Let's Bet the Best: Analysis of English Premier League dataset

Vamshi Kiran Reddy (Team Lead) (A20423990)
vkesireddy@hawk.iit.edu

Pranav Makkar (A20426151)
pmakkar@hawk.iit.edu

Project Description and Research Goal

The Premier League also referred to as English Premier League (EPL) is the top level football league. In this league, 20 teams (clubs) participates and each team plays 38 matches throughout the season. In today's world every football fan is excited to see their favourite team win the matches, leagues, and also champion leagues. But what if it is being told that we can estimate some facts and predict which team going to win the match or league, it would be quite impressive. In this project, we are trying to extract meaningful and interesting insights from readily available dataset of English Premier League. We also aim to build a model that could make some predictions about the results of the upcoming matches.

What this project seeks to address

- What factors might increase the chance of a team winning a game?
- Does playing at home increases chances of winning a match?
- To make a prediction on which team is likely to win next season?
- What are the strengths and weakness of a particular team?
- Is there any correlation between team winning a league and the team manager?
- Based on half time result, can we determine which team is going to be the winner?
- Based on statistics of a particular team, can we determine which team will win Premier League?
- How Leicester City won the 2015-2016 season? What differently did they do from previous few seasons to win the cup?

Proposed Methodology

- Data Cleaning
    - To make sure that data in the provided data set are having a proper recognizable data type.
    - Filling in null or empty data points with some value, where it can be mean, median or zero.
    - Adding missing values in the data set.
    - Detection and removal of any Outliers.
- Data Visualization and further exploration (using Tableau)
    - Dashboard of EPL season wise winners, standings, number of goals.
    - Using correlation plots to determine relationship between various factors.
    - Depicting the season results based on home, away and drawn matches using Bar charts.
- Apply Various Statistical models
    - Making use of Logistic regression model
    - Using k-Nearest Neighbour prediction method
    - Decision trees
    - Ensemble Model as this will improve predict better results by combining results of above mentioned models.

  And also some new models as we will learn in class, we will try to implement it as well.

Metrics to Measure Analysis

- As we are trying to develop a predictive classifier model, we can use following metrics to validate the results
    - Confusion matrix
    - F1 score
    - Area Under Curve and ROC

Relevant Studies and related work

As we have taken this project from Kaggle, there are some competitors who are working or worked on the same topic. In a relevant article [5] where author proposed a predictive model that allow us to make good prediction about the Football world cup. They build a predictive model by using Decision trees, which gives an accuracy of 79%. Other mentioned references just gave us the overall idea of how we are going to address our problem.

1. Analysis of Football data and Prediction of Results. https://medium.com/@sathieswarantanabal/analysis-of-football-data-and-prediction-of-results-6244fc5b4092
2. H Ruiz, P Power, X Wei, and P Lucey. "The Leicester City Fairytale?": Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons. *KDD 2017, August 2017, El Halifax, Nova Scotia Canada*
3. J Navarro, L Fraduab, A Zubillagac, P Forda and A.P. McRoberta. Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. *Journal of Sports Sciences, 2016. Vol. 34, No. 24, 2195–2204.*
4. PREMIER LEAGUE: Exploratory Data Analysis [2000 to 2016]: https://steemit.com/sports/@favcau/premier-league-exploratory-data-analysis-2000-to-2016.
5. 2018 World Cup Predictions using decision trees (June 2018): https://www.datasciencecentral.com/profiles/blogs/2018-world-cup-predictions-using-decision-trees.

Data set and Sources

1. Results of each match played from Season 2006-2007 to 2017-2018:
   Description: This data contains the results of 4560 Premier League (n=4560) matches played in 12 seasons i.e. 380 matches per season. Each of the result have a date of match being played, home team, away team, number of goals scored by home and away team, final result, the season year, Half time goals, half time result, Number of shots, number of fouls committed. (p=20).
   Data Source: Football Dataset http://www.football-data.co.uk/englandm.php.

   Note: Manager Information is missing this dataset.

Date	Match Date(dd/mm/yy)
HomeTeam	Home Team
Away Team	Away Team
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals

FTR	Full Time Result
HTHG	Half Time Home Team Goals
HTR	Half Time Away Team Goals
Referee	Half Time Result (H=Home Win, D=Draw, A=Away Win)
HS	Match Referee
AS	Home Team Shots
HST	Away Team Shots
AST	Home Team Shots on Target
HF	Away Team Shots on Target
AF	Home Team Fouls Committed
HC	Away Team Fouls Committed
AC	Home Team Corners
HY	Away Team Corners
AY	Home Team Yellow Cards
HR	Away Team Yellow Cards
AR	Home Team Red Cards

2. Statistics of each team played from Season 2006-2007 to 2017-2018:
Description: This contains the statistics of every team played in all 12 seasons. This set has team information which played in all seasons of English Premier League (n=240) and with each team it has number of wins, losses, goals, total yellow & red cards issued to the team, stats related to attack of the team, stats related to defence of the team, Team play stats, and other stats (p=41).
Data source: Statistical Information from Premier League website. www.premierleague.com And Premier League Dataset on Kaggle.
Note: Columns containing null value: Save, head_clearance, total_through_ball, backward_pass, big_chance_missed, dispossessed

Team	Team name
Wins	wins
losses	losses
goals	Goals scored
total_yel_card	Yellow cards
total_red_card	Red cards
total_scoring_att	shots

ontarget_scoring_Att	Shots on target
hit_woodwork	Shots that hit the bar/post.woodwork
att_hd_goal	Goal from headers
att_pen_goal	Goals from penalties
att_freekick_goal	Goals from free-kicks
att_ibox_goal	Goals from inside the box
att_obox_goal	Goals from outside the box
goal_fastbreak	Goals from counter attacks
total_offside	offsides
clean_sheet	Clean sheets
goals_conceded	Goals conceded
saves	saves
Outfielder_block	blocks
interception	interceptions
total_tackle	tackles
last_man_tackle	Last-man tackles
total_clearance	Headed clearances
own_goals	Own goals
penalty_conceded	Penalties conceded
pen_goals_conceded	Goals conceded from penalties
total_pass	passes
total_through_ball	Through balls
total_long_balls	Long balls
backward_pass	Backward passes
Total cross	Crosses
corner_taken	corners
touches	touches
big_chance_missed	Big chances missed
clearance_off_line	Clearances off-the-line
Dispossessed	Dispossessed

penalty_save	Penalties saved
total_high_claim	High claims by goalkeeper
Season	Corresponding season of statistics.
total_clearance	clearances

Data processing and pipeline

- Cleaning of null values from Match Results dataset.
- Change data types
- Adding Manager Details in Match result dataset.

Data Stylized facts

- For Visualization: Bar charts, Line graph, Histograms Boxplots, Density Plots, and Correlation Plot.
- For Dimensionality reduction: Principal Component Analysis (PCA): It is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in large set

Model selection

- Feature Selection Requirements: Based on output from Principal component Analysis (PCA) we are going to determine which features have strong relationship with the final outcome.
- As our project is based on building a model for predicting Match Results i.e. Win/Loss/Draw, we are going to choose Classification approach.

Tools

a. Software IDE
    i. RStudio or
    ii. Jupyter Notebooks
b. R Libraries
    i. tidyverse
    ii. ggplot2
    iii. dplyr
    iv. gridextra
    v. reshape2
    vi. corrplot
c. Project Management and Source Control
    i. GitHub
    ii. Slack
d. Data Visualization tool
    i. Tableau