Machine Learning

News Popularity Prediction Explanatory Data Analysis

Mobin Nesari
Dr. Hadi Farahani
March 12, 2023

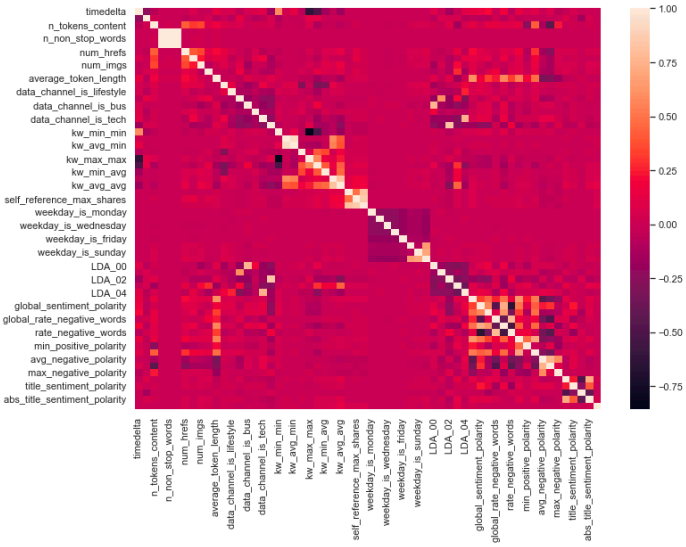
Introduction:

The news popularity prediction dataset contains information on news articles and their social media engagement. This dataset is widely used by researchers and machine learning practitioners to develop models that can predict the popularity of news articles based on a range of features. By analyzing this dataset, researchers can gain insights into the factors that contribute to the success of news articles on social media and improve their ability to predict which articles are likely to go viral. This database contains 60 columns and 39644 rows which represent each news record. Description of dataset is saved in 'description.csv' next this file.

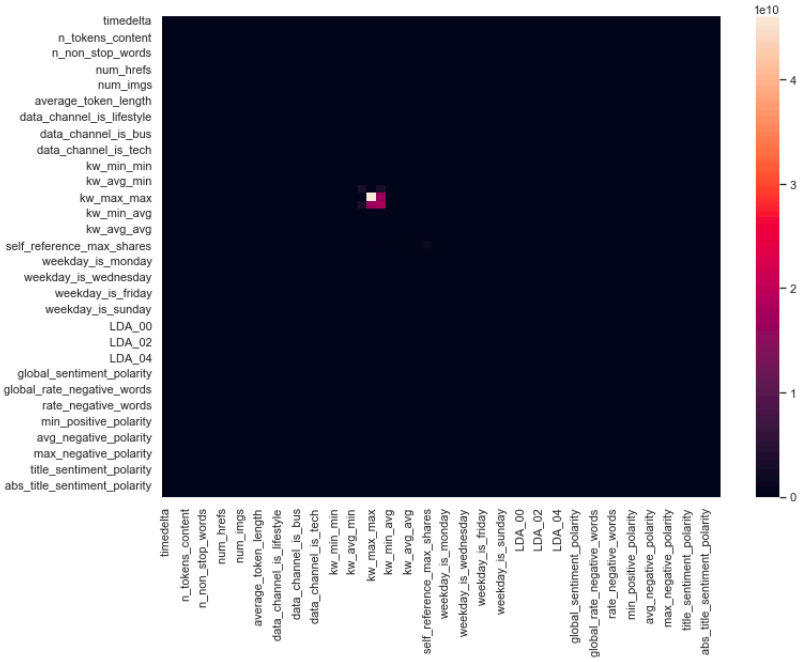
Missing Values:

Upon inspection, it was observed that the dataset does not contain any missing values. This is an important finding as missing data can cause issues when building predictive models and can lead to biased or inaccurate results. The absence of missing values indicates that the dataset is complete and can be used for further analysis and modeling with confidence.

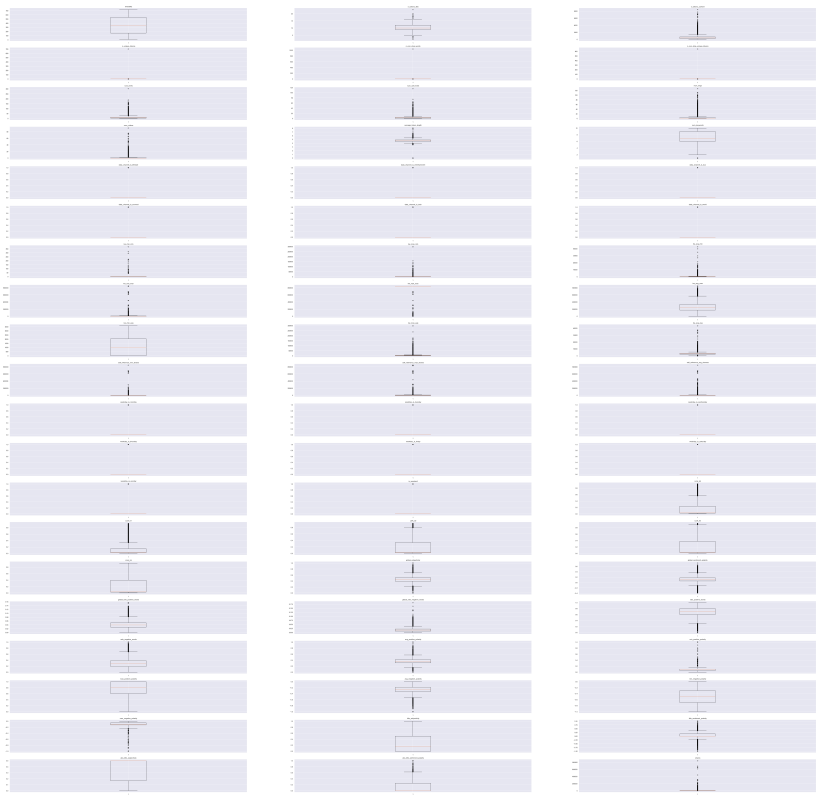
Correlation Matrix:



Covariance Matrix:

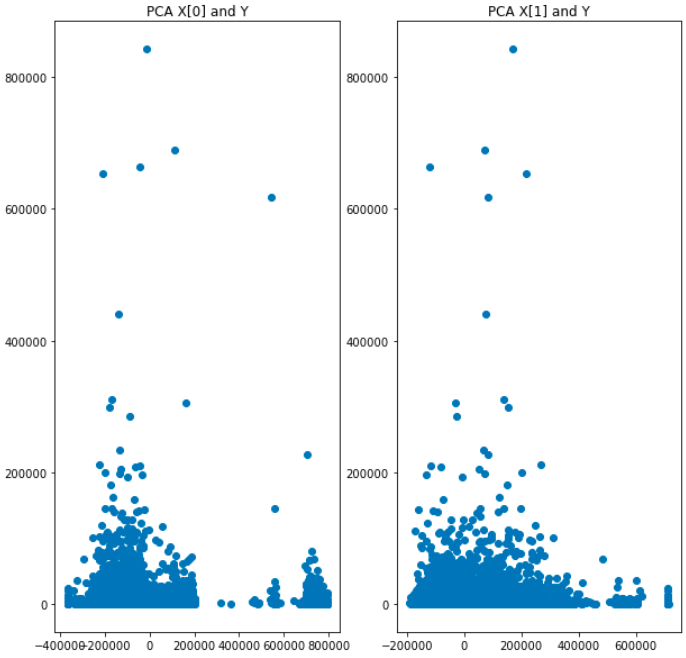


Feature Analysis:

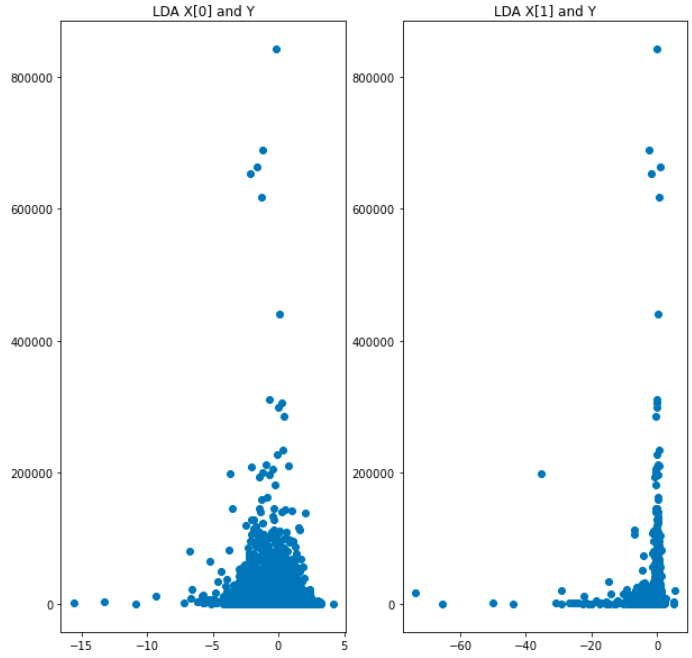


Dimensionality Reduction:

For this dataset, Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) have been applied on dataset and set to get only two components from 59 features.



PCA Dimension Reduction



LDA Dimension Reduction

Statistical Test:

Five hypothesis tests have been considered for this dataset. First test is one sample t-test with base hypothesis that mean of videos numbers is 2 for news. This test accepted with t-stat = -36.35866 and p-value = 9.3795×10^{-285} .

Second test is two sample t-test with base hypothesis that if mean shares of weekdays articles has significant difference from the mean number of share for news article published on weekends. This test also accepted by t-statistics = -3.3769 and p-value = 0.00073.

Chi-Square test is applied on dataset. Tested whether there is a significant association between popularity (shares >= 1400 vs shares < 1400) and the number of tokens in the content of news articles (binned into six categories). The test has been approved due to small value of p (7.6690×10^{-15}).

ANOVA test has been applied on this dataset. We tested whether there are significant differences in the mean number of shares among news articles with different numbers of images (0, 1, 2, 3, or 4+). F-statistics of this test is 59.7319 and p-value is 0.0002992.

Pearson's Correlation Coefficient Test is applied on this dataset. Tested whether there is a significant linear relationship between the number of tokens in the title and the number of tokens in the content of news articles. Correlation coefficient is 0.01815 and p-value. Is 0.0002992.