**Section –B**

**Data Prep:**

1. Quality checks performed / Errors found
2. A large number of values (approx. 50%) were missing in “type\_of\_shot” and “type\_of\_combined\_shot” but EDA shows that these values are important for predictions.
3. Repeated Columns with different values are also to be seen.
4. The “home/away” column doesn’t clearly define where the match is happening
5. The values of “team\_id” and “team\_name” are same for the entire dataset.
6. “match\_event\_id” doesn’t show relevance with “match\_id“ and this doesn’t provide any information
7. Data preprocessing steps
8. “shot\_id\_number” is to be interpolated linearly.
9. “team\_id” and “team\_name” are same for the entire dataset, hence dropped
10. The repeated columns were dropped after extracting information from them (look for “same\_columns” function in code.
11. There is a clear one-to-one correspondence between “match\_id” and the following columns:
12. “home/away”
13. “lat/lng”
14. “date\_of\_game”
15. “shot\_basics”

5. There was a clear pattern between “location\_x” , “location\_y” and the following columns:

1. “distance\_of\_shot”
2. “range\_of\_shot”
3. “area\_of\_shot”
4. “shot\_basics”

**EDA:**

Feature generation:

Graphs were plotted to see a clear relation between columns and then its values were predicted using KNN.

Exploratory data analysis:

There was a clear pattern between “location\_x”, “location\_y” and the certain columns which included “distance\_of\_shot”, “range\_of\_shot”, “area\_of\_shot” and “shot\_basics”.

Graphs with colors varying according to the above columns were plotted against “location\_x” and “location\_y” to see the trends.

**Model building**:

Xgboost is been used as the model because of the large number of columns that were produced due to categorical encoding. Also XgBoost can be trained with NULL values which provides a upperhand as the values of “type\_of\_shot” have a large say in the predictions.

**Conclusion**:

Important Features:

The top 5 most important features were or significant variables after encoding the variables were determined by xgboost.XGBClassifier.model\_importances\_.They are:

1. “shot-39” (categorical data)

2. “distance\_of\_shot”

3. “shot-44” (categorical data)

4. “24+ft” (categorical data)

5. “shot-17” (categorical data)

The other conclusions drawn were that many other classifiers would fail to do a better performance since in other classifiers the variables should be independent of one other which clearly isn’t the case here.

P.S:- I would like to extend my sincere regards to the team behind this data science hackathon it was a tool to provide young data science enthusiasts to tackle a real world dataset where the data isn’t clean and most the part involves cleaning of data. Kudos.