English Premier League – Football Prediction

Term Project – CS677- Wasupol Tungsakultong

**Motivation and Objectives**

Football is the most famous sport in the world and has always been followed by large number of people. In recent years, there are many types of data in games and it always rapidly change over the time like play-by-play data collection.

In the role of Data Science, the football industry has many possibility of uses and applications

* Match strategy, tactics
* Player’s performance and styles
* Training
* Injury prediction
* Performance management and prediction
* Many more

Data Science in football can evaluate a team’s performance in games and predict the result of future games based on data.

In this project, will mainly focus on match played and goal scores and will predict for this season (2020/2021) who will be the champion.

**Data Preparation and Explorer**

Finding reliable and quality football data in public is quite challenging because the well-known football data providers do not make data available publicly and most of the public data was not good enough to make models for prediction.

I found the best data in <https://www.kaggle.com/saife245/english-premier-league> which contained interesting points like bets odds and match results in both half-time and full-time. So, I decided to use previous 6 seasons as training data and predicting the current season.

For current season fixture based on <https://fixturedownload.com/results/epl-2020>

There are some Data Definition and parameters that need to made in this project.

* Define Each Team names to unique digits like {“Arsenal”: “1”, “Man United”: 2}
* Define Home wins, away wins or Draw as [1,2,3], If match is not started, it will be [0]
* Calculate the Home wins percentage, away wins percentage for each team
* Game score used for predicting the winner

**Choice of Models**

I would like to compare between all models that I learnt from the class to make a models to predict this season matches and find the champion.

**Data Features**

* FTHG = Full Time Home Team Goals
* FTAG = Full Time Away Team Goals
* FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)
* Team Digits
* Home wins percentage
* Away wins percentage

Match prediction model

* First calculate all team’s win in home and away and find percentage.
* Convert Team name to digits.
* Train regression models from data feature and outputs as Home Goal Score and Away Goal score prediction in each match.
* In meantime, I also train regression models to predict match outcome.
* Find the winner from each match from prediction scores.

Model choices

* Logistic Regression
* k-Nearest-Neighbors
* Gaussian Naive Bayes
* Random Forest Classifier
* Decision Tree Classifier
* SVM Regressor with Linear kernel
* LDA
* QDA

**Implementation**

* Calculate win percentage in both Home and Away team

A screenshot of a computer

Description automatically generated with medium confidenceText

Description automatically generated

* Apply Winner Team and Winner Label

A screenshot of a computer

Description automatically generated with medium confidence

* Apply Team name digit

Text

Description automatically generated

* Predict match scores outcome

Text

Description automatically generated

**Results**

Match winner’s outcome, We can see that we obtain the best accuracy by using SVM model in linear, which we will choose in order to build our winner classification models

![Chart, bar chart

Description automatically generated]()

Match Home score and Match Away score (Goal)

![Chart, bar chart

Description automatically generated]()![Chart, bar chart

Description automatically generated]()

We obtain the highest accuracy from Random Forest for Home Goals while SVM for match score for Away.

Use prediction results to find the champion

From the result of the modes, I decided to use Random Forest for Home Goals and Away goals then calculate the winner by number of wins with the fixture of current season.

Background pattern

Description automatically generated

From prediction, Man City still be the champion of this season, which is very close to currently season performance because Man City is leading Man United by 8 points.

**Evaluation**

**The Root Mean Squared Error** is a metric that, compared to the Mean Absolute Error, is not interpretable. However, the RMSE gives a larger weight to predictions that are far away from the actual value

![Chart, bar chart

Description automatically generated]() ![Chart, bar chart

Description automatically generated]()

As you can see that in both prediction of Home and Away goals, the lowest RMSE is Random Forest which I choose for champion prediction.

**Conclusion**

* Finding data is very hard and challenges
* There are more parameters that can be assign from better predicting models
  + Number of Goal score win (such as win by 2 goals or more)
* There still are many factors that need to be consider in real football game
  + Players centric stats
  + Games stats, like corners, shots, tackles, expected goals and etc.
  + And Many more.