Football Leagues, Goals and Machine Learning

Capstone Project by Sibi Rajendran

INTRODUCTION

Every country in Europe has a club based football league. Each league has 20 teams which play two matches with every other team in their league - one at their home stadium and the opponent's home stadium. Each such match has three possible outcomes - the home team wins, the match ends in a draw or the visiting team wins.

Among these leagues, the English Premier League is the most popular one, being watched by an estimated figure of 5 billion people across the globe. In a season, since each team plays two games against every other team, there are a total of 380 games.

Given such a format, it is natural that there are several online fantasy leagues, betting agencies and pundits who try to predict the outcome of each match. In this project, an attempt has been made to find out the factors that affect the outcome of a match and also to predict the results of any fixture by using these factors.

DATA

The data has been obtained from two sources.

1. Hugo Mathien's European Soccer Database in Kaggle https://www.kaggle.com/hugomathien/soccer
2. Football-data.co.uk http://www.football-data.co.uk/data.php

The dataset from Kaggle was obtained as a SQL database and it contained data for all leagues from the 2008/09 season till 2015/16. The second data set contained csv files that had details about each match as separate observations. The Kaggle dataset was primarily used for the data story as it made comparison between different leagues easier and the other dataset was used for further analysis of the English Premier League as it contained more in game features (corners, shots on target etc.).

Data Dictionary

European Soccer Database

Leagues Dataset : Obtained by merging leagues table and the country table

league_id	Unique id for each league
league_name	Name of the league
country_name	Name of the country
country_id	Unique id for each country

Matches Dataset :

id	Unique id for each match
league_id	Unique id for the league
home_team_api_id	Id for the home team
home_team_long_name	Name of the home team
away_team_api_id	Id for the away team
away_team_long_name	Name of the away team
home_team_goal	Number of goals scored by the home team
away_team_goal	Number of goals scored by the away team
season	The season in which this match happened
total_goals	Total number of goals scored

Football-Data : Detailed description about each column in the second data set is given in the notes.txt file in the data folder along with the project.

As an overview, some of the columns are :

HomeTeam : Name of the team playing home
AwayTeam : Name of the visiting team
FTR : Full Time Result
FTHG : Full Time Home Goal
FTAG : Full Time Away Goal
HC : Number of corners taken by the home team
AC : Number of corners taken by the away team

TOOLS USED

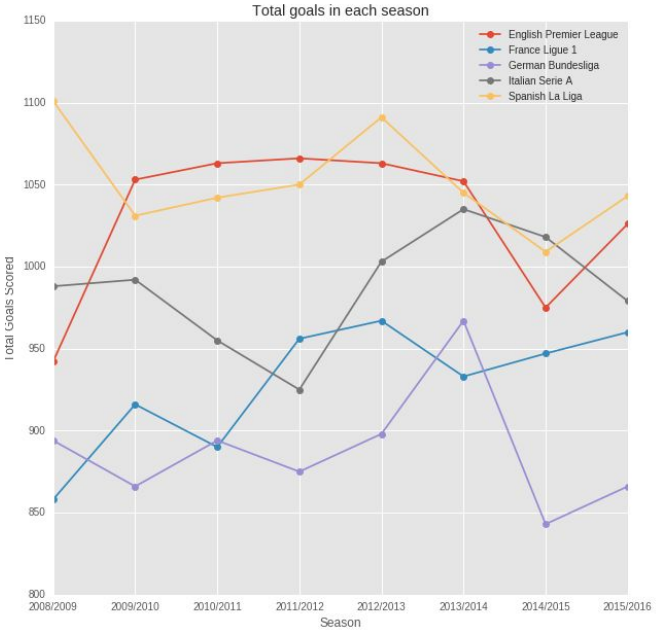
Pandas : Loading the data, data wrangling and manipulation, feature engineering.
Scikitlearn : Libraries for classifiers, model evaluation, metrics, cross-validation
Matplotlib and Seaborn : Data visualization

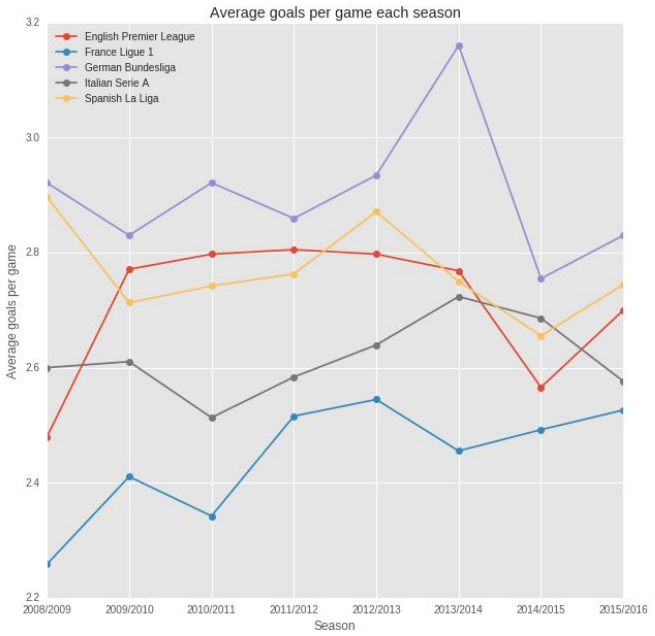
DATA WRANGLING

Using sqlite3, the European Soccer Database was imported and the required information (countries, matches, leagues, teams) was read using panda. The database contains information about matches that have already occurred and hence there were no missing values in any of the columns. Furthermore, to analyse statistics about each league separately, we subset them into different countries. Also, a result column is added to the dataframe; convention is that 1 represents a home win, 0 represents a draw and a -1 represents an away

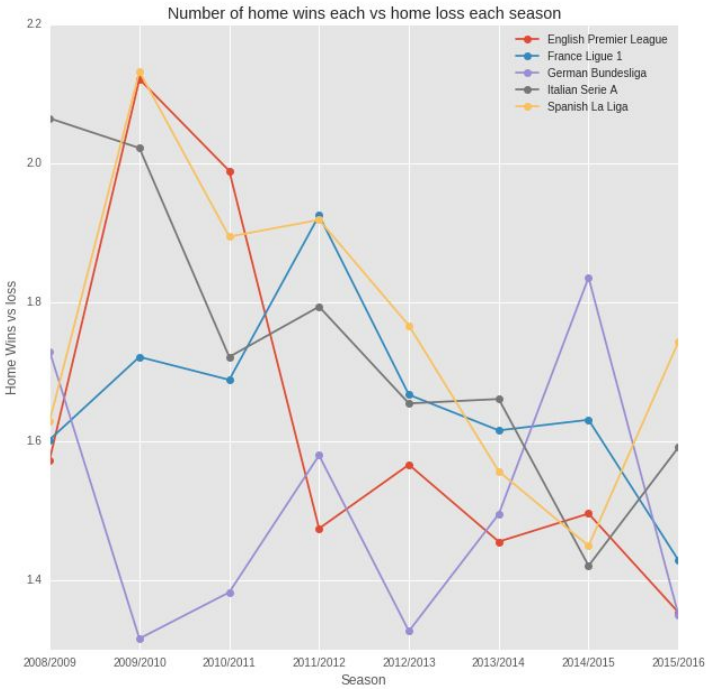
EXPLORATORY ANALYSIS

To start off, the number of goals scored each season by all teams put together in league is plotted. This would give a fair sense of the playing style of a league. Also, a graph depicting the average number of goals per match was plotted. It was seen that teams in the Spanish League consistently scored more goals than teams in other leagues. Also, since the total number of goals might not be a good indicator of the league as teams in the German league play considerably fewer matches, the average number of goals per game is a better indication. In the second graph plotting this metric, we see that the German league teams score a minimum of two goals in each game. Notably, in the 2013/14 season, teams in the German league scored more than 3 goals on an average.





INVESTIGATION HOME TEAM ADVANTAGE

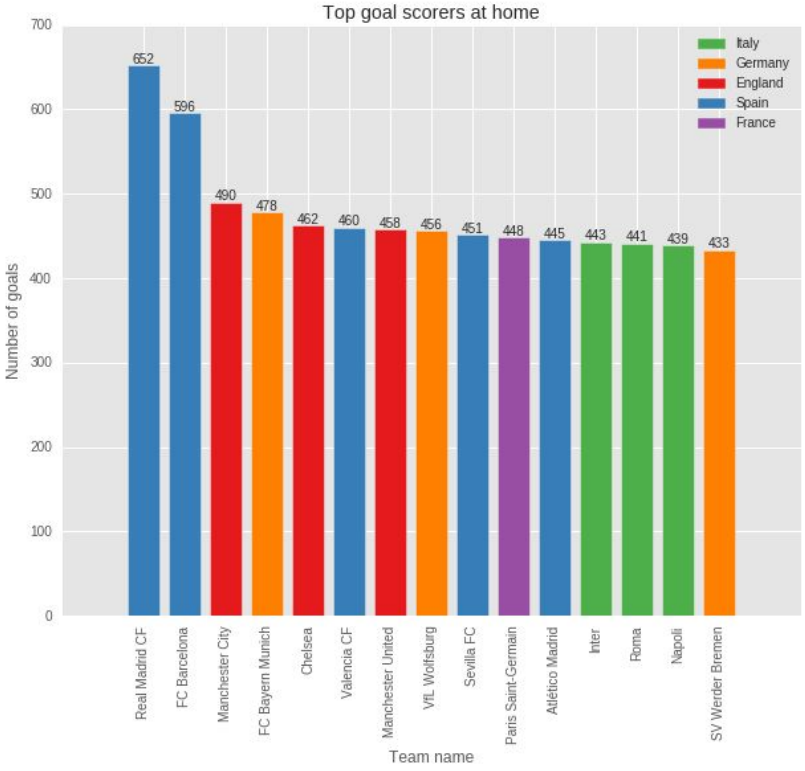


The question that I really wanted to answer with the data was to check the veracity of the famed 'Home Advantage'. I wished to validate this theory by simple analysis. Also, I had this hunch that, in recent times, teams in the English Premier League were losing this home advantage. Both of these questions could be answered by plotting graphs that show the trends in the number of goals and number of wins at home against playing away.

By plotting this graph, it was clearly seen that the ratio between home wins and home loss was steadily decreasing in the English Premier League. This confirms that the effect of playing home is decreasing in EPL.

To sum up this analysis, the strongest teams in Europe based on their home advantage was also plotted.

The Spanish dominance was again seen in this graph as 5 in the top 15 teams that have the most advantage at home were Spanish. The two giants, Real Madrid and Barcelona enjoyed a far greater home advantage than their closest rival Manchester City when plotting the number of goals each team scored at home.



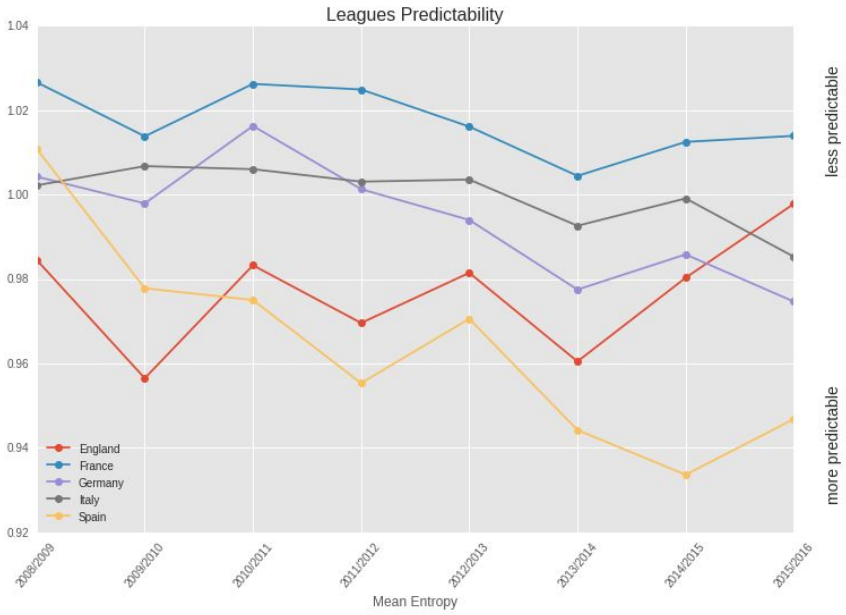
PREDICTABILITY ACROSS LEAGUES

It is well known that football is a highly unpredictable sport so much so that in the 2015/16 season, Leicester City won the EPL beating odds of 5000/1. But, really, how unpredictable is it anyway? For this, we turn to the second dataset that contains odds of a home win, draw and away win for each match given by Bet365, a popular betting agency. I believe that these odds capture information that cannot be quantified - information such as a general sentiment towards a team's win, arrival of a new star player, change of manager etc.

We convert these odds to probabilities and calculate its entropy (a measure of randomness used in information theory). We thus try to gauge the randomness in the occurrence of the three outcomes.

From the graph, it can be inferred that on the whole, the English Premier league with an entropy of 1 is quite unpredictable and the Spanish league has been the most predictable in recent seasons.

In the notebook, an analysis of each team's predictability has also been carried out. It was observed that star teams like Barcelona, Real Madrid, PSG were much more predictable than the other teams.



PREDICTING SCORES USING POISSON REGRESSION

As I am follow the EPL closely, I chose this league for further analysis and prediction. The goal is to predict the score of new fixture using historic data, viz. Previous season's data. After reviewing the existing literature, I tried predicting the scores using Poisson Regression.

For this, a few statistics were calculated for each team from the previous season's data. Specifically, each team's attacking strength, defensive strength at home and away were calculated.

In any season, since each team plays 19 games at home,

Attacking strength at home (HAS) = (Goals scored at home / 19) / Average Number of goals at home
Defensive strength at home (HDS) = (Goals conceded at home / 19) / Average Number of goals conceded at home

Using these, we could now calculate the average number of goals a team would score in a given fixture.

Average number of goals scored by the home team =
Home team's home attacking strength * Away team's away defensive strength * Avg goals that season

Average number of goals scored by the away team =
Away team's away attacking strength * Home team's home defensive strength * Avg goals conceded that season

These two metrics give us the average number of goals scored by the two teams in that particular match. Using the Poisson formula, we could calculate the probability of a team scoring any number of goals.

For example,

	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

To calculate the probability that the expected score is 2-2, we simply multiply the probability that team_1 scores 2 goals and team_2 scores 2 goals. In this case, it comes out to 2.57%.

Similarly, if we want to calculate the possibility of a draw, we calculate the probability of each draw first (0-0, 1-1, 2-2) etc. and add them all together.

Straightaway, such an analysis lends itself to efficient betting. There are different kinds like home win, draw, away win, over 2.5 goals, under 2.5 goals etc. We could calculate the probabilities of each of this happening through Poisson Regression.

Performing this analysis for different leagues, it was observed that for a few teams (the strong ones), the expected goal metric gives accurate prediction of results. However, for other teams, it keeps wavering due to the inherent unpredictable nature of football.

PREDICTION AS A MULTICLASS CLASSIFICATION PROBLEM

Instead of trying to predict the scores in a given fixture, we can try and predict the final outcome - home win, draw or a home loss. This becomes a multi class classification problem and we can try out different machine learning methods to model the data that we have. For a multi class classification problem, the classifiers that work are decision trees, random forests, gradient boosted trees, KNN classifier and naive bayes.

Since there are three possible results to any given fixture, with a Random Benchmark Model where we randomly pick one of the three outcomes as the final result, we will end up with an overall accuracy of 33%. On careful feature engineering, we will try to improve the model to give a much better accuracy.

From the second dataset, we extract the features that we need. This includes the goals scored by both the teams at full time, number of corners, number of shots on target. The label is the full time result which is H, D or A denoting a home win, draw or away win. We transform this into 1, 0 and -1 for the scikitlearn classifiers to work.

Initially, we add just the teams' attacking strength, defensive strength as features. In the second try, we also add the number of shots on target and the number of corners taken by both teams. This might be a good indication of teams scoring more goals and winning the game. We see an increase in the accuracy of the classifier after these new features are added into the model.

While it may be apparent that more the number of shots on goal, more is the chance that a team might win, it isn't the case with number of corners as well. But this shows that teams getting more corners partly had more wins than the rest. In reality, corners only lead to better creation of chances - a direction relation to the number of shots on target and hence, goals. However, since we do not have free access to data concerning the number of chances created or other in game features, number of corners prove to be a good enough metric for our analysis.

FEATURE ENGINEERING

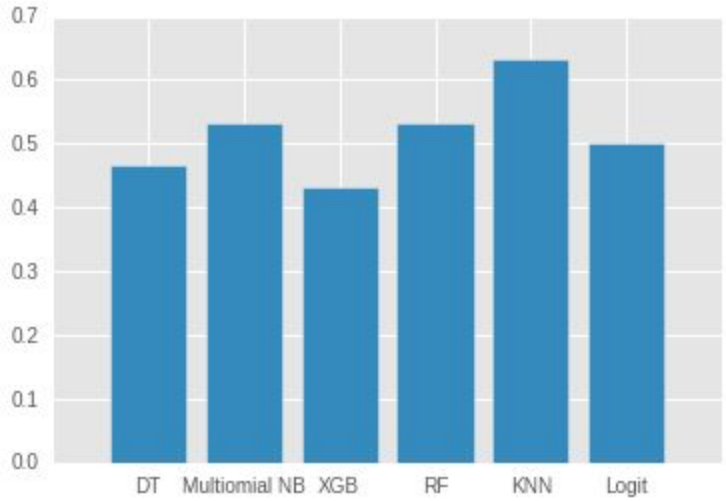
A team's form

Intuitively, a team that has been consistently performing well in the past few games tends to perform well in the next game as well. This can be attributed to hidden factors such as a good chemistry between the players, high morale in the team etc. To model this feature, we introduce the concept of 'n-recent stats.

N-recent stats : In these features, we include the average number of shots, corners and goals in the last n-matches. We include these statistics for both the home and away teams in any fixture.

pastHS = average number shots by the home team in the last n-games
pastHC = average number corners by the home team in the last n-games
pastHG = average number goals by the home team in the last n-games

Using these features for both the home and away team and combining them with the already existing attacking and defensive strength features, we get our training dataset.



Result : K-Nearest Neighbors classifier with n-neighbors set to 13 performed the best with a prediction accuracy of 0.60 and n-games set to 4.

On tuning the XGBoost classifier, we were also able to achieve an accuracy of around 60%.

Note that although we were able to use cross validation in the initial formulation of the problem, we cannot use cross validation to test these results as matches occur sequentially and the n-recent stats depends on the previous n-games. Hence, it is meaningless to predict games that are already over using stats from succeeding matches!

To avoid this problem, we varied the number of games in the training data and testing data to calculate how well the model performs.

Home Advantage

The other factor that always comes into play when trying to predict the outcome of a match is which of the two teams is playing home. Instead of explicitly creating a separate feature to contain this information, we take the difference between the the n-recent performance stats of the home team and the the n-recent performance stats of the away team. This order inherently stores the information about which team is playing home and which team is playing away.

Similarly, we were able to achieve 60% accuracy with the KNN classifier and there did not seem to be a huge difference after adding this feature for the English Premier League. However, an increase was noted for the Spanish league - indicating that there is a marked home advantage in the Spanish league.

CONCLUSION

Pinnacle Sports ran a competition with Mark Lawrenson, a football pundit for the exact season that predicted in the analysis above. Mark Lawrenson had a prediction of around 55% and Pinnacle did marginally better The model we built with 60% does slightly better than these two and way better than the Random Benchmark Model (33% accuracy) which is where started the process.

In conclusion, using only historic data for a season, we can use the above classifiers (KNN or XGB) and predict the results of the matches in an upcoming week with an accuracy of around 60%. Personally, I tried this out for the 2015/16 season's boxing day matches and correctly predicted 7/10 matches in that week!

Finally, with more information about each match like the number of attacking moves, number of through passes (Packing) etc., we will be better poised to build a more accurate model.

LIMITATIONS AND FURTHER RESEARCH

Any match is highly unpredictable and there are several other factors that come into play during a game. Players' form, strategic nuances in formation, player injuries and fatigue level etc constitute a major part of a team's performance in a match. Data about such characteristics is not freely available. Including these features would help us get a better understanding of how to more accurately model the data available and also which factors contribute more to a team's victory.

Further research and analysis could be carried on by applying these models in other leagues and seasons. Also, a weighting scheme could be be developed for the Poisson Regression to incorporate the features engineering in the latter part of the project.