Sahaj software | inspiring brilliance

Guidelines

At Sahaj, we strive to build high-quality software that has strong aesthetics (is readable and maintainable), has an extensive safety net to safeguard quality, handles errors gracefully and works as expected, without breaking down.

We are looking for people with data science knowledge coupled with pragmatism to deliver the implementation to business production environments. The data scientist should understand the domain and build models that have the ability to deal with the real-life constraints.

Following is a list of things to keep in mind, before you submit your solution, to ensure that your model focuses on attributes, we are looking for -

- Have you understood all the variables in the data?
- Have you followed best practices to make sure your model is robust?
- Have you thought about how the model would evolve, over a period of time, in production, when more data is available?
- Have you made an effort to make your code readable and robust?
- Have you thought about the biases that can be there in the data?

There are numerous ML models with their strengths and weakness. Trying different models is not the most important thing in the data science application. Significant value is derived when the time is spent on selecting good data and features than models.

Problem Statement

Your goal is to build a model to predict the outcome of a football match, given data for the past 9 years. All the football matches from 2009 to 2017 are covered in the dataset.

Based on the dataset provided, your goal should be to come up with an optimal solution to predict if a Home Team would win or lose or draw a game (column name *FTR*) for the year of 2017-18.

We are looking to compare multiple approaches and choose the one that performs the best.

Use of external data sources is encouraged, with some recommendation. Of course, the actual results of these matches can be easily downloaded from the web. However, this problem statement is intended to be for fun and learning. You can choose to enrich the dataset by using other publicly available European football league datasets, e.g. http://www.football-data.co.uk/ or http://football-data.mx-api.enetscores.com/

The train and the test dataset is not randomly sampled for testing and training.

About the Data

The data is collected from http://www.football-data.co.uk/ and consists of different leagues.

Column Details

Name	Description
HomeTeam	Home Team
AwayTeam	Away Team
FTR	Full-Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals

HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
AC	Away Team Corners
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HC	Home Team Corners
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards
Date	On which day the match was played
league	Under which league the match was played

Expectations from the submission

1. Data Analysis showing your discovery of the dataset.
2. Choice of the model(s).
3. Model validation framework that ensures the model can work for future years.
4. Feature engineering and feature selection framework.
5. Please ensure your model is stable - ensure you have done cross-validation.
6. An ipython notebook or python executable file of the data analysis.
7. An ipython notebook or python executable file to generate the model you have chosen.
8. A CSV containing predictions for each row in the *test.csv*