***ASSIGNMENT-PREDICTING THE IMPACT OF A TWEET***

**FINDINGS FROM THE ANALYSIS**

* There were no NULL values in the Dataset.
* The records are uniquely identified by the attribute 'ID' which is the tweet URL.
* There were 47512 unique records and 2488 duplicate records.
* There was one categorial feature namely ‘Media Type’.
* The metric used for model evaluation is mean absolute error and r2 score*.*
* The performance of various models are given below

|  |  |  |
| --- | --- | --- |
| **MODEL** | **MEAN ABSOLUTE ERROR(Test Error)** | **R2 SCORE** |
| 1)Linear regression model | 0.3574 | 0.9999 |
| 2)Decision Tree model | 768.0290 | 0.9982 |
| 3)XGBoost model | 676.1762 | 0.9988 |
| 4)XGBoost with Hyper parameter optimization | 0.3519 | 0.9999 |

**SUMMARY**

* Among the four models ,linear regression model and xgboost with hyper parameter optimization model give good results as the mean absolute error was low and the r2 score was pretty good.
* But the best model to solve the given problem would be linear regression model because it consume less time , has good results and is simple.
* The decision tree model has low bias (less training error) and high variance (high test error) and the variance can be reduced by using some boosting techniques.
* With xgboost model we were able to reduce the variance ,but the variance was still high and hence we go for xgboost with hyperparameter optimization .
* With hyperparameter optimization we optimize the best parameters and we trained the model with the best parameters so as to get a generalized model.
* The learning rate choosen were 0.05 ,0.1 ,0.15 ,0.20 and the best was 0.15.