Machine Learning Engineer Nanodegree

Capstone Proposal

Raul Britto
May 16th, 2019

Proposal

Domain Background

Currently, sporting bets companies have become very popular with marketing campaigns on websites, TV channels, and disclosure sports groups. Hence, the betting volume and total value of bettings have increased with the years. This growing of sporting bets companies is based on the increase of data and statistics sporting, increased of sports transmission, publicity and the possibility of sports betting being made from home on the Internet.

Thus more and more people are encouraged to do sporting bets and follow sports that they love it. For example, It is estimated that Brazilian annually about R$ 2 billion in sports betting [1]. However, most parts of them did bets without using any prediction model or statistical model and these companies earn more and more money with amateur gamblers who betting as a hobby. According to a survey published by the Economist magazine, Brazilian bettors lost US $ 4.1 billion in 2014 on lottery and betting sites [2].

Problem Statement

The goal of this project is to create a sporting bet model to soccer matches, specifically in the Premier League [3], using data from the last ten seasons and statistics such as number of goals, corners and yellow cards for each team to predict if the match on analysis will be a draw or if will has a victorious team.

Datasets and Inputs

The most part of data were collected on Football Data website [11], where it is available to users details and data about soccer games of several championships and odds given by sportsbook. The files are separated by seasons and specifically were collected data from the season 2009/2010 to the season 2018/2019.

The data contained in these files are:

- *Away Team*;
- *HomeTeam*;
- *Referee*;
- Date of match (later separated into seasons with the *Season* column);
- Total goals in the first half for each team (*HTHomeGoals / Away*);
- Which team won in the first half (*HTResult*);
- Total goals in the match by team (*FTHomeGoals / Away*);
- Which team won in time the match (*FTResult*);
- Number of corners (*HomeTeamCorners/Away*);
- Number of shoots (*HomeShots/Away*);
- How many of these shoots were in the target (*HomeTeamShotsTarget/Away*);
- Number of yellow and red cards (*HomeTeamRedCards/Away* and *HomeTeamYellowCards/Away) ;*
- Number of fouls by team (*HomeTeamFouls/Away*);
- Quotas from betting websites 365 Bet [6] for win of each team or draw (*B365H, B365A, B365D*);

Other data were collected from the official website of the Premier League [3], saved and manipulated. These data were added to the database after a treatment, the average of each team was calculated from the relevant statistics. Below it is follows the description of the data and its respective attribute added to the base:

- Goals of each team per season (*MeanGoalsHome / Away*);
- Goals conceded from each team per season (*MeanGoalsConHome / Away*);
- Corners by each team per season (*MeanCornersHome / Away*);
- Yellow cards for each season (*MeanCardsHome / Away*);
- Shoots of each team per season (*MeanShotsHome / Away*).

Solution Statement

The goal of the work is to predict the winner of the match or if the game will have a draw. Because of this, classification algorithms will be used, predicting the classes: victory, draw and defeat. This allows the method to be compared with others that predict the same output or those that predict the teams' goal numbers, a value that can be categorized as proposed here.

Benchmark Model

How the 2018/2019 season has finished recently, the results of the last ten games will be used to compare the model proposed’s prediction with the predictions taken from the benchmarker, the website Footballpredictions.com [5]. The results given by the website will be compared with the result from regression and classification models, so all the predictions,

be they given by regression, classification or benchmarker output will transformed into a classification output to create a commun way to compare the results.

Evaluation Metrics

To evaluate model performance, one would classify match results into home wins, away wins and draws and then look at the number of matches that the model has correctly identified, using a standard classification matrix. There is unlikely to be a great degree of imbalance in the class values for the dataset, although given the commonly observed home advantage phenomenon, one is likely to see a slight skew in favor of home wins. In this case, classification accuracy is a reasonable measure of evaluation. In cases where the data is highly imbalanced, ROC curve evaluation may be more appropriate[7].

It is important to preserve the order of the training data for the sport prediction problem, so that upcoming matches are predicted based on past matches only. Using cross-validation generally mean shuffling the order of the instances and therefore is not an appropriate means of splitting the data into training and testing, for the sport result prediction problem. A held-out training test split is more appropriate, with the order of the instances being preserved.

Project Design

Above will be explained the project's workflow:

Data Collection

As mentioned the database was collected on Football Data website and from the official website of the Premier League. The data were manipulated and put together in files by season.

Data Preprocessing

The data will be treated together joining all data available. It will be evaluated which attributes will be manainted and which will be discarded. For instance: number of fouls by teams and odds given by other sportsbook, only it will be considering odds provides by 365bet.

After this, it will be calculated the correlation between the attributes selected and one-hot-encoding to the categorical attributes as teams name. It will be tried find outliers in the dataset as well.

Data Splitting

The data will be splitted as explained in the article [] to validate the model. Matches more recents will be inserted into the model one more time, because matches played ten years ago have less leverage in recents games and a team has loss of roster over the years.

Model training and evaluation

Some supervisioned learning algorithms will be tested and evaluated as Decision Tree, XGB Model, Neural Networks and Ensemble methods and finally the best algorithm will have its parameters optimized.

References

[1] Mercado de apostas esportivas movimenta R$ 2 bilhões no Brasil, 2018. https://www.terra.com.br/noticias/dino/mercado-de-apostas-esportivas-movimenta-r-2-bilhoes-no-brasil-segundo-pesquisa,5e91353bb264cfb927b0b93d8a94e1a397u8wbih.html

[2] Brasileiros perderam US$ 4,1 bilhões em sites de apostas e loterias, 2015. https://epocanegocios.globo.com/Informacao/Resultados/noticia/2015/09/brasileiros-perderam-us-41-bilhoes-em-sites-de-apostas-e-loterias.html

[3] Premier League Live Scores, Stats & Blog. www.premierleague.com/stats/top/clubs

[4] Your Football Betting Odds, Football Results, Free Bets, Football Scores Football Betting, Scores & Results Service. http://www.football-data.co.uk/

[5] Footballpredictions.com Premier League Predictions https://footballpredictions.com/footballpredictions/premierleaguepredictions/

[6] Bet 365. https://www.bet365.com/

[7] A machine learning framework for sport result prediction https://www.sciencedirect.com/science/article/pii/S2210832717301485