HDSC June ’23 Premiere Project by Team Hadoop

Overview

Predicting the outcome of football matches has long been a subject of great interest and excitement among football enthusiasts, sports analysts, and betting enthusiasts. The ability to accurately forecast match results not only adds to the thrill of the game but also holds significant implications for various stakeholders involved.

The goal is to develop a prediction algorithm that can leverage historical data and current information to provide reliable insights into the probable outcome of future football matches. Such an algorithm would enable users to make informed decisions, whether for betting purposes, fantasy league competitions, or simply for enhancing their understanding and enjoyment of the sport.

Aim and Objective

The aim of this project is to develop a machine learning-based system that accurately predicts football match outcomes, leveraging historical data and advanced analytical techniques. The goal is to provide users with reliable insights into future match results, enhancing their understanding, enjoyment, and decision-making in relation to football.

Data Source

The dataset was obtained from Kaggle via the link : https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017/code

Data PipeLine

- Business Understanding
- Data Understanding
- Exploratory Data Analysis
- Data Preprocessing
- Feature Encoding
- Data Splitting
- Model Training and Evaluation

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial initial step in the data analysis process. It involves examining and understanding the structure, patterns, and characteristics of the dataset at hand. By exploring the data, we can uncover valuable insights, identify trends, detect anomalies, and gain a deeper understanding of the variables and relationships within the dataset

During our Exploratory Data Analysis, the following were some of the Insights uncovered.

1. According to the correlation matrix, there is a -0.14 weak negative connection between the home and away scores. This implies that a change of one goal in the away score causes a change of 0.14 goals in the home score.

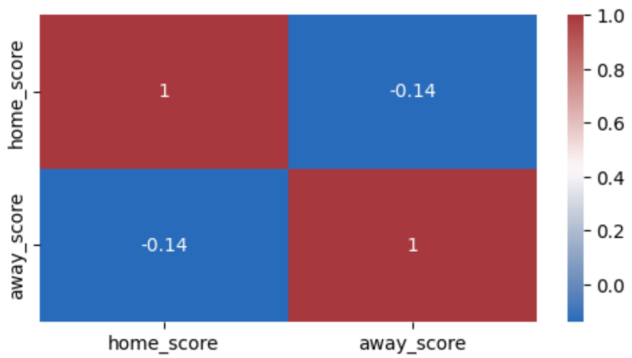


Fig 1 : Correlation heat map

2. The top 3 home teams with the highest number of matches played include Brazil, Argentina & Mexico with 595, 580 and 550 respectively.

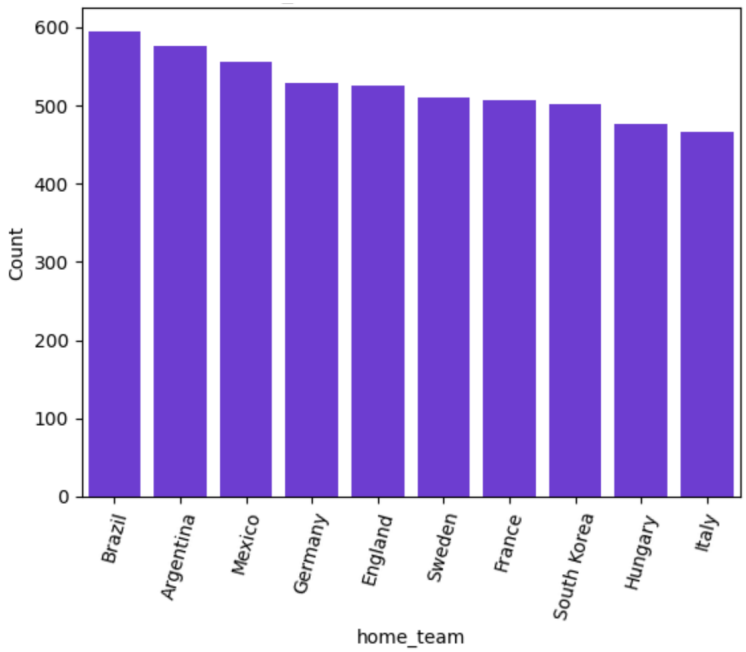


Fig 2 Showing the Top 10 Home Team with Numbers of Matches Played

3. The top 3 away teams with the highest number of matches played are Uruguay, Sweden and England

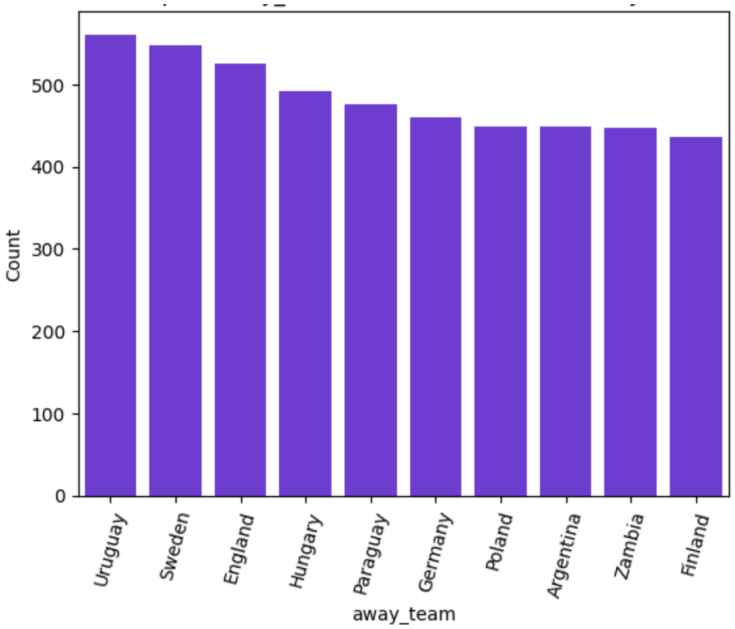


Fig 3 Showing the Top 10 Away Team with Numbers of Matches Played

4. Friendly has the highest number of matches played followed by FIFA World Cup qualification

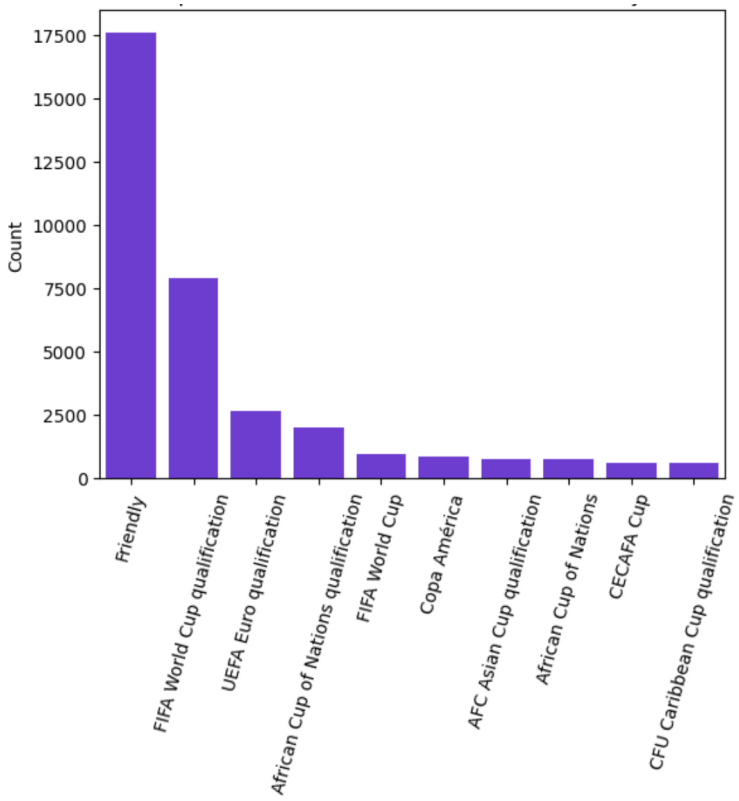


Fig 4 Showing the Top 10 Tournaments with Numbers of Matches Played

5. 49% of the total matches were won by home team, followed by 28% won by the away team

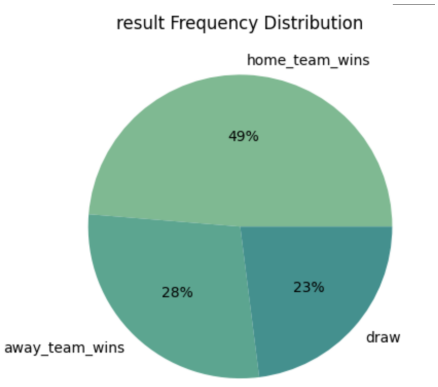


Fig 5 Showing the total games won by home, away and draw

Data Preprocessing

The following was done during the data preprocessing stage:

- The date column was dropped was dropped because it was irrelevant as we do not plan to carry out a time series analysis as the time spans are inconsistent.
- In our dataset, the distribution of the target column reveals a class imbalance issue. The home_team_wins class accounts for 48.71% of instances, while the draw and away_team_wins classes comprise 23% and 28.29% respectively. This imbalance may lead the model to overemphasize the trends of home team wins during training, resulting in poor performance on unseen data without similar patterns. To mitigate this, we will address the class imbalance by employing oversampling techniques to increase the representation of the minority classes.
- Data normalization was perfomed to ensure that all features have comparable ranges, preventing any particular feature from dominating the learning process. By bringing the features to a consistent scale, we enable the machine learning model to give equal consideration to each feature, avoiding biases that may arise due to variations in their original scales. Normalization plays a crucial role in creating a balanced learning environment, enhancing the model's ability to learn meaningful patterns and make accurate predictions.

Feature Encoding

In order to facilitate the modeling process, it is necessary to represent all variables in numerical form. As a result, we will apply specific encoding techniques to transform the data appropriately. For the "neutral" and "results" columns, we will utilize label encoding, while for the remaining columns, we will employ one-hot encoding. Label encoding assigns unique numerical labels to different categories in a column, while one-hot encoding creates binary columns to represent each category separately. By employing these encoding methods, we ensure compatibility with the machine learning algorithms that require numerical input for modeling purposes.

Data Splitting

To ensure that our model does not overly adapt to the training data and subsequently underperform on unseen data, we have implemented a data splitting strategy. The dataset has been divided into three distinct subsets: the training set, the testing set, and the validation set.

Model Training and Evaluation

Model Training: Machine learning revolves around understanding the patterns and behaviors exhibited by a dataset and then testing this understanding on new data. To accomplish this, the dataset was divided into three distinct sets: the training dataset and the testing dataset and validation dataset.

Baseline Model : In our project, we have chosen the K-Nearest Neighbors (KNN) algorithm as our baseline model. The KNN algorithm will serve as a foundational model from which we can assess the performance and effectiveness of more advanced techniques or models.The KNN model will provide a reference point to measure the progress and advancements made in our machine learning project.
We observed the following accuracy results: the training accuracy of the KNN model is 0.71, while the validation accuracy is 0.47.

It is evident that the KNN model demonstrates a higher accuracy on the training set compared to the testing set. This discrepancy indicates the presence of overfitting, whereby the model has learned the training data too well and struggles to generalize to new, unseen data.

We also used the **Random Forest Classifier** and **XGBoost Classifier** which gave the following results:

Random Forest Classifier: Training Accuracy : 1.0 , Validation Accuracy: 0.98. Which implies that the RF model performs well on training set and testing set.

XGBoost Classifier: Training Accuracy : 1.0 , Validation Accuracy: 1.0

The best performing model is the xgboost with a 100% accuracy on both the train and validation set. For that reason we will use it for our model evaluation.

Model Evaluation

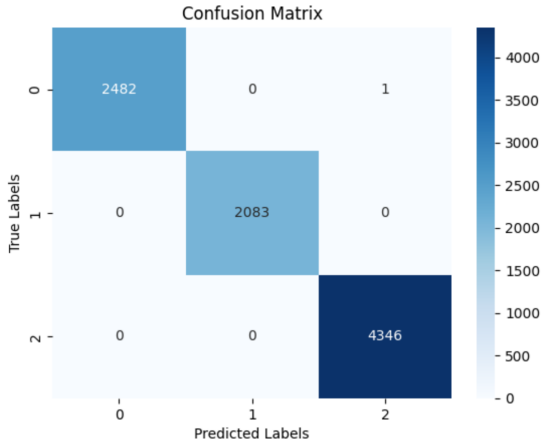
During this phase we assessed the performance and effectiveness of a machine learning model. It involved measuring how well the model performs on the test dataset or how accurately it can make predictions on new, unseen test data.

The results for the Model Evaluation using **XGBoost Classifier** which gave the highest accuracy on the training and testing accuracy are;

Training Accuracy : 1.0
Test Accuracy: 0.99.

Which implies that on the unseen test data, the model still performed exceptionally well. This demonstrates that the model has not been overfit.

Confusion Matrix:



Upon analyzing the confusion matrix, it is evident that only one result has been incorrectly classified, resulting in an overall accuracy of approximately 99%. The high accuracy indicates that the model is performing exceptionally well in correctly predicting the majority of instances. However, it is crucial to further investigate and understand the misclassified result to identify any potential patterns or underlying reasons for the misclassification. By examining such cases in detail, we can refine the model or adjust our approach to improve its performance and reduce misclassifications in future predictions.