Predicting Results using Logistic Regression & Decision Tree

by

Aradhya Mehta

A Project Report Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in
Computer Science

Supervised by
Dr. Carol Romanowski

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

May, 2016
Approval

Name: Aradhya Mehta

Degree: Masters in Computer Science

Project: Predicting Results using logistic regression & decision tree

Advisor: Dr. Carol Romanowski
Associate Professor

Signed: ____________________________

Date: ______________________________

Computer Science @ RIT
Acknowledgements

I would like to thank my advisor, Dr. Carol Romanowski for the guidance and advice during the entire duration of the implementation of the project. She has been helpful throughout the semester and has been providing consistent feedback which made the work easy. I would also like to thank Prof. Zack Butler for all his help during the colloquium. His feedback every alternate week was helpful and made me improve on many aspects of the project. Finally I would like to thank my friends for all the support and also providing valuable feedback during the duration of the project.
ABSTRACT

Predicting Results using logistic Regression & Decision Tree

Aradhya Mehta

Supervising Professor: Dr. Carol Romanowski

Soccer is a sport that could be very useful in data mining when it comes to predicting results. The paper aims to show the different ways that could be used to predict results based on previous matches. Most of the work that has been carried in such a field has been for multiple seasons which is not the most suitable approach since with every season the fortunes of a team could change and those past results from an inconsequential season should have no bearing on it. The paper aims to rectify that by predicting results of the same season based on those same season’s results. The accuracy of the prediction greatly depends on the set of features that have been selected for prediction from the dataset. The paper makes use of multiple approaches to select the best set of features. Logistic Regression and decision tree are algorithms that will be use to predict results using R. All of the feature selection approaches have then been compared to other similar work carried out by authors using the same approaches. Improvements have also been suggested that could aid in increasing the accuracy of the results even further.
Contents

Approval ........................................................................................................... ii
Acknowledgement ............................................................................................ iii
Abstract ............................................................................................................. iv

1 Introduction ..................................................................................................... 1
2 Related Work .................................................................................................. 2
3 Dataset and Data collection .......................................................................... 3
4 Approach ......................................................................................................... 4
   4.1 Feature Selection ....................................................................................... 4
       4.1.1 Elimination Approach ................................................................. 4
       4.1.2 Calculated Elimination ............................................................... 5
       4.1.3 Single Team Prediction ............................................................... 6
   4.2 Classifier Algorithms ............................................................................... 7
       4.2.1 Logistic Regression ........................................................................ 7
       4.2.2 Decision Tree ................................................................................ 7
5 Results ............................................................................................................ 9
6 Evaluation Compared .................................................................................... 11
7 Future Work ................................................................................................... 13
8 Conclusion ....................................................................................................... 14
9 References ...................................................................................................... 15
List of Tables

5.1. Accuracy Results .................................................. 9
5.2. T-score Results .....................................................10
List of Figures

4.1. Elimination Approach .............................................. 5
4.2. Calculated elimination approach ............................... 6
4.3. Probability during decision tree ................................. 8
5.1. Graph Comparison with results of other authors .......... 11
Chapter 1
Introduction

Soccer is one of the most popular sports in the world. There is keen interest from all corners to be able to predict what the results of future matches will be based on the team’s previous results and form. These heat-of-the-moment predictions are made by both fans and also betting companies. Keeping all the emotions aside and predicting the results for future games is where the field of data mining could be used and worth exploring. Soccer has seen good progress with regards to statistics and although the importance of statistics is not as important as American Football or Baseball, organizations and soccer teams are using more and more data mining methods to make sure that they have an upper-hand on their opponents by knowing every statistic about them. The aim of this project is to predict the result of soccer games using data from the previous season with accuracy being the main factor for evaluation. The prediction is to be done on the soccer matches from the English premier league around the world consisting of a single season.

Predicting the results could differ drastically based on the information or the data that the system is provided. The prediction could be different if only the team’s performances are taken into account, they could be different if player performances are taken into account and they could be different if the weather conditions are taken into account. The data could be very vast for prediction. In addition to the usual features such as Home team score, away team score etc. there are many other features related to the game which could make the prediction more accurate. Features such as cards accumulated by the team, tendency to score within a certain period of the game, also individual player performances, all these could be used to improve the accuracy of the prediction more. For this project though, we will not be considering individual performances as the data set becomes very vast and difficult to deal with. From the features that will be considered for the project as well, an important part is to select the right set of features since the accuracy of the prediction is largely dependent on the set of features. The features that will be used for selecting the best set of features in the data collection chapter 3. How the different approaches were used to select the features have been selected has been explained in chapter 4.1. The classifier algorithms Logistic Regression and Decision tree will then be implemented on the data set and the accuracy of the results will then be evaluated and compared with Haghighat [2], Buursma [3] and SingleTeam [4] which have been explained in chapter 4.2 and chapter 6 respectively. Since there were three different approaches, it was important to understand the differences between the three of them and that has been explained in chapter 5. There are some weaknesses in this project and can be worked on in the future and they have been explained in chapter 7.
Chapter 2
Related Work

Some of the steps of this project have been put in place because of ideas from other similar work of predicting results. Haghighat [2] also has worked on predicting results although the authors work is more focused towards plenty of sports rather than just soccer. This paper manage to compare its elimination approach to the approach followed by Haghighat. Although the approach employed for selecting the features is more or less the same, the algorithms that have been used to correctly predict the results are different. This paper makes use of the logistic regression and decision tree algorithm whereas Haghighat employs many other algorithms such as the artificial neural network, support vector machine algorithm. All of these approaches have then been compared with accuracy being the metric for comparison.

Another paper, known as Buursma [3] has also predicted results in soccer. The calculated elimination approach followed in this paper has its idea taken from Buursma. This paper, like Haghighat also employs plenty of algorithms to predict the results so those results have been compared with the logistic regression and decision tree algorithms implemented by this paper.

Finally, SingleTeam [4] is another paper that has been doing its job of predicting results as well over a couple of seasons. The approach followed by this paper though, is considerably different to the approach followed by SingleTeam. Unlike SingleTeam, this paper does not have any external factors that affect the prediction in a positive or negative way. SingleTeam also look at player performance’s along with team performances whereas this paper just looks at player performance over the last couple of seasons.
Chapter 3
Dataset and Data Collection

The data was collected from a soccer website known as football-data.co.uk [1], where we managed to get the dataset for the English premier league season 2010/11. There was no real need to clean the data since most of the data taken from that website was already clean. Some of the features mentioned below are the ones that play an important part in this project and these features being mentioned here gives a brief overview of what the dataset looks like.

- **Home Goal** = “goals scored by the home team”
- **Away Goal** = “goals scored by the away team”
- **Home Shots** = “Total number of shots by the home team”
- **Away Shots** = “Total number of shots by the away team”
- **Full Time Result** = “Final result of the game”
- **Home Shots on target** = “Number of shots on goal”
- **Away Shots on target** = “Number of shots on goal”
- **Home Team Corners** = “Corners obtained by the home team”
- **Away Team Corners** = “Corners obtained by the away team”

The Full Time Result feature outputs the result as either ‘H’, ‘A’, ‘D’, which stand for home win, away win or a draw respectively. This feature is only present in the training dataset but is not present in the test dataset since the prediction will be done on the test dataset. These are the features that have been use for predicting the results of the matches. Not all of these features have been used for all of the three approaches that have been explained in chapter 3 but a combination of some of these features have been used for the three approaches.
Chapter 4

Approach

4.1. Feature selection

The feature selection approach is the most important step with regards to this project. Selecting the right set of features is very critical to increase the chances of correct prediction of the matches. For this project, there were a lot of papers read and research done in order to pick the best approach but unfortunately I couldn’t pick one approach to be the one which stood out more than the others. So I decided to use all the three feature selection approaches and then compare its results to the other author’s work the three approaches that were picked were by all different authors, i.e. Haghighat [2], Buursma [3] and SingleTeam [4]. The initial two changes have pretty much the same working flow bar a minor but important change in the second approach. The three approaches have been explained in more detail below:

4.1.1. Elimination approach

This idea behind this approach was taken from Haghighat [2]. Figure 3.1 below shows the flow behind this approach. The dataset initially had 57 features. Most of those features, 34 of them to be exact were of no use to the project since they were features that were related to betting odds etc. and were out of the scope of the project. After these unnecessary features were discarded, we are left with 23 features out of which some of them will become a feature set that will help us predict the soccer results. Now every feature was removed from the dataset one after one and the dataset was loaded in WEKA. Using the Multi class classifier in WEKA, we check for the accuracy of the dataset. If the accuracy of the dataset increases or stays the same after a feature is removed, then that feature is discarded for good. If the accuracy of the dataset decreases after removing a certain feature from the dataset, then the feature is kept in the dataset. This process is repeated until all of the features have not been removed once from the dataset and checked for accuracy with the above approach maintained for all of the features. Once some of the features have been discarded, we are left with a set of features which become the primary features of the dataset, i.e. the most important features, on the removal of which the prediction could become weak.

Once these primary features of the dataset have been decided, the eliminated features are added one by one into the dataset again. On adding every feature, the accuracy is checked for using the multi class classifier algorithm in WEKA. If the accuracy increases on adding a certain feature, then the feature is added back into the dataset although it does not become part of the primary set of features. If the accuracy stays the same or decreases on adding a certain feature to the dataset, then the feature is discarded and not added to the dataset. With this approach, we ended up with 7 features in total in the dataset that will help us in predicting the results of the dataset. The reason
behind using WEKA for this approach was that the final goal was to be able to predict correctly and implementing that step was much more important to this, which was more of manual approach.

![Diagram of Elimination Approach](image)

**FIGURE 4.1: ELIMINATION APPROACH**

### 4.1.2. Calculated elimination

This approach, as mentioned before is more of using a step before implementing the elimination approach. The idea behind this approach was taken from Buursma [2]. Figure 4.2 below shows the calculated elimination approach flow. Here before the main flowchart that is followed in the above approach is made use of, there is one more important step that needs to be worked on that makes this approach different. In this approach, we only pick the features based on the statistics of the last ‘x’ amount of games, where x is the number of matches. Unlike the elimination approach, where the results of an entire season, i.e. 38 games are considered, this approach will consider results based on only the last ‘x’ amount of games. Again using the multi-class classifier in WEKA, accuracy for different ‘x’ are checked. We set x at 5, 10, 13, 15, 18 and 22 and check for the accuracy. We saw that the accuracy becomes constant at 15 to 18 so we set x as 15. Once we have found x, the statistics for the last 15 games are gathered in the dataset and then the same approach as the elimination approach is followed. We assumed that some features would change because of not counting the entire set of games but to our surprise, the features stayed the same after checking for accuracy and following the same flow. Having a different dataset with the same features might have helped but then comparing it for the sake of the project would have become very difficult.
4.1.3. **Single team prediction**

This approach is considerably different to the previous two approaches since it focuses on just a single team’s performance over the course of two seasons. Although the dataset that has been used is the same, only the matches that concern the team ‘Manchester United’ are taken into consideration. Unlike the previous two approaches, this time the final feature set contained comparatively more features which helped with the prediction. This is in contrast with the authors from SingleTeam [4], where the authors, along with making use of team performance features also make use of player performance which help them with the prediction in more detail. This approach that I made use of was very basic in this project since the dataset is only a few seasons long and the prediction is based on entirely one team irrespective of the opponent or their form. The entire purpose behind this approach was to understand how different the approaches are with regards to its feature selection.
4.2. Classifier algorithms

4.2.1. Logistic regression

Logistic Regression is a regression algorithm where $y = f(x)$ is a curve that has to be fit in the model. For such a regression curve, ‘$y$’ has to be predicted with the help of certain values of ‘$x$’. Logistic regression is known to give the best accuracy compared to the other prediction algorithms because of its ability to make its calculations very simple. These best results are usually present in binomial logistic regression where there are two results present, i.e. a 0 or 1, a yes/no etc., whereas in this project we will be implementing the multinomial logistic regression where we will be having more than two outputs. Both of the algorithms to predict the results was implemented in R. R makes it easy to fit the regression curve because of its glm() function, i.e. the generalized linear model. The glm() formula can be given as “Result <- glm(Y ~ X1 + X2 + … + Xn, data = “…”, family = “…”)” where $y$ in terms of this project is Full Time Result and the different $X$ are Home Goal, Away Goal, Home Shots, Away Shots etc. This is how glm() was used for the initial approach, i.e. the elimination approach. Once the model has been fit, the prediction of the results can be done using an predict() function in R, where we can get the accuracy of the dataset.

4.2.2. Decision tree

Decision tree is the other algorithm that has been used to predict the results once the features have been selected using the different feature selection approaches. Decision tree in R basically makes decisions for the user for the data that needs to be predicted. Like the glm() function in logistic regression, the decision tree makes use of an rpart() function, i.e. the recursive partitioning and regression to make a rough decision for the prediction. The rpart() function also has a similar formula to the glm(), i.e. “rpart(y ~ x1 + x2 + x3, data = “…”, method = “…” )”. Once the rpart() is run on the dataset, figure 4.3 is generated below with the help of R’s visualization packages.
As can be seen from the figure, the rpart() function gives us the basic probability of what the results could look like if certain features were there. After the rpart() function gives us the probability with the help of R’s visualization packages, the predict() function predicts the result of the matches. The results are written onto a .csv file and the accuracy is then calculated.
Chapter 5
Results

Since I could not make use of just one approach for the feature selection, the results for all the three approaches had to be calculated. Apart from the accuracy percentage which showed the correctness of the results, I also managed to calculate the t-score, which basically is a way to know the difference from the mean or how further that score is to the mean.

The table below shows the accuracy and t-score for all the three approaches.

<table>
<thead>
<tr>
<th></th>
<th>LOGISTIC REGRESSION</th>
<th>DECISION TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elimination Approach</td>
<td>56.34 %</td>
<td>57.89 %</td>
</tr>
<tr>
<td>Calculated Elimination</td>
<td>65.47 %</td>
<td>62.36 %</td>
</tr>
<tr>
<td>Single Team Prediction</td>
<td>69.12 %</td>
<td>68.42 %</td>
</tr>
</tbody>
</table>

The above table shows that the elimination approach is not one of the approaches that should be made use of to select the features. The flowchart that has been explained for this approach in the feature selection chapter itself probably shows that this approach is more of a brute-force approach to select the features. Eliminating features and then checking for accuracy and then adding them back if the accuracy increases does not seem like the best approach in itself in theory. The calculated elimination feature selection approach in comparison makes more sense and an approach that can be used to select the feature. The reason for this is that the calculated elimination approach, although following the same approach as the previous one has a limit into how further into the history of the games should we look at to get the best features. In this approach, we first get a number to look into the games and then follow the flow to get the best set of features. This increases the prediction accuracy since only the more recent results of the team are used to get a prediction. So it is understandable that the prediction accuracy is better than the elimination approach. The third approach results are more different and higher than the initial two because the approach to select the features is limited to a single team (the opponents form also does not matter) and so since the dataset is more inclusive, the results and its accuracy will increase since the results will favor the single team’s previous results. So, according to the accuracy results, we could assume that the calculated elimination approach is the one that should be used for selecting the best features in order to predict the result. But to better compare these three approaches, we could understand and calculate t-score. This term, t-score basically means the ability to know as to how far the value is from the mean score. A t-score is basically a general score, given a mean value,
which in this project is 380. The table below shows the t-score of the three approaches using both the algorithms, i.e. logistic regression and decision tree.

<table>
<thead>
<tr>
<th>Approach</th>
<th>t-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elimination Approach</td>
<td>0.283</td>
</tr>
<tr>
<td>Calculated Elimination</td>
<td>0.536</td>
</tr>
<tr>
<td>Single Team Prediction</td>
<td>0.037</td>
</tr>
</tbody>
</table>

**TABLE 5.2: T-SCORE RESULTS**

According to the t-score, Approach 3 (single team prediction) is the most useful and closest in interpretation for the correct results but one issue with the third approach is that the mean value was different compared to the initial two approaches. So there can be no guarantee that if the mean values in the third approach increase (it is very low right now), the difference in accuracy could also increase which may make it a not so suitable approach then. According to the t-score though, the different between the initial two approaches is very minimal, which seems fair since majority of the approach is similar and all of the features are same, so only the values in the dataset might differ by a bit.

So overall, although the results generated were not as high as we expected because of the fact the dataset was not reliant on other external features to make a difference to the result; the accuracy is still decent in value and probably using a more inclusive dataset or using a dataset which has more important or more external features would improve the prediction accuracy by quite a good number.
Chapter 6
Evaluation Compared

The results mentioned above in this project will be compared to three other papers that I have used for research. All of these papers have done similar work (not exactly same) in this field and these authors were the closest I could find to compare with the results of my project. The graph below compares my results with the other results generated by the authors using similar approaches.

NOTE: All the authors have not made use of all the feature approaches mentioned in this paper. The calculated elimination approach was followed by only one other author (Buursma). Similarly, only the one author has used the single team approach that was similar to my approach. Only the elimination approach has been the approach that has been used by more than one author (Buursma, Haghighat), the results of which have been shown on the graph.
The graph above compares the results I managed to generate through my implementation and approach to the results generated by other authors. As can be seen by the graph, the elimination approach is very close regarding my results and the author’s results. Since the approach is very simplistic, it is understandable that most of the results are very close although the feature actually are different. Similarly, the other two approaches are also very similar in results when compared to the results to the other authors. Only the single team prediction has results which are considerably lower to this project’s approach. This is because of the fact that external performances probably lower the results when dealing with a single team and the author’s work does that whereas my project just looks at the team’s performance history. There have been more authors who did work on such things but it was not suitable to compare their results since the dataset or approach selected was not the same as the one used in this project.
Chapter 7

Future Work

- There is quite a lot that can be improved in this project. The dataset as of now contains only 54 features, most of which have been discarded since they are not at all relevant to the dataset. This can certainly be worked on and a dataset could be more inclusive even though it has a lot of features. One of the steps that can be improved in the project is the use of external factors to improve the accuracy. The dataset in this project only takes care of the team’s performances and its external factors such as number of yellow cards, number of shots on target etc. The dataset though, does not take into account external factors such as player performances, weather condition etc. The dataset could include certain features such as player impact, player history in a certain games, i.e. has he made impact against the opposition or not etc. Factors such as these will tremendously help in increasing the accuracy while predicting the results for a game or for bunch of games. Similarly, weather condition features could also improve the prediction results and in turn improve the accuracy of the prediction. Features such as whether a team plays better at night or in the evening, whether the team plays better in dry conditions, or moderate or rainy or snowy conditions. Like these, many factors could impact the results of the prediction could affect the prediction of a particular match. The use of external factors is certainly one aspect that would surely improve the results of the project.

- Another aspect of the project that can be worked on in the future is the better selection of feature to get the prediction accurate. Although all of the three feature selection approaches give decent results, there could be better approaches to select the features or someone could actually design an approach that selects the results better. The current approaches that I made use of in this project feel too much a brute-force way of trying to select through accuracy of the dataset. There certainly has to be better ways of selecting the features that could help with increasing the accuracy of the results.

- The scope of the project could also be increased if the authors have more time to work on the project. Like soccer, a prediction system could be made for many other sports, especially the more statistics heavy sports such as football, baseball etc. Since these sports are more inclined to statistics compared to soccer, having more features to work with will certainly give better results simply by using the approaches used in this project.

- Another aspect that can be improve upon is the test dataset. The split that I did for the training and the test dataset was manual, i.e. for 380 matches, I split the values to 220 training matches and 160 test matches. This manual splitting could have potentially affected the results. Some k-cross validation in the future would be very useful and a neutral split would lead to better results.
Chapter 8

Conclusion

Through this project, we were able to predict the results of soccer matches using previous data. Feature selection was the most important step before predicting the results and so because no clear useful approach, we made use of three approaches, i.e. elimination approach, calculated elimination and single team prediction to predict the results of the matches. Logistic regression and decision tree were the two algorithms that were used to predict the results using R. The results were somewhat disappointing but expected with most of the accuracy of the results coming in the 55 % - 65 % range. Better feature selection approaches or making use of features of external factor such as player performance, weather condition etc. might have resulted in better predictions occurring. There is a lot that can be worked on in the future such as external factors, having better ideas for feature selection, having a dataset which spans multiple seasons instead of a single one are some of the major pressing concerns that could be improved upon.
Chapter 9

References

[1] Dataset: football-data.co.uk [Online; accessed 02-May-2016]


