Predicting the Outcome of a Football Transfer Rumour: An Analysis of Sports Media Unreliability

by

Titouan Savigny

Student ID: 657324

Supervisor: Dr. Nuno Almeida Camacho

Second Assessor: Dr. Flavius Frasincar

DSMA MSc Thesis: Data Science and Principles of Marketing

Erasmus University, Rotterdam

August 2023

Abstract

Unreliability in news and media has always been a cause for concern and whilst traditionally we can fact check statements, when it comes to football transfer rumours, the task is slightly different. Rumours inherently contain a degree of incertitude and therefore it is difficult to determine whether they are true or false. This thesis delves into the credibility of media sources concerning football transfer rumours, an area where speculation, hope, and sometimes, misinformation reign supreme. Drawing from a data set of transfer rumours published by a variety of media outlets reporting on the winter transfer window of 2020 until the summer window of 2022, this study aims to determine (1) if there is a disparity in reliability between different outlets and (2) if we can accurately predict whether a rumoured transfer will become true. To create our model, we test a variety of machine learning methods such as Random Forest, Logistic Regression and Support Vector Machines. Preliminary findings suggest a notable difference in reliability among sources and that Random Forest provides the most accurate results. The study highlights the need for more accurate fact-checking mechanisms and emphasizes the potential consequences for fans, clubs, and players.

Table of Contents

1 Introduction

In an era marked by rapid technological progress and the rise of digital platforms, online media has become an integral part of our lives, shaping our perceptions, opinions, and decisions. Social media has revolutionised how news is distributed with more and more people consuming their news online rather than through traditional outlets such as a newspaper (Nilsen, 2015). Although this has rendered news a lot more accessible, this accessibility has also opened the door for an ever-growing problem: Fake News. The increasing prevalence of media in our daily lives, with a record high of 72% of people reporting consuming online news in 2022[1], has raised concerns about its reliability. The ability to distinguish accurate information and biased content has become an essential skill.

This thesis aims to explore the concept of media reliability in the digital age, specifically by looking at sports media and their coverage of the transfer market. Media reliability refers to the trust given by the public in media outlets to deliver information accurately and objectively. It assesses the extent to which media outlets respect journalistic principles such as truth and honesty. The goal is to critically analyse the factors which affect the reliability of transfer rumours in an attempt to build a classification model which can determine whether the transfer mentioned in a rumour will materialise or not. Additionally, it will observe the potential effects these have on football clubs, players, and agents but also on customer perception of various news outlets. To do this, we build a dataset of previously published rumours with the outcome variable being whether the rumoured transfer happened or not. The independent variables contain information such as the nationality of a player, player agent, age, the outlet, position or market value, all factors which can potentially affect whether a rumour is true or not. We then use this information and build models to predict the probability that a further prediction will be accurate and additionally gain insights on which outlets are more or less reliable.

The web has significantly decreased the barriers to journalism by lowering production and distribution costs (Flaxman, 2016). This, with the rise of social networks and online publishing, has given a voice to many aspiring journalists. However, this has also come with negative effects, most noticeably the rise and prevalence of fake news. Fake news is a term that hit the mainstream media during the 2016 election. Ever since, it has been a major concern, especially as it tends to spread much faster and touch more people than the truth (Vosoughi, 2018). We distinguish three main motivators for fake news: pushing a political agenda, monetary incentives, and social recognition (Kalsnes, 2018). In the context of professional journalism in the transfer market, the drivers are mainly monetary.

[1] *Consumption of online news rises in popularity*. Consumption of online news rises in popularity - Products Eurostat News - Eurostat. (2022, August 24). https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220824-1

Interestingly it is not only the news outlet that profits from rumours but also clubs, players, and their agents.

Football transfers are a crucial aspect of the sport and are widely covered by media outlets. Every upcoming transfer window, speculation on the next potential big transfer runs amok. However, despite their importance, transfer news is often unreliable and filled with speculations, false rumours, and false information. Football transfers represent a huge industry where billions are in play every transfer window. In the winter 2023 premier league window alone a total of over 815 million pounds were spent on transfers, smashing the previous record of 230 million[2]. With so much at stake, it is important for clubs to know which rumours are true or false to make better-informed decisions during the bargaining process.

One of the main issues with transfer rumours is that they are inherently never a certainty. This is unlike traditional news, which can be fact-checked as true or false. With rumours, it is very complicated to determine whether the rumour is founded on reliable intel or made up to drive clicks or for an alternative reason (Pendleton, 1998). This means it is very hard to hold accountable a news outlet for spreading false transfer rumours and currently the assessment of reliability for transfer rumours comes from personal or group experience gained over time. Although verifying sources remains almost impossible, especially since a primary source like a club can also spread false rumours, we wish to explore if there is a way of reliably predicting if a transfer rumour becomes true and use this as a proxy for determining if a news outlet is reliable or not. This thesis will aim to bring empirical evidence and change how we assess the reliability of transfer journalists and sports media.

The literature review investigates the consequences of unreliable media and the many implications of fake news, the main ones being the continuous decline of trust in news media (Allcott & Gentzkow, 2017). A decline in trust is often synonymous with a decline in readership. However, this decline in trust is often delayed as the validity of articles cannot be checked in an instant, often allowing for their propagation. In the context of the transfer market, rumours and fake news can have heavy implications for the bidding process in which certain parties are better informed. This is an example of asymmetric information, and it leads to market imperfections (Mailath & Postlewaite, 1990). Understanding these consequences is important in order to render the transfer market more equitable and able to address issues. Furthermore, uncovering reliability in this sector is critical to promote more accurate reporting. Additionally, we will take a look into the potential identifiers of unreliable news such as the news outlet or the player agent and how they can impact the reliability of a transfer rumour.

[2] White, T. (2023, February 1). *Premier League clubs smash January transfer fee record*. The Independent. https://www.independent.co.uk/sport/football/premier-league-transfer-window-spending-chelsea-b2273361.html

There exists no already pre-existing formatted dataset about transfer rumours verifying their validity, therefore we need to create one. With technological developments, we store and record vast quantities of data, hence all the information required is available, but not formatted coherently to perform statistical analysis. To obtain a data set, we will perform web scraping on various websites. Firstly, to get our rumours, we use the BBC sports gossip page[3], which aggregates rumours from various news outlets or journalists. Then, we must perform natural language processing tasks, such as named entity recognition or data parsing, to extract the meaning out of the text and determine whether the transfer did happen, as well as which agents are involved in the rumour. Additionally, we create data sets to support our analysis; for example, we need information on all transfers which did happen during our specific time period to cross-check our rumours. We also create a database of all the players from the top leagues within a given year, with key information such as age, nationality, or market value, to use as additional independent variables for our analysis. Once we have completed all these tasks, we have a final data set that is suitable for predictive modelling and rumour classification.

This thesis aims to create a predictive model which would allow individuals to assess the reliability of a transfer journalist or news outlet, allowing them to better curate their news outlets for transfers. This would hold news outlets accountable for the rumours they release and pressure them to be more diligent in their sources rather than releasing anything for sensationalism and clicks. Furthermore, clubs and agents alike tend to withhold information or leak false information about potential transfers, leading to asymmetric information which allows them to leverage their position and drive-up transfer prices. Therefore, this model would also be useful for football clubs as they would be able to make better-informed decisions in the transfer market.

To create this model, this paper will look into methods for assessing credibility and web scraping data. As observed in the paper "A survey on fake news and rumour detection techniques" (Bondielli & Marcelloni, 2019), credibility techniques generally use the words in the article or tweets, and analyses features such as the presence of swear words to assign a credibility score. In addition, they may look at contextual information such as the user, their follower count, the comments or how active they are. More advanced techniques try to identify the message and compare it to other credible sources to get an even better credibility score. However, there is little literature proposing to use the features of past articles to predict a new article's reliability, which the proposed model offers to do. Of course, rumour detection is very context-based, as every dataset is different. This proposed method is more suitable for transfer rumours due to every rumour having repeating features, but it is conceivable that this could be applied in other rumour detection scenarios.

[3] BBC. (n.d.). *Saturday's Gossip*. BBC Sport. https://www.bbc.com/sport/football/gossip

There is no best model in predictive modelling, and whilst some may tend to outperform others, the optimal model will vary case by case. Different models often involve a trade-off between interpretability and accuracy, it is therefore important to test a variety of models to find the one that is best suited for our research purpose. Our data set contains a binary outcome and many categorical variables. This means it is not supported by all types of classification models. Given this, we decided to test four different types of models: Decision Trees, Random Forest, Logistic Regression and Support Vector Machines. We hyper-tuned the parameters and compared different variations of models, such as lasso regression, to obtain the best accuracy possible. From these models, we gained valuable insights into what factors most influence the outcome of a rumour, as well as creating a model able to reach an accuracy of 77% given a 67.8% no information rate (This represents the proportion that the majority class represents in the data set).

Through an analysis of existing literature, empirical research, and predictive modelling this thesis aims to contribute to the ongoing discourse on media reliability and its implications. The ambiguity in truth that a rumour inherently holds often allows media outlets to shield themselves from any repercussions. Through looking at their past rumours and potential factors affecting credibility, we wish to determine which outlets are more reliable. This research seeks to shed light on these actors which may tend to release articles for sensationalism rather than seeking the truth; allowing readers to better understand which sources are more reliable or not. This has the potential to change their source for consuming transfer rumours, applying pressure on news outlets to be more diligent as to what they decide to post. We aim to answer the question: “Can we accurately predict the likelihood that a rumour is false or inaccurate?”. If not, we wish to see the factors which tend to influence the veracity of a rumour, as well as being able to compare various news outlets and the accuracy of their reporting overall.

2 Literature Review

The literature review focuses on understanding how media and media reliability has evolved and the increased digitalisation of our society. It will lay the foundations for understanding the importance of media reliability and the impacts it may have. It will delve into how our relationship with media has progressed and the new challenges that come with it.

2.1 Media and Media Reliability

Media reliability is a crucial aspect of journalism that influences the credibility of news outlets and public perception. The following literature review explores the importance of media reliability, highlighting the impact on media outlets, the audience, and society as a whole.

Trustworthiness is essential to the success of news outlets. A study by the Pew Research Center[4] found that public trust in media organisations has declined in recent years, whilst also highlighting that news outlets with a reputation for accuracy and impartiality tend to be more successful. Trust in news media is strongly related to news consumption as people who believe a source to be more credible are more likely to consume its content regularly and are more willing to pay for it (Newman et al., 2018). The importance of media reliability extends beyond media organisations and the audience, with significant implications for society. A study published in the Journal of Communication (Stromer-Galley, 2004) found that media reliability is essential in promoting democratic values, including transparency, accountability, and the right to information.

Given these implications, it follows that a news corporation would at least want to be seen as a credible source to drive traffic and increase popularity. However, we must first define credibility to then determine the steps needed to be credible. Fogg and Tseng (1999) discern four different types of credibility. The first is reputed credibility which comes from the source labels. For example, the label of being a professor with a PhD. Presumed credibility is based on pre-existing assumptions about the source. Surface credibility is derived from the reader's first impression; for example, if a website interface is clean and easy to navigate, we may believe the source to be more trustworthy due to its pleasing aesthetics. Finally, the most reliable method for credibility judgements, experienced credibility arises from the reader's past experiences and their personal judgments about the source.

We can further examine the factors that influence a reader's perception of credibility. Wathen and Burkell (2002), drawing from previous literature, concluded that 3 main factors can affect news credibility: The *source*, which is the news provider, possesses a certain expertise, knowledge,

[4] Rosenberg, S. (2021, July 27). Trust and distrust in America. Pew Research Center - U.S. Politics &amp; Policy. https://www.pewresearch.org/politics/2019/07/22/trust-and-distrust-in-america/

credentials and acquired trustworthiness over time. The *message* which represents the content. Finally, the *medium*, which is how the message is delivered by the source. For example, the interface design or usability. On the other hand, there are the factors that the source cannot directly control but only influence which are the receiver's perception of the source, their prior expertise on the subject and their social location.

Overall, media reliability is a crucial aspect of news with wide-ranging implications as prioritising accuracy and impartiality in reporting enhances reputation and attracts a loyal audience.

2.2 Truth in the Digital Era

The rise of digital news has been significant, as people are increasingly relying on digital devices as their news sources. In 2020, over 80% of Americans reported getting their news online[5]. Although this increases news accessibility overall, it also presents many dangers.

One of the major concerns is the accuracy of digital media and the ease of publishing fake news, especially on social media where anyone can be a self-proclaimed journalist. A study by Tandoc, Lim, and Ling (2018) found that misinformation spread six times faster than accurate news on Twitter. The viral spread of misinformation can often be explained by social media algorithms that promote content that generates the most engagement. This is why fake news thrives so much as it is designed to be sensational and attract clicks (Zannettou et al., 2019). These concerns have been further pushed in recent years with the development of generative artificial intelligence. Previously, a human was necessary to have a general idea of what people will click on, what will captivate readers and then have to create an article. Now, with the recent developments in AI and programs, such as Chat-GPT, that have become widely available, the ease to create and propagate a story may cause an increase in fake news.

The rise of Internet news has also brought a new way of interpreting and assessing its reliability. In our present era, most news websites and social media posts will have comment sections where we can read how others have responded to a given article. In their paper, Heinbach, Ziegele and Quiring (2018) found that user comments had a significant impact on the persuasiveness of news articles over time. Articles with positive comments were more persuasive than articles with negative comments or no comments at all. This effect was particularly strong for articles with low-credibility sources, where positive comments from other readers could offset the negative effects of the low-credibility source and increase the persuasiveness of the article. However, over time, the persuasiveness of articles from low-

[5] Shearer, E. (2021, January 12). *More than eight-in-ten Americans get news from Digital Devices.* Pew Research Center. https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

credibility sources tends to increase to the level of articles with high-credibility sources as people forget the source of the message. This issue is further worrying when we look at the increasing number of bot accounts on social media. A study from 2017 by Varol et al. found that there were an estimated 9-15% of bot accounts on Twitter. These bots can then be used to like, share, and comment on posts to artificially drive up the interaction and popularity of a post (Lazer et al., 2018). This adds another dimension that needs to be analysed when detecting fake news.

There are many drivers for the spread of misinformation, and three of the main reasons are political, financial, and social motives (Kalsnes, 2018). False information can be used for monetary gains by driving traffic to a certain website with, for example, a clickbait article. Higher traffic means a higher valuation for ad placement or an increase in user subscriptions, showing a potential monetary value from increased traffic. This was the case during the 2016 U.S. presidential where teenagers from Macedonia published many false stories about American presidential candidates to cash out on advertising from their generated content[6]. This phenomenon tends to snowball as interactions from a post lead to further interactions in a self-perpetuating cycle (Tandoc et al., 2018).

Alternatively, it can be a powerful tool to control people and guide their thoughts. People tend to identify with particular groups and consume ideologically homogeneous news. This creates an echo chamber that reinforces their pre-existing beliefs rather than diversifying their news content contributing to political polarisation (Bakshy et al., 2015). With this in mind, we have a certain understanding of consumer behaviour which means we can produce content to cater to targeted audiences and manipulate a narrative or their beliefs.

With so many incentives to create fake news and the ability to spread false information, the protocols to detect fake news have tended to lag behind. The numerous challenges in detecting it including the difficulty of verifying the authenticity of sources, the prevalence of deepfakes, and other forms of advanced digital manipulation (Bondielli & Marcelloni, 2019) contribute significantly to this challenge. The article goes on to discuss existing techniques for detecting fake news and highlights the need for further research as current methods are often not sufficient.

In general, the emergence of the internet and social media has created many opportunities for parties wishing to utilise fake news for monetary, political, or social gains which has created a significant market for fake news. On the other hand, our ability to detect fake news is trailing behind.

[6] Kirby, E. J. (2016a, December 5). *The city getting rich from fake news*. BBC News. https://www.bbc.com/news/magazine-38168281

2.3 Truth Detection Models in the Digital Age

With the rapid rise of fake news, the demand for fake news detection has also increased. Over the years, many different approaches have been developed to address fake news employing both manual and automated techniques. There exists no universal approach to solve this issue and in this section, we will delve into some of the techniques which have been and are currently being employed.

2.3.1 Machine Learning Methods

The use of machine learning algorithms is a common method for detecting fake news. These techniques look for patterns and common signs of misinformation by analysing the content of text data such as social media posts, and other sources. They frequently rely on factors like the language used, how reliable the cited sources are, and how consistent the information is among them. More advanced algorithms will have the ability to also look at contextual information such as the comments, the number of followers of the account, how new the account is and other contextual information which may influence credibility (Bondielli & Marcelloni, 2019). By training these algorithms on large datasets of known fake and genuine news, they can learn to distinguish between reliable and unreliable information and classify it.

One of the issues with fake news detection techniques is that there exists no benchmark dataset to train and compare various approaches. The datasets will vary widely depending on which domain or issues we are studying. It is therefore complicated to compare and assess which method is better. Ultimately, the best-performing model will vary on a case-by-case basis. Therefore, we must test different methods to find the best approach.

Some studies will focus purely on the text data, the content, and analyse it through natural language processing techniques (NLP). This means the model will be applicable to any scenario and is not limited by the need to have contextual information. Instead, we wish to examine the language used, such as whether it is sensationalist or contains inflammatory language. We may look at other factors such as the structure of the phrase, linguistic cues, and the number of different words (adjectives, nouns, verbs). More advanced techniques will look into sentiment analysis or opinion mining to gain an understanding of the emotions expressed throughout the text and attempt to use these as a variable to determine veracity. By examining such techniques, we can extract information from the text and see any patterns within fake news or a legitimate statement (Faustini & Ferreira, 2020). In this paper, researchers found Support Vector Machines and Random Forest to be most effective in classifying fake news. However, using text data has its limitations. Fake news is adapting its writing style to match more credible sources, limiting the capacity of these methods. Furthermore, fake news tends to be in a short-term format, especially for social media, and therefore the text data available is limited. It is important to look at other methods which maybe take external factors or content to build a better model, which seems to be the new approach to such tasks (Bondielli & Marcelloni, 2019).

The machine learning methods employed specifically when tackling classification tasks are numerous. To list a few there are Support Vector Machines, Random Forest, Logistic Regression, and Hidden Markov Models amongst others. As previously explained, due to the diverse nature of the data, it's difficult to compare and contrast these methods. Different techniques might outperform others depending on the specific dataset.

2.3.2 Other Approaches

One of the original ways of verifying fake news is fact-checking. Fact-checking can be manual or automatic. With manual fact-checking, it can be expert based in which a knowledgeable person on the subject confirms or denies the truth of a statement. It can also be crowd-sourced where we rely on the collective sentiment towards a statement. There are fact-checking organizations whose role is to investigate articles and determine their veracity. They employ expert opinions and investigate reliable sources to assess the credibility of information. Automatic fact-checking, a relatively new concept, closely ties in with machine learning and natural language processing techniques (Zhou & Zafarani, 2020).

Other methods include concepts such as crowdsourcing. By leveraging collective intelligence, we can help determine and identify fake news. There are many ways to do so. For example, crowdsourcing can be very helpful when creating a data set of fake news to be used as training data for machine learning models. Readers annotate data, highlighting misleading sentences or portions they presume to be false, or they may simply indicate that they believe the entire article to contain lies. By gathering the collective sentiment, which often tends to be correct, we can assess credibility.

In conclusion, despite the promising results from various methods, detecting fake news remains a complex challenge. Misinformation is constantly evolving. To combat this and promote accurate and reliable information, researchers and organizations need to develop more effective truth detection approaches.

3 Research Background: The Football Transfer Market

The football transfer market is an enormous business, in the summer 2022-2023 season alone, a total of 4.3 billion euros were spent [7]. With so much at stake, many fans await the opening of the transfer window each year to see who their club is going to acquire, and who will be the next big money move.

Before discussing how rumours affect the transfer market, we must understand the inner workings of the transfer market. Transfer windows open twice a year: during January and July/August. There are two types of players: Free agents, whose contracts have run out and who are free to sign with whichever club wants them. Signed players, who already have a club, in which case the buyer club would first have to negotiate a transfer fee for the player. Once the transfer fee has been agreed upon, the player must agree to personal terms with the buying club and only then is the transfer completed (Carmichael & Thomas, 1993). Behind the scenes, the player agents work to ensure their player gets the best contract possible. Their role is to advise the player on his career options, contract negotiations, and in general help the player further themselves in their career. However, since they are often paid a percentage of the transfer fee, they have an incentive to drive up the price of their player to get a bigger cut. This is where rumours can play a large role since rumours which show interest in a player from multiple clubs will often drive up the player's price (Kelly & Chatziefstathiou, 2017).

In general transfer rumours play a huge role in the transfer window, clubs are often reliant on these to assess which players are available, at which price and figure out which other clubs are interested. Larger clubs tend to have more financial resources, connections and scouts which gather information for them. They are well-informed as to which clubs are interested in who and at which price points. On the other hand, smaller clubs do not have such resources and will often have to rely on transfer rumours, at least partially, in their bargaining process to understand which clubs are interested in a player and at what price. This can lead to a problem of information asymmetry in the bargaining process which greatly advantages larger clubs. On the other hand, the selling club can also profit from asymmetric information. By propagating fake news and feigning that there is more interest for a player than the reality, we artificially increase demand for the player and increase his price. This is an example of asymmetric information, and it leads to market imperfections (Mailath & Postlewaite, 1990) which can harm the market of football players and especially clubs with less financial power. The party with an informational advantage can bid below the expected value which translates into less value for the auction. The uninformed party then must bid conservatively to not suffer from the winner's curse. We

[7] Tripathi, S. D. (2023, January 18). *Top Ten biggest single transfer window spending clubs in football*. FootTheBall. https://www.foottheball.com/football-top-10/biggest-single-transfer-window-spending-clubs-ever-in-football-history/

can parallel this with bigger clubs being able to buy players at a discount since they know which clubs are interested in the player and what price they are willing to pay. Then, smaller clubs are not willing to engage in a bidding war ultimately hurting the selling club. This will only help accrue the difference between large clubs with more financial power as they are less affected by the premium they must pay in terms of wages and the transfer fee. Additionally, since club performance is strongly correlated with spending (Hall et al., 2002), this will lead to even larger gaps in performance and lead to uncompetitive leagues.

Rumours can also be a great source of content for news outlets, especially during the summer season when there are fewer games and transfer rumours can make for sensational headlines. The newspapers will usually reference some source but with little to no verification[8]. All they want is the ability to create an article which will draw the reader in and since the transfer market is only speculation, they are technically not lying. However, over the recent years, transfer news trust has switched over more and more to independent journalists such as Fabrizio Romano (As of March 2023) who on Twitter alone has 14 million followers as they are often perceived to be more reliable.

Overall, rumours play a very important role in the football transfer market, not only for clubs but also for news outlets. Being able to assess their reliability could help even the playing field during the bargaining process and allow for smaller clubs to acquire players for a fairer price or at least know to not waste resources on certain targets.

[8] Geoffroy, R. (2022, August 7). *Football: Behind the transfer market, the flourishing business of online rumours*. Le Monde.fr. https://www.lemonde.fr/en/les-decodeurs/article/2022/08/07/football-behind-the-transfer-market-the-flourishing-business-of-online-rumours_5992780_8.html

4 The Data Collection & Data Preparation Process

An abundance of data exists in our current world; however, it is not always formatted in a comprehensible and interpretable manner. In this section, we investigate the world of collecting data, specifically within the realm of news and football. We explore the various methods and techniques employed by researchers, journalists, and data analysts to gather, assess, and interpret vast amounts of existing information. In the context of our thesis, this can be separated into two parts. The first one is gathering data from news sources. The second is understanding the data which consists of natural language processing techniques to extract the meaning of our rumours.

4.1 Data Collection

There exists a multitude of methods to obtain and collect data such as surveys, automatic digital collection, or manual collection. For each problem, if we wish to analyse some data, we must first gather it and ensure that it's suitable to answer our hypotheses. In today's digital age, data is continuously recorded. This abundance of data is great for researchers and data analysts but does not always come pre-formatted and ready for use. To obtain a final and interpretable data set we must first go through several steps, especially in text analytics where we want to identify the meaning of the text (Ittoo et al., 2016). This section will go over the literature and our methodology for gathering and cleaning unstructured data.

` For our problem, to be able to perform any sort of predictive modelling, we must first gather a data set. There is no pre-existing collection of rumours, or at least not formatted in a table and ready to use. However, the data is omnipresent throughout the internet and accessible. We must therefore collect it first through web scraping techniques. In addition to this, there is complementary data we need to obtain such as all the transfers which have happened as well as player information that will be used in verifying if a rumour is true.

4.1.1 Web scraping

With the emergence of the internet and the development of computers, web scraping has become an essential tool for extracting data from websites to facilitate research, analysis, and automation tasks. The objective is to find, extract and aggregate information from the internet. It becomes especially useful when trying to transform unstructured data or information from a website to a structured data frame which can then be used for further analysis (Vargiu & Urru, 2012).

There exists a multitude of different web scraping techniques with each their advantages. The first one we will look at is HTML Parsing. Most web scraping tasks employ this method with popular libraries like BeautifulSoup or Scrapy which enable users to navigate the html tree of a website. From this, we can identify the location of the data we wish to extract and hence collect it. This is the method we employ for our thesis. The second method is Application Programming Interfaces (APIs) based scraping. APIs are interfaces that simplify the retrieval of data, often without any coding, and can be created by the website itself such as the Twitter API or by third-party programs. Utilizing APIs provides a more structured and simple approach to data extraction compared to parsing HTML (Glez-Peña et al., 2013). Finally, we can use headless browsers, such as Selenium to simulate real user interactions with web pages. This method is practical when scraping for dynamically generated content or iterating over many different pages however raises legal concerns since we are simulating human behaviour.

Web scraping raises ethical and legal concerns, such as respecting terms of service, and protecting user privacy. From a legal perspective, we must also comply with copyright laws and intellectual property rights. Some websites have already put in place measures to prevent or restrict scraping through IP blocking or implementing CAPTCHA systems. Researchers and practitioners must be mindful of these concerns and act accordingly (Landers et al., 2016).

4.1.2 The Rumours

There are a few different approaches to gathering rumours. The first one is going on news websites such as The Guardian or Sky Sports and collecting all their news articles about football and rumours. The issue is that if we wish for a big sample size of news outlets this would mean a lot of different websites to web scrape with each their way of formatting and formulating rumours, rendering the text scraping, and processing tasks much more complicated. Another option is going through Twitter and looking at sports journalists'/outlets' tweets. The issue with this method is that it would mean we collect tweets that are not rumours since users don't only post rumours. Therefore, ultimately, we decided to use the BBC Sports gossip page[9]. This page is practical and compatible with our tasks for the following reasons. The gossip page gathers rumours from various outlets and journalists whilst always following a similar format in which the original source is specified. The rumour is accompanied by a URL link to the original article, and it is summed up in one or two lines which indicate which parties are involved.

One issue with this method is accessing previous versions of the page to be able to see what rumours existed, for example, in the summer of 2022. To address this issue, we employ the help of an archive website collector: Wayback Machine[10]. This website allows you to access older versions of a certain page. Therefore, we can now look at the gossip page during the dates of Summer 2022 and web scrape this information.

[9] BBC. (n.d.). *Saturday's Gossip*. BBC Sport. https://www.bbc.com/sport/football/gossip

[10] Wayback Machine. (n.d.). https://web.archive.org/

The scraping process involves three steps. We first gather a list of all the URLs of the BBC gossip pages from different dates. We then use the package BeautifulSoup on Python to create functions that will identify the rumours and links in the text. Through the use of a for loop, we apply these functions to each URL and obtain a data set with each rumour as well as the link associated with the rumour. We also record the year and season in which we are scraping. Ultimately, we obtain a total of 3353 rumours which came from the transfer windows spanning from 2020 to 2022.

4.1.3 Selection Bias

An important concept to evoke given the context of how we gather our data is selection bias. Porta defines selection bias as "bias in the estimated association or effect of an exposure on an outcome that arises from the procedures used to select individuals into the study or the analysis". Selection bias occurs when the process of selecting a sample from a population results in a sample that is not representative of the entire population. It arises when certain individuals or groups are more or less likely to be included in the sample based on specific characteristics or factors, leading to a distortion in the estimates or inferences drawn from the data. In our case, the fact that someone from the BBC selected which gossip to post means the samples we have for each various outlet are not necessarily representative of the outlet. Therefore, any inference we make on the reliability may be wrong and the model is more suited for rumours that the BBC retransmits rather than the actual outlets. If we wished to avoid selection bias completely, we would need to select a specific number of outlets and gather all their rumours in the same periods of time. Given the non-homogeneity of the web articles, this task requires extensive text processing skills, hence why we chose to go with the BBC gossip page.

4.1.4 Complementary Data

The rumours data set is not sufficient for our analysis. We wish to know if a transfer rumour went through or not. To verify this, we need data on all the transfers that happened so that once we have extracted the player and interested clubs, we can cross-check if the rumour did happen. Fortunately, this data set already exists and we will therefore be using the work of user *ewenme* on GitHub[11] as our transfer data set, which was web scraped from the site TransferMarkt.

The second part of complementary data we need is player information. This consists of variables such as their agents, market value, age, nationality, or time left on the contract. All these variables have potential predictive power in determining if a rumour is true and will therefore be useful in optimising the models. To perform this task, we use the same site as for the transfers, TransferMarkt, to not have any issues with differences in spelling. For each football league, we go to their dedicated

[11] GitHub, ewenme, https://github.com/ewenme/transfers

page. The page[12] (See footnote for example) then shows us all the clubs involved in this league. From this, we obtain the URL for each club which gives us the full squad list as well as information on the players. From there, we scrape all the information and repeat this for each league. We can also filter throughout the seasons to obtain the same information but for previous seasons. Since our rumours start in the summer of 2020, we gather the data from 2020 to 2022.

4.2 Data Preparation

Once we've collected the data, we must pre-process and format it, making it suitable for applying natural language processing techniques. Currently, our dataset indicates nothing other than the raw text we've extracted. To be able to perform statistical analysis, we must extract the meaning of the text through pre-processing and then natural language processing techniques.

4.2.1 Pre-Processing

Pre-processing the data involves several steps. It will vary depending on your original data and what further work you want to perform. The first step is data cleaning. This consists of handling any missing values, outliers, or noise in the data. We can address these issues by, for example, deleting rows with null entries or observations which seem to contain an error. In the context of text analytics, we remove punctuation, lowercase our text, handle special characters, and tokenise the text into a list of words.

Our rumours data comes with four columns, the rumour, the year, the season, and its hyperlink. To pursue our analysis, we must first clean up the text and extract information from it. Firstly, we want to record the outlet which published the article. The BBC page writes this in the same format for every rumour, it is at the end of the article and in brackets. Therefore, to extract this information, we look for the last point and then extract the text in brackets to get the name of the original source. We must then format the text. To do so, we remove any punctuation, convert the words to only contain letters from the English alphabet, and lowercase all the letters. We then tokenise the text which consists in separating the text into a list of words.

[12] https://www.transfermarkt.com/premier-league/startseite/wettbewerb/GB1

The players data set, on the other hand, requires a lot more pre-processing. If we look at *Table 4.1*, we can observe what the data looks like after we scrape it:

Table 4.1: 2022 Players Data Example – Lionel Messi

Player	**League**	**Position**	**Birth Date**	**Country**	**Join Date**	**Market Value**	**Club**	**Player Agent**
Lionel Messí	Ligue 1	Right Winger	Jun 24, 1987	Argentina	Aug 10, 2021	€50 .00m	PSG	Relatives

We wish to use these as independent variables for our predictive models; however, to do so we must reformat them. Firstly, we need to convert any date variable into a continuous variable. Date of birth becomes age in years and the join date (When they joined the club) is converted into days since they joined the club. We do this by taking the difference in days between the 1st of August or the 1st of January of the given year that we are looking at and the join date. This implies that our join column will not be exactly accurate but will be off by a maximum of 30 days. However, we consider this to be insignificant as it should balance out. Additionally, we ensure all columns are of the right data type by converting any numeric columns to numeric. Another important step is simplifying the names. Different alphabets have different characters that are at risk of causing problems when merging with the rumours data set. Therefore, for any player, to ensure spelling is homogenous across datasets we convert the writing to the same alphabet. Finally, we group the different positions into three different ones: Attacker, Midfielder and Defender. We do so to limit the number of options for the categorical variable and since we do not need the position to be highly specific, we mainly want to see the difference for these 3 options. We can see how the cleaned table looks in *Table 4.2:*

Table 4.2: 2022 Cleaned Players Data Example – Lionel Messi

Player	**League**	**Position**	**Age**	**Country**	**Join**	**Market Value**	**Club**	**Player Agent**
Lionel Messi	Ligue 1	Attacker	35.0	Argentina	325.0	50.0	PSG	Relatives

4.2.2 Named Entity Recognition

Named Entity Recognition (NER) is a natural language processing task that involves identifying and categorizing named entities, such as names, locations, organizations, and dates, within text data (Marrero et al., 2013). In the context of our thesis, this means identifying the clubs and agents in the rumours. There exist different approaches to perform this, we will look at two different approaches.

Rule-Based Approach: Rule-based NER techniques rely on predefined patterns and linguistic rules to identify named entities. These approaches often utilize regular expressions, string matching, and custom rules based on specific domain/ language expertise. While rule-based methods can achieve high precision, they are often limited to certain domains and languages whilst also requiring a lot of pre-existing knowledge to design the rules.

Dictionary-Based Approaches: This approach identifies named entities in the text by using pre-existing dictionaries. In this method, a dictionary containing a collection of words or phrases associated with specific entity types is used to match and classify entities in the text. We can manually create this dictionary or use pre-existing ones depending on our needs. This method has the advantage of being simple to implement and allows a high degree of customisation leading to high precision scores. However, it relies heavily on the accuracy and quality of the dictionary which needs to be updated every time a new scenario or entity comes up.

As previously discussed, we gathered a data set of all the players along with their clubs. Using this data, we decide to see how well the dictionary-based approach performs. We, therefore, create two dictionaries, one for all the clubs and one for all the players. The issue with this approach is that the spelling for a club or player may vary. For example, the article may say “Chelsea”, but the dictionary says “Chelsea FC” which means it won’t be recognised as a club entity in our NER. This is especially true for rumours in which the club names tend to be abbreviated or simplified. This often comes from a second word such as FC or United being added on. Therefore, to remedy this, in the dictionary we simplified the club names as much as possible. For example, Leeds United becomes Leeds or we remove any FC. Additionally, we add nicknames of common clubs in the dictionary such as “Gunners” for Arsenal.

To test the effectiveness of this approach, we manually label 50 entries and then run the NER. After comparing the manual entries from the NER results, we obtain a 94% success rate in identifying the clubs. We tried the rule-based approach but there is no existing method that could successfully identify the club names and the ones which we did try, had a sub-50% success rate.

With regards to the player names, a similar issue arises since they can be spelt differently in a rumour than in the dictionary. A potential fix for this is using a pre-existing rule-based approach that is designed to identify names. However, the rules employed often have to do with two consecutive words with capital letters and since clubs are noted with capital letters it often confuses a club with a name. Additionally, football players come from all around the world and the rule-based approach is not suited for the diversity of names. Nevertheless, we wished to compare the two approaches and tested the accuracy on the labelled data set. The rule-based approach came out with less than 40% accuracy compared to 90% for the dictionary-based approach. Therefore, we concluded that the dictionary-based approach was best suited for our analysis.

Ultimately, we decide to cut our dataset to entries where only one player is recognised. This eliminates all the entries where our NER failed to identify a name, increasing our accuracy for NER on the final data set. It also eliminates the case where there was more than one player mentioned which could evoke a player swap or other scenarios which would be more complicated to analyse. Furthermore, we decide to only preserve entries where at least two clubs are mentioned since if only one club is mentioned, we either failed to identify one of the clubs or the rumour is not about a transfer. This comes at the cost of losing 1433 rumours but ends with a 100% accuracy in identifying the clubs and 98% for the players involved.

4.3 Extracting the meaning

We now have a dataset with the mentioned clubs and players. If we use the assumption that every rumour is about a potential transfer, we could simply cross-check if the transfer happened by checking if there was a transfer from the players club to one of the interested clubs using the transfers data set. However, approximately 10% of rumours, are denying rumours or informing the reader that a club is no longer interested. To address this issue, we look at three different methods.

4.3.1 Dependency Parsing

Dependency parsing is a linguistic technique used to analyse and represent the syntactic structure of a sentence. It involves identifying the relationships between words and hierarchically organizing them. To do so, we tokenise the text, which consists of splitting it up into individual words. We then parse the text by assigning a label, typically defined in terms of grammatical functions, such as subject or order. Using various algorithms and models such as rule-based approaches, we analyse and generate dependency structures for sentences. These relationships, known as dependencies, capture the grammatical and semantic connections within a sentence. By constructing a dependency tree, dependency parsing shows the role of each word and its dependence on other words in the sentence.

4.3.2 Pattern-Matching

Text analytics searches for specific patterns or sequences in textual data. It extracts relevant information based on predefined patterns or rules. Pattern-matching techniques analyse and extract structured information from unstructured text. They recognize patterns representing entities or relationships. These techniques build on NLP methods like NER or dependency parsing. Different methods, like regular expressions or statistical models, can be used for pattern matching. Overall, it uncovers recurring relationships between words in text data and provides insights into their meaning.

4.3.3 Sentiment Analysis

Sentiment analysis is a technique used to determine the sentiment of any text expressing opinions or emotions such as reviews or customer feedback. The objective is to classify the sentiment expressed in the text as positive, negative, or neutral (Zhang et al., 2018). This involves several steps. Firstly, the text is pre-processed by removing stop words, punctuation, and other noise. We then proceed with feature extraction which consists in identifying the words which will affect our sentiment score. For example, the word happy is assigned a positive score. Contextual clues, intensifiers and negations will also affect the final sentiment score. There exist various methods for calculating sentiment scores, the more traditional approach involves using predefined dictionaries which associate words and patterns with a certain polarity. Once this is complete, we get a sentiment score and, if we wish, we can also identify the emotions from the text. By analysing the sentiment, organizations and individuals can gain insights into public opinions, customer satisfaction, and identify emerging trends or issues. It allows for marketeers and organisations to make data-driven decisions and understand customers better.

Typically, sentiment analysis is used for understanding customer reaction to a product or public sentiment, however, we decide to use it in the context of our analysis. Whilst the other methods are more traditionally used for such tasks, sentiment analysis is simpler to apply and still had the potential to be effective. We experimented by splitting our rumours on club names and using the subsequent words to determine a sentiment score. A positive or neutral score would indicate that the club is interested in buying or selling a player meanwhile a negative one would show the opposite. To test this, we tokenise our data and using our identified clubs we take the 3 words which follow a mentioned club. We tried a range of different following words (2-6), but 3 ended up giving the best results on the validation sample. From this, we run our sentiment analysis and use the polarity score to determine if the rumour is still active or is being disproven. If for either of the clubs, we identified a negative score, we consider the rumour to no longer be active. In the end, this allows us to go from 10% of rumours with no actual transfer being proposed to 4%, as we still fail to identify the meaning in some rumours due to the language employed. Based on these results, we decided to adopt this method.

4.4 Data Consolidation

Once we have pre-processed our data sets, we can then consolidate our data. We wish to have additional information to increase the predictive power of our models and must therefore merge our rumours to the players dataset as well as cross-checking with the transfers data whether our rumoured transfer went through or not.

The first step is to merge the player's data with the rumours. As previously established, the names both come from the TransferMarkt website and therefore we merge on the player's name column without any errors. We do so by doing a left join onto our rumours data set and hence obtaining all the player data for the player in the rumour. The only subtility with this is that the player data is separated into different tables for the different years, so we must filter by the year column to perform the merge.

The final step to creating the dataset is determining our outcome variable – Whether the transfer did or did not happen. The objective for this section is, given the rumour and the clubs/ player we identified, to be able to determine if the rumoured transfer has materialised and if yes, the outcome will be True. We have a dataset of all the transfers. Therefore, for each rumour, we verify if a transfer has occurred for the player to one of the interested clubs. Firstly, we identify the player's name from the rumour and then look if, during the time period of the rumour, the player was transferred. If there is no transfer then the rumour is false, if there is, we must now check if the destination club is part of the interested clubs. This can be problematic as we run into naming issues with regard to the clubs. The clubs in the rumours are spelled differently than in the TransferMarkt database. To solve this issue, we check for two criteria. The first one is if the interested club name is contained within the destination club name, or vice versa. If yes, we consider the two clubs to be the same and the rumour is true. The second criteria is a word similarity checker from the Spacy library. For each interested club, we run a word similarity test with the destination club, if a score of over 0.8 is found, we consider the two names to represent the same club and the rumour is then true. With this information, we are ultimately able to create the variable "Outcome" which is a boolean variable, reading True if the transfer from a rumour did happen and false otherwise. We obtain 100% accuracy with our validation sample. Overall, we end up with a dataset of 1638 entries and 8 independent variables. The entire process is illustrated in *Figure 1*, Appendix, and an example of the final table which will be used in our analysis can be seen in *Table 4.3*. We can note that the player's name is not mentioned nor the year, this is because they do not help create our predictive model.

Table 4.3: Example of Final Data Set – Timo Werner 2021

Outlet	**League**	**Position**	**Age**	**Country**	**Join**	**Market Value**	**Player Agent**	**Outcome**
Express	PL	Attacker	25	Germany	1635	65.0	Sports360	False

5 Methodology – Rumour Classification and Predictive Modelling

There exist numerous different methods of performing predictive modelling with no single method always being on top. Different industries, the nature of the data or other factors such as the objective of your research will influence which model is best suited and this model will vary case by case. These techniques utilise statistical and machine learning techniques, as well as data mining approaches, to gain insights from data. From simple linear regression models to more complex methods like Random Forests or neural networks, each offers different insights. As model complexity increases, interpretability often diminishes. Therefore, it's valuable to explore various models not just for performance but also for interpretability. In our methodology, we employ techniques like Random Forest and Logistic Regression to analyse our data.

5.1 Decision Trees

Decision trees are a type of supervised machine learning algorithm which have the advantage of being simple to interpret as they are easily representable (Kotsiantis, 2013). We start at a root node which represents the entirety of the population, the data set. We then split the root node into two or more child nodes based on one of the features, the independent variables of the data set. For example, in our data set this could be whether the market value of our player is over 30 million. The condition to split on is chosen based on which condition minimizes the Gini index, which calculates the purity of a node by looking at the proportion of misclassified classes after a split. We repeat this process until a certain stopping criterion has been reached, this could be a maximum number of splits or when the nodes are pure (Loh, 2011). Decision trees are easy to understand and visualize, however, they are prone to overfitting which means they are mainly suitable to predict the training data set. To mitigate this risk, we can use techniques such as pruning to simplify the tree and reduce the risk of overfitting. Pruning consists in applying a penalty parameter which increases the more splits we do down the decision tree to not have an overly long tree that overfits our data. To test which complexity parameter (cp) is best, we run the model with various values for the cp and chose the optimal cp based on accuracy and tree length.

Other than interpretability, decision trees have many other advantages. Decision trees can handle non-linear relationships between features and the target variable. Furthermore, they can handle both categorical and numerical features. They can also naturally handle missing values and outliers, meaning we require fewer data pre-processing and loss of data, simplifying the modelling process. Finally, by evaluating the reduction in impurity or information gain associated with each feature, decision trees can rank the features based on their predictive power. This means we can see which variables are most important but also the ones which affect the prediction little to none (Fabricius & De'ath, 2000). However, they tend to underperform in classification accuracy compared to other machine learning models which is often a crucial aspect of research. We can still use the idea behind classification trees to create more complex models such as Random Forest or XG Boost which both utilise decision trees as a foundation for building the model.

5.2 Support Vector Machines

Support Vector Machines (SVM) were first introduced into the literature in 1995 by Vapnik in his paper: Support-Vector Networks. In his paper, he defines a new machine-learning method used for classification problems that implement the idea that input vectors are non-linearly mapped to a high-dimension feature space. This concept has since involved into SVM which can be used in both regression and classification problems. SVM work by finding an optimal hyperplane that separates data points of different classes. It can handle both linearly separable and non-linearly separable data by using various kernel functions, such as the polynomial or linear kernel, to transform the input space into a higher-dimensional feature space (Karatzoglou et al., 2006). For our models, we compare the performance of three different kernels: Linear, Radial and Polynomial.

The linear kernel is the most basic, it computes the product of the input features to find the linear boundaries in the original feature space. It is ideal when the data is linearly separable, and the number of features is large compared to the number of samples.

The Radial Basis Function (RBF) kernel maps the data into an infinite-dimensional space and is often applied for non-linearly separable data. It measures the similarity between data points based on their distance to the support vector. It can capture complex relationships and allows SVM to create non-linear decision boundaries however hyper tuning the model parameters can be challenging.

The polynomial kernel functions similarly to the RBF. It maps the data into a higher-dimensional feature space using polynomial functions. It is best suited when the data has polynomial relationships between features. It allows for curved relationships rather than simple linear patterns.

One of the main advantages of SVM is its ability to handle data with many independent variables, it performs well even when the number of features is much larger than the number of samples. It is also effective in handling datasets with a small number of training samples since it focuses on the data points closest to the decision boundary, rather than the entire dataset. Another advantage is the kernel trick which allows SVM to implicitly map the input data into a higher-dimensional feature space, where it becomes easier to find a linear separation. This flexibility makes SVM suitable for a wide range of complex datasets.

On the other hand, SVM performance is very sensitive to hyper tuning the parameters and choosing the appropriate kernel function. This can be a complicated and computationally heavy task, especially with a high level of dimensions, and affects overall model performance (Noble, 2006). To ensure we chose the best kernel function, we test three different options and by comparing their accuracy scores, we can determine the best-performing kernel. Furthermore, unlike decision trees or linear models, SVM does not provide direct interpretability in terms of feature importance or decision rules.

5.3 Logistic Regression

Logistic Regression is a statistical algorithm used for classification tasks with a binary outcome variable, where the objective is to predict the probability of an instance belonging to a particular class. It models the relationship between the independent variables and the outcome using the logistic function. It estimates the parameters that best fit the data by maximizing the likelihood of the observed outcomes (Lavalley, 2008).

Logistic Regression has the advantage of being simple and interpretable. The model's output represents a percentage change in the probability of obtaining the positive class, and the coefficients associated with each feature provide insights into the direction and magnitude of their impact on the predicted probability. We can further look at the significance scores to see which variables are most important in the predictions allowing for easy interpretation of feature importance. It is also computationally efficient and performs well on datasets with a large number of features (Stoltzfus, 2011). Additionally, it can handle missing values and multicollinearity issues through techniques like regularization.

However, Logistic Regression also has its limitations. Firstly, it assumes a linear relationship between the input features and the log-odds of the outcome, which may not always be true. In such cases, the inclusion of interaction terms may be required to capture non-linearities. Another limitation is that it is primarily designed for binary classification problems although this is not an issue for our specific research question. Furthermore, it is sensitive to outliers which can disproportionately influence the estimated coefficients and impact the model's predictions.

5.3.1 Lasso Logistic Regression

Lasso Logistic Regression, is a variant of Logistic Regression that incorporates a penalty term to the Logistic Regression function, encouraging some of the model's coefficients to be exactly zero, essentially performing feature selection. This allows us to automatically identify and exclude irrelevant or less important predictors, leading to a more interpretable and potentially more robust model (Meier et al., 2008). This is especially interesting with our data set containing many categorical variables which means we have many predictor variables, many of which are unimpactful.

Lasso Logistic Regression is excellent for high-dimensional datasets since it automatically performs feature selection by setting irrelevant or redundant predictor coefficients to zero. This reduces the risk of overfitting and enhances model interpretability. Additionally, it mitigates the risk of multicollinearity among predictors, making the model more stable and less sensitive to small changes in the data.

On the other hand, lasso Logistic Regression's feature selection process might be biased when predictors are highly correlated, leading to the selection of one predictor over another, even if both are informative. It also requires tuning the regularization parameter (lambda) through techniques such as cross-validation to find the best lambda possible.

Lasso Logistic Regression is a valuable technique for feature selection and regularization, particularly in scenarios with high-dimensional data and a need for interpretability. Given the high dimensionality of our data, we decide to test it. To optimise our results, we make sure to tune for the best lambda value possible and then interpret the results.

5.4 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It was first developed by Breiman, L. in his 2001 paper: Random Forests In his paper, Breiman introduces the idea of randomness in the construction of individual trees, combining their predictions to obtain a more robust and accurate model. By doing so, it also helps in reducing overfitting and increasing predictive accuracy.

In Random Forest, we create *n*, generally around 500, decision trees which are built on random subsets of the training data, known as bootstrap samples. For each decision tree, we randomly select *x* features to be used as a condition for a node split. This randomness introduces diversity among the trees, ensuring that each tree is different resulting in less overfitting (Biau et al., 2016). During prediction, each tree in the forest independently produces a prediction, and the final prediction is determined by majority voting, meaning whichever prediction was most common out of the *n* trees, is selected as the outcome. In regression tasks, we average the results across all trees. We can further optimise the model by testing for which values of *x* and *n* the model performs best. We apply this concept in our model, by using cross-validation and tuning the number of variables to use for each decision tree, we can observe the accuracy for different values of *x (See Figure 5, Appendix)*. In our case the optimal value is 8 for the number of variables, however, if computational power is a concern, we can use 6 as the difference in accuracy is not too important.

Random Forests have many advantages as a predictive method. One of the significant advantages is its robustness against overfitting. It reduces the variance in the data and predictions by using multiple trees. Similarly to decision trees, the notion of variable importance is still present, and we can use it to further our understanding of how the model works (Belgiu et al., 2016). It can also handle missing values and outliers without extensive data pre-processing. This versatility and adaptability enable it to be applied to large-scale problems. Given the nature of our data set which is relatively small with a majority of categorical variables, this adaptability ensures a good performance. Previous studies (Cha et al., 2021), which looked at the performance of Random Forest and Gradient Boosting Machines on small datasets with many categorical variables showed that Random Forest outperformed its counterpart with many of the models showing excellent predictive performance.

However, the interpretability of Random Forest is lower compared to individual decision trees because it becomes more challenging to understand the underlying decision rules when multiple trees are combined. It also tends to be computationally expensive, especially when dealing with large datasets or a large number of trees. Furthermore, Random Forest may not perform well on datasets with imbalanced outcome variables, as it tends to favour the majority class due to majority voting. Our final data set has a slight class imbalance with 68% of rumours not becoming true. In such cases, we can employ techniques like class weighting or data resampling to adapt our model and more accurately predict both outcomes rather than the dominant one. For our model, we decide to compare the results of a class weighted against a normal Random Forest.

5.5 The Issue of Categorical Variables

One of the major issues with the data set is that it has categorical variables with many levels. We say that the features have high cardinality. For example, we have over 130 unique outlets. This means that some of these levels are observed once or very few times and sparsity can lead to difficulties in estimating reliable statistical relationships within the models. Rare levels in the independent variables can introduce noise hindering the model's ability to generalize unseen data and lead to overfitting. Finally, it reduces interpretability as it becomes more and more complicated to discern how the model operates as the number of variables increases.

To address this issue we can use methods such as similarity encoding (Cerda et al., 2018), which tries to group categorical variables which are similar based on factors such as collinearity. However, this is not very plausible for our data set, especially since one of our biggest categorical variables is the outlets. If we wish to analyse and interpret a specific outlet's reliability, we lose a lot of interpretability by having it grouped with other distinct outlets. Instead, we adopt a similar idea and group based on factors such as language. For the outlet data set, we make sure that different spellings of outlets are not counted as distinct and instead group them. We also grouped by language for Dutch and Portuguese outlets as they had few occurrences, as well as grouping independent journalists.

Once we have grouped the variables as we see fit, we still have categorical variables which have levels that occur only once, especially with regard to agents and outlets. This is an issue since when we run certain models, such as decision trees, and then use the output to create predictions, we are at risk of encountering a data point that did not exist in the training data and therefore the model will throw an error. Therefore, the strict minimum is that a categorical variable is present at least twice and that during the split into training and testing data, each set contains one of the entries. To ensure we have this condition, we must impose a certain threshold. A threshold represents a minimum number of occurrences for a level of a categorical variable. To implement this, we can use various methods which work based on finding similarities and patterns between the unique values of categorical variables, ultimately grouping them (Moeyersoms & Martens, 2015). Alternatively, the simplest approach, which we employ, is to have a minimum occurrence threshold, and if a level does not meet this criterion, we group it into a new category called "Other". Although this method tends to underperform compared to more complex encoding methods (Pargeant et al., 2022), it is simple to implement and allows us not to group distinct outlets, making it ideal for our thesis.

Finally, we wanted to see the impact that changing the threshold for minimum occurrences of a certain level has on the model performance. Varying the threshold means we have a trade-off. As our threshold increases, the number of unique levels decreases. Having fewer levels makes the model more interpretable as there are fewer factors to take into account, however, it leads to information loss. We lose the ability to interpret low occurring levels and make predictions on them. To determine which model is best we test for four different thresholds: $n > 4$, $n > 9$, $n > 14$ and $n > 19$ where n is the number of minimum occurrences for a particular level. Then, we chose the optimal value for n, based on model accuracy and interpretability.

6 Results

To choose a final model, which is best suited for our research purpose, we first test various models and analyse them to see what insights we can obtain. Although accuracy, which represents the proportion of correctly classified outcomes, is a key metric in determining predictive power, other key metrics such as sensitivity, specificity, or the no-information rate[13] must be taken into consideration when choosing the model. The interpretability of variables and the model is another key component that we must factor in. Additionally, we wish to compare if there is a difference between different thresholds for the minimum occurrences of our categorical variables.

6.1 Categorical Variables Threshold

As previously discussed, the data set contains categorical variables with many levels. In this section, we test the accuracy of the three different chosen models (Random Forest, SVM and Logistic Regression) under different thresholds to determine which one to use for further analysis. It is important to interpret the results on different machine learning models, as different thresholds will yield different results and the optimal value for *n* may vary (Pargeant et al., 2022). We start with the threshold *n* > 4. We can't start too low since if, for example, we have a level which only occurs twice, there is a possibility that our test and training data have variables that only appear in one of the data sets. To ensure we don't have such problems we start with *n* > 4. *Table 6.1* and *Figure 2*, *Appendix,* show the accuracy results of our models based on the various thresholds imposed.

Table 6.1: Accuracy per threshold of minimum occurrences (n) of categorical variables (rows), and per model (columns)

Threshold	Random Forest	SVM (Kernel)	Logistic Regression
n > 4	75.8	71.3 (Radial)	71.9
n > 9	77.3	70.3 (Poly)	73.8
n > 14	77.1	72.5 (Radial)	73.2
n > 19	78.0	72.6 (Poly)	72.3

Note: The no-information rate is equal to 67.9%.

The kernel in parenthesis is the best performing kernel for the threshold out of Linear, Radial and Polynomial

[13] Sensitivity: The proportion of true positives correctly identified by a test.

Specificity: The proportion of true negatives correctly identified by a test.

No-information Rate: The probability of correct prediction by a model when no predictor information is used.

As we can see from the table, the best-performing model is Random Forest whilst SVM and Logistic Regression perform similarly. Random Forest is best suited for predictions and will likely be our final model. With regards to the optimal threshold, if we only look at accuracy then n > 19, is the best-performing threshold. However, this is also the threshold that converts the most information into the "Other" level and therefore limits our ability to make future predictions on low-occurring levels as well as the number of outlets we can analyse.

To decide between these thresholds, we must see how they will affect our interpretability, for two reasons. First, thresholds influence the number of unique values per variable, influencing interpretability. The more levels, the harder it is to interpret. For instance, if we look at *Table 1* in the Appendix, we can observe the number of different levels per categorical variable as well as total levels. For *n > 9*, we obtain a total of 110 levels compared to 84 for *n > 14* and 71 for *n > 19.* If we use the smaller threshold, we risk having too many variables to interpret and risk making conclusions about an outlet's accuracy based on a small sample size which will likely be overfitted to the particular data set we obtained. Additionally, computational complexity increases with higher dimensionality. On the other hand, choosing a large threshold means we end up with a very large "Other" level for categorical variables like agent which has a large number of values with under 20 appearances in the data. This leads to the agent variable being less important in the model and potentially we lose out on some information that can help make predictions.

Second, imposing thresholds limits our ability for potential predictions to only the variables we retain. For example, we can no longer predict if the league is the Scottish Premiership with the threshold n>19 for "League" since this threshold transforms these observations to "Other". Given our objective is to analyse media reliability, we also wish to make sure we have enough observations for each of the outlets that give us a reliable interpretation of their coefficients and importance in the model. With this in mind, we decided to set the threshold to 19. This threshold obtains the best result in the Random Forest model, which itself outperforms the other models. Although we lose interpretation of certain outlets, we still have a total of 26 outlets compared to 40 if we used a threshold of 5. This is already a large number of outlets and sufficient for the purpose of this thesis which is to observe if reliability varies between different media outlets. To ensure we don't end up with a too large "Other" level for our categorical variables, we looked at how thresholds influenced the loss of unique observations. *Table 1* in the Appendix shows that the number of distinct observed levels at various thresholds decreases most rapidly with regard to the "Agent" category, therefore we decided to modify the agent threshold to n > 9. Ultimately, we have a data set with a no-information rate of 67.9%. *Table 6.2* shows accuracy metrics for the 8 different models we train our data on and *Figure 3* in the Appendix shows the bar chart representation of the results.

Table 6.2: Performance Metrics (%; in columns) of Alternative Models (rows~~)~~)

Model	Accuracy	Specificity	Sensitivity	Balanced Accuracy
Decision Tree	66.7	49.5	74.7	62.2
Random Forest	77.1	45.7	91.9	68.8
Class Weighted RF	73.4	69.5	75.2	72.3
SVM – Radial Kernel	69.4	31.4	87.4	59.4
SVM - Polynomial	72.6	23.0	96.0	59.5
SVM - Linear	68.5	20.0	91.4	55.7
Logistic Regression	72.0	30.1	91.5	61.2
Lasso	68.2	1.10	99.6	50.7

Note: Balanced Accuracy represents the average between specificity and sensitivity Categorical variable min occurrence, n >19 except for" Agent": n > 9, no-information rate = 67.9%

6.2 Analysis

Following our decision for the threshold, we wish to look into the models and see how they operate to gain further insights into which variables most affect our models and predictions. This section will explore interpretability methods such as variable importance plots and will try to understand which outlets are most reliable as well as observing if other variables also have a significant effect.

6.2.1 Decision Tree

The unpruned decision tree achieves an accuracy of 66.7%, which is inferior to the no-information rate, whilst the pruned decision tree recommends an infinite complexity parameter (*See Figure 4, Appendix*). This means the pruned tree has a length of 0 and simply recommends classifying the outcome as the majority class. A recommended complexity parameter of infinity indicates that the tree does not successfully learn the underlying patterns or relationships between the features and the target. This is especially true considering the unpruned decision fails to beat the no-information rate. Instead, we must look at other methods if we wish to gain further insights.

6.2.2 SVM

The SVM models underperform compared to Random Forest. The best model obtains a 72.6% accuracy and uses the polynomial kernel. The linear kernel has the worst results, possibly indicating that our original feature space is not linearly separable. On the other hand, the polynomial and radial kernel transform the data into a higher-dimensional feature space to capture more complex relationships and patterns in the data that cannot be linearly separated. This allows them to perform better than the linear kernel which is more limited.

6.2.3 Random Forest

As seen previously, Random Forest appears to be the most accurate model. To have the best-performing model we first train the model whilst varying the number of randomly selected predictors chosen for each decision tree. We test for values between 1 and 9 and obtain the cross-validated accuracy based on the number of predictors (*see Figure 5, Appendix*). The accuracy stabilises around 6 predictors but 8 has the best accuracy and is therefore used to create the Random Forest that has 77.1% accuracy in predicting the test data. However, accuracy does not tell the whole story and when we look at the specificity, which is the percentage of correctly predicted true outcomes (The transfer does happen), we only have 45.7% accuracy. This means the model greatly improves from the no-information rate, however, it still favours predicting the false outcome with a 91.9% sensitivity. To attempt and remedy this issue, we decide to employ class weighted Random Forest which allows us to emphasize the minority class to have a better-balanced accuracy.

6.2.4 Class Weighted Random Forest

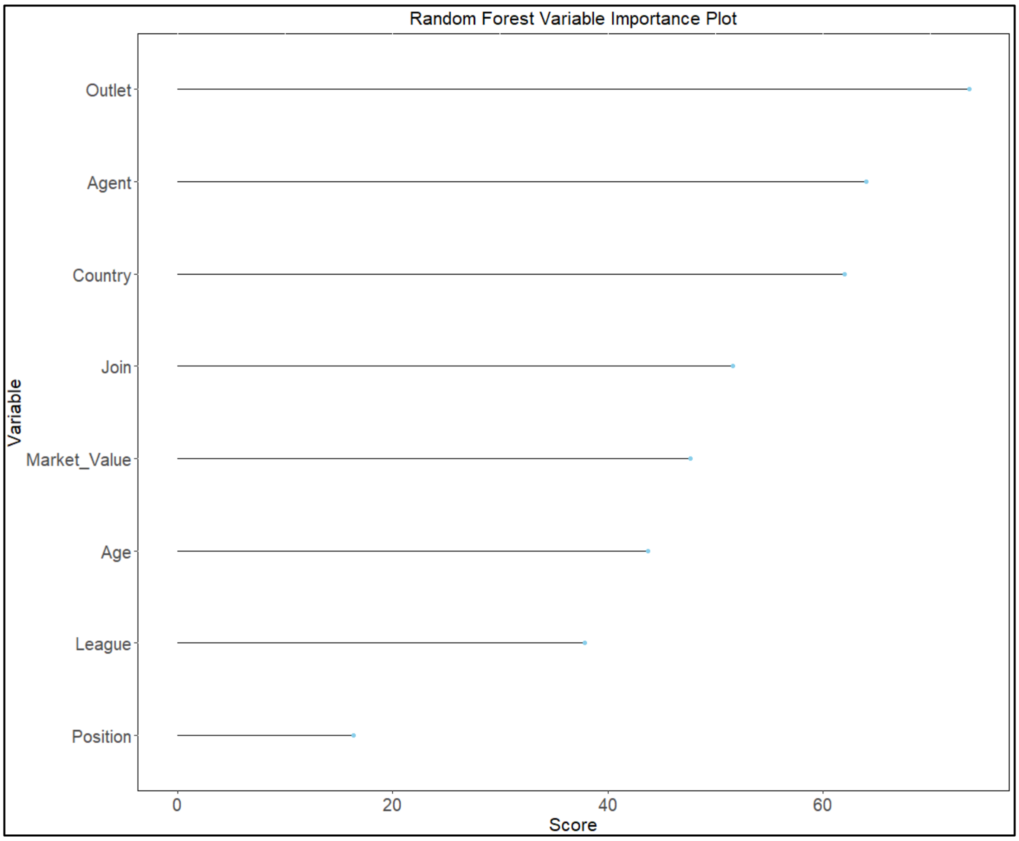
In Random Forest, class weights are used to adjust the importance of different classes in the training process. By default, equal weights are assigned to all classes, meaning that it treats all classes as equally important. Class weights work by adjusting the loss function used to train a classification model such that the model prioritises the classification of the minority class. Employing this method allows us to reach a 73.4% accuracy but most importantly, we now have a specificity of 69.5% compared to the 45.7% of the normal Random Forest. This is especially interesting as we often wish to see if a rumour will come true since it confirms the idea that the clubs involved have a genuine interest in the player. In general, the reader is more interested in rumours which materialise than does that don't. Considering we lose in accuracy; we have a trade-off between overall accuracy and specificity. It is difficult to determine which is better as it depends on how much we value the ability to predict that an outcome will come True. For our research, we believe a 3.7% difference in accuracy does not justify the more balanced accuracy and therefore decide to use the Random Forest as the subsequent model.

We can now look at the variable importance plots and partial dependence plots to see which variables have the most impact on the model and how they affect the model

6.2.5 Random Forest Variable Importance and Partial Dependence Plots

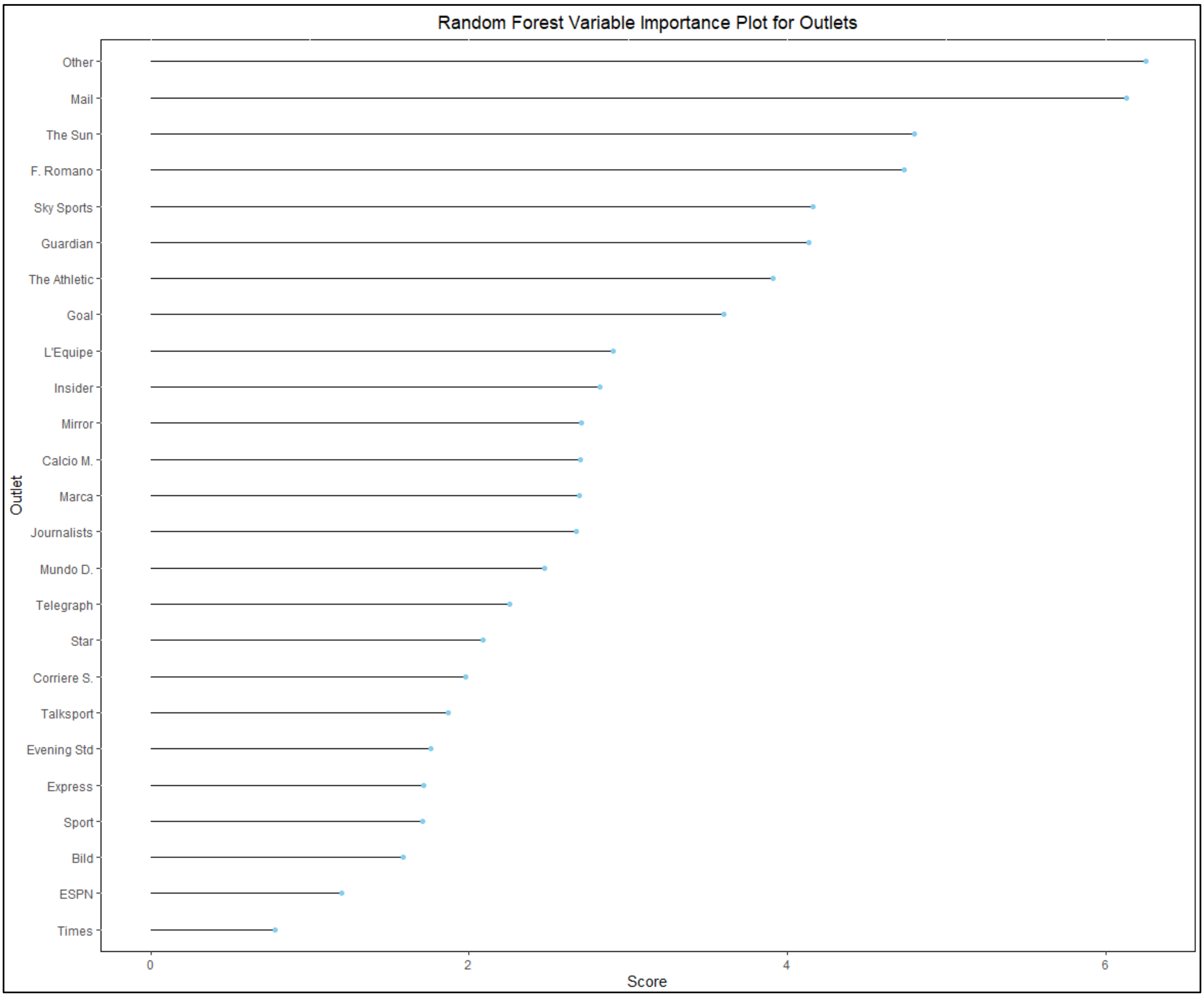
If we look below at *Figure 6.1*, we can observe the variable importance plots for our different variables. Variable importance for classification in Random Forest is calculated by first taking the accuracy of each tree on the out-of-bag sample which is the baseline accuracy. We then permute the predictor variables and retain their accuracy. The importance of the predictor variable is calculated as the difference between the baseline accuracy and the accuracy obtained using the permuted variable. A larger score indicates that the variable plays a larger role in purifying the nodes of the trees. From *Figure 6.1*, Outlet is the most important variable, followed by Agent. Country and the three continuous variables have similar importance whilst League and especially Position trail behind.

Figure 6.1: Variable Importance Plot for Random Forest



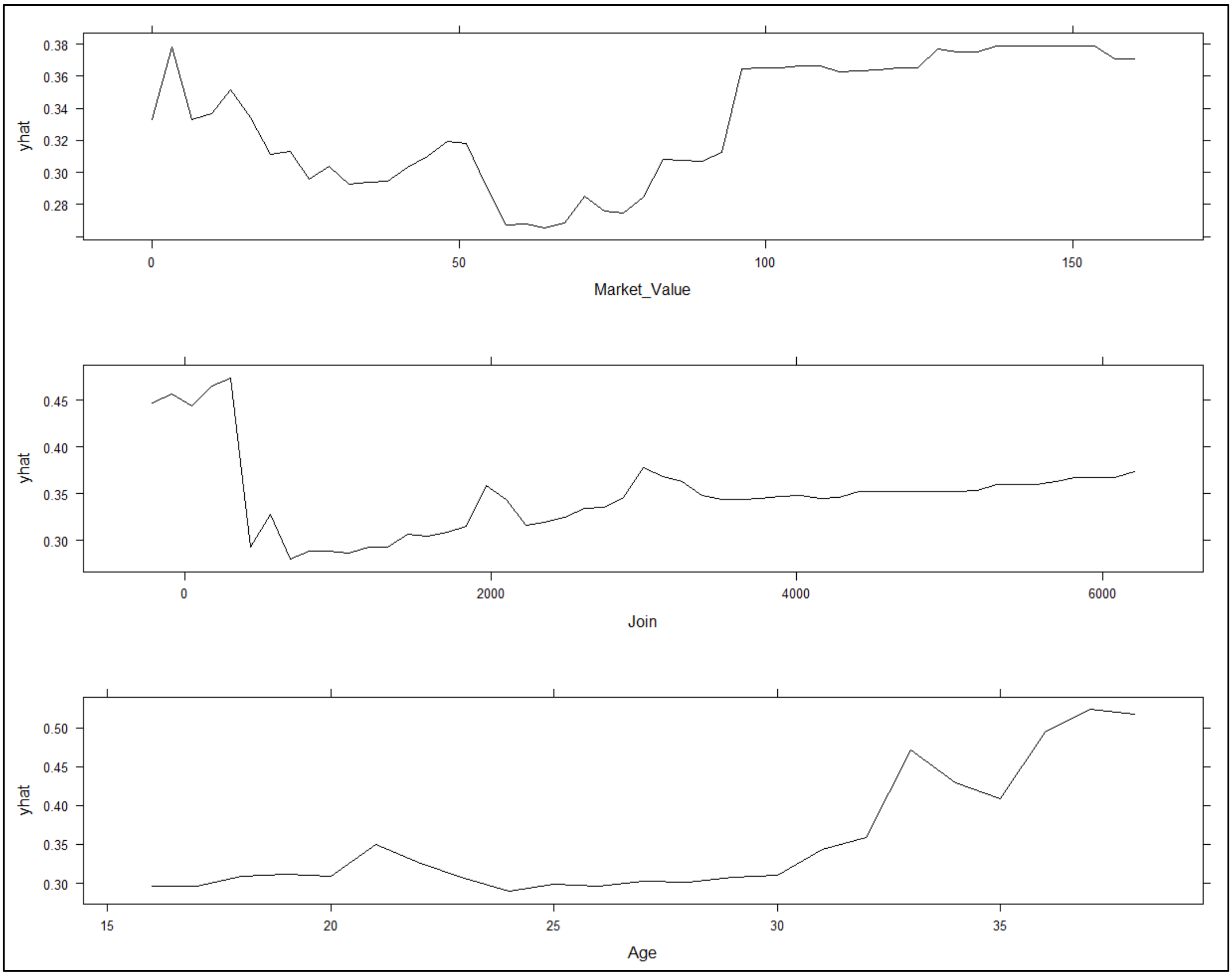
Although we are able to observe the amplitude of how the categories affect the model, we can’t see how individual outlets impact the model. *Figure 6.2* plots the variable importance for the various levels of the “Outlet” variable. The three most important outlets are Fabrizio Romano, The Daily Mail and The Sun. Prevalence in the data set will tend to create more variable importance, especially when other levels of the categorical variable have little occurrences since there is little data to train these levels on. This also explains why “Other” is so important as it is the most common outlet. If we look at *Table 2* in the Appendix, we see that The Daily Mail and The Sun are both part of the top 3 most occurring outlets, however, F. Romano is 12th. This indicates that, despite his relatively low number of occurrences, he must have a very positive or negative impact on the outcome to explain such high importance.

Figure 6.2: Random Forest Variable Importance Plot for Outlets



These findings indicate that outlet is an important variable in predicting accuracy and shows us how the importance of the various outlets varies, however, it does not show how they affect reliability. Do they positively or negatively affect the outcome variable? Through observing the partial dependence plots, we show in which direction our dependent variables affect the outcome. *Figure 6.3* shows the partial dependence plots for the continuous variables Join, Market Value and Age.

Figure 6.3: Partial Dependence Plots for Market Value, Join and Age



Partial dependence plots show how the predicted outcome changes as the chosen predictor varies while keeping all other variables constant. It shows the marginal effect of the predictor variables (Friedman, 2001). With categorical variables, for each of the levels, we replace all instances to have the same level. The Y – axis then represents the probability of the True class. We must be careful when interpreting these graphs as the distribution of the data is not even. For example, for “Join”, at the beginning where there is a steep slow upward followed by a rapid decline, this represents players who have joined their club for less than 200 days. These players are very rare as new players are not often mentioned in transfer rumours, unless for a loan move, and therefore the slope can easily be impacted by a few outliers. A similar phenomenon can be seen with players who have a market value of over 70 million euros, where there is a steep upward trend then the curve flattens. When we look at the data, only 93 players are over 70 million euros. Therefore, the partial dependence plot is strongly affected by a few points but not necessarily representative of the overall trend. For “Join” and “Age”, the general trend seems to be that as player longevity at the club or age increases, the likelihood of the rumour being true increases. This reasoning appears logical. As longevity increases, players are more willing to leave and undertake a new path in their career whether it be for personal accomplishments, monetary gains or joining a better club. With regards to market value, as the player's value increases from 0 to 70 million, the probability of the rumour being true decreases. This supports the idea that outlets will try to print more sensationalist headlines by talking of higher valued players which attract more readers. This could also mean higher market value players and their agents are more likely to spread news regarding their transfer to drive up interest and try to negotiate better contracts.

Figure 6 in the Appendix shows the partial dependence plot for the categorical variable Outlet whilst *Table 6.3* orders the values from the plot in a table.

Table 6.3: Probability of Positive Class per Outlet from the Partial Dependence Plot

Outlet	**Probability Positive Class (%)**
Guardian	44.97
F. Romano	42.56
Goal	37.48
Talksport	37.00
Sport	36.64
Mail	35.72
Sky Sports	34.40
ESPN	34.19
Star	34.15
Express	33.31
Telegraph	33.19
Journalists	33.17
Evening Std	32.69
The Athletic	32.63
Times	31.50
Mundo D.	31.42
L'Equipe	31.41
Bild	31.08
Mirror	30.92
90 Min	30.27
Calcio M.	30.12
Corriere S.	29.41
Marca	29.01
Insider	28.91
The Sun	28.21
Other	28.19

Table 6.3 shows the predicted probability of the positive class given the different outlets, ordered from most to least likely. A higher probability means a higher chance of the transfer occurring and therefore is an indicator of a more reliable news media. We can see that the most unreliable outlets are Marca, Insider and The Sun. Marca and The Sun are notorious for producing a lot of headlines and are commonly known for not being reliable which the plot supports. At the top of the table we find Fabrizio Romano, the Guardian, and Goal who are the most likely to have the outcome be true and show they are more reliable sources of information. Fabrizio Romano, a journalist who is well respected and is even often held as the standard for reliability obtains the second highest probability for the true class, with an over 40% chance of the outcome being True if he reports the rumour. On the other hand, The Sun has a probability inferior to 29%, highlighting a vast disparity that exists between different outlets. Although perceived reliability would have come to similar conclusions, we show empirical evidence, which is more damning and likely to pressure outlets to strive for more accurate reporting of football transfers.

Overall, our findings indicate that outlets play an important role in determining the outcome of a rumour. Fabrizio Romano and The Guardian are found to be the most reliable source of information whilst tabloids like The Sun and Marca are the least reliable. This confirms already pre-existing common beliefs which had not necessarily been backed through data. Additionally, longevity variables such as age and time since the player joined the club are positively correlated with the outcome being true. On the other hand, as the market value increases, up to 70 million, the probability will tend to decrease, possibly indicating that outlets will have more false rumours about highly valued players which are more likely to create a buzz and attention.

6.2.6 Logistic Regression

Similarly to SVM, Logistic Regression underperforms compared to Random Forest. Unpenalized Logistic Regression obtains 72% accuracy whilst Lasso regression gets 68.2%. To interpret the model, we need to look at the model variable coefficients. *Figure 7* in the Appendix shows the top 15 largest coefficients by absolute value. However, simply looking at these values is not enough, we also need to look at the significance level. The significance level represents the probability of incorrectly rejecting the null hypothesis, which is that the coefficient is not significantly different from 0. In *Table 6.4*, we record all the coefficients which are statistically significant at the 10% level. It is also important to note that there is a reference category for each categorical variable which are: "Midfielder" – Position, "Wales" - Country, "The Times" - Outlet, "Wasserman" – Agent, "Serie A"-League. All other levels of the categorical variable are compared to this category and their coefficients represent their relative effects on the outcome compared to the reference category. This implies that we can't compare coefficients across different variables because the reference category is different.

Table 6.4: Logistic Regression Output Table

Variable	Estimate	Std. Error	Z value	P-Value	Sig. Level
Age	-0.040	0.019	2.05	0.030	*
Guardian	1.43	0.82	-2.06	0.082	.
Elite Project Group	1.27	0.59	-2.17	0.032	*
Rafaela Pimenta	1.54	0.60	-2.50	0.010	*
Attacker	0.35	0.17	-2.70	0.038	*
Defender	0.45	0.17	-2.30	0.0084	**
Morocco	-1.52	0.92	-1.66	0.097	.

*Signif. codes: '***' p < 0.001, '**' p < 0.01, '*' p < 0.05, '.' p < 0.1*
Intercept Estimate: -2.7
Note: This regression output table only contains coefficients that are significant at the 0.1 level minimum. There are 97 variables otherwise and the table becomes uninterpretable.
The **standard error** measures the variability or uncertainty in the estimated regression coefficient.
The **z-score** is a measure of how many standard errors the estimated coefficient is away from zero.
The **p-value** represents the probability of obtaining the observed z-score. A small p-value (usually below the significance level of 0.05) indicates that the coefficient is significantly different from zero, suggesting that the predictor variable has a statistically significant effect on the outcome variable.

When we look at statistically significant coefficients, there are few variables that remain. The negative intercept is representative that the majority class is that the transfer will not happen and therefore if we have no information on the rumour and its context, the outcome is more likely to be false. With regards to the outlets, only The Guardian is found to be statistically significant. This supports the notion that it is one of the most important variables in model prediction and the fact it was the most important outlet in Random Forest. Interestingly, in this model, the positions are utilised, and it appears that if the player is an attacker or defender, then the transfer is more likely to happen than if he was a midfielder.

Logistic Regression has the advantage of being a lot more interpretable than Random Forest as we can see which coefficients are statistically significant and can calculate the percentage change that a certain value of a variable will have on the probability of the outcome being true. It also shows linear relationships which are easily interpretable. However, it fails to match the accuracy of Random Forest, which potentially shows that there are little linear relationships between the outcome variable and the independent variables.

7 Conclusion

This thesis aimed to shed light on the issue of media reliability in the domain of football transfers. The transfer market is an enormous enterprise with billions of euros at play in every single transfer window. With so much at stake, the coverage of the transfer market attracts millions of readers and watchers creating an incentive to be the news outlet that breaks the next big story. However, unlike traditional media where an article or piece of information can be fact-checked, with transfer rumours we can rarely verify the sources that the sports outlet uses. Especially since neither the sources nor the outlet is willing to give that information as it puts them at risk of losing their sources of information. Furthermore, the nature of rumours is that they are uncertain and only evoke a possibility. Therefore, it is almost impossible to understand when a rumour is based on reliable intel and when it is only for clicks and views. Through this thesis, we hoped to bring empirical evidence and show which outlets are most reliable by creating a machine learning model capable of predicting the outcome of a rumour, which is represented by whether the transfer happens or not. Through this, we can see how different outlets impact the model and see which outlets positively affect the outcome of the transfer happening, using this as a reliability measure.

The first part of this study consisted in building two datasets: one consisting of rumours from the BBC gossip page and another detailing information about players. Although the rumours data set can be improved upon and suffers from selection bias, the player's data has the potential to contribute to further studies on the topic of the footballing world. One of the main challenges was extracting the meaning behind a rumour. Through the use of named entity recognition, we successfully manage to identify almost every player and club in the rumours. Our findings showed that using a dictionary-based approach was much simpler and still effective compared to a rule-based approach. Moreover, by applying sentiment analysis to the words following the club names, we were able to assess if there was genuine interest for a transfer in a rumour. Although not perfect, this method's simplicity and efficiency made it viable for our analysis and demonstrated that sentiment analysis can have a wider user than the more traditional domains we typically associate it with.

Once the data-gathering process was completed, we proceeded to try out a multitude of machine-learning models to understand the underlying patterns in our data. Our findings showed that Logistic Regression was poor at predicting the outcome, potentially indicating a lack of linear relationships between the independent variables and the outcome. In contrast, Random Forest was the best model and managed to improve from a 67.9% no-information rate to a 77% accuracy in predicting whether a rumour would materialise or not. When we look at the variable importance plot, we observe that “Outlet” held the most importance. This indicates that there exists an important difference in the accuracy of rumours reported by different outlets as their presence greatly affects the outcome variable. Agent was the second most important variable also showing the impact that they can have on rumours.

However, we must be careful when interpreting this category as it had many levels with few occurrences. We then proceeded to analyse the partial dependence plots. The presence of the top outlet results in a 45% chance of the outcome being true. On the other hand, the worst outlet has a 28% probability. These findings confirmed pre-existing held beliefs about sports outlets' reliability. Journalist Fabrizio Romano, who is known for his reliability, is found to be the second most reliable outlet. On the other hand, popular tabloids such as Marca or The Sun, which have a reputation for printing fake news, achieved the worst accuracy.

Overall, this study brought empirical data and evidence that there does exist a disparity in the accuracy of rumours reported by different outlets. However, the presence of selection bias and the inability to extract the degree of certitude that a rumour is evoking are major limitations of this study. We are able to conclude that there are disparities but cannot make assertations about the reliability of specific outlets without further research.

7.1 Implications

By shedding light on the disparity in the accuracy of transfer rumours from various outlets, we aim to try and encourage more accurate reporting. Fans can use this thesis as a basis for understanding which outlets are most reliable. If we were to push the analysis further and focus on specific outlets, we could gain even further insights and see the interaction effects between outlets and other factors such as the club. This means we would know when to believe an outlet based on the context of the rumour and not just the outlet's reputation. Ultimately, the goal would be for fans to have the ability to curate their news and select the outlets which best suit their interest. If fans do indeed value accuracy over pure speculation, this would reward outlets which are more diligent in their reporting and apply pressure for outlets to have more credible sources before reporting on a rumour.

Unfortunately, we do not believe our model to be good enough for clubs to utilise and make decisions worth millions of dollars based on it. Similarly to the fans, they can use it as a proxy for assessing the accuracy of a rumour and have empirical data to support it rather than only pre-existing experiences. However, due to the limitations of the model, it would be unreasonable to take the model output as fact. Despite this, we hope that the model can still help clubs to gain a better understanding of the state of the transfer market and decrease the imbalance between the big and small clubs.

7.2 Limitations and Further Research

While we successfully created a database of articles to extract insights into the reliability of media outlets, our research does have certain limitations and potential for improvement.

Firstly, the most significant limitation is selection bias. This bias means we cannot confidently make conclusions on a particular outlet's reliability, as the articles from that outlet were selected by another party and we cannot assert that this chosen sample is representative of the outlet. As a result, drawing conclusions about these outlets is challenging, and we make the assumption that our samples represent the various newspapers or journalists we mention. To address this issue, we'd need to select specific outlets and gather all their rumours within a designated timeframe. This approach would ensure a representative sample, allowing conclusions about reliability without the influence of selection bias.

The second limitation is our dataset's size. Due to the vast number of categorical variables and the need for a certain number of occurrences to make them meaningful, we often have to group levels with few occurrences, resulting in information loss. While gathering more data could prevent this, the biggest cause of data loss in our process was the natural language processing steps inherently resulting in data loss. For instance, we retain only articles that identify one player and more than one club, leaving out many rumours that mention multiple players or just one club. The methodology could be improved upon to be able to recognize multiple potential transfers within a rumour or detect player swap rumours mentioning two players. The end goal would be to be able to perfect the text processing techniques to recognise and identify all the players, clubs, and potential scenarios that a rumour can evoke.

In our predictions, we heavily rely on external player information and utilize little contextual or content details. For contextual information, considering fan reactions or comments on rumours might bring additional insights. Another piece of key information is whether the same rumour has already been posted. If we could track the first outlet to break the news, we could identify deeper connections such as interaction effects with clubs or players Furthermore, we could delve deeper into the content of the rumour. Rumours often express a different degree of certitude, sometimes we may be saying a transfer is close to happening and sometimes we may simply report that a club is keen on signing a player but has not made any offers or approaches. By identifying these differences, we would account for outlets that tend to publish rumours which are less likely to happen as long as we can identify that they are themselves acknowledging this in the text. We could then create a category that tracks the degree of certitude expressed obtaining a fairer assessment of the outlet's reliability.

Lastly, our model choices present another limitation. Although we experimented with multiple models, we can't definitively say that Random Forest is the optimal model for this data type. Techniques like Neural Networks or XGBoost may yield superior results. We didn't test all machine learning methods, so we are unable to conclude that we've selected the best model to predict a transfer rumour's validity.

Appendix

Figure 1: Flowchart of data gathering process

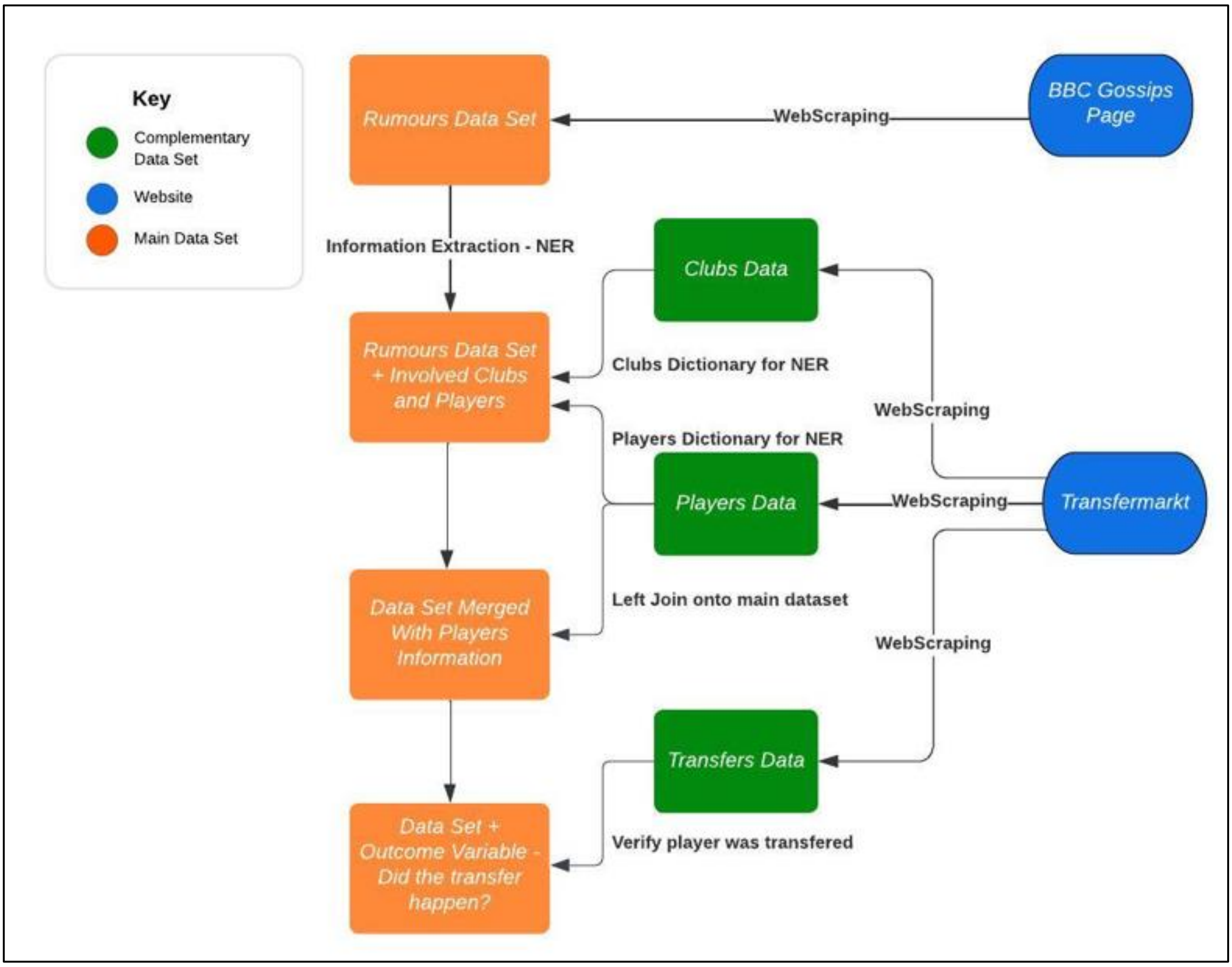

Key
Complementary Data Set
Website
Main Data Set
Rumours Data Set
WebScraping
BBC Gossips Page
Information Extraction - NER
Clubs Data
Rumours Data Set + Involved Clubs and Players
Clubs Dictionary for NER
WebScraping
Players Dictionary for NER
Players Data
WebScraping
Transfermarkt
Data Set Merged With Players Information
Left Join onto main dataset
WebScraping
Transfers Data
Data Set + Outcome Variable - Did the transfer happen?
Verify player was transfered

Table 1: Number of levels depending on the threshold n

Threshold	League	Country	Outlet	Agent	Total
n > 0	11	60	139	253	463
n > 4	11	42	40	82	175
n > 9	11	32	36	31	110
n > 14	10	25	30	19	84
n > 19	10	21	26	14	71

Figure 2: Machine Learning Model Accuracy (%) based on categorical variable threshold

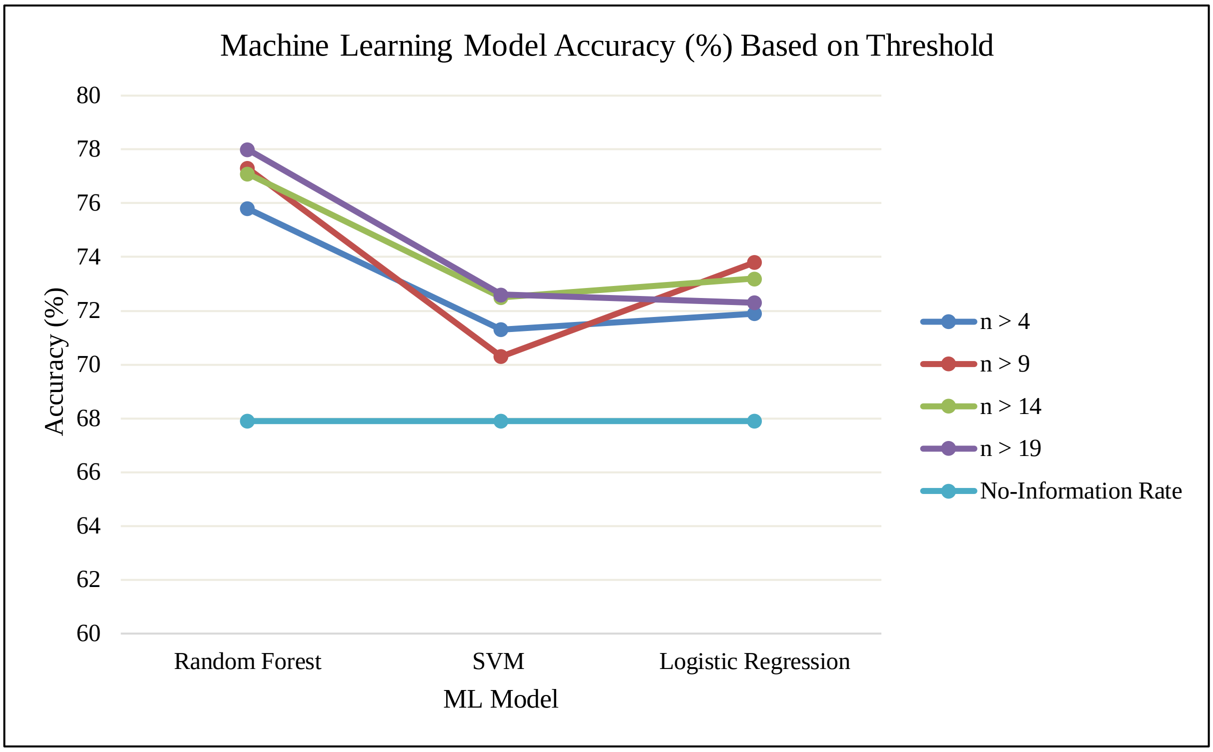


Figure 3: Performance metrics (%) of various ML models

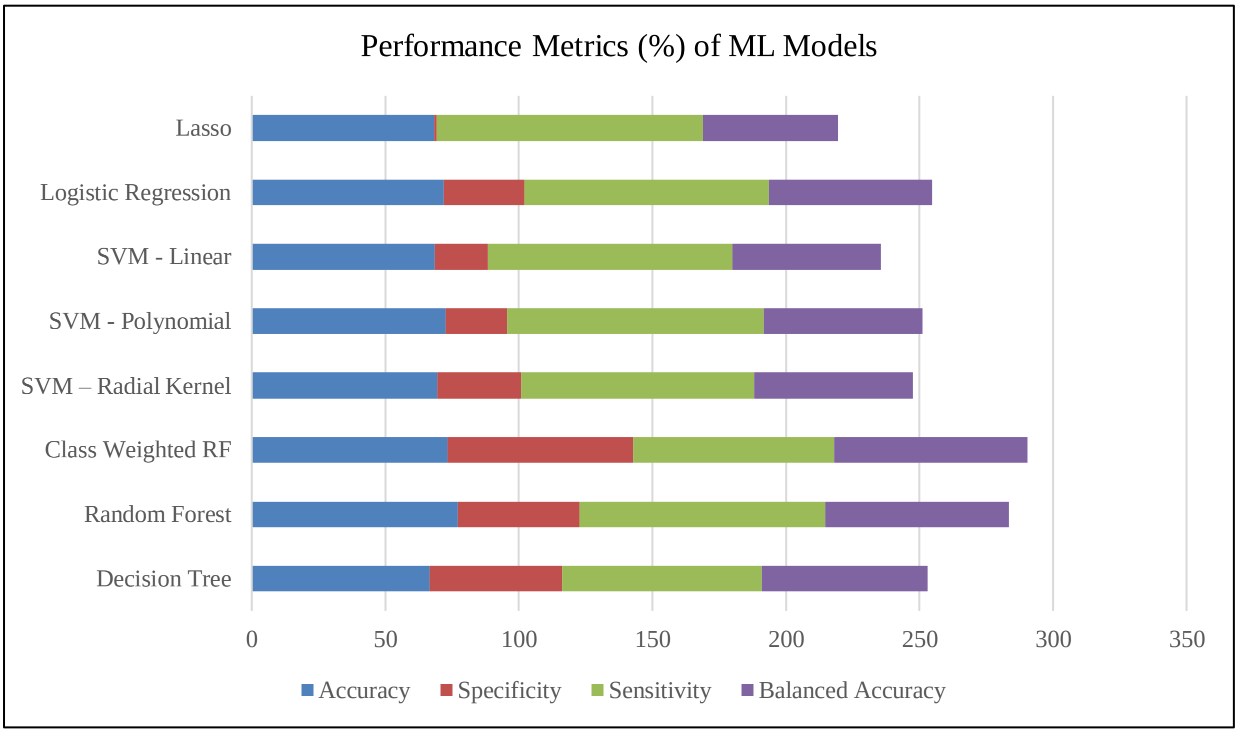


Figure 4: X-val relative error compared to complexity parameter value

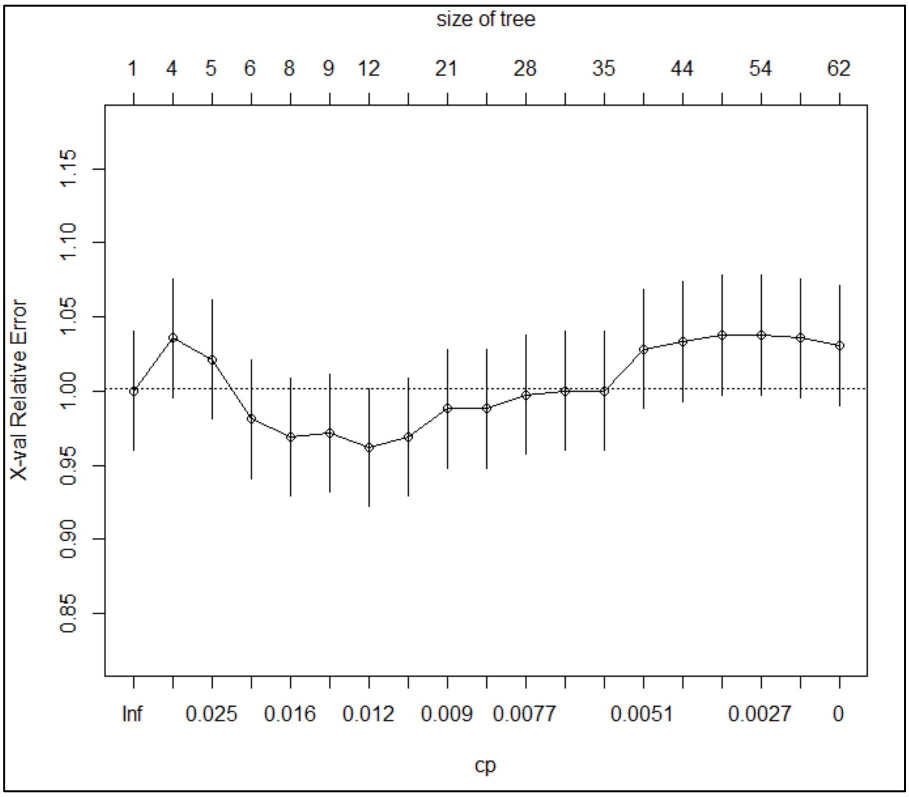


Figure 5: Cross-Validated Accuracy of RF based on number of randomly selected predictor

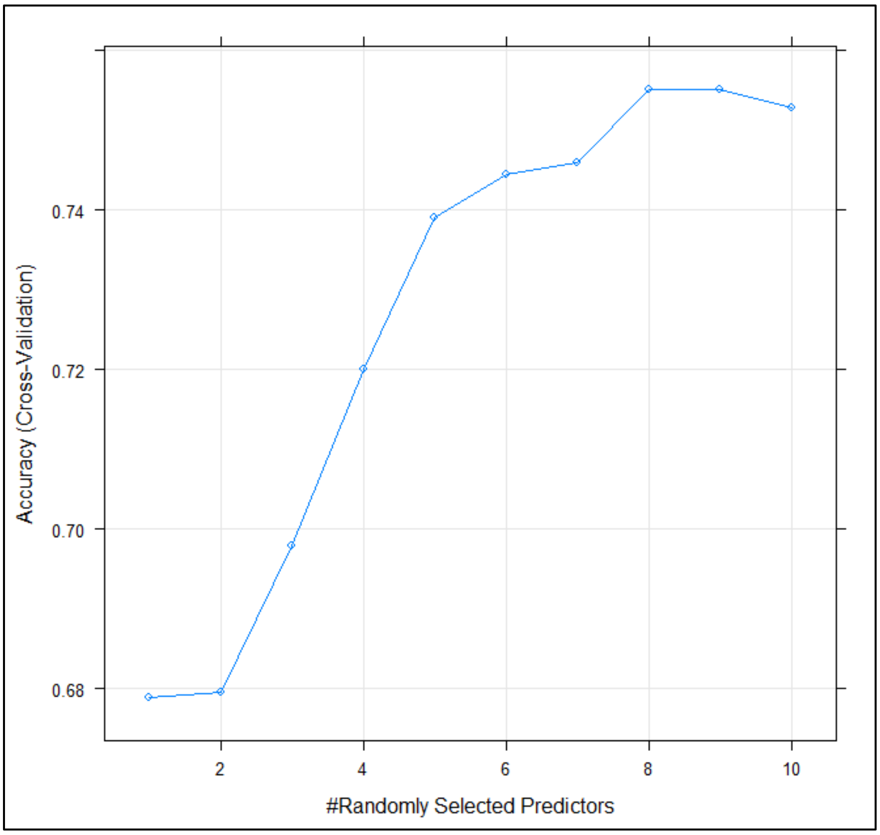


Figure 6: Random Forest Partial Dependence Plot for Outlets

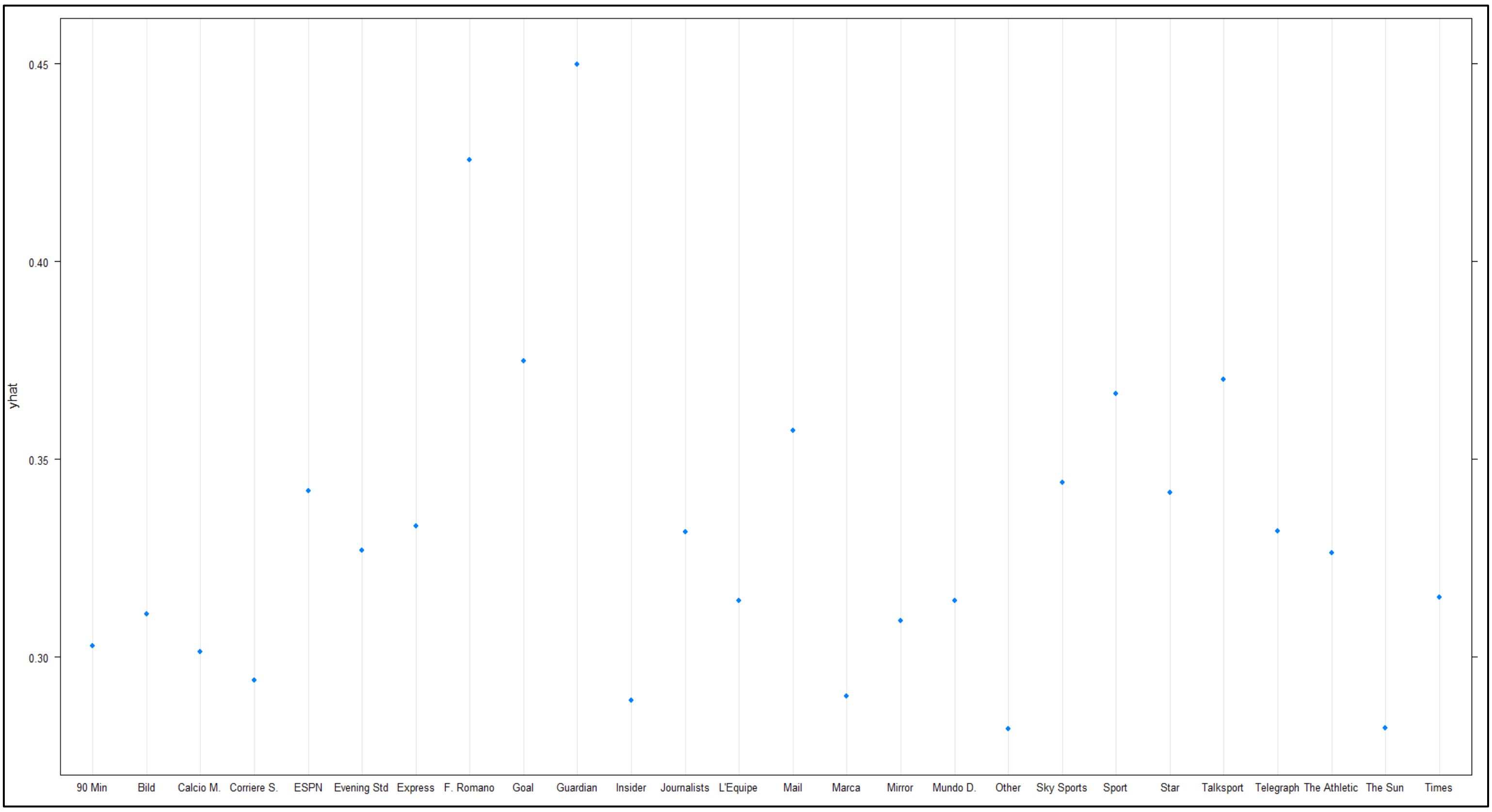


Table 2: Number of occurrences per Outlet in final Data Set

Outlet	**Number of Occurrences**
Other	287
Mail	146
The Sun	123
Sky Sports	93
The Athletic	92
Insider	73
Mirror	72
90 Min	64
Marca	63
Calcio M.	62
L'Equipe	58
F. Romano	55
Corriere S.	53
Journalists	45
Goal	42
Telegraph	40
Mundo D.	39
Guardian	30
Bild	30
Evening Std	29
Star	28
Times	27
Express	25
Sport	22
ESPN	20
Talksport	20

Figure 7: Top 15 Coefficient Scores Logistic Regression

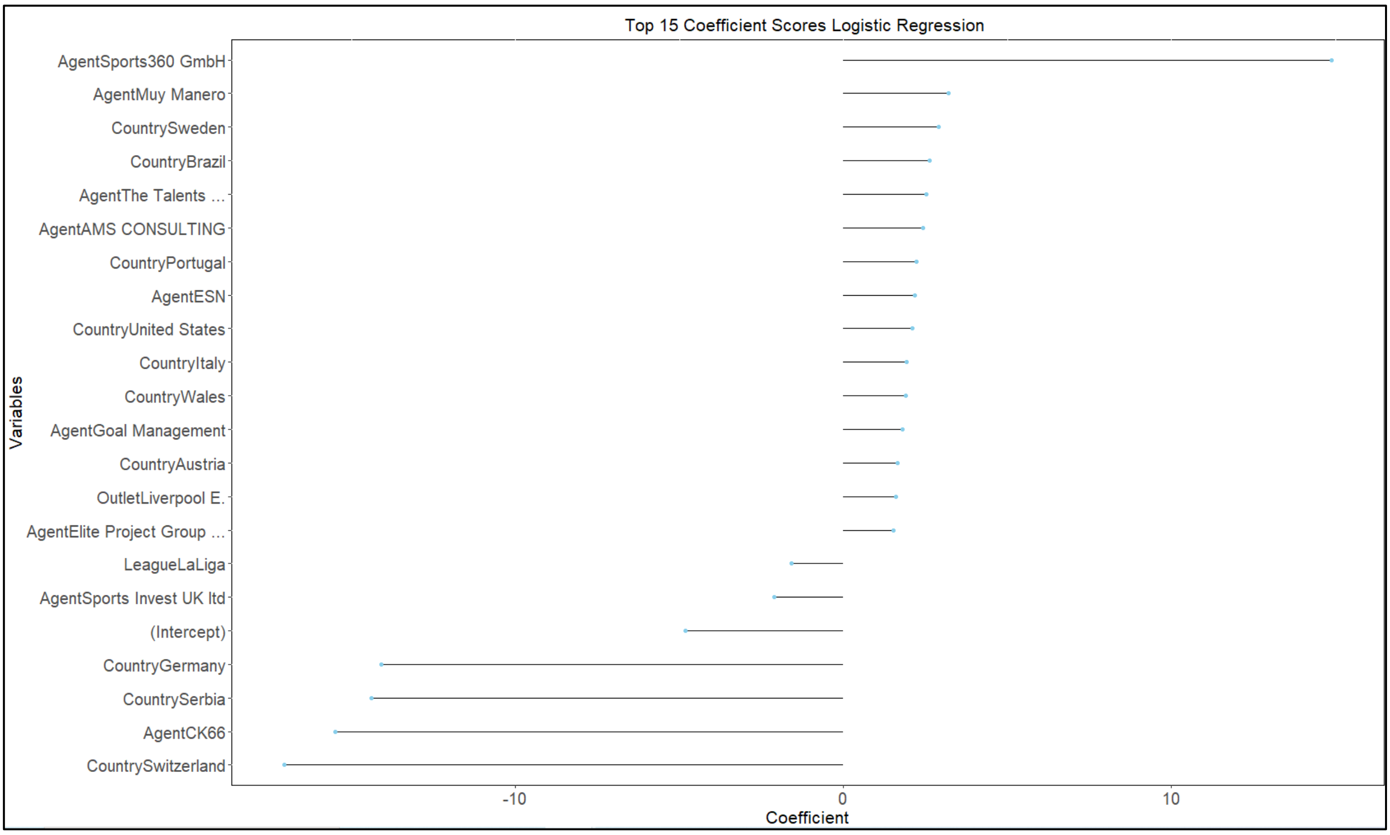
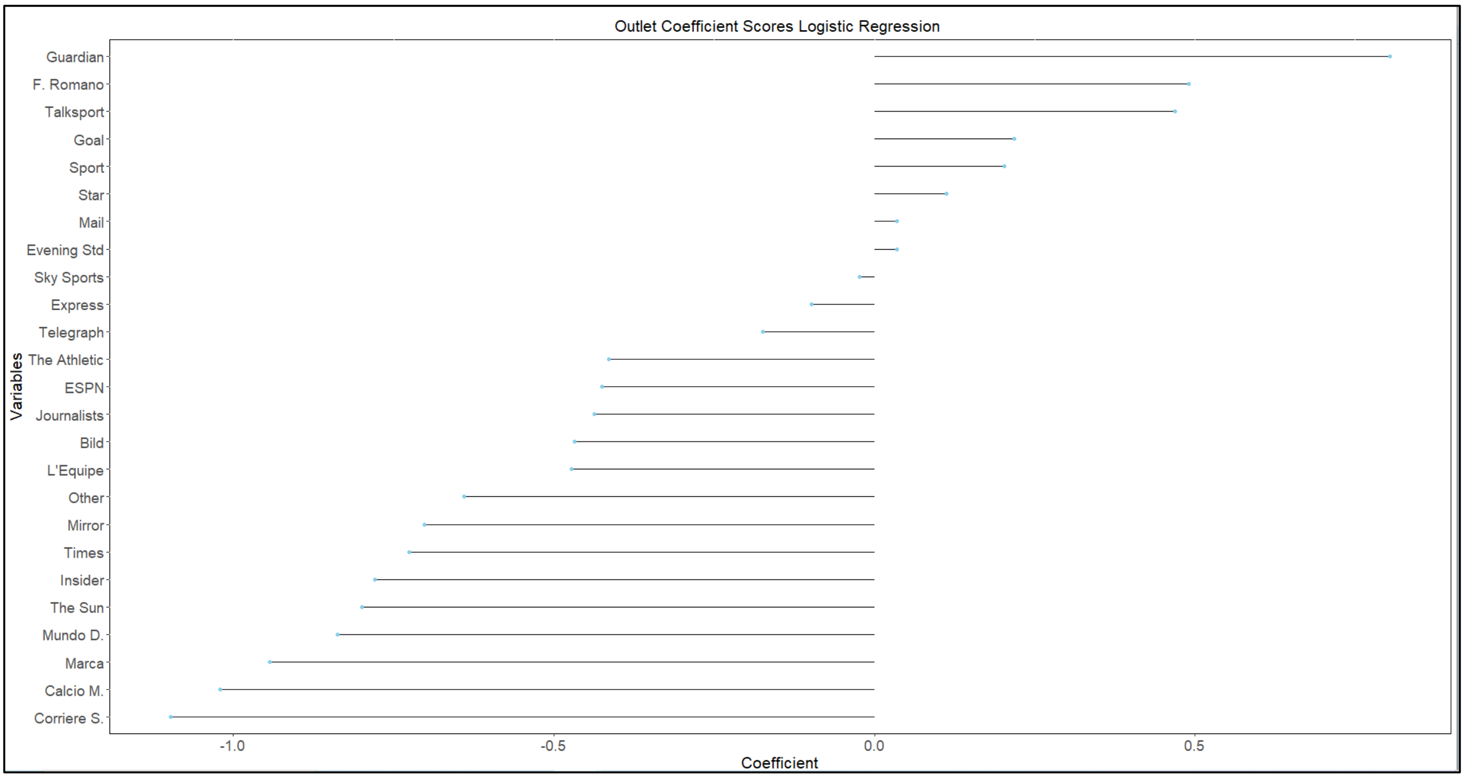


Figure 8: Outlet Coefficient Scores Logistic Regression



Bibliography

Allcott, H., & Gentzkow, M. (2017a). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130.

Belgiu, M., & Drăguţ, L. (2016). Random Forest in remote sensing: a review of applications and future directions. *Isprs Journal of Photogrammetry and Remote Sensing, 114,* 24–31.

Biau, G., Scornet, E. (2016). A Random Forest guided tour. *TEST* 25, 197–227

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497,* 38–55.

Breiman, L. (2001) Random Forests. *Machine Learning* 45, 5–32

Carmichael, F., & Thomas, D. (1993). Bargaining in the transfer market: Theory and evidence. *Applied Economics*, *25*(12), 1467–1476.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8–10), 1477–1494.

Cha, G.-W., Moon, H.-J., & Kim, Y.-C. (2021). Comparison of Random Forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. *International Journal of Environmental Research and Public Health*, *18*(16)

Chen, C.-D., & Kutan, A. M. (2016). Information transmission through rumours in stock markets: A new evidence. *Journal of Behavioural Finance*, 17(4), 365–381.

Cortes, C., Vapnik, V., (1995). Support-vector networks. *Machine Learn*ing 20, 273–297

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178–3192.

Faustini, P. H. A., & Covões Thiago Ferreira. (2020). Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158.

Flaxman, S., Goel, S., & Rao, J. M. (2016a). Filter bubbles, Echo Chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320.

Fogg, B. J., & Tseng, H. (1999). The Elements of Computer Credibility. In CHI'99 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 80-87). New York: ACM.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5).

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797.

Hall, S., Szymanski, S., & Zimbalist, A. S. (2002). Testing causality between Team Performance and Payroll. *Journal of Sports Economics*, 3(2), 149–168.

Heinbach, D., Ziegele, M., & Quiring, O. (2018). Sleeper effect from below: long-term effects of source credibility and user comments on the persuasiveness of news articles. *New Media & Society*, *20*(12), 4765–4786.

Hendricks, K., & Porter, R. H. (1988). An empirical study of an auction with asymmetric information. *The American Economic Review*, 78(5), 865–883.

Herrero-Gutiérrez, F.-J., & Urchaga-Litago, J.-D. (2021). The importance of rumours in the Spanish Sports Press: An Analysis of news about signings appearing in the newspapers Marca, as, Mundo Deportivo and sport. *Publications*, *9*(1), 9.

Ittoo, A., Nguyen, L. M., & van den Bosch, A. (2016). Text analytics in industry: Challenges, Desiderata and Trends. *Computers in Industry*, 78, 96–107.

Kalsnes, B. (2018, September 26). Fake news. *Oxford Research Encyclopedia of Communication.*

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software*, *15*(9), 1–28.

Kelly, S., & Chatziefstathiou, D. (2017). 'Trust me I am a football agent'. the discursive practices of the players' agents in (un)professional football. *Sport in Society*, 21(5), 800–814.

Kotsiantis, S.B. (2013). Decision trees: a recent overview. Artificial Intelligence Review 39, 261–283.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, 21(4), 475–492.

LaValley, M. P. (2008). Logistic Regression. *Circulation*, *117*(18), 2395–2399.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The Science of Fake News. *Science*, 359(6380), 1094–1096.

Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering, 34*(1).

Local journalism: The decline of newspapers and the rise of digital media: The decline of newspapers and the rise of Digital Media. (2015). I. B. Tauris & Company.

Loh, W. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Mailath, G. J., & Postlewaite, A. (1990). Asymmetric information bargaining problems with many agents. *The Review of Economic Studies*, *57*(3), 351.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards &amp; Interfaces*, *35*(5), 482–489.

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for Logistic Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(1), 53-71.

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, *72*, 72–81.

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2018). *Reuters Institute Digital News Report 2018*. Reuters Institute for the Study of Journalism.

Noble, W. (2006). What is a support vector machine? *Nature Biotechnology* 24, 1565–1567

Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, *37*(5), 2671–2692.

Pendleton, S.C., (1998), Rumour research revisited and expanded, *Language & Communication*, *18*(1), 69–86.

Porta, M. S. (2014). *A dictionary of epidemiology*. Oxford University Press.

Stoltzfus, J. C. (2011). Logistic Regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104.

Stromer-Galley, J. (2004). On-line interaction and why candidates avoid it. *Journal of Communication*, 54(3), 467–482.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2017). Defining "fake news." *Digital Journalism*, *6*(2), 137–153.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). How does fake news spread on Twitter? Three patterns of information diffusion. Journalism Practice, 12(4), 456-471.

Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artificial Intelligence Research*, 2(1).

Varol, O., et al. (2017) in Proceedings of the 11th AAAI Conference on Web and Social Media (*Association for the Advancement of Artificial Intelligence, Montreal*), pp. 280–289.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and fake news online. *Science*, *359*(6380), 1146–1151.

Wathen, C. N., & Burkell, J. (2002). Believe it or not: factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology*, 53(2), 134.

Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: a meta-analysis of effect size. *Journal of the Academy of Marketing Science: Official Publication of the Academy of Marketing Science*, 21(2), 101–112.

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The web of false information: rumours, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, *11*(3).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4).

Zhou, X., & Zafarani, R. (2020). A survey of fake news. *ACM Computing Surveys*, *53*(5), 1–40.