Hi. My name's Ryan, and this blog demonstrates some of the projects I've worked on. These projects ideas mainly came about with a desire to improve my coding and statistical knowledge, but hopefully along the way some interesting conclusions can be found.

Predicting Football Player's Transfer Fee Within Europe's 3 Top Leagues

October 15, 2020

Transfer fees in football have often been the subject of derision by the general public due to vast quantities of money clubs are more than willing to spend. However, even those who accept these lofty prices are sometimes puzzled by the price tag of certain players. Was Harry Maguire really worth £80 million? Neymar, £200 million? How about David Luiz, who throughout his career clubs have forked out over £100 million for?

This project aims to investigate what are the factors that go into the transfer fee of a football player. This is done by taking the statistics of players, for example the amount of clearances per game, goals per game, statistics of that form, and seeing how they relate with the transfer fee.

This is what many people do with football in order to justify the cost of a newly bought unknown player. The first port of call is often to look at their career stats on Wikipedia and mentally work out a goals per game ratio, if the new player was a striker, to see if that justifies the transfer fee. In this process, we are basically saying "I think goals per game is the most important factor of what dictates the transfer fee, so let me see if the player is really worth what we bought him for by looking at his goals per game".

In essence, this project emulates that process, except using more stats than goals per game – we'll be working with over 100 different types offensive (such as goals) and defensive (such as tackles) statistics. This means we can look at more stats, other than goals per game, and form an opinion on a player's transfer fee. What we're doing is finding the statistics that historically look to be related with the price of a player. This is done by taking players' stats over a long period of time, and also their transfer fees, to see if there is a relationship between the two.

Maybe for Centre Backs, a defensive position, we'll see that the transfer fee is highly related to the

SHARE

Labels

Analysis
Football
Statistics

amount of tackles per game, and then if a club buys a new player, we can find the stat for tackles per game for the new player and form an opinion on their transfer fee.

So, what are the specific statistics that dictate the transfer fee of the player? That's what we aim to find out.

Method

In order to start this statistical analysis, we need 2 things.

1. The performance statistics of players across a large time frame.
2. The fees of players across a large time frame.

First we begin with the performance statistics.

All performance statistics will be taken from whoscored.com.

(https://www.whoscored.com/Players/97752/History/Paul-Pogba -> History -> Detailed).

Teams from the top leagues in England, Spain and Italy will be used. A list of each player from each team will be generated (so a squad list as of July 2020).

The data for each of those players will then be taken (web scraped using python) from the whoscored website and all statistics will be used on a "per game" basis, for example goals per game, saves per game, tackles per game.

How do we get the fees?

Transfer fees will be taken from transfermarkt.com (again, web scraped using python).

That's pretty straight forward, however there is the question of inflation.

Cristiano Ronaldo was sold for £80 million but if he was that age right now and with his performance stats, he'd go for much more than that.

We need a fair comparison between the transfer fees we see now, in 2020, and the fees we saw in the past.

In order for inflation to not skew the results of the analysis, inflation is adjusted for. This was done using a football inflation index which was created by totallymoney.com (https://www.totallymoney.com/content/transfer-index/).

Adjusting the transfer fees of the past, we can calculate what the fee would be in today's money.

Using the inflation rate, Ronaldo's £80 million transfer would be £200 million in today's money, still pretty cheap in my opinion.

Now we investigate the relationship between the two pieces of data.

We'll be taking all the performance data of each player before their transfer, and averaging it out through the years. Each player will then have his career performance stats up until point of transfer. We'll then investigate the relationship between these stats and the transfer fee using different types of statistical analysis.

How we do we do this?

The data consists of 88 independent variables, of which around 5 to 10 variables will be selected to enter an OLS regression.

We use 3 different types of feature selection (selection of variables) to whittle the variables down.

1. Correlation matrix
2. Lasso Regression
3. Random Forest

Of the 3 methods, the variables with the strongest relationship with the transfer fee (outcome) will be considered.

It involves considering how strongly they are related, and also selecting variables such that the variables are not correlated with each other. For example, if "number of goals per game" and "number of goals inside the penalty area per game" are both selected by the above methodology, the most suitable variable will be chosen. Avoiding this correlation is a way to satisfy the assumptions of the OLS.

Outcomes of these variable selection can be seen in the appendix.

OLS post diagnostics

After each regression the validity of the OLS assumptions will be tested and adjusted depending on the test results using the following methods:

1. **Linearity between predictors and outcome**: predicted values vs actual values plotted to see if they lie on a diagonal line
2. **Normality of error terms:** Anderson-Darling test and histogram plot
3. **Multicollinearity amongst predictors:** Variance inflation factor, values over 10 to be removed from analysis. Correlation heatmap also used.
4. **No autocorrelation of error terms:** Durban-Watson Test
5. **Homoscedasticity:** Residuals plot to see if variance appears uniform

These diagnostics can be seen in the appendix.

The Data

The performance data as stated previously is taken from the whoscored.com website. Each stat is averaged out on a per game basis from the players' whole career before the data of transfer.

1530 players have been web scraped from the top 3 leagues in Europe. Players who have had a transfer in the last 15 years will be included in the analysis, including players with more than 1 transfer.

To get an impression of what the data looks like, here is an example of what the data looks like, showing "goalTotal" (amount of goals scored per 90 minutes) by Cristiano Ronaldo. In this year he was transferred to Juventus.

Name	Position	Height	Nationality	Games_Played_Season	goalTotal	Inf_price
Cristiano-Ronaldo	Midfielder (Left)	187	Portugal	438.5	0.969213227	119,340,000

In total there are 88 variables that will go into predicting the transfer value.

Data considerations

Groups

To be considered for the data analysis the player must have played on average at least 5 games a season. This is to avoid any anomalies of data that may come about by playing a low number of games.

Players will be grouped by their playing position since different positions require different attributes which the game determines as valuable. Each player can have multiple positions so it's necessary to state how a player's position is defined.

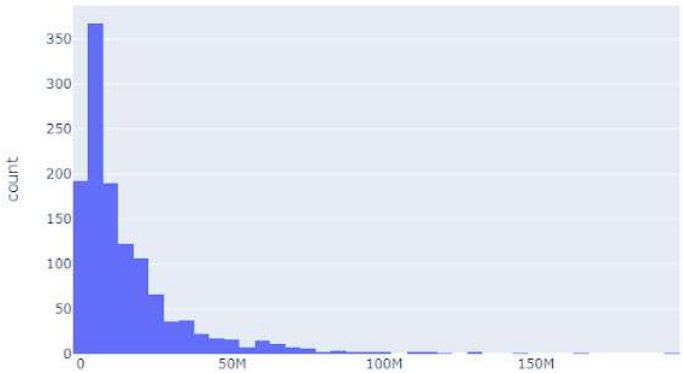
A regression will be done on each of these positions.

These positions are defined as follows:

1. **Goalkeeper**.
2. **Centre Back** - the player can only have the position listed as centre back and not any other position.
3. **Full back** - the player can have any other position.
4. **Defensive midfielder** - the player's position is Defensive midfielder and cannot have any other position in order to prevent more attacking statistics influencing the analysis.
5. **Centre midfielder** - cannot have forward, left, right or defensive in their positions. This to prevent valued statistics that represent other positions.
6. **Winger** - Must contain left or right in their position title but not forward
7. **Forward** - Contains forward in their position and any other position is allowed

Transforming the outcome variable

The outcome (dependent variable) transfer fee variable is seen to not be normally distributed which violates the assumption of the OLS.



Inf_price

For this reasons the transfer fee is log transformed in order to satisfy this assumption and used in the regression.

Results

Centre Back

n= 166
R-squared = 0.275

Variable	Variable meaning	Coefficient	Percent increase of transfer fee	P value
shotsTotal	Total amount of shots	0.80	123%	0.03
shortPassAccurate	Accurate short passes	0.01	1%	0.001
goalPenaltyArea	Goals inside penalty area	2.86	1642%	0.038
foulCommitted	Fouls Committed	-0.35	-30%	0.061
passLongBallInaccurate	Inaccurate long balls	-0.20	-18%	0.231
dispossessed	Player dispossesed by opponent	1.23	242%	0.103
Games Played Per Season	Games player per season average	0.03	3%	0.276
Games_Played_Season	Total games played up until transfer	0.00	0%	0.008

For all regressions games played person and total games played up until transfer are put into the regression in order to take out the effect of these variables - it is obvious that the amount of games a player plays is dictates by their perceived value.

The way to read this table is by looking at a variable and seeing what percent increase of transfer fee it's associated with. For example, for shotsTotal, for an increase in 1 unit of shots taken per 90 minutes, all other variables being the same, the player's transfer fee increases by 123%. Because we are often working with statistics that are under the value of 1 (such as 0.5 goals per 90 minutes) the percentage increase due to an increase in 1 in the statistic (e.g. 0.5 goals per game to 1.5 goals per game) we see in value will be large, because 0.5 to 1.5 goals per game is a huge difference.

The p value tells us how strong the relationship of the variable is with the transfer fee. Some variables in this analysis come out not as significant (<0.05) and in my commentary I try to touch on the significant variables more.

We can see for centre backs goals inside the penalty area increases their value by 1642% for 1 unit increase per 90 minutes. Strangley, the frequency of being dispossessed, a bad thing, actually relates to a higher transfer fee. Maybe this equates to more time on the ball, which can be a positive trait for a centre back. Not committing fouls and making accurate long balls is important for a centre back.

Full back

n= 284
R-squared = 0.155

Variable	Variable meaning	Coefficient	Percent increase of transfer fee	P value
penaltyTaken	Penalty taken	12.18	19518339%	0.056
dribbleWon	Successful dribbles	0.60	83%	0
assistThroughball	Through ball assists	5.85	34589%	0.519
keyPassShort	Key pass short	0.17	18%	0.587
dispossessed	Player dispossesed by opponent	-0.18	-17%	0.403
shotCounter	Shots on counter attack	-1.91	-85%	0.268
Games Played Per Season	Games player per season average	0.04	4%	0
Games_Played_Season	Total games played up until transfer	0.00	0%	0.692

Taking a penalty means a full back is probably more technical on the ball and a leader type so shows up in this regression. Assists and dribbles are more important for a full back.

Defensive Midfielder

n= 122
R-squared = 0.293

Variable	Variable meaning	Coefficient	Percent increase of transfer fee	P value
interceptionAll	Interceptions	0.77	116%	0.00
shortPassAccurate	Accurate short passes	0.02	2%	0.02
outfielderBlockedPass	Blocked passes	-0.45	-37%	0.06
shotOpenPlay	Shots from open play	0.37	45%	0.32
penaltyTaken	Pentalty taken	9.80	1797153%	0.09
dribbleWon	Successful dribbles	0.72	104%	0.00
goalRightFoot	Goals with right foot	-1.53	-78%	0.67
foulCommitted	Fouls committed	0.01	1%	0.96
shortPassInaccurate	Inaccurate short pass	-0.03	-3%	0.63
Games Played Per Season	Games played person average	0.00	0%	0.91
Games_Played_Season	Total games played up until transfer	0.00	0%	0.71

For a defensive midfielder dribbles and interceptions are important, but strangely blocked passes and goals means the player is actually worth less. The blocks could be due to a defensive midfielder of not great quality playing for a lesser team, therefore having to do more blocks.

Winger

n= 364
R-squared = 0.246

Variable	Variable meaning	Coefficient	Percent increase of transfer fee	P value
dribbleWon	Successful dribbles	0.40	50%	0.00
assistThroughball	Through ball assists	8.05	314252%	0.01
shotCounter	Shots on counter attack	0.59	81%	0.57
shortPassAccurate	Accurate short passes	0.04	4%	0.00
goalObox	Goals outside the box	-2.23	-89%	0.20
keyPassShort	Short key passes	0.12	12%	0.54
Games Played Per Season	Games played person average	0.03	3%	0.00
Games_Played_Season	Total games played up until transfer	0.00	0%	0.46

For a winger, dribbling and assisting is very important, also short key passes.

Centre Midfielder

n= 113
R-squared = 0.284

Variable	Variable meaning	Coefficient	Percent increase of transfer fee	P value
keyPassThroughball	Through ball key passes	4.38	7852%	0.02
shotsTotal	All types of shot	0.63	87%	0.00
shortPassAccurate	Accurate short passes	0.03	3%	0.04
assist	Assists	-3.55	-97%	0.03
penaltyTaken	Pentalty taken	-9.97	-100%	0.02
goalPenaltyArea	Goals inside the penalty area	3.35	2752%	0.10
Games Played Per Season	Games played person average	0.00	0%	0.77
Games_Played_Season	Total games played up until transfer	0.00	0%	0.46

The central midfielder must make a lot of key through ball passes and take a good amount of shots in order to be valuable. Making assists is actually related to a lower value, this quirk is potentially due to the low base in this regression.

Forward

Forward
n= 331
R-squared = 0.33

Variable	Variable meaning	Coefficient	Percent increase of transfer fee	P value
shotsTotal	All types of shot	0.41	50%	0.00
Games_Played_Season	Total games played up until transfer	0.00	0%	0.00
goalTotal	All types of goal	1.19	228%	0.02
Games Played Per Season	Games played person average	0.01	1%	0.14
keyPassShort	Short key passes	0.35	41%	0.04
shotSetPiece	Shots from set piece	-0.21	-19%	0.44

penaltyTaken	Pentalty Taken	-1.01	-64%	0.48
shortPassAccurate	Accurate Short Passes	0.03	3%	0.00
assistOther	Assists that are not from corner, cross, throughball	1.86	540%	0.07

The forward, as expected, is required to score goals and take shots.

Keeper

n= 76
R-squared = 0.332

Variable	Variable meaning	Coefficient	Percent increase of transfer fee	P value
clearanceTotal	All types of cleanrace	-1.15	-68%	0.00
Games Played Per Season	Games played person average	0.03	3%	0.01
Games_Played_Season	Total games played up until transfer	0.00	0%	0.66
passFreekickInaccurate	Innacurate freekick	-0.29	-25%	0.13
challengeLost	Lost challenges (challenge with another player)	-3.65	-97%	0.05
redCard	Red cards	-4.27	-99%	0.77
saveObox	Saves outside of the box	0.11	12%	0.76
savePenaltyArea	Saves inside penalty area	-0.03	-3%	0.94
outfielderBlockedPass	Blocked pass	4.89	13231%	0.19

Clearances are related with lower transfer values for keepers, so is losing challenges. This, like defensive midfielders, is probably due to lower quality keepers playing for lower quality teams therefore having to make a lot more clearances.

Conclusion

Despite having a player base of 1515 players the analysis could benefit from having more players, since of those players a transfer must occur and we subset these players into different positions.

Some of the results are the opposite of what we expect. For example, the amount of clearances a keeper makes decreases their value. Or perhaps we do expect that, but then the solution has to be we must take into account into the analysis players playing for lesser teams, especially defensive players. Lower quality defensive players playing for poor teams are probably going to have high defensive stats in clearances and tackles due to them being attacked a lot. If this analysis was to be done again, perhaps the players' average team points should be entered into the regression to account for this.

The analysis itself probably' isn't as insightful as was hoped. Predicting players' values through statistics is a known difficulty in the football world, and the type of data I scraped probably isn't powerful enough to get a solid conclusion.

Doing this analysis, it became apparent that a classification analysis would be better - essentially describing what are the traits of each position. Whether this would unveil something we don't already know though is in question.

The project itself has been beneficial, forcing me to learn python, web scraping and a slew of statistical techniques such as Random Forest, Lasso regression and regression post-diagnostics.

The appendix is yet to be added but I plan to do that on a future date.

Ryan Pollard - 2020

LABELS: ANALYSIS, FOOTBALL, STATISTICS SHARE

Comments

HH · October 21, 2020 at 12:25 PM

Excellent work Ryan

REPLY DELETE

Enter your comment...

Popular posts from this blog

Age ain't nothing but a statistic

October 22, 2020

I'm now safely settled into my fourth decade, my 30s, a decade seemingly defined by thinking about if nutmeg would be a good addition to porridge and washing dishes almost adequately. Accompanying me in this decade is the heightened awareness of the degradation of the already battered corpse that I reside in. When I now play ...

SHARE 1 COMMENT READ MORE

About Me

Ryan Pollard

BSc Mathematics. Interested in using Statistics to come to interesting conclusions.

VISIT PROFILE

Archive

Labels

Report Abuse