

WEB AND SOCIAL MEDIA ANALYTICS

GROUP ASSIGNMENT

TEXT ANALYTICS ON SHARK TANK EPISODES

M. PREM MURALI

M. ANNAPOORANI

MANISH MURALIDHARAN

MANIKANDARAJ

PROBLEM STATEMENT:

A dataset of Shark Tank episodes with 495 entrepreneurs making their pitch to VC sharks is given and an analysis is to be done using the description given by the entrepreneurs and the important factors are to be found out using text mining and other classification techniques.

DATA EXPLORATION AND CLEANING:

The data containing all 495 records of entrepreneurs are loaded and the initial exploration of data is done. As explained in the problem statement, only the description field is to be used for the initial analysis and hence the corpus for text mining is built using the description field of the data and the analysis is done.

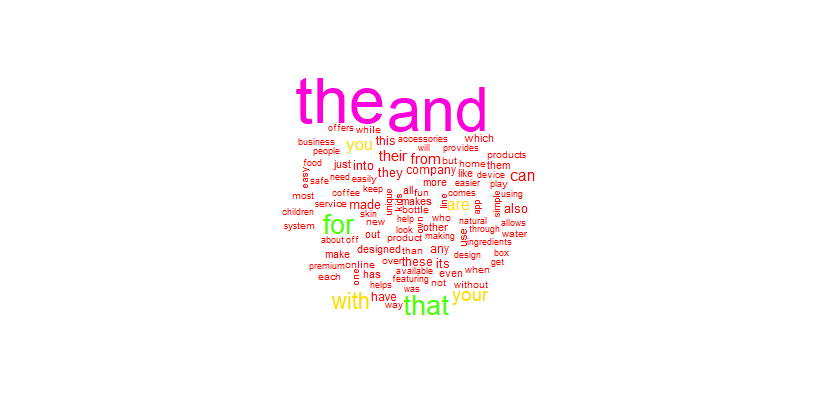
For cleaning up of the data, below functions are used to clean up the description text for analysis.

To apply the text mining techniques, a corpus has to be built using the description content available in the data. Then the below mentioned clean up functions are applied to the built corpus.

*To lower:* This function is used to change all the characters to lower case in the description data.

*Remove Punctuation:* Punctuation available in the description field has been removed using the function in the corpus.

*Word cloud of the description data after the initial clean up:*



Looking at the word cloud above, it seems the obvious words such as “*the, and, for, with, that”* are appearing mostly which is common since the description will have more words such as this and is to be removed from the corpus for text mining and analysis to find the frequency of keywords appearing in the text.

*Removing words:* Using the stop words function in text mining library, words which are not to be included as part of text mining but appearing frequently in text. There is a common dictionary in R library which removes the common words from the text. The words “*the, and”* are added to the library and the corpus is cleaned up by removing these stop words from the description field in the data.

*Strip White space:* Double space available in the description text is removed using the strip white space option available in the text mining library.

*Stemming:* Stemming is used to reduce the word to its word stem so that the words can be grouped together with its root version which is helpful for text analysis.

The final word cloud after applying all the cleanup functions is shown below.

*Word cloud of the description data after applying clean up functions:*



*Document Term Matrix:*

Once the clean up functions are applied, Document Term Matrix is built using the Document Term Matrix function in R; it describes the frequency of terms that occur in a collection of documents. Rows correspond to documents in the collection and columns correspond to terms.

*Removing sparse terms:* Sparse terms are the terms which have very less frequency in the entire corpus and they are removed in the corpus using the sparse function.

The corpus is now converted to a data frame with the words appearing as the column names and then the dependent variable as mentioned in the problem statement *deal* is chosen and further the models requested are built.

Results of the Document Term matrix showing the non sparse entries and the maximal term length is shown below.

<<DocumentTermMatrix (documents: 495, terms: 3506)>>

Non-/sparse entries: 9600/1725870

Sparsity : 99%

Maximal term length: 21

Weighting : term frequency (tf)

Out of the 495 entries, Deal as the dependent variable is appearing in the data matrix and results are shown below

FALSE TRUE

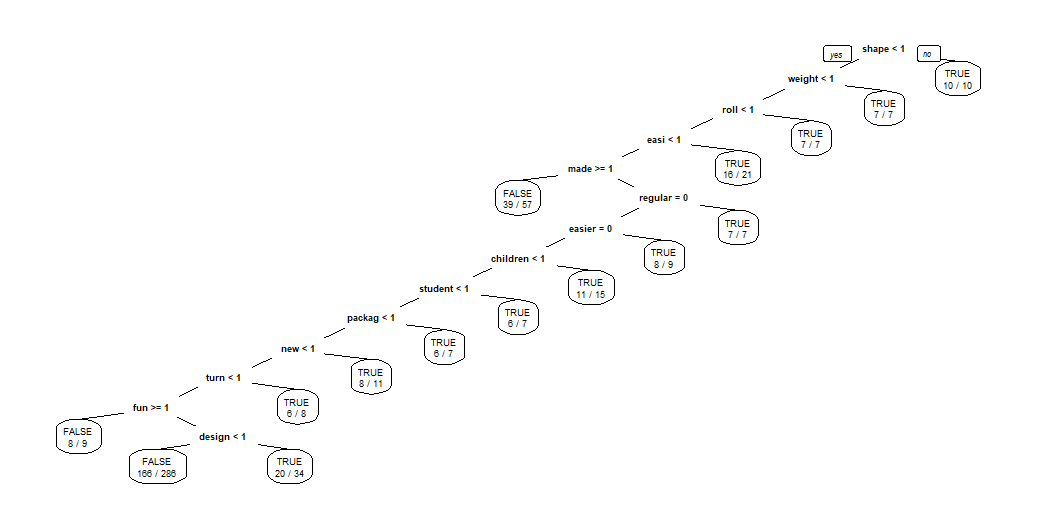
244 251

Using the final data frame, CART, Logistic Regression, Random Forest models are built initially and also ratio is calculated by dividing the asking for column by valuation column. Models are built before and after adding the ratio column to the final data frame and accuracy is compared to identify the best suitable model.

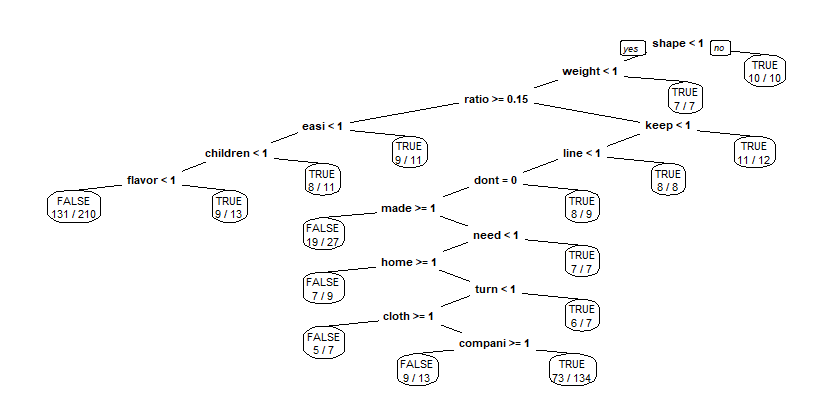
RESULTS:

CLASSIFICATION TREE (BEFORE AND AFTER):

BEFORE:



AFTER:



CONFUSION MATRIX OF THE MODELS:

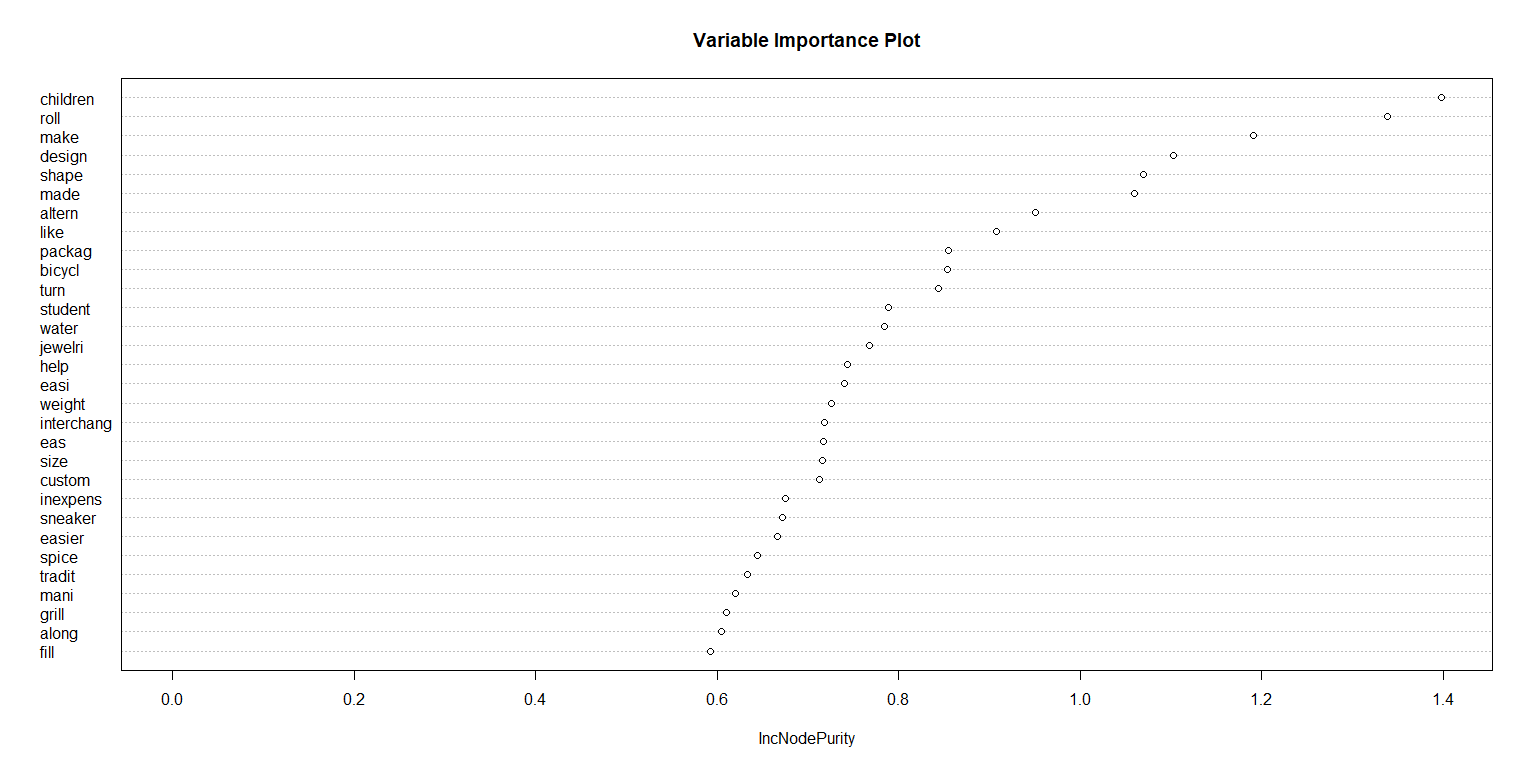
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CART MODEL | |  |  |  |  |  |
| BEFORE |  |  |  | AFTER |  |  |
|  | FALSE | TRUE |  |  | FALSE | TRUE |
| FALSE | 213 | 31 |  | FALSE | 171 | 73 |
| TRUE | 139 | 112 |  | TRUE | 95 | 156 |
|  |  |  |  |  |  |  |
| LOGISTIC REGRESSION | | |  |  |  |  |
|  |  |  |  |  |  |  |
|  | FALSE | TRUE |  |  | FALSE | TRUE |
| FALSE | 244 | 0 |  | FALSE | 244 | 0 |
| TRUE | 1 | 250 |  | TRUE | 0 | 251 |

ACCURACY OF THE MODELS:

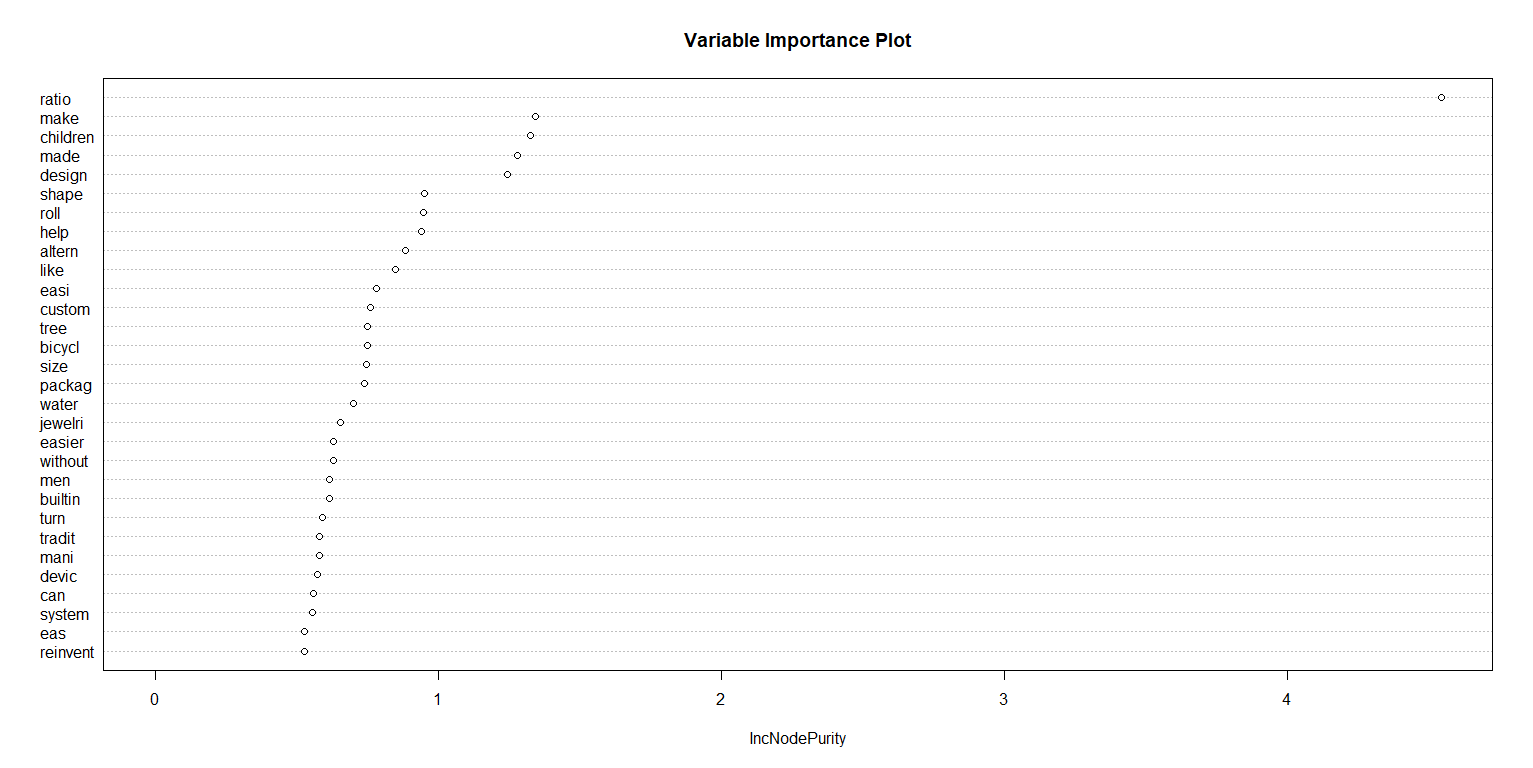
|  |  |  |
| --- | --- | --- |
| ACCURACY | BEFORE | AFTER |
| CART MODEL | 0.656566 | 0.660606 |
| LOGISTIC REGRESSION | 0.99798 | 1 |
| RANDOM FOREST | 0.53535 | 0.55757 |

VARIABLE IMPORTANCE PLOT (BEFORE AND AFTER):

BEFORE:



AFTER:



INTERPRETATION OF THE RESULTS:

* Text mining is done on the given data and the non-sparse entries available in the data is very less that indicates the terms are repetitive in most of the descriptions and they do follow a pattern.
* From the models built, we are able to understand better on the significance of analyzing the text data and here deal is considered as the dependent variable and for that the significant parameters and the results of the model are obtained.
* Logistic regression model gives an overall accuracy of 99% before including ratio as a variable and 100% after including ratio as a variable. Even though the accuracy is great, this cannot be considered as a suitable model for the dataset since this is overfitting of the model and is always problematic.
* Random Forest model gives an accuracy of 53% before including ratio as a variable and 55% after including ratio as a variable. This can be improved further for more accuracy by refining the model using grid search function.
* CART model gives an accuracy of 65% before adding the variable ratio and gives an accuracy of 66% after including the ratio column in the model. Out of the 3 models, this seems to be sophisticated model with the accuracy being good and also the model is sustainable. It can also be improvised further by refining the model.
* Variable importance plot of the Random Forest model shows that the word children is of most significance in the description which gives an insight and after adding ratio as the variable it is the most important parameter and can be considered.
* By using other dependent variables, we will be able to analyze the text data better and provide further significant results.