Online News Popularity Prediction

Aditya Jain
aditya21511@iiitd.ac.in
2021511

Shrey
shrey21562@iiitd.ac.in
2021562

Vasu Kapoor
vasu21573@iiitd.ac.in
2021573

Vinayak Sharma
vinayak21574@iiitd.ac.in
2021574

***Abstract*— Online news articles that have a high popularity score are more likely to be shared and commented on by the readers. The popularity of online news articles depends on many factors, such as the topic, the headline, the keywords, the images, and the text quality. A model that can predict the popularity of an article before it is published can be very useful for media professionals who want to create engaging and relevant content for their audience.**

1. Introduction

The UCI Online news popularity dataset is a valuable resource in the field of digital journalism. In an era dominated by the internet and social media, understanding what makes the online news articles popular and shareable is very important for editors, media networks, and content creators. 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.

2. Literature Survey

The dataset employed in our study comprises news articles originating from the website www.mashable.com. This dataset was sourced from Fernandez et al. [1] and is publicly accessible at Kaggle. The prediction of news article popularity has garnered significant attention, with readily available datasets on platforms like Kaggle and numerous research papers contributing to this domain.

[1] "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News" by Fernandez et al., employed various machine learning models, along with Natural Language Processing (NLP) and Intelligent Decision Support Systems (IDSS), to draw their conclusions. Their research revolved around predicting the popularity of news articles through a binary classification task.

[2] "Online News Popularity Prediction" by Feras Namous et al. In their work, they posed a classification problem using the same dataset and applied eleven data mining algorithms. Notably, their research highlighted Random Forest and MLP as the top-performing models, achieving an accuracy of 65% while advocating for further improvements in this area.

[3] "Online News Popularity Prediction" by Shuo Zhang also employed a three-layer neural network and implemented bimodal distribution removal techniques in their analysis of popularity classification. Their research culminated in achieving a classification accuracy of 70%.

3. Dataset & Pre-Processing

The UCI Online News Popularity dataset underwent prior preprocessing by Fernandez et al. [1], involving the utilization of Natural Language Processing (NLP) techniques and HTML tag analysis. Each URL in the dataset serves as a unique identifier, corresponding to news articles published on The Mashable website over a two-year period spanning from 2015 to 2017. 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. To summarize, 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.

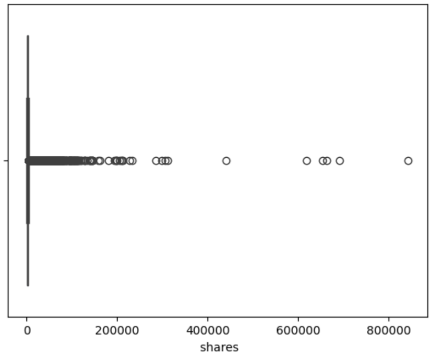

Fig.1: Box plot of target attribute (i.e. 'shares')

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*

Table 1: Top 10 features based on Mutual Information

A. Collinearity

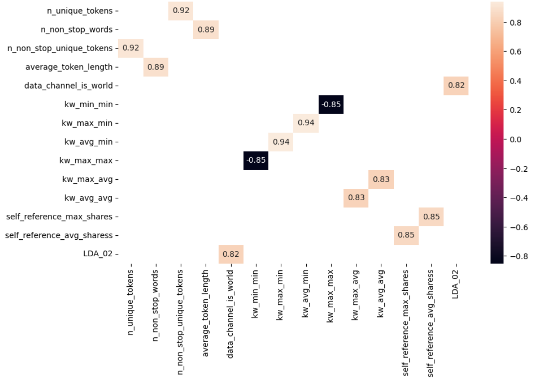

Fig. 2: Correlation Heatmap of attributes having absolute value of correlation index at least 0.8.

A correlation heatmap analysis was conducted to assess the relationships between all pairs of features within the dataset. The results revealed that certain feature pairs exhibited a strong correlation, indicating that they contribute similar information to the dataset. Therefore, it is preferred to remove one of the features from each highly correlated pair to reduce dimensionality and prevent redundancy. Specifically, when the correlation coefficient between a pair of features exceeds 0.85, one of them can be filtered out, simplifying the dataset and maintaining its essential information. This approach helps enhance the efficiency and interpretability of further data analysis and machine learning tasks.

B. Outlier detection

Outliers detection is essential for data quality assurance. Hence, the Local Outlier Factor algorithm is used for detection of the outlier. But it turns out that it is not detecting the outliers that need to be removed. These outliers are analyzed through the data analysis done.

C. Dimensionality reduction

Based on the mutual information analysis, we can deduce that certain variables only contribute to the overall information in the dataset. Therefore, a dimensionality reduction appears necessary. Consequently, we opted for Principal Component Analysis (PCA) to achieve dimensionality reduction. Specifically, we retained the four highest-ranked components from the PCA process, as they collectively captured 99.6% of the variance present in the original dataset.

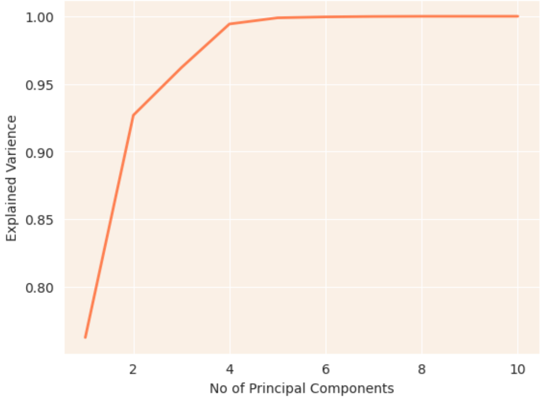

Fig. 3: Explained Variance against no of PCA components.

4. Methodology

We used various regression techniques for our predictions over the dataset. Our results include errors and R2 score. From our data analysis, we used 4 components for the PCA transformation.

A. Linear Regression

- *Simple Linear Regression*: was used at a baseline model, since our dataset is regressive. In this technique, the predicted value is a linear relation of the features. However, the results found are not satisfactory.

- *Lasso Regression*: It is a derivative of the linear regression model where it performs regularization as well as feature selection. The results observed were the same as simple linear regression.

- *Ridge Regression*: It is also a derivative of the linear regression model that performs regularization. However, the results observed were the same as linear regression.

B. Bagging

It is an ensemble method that uses various models and takes the average prediction. It is used to lower the variance of the model by training multiple models on different subsets of the data. However, the results were not meaningful as it overfitted the data: Poor performance under Linear Regression and overfitting under Decision Regressor for multiple number of estimators with Grid Search being the pivot.

C. Multi-Layer Perceptron

MLP is a 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.

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

Table 2: Performance metrics for different models

5. Analysis

The R2 score is a statistical measure that brings out the proportion of variance in the dependent variables as explained by the independent variables. Hence, it is studied to compare performance among different models.
Beginning with Linear Regression, we examined the underlying assumptions behind linear regression.

A. Non-Collinearity between Independent Variables

We opted to employ Principal Component Analysis (PCA) as a preliminary step in our pre-processing stage. PCA, inherently, undertakes the transformation of the original feature set into a set of orthogonal features, organized in ascending order of their respective cumulative explained variances. Consequently, our approach enabled us to simultaneously achieve dimensionality reduction and feature independence.
Four highest components were picked from PCA since they inherited 99.6% variance from the original dataset.

B. Linear Dependence of Dependent Variables on Independent Variables

This can be manually illustrated by scatter-plotting the obtained mutually independent components against the target variable.

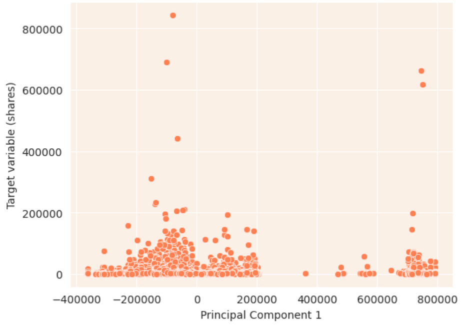

Fig. 4: Scatter plot between Target Variable and column 1 of the data with PCA=4

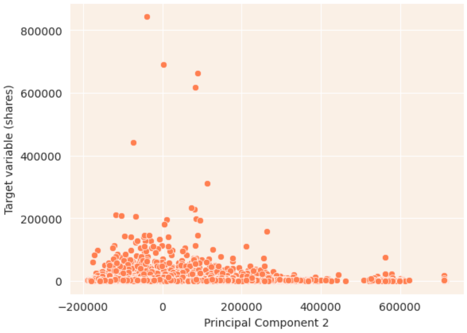

Fig. 5: Scatter plot between Target Variable and column 2 of the data with PCA=4

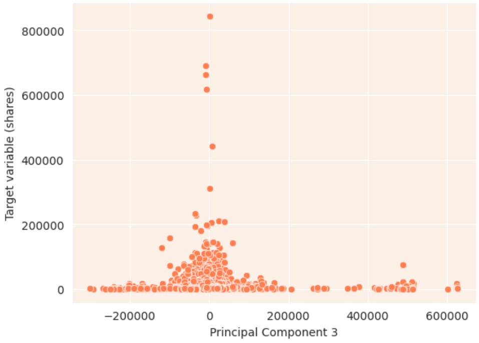

Fig. 6: Scatter plot between Target Variable and column 3 of the data with PCA=4

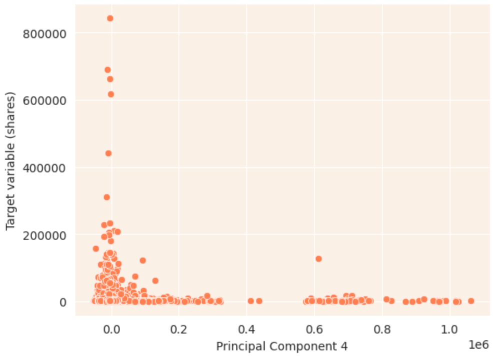

Fig. 7: Scatter plot between Target Variable and column 4 of the data with PCA=4

As inferred from the graphical representations, none of the extracted components demonstrate a linear association with our target variable. Consequently, this discrepancy stands as a noteworthy departure from our initial assumption. Hence, this is one of the primary factors contributing to the observed underperformance of our linear models.

C. Normality of Residuals

The presence of normally distributed prediction errors, or residuals, signifies their symmetric distribution centered around zero. This observation emphasizes the model's capacity to encapsulate the core characteristics and trends of the dataset.

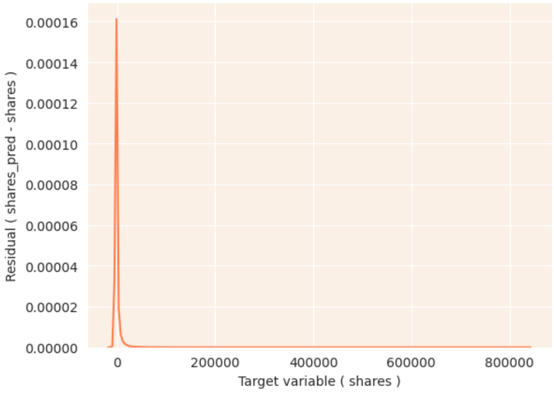

Fig. 8: Density plot for residual vs target variable

The majority of the prediction error is within 10,000. Hence, the distribution is left-skewed and leptokurtic in nature. For further assurance, we drew a Q-Q plot compared with the standard normal distribution.

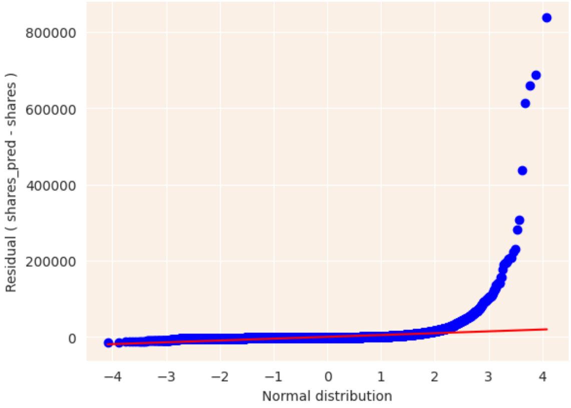

Fig. 9: Similarity between normal distribution and Residual

As observed, the distribution exhibits sharp deviations at the extremes, a common trait of a leptokurtic distribution. Hence, the residual distribution does not adhere to the normality assumption. Consequently, we can confidently assert that the dataset is not conducive to linear regression modeling.

D. Homoscedasticity of Residuals

Heteroscedasticity refers to different variances for a single feature.
On plotting scatter plots of residuals with features, similar scatter plots were obtained.

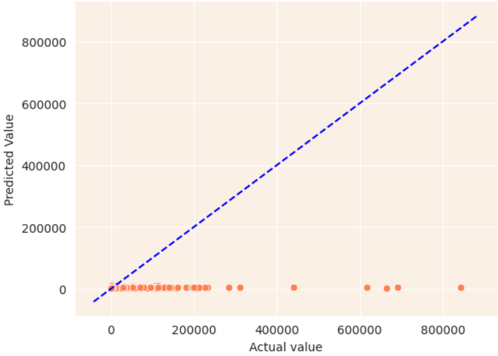

Fig. 10: Scatter plot between predicted value and actual value.

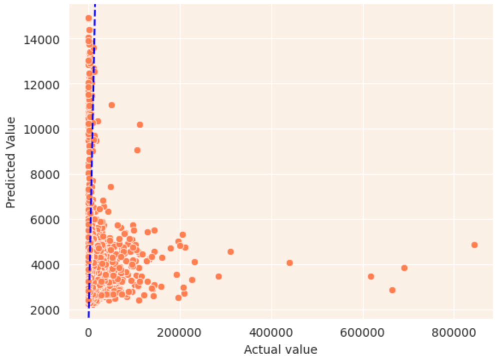

Fig. 11: Scatter plot between predicted value and actual value (different scaling)

Overall, linear regression is found to underperform on the given dataset.

7. Conclusion

In a comprehensive analysis, we deduced why linear regression demonstrates suboptimal performance on the provided dataset. By inferring from the performance metrics derived from linear regression, we can confidently assert that the alternative models previously discussed in our methodology exhibit similar suboptimal performance.
This can be attributed to extremely high values in our target variable, which still require comprehensive handling through outlier detection algorithms. As a potential solution, we can consider removing data points where the target variable surpasses a defined threshold, such as the whisker threshold identified through boxplot analysis. This action could lead to an enhancement in model performance on the remaining dataset. However, it's essential to acknowledge that these removed points may still introduce substantial errors when applied to the testing data.

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. 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.

8. References

[1] Shuo Zhang, Online News Popularity Prediction
[2] Feras Namous, Ali Rodan, Yasir Javed, Online News Popularity Prediction
[3] Ijraset, Online News Articles Popularity Prediction System
[4] Fernandez et al. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News
[5] News articles from www.mashable.com