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.

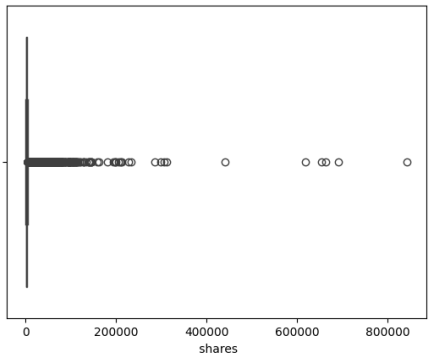

Fig.1: Box plot of target attribute (i.e. 'shares')

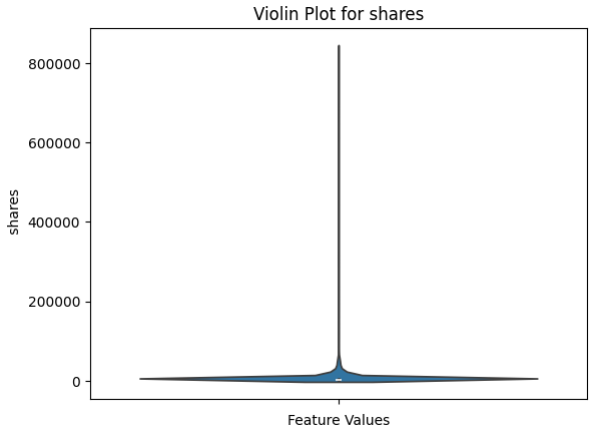

Fig. 2: Violin plot of target attribute (i.e 'shares')

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

Since the target variable is so highly skewed, we cannot apply Regression models because they will give a high error at those skewed points as discused in the Mid sem Report. We are converting the dataset into classification by making the target variable 1 if the value is more than the median value of the shares, which is 1400, and 0 if the value is less than the median value. We have taken the median value because it divides the data into two equal parts. So we call the news popular if the target variable is predicted as 1, else we say it is unpopular.

A. Collinearity

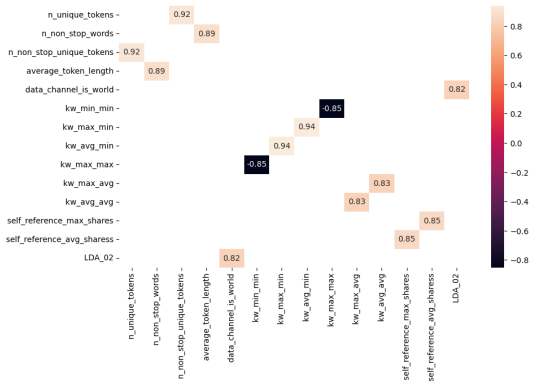

Fig. 3: Correlation Heatmap of attributes having absolute value of correlation index at least 0.8.

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 0.8, 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. We used the Local Outlier Factor algorithm and Isolation Forest algorithm for detection of the outliers. But it turns out that it is not detecting the outliers that need to be removed. These outliers are analyzed through data analysis.

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. We already used the Heatmap to remove the high correlated features. Consequently, we opted for Principal Component Analysis (PCA) to achieve dimensionality reduction. But after using PCA our accuracy was getting reduced due to information loss. We also used RFE and to get the top features and also used F-score to

calculate feature importance but even after using them to take the top features, the accuracy of models is getting reduced.

4. Methodology

We used various classification techniques for our predictions over the dataset.

A. Logistic Regression.

Logistic Regression is a method used for classification problems where the goal is to predict the probability of an instance belonging to a particular class. Logistic Regression uses the Sigmoid Function to modal the probability that the particular input belongs to a particular class.

B. Decision Tree Classifier

Decision Tree works by partitioning the dataset based on the value of features. At each node of the tree, a decision is made based on a feature's value, leading to the next node or a leaf node where the final prediction is made. Decision Trees are used to capture nonlinear relationships between features and the target variable. For splitting, 'gini-impurity'

C. Random Forest Classifier

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees to create a more robust model. It's an extension of bagging with an additional element of randomness introduced during tree-building. 100, 300, and 500 Decision Trees were used to get the best accuracy

D. Bagging

Bagging (Bootstrap Aggregating) is an ensemble learning technique that involves training multiple instances of the same learning algorithm on different subsets of the training data. We used Logistic Regression as the learning algorithm. The primary goal of bagging is to reduce overfitting and improve the stability and accuracy of the model.

E. AdaBoost

AdaBoost (Adaptive Boosting) is an ensemble learning method. AdaBoost works by combining the predictions of many weak learners to create a strong learner. We used Logistic Regression as our weak learner. It assigns weights to each instance in the dataset and focuses on misclassified instances in subsequent iterations, giving more emphasis to the difficult-to-classify examples. 50, 100 estimators were used to find the best accuracy,

F. Gradient Boosting Classifier

Gradient Boosting is an ensemble learning technique that builds a robust predictive model that is formed by combining weak learning models (generally, decision trees). Gradient boosting is adequate for capturing the complex patterns in the data. 100, 300, 500 estimators were used to get the best accuracy.

G. MLP classifier

The MLP Classifier (Multi-Layer perceptron Classifier) is an artificial neural network-based classification algorithm. The MLPClassifier is a feedforward neural network, meaning that information moves through the network layer by layer without cycles or loops.

H. Bernoulli Naive Bayes

Bernoulli Naive Bayes is a variant of the Naive Bayes algorithm, specifically designed for binary (two-class) classification problems. It also has a strong conditional independence assumption. It also assumes that the distribution of the target column is a bernoulli distribution.

I. Stacking Classifier

Stacking, short for stacked generalization, is an ensemble learning technique that involves combining the predictions of multiple base models to improve overall predictive performance. In stacking, a meta-model is trained to predict the target variable based on the predictions made by the individual base-models. In this analysis, random Forests and Gradient Boosting were used as base models since they learnt the data well as compared to others.

5. Results and Analysis

The dataset was first approached as a regression problem. It was found that the results of the regression problem were not even close to satisfactory, and an in-depth analysis of the data was provided on why the dataset failed as a regression problem.

In summary, the dataset failed to show any linear independence or normal behavior or any linear separability for various algorithms to provide an optimal solution.

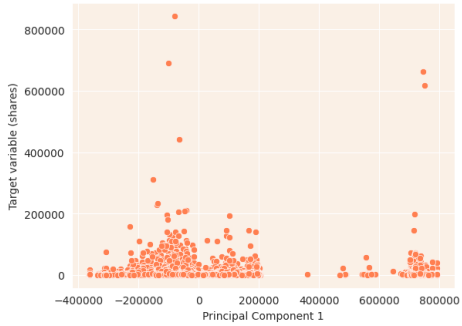

Fig. 4: Scatter plot between Target Variable and column 1 of the data with PCA=4

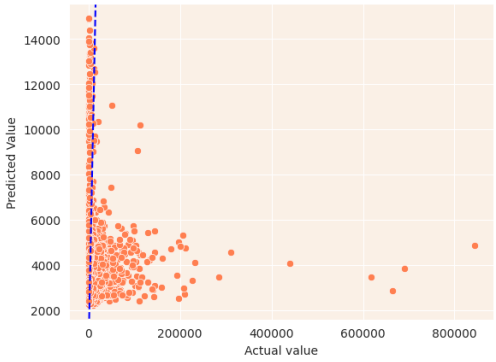

Fig. 5: Scatter plot between predicted value and actual value (different scaling)

As a result, the models generated contained values that deviated largely from their true values that ruined the overall performance of the model. A comprehensive analysis of PCA features (Fig 4 and Fig 5) and feature importance (Table 1) led to a conclusion that approaching the dataset as a regression problem does not imply much about the shares of the news articles even after pre-processing, including outlier detection and Dimension Reduction.

MODELS	TRAINING	TESTING
	Accuracy	Accuracy
Logistic	65.2	65.3
MLP	62.3	65.3
Bagging (LogReg)	65.5	65.13
Random Forest(500)	100	68.1
AdaBoost (LogReg)	64	63
Gradient Boosting	71	67.1
SVM	66.3	66
Naive Bayes	63.17	63.13
Stacking Classifier	90	67

Table 2: Performance metrics for different models

In comparison to the performances of other paper, the performance of logistic regression was improved from 55.9% (Ali Rodan et al.) and 63.04% (Riya Talwar et al) to 65.2%

In comparison to the performances of other paper, the performance of Random Forest was improved from 65.8% (Ali Rodan et al.) and 63.04% (Riya Talwar et al) to 68.1%.

Gradient Boosting was also observed to give a good performance as compared to the other machine learning algorithms,

However, Decision Trees and Random Forests also seem to overfit the data.

Stacking Classifiers using a combination of both Random Forests and Gradient Boosting with a predictor as Logistic Regression also provided an optimal performance with respect to other models.

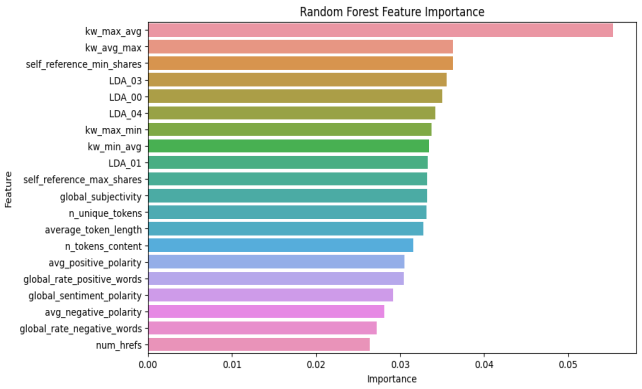

Fig. 6: Top 20 features based on Random Forest's performance

Using Random Forest which provided the best performance, feature importance of the algorithm was exploited and analyzed to figure out all the important features of the data that could potentially be highly influential in determining the popularity of a new article.

7. Conclusion

In a comprehensive journey, we deduced why regression-oriented models demonstrate suboptimal performance on the provided dataset by inferring from their performance metrics.
Therefore, a more pragmatic approach involved implementing a threshold to transform these extreme values into discrete labels. Consequently, our continuous target variable has undergone a discretization process. This approach served 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 insights and 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.

Recommendations for stakeholders are highlighted as follows:

- Publishers: The statistical summarization of articles, avoiding the need for their internal contents, proves adequate for predicting popularity pre-publication. Entities prioritizing reputation and trust can confidently rely on our methodology for enhanced security and reliability assurance.
- Writers: In alignment with our findings, to argue the likelihood of achieving popularity, authors are advised to incorporate references to already-popular articles, enhance subjectivity in titles, and integrate more visual elements such as photos and videos. Additionally, it is recommended to refrain from using negative words, while avoiding multi-topic discussions whenever possible.

Furthermore, as cited earlier, we take pride in highlighting our notable performance in the literature, particularly within non-deep learning architectures. Our commitment extends towards future explorations in this domain, along with analysis of prevailing and emergent high-impact trends on top of popularity prediction.

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