**ZS DATA SCIENCE CHALLENGE – 2019**

**APPROACH –**

* **Data Preparation –** 
  + After going through the dataset, we got to know that there are several unnecessary and duplicate columns with unnecessary data are present in the dataset file. As example, duplicate remaining\_min column has values like 112.64 but a football match’s duration is 90 minutes. Similarly, remaining \_sec column has ambiguous values like 122.608, but it’s value should be within 60. Similar are the cases with power\_of\_shot,knockout\_match and distance\_of\_shot columns. They are containing wrong values.
  + There is also an extra column containing index which we don’t need
  + Date\_of\_game and match\_event\_id columns are not giving any important information for prediction of goal so we need to drop it.
  + There are several missing values in the dataset. As example, in shot\_id\_number column several values are missing. It can be seen that missing values can be appropriately filled by linear interpolation in forward direction. In several other columns, it can be seen that above and below values to the missing values are same, so they can be filled with previous or next values (game\_season, lat/lng, power\_of\_shot, knockout\_match and home/away columns).
  + Splitting of data into training and test data is being done on the basis of is\_goal column values. The rows where is\_goal values are missing are being put into test dataset.
  + Label Encoding is done for different types of text data and other values which should be present as categorical value for training purpose.
  + One Hot Encoding is done for to remove the confusion of hierarchy among categorical variables after label encoding.
  + Standard scaling is done for feature scaling.
* **Exploratory Data Analysis**
  + I have split the lat/lng feature to different latitude and longitude features for easy interpretation and training.
  + From Correlation map, it is visible that is\_goal has a better correlation with range of shot and negatively correlated with distance of shot. Also, it is positively correlated to remaining minutes and remaining seconds.
* **Model Building**
  + I have used Gradient Boosting Regression technique which produce machine learning prediction model in the form of an ensemble of weak decision models, as decision trees. It minimize the loss of the model by adding weak learners using a gradient descent like procedure. It has ability to sample the data rows and columns before each boosting iteration. This technique is usually effective because it results in more different tree splits, which means more overall information for the model , which is needed for predicting goal from our dataset. Also, there are very few features which are directly affecting the prediction of goal, so in that case, this model works better than other models. This model ideally outperforms models like LinearRegressor and k-nearest-neighbors models.
* **Conclusion**
  + We can observe that goals done by Cristiano Ronaldo depends on several features including range of shot, distance of shot , shot basics and area of shot. He can goal better from short distance in range less than 8 feet. Also, performance is better on the home ground. As the time of the match increases, he has more chance to convert the ball into goal(mostly remaining time 7-10 minutes).