A multi-task learning approach to predict the outcome of a football game using human experts input

Contents

In this work, we propose to exploit a new type of data to capture the game environment. We will extend the methodology described in and build a multi-head multi task deep learning network able to digest both Text data provided by human football experts and more standard time series tabular data. The results will be compared to traditional and machine learning approaches described in the literature.

Literature review

Data analytics has been adopted by most industries, including football. However, the availability of football match data has increased in recent years, allowing anyone involved in football to gain hidden knowledge and predict outcomes.

In our literature review, we will concentrate on models that predict the outcomes of football matches based on various baselines. These baselines are statistical and machine learning approaches that rely on statistical machine learning based on historical data about individual teams.

Statistical approaches

The very first generation of football-related models focused on the distribution of the number of goals scored during a football game.
Starting by *(Maher, 1982)* included a general model that uses an independent poisson distribution to approximate a team's scoring frequency. However, this model assumes that each team's home, away, attack, and defense parameters are constant.
(Dixon & Coles, 1997) claimed in their paper that the previous model has certain limitations, such as low score results (0-0, 1-0, 0-1, and 1-1) being inherently unreported and each team's scoring and defending abilities being regarded as constant over time.
They proposed a new model with some specific enhancements. This model is a **bivariate poisson distribution** with parameters based on past performance for the number of goals scored by each team. This implies that recent matches have a greater impact on strength estimates.
Unlike previous models that used a **Poisson distribution**, *(Boshnakov et al., 2017)* proposed a new approach that uses an independent and identically distributed **Weibull distribution** for the number of goals scored by the home and away teams in a match. Furthermore, they allowed for dependence between the goals scored by the two teams by using a copula to generate a bivariate distribution with positive or negative dependence. They compared this **bivariate Weibull** model to the others and discovered that it provides a good fit to **English Premier League** data from 2006/2007 to 2015/2016 seasons and achieves positive betting returns.

Bookmakers predictions

Making bets on the outcome of football matches has a long history around the world, and typically entails selecting matches that are thought to be the most likely to end in a draw. Bookmakers provide odds on a match's various outcomes and the simplest version of this relies exclusively on the results which can be a win by either the team playing at home or the team playing away, or a draw. More complex bets on the score or the half-time and full-time results are also available.
Football match predictions is a very complex problem in which the outcome of the game is predicted based on the teams' previous performance and the relative abilities of the players in the teams. This implies that the team with the best players should win.
Bookmakers set their odds based on this challenge and employ sophisticated pricing models that assign "odds" to different outcomes in order to maximize their chances of profit.

According to *(Beal et al., 2019)*, bookmakers' accuracy was around 67% for **American football**, 74% for **basketball**, 64% for **cricket**, 61% for **baseball**, and only 54% for **football** during the 2017/2018 season. This is due to the fact that the frequency of goals is far lower than the frequency of points scored in the other sports.

Machine Learning approaches

To date, **probabilistic methods** have yielded limited results and appear to have reached a plateau in terms of accuracy because team performance is dependent not only on team abilities but also on a variety of dynamic factors such as team configurations, player health, match location, weather, team strategies, and other external factors. As a result predicting football match outcomes becomes a very complex computational problem.
Among the **machine learning** methods interested in football match predictions, we highlight **sentimental analysis** of social media platforms. These studies focus on opinion aggregation and underscore the importance of hidden information contained in the sentiment of fans publications.
True, fans tweets are unlikely to influence game outcomes, unless they are used to exceptionally motivate or demotivate a team. For example, supporters use social media platforms like **Twitter** to express their personal feelings, primarily about the team they are following and their next opponent's strengths, weaknesses, and prospects, and this can sometimes help establish betting odds.
This method is used by *(Schumaker et al., 2016)* to predict **English Premier League** matches by analyzing fans tweets during the final three months of the 2013–2014 from February 16 through May 11 2014. It achieved an accuracy of 50%,this system has only two possible outcomes: home team win or away team win. If the model's number of Home or Away tweets was zero, the match was not considered for that model. However, it was discovered that sentiment was unable to recognize a draw outcome, so this category was dropped.
As well as that, *(Baboota & Kaur, 2019)* investigates the application of machine learning techniques to predict football match outcomes and compares the results to bookmakers. They employ feature engineering and exploratory data analysis to identify the feature set that contains the most important factors for predicting match outcome. The authors put **Gaussian naive Bayes**, **SVM**, **Random forest**, and **Gradient boosting** to the test. They used training data from 2005 to 2014 in the **EPL** (English Premier League) and discovered that the **Gradient boosting** method performed the best with an accuracy of 56.7 %

Moreover, to address weaknesses described above, recently, a new technique has emerged *(Beal et al., 2020)* that incorporates human expertise and judgment, such as media information, rather than just basic performance statistics, which has helped improve prediction accuracy.
These new baselines used a dataset of 6 seasons of **English Premier League** games from 2013/14 to 2018/19, including football match data, the **Guardian** previews and predictions from bookmakers odds for 1770 games, employing both statistical machine learning techniques and **Natural Language Processing**.
When **NLP** methods, **statistical approaches**, and **bookmakers predictions** were compared to the ensemble learning approach which combines the first three techniques and uses a **Random forest classifier**, the results revealed that these methods could be improved.
It achieved an accuracy of 63.2 %, a 10.8 % increase over **the bookmakers'** accuracy(52.43%), 4.1 % more than *(Dixon & Coles, 1997)* (59.11%) and a 13 % improvement over the sentiment analysis approach in *(Schumaker et al., 2016)*.
Experiments revealed that using the ensemble model increases the likelihood of predicting draws and longshot results. This is especially true when the text vectors model identifies more longshots (38.9 %) by taking into account human input, whereas the first three models are typically poor at predicting these events.

Our contribution

In most **machine learning** situations, we are only concerned with one problem at a time. Whatever the task, the problem is typically outlined as using data to solve or optimize a single metric at a time. However, this approach will eventually reach a performance limit, which is often due to the size of the dataset or the model's ability to derive meaningful representations from it.
On the other hand, **multitasking learning** is a machine learning approach in which we attempt to learn multiple tasks at the same time while optimizing multiple loss functions.Instead of training separate models for each task, we expect a single model to learn how to perform all tasks simultaneously.
The model uses all of the data available in the various tasks to learn generalized representations of the data that are useful in multiple contexts during this process.
Our contribution to this project is to extend the methodology described in *(Beal et al., 2020)* which consists of developing a model that ensembles three separate models trained separately and to build a **multi-head**, **multi-task** deep learning network *(Vafaeikia et al., 2020)* capable of capturing the game environment as well as more standard time series tabular data. This dataset includes, until now, 3880 **English Premier League** match previews from 2009 to the present.
The findings will be compared to traditional and machine learning approaches described in the literature, and the following question will be addressed:
Do football experts' human analysis influence the predictions of football matches?

References

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003
Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2020). Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model. *arXiv:2012.04380 [Cs]*. http://arxiv.org/abs/2012.04380
Beal, R., Norman, T. J., & Ramchurn, S. D. (2019). Artificial intelligence for team sports: a survey. *The Knowledge Engineering Review*, *34*. https://doi.org/10.1017/S0269888919000225
Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*(2), 458–466. https://doi.org/10.1016/j.ijforecast.2016.11.006
Dixon, M. J., & Coles, S. G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 265–280. https://doi.org/10.1111/1467-9876.00065
Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118. https://doi.org/10.1111/j.1467-9574.1982.tb00782.x
Schumaker, R. P., Jarmoszko, A. T., & Labedz, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems*, *88*, 76–84. https://doi.org/10.1016/j.dss.2016.05.010
Vafaeikia, P., Namdar, K., & Khalvati, F. (2020). A Brief Review of Deep Multi-task Learning and Auxiliary Task Learning. *arXiv:2007.01126 [Cs, Stat]*. http://arxiv.org/abs/2007.01126

The Guardian Scraper

> Scraping Premier League Previews from the Guardian.

To reach the aim of our project, we have begun the first task, which is to collect football experts' comments and data from English Premier League matches.
For this task, we will use match previews from "The Guardian." It goes back far enough, from 2009 to today, to allow us to integrate deep neural networks.
Indeed, "The Guardian's" football experts publish previews every week, usually two or three days before the matches.
In this regard, we began by creating a data extraction tool that will allow us to extract this information on a regular basis.
The information to be extracted is as follows:

- The names of the competing teams.
- The date of the game
- The identity of the referee
- The stadium's name
- Sports odds that will be converted to decimal format
- The football expert's text
- The text's author.

A preliminary examination of the Guardian's website

Issues	Solutions
4 possible formats for previews(old format, new format,Cup's format and a particular format)	Select the appropriate html tags
Preview titles are not the same (we can find Squad Sheets or match preview)	Pick only the names of the teams and eliminate the rest
The date of the match is not always available	Pick the preview date
The order of the elements and labels are not the same	Using regex patterns to get information
Missing values for betting odds	We treat the general case separately and we set up specific regex patterns for these particular cases
Odds format is different	We treat the general case separately and we set up specific regex patterns for these particular cases
We can find non-numeric values for Odds like (Evens,evens,Eve,odds-on)	Replace evens by 1-1
There are some previews that don't have author and text	For previews that have no text, we put None (not available)
The existence of previews for the FA CUP,Carabao Cup,Champions league,World Cup	Filter previews by checking if the match exists in "Opta" database, and pick only Premier League match
We are not sure if the names of the teams are the same as the ones in Opta	Set up a dictionary or check manually to map teams to their IDs
When we send many requests, the guardian server blocks your IP address, which is interpreted as a DDOS attack	Do a sleep of a random x seconds between requests or change IP address and work with rotating proxy

The guardian previews are spread across several pages, so our extraction tool will go through all of them, extracting the previews as it goes.
Indeed, the scraper extracts information from **Premier League** match previews using two methods:
The first method is to use the api provided by **The Guardian's** website, by introducing a connection key and a well-parsed portion of the link.
For example, a link is formatted as https://www.theguardian.com/football/2022/feb/12/newcastle-aston-villa-match-preview-premier-league, **the guardian** API only requires "football/2022/feb/12/newcastle-aston-villa-match-preview-premier-league."
However, this method is not always reliable because the API is not always functional for some previews. So, in this case, our scraper employs its second method, which is the traditional approach of scraping the html format of the pages.
We identified some special cases by analyzing the website; there are some changes in form and content, so we attempted to find a set of regular expressions that dealt with generic and special cases in order to locate the various information.
Following the data extraction stage, information will be normalized and cleaned, and betting odds extracted from the previews will be converted to a decimal format.
Then, to ensure that we are only extracting **English Premier League** matches, we map the extracted preview by its equivalent in a database called **Opta** by introducing the names of the two opposing teams, the competition id, and the closest game date to the date of the preview. This mapping returns the match id and the date of the game. However, **Opta** is a trustworthy database provided by **Opta Sport**, the pan-European leader in the supply of sports information data. Moreover, prior to this phase, we were unsure of the correspondence between the names of the teams written by **the guardian** and the names of the teams in **Opta**. For example, a single team can be referred to in several ways, such as **Manchester United, Manchester Utd, Man United, Man Utd** and so on.
To address this issue, we created a dictionary containing the various possible names of **103 English teams** competing in the first and second divisions. This dataset was created in two ways: the first using **Opta**, which provides the various team aliases, and the second manually, in which we added other nicknames that were not provided.
Finally, once the mapping process is done, this information will be gradually stored in a **Mongodb** database, specifically in a collection called **Previews**.
It should be noted that when the scraper is launched, it checks the last date stored in the **Previews** collection to determine at what level it will stop.
This scraper has extracted data from 2009 to the present and is **operational in a production environment** .
Our package is **open source** and freely available to all developers on **Github** and **Pypi** (default software repository for Python developers to store created Python programming language).

Example of the scraper output

	gameId	homeTeam	awayTeam	textPreview	author	venue	referee	odds	oddsHomeTeam	oddsAwayTeam	oddsDraw	gameDate	previewDate
0	695119	West Bromwich Albion	Everton	Pepe Mel takes charge of West Bromwich Albion ...	Harvey Taylor	The Hawthorns	M Oliver	[12-5, 5-4, 12-5]	3.400000	2.25	3.4	2014-01-20 20:00:00	2014-01-17 18:47:09
1	695118	Swansea City	Tottenham Hotpsur	Tottenham travel to south Wales with cause for...	Harvey Taylor	Liberty Stadium	M Atkinson	[11-5, 7-5, 12-5]	3.200000	2.40	3.4	2014-01-19 13:30:00	2014-01-17 18:42:00
2	695115	Manchester City	Cardiff City	Ole Gunnar Solskjaer may like to remind his Ca...	Jamie Jackson	Etihad Stadium	N Swarbrick	[1-7, 22-1, 17-2]	1.142857	23.00	9.5	2014-01-18 15:00:00	2014-01-17 18:36:58
3	695113	Crystal Palace	Stoke City	Tony Pulis spent seven years in charge of Stok...	Harvey Taylor	Selhurst Park	M Clattenburg	[11-8, 13-5, 12-5]	2.375000	3.60	3.4	2014-01-18 15:00:00	2014-01-17 18:31:38
4	695116	Norwich City	Hull City	Norwich will hope to end their six-game winles...	Harvey Taylor	Carrow Road	H Webb	[7-5, 9-4, 12-5]	2.400000	3.25	3.4	2014-01-18 15:00:00	2014-01-17 18:18:33

awayTeam	textPreview	author	venue	referee	odds	oddsHomeTeam	oddsAwayTeam	oddsDraw	gameDate	previewDate	previewLink
Everton	Pepe Mel takes charge of West Bromwich Albion ...	Harvey Taylor	The Hawthorns	M Oliver	[12-5, 5-4, 12-5]	3.400000	2.25	3.4	2014-01-20 20:00:00	2014-01-17 18:47:09	https://www.theguardian.com/football/2014/jan/...
Tottenham Hotpsur	Tottenham travel to south Wales with cause for...	Harvey Taylor	Liberty Stadium	M Atkinson	[11-5, 7-5, 12-5]	3.200000	2.40	3.4	2014-01-19 13:30:00	2014-01-17 18:42:00	https://www.theguardian.com/football/2014/jan/...
Cardiff City	Ole Gunnar Solskjaer may like to remind his Ca...	Jamie Jackson	Etihad Stadium	N Swarbrick	[1-7, 22-1, 17-2]	1.142857	23.00	9.5	2014-01-18 15:00:00	2014-01-17 18:36:58	https://www.theguardian.com/football/2014/jan/...
Stoke City	Tony Pulis spent seven years in charge of Stok...	Harvey Taylor	Selhurst Park	M Clattenburg	[11-8, 13-5, 12-5]	2.375000	3.60	3.4	2014-01-18 15:00:00	2014-01-17 18:31:38	https://www.theguardian.com/football/2014/jan/...
Hull City	Norwich will hope to end their six-game winles...	Harvey Taylor	Carrow Road	H Webb	[7-5, 9-4, 12-5]	2.400000	3.25	3.4	2014-01-18 15:00:00	2014-01-17 18:18:33	https://www.theguardian.com/football/2014/jan/...

Task 2 NLP Review

After completing our first task, which was to extract previews from the Guardian website, we began our second task, which was to reproduce the results of the *(Beal et al., 2020)* article by processing and analyzing the texts that we had extracted. The article's authors proceeded as follows:

- **Information extraction**: they extracted the main features of each sentence in the article's text.
- **Allocation of Text Context**: each sentence is assigned to a team.
- **Text Vectorisation**: they converted the sentences into vectors using a Count Vectorizer technique to have a numerical representation of the words in a sentence.
- **Prediction**: Once the feature set for each game is formed, they trained a Random Forest model using historic data and the numerical representation of the words in the sentence.

Understanding the fundamental concepts and learning the "spacy" tool were required to properly assimilate these operations.
Spacy is an open source Python library for natural language processing that can be used to extract information from text.

The following techniques were used in this task:

- Named Entity Recognition
- Spacy Entity Ruler
- Count vectorization

Named-Entity Recognition (NER)

The task of identifying and categorizing key information in text is known as Named Entity Recognition (NER). It is also known as entity extraction or identification. Each detected entity is assigned to a predefined category. An NER model, for example, may detect the word “Mark” in a text and classify it as a “Person.”

example :

```
Pepe Mel takes charge of West Bromwich Albion for the first time and the part-time novelist will be hoping for a fan
tasy start against opponents managed by a fellow Spaniard. Roberto Martínez's side have impressed this campaign part
ly due to the influence of Ross Barkley, who is out for some time with a broken toe. The new signing Aiden McGeady c
ould feature for the visitors. Harvey Taylor
```

```
doc = nlp(text)
# check for entities
for ent in doc.ents:
    print(ent,ent.label_)
```

```
Pepe Mel PERSON
West Bromwich Albion ORG
first ORDINAL
Spaniard NORP
Roberto Martínez's PERSON
Ross Barkley PERSON
Aiden McGeady PERSON
Harvey Taylor PERSON
```

In our case, identifying the names of the teams in the previews is a critical task that will allow us to extract the main features in our text.
This operation is not possible with the standard spacy NER because of errors in entity detection; spacy can consider a team name to be a person and vice versa.
To address this issue, we decided to build a model that will allow us to detect our own entities.

Train a model to detect custom entities

Before we implemented our model, which will allow us to automatically detect the names of the teams, we fed it a training dataset with labels generated by an external text annotation tool.
This annotation identifies the custom entities that our model will learn during its training.

Pepe Mel COACH takes charge of West Bromwich Albion WEST BROMWICH ALBION for the first time and the part-time novelist will be hoping for a fantasy start against opponents managed by a fellow Spaniard. Roberto Martínez COACH 's side have impressed this campaign partly due to the influence of Ross Barkley PLAYER , who is out for some time with a broken toe. The new signing Aiden McGeady PLAYER could feature for the visitors. AWAY TEAM Harvey Taylor

Ole Gunnar Solskjaer COACH may like to remind his Cardiff City CARDIFF CITY players of their 3-2 win in the reverse of this fixture in August , yet that was in Manuel Pellegrini COACH 's second game in charge and since then the manager has finely tuned Manchester City MANCHESTER CITY into a near-terrifying prospect , having amassed 99 goals in all competitions. It would be a major shock if Cardiff CARDIFF CITY were even to garner a point. Jamie Jackson

we test again

```
# check again the new entities
doc = nlp(text)
# check for entities
for ent in doc.ents:
    print(ent,ent.label_)
```

```
Pepe Mel takes COACH
West Bromwich Albion for WEST BROMWICH ALBION
Roberto Martínez COACH
Ross Barkley PLAYER
Aiden McGeady PLAYER
```

Entity ruler

Using token-based rules or exact phrase matches, the entity ruler allows us to add spans to the Doc entities. It can be used in conjunction with the statistical EntityRecognizer to improve accuracy, or it can be used on its own to implement a rule-based entity recognition system.

We took advantage of the dataset that we have which contains the teams and their different names. In this sense we have linked each name or nickname of a team to its main entity

	ID	optaId	name	symid	shortClubName	optaName	whoScoredName	sofifaName	statsName	inStatH
0	9e78bbc137fd00c66162080bc9e987e67297643dc50616...	3	Arsenal	ARS	Arsenal	Arsenal	Arsenal	Arsenal	Arsenal	Arsen
1	0ef9883721814dd09038659130c61c76f18976cb7b8e86...	47	Portsmouth	POR	Portsmouth	Portsmouth	Portsmouth	Portsmouth	Portsmouth	Portsr
2	eb89c068ca204a72408360450847a990c97c5b5ff0ec9f...	110	Stoke City	STK	Stoke	Stoke City	Stoke	Stoke City	Stoke City	
3	c1a486f8ca465e58b6301f038e754058986187d454110c...	56	Sunderland	SUN	Sunderland	Sunderland	Sunderland	Sunderland	Sunderland	Sunde
4	0db353094ccf93e0005cf378ea862b56e77cacc57b7c5e...	111	Wigan Athletic	WIG	Wigan	Wigan Athletic	Wigan	Wigan Athletic	Wigan Athletic	W At

For example the nickname Spurs is now detectable in the text that is linked to the Tottenham Hotspur entity

```
# example
doc = nlp("Spurs go go")
for ent in doc.ents:
    print(ent.text,"-------->",ent.label_)
```

```
Spurs --------> TOTTENHAM HOTSPUR
```

Get the names of the coaches

We also noticed that in most of the previews, we find the names of the managers but not the names of the teams, so to ensure the extraction of information, we used a database that we have that contains a history of managers for each team.
As a result, it is now easier to identify the section of the text that refers to one of the two teams.

Example of the dataset

	Team	Manager	country	startDate	endDate
0	Arsenal	Arsène Wenger	France	1996-10-01	2018-06-30
1	Arsenal	Unai Emery	Spain	2018-07-01	2019-11-28
2	Arsenal	Freddie Ljungberg	Sweden	2019-11-29	2019-12-21
3	Arsenal	Mikel Arteta	Spain	2019-12-22	2023-06-30
4	Portsmouth	Harry Redknapp	England	2005-12-07	2008-10-25

The final output: for each preview, we have the coaches of each team.

venue	referee	odds	oddsHomeTeam	oddsAwayTeam	oddsDraw	gameDate	previewDate	previewLink	homeTeamCoach	awayTeamCoach
The Hawthorns	M Oliver	[12-5, 5-4, 12-5]	3.400000	2.25	3.4	2014-01-20	2014-01-17 18:47:09	https://www.theguardian.com/football/2014/jan/...	Pepe Mel	Roberto Martínez
Liberty Stadium	M Atkinson	[11-5, 7-5, 12-5]	3.200000	2.40	3.4	2014-01-19	2014-01-17 18:42:00	https://www.theguardian.com/football/2014/jan/...	Michael Laudrup	Tim Sherwood
Etihad Stadium	N Swarbrick	[1-7, 22-1, 17-2]	1.142857	23.00	9.5	2014-01-18	2014-01-17 18:36:58	https://www.theguardian.com/football/2014/jan/...	Manuel Pellegrini	Ole Gunnar Solskjær
Selhurst Park	M Clattenburg	[11-8, 13-5, 12-5]	2.375000	3.60	3.4	2014-01-18	2014-01-17 18:31:38	https://www.theguardian.com/football/2014/jan/...	Tony Pulis	Mark Hughes
Carrow Road	H Webb	[7-5, 9-4, 12-5]	2.400000	3.25	3.4	2014-01-18	2014-01-17 18:18:33	https://www.theguardian.com/football/2014/jan/...	Chris Hughton	Steve Bruce

Previews Preprocessing

First of all, beginning with the tokenization step, which is the task of chopping up texts into pieces in order to remove stop words such as (the,a,an,so,what..). We also removed all punctuation because it isn't important in the text, and then we used a lemmatization technique that allows for lexical processing, such as (runs, running,ran) => run.

Example of text

```
West Ham have never won successive games in the Premier League under Sam Allardyce. Now would be a good time to star
t. Last week's victory at Cardiff lifted West Ham out of the bottom three but they remain in a precarious position a
nd Upton Park is bound to be gripped by nerves against Newcastle, who have lost their past four matches. The welcome
news for West Ham is that Andy Carroll will be on the bench again as he continues his comeback from injury and he sh
ould be joined there by Ravel Morrison, despite talk linking him with Fulham. Jacob Steinberg
```

```
for p in normalized_preview:
    doc = nlp(p)
    displacy.render(doc,'ent')
    print("-------------")
```

West Ham WEST HAM UNITED never win successive game Premier League Sam Allardyce

---

good time start

---

week victory Cardiff CARDIFF CITY lift West Ham WEST HAM UNITED remain precarious position Upton Park STADUIM bind grip nerve Newcastle NEWCASTLE UNITED lose past match

---

welcome news West Ham WEST HAM UNITED Andy Carroll PLAYER bench continue comeback injury join Ravel Morrison PLAYER despite talk link Fulham FULHAM

We continue our normalization and move on to the next step, which is detecting the names of the two teams, the names of the coaches, and changing their names by hometeam, awayteam, homecoach, and awaycoach.The reason for this is so that our model's predictions can generalize.
We noticed that the words 'hosts,' 'home side,' and 'visitors,' which refer to the home team and away team, are frequently used in the previews, and they have been changed.

We take the same example:

```
West Ham have never won successive games in the Premier League under Sam Allardyce. Now would be a good time to star
t. Last week's victory at Cardiff lifted West Ham out of the bottom three but they remain in a precarious position a
nd Upton Park is bound to be gripped by nerves against Newcastle, who have lost their past four matches. The welcome
news for West Ham is that Andy Carroll will be on the bench again as he continues his comeback from injury and he sh
ould be joined there by Ravel Morrison, despite talk linking him with Fulham. Jacob Steinberg
```

Preview after cleanning

```
['West Ham never win successive game Premier League Sam Allardyce',
 'good time start',
 'week victory Cardiff lift West Ham remain precarious position Upton Park bind grip nerve Newcastle lose past match
',
 'welcome news West Ham Andy Carroll bench continue comeback injury join Ravel Morrison despite talk link Fulham',
 'Jacob Steinberg']
```

Preview after normalization

```
['hometeam never win successive game Premier League homecoach',
 'week victory Cardiff lift hometeam remain precarious position Upton Park bind grip nerve awayteam lose past match
',
 'welcome news hometeam Andy Carroll bench continue comeback injury join Ravel Morrison despite talk link Fulham']
```

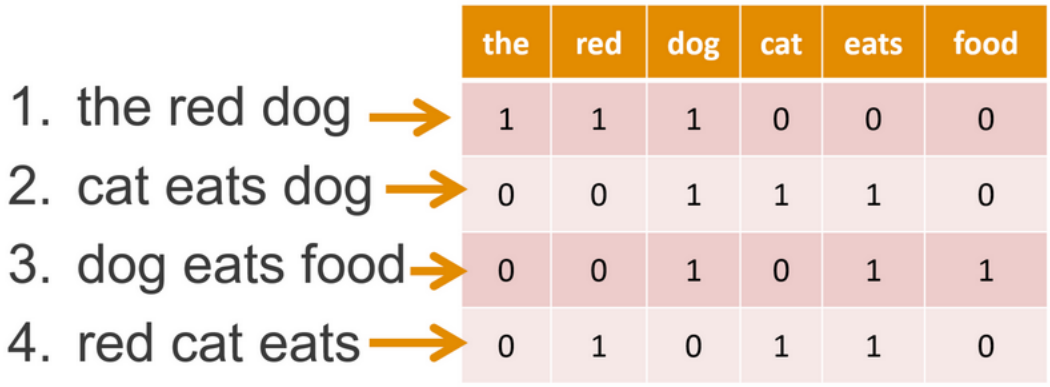
Allocation of texts

This section consists of assigning each sentence to the appropriate team. In a preview, for example, the journalist may discuss squad A or team B. As a result, we will have three columns: one for sentences about team A, another for sentences about team B, and a third for sentences about both teams at the same time.

Modeling

Vectorization

When the text processing and allocation phases are completed, it is time to begin the modeling phase.
However, our model will not be able to understand these raw texts, so we must convert them into vectors, which are digital representations of these character strings.Here, the goal is to extract some textual feature so that the model can learn.
Among the vectorization techniques, we highlight the bag of words: it is a very simple technique that calculates the vectors of a text based on the frequency of vocabulary words. It is simple to interpret and only refers to the frequency of vocabulary words in a given document.
As a result, articles, prepositions, and conjunctions that do not contribute much to meaning are just as important as adjectives or verbs.
An example for the count vectorization technique:



There are other techniques that, in general, work better in machine learning models to address this issue such as TF-IDF: term frequency-inverse document frequency.
The idea behind the TF-IDF approach is that words that appear less frequently in all documents but more frequently in individual documents contribute more to classification. these terms can be calculated as follows:

```
TF  = (Frequency of a word in the document)/(Total words in the document)

IDF = Log((Total number of docs)/(Number of docs containing the word))
```

It should be emphasized that for this work, we will utilize the count vectorizer approach to vectorize the preview texts.
This function comprises certain hyperparameters that must be fixed and find the best combinations in order to increase the quality of the vectors.
Among these hyperparameters, we can find:

- stop_words: CountVectorizer provides a predefined set of stop words; in our case, we can specify 'english.'
- ngram_range: the number of word combinations to consider, for example (1,1) takes only tokens, whereas (1,2) specifies that we want to consider both unigrams (single words) and bigrams (combination of 2 words)
- min_df:stands for minimum document frequency; it disregards words with a document frequency that is strictly lower than the specified threshold.

Get target values

To enable our model to train and make predictions, we must first provide the target values (the outputs of the matches).

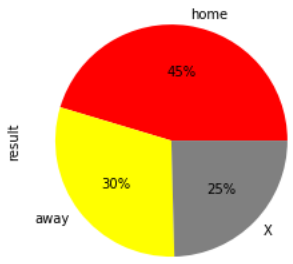
We have set two target values:

- The outcome of a match: home win, away win, draw
- The goal difference: the difference in goals scored

	oddsAwayTeam	oddsDraw	gameDate	previewDate	previewLink	homeTeamCoach	awayTeamCoach	normalizedText	goalDifference	result	score
	2.25	3.4	2014-01-20	2014-01-17 18:47:09	https://www.theguardian.com/football/2014/jan/...	Pepe Mel	Roberto Martínez	homecoach take charge hometeam time time novel...	0	X	1-1
	2.40	3.4	2014-01-19	2014-01-17 18:42:00	https://www.theguardian.com/football/2014/jan/...	Michael Laudrup	Tim Sherwood	awayteam travel south Wales cause optimism.hom...	-2	away	1-3
	23.00	9.5	2014-01-18	2014-01-17 18:36:58	https://www.theguardian.com/football/2014/jan/...	Manuel Pellegrini	Ole Gunnar Solskjær	awaycoach like remind awayteam player win reve...	2	home	4-2
	3.60	3.4	2014-01-18	2014-01-17 18:31:38	https://www.theguardian.com/football/2014/jan/...	Tony Pulis	Mark Hughes	homecoach spend year charge awayteam turn midt...	1	home	1-0
	3.25	3.4	2014-01-18	2014-01-17 18:18:33	https://www.theguardian.com/football/2014/jan/...	Chris Hughton	Steve Bruce	hometeam hope end game winless run awayteam wi...	1	home	1-0

The proportions of the results

It is worth noting that the class distribution of English Premier League games that we have from 2009 to 2022 is 45% home wins, draws 25% and away wins 30%.



Split previews into train and test dataset

Before setting up a machine learning model, we must divide our previews into train and test data. the train dataset is used to train the machine learning model and the test dataset is to assess the fit, which is data that the model has never seen before. To accomplish this, we will split the data into 70% train and 30% test without applying a shuffle to avoid distorting the temporal order of the matches.

The classifier

There are various methods in machine learning for dealing with classification or regression problems that are highly fascinating to try.
In this work, we will use a random forest classifier that takes vectorized texts as input to predict the outcomes of football matches(Home win, Away win, Draw).
A random forest's fundamental notion is to aggregate a large number of individual decision trees into a single model that function as an ensemble. All individual tree projections are pooled, and the class with the highest votes becomes our model's prediction.
In addition, we can experiment with several hyperparameters in the Random forest classifier to increase model performance, such as:

- n_estimators: the number of trees that the classifier will consider.
- max_depth: the longest path from the root node to the leaf node.
- min_sample_split: the minimal amount of observations required to split any given node.
- Criterion: a function that determines how good a split is. we can experiment with (gini,entropy) values.

Model Evaluation

Access the true performance of a model is key in its validation step. It allows the modeller to anticipate the capacity of the model to generalise and keep similar predictive power to what has been observed in the training/validation phase.
Predicting the outcome of a football game is no exception and usually the same step used when validating any classification model can be followed.
Having said that, predicting the outcome of a football game has 2 particular aspects:

- Existence of a solid Benchmark producing prediction: the betting market
- Predictions can be used in a direct investment strategy where economical outcome can be simulated/observed

Because accuracy and precision may not always indicate model capability, there are alternative more effective criteria for measuring model performance for our purpose. In this regard, we have created a R package that will enable us to set up the following metrics:

- Log loss: For each occurrence, log loss is the negative average of the log of corrected estimated probabilities. It considers the predictability of the result. Each estimated probability is compared to the actual class output value (0 or 1), and a score is computed that penalizes the probability based on the difference between the expected and actual values. The penalty is logarithmic, with a low score for little variations (0.1 or 0.2) and a high score for major differences. $log\ loss = -1/N \sum_{i=1}^{N} (log\ (P\,i\,))$
- Brier Score: It is an evaluation metric that is used to check the goodness of a predicted probability score. It is very similar to the mean squared error, but it is only applied to prediction probability scores with values ranging from 0 to 1. It is also similar to the log-loss evaluation metric, but the only difference is that it is gentler in penalizing inaccurate predictions than log loss. The best has a score of 0.0, while the worst has a value of 1.0.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

- Residual diagnosis: It is the discrepancy between the observed and estimated values. They're a diagnostic tool for evaluating the quality of a model. It aids in the visualization of errors distribution.
- Calibration Plot: In general, we anticipate the class value that has the best probability of being the true class label for any classification task. However, there are situations when we need to estimate the likelihood of a data instance belonging to each class label. It can assist us assess how decisive a classification model is and grasp how'sure' a model is when predicting a class label.The ideal calibrated model's curve is a linear straight line traveling linearly from (0, 0).
- Tailored scoring rules: The key notion is that getting a higher score than your benchmark isn't enough (market). You must outperform it by a comfortable margin that allows you to benefit. To put it another way, we're comparing model forecasts to those of bookies.
- Trading simulation strategy : We provide the necessary investment instruments for evaluating our model's success. To that end, we provide functions for calculating the amount invested each transaction as well as the projected return on each transaction.

Reference

Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2020). Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model. *arXiv:2012.04380 [Cs]*. http://arxiv.org/abs/2012.04380

By Meher Kharbachi