Predicting European Soccer Matches

Sahil Goyal

Machine Learning Engineer Nanodegree Capstone

Abstract

Football (or soccer) is the biggest global sport and is a fast-growing multibillion dollar industry; with an estimate of 27 billion dollars in terms of annual revenue for the football clubs [1]. Prediction of soccer matches is a tough problem. Predicting the exact scoreline is a near impossible task. On the field often defies on paper. So much so, in the 2011-12 season of the English Premier League (EPL), it was not known who the champions would be until the last five minutes of a nine month long season [3]. In the 2015-16 season of EPL, at the beginning of the season it was more likely that Kim Kardashian would become the US president, than the eventual team Leicester winning the title [4].

Keywords: Football, Machine Learning, Feature Engineering

1. Introduction

With more and more money pouring into the sport, the betting industry for predicting the outcome of matches is worth a billions pounds every year [2]. The ridiculous amount of money can be summed in the chart below, that shows the amount of money clubs have paid to secure a football player's services (I being a mere mortal find myself counting the number of zeros in these huge sums of money). // TODO: Add Image

TODO: ADD IMAGE; ADD REFERENCES

- Bullet point one
- Bullet point two

1. Numbered list item one
2. Numbered list item two

1.1. Personal Motivation

Quisque elit ipsum, porttitor et imperdiet in, facilisis ac diam. Nunc facilisis interdum felis eget tincidunt. In condimentum fermentum leo, non consequat leo imperdiet pharetra. Fusce ac

1.2. Problem Statement

Donec eget ligula venenatis est posuere eleifend in sit amet diam. Vestibulum sollicitudin mauris ac augue blandit ultricies. Nulla facilisi. Etiam ut turpis nunc. Praesent leo orci,

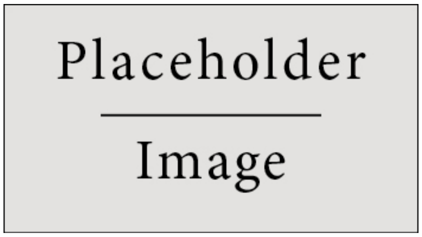

Figure 1: Figure caption

Integer risus dui, condimentum et gravida vitae, adipiscing et enim. Aliquam erat volutpat. Pellentesque diam sapien, egestas eget gravida ut, tempor eu nulla. Vestibulum mollis pretium lacus eget venenatis. Fusce gravida nisl quis est molestie eu luctus ipsum pretium. Maecenas non eros lorem, vel adipiscing odio. Etiam dolor risus, mattis in pellentesque id, pellentesque eu nibh. Mauris nec ante at orci ultricies placerat ac non massa. Aenean imperdiet, ante eu sollicitudin vestibulum, dolor felis dapibus arcu, sit amet fermentum urna nibh sit amet mauris. Suspendisse adipiscing mollis dolor quis lobortis.

$$e = mc^2 \tag{1}$$

1.3. Datasets and Inputs

The dataset I will be using for this task is the European soccer database over at Kaggle [6]. The database contains data from 25,000 matches; 10,000 players for the years 2008-2016, spread over 11 European countries. The matches are defined well in terms of data, with information about fouls, shots on target, possession etc. The dataset also contains betting odds from 11 different betting companies.; which will help me in evaluating my models.

The dataset contains attributes for both teams and matches, that I will be using to come up with a regression model to obtain a skill level rating.

The dataset is perfect for the questions I am trying to answer. Note that I intend to select one country's league first (probably the English Premier League because of familiarity), and then try to generalize it for different countries.

1.4. Overview of time spent

Task	**Percentage of time spent**
Data Loading and Familiarity	10%
Feature Generation using domain knowledge	50%
Model Fitting and Optimization	30%
Results, Visualizations and Discussion	10%

Table 1: Table caption

2. Exploring the data

2.1. Exploratory Visualizations

2.1.1. SQLite

2.1.2. The base classifier: Analyzing the home advantage

2.1.3. Players' ratings histogram and visualization

3. Feature generation

3.0.1. Idea behind dividing players into midfield, defender etc...

3.0.2. NaN filling and logic

Mention imputer