**Data Prep:**

* Quality checks performed / Errors found: The data had a lot of null values. For example, the ‘power\_of\_shot’ column, ‘distance\_of\_shot’ column, etc. Columns containing string values also had some missing fields.
* Data preprocessing steps: The null values (NaN) were replaced by the mean of all the values of the respective columns for the integer fields while the string fields were replaced by the most occurring value (mode).

**EDA:**

* Feature Generation:

1. Power of shot (integer data) – To determine the force with which the shot was played.
2. Distance of shot (integer data) – To determine the distance from which the shot was played.
3. Remaining time (calculated as remaining\_min\*60 + remaining\_sec) – Time plays an important role in determining whether the shot was successful or not. It tells us about the mental and emotional state of the player.
4. Area of shot (calculated by mapping the areas given to integer type data) – To determine the area from which the shot was played.
5. Range of shot (calculated by mapping) – To determine the range of the shot.
6. Shot basics (calculated by mapping) – To determine what kind of shot was played.
7. Location x and Location y (integer data) – Tells the exact location from where the shot was played.

* Exploratory data analysis: Graphs provided in the notebook

**Model building:**

* Model choice: I started with basics models such as Logistic Regression, SVM, KNN. They gave a somewhat accurate score. However, more training and exploration of different algorithms was required. Many other models such as Random Forest, XGBoost, Neural networks, Ada Boost, etc were trained to determine the mean absolute error and cross validation score. Out of these the two best algorithms were chosen. These were Random Forest and XGBoost.

A histogram was plotted for both the outputs to determine the spread of predictions and overfitting. On submitting, XGBoost gave an accuracy of 83% (This was done without taking in consideration the location features). However, Random Forest gave an accuracy of 82.6% (with the location features).

**Conclusion:**

* Important Features: As of now, the features having the most significant role in determining whether it was a goal or not have been used. It was also taken care that the feature had a less number of nulls to avoid irrelevant data.

Features are now selected using feature selection techniques such as Univariate Selection and ExtraTreesClassifier. Features having importances of above 0.05 have been selected and the model is trained on these new features.