Project Milestone: Prediction of Football Match Results with Machine Learning

COMP3020

I. Team members

1. Hoang Trung Thanh - V202100 - 21thanh.ht@vinuni.edu.vn
2. Tran Anh Vu - V202100569 - 21vu.ta@vinuni.edu.vn
3. Nguyen Canh Huy - V202100401 - 21huy.nc@vinuni.edu.vn

II. Context

1. Motivation

We have chosen this project due to our passion for football and the intriguing challenge of predicting match outcomes in a sport known for its unpredictability. This project offers an opportunity to hone our skills in handling extensive datasets, selecting relevant features, and developing machine learning models amidst noisy data.

2. Background

In the realm of football match result prediction, several models have demonstrated acceptable accuracy. For instance, Cañizares et al. (2017) achieved 61%, Prasetio et al. (2016) 69.5%, and Rodrigues (2022) 65.25%. These models rely on various data such as goals, shots, corners, and player statistics from top European football leagues, covering seasons from 2010/2011 to 2018/2019

II. Method

1. Data Process

- We used the Dataset of the English Premier League (EPL) from 2000 to 2022. After processing the data, we kept some of the key components such as: HomeTeam, AwayTeam, Referee, FullTime and HalfTime score, etc.
- All data is in csv format, ready for use within standard spreadsheet applications. (http://www.football-data.co.uk/matches.php)

2. Training Model using SVM

- We try to use a baseline SVM to classify the result of a match between three classes: Draw, Home Win and Away Win. Because there's a lot of data in the matches that are supposedly linearly proportional to the result such as: The cost of the whole team, the placement, the betting rate, etc. so we intend to use SVM as it is a great choice of linear models for prediction of linear data.

3. Training model using Random forest

- We also tried the Random Forests Classifier because this method can perform well in high-dimensional datasets with a large number of features. They automatically perform feature selection by considering subsets of features in each tree, which can be beneficial when dealing with many features. Moreover, Random Forests are capable of capturing complex, nonlinear relationships in the data, which matches with the case of soccer result prediction.

III. Preliminary Experiments:

- **Split train-validate dataset:** We split the processed data into train and validate datasets based on the date of samples, matches used for the training set will always happen before those put in validating sets because it makes more sense if we use results of the past to predict matches in the future. We considered multiple dates as splitting points to find the optimal value
- We tried several models for the classifier, and then collected their accuracy (we did not consider f1_score because our data is balanced: 23% Away Team win - 35% Draw - 42% Home Team win). From these trials, we collect a table of accuracy scores:

Model	Date of Split			
	2010-01-01	2012-01-01	2015-01-01	2020-01-01
K-nearest neigbors (k = 100)	0.58	0.51	0.54	0.55
Perceptron	0.50	**0.44 (worst result**	0.49	0.47
SVM	0.54	0.52	0.54	0.56
XGBoost	0.53	0.50	0.52	0.56
Random Forest (number of estimators = 300)	**0.58 (best result)**	0.54	0.54	0.57

IV. Challenges

1. The uncertainty nature of the data

- One of the main challenges is the uncertainty nature of the data, as soccer is naturally a chaotic sport. Of course, a stronger team usually wins against a weaker team. However, there is not a function that maps the features to the result. Two matches with seemingly identical features could have very different outcomes, so there are noises that we need to handle.

2. The complexity of the features

- In football, even pre-match there are a lot of features that we have to consider to be taken into our model, from the cost, the power of each player on the field, to the betting rate of each vendor. We cannot take every feature into our model, because it could easily led to our model being overfitted, as we do not have infinite data. As such, one of the main obstacles that we need to clear is feature selection and feature engineering.

V. Next Steps

1. **Better feature selection and feature engineering**

- The feature in our current model is only taken directly from the table, and with some preprocessing such as one-hot encoding and label transforming, and we only take into account the most direct implications of team winning such as betting rate for each team. We will try to find the new relationship between other hidden features and the result to have a better prediction of the match

2. **Hyperparameter tuning & trying other models**

- In this report, we have not touched some models such as neural network. We will try to optimize our models by hyperparameter tuning, as well as trying other kinds of models like neural networks to see the efficiency of each model in this problem that we are tackling.

VI. Contribution

Hoang Trung Thanh	
1	Trained model using SVM method
2	Data cleaning & preprocessing & feature selection
3	Finalizing the milestone report
Tran Anh Vu	
1	Trained model using Random Forest method
2	Finalizing the milestone report
Nguyen Canh Huy	
1	Data finding and data crawling
2	Finalizing the milestone report

VII. References

Rodrigues, F., & Pinto, A. M. G. (2022). Prediction of football match results with Machine Learning. Procedia Computer Science, 204, 463–470. https://doi.org/10.1016/j.procs.2022.08.057