**Football Transfer Market**

A transfer is the process undertaken when a player under contract switches teams in professional football. The registration of a player is transferred from one football club to another. Players may often only be moved during a transfer window and in accordance with the regulations established by a regulatory organization. A transfer fee is an agreed-upon sum of money that a team is willing to pay to another club that has the player's exclusive contract playing rights. When a player switches clubs, the terms of their previous agreement are terminated, and they and their new destination club negotiate the terms of a new agreement.As a result, the transfer fee serves as monetary compensation for the mutually agreed-upon early termination of a professional football player's contract. Transfer fees are based on a player's present football skills, potential for the future, the length of the current contract, the amount of the future pay owing, and the clubs' readiness to reach an agreement on a supply-and-demand-based economic equilibrium.

Over the past ten years, there has been a significant growth in the amounts at risk on the market for football players. According to football-observatory.com :

At big-5 league level, the investments in transfer indemnities have grown from €1.5 billion in 2010 to a new record of €6.6 billion in 2019 (+340%). During this period, big-5 league clubs have recorded a cumulative deficit of €8.9 billion. English Premier League clubs alone have a total net negative balance of €6.5 billion, with a record deficit for Manchester City (€1.1 billion). (Poli et al.)

Primary Research Question:

* Predicting the monetary transfer market fee of players that have been bought/sold in the 2022 summer transfer market.

Secondary Research Questions:

* Predicting the best goalscorer and best playmaker using the player stats of the 2022 season.

**Overview of Data**

The final dataset comprises of two datasets ,

* <https://www.kaggle.com/datasets/shushrutsharma/top-5-football-leagues?select=Fullmetadata.csv>
* <https://www.kaggle.com/datasets/ruslanhuretski/202223-football-summer-transfer-window>

Definitions of terms:

* Transfer fee: In football, a transfer is the transferring of a player from his existingfootball club to another. Transfer fee is the financial compensation paid from an interested club to the club that possesses the player's exclusive contracted playing rights.
* Best goalscorer: There are many ways to find the best goalscorer but the simplest and most basic is the player who scores the most amount of goals in a season.
* Best playmaker: There are several ways to find the best playmaker but for this research purpose the player with the most assists will be the best playmaker.
* Assists :An assist is a pass or a cross made by a player that aids in the scoring of a goal. An assist is often given to a player for passing or shifting the ball to the scorer.
* Key passes:The last pass or pass-cum-shot that results in the receiver of the ball attempting to score but failing.
* Loan: In sports, a loan allows a player to temporarily play for a club other than the one to which they are currently signed. Loan agreements can run anything from a few weeks to a whole season, and they can even span numerous seasons at a time.
* xG : Expected Goals is a statistic that calculates the likelihood of a shot resulting in a goal. It estimates the possibility of a goal based on past data from thousands of shots with comparable attributes.
* xA: It takes various things into account, including the type of pass, pass end-point, and pass length. Adding up a player's or team's anticipated assists provides us an idea of how many assists a player or team should have earned based on their buildup and attacking performance.
* Npg: This term refers to the number of goals scored by a player without including those scored from the penalty spot. This will provide a more balanced perspective of a player's goal scoring abilities.

**Data Cleaning**

Step 1) Merging datasets

The first step of the data cleaning process was to merge the two datasets and create one single dataset where all the independent variables from both datasets are connected with the corresponding value of the dependent variable. The merge was done using the VLOOKUP function on excel using the player names from both to merge. The VLOOKUP function was able to merge 80% of the data but there were few missing/null and incorrect values that needed to be ironed out. The reasons for incorrect data points were incomplete data, such as value for player names involving only first name instead of full name . The second reason was the presence of signs in the player name column, especially for Spanish players which have the sign ~ (tilda) in their name. Lastly few data points had commas,full stop or dollar signs that needed to be removed manually.

Step 2) Removing null values

After successfully merging the dataset I checked for null values and removed all the data points containing them.

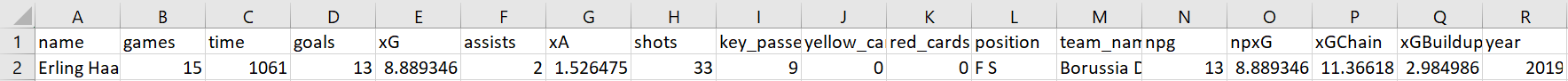
Step 3)Removing unwanted columns

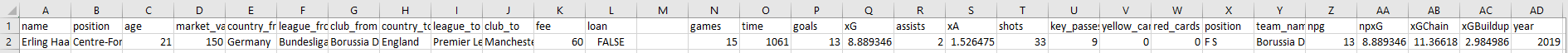
Before any form of EDA I removed all unwanted and duplicate columns such as player\_id, club\_name,league\_from,etc. Most of these were duplicate variables that were present in both the datasets and after merging I removed the unwanted ones.

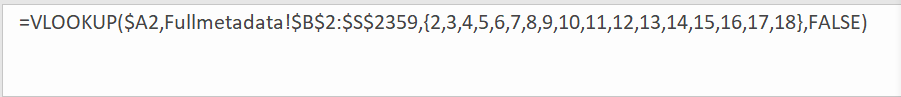
Step 4) Assigning data type

Next step was to assign appropriate data types to all the variables. All values involving numeric values were converted to int or num. Variables involving char were reloaded to R using “as.factor” to convert into factor w/levels.

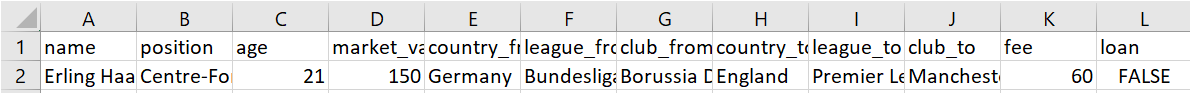
Initial dataset:





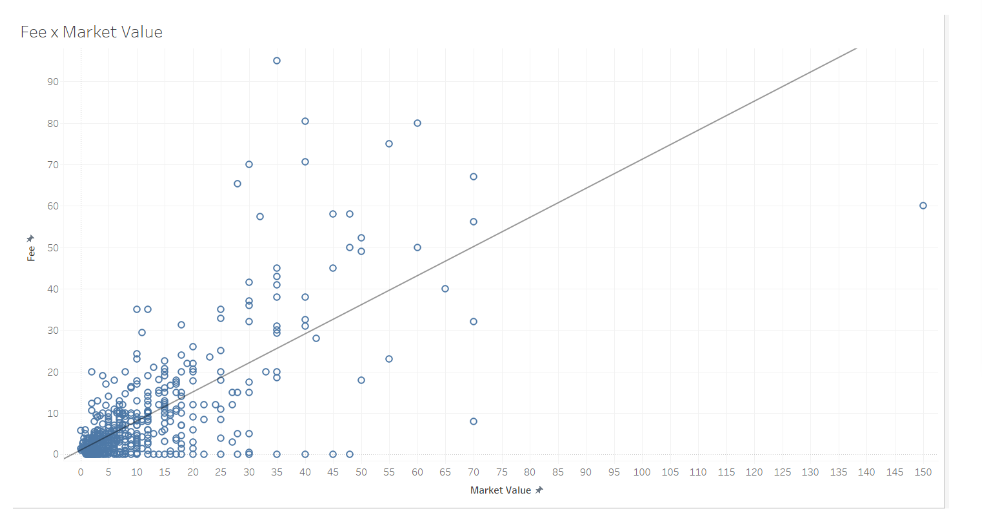


Final dataset:



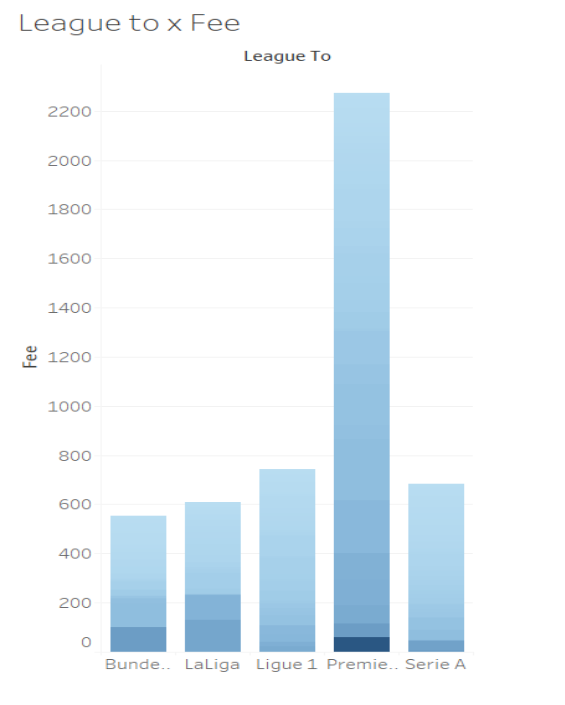
**EDA**

* Fee x Market value



The graph shows Fee on y-axis and Market value on x-axis. We can see clearly that there is disparity between the two variables. There are numerous cases where the potential market value is either less or more than the transfer fee of the player. I aim to answer the key variables that can bring more clarity to the factors causing such discrepancy in transfer fee value.

* League\_to x Fee



The graph shows fee on the y-axis and the top 5 leagues in Europe on the x-axis. It is a well known fact that clubs in the Premier League (First division of England) often over pay for players and are victim to massively inflated transfer fee values. The graph is another proof of the competitive nature of the league.

**Overview of Modeling**

I created a train-test-valid split.

Train : 80%

Test:10%

Valid: 10%

First I train the model using the training dataset and then check the performance of the model on the valid dataset. After I have evaluated the performance of all the models, the most accurate model is tested with the test dataset.

Types of models used:

* *Linear Regression*

A linear regression model regression uses the significance of another variable to predict the values of one variable. It shows the relationship between a continuous dependent variable and several independent variables using this method of analysis.The independent variables may be categorical (position,league\_from) or continuous (goals,time).

* *Ridge and LASSO Regression*

This regularization approach, also known as the penalized regression method, is implemented in feature selection by the shrinkage method. Least Absolute Shrinkage and Selection Operator, or Lasso, is a regularization and model selection technique.

By incorporating a penalty component, ridge regression constrains the coefficients in a manner similar to lasso regression. Ridge regression, on the other hand, uses the square of the coefficients whereas lasso regression uses their magnitude.

* *Regression Tree*

Regression trees are decision trees with continuous target variables rather than class labels on the leaves. The decisions are explained, various consequences can be seen, and potential events can be predicted using a regression tree. You can choose the best possible choice with the help of the analysis. Regression trees' capability to be readable by individuals is its key benefit. In addition to predicting the values of targets' attribute values, regression trees also describe which attributes are utilized and how they are used to arrive at the predictions.

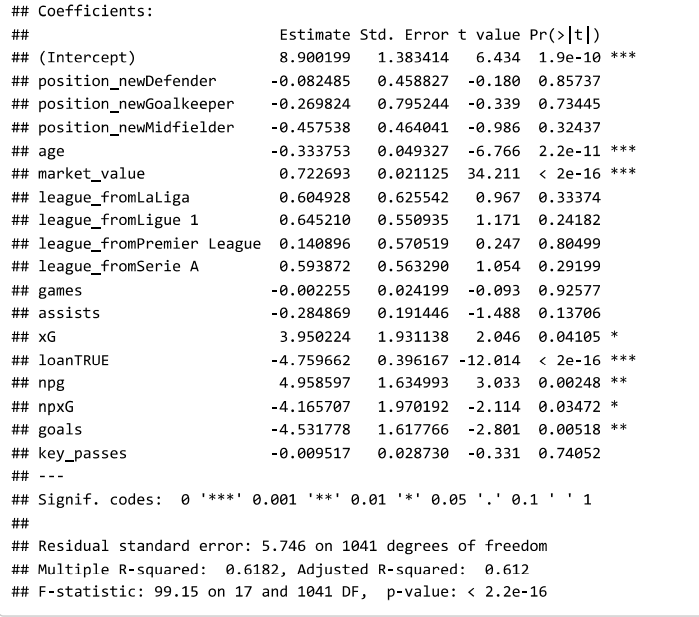
Comparison

Best Linear regression model:

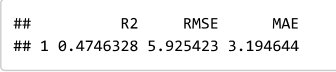
model1<-lm(fee~position\_new+age+market\_value+league\_from+games+assists+xG+loan+npg+npxG+goals+key\_passes, data =train)

summary(model1)

Output:



Performance of LR model on valid dataset:



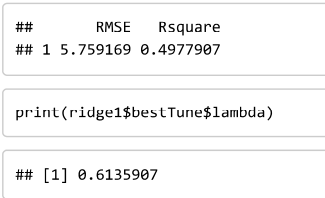
Best Ridge model:

ridge1<-train(fee~position\_new+market\_value+league\_from+league\_to+loan+games+time+assists+shots+key\_passes+red\_cards, data =train, method = "glmnet",

trControl = trainControl("cv", number = 10),

tuneGrid = expand.grid(alpha = 0, lambda = lambda))

Output:



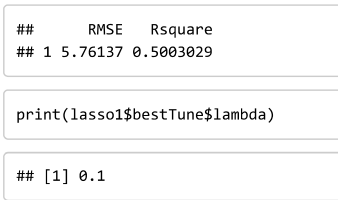
Best LASSO model:

lasso1<-train(fee~position\_new+market\_value+league\_from+league\_to+loan+games+time+assists+shots+key\_passes+red\_cards, data =train, method = "glmnet",

trControl = trainControl("cv", number = 10),

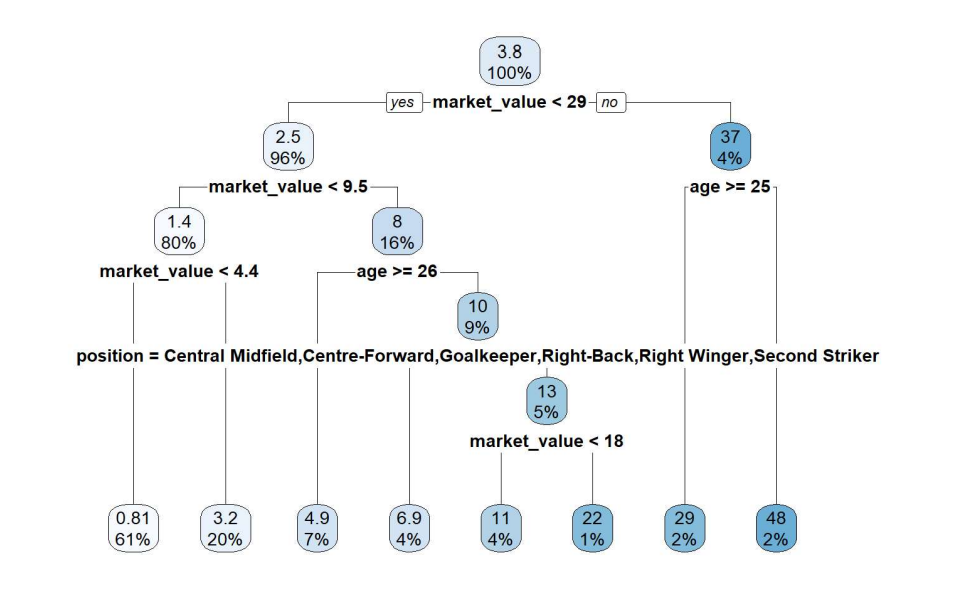
tuneGrid = expand.grid(alpha = 1, lambda = lambda))

Output:



LASSO regression had the best R Squared value (0.5) out of the three. Linear regression had a good r squared value on the train data but the R squared value on unseen data,valid dataset, was lower.

Regression Tree:



Feature Engineering:

For every regression model I started by including all the independent variables in the model and then removing the most insignificant variable one by one. I reiterated this process several times until I reached an efficient r square value and all the variables were significant.

**Future work**

In the future I would like to assess the performance of the players using more variables such as passes, passing accuracy and defensive performance. I would like to create a algorithm that can predict the best player available in the market according to the constraints inputted by the club. If a specific club is looking for a player that has very high goals , is fit (i.e. plays games regularly and has a good track record with injuries) and young with lots of potential under a certain budget the algorithm will predict the best player for that club.

**Limitations**

There were several variables that I was personally interested in while searching for a suitable dataset. Variables that help determine the performance of players playing in the defensive,midfield or even goalkeeper position. The vast majority of dataset available are focused on variables that help in assessing the performance of attackers, variables such as goals,assists,xG and xA. These provide a very incomplete image about the players that are responsible for defensive duties instead of offensive.

**Conclusion**

The project revealed interesting insights into the key variables that are significant in predicting the transfer fee of a player. Having in-depth knowledge about the sport I anticipated a few variables that would be key in determining the transfer fee of the model. It was not a surprise to see variables such as goals,age,assists,xG and xA as key variables. There were few surprising variables such as yellow cards or red cards that played a significant role in training the final model for regression. The EDA also showed that initial hypothesis regarding assumptions in the football transfer market were true. Assumptions such as premier league clubs experiencing inflation in the transfer market and goalkeepers having the least transfer fee while attackers having the most average transfer fee for a player.

**References**

Poli, Raffaele, et al. “Monthly Report Reveals Growing Football Transfer Market Inflation - CIES Football Observatory.” *Monthly Report Reveals Growing Football Transfer Market Inflation - CIES Football Observatory*, Sept. 2019, football-observatory.com/Monthly-Report-reveals-growing-football-transfer.