**Project Report – by Tien Dinh**

1. Dataset

My dataset is called: European Soccer Database 25k+ matches, players & teams attributes for European Professional Football. The dataset is from Kaggle.com:

<https://www.kaggle.com/datasets/hugomathien/soccer>

This is the EA sport FIFA soccer database for data analysis for machine learning. The creator’s name is HUGO MATHIEN. We don’t know much about the guy’s history, but we do know that he is a dedicated soccer fan and a hardworking database engineer. There is no motivation listed in the description, but I believe this data was created as historical record of European soccer team/player/matches, either out of curiosity or true passion. And because of it, I believe this record project is no funded. The original data source was collected from multiple trusted soccer website, including FIFA’s contribution themselves. To list a few:

* <http://football-data.mx-api.enetscores.com/> : scores, lineup, team formation and events
* <http://sofifa.com/> : players and teams attributes from EA Sports FIFA games. FIFA series and all FIFA assets property of EA Sports.

They offer 25,000 matches, over 10,000 real players and top 11 European Countries with their lead championship league. Within the league, they offer starting line-up, record of matches events. There are also players, players attributes, team and team attributes for each individual player and team. The two player tables include stats such as rating, strong\_foot, crossing, defensive\_work\_rate, … The two team tables feature team name/id, buildUp, chanceCreation, defenseTeamWidth, … The data will be useful for evaluating player’s rating, team’s strength, team’s winning chance and more.

From the evidence gathered above, from my opinion, the table is used to keep a true historical record of European soccer matches, players and teams in the past. All the data is useful for different purposes. Soccer enthusiasts can look up the database and check how much their favorite team won in the past, by how many goals, starting line-up. Teams and coaches can look up player’s rating and decide who to pursue in the future. There is also an interesting description of betting odds on the website, however this table is taken down due to some rule violations. Overall, this data is very useful for keeping track of European soccer in the past.

The data is distributed publicly on Kaggle website (linked on top) with Open Data Commons Open Database License (ODbL) v1.0. The author started collecting data from 2008 until 2016 (most recent update is 16th Oct 2016). There is no restriction on usage, and it is solely dependent on the users. Reading the history of data, it was well-maintained and active from 2008 until 2016. After 2016, the author stopped working on the project, but the data is still opened to public until today.

The data has a historical impact on European soccer from the past till 2016. It keeps true record of matches, players and team for users who wanted to investigate in soccer. I believe this is the best soccer data set systematically collected on Kaggle (The data were given golden with 3000+ upvotes). Unfortunately, they don’t have recent 2022 update, but 2016 is a close timeline where the peak of current football generation was.

2. Summary of data

3. Question, analysis, visualization results, model, discussion.

4. Impact

My initial idea is to base on the line-up, league, match events and betting odds from 2008 to 2016, make a prediction of who is the most likely to be the champion in English Premier League in 2017, then compare it with the real 2017 record. I would like to use the same data to make predict who will be wining this year 2020 English Premier League, but the result will likely be inaccurate due to a gap in data.

Some hypotheses:

* The player stats will most likely affect the team performance (goal/match, assist/match, hour played, number of time that they’re in the starting line-up)
* The team history will also affect the team performance (league title, standings last few years, ..)

Some questions:

* Which of the quality in a player that affect the team performance the most? Example: Central Back will be number of tackles, number of hours played, but Central Forward will be goal scored, goal assisted. This will be different across positions. We need to rank the factor based on the importance, the significance of the contribution in the model.
* How do I address the team history with the team performance? Will that be the case where the winning team last year should also be wining this year?
* Other factors, like betting odd (this one is new). Do these factors also contribute to the winning chance of a team? Higher bet is favorable to win but sometimes the underdog can make miracle happen. Like 2016 with Leicester City winning the Premier League out of nowhere. How can we investigate this scientifically?

4. Models

This problem that I investigate is a predicting win/lose odd for a soccer team in a league. This will be categorized as supervised machine learning model on the training data and classify the probability as either Won or Lost. On top of my head, I think I can use:

* Naïve Bayes: a classification algorithm based on Bayes Theorem, that determines the probability of winning or losing an Opportunity given that each predictor variable has taken on a certain value.
* Logistic Regression: A generalized linear model but with a binary outcome, which is transformed into a probability using the sigmoid function. The ‘weight’ for each predictor variable is determined by the model in order to reduce the error between the actual and predicted values.
* Extreme Gradient Boost: Builds an ensemble of Decision Trees in a sequential manner, where the residuals of each model are fit in the subsequent model.

This will be susceptible to change during working with the data. Naïve Bayes might be simple but can work best out of the three. There might also be more models that I can test in the future that is not included here.

Conclusion:

My passion is soccer (European football). I would like to use data science to investigate the matches, teams, results systematically to give me a guess of who will be winning the next league title. Right now, my goal is to develop a model that works for year 2017-2018 since we only have data from 2008-2016. This will be a supervised learning task because we have history and real data in comparison. Later, if possible, I would like to predict this year’s 2022 wining champion but lack of 2017-2021 data might impact the result.

> test2 <- testMergePlayer

> View(test2)

> View(test2)

> test2$date<- substr(test2$date, 1, 4)

> test2 <- test2[test2$date >= "2015" ]

Error in `[.data.frame`(test2, test2$date >= "2015") :

undefined columns selected

> test2 <- test2[test2$date == "2016" | test2$date == "2015" ]

Error in `[.data.frame`(test2, test2$date == "2016" | test2$date == "2015") :

undefined columns selected

> test3 <- test2[test2$date == "2016" | test2$date == "2015",]

> View(test3)

test2016$date<- substr(test2016$date, 1, 4)

> fit1 = lm(test3$overall\_rating~.-test3$player\_api\_id-test3$player\_fifa\_api\_id-test3$player\_name-test3$birthday-test3$date, data = test3)

> fit1 = lm(test$overall\_rating~. , data=test[ , -which(names(test) %in% c("player\_api\_id","player\_fifa\_api\_id", "player\_name", "birthday", "date", "potential"))] )

> fit1 = lm(test$overall\_rating~+height+weight+preferred\_foot+ attacking\_work\_rate+ defensive\_work\_rate+ crossing+ finishing+ heading\_accuracy+ short\_passing+ volleys+ dribbling + curve+ free\_kick\_accuracy+ long\_passing+ ball\_control+ acceleration+ sprint\_speed + agility+ reactions+ balance +shot\_power+ jumping+ stamina+ strength+long\_shots+ aggression+ interceptions+ positioning+ vision +penalties+ marking+ standing\_tackle+ sliding\_tackle+ gk\_diving+ gk\_handling+ gk\_kicking+gk\_positioning+ gk\_reflexes, data = test)

summary(fit1)

confint(fit1)

mod2 <- lm(y ~ . - x1 - x3, data) # Remove certain predictors from model

**#count unique values for each variable**

**sapply(lapply(df, unique), length)**

*# Or a chunk at a time*

res <- dbSendQuery(con, "SELECT \* FROM mtcars WHERE cyl = 4")

**while**(!dbHasCompleted(res)){

chunk <- dbFetch(res, n = 5)

print(nrow(chunk))

}

dbClearResult(res)

dbDisconnect(con)