# Data Pre-processing

After analysing the data, we observe that the ‘subscriber count’ and all the ‘bytes transfer’ data were skewed. So, we transformed the data using the following formula:

**X’ = log(1+x)**

After this transformation, our data becomes gaussian, as it should have been so as to optimise the efficiency of our model.

Further normalizing the data using two methods – Min-Max normalization and Standardisation, with and without using the above transformation, respectively.

# Model Selection

We applied many different models – Random Forest Classifier, ADA Boost Classifier, Extra Trees Classifier, Logistic Regression, Decision Tree Classifier, MLP Classifier. Then we ensembled the model with our top four models whose MCC score were greater than 0.7.

These models were used for ensemble.

# Overfitting