1. Introduction

Predicting the outcomes of European football matches, deciding whether a team wins, draws, or loses, by using data science and machine learning, is a complex task.

For my 'final work' at Erasmushogeschool Brussel, studying applied informatics, I took it upon myself to take a deep dive into this daunting issue, aiming on achieving comparable results than existing models. My initial goal was to produce a model that could achieve a steady average accuracy of at least 75%. Besides the model's performance, as an aspiring data scientist, the main priority of my final work was the AI knowledge acquisition itself.

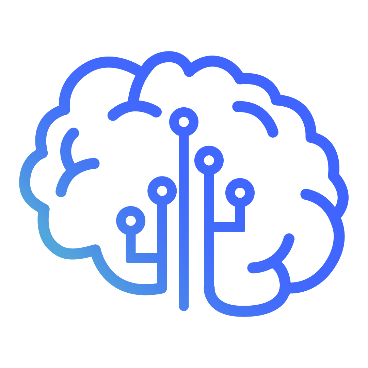
The process of developing a comprehensive model, built in Jupyter Notebook using Python, and performing at a satisfying level towards my preset goals, turned out to be an iterative process of analysing, developing and testing, and proved to be a challenging endeavour at times frustrating and intimidating.

Starting from a baseline model with minimal hyperparameter tuning, and without any feature engineering whatsoever, the progress made by extensive data analysis and thorough feature development was impressive.

The final outcome of my work is an ensemble[[1]](#footnote-1) model, which I named 'football brain', contrived of 7 distinctive classification models, performing in a binary classification strategy, displaying an average accuracy of around 83%, surpassing the expectations of myself and the people I shared my findings with.

An important note to make is that the development of this project will not cease after finishing my degree. The project will continue to be improved in the future. New engineering techniques will be tested, datasets will continue to be analysed, improved and expanded, and the path of moving the whole project to an 'MLOps[[2]](#footnote-2)' environment will be investigated.

Happy to share that during the process of developing the project, interest has grown from an independent investor, who is interested in monetizing the project by packaging it into a marketable online application.



1. Methodology
   1. Data Collection

During the initial phase of development, the collection of data was a critical task, undertaken by scraping from the football statistics website www.fbref.com. FBref, renowned for its comprehensive tracking of football team and player statistics globally, presented a significant source of valuable data. However, the website did not offer a straightforward method for downloading data in CSV format. Consequently, an intricate web scraper was constructed in Jupyter Notebook, designed to gather match data on a weekly basis. This process involved considerable effort to ensure the accuracy and reliability of the scraped data.

Upon thorough evaluation and deliberation, it became evident that the datasets obtained from FBref.com were insufficient for my needs. The limitations of the data, coupled with the complexities involved in maintaining the scraper, led to the decision to discontinue using these datasets.

A screenshot of a web page

Description automatically generatedIn the search for a more suitable data source, I discovered www.football-data.co.uk, a website that not only mirrored the statistical depth of FBref.com but also offered additional insights. Notably, next to offering ready to download CSV files for competitions of over 25 countries, this alternative source provided bookmakers' odds for both past and upcoming matches. These odds are instrumental in identifying which predictions have the potential to be most profitable, thereby enhancing the predictive models' effectiveness. The availability of such critical data points could significantly influence the accuracy and financial viability of our predictions.

* 1. Types of Data Collected

The datasets provided by football-data.co.uk encompass a vast amount of information collected for every match. However, not all of this data is relevant to the project. Therefore, only a selection of the data that is most pertinent to our analysis is used.

Div = League Division

Date = Match Date (dd/mm/yy)

Time = Time of match kick off

HomeTeam = Home Team

AwayTeam = Away Team

FTHG = Full Time Home Team Goals

FTAG = Full Time Away Team Goals

FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)

HTHG = Half Time Home Team Goals

HTAG = Half Time Away Team Goals

HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

Referee = Match Referee

HS = Home Team Shots

AS = Away Team Shots

HST = Home Team Shots on Target

AST = Away Team Shots on Target

* 1. Data Cleaning and Preprocessing

sdsds

* 1. Handling Missing Data

The limited subset of data used from the datasets provided by football-data.co.uk had little to no missing information. The only columns with a minimal amount of missing data were HS, AS, HST, and AST. To address these gaps, I employed a technique called imputation, where the empty values in these columns are filled with their respective means. This method ensured that my dataset remained complete and reliable for analysis and training.

* 1. Feature Selection and Engineering

xcxcxc

1. A machine learning approach that combines multiple individual models to improve predictive performance and robustness compared to any single model alone. [↑](#footnote-ref-1)
2. The practice of streamlining and automating the deployment, monitoring, and management of machine learning models in production to enhance reliability and efficiency. [↑](#footnote-ref-2)