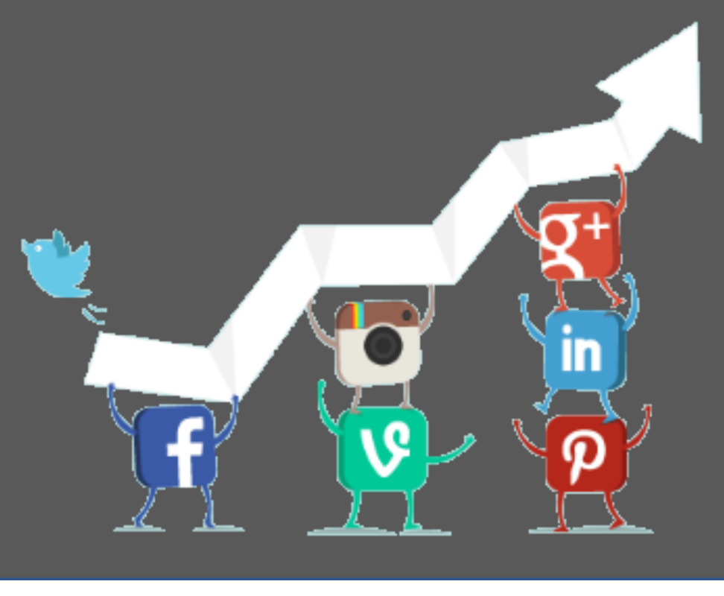


CPSC-6300 Applied Data Science

ONLINE NEWS POPULARITY

Shivam Pandit, Nandini Krupa Krishnamurthy, Rohit Chundru

Goal

The Project aims at finding the best classification model that fits the *Online news popularity* dataset. The chosen methods for fitting the model were selected incrementally for increasing model accuracy.

- K-Nearest Neighbors (KNN)
- Classification and Regression Tree (CART)
- Naive-Bayes (NB)
- C5.0

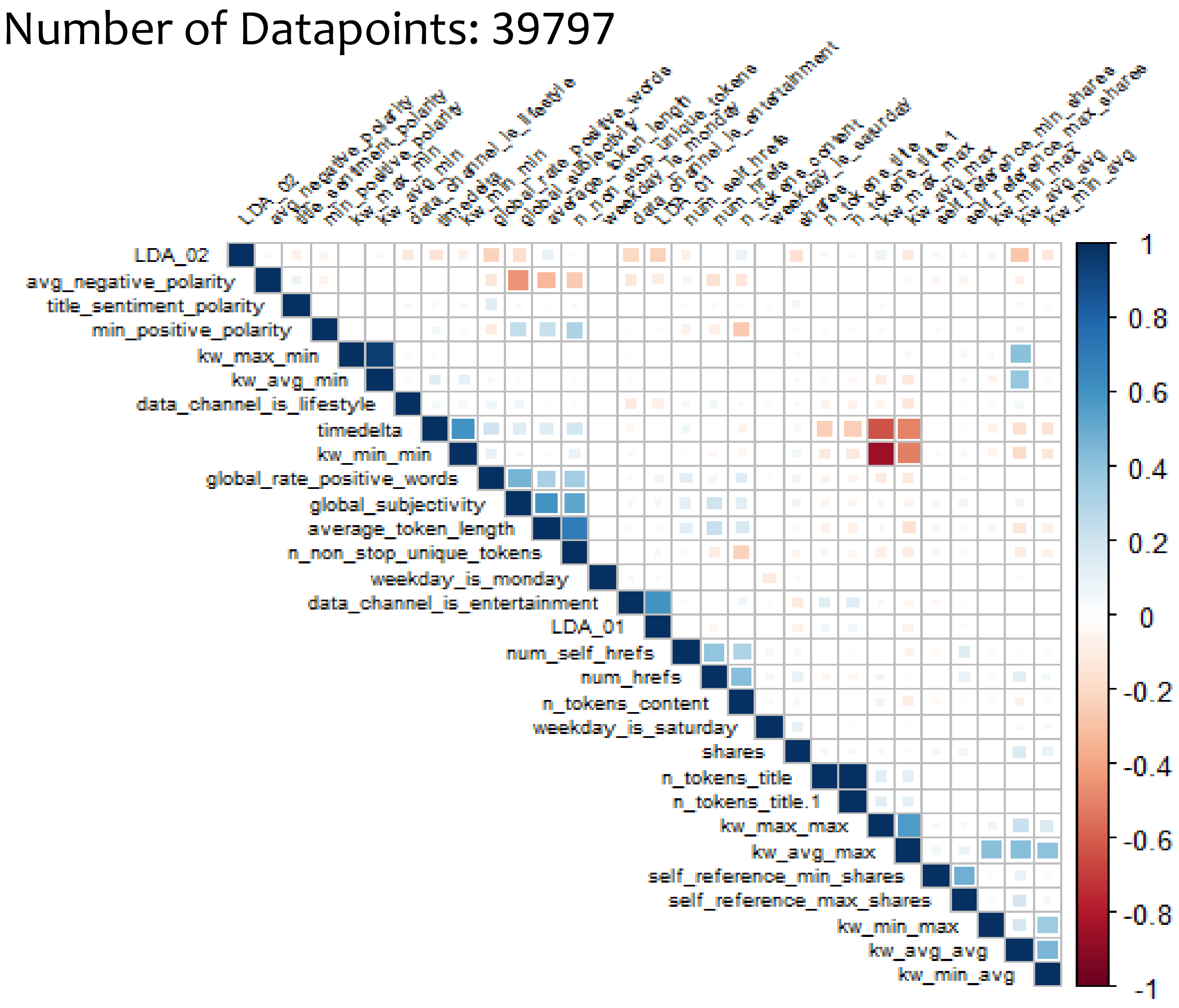
Approach

- Exploratory Data Analysis (EDA) is performed to clean data and remove outliers.
- Forward stepwise regression is used for variable selection which gave us 30 predictive variables.
- Supervised classification is performed on 30 variables that were initially 61.
- Since shares field had skewness, log transformation is used.
- Data is split into Test and Training sets.
- Naive Bayes is implemented.
- KNN is implemented.
- CART is implemented.
- C5.0 is implemented.
- Accuracies of implemented models is compared.

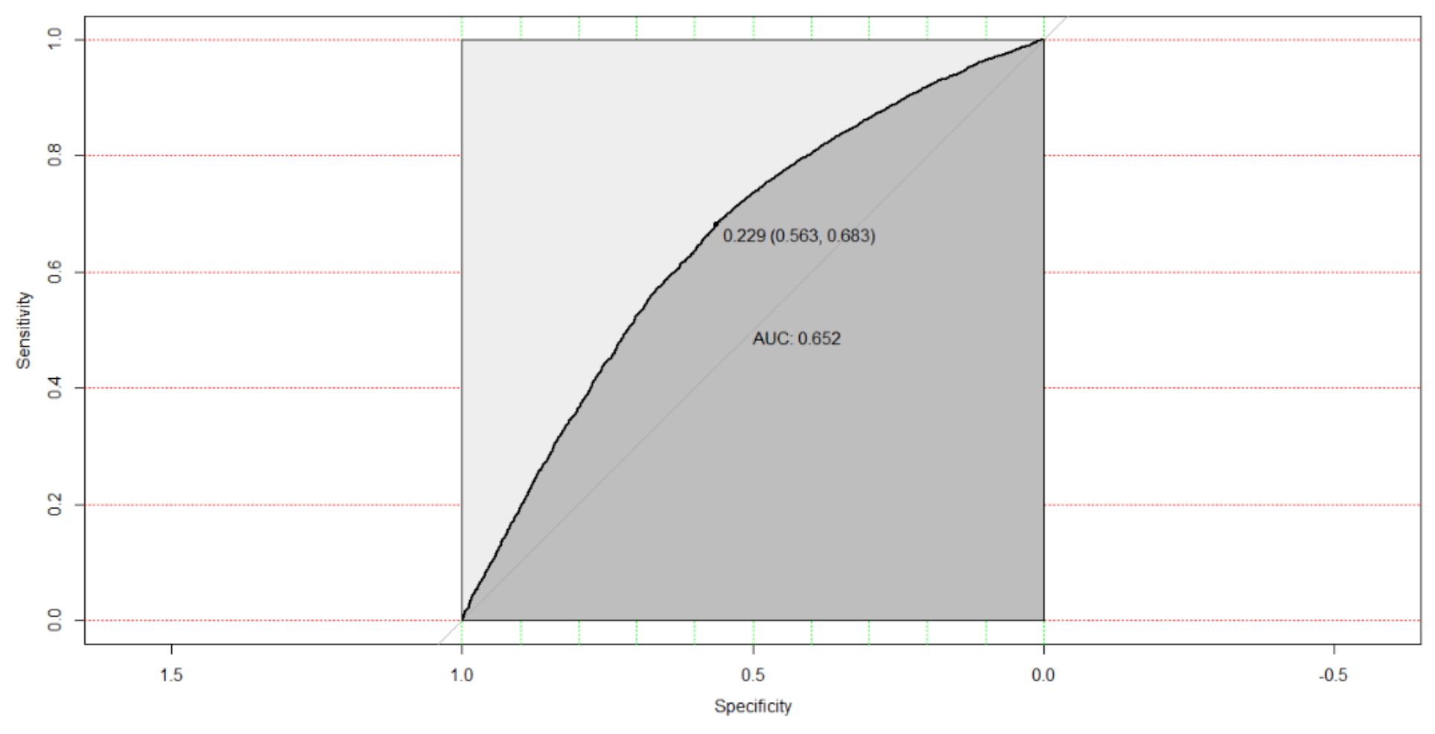
Dataset Information

The dataset – Online News Popularity – is available at https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#. It possess a set of features about articles published by *Mashable* in a period of two years.

- Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)
- Number of Datapoints: 39797

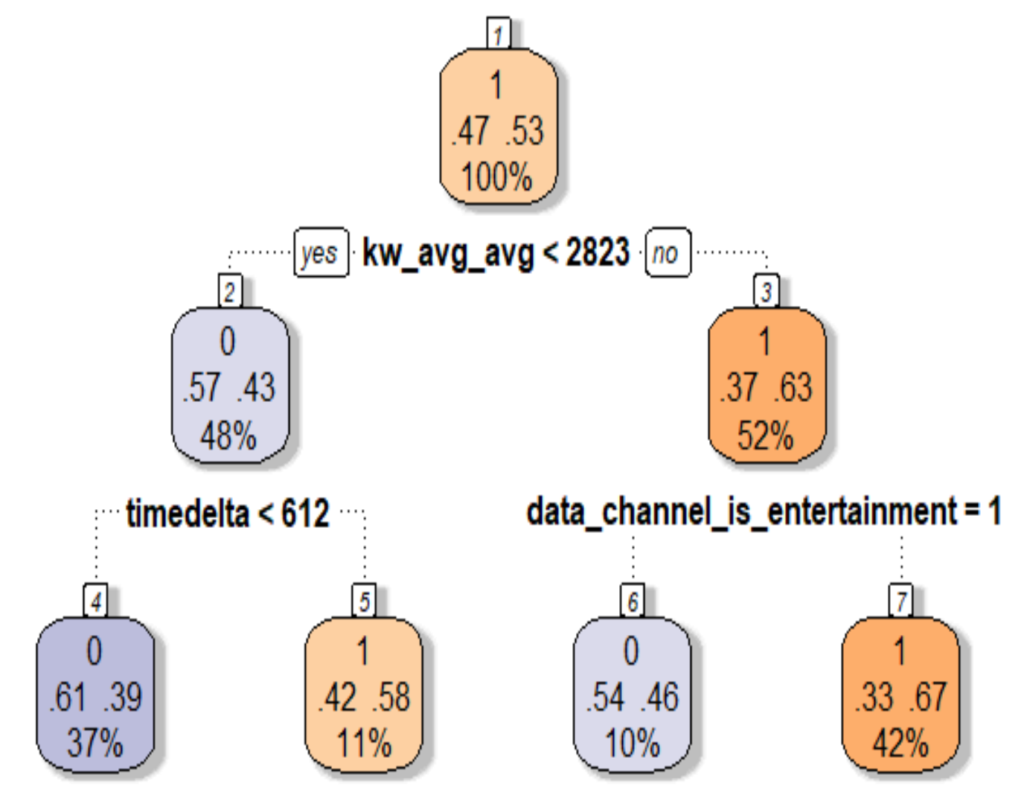


NB



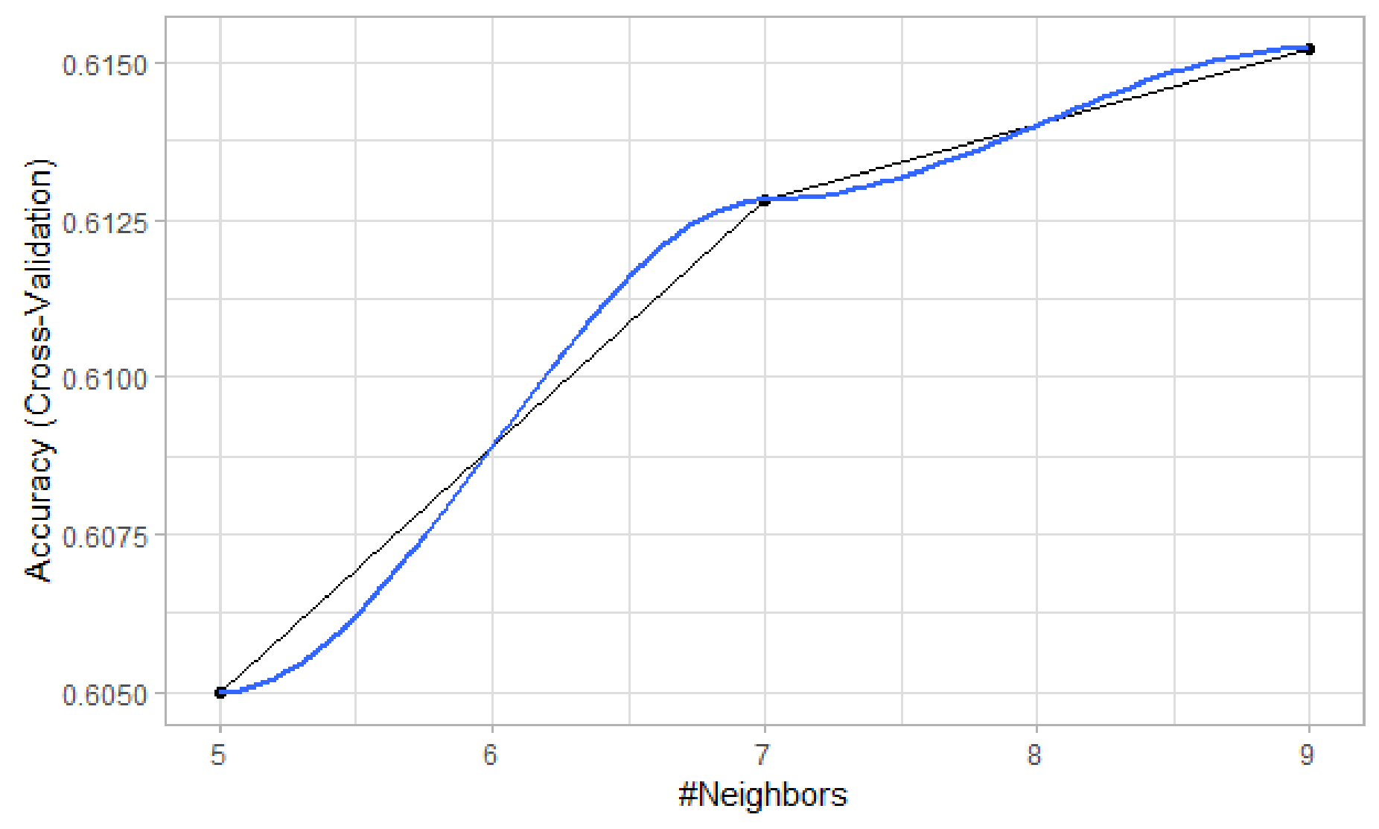
Confusion Matrix		
	Predicted No	Predicted Yes
Actual No	4365	4008
Actual Yes	1243	3906
Sum	5608	6284

CART



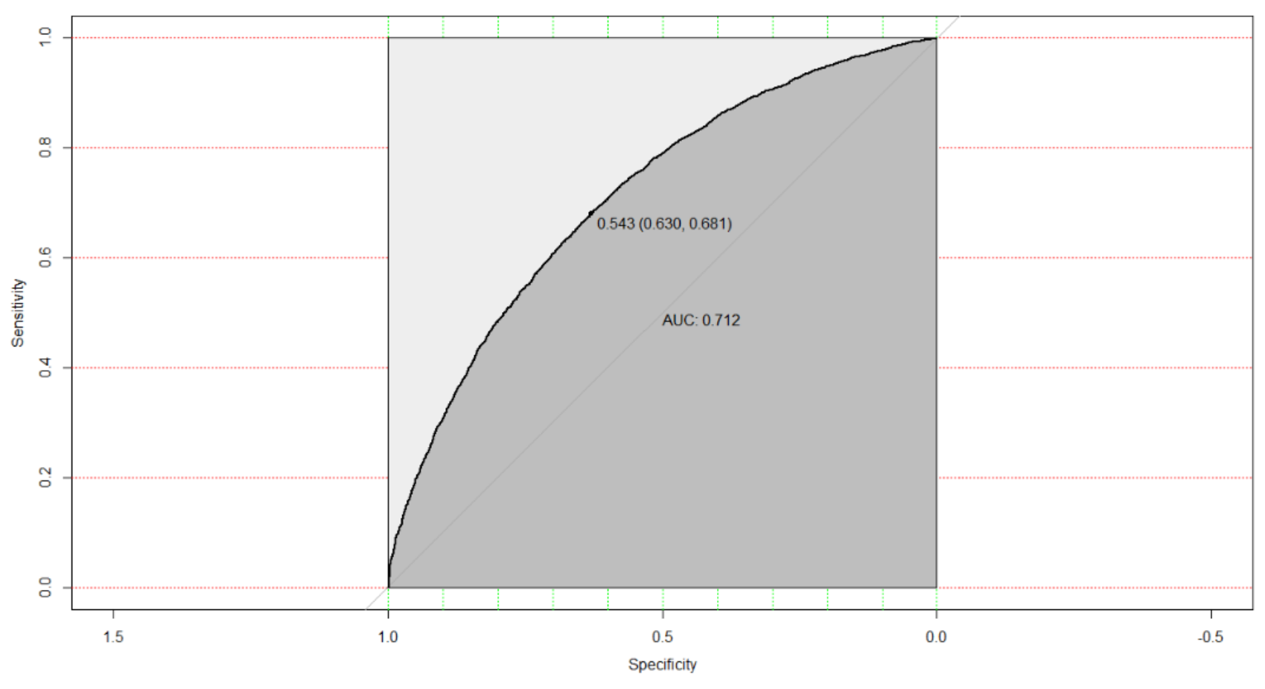
Confusion Matrix		
	Predicted No	Predicted Yes
Actual No	3513	2378
Actual Yes	2095	3906
Sum	5608	6284

KNN



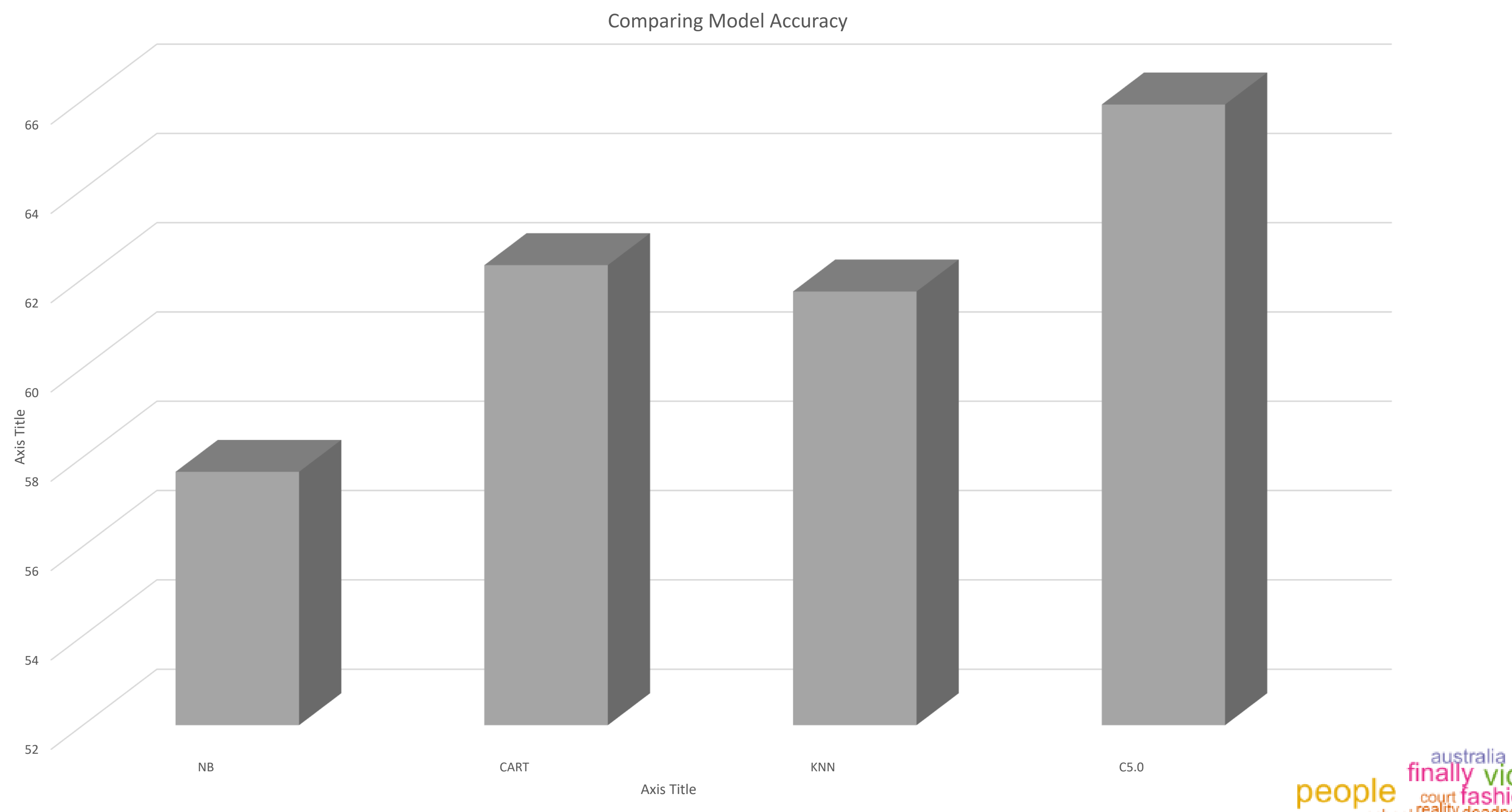
Confusion Matrix		
	Predicted No	Predicted Yes
Actual No	3341	2286
Actual Yes	2267	3998
Sum	5608	6284

C5.0



Confusion Matrix		
	Predicted No	Predicted Yes
Actual No	3129	1819
Actual Yes	2356	4589
Sum	5485	6408

Accuracy / Model Performance



Observations and Conclusion

- Out of the four methods implemented, C5.0 is the best to fit the model. This gives highest accuracy of 65.9%.
- The data set on a whole gives average accuracy of 61.895% which shows that the dataset is inconsistent indicating that irrelevant information been used.
- Therefore, this data is insufficient to predict the number of shares for a news article considering its popularity.