IE7300 Statistical Learning for Engineers SEC 03 Spring 2023

Online News Popularity

Group 6

Project Proposal Report

Professor: Ramin Mohammadi

Group Members: Phanindra Raja Varma Gadiraju
Venkata Sasank Jonnalagadda
Sai Sruthi Kalidindi
Henna Shah

Problem Setting and Definition:

The Online News Popularity dataset includes details about numerous articles that were posted over a two-year period on the well-known news and entertainment website Mashable. The dataset contains a wide range of information about each article, including the title, length, day published, article category and many others.

Our goal is to build a machine learning model that can forecast how many shares an article will likely get on social media sites. The popularity of the article or the quantity of times it has been shared on social media sites like Facebook, Twitter, LinkedIn, and Google+ in this case serves as the target variable.

In the context of internet content and media, the popularity of pieces can be predicted. Website owners, content producers, and marketers can better understand their audience's preferences and interests by knowing which articles are likely to garner more social media shares. The website's total reach and engagement can be increased by using this data to produce content that is more likely to be shared.

Attribute Information:

Here is the brief description of the Attributes:
Total number of attributes are 61 in which 58 predictive attributes, 2 non-predictive, 1 goal field.

Non-Predictive Attributes:

1. Url
2. Timedelta

Few Predictive Attributes:

1. n_tokens_title: Number of words in the title
2. n_tokens_content: Number of words in the content
3. n_unique_tokens: Rate of unique words in the content
4. average_token_length: Average length of the words in the content
5. global_subjectivity: Text subjectivity
6. global_rate_positive_words: Rate of positive words in the content
7. global_rate_negative_words: Rate of negative words in the content
8. title_subjectivity: Title subjectivity

9. weekday_is_monday: Was the article published on a Monday? (Similar columns with all the days in the week)

Target Attribute:

1. Number of shares

Since the target variable is of type continuous variable, we are implementing regression models. We are also planning to implement classification models by choosing a certain threshold and classifying if the article will have good reach or not.

Action Plan:

1. Visualizing Data to draw insights
2. Performing Exploratory Data Analysis
3. Model Implementation
4. Model Evaluation

Models which we are planning to implement are:

1. Linear Regression
2. Logistic Regression
3. Naïve Bayes
4. Decision Tree
5. Random Forest
6. SVM