We have 2 files in dataset Train.rpt & Test.rpt , convert it into .csv files.

EDA

1. **finding null values:**

The train data contains multiple null values.

**isnull().any().sum()** this function returns the no of null values in train data.

I have handled null values by using **fillna(0):** This function replace Nan values with 0.

1. **Low\_memory=False**

I have used this while reading the data as file is very large and have mixed type data.

1. **Encoding categorical variables**

As the data is large mapping and one hot encoding will not be a good choice. So I have used

**Pd.categorical().codes()**

It returns codes are an array of integers which are the positions of the actual values in the categories array

1. **Splitting into train and test data:**

As 80% training and 20% testing

1. **Modelling**

Here, I have tested for all models and then selected xgboost as a final model.

Xgboost model gives the good result as compared with all models. I have achieved this by using hyperparameter tuning.

Here the parameters are max\_depth, n\_estimators, learning rate.

1. **Prediction**

I did the prediction on test.csv which gives r2 score=0.94, rmse=125.67, and mae=90.41

The performance metrics used are r2 score, rmse, mae