Cian Thornberry

21338671

An investigation into The Football Transfer Market

Introduction:

For my project I chose to explore data relating to the transfer market in football. I was particularly drawn to this topic as being a football fan I am aware of the power of the transfer market and how it has developed over the last few years. The last 20 years have seen a dramatic increase in the value of players and the total amount spent each season on transfers has increased phenomenally. A problem many people see with modern football is that the influence of money within the game may have led football to lose a bit of its soul as the game has become dominated by billionaire owned clubs dominating the transfer market at the expense of some of footballs smaller clubs. I was interested to explore the data relating to changes in the transfer market over the last 20 years and see if I could find any significant results.

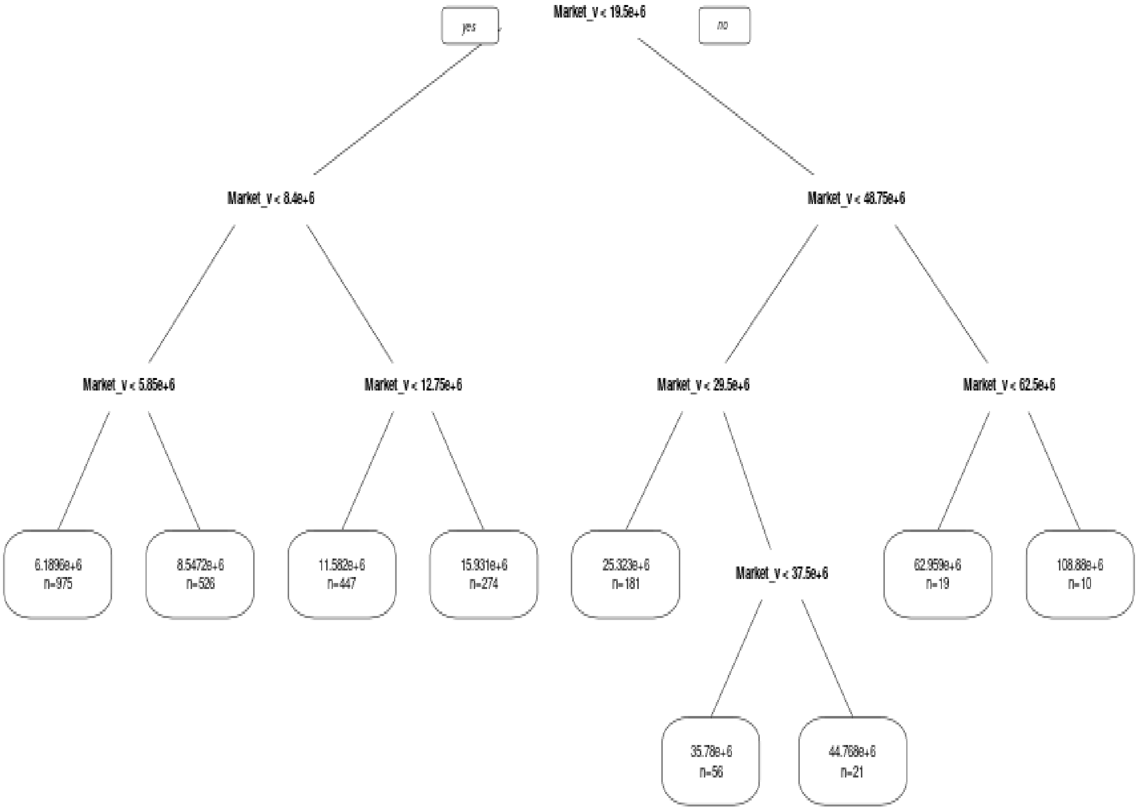
My Dataset:

My data set examines the top 250 football transfers starting from the season 2000-2001 and up to 2018-2019. Originally my dataset had information from all the leagues across the world but after seeing that some of the leagues only had a small number of data, I decided to restrict my dataset to only containing information from Europe's top 5 leagues.

https://www.kaggle.com/datasets/vardan95ghazaryan/top-250-football-transfers-from-2000-to-2018

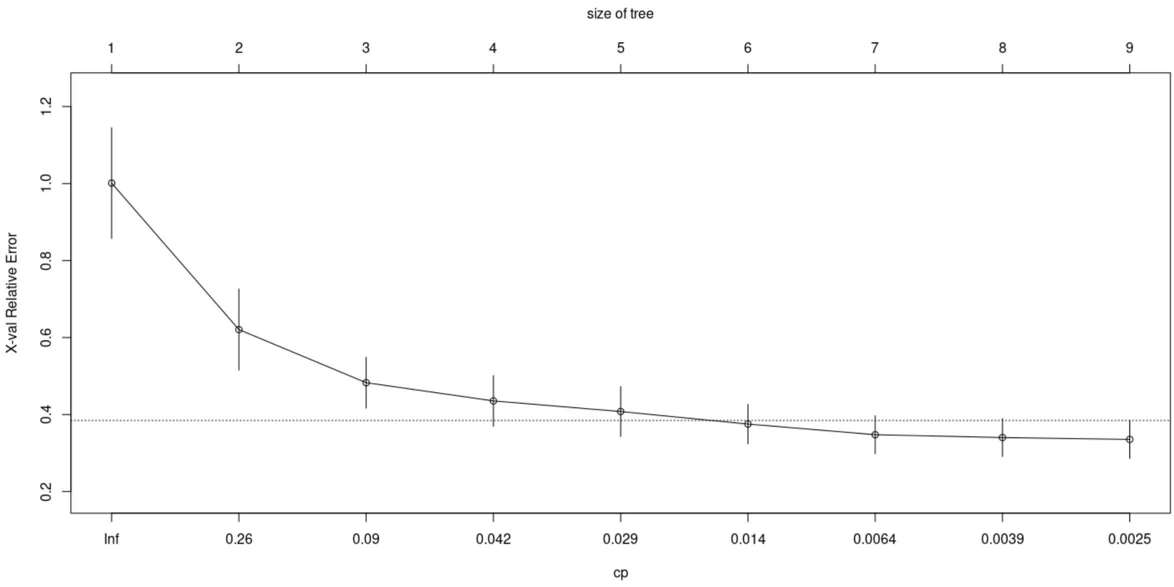
As my dataset is a regression-based model, I built a regression tree to explore the relationship between the variables. Regression trees involve dividing the predictor space into a number of basic sections or strata. We normally utilize the mean or the mode response value for the training observations in the region that an observation belongs to in order to produce a prediction for that observation. Logistic Regression and trees differ in the way that they generate *decision boundaries*, Regression Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single line to divide the space exactly into two.

I started my analysis by using Market Value to predict a player`s final transfer fee. The figure below shows a pruned regression tree to fit this data. It consists of a series of splitting rules, starting at the top of the tree. The top splits assign observations with a market value of less than 19.5e+6 to the left and observations with a market value of greater than 19.5e+6 to the right. The regression tree continues to split the data this way and provides the predicted transfer fee for the players in each category using the mean response variable given in the terminal nodes. For example, players with a market value less than 5.85e+6 the expected transfer fee is 6.1896e+6, i.e. 6.1896^6 = 56,230.96.



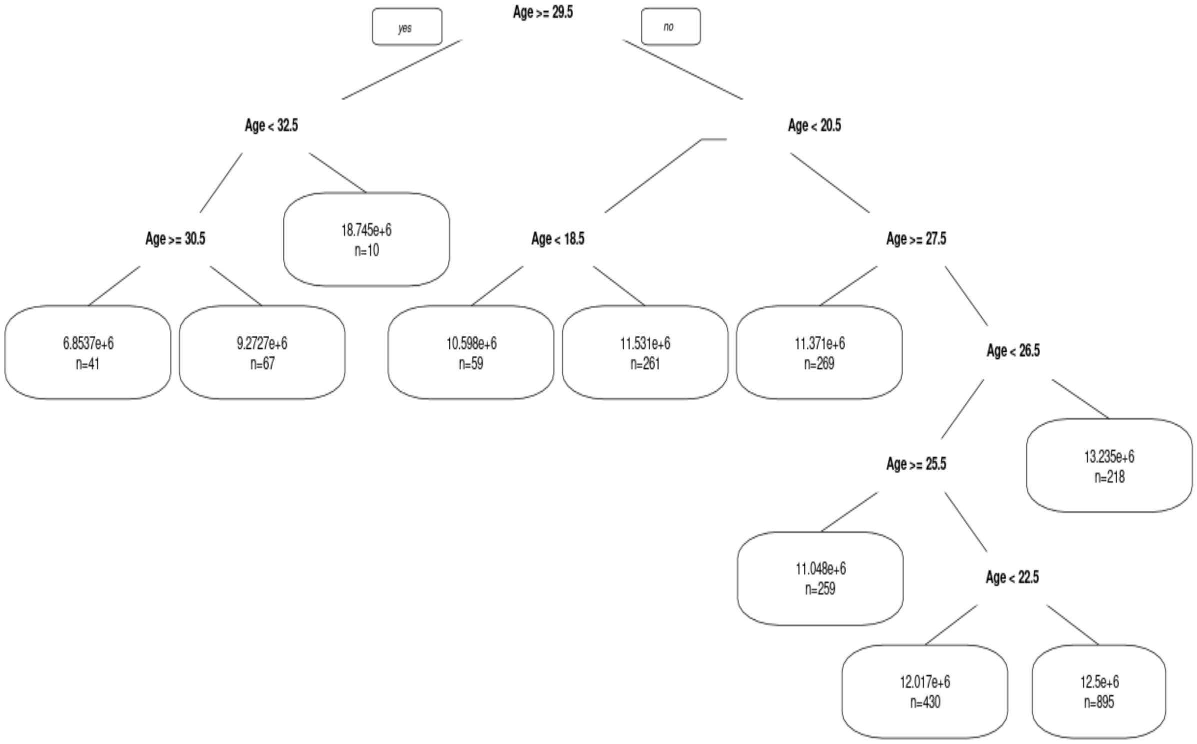
Using the predict tool in R we can use the data to make a prediction of our final transfer fee. The prediction result for this data is 6,189,631 by calculating the mean square error (MSE) for the data we can see how accurate our prediction is 4.568242e+13, i.e., 377,344,766.1. The MSE score for this data suggests using Market value to predict a final transfer fee may not be an accurate model, this contradicts the results from my previous project where my correlation between the two was 0.84.

To control the size of my regression tree I created a variable called best and coded this to find the level of cp for which pruning was optimal.



A good choice of cp for pruning is often the leftmost value for which the mean lies below the horizontal line.

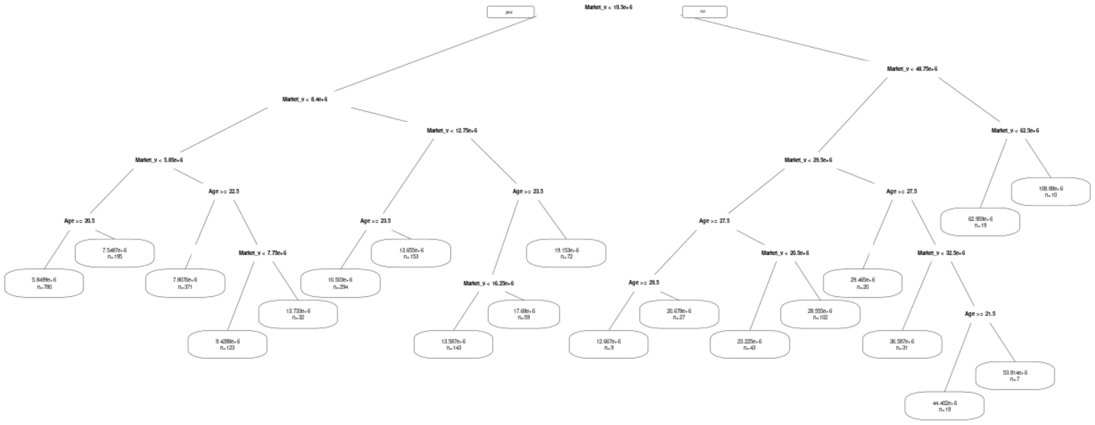
The next part of my analysis was to see how Age could be used to predict a player`s final transfer fee. The figure below shows a pruned regression tree to fit this data. The top splits assign observations with an age of greater than 29 to the left and less than 29 to the right. The regression tree continues to split the data this way and provides the predicted transfer fee for the players in each category using the mean response variable given in the terminal nodes. For example, players whose age is less than 22.5 have an expected transfer fee of 12.017e+6 i.e.3,011,454.925.



Using the predict tool in R we can use the data to make a prediction of

Our final transfer fee. The prediction result for this data is 12,500,034 by calculating the mean square error (MSE) for the data we can see how accurate our prediction is 1.431067e+14, i.e., 151.1. The MSE score for this data suggests using Age to predict a final transfer fee is an accurate model, this result is in line with my result from project two where I found the correlation between age and final transfer fee to be -.6.

The final part of my analysis was to join the previous two variables, Market value and Age, and see how the two variables together could be used to predict a player`s final transfer fee. The Regression tree divides the observation into groups depending on age and market value. The regression tree continues to split the data this way and provides the predicted transfer fee for the players in each category using the mean response variable given in the terminal nodes. For example, players over 23 who have a market value of between 12.75e+6 to 16.25e+6 are predicted to have a final transfer fee of 13.587e+6, i.e. 6,291,315.318.



Using the predict tool in R we can use the data to make a prediction of

Our final transfer fee. The prediction result for this data is 13,235,367.
by calculating the mean square error (MSE) for the data we can see how accura te our prediction is， the MSE for this data is 1.629275e+13, i.e., 570.07.
The MSE score for this data suggests using Age and Market value together is a good model for predicting final transfer fee.

Conclusion:

From undergoing two projects with this dataset using first logistic regression and then regression trees I can see that for this particular dataset regression trees are a more appropriate tool for analysis compared to logistic regression. Comparing the two approaches used in the projects regression trees has given me a more in-depth analysis of the data, the multiple subgroups created by the trees allow for a much more in depth analysis compared to using the logistic regression approach and creating a single line to describe the data.