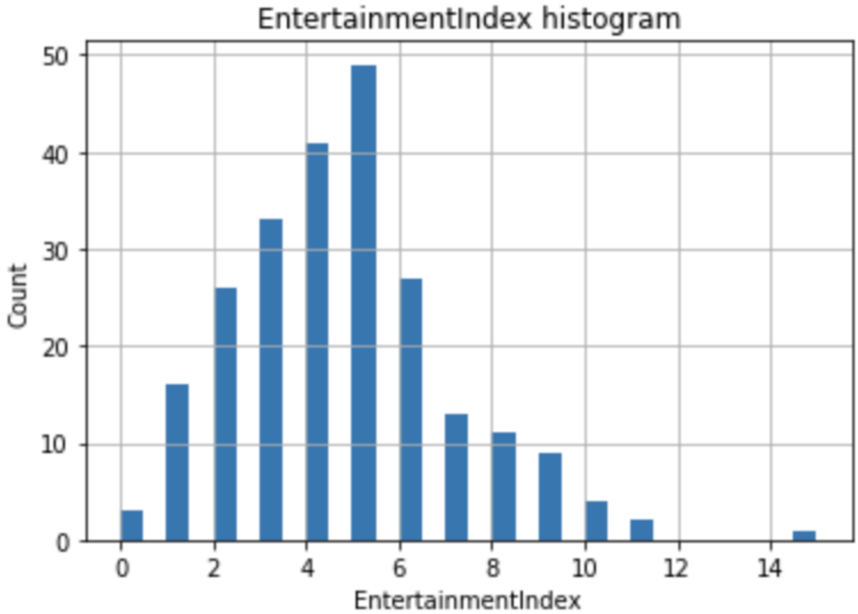


**Fig. 1.** *EntertainmentIndex* histogram for matches of Russian Premier League 2019/2020

Also note that we remove from the selection all matches that ended with a technical defeat of one of the teams, as well as the match that ended with the defeat of one of the teams that fielded the youth team. It was due to the coronavirus pandemic and the prediction of these matches results has nothing to do with proposed method. It also raises the question of how to collect historical data for the first rounds, when a sufficient number of matches have not yet been played. For this purpose, we suggest taking into account the last matches of the previous season. The indicator of teams that have moved up to the Premier League from the Russian Football National League (FNL) is counted on a par with the performance of Premier League clubs. Our study also considered other solutions to these problems – averaging the results of the last matches of the previous season, multiplying the indicators of the FNL championship by some coefficients of the ratio of individual indicators of the FNL to the RPL. There was analyzed distribution of matches entertainment in different seasons and divisions, so amendments were made in this regard. There was also a study of changes in indicators during the endings of the Russian Premier League Championships. However, the above amendments did not lead to an improvement in the quality of the model, so it was decided not to complicate the formation of data unnecessarily. The decision to choose a period equal to the length of one season is certainly not obvious. In this regard, other options were considered – to take several seasons, half of the season, the summer-autumn part and the spring part separately (in Russia, the major part of the winter football is not played due to climatic conditions). However, these changes also did not lead to a significant increase in quality.

### 4.2 Hypothesis testing and model training

We apply the method described in section 3.2 for the obtained sample. To estimate criterion confidence level the p-value estimation is usually used. It is the probability of obtaining test results at least as extreme as the results observed, under the assumption that the null hypothesis is correct. In addition to p-value, the ols function returns an estimate of the confidence interval, which can also be viewed [35], but for this article we will limit ourselves to p-value. If the p-value is small (usually p-value < 0.05), then the null hypothesis is rejected, which in the context of our problem means that the feature is significant. For our sample, the lowest p-value is achieved for such features as “Number of saves by the goalkeeper per match” and “Number of penalties scored". In general, this is quite expected, since these signs should significantly affect the entertainment of the match based on common sense and the logic of building the metric.

Let's train CatBoostRegressor on the available sample and configure the hyperparameters "maximum depth" and "the number of iterations" by using grid search cross-validation. Since the sample is small and consists of only 236 objects, there is a high risk of overfitting. Too much depth allows the model to accurately remember the training sample, and if it is not enough, there is a risk of underfitting. Along with limiting the depth, the following approach is often used to combat overfitting in boosting: training process stopped after some iteration, without allowing the model to retrain. The optimal parameters were selected by cross-validation as follows: depth 9, the number of iterations – 70. The MAE error value on the test set is 1.700, which is good considering the high nondeterminism of football indicators. Besides that, can be viewed as the contribution of individual factors in the CatBoostRegressor model, which is stored in the feature_importances_ field. The sum of all features indicators contribution is equal to 100. Table 2 shows the top 5 most significant features for the CatBoost algorithm.

**Table 2.** The most significant factors for entertainment prediction.

| Feature description | Feature significance (in a percentage) |
|---|---|
| Average overall ranking position of the competing teams in the standings | 8.536 |
| The number of missed goals | 8.324 |
| The number of corners | 8.293 |
| The number of saves | 7.160 |
| The number of offsides | 6.255 |

## 5 Discussion

Firstly, we will discuss the most significant factors identified by the model. Note that both the student's T-test and CatBoost gave a high score of significance for the number of saves. Indeed, goalkeeping saves to make the game more entertaining. Their number depends on both the goalkeeper game and the style of the playing teams. Of course, a lot of dangerous shots make the game spectacular, and their abundance is usually due to either good attacking actions (which usually persist for several matches and depend on the style of play and performers), or poor defensive play (which is also a long-term problem for the team). The number of goals conceded can be explained by exactly the same factors. The most significant factor is the overall rating of the teams in the standings. This is logical and consistent with the existing approach to choosing the most interesting matches to show to viewers – the higher the team, the better and more spectacular it plays. Of course, there are many exceptions, but they only confirm the existence of the rule. The number of corners and offsides directly depends on the activity of the attacking groups of opponents. From this viewpoint, everything looks very plausible. The least significant factor is the number of fouls (0.04%). Indeed, it is difficult to understand whether a large number of fouls indicates the intensity of the game, the players do not keep up with their opponents, or whether tactical fouls are committed.

Secondly, let’s discuss the accuracy of the algorithm's prediction. Of course, MAE equal to 1.7 is not too small, but on the other hand, the entertainment of the match is quite difficult to predict, because it depends on many factors. It is not uncommon for top clubs to play “tactical chess” and end their meetings with a boring 0-0. But sometimes it also happens in other ways, when low-scoring teams from the lower half of the standings arrange a scoring show. There was no overfitting during the model training, because the accuracy of the training sample was always tracked. The dependence of the feature descriptions for different objects is small, because we take into account the statistics of the last three matches during feature compiling, and therefore it is impossible to extract information about how one particular match ended from these features.

Thirdly, let's look at the business part that we started the article with. From a business point of view, it is more important to arrange matches by entertainment in each round. It is also difficult to come up with a single metric that would show the efficiency of the described algorithm. Let's enter the following indicator – the percentage of rounds in which the algorithm correctly predicted at least one match from the top 2 in terms of entertainment-and calculate its value for the sample. Technically, we will implement this in the following way: we will train the algorithm on the 29 rounds and predict the values of the entertainment index to the remaining one. For the sample under consideration, this indicator is 60%. Given that you need to choose two matches out of eight, this is a good result.

## 6 Conclusion

This article introduced the Entertainment Index algorithm for predicting whether a match is interesting for viewers. This algorithm mathematically describes the number of events interesting to the viewer on the football field, taking into account both the quantitative indicators of basic statistics and the numerical expression for the plot elements of the game. We created a machine learning model based on CatBoost, which predicts the entertainment of Russian Premier League matches with good quality. The most important factors affecting the entertainment of the match were highlighted. The subject area of the problem was determined and the application of the algorithm for the problems of recommendation and ranking of football matches by the broadcaster was shown. The proposed approach scales very well and can be used to solve similar problems in other regular football championships.

## References


1. Arabzad, S., et al.: Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. In: Journal of Applied Research on Industrial Engineering, vol. 1., № 3, pp. 159-179. (2014)
2. Rudrapal, D., et al.: A Deep Learning Approach to Predict Football Match Result. In: Computational Intelligence in Data Mining. Advances in Intelligent Systems and Computing, vol. 990, Springer, Singapore (2020)
3. Prasetio, D., Harlili, D.: Predicting football match results with logistic regression. In: 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), George Town, pp. 1-5 (2016)
4. Hvattum, L., Arntzen, H.: Using ELO ratings for match result prediction in association football. In: International Journal of forecasting, vol. 26, №. 3, pp. 460-470 (2010)
5. Bunker, R., Thabtah, F.: A machine learning framework for sport result prediction. In: Applied computing and informatic, vol. 15, №. 1, pp. 27-33 (2019)
6. Owramipur, F., Eskandarian, P., Mozneb, F.: Football result prediction with Bayesian network in Spanish League-Barcelona team. In: International Journal of Computer Theory and Engineering, vol. 5, №. 5, pp. 812 (2013)
7. Igiri, C., Nwachukwu, E.: An improved prediction system for football a match result. In: IOSR Journal of Engineering (IOSRJEN), vol. 4, №. 12, pp. 12-20 (2014)
8. Aldous, D.: Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability? In: Statististical Sciene, vol. 32, №. 4, pp. 616–629. (2017)
9. Joseph A., Fenton N., Neil M.: Predicting football results using Bayesian nets and other machine learning techniques. In: Knowledge-Based Systems, vol. 19, №. 7, pp. 544-553 (2006)
10. Hucaljuk, J., Rakipović, A.: Predicting football scores using machine learning techniques. In: 2011 Proceedings of the 34th International Convention MIPRO, IEEE, pp. 1623-1627 (2011)
11. Egebjerg, N. et al.: Big social data analytics in football: Predicting spectators and TV ratings from facebook data. In: 2017 IEEE International Congress on Big Data, IEEE, 2017, pp. 81-88 (2017).
12. Buraimo, B., Simmons R.: A tale of two audiences: Spectators, television viewers and outcome uncertainty in Spanish football. In: Journal of economics and business, vol. 61, №. 4, pp. 326-338 (2009).
13. Entertainment rating. Are the top matches of the English Premier League good? Available at: https://www.soccer.ru/blogs/record/958776/chelsi-tottenhem (accessed 12.10.2020)
14. The most entertaining football is in Holland, the best is in England. Only digit, available at: https://www.sports.ru/tribuna/blogs/instatfootball/1939406.html (accessed 12.10.2020)
15. Which League has the most entertaining football? We calculate it using mathematics, available at:

https://matchtv.ru/football/rpl/matchtvnews_NI791236_V_kakoj_lige_samyj_zrelishhnyj_futbol_Vychislajem_s_pomoshhju_matematiki (accessed 12.10.2020)

16. Methodology for calculating the entertainment rating of a match (football), available at: http://euro-rating.com.ua/rejtingi-v-sporte/rejting-zrelishhnosti-matcha/metodologiya (accessed 12.10.2020)
17. Bond, A., Addesa, F.: Competitive intensity, fans' expectations, and match-day tickets sold in the italian football serie A, 2012-2015. In: Journal of Sports Economics, vol. 21, №. 1, pp. 20-43 (2020)
18. Reade, J., Schreyer, D., Singleton, C.: Stadium attendance demand during the COVID-19 crisis: Early empirical evidence from Belarus. In: Applied Economics Letters, pp. 1-6 (2020)
19. Galitsyn, S.V. et al.: Neural networks application for attendance prediction of football matches. In: Uchenye zapiski universiteta imeni P.F. Lesgafta, №. 10 (176), pp. 76-82 (2019)
20. Uribe, R., et al.: Determinants of football TV audience: The straight and ancillary effects of the presence of the local team on the FIFA world cup. In: Journal of Business Research (2020).
21. Meier, H., Schreyer, D., Jetzke, M.: German handball TV demand: did it pay for the Handball-Bundesliga to move from free to pay TV? In: European Sport Management Quarterly, pp. 1-18 (2019).
22. HSE participates in the development of the Russian Premier League calendar for the second year in a row, available at: https://www.hse.ru/news/science/382737926.html (accessed 12.10.2020)
23. statsmodels.formula.api.ols, available at: https://www.statsmodels.org/stable/generated/statsmodels.formula.api.ols.html (accessed 12.10.2020)
24. Tianqi, C., Carlos, G.: XGBoost: A Scalable Tree Boosting System. In: arXiv preprint arXiv:1603.02754 (2016).
25. Ke, G. et al.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA (2017).
26. Gradient Boosting Machine, available at: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html (accessed 12.10.2020)
27. Hancock, J., Khoshgoftaar, T.: Survey on categorical data for neural networks. In: Journal of Big Data 7, 28 (2020)
28. Dorogush, A., Ershov, V., Gulin, A.: CatBoost: gradient boosting with categorical features support. In: arXiv preprint arXiv:1810.11363 (2018).
29. Prokhorenkova, L., et al.: CatBoost: unbiased boosting with categorical features. In: Advances in neural information processing systems, pp. 6638-6648 (2018).
30. Potdar, K., et al.: A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. In: International Journal of Computer Applications (0975 – 8887) vol. 175, №. 4, (2017)
31. Huang, G., et al.: Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. In: Journal of Hydrology, vol. 574, pp. 1029-1041 (2019)
32. Daoud, E.: Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset, available at: https://publications.waset.org/10009954/comparison-between-xgboost-lightgbm-and-catboost-using-a-home-credit-dataset (2019)

33. Kang, P., et al.: Catboost-based framework with additional user information for social media popularity prediction. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2677-2681 (2019).
34. Information portal. Football news in Russia and the world, available at: https://news.sportbox.ru/Vidy_sporta/Futbol (accessed 12.10.2020)
35. Cumming, G.: Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. In: Perspectives on Psychological Science, vol. 3(4), pp.286-300 (2008)