Author name: Ankit Prasad  
Date: 05-08-2021

**Report on tweet viral prediction**

**Data Pre-processing:**

1. In the first step I checked if the dataset is having any missing values or not. I did not found any missing values.
2. Then checked for the unique values in each column and data type of each column.
3. In third step I plotted the correlation of each column with each other. From the plot I checked for impact column which is our main column to be predicted if a tweet goes viral or not.
4. I removed all the columns which were significantly not related to impact column.
5. For model to train and validate I split my dataset into training and testing dataset.

**Data Modelling:**

I explored different regression techniques on train and test dataset. At the very first attempt I used all the models with all its default hyper-parameters. Results are as follows:

1. **Adaboost regression:** r2 score is: 0.7350874474413964

Mean absolute error is: 52854.99055323718

Wall time: 3.2 s

1. **Extra trees regression:** r2 score is: 0.9993120331762226

Mean absolute error is: 211.67559999999997

Wall time: 12.5 s

1. **Random Forest regression:** r2 score is: 0.9991049856128338

Mean absolute error is: 420.06259200000005

Wall time: 23.9 s

1. **XG boost regression:** r2 score is: 0.9992741322138515

Mean absolute error is: 694.9607264678955

Wall time: 3.18 s

1. **Linear regression:** r2 score is: 0.9999999992634825

Mean absolute error is: 0.2776949395134149

Wall time: 82 ms

1. **Neural Network:** r2 score is: 0.2725128415084551

Mean absolute error is: 26679.89909385214

Wall time: 55.2 s

Note: In neural network accuracy can be increased by increasing hidden layers and neurons. But execution time will also increase. In this case I already got very good results with traditional ml algorithms. So I did not dig deep in neural networks.

**Best model:**

For this scenario linear regression is the best model undoubtedly. Its wall time is also minimum among all and accuracy is also the highest.

R2 score is: 0.9999999992634825

Mean absolute error is: 0.27769493951341495

Explained variance score is: 0.9999999992640659

Mean squared error is: 8.265401442597105

Wall time: 53 ms

How I defined the best model?

I choose three parameters to define best model:

1. R2 score should be near to 1 (It means model is well fitted).
2. Mean absolute error (Should be as low as possible).
3. Wall time (Should be minimum).