Final Project Report

IST687 - Group B2

9/26/2018

## Table of Contents

### INTRODUCTION

1. Project Background and Description
2. Project Scope and Context of this Analysis

### BUSINESS QUESTIONS

1. What are the Business Questions?

### DATA ACQUISITION, CLEANING, TRANSFORMATION, MUNGING

1. Describe your data acquisition process
2. What data did you select, all, subset, why
3. What was your initial quality assessment
4. What fields/variables did you finally decide on, why
5. Provide a data dictionary
6. Provide data descriptive statistics, rows, str
7. Did you have to do any cleansing, describe
8. Interesting findings

### DESCRIPTIVE STATISTICS

1. Provide demographic statistics – Location
2. Any early observations, nuggets of interest, interpretation, interesting findings
3. Graphs, charts, tables, visuals, text

### USE OF MODELING TECHNIQUES

1. Linear modeling
2. Support vector
3. Provide key statistics of interest and interpretation for each model

### OVERALL INTERPRETATION OF RESULTS/ACTIONABLE INSIGHTS

### REFERENCES

### APPENDIX – RStudio CODE

# FINAL PROJECT REPORT

## IST687 - GROUP B2

### INTRODUCTION

1. Project Background and Description  
   This project is an exercise in taking a dataset from the 2017-2018 English Premier League season and tries to gauge what factors are able to predict which team is likelier to win; what factors are likely to separate winning teams from losing teams, and what can teams from the bottom of the chart learn from teams at the top.
2. Project Scope and Context of this Analysis  
   The scope of this project encompasses data gathered from the 2017-2018 English Premier League season, with goals scored, goals allowed, shots taken, shots allowed, win record, and other soccer metrics. The data shows whether a team won, lost, or drew a game, how many shots did they take and allow, how many fouls they commited, how many goals they scored (both at half-time and end of regular time), and whether they received any yellow or red cards. Obviously, we expect that teams that score more goals tend to win more but, by looking at actual goal differential, we might find that this is not the case. We also want to see whether there are other factors that are pushing a teams’ likelihood to win - such as shots taken, shots on goal taken, and corner kicks - as against to allowing an opposing team more chances to score.

### BUSINESS QUESTIONS

1. What are the Business Questions?

### DATA ACQUISITION, CLEANSING, TRANSFORMATION, AND MUNGING

1. Describe your data acquisition process.  
   The data was acquired from www.football-data.co.uk, a website that provides free data on soccer coverage from around the world. For this project, we downloaded the csv file for the English Premier League 2017-2018 season. We grabbed the csv url and loaded it into RStudio using the following command:

url <- "http://www.football-data.co.uk/mmz4281/1718/E0.csv"  
prem <- prem <- read.csv(url, stringsAsFactors = FALSE)  
head(prem)

## Div Date HomeTeam AwayTeam FTHG FTAG FTR HTHG HTAG HTR  
## 1 E0 11/08/17 Arsenal Leicester 4 3 H 2 2 D  
## 2 E0 12/08/17 Brighton Man City 0 2 A 0 0 D  
## 3 E0 12/08/17 Chelsea Burnley 2 3 A 0 3 A  
## 4 E0 12/08/17 Crystal Palace Huddersfield 0 3 A 0 2 A  
## 5 E0 12/08/17 Everton Stoke 1 0 H 1 0 H  
## 6 E0 12/08/17 Southampton Swansea 0 0 D 0 0 D  
## Referee HS AS HST AST HF AF HC AC HY AY HR AR B365H B365D B365A  
## 1 M Dean 27 6 10 3 9 12 9 4 0 1 0 0 1.53 4.5 6.50  
## 2 M Oliver 6 14 2 4 6 9 3 10 0 2 0 0 11.00 5.5 1.33  
## 3 C Pawson 19 10 6 5 16 11 8 5 3 3 2 0 1.25 6.5 15.00  
## 4 J Moss 14 8 4 6 7 19 12 9 1 3 0 0 1.83 3.6 5.00  
## 5 N Swarbrick 9 9 4 1 13 10 6 7 1 1 0 0 1.70 3.8 5.75  
## 6 M Jones 29 4 2 0 10 13 13 0 2 1 0 0 1.62 4.0 6.50  
## BWH BWD BWA IWH IWD IWA LBH LBD LBA PSH PSD PSA WHH  
## 1 1.50 4.60 6.75 1.47 4.5 6.50 1.44 4.40 6.50 1.53 4.55 6.85 1.53  
## 2 11.00 5.25 1.30 8.00 5.3 1.35 10.00 5.00 1.30 10.95 5.55 1.34 10.00  
## 3 1.22 6.50 12.50 1.22 6.2 13.50 1.25 5.75 15.00 1.26 6.30 15.25 1.25  
## 4 1.80 3.50 4.75 1.85 3.5 4.30 1.80 3.40 4.60 1.83 3.58 5.11 1.80  
## 5 1.70 3.60 5.50 1.70 3.7 5.00 1.67 3.60 5.25 1.70 3.83 5.81 1.70  
## 6 1.57 4.00 6.00 1.65 3.8 5.30 1.60 3.70 6.00 1.64 3.94 6.35 1.62  
## WHD WHA VCH VCD VCA Bb1X2 BbMxH BbAvH BbMxD BbAvD BbMxA BbAvA  
## 1 4.2 6.00 1.53 4.50 6.50 41 1.55 1.51 4.60 4.43 6.89 6.44  
## 2 4.8 1.33 10.00 5.50 1.33 40 11.50 10.10 5.60 5.25 1.36 1.32  
## 3 5.5 13.00 1.25 6.25 15.00 41 1.27 1.24 6.55 6.06 15.50 13.67  
## 4 3.3 5.00 1.83 3.60 5.00 41 1.86 1.81 3.65 3.50 5.11 4.82  
## 5 3.5 5.50 1.70 3.80 5.75 40 1.71 1.69 3.85 3.69 6.00 5.50  
## 6 3.6 6.00 1.65 4.00 5.50 41 1.66 1.61 4.05 3.84 6.50 5.98  
## BbOU BbMx.2.5 BbAv.2.5 BbMx.2.5.1 BbAv.2.5.1 BbAH BbAHh BbMxAHH BbAvAHH  
## 1 37 1.65 1.61 2.43 2.32 21 -1.00 1.91 1.85  
## 2 35 1.70 1.63 2.40 2.27 20 1.50 1.95 1.91  
## 3 36 1.71 1.66 2.33 2.23 20 -1.75 2.03 1.97  
## 4 36 2.19 2.11 1.79 1.72 18 -0.75 2.10 2.05  
## 5 35 2.17 2.08 1.80 1.76 19 -0.75 1.94 1.90  
## 6 36 2.17 2.08 1.80 1.75 19 -0.75 1.83 1.78  
## BbMxAHA BbAvAHA PSCH PSCD PSCA  
## 1 2.10 2.02 1.49 4.73 7.25  
## 2 2.01 1.96 11.75 6.15 1.29  
## 3 1.95 1.90 1.33 5.40 12.25  
## 4 1.86 1.83 1.79 3.56 5.51  
## 5 2.01 1.98 1.82 3.49 5.42  
## 6 2.16 2.10 1.56 4.25 6.85

1. What data did you select, all, subset, why?  
   We decided to focus on just a single season of data, selecting data for both the home and away team regarding whether the home team one, lost, or drew, how many goals each team scored both at half-time and end of regulation, how many shots each team took both in general and on goal, corner kicks, fouls committed, yellow and red cards received, and other data such as the date, and referee. We also had different betting odds for all teams and for each match but we decided to drop these columns as they weren’t part of the scope of the project.
2. What was your initial quality assessment?  
   Our first quality assessment was to check that there were no missing values within the dataset.  
   After that, we made sure that the data types were correct, i.e. strings weren’t stored as factors, numbers were numeric, and dates were stored as dates.  
   We also made sure that there were no NULL values or NA's, which fortunately there weren’t.  
   Finally, after the assessment was made, we decided to transform the dataset in order to have one row per team per date, renaming the columns, dropping the betting line columns, and adding a flag to signal whether the team in question was the home or away team.

# Select Home Team and Home team data.  
premHomeTeams <- prem %>% dplyr::mutate(HomeAway = "Home") %>% dplyr::select(Div,   
 Date, Team1 = HomeTeam, Team2 = AwayTeam, HomeAway, GS = FTHG, GA = FTAG,   
 FTR, HTGS = HTHG, HTGA = HTAG, HTR, Referee, ShotsTaken = HS, ShotsAllowed = AS,   
 ShotsOnTarget = HST, OppShotsOnTarget = AST, FoulsCommitted = HF, FoulsReceived = AF,   
 CornerKicks = HC, OppCornerKicks = AC, YellowCards = HY, OppYellowCards = AY,   
 RedCards = HR, OppRedCards = AR)  
  
# Select Away Team and Away team data.  
premAwayTeams <- prem %>% dplyr::mutate(HomeAway = "Away") %>% dplyr::select(Div,   
 Date, Team1 = AwayTeam, Team2 = HomeTeam, HomeAway, GS = FTAG, GA = FTHG,   
 FTR, HTGS = HTAG, HTGA = HTHG, HTR, Referee, ShotsTaken = AS, ShotsAllowed = HS,   
 ShotsOnTarget = AST, OppShotsOnTarget = HST, FoulsCommitted = AF, FoulsReceived = HF,   
 CornerKicks = AC, OppCornerKicks = HC, YellowCards = AY, OppYellowCards = HY,   
 RedCards = AR, OppRedCards = HR)  
  
# Bind datasets into a new dataset.  
premTransformed <- premHomeTeams %>% bind\_rows(premAwayTeams)  
head(premTransformed)

## Div Date Team1 Team2 HomeAway GS GA FTR HTGS HTGA  
## 1 E0 11/08/17 Arsenal Leicester Home 4 3 H 2 2  
## 2 E0 12/08/17 Brighton Man City Home 0 2 A 0 0  
## 3 E0 12/08/17 Chelsea Burnley Home 2 3 A 0 3  
## 4 E0 12/08/17 Crystal Palace Huddersfield Home 0 3 A 0 2  
## 5 E0 12/08/17 Everton Stoke Home 1 0 H 1 0  
## 6 E0 12/08/17 Southampton Swansea Home 0 0 D 0 0  
## HTR Referee ShotsTaken ShotsAllowed ShotsOnTarget OppShotsOnTarget  
## 1 D M Dean 27 6 10 3  
## 2 D M Oliver 6 14 2 4  
## 3 A C Pawson 19 10 6 5  
## 4 A J Moss 14 8 4 6  
## 5 H N Swarbrick 9 9 4 1  
## 6 D M Jones 29 4 2 0  
## FoulsCommitted FoulsReceived CornerKicks OppCornerKicks YellowCards  
## 1 9 12 9 4 0  
## 2 6 9 3 10 0  
## 3 16 11 8 5 3  
## 4 7 19 12 9 1  
## 5 13 10 6 7 1  
## 6 10 13 13 0 2  
## OppYellowCards RedCards OppRedCards  
## 1 1 0 0  
## 2 2 0 0  
## 3 3 2 0  
## 4 3 0 0  
## 5 1 0 0  
## 6 1 0 0

1. What fields/variables did you finally decide on, why  
   The original dataset had 65 columns and 380 rows. The variables included regarded the league, date, teams playing, referee, match results and goals at half-time and at the end of regulation, statistics for the match, and betting odds for different sites. We transformed this data so as to end with 24 columns and 780 rows, setting on the following columns:

* Match Information: Div, Date, Team1, Team2, HomeAway, Referee
* Match Results: GS, GA, FTR, HTGS, HTGA, HTR
* Match Statistics: ShotsTaken, ShotsAllowed, ShotsOnTarget, OppShotsOnTarget, FoulsCommitted, FoulsReceived, CornerKicks, OppCornerKicks, YellowCards, OppYellowCards, RedCards, OppRedCards

colnames(premTransformed)

## [1] "Div" "Date" "Team1"   
## [4] "Team2" "HomeAway" "GS"   
## [7] "GA" "FTR" "HTGS"   
## [10] "HTGA" "HTR" "Referee"   
## [13] "ShotsTaken" "ShotsAllowed" "ShotsOnTarget"   
## [16] "OppShotsOnTarget" "FoulsCommitted" "FoulsReceived"   
## [19] "CornerKicks" "OppCornerKicks" "YellowCards"   
## [22] "OppYellowCards" "RedCards" "OppRedCards"

1. Provide a data dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Column Name | New Names | Definition |
| 1 | Div | Division | Football league division. |
| 2 | Date | Match Day | Day match was played. |
| 3 | Team1 | Team | First team of the match. |
| 4 | Team2 | Opponent | Opposing team of the match. |
| 5 | HomeAway | Home-Away Flag | Team1 Home-Away identifier. |
| 6 | GS | Goals Scored | Goals scored by Team1. |
| 7 | GA | Goals Allowed | Goals allowed by Team1. |
| 8 | FTR | Full Time Result | Result at the end of regulation. |
| 9 | HTGS | Half-time Goals Scored | Goals scored at half-time by Team1. |
| 10 | HTGA | Half-time Goals Allowed | Goals allowed at half-time by Team1. |
| 11 | HTR | Half-time Result | Result at half-time. |
| 12 | Referee | Game Referee | Referee for the match. |
| 13 | ShotsTaken | Shots taken | Shots taken by Team 1. |
| 14 | ShotsAllowed | Shots allowed | Shots allowed by Team 1. |
| 15 | ShotsOnTarget | Shots on target | Shots on goal taken by Team 1. |
| 16 | OppShotsOnTarget | Shots on target allowed | Shots on goal allowed by Team 1. |
| 17 | FoulsCommitted | Fouls committed | Fouls committed by Team 1. |
| 18 | FoulsReceived | Fouls received | Fouls committed by Opposing Team. |
| 19 | CornerKicks | Corner kicks | Corner kicks taken by Team 1. |
| 20 | OppCornerKicks | Opponent corner kicks | Corner kicks taken by Opposing Team. |
| 21 | YellowCards | Yellow cards | Yellow cards given to Team 1. |
| 22 | OppYellowCards | Opponent yellow cards | Yellow cards given to Opposing Team. |
| 23 | RedCards | Red cards | Red cards given to Team 1. |
| 24 | OppRedCards | Opponent red cards | Red cards given to Opposing Team. |

1. Provide data descriptive statistics, rows, str  
   The complete dataset has 24 variables and 780 observations - two per match. The dataset does not have any NA’s, meaning that there will be no additional data dropped.  
   The following shows the dimensions and structure of the dataset.

dim(premTransformed)

## [1] 760 24

str(premTransformed)

## 'data.frame': 760 obs. of 24 variables:  
## $ Div : chr "E0" "E0" "E0" "E0" ...  
## $ Date : chr "11/08/17" "12/08/17" "12/08/17" "12/08/17" ...  
## $ Team1 : chr "Arsenal" "Brighton" "Chelsea" "Crystal Palace" ...  
## $ Team2 : chr "Leicester" "Man City" "Burnley" "Huddersfield" ...  
## $ HomeAway : chr "Home" "Home" "Home" "Home" ...  
## $ GS : int 4 0 2 0 1 0 3 1 4 0 ...  
## $ GA : int 3 2 3 3 0 0 3 0 0 2 ...  
## $ FTR : chr "H" "A" "A" "A" ...  
## $ HTGS : int 2 0 0 0 1 0 2 1 1 0 ...  
## $ HTGA : int 2 0 3 2 0 0 1 0 0 0 ...  
## $ HTR : chr "D" "D" "A" "A" ...  
## $ Referee : chr "M Dean" "M Oliver" "C Pawson" "J Moss" ...  
## $ ShotsTaken : int 27 6 19 14 9 29 9 16 22 6 ...  
## $ ShotsAllowed : int 6 14 10 8 9 4 14 9 9 18 ...  
## $ ShotsOnTarget : int 10 2 6 4 4 2 4 6 6 3 ...  
## $ OppShotsOnTarget: int 3 4 5 6 1 0 5 2 1 6 ...  
## $ FoulsCommitted : int 9 6 16 7 13 10 14 15 19 6 ...  
## $ FoulsReceived : int 12 9 11 19 10 13 8 3 7 10 ...  
## $ CornerKicks : int 9 3 8 12 6 13 3 8 11 5 ...  
## $ OppCornerKicks : int 4 10 5 9 7 0 3 2 1 7 ...  
## $ YellowCards : int 0 0 3 1 1 2 0 3 2 1 ...  
## $ OppYellowCards : int 1 2 3 3 1 1 3 1 2 2 ...  
## $ RedCards : int 0 0 2 0 0 0 0 0 0 1 ...  
## $ OppRedCards : int 0 0 0 0 0 0 0 0 0 0 ...

1. Did you have to do any cleansing, describe  
   No, the dataset was already cleaned when we downloaded it.  
   However, we did need to transform it as the data wasn’t in the format we desired. We transformed it so that we our analysis and data modeling would be better adjusted and easier for the later stages of the project.
2. Interesting findings

### DESCRIPTIVE STATISTICS

1. Provide demographic statistics – Location
2. Any early observations, nuggets of interest, interpretation, interesting findings
3. Graphs, charts, tables, visuals, text

### USE OF MODELING TECHNIQUES

1. Linear modeling
2. Support vector
3. Provide key statistics of interest and interpretation for each model

### APPENDIX - RStudio Code

# Load the required packages  
require(dplyp)  
require(ggplot2)  
  
# Load the dataset into R  
url <- "http://www.football-data.co.uk/mmz4281/1718/E0.csv"  
prem <- prem <- read.csv(url, stringsAsFactors = FALSE)  
  
# Data transformation Select Home Team and Home team data.  
premHomeTeams <- prem %>% mutate(HomeAway = "Home") %>% select(Div, Date, Team1 = HomeTeam,   
 Team2 = AwayTeam, HomeAway, GS = FTHG, GA = FTAG, FTR, HTGS = HTHG, HTGA = HTAG,   
 HTR, Referee, ShotsTaken = HS, ShotsAllowed = AS, ShotsOnTarget = HST, OppShotsOnTarget = AST,   
 FoulsCommitted = HF, FoulsReceived = AF, CornerKicks = HC, OppCornerKicks = AC,   
 YellowCards = HY, OppYellowCards = AY, RedCards = HR, OppRedCards = AR)  
  
# Select Away Team and Away team data.  
premAwayTeams <- prem %>% mutate(HomeAway = "Away") %>% select(Div, Date, Team1 = AwayTeam,   
 Team2 = HomeTeam, HomeAway, GS = FTAG, GA = FTHG, FTR, HTGS = HTAG, HTGA = HTHG,   
 HTR, Referee, ShotsTaken = AS, ShotsAllowed = HS, ShotsOnTarget = AST, OppShotsOnTarget = HST,   
 FoulsCommitted = AF, FoulsReceived = HF, CornerKicks = AC, OppCornerKicks = HC,   
 YellowCards = AY, OppYellowCards = HY, RedCards = AR, OppRedCards = HR)  
  
# Bind datasets into a new dataset.  
premTransformed <- premHomeTeams %>% bind\_rows(premAwayTeams)  
  
# Check the structure of the data, head, and see that there are no NAs.  
str(premTransformed)  
head(premTransformed)  
summary(premTransformed)  
  
# Column names need to be changed to better understand the data.  
colnames(premTransformed)