CS 7290: Causal Modelling in Machine Learning

Causality Analysis on transfer prices of Soccer players

1. Introduction and Motivation:

Football or Soccer is the world's largest sport, it has the highest industry share among all sports. It directly addresses 43% of the whole global financial sports market which has an industry value of nearly $600 billion. This staggering amount is due to exorbitantly high TV deals, and increasingly rich owners. Besides goals and silverwares, soccer fans find transfer stories exciting. Transfers involving top players with high market value never failed to hit the headlines. Market value varies greatly for different players, different areas and different periods of time.

Since early 2016 football industry is experiencing an acute form of hyper-inflation. Specifically, we are seeing a classic case of what economists call "demand-pull inflation". Because clubs have more money to spend, you would assume that clubs can now spend on better players, however, that's not necessarily true. While there is an increase in money supply, there is no increase in the supply of top quality, world-class footballers. Because of this, top class footballers are now worth even more, resulting in "demand-pull inflation". Demand-pull inflation is when aggregate demand outpaces aggregate supply in the market. Clubs have more available funds, as business spending increase (an increase in aggregate demand), but there is no increase in aggregate supply or world class players. Football has no governing body that monitors this ever-growing inflation, this could be good or bad, depending on how you see it. Its good cause clubs make more and more money, but things could turn to the worse as the model is not sustainable.

In the world of soccer, a German website, transfermarkt.de, is the authority in judging market value of soccer players. This website records detailed information for major soccer players and evaluates their value based on data analysis, as well as opinions of experts. The values are not obtained by applying straightforward algorithms. Factors from all aspects are taken into considerations to decide the digits of a market value. There are many models out there which

predict the market value of players based on many variables, but it is a rare sight in the sport to see a good player bought at the market price. Over the past 5 years the inflation has increased the difference between market value and actual transfer price. This inflated market has led clubs into taking innovative transfer strategies which differ from the traditional ones, clubs have started investing in young players and their potential, in order to avoid an even more inflated rate in the future.

Our motivation is to model this inflation in prices and try to accurately classify players in price brackets. We want to model the negotiated price of these players based on many variables. We believe the prototype model we build could be eventually used to help teams understand the market better and find undervalued players.

2. Initial Assumptions about the model:

For this project, our target variable is the Real-World transfer fee (T) of players who have been involved in an inter-club transfer from 2008 - 2019. Given our domain knowledge about soccer, the following is a list of factors that we believe are the most influential towards T:

- *A = Player's Age*
- *N = Player's Nationality*
- *Pos = Position*
- *OC = Origin Club (The club that the player transferred out of/from)*
- *OL = Origin League (The league in which the Origin Club participates)*
- *AC = Arrival Club (The club that the player transferred to)*
- *AL = Arrival League (The league in which the Arrival Club participates)*
- *Ovr = Player's Overall Rating in the previous year*
- *Pot = Player's Potential Rating in the previous year*
- *Y = Year (Year of Transfer)*
- *T = Transfer Price*

From our initial assumptions of the model, the following would be the list of (causal) paths:

- Age → Transfer Price
- Nationality → Transfer Price
- Year of Transfer → Transfer Price
- Position → Overall Rating → Transfer Price
- Position → Potential Rating → Transfer Price
- Origin League → Origin Club → Transfer Price
- Arrival League → Arrival Club → Transfer Price

Given these variables, the following are our initial assumptions about the model, expressed as a DAG using bnlearn:

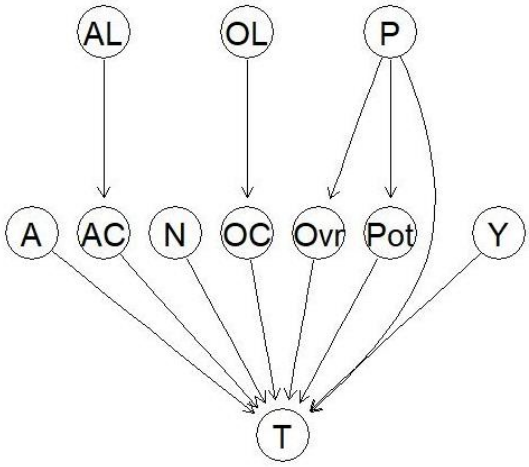


Fig 1: Initial DAG encoding assumptions about the model in bnlearn

To test the conditional independencies that exist in our initial assumptions of the model, we transferred the DAG to Causal Fusion:

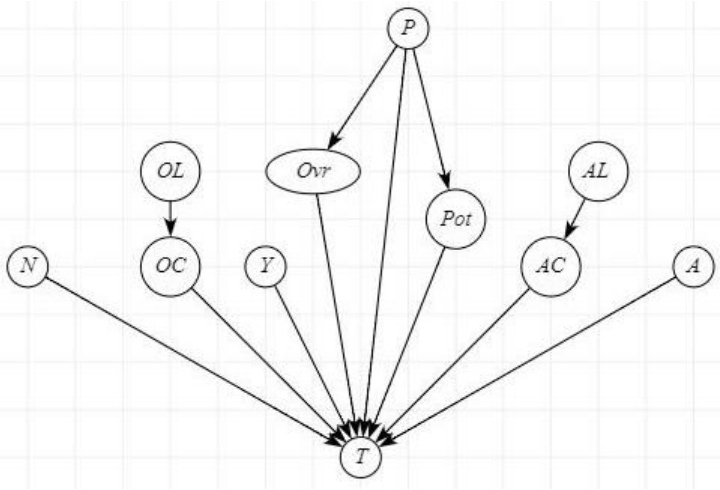


Fig 2: Initial DAG encoding assumptions about the model in Causal Fusion

Now that we have an initial model with our assumptions, the next step is to look at the D Separation Statements in the DAG and Conditional Independence Statements in the Data and try to understand how this model holds up when we look at the Global Markov Property Assumption and the Faithfulness Assumption.

3. D Separation and Conditional Independence:

The following is the list of conditional independencies that Causal Fusion generates under Testable Implications.
These CIs have been color coded based on our domain knowledge about soccer, without looking at the data:

- Green = True
- Red = False (Variables are already independent),
- Yellow = Not sure
- Orange = False (Unobserved Confounder detected)

$A \perp\!\!\!\perp AC$	$AC \perp\!\!\!\perp N$	$A\ L \perp\!\!\!\perp Pos$	$N \perp\!\!\!\perp O\ C$	$OL \perp\!\!\!\perp Pot$
$A \perp\!\!\!\perp AL$	$OC \perp\!\!\!\perp Pot \mid OL$ $OC \perp\!\!\!\perp Ovr \mid OL$	$AL \perp\!\!\!\perp Ovr$	$N \perp\!\!\!\perp O\ L$	$OL \perp\!\!\!\perp T \mid A,AC,N,OC,Ovr,P,Pot,Y$
$A \perp\!\!\!\perp Ovr$	$AC \perp\!\!\!\perp OL$	$AL \perp\!\!\!\perp Y$	$O\ L \perp\!\!\!\perp Y$	$OL \perp\!\!\!\perp Ovr$
$A \perp\!\!\!\perp N$	$AC \perp\!\!\!\perp Ovr \mid P$	$AL \perp\!\!\!\perp OL$	$N \perp\!\!\!\perp Pos$	$Ovr \perp\!\!\!\perp Pot \mid P, A$
$A \perp\!\!\!\perp OC$	$A\ C \perp\!\!\!\perp P$	$AL \perp\!\!\!\perp Pot$	$N \perp\!\!\!\perp Y$	$Ovr \perp\!\!\!\perp Y$
$A \perp\!\!\!\perp OL$	$A\ C \perp\!\!\!\perp Y$	$N \perp\!\!\!\perp Ovr$	$Pos \perp\!\!\!\perp Y$	$AL \perp\!\!\!\perp T \mid A,AC,N,OC,Ovr,P,Pot,Y$
$A \perp\!\!\!\perp Pot$	$AC \perp\!\!\!\perp Pot \mid P$	$N \perp\!\!\!\perp Pot$	$OC \perp\!\!\!\perp P$	$Pot \perp\!\!\!\perp Y$
$A \perp\!\!\!\perp P$	$A\ L \perp\!\!\!\perp N$	$OC \perp\!\!\!\perp Ovr \mid P$	$OC \perp\!\!\!\perp Y$	$AC \perp\!\!\!\perp Pot \mid AL$
$A \perp\!\!\!\perp Y$	$AL \perp\!\!\!\perp OC$	$OC \perp\!\!\!\perp Pot \mid P$	$O\ L \perp\!\!\!\perp P$	$AC \perp\!\!\!\perp Ovr \mid AL$

Conditional Independencies in the Initial DAG that are false, as there was evidence of an unobserved confounder or direct influence:

- *$A \perp\!\!\!\perp A\ C$: Age is conditionally independent of Arrival Club.*
  The Arrival Club could have a good number of players that are either inexperienced and young or players nearing retirement, which would mean that the club could be seeking players ONLY in a particular Age group.

- *$A\ C \perp\!\!\!\perp P$: Arrival Club is conditionally independent of Position*
  The Arrival Club could have certain positions that they want to fill considerably more than others and could be seeking ONLY players who play in a specific position.

- *$A\ C \perp\!\!\!\perp Y$: Arrival Club is conditionally independent of Year of Transfer*
  The Arrival Club could see a change in management/coach or could receive an influx of investment to ramp up player purchase in a certain season/year.

- *$A \perp\!\!\!\perp Ovr$: Age is conditionally independent of Overall Rating*
  This is not true as domain knowledge suggests Age influences Overall Rating.

- *A ⫫ Pot: Age is conditionally independent of Potential Rating*
  This is not true as domain knowledge suggests Age influences Potential Rating.

Furthermore, we decided to add one more variable in the model called Appearances, i.e. the total number of matches the player has played up until the year of transfer. This gives an idea about how reliable the player is for the club he plays for and whether the player is injury prone or not.

Addition of unobserved confounders, and directed edges gives us the following DAG:

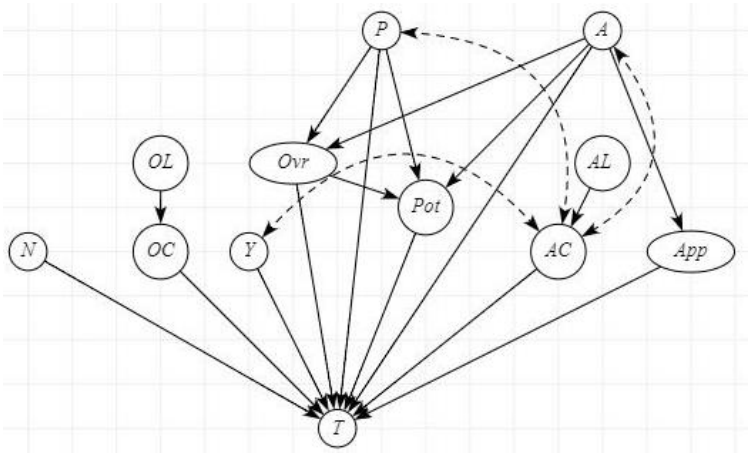


Fig. 3 New DAG on addition of unobserved confounders and directed edges

Now that we have a final DAG, we move on to the D Separation tests and the Conditional Independence tests.

3.1 D Separation tests

D Sep tests performed manually, resulting in the answer as TRUE for all sets in consideration.

3.2 CI tests using Causal Fusion

CI tests on Causal fusion fail with the following message.

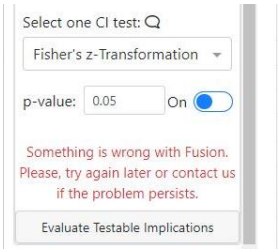


Fig 4. Causal Fusion CI tests error

3.3 CI tests using Bnlearn

CI tests were done in bnlearn by keeping the p-value = 0.05

Testable Implication	**D Separation in DAG**	**CI in data**
A ⫫ AL	*TRUE*	*FALSE - p-value = 6.916e-05*
A ⫫ N	*TRUE*	*FALSE - p-value = 0.02091*
A ⫫ OC	*TRUE*	*FALSE - p-value < 2.2e-16*
A ⫫ OL	*TRUE*	*FALSE - p-value < 2.2e-16*
A ⫫ P	*TRUE*	*FALSE - p-value = 1.151e-05*
A ⫫ Y	*TRUE*	*FALSE - p-value = 0.03622*
AC ⫫ N	*TRUE*	*FALSE - p-value = 0.03109*
AC ⫫ OC | OL	*TRUE (AC ⫫ OC)*	*FALSE - p-value = 0.01459*
AC ⫫ OL	*TRUE*	*FALSE - p-value = 0.009141*
AC ⫫ Ovr | A,P	*TRUE*	*TRUE - p-value = 0.9918*
AC ⫫ Pot | A,P	*TRUE*	*TRUE - p-value = 0.9309*
AL ⫫ N	*TRUE*	*FALSE - p-value = 2.906e-15*
AL ⫫ OC	*TRUE*	*FALSE - p-value = 6.89e-13*
AL ⫫ OL	*TRUE*	*FALSE - p-value < 2.2e-16*
AL ⫫ Ovr	*TRUE*	*FALSE - p-value = 0.001799*
AL ⫫ P	*TRUE*	*TRUE - p-value = 0.7554*
AL ⫫ Pot	*TRUE*	*FALSE - p-value = 0.0004557*
AL ⫫ Y	*TRUE*	*FALSE - p-value = 6.102e-07*
N ⫫ OC	*TRUE*	*FALSE - p-value = 1.449e-13*
N ⫫ OL	*TRUE*	*FALSE - p-value < 2.2e-16*
N ⫫ Ovr	*TRUE*	*TRUE - p-value = 0.09193*

N ⫫ P	*TRUE*	*FALSE - p-value = 1.007e-06*
N ⫫ Pot	*TRUE*	*TRUE - p-value = 0.4972*
N ⫫ Y	*TRUE*	*TRUE - p-value = 0.5188*
OC ⫫ Ovr | OL	*TRUE (OC ⫫ Ovr)*	*FALSE - p-value < 2.2e-16*
OC ⫫ P	*TRUE*	*TRUE - p-value = 0.7888*
OC ⫫ Pot | OL	*TRUE (OC ⫫ Pot)*	*FALSE - p-value = 4.214e-13*
OC ⫫ Y	*TRUE*	*FALSE - p-value = 0.01695*
OL ⫫ Ovr	*TRUE*	*FALSE - p-value < 2.2e-16*
OL ⫫ P	*TRUE*	*TRUE - p-value = 0.9705*
OL ⫫ Pot	*TRUE*	*FALSE - p-value = 2.658e-07*
OL ⫫ T | A,AC,App,N,OC,Ovr,P,Pot,Y	*TRUE (OL ⫫ T | OC)*	*TRUE - p-value = 1*
OL ⫫ Y	*TRUE*	*FALSE - p-value = 0.03277*
Ovr ⫫ Y	*TRUE*	*TRUE - p-value = 0.148*
P ⫫ Y	*TRUE*	*FALSE - p-value = 0.02751*
Pot ⫫ Y	*TRUE*	*TRUE - p-value = 0.4934*
AL ⫫ App	*TRUE*	*FALSE - p-value = 7.998e-05*
AC ⫫ App | A	*TRUE*	*TRUE - p-value = 0.8463*
AL ⫫ T | A,AC,App,N,OC,Ovr,P,Pot,Y	*TRUE (AL ⫫ T | AC)*	*TRUE - p-value = 1*
App ⫫ N	*TRUE*	*TRUE - p-value = 0.3461*
App ⫫ OC | OL	*TRUE (App ⫫ OC)*	*FALSE - p-value = 4.205e-06*
App ⫫ OL	*TRUE*	*FALSE - p-value = 2.637e-15*
App ⫫ Ovr | A	*TRUE*	*FALSE - p-value = 1.942e-15*
App ⫫ P	*TRUE*	*FALSE - p-value = 0.002604*
App ⫫ Pot | A	*TRUE*	*TRUE - p-value = 0.9131*
App ⫫ Y	*TRUE*	*FALSE - p-value = 0.001267*

Table 1. Results of D Sep and CI tests on the Testable Implications Set from Causal Fusion

3.4 Summary of D Separation and CI tests

If we only consider the testable implications set from Causal Fusion, here are the results*:*

- *% of True CI statements that are also True D Sep statements =* ***33%***
- *% of True D Sep statements that are also True CI statements =* ***100%***

Now, if we look at all possible cases including redundant statements:

- *% of True CI statements that are also True D Sep statements =* ***49%***
- *% of True D Sep statements that are also True CI statements =* ***95%***

In both the above cases, we can confidently say that our model does very well on the Faithfulness Assumption!

We believe that by doing a more detailed exploratory analysis into more cases of unobserved confounders, we'll be able to make our model perform better in terms of the Global Markov Property assumption.

4. Data Overview:

The original data was pulled from the Github repo: https://github.com/ewenme/transfers.

Data includes selected variables for each of the following national soccer leagues:

- English Premier League
- English Championship(excluded because it is a 2nd division league)
- French Ligue 1
- German 1.Bundesliga
- Italian Serie A
- Spanish La Liga
- Portugese Liga NOS
- Dutch Eredivisie
- Russian Premier Liga

Common variables:

- club_name (club: Arrival_club)
- player_name (player name)
- age (player age at time of scrape)
- position (player position)
- club_involved_name (other club involved in transfer: origin club)
- fee (raw transfer fee information)

- transfer_movement (transfer in/out)
- fee_cleaned (numeric fee, GBP)
- league_name (league: Arrival_league)
- year (year)
- season (season, interpolated from year)

The data was scraped from transfermarkt.com, which is a reliable source for market value of football players. The goal behind the causal analysis is to model the actual transfer price of the players, which is usually relatively higher than their market value.

4.1 Pre-processing Variables

Since we are working on an idea which has very little literature, we understood that it is imperative to spend a lot of time studying and processing our variables based on intuition and data trends.

4.1.1 Leagues

The data on leagues is very important to our model, the top 5 leagues we are concentrating on are

- English Premier League
- French Ligue 1
- German 1.Bundesliga
- Italian Serie A
- Spanish La Liga

The decision to only concentrate on these leagues is made using intuition and data. Being avid football fans, we have seen teams from these leagues perform on a different level in comparison to other leagues, our data also shows the transfer movement and transfer amounts spent by these leagues is much higher in comparison to the other leagues in Europe. All the data on other European teams has been classified into the term ‘Others’.

league	transfer_price
Premier League	10503.06
Primera Division	4883.35
Serie A	4720.34
1 Bundesliga	2902.72
Ligue 1	2747.35
Other	1820.30

However the decision to only use these leagues also hinged on the fact that the more categories we had the more CPT tables we ended up building, the number of CPT tables was not the issue, we were worried the more noise we have the harder it would be to find a causal signal in our dataset.

Final categorization for leagues:

- English Premier League (Premier league)
- French Ligue 1 (Ligue1)
- German 1.Bundesliga (1.Bundesliga)
- Italian Serie A (Serie A)
- Spanish La Liga (Premiera Division)
- Others (Portugese Liga NOS, Dutch Eredivisie, Russian Premier Liga)

Now another issue we faced at this point was we did not have data on the league of the club the player was transferring from**,** we call this variable origin_league throughout the project. The term origin_league signifies which league the player was transferred from, having data for this column was important cause this will let us analyze transfer movement between leagues and how it affected the price of the players. The data from the Github repository only had data on the arrival_league**,** which was data on the league the player got transferred into. We then used the data on arrival_club ,arival_league and origin_club to generate data for origin_league**.** Since we already had a list of clubs and leagues we just had to use that data to generate data on the origin_league. However, the names of the clubs were not standardized, which meant we had multiple name versions for each club, so we then resorted to using the fuzzywuzzy python package which uses Levenshtein Distance to calculate the differences between sequences. Using a threshold of 80 for partial ratio we generated values for origin_league, which we then manually checked to correct anomalies. This completed the preprocessing of the variable's origin_league and arrival_league.

Notebook - generating_from_league.ipynb

4.1.2 Clubs

Our data had roughly 300 different unique value for clubs. It was very important to categorize these unique values into different tiers to try and understand what kind of influence the size of a club has on the transfer price of players. There is a lack of rules surrounding transfers in the world of soccer and that is the reason that a lot of small clubs hold richer clubs for ransom over transfer fee of soccer players. We decided to break down clubs based on how much these clubs have spent over the years on transfers. Building a Dataframe on the net spend of clubs we manage to understand our data better we got to see what clubs have been more active in the transfer windows and which clubs have been breaking even. The initial attempt to classify these clubs was based on the net spend of the clubs, net spend was a variable calculated by subtracting transfer fees earned and transfer fees spent. This variable was not very accurate in classifying the clubs as clubs sign different sponsorship deals and some of the clubs are in debts, hence they resort to selling players in order to reach financial stability. So, we decided to classify clubs purely based on their spending power (transfer fees spent) we divided clubs into 4 tiers based on the amount of money they spent from 2009 to 2019.

- Tier 1: Clubs that spent more than 300M Euros
- Tier 2: Clubs that spent more than 100M Euros but less than 300M Euros
- Tier 3: Clubs that spent more than 20M Euros but less than 100M Euros
- Tier 4: Clubs that spent less than 20M Euros

Here is a list of the top 5 clubs:

	club_name	fee_spent	fee_earned	net_spend	Tier
0	Manchester City	1504.897	421.885	1083.012	Tier_1
1	FC Barcelona	1331.170	681.450	649.720	Tier_1
2	Real Madrid	1226.930	701.550	525.380	Tier_1
3	Chelsea FC	1213.760	614.694	599.066	Tier_1
4	Juventus FC	1180.153	690.653	489.500	Tier_1

Using this classification, we managed to encode the variables arrival_club and origin_club.

Notebook - encoding_club_tier.ipynb

4.1.3 Positions

The position the player play's quotes the price range he deserves. Forwards have been traditionally expensive ever since the dawn of the sport, they give you goals, and goals win you matches. But this tradition has transformed in the past decade clubs started spending large amounts of money on midfielders, defenders and goalkeepers, they have bought into the saying that 'good teams win games.

position	transfer_price	counts	average_transferprice
LW	3906.297	521	7.497691
RW	3123.082	473	6.602710
CF	8307.095	1299	6.394992
CAM	2464.467	398	6.192128
CM	4496.234	789	5.698649
CDM	2873.077	537	5.350236
CB	5502.631	1105	4.979757
RB	1625.949	382	4.256411
LB	1824.265	432	4.222836
GK	1400.728	376	3.725340

CF – Centre forwards
CB – Centre backs
CM – Central Midfielders
LW – Left wingers
RW – Right wingers
CDM – Centre defensive midfielders
CAM – Centre attacking midfielders
LB – Left back
RB – Right back
GK – Goalkeeper

The above table shows us that left wingers have demanded more money on an average, however centre forwards are the most transferred position followed closely by centre backs.

We initial had great designs in place to classify positions and the above table was built after classifying the noisy position data we initially had, but we decided to keep the classification simple as more the number of categories the more possible combinations the model will generate.

Position classification:

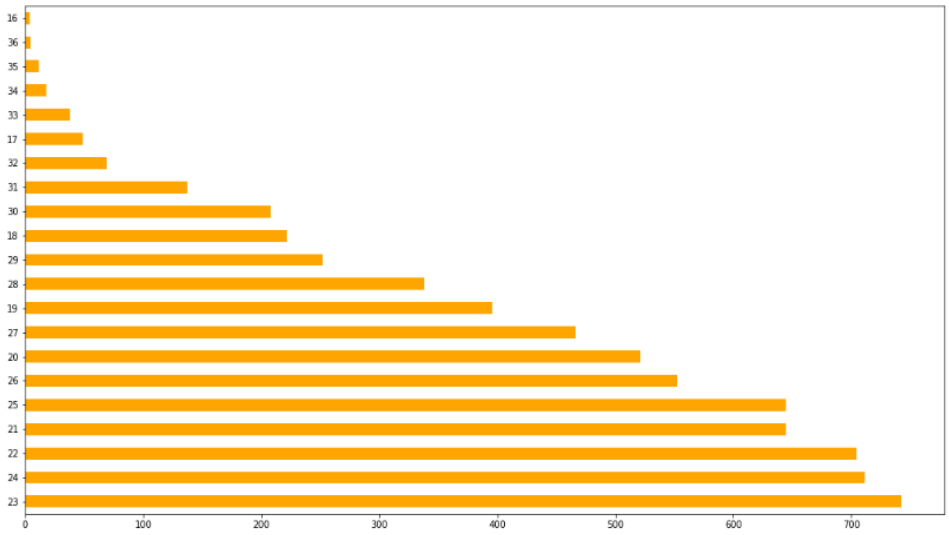
- F – CF, LW, RW
- M – CM, CDM, CAM
- D – CB, LB, RB
- GK

Notebook - categorizing_positions.ipynb, final_preprocessing.ipynb(has the final classification)

4.1.4 Age

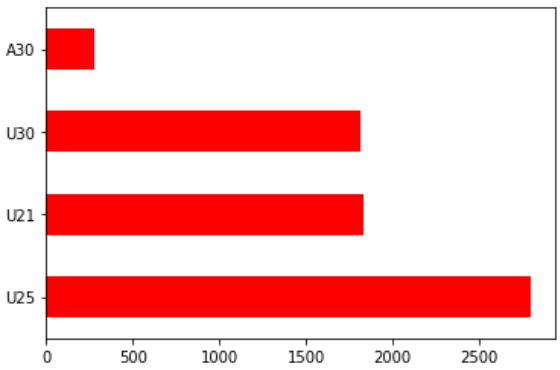
The career life of a soccer player spans from 16-35 on an average, and on an average a player peaks between the age of 27-31, this is when the player has a good combination of fitness and experience. Over the years players in their peak have demanded large sums of transfer prices, but this dynamic has shifted in the last decade , teams are willing to take a risk on young players, the transfer of Portuguese teenage sensation João Félix for a sum 138.60 million euros is a prime example, at the age of 20 he holds the record for the 4th highest transfer fees ever paid for a player in soccer history. This trend does not seem to be slowing down, over the last 10 years the transfer record has been broken many times. The age of a player has increasingly become a key factor in the transfer price, and we want to model it efficiently, for this very reason we decided to build two different classifications.

Visualizing data



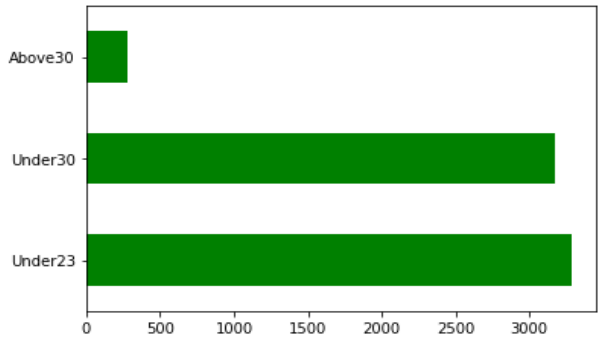
Age grouping 1:

- U21: 16-21
- U25: 21-25
- U30: 25-30
- A30: 30+



Age grouping 2:

- Under23: 16-23
- Under30: 23-30
- Above30: 30+



Notebook - Categorizing_age.ipynb

4.1.5 Year

Time becomes a very important variable in our data because we want to capture the changing trends in transfer prices. You can never accurately pinpoint when the inflation of prices started, cause a lot of other exogenous variables influence that, but through intuition we firmly believe that in 2016 the world record transfer of 23 year old French midfielder Paul Pogba for a fee of 105 million euros was the beginning of the inflation. There are several reasons we believe this intuition, one of them being that the previous highest transfer fee for a midfielder was 77.5 million euros and that player was Zinedine Zidane who was then already a world cup winner, euros winner and champions league winner, he had three of the highest team honors in footballing history, he was also at the ripe age of 29. All these facts above tell us that there was a shift in the market and teams started valuing players differently in this decade, Manchester united took a bet on a young dynamic French midfielder and this set of a chain reaction in the market as more teams started to taking a chance on the potential of young players. Based on this hunch, we classified the years columns in the dataset to before and after 2016.

4.1.6 Transfer Prices (Target variable)

This is our target variable, so we took extra care classifying it, we took two approaches to classify it based on our data.

Grouping 1

- Above80M: transfers above 80M
- 80Mto40M: transfers between 40M to 80M
- 40Mto10M: transfers between 10M to 40M
- 10Mto1M: transfers between 1M and 10M
- Below1M: transfers below 1M

Grouping 2

- Above60M: transfers worth 60M+
- 60Mto20M: transfers between 20M to 60M
- 20Mto5M: transfers between 5M to 20M
- Below5M: transfers below 5M

However, after building our initial models we realized that the transfer prices below 5M dominated our dataset and hence it was harder to find the real signal. The intuition here is that the transfers below 5M are usually closer to the market value, even if they are not the difference between the market value and transfer price is really small in comparison to the big prices. We wanted to look at the big prices under a microscope and find the cause for the rise, so we decided to remove all the extra noise in our data(transfers below 5M), this greatly reduced our dataset from 6300 rows to 1900 rows. But we believe this dataset would do a better job of capturing the relationships.

Notebook - transfer_price_categorization.ipynb, final_preprocessing.ipynb(final classification)

4.2 Data Augmentation

Commented [JC1]: Shreyans update this add notebooks and links as well

We decide to augment our current dataset with more third-party data, we wanted to scrape for the following variables. Packages used: Selenium, BeautifulSoup

- Nationality
- Height
- Goals (biased for attackers)
- Appearances
- FIFA Overall (year of transfer)
- FIFA Potential (year of transfer)

The two source we scraped from were:

- Wikipedi.org
- fifaindex.com

4.3 Classifying Scraped variables

Notebook - categorizing_scraped_data.ipynb, nationality_categorization notebook

Commented [JC2]: Add notebook here for categorizing

4.3.1 Goals and Appearances (Wikipedia)

This data was scraped from the Wikipedia info box, it had data for every season the player played, we needed to take the cumulative sum of the variables based on the year, so we would get his total number of appearances and goals before the transfer happened. This data could help us understand the players performance better, the appearances along with age can tell us more about the injuries the player faced, if it is compared to the average number of appearances for that age we can tell the number of missed games and if that has an effect on the transfer price(it would be ideal to have data on player injuries, but we did not find a reliable source). The total goals give us a track record of how many goals the player has scored for his age, this metric is biased more towards the goal scorers and we scraped it for experimentation purposes.

Classifying Appearances

- 300 above: 300above
- 150 to 300: 150to300
- 50 to 150: 50to150
- below 50: below50

Classifying goals

- above 100: above100
- 50 to 100: 50to100
- 20 to 50 :20to50
- below 20: below20

4.3.2 Height (Wikipedia)

This would be one of the weaker variables in the model, because height has value only to some positions in football. The positions that have an advantage with height usually are strikers, center backs, goalkeepers, and central defensive midfielders, having said that we have seen average

height players perform well in this position too. Centre backs and keepers on an average tend to be 6 feet.

Groups for height:

- 6.5 feet above: 6.5above
- 6 to 6.5: 6to6.5
- 5.5 to 6: 5.5to6
- below 5.5: below5.5

4.3.3 Nationality (FIFA Database)

Commented [JC3]: Mohit update this

Nationality was turned into a Categorical Distribution using the Python package – pycountry-convert. The countries that are part of the same continent are grouped together resulting in a total of 6 categories:

- AF: Africa
- AS: Asia
- EU: Europe
- N_A: North America
- OC: Oceania
- SA: South America

4.3.4 Overall and Potential Ratings

Overall and Potential ratings are given on a scale on 0 – 100. These ratings were scraped for all players in the FIFA dataset from the year 2008 to 2020. We categorized the ratings into the following four categories:

- below65
- 65to74
- 75to84
- 85above

Merging data notebook - data_merger.ipynb

5. Implementing the Model using pgmpy and Pyro

We learned a Bayesian Network from the dataset using the library Pgmpy. We used this library as a replacement of Bnlearn for easy transference of CPT into Pyro.

5.1 Generating CPTs for variables using pgmpy

First, we created the BayesianModel by defining the nodes and edges of the DAG. The BayesianEstimator was used to learn the Conditional Probability Distribution from the dataset with an equivalent sample size of 10.

5.2 Modelling in Pyro based on DAG and CPTs from pgmpy

Using the package pgmy we were able to generate the CPT's in python, which we serialized and imported into our pyro model. The model was built on the following DAG

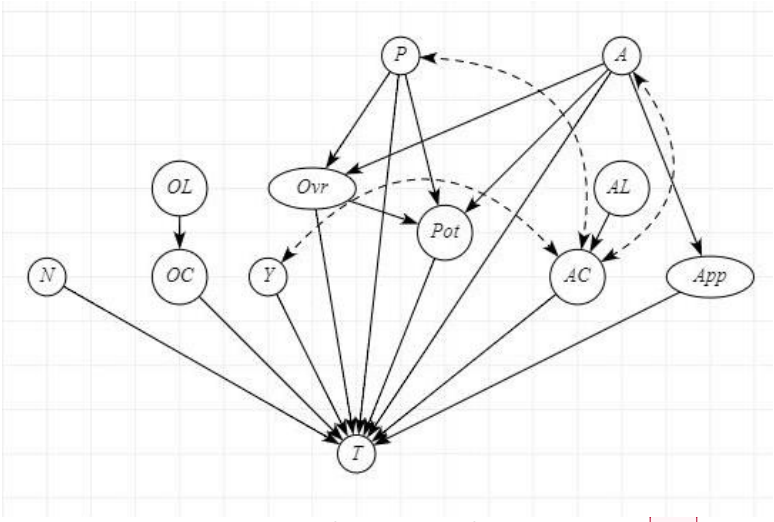


Fig. 3 New DAG on addition of unobserved confounders and directed edges

Commented [JC4]: Snapshot of model

In [25]:
```
def pyro_model():

    Age = pyro.sample("A", dist.Categorical(probs=Age_probs))
    Position = pyro.sample("P", dist.Categorical(probs=Position_probs))
    Nationality = pyro.sample("N", dist.Categorical(probs=Nationality_probs))
    Year = pyro.sample("Y", dist.Categorical(probs=year_probs))
    Arrival_league = pyro.sample("AL", dist.Categorical(probs=arrival_league_probs))
    Origin_league = pyro.sample('OL', dist.Categorical(probs=origin_league_probs))
    Arrival_club = pyro.sample('AC', dist.Categorical(probs=arrival_club_probs[Arrival_league]))
    Origin_club = pyro.sample('OC', dist.Categorical(probs=origin_club_probs[Origin_league]))
    Overall = pyro.sample('Ovr', dist.Categorical(probs=overall_probs[Position][Age]))
    Potential = pyro.sample('Pot',dist.Categorical(probs=potential_probs[Position][Overall][Age]))
    Appearances = pyro.sample('App',dist.Categorical(probs=app_probs[Age]))

    transfer_price = pyro.sample('TP', dist.Categorical(probs=transfer_price_probs[Year][Potential][Positi
on][Overall][Origin_club][Nationality][Appearances][Arrival_club][Age]))

    return{'A': Age,'P': Position,'N': Nationality,'Y': Year,'AL': Arrival_league,'OL':Origin_league,'AC':
Arrival_club,'OC':Origin_club,'Ovr':Overall,'Pot':Potential, 'App':Appearances, 'TP':transfer_price}

print(pyro_model())
```

```
{'A': tensor(1), 'P': tensor(1), 'N': tensor(5), 'Y': tensor(0), 'AL': tensor(5), 'OL': tensor(1), 'AC': t
ensor(0), 'OC': tensor(3), 'Ovr': tensor(0), 'Pot': tensor(1), 'App': tensor(2), 'TP': tensor(0)}
```

6. Experiments and Interventions to test the model

6.1 Intervening on south American forwards

Intuition: We know that South American forwards are coveted by football clubs around the world, from the like of Mardonna and Pele to Neymar and Suarez. We believe that the price of a south american forward is higher than the average forward.

We intervened on the variables Nationality and Position in our model.

- Nationality: South America
- Position: Forward

Testing: We ran a quick analysis on the forwards in our data and calculated the number of transfers that were above 20M, this value stood at 18%. After applying the above intervention, we calculated the number of transfers above 20M the percentage rises to 31%, this result is in line with our intuition.

6.2 Transfer between English teams

Intuition: There is an assumption here that most players transferred in the Premier League (richest league) are English, we believe that the transfer price for English players is abnormally high, due to the requirement of filling up homegrown quota in the league.

We intervened on the variable's arrival league and origin league.

- Origin league: Premier league
- Arrival League: Premier league

Testing: We combed the data and calculated the percentage of transfer above 20M in the entire dataset this value stands at 17.8%. Now after intervention we calculated the percentage of sample transfers above 20M, this value rose to 41.3%, this significant increase in percentage aligns with our intuition.

6.3 Intervening on year to see inflated probabilities for price brackets

Intuition: We believe inflation spiked the value of all transfers after 2016.
We intervened on year in two different models:

- Year: Before 2016 (model 1)
- Year: After 2016 (model 2)

Testing: We subtracted the percentage of transfers above 20M in both the models and saw an increase of 3% in transfers above 20M. This is in line with our intuition however the value is lower than expected.

6.4 Intervening on transfers between tier 1 clubs

Intuition: Players moving from high tier clubs to high tier clubs demand larger transfer sums.
We intervened on the following variables:

- Origin club tier: Tier1
- Arrival club tier: Tier1

Testing: The percentage of transfers that were above 20M in our data was 17.8%, intervening on origin club tier and arrival club tier our new percentage in the samples for transfers above 20M stood at 43.6% this is in line with our intuition.

6.5 Intervening on young and high potential stars to test intuition about our transfer strategy.

Intuition: We believe young players with a high potential fetch a large transfer price.
We intervened on the variables age and potential

- Age: under23
- Potential: 85above

Testing: The percentage of transfers that were above 20M in our data was 17.8%, intervening on Age and Potential our new percentage in the samples for transfers above 20M stood at 59% this is in line with our intuition.

6.6 Causal effect of variables on transfer price

The variables we have analyzed are

- Year: causal effect between before 2016 and after 2016 = 0.006(after – before)
- Potential: causal effect '85 and above' compared to '65 to 74' = 0.029
- Overall: causal effect between '85above' and 'below65' = 0.03
- Age: causal effect between 'above 30'and 'under30' = 0.1
- Arrival club: causal effect between Tier1 and Tier 3 = -0.016
- Origin club: causal effect between Tier1 and Tier 3 = 0.05

1. Counterfactual Analysis

- Calculating probability of transfer price in the range 20M – 60M
- Conditioned on potential being below 65 = 0.29
- Counterfactual query: if potential were above 85 what would the probability to price bracket 20-60M be = 0.38

7. Future Developments

This project could be improved in several ways:

- Accounting for Unobserved Confounders:

  Player contracts - Often good players on the last year of contract go for a relatively small price even during inflation and vice versa. This variance is not explained by our data, but this could be the next variable addition to our causal model.

  Market Value – This is the calculated value by the German website transfermarket, we need this variable before the transfer goes through. There was some difficulty in scraping this variable as it is time sensitive. This could be another variable we could add.

  This could be the future steps we could take for this project to build an even more robust model.

- Building separate DAG's for each position:

We initially planned on building a Forward only DAG which only modelled transfers for forwards, this is the reason we scraped goals for each player, however due to time constraints and other issues we were not able to complete this task.

We also planned on building a Defender only DAG which only modelled transfers for defenders, this is the reason we scraped height for each player as the average height of defenders is 6 feet, however due to time constraints and other issues we were not able to complete this task.

8. Conclusion

In this project we built a model for the inflated transfer prices in soccer. The data was preprocessed based on our beliefs and trends in data, we tested the initial assumptions of our DAG and finalized a design which met our DAG specific tests. Using the DAG, we designed a Bayesian model which aligned with our data generation beliefs, we also implemented experiments using concepts of intervention and counterfactuals to test our intuitions about the transfer market. This project highlights several factors that play an important role in players' market value, also figures out a good model for making predictions based on these factors.

9. References

1. Transfer Market website
2. Wikipedia
3. https://github.com/ewenme/transfers