Dylan Ellis and Colby Wight
Data Science Incubator CS 5830/6830
Presentation Slides

Project 3 - Web Scraping and Soccer Transfer

Introduction

Soccer is one of the world's most popular sports. The soccer transfer window is a time when teams can buy and sell players. If a team wants to buy a player there is usually a fee associated with that transfer. With the January transfer window just closing there has been a lot of talk about who went where. Our analysis will focus on the players that have changed teams. We want to know if the players being talked about on the various internet sites are more likely to be transfered. We also want to know if there is a relationship to the player's actual value and their transfer fee. As player transfers can go for millions of dollars, making the right trades for the right price is crucial to a soccer club's success.

Dataset

The first step was finding a set of player names to work from. To find an initial set of player names we downloaded a CSV through Kaggle. In cleaning this data set we had to split the players full names into first and last names. Some players only go by one name, e.g. Neymar, so these players were eliminated from the data set to keep the cleaning simple. At this point we scraped 6 popular soccer transfer news sites: Sky Sports, Onefootball, 90min, newsnow, talksport, and sbnation for the names of the players left in the list. We then summed the number of occurrences of each players' names to find how often they were being talked about.

The next problem that arose was players who had more than two words in their full name, e.g. David De Gea or Frenkie de Jong, to handle these names the sums were sorted then examined in a CSV file. Any names that were not full last names and had counts greater than 25 were eliminated from the data set.

We needed to know what players actually transferred and what their transfer fee and market value was. We found a website called TransferMarkt (https://www.transfermarkt.co.uk/) that had a nice table with this information. We were able to turn this into a data table with the player name, age, transfer fee, and market value as columns.The transfer fee and market value columns required the removal of £ character and m denoting million. The cleaning became more difficult as the transfer fees got smaller since the m for million changed to k for thousand. There were also players in the table who were loaned instead of transferred, meaning they had no transfer fee. These players had their transfer set to 0 and were removed from the data set. We were able to get the names, age, market value and transfer fee of the 75 most expensive players that were transferred during this transfer window.

Analysis Technique

In order to see how strong of a relationship there was between a players market value and how much they were transferred for, we used a scatter plot of the two values and found the

pearson correlation coefficient with its accompanying p-value. This allowed us to visualize the general trend, determine the correlation, and determine our confidence in the correlation.

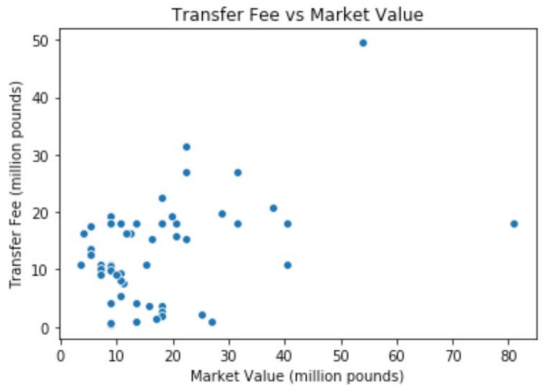
To further explore the relationship between market value and actual transfer fee, we created histograms of each then found the standard deviation and mean. We conducted a t-test to compare the two means and determine the significance of the difference between them.

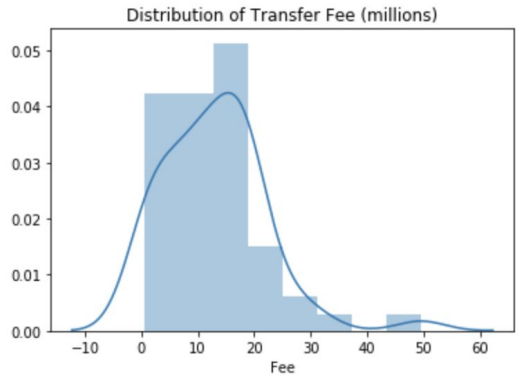
We wanted to analyze the relationship between age and the market value to see if young players are worth more on the transfer market than older players are. We did this by first generating a scatter plot of the age VS market value and the age VS transfer fee. We then split the data into two classes, younger players (below 25) and older players (over 25) and found the mean transfer fee for each class of player. We compared the means using a t-test to see if there was a statistical difference between the means.

Finally, we wanted to see if players who were transferred recently were talked about more in the news than players who hadn't been transferred recently. We did this by using the requests package to scrape 6 popular soccer news websites for instances of the names of both players who had been transferred recently as well as players from the data set found on kaggle.

Results

We learned a few things when analyzing the relationship between player transfer fee and their market value. As you might expect, we found that there is a positive relationship here. The more a player is worth (their market value) the more they will typically cost to transfer. This relationship was seen with the top 75 transfers. However, correlation was not as strong as you might think at .42. With a p-value of 0.001 we can be confident that this positive relationship exists here as seen in the figure below.

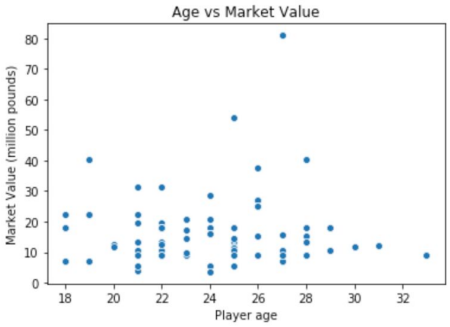


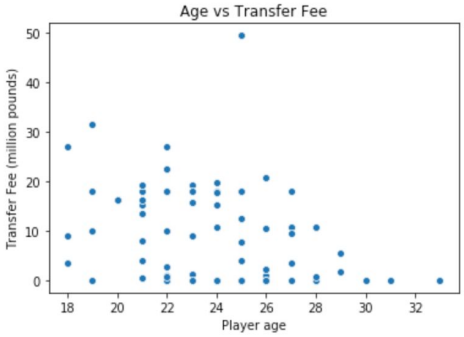


When we looked at the scatterplot with only the top 25 players, the correlation was not so obvious. Why might this relationship not be as strong as you would think? To further explore this correlation, we found that the mean player market value was 17.13 m and the mean player transfer fee was 9.30 m. This tells us that the how much money a player gets transferred for is not the same as what they are worth, in fact they usually get transferred for much less than they are worth. There is one player who explains this trend very well, Christian Eriksen. Eriksen is out of contract in the summer and doesn't want to sign a new one. This means that the club wants to make any money they can by selling him in the winter before his contract expires. Other teams aren't willing to pay his full market value because they know that if they just wait until the

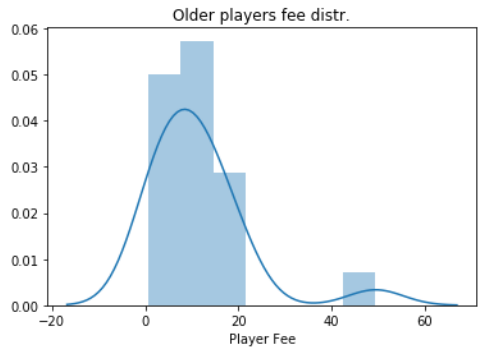
summer they can have him for free. Eriksen is an extreme example of this phenomena, but often the only players who are transferred in January are wanting to leave their clubs and therefore, go for less than their market value.

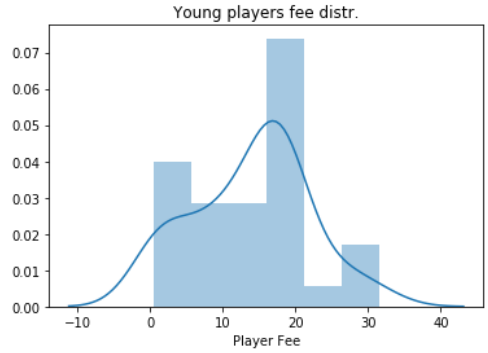
We also saw the difference age has on a players transfer fee. While we don't see a huge difference in players market value and their age, we do see that younger players tend to have a higher transfer fee than older players. The teams buying players seem to put a premium on younger players. If you are a coach who has their eye on an older player you may just be able to get a good deal on them.





While in the plots above and the histograms below it looks like there may be a difference between age and fees, our age analysis does not strongly support that. We saw that players 25 and older had a mean transfer fee of 7.07 million pounds with a large standard deviation of 10.17. Players younger than 25 did have a higher average transfer fee at 10.95 million pounds with a standard deviation of 9.07. The p-value for the t-test, looking at whether the means are significantly different was 0.086. With this p-value and the data we have, we cannot strongly say that younger players have higher fees than older players. We may have our suspicions but it would be wise to get more data on transfers before jumping to conclusions.





Finally, we wanted to see if players who had been transferred recently were talked about more than players who had not. We checked for instances of player names in article titles from 6 large soccer news outlets. We found that most of the players who had recently been transferred didn't even occur once on the news sites. This is likely because there has only been one round of games since the transfer window closed so a lot of recently transferred players haven't made their debut yet. The two players who had multiple mentions of their name were Erling Haaland, who has scored 7 goals in his first 3 games with Dortmund, and Bruno Fernandes, who now plays for one of the largests clubs in the world with Manchester United.