**Exploratory Data Analysis: Steps taken and observations**

|  |  |
| --- | --- |
| **Step taken** | **Observation** |
| Check the structure of the train dataset | 28786 rows and 61 columns found |
| Check the data types of the train dataset | No categorical variables found. Label Encoding is not required |
| Check if there are any missing values in any columns | No missing values were found in train (and test) |
| Check the nature of the target column | Target was heavily right skewed and hence log was taken |
| Explore few variables to understand their correlation with the target:   * n\_tokens\_content * num\_imgs * num\_videos * num\_hrefs | * The more the words in the document, lesser the shares (negatively correlated). * More images are resulting in more shares but beyond 20 images is not resulting in increased shares. Not a strong linear correlation. * Just like images, shares increase with videos but, beyond 10-12 videos is resulting in increased shares. Not a strong linear correlation. * Same for hyperlinks. Beyond 10 hyperlinks are not yielding increased shares even though till 10 it is positive correlation. Overall, not a strong linear correlation. |
| Check the distribution of articles on number of words in title | Most articles have title around 10 words. Range is 5-15 |
| Check the distribution of articles on weekdays Vs weekends | More articles are published during weekdays than weekends |
| Check the correlation between the features and the target | Linear correlation between the features and target is very low |
| Plot the histogram of all the features to check their distribution | Heavily right skewed features found: n\_tokens\_content, num\_hrefs, num\_self\_hrefs, num\_imgs, num\_videos  Log transformation done to make them more normally distributed (log10) |

**Note:**

1. There is no strong linear correlation between the features and the target. Hence tree-based models would perform better than linear models.
2. Standardization of data was done using **MinMaxScaler** as the columns have very different scaling (some are ratios, some are absolute values).
3. **RandomForestRegressor** was used to find the feature importance. ‘n\_non\_stop\_words’ did not have good ‘importance’ score and hence it was not used for model training.