# Football Outcome Prediction

## 1. Approach to the dataset

The dataset contains records corresponding to matches played between 2009 to 2017 by 5 leagues: Bundesliga, Ligue 1, La Liga, Serie A and Premier League. Matches are played in seasons and each season is from August to May (next year). Since the possible outcomes are 3 (H, A & D), it is a multi-class classification problem.

For building football outcome prediction algorithm, I have used only 2 seasons of data for training: 2015-16, 2016-17. This is because data which is more than 2years old may not be a real representative as team players/ coaches could have changed. If we had player information, it was possible to validate this. I have taken an assumption that for 2years, team structure remains almost the same.

I had tried to train model using only the latest season records also. It gave marginally lesser accuracy compared to using 2 seasons data. Hence, I have used 2 seasons data for the final submission.

Another approach I tried was to build a separate model per league as the teams are specific to individual leagues. Hence, combining all the data of all leagues may not be a good idea. But this approach did not yield good accuracy and hence did not use it in final submission.

## 2. Choice of the model(s).

Models considered are linear models: Logistic Regression (multinomial), Ridge Regression, KNN, SVC and Random Forest Classifier.

## 3. Model validation approach (used locally).

All the models were cross validated in train data. Logistic Regression gave the best result in that. Also, it gave the best score in test accuracy, precision and recall compared to all other models.

I could achieve only 68% accuracy even after using Logistic Regression, but it could improve if I can add more external features.

## 4. Choice of features used for the model creation.

Other than all the features provided in the train dataset, I have used some external data as well. From http://www.football-data.co.uk I have taken betting odds data and HTR(Half Time Result). Also, I have created two reference files using data available online. They are:

- **List_Of_Champions.csv:** Season and the defending champion
- **Team_Standings.csv:** Season and team standings (at the end of the season)

For every record, a lookup is done on List_Of_Champions.csv to get the defending champion (last season's winner). This is used to create 2 new features: Is Home Team the defending champion, Is Away team the defending champion. Similarly, for every record, a lookup is done on Team_Standings.csv to get the current rank of the Home and Away teams.

Using past seasons data in train, 6 new features are added: **Last_yr_H, Last_yr_A, Last_yr_D, Past_2yrs_H, Past_2yrs_A, Past_2yrs_D**. These features contain the total number of H, A and D results achieved when the same Home Team and Away team played in the last season and the last 2 seasons. For eg:, since the training data used is 2015-16 and 2016-17, features are calculated as

- **Last_yr_H, Last_yr_A, Last_yr_D**: Using 2015-16 data
- **Past_2yrs_H, Past_2yrs_A, Past_2yrs_D**: Using 2014-15 and 2015-16 data