Online News Popularity Prediction

Aditya Jain - 2021511
Shrey - 2021562
Vasu Kapoor - 2021573
Vinayak Sharma - 2021574

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

- UCI Online news popularity dataset is a valuable resource in the field of journalism.
- Online news articles that have a high popularity score are more likely to be shared and commented on by the readers.
- In an era dominated by the internet and social media, understanding what makes the online news article popular and shareable is very important for editors, media networks, and content creators.

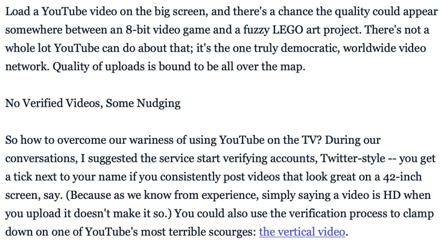
Motivation

- **Factors** affecting the popularity of Online News Article
  - Topic
  - Headline of the article.
  - Images
  - Text Quality

- The task that this dataset aims to address is the prediction of news articles popularity by looking at the number of shares it receives on social media.

Literature Review

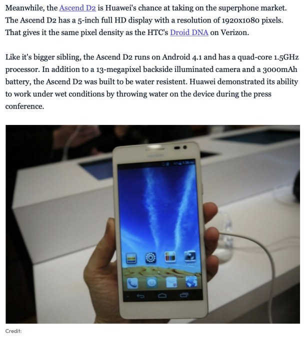
- "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News" by Fernandez, Vinagre and Cortez.
- NLP algorithms and IDSS.
- Proposed a Binary Classification problem.
- Search methods on Random Forests, AdaBoost and KNN for best performance



Load a YouTube video on the big screen, and there's a chance the quality could appear somewhere between an 8-bit video game and a fuzzy LEGO art project. There's not a whole lot YouTube can do about that; it's the one truly democratic, worldwide video network. Quality of uploads is bound to be all over the map.

No Verified Videos, Some Nudging

So how to overcome our wariness of using YouTube on the TV? During our conversations, I suggested the service start verifying accounts, Twitter-style -- you get a tick next to your name if you consistently post videos that look great on a 42-inch screen, say. (Because as we know from experience, simply saying a video is HD when you upload it doesn't make it so.) You could also use the verification process to clamp down on one of YouTube's most terrible scourges: the vertical video.

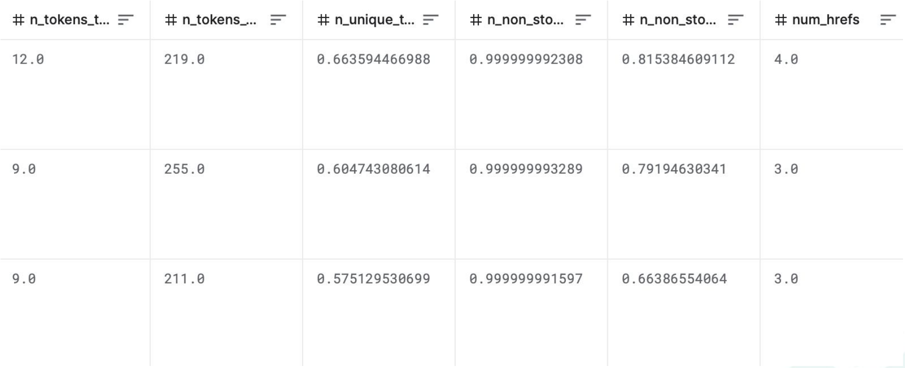


Meanwhile, the Ascend D2 is Huawei's chance at taking on the superphone market. The Ascend D2 has a 5-inch full HD display with a resolution of 1920x1080 pixels. That gives it the same pixel density as the HTC's Droid DNA on Verizon.

Like it's bigger sibling, the Ascend D2 runs on Android 4.1 and has a quad-core 1.5GHz processor. In addition to a 13-megapixel backside illuminated camera and a 3000mAh battery, the Ascend D2 was built to be water resistent. Huawei demonstrated its ability to work under wet conditions by throwing water on the device during the press conference.

Credit:

NLP ALGORITHMS



# n_tokens_t...	# n_tokens_...	# n_unique_t...	# n_non_sto...	# n_non_sto...	# num_hrefs
12.0	219.0	0.663594466988	0.999999992308	0.815384609112	4.0
9.0	255.0	0.604743080614	0.999999993289	0.79194630341	3.0
9.0	211.0	0.575129530699	0.999999991597	0.66386554064	3.0

Literature Review

- "Online News Popularity Prediction" by Feras Namous, Ali Rodan.
- Classification problem.
- Extracted 20 best features of the dataset using fischer score.
- 11 ML algorithms.
- Random Forests - 66%, MLP - 65%

Dataset description

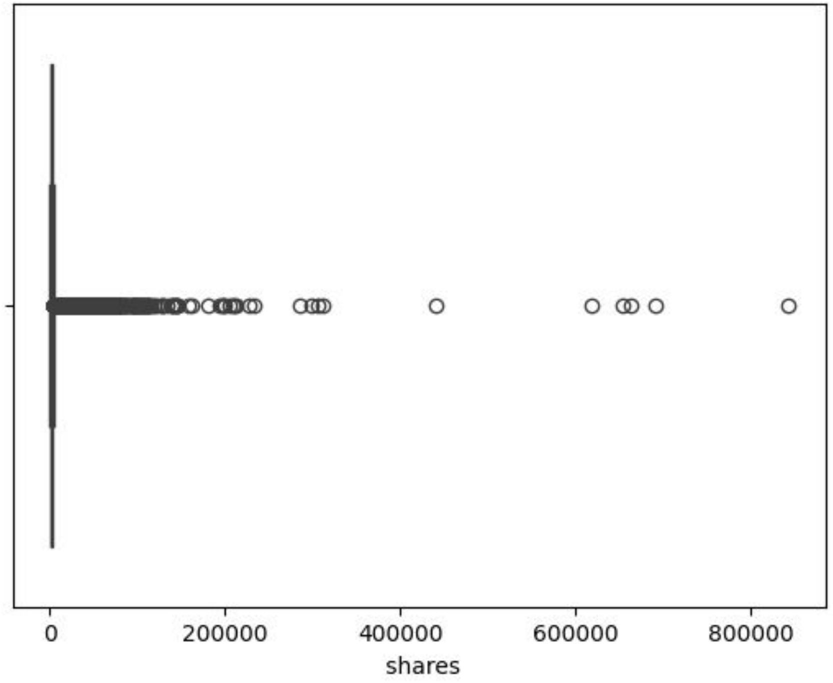
- The UCI Online News Popularity dataset underwent prior preprocessing by Fernandez et al.

- Multiple features were derived from the articles using NLP processes, encompassing aspects such as title subjectivity, Latent Dirichlet Allocation (LDA), the frequency of positive and negative words, release date, article type, data channel, image count, and more.

Dataset description

- The dataset comprises a total of 58 predictive attributes, alongside two non-predictive attributes, and one target attribute, which is denoted as "shares."

- It is important to note that the target variable is highly skewed.



Dataset description

Collinearity:

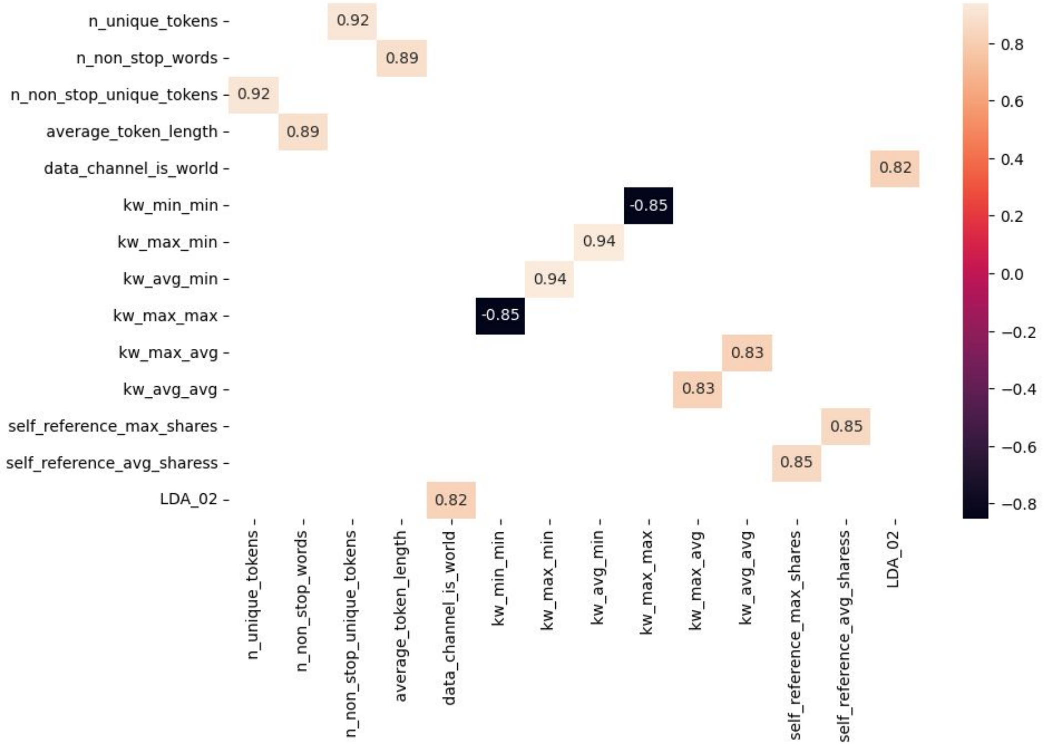
Used Correlation Heatmap Analysis Conducted.

Key Findings:

- Strong Correlations Identified:
  - Indicates Redundancy in Certain Feature Pairs.
  - Correlation Coefficient Threshold: 0.85.

Action Taken:

- Feature Removal Strategy:
  - Removed one Feature from Highly Correlated Pairs.
  - Enhances Efficiency and Prevents Redundancy.



Dataset description

Outlier detection:

Importance of Outlier Detection:

- Quality Assurance: Essential for Data Integrity.
- Algorithm Used: Local Outlier Factor (LOF) Algorithm Employed.

Challenges Faced:

- Issue Identified: Outliers Not Detected as Expected.

Dataset description

Dimensionality Reduction

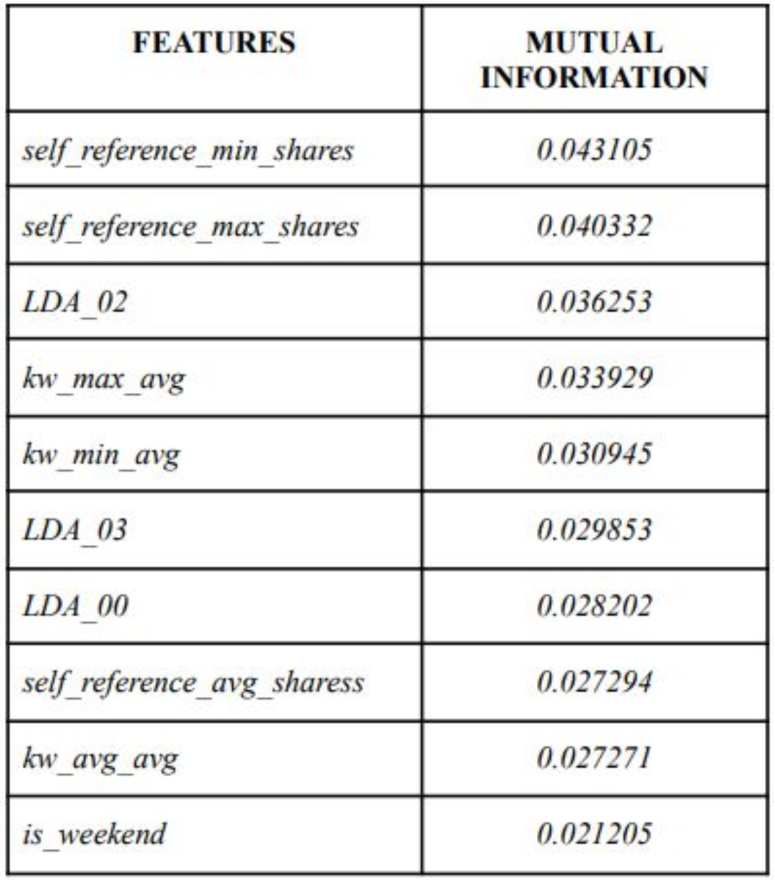
- Mutual Information Analysis: Identified Variables with Limited Contribution.

Approach:

- Method Used: Principal Component Analysis
- Outcome: Retained Top Four Components Capturing 99.6% Variance.

Significance:

- Reducing the features while preserving maximum information and Interpretability



FEATURES	MUTUAL INFORMATION
self_reference_min_shares	*0.043105*
self_reference_max_shares	*0.040332*
LDA_02	*0.036253*
kw_max_avg	*0.033929*
kw_min_avg	*0.030945*
LDA_03	*0.029853*
LDA_00	*0.028202*
self_reference_avg_sharess	*0.027294*
kw_avg_avg	*0.027271*
is_weekend	*0.021205*

Methodology

- Various regression techniques are used for predictions over the dataset.

Methodology

- Various regression techniques are used for predictions over the dataset.

- Regression Techniques:
  - Linear Regression
  - Bagging
  - Multilayer Perceptron

Methodology

- Various regression techniques are used for predictions over the dataset.

- Regression Techniques:
  - Linear Regression
  - Bagging
  - Multilayer Perceptron

- Evaluation Metric:
  - MAE (Mean Absolute Error)
  - R2 Score

Methodology

- **Linear regression:**
  - **Simple Linear Regression:** Used as a baseline model. The predicted value is a linear relation of the features.

Methodology

- **Linear regression:**
  - **Simple Linear Regression:** Used as a baseline model. The predicted value is a linear relation of the features.
  - **Lasso Regression:** Derived from linear regression. It performs regularization as well as feature selection.

Methodology

- **Linear regression:**
  - **Simple Linear Regression:** Used as a baseline model. The predicted value is a linear relation of the features.
  - **Lasso Regression:** Derived from linear regression. It performs regularization as well as feature selection.
  - **Ridge Regression:** Derived from linear regression. It performs regularization.

Methodology

- **Linear regression:**
  - **Simple Linear Regression:** Used as a baseline model. The predicted value is a linear relation of the features.
  - **Lasso Regression:** Derived from linear regression. It performs regularization as well as feature selection.
  - **Ridge Regression:** Derived from linear regression. It performs regularization.

  However, the result found are not satisfactory.

Methodology

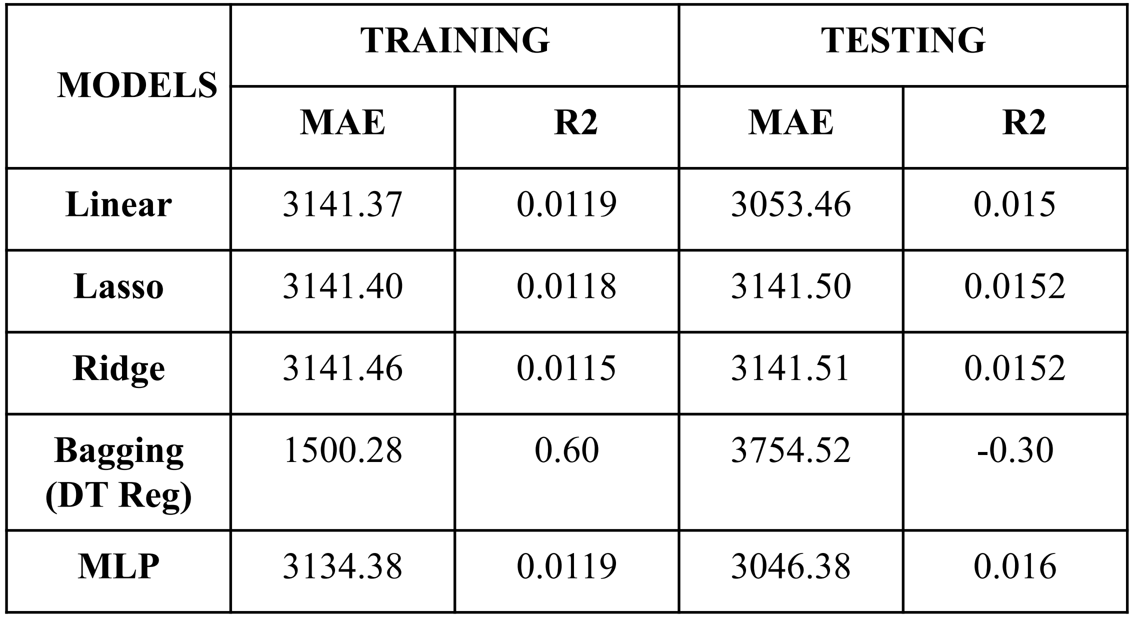
- **Bagging:** An ensemble method that uses various models and takes the average prediction. It is used to lower the variance of the data. However, the results were not meaningful as it overfitted the data.

Methodology

- **Bagging:** An ensemble method that uses various models and takes the average prediction. It is used to lower the variance of the data. However, the results were not meaningful as it overfitted the data.

- **Multi-Layer Perceptron:** MLP is type of ANN. They are feed forward neural networks which are composed of several layers of nodes in a unidirectional way. The results are however not satisfactory at all even after hyperparameter tuning.

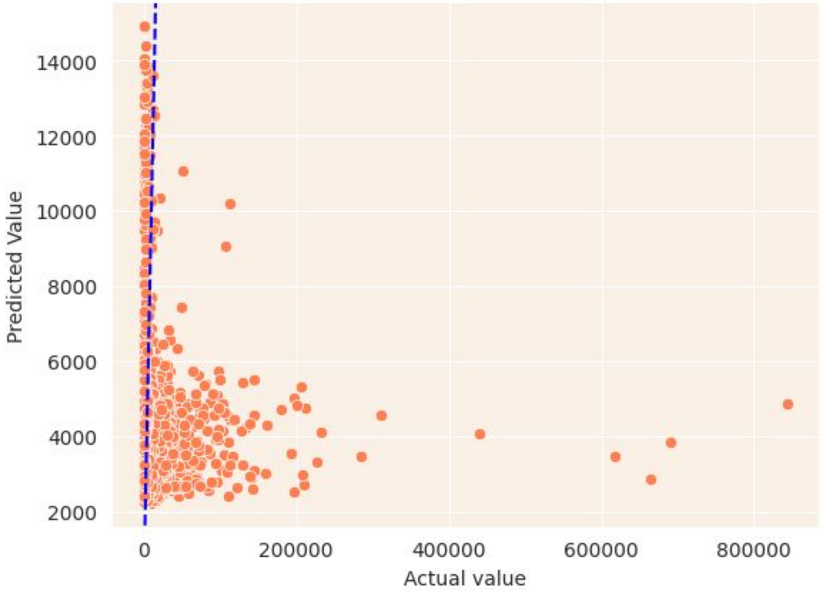
Results and Analysis



MODELS	TRAINING		TESTING	
	MAE	R2	MAE	R2
Linear	3141.37	0.0119	3053.46	0.015
Lasso	3141.40	0.0118	3141.50	0.0152
Ridge	3141.46	0.0115	3141.51	0.0152
Bagging (DT Reg)	1500.28	0.60	3754.52	-0.30
MLP	3134.38	0.0119	3046.38	0.016

Results and Analysis

MODELS	TRAINING		TESTING	
	MAE	R2	MAE	R2
Linear	3141.37	0.0119	3053.46	0.015



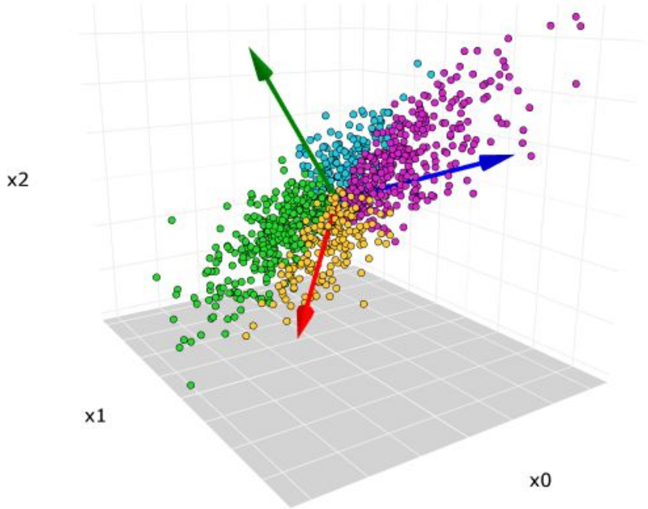
Scatter plot between predicted value and actual value

Major underlying assumptions behind linear regression:

- Non-Collinearity between Independent Variables
- Linear Dependence of Dependent Variables on Independent Variables
- Normality of Residuals

Results and Analysis

MODELS	TRAINING		TESTING	
	MAE	R2	MAE	R2
Linear	3141.37	0.0119	3053.46	0.015



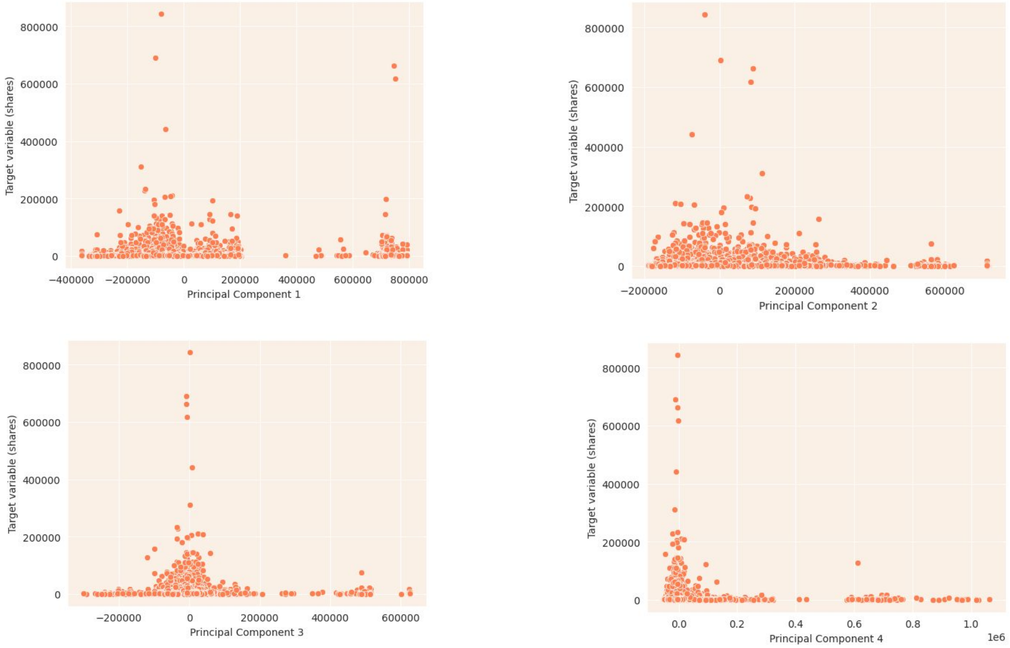
Orthogonal PCA vectors

Major underlying assumptions behind linear regression:

- Non-Collinearity between Independent Variables
- Linear Dependence of Dependent Variables on Independent Variables
- Normality of Residuals

Results and Analysis

MODELS	TRAINING		TESTING	
	MAE	R2	MAE	R2
Linear	3141.37	0.0119	3053.46	0.015



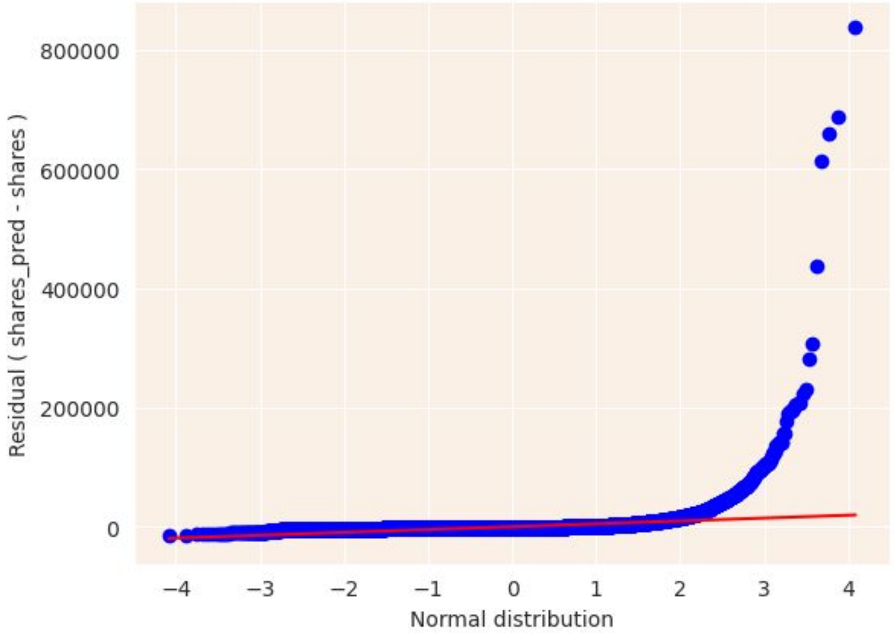
Scatter plot between Target Variable and PCA vectors

Major underlying assumptions behind linear regression:

- Non-Collinearity between Independent Variables
- Linear Dependence of Dependent Variables on Independent Variables
- Normality of Residuals

Results and Analysis

MODELS	TRAINING		TESTING	
	MAE	R2	MAE	R2
Linear	3141.37	0.0119	3053.46	0.015



Q-Q Plot between normal distribution and Residual

Major underlying assumptions behind linear regression:

- Non-Collinearity between Independent Variables
- Linear Dependence of Dependent Variables on Independent Variables
- Normality of Residuals

Results and Analysis

MODELS	TRAINING		TESTING	
	MAE	R2	MAE	R2
Linear	3141.37	0.0119	3053.46	0.015
Lasso	3141.40	0.0118	3141.50	0.0152
Ridge	3141.46	0.0115	3141.51	0.0152
Bagging (DT Reg)	1500.28	0.60	3754.52	-0.30
MLP	3134.38	0.0119	3046.38	0.016

⇒ Suboptimal Performance

According to the performance metrics.

Results and Analysis

MODELS	TRAINING		TESTING	
	MAE	R2	MAE	R2
Linear	3141.37	0.0119	3053.46	0.015
Lasso	3141.40	0.0118	3141.50	0.0152
Ridge	3141.46	0.0115	3141.51	0.0152
Bagging (DT Reg)	1500.28	0.60	3754.52	-0.30
MLP	3134.38	0.0119	3046.38	0.016

This can be attributed to extremely high values in our target variable.

Therefore, a more pragmatic approach involves implementing a threshold to transform these extreme values into discrete labels. Consequently, our continuous target variable undergoes a discretization process.

Results and Analysis

This approach serves two primary objectives:

- It facilitates the evaluation of different models based on accuracy metrics. By correctly classifying extreme values into their respective labels, all models can be compared based on the classification of non-extreme data points, enabling meaningful improvements.

- It provides more informative feedback, especially for news writers, as it predicts a label (e.g., "popular" or "unpopular") associated with each data point rather than an exact value that has less significance.

Timeline

Week 1-2: Exploratory Data Analysis

Week 3: Feature engineering

Week 4: Pre-processing for Regression

Week 5: Model Generation for Regression

Week 6: Threshold generation for discretization and Preprocessing for Classification

Week 7: Model Generation & Hyperparameter Tuning for Classification

Week 8-9: Model Evaluation & Ensembling Methods

Week 10: Comparisons and Conclusions

Team member’s contribution

Contribution till the mid project:

- ***Aditya Jain:*** Data Analysis, Model Generation, Report
- ***Shrey:*** Data Analysis, Pre-processing, Report
- ***Vasu Kapoor:*** Data Analysis, Model Generation, Literature review, Report
- ***Vinayak Sharma:*** Pre-processing, Report, Model generation, Data Analysis, Literature review