**Football Game Prediction**

Dataset contains (370,106) it contains a lot of features as compared to samples of datset which made us to do lot of EDA.

There is no duplicate values and less missing values.

Topics Below:

1. EDA Performed
2. Description after EDA
3. Model Selection and metrics selection
4. Final Results and score
5. Future Work(feature selection using Dimension Reduction Algorithms,Ignore 5-6 features,Cross Validation to get best hyperparameter )etc.

**EDA Performed:**

1. We check if the Team have Home ground advantage or not and we clearly Observe from Bar chart that Home Team Have Advantage.
2. We classified the Date into Weekend or Weekdays and check the result with outcome of the match and some teams wins more matches and have some impact on weekends.
3. There is no relation between Full Time Goal and Half Time Goal.
4. We have lots of features of odds, I wrote a function to create them in weightage winning percentage and then compare with the match winning and this show great stats with Outcome and almost all the bookmakers company have same value,So to reduce the feature we took max which gives the same results as all does.
5. We then check the over/under2.5 goals and with the number of goals as well as outcome and no difference comes into this so we remove this feature.
6. Asian Handicap is hard feature to decode and also when decoded it is one of the import feature to determine the outcome.
7. Asian Handicap have three lines,full line ,half line ,Quarter line: We make a goal difference(Goal Supremacy) with respect to home team then we check.
8. If people invested money or if the odds are in favour of some team then it is more chance of winning or considered as strong team.
9. So we divided this into three category: Full Time, Half Time,Quarter Time and considered we put 1Rs. Money into odds. And then see wheather we make good money or loose and according to that we make a binary feature for that .
10. Almost all the Odds feature have Critical Odds which might we change or impact on Outcome.
11. So we check the odds with ciritcal one and the once which not match with previous will give a 0 else 1.

**Desrciption after EDA:**

1. After Completion of EDA we have 15 features left and all are important so we did not apply PCA. And all features are in binary

**Model Selection:**

1. The class is highly imbalanced so we used Stratified Kfold Validation.
2. For metrics instead of accuracy I used MCC and ROC\_AUC score
3. I don’t think Linear models would work good in newly formed dataset because we can see a lot of dependent feature also present in our dataset. But we will apply linear models to check our assumption.
4. Ensemble techniques might work best on this dataset.

**Final Result and score:**

Model ROC\_AUC MCC

0 LogReg 0.778367 0.551581

1 RF 0.808571 0.611430

2 KNN 0.517551 0.038005

3 SVM 0.500000 0.000000

4 GNB 0.808163 0.597137

5 XGB 0.818776 0.637551

We can clearly see that our assumption is correct. Ensemble Algo works well

**Future Work:**

1. Asian Handicap odds might be not that good,I understand the terms but think in relation with money.
2. Didn’t analyze the normal odds and Asian Handicap odds
3. Need to use Over/Under 2.5 goal in better way
4. Although here PCA is not necessary because I already reduce all features but still need to consider Dimension Reduction Algorithms.
5. Split the data into Home and Away team then analyse the datset
6. Split the data into draw,win and loose
7. Not used features like Red card Yellow Card Referee Time which might lost of information.