

Documentation

- Quality training dataset :

Quality of Data set is good but there is more missing values in the data which impact on models.

- Errors found in the training dataset

No , I have not find any error in the Dataset.

- Data Preprocessing Steps :

1. Convert the all categorical variable into the integer.
2. Then converted into the numerical matrix
3. Apply Median on the missing values and some are removed their entire column
4. Apply Xgboost model
5. Predict the desire output

- Model Used :

I used Xgboost model to predict the desire output because it is very fast and efficient model for the categorical large dataset.

There are many independent variable in the data and each has its impact on the model so that's why I used Xgboost model.

It work just like random forest model but it is very improved model.