Saint-Petersburg State University
Mathematics and Mechanics Faculty

Albina Yezus

Predicting outcome of soccer matches using machine learning

Term paper

Scientific adviser:
Alexander Igoshkin, Yandex Mobile Department

2014

1 CONTENTS

2 Abstract 3
3 Introduction 4
4 Problem statement 4
5 Approach 4
5.1 Choosing match set that is to be analyzed 4
5.2 Deciding on key features. 5
5.3 Data extraction 5
5.4 Testing various machine learning algorithms 5
5.5 Improving implemented algorithm 5
6 Algorithms 5
6.1 Match data 5
6.2 Key features 6
6.2.1 Feature formulas 6
6.2.2 Plots 7
6.3 Data extraction 8
6.4 Machine learning algorithms 9
7 Technology 9
7.1 Programming language 9
7.2 Environment 10
7.3 Libraries 10
8 Difficulties 10
8.1 Difficult data extraction 10
8.2 Layout changes 10
8.3 Feature evaluation 11
8.4 Classifier evaluation 11
9 Results 11
10 Further research 11
11 Related work 11
12 Conclusion 12
13 Links 12
14 References 12

2 ABSTRACT

In this study the methods of machine learning are used in order to predict outcome of soccer matches. Although it is difficult to take into account all features that influence the results of the matches, an attempt to find the most significant features is made and various classifiers are tested to solve the problem.

Keywords: machine learning, sporting prediction, soccer, data mining

3 INTRODUCTION

The aim of this work is to see if it is possible to predict the outcome of sport games with good precision. It is to be done by analyzing soccer matches of various football leagues. Firstly, it is crucial to choose features that seem to be significant carefully and analyze their influence on matches outcome. Secondly, using machine learning methods, such as KNN, Random Forest, logistic regression, SVM and others, the model is to produce an output representative of the probable outcome of the match.

Several attempts were made to create a model that would be able to predict the outcome of the games with good precision; however, it has appeared to be utterly difficult to succeed in this field. Thus, the study is considered to be successful if it predicts the outcome with the precision of 70%.

4 PROBLEM STATEMENT

To create a model that predicts outcome of soccer matches with sufficient accuracy. Sufficient accuracy means one of following:

1. 70% accuracy when predicting results of match set;
2. Making profit betting against bookmakers.

5 APPROACH

It is possible to divide all work into two steps:

1. Choosing match set that is to be analyzed;
2. Deciding on key features;
3. Data extraction;
4. Testing various machine learning algorithms;
5. Improving implemented algorithm.

These steps may be intermingled, as feature's significance can be evaluated basing on the results of algorithms that are applied later.

5.1 CHOOSING MATCH SET THAT IS TO BE ANALYZED

The problem of choosing between different leagues is not an easy one as the following problems should be avoided:

1. Unfair refereeing;
2. Match-fixing;
3. Difficult or impossible data extraction.

5.2 DECIDING ON KEY FEATURES.

In order to create bedrock for future research the initial model is created. The features of this model are chosen to fit the following conditions

1. What seems legit?
2. What is possible to extract?
3. What really matters?

5.3 DATA EXTRACTION

Data extraction task is solved by using data from websites that store information about league which is chosen to be observed. The following information should be available:

1. Match information;
2. Season information;
3. Result table for each moment in the season.

The data is extracted by parsing html source of pages with necessary information.

5.4 TESTING VARIOUS MACHINE LEARNING ALGORITHMS

As the problem is one of classification problems where result equals one of 3 values {0, 0.5, 1} which mean loss of the fist team, draw match and win of the first team respectfully, several methods can be applied. The methods used in this term paper are:

1. K nearest neighbors;
2. Random forest;
3. Logistic regression;
4. Support vector machine.

5.5 IMPROVING IMPLEMENTED ALGORITHM

It may be useful to improve implemented algorithm in accordance to the stated problem. As it is not yet decided which algorithm suits the problem best, this part of the research is not done yet.

6 ALGORITHMS

6.1 MATCH DATA

The subject of the research is English Premier League as it seems to lack problems stated before.

6.2 KEY FEATURES

6.2.1 Feature formulas

The initial set of features can be divided into two groups:

1. Static;
2. Dynamic;

Where static features are given for each team and do not depend on the rival and dynamic features depend on both teams and represent their correlation.

All features are normalized, which means they are from the interval [0; 1].

6.2.1.1 Static features

1. Form:

$$\frac{1}{10}\sum_{k=1}^{5} res_k$$

where res_k - result of the –kth match (value in {0, 1, 2});

2. Concentration:

$$1 - 2 * x$$

where x – the nearest match lost to the weak team (the difference between the current team and that team >= 7);

3. Motivation:

$$\min(\max\left(1 - \frac{dist}{3 * left}, derby, \frac{tour + dist}{2}\right), 1)$$

where:

- derby – 1 if match is a derby and 0 otherwise;
- dist – distance to the nearest “key position”;
- left – tours left in the season;
- tour – 1 if left < 6 and 0 otherwise;
- key position – value in {1, 2, 3, 4, 5, 6, 17, 18}.

6.2.1.2 Dynamic formulas

4. Goal difference:

$$\frac{1}{2} + \frac{dif}{2 * \max\_dif}$$

where:

- dif – difference between goals;
- max_dif – maximal difference between goals;

5. Score difference:

$$\frac{1}{2}+\frac{dif}{2*\max\_dif}$$

where:

  - dif – difference between scores;
  - max_dif – maximal difference between goals;

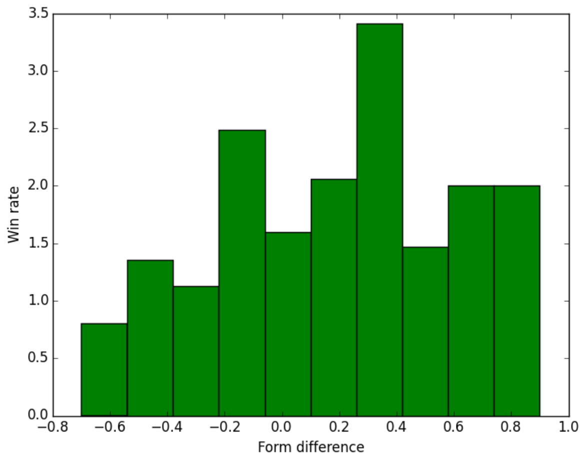
6. History:

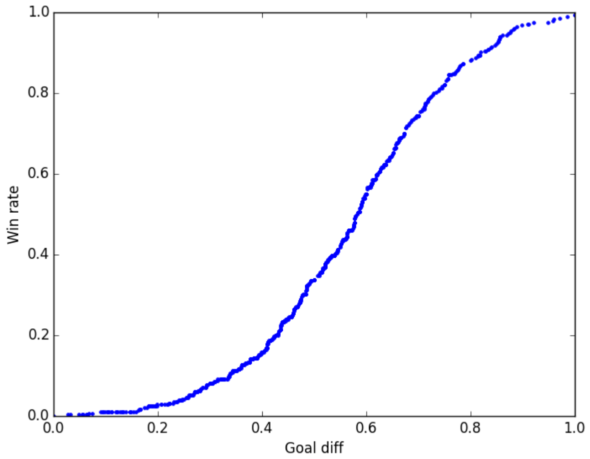
$$\frac{p1+p2}{2}$$

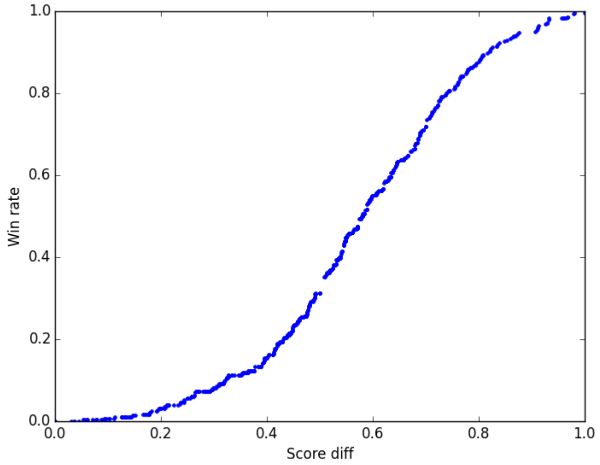
where p_i – result of –i^{th} match (value in {0, 0.5, 1}).

6.2.2 Plots

As it is difficult to estimate feature significance at this point, it may be possible to reveal dependences based on plots. The results are not very encouraging. Still, they are interesting to look at.







It is not difficult to see that the higher the score difference and goal difference, the more matches are won. On the other hand, it is difficult to find correlation between form difference and match outcome, so it is a good thing we have separated those two features.

6.3 DATA EXTRACTION

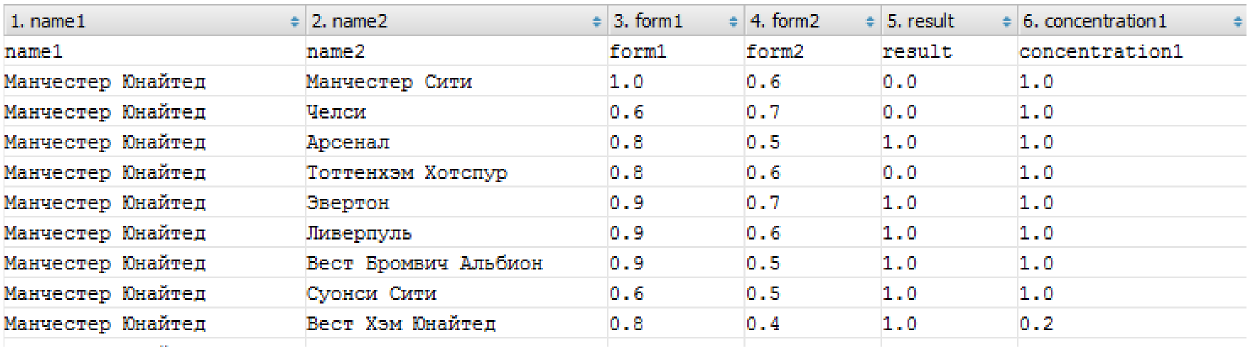
The information was extracted from the following websites:

1. http://www.championat.com/
2. http://www.statto.com/

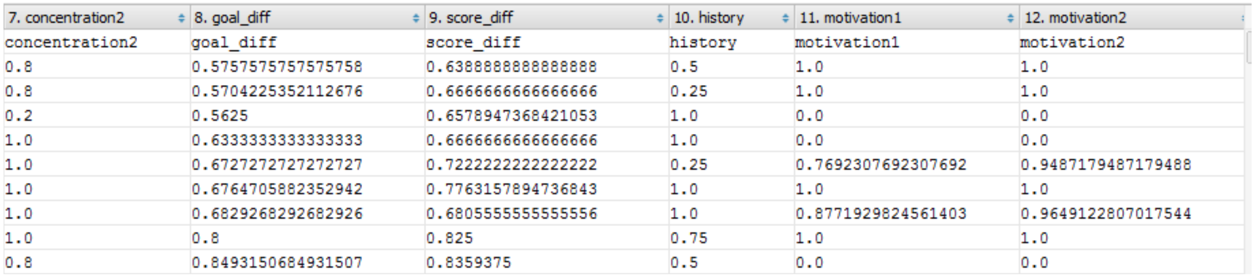
From *champinat.com* it was easy to loop over matches for each season and get information about form, concentration and history by parsing page with match information.

Statoo.com was useful as it has result table for each date of the season, so information based on scores, positions etc was extracted from there.

The collected dataset has the following representation:



1. name1	2. name2	3. form1	4. form2	5. result	6. concentration1
name1	name2	form1	form2	result	concentration1
Манчестер Юнайтед	Манчестер Сити	1.0	0.6	0.0	1.0
Манчестер Юнайтед	Челси	0.6	0.7	0.0	1.0
Манчестер Юнайтед	Арсенал	0.8	0.5	1.0	1.0
Манчестер Юнайтед	Тоттенхэм Хотспур	0.8	0.6	0.0	1.0
Манчестер Юнайтед	Эвертон	0.9	0.7	1.0	1.0
Манчестер Юнайтед	Ливерпуль	0.9	0.6	1.0	1.0
Манчестер Юнайтед	Вест Бромвич Альбион	0.9	0.5	1.0	1.0
Манчестер Юнайтед	Суонси Сити	0.6	0.5	1.0	1.0
Манчестер Юнайтед	Вест Хэм Юнайтед	0.8	0.4	1.0	0.2



7. concentration2	8. goal_diff	9. score_diff	10. history	11. motivation1	12. motivation2
concentration2	goal_diff	score_diff	history	motivation1	motivation2
0.8	0.5757575757575758	0.6388888888888888	0.5	1.0	1.0
0.8	0.5704225352112676	0.6666666666666666	0.25	1.0	1.0
0.2	0.5625	0.6578947368421053	1.0	0.0	0.0
1.0	0.6333333333333333	0.6666666666666666	1.0	0.0	0.0
1.0	0.6727272727272727	0.7222222222222222	0.25	0.7692307692307692	0.9487179487179488
1.0	0.6764705882352942	0.7763157894736843	1.0	1.0	1.0
1.0	0.6829268292682926	0.6805555555555556	1.0	0.8771929824561403	0.9649122807017544
1.0	0.8	0.825	0.75	1.0	1.0
0.8	0.8493150684931507	0.8359375	0.5	0.0	0.0

6.4 Machine learning algorithms

For the algorithms described below implemented classifiers from python library were used.

The methods applied with the resulting precision:

1. K nearest neighbors: 0.558
2. Random Forest: 0.634

As these methods are academic and are not usually used in real researches, not much effort was put into them. Considering this fact, the results are quite encouraging.

7 Technology

7.1 Programming language

Python is the main programming language of the project as there are many useful libraries included, that simplify data extraction and training/testing machine learning classifiers much easier.

The following versions of Python were used:

1. Python 3.3;
2. Python 3.4.

Python 3.3 was used as it is the latest available version on Wakari.

Python 3.4 was used as it is the latest available version of Python.

7.2 ENVIRONMENT

There are quite few GUIs for Python. As this paperwork is the result of collaboration, there is a need of something that allows several people to have access to the source code. Thus the following environments were used:

1. Wakari;
2. PyCharm.

Wakari allows to create notebooks in which it is easy to launch various part of the code and write report on the go, so it is handy for writing classifiers.

PyCharm simplifies git sync and library installation, supports TODO syntax and has many other useful features.

7.3 LIBRARIES

The following libraries for Python were used:

1. For extraction:
    a. Selenium;
    b. Urllib;
    c. BeautifulSoup;
2. For maching learning algorithms:
    a. Sklearn;
3. For data analysis and everywhere else:
    a. Numpy;
    b. Pandas;
    c. Matplotlib;
    d. Etc.

8 DIFFICULTIES

During the work on the research several problems were encountered.

8.1 DIFFICULT DATA EXTRACTION

Although chosen features required simple information to be extracted, it was not always a straightforward task. The main problem was with *statto.com*, where information is stored in javascript and cannot be accessed simply by following the link. It was necessary to send a request that changes the page layout so that it can be parsed. This problem has been solved by applying selenium library that simplifies the process.

8.2 LAYOUT CHANGES

In the middle of the work *championat.com* layout changed. This difficulty was overcome by adjusting code to new layout.

8.3 Feature evaluation

As only a couple of classifiers were applied, it is difficult yet to evaluate significance of each feature. Moreover, no methods were applied to select the most significant once. This problem is to be solved in the future by looking into classifiers' results with less / other / modified features and applying various methods, such as Lasso.

8.4 Classifier evaluation

It is not yet clear how to try the classifiers against the bookmakers. This problem is to be solved in the future when better classifier is found.

9 Results

The following result has been got:

1. Dataset with 9 features and 640 objects;
2. Scripts for simple data extraction;
3. Academic machine learning classifiers applied;
4. Results of classifier over 60%.

10 Further research

This work has recently been started and although there are already some results obtained, there is still a long way ahead. The directions of the further investigation are:

1. **Varying data.** By this we understand varying size of the dataset, number of features and formulas of features.
2. **Training and testing classifiers.** As only academic classifiers were yet tested, it is necessary to apply other methods, such as SVM and Logistic regression.
3. **Adjusting the best classifier.** This means either rewriting it in accordance to the stated problem or writing some kind of combination of different classifiers.
4. **Beating bookmakers.** This is the ultimate goal of this paperwork. The point is to find an algorithm that would be able to make profit by betting.

11 Related work

As it was mentioned before, attempts to create a model that would successfully predict outcomes of sport events were made before.

Concerning football one piece of research (Hamadani, 2006) offered approach similar to the current study. In that study 3 seasons were tested and it was discovered that each season a different set of features had more significance which means either that football is an unstable game or that optimal features lie somewhere between those found by the study.

Another study (Blundell, 2009) has proved that American Football matches can be accurately modelled using features within a regression model. It was also discovered that simple logistic model could achieve just as accurate forecasts compared to some more complex alternatives.

However, all aforementioned studies researched long periods without taking into account a third factor. This study, on the other hand, focused on a particular sport league without limit to matches in this league only – results of the players in other leagues are also to be taken into account. Thus the results of this study are expected to be more accurate.

12 Conclusion

Machine learning methods can be applied to different fields, including sports. On the example of English Premier League it is shown that it is possible to find a classifier that predicts the outcome of soccer matches with the precision of more than 60%. However, there is still a lot of work to be done and the research will be proceeded.

13 Links

The source code may be found by the following links:

- Work related to machine learning: https://wakari.io/sharing/bundle/Albinutte/Project;
- Work related to data extraction: https://github.com/Albinutte/football-prediction.

14 References

[1] Predicting the outcome of NFL games using machine learning, Babak Hamadani, cs229 - Stanford University (2006)

[2] Numerical Algorithms for Predicting Sports Results by Jack David Blundell, School of Computing, Faculty of Engineering (2009)

[3] Predicting football results using Bayesian nets and others, machine learning techniques, A. Joseph ¤, N.E. Fenton, M. Neil (2006)

[4] Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line, Jim Warner (2010)