**ZS Data science challenge**

# RONALDO

## Data preparation:

1. match\_event\_id 10.0 type <class 'numpy.float64'>

location\_x 167.0 type <class 'numpy.float64'>

location\_y 72.0 type <class 'numpy.float64'>

remaining\_min 10.0 type <class 'numpy.float64'>

power\_of\_shot 1.0 type <class 'numpy.float64'>

knockout\_match 0.0 type <class 'numpy.float64'>

game\_season 2000-01 type <class 'str'>

remaining\_sec 27.0 type <class 'numpy.float64'>

distance\_of\_shot 38.0 type <class 'numpy.float64'>

is\_goal nan type <class 'numpy.float64'>

area\_of\_shot Right Side(R) type <class 'str'>

shot\_basics Mid Range type <class 'str'>

range\_of\_shot 16-24 ft. type <class 'str'>

team\_name Manchester United type <class 'str'>

date\_of\_game 2000-10-31 type <class 'str'>

home/away MANU @ POR type <class 'str'>

shot\_id\_number 1.0 type <class 'numpy.float64'>

lat/lng 45.539131, -122.651648 type <class 'str'>

type\_of\_shot shot - 30 type <class 'str'>

type\_of\_combined\_shot nan type <class 'float'>

match\_id 20000012 type <class 'numpy.int64'>

team\_id 1610612747 type <class 'numpy.int64'>

remaining\_min.1 10.0 type <class 'numpy.float64'>

power\_of\_shot.1 1.0 type <class 'numpy.float64'>

knockout\_match.1 50.608000000000004 type <class 'numpy.float64'>

remaining\_sec.1 54.2 type <class 'numpy.float64'>

distance\_of\_shot.1 38.0 type <class 'numpy.float64'>

These were the types of different columns, some features like team\_name were not that relevant and most of the features were strings, and some of them had to be changed into features by creating a dictionary with real numbers and mapping them back in the data frame.

And most of the columns have null values, columns like power\_of\_shot has their missing values in complementary columns like power\_of\_shot.1

## EDA:

Feature generation, all given integers were just normalised, where as

* location has been turned into a feature by using a k means cluster, so that they can be turned into real numbers in a range
* as some of the values in some columns are present in other complementary columns, both of them have been merged and most logical ones have been taken as features
* and things like the game\_season are just continuous in their values across different rows, so I used ffill not to lose some valuable rows

## Model Building:

Different models were built. Following are a few:

* logistic regression
* svm with linear kernel
* svm with rbf kernel
* decision trees
* random forest
* XGB classifier

And again those were tuned according to their hyperparameters, after looking at the probability distributions of different models, Gaussian SVM has been chosen to predict

## Conclusion:

I had very little time to spend on this, I have not quantitatively seen which features are the prominent ones, but I think, the following are prominent features.

* power\_of\_shot
* game\_season
* distance\_of\_shot
* type\_of\_shot (which means it’s a mix of type\_of\_shot and type\_of\_combined\_shot)

please find the ipynb file here.