Football Leagues, Goals and Machine Learning

•••

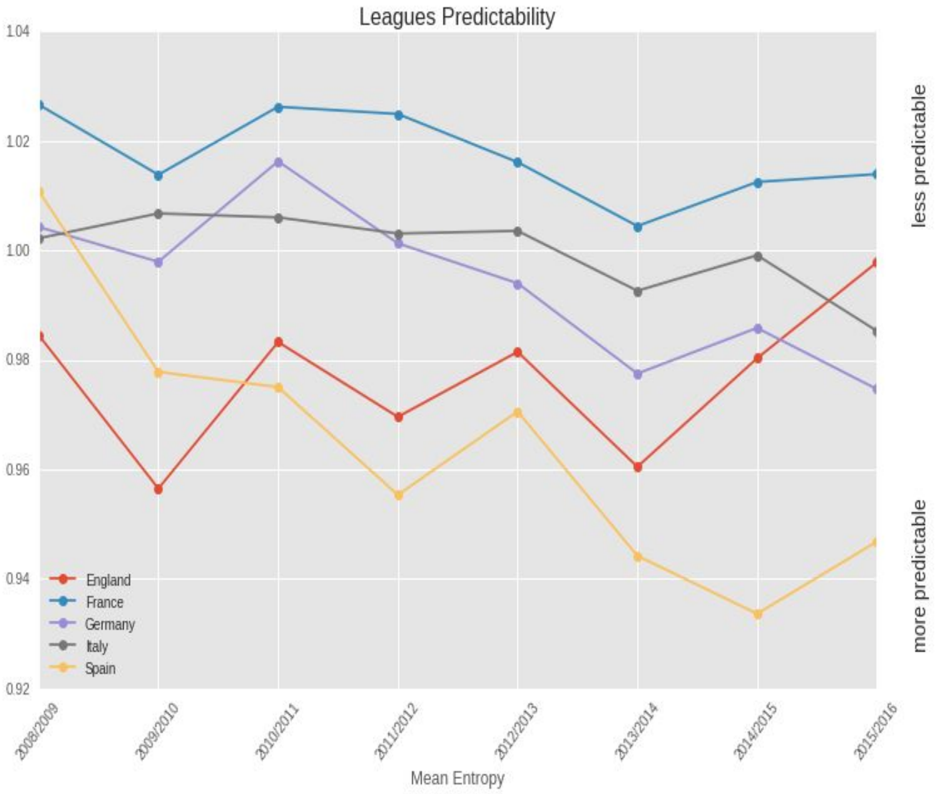
Capstone Project by Sibi Rajendran

Football Leagues

Each county has its own major football league - 20 teams in each league.

Most famous being the English Premier League (EPL) 5 billion viewers in 2012/13 season.

As a sport, it is highly unpredictable - Leicester City won the EPL against odds of 5000-1.



Project Goals (no pun intended)

Use available data to analyse ‘home advantage’ and goal trends across Europe.

Use available data to predict number of goals scored by each team in each fixture.

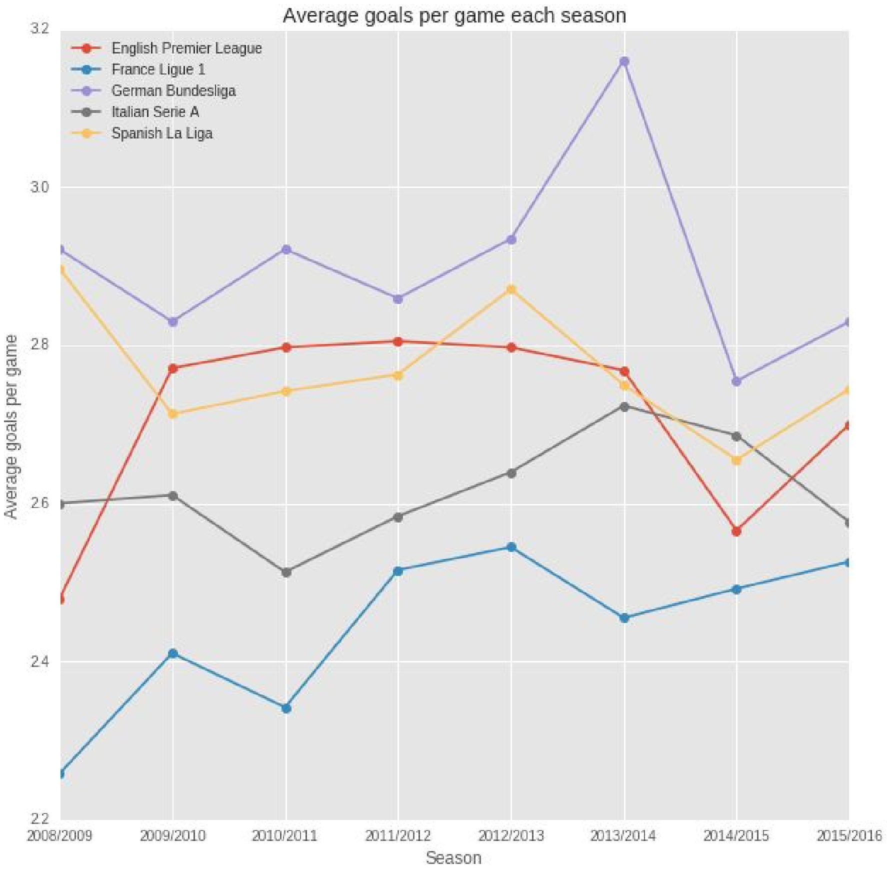
Develop features from historic data to predict the final outcome of a match.

Goal Trends

Teams in Germany and Spain have consistently scored a lot more goals than other leagues.

On an average, German league teams score 2 goals in each game!

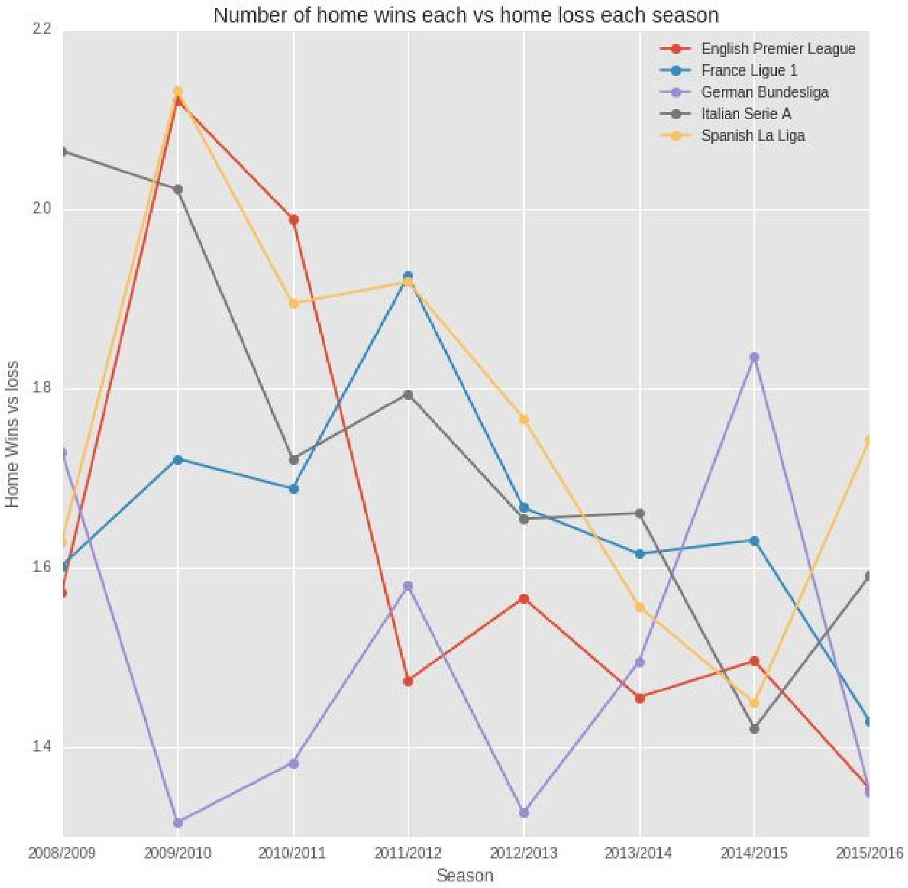
Direct correlation to each league's playing style (England - defensive, Spanish - attacking).



Home Advantage

Spanish teams again enjoying more advantage at home recently.

Steady decrease in home advantage at EPL in the past few seasons - another question answered.

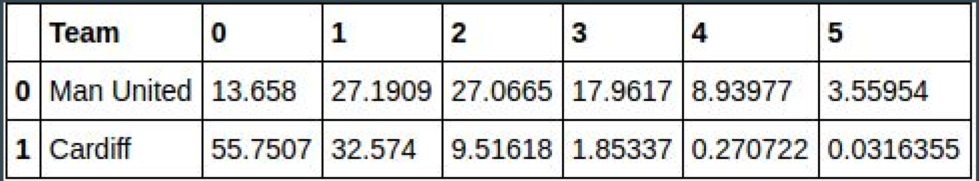


Predicting Scores

Use historic data to come up with each team's attacking strength and defensive strength - at home and away.

Subsequently, use Poisson Regression to calculate the probability of a team scoring n-goals in a particular fixture.

Example : Can be used to calculate scenarios like a team scoring more than 2 goals by simple arithmetic.



	Team	0	1	2	3	4	5
0	Man United	13.658	27.1909	27.0665	17.9617	8.93977	3.55954
1	Cardiff	55.7507	32.574	9.51618	1.85337	0.270722	0.0316355

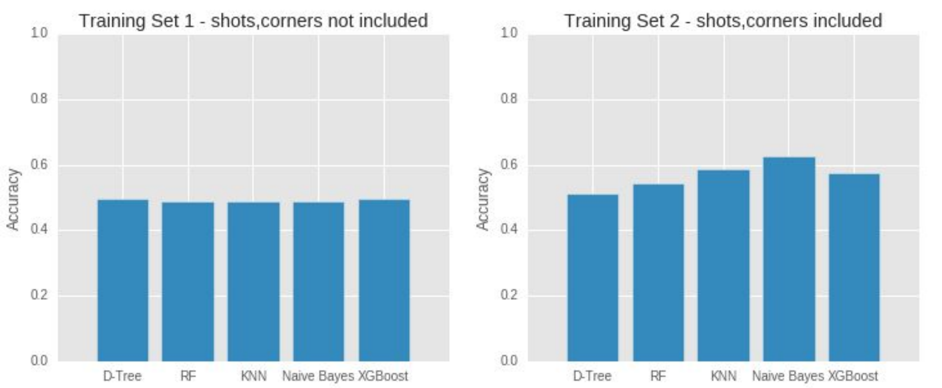
Advantages : Simple model, easy to use and predict
Disadvantages : Not very robust, doesn't consider several other factors.

Prediction of Outcomes - Classification Problem

There are three possible outcomes to a game : Home win, draw, away win.

Randomly predicting gives us an accuracy of 33% <- Benchmark.

Now, convert the problem to a multi class classification problem.
Use decision trees, KNN, XGB, Naive Bayes classifiers to model the data - attacking and defensive strength, corners, shots on target etc.



Feature Engineering

1. **N-recent performance**

Average of number of goals scored, corners taken, shots on target in n-recent games for both home and away teams.

This is a measure of performance or form in recent games.

2. **Home Advantage**

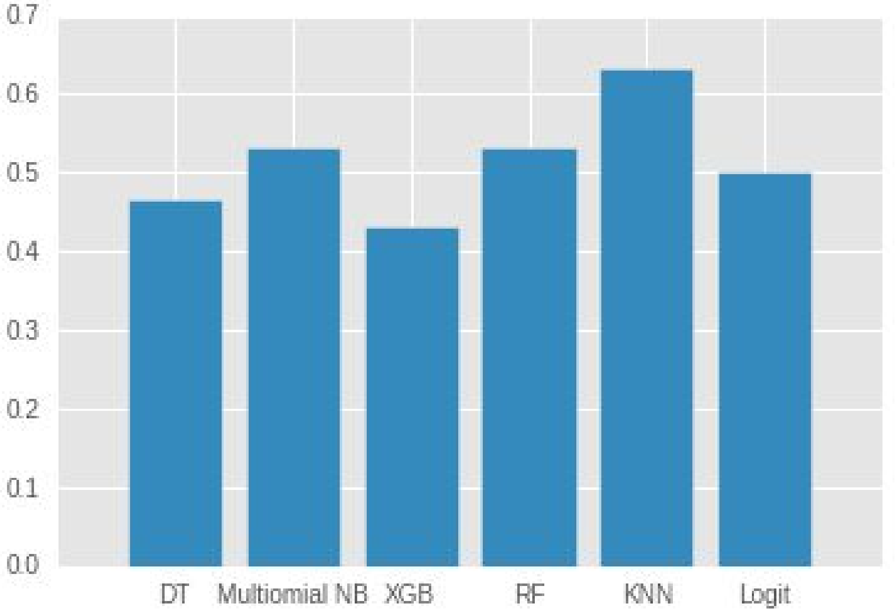
Difference of n-recent performance stats for home team and away team.
This order inherently captures home advantage along with other stats in the data.

Results

KNN performs the best with a accuracy of 63.3% after tuning and considering the 4 recent games in n-recent performance and n-neighbours = 13.

XGB can also be tuned to give a similar performance but too much fluctuation in accuracy for a slight change in n_estimators.

Including home advantage does not improve the model by much in the EPL but gives better results in the Spanish league.



Conclusions and Limitations

Mark Lawrenson, football pundit, had an accuracy of 55% in the data set we worked with. 60% using our model is well above the random benchmark and marginally better than the Mark Lawrenson.

Data doesn’t include several other factors and statistics - not freely available for use. OPTA - a company that tracks such info but need to pay to obtain data.

More feature engineering on obtaining such data would lead to a more accurate classifier.

THANK YOU