Football Transfer Values Analysis

By Filippe Fontenele

Summary

Football is often referred as the world game. Football players are some of the highest paid athletes in the world but the goal of this project is finding what are the main predictors for their transfer values. In this notebook, I will use a sample data from players around the world and through data analysis go through their on-field performances, attributes, different transfer fees to reach a machine learning model that will satisfy our question.

Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

Business Problem

This project is aimed at helping football clubs when they are scouting potential new signings in the transfer market.

Football is as big as a business as any other and players are an asset so before investing in an asset, clubs are better off having a tool to assess their investment.

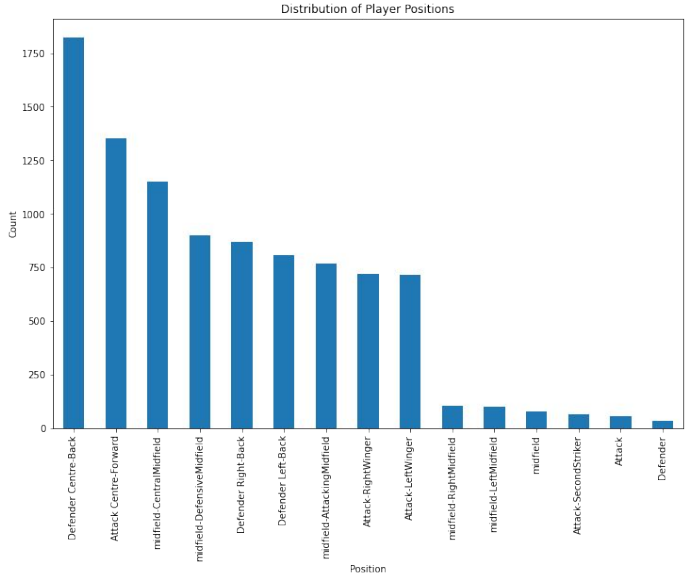
We are still talking about humans that are subject to their environment, but we are confident that by the end of our analysis, we will have a model that will help football clubs around the world make better signings.

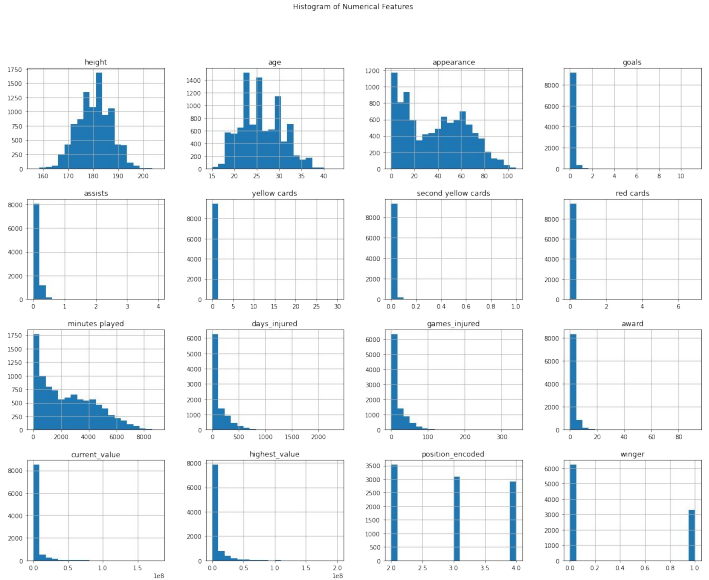
Data

In this case we have a database comprised of information such as age, height, playing position, as well as professional statistics like goal scoring, assists (in 2 seasons 2021-2022 and 2022-2023), injuries, along with total individual and team awards in their career.

Methods

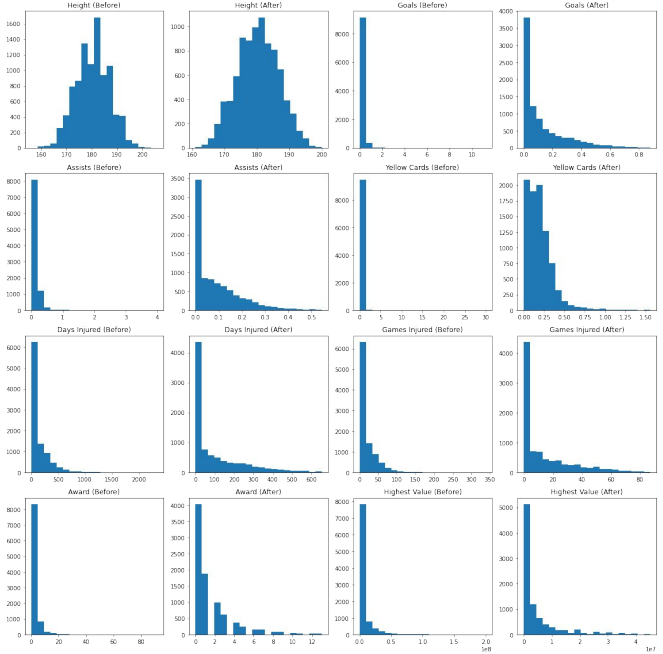
After obtaining the dataset, the project started by analyzing the available features in the dataset that could help us predict our target variable `current_value`. The data already had no null values and has a `position_encoded` column that saves us a step in our project.





Methods

Another key methods used in this football player database project were: handling outliers using z-scores, performing cost-complexity pruning on a decision tree regressor to find the optimal model complexity, evaluating the model's performance on both log-transformed and original scales of the target.



Results

We used Decision Trees for this project. In terms of efficiency, decision trees are generally considered to be efficient models, especially compared to more complex machine learning algorithms. Some key points about the efficiency of decision trees are their training speed, prediction speed and interpretability.

- Initial Model:
Mean Squared Error (MSE): 13,967,556,753,283.84
R-squared: 0.57

The high MSE value indicates that the model is not making very accurate predictions on the test set but the R-squared value is able to explain about 57% of the variance in the target variable.

Results

- Best Pruned Model:
Mean Squared Error (MSE): 8,513,281,671,558.86
R-squared: 0.74

This diagram represents the structure of our best predictive model, achieve by doing Cost-Complexity Pruning (Post-Pruning) technique.

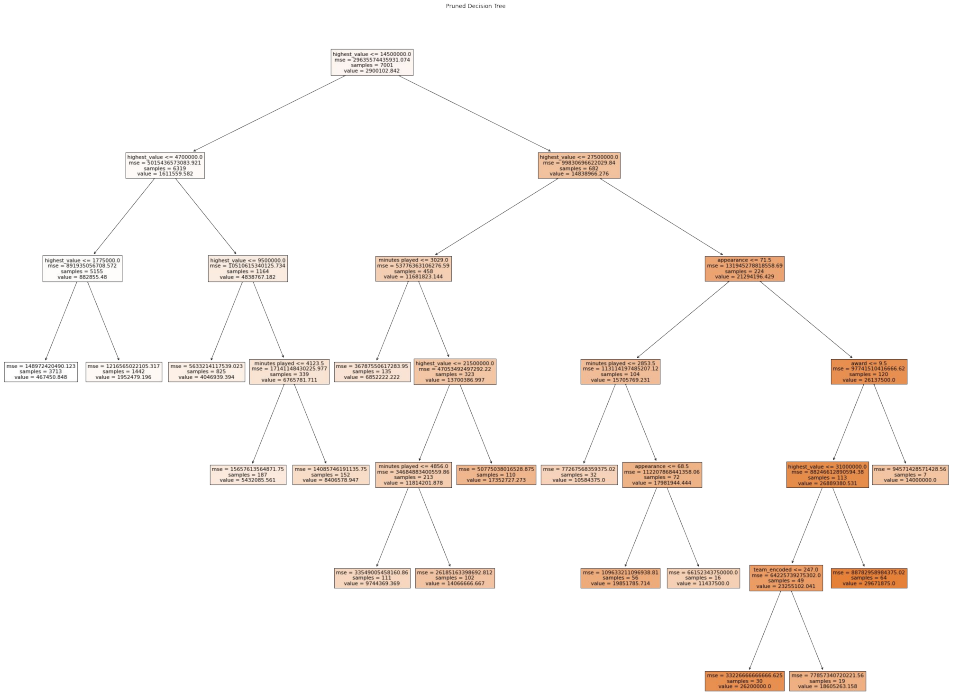
The top node (root) represents the initial split, and subsequent nodes represent further splits on other features.

The leaf nodes at the bottom represent the final predictions.

The path from the root to a leaf node defines a sequence of decisions that lead to a particular prediction.

The color coding indicates the range of predicted values within each leaf node.

By visualizing the decision tree, we can better understand how the model is making its predictions and identify the key factors driving the player value estimates.



Results

- Final Model (Log-Transformed):
Mean Squared Error (log-transformed scale): 0.83
R-squared (log-transformed scale): 0.84

On the log-transformed scale, the final model had an MSE of 0.83 and an R-squared of 0.84, suggesting that it performed well in explaining the variance in the log-transformed target variable.

Conclusions

In this project, we set out to develop a model that could help predict the current value of football players. We had a lot of data to work with, including information on the players' appearances, minutes played, awards, assists, goals, positions, and the teams they play for.

After going through several iterations of the model, we were able to achieve some promising results. Our best model was able to explain about 74% and the log-transformed one about 84% of the variation player values. This means those models were quite good at predicting the relative differences in player values.

However, when we looked at the model's performance on the actual, original player values (not the log-transformed ones), the results were not as strong. The model only explained about 61% of the variation in the original player values, and the errors were quite large.

Thank You!

Email: fmfontenele@gmail.com
GitHub: @f83codered
LinkedIn: linkedin.com/in/filippe-fontenele/