**Data Prep:**

1. The csv file is read as a dataframe and there are some unnecessary columns which are dropped.
2. The data has many None values. It even had None in some important fields like the ‘shot\_id\_number’. This field is completely replaced by list of consecutive numbers from 1 to the length of the data.
3. The ‘shot\_id\_number’ was converted into the index of the dataframe for convenience.
4. The ‘home/away’ column has been split into two columns, with one as ‘home’ and other as ‘away’. For a match played at the home ground, the column ‘home’ records the opponent team name and column ‘away’ records None. For matches played away, the ‘away’ column records the venue of match and ‘home’ column records None.
5. The ‘lat/lng’ column is also separated into two columns, ‘lat’ and ‘lng’ respectively and converted into integers for our convenience.
6. Both ‘type\_of\_shot’ and ‘type\_of\_combined\_shot’ are mixed together as they converse each other.
7. The date of each game is expressed in string format, so they are converted into the number of days before today that the match was played.
8. Some of the columns which have strings as their values are label encoded for convenience.
9. The other None elements were filled using the pandas inbuilt function fillna and using the method ffill to fill the values with adjacent values.

**EDA:**

1. The additional features generated were already mentioned in the data preparation step.
2. The data was scaled for better coefficient formation and data uniformity for prediction purposes.
3. Using the correlation matrix on the data, we have identified the important fields required for our model.
4. Observed that 'match\_event\_id','location\_y','power\_of\_shot','distance\_of\_shot', 'area\_of\_shot', 'shot\_basics','range\_of\_shot','distance\_of\_shot.1','is\_goal' .

**Model building:**

Applied different models like Linear Regression, Ridge Regression, Support Vector Machine, Random Forest Regressor and ensemble methods and concluded that Support Vector

Machine gave better output as compared to the others.

As we need probability of the shot, the Logistic Regression plays a very good role in predicting such values between 0 and 1.

Predicted the test data on this model.

**Conclusion:**

Some data which was in text format was converted into relevant formats and other preprocessing was done.

The None fields were filled for the machine learning model to learn and predict on the data.

Many Machine Leaning models were tried and concluded that Logistic Regression performed better than the rest.