## ZS Associate Data Science Challenge – Cristiano Ronaldo

**Approach used to solve the problem:**

1. Data Preparation:

* Quality Checks Performed:

While looking at the data it seems that data has lot of inaccuracies like many columns have lot of null values also target variable ‘is\_goal’ also have null values. So what I do is remove rows which has target variable as null.

* Data Preprocessing Steps :

To preprocess data first of all I separate the data into different types i.e object into category and int into float. After that for float type features for filling missing values in them I use mean imputation technique and for categorical features I introduced another category for filling missing values.

Also I draw correlation plot to find which features are co-related and I remove those features it found that match\_event\_id and match\_id are co-related so I remove them.

Also for categorical features I used one hot encoding approach to convert them to numerical.

2. EDA

* Feature generation:

After all the preprocessing done I am left with total of 1794 total features so to identify among them which features are of high importance I build a random forest classifier and perform hyper-parameter tuning of max\_depth and n\_estimators and with the help of the classifier I am able to determine which features are useful based on their relative importance. So, from total of 1794 features I picked 100 features.

These are:

'distance\_of\_shot', 'type\_of\_shot\_shot - 39', 'type\_of\_shot\_shot - 44',

'type\_of\_combined\_shot\_shot - 4', 'location\_x', 'remaining\_sec.1',

'remaining\_sec', 'location\_y', 'remaining\_min.1',

'shot\_basics\_Goal Area', 'distance\_of\_shot.1', 'remaining\_min',

'power\_of\_shot.1', 'knockout\_match.1', 'type\_of\_combined\_shot\_shot - 5',

'power\_of\_shot', 'type\_of\_shot\_shot - 17', 'type\_of\_shot\_shot - 49',

'type\_of\_shot\_shot - 38', 'lat/lng\_42.982923, -71.446094',

'type\_of\_combined\_shot\_shot - 0', 'type\_of\_shot\_shot - 15',

'type\_of\_shot\_shot - 52', 'type\_of\_combined\_shot\_shot - 1',

'knockout\_match', 'game\_season\_2007-08', 'home/away\_MANU @ SAS',

'game\_season\_2008-09', 'game\_season\_2005-06', 'type\_of\_shot\_shot - 4',

'type\_of\_combined\_shot\_type\_of\_combined\_shot', 'lat/lng\_lat/lng',

'range\_of\_shot\_24+ ft.', 'game\_season\_game\_season',

'game\_season\_2001-02', 'lat/lng\_29.444994, -98.524120',

'type\_of\_shot\_shot - 12', 'home/away\_home/away',

'type\_of\_shot\_shot - 36', 'game\_season\_1998-99', 'game\_season\_2000-01',

'game\_season\_2012-13', 'type\_of\_shot\_shot - 32',

'shot\_basics\_shot\_basics', 'game\_season\_2003-04',

'home/away\_MANU @ HOU', 'date\_of\_game\_date\_of\_game',

'lat/lng\_29.740325, -95.365762', 'area\_of\_shot\_Center(C)',

'home/away\_MANU vs. MEM', 'game\_season\_2002-03', 'game\_season\_2009-10',

'game\_season\_2011-12', 'range\_of\_shot\_Less Than 8 ft.',

'type\_of\_combined\_shot\_shot - 3', 'home/away\_MANU vs. MIN',

'home/away\_MANU vs. DEN', 'home/away\_MANU @ GSW', 'game\_season\_2010-11',

'shot\_basics\_Goal Line', 'type\_of\_shot\_shot - 23',

'game\_season\_2004-05', 'team\_name\_Manchester United',

'game\_season\_1999-00', 'home/away\_MANU vs. SAC', 'team\_name\_team\_name',

'area\_of\_shot\_area\_of\_shot', 'home/away\_MANU vs. HOU',

'date\_of\_game\_2000-04-27', 'home/away\_MANU vs. IND',

'game\_season\_2006-07', 'home/away\_MANU @ CLE', 'home/away\_MANU vs. NYK',

'game\_season\_1997-98', 'home/away\_MANU vs. SAS', 'game\_season\_2015-16',

'home/away\_MANU @ DEN', 'home/away\_MANU vs. PHX',

'range\_of\_shot\_range\_of\_shot', 'type\_of\_shot\_shot - 31',

'game\_season\_2014-15', 'lat/lng\_37.754130, -122.437947',

'lat/lng\_33.552026, -112.071667', 'home/away\_MANU vs. DAL',

'type\_of\_shot\_shot - 43', 'lat/lng\_43.717098, -79.395917',

'home/away\_MANU vs. WAS', 'home/away\_MANU vs. ORL',

'lat/lng\_39.739968, -104.954013', 'home/away\_MANU @ UTA',

'home/away\_MANU vs. NOH', 'home/away\_MANU vs. LAC',

'lat/lng\_39.993941, -75.143458', 'home/away\_MANU @ PHX',

'home/away\_MANU @ MIL', 'lat/lng\_38.567296, -121.456638',

'date\_of\_game\_2003-02-01', 'lat/lng\_43.062206, -87.944754',

'home/away\_MANU vs. UTA', 'lat/lng\_41.484971, -81.671552'

* Exploratory Data Analysis:

I have performed EDA on the following features:

1. is\_goal

2. area\_of\_shot

3. range\_of\_shot

4. shot\_basics

5. game\_season

For is\_goal feature I have draw a bar plot which tells the success rate of goal. Out of total 24,429 it shows that ronaldo has succeded in 10,879 times while in remaining 13,550 he fails. So, success rate percentage is about 44.5.

For area\_of\_shot feature it is found that most famous area in which ronaldo likes to play is

Center.

For range\_of\_shot feature the most famous range is less than 8 ft.

For shot\_basics feature it is found that most famous shot ronaldo likes to play is Mid Range

followed by Center.

For game\_season feature it is found that most games ronaldo plays in the year 2005.

3. Model Building:

* Model Choice:

I have used bagging model Random Forest( a tree based model) as my preferred choice because the dataset looks like non-linear and for non-linear data the tree based model outperformed all the other models. By this model I am able to get an accuracy of 63.2%, auc of 0.604 and mean absolute error of 0.367.

Instead of this I have also used deep learning models like MLP( multi layer perceptron) and CNN( convolutional neural network) to increase my score. Through this my score increased a little but not significantly.

4. Conclusion:

* Important Features:

The top 5 most significant features are:

1. distance\_of\_shot

2. type\_of\_shot\_shot – 39

3. remaining\_sec .1

4. location\_y

5. location\_x

For getting the 5 most important features I build a random forest classifier and perform hyper-parameter tuning of max\_depth and n\_estimators on this classifier and with the help of the classifier I am able to determine the top 5 most important features based on their relative importance.