Predict Value of Top Five Leagues' Clubs

Overview:

In this project, I plan to predict the value of the clubs in the top five leagues includes Saudi League. By using different distributions, I will build a module that reads data with specific features and observations then gets a prediction of the value of the club.

Goal:

Predict the value of the club by specific features and define features that are not affected to exclude and which of them have an effect to include then predict the value.

Datasets:

All the Source data coming from the Transfer Market website (https://www.transfermarkt.com) and I will get some information from another source in internet which is Koora website (https://www.kooora.com/) to explore the other features, but before that, I must clean the data, then change the all data that have categorical type to be numerical, to make sure that I can handle this data then I will do engineering for some data for example calculate the point per game and value per player also the column figure the value per the size of the stadium add the columns that I need while I study the data, what I will predict is the value of the club based on the all the features.

Features:

Below is the description of features and the prediction:

Value price: it will be the main data that I will predict, and I will round the number to reflect millions, and the data type is float.

Squad size: total number of the players in each club, and the data type will be int.

Average age: the age average for all players excludes the previous seasons, and the data type is a string.

Foreigners: each league has a specific rule about the players who are not from the same country of the league so I will consider this factor, and the data type will be int.

League Since: Represents the years that the club is in this league, and the data type will be int.

League Name: I will define the name of each league, and I will convert this column from string to be int.

Stadium: Show the total size of a stadium. The data type will be int.

Total Cups: sum all the cups that get it by this club, the data type will be int.

Goals: Total goals of this club per number of years in the league. This data will be engineering data, and the data type will be int.

History points: Total point of this club per number of years in the league, this data will be engineering data, and the data type will be int.

Tools:

To analyze the data and predict the value of clubs, I use different tools like Chrome inspection, Chrome driver, Notepad ++, Jupyter, Python as programming language, Anaconda. Also, I use a different library from python for Example requests, urllib, numby, panda, matplotlib, SciPy, statistics, seaborn, Beautifulsoup, selenium, re, sklearn.

Conclusion:

What I expected after study the define the module that predicts the value of the club and define which features have a high effect on the club value, also which of them have not to affect. Also, define which regression with minimum error.