Online NEWS Popularity

DA 204o: Data Science in Practice – Course Project

- **Rajat Chaudhary** rajatc@iisc.ac.in
- **Srinivas Shavukapu Kattegummula** srinivassk@iisc.ac.in
- **Sonu Goyal** sonugoyal@iisc.ac.in
- **Yuvaraj G** yuvarajgopi@iisc.ac.in

Image source: Internet

Table of contents

Motivation & Problem Statement

- **Problem Overview :**

  Media companies struggle to reliably predict which articles will go viral due to complex factors like keywords, sentiment, timing, and social signals.

- **Importance :**

  Accurate predictions drive optimized content strategies, boost engagement, increase ad revenue, reduce effort on low-impact articles, and enable data-driven editorial decisions.

- **Project Goals :**

  Develop a model to predict article popularity (shares or viral status) and identify key virality factors. Deploy a real-time prediction system for immediate use.

- **Data Science Solution :**

  Leverage machine learning and explainability methods to forecast viral content, uncover critical drivers, and provide actionable insights for editorial teams.

Data Collection and Preparation(1/3)

- Data source(s) (where it's from, how it was collected)
  - UCI Machine Learning Repository - Online News Popularity - UCI Machine Learning Repository
  - Extracted from Mashable articles published over a period, including metadata, content features and social engagement metrics.
- Description of the data (features, size, format)
  - 58 features, ~39k observations. Tabular data(structured)

Table 1. Description of Data

Attribute	Details
Size	Approximately 39,000 observations
Features	58
Feature Type	Integer and Real
Dataset Characteristics	Multivariate
Format	Tabular data

Table 2: List of attributes by category.

Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)
Article text polarity score and its absolute difference to 0.5	ratio (2)

Target	Type (#)
Number of article Mashable shares	number (1)

Data Collection and Preparation(2/3)

- Preprocessing
  - Train/test split (80/20)
  - Cleaned column names
    - ✓ Removed leading/trailing whitespaces
  - Dropped irrelevant columns
    - ✓ url , timedelta, kw_min_min(>60% outlier)
  - Outlier correction(Table 3)

Rate cannot be >1

COLUMN	ROWS > 1
n_unique_tokens	1
n_non_stop_words	1
n_non_stop_unique_tokens	1

bad data rows are :

	n_unique_tokens	n_non_stop_words	n_non_stop_unique_tokens
31037	701.0	1042.0	650.0

COLUMN	REPLACED	OLD MEAN	IMPUTED VAL	NEW MEAN
n_unique_tokens	1	0.5520	0.5299	0.5299
n_non_stop_words	1	1.0028	0.9700	0.9700
n_non_stop_unique_tokens	1	0.6930	0.6725	0.6725

Count cannot be < 0

COLUMN NAME	COUNT < 0	% OF TOTAL
kw_min_min	18399	58.01%
kw_max_min	0	0.00%
kw_avg_min	652	2.06%
kw_min_max	0	0.00%
kw_max_max	0	0.00%
kw_avg_max	0	0.00%
kw_min_avg	6	0.02%
kw_max_avg	0	0.00%
kw_avg_avg	0	0.00%

COLUMN	REPLACED	OLD MEAN	IMPUTED VAL	NEW MEAN
kw_max_min	0	1175.48	N/A	1175.48
kw_avg_min	652	316.06	322.72	322.72
kw_min_max	0	13578.15	N/A	13578.15
kw_max_max	0	753332.38	N/A	753332.38
kw_avg_max	0	259157.53	N/A	259157.53
kw_min_avg	6	1120.29	1120.50	1120.50
kw_max_avg	0	5681.05	N/A	5681.05
kw_avg_avg	0	3140.96	N/A	3140.96

Table 3 Outlier correction

Data Collection and Preparation(3/3)

- Targeted popularity
  - Multi-class classification problem using dynamic thresholds based on the training set to focus on virality rather than exact share counts.

Table 4. Popularity Levels Defined by Training Set Percentiles

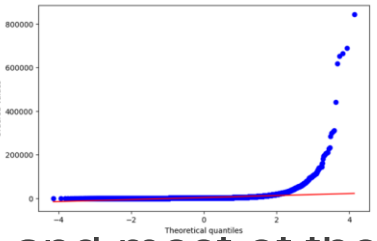
Class	Shares Range (based on training percentiles)
Niche	$0 \leq S < P_{50}$
Performer	$P_{50} \leq S < P_{95}$
Trending	$P_{95} \leq S < P_{99}$
Viral	$S \geq P_{99}$

Note:

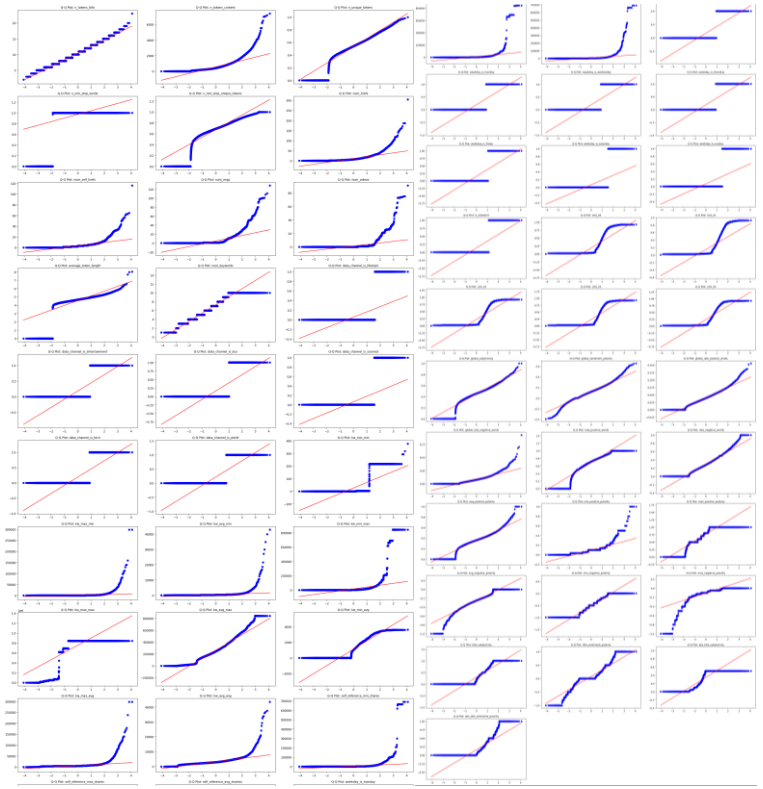
- After thorough analysis done with regression models, decided to map the shares as per industry standard classes (Refer report for Regression model evaluation and details)

Exploratory Data Analysis (EDA)

Target variable (Shares) : Shares column is highly skewed with a long tail, indicating that a small number of articles receive very high shares.

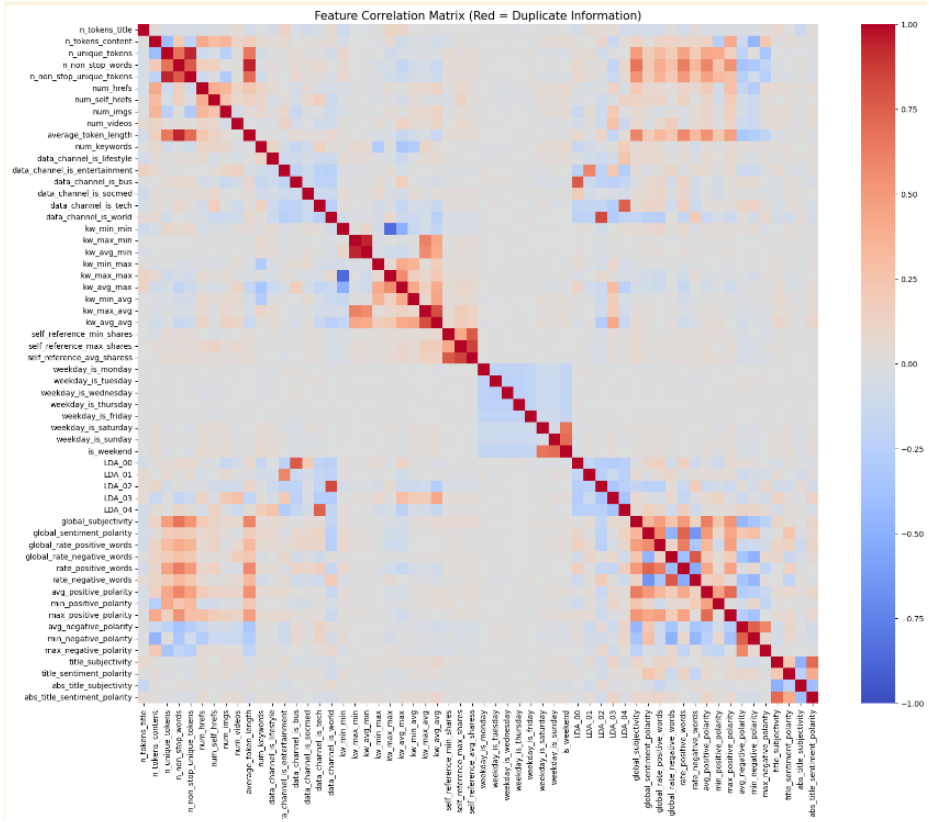


Very few features are having linear relationship and most of the features are non-linear, which directed us not use linear algorithms



Feature Engineering

- Correlation value:
  - Positive (> 0): An increase in this feature results in higher shares.
  - Negative (≤ 0): An increase in this feature leads to fewer shares.
  - The dark red box close to the diagonal indicates that these features are duplicates of each other.



Evaluation strategy

- The main objective is to predict popularity (**viral/trending**), so we have focused on a macro-level evaluation approach rather than overall model accuracy.
- The **primary** evaluation metric used is **hit rate** (predicted value versus actual value).
- **Secondary** evaluation metrics, including **accuracy, precision, and F1 score**, are considered for the full model to further fine-tune its performance.

Model(Algorithms)

- Regression (Explored to check the adaptability of model) *
  - Random Forest Regressor
  - Gradient Boosting Regressor
  - Lasso Regression
  - Support Vector Regressor (SVR)
  - K-Nearest Neighbors Regressor (KNN)
  - Multi-layer Perceptron Regressor (MLP)

Model	MAE	MSE	RMSE
Random Forest Regressor	0.6839	0.8190	0.9050
Gradient Boosting Regressor	**0.6529**	**0.7474**	**0.8645**
Lasso Regression	0.7077	0.8573	0.9259

Model	CV Score (Neg MSE)	MAE (Original Scale)
SVR	-0.8138	2333.99
KNN	-0.8525	2439.32
MLP	-0.7950	2362.38

- Classification (Final model)

Table 5. Classifier Performance Summary (All Models)

Classifier	Accuracy	Category	Total Actual	Correct Hits	Hit Rate (%)	Precision	F1-Score
XGBoost (Dynamic Tiers)	**58.57%**	Niche (< 50%)	4075	2633	64.61%	67.63%	0.6609
		Performer (50–95%)	3466	1945	56.12%	58.13%	0.5711
		Trending (95–99%)	318	62	19.50%	9.75%	0.1300
		Viral (> 99%)	70	4	5.71%	7.41%	0.0645
LightGBM (Balanced)	**62.27%**	Niche (< 50%)	4075	2798	68.66%	65.88%	0.6724
		Performer (50–95%)	3466	2133	61.54%	58.57%	0.6002
		Trending (95–99%)	318	6	1.89%	15.38%	0.0336
		Viral (> 99%)	70	0	0.00%	0.00%	0.0000
Random Forest (Balanced)	**63.12%**	Niche (< 50%)	4075	2739	67.21%	68.03%	0.6762
		Performer (50–95%)	3466	2264	65.32%	58.14%	0.6152
		Trending (95–99%)	318	2	0.63%	28.57%	0.0123
		Viral (> 99%)	70	0	0.00%	0.00%	0.0000
AdaBoost (Balanced Tree)	**42.92%**	Niche (< 50%)	4075	2353	57.74%	64.75%	0.6105
		Performer (50–95%)	3466	918	26.49%	55.43%	0.3585
		Trending (95–99%)	318	94	29.56%	6.47%	0.1062
		Viral (> 99%)	70	38	54.29%	3.20%	0.0605
Naive Bayes (Transformed)	**57.76%**	Niche (< 50%)	4075	3327	81.64%	58.66%	0.6827
		Performer (50–95%)	3466	1223	35.29%	61.71%	0.4490
		Trending (95–99%)	318	25	7.86%	11.79%	0.0943
		Viral (> 99%)	70	5	7.14%	7.94%	0.0752

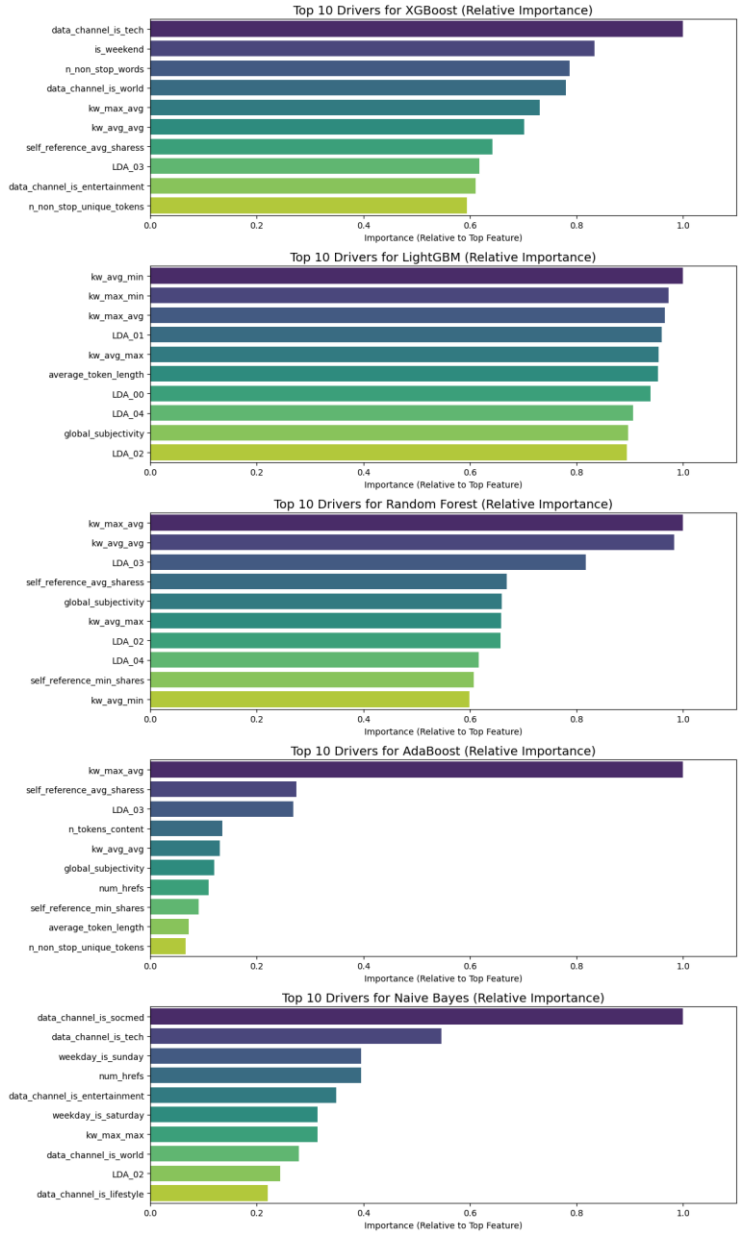
- Used OPTUNA for hyperparameter tuning.

*Note:

- After thorough analysis done with regression models, decided to go for Classification model. Please refer report for more details.

Insights

- Because NEWS can spread rapidly for multiple reasons, we analyzed the relative importance of feature contributions to identify the top contributing feature in each model.
- Based on the contribution scores, kw_avg (content keyword) is a key factor influencing NEWS virality.
- Secondary features are Global subjectivity, weekend, type of media(social and entertainment)



Limitations and Future improvements

- Overall model accuracy is low because the features for NEWS/Social media popularity will be very dynamic. We will adapt the advanced text embeddings to richer semantic understanding of article/content.
- Apply time series or sequence models.

Thank You