Student: 105502506 106502549
Professor: Hung-Hsuan Chen
Course: CE6143-Introduction to Data Science
Date: 2020.1.7

Final Project: Football Game Result Predictor

Being a football fan growing up, we always have that one team we support the most and wish it win every single game. Most of the football-themed video games rate teams, players, coaches and strategies. We would play these games as many times as possible, only looking forward to beat the opponent one time. However, the reality seldom lives up to the expectation.

With the opportunity to develop a Data Science based project, we decided to create a game result predictor. In order to get enough data to train for a model, we choose a nice collected database from Kaggle coming in contents shown below:

- Database: https://www.kaggle.com/hugomathien/soccer
- Country 11 x 2
- League 11 x 3
- Match 26.0k x 115
- Player 11.1k x 7
- Player_Attributes 184k x 42
- Team 299 x 5
- Team_Attributes 1458 x 25

Data frame Example:

id	country_name	league_name	season	date	home_team	away_team	home_team_goal	away_team_goal
4769	France	France Ligue 1	2008/2009	2008-08-09 00:00:00	AJ Auxerre	FC Nantes	2	1
4770	France	France Ligue 1	2008/2009	2008-08-09 00:00:00	Girondins de Bordeaux	SM Caen	2	1
4771	France	France Ligue 1	2008/2009	2008-08-09 00:00:00	Le Havre AC	OGC Nice	1	0
4772	France	France Ligue 1	2008/2009	2008-08-09 00:00:00	Le Mans FC	FC Lorient	0	1
4774	France	France Ligue 1	2008/2009	2008-08-09 00:00:00	AS Monaco	Paris Saint-Germain	1	0

After calculating, we use every team's score records of which it's home team or away to predict result. "Home Home Average Goals" means the average home goals of the home team in this match, so on and so forth.

AwayTeam	Away Team Goal	Home Team Goal	HomeTeam	Home Home Average Goals	Away Home Average Goals	Home Away Average Goals	Away Away Average Goals
FC Nantes	1	2	AJ Auxerre	0.947368	1.526316	0.631579	0.631579
SM Caen	1	2	Girondins de Bordeaux	1.894737	1.684211	0.631579	1.052632
OGC Nice	0	1	Le Havre AC	0.894737	1.157895	1.789474	0.947368
FC Lorient	1	0	Le Mans FC	0.947368	1.421053	1.105263	1.368421
Paris Saint-Germain	0	1	AS Monaco	1.368421	1.105263	1.210526	1.052632

We use the three training methods we learned this semester to train models: KNN, Gaussian and Logistic Regression. The accuracies are 71.48%, 64.27% and 65.29%. But of course, these are just the simple models we trained for final project and practicing. The actual accuracies could be way lower and way less predictable, while there are more factors (like players' ability, what strategy used, …) need to be considered into the predictions.

References:

https://www.kaggle.com/c/football-data-challenge
http://odds.football-data.co.uk/
https://www.kaggle.com/airback/match-outcome-prediction-in-football
https://towardsdatascience.com/pandas-for-football-analysis-42c23b252995