**ZS Data Scince Challenge - CRISTIANO RONALDO**

As we all know, Cristiano Ronaldo is a legend in football world. given the dataset of Cristiano Ronaldo’s “attempts” in the goal target of his all recorded and unrecorded matches, predict if he has scored a goal or not.

Authored by: Sadanand Vishwas

A Machine Learning model to predict if a shot taken by Cristiano Ronaldo is a goal or not.

Cristiano Ronaldo’s goal Prediction

A Machine Learning model to predict if a shot is goal or not.

Problem Statement: As we all know, Cristiano Ronaldo is a legend in football world. He has played thousands of matches and score hundreds of goals. Now, given the dataset of Cristiano Ronaldo’s “attempts” in the goal target of his all recorded and unrecorded matches, predict if he has scored a goal or not.

Tools & Technologies used:

1. Python 3.7
2. NumPy
3. Scikit-learn
4. Pandas
5. Anaconda 3.0 on Windows 10 Home (1903)

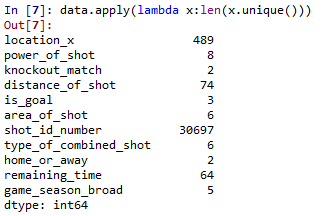
Solution Approach:

1. Data Preprocessing:

I’ve taken the dataset and converted it to the pandas DataFrame, now after analyzing the data we found that there are null values are inserted in some columns.



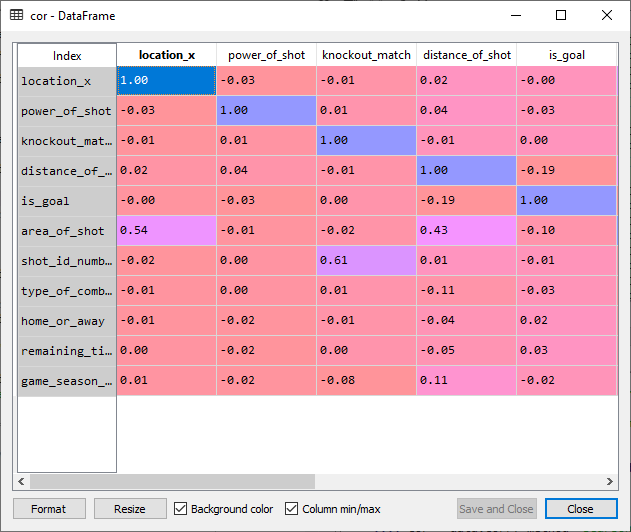
So, I filled the null values with “mean” or by “ffill”/ “bfill” method. For some categorical data they contained many types of unique values, so I created some broad categories for them. After this the data was like following:



After encoding the categorical data columns and removing the unnecessary columns we got the cleaned and preprocessed data.

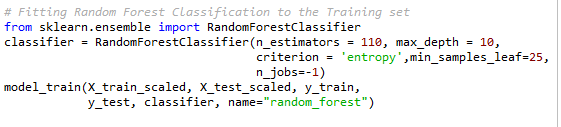
1. Exploratory data analysis:

I used correlation matrix for analyzing the significance of each variable, then removed the variables with least significance or with high dependence on another variable.



1. Model building:

As we predict if a certain shot is a goal or not, so I’ve taken different classification models and applied them to the cleaned data. For better prediction (is goal or not for the submission dataset). I used correlation matrix and confusion matrix to evaluate the models. After training the model with training set data and evaluating it for test set data, I optimized the accuracy as much as possible and then taken the predication for the submission dataset and stored the results.



Conclusion:

After running the data set on different classification models, I reached the conclusion that the “Random Forest Classification” model gives highest accuracy of 60.60% with mean absolute error of 39.39%.



Top 5 most significant variables:

1. distance\_of\_shot
2. area\_of\_shot
3. power\_of\_shot
4. type\_of\_combined\_shot
5. remaining\_time