CSE 532 Project Update 1: Soccer Analytics

Akhilesh Soni, soni6@wisc.edu

Github repository: https://github.com/soniakhilesh/soccer-analytics
Dataset: https://www.kaggle.com/saife245/english-premier-league

In first phase of the project, we started working on data preparation, feature engineering, and running a model to check sanity of the data. The goal of this project is to predict the match outcome (Win, Loss, Draw) using the performance of the teams in recent matches and previous encounters among them.

Data Engineering We extracted following features from the data:

- Strength of a team: We measure strength of the team by it's position (standing) in the last year's leagure. Every season, top 17 teams are qualified for next year's league and bottom three teams in the table are relegate, thus giving a chance to three new teams to enter the league. Hence, sometimes it happens that we do not have information of team's performance in the last season's league. In such cases, we assign a position of 30 to the teams which did not play the league last season i.e. they didn't qualify.
- Current form: Current form of the home and away teams: Each season of EPL goes for almost 8-9 months. Hence, players go through different phases of performance and injuries which also affects the team's performance as a whole. To account for this behavior, we measure current form of the team by calculating exponential decay function of number of shots on target in the last 5 games with the most recent game getting the highest weight.

  $$Form = \sum_{i=1}^{5} ST * (\exp^{-\alpha i})$$

  where ST stands for shots on target.

  We believe that shots on target is a better measure of performance than goals as it's a better measure of team's attacking tactics than goals.
- Previous encounters: If Team A (Home) is facing Team B, we all account for all previous encounters of those teams. In particular, we give a point of +1 to Team A for a win in the past, 0 for a draw, and -1 for a loss. We then take an average of these scores. Recent performance: To account for team's recent performance, we also consider an average of the following stats over the last three games.

- FTHG = Full Time Home Team Goals
- FTAG = Full Time Away Team Goals
- FTR = Full Time Result
- HTHG = Half Time Home Team Goals
- HTAG = Half Time Away Team Goals
- HTR = Half Time Result
- HS = Home Team Shots
- AS = Away Team Shots
- HST = Home Team Shots on Target
- AST = Away Team Shots on Target
- HHW = Home Team Hit Woodwork
- AHW = Away Team Hit Woodwork
- HC = Home Team Corners
- AC = Away Team Corners
- HF = Home Team Fouls Committed
- AF = Away Team Fouls Committed
- HY = Home Team Yellow Cards
- AY = Away Team Yellow Cards
- HR = Home Team Red Cards
- AR = Away Team Red Cards

Note that we calculate this separately for a home or an away game. This implies that if a team is playing a home game, we only consider last 3 home games to measure the recent performance of team in home games. Similarly for an away game, we only consider last three away games of the team. We converted the targets (Win, Loss, Draw) to numerical features which are equivalent to +1,0,-1 respectively and use OnevsRest strategy to handle multi-class classification. We use the sklearn.multiclass tools to handle this.

We wrote the cleaned data after extraction above listed features with the required features and target variables to the file finalData.csv in data/archieve/Datasets/finalData.csv

Our preliminary data engineering can be found in src/data.ipynb

Preliminary Experiments We split the data into training-testing data with test size constituting 30% of the data. We scaled the data to 0-1 in order to avoid scaling issues which can arise as our features are on a different scale. We tested performance of Support Vector Machine on this a multi-classification problem (Win, Loss, Draw),using a one vs rest strategy. Preliminary experiments showed AUC score in the range of 0.65-0.70 on training and testing set. We also performed cross validation and hyper-parameter tuning to get best posssible performance from SVM.

Next Steps

- We plan to spend some more time in feature engineering to see if it's possible to extract better features. In particular, we plan to improve our form feature which currently is based solely on number of shots. Moreover, we don't yet what's the best choice of α in the form function listed above. It might be possible to construct two type of form features: *aggressive form* based on goal scored, shots on target, goals scored, etc and *defensive form* based on goals conceded, fouls committed, corners conceded, etc. Our aim is to fit a regression model separately to calculate defensive and aggressive form based on performance in last 5 matches.
- This will be followed by testing performance of k-means and Neural Network models on our problems. We also plan to get a reference of state of the art methods (like betting websites) to analyse how good is our model performing.
- If time permit, we also plan to explore supervised PCA on the refined dataset to identify most important information in the dataset and study performance of different models.