Members

Outline

Project Overview

Business Problem

Data Understanding & Cleaning

Exploratory Data Analysis

Modelling & Evaluation

Conclusions & Recommendation

Project Overview

The Football Predictive Model aims to harness the power of data analytics and machine learning **to enhance decision-making and strategy in the world of football(soccer).**

By leveraging historical performance data, player statistics, and various game-related factors this model seeks to provide:

- ➤ Valuable insights
- ➤ Predictions
- ➤ Recommendations to our **stakeholders.**

Business Problem Statement

Assist betting enthusiasts in making more informed decisions.

Our Solution

Our goals:

1. Create a predictive model to accurately predict match outcomes.
2. Determine the most influential factors that determine match outcomes.
3. Give our stakeholders data-Driven insights influencing match outcomes.
4. Continuously update and improve the model's accuracy over time.

Data Understanding

Our data is scraped from a website covering Premier League match results and player performance, and team rankings from 2011 to 2023

It consists of **4700** rows and **46** columns, with three data types:

- **1** Integer
- **31** Float
- **14** Object columns.

Data Cleaning

I	II	III	IV

- Converting date and time objects to **datetime**

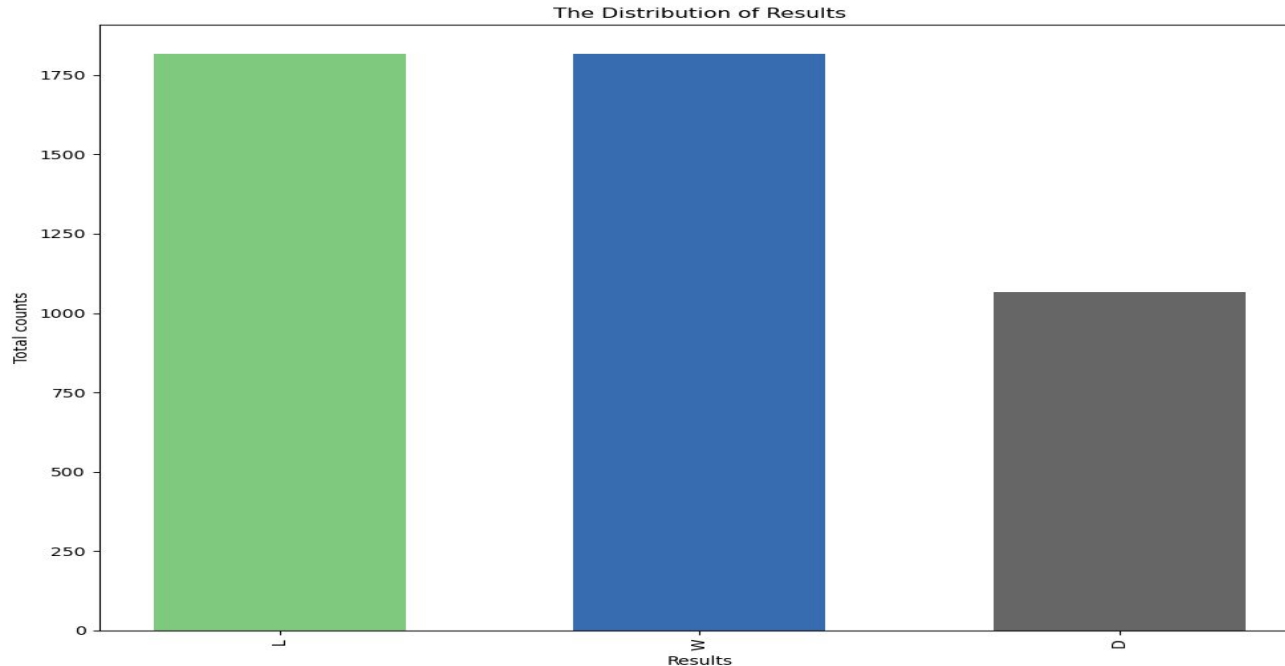- Handling missing values in the; "Notes" , "Attendance", "Distance" column

- Checking for duplicates (**none found**)

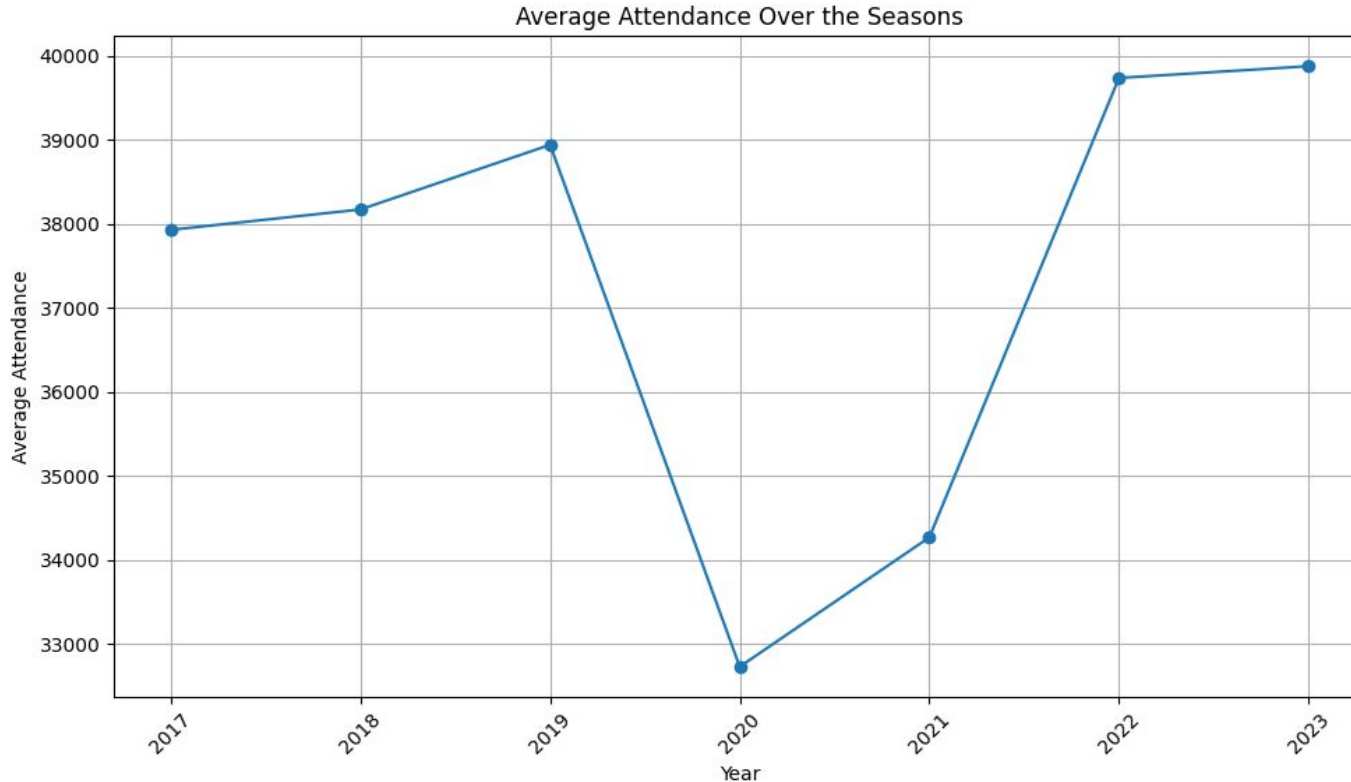- Removing unnecessary columns.

Exploratory Data Analysis
1.Univariate Analysis

The Distribution of Results

Distribution of Results

This is a visualisation of the results of all the teams during all of the seasons.
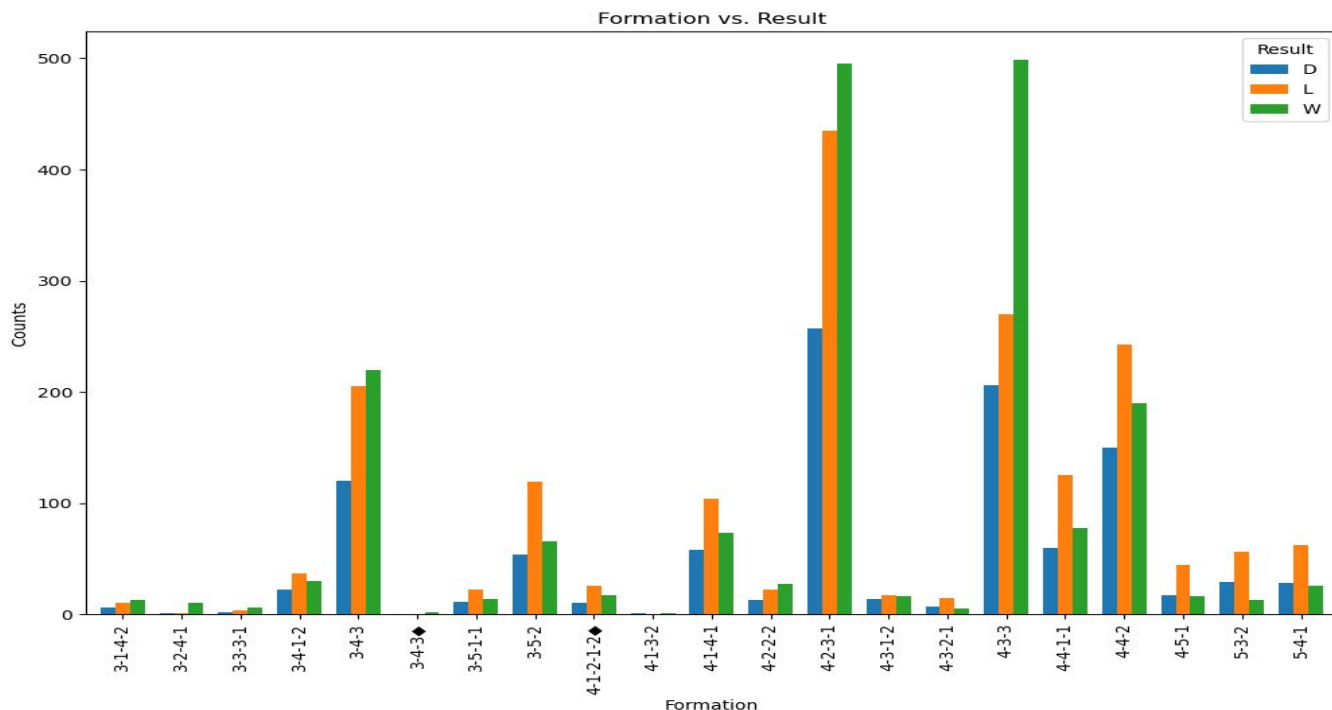
Average Attendance over the seasons

Average Attendance Over the Seasons

- These is a representation of the average attendance across the seasons,

- During 2020 and 2021 the average attendance reduced due to the covid19 pandemic.
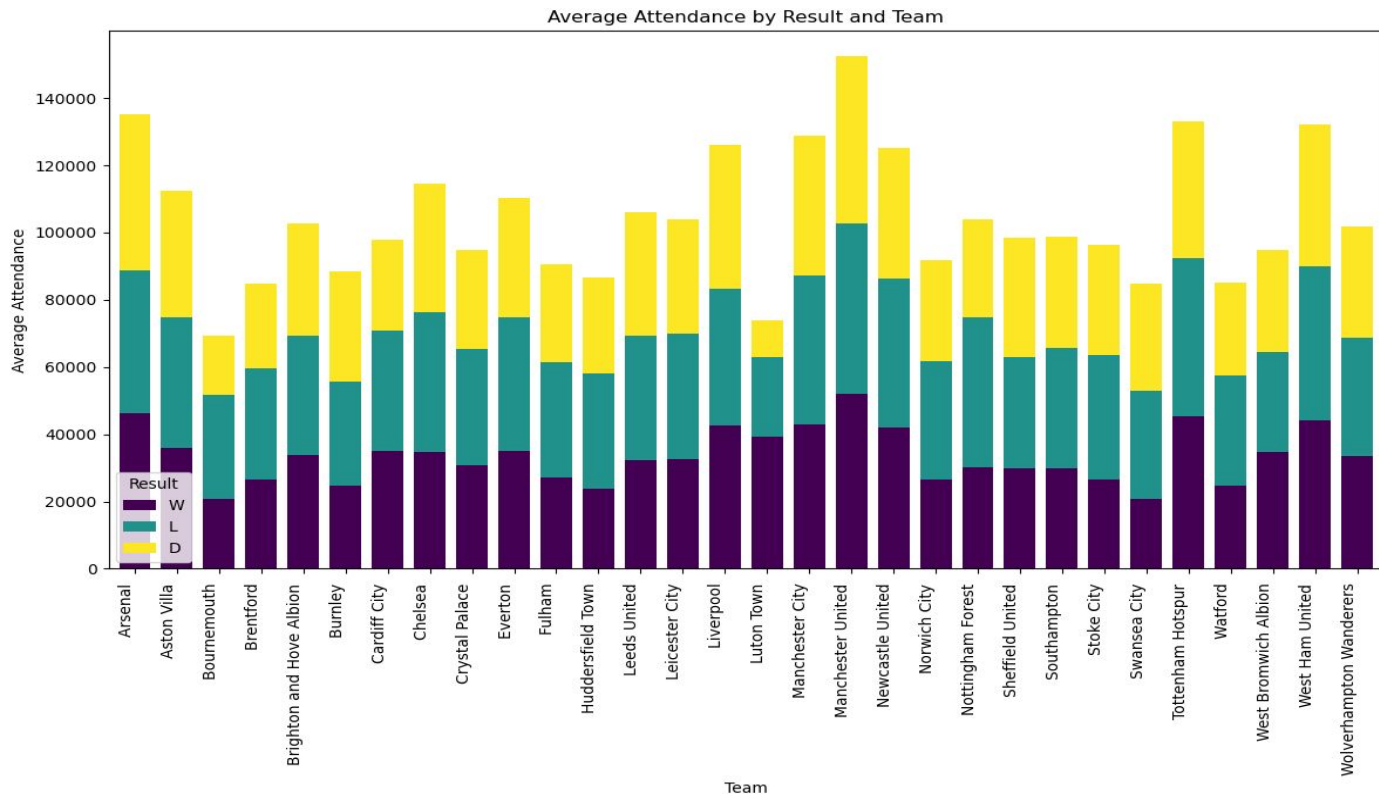
2.Bivariate Analysis

Checking the performance when using certain Formations

The performance of teams on specific formations varies and this graph allows us to see the formations that have the best performance.

Checking the Effect of Attendance on Team performance

Average Attendance by Result and Team

This is a visualisation of the performance of teams during the seasons as compared to the average match attendance.

Modelling

To prepare our data for modelling we;

Handle multicollinearity

Label Encoding

Standardizing data

Dimensionality Reduction

Handling Class Imbalance

Our metric of success is **precision** using this metric we were able to choose the model that performed the best out of our **three main** models:

● Logistic Regression	Accuracy: 57.91% Precision: 58.17% Recall: 57.91% F1-score: 57.57%
● Neural Network	Accuracy: 69.18% Precision: 71.29% Recall: 69.18% F1-score: 68.40%
● Gradient Boosting	Accuracy: 71.64% Precision: 73.33% Recall: 71.64% F1-score: 71.11%

Model Evaluation

ROC Curves
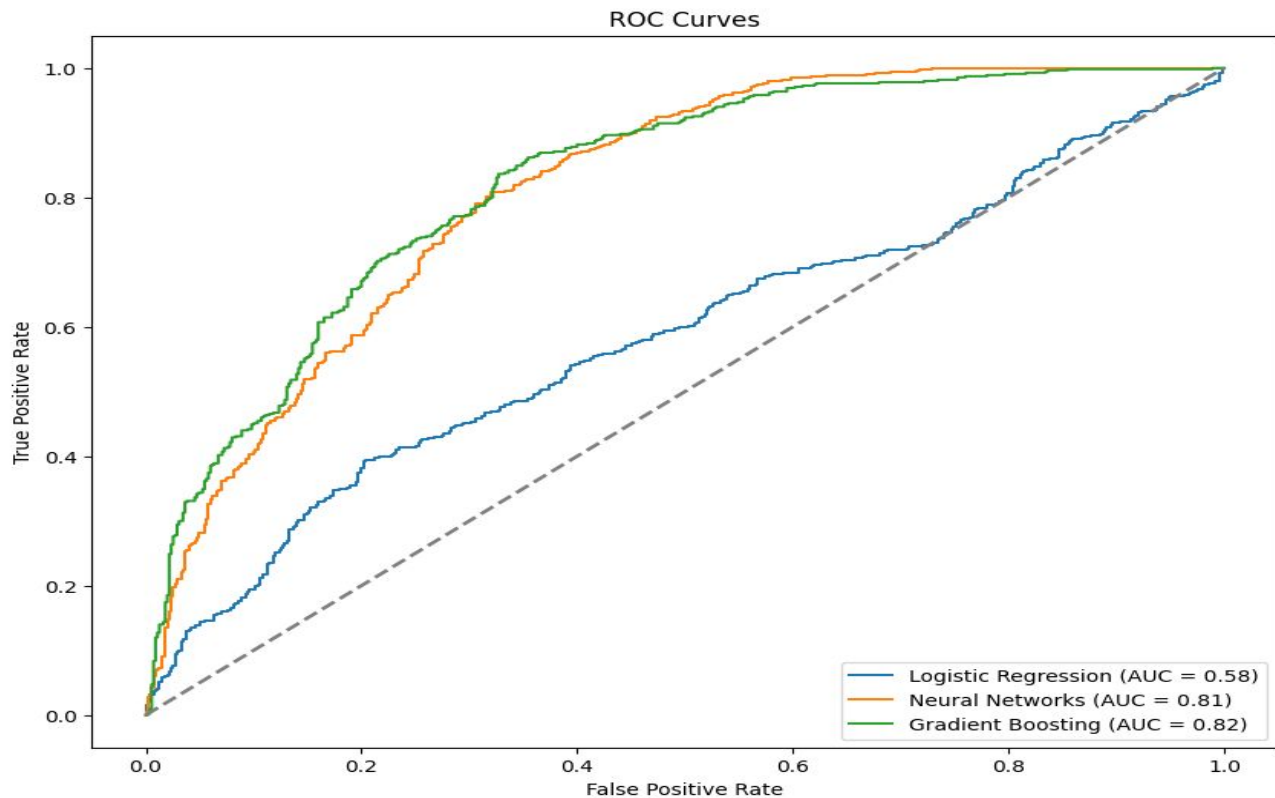
Model Selection

We choose the **Gradient Boosting model** since it had a precision score of **73.33** suggesting accurate predictions and a high proportion of true positives.

The model also performed the best on all the other metrics.

Conclusion

❖ The most influential features from our model were; Venue, Goals a team scored, Goals scored against the team, the number of shots made and the attendance of the match.

❖ Teams playing at home (In their stadium) tend to have more wins.

❖ The most popular formation were the 4-3-3 formation with the highest number of wins, 4-2-3-1 formation and the 4-4-2 formation with the lowest win rate among the three.

Recommendations

❖ We recommend a deeper analysis of some of the other features (eg. **player injuries/statistics , coaching styles, weather and geopolitical factors etc**) that affect a football match outcomes.

❖ Building a better data collection system to capture the various factors that affect match outcomes which will help in building more robust models

❖ We recommend that the users of our model should bet responsibly noting that betting can lead to gambling addiction.

Limitations & Future Work

❖ One limitation of this project is the availability and quality of the data used. Predicting the outcome of a football match can be challenging due to the complex nature of the game and the numerous variables that can affect the result.

❖ To continuously update and improve the model's accuracy over time.

Model Deployment

We utilised Streamlit to deploy our model whereby we created a web application which will take into consideration the user selecting the two playing teams and try to predict.

The expected outcome is either win or loss/draw. Which is based on the home team

Thank You!