

**Random Forest Project - Predicting Popularity of Online News**

[Document subtitle]



[Date]

[Company name]

[Company address]

**Group Members**

Zarafsheen Irfan (2968)

Eihtisham Amjad ()

Sawaira Arooj (2980)

Table of Contents

[Introduction: 1](#_Toc103949495)

[Data Preparation: 1](#_Toc103949496)

[Reasons for feature selection, model parameters or hyper parameters: 3](#_Toc103949497)

[Feature Selection: 3](#_Toc103949498)

[Model Parameters: 3](#_Toc103949499)

[Model Hyper parameters 4](#_Toc103949500)

[Training and Hyper Tuning of Parameters: 4](#_Toc103949501)

# Introduction:

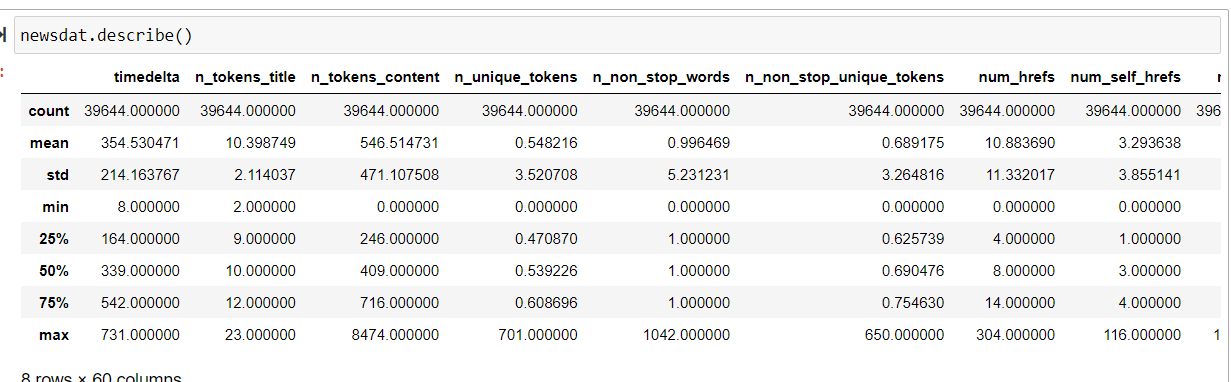
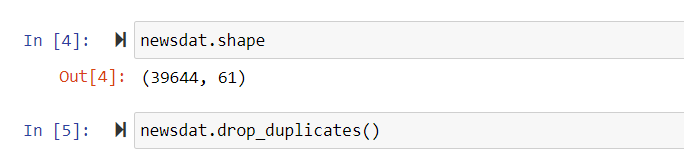
With the help of Internet, the online news can be instantly spread around the world. Most of peoples now have the habit of reading and sharing news online, for instance, using social media like Twitter and Facebook. Typically, the news popularity can be indicated by the number of reads, likes or shares. For the online news stake holders such as content providers or advertisers, it’s very valuable if the popularity of the news articles can be accurately predicted prior to the publication. Thus, it is interesting and meaningful to use the machine learning techniques to predict the popularity of online news articles. Various works have been done in prediction of online news popularity. Popularity of news depends upon various features like sharing of online news on social media, comments of visitors for news, likes for news articles etc. It is necessary to know what makes one online news article more popular than another article. Unpopular articles need to get optimize for further popularity.

# Data Preparation:

As early in Data Preprocessing we analyzed that our data has no missing values and no duplicate values but we have some outliers which has been found out in or exploratory analysis. But processing the data for our 2 models has been different. For Model1 there is no change in data as we have the whole data in numerical so need not to be having a one hot encoding but as the range of the data in different columns in very variance manner we will perform normalization which we will discuss in our later section of report. But in Second model where we need to classify the articles into different categories we have choosed neural networks and need to change the every different categories into one category and needed to perform one hot encoded which will finally place into a vector of one column. Then we performed cross-tabulation of every category in our data set and found out the distribution of categories in our data set. The below table shows the number of articles placed in each category:

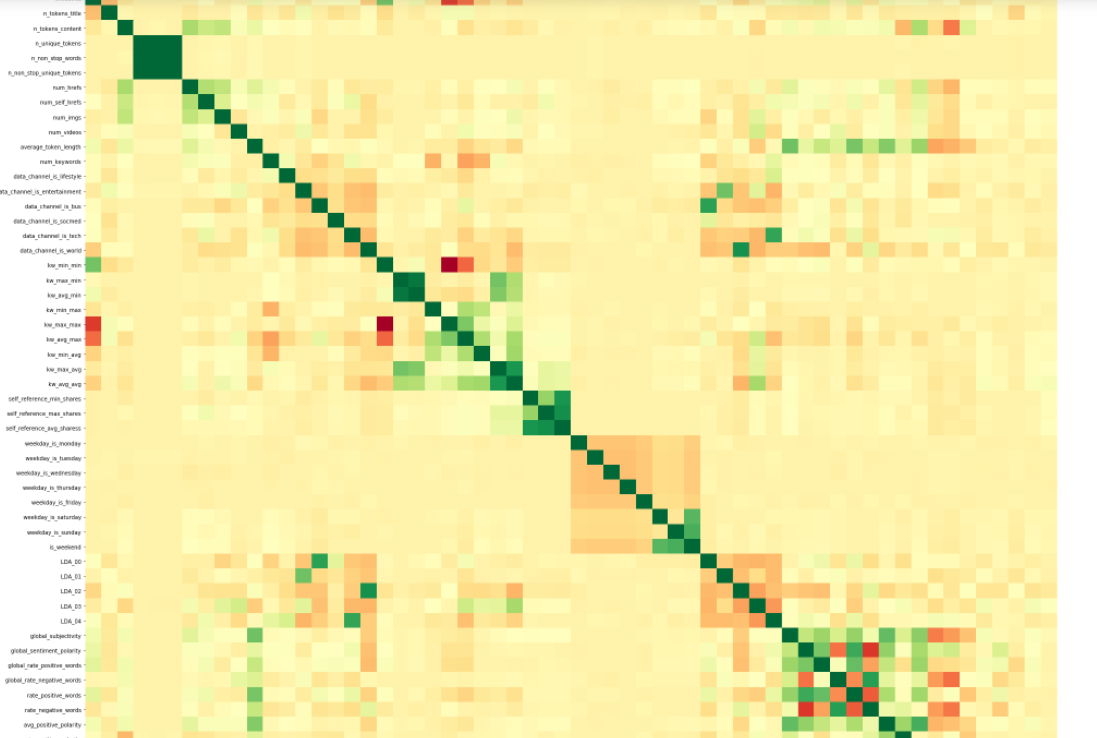
Category Number of articles in that category World 8427 technology 7346 Social media 2323 Entertainment 7057 Lifestyle 2099 Business 6258

As we can observe the data classification if categories with Social media and life style has less variety of assumptions. We are dropping time delta and URL which is metadata of articles which won’t be any use in our regression model.

Reason for choosing the algorithm:

For this project we have choose Random forest Regression as our Model. Because as we observed in Data Preprocessing we had Low bias and by Descriptive Statistics we had observed High variance of data and above that we have some anomalies which we need to consider and handle it. By considering the above all conditions we have selected Random forest regression as predicting our shares.

# Reasons for feature selection, model parameters or hyper parameters:

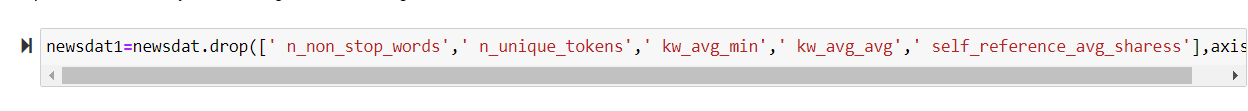


## Feature Selection:

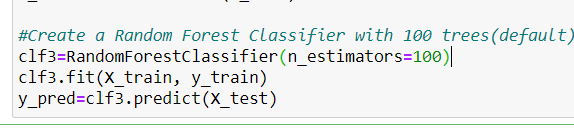
By the above correlation graph we can clearly say that these features: • number of unique words and number of non-stop-words and number of non-stop-unique tokens • Kw-avg- min and kw-max-min These are strongly correlated and linearly dependent which makes us to assume that these features are so linearly dependent that any one of the strong correlated feature can be used and excluding the other features won’t affect the model and will be indirectly helpful in our model by not allowing to do overfitting.

## Model Parameters:

We dropped following parameters because they were least co related with shares.



## Model Hyper parameters



Estimators with highest accuracy is 100.

# Training and Hyper Tuning of Parameters:

After performing feature engineering we train our model and evaluate our model with different min samples split which has been done by keeping in the loop and recorded the Root mean square error value. The below figure give us at what min\_samples\_split we get our efficient model that is less RMSE.

