**II. Dataset & Feature Selection**

Our dataset is given by UCI machine learning library, initially procured and pre-processed by K.Fernandes and his team. It extricates 59 properties (as numerical qualities) portraying distinctive parts of every article, from a sum of 39644 articles distributed in the most recent two years from Mashable site. The dataset contains data around 39644 articles distributed by Mashable in a two-year period, from January 7 2013 to December 28 2014. The area of interest is the quantity of shares for every article, which demonstrates the level of its general prevalence crosswise over varied groups. The information likewise accompanies 61 highlights, either numerically or categorical. A few of these components ought not out of the ordinary as imperative elements adding to the prominence of an article. The sum total of what components have been classified into 3 sorts: number - a whole number in this setting, proportion - a float number inside the scope of [0, 1], boolean variable and ostensible nominal variable with the typical 1-of-C encoding.

The full list of capabilities is basically ordered as in the below mentioned Table.

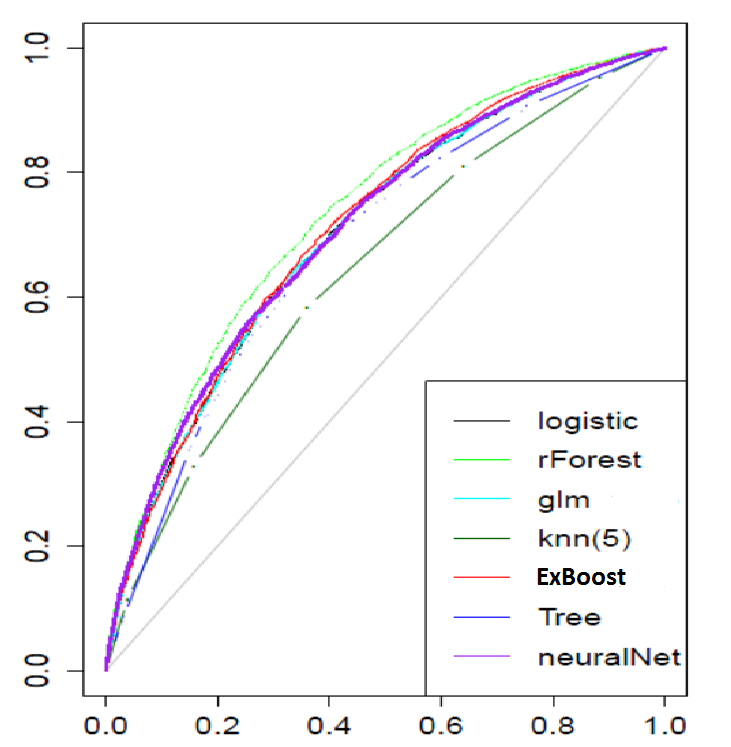
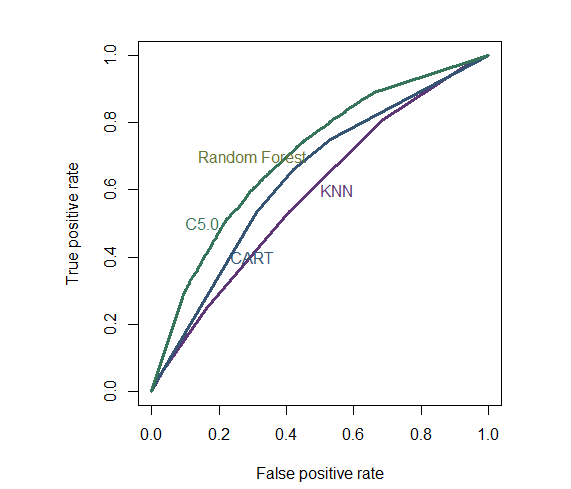


Dimensions: 39644 \* 58

Conclusion: Finding news popularity with respect to number of shares   
shares>1400 -- Popular  
shares < 1400 -- Not popular

For feature selection analysis, I considered first 30,000 rows as the train subset and next 9644 rows as test subset. As a pre-processing part we don’t need the **url** and **timedelta**, since they are not significant. Also, I replaced variable named **weekday\_is\_** with **weekday\_name**. This way I eliminated 6 columns as well got rid of some other unusual information. In the below table we can see the analysis done and can conclude that Random forest and eXtreme boosting will perform well on this dataset for predicting the shares. I used R for this purpose.

|  |  |  |
| --- | --- | --- |
| **Classifiers** | **Accuracy** | **Kappa** |
| Logistic Regression | 62.78% | 39.51% |
| Neural Networks (15 hidden Units) | 61.76% | 38.55% |
| Decision Tree | 63.52% | 28.78% |
| Naive Bayes | 62.87% | 11.79% |
| Generalized Linear Model | 65.40% | 30.26% |
| Boosted Logistic Regression | 60.84% | 30.22% |
| eXtreme Gradient Boosting | 67.04% | 21.19% |
| kNN(k = 5) | 56.34% | 12.56% |
| CART (Classification and Regression Trees)(Trials = 5) | 61.92% | 23.98% |
| C5.0 | 65.14% | 30.51% |
| Random Forests | 66.12% | 32.24% |



Also from the above ROC curves we can see that Random Forests and ExBoost is performing

sufficiently well when we consider Area under the ROC curve.