Software Design Specification

for

Predicting outcomes and scores for EPL

Version 1.7

Prepared by

Manmeet Singh Chawla

Bhagyalakshmi Patchipalu

Ramontal Philemon L

Team Euro Soccer

November 29, 2018(1st draft)

December 09, 2018(revision)

Table of Contents

Table of Contents ii

Revision History ii-iii

1. Introduction 1

1.1 Goals 1

1.2 Statement of scope 1

1.3 Model context 1

1.4 Major constraints 1

2. Data design 1-2

2.1 Data sources 2

2.2 Input data 2

2.3 Output data 2

2.4 Variable description 2

2.4.1 Independent Variables 2

2.4.2 Dependent Variables 3

2.5 Pre-design analysis 4

2.6 Tidying procedure 4

3. Model Architecture 4

3.1 Type of Model 4

3.2 Training set 4

3.3 Testing set 4

4. Approach 4

4.1 Implementation Details 4

5. Testing Strategy 5

5.1 Classes of tests 5

5.2 Expected response 5

5.3 Performance bounds 5

5.4 External review and validation 5

6. References 5

Appendix A: Glossary 6

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Date** | **Reason for Changes** | **Version** |
| Bhagyalakshmi Patchipalu | 11/30/18 | Input Info | 1.0 |
| Ramontal Philemon L | 12/04/18 | Input info/Update | 1.1 |
| Manmeet Singh Chawla | 12/05/18 | Model and Approach | 1.2 |
| Bhagyalakshmi Patchipalu | 12/05/18 | Edited and updated sections 2.4 | 1.3 |
| Ramontal Philemon L | 12/05/18 | Final update | 1.4 |
| Bhagyalakshmi Patchipalu | 12/08/18 | Revising oversights of the document | 1.5 |
| Philemom Ramontal | 12/09/18 | Revising document | 1.6 |
| Manmeet Singh Chawla | 12/09/18 | Final update | 1.7 |

# Introduction

Predicting the football match outcomes and scores for English Premier League. The objective of this model is to predict the outcomes and scores for English Premier League (EPL) by using Data Mining and Machine learning techniques. By using these approaches, the most optimal model will be able to provide relative accuracy.

## Goals

The goal of this analysis is to predict football match results and scores in the English Premier League by using four different models they include Logistic Regression, Decision Tree Classification, Support Vector Machine and Artificial Neural Network (ANN) however, the most optimal model was proven to be Artificial Neural Network (ANN). For predicting Home and Away team scores, Multiple Linear Regression model was the optimal choice.

## Statement of scope

There are various model options for this problem. However, 4 models have been chosen to test the data set of all EPL teams for the past 10 years. The model with the highest percentage of accuracy will be the most optimal for this project. The allotted time for this project is sufficient to reach a reasonable conclusion. The project requires a data set of at least 1000 rows of data. Such a model can be implement for future matches in the EPL and can deliver positive results for EPL clubs.

## Model context

This project consists of creating models whose result will predict the outcome of future EPL matches. The models created will use past results for comparative analysis with future matches. A decision tree classification, logistic regression, support vector machine and artificial neural network models will be used for this analysis. The data will be tested in each model to determine which will provide the most accurate results.

## Major constraints

The human element can turn a game upside down which will also affect the model. If a coach decides to sit a player for a game this can affect the outcome. Secondly, if a player is injured and cannot participate this will also affect the accuracy of this predictive model.

# Data design

A preliminary assessment of the data will be done to assess the various variable. The dependent variable is the outcome/score of a given match. Therefore, there are multiple independent variable which include

* FTHG vs FTAG
* Home shot on target vs Away shot on target
* Home team fouls vs Away team fouls
* Home team corners vs Away team corners

## Data sources

Data collected from the website: <https://datahub.io/sports-data/english-premier-league#data>

The data is structured in a CSV file format, so it can be imported into the model.

(Nivard van Wijk, December 2012.) Based on data analysis of multiple seasons of the English Premier League the author concluded that (i) There is a signiﬁcant home ground advantage, (ii) The number of goals scored and a Poisson distribution can describe both home and away team.

(Aditya Srinivas Timmaraju.) One of the things that makes predicting outcomes tricky is the significant incidence of draws as compared to other sports. One of the key challenges of this problem is the high incidence of drawn games in EPL.

(Francisco Louzada, October 2013.) In football, and in any sports competition, there is a strong interest in knowing which team or which player the champion at the end of the season shall be. Besides this, the result of a match, the chance to qualify for a specific tournament, the chance of being relegated, the best attack, the best defense, among others, are also subject of interest.

## Input data

Input data consists of game result, home and away team goals, fouls and corner shots.

## Output data

Confusion matrix of the result.

## Variable description

### Independent Variables

### Home Team Shots on Target (HST)

### Description: Shot taken by home team within the target area.

#### Data type: Numeric

### Away Team Shots on Target (AST)

### Description: Shot taken by away team within the target area.

#### Data type: Numeric

### Home Team Fouls (HF)

### Description: Home team is aggressive defensively

#### Data type: Numeric

### Away Team Fouls (AF)

### Description: Away team is aggressive defensively

#### Data type: Numeric

### Home Team Corners (HC)

### Description: Corners kicks taken by Home team

#### Data type: Numeric

### Away Team Corners (AC)

### Description: Corners kicks taken by Away team

#### Data type: Numeric

### Dependent Variables

#### Full Time Results (FTR)

#### Description: It represents Win//Loss/Draw

#### Data type: Numeric

#### Full Time Home Team Goal (FTHG)

#### Description: No. of goal scored by the Home team.

#### Data type: Numeric

#### Full Time Away Team Goal (FTAG)

#### Description: No. of goal scored by the Away team

#### Data type: Numeric

## Pre-design analysis

All the Independent variables were checked for the statistical significance and the least significant variables were eliminated through the backward elimination process.

## Tidying procedure

There are no missing values in the data set. Few variables which a statistical significance did not have were eliminated from the data set.

# Model Architecture

## Type of Model

For predicting the result, Artificial Neural Network model is used. And for predicting the scores, Multiple linear regression model is used.

## 3.1 Training set

Data set is split in to 80% to create the training set i.e. (800 rows of 19 variables), because subsample random selections of the training data, train the classifier and record the performance on the validation set.

## Testing set

Data set is split in to 20% to create the testing set i.e. (201 rows of 19 variables), because the models performance statistic will have greater variance.

# Approach

## Implementation Details

* Computer languages: R Studio 1.1.463
* Libraries: readr (for reading the csv file), (e1071) (SVM), kernlab (assist with SVM feature selection), tidyverse (Data Visualization), h2o (For neural networks), caTools, rpart (decision trees)

# Testing Strategy

After training each model with the training test. The model will predict the test set and the confusion matrix will be created to compare with the actual data. The confusion matrix will allow a clear view of the models accuracy. Each test set will give a level of accuracy so the optimal model will be identified.

## Classes of tests

## Expected response

The model should predict scores of matches with relative certainty.

## Performance bounds

We would like to have the analysis ran within 12 hours. Anything above 12 hours would be unacceptable for this analysis. We should be able to process 1000 rows of data within the allotted timeframe.

## External review and validation

The results from the test set will be compared with the actual scores and the result from the data set.

# References

##### Nivard van Wijk. (December 2012). Soccer Analytics Predicting the outcome of soccer matches.

##### Retrieved from <https://beta.vu.nl/nl/Images/werkstuk-wijk_tcm235-315158.pdf>

##### Aditya Srinivas Timmaraju. Game ON! Predicting English Premier League Match Outcomes

##### Retrievedfrom [http://cs229.stanford.edu/proj2013/TimmarajuPalnitkarKhannaGameON!PredictionOf EPLMatchOutcomes.pdf](http://cs229.stanford.edu/proj2013/TimmarajuPalnitkarKhannaGameON!PredictionOf%20EPLMatchOutcomes.pdf)

* Francisco Louzada. (2014). “Predicting Match Outcomes in the English Premier League: Which Will Be the Final Rank?”

Retrieved from <https://www.coursehero.com/file/32926170/2JDS-1157-3pdf>

Appendix A: Glossary

|  |  |
| --- | --- |
| EPL | English Premier league |
| FTR | Full Time Result |
| FTHG | Full Time Home Team Goals |
| FTAG | Full Time Away Team Goals |
| HST | Home Team Shots on Targets |
| AST | Away Team Shots on Targets |
| HF | Home Team Fouls |
| AF | Away Team Fouls |
| HC | Home team Corners |
| AC | Away team Corners |
| CSV | Comma Separated Values |
| SVM | Support Vector Machine |
| ANN | Artificial Neural Network |