CSE 532 Project Proposal: Predicting soccer match result using ML

Akhilesh Soni, soni6@wisc.edu

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

Introduction

Soccer is a game of coordination. Each side consists of 11 players with 3 allowed substitutions. A game is played in two intervals of 45 minutes each with a 15 minutes break in between. The team which scores more goals is declared as the winner. Each goal, more often than not, involves a highly complex interaction among players amidst uncertainty of the situation. Sometimes goal results from individual brilliance and sometimes from the well coordinated passes among players of the attacking team. Of course, each good attack does not result in a goal and often there are goals scored against the run of play. However the game stats do give a sense of which team might be dominating the game and is more likely to win. We aim to study if we can predict result of the match from the historic game stats of the two teams against themselves and against other teams. In this project, we keep our focus on English Premier League or mostly refereed to as EPL. In a typical match of EPL, one team hosts another team. This means that each match involves one team playing "Home" and the other team play "Away" game. Each season of EPL runs from August to May of following year. There are 20 teams and each team plays 38 games: 19 home and 19 away.

Dataset

EPL dataset for all matches over the last 20 seasons is available here. The data along with final result (Win/Loss/Draw) consists of statistics like shots taken by home and away team, yellow or red cards conceded, score at half time,possession, number of off-sides, number of free kicks conceded, etc. These stats are a good representation of the match summary and hence can be used to predict the likelihood of home team winning. We also have standings of each team over the last 20 seasons in EPL. This will be a good representative of the overall strength of the team over the last few seasons. There are roughly 30 features which consists of goals, off-sides, fouls, free kicks, cards, shots, corners etc. Each season of EPL consists of 380 games. We plan to use approximately 20 seasons of data which would be equivalent to around 6000 data points roughly. Each game comes with half and full time stats and target label being whether home team won or not.

Project Scope

The goal of this project is to predict the match outcome using the historic features listed. In particular, we plan to consider the recent performance of the home and away teams in last few (say 10 games) at home and away venues against other teams in the league. We also consider the history of encounter of the playing teams against each other and their match stats in those encounters. Another feature would be to include standings of teams over last few seasons (say 5). We also plan to accommodate a feature to consider how far are we into the season. Our goal is to find the probability of home team winning a match. For our purpose, we plan to use around 15 seasons of

data to train and predict the outcome of the next season and thus predicting the league winner. We will do this on a rolling basis and do predictions for 5 seasons i.e. use the data for the last 15 seasons to train and predict next season and repeat this 5 times on a rolling basis to evaluate model accuracy. A team is awarded 3 point for a win, 1 for a draw. and none for losing. This well help in ranking the teams at end of the league.
We also plan to train our model with half game time statistics along with the historic data and then predict likelihood of home team winning the game. This can be used in real time for getting the odds of home team winning midway through the game.

Algorithms

We plan to make use of following algorithms and compare their performance on the testing dataset.

1. k-nearest neighbors: Tuning how many neighbors (k) do we need
2. Kernel SVM: Deciding type of kernel and penalty parameters
3. Neural Networks: Experimenting with different number of layers

We plan to run two types of experiments and model training:

- Historic game stats: We train our model on the match statistics for last 15 seasons and predict outcome of next season.
- Historic with half game stats: We train our model with the first half game data along with the historic data. We then investigate if we are able to predict the end game results.

Our first model comparison will be based on how many times in the season is the model able to predict the match outcome correctly. We will also draw insights into the team positions in the league predicted by the model at end of the season.

Our second comparison will be for the model trained with historic as well as half time game stats. We will compare on how many times is the model able to correctly predict end game results midway through the game.

Timeline
The roughly proposed timeline of the project is given in Table 1
Total project duration is Oct 22-Dec 12: 8 weeks

	Description
Week 0	Identify topic, find dataset, and write project proposal–**Oct 22**
Week 1,2	Data cleaning and feature engineering, Dimensionality reduction
Week 3	Build a pipeline and implement k-means and analyze results– **Nov 17**
Week 4	Implement SVM and do Cross Validation
Week 5	Implement Neural networks and do CV–**Dec 1**
Week 6	Start documenting, analyse results, compare performance of models (error rates)
Week 7	Get insights into important features, discuss, and document the report–**Dec 12**

Table 1: Project Timeline