University of Washington

ME 599D

Data-Driven Modeling

Project Proposal

BingWen Hu

Dau Cheng

*Introduction*

The dataset and original idea of this project is retrieved from a football prediction competition on kaggle presented by [Octosport](https://www.octosport.io/) and [Sportmonks](https://www.sportmonks.com/). ” The competition is about predicting the probabilities of more than 150000 match outcomes using the recent sequence of 10 matches of the teams.”[1]. And our goal is to apply machine learning models on the given dataset and try to compare the result with the bookmakers odds.

There are several sports game prediction related research and a common mentioned difficulty is data collection. Fortunately, the dataset is given for the competition and makes this project practical. However, we still need to deal with the problem of missing data and imbalanced classes. An naive solution is to handle the problem with Sklearn library SimpleInputer, but Maupong’s survey(2021) provides more options for us to discuss in the project. Another solution is to utilize a Masking layer on top of LSTM which can prevent missing values with a mask value. Another challenge in this project is to find the right machine learning model to best fit the data. Many methods had been applied, including logistic regression[3], Naive Bayesian classifier, SVM[4] and neural network[5]. In this project, we aim to start with Decision Tree (DT) algorithms[6] and discover more using Long Short-Term Memory (LSTM) [7] networks is applicable.

*Approach*

We plan to use Python as our main programming language with Pandas, numpy and matplotlib package for data pre-processing and visualization. Scikit-Learn、 LightGBM(DT)、Tensorflow(LSTM) are leveraged for training our machine learning models.

Simple Test Case example:

Home\_team\_ID: 000

Away\_team\_ID: 001

League\_ID: 001

Home\_team\_coach\_ID: 9605704

Away\_team\_coach\_ID: 30866250

(Right side are result of 10 games, will be separate to 10 columns)

Home\_history\_is\_play\_home: 0100111010

Away\_history\_is\_play\_home: 1101101010

Home\_team\_goal: 3201201310

Away\_team\_goal: 0121032101

Home\_team\_rating: 3157652161

Home\_team\_oponent\_rating: 2121532171

Away\_team\_rating: 3157652161

Away\_team\_oponent\_rating: 2121532171

Example output:

| id | home | draw | away |
| --- | --- | --- | --- |
| 000001 | 0.333 | 0.333 | 0.333 |

*References*

1. [Football Match Probability Prediction | Kaggle](https://www.kaggle.com/competitions/football-match-probability-prediction/overview) ([www.kaggle.com/competitions/football-match-probability-prediction/overview](http://www.kaggle.com/competitions/football-match-probability-prediction/overview))
2. Emmanuel, T., Maupong, T., Mpoeleng, D. et al (2021). A survey on missing data in machine learning. *J Big Data 8, 140*
3. D. Prasetio and D. Harlili (2016). Predicting football match results with logistic regression *International Conference On Advanced Informatics: Concepts, Theory And Application, 1-5*
4. Rahul Baboota, Harleen Kaur (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League, *International Journal of Forecasting, 35(2), 741-755*
5. K. Huang and W. Chang (2010). A neural network method for prediction of 2006 World Cup Football Game *The 2010 International Joint Conference on Neural Networks, 1-8.*
6. Alfredo, Y.F., & Isa, S.M. (2019). Football Match Prediction with Tree Based Model Classification. *International Journal of Intelligent Systems and Applications.*
7. Qiyun Zhang, Xuyun Zhang et al. (2021). Sports match prediction model for training and exercise using attention-based LSTM network*, Digital Communications and Networks*