Premier League Predictor

Using Data Science to predict matches in the English Premier League

Premier
League

Overview

We will try to predict future matches results based off data from the past 15 seasons of the premier league.

The Team

- Hushi Aujla
- Matt Overstreet
- Danny V
- Carlos Santiago

Questions to answer

We will analyse epl match data with machine learning to

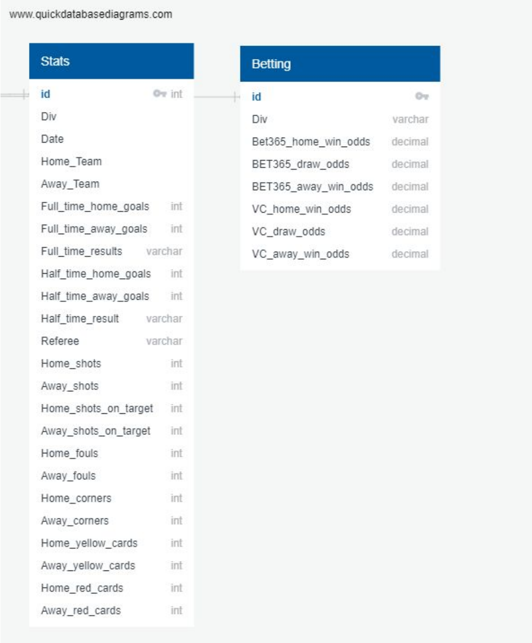
- Predict Full time results
- See if there is a Home team advantage
- Discover which stats are key measures to finding the full time result

Technologies Used

- Python
- Jupyter Notebook
- Github
- PostgreSQL
- Google slides
- Tableau
- Microsoft Excel
- Slack
- Zoom

Data Mining and Cleaning

Data: https://football-data.co.uk/



www.quickdatabasediagrams.com

Stats	
id	int
Div	
Date	
Home_Team	
Away_Team	
Full_time_home_goals	int
Full_time_away_goals	int
Full_time_results	varchar
Half_time_home_goals	int
Half_time_away_goals	int
Half_time_result	varchar
Referee	varchar
Home_shots	int
Away_shots	int
Home_shots_on_target	int
Away_shots_on_target	int
Home_fouls	int
Away_fouls	int
Home_corners	int
Away_corners	int
Home_yellow_cards	int
Away_yellow_cards	int
Home_red_cards	int
Away_red_cards	int

Betting	
id	
Div	varchar
Bet365_home_win_odds	decimal
BET365_draw_odds	decimal
BET365_away_win_odds	decimal
VC_home_win_odds	decimal
VC_draw_odds	decimal
VC_away_win_odds	decimal

Database: postgresSQL

Data Cleaning:

- Dropped columns with NaN data
- Dropped betting data
- Chose to use data from only the last 15 season

Machine Learning - Analysis Phase

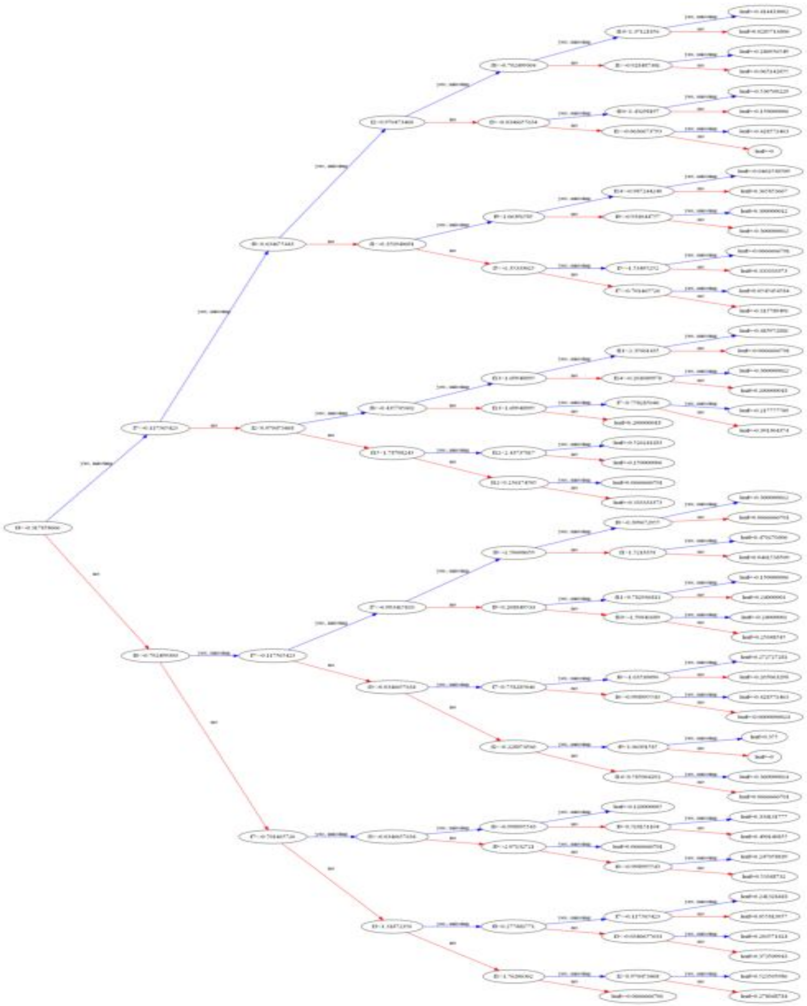
XGBoost advantages:

- Works well with small and medium data sets
- Highly flexible
- reduces overfitting in decision trees and helps to improve the accuracy

- Defined Features
  - Half time result
  - Total shots
  - Shots on target
  - Fouls committed
  - Corners taken
  - Yellow cards
  - Red cards

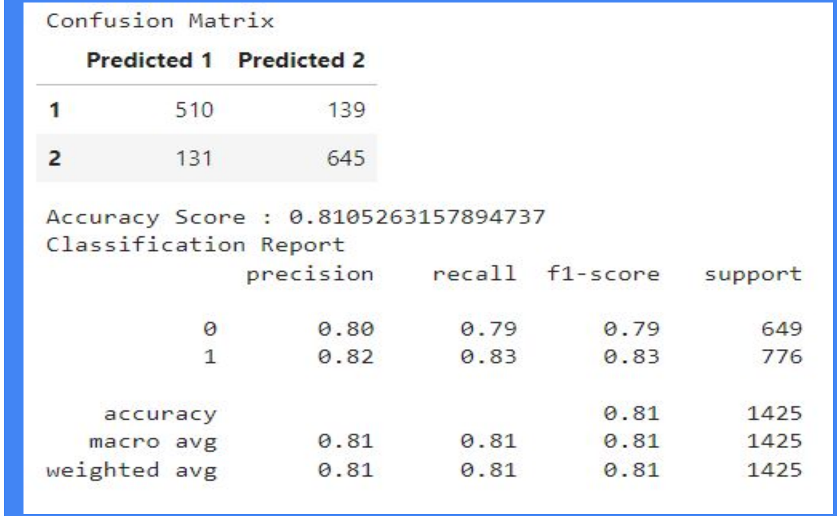
Defined Target: Full time result

Data was split into training(75%) and testing sets (25%)



Machine Learning Results

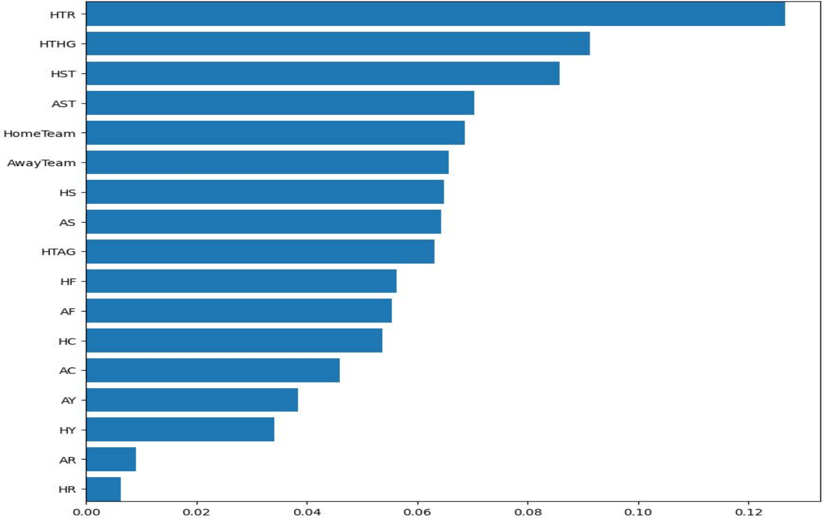
- Model predicts match results with an accuracy of .8105
- Half time result and half time home goals where the most important features in our model
- Followed by shots on target



Confusion Matrix

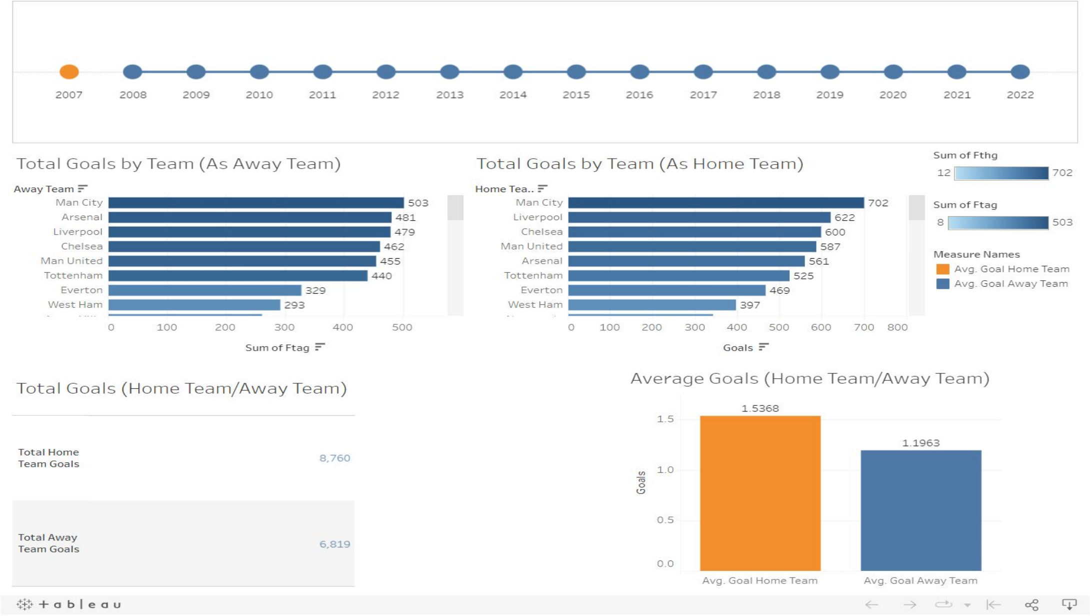
	Predicted 1	Predicted 2
1	510	139
2	131	645

```
Accuracy Score : 0.8105263157894737
Classification Report
              precision    recall  f1-score   support

           0       0.80      0.79      0.79       649
           1       0.82      0.83      0.83       776

    accuracy                           0.81      1425
   macro avg       0.81      0.81      0.81      1425
weighted avg       0.81      0.81      0.81      1425
```



Dashboard

We created a dashboard so viewers can the total amount of goals per team as a home and away team. This data can also be filtered by year.



Total Goals (Home Team/Away Team)

Total Home Team Goals	8,760
Total Away Team Goals	6,819

Dashboard can be found here

Outcomes

- Develop a machine learning model to predict match results for the English Premier League
- Create a dashboard to visualize the data and display the home team advantage
- Discover the most important indicators of full time result
    - Half time result and half time home goals appeared to be the most important indicator of a win
    - Followed by shots on target

Recommendations for further work

In the future we could:

- Look at weather data to see how conditions affect the outcome
- Include financial data from teams to analysis how budget effects game results
- Repeat the analysis for other soccer leagues around the world

The End!