FOOTBALL PREDICTIONS

AKSEL KAASIK, TOMI LAHE

INTRODUCTION

Football is one of the most popular sports globally, captivating millions of fans with its unpredictability and thrilling outcomes. As passionate sports enthusiasts, we were inspired to explore whether we could predict various aspects of a football match, such as the result, goals scored, and other key features. By diving into historical match data, we aim to uncover patterns and factors that influence match outcomes beyond just team strength.

This project involves processing and analyzing European football match data to extract meaningful insights and create predictive models. Our goal is not only to understand the nuances of football better but also to leverage machine learning techniques to predict match results and performance metrics with accuracy. Success in this project could provide valuable tools for fans and analysts while also offering a deeper appreciation for the complexities of the beautiful game.

DESCRIPTION OF THE DATA

Our data was taken from Kaggle [1], the creator of the dataset is Mart Jürisoo. The dataset contains 47917 different men football matches starting from 1872. After filtering for European matches we are left with 12 878 matches. The matches have various features like date, home team, away team, name of the tournament, location of the match and team scores.

OUR APPROACH

To predict football match outcomes, we took a multi-step approach:

1.DATA COLLECTION AND PREPROCESSING

- We gathered extensive data on football matches, including match results, teams, scores, and tournament details. Data was cleaned and filtered to focus only on European matches.

2.FEATURE ENGINEERING

- Key features, such as team performance, home vs. away advantage, and tournament type, were engineered to better capture the dynamics that influence match results.

3.MODEL SELECTION AND TRAINING

- We employed a variety of machine learning models, including Random Forest, HistGradientBoosting, and XGBoost, to predict match outcomes. We used a combination of classification algorithms with hyperparameter tuning for optimal performance.

4.ENSEMBLE LEARNING

- To enhance predictive accuracy, we implemented an ensemble approach by combining the predictions of multiple models using a Voting Classifier. This allows for the integration of different strengths from each model.

5.EVALUATION AND REFINEMENT

- We evaluated model performance using accuracy, classification reports, and confusion matrices, ensuring robustness and fairness across different match types and teams. Model refinement was performed based on the insights gathered from these evaluations.

6.EXPLORATORY DATA ANALYSIS (EDA)

- We analyzed trends in match outcomes, focusing on factors like the impact of neutral venues, the number of goals in friendlies vs. competitive tournaments, and team-specific patterns in major events like the UEFA Euro.

Through this structured approach, we have developed a predictive model somewhat capable of analyzing football match outcomes, accounting for key influencing factors and tournament dynamics.

DATA SCIENCE METHODS USED

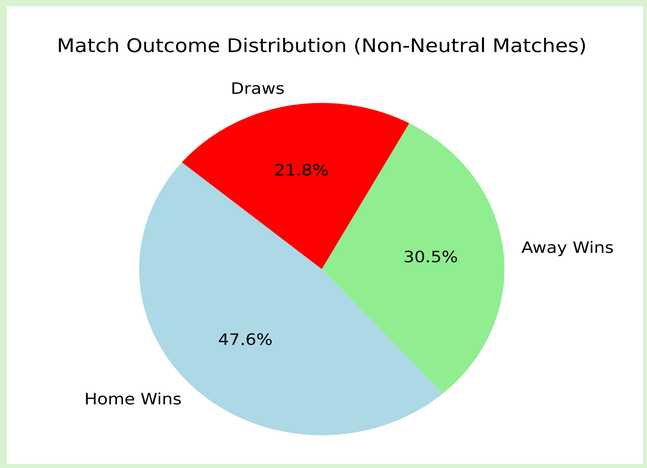
We filtered data, used feature engineering to add some important features. Also we tried different classification algorithms such as Random Forest Classifier, HistGradientBoostingClassifier and XGBoost. To improve model performance and to avoid overfitting and underfitting we did some hyperparameter tuning.

SOURCES

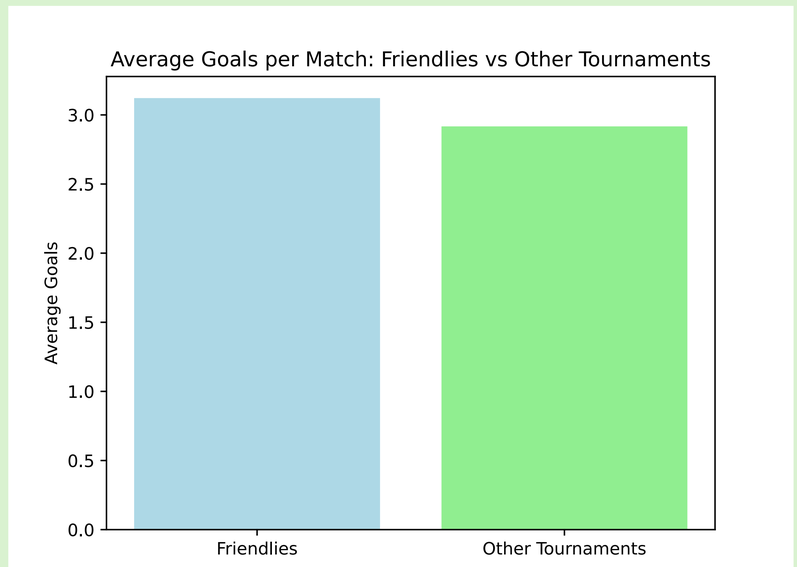
https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017?select=results.csv

RESULTS

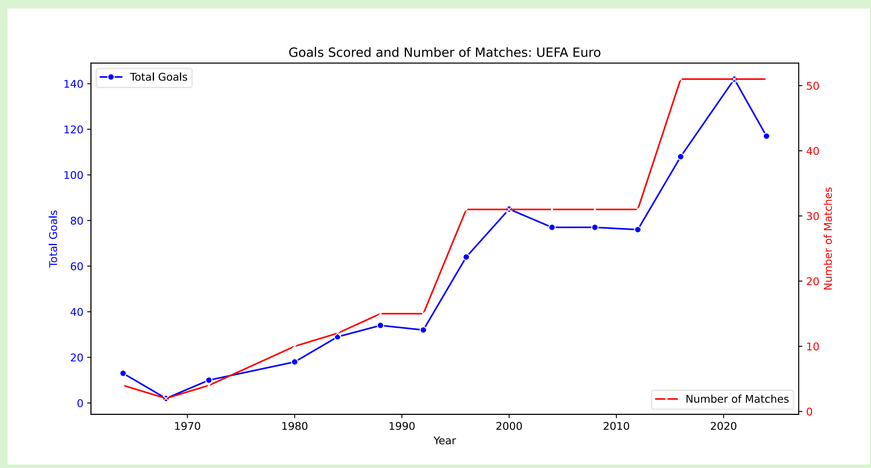
One of the factors we wanted to assess was the home advantage. A home advantage is when the match is being played at the home ground of one of the teams. Our analysis showed that from the dataset 47.6% of the times the home team won the match which shows that the home advantage plays a big role in the match outcome. This is why in league games and tournaments with knockout stages both teams play twice to nullify the home advantage.



Another vital part of a football match is of course goals. We wanted to see if teams are more likely to score in different types of tournaments. We compared competitive tournaments and to our surprise friendly matches tend to have more goals compared to competitive matches. This is likely due to taking less risks while playing in tournaments when being in the lead.



Finally we turned our heads to the crown jewel of European football, the UEFA European Football Championship. When playing around with our data we discovered that the amount of goals scored in UEFA Euro tournaments has been rising since the first tournament in 1960. We wanted to see if the rise in goals has been due to football becoming more fast-paced and better in attacking quality or if it's root lies somewhere else. We then graphed it side by side with the amount of matches in each tournament and revealed that the rise in goals has been the result of the number of participants increasing.



Our model performances were lower than expected. All models had an accuracy of **57% – 58%** and struggled in predicting draws correctly. A draw is the hardest outcome to predict and was also represented less in our dataset compared to wins and losses. Model performance could be improved by engineering more features that take into account lineups and player statistics but it wasn't in the scope of our current project.