***Project Proposal***

***Social Media Shares Prediction***

Shiva Chaitanya Goud Gadila

Kent State University

**Abstract**

The main aim of this project is to estimate the number of shares a social media post or article can get. This particular dataset is taken from the Machine Learning Repository of University of California, Irvine Online News Popularity Data Set (UC Irvine). It contains 61 columns and 39,644 observations including different factors like URL, number of tokens in the title and content, number of images, videos, audience reactions, comments and many more, which could significantly affect the engagement rate of content creators on social media.

Social media has become a huge part of our day-to-day life and a social media influencing/content creators’ job is the most in-demand job among youngsters. Every article, picture or video shared by these influencers earns them thousands of dollars and in such a case, it is important to track their engagement ratio including number of likes, comments, and shares.

This project gives a full scope to use both regression and classification techniques like multi-variate regression, decision trees, and other machine learning algorithms including clustering to explore and analyze the dataset thoroughly, and then implement a model to predict the number of shares of each article posted. Before proceeding to the analysis, the dataset will be examined to know the percentage of missing values and different techniques like imputation with Mean/Median/Mode or removing rows with null values will be performed and check for duplicate records, and errors would be done to ensure data quality. During the initial analysis, correlations among the variables will be checked to avoid columns having high correlation, distribution of data will be plotted to check if there is any skewness in the data and outliers will be identified to handle them in an effective way.

The dataset will be divided into training and test data, after performing data cleaning and exploratory data analysis. All numeric variables will be normalized before splitting the data. A suitable model will then be chosen to be developed on training data and after measuring the model’s performance using confusion Matrix, RMSE, R square, AUC etc. It will be deployed onto the test dataset on which the predictions of number of shares will be made. Even though the URL of each article is mentioned in the dataset, it will not be considered for analysis due to both confidentiality and complexity reasons. However, final predictions will be tied back to know the exact article which might receive such a number of shares.

Dataset Link: <https://archive.ics.uci.edu/dataset/332/online+news+popularity>

References:

<https://code.datasciencedojo.com/datasciencedojo/datasets/tree/master/Online%20News%20Popularity>