



University of Pennsylvania
ScholarlyCommons

Joseph Wharton Scholars


Wharton Undergraduate Research

5-2020

Transfer Pricing: An Analysis Of The Impact Of Player Brand Value On Transfer Fees In European Football

Karl Valentini

Follow this and additional works at: https://repository.upenn.edu/joseph_wharton_scholars

 Part of the [Marketing Commons](#), and the [Statistics and Probability Commons](#)

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/joseph_wharton_scholars/94
For more information, please contact repository@pobox.upenn.edu.

Transfer Pricing: An Analysis Of The Impact Of Player Brand Value On Transfer Fees In European Football

Abstract

In the past decade, the English Premier League has dramatically grown its commercial appeal and global audience. Yet, there has been little research on how the league's growth could change the dynamics of the labour market for players. This study provides a unique framework for investigating player brand value through media coverage and exposure in order to determine its impact on player worth. Measuring player worth through the transfer fees that clubs pay to acquire players, this study provides insights on which factors are most predictive of transfer fee amounts. Throughout this study's analysis, machine learning techniques are implemented and compared in order to assess the impacts of player brand value and determine a best fit model for predicting transfer fees. This study finds evidence that player brand value affects transfer fees and that the affect changes over time, but it concludes that more research is still needed.

Keywords

Branding, Media, Machine Learning, Football, Transfer Fees, Sports

Disciplines

Business | Marketing | Statistics and Probability

**TRANSFER PRICING: AN ANALYSIS OF THE IMPACT OF PLAYER BRAND
VALUE ON TRANSFER FEES IN EUROPEAN FOOTBALL**

By

Karl Valentini

An Undergraduate Thesis submitted in partial fulfillment of the requirements for the

JOSEPH WHARTON SCHOLARS

Faculty Advisor:

Eric T. Bradlow

K.P. Chao Professor, Marketing, Statistics, Education and Economics

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MAY 2020

Abstract

In the past decade, the English Premier League has dramatically grown its commercial appeal and global audience. Yet, there has been little research on how the league's growth could change the dynamics of the labour market for players. This study provides a unique framework for investigating player brand value through media coverage and exposure in order to determine its impact on player worth. Measuring player worth through the transfer fees that clubs pay to acquire players, this study provides insights on which factors are most predictive of transfer fee amounts. Throughout this study's analysis, machine learning techniques are implemented and compared in order to assess the impacts of player brand value and determine a best fit model for predicting transfer fees. This study finds evidence that player brand value affects transfer fees and that the affect changes over time, but it concludes that more research is still needed.

Keywords

Branding, Media, Machine Learning, Football, Transfer Fees

Disciplines

Statistics, Marketing, Business, Sports

Table of Contents

<i>SECTION ONE: INTRODUCTION</i>	4
1.2 Research Question	4
1.2 Research Contributions	6
1.3 Hypotheses	8
<i>SECTION TWO: BACKGROUND AND LITERATURE REVIEW</i>	9
<i>SECTION THREE: DATA COLLECTION AND ANALYSIS</i>	15
3.1 Data Overview	15
3.2 Data Collection	17
3.3 Exploratory Data Analysis	21
<i>SECTION FOUR: RESEARCH METHODOLOGY</i>	29
4.1 Overview of Statistical Modeling Frameworks	29
4.2 Multiple Regression	30
4.3 LASSO Regression	33
4.4 Decision Trees	35
4.5 Limitations	39
<i>SECTION FIVE: RESULTS AND DISCUSSION</i>	40
5.1 Comparing Regression Models	40
5.2 Comparing Decision Trees	57
5.3 Discussion of Results	64
<i>SECTION FIVE: CONCLUSION AND SUMMARY</i>	71
5.1 Concluding remarks	71
5.2 Areas for further research	71
<i>WORKS CITED</i>	73
<i>APPENDICES</i>	77
Appendix I	77
Appendix II	77

SECTION ONE: INTRODUCTION

Throughout the world, people enjoy and consume sports for leisure. As fan bases for some sports have grown into the hundreds of millions, or even billions, the industry of professional sports is estimated to be worth approximately \$91bn (Wetzel, Hattula, Hammerschmidt, and van Heerde 2018). Since the turn of the decade, the commercial success of European football, one of the leading international sports, has enabled major European football leagues and their teams to procure increasingly lucrative television, broadcasting and sponsorship deals (Ahmed 2019). Over that time, England's Premier League has solidified its status as the most commercially successful football league in the world.

1.2 Research Question

With its commercialization, the economics of football can be likened to those of a business with its own economic model and relations with the wider economy. As teams employ players and profit from paying fans, their behaviors parallel with those of corporations. Much like any financial business, a football team seeks to grow revenue and profits, has employees, and provides a product, in this case football matches, for paying customers (Muller, Simons and Weinmann 2017). Drawing on the fact that football players operate in a unique labour market system, this study examines how football teams act as economic agents in the labour market for their main employees: professional football players. Although there are many aspects underlying the relationships between players and clubs, this paper focuses on how clubs determine the worth of players when engaging in transfers to buy and sell them. In particular, this study aims to understand how much value football teams place on brand association – the value beyond objective football performance – when buying players. For that aim, the motivating research

question is whether a player's personal brand translates to clubs paying an extra premium in the transfer fee to acquire that player above the amount paid for sporting performance and acumen.

This paper studies football transfers and the English Premier League for several reasons. First, the labour market in sports differs from the common labour market because employees (i.e. players) are tied to contracts that limit their ability to move between firms (i.e. teams). Within sports, football is also unique because the player-reservation system includes player cash sales, unlike the systems in professional American sports which are restricted to player-for-player swaps (Dobson and Gerrard 1999; Dobson, Gerrard and Howe 2000). The cash amount that a buying club pays a selling club to buy the rights to a player's contract is referred to as a transfer fee and this fee sets a market price to a player's economic worth (Muller et al. 2017). Assuming that teams act as rational economic agents when completing transfers, they should be paying no more than the value of total profits a player can bring. From a sporting perspective, on-field performance is important for driving wins and championships, which boost popularity and profits. However, there are other non-field contributions that can also bring a club greater popularity, and with that, profits. Even if some players do not necessarily help their teams win, they can be fun to watch, popular with customers (i.e. the fans), savvy on social media and a draw for advertisers and sponsors who want to be associated with specific players. From that perspective, transfer fees should represent both the combined on-field and off-field value a player provides to a team. Accordingly, studying transfer fees offers an opportunity to directly measure in monetary terms how employers value their employees in a unique market setting.

Second, the institutional features of the English Premier League make it appropriate for addressing the marketing question of brand identity for a sponsoring firm, which in this case is the football club buying a player. The league's latest media rights deal for the next three years

generates an estimated £9.2bn for clubs and provides global exposure (Ahmed 2019). The deal's size reflects how successfully the Premier League and its teams have marketed themselves across the world. While clubs used to depend on having large stadiums to fill with local fans and generate large gate receipts, Premier League clubs now depend largely on media rights and sponsorships for the majority of their revenues compared to teams in other leagues, while their fanbases now span the globe. For clubs to continue generating prodigious amounts of revenue from these new sources, they must rely on marketing and branding themselves, and by extension their players, in order to add to their global customer bases. As a result, a player's brand value is more important in today's game than it ever has been in the past. It can bring direct economic benefits for the player and his club. Accordingly, club perceptions of players can have important implications for sports marketing and marketing more generally. As profits have swelled, Premier League clubs have also been the biggest spenders in the transfer market. During the last summer transfer window, over half of the clubs in the Premier League broke their respective transfer records, with total spending of Premier League clubs exceeding £1bn for the fourth successive summer (Ahmed 2019). Since the sums involved are so large, recruitment strategy impacts on-field performance and the off-field performance of the bottom line. The decisions to buy and sell players become economic ones for Premier League teams, and so studying transfer fees is important for learning about how economic agents behave.

1.2 Research Contributions

By considering players and teams as economic agents, this paper aims to contribute to marketing and sports strategy research more generally. While this study focuses on the brand value of individual football players, the notion of individual brand value in business is an important one. It is now common practice for firms, across industries, to use celebrity

endorsements, and academics have been studying how firms use celebrities and influencers in their marketing efforts (Ding, Molchanov and Stork 2010; Bartz, Molchanov and Stork 2013). However, studying celebrity endorsements is difficult because the mechanisms through which a celebrity influences a purchase decision can be opaque and tying purchase decisions directly to the characteristics of celebrity endorsers is hard. On the other hand, with football, there are many prices going back and forth between clubs in the form of transfer fees. In that sense, the transfer market can be likened to the stock market in which individual stocks are traded so that market-clearing equilibrium prices can be found. Thus, there is an opportunity to measure more concretely how an individual brand can contribute to the success of a larger business, which for this study is a football team. The possible implications for understanding how to value individual brands is what makes this research so important. Moreover, while large profits usually follow sporting success, there is a confluence of many factors that drives profits, and the relative importance of individual factors can be unclear. By isolating for off-field characteristics that affect transfer pricing, this research seeks to contribute to the understanding of which factors drive profitability across entertainment-based and media-centric businesses, such as sports franchises. Accordingly, it also aims to contribute to the knowledge of transfer pricing, which has already been studied in a variety of ways within the context of football specifically (Arai, Ko & Ross, 2013; Torgler and Schmidt 2007; Bryson, Frick and Simmons 2012; Dobson et al. 1999; Dobson et al. 2000).

In addition, this paper will contribute to the field of sports analytics. Building on the existing literature, this study provides new methods for estimating a football player's economic worth by applying better statistical estimation techniques than most past studies have used. It also introduces and examines new factors, such as those related to branding, that can be of

relevance for evaluating a player's value. Both of these contributions are important because there are several features which make prediction particularly complex in football, as compared to other sports. Unlike in racquet sports, like tennis, football matches have a designated duration, officially lasting ninety minutes. There is also only one type of scoring event and it yields one point, whereas in American football, for example, it is possible to get one, two, three or six points depending on the different types of scoring events (Herbinet 2018). Unlike most other sports, the outcomes of matches can also take three forms: a draw, a win or a loss. Finally, part of what makes football, and accordingly a player's monetary value, so difficult to predict is that the game is so low scoring. Often, the match is decided by only one goal, which adds a high degree of randomness to outcomes (Herbinet 2018).

1.3 Hypotheses

In addressing the main research question discussed above, this study posits as its main hypothesis that:

H1: A player's brand value should have a statistically significant effect on the transfer fee amount that a club pays to buy a player.

A corollary hypothesis is that:

H2: The effect of player brand value on transfer fees should become stronger over time.

In evaluating these hypotheses, this paper empirically tests how many and which characteristics influence the transfer fees paid for players, as well as the magnitudes thereof. In doing so, this study evaluates how characteristics associated with a player's ability to generate brand translate to the transfer fee paid, after controlling for objective sporting performance.

To help deal with the complexity of estimation in football, this study provides an updated dataset covering several important metrics related to the Premier League and aggregates data on

player, team and media characteristics, which can be used for a variety of different types of statistical studies. Given the existing literature mostly uses data from the early 2000s, and few studies include data on the Premier League, this new dataset can be particularly useful. Moreover, since few studies have examined football player brand value, this study establishes a new method for estimating brand value by equating it to media exposure. Celebrity endorsers garner media attention through their personal brands, which they use to showcase the products they are attached to (Bartz et al. 2013). Football players, as celebrity endorsers of the teams they play for, should also experience a link between their personal brands and media coverage. As a result, this study proposes media exposure, as measured by a player's relative popularity in the media, to serve as a proxy for player brand value.

Finally, whereas past research has largely been limited to the use of linear regression models, this study tests a variety of machine learning models to determine which statistical models provide the best predictive capabilities. A model is deemed successful if it can closely predict a player's transfer fee, measured against the real transfer fee paid by clubs. Machine learning models have started to be applied in football, and sports, particularly within the context of sports betting, but not as much by academics yet (Herbinet 2018). The use of machine learning models fits into the existing literature on using machine learning models to predict outcomes, while expanding on its use by considering applications in sports. Its use addresses the desire to apply data science techniques in sports analytics and could embolden more academics to use machine learning models in their research of sports, marketing or business more generally.

SECTION TWO: BACKGROUND AND LITERATURE REVIEW

Since the 1950s, researchers have studied American professional team sports, such as baseball, basketball and football, for a variety of economic implications. Studying team sports

and league structures allows for research on competitive labor markets, labor-management, economic rewards and competitive firm behavior. Only recently have academics turned to football. The literature can be divided into three parts: player value in football, brand value in football and brand value as a whole.

The first, and largest, group of studies looks at player economic value as the dependent variable. Economic value can be studied according to market value, remuneration or transfer fee paid. Whereas a transfer fee represents the actual amount paid, market value represents the estimated worth of a player at any given time (Muller et al. 2017). Most studies have considered different player characteristics, team performance characteristics, selling club characteristics and buyer club characteristics in economic analysis (Torgler and Schmidt 2007; Bryson, Frick and Simmons 2012; Dobson et al. 1999; Dobson et al. 2000).

Player characteristics are inevitably linked to market value and remuneration. The literature confirms that measures of productivity, like goals scored, games played (i.e. league or international appearances) and assists, have positive linear links to remuneration (Bryson et al. 2012; Frick 2007). Forwards and midfielders also earn relatively more than defenders (Bryson et al. 2012; Frick 2007). In some cases, forwards can earn as much as 130 percent more than goalkeepers (Frick 2007). Likewise, age, experience and contract duration have also all been shown to have positive links to measures of economic value (Bryson et al. 2012; Frick 2007). The impacts of contract length, in particular, merit further study, given how difficult it has been to examine historically (Frick 2007). To test some of these findings, researchers have also looked at non-league, semi-professional football players, and the results support the findings on professional players (Dobson et al. 2000).

To delve deeper, some studies have also looked at even more specific player characteristics. For example, one study examined players' two-footedness by analyzing data on 1,991 players from the top five European football leagues during the 2005/2006 season. Controlling for demographics, player position, player output and a club's ability to pay, the researchers found that two-footed players command a significant remuneration premium (Bryson et al. 2012). Intuitively, the findings make sense since two-footedness can allow for better ball control, passing and shooting accuracy, tackling, movement with the ball, use of space on the field, and positional flexibility (Bryson et al. 2012). Nonetheless, few studies have had success predicting market value or a transfer fee based on player characteristics alone. Many studies have also evaluated buyer and seller club characteristics. One study showed that player market value has strong links to variables related to the sizes of the buyer and seller clubs, such as ground capacity and average league attendance in the prior season (Dobson et al. 1999). When buyer and seller club characteristics influence transfer or market value, then researchers suggest that there can be evidence of monopoly rents (Dobson et al. 1999). These rents occur when a club sells a player for a price between the selling club's minimum reserved price and the buying club's maximum price (Dobson et al. 1999). In studying monopoly rents, researchers suggest it is informative to study different segments of players. For example, monopoly rents are more likely when clubs are selling top players and the statistical significance of player characteristics predicting market value can vary depending on the segment (i.e. top or lower level players) (Dobson et al. 1999).

The study of monopoly rents extends to monopsony rents as well (Garcia-del-Barrio and Pujol 2006). Looking at players in the Spanish first division during the 2001/2002 season, a group of researchers estimated market values based on player performances and economic

contributions to their clubs. Economic contribution is measured by counting the number of unique internet links after a Google search. The study finds that economic factors are important determinants of a player's market value. Corroborating other studies, it also finds that while selling clubs can extract monopsony rents from middling players, the aggregate rents are nil because any excess is given to the top players and these top players command premiums in market value (Garcia-del-Barrio et al. 2006). Given how few studies link economic factors to player value, the last study is particularly informative for this paper.

Besides player or club characteristics, a newer form of valuing players has emerged consisting of crowd-sourcing market values. Several websites, notably Transfermarkt, allow fans and users to input their belief of a player's market value. Studies show that crowd-sourced market values can be highly predictive of player transfer values, highlighting the potential wisdom of crowds (Herm, Callsen-Bracker and Kreis 2014; Müller, Simons and Weinmann 2017). In testing crowd-sourced values, one study also evaluated whether agents influence market value. The findings were inconclusive, but the topic merits further investigation (Herm et al. 2014). Expanding on the wisdom of crowds, another study examined player performance and player-popularity metrics over six seasons leading up to the 2014/2015 season across Europe's top five leagues. For transfers below €18 million, representing around 90 percent of all transfers, their multi-level regression model proved more accurate, while above the threshold the crowd-sourced values proved more accurate (Müller et al. 2017). The results suggest that both data analytics and crowd-sourced models can be useful for estimating different players' market values.

The second group of studies deals with brand value in football. From a commercial standpoint, teams generate revenue from fans, the media, communities which support a club and

sponsors (Bauer, Sauer and Schmitt 2005). One study looks only at fans' attendance at games as a measure of economic success and uses Keller's modified customer-based brand equity model to determine a brand equity value for each club in the German Bundesliga (Bauer et al. 2005). Its results show that clubs with strong brand equity benefit from economic success, as measured through fan attendance (Bauer et al. 2015). Another study takes a more varied approach by looking at how brand equity is a factor of a club's recruitment spend, winning percentage and publicity in the media. Brand equity is then also measured according to its impact on fan attendance at clubs (Wetzel et al. 2018). In terms of the three brand factors, the findings show that an additional €1 million spent on recruitment translated to an attendance boost of 43,330 fans during a season; a 10 percentage point increase in the winning percentage translated to 10,791 more fans; and increasing media mentions by 100 mentions brought 38,700 more fans during the season. Controlling for other factors, older clubs also have greater economic success than younger ones, which provides evidence for why some of the same clubs have dominated throughout the history of modern European football (Wetzel et al. 2018). Finally, by using recruitment spend as a factor, the study ties a club's brand equity to the transfer fees paid for players, which is similar to what this paper seeks to do.

Finally, the wider literature on branding should also be considered. Among the many tactics a company can employ as part of branding efforts is the use of celebrity endorsers. One of the most recent papers on celebrity endorsement uses an event study to evaluate whether celebrity endorsement announcements are linked to abnormal stock market returns. Looking broadly, the study does not find statistically significant results, but does show positive results for the endorsement of technology industry products (Ding et al. 2010). In their paper, the researchers discuss a lot of the earlier literature, and their framework and findings are largely in-

line with the literature described (Kamins et al. 1989; Choi and Rifon 2007). When measuring the effects of an individual's brand, like a celebrity's, the effects need not always be positive. To that end, another event study looks at the impacts on firms when their celebrity endorser is disgraced, finding negative and statistically significant abnormal returns (Bartz et al. 2013). Tellingly, when a firm terminates its association with a disgraced celebrity, there is no evidence of abnormal returns (Bartz et al. 2013). The research on celebrity endorsement is informative for how professional athletes can impact firm value for football teams. In many respects, football players are celebrities themselves. And so, it should be examined whether and how a player can provide an uplift to a team's brand after getting transferred just as companies use celebrities to boost their brands.

Although research in football has been growing, some of the studies that show important findings include some variables that are now outdated and have significant limitations. For example, Dobson et al.'s (2000) study on non-league players measures a buying club's influence on player value according to only stadium attendance and the proportion of seats versus standing room in a stadium. Likewise, Bauer et al.'s (2005) study on customer-based brand equity defines economic success only through fan stadium attendance. Today, stadiums only have seats, and fan attendance alone fails to incorporate fans who support clubs remotely through watching games on television or the companies who buy sponsorship packages (Dobson et al. 2000). The studies on brand equity are also only team-based and do not account for the potential brand-equity values of individual players. Today, the popularity of some players can far exceed their team's. Nonetheless, the studies provide a strong platform for continued research. This paper will look to build on the measures that past studies have used to estimate player market value while also updating them for the current environment of European football.

SECTION THREE: DATA COLLECTION AND ANALYSIS

3.1 Data Overview

The data for this study comes from three publicly available online sources, which will therefore easily allow replication by other scholars. The first is *Transfermarkt*, a German-based football news website (<https://www.transfermarkt.com/>). On the website, there are data tables covering several important variables for this study, including player transfers, club transfer spending, club league performance and club season attendance figures. The second is the *Kaggle European Soccer Database* available on *Kaggle Data Science* (<https://www.kaggle.com/hugomathien/soccer>). The database provides the EA Sports FIFA video game's player ratings, such as a player's overall rating or potential, which are used to evaluate and control for player ability. The makers of FIFA, the popular football video game, use over five million data points gathered by a network of around 9,000 members, including coaches and professional scouts, who watch players and evaluate their on-field attributes (Saed 2016). The third is the Open Platform API of *The Guardian*, a British daily newspaper. Since the Guardian has prominent football coverage, the API provides access to the number of media mentions each player and club in the dataset receives during given timeframes. Combined, these three sources provide data on a variety of football player and team characteristics as well as on media practices. Thus, by having player transfer values as the outcome variable to be predicted, firm spending and player performance as controls, and media mentions as a variable of interest, this study can model the primary research hypotheses.

Table 1 provides a brief description of the role and purpose of the relevant variables collected and used in this study.

Table 1. Description of the data collected and the dataset's variables

Variable	Source	Description
Player Height	<i>Kaggle European Soccer Database</i>	A player's height in centimetres
Player Weight	<i>Kaggle European Soccer Database</i>	A player's weight in pounds
Overall Rating	<i>Kaggle European Soccer Database</i>	The rating of a player's overall ability as assigned by the EA Sports game FIFA
Potential	<i>Kaggle European Soccer Database</i>	The rating of a player's potential as assigned by the EA sports game FIFA
Market Value	<i>Transfermarkt</i>	The source website hosts a forum where fans can input their estimates of a player's worth on the transfer market
Age	<i>Kaggle European Soccer Database</i>	A player's age
Points	<i>Transfermarkt</i>	The cumulative points obtained by a buyer club in the four seasons leading up to the transfer window in which the club purchases a player
Avg. Attendance	<i>Transfermarkt</i>	The average attendance of buyer club's fans at its stadium during the season immediately before the transfer window in which the club purchases a player
Expenditures	<i>Transfermarkt</i>	The cumulative spending on transfers to purchase players by a buyer club in the four seasons leading up to the transfer window in which the club purchases a player

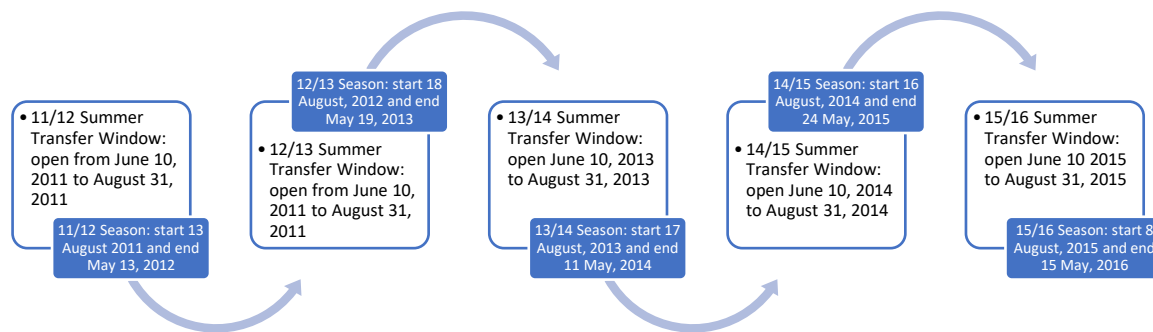
Player Media	<i>The Guardian API</i>	A count of the number of times a player is mentioned in a Guardian article in the six months before the transfer window in which the player is bought
Club Media	<i>The Guardian API</i>	A count of the number of times a buyer club is mentioned in a Guardian article in the six months before the transfer window in which the club buys a player
Fee	<i>Transfermarkt</i>	The transfer fee a club pays to buy the right to a player
Team ID	None	A unique number given to each buyer club
Season ID	None	A number from 1 to 5 for each of the five seasons in the dataset
English Seller	<i>Transfermarkt</i>	Denotes whether the club selling a player bought by a Premier league club in the transfer windows in the dataset is an English club or a foreign club (Yes: for an English Seller / No: otherwise)

3.2 Data Collection

The data collection process consists of a combination of web scraping and manual coding. The website *Transfermarkt* has detailed data on the major European football leagues. In order to access, clean and transfer the online data tables to usable datasets, this study uses the R coding language to scrape the contents of the data tables on the relevant webpages for this study. This study obtains data tables from the different webpages and builds datasets for each transfer window leading up to each of the five seasons from 2011/2012 to 2015/2016. For example, the relevant transfer window for the 2011/2012 seasons represents the time period between June

2011 and August 2011 when Premier League football clubs bought players leading up to the 2011/2012 Premier League season. On *Transfermarkt*, the relevant webpages for this study are the pages for the English Premier League and the English Championship covering the cumulative number of points earned by clubs, attendance figures, transfer income and expenditures, and yearly player transfers. Accordingly, from the player transfers page, this study downloads the data covering all the players bought or sold, excluding loans, by Premier League clubs during the summer transfer window leading up to each season studied. Figure 1 illustrates the sequence of events for each premier league season.

Figure 1. Timeline of English Premier League events



The data collected consists of the list of player transfers featuring a Premier League club as either a buyer or seller of a player during each of the summer transfer windows listed in the table above. After downloading the table of transfers in R, this paper then applies several filters. First, this study only selects transfers featuring a Premier League club as the buyer since this paper evaluates the factors leading clubs to buy players, not sell players. Of those transfers, this study only selects transfers of players with market values above one million euros at the time of transfer and who were bought for fees above one million euros. After applying these filters, the

final dataset of transfers consists of 45 percent of all the players bought or signed by English Premier League clubs over the five selected seasons. Of this list of players transferred to English clubs, this paper notes that a further seven transfers are excluded from this study because the players bought are not represented in the FIFA football database obtained through Kaggle. A histogram of the relative proportion of transfers for players bought above the one million euros threshold can be found in Appendix I. An illustration of the data table obtained from *Transfermarkt* for the list of transfers in one of the summer transfer windows studied can be seen in Table 2.

Table 2. Illustration of the data collected from Transfermarkt on player transfers

Player Name	Age	Position	Selling Club	Fee	Market Value	Buying Club	English Seller
Santi Cazorla	27	CM	Malaga FC	19.00m	20.00m	Arsenal FC	N
Lukas Podolski	27	SS	FC Koln	15.00m	20.00m	Arsenal FC	N

In the final dataset, the seller clubs are coded Y if they are English clubs and N otherwise in order to control for the effects of buying a player from a foreign league. For example, given the international coverage of the Premier League, the league's players may benefit from more media coverage and popularity, and thus higher brand value, than players from foreign leagues would. In relation to transfer fees, some English clubs may also pay more for players from certain leagues. For example, since English clubs are all relatively wealthy, they are not under the same obligations to sell their players, which means they could ask other Premier League buyers to pay premiums to buy their players (Chibber 2018).

In addition to the list of summer transfers, this study also collects three other datasets from *Transfermarkt* for each of the five seasons studied: each buyer club's cumulative points, cumulative spending and average attendance. For each season dataset, the cumulative points and spending totals are taken from the four-year period timeframe leading up to each summer transfer window. Meanwhile, for attendance, this study obtains the average attendance of each buyer club during each season prior to the seasons studied.

Moreover, this study constructs data on individual player performance by accessing the FIFA player ratings on the *European Soccer Database*. Since the Kaggle file is a database.sqlite file, this paper relies on SQLiteStudio to access, clean and manipulate the data as well as to transfer the dataset into R for analysis. Specifically, this study incorporates the Player, Team and Player Attributes data tables into the main datasets. In the Player Attributes dataset, a player, who is attributed a unique player API number, could have several entries of FIFA performance metrics at different dates over the course of a year. When constructing the dataset for each season, this study filters the Player Attributes data to only include the first entry of metrics for the year leading up to the season studied. For example, for the dataset on the 2011/2012 season, this study filters for the first entry of performance metrics in 2011 for each player studied in the dataset. In doing so, this paper matches the player's on-field ability to the time of his transfer.

Finally, the last step of the data collection process consists of manually coding the number of media mentions for each player studied. Using the Guardian Open Platform API, this study searches the Guardian's content for a player's name, and then filters for the "football" section and the six-month period leading up to the transfer window the player is bought in. For example, for the 2011/2012 dataset, this study searches for a player's name, such as "Alex Oxlade-Chamberlain", restricts the search to the "football" section and sets the from-date to

2011-01-01 and the to-date to 2011-06-01. After the search, this study inputs the number for “total”, representing the total number of articles the player searched has appeared in during the given timeframe, in the entry for that player in the dataset. This study repeats the process for each player and for each of the Premier League clubs during the seasons covered in the final dataset.

Table 3 presents an illustrative example of a row entry in the final dataset. The final dataset contains all of the filtered players who were bought by English Premier League clubs between the 2011/2012 and 2015/2016 seasons. It contains 378 observations and 15 variables.

Table 3. Sample row entry from the final dataset

Player Height	Player Weight	Team ID	Overall Rating	Potential	Fee	Market Value	Age
175.26	154	9825	69	82	13800	2500	17

Season ID	Team Points	Average Attendance	Team Expenditures	Player Media	Club Media	English Seller
1	433	60.023	121100	24	1168	Y

3.3 Exploratory Data Analysis

As mentioned above, the final table’s dimensions are 378 by 15 for the 378 players and 15 variables studied. Of these 15 variables, 14 of them are predictor variables while the “Fee” variable is the main outcome variable. Table 4 provides an illustration and summary of the variables and the possible values for the observations studied.

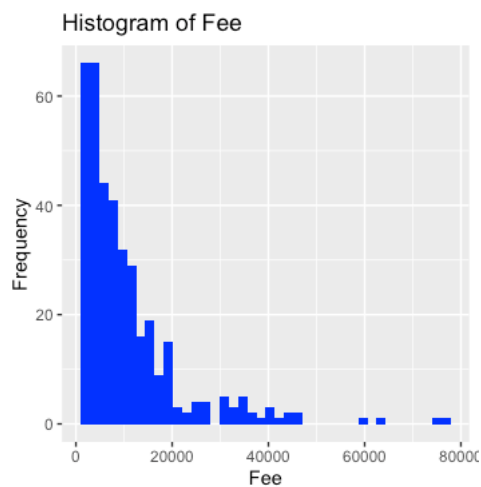
Table 4. Summary of final dataset's variables

Numeric Variables	N	Mean	Min	25 th	Median	75 th	Max
Player Height	378	182.2	165.1	177.8	182.9	188.0	200.7
Player Weight	378	168.9	128.0	157.0	170.0	181.0	207.0
Overall Rating	378	75.5	0.0	73.0	76.0	79.0	87.0
Potential	378	79.8	0.0	77.0	80.0	83.0	91.0
Market Value	378	8544.0	1000.0	3500.0	6000.0	10375.0	50000.0
Age	378	24.7	17.0	23.0	25.0	27.0	35.0
Points	378	214.2	0.0	93.0	213.0	332.0	433.0
Avg. Attendance	378	35444.0	10265.0	25086.0	34871.0	43171.0	75530.0
Expenditures	378	184580.0	6820.0	59040.0	113390.0	277910.0	651760.0
Player Media	378	20.0	0.0	1.0	7.0	25.0	260.0
Club Media	378	632.1	114.0	350.0	498.0	995.0	1429.0
Fee	378	10615.0	1000.0	3750.0	7350.0	12838.0	76000.0
Categorical Variables		Counts					

Team ID (Factor with 30 levels)	-	8455: 25	8650: 25	8456: 24	8472: 22	10252: 22	Other: 236
Season ID (Factor with 5 levels)	-	1: 65	2: 68	3: 73	4: 82	5: 90	
English Seller (Factor with 2 levels)	-	Y: 162	N: 216				

Figure 2 provides a histogram of the outcome variable “Fee”. The frequency of transfers is on the y-axis and is a function of the given transfer fee amounts on the x-axis.

Figure 2. The distribution of transfer fees in the final dataset

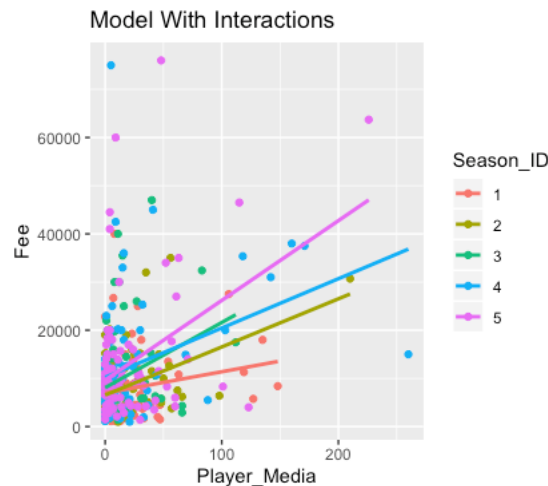


The histogram is right-skewed, which suggests that a logarithmic transformation of the “Fee” variable could lead to a more appropriate set of regression and machine learning analyses. Most

of the data is concentrated below the level of 20 million euros and there are a few outliers at higher values. The minimum equates to one million euros while the max stands at 76 million euros. The median is approximately 7.4 million euros while the mean is an estimated 10.6 million euros as it is pulled above the median by the outliers.

Since the hypotheses H1 and H2 investigate how a player's media popularity affects transfers fees, this paper examines fee levels relative to the predictor variable "Player Media", which is the proxy for player popularity and brand value. This study performs the analysis overall for H1 and across time for H2. Figure 3 plots the relationships between transfer fees and player media mentions across the five seasons studied.

Figure 3. The regression of "Fee" as a function of "Player Media" across five seasons

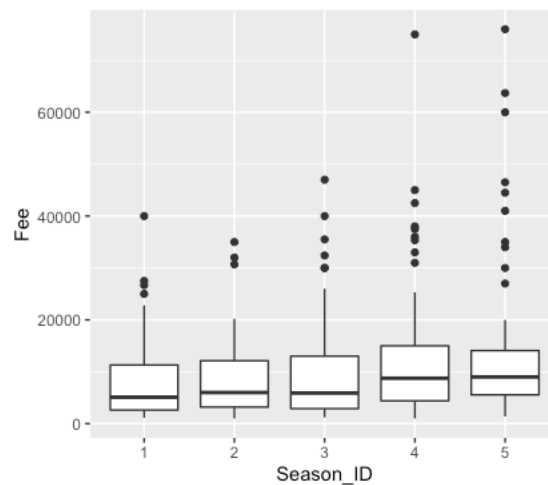


In the plot, the linear relationships between player media mentions and transfer fees across the five seasons studied are positive. These positive trendlines suggest that the more a player is mentioned in the media, and thus the greater his brand value, then the higher transfer fee he can command from a prospective buyer club. The five trendlines across seasons also show that the lines become steeper over time. For each season, its respective line becomes steeper as does its rate of change compared to the line and rate of the prior season moving chronologically from

season one to season five, which suggests that the effect of player coverage on transfer fees increases over time. While this plot can be informative for answering the two hypotheses, this paper still rigorously tests the relationship between player media exposure and transfer fees through several statistical models that are outlined below.

Given this study includes a time factor, it also constructs boxplots to show how transfer fees vary directly across seasons without any other interaction terms. Figure 4 below provides a summary of the differences in fees across seasons.

Figure 4. Boxplot summary of transfer fees across five seasons

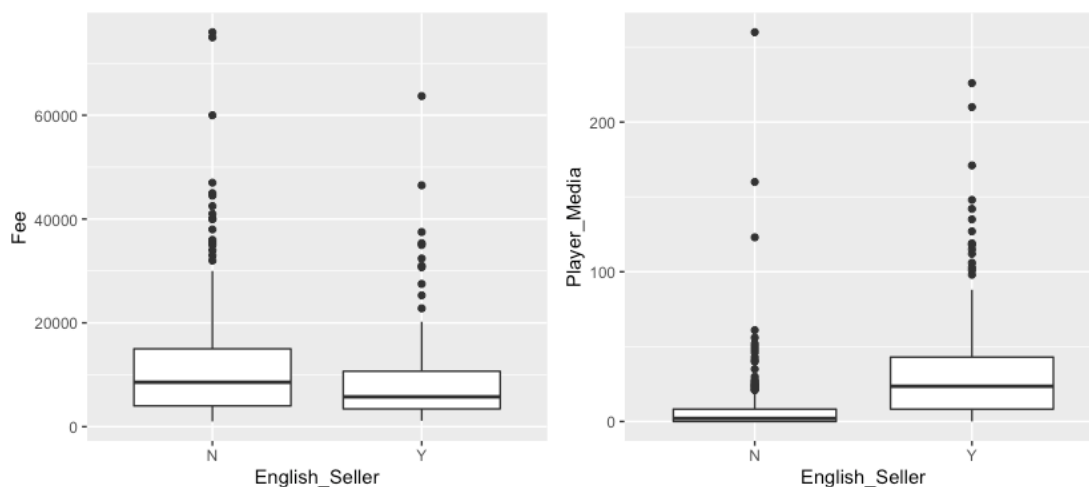


As the seasons progress from 2011/2012 to 2015/2016, or ID 1 to 5, the boxes begin to slightly shift upwards and have higher numbers of large outliers in seasons 4 and 5. These boxes and the average fee levels suggest that transfer fees have grown larger over time, which corroborates media reports on club transfer spending (Ahmed 2019). For the outliers, this paper observes an increase in quantity and in value over time, which would fit with the narrative that clubs are regularly breaking new transfer records each new season (Ahmed 2019). As discussed previously, over the last decade, clubs have continued to grow richer from television rights

contracts and lucrative sponsorships, and this graphic suggests that the trickle-down effect of increased wealth manifests itself in higher amounts of transfer spending.

Another categorical variable to visualize is the predictor “English Seller”, which represents whether or not a player was bought from an English team. This study visualizes the relationship between the categorical variable and the main two variables of interest, namely the outcome “Fee” and the predictor “Player Media”. Figure 5 shows boxplots with both variables modeled as functions of the categorical predictor “English Seller”.

Figure 5. Boxplots of “Fee” and “Player Media” as functions of whether or not the seller in a transfer is an English club

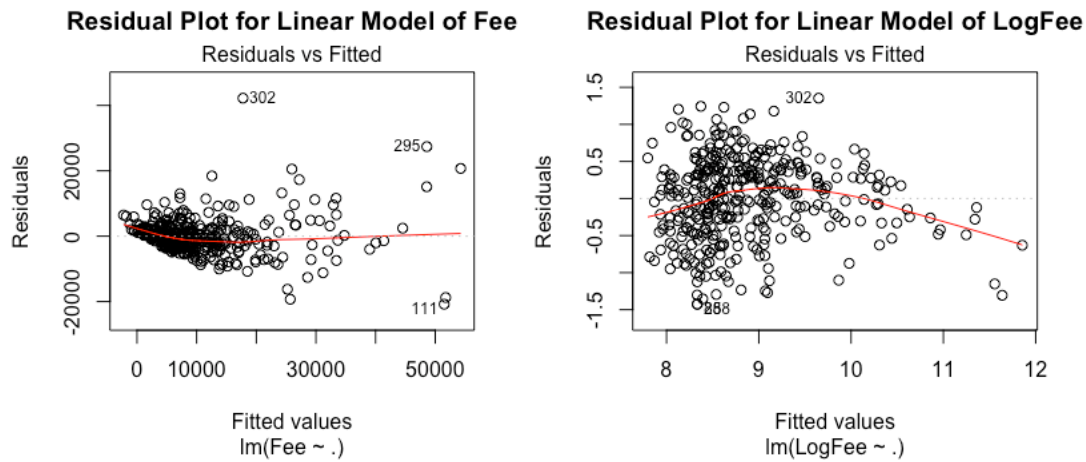


From the first boxplot, this study observes that there seems to be a greater range of transfer fee values paid for players arriving from foreign clubs than from English clubs, but the difference is not large enough to make definitive conclusions. In the second boxplot, this study observes that players already plying their trade in the English Premier League garner more media coverage than foreign-based players before their transfers to Premier League clubs. The relationship points to a potential limitation in the “Player Media” predictor variable, which could affect its ability to approximate for brand value and predict transfer fees. This paper explores this relationship further in the Limitations section 4.5 below.

To examine whether to transform any variables, this paper constructs a correlation heatmap seen in Appendix II. It shows the results from an analysis of pairwise correlations between the numeric variables in the dataset. This set of numeric variables excludes the categorical variables of “Player ID”, “Team ID”, “Season ID” and “English Seller”. The heatmap suggests that there are relatively strong correlations between the outcome fee variables and the variables for average attendance, club media mentions, player market value, overall rating, cumulative club points, and player potential. For the main predictor of interest, namely “Player Media”, its correlation with most of the other variables in the dataset is consistent, but ultimately relatively weak. The collinearity of the variables can affect the results of the statistical models fitted, which is why it is discussed further in the Results section.

Moreover, this study fits regressions and examines the residuals to test the standard regression assumptions and to inform the choices for fitting the regression models detailed in the Research Methodology section. To that end, this paper uses residual plots to evaluate issues such as the non-linearity of the response-predictor relationships and non-constant variance of error terms (James et al. 2017, 91-92). A residual is denoted by the equation: $e_i = y_i - \hat{y}_i$, and given the presence of multiple predictor variables, this paper plots the residuals versus the predicted values \hat{y}_i (James et al. 2017, 93). Residual plots are shown in Figure 6 with the leftmost plot representing a regression of “Fee” and rightmost illustrating the regression of “LogFee”.

Figure 6. Residual plots for “Fee” and its transformation “LogFee”



The leftmost plot representing the linear regression of the outcome variable on its original scale shows a pattern with most of the residuals clustered together to the left of the plot, which suggests some non-linearity. This study also evaluates the assumption of constant variance and notices that in the leftmost plot the residuals seem to fan out to the right. Given that the magnitudes of the residuals increase as the fitted values grow larger, the regression on the original scale shows signs of heteroscedasticity (James et al. 2017, 95).

To deal with these potential issues, this paper uses the log function to transform the outcome variable, creating the outcome variable “LogFee” (James et al. 2017, 95). The rightmost plot shows the residuals for that transformation. Although there is perhaps still evidence of a discernible pattern, there is equal variance of the residuals across the x-axis, as the spread of the residuals looks contained within the limits of -1.5 and 1.5. Thus, this paper observes that the residual plot for “LogFee” is more homoscedastic than the plot for “Fee”. As a result, this study uses the transformation “LogFee” when fitting its principal statistical models.

SECTION FOUR: RESEARCH METHODOLOGY

4.1 Overview of Statistical Modeling Frameworks

This study relies on several advanced statistical models and machine learning techniques applied through the R coding language. Following the EDA stage, this study transforms the outcome variable “Fee” to the logarithmic scale, replacing the outcome variable “Fee” with a new one called “LogFee”. The rest of the 14 predictor variables outlined in the data collection process stay the same. This study then adds two interaction terms to measure the time effect of media mentions on transfer spending, which consist of multiplying the “Season ID” variable with the “Player Media” and “Club Media” variables. Using this dataset, this study fits statistical models using multiple regression, LASSO regression and Decision Trees frameworks to compare and contrast their performance, as well as to assess the robustness of their findings.

In fitting these models, this study performs out-of-sample validation to test the predicted values. This study splits the 378 observations to create two sets: a training set and a validation set. The training set contains 70 percent of the observations and the validation set consists of the remaining 30 percent of observations. In this cross-validation process, the model is fit on the training set, and then the resulting model is applied to the validation set to obtain predicted values for the remaining observations (James et al. 2017, 176). From the validation set, this study obtains an estimate of the test error rate through the Mean Squared Error, or MSE (James et al. 2017, 176). Finally, after repeating this process for each of the statistical models, this study compares the performance of the models to find the one with the lowest MSE. The cross-validation process helps test the validity of the statistical model chosen and prevent the possibility of overfitting, which can occur when performing in-sample validation or statistical analysis. The model with the lowest out-of-sample MSE is then selected as the best fit model and

is examined further in the Results section. Results from all of the analyses and the accompanying R code are available upon request.

4.2 Multiple Regression

The multiple regression approach is informative for evaluating the linear relationship between an outcome variable and several possible predictor variables. This study's dataset includes 14 possible predictor variables, which can affect the “LogFee” outcome variable. To test the impact of the different predictors, this study fits several multiple regression models with different specifications.

The first model fitted is a regression that isolates the predictor variable “Player Media”, which represents the number of mentions a player receives in the media prior to getting bought. The linear regression can be described as follows:

$$(1) \text{LogFee}_i = \beta_0 + \beta_1 \cdot \text{PlayerMedia}_i + \epsilon_i$$

The index i denotes the different in-sample observations. Following that regression, this study then runs a similar regression by adding “Club Media” as a second predictor variable:

$$(2) \text{LogFee}_i = \beta_0 + \beta_1 \cdot \text{PlayerMedia}_i + \beta_2 \cdot \text{ClubMedia}_i + \epsilon_i$$

These two regressions seek to isolate the potential effects of media coverage on the transfer fees Premier League clubs pay for players. However, this paper then seeks to examine how these predictor variables interact with the rest of the potential predictor variables. Thus, this paper fits a third model with the 14 possible predictor variables for the outcome variable of “Fee”. This multiple regression includes three categorical variables, namely “Team ID”, “Season ID” and “English Seller”, and their fixed effects are denoted by α_i , δ_i and θ_i respectively. The third multiple regression equation is as follows:

$$\begin{aligned}
(3) \text{ LogFee}_i = & \beta_0 + \beta_1 \cdot \text{PlayerMedia}_i + \beta_2 \cdot \text{ClubMedia}_i + \beta_3 \cdot \text{PlayerHeight}_i + \beta_4 \cdot \\
& \text{PlayerWeight}_i + \sum_i \alpha_i \text{ Team ID}_i + \beta_5 \cdot \text{OverallRating}_i + \beta_6 \cdot \text{Potential}_i + \beta_7 \cdot \\
& \text{Age}_i + \beta_8 \cdot \text{MarketValue}_i + \sum_i \delta_i \text{ SeasonID}_i + \beta_9 \cdot \text{Points}_i + \beta_{10} \cdot \\
& \text{AvgAttendance}_i + \beta_{11} \cdot \text{Expenditures}_i + \sum_i \theta_i \text{ EnglishSeller}_i + \epsilon_i
\end{aligned}$$

These three models are informative for evaluating how player media mentions affect transfer fees, addressing the first hypothesis. To test the second hypothesis, this paper also fits the following three models, which introduce the predictor “Season ID” as an interaction term to the first three regression models. These regression models can be described as follows:

$$\begin{aligned}
(4) \text{ LogFee}_i = & \beta_0 + \beta_1 \cdot \text{PlayerMedia}_i + \sum_i \delta_i \text{ SeasonID}_i + \beta_2 \cdot \text{PlayerMedia}_i \cdot \\
& \sum_i \delta_i \text{ SeasonID}_i + \epsilon_i \\
(5) \text{ LogFee}_i = & \beta_0 + \beta_1 \cdot \text{PlayerMedia}_i \cdot \sum_i \delta_i \text{ SeasonID}_i + \beta_2 \cdot \text{ClubMedia}_i \cdot \\
& \sum_i \delta_i \text{ SeasonID}_i + \beta_3 \cdot \text{PlayerMedia}_i + \beta_4 \cdot \text{ClubMedia}_i + \sum_i \delta_i \text{ SeasonID}_i + \epsilon_i \\
(6) \text{ LogFee}_i = & \beta_0 + \beta_1 \cdot \text{PlayerMedia}_i + \beta_2 \cdot \text{ClubMedia}_i + \beta_3 \cdot \text{PlayerHeight}_i + \beta_4 \cdot \\
& \text{PlayerWeight}_i + \beta_5 \cdot \sum_i \alpha_i \text{ Team ID}_i + \beta_6 \cdot \text{OverallRating}_i + \beta_7 \cdot \text{Potential}_i + \beta_8 \cdot \\
& \text{Age}_i + \beta_9 \cdot \text{MarketValue}_i + \beta_{10} \cdot \sum_i \delta_i \text{ SeasonID}_i + \beta_{11} \cdot \text{Points}_i + \beta_{12} \cdot \\
& \text{AvgAttendance}_i + \beta_{13} \cdot \text{Expenditures}_i + \beta_{14} \cdot \sum_i \theta_i \text{ EnglishSeller}_i + \\
& \beta_{15} \cdot \text{PlayerMedia}_i \cdot \sum_i \delta_i \text{ SeasonID}_i + \beta_{16} \cdot \text{ClubMedia}_i \cdot \sum_i \delta_i \text{ SeasonID}_i + \epsilon_i
\end{aligned}$$

These models show how the relationship between media mentions and transfer spending evolves over time to evaluate whether the variables increase together from one season to the next.

According to the properties of an Ordinary Least Squares regression, the coefficient values for these multiple regression models are the result of minimizing the sum of square residuals, or RSS. Accordingly, for multiple regression models, the Mean Squared Error for a model with p as the number of predictor variables is as follows:

$$(7) \text{MSE} = \frac{RSS}{n-p-1}$$

The MSE provides the basis for comparing the results of the six regression models discussed above. A summary of the results for the regression models (1) through (6) is shown in the Results section.

Moreover, this study utilizes methods for evaluating a model's accuracy in order to find the model with the lowest test error (James et al. 2017, 210). To that end, this paper relies on two penalization methods to account for the effects of overfitting: Mallow's C_p statistic and the Bayesian Information Criterion, or BIC. Both the C_p and the BIC provide estimates of the test error and are applied to the regression (6) model with an adjustment made to only include numeric values. These penalization methods deal with overfitting by making an adjustment to the training error when estimating the test error (James et al. 2017, 210).

First, this study implements Mallow's C_p method. The formula for Mallow's C_p is as follows:

$$(8) C_p = \frac{1}{n} (RSS + 2d \cdot \hat{\sigma}^2)$$

In the formula, the variable d represents the number of predictor variables. In applying Mallow's test statistic to the regression model of "LogFee", the purpose is to minimize the C_p in order to find the model that reduces the amount of prediction errors. As seen by its formula, the C_p statistic adds the penalty term $2d \cdot \hat{\sigma}^2$ to the RSS. As the number of predictor variables increases, the penalty term increases as well to counterbalance the reduction in the training RSS. In modifying the RSS, the penalty accounts for how the training error can underestimate the test error (James et al. 2017, 211).

Another criterion to examine model fit is the Bayesian Information Criterion (BIC). The formula for the BIC statistic is as follows:

$$(9) \text{ BIC} = \frac{1}{n} (RSS + d \cdot \ln(n) \cdot \hat{\sigma}^2)$$

In the formula, the new variable n represents the number of observations. Similar to Mallows's C_p , the BIC statistic adds a penalty term as well: $d \cdot \ln(n) \cdot \hat{\sigma}^2$. Given C_p 's penalty has a multiplier of $2d$, the BIC's penalty for adding predictors is even larger than the C_p 's since whenever n is greater than seven the multiplier $d \cdot \ln(n)$ is greater than $2d$ (James et al. 2017, 211). This larger penalty term leads the BIC approach to fit smaller models with fewer predictor variables than the C_p approach would (James et al. 2017, 211).

To minimize C_p and BIC, this study uses the `leaps()` package and the `regsubsets()` algorithm in R. The implementation of both statistical methods for the OLS linear model (6) proves instructive for variable selection (James et al. 2017, 78). Accordingly, this paper narrows the selection of variables to construct new best fit linear models. The end result is two new linear models fit with variables selected through the C_p and BIC criteria tests. Modelling these regressions generates two additional out-of-sample MSE values, which this paper uses to evaluate and compare model fit among the regression models. An analysis of the findings for the two penalized regressions can be seen in the Results section.

4.3 LASSO Regression

While the two penalization methods described above are informative for fitting linear regression models, they consist of subset selection, which leads to fitting models that include only a subset of predictor variables (James et al. 2017, 214). However, this paper also wants to generate and examine a model including all of the possible predictor variables while still adjusting for and limiting overfitting. Thus, this study applies Least Absolute Shrinkage and Selection Operator (LASSO) regression methodology. The LASSO is a shrinkage method that

shrinks regression coefficient estimates towards zero, which in turn reduces their variance (James et al. 2017, 215). The formula for the LASSO equation is as follows:

$$(10) \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j|$$

In the formula, the left side is equivalent to the RSS problem while the second part represents the LASSO penalty term. In the penalty term, the lambda, or λ , is called the tuning parameter, and the choice of its value determines the best model fit. The optimal tuning parameter value can be found using cross-validation, as this paper selects the value corresponding to the lowest cross-validation error (ISLR 219). As the tuning parameter becomes larger, then the penalty term shrinks some of the predictor coefficients to be exactly zero. Similar to the C_p and BIC approaches, the LASSO effectively results in a smaller model which selects only a portion of the total set of predictor variables (James et al. 2017, 219). However, in some ways, the LASSO is more instructive than both penalization methods since the tuning parameter provides more flexibility. By adjusting the value of the tuning parameter, it is possible to select subsets of predictor variables of varying sizes when fitting a model, and the LASSO's flexibility still allows for generating a model including all possible predictor variables (James et al. 2017, 220).

This study uses the glmnet package in R, with the functions glmnet() and cv.glmnet(), to obtain LASSO solutions. As highlighted above, the tuning parameter value is largely responsible for driving LASSO solutions, and so this study uses K-Fold Cross-Validation to obtain an optimal value for λ (James et al. 2017, 227). The cross-validation process consists of randomly dividing the 378 observations into K sections, also known as K folds (James et al. 2017, 181). Next, this paper trains the LASSO model on K – 1 of the folds and then fits predicted values on the hold out fold in order to calculate an MSE value corresponding to that fold (James et al. 2017, 181). The process is repeated for each of the K folds, with a different test fold each time.

After performing the process for K folds, this paper obtains K MSE values, which are averaged out to estimate the testing error. For this study, the K parameter is set to $K=10$ and so cross-validation is performed using 10 folds. The selected value of λ , the testing error and the results from the LASSO equation can be found in the Results section.

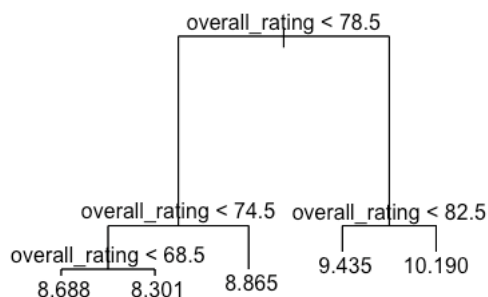
After obtaining the λ value and inputting it into the LASSO equation, this paper obtains a select number of LASSO variables with non-zero coefficients. This paper then uses these variables to fit a linear model. The linear model also outputs coefficient estimates, which can be compared to the LASSO estimates. In keeping with the general cross-validation process outlined in section 4.1, this paper runs the linear model on separate training and testing sets, and then finds the MSE. The results are outlined in the Results section.

4.4 Decision Trees

The classical regression and regularized regression approaches discussed above are established and useful. However, decisions trees provide another intuitive avenue for prediction that is not based on a linear model. Decision tree methods deal well with non-linear relationships and qualitative predictors, such as this study's "Season ID", "Team ID" and "English Seller" variables (James et al. 2017, 15; Brid 2018). This study explores decisions trees in several ways to build prediction models. Specifically, this paper applies single regression trees with and without pruning, trees with bagging and random forest. For these four methods, this study repeats them for three distinct iterations each according to three variable specifications. The first specification, tree (1), only uses the "Player Media" predictor. The second, tree (2), uses "Player Media" and "Season ID". Lastly, the third specification, tree (3), includes the 14 original predictor variables. It is not possible to model the interaction terms directly for the decision trees.

Decision tree methods are supervised learning methods that build predictive models using a tree-like structure (James et al. 2017, 303; Brid 2018). These methods can vary between producing a single tree or combining larger numbers of trees to make predictions. The trade-off for adding more trees to make predictions largely centers around increased difficulty with interpretation versus added predictive accuracy (James et al. 2017, 303). The basic framework for constructing a tree consists of dividing the predictor space into boxes connected by branches in order to then select the boxes with the lowest RSS values (James et al. 2017, 306). The process of selecting these boxes is known as recursive binary splitting. At the start, all of the observations are contained within one box, which makes up the entire predictive space. A predictor and a cut-off point are then found to minimize RSS, which serves to split the box into two branches each with its own new box (James et al. 2017, 307). This process is iterative since after two boxes are produced one of them may get split again to further reduce RSS, meaning splitting can occur several times before a set criterion is met (James et al. 2017, 307). Possible criteria include selecting the minimum number of observations and minimum deviance needed in each node. The criteria can determine how many nodes a tree has. A regression tree fitted for the outcome variable “LogFee” and the predictor variable “Overall Rating” show in Figure 7 below provides an illustrative example:

Figure 7. Illustrative example of a regression tree



As seen in Figure 7, the regression tree is upside down. It can have several branches and boxes, otherwise known as nodes. The nodes with no branches attached are called terminal nodes, which can be likened to a tree's leaves (James et al. 2017, 304). The above tree has five terminal nodes, splits the variable "Overall Rating" at four separate points and has a residual mean deviance of 0.5193.

However, fitting a single tree can lead to overfitting. To get as low of a testing error as possible for a tree, this study uses pruning to obtain several subtrees and then uses K-Fold Cross-Validation, as discussed above, to obtain the best subtree that minimizes the testing error (James et al. 2017, 309). To implement pruning, this paper relies on the package "prune.tree" and the function "cv.tree" in R. Equipped with these techniques, this study fits three single regression trees denoted (1), (2) and (3) for the three different types of variable specifications discussed above with and without pruning. The process results in six trees in total. This study uses out-of-sample testing following the guidelines in section 4.1 to determine the MSE for each tree. Since the process is performed with and without pruning, the MSE values are compared to observe the effects of pruning. A summary of the final outputs can be found in the Results section.

In addition to fitting single trees, this study also uses methods to fit and combine multiples trees, such as bagging and random forest. While relying on a single tree can lead to high variance, the process of averaging out multiple uncorrelated trees can lower the variance and MSE of the resulting combined equation (James et al. 2017, 317). In order to perform regression with multiple trees this paper uses the "randomForest" package in R to perform both bagging and random forest. First, the principle of bagging is applied to reduce variance. The process entails using the process of bootstrapping to construct a number B of datasets by repeatedly sampling with replacement from the original dataset (James et al. 2017, 189). With

these distinct datasets, this study then creates B large and un-pruned single regression trees and then bags these trees so as to average their results (James et al. 2017, 317). The equation is described as follows:

$$(11) \hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x)$$

For this study, the tuning parameter for the number of trees is set to 100. As a part of the random forest package, bagging represents a specific method for generating estimates from non-random trees when the tuning parameter m is set to the total number of predictors, or $m = p$ (James et al. 2017, 329). Thus, this study performs bagging to fit another three decision tree models with bagging according to the predictor specifications outlined above. Depending on the fit, the $mtry$ is set to $mtry=1$, $mtry=2$ or $mtry=14$ depending on how many predictors are in the bagging model which need to be considered as candidates for each split that occurs on a node (James et al. 2017, 329). By averaging the results over 100 trees, the overall variance and testing error decrease. The results of performing bagging are reported in the Results section.

Finally, after performing bagging, this study performs random forest since random forests can often produce less variance and a smaller testing error than bagged models (James et al. 2017, 320-321). A random forest model requires making different assumptions for the tuning parameter m . Whereas for bagging $m = p$ predictors, this paper now follows the default setting for random forest regression trees by setting $m = \frac{p}{3}$ to generate random forest models (James et al. 2017, 329). Given bagging considers all variables for splits to minimize RSS, the trees can be quite similar, leading to predictions that can be highly correlated with each other (Brownlee 2019; James et al. 2017, 320). In setting a smaller value for the parameter m , the random forest model uses a smaller random sample of predictors for splitting (Brownlee 2019). Thus, the model generates predictions that are less highly correlated with each other than those produced

during bagging, which in turn should lead to a greater reduction in variance when averaging out the predictions (James et al. 2017, 320). Similar to the two other decision tree methods outlined, this study fits three random forests according to the predictor specifications defined above. For both bagging and random forest, training and test sets are used to make predictions and to obtain MSE values for each model. These values provide a basis for comparison. For all of the decision tree methods discussed above, their outputs are analyzed and discussed in the Results section.

4.5 Limitations

Although steps are taken to ensure as much accuracy as possible, there remain a few limitations to highlight. First, this study seeks to understand the relationship between a player's transfer fee and his brand value. Based on prior literature on the brand value of clubs, this paper settles on using a player's media attention, or coverage, as an approximation of brand value, but the data on media coverage is sparse and limited (Wetzel et al. 2018). Given the paucity of options available, this paper only uses one media source, *The Guardian*, to obtain data on media coverage. While the website's UK focus is helpful for covering the Premier League, it can lead to less favorable or extensive coverage of the players bought from other leagues outside of the UK. The boxplot in Figure 5 of the Exploratory Data Analysis section 3.3 highlights how *The Guardian* may favor coverage of players already based in the UK in the months leading up to their transfers. In evaluating the two side-by-side boxes, this paper finds that players bought from foreign clubs, represented by the "N" box, receive on average fewer media mentions than those purchased from English clubs. A possible explanation is that the Guardian covers Premier League games more in-depth than games from foreign leagues, and so players already on Premier League teams earn more media mentions from Premier League match reports than their foreign counterparts. Moreover, the search process for obtaining counts of player and club media

coverage can be difficult to validate. From the API search results, this study tabulates the total number of articles the search appears in, but it cannot access the articles to verify the contents for information pertaining to a player's transfer or his brand. To implementation of several filters to search through the Guardian's API helps mitigate this potential issue.

Second, there are so many possible factors that determine a transfer fee that this paper cannot rule out the possibility of confounding or omitted variables. One variable that this study does not consider is a team's positional need. For example, a team, even a highly performing one, may have a strong desire to buy players to fit a specific tactical set-up and overpay for the players that can meet the needs of their tactical plan. Another possible omitted variable of interest is the stature and wealth of the selling club. While this study does account for whether the seller is an English or foreign club, it does not assess the league standing, prestige and wealth of the seller, like it does for the buyer, in the dataset. These possible limitations could serve as areas for further research. Further study of transfer fees should include detailed comparisons of both seller and buyer clubs to understand the negotiation dynamics affecting transfer fees.

SECTION FIVE: RESULTS AND DISCUSSION

5.1 Comparing Regression Models

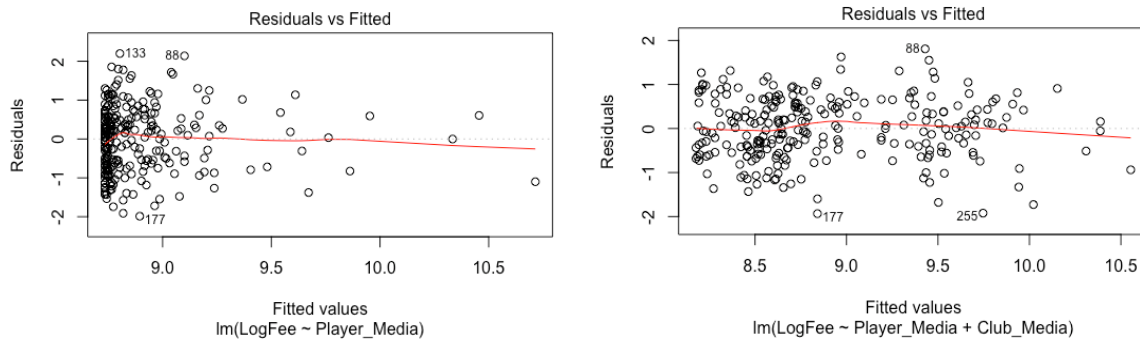
In the Methodology section, this paper discusses the implementation of several regression models. These models include the standard regressions (1) – (6), the two penalized regressions and the LASSO regression. In this section, this paper compares the out-of-sample validation results for each regression to evaluate and determine the best model fit. This study considers the model generating the lowest MSE from the test set, which represents 30 percent of the dataset's total number of observations, as the best fit regression model.

First, this study evaluates the classical regression models and summarizes the results. To address the hypothesis H1 on the impact of brand value on transfer fees, this paper initially models the media coverage predictor variables individually to isolate for their predictive capabilities. Table 5 provides summaries of the regressions (1) and (2) on the training dataset. Figure 8 plots their residuals side-by-side.

Table 5. Summary of regressions (1) and (2)

Regression 1: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	8.734958	0.059581	< 2e-16 ***
Player Media	0.007614	0.001454	3.37e-07 ***
MSE: 0.8137478		Adjusted R-squared: 0.09474	
Regression 2: Coefficients			
(Intercept)	8.0282665	0.0832500	< 2e-16 ***
Player Media	0.0040600	0.0012632	0.00147 **
Club Media	0.0012128	0.0001144	< 2e-16 ***
MSE: 0.5508651		Adjusted R-squared: 0.3625	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05			

Figure 8. Residuals of “LogFee” as functions of “Player Media” as well as “Player Media” and “Club Media”



In both of the regressions summarized above, the predictors for media coverage, “Player Media” and “Club Media”, are positive and statistically significant. Although the adjusted R-squared statistic is comparatively low in the first model, it corresponds to a correlation value of $r=0.3$. In the second regression, both the adjusted R-squared and correlation statistic increase, with $r=0.6$. These results suggest that player brand value, and brand value in football more generally, does impact the amount clubs pay for transfers. When viewed on the original scale, the effect size is also quite large, as one additional player media mention in model (1) translates to a 0.7 percent increase in a player’s transfer fee on the original scale. In terms of model fit, as seen in the plots of residuals, the fitted values seem relatively evenly spread and each model’s fit looks strong. Adding the “Club Media” predictor also improves model fit, as the MSE value improves from $MSE=0.81$ to $MSE=0.55$. Despite their results, these two models are not enough to confirm the first hypothesis H1. Rather, it is important to observe their significance compared to other variables as well and as part of a better fit model with more predictive accuracy.

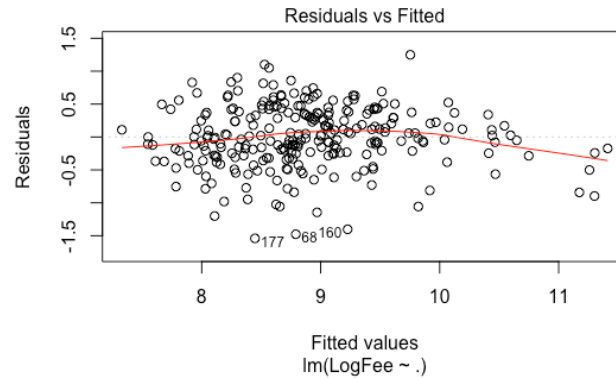
For further evidence, this study generates results using the 14 predictor variables. Modelling all of the possible predictor variables together serves to analyze the variables that

have the most substantial effects on predicting transfer fees and provides a basis for comparing the magnitude and significance of the effects, or lack thereof, of the two media coverage predictor variables. A summary of the results for regression (3) can be seen below in Table 6. For the purpose of brevity, only the variables deemed statistically significant are included. The residuals are plotted in Figure 9.

Table 6. Summary of regression (3)

Regression 3: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	6.929e+00	1.247e+00	7.97e-08 ***
Overall Rating	4.590e-02	1.702e-02	0.00756 **
Potential	-3.246e-02	1.609e-02	0.04493 *
Market Value	5.579E-05	6.042e-06	< 2e-16 ***
Age	-9.478e-02	1.855E-02	7.10E-07 ***
English_SellerY	1.910e-01	8.57e-02	0.02689 *
MSE: 0.3479561		Adjusted R-squared: 0.658	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05			

Figure 9. Residuals of “LogFee” as a function of 14 predictor variables



Based on these results, the full model of regression (3) is a better fit for the data than the regressions (1) – (2) given the increase in the adjusted R-squared statistic and the lower MSE value of $MSE=0.35$, which is used to determine the best model fit. The improved model fit suggests that there are variables besides those related to player brand value that can be informative for predicting transfer fees. An examination of the summary reveals five statistically significant variables, but the player and club media coverage variables are not among those five predictors. These results challenge the first hypothesis (1), suggesting that player brand value may not be well-suited for predicting transfer fees. Yet, since the two media variables were significant in regressions (1) and (2), they merit further investigation. It is possible that the discrepancy is due to collinearity between the media coverage variables and other variables in the dataset, which this study examines and discusses in section 5.3.

In evaluating the second hypothesis H2 dealing with the time effects of media mentions on transfer fees, this study adds the predictor “Season ID” as an interaction term with the two media coverage predictors and performs regressions (1) – (3) again. Table 7 provides summaries

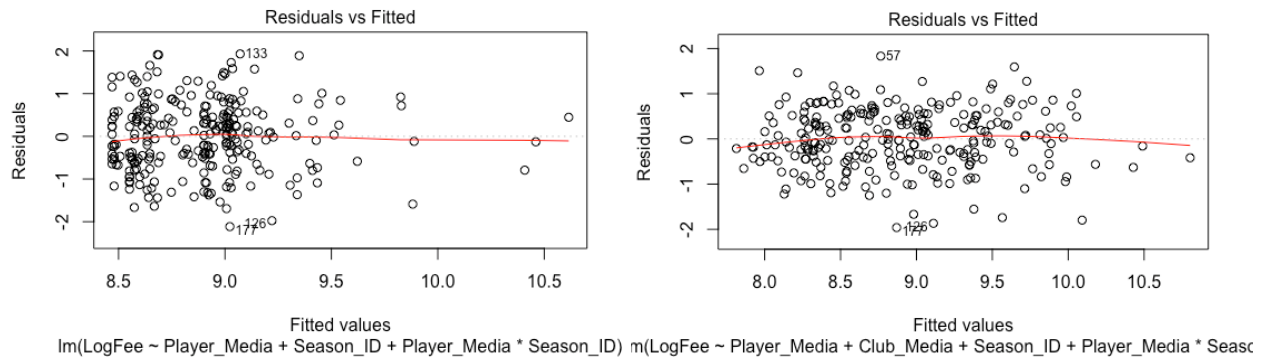
of the regressions (4) and (5). The table only shows the statistically significant coefficients.

Figure 10 plots their residuals side-by-side.

Table 7. Summary of regressions (4) and (5)

Regression 4: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	8.6345981	0.1541380	< 2e-16 ***
Player Media	0.0066706	0.0034781	0.0562
Season ID5	0.3725906	0.1960989	0.0586
MSE: 0.8042676		Adjusted R-squared: 0.1155	
Regression 5: Coefficients			
(Intercept)	7.5098604	0.198068	< 2e-16 ***
Player Media	0.0051125	0.0028738	0.0765
Club Media	0.0014945	0.0002451	4.1e-09 **
Season ID4	0.6421868	0.3871799	0.0262 *
Season ID5	0.7804913	0.2774576	0.0053 ***
Club Media: Season ID4	-0.0005623	0.0003387	0.0981
MSE: 0.5328813		Adjusted R-squared: 0.4009	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05			

Figure 10. Residuals for “LogFee” after the introduction of two interaction terms



For regression (4), its results differ from the corresponding regression (1) since “Player Media” is now only significant at the 0.1 level whereas in (1) it was significant at the 0.001 level.

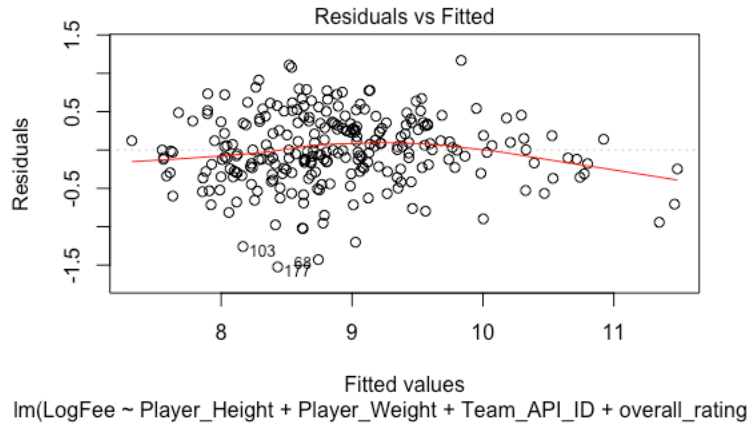
Moreover, none of the interaction terms are significant. Regression (5) provides a better fit than (4) and more closely matches the results from its corresponding regression (2) since “Club Media” is still significant at least at the 0.01 level and one of the interaction terms is nearly significant at the 0.1 level. Similar to the first two regression models, the introduction of variables related to club media mentions improves model fit, as the MSE value decreases from $MSE=0.8042676$ to $MSE=0.5328813$. Yet, taken as a whole, neither model provides strong evidence for a time effect in the relationship between media mentions, for both players and clubs, and transfer fees. As a result, there is no conclusive evidence for confirming the hypothesis H2 that player media mentions and transfer fees both increase together over time from one season to the next.

After evaluating the models for media mentions across time individually, this paper considers the results for the full model in regression (6) with 14 predictor variables and the two interaction terms for media coverage over time. A summary of regression (6) is detailed in Table 8. Figure 11 plots the residuals.

Table 8. Summary of regression (6)

Regression 6: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	7.038e+00	1.274e+00	9.80e-08 ***
Overall Rating	4.060e-02	1.690e-02	0.0172 *
Potential	-3.246e-02	1.609e-02	0.04493 *
Market Value	6.100e-05	6.292e-06	< 2e-16 ***
Age	-9.255e-02	1.838e-02	1.03e-06 ***
Season ID4	7.091e-01	2.960e-01	0.175 *
Club Media	8.641e-04	5.226e-04	0.0997
English_SellerY	1.949e-01	8.784e-02	0.0275 *
Season ID2: Player Media	-7.644e-03	3.260e-03	0.0200 *
Season ID5: Player Media	-5.558e-03	3.135e-03	0.0777
Season ID4: Club Media	-6.419e-04	3.315e-04	0.0542
MSE: 0.3461359		Adjusted R-squared: 0.6676	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05			

Figure 11. Residuals for “LogFee” as a function of 14 predictor variables and two interaction terms

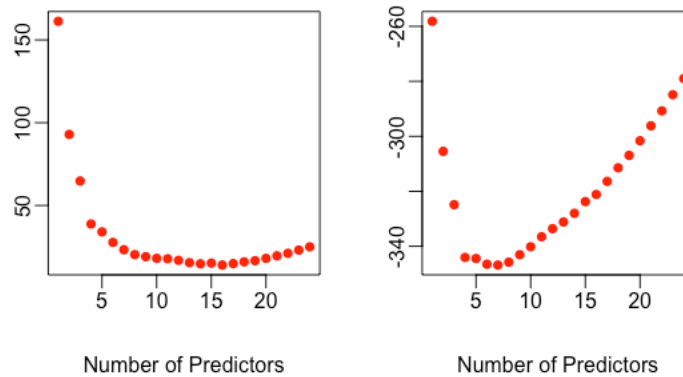


Similar to regression (3), the full model with the interaction terms suggests other variables are better suited than the media mentions variables for predicting transfer fees. Once again, the predictors “Market Value”, “Age”, “English SellerY” and “Overall Rating” are highly statistically significant, while neither of the two variables for media mentions are. However, this model provides better evidence for H1 and H2 than regression (3). The “Club Media” variable is significant at the 0.1 level, which is already an improvement. Moreover, one of the interaction terms, “Season ID2: Player Media”, is statistically significant and some of the other interaction terms are significant at the 0.1 level. Although the effects are not strong, these results provide at least some evidence for hypotheses H1 and H2 that player media coverage impacts transfer fees and that the impacts change over time.

According to the C_p and the BIC methods, this study fits two penalized regressions. The methods are instructive for variable selection, and the results from the C_p and the BIC criteria inform the selection of variables when fitting two linear regressions based on formulas (8) and

(9). In Figure 12, the leftmost plot shows the C_p as a function of the number of predictors and the rightmost plot shows the BIC as a function of the number of predictors in the regression model.

Figure 12. C_p and BIC error as functions of the number of predictors



According to the plot for Mallows's C_p criterion, the red dot representing the desired lowest possible C_p value corresponds to a model with 16 predictor variables. A model of that size should minimize prediction errors. Likewise, the same plot for the BIC criterion shows that prediction errors are minimized for a model with seven predictor variables. Prediction errors increase much more dramatically as more predictors are added for the BIC statistic than for the C_p statistic. Given the BIC has a larger penalty than C_p , the results make sense and the BIC fits a smaller model than the C_p does.

As discussed in the Methodology section, after establishing which predictors lead to the lowest C_p and BIC prediction errors, this study fits penalized regression models for the outcome variable "LogFee". This paper then performs cross-validation with the same training and test sets used for the other regression models. Tables 9 and 10 provide summaries for the C_p -fitted and BIC-fitted linear regressions and only include the statistically significant variables.

Table 9. Regression based on the C_p statistic (8)

C _p Regression: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	6.966e+00	5.989e-01	< 2e-16 ***
Player Weight	5.608e-03	2.067e-03	0.007132 **
Market Value	5.730e05	5.579e-06	< 2e-16 ***
Age	-5.870e-02	1.107e-02	2.57e-07 ***
Season ID4	6.790e-01	2.0282-01	0.000946 ***
Season ID5	1.056e+00	2.001e-01	2.89e-07 ***
Club Media	9.300e-04	2.383e-04	0.000213 ***
Season ID2: Player Media	-7.306e-03	3.120e-03	0.02008 *
Season ID5: Club Media	-7.461e-04	2.805e-04	0.008350 **
MSE: 0.347701		Adjusted R-squared: 0.6987	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05			

Table 10. Regression based on the BIC statistic (9)

BIC Regression: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	8.309e+00	4.598e-01	< 2e-16

Market Value	5.614e-05	5.325e-06	< 2e-16 ***
Age	-5.712e-02	1.101e-02	4.40e-07 ***
Club Media	6.480e-04	1.1074e-04	5.76e-09 ***
Season ID4	3.275e-01	1.235e-01	0.00853 **
Season ID5	5.992e-01	1.222e-01	1.71e-06 ***
Player Media	4.283e-03	2.063e-03	0.03890 *
Season ID2:	-7.208e-03	3.098e-03	0.02078 *
Player Media			
MSE: 0.3415589		Adjusted R-squared: 0.6838	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05			

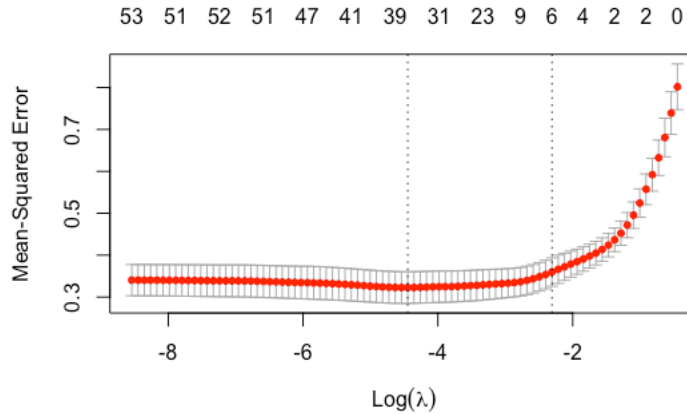
Comparing both resulting penalized regressions, this paper observes, as expected, that the BIC regression is smaller and that all of its selected seven predictors are statistically significant whereas not all of the 16 predictors for the C_p are significant. Some of the significant predictors, such as “Market Value” and “Age”, overlap with the significant variables of the regressions (1) – (6) and include positive coefficients. These results reinforce how these variables should be strong predictors of how much clubs should be willing to pay for a given player. As they relate to the hypotheses, these two penalized regressions include stronger evidence for H1 and H2 than the standard regression models. In both penalized regressions, the variable “Club Media” is highly significant, and in the BIC-regression, the “Player Media” variable is significant too. Across both models, the coefficients for both variables are also positive. These results suggest that media coverage affects transfer fees. Accordingly, they provide evidence for supporting the hypothesis H1 that greater player brand value, as measured by media coverage, is positively linked to higher

transfer fees. Although not shown in the tables, both penalized regressions include all of the interaction terms related to player and club media mentions across seasons. Some of the interaction terms for both player and club media mentions are also statistically significant in the two penalized regressions. Since variables selected by penalization regressions represent the ones that minimize prediction errors, the inclusion of all of these interaction terms and the statistical significance of some of them suggest that it is useful to measure media mentions over time when evaluating transfer fees. Although this paper cannot conclusively confirm the direction and full impact of the relationship, it finds support for hypothesis H2 insofar that there is a statistically significant difference between player media coverage and the amount clubs pay for players across distinct seasons. Despite the promise of these results, this paper notes that there is a discrepancy between how “Club Media” and “Player Media” are significant in the penalized regressions, but not in the unconstrained full model regressions (3) and (6). Once again, the discrepancy may point to collinearity between certain variables, which is explored in section 5.3 of the Results. Given these results, a comparison of both models requires looking at their respective MSE values. Using the C_p approach, $MSE=0.347701$, and for BIC, this paper finds $MSE=0.3415589$. Somewhat surprisingly, the BIC framework does not reduce the MSE much more than the C_p framework does, and both penalized MSE values are quite close to the MSE values resulting from the full model regressions (3) and (6).

While penalization methods are one form of variable selection, shrinkage methods provide another avenue for narrowing the set of predictors. Applying the LASSO approach involves obtaining results for a set of select variables and then performing linear regression. First, as discussed during the Methodology section, this study uses cross-validation to determine the optimal value for the tuning parameter λ . Figure 13 provides an illustration of the MSE on

the y-axis. The top end of the x-axis lists the number of non-zero coefficients and the bottom end the optimal λ value at given values for the MSE.

Figure 13. MSE as a function of the tuning parameter and the number of non-zero coefficients for a LASSO equation



The above plot includes several features that show the results obtained using the LASSO approach. First, the progression of the red points shows that the cross-validation errors vary according to different values of lambda and that the largest values of lambda correspond to the largest error values. Of the two vertical lines, the first denotes the lambda value producing the smallest mean cross validation error, which is the lambda used to fit the LASSO equation. After running the cross-validation process, this value is $\lambda = 0.01285$ with a standard error of 0.3276 and 39 non-zero coefficients. Using these results, this study fits a LASSO equation and obtains values for the 39 coefficients. This paper then uses these initial LASSO results to fit another linear model based on the predictor variables corresponding to the coefficients generated in the LASSO equation. The formula for that LASSO linear regression is as follows:

$$(12) \text{LogFee}_i = \beta_0 + \beta_1 \cdot \text{PlayerMedia}_i + \beta_2 \cdot \text{ClubMedia}_i + \beta_3 \cdot \text{PlayerHeight}_i + \beta_4 \cdot \text{PlayerWeight}_i + \beta_5 \cdot \text{Team ID}_i + \beta_6 \cdot \text{OverallRating}_i + \beta_7 \cdot \text{Age}_i + \beta_8 \cdot \text{MarketValue}_i +$$

$$\beta_9 \cdot Points_i + \beta_{10} \cdot AvgAttendance_i + \sum_i \theta_i EnglishSeller_i + \sum_i \delta_i SeasonID_i + \beta_{11} \cdot PlayerMedia_i \cdot \sum_i \delta_i SeasonID_i + \beta_{12} \cdot ClubMedia_i \cdot \sum_i \delta_i SeasonID_i + \epsilon_i$$

The resulting statistically significant coefficient estimates from the LASSO equation and the linear regression are compared in Tables 11 and 12.

Table 11. Summary of LASSO equation coefficient estimates

LASSO Equation: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	8.309e+00	4.598e-01	< 2e-16
Market Value	6.640e-05	6.311e-06	< 2e-16 ***
Age	-6.730e-02	1.1214e-02	8.73e-08 ***
Adjusted R-squared: 0.679			
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05			

Table 12. Summary of regression coefficients using LASSO predictor variables

LASSO Regression: Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	6.679e+00	1.012e+00	1.69e-10 ***
Overall Rating	1.432e-02	5.795e-03	0.01399 *
Market Value	6.518e-05	5.340e-06	< 2e-16 ***
Age	-6.366e-02	1.000e-02	6.69e-10 ***
Season ID4	4.733e-01	2.333e-01	0.04330 *

Season ID5	6.673e-01	2.751e-01	0.01581 *
English SellerY	1.911e-01	7.318e-02	0.00942 **
Season ID2: Player Media	-7.193e-03	2.974e-03	0.01614 *
MSE: 0.3107289		Adjusted R-squared: 0.6677	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05			

An interpretation of the coefficients shows that once again the predictors “Market Value” and “Age” are highly significant across both models. Since they appear significant across the different regression models, this paper concludes that both variables are strong predictors of a player’s transfer fee. Although neither table above shows significant coefficient estimates for the two individual media coverage predictor variables, this paper observes that one of the interactions for player media mentions across seasons is statistically significant. Since the same interaction term is significant in some of the other regression models, the results provide further support, albeit to a small degree, for the second hypothesis H2. Likewise, the LASSO results for the individual season four and five coefficients match the results from several of the other regressions, which find these coefficients statistically significant. Thus, at the very least, the findings of the EDA phase on transfer fees over time in section 3.3 can be re-considered, and this study can conclude from the LASSO approach that the season or transfer window a player is bought in can affect the transfer fee. Comparing the predictions from the testing set to the actual transfer fees, this study finds MSE=0.3107289 for the LASSO regression. The low MSE value points to a strong model fit, as does the high adjusted R-squared value. Despite not being able to

fully confirm the hypotheses, this study confirms that the LASSO regression provides evidence to support H1 and H2 as well as a strong model capable of predicting player transfer fees.

After performing and obtaining results for all of the regression models, this paper directly compares the results for model fit. Table 13 below aggregates the MSE values for each of the regressions performed.

Table 13. Summary of regression MSE values

Type	MSE
Regression (1)	0.8137478
Regression (2)	0.5508651
Regression (3)	0.3479561
Regression (4)	0.8042676
Regression (5)	0.5328813
Regression (6)	0.3461359
Regression C_p	0.347701
Regression BIC	0.3415589
Regression LASSO	0.3107289

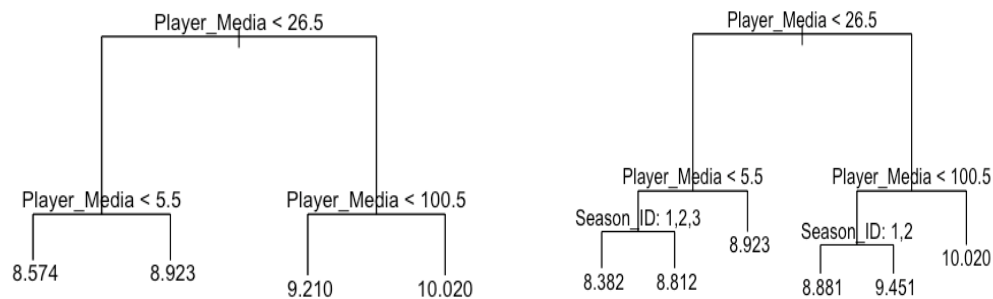
In comparing results, this study finds that the LASSO approach yields the regression with the lowest MSE and thus the best model fit. The LASSO model fit does not necessarily confirm the hypotheses H1 and H2, but it does suggest that the LASSO approach is useful for building a model capable of predicting player transfer fees with a high degree of accuracy.

5.2 Comparing Decision Trees

In the Methodology section, this paper details the process for the implementation of four separate decision tree methods. Once again, this study uses out of sample validation for all of these decision trees to generate the resulting MSE values for model comparison. Based on the methodology, this paper generates 12 sets of results and MSE values, which represent the implementation of the four methods three different times each to account for three types of model specifications (1) – (3).

First, the results for a single tree implementation are shown. The first single tree (1) only fits the predictor “Player Media” while the second, single tree (2), adds “Season ID” as a second predictor. These two trees serve to isolate for the effects of the main predictors dealing with media exposure, which is the focus of the two hypotheses H1 and H2. Figure 14 below shows the two resulting single trees.

Figure 14. Single trees (1) and (2)

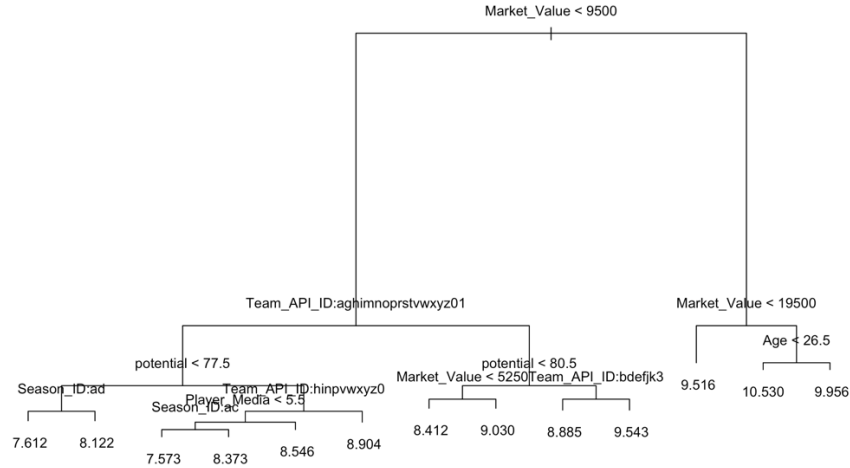


For both trees, the number of player media mentions has the same effect, dividing the trees into two branches at approximately 27 media mentions. The right branch represents players with higher amounts of media mentions compared to the left branch, and a comparison of the terminal nodes for both branches shows that players with more media mentions have higher transfer fees,

which aligns with the first hypothesis H1. Meanwhile, for both branches in the second tree, the season parameter groups seasons 4-5 together in their own group, which is always separate from the group containing seasons 1-2. The branches with seasons 4-5 correspond to the nodes with the highest transfer fees paid. Of all of the end nodes, the node corresponding to the highest transfer fees follows from the branches with the highest number of player media mentions and corresponding to the latter seasons 4 and 5. Thus, these results help confirm both hypotheses H1 and H2. Both trees show that a greater number of media mentions corresponds to a higher transfer fee, and that the transfer fee increases as the number of player media mentions increases and as the data points move to the latter seasons of the dataset, namely seasons 4 and 5. In terms of MSE values, first tree illustrated has $MSE=0.8045059$ while the second results in $MSE=0.8160457$. Both MSE values are higher than the ones obtained in the regression models. Although this study performs pruning for both of these trees (1) and (2), it does not elaborate on the results since they are similar to the original single regression trees illustrated above. Table 14 below presents the resulting MSE values from the trees (1) and (2) with pruning.

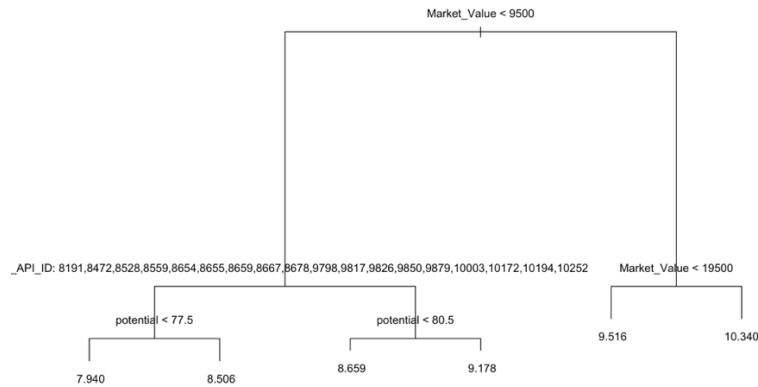
In order to understand the effects of the main predictor variables relative to other possible predictors for the main outcome variable, this study fits the third type of single tree (3) with the 14 predictor variables in the dataset. Since several predictor variables are fitted, the results for the tree with and without pruning are shown to illustrate the impacts of pruning. The results for single tree (3) are shown in Figures 15 and 16 below.

Figure 15. Single tree (3) with 14 predictor variables without Pruning



The single regression tree without pruning has 13 terminal nodes, uses six predictor variables for tree construction, resulting in $MSE=0.3835501$. The first variable used for minimizing RSS in the tree is “Market Value”, and the right branch suggests that players with higher market values command higher transfer fees. After a player’s market value, his age becomes the next metric of differentiation in the right branch. The split along “Age” for the right branch reveals that young players, specifically those younger than 26.5 years old, with market values above 19.5 million euros are bought for the highest transfer fees. On the left branch, the first additional split is along the lines of players’ respective teams. However, the left path seems overly complex and difficult to interpret. In that regard, pruning helps make the single tree easier to evaluate, and its effect can be seen in Figure 16 below.

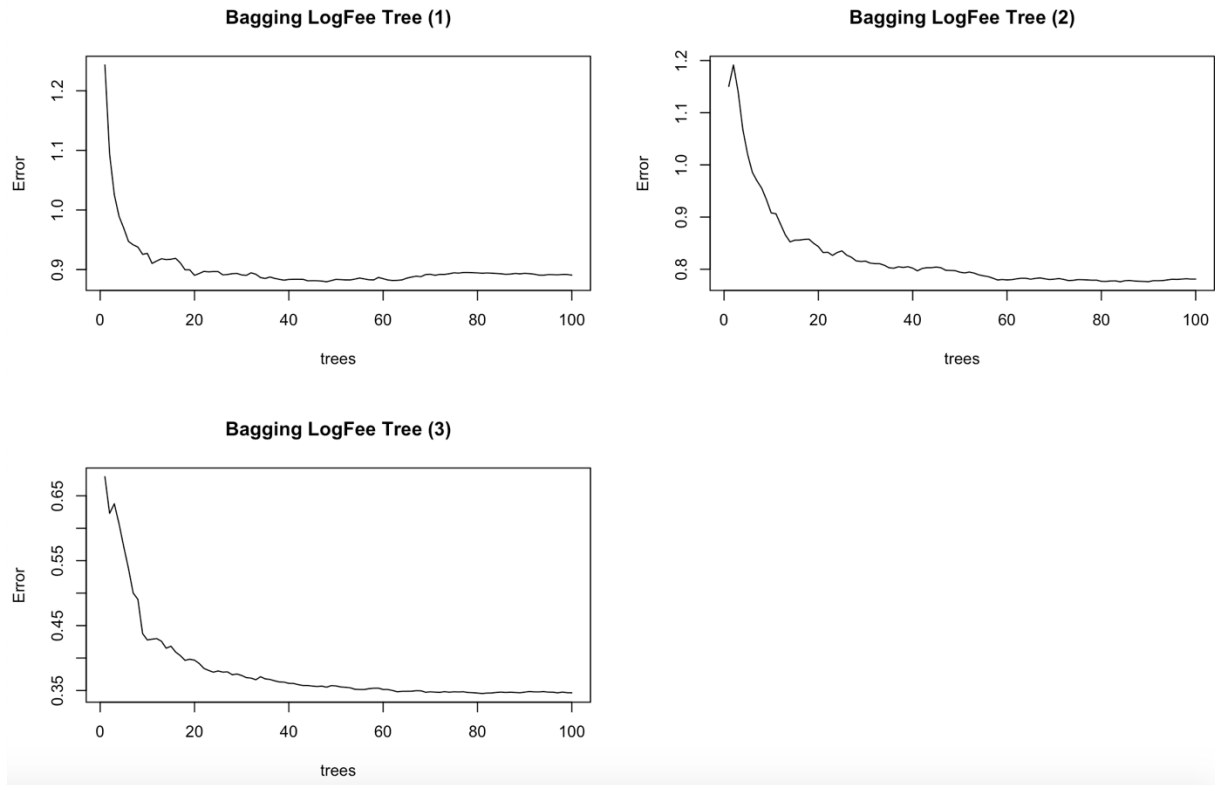
Figure 16. Single tree (3) with 14 predictor variables with Pruning



Comparing the single tree (3) with and without pruning confirms that pruning cuts down on the number of splits and terminal nodes, making it easier for interpretation. Pruning also slightly improves accuracy and reduces testing error, resulting in $MSE=0.3687932$. The pruned tree has six terminal nodes and only uses three variables in tree construction: “Market Value”, “Team ID” and “Potential”. Once again, the first variable to be split is “Market Value”, which reinforces its importance as a predictor for transfer fees. Since the left side of the tree is now easier to analyze, this study finds that the high potential branches along each split of potential lead to nodes with higher transfer fees. From a football perspective, the use of a player’s potential to evaluate his transfer value makes sense since clubs want to buy players who can improve and appreciate in value.

Next, this paper analyzes the results from using multiple regression trees. The first method implemented for multiple regression trees is bagging. Figure 17 below presents the results from using bagging for the three types of tree specifications (1) – (3).

Figure 17. Implementation of Bagging for the three types of tree specifications (1) – (3)

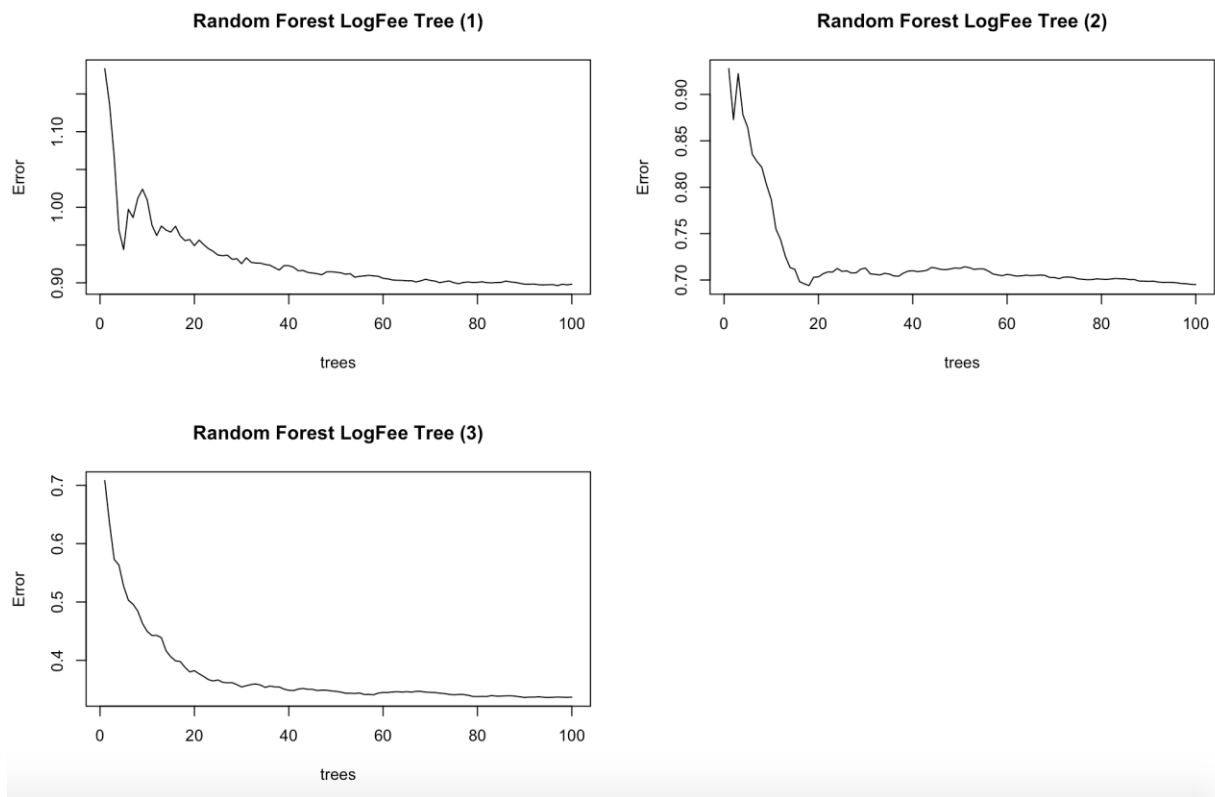


Throughout the three plots, the testing error decreases substantially as more trees are bagged in the model, until levelling off. Although 100 trees are bagged across the three different specifications, the results vary since the testing error decreases substantially after around only 20 trees for the first bagged model whereas the latter two models require approximately 60 trees to reach similar testing errors. With bagging, the model gains accuracy. For the first two models with bagging, their MSE values do not differ greatly from the MSE values of the single trees with the same model specifications. However, the third bagged model with all fourteen predictors has $MSE=0.2748568$, which is significantly lower than the $MSE=0.3687932$ of the single tree (3) with pruning for all 14 predictors. The trade-off to greater accuracy is less interpretability. Since so many trees are fitted, the results from the bagging method are less explicit and harder to draw inferences from. For example, this paper cannot observe different

branches, splits and nodes to draw conclusions. As a result, the path to “LogFee” values, how variables interact with each and how they interact with the outcome “LogFee” are all unclear. Nonetheless, the bagging method provides an avenue for improving predictive accuracy.

Finally, this study implements decision trees using the random forest framework, which also entails using multiple regression trees. The results from applying random forest to each of the three types of tree specifications (1) – (3) are illustrated in Figure 18.

Figure 18. Implementation of Random Forest for the three types of tree specifications (1) – (3)



Similar to the results for bagging, these results confirm that the random forest framework reduces testing error for prediction as more trees are added to an implementation. Comparing bagging and random forest demonstrates that the corresponding bagging and random forest plots for each of the three specifications have their testing errors level off at different numbers of trees. For example, specification tree (2) starts to level off at around 60 trees with bagging whereas it

reaches its low point after approximately 20 trees with random forest. Thus, although both frameworks use multiple regression trees, their results differ. A comparison of MSE values shows that random forest does not significantly reduce testing error relative to the single tree method for the first two model specification trees (1) and (2). However, compared to the single tree with pruning for all 14 predictor variables, the corresponding random forest implementation significantly reduces MSE from $MSE=0.3687932$ to $MSE=0.2724107$. The lower MSE value reinforces how the random forest approach yields greater accuracy and predictive power. Yet, like with bagging, it is difficult to interpret the relationships between the predictor variables as well as between the predictors and the outcome. Based on both the bagging and random forest results, this study confirms that using multiple regression trees involves a trade-off between greater predictive accuracy and less model interpretability.

Since this study implements four decision tree methods and applies each of them across three distinct model specifications, it obtains 12 MSE values, which are summarized in Table 14 for direct comparison.

Table 14. Comparison of MSE values achieved using Single Trees, Single Trees with Pruning, Bagging and Random Forest

MSE	Single Tree	Single Tree with Pruning	Bagging	Random Forest
Specification (1)	0.8045059	0.8045059	0.8271133	0.824194
Specification (2)	0.8207835	0.8160457	0.8202164	0.7393126
Specification (3)	0.3835501	0.3687932	0.2748568	0.2724107

The table above shows that the effect of using different types of decisions trees on the MSE depends on the inputs. For example, the MSE increases after using multiple regression trees

through bagging and random forest compared to the implementation of single trees for specification tree (1) with only “Player Media” as the input. In contrast, for specification (3), which includes all 14 predictors as inputs, the MSE decreases going from single to multiple regression trees. As seen in the table, the winning model showing the best predictive capabilities with the lowest MSE value is the random forest model implemented with all 14 of the predictor variables in the dataset. Yet, in using that approach to make predictions for transfer fees, this paper cannot discern the precise impact or significance of a player’s media coverage and brand value on the transfer fee. This study also cannot evaluate how the relationship between player brand value and transfer fees evolves over time.

5.3 Discussion of Results

This study provides several insights for both researchers and practitioners who are interested in marketing and branding in sports and entertainment more widely. Taken as a whole, the results provide evidence that media coverage, particularly player media exposure, has a positive economic effect on the economic worth of football players and that the effect changes over time. For example, the linear regressions and single regression trees that isolate for the main predictor variables related to H1 show that the player and club media coverage variables are both positively and significantly linked to transfer fees. Likewise, the BIC regression finds similar results. The BIC model also finds that the interaction term for player media mentions over time is statistically significant for transfer fees. However, the strength of the evidence varies greatly depending on the statistical models. Although some models include “Player Media”, “Club Media” or their interaction terms as highly significant variables, there is a lack of consistency in the effects across models. Thus, this study finds support for H1 and H2, but does not conclusively confirm the two hypotheses.

The results also provide useful information on the application and suitability of different statistical models for making economic predictions in sports and entertainment. Through the collection of data on different variables and statistical model implementations, this study provides guidance on which factors most affect a player's economic value and a club's willingness to pay for a player. First, this study demonstrates that random forest models are robust for making economic predictions, such as predicting a player's transfer fee, since it is the best fit model with the lowest MSE. However, since it is difficult to interpret the effects of specific variables with random forest, this discussion focuses on the potential implications of the LASSO regression, which is the next best fit model with the second lowest MSE.

In drawing potential conclusions from the LASSO regression, this paper examines the variables of statistical significance. These variables include the following: "Overall Rating", "Market Value", "Age", "Season ID", "English Seller" and the interaction between "Season ID" and "Player Media". Of these variables, the "Market Value" and "Age" variables are the ones which are most consistently significant across the statistical models. Intuitively, the significance and negative coefficient of the "Age" predictor makes sense since football players typically start their careers young, some as young as 16 or 17 years old, and reach their peak performance levels in their early to mid-twenties. A player's career at the elite level rarely lasts past the age of 30 given the sport's physical demands. Thus, clubs should be willing to pay less for players as they grow older. Likewise, the strength of the "Market Value" predictor seems reasonable. As discussed in the Methodology section, the variable represents a player's real-time economic value at any given time generated from fans' crowd-sourced estimates of a player's economic value on the football website *Transfermarkt*. While clubs may benefit from having more information on players than fans, their internal valuations are likely not too far off from those of

fans, which the data and results seem to confirm. The results showing the variable's significance for predicting transfer fees also corroborate and further validate past studies, which also find that crowd-sourcing estimates of a player's value can help predict how much a club actually pays for a player (Herm, Callsen-Bracker and Kreis 2014; Müller, Simons and Weinmann 2017).

Continuing the discussion of statistically significant variables in the LASSO model, this paper can justify the significance of the "Overall Rating" variable. In the dataset, the variable is the only one that provides an estimate of a player's current ability. Since teams are in the habit and business of trying to win games, they need to build their squads around the best players they can get. Thus, naturally, teams should base a large part of their player acquisition strategy and willingness to pay for a player on the player's footballing ability and performance levels. The competition to buy the best players should also drive transfer fees up. Although not as straightforward as the one for "Overall Rating", the possible explanation for the significance of the "English Seller" variable also makes sense. Specifically, the variable is significant for the category "Y", which is used to represent the transfers when a Premier League club buys a player from a fellow English club. Media commenters commonly postulate that English clubs charge premiums for their players when selling to fellow English clubs to compensate and offset for the possibility of strengthening their direct rivals. Clubs may also prefer buying, and could even be willing to pay premiums for, players with Premier League experience since these players could start contributing to their new teams without needing to adapt to a new league and its unique physical demands. Finally, the remaining two significant variables in the LASSO regression, namely "Season ID" and the interaction between "Season ID" and "Player Media, pertain to the second hypothesis (2). The justification for both variables, and accordingly for the second hypothesis H2, is that the Premier League's commercialization has increased over time and is

projected to continue growing (Ahmed 2019). As the Premier League commercializes and reaches a global audience, its clubs can profit more from the popularity of their brands and their players' brands, and thus make more money available to buy players. Likewise, as more people follow football, players can boost their own personal brands, such as by attracting large amounts of followers on social media platforms, and profit from their brands through sponsorships. The convergence of these trends should lead to clubs paying higher amounts for popular players who can boost club brand profiles and thus club profits. While these variables are significant for the LASSO regression in this study, they should be further investigated within the current sports and entertainment climate.

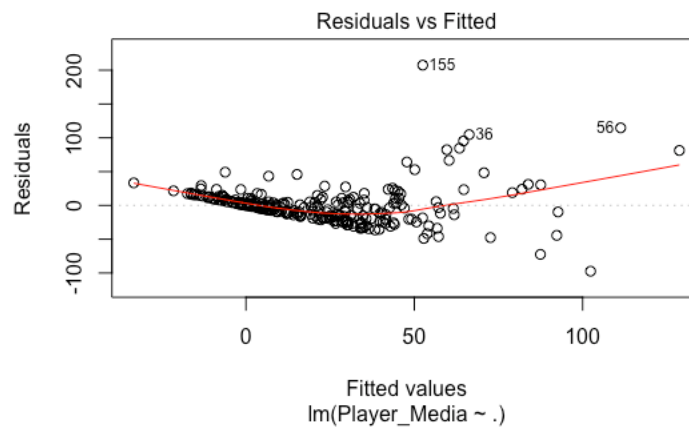
As discussed in section 5.1, there remain a few issues with the results, which this paper seeks to investigate further. First, the results show a discrepancy in the effect of the main predictor variable, "Player Media". The variable is significant for "LogFee" when fitted individually, but it loses some of its significance when fitted alongside other predictor variables. Moreover, this study also looks to understand why the two media coverage predictors carry greater significance in the constrained regressions than in the unconstrained ones. Both of these issues point to potential problems with collinearity between some of the predictor variables. Since the study focuses specifically on player brand value, which is proxied through "Player Media", this study investigates the effect of this variable and its relationship to other variables by performing two additional statistical models with "Player Media" as the outcome variable.

In the first model, this study performs linear regression (3), but replaces the outcome variable with the "Player Media" variable, removing any variables directly related to transfer fees. The results are shown in Table 15, which only includes the statistically significant predictors. Figure 19 shows the residuals plot.

Table 15. Summary of regression with “Player Media” as the outcome variable

Player Media			
Regression:			
Coefficients			
	Estimates	Std. Error	Pr(> t)
(Intercept)	-1.253e+02	7.283e+01	0.0868
Market Value	1.926e-03	3.222e-04	9.12e-09 ***
English Seller	3.545e+01	4.546e+00	2.53e-13 ***
MSE: 688.3003		Adjusted R-squared: 0.3539	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05			

Figure 19. Residual plot for “Player Media” as the outcome variable



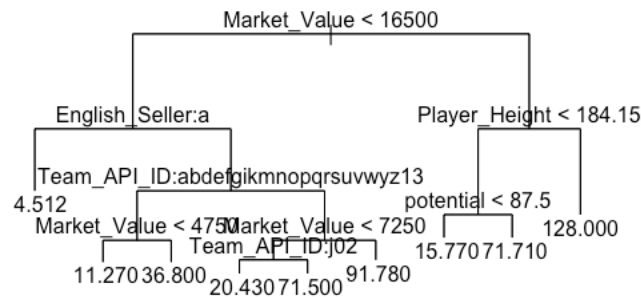
Looking at regression’s summary reveals that only two variables are significant and both of them are significant at the highest level of 0.001 significance: “English Seller and “Market Value”.

The residuals plot also shows a fanning out pattern as the fitted values increase, which reflects poor model fit. Given these results, it is possible that both of these variables bias the main

predictor, “Player Media”, and its effects. For example, possible collinearity between the main predictor and these significant variables could lead to variable omission bias. Since both “Player Media” and “Market Value” are numeric variables, this paper finds their correlation of $r=0.38$, which is significant enough to potentially cause issues in a regression analysis.

In order to find further support for the explanation, this study also performs a new single regression tree without pruning by making “Player Media” the outcome variable and removing the “LogFee” variable from the training and validation datasets. Although the random forest approach may yield greater predictive accuracy, the single tree allows for visualizing the path of how variables interact with “Player Media”. Figure 20 shows the results for the regression tree.

Figure 20. Single regression tree with “Player Media” as the outcome variable



Although the regression tree may seem complex, its features provide support for the explanation provided above in the multiple regression analysis. As the variable first used to minimize RSS and split the tree, the “Market Value” is strongly linked to “Player Media”. Likewise, the next split on the left branch occurs with the “English Seller” variable, which establishes its importance for interpreting player media coverage in the regression tree. Since these two variables are the same as the two statistically significant variables in the multiple regression, they

are likely closely linked to the study's main predictor of "Player Media". Thus, the results for the regression tree are consistent with the results obtained from performing a multiple regression.

Although the relationships between these three variables should be studied further in future research, the relationships make sense. Players with high market values are often the best players. In turn, fans are drawn to following the best players, which makes the media more likely to cover and discuss those players. The construction of both variables for the dataset could also play a role in their collinearity. The variable "Market Value" comes from a website that aggregates fans' opinions on a player to generate an estimate of his potential worth on the transfer market. Fans could contribute more inputs and opinions for a player's worth if the player is already popular, which would make the estimates for player market values partly reliant on player popularity. Similarly, the media is also likely to comment on and cover popular players more than less popular ones in order to generate fan engagement. Finally, the connection to the "English Seller" variable is first illustrated during the Exploratory Data Analysis stage of section 3.3, which shows how players from foreign leagues receive far fewer media mentions than players who are bought from fellow English clubs. As discussed in section 4.5, the nature of the bias in the data makes sense since the only source used for obtaining media mentions is an English media outlet. However, despite its potential limitations in this study, the variable on player media coverage and popularity should still be considered as a potential proxy for player brand value and as a possible factor for determining transfer fees in future research. If it is possible to obtain media mentions from a variety of different media outlets and to control for potential country effects, then the variable could become more robust and show significant predictive capabilities.

SECTION FIVE: CONCLUSION AND SUMMARY

5.1 Concluding remarks

In the past decade, the English Premier League's global profile and wealth have grown exponentially. While researchers have previously explored transfer fees in football, few studies examine transfer fees within the context of the commercialization taking hold of the sport. This study implements and evaluates several different statistical models which rely on machine learning techniques to determine the effect of football player brand value on transfer fees. This study's results provide evidence that a player's brand value, as measured by the amount of media coverage received, can positively impact how much a club would pay to buy that player. The findings also suggest that the impact can vary over time. Since the findings can vary depending on the different statistical models this study tests, they are not conclusively confirmed, and further research is suggested. This study's findings serve as a foundation for further research on how brand value impacts the economic aspects of sports and entertainment, such as the question of how much to pay for talent. While this study is focused on football, the approach and findings should also be considered by researchers and practitioners investigating how to value individual performers within business organizations as well.

5.2 Areas for further research

Within the context of football, there remains a need for further research on several aspects related to transfer fees and the measure of a player's worth. While this study finds some evidence of a branding or media effect for players, more research is needed to consider the impacts of media coverage across a variety of media sources for players in different leagues and countries. For that research, the collection of new media-related data or of any data that could approximate an individual's personal brand is critically important. Moreover, this paper

encourages further investigation into how negotiation dynamics between clubs can affect how a player is valued. This study also does not include contract value as a variable given that the data on player contracts is not publicly available. If it is possible to eventually obtain such data, then understanding the factors impacting player contracts will be important. Although clubs are not the focus of this study, researchers should investigate how club branding affects different aspects of a club's football operations, including but not limited to the player acquisition strategy. Many sports may have different structures to football, but the research on branding in football should be expanded to other sports as well. Beyond football, researchers and practitioners should continue investigating the economic impacts of influencers or any other individuals who can use their personal brands to provide economic value. In a similar vein, further research on organizational behavior should be conducted using insights from sports research to evaluate which factors businesses use to evaluate, hire and retain their employees.

Moreover, despite growing interest in machine learning, few researchers studying sports, marketing or entertainment implement machine learning techniques in their studies. This study can provide guidance on which machine learning methods are informative and on how best to implement them. Future research should aim to utilize methods such as penalization or shrinkage for regression or decisions trees since they can lead to high degrees of predictive accuracy. The machine learning methods can go beyond the scope of the statistical models covered in this study. For example, Neural Networks and Principal Component Analysis are two other machine learning techniques that can and should be applied to draw insights for sports, marketing and entertainment. Ultimately, both practitioners and researchers alike should explore how to best utilize machine learning techniques and machine learning's possible applications.

WORKS CITED

- Ahmed, Murad. 2019. "Premier League Transfers: The Rational Thinking Behind The Big Spending | Financial Times". *Ft.Com*. <https://www.ft.com/content/8873b1cc-ba7b-11e9-8a88-aa6628ac896c>.
- Bauer, Hans H., Nicola E. Sauer, and Philipp Schmitt. 2005. Customer-based brand equity in the team sport industry. *European Journal Of Marketing* 39 (5/6): 496-513. doi:10.1108/03090560510590683.
- Bartz, Sherry, Alexander Molchanov, and Philip A. Stork. 2013. "When A Celebrity Endorser Is Disgraced: A Twenty-Five-Year Event Study". *Marketing Letters* 24 (2): 131-141. doi:10.1007/s11002-013-9229-2.
- Brid, Rajesh. "Decision Trees—A Simple Way To Visualize A Decision." Medium. 2018. <https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb>.
- Brownlee, Jason. "Bagging And Random Forest Ensemble Algorithms For Machine Learning." Machine Learning Mastery. 2019. <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>.
- Bryson, Alex, Bernd Frick, and Rob Simmons. 2012. The returns to scarce talent. *Journal Of Sports Economics* 14 (6): 606-628. doi:10.1177/1527002511435118.
- In this study, a group of researchers looks at whether players who are two-footed
- Chibber, Kabir. "The Premier League Premium." Quartz. 2018. <https://qz.com/1475058/premier-league-premium-how-footballers-benefit-from-playing-in-england/>.

- Choi, Sejung Marina, and Nora J. Rifon. 2007. "Who Is The Celebrity In Advertising? Understanding Dimensions Of Celebrity Images". *The Journal Of Popular Culture* 40 (2): 304-324. doi:10.1111/j.1540-5931.2007.00380.x.
- Ding, Haina, Alexander E. Molchanov, and Philip A. Stork. 2010. "The Value Of Celebrity Endorsements: A Stock Market Perspective". *Marketing Letters* 22 (2): 147-163. doi:10.1007/s11002-010-9117-y.
- Dobson, Stephen, and Bill Gerrard. 1999. The determination of player transfer fees in English professional football. *Journal Of Sport Management* 13 (4): 259-279. doi:10.1123/jsm.13.4.259.
- Dobson, Stephen, Bill Gerrard, and Simon Howe. 2000. The determination of transfer fees in English nonleague football. *Applied Economics* 32 (9): 1145-1152. doi:10.1080/000368400404281.
- Frick, Bernd. 2007. The football players' labor market: empirical evidence from the major European leagues. *Scottish Journal Of Political Economy* 54 (3): 422-446. doi:10.1111/j.1467-9485.2007.00423.x.
- Garcia-del-Barrio, Pedro, and Francesc Pujol. 2006. Hidden monopsony rents in winner-take-all markets—sport and economic contribution of Spanish football players. *Managerial And Decision Economics* 28 (1): 57-70. doi:10.1002/mde.1313.
- Herbinet, Corentin. 2018. "Predicting Football Results Using Machine Learning Techniques". *Imperial.Ac.Uk*. <https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-professional-football-matches.pdf>.

- Herm, Steffen, Hans-Markus Callsen-Bracker, and Henning Kreis. 2014. When the crowd evaluates football players' market values: accuracy and evaluation attributes of an online community. *Sport Management Review* 17 (4): 484-492. doi:10.1016/j.smr.2013.12.006.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction To Statistical Learning*. 7th ed. Reprint, New York: Springer, 2017.
- Kamins, M., Brand, M., Hoeke, S. and Moe, J. (1989). Two-Sided versus One-Sided Celebrity Endorsements: The Impact on Advertising Effectiveness and Credibility. *Journal of Advertising*, 18(2), pp.4-10.
- Malik, Mukul. 2019. "Basics Of Neural Network". *Medium*. <https://becominghuman.ai/basics-of-neural-network-bef2ba97d2cf>.
- Müller, Oliver, Alexander Simons, and Markus Weinmann. 2017. Beyond crowd judgments: data-driven estimation of market value in association football. *European Journal Of Operational Research* 263 (2): 611-624. doi:10.1016/j.ejor.2017.05.005.
- "Predicting Football Matches Using EA Player Ratings And Tensorflow". 2018. *Medium*. <https://towardsdatascience.com/predicting-premier-league-odds-from-ea-player-bfdb52597392>.
- Saed, Sherif. 2019. "EA Explains How FIFA Player Ratings Are Calculated - VG247". *VG247*. <https://www.vg247.com/2016/09/27/how-ea-calculates-fifa-17-player-ratings/>.
- Torgler, Benno, and Sascha L. Schmidt. 2007. What shapes player performance in football? empirical findings from a panel analysis. *Applied Economics* 39 (18): 2355-2369. doi:10.1080/00036840600660739.
- Wetzel, Hauke A., Stefan Hattula, Maik Hammerschmidt, and Harald J. van Heerde. 2018. Building and leveraging sports brands: evidence from 50 years of German professional

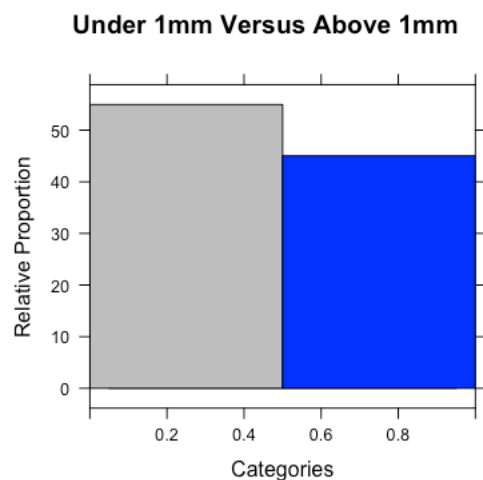
football. *Journal Of The Academy Of Marketing Science* 46 (4): 591-611.

doi:10.1007/s11747-018-0580-y.

APPENDICES

Appendix I

The of the full dataset illustrates the proportion of transfers of players bought above the one million euros threshold relative to the total number of transfers of players bought by English Premier League clubs over the five seasons studied



Appendix II

Correlation Heatmap With Numeric Variables

