Data Preparation :

1. Renamed the columns that were repeated by a suffix 1 to the columns ( remaining\_min1,remaining\_sec1, power\_of\_shot1, etc.)
2. Put the is\_goal column at the end of the data
3. Filled some of the missing values by merging the columns(remaining\_min,remaining\_sec, power\_of\_shot, etc.)
4. Performed winsorization to clip the values in the merged columns
5. Dropped the repeated columns(remaining\_min1,remaining\_sec1, power\_of\_shot1, etc.)
6. Dropped the columns that may be trivial for goal prediction ('shot\_id\_number\_1','team\_name','team\_id','match\_event\_id','lat/lng')
7. Filled the missing values of categorical data using back filling method ('match\_id','team\_id','lat/lng','team\_name','match\_event\_id','game\_season','date\_of\_game')
8. Filled the missing numerical data column values using mean and median(‘location\_x’,’location\_y’,’power\_of\_shot’ ,etc)
9. Label encoded the categorical data
10. Converted the ‘date\_of\_games’ to day, month and year as separate columns
11. To label the home or away team 1 home while 0 represents away splitting the column values of 'home/away' to check for '@' or 'vs.'

EDA : I have not generated any new features.

Exploratory Analysis : Plotted the histogram of the data with all the columns to check where the value for the column lies. Also plotted a heatmap of correlation matrix to check for the columns with which the ‘is\_goal’ is more correlated. From the heatmap, the conclusion was that ‘is\_goal’ was more correlated with ‘remaining\_sec’ and ‘range\_of\_shot’. Rest columns did not have much correlation with ‘is\_goal’

Model Building : I created a neural network with three dense layers with units as (16,8 and 4 in respective layers) and a final output layer with one activation unit. The dense layers has the ‘relu’ activation while the output unit had a ‘sigmoid’ activation function. Used ‘binary\_crossentropy’ as loss function and adams optimizer for training . Batch size : 256, epochs = 10, metrics = ‘accuracy’. Since the no. of parameters were 16 and the exploratory analysis did not gave any meaningful results. Neural Networks seemed a good step to me.

Conclusion : Needs more exploratory analysis to figure out the variables that are affecting the ‘is\_goal’ variable most.