Project Plan

1. Explore Data
   1. Columns.
   2. Rows.
   3. Columns – Data Types, conversion if needed.
   4. Missing Values / Outliers check.
   5. Feature Engineering
      1. Eliminate features that are irrelevant for the prediction.
      2. Try to convert all other non-numeric features into numeric. The one’s which are not convertible by any means – eliminate.
      3. Post numeric conversion –
         1. Use Data-Visualization techniques to understand correlation and dependencies of the target.
         2. Plot Correlation Matrix to get an idea
   6. Data Preparation for model training
      1. Final feature selection
      2. Train-Test Split (80-20)
2. Base Model
   1. Train the first base – Decision Tree (Default settings)
   2. Feature importance analysis
   3. Performance metrics analysis –
      1. Root mean squared error.
3. Model 1: Linear regression
4. Model 2: Neural Network

Execution

* Total number of columns = 15 Features + 1 target column (As mentioned in the document provided)
  + Hence, eliminated first 2 columns.
* Changing the column name ‘Post Contet’ to ‘Post Content’ for better understanding.
* Handling Date-time
  + Setting Date-time column as the index. (Helps to understand time dependency of features using Time Series plots)
  + Converting data type to -> datetime64 using pandas
* Missing Values and Outliers Detection
  + No missing values found – All 15 column have 50k non-null values.
  + Outliers are practically not possible as a certain post/tweet might get viral and have extremely high retweets, likes, shares – Which actually is not an outlier.
  + Number of shares, likes, etc cannot be negative and it has been checked. No negative values found.
* Columns Data types
  + All columns are numeric except
    - Post Content
      * Len of the post (Already extracted)
      * Language of the post can be a good feature to extract. Langdetect module in python is used.
      * Number of words in the post.
      * Number of new lines (\n) used in the post.
      * Now the newly created 2 features are numeric.
      * 1 feature (language) is object but, will be label encoded.
    - Media Type
      * One-hot encode
* A new feature is added – Hour (Represents time (Hour) of the day)
* Post Processing: Number of Features = 18 + 1(Target)
* Base Model = DecisionTreeRegressor was trained.
  + The RMSE achieved was 3624
  + But the feature importance showed only 3 main features contributing to the prediction: Likes, Shares and Comments.
  + Training time = 440 ms
* Model 1 = Linear Regression
  + RMSE = 11.53 (With all features)
  + RMSE = 11.53 (With only 3 main features found in the feature importance) (Very minor change was observed which confirms that these 3 features are the most important)
  + Training time = 76 ms
* Model 2 = Shallow Neural Network
  + **Exponential Decay Scheduler** was used and tuned to get the best results in least time. Exponential Decay is used so that the learning rate is reduced as the loss moves towards the optimal small value.
  + **Callback** function for early stopping was used. The main motive is to restore the best weights.
  + Training time = 22 sec
  + RMSE = 37

Conclusion:

* + Linear Regression Performed the best.
  + Linear Regression also is the fastest with the training time 76 ms.
  + Best achieved RMSE: 11.53