PREDICTING WINNER BETWEEN INTER MILAN AND JUVENTUS USING LOGISTIC REGRESSION

Rohit Sen

Project URL:
https://github.com/senrohit/CSE482Project

ABSTRACT

Football is the most watched and most popular sport in the world. Everywhere around the world people always want to know the outcome of a game before it happens and numerous websites do analysis by analyzing the teams and try to predict the outcome. The goal of this project is to create a model based on numerous seasons of Seria A league data containing a large amount of games and use the model to predict a game of the current season. I formulated the problem as a ternary classification problem and applied Logistic Regression to solve it. After creating the model, the F1 score and accuracy score for test set came out to be: 0.4587, 0.6067 respectively. If we look at the accuracy, it can be said that it is possible to predict the winner of the game.

1. INTRODUCTION

For introduction, you need to include the following information:

1. As one of the most popular sports on the planet, football has always been followed very closely by a large number of people. In recent years, new types of data have been collected for many games in various countries, such as play-by-play data including information on each shot or pass made in a match. The collection of this data has placed Data Science on the forefront of the football industry with many possible uses and application like match strategy, tactics, analysis, betting, etc. In particular, the betting market has grown very rapidly in the last decade, thanks to increased coverage of live football matches as well as higher accessibility to betting websites.

2. This project aims to predict the possible winner of a Seria A league game by using a prediction method, namely Logistic Regression, where a model is built from data and that model is used to evaluate how accurate the f1 score and accuracy of the prediction is. This has become possible thanks to the large amount of data that is now being recorded in football matches. The main approach that is taken is to build a model from the data in order to better understand a team's performance and thus to generate better predictions for the future. We will use data for shots taken, goals scored, which team is home or away, amount of red cards, yellow cards, etc in order to reduce the impact of luck on the final prediction result.

3. The data was collected from http://football-data.co.uk/data.php . Four seasons- 2015-16, 2016-17, 2017-18, 2018-19 were used as the raw data. The total amount of games in the raw data appeared to be 1470, and each row contained attributes such as: Date, HomeTeam, AwayTeam, Home team goals, away team goals, shots of each team, corners of each team, fouls committed by each team and numerous others, which helps us create features related to team form and others that will help with the prediction.

4. There were some challenges that were encountered while trying to parse through and analyze the data. It was helpful to parse data as time because time is easier to deal with. Some NaN values were encountered which were converted to empty strings for ease of parsing. I initially thought ternary classification was a little puzzling, but Logistic Regression helps with it. I was also initially having trouble with splitting the data into test (season 2018-19) and train(rest of the seasons) from the main dataset of all the seasons but manually splitting it with slicing got rid of the problem.

5. The findings of the project are that it is possible to predict the outcome of a game as there were good accuracy and f1 scores for our data. It is not possible to correctly predict the outcome of a game all the time, but Logistic Regression helps us with a good accuracy for the prediction.

2. DATA

The data was collected from http://football-data.co.uk/data.php . Four seasons- 2015-16, 2016-17, 2017-18, 2018-19 were used as the raw data. The total amount of games in the raw data appeared to be 1470, and each row contained attributes such as: Date, HomeTeam, AwayTeam, Home team goals, away team goals, shots of each team, corners of each team, fouls committed by each team and numerous others. All the separate dataframes of the different seasons was merged together by using pd.concat.

The data that was collected includes all the 38 games played by each of the 20 teams for the last 4 seasons of the Italian Seria A league. There were some NaN values present in the dataset which were replaced by an empty string for ease of parsing. There were a lot of unnecessary betiing related odds columns which were all discarded and only information related to the game was extracted. The raw data had 1470 rows and more than 40 features.

CSE881-2015, Month 1–2, 2004, City, State, Country.

The NaN values that were encountered were converted to empty strings for ease of parsing. The data from the four seasons were merged together by using pd.concat. Date was parsed as time and then back to date again because the formats of the dates for some of the data were different. For some features that were created by me, I scaled them, centered to the mean and component wise scale to unit variance. Using the fact that a team won,lost or drew a game, a points tally was created for each team to understand each teams standing in the league table. Five previous games for each team are each a feature. Goal difference of teams were calculated by subtracting goals conceded from goals scored. Team form was created from the previous game results of teams. The difference in points between two teams was created by subtracting Home team points and away team points, and same was done to create difference in form points. A matchweek was also created for every game played during a particular matchweek uptil matchweek 38 i.e, last matchweek of the season. HomeTeam and AwayTeam Losing Streak and winning streak were added to the final dataframe too. Before the data was passed through a prediction technique, all the categorical variables were converted into dummy variables.

A very large amount of work was done to calculate and add a total of 36 features to our main dataset. The set of predictor attributes are: ['HTP', 'ATP', 'HM1_D', 'HM1_L', 'HM1_W', 'HM2_D', 'HM2_L', 'HM2_W', 'HM3_D', 'HM3_L', 'HM3_W', 'AM1_D', 'AM1_L', 'AM1_W', 'AM2_D', 'AM2_L', 'AM2_W', 'AM3_D', 'AM3_L', 'AM3_W', 'HTGD', 'ATGD', 'DiffFormPts']. HM- Home, L- Loss, D – Draw, W- Win, GD – Goal Difference, HTP – HomeTeam points, ATP – AwayTeam points. The target attribute is – 'FTR' which stands for full time result and it can either a W(Win), D(Draw), L(Loss).

There were a total of 36 attributes that were created for the dataset. The size of the final dataset after preprocessing is – 1350 rows and 24 columns. The dataset size is 182 kilobytes.

3. METHODOLOGY

The final dataset was manually split into training and test set. The test set contains the games of the current season which is obtained by simple array slicing and the training set contains the games of all of the rest of the seasons. Then Logistic regression was performed with the training and test sets to output the f1 score and accuracy of the prediction.

- Project_bigdata.ipynb: this is the Jupyter notebook file that I wrote to collect data from the csv files and preprocess it and use logistic regression on the data.

4. EXPERIMENTAL EVALUATION

This section describes the experimental setup and results you obtain.

4.1 Experimental Setup

This section should include:

1. Computing platform: Jupyter Notebook
2. I used Logistic Regression to calculate the f1 score and accuracy score on the training and test data sets.

4.2 Experimental Results

The f1 score and the accuracy score obtained from using logistic regression to train a model and then predict the test are as follows:

F1 score and accuracy score for training set: 0.6369 , 0.6905.

F1 score and accuracy score for test set: 0.4587 , 0.6067.

The scores indicate that a successful prediction of a HomeTeam win can be possible using the data.

5. CONCLUSIONS

The overall result of this project is fair considering the amount of data used. Some ideas for improvement are to add more data and to use different kinds of techniques such as Support Vector Machine, Poisson Distributions, etc which might help increase the f1 score and the accuracy score.

6. REFERENCES (at least 3 references)

1) B "Predicting Sports Winners Using Data Analytics with Pandas and Scikit-Learn." Performance by Robert Layton, *YouTube*, YouTube, 3 Aug. 2015, www.youtube.com/watch?v=k7hSD_-gWMw.
2) HERBINET, Corentin. "Predicting Football Results Using Machine Learning Techniques." *Imperial College London*, 2018, www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf.
3) Sheehan, David. "Predicting Football Results With Statistical Modelling." *dashee87.Github.io*, 4 June 2017, dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling/.