Machine Learning Engineer Nanodegree

Capstone Proposal

Philip Burroughs

March 23rd, 2019

Proposal: Classifying Football (Soccer) games to have over 2.5 Goals

Domain Background

The domain I have chosen is the European football leagues. Football can only exist because of its fans, without the huge fan base the beautiful game would suffer and not be the world-wide phenomenon it is today. Negative stories often plague the sport through the media and the Guardian newspaper once reported a dip in early season viewings. This is huge concern for me personally as I love the sport and have lived and breathed the unity it can bring to a nation. I am one of many who would tell you the best moment of their life came on July 11th, 2018 when Kieran Trippier scored ***that*** freekick vs Croatia in the world cup semi-finals. The world cup also brought with it a huge economic boost to the UK due to the unity & excitement of a nation who loves the sport.

To protect the sport and keep it as popular as it is today we need to keep bringing new fans into the sport as well as keeping the current fans as excited as ever. With an increased price to watching football games hitting the nation and TV wars over which companies will show what games it is becoming more essential than ever that the few games we do get see on the TV need to be exciting and take the viewers through the emotional storylines that football often delivers. The key way to do this is to show the games with the most goals. I can tell you for a fact when I heard Man City were beating Chelsea 4-0 after half an hour my TV was straight on and I made sure every friend I had knew what unimaginable storyline was unfolding in the north of England that Sunday afternoon. The aim of this project is to build a machine learning algorithm that can predict the higher scoring games then companies could use this to show the more exciting games to the public and ultimately help to make more people fall in love with the sport.

https://www.theguardian.com/football/2016/oct/24/sky-sports-bt-sport-people-switching-football-off
http://www.sportbible.com/football/news-reactions-remembering-when-kieran-trippier-scored-that-free-kick-vs-croatia-20190111
https://www.bbc.co.uk/news/business-44711254

Problem Statement

The problem I am trying to solve is to not watch low scoring games. With the average goals per game in top flight football landing between 2-3 goals I want to classify the games that will end with 3 or more goals in them. This easily measurable by taking the accuracy and the precision score of the models. The problem is replicable and occurs hundreds of times a week as there are constantly games being played and shown to the public.
https://www.planetfootball.com/in-depth/investigating-europes-four-major-leagues-exciting/

Dataset & Inputs

I pulled my data from http://www.football-data.co.uk/data.php

I took the CSV's of the "main leagues" from this website. I stored all of this in my own SQL server. In SQL I then made some adjustments to the data. I took away what wasn't needed, such as betting odds, yellow cards per game, the referee and other columns like this. In SQL I then applied a back test to the data and made some new columns which was the average goals the home teams had scored and conceded that season. I also applied the division name and the last number of games that ended under or over 2.5 goals for the home and away team.

This then leaves the data set with is attached with this file. "football_data.csv". The columns remaining that will be used to drive the models are: Division, Team Name, Home goals average goals scored & conceded: Away team average goals & conceded: The last 5 games home team games under and over 2.5 goals as well as the last 5 games away team games that ended under or over 2.5 goals. I believe these characteristics are relevant for the problem I am trying to solve as looking at historical scoring patterns and average goals scored per team should help to forecast the future.

Solution Statement

The solution to the problem will be to identify games that will end with over 2.5 goals. Accuracy and Precision can be used to score the model and the higher the precision of the model the better at identify high scoring games the model will be. We do not need to correctly predict every game as not every game is shown publicly, but when we do pick a game we want to be sure that the game will not disappoint so that is why the Precision score will be a good measure for this.

Benchmark Model

It is often hypothesised and well shown that football final scores follow a Poisson distribution when plotted. This has then led to many models being made, especially in the betting industry, to be built using the Poisson distribution to work out the probability of a game ending with X number of goals in. I am going to average the 4 characteristics (Home goals scored & conceded, Away goals scored & conceded) and then use the Poisson distribution equation to generate a probability. If this probability is >= 50% then the model will predict over 2.5 goals, if not it will then predict under 2.5 goals. This can then be scored used accuracy and precision just like the machine learning model will. They can then be compared.

https://en.wikipedia.org/wiki/Poisson_distribution

https://blog.annabet.com/soccer-goal-probabilities-poisson-vs-actual-distribution/

Evaluation Metrics

As mentioned previously, the Precision metric will be used as an evaluation metric to score the models. Precision is the number of true positives found out of all the number of points flagged as a positive. This is essentially the percentage of all the games highlighted as ending over 2.5 goals, how many did. The model with the highest Precision will be the best at picking games that end over 2.5 goals.

Project Design

Workflow:

1. Import the data into Python, create two data frames. One data frame just the top divisions, the other containing all the divisions. Maybe the model will learn better from all the divisions instead of just a select few. During this step conduct any further data cleansing.
2. Carry out some initial analysis and visualisations. Do all divisions have the same number of average goals? Does this change season per season? We need to understand what differences could affect the models.
3. Build the benchmark model, Poisson distribution, and apply it to the data sets and score the model. It will be interesting to see how the performance changes per division.
4. Prepare the data for entry into the machine learning algorithms. One-time hot code the categoric variables and normalise the numeric variables so they all lie between 0-1. I do not think this will make a huge difference as the football goals are all averages of integers that follow the same scale but its still better to apply this.
5. Train and compare the initial machine learning algorithms. If there are no huge restraints causing issues, such as training times, then start to tune the parameters of the models and keep track of improvements using the Precision score.
6. Settle on the best machine learning algorithm built and then compare the scores of this algorithm to that of the benchmark model.
7. Publish the results, Poisson distribution is still the best model we have for identifying games that will end over 2.5 goals ***OR*** machine learning algorithms can help to identify the more exciting games so publish the model and parameters used.

Seeing as I treat each season as different for the averages the data might have to be cut down further and only start looking at teams that have played 5 or more games that season so we start to get more meaningful averages. The classification algorithms I will be exploring are: SVM, Decision Tree, Random Forests and Logistic Regression. PCA analysis might be used as well seeing as we have a lot of variables in the data set. Also, the exploration of characteristic importance might be conducted to help trim down the data further without losing much from the overall score.

Poisson Distribution Curve Code:

1 - (avg_hg_scored ^ 2 * (2.71828 ^ (-1 * avg_hg_scored)) / (2!)) + (avg_hg_scored ^ 1 * (2.71828 ^ (-1 * avg_hg_scored)) / (1!)) + (avg_hg_scored ^ 0 * (2.71828 ^ (-1 * avg_hg_scored)) / (0!)) = probability of 3+ goals based off average home goals scored.

Repeat the above step 4 times for the 4 characteristics then average the probabilities to get the Poisson probability of a game ending with 3+ goals.