

Online news popularity prediction



Purdue University
Elena Gomez
35459621

INDEX

1. Abstract
2. Overview and Motivation
3. Literature Review
 - 3.1 Relevant Work
 - 3.2 Contributions of the project
4. Technology details
5. Performance evaluation
6. Conclusion



1.ABSTRACT



1.ABSTRACT



- **Objective:** Predicting online news popularity using supervised learning.
- **Dataset:** Database of online news articles from the New York Times.
- **Approach:** The project treated the problem as both regression and classification, studying the dataset using visualization techniques and thorough data preprocessing.
- **Techniques:** The project applies different classification and regression techniques, such as logistic regression, decision trees, random forest, simple linear regression, and multiple regression.
- **Results:** The project compares the performance of the techniques by analyzing different metrics, focusing mainly on accuracy, and selects the best model. The usefulness of the model in the real world for decision-making is also discussed.
- **Keywords** data science, online news, popularity, classification, regression, decision-making.

2.OVERVIEW AND MOTIVATION



■ New Visitor ■ Returning Visitor



2.OVERVIEW AND MOTIVATION



- Advertising is an effective means for companies to expand their reach and promote their products or services to a larger audience.
- Selecting the optimal placement for advertisements can be a complex decision due to various factors, such as advertising on web pages or articles that are not popular.
- To avoid wasting money, the project aims to train models that predict if an article is popular or not based on quantitative and qualitative features, including the number of images, videos, publication date, and article topic.
- The most important variable for determining popularity is the number of shares, which will be used as the target variable for prediction.
- The problem will also be treated as a classification problem by distinguishing between "popular" and "unpopular" articles.

3.LITERATURE REVIEW

3.1 RELEVANT WORK



Four papers have been published between 2010 and 2021 that propose different models to predict the popularity of online news articles.

- In the first paper, a **logistic regression model** is used to predict the number of votes that a news article would receive based on the number of votes received by similar news articles in the past.
- The second paper proposes a **support vector regression model** that predicts the popularity of news articles based on the number of views, comments, and shares received by similar news articles in the past, as well as the content and sentiment of the articles.
- The third paper uses a **decision tree algorithm** to analyze the news source and topic of news articles and found that these factors play an important role in the popularity of news articles.
- In the fourth paper, a **random forest regression model** is trained to predict the popularity of news articles based on the number of shares, likes, and comments received by similar news articles on social media platforms.

3.2 CONTRIBUTIONS OF THE PROJECT



CLASSIFICATION TECHNIQUES

- Logistic regression
- Decision trees. (We will study with which hyperparameter the performance of the model is better).
- Random forest. (We will study with which hyperparameter the performance of the model is better).

REGRESSION TECHNIQUES

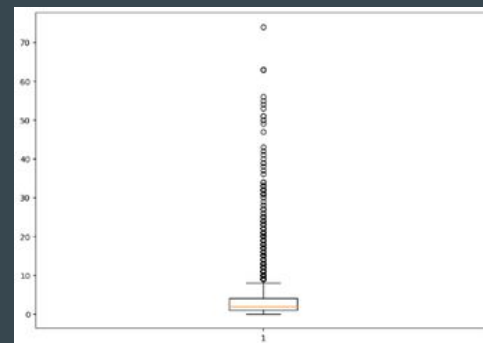
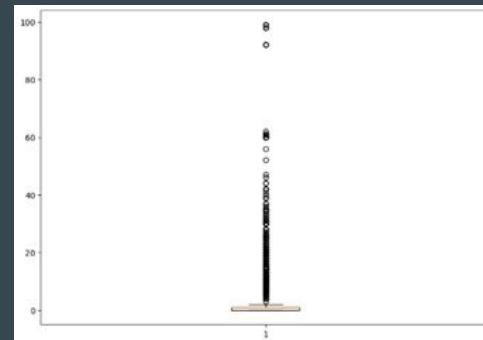
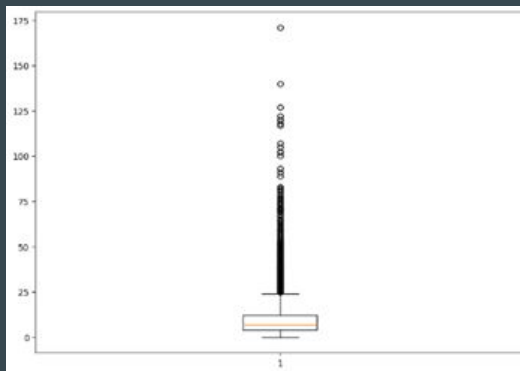
- Simple linear regression
- Multiple regression
- I planned to apply also forward regression, backward regression, and stepwise regression, but after evaluating the performance of simple linear regression and multiple regression, I decided not to train these ones. (We will see this later)

4. TECHNOLOGY DETAILS

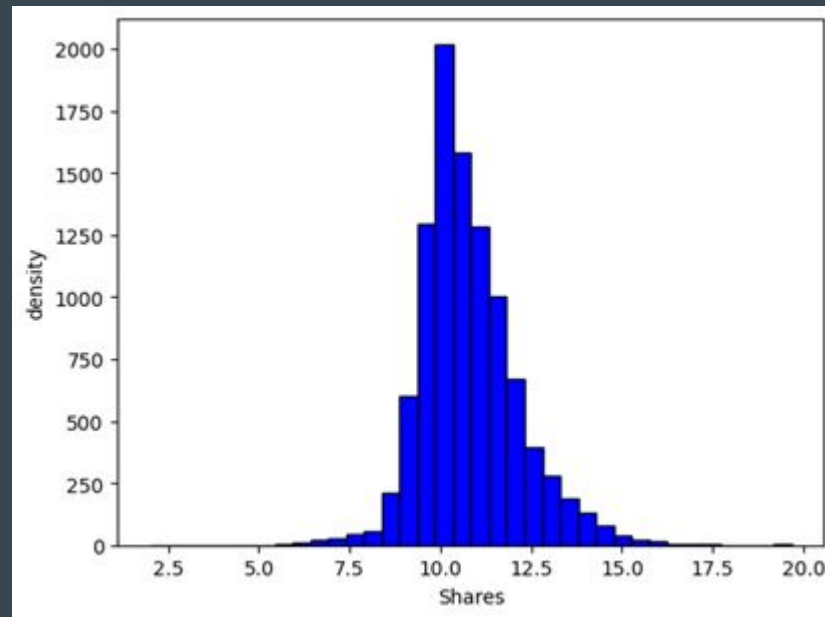
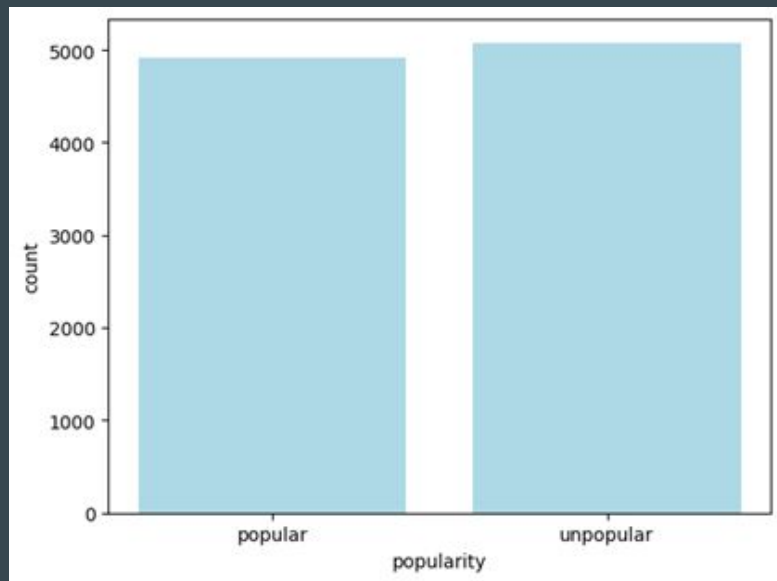


DATA PREPROCESSING

- Missing values
- Irrelevant variables
- Converting variables
- Creating new variables
- Outliers

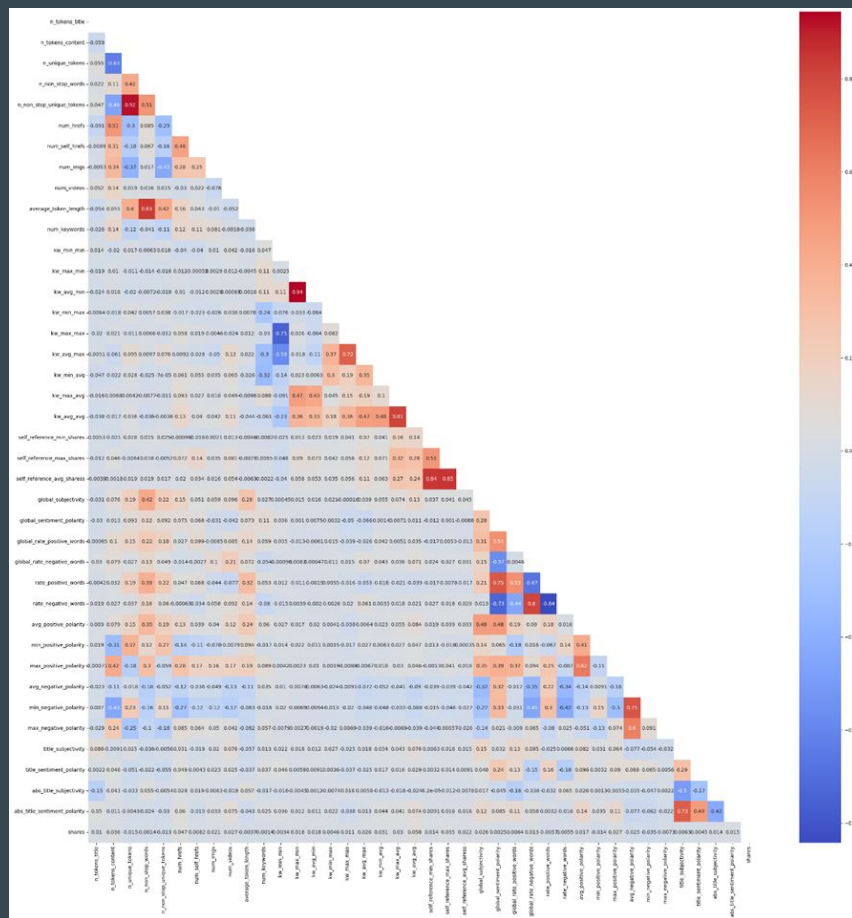


VISUALIZATION TOOLS



CLASSIFICATION TECHNIQUES

- Firstly, I removed correlated variables.
- I splitted the data into train and test.

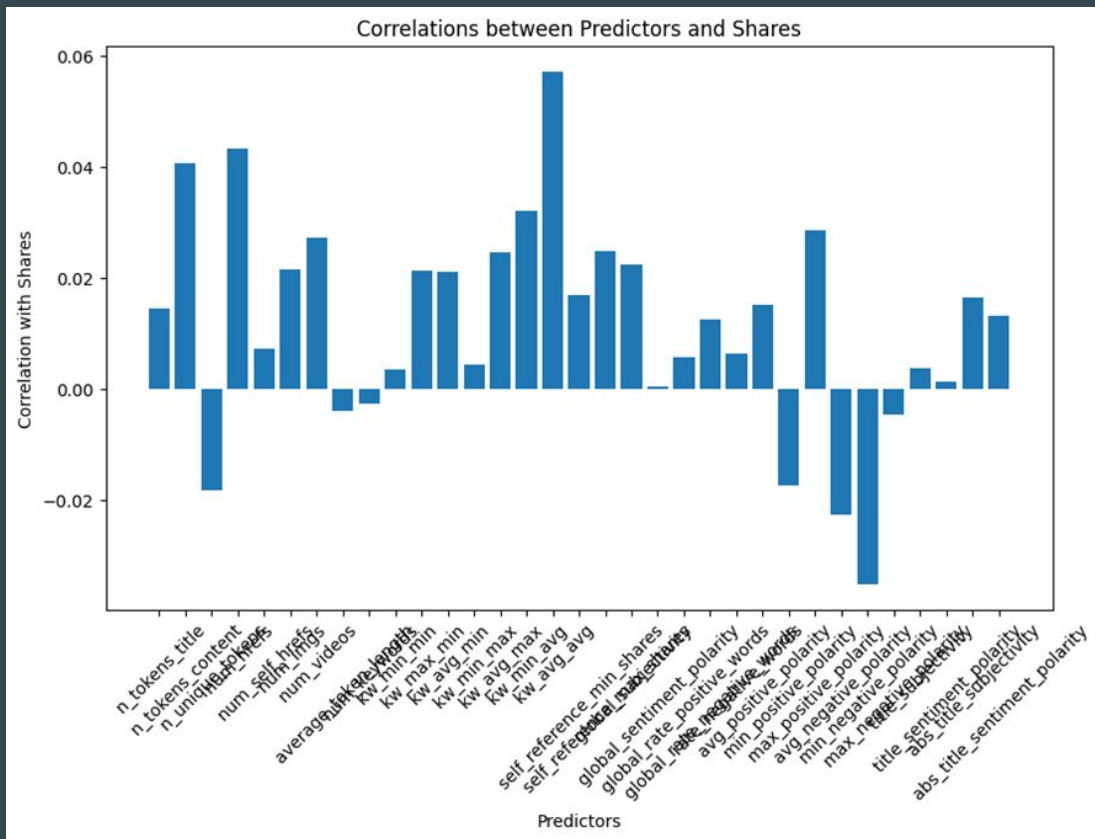


CLASSIFICATION TECHNIQUES

- Logistic regression
- Decision trees (grid search)
- Random forest (grid search)

REGRESSION TECHNIQUES

- Correlations



REGRESSION TECHNIQUES

- Simple linear regression
- Multiple linear regression

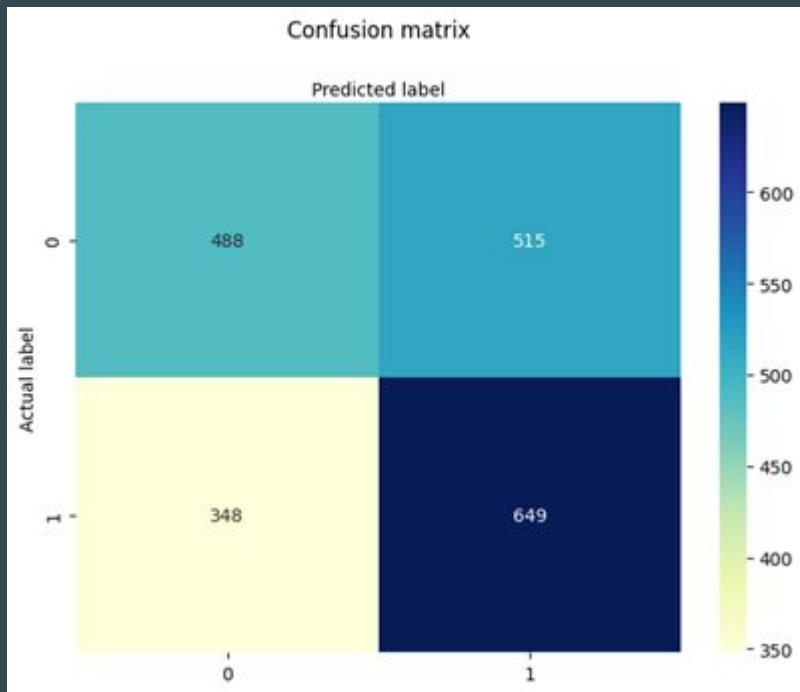
5. PERFORMANCE EVALUATION



■ New Visitor ■ Returning Visitor



LOGISTIC REGRESSION



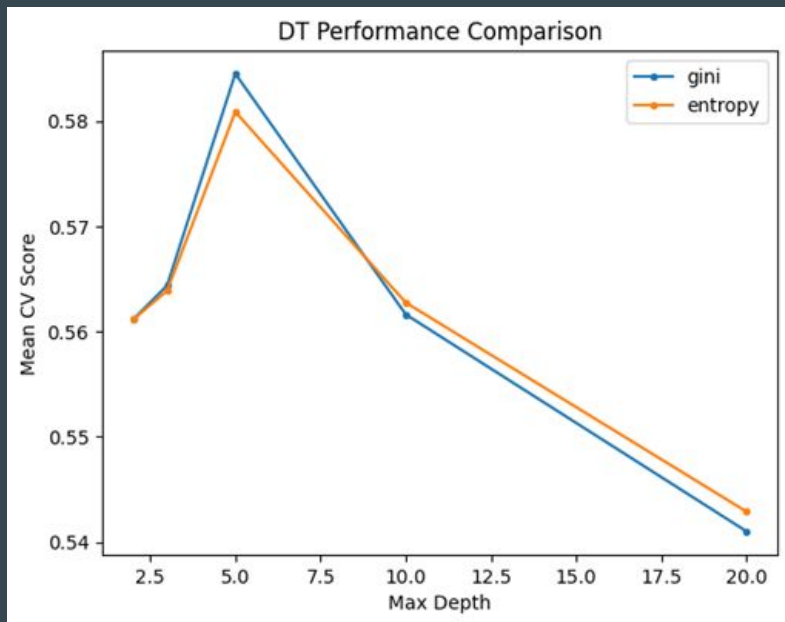
×	precision	recall	f1-score	support
popular	0.58	0.49	0.53	1003
unpopular	0.56	0.65	0.60	997
accuracy			0.57	2000
macro avg	0.57	0.57	0.57	2000
weighted avg	0.57	0.57	0.57	2000

DECISION TREES (parameter by default)




	precision	recall	f1-score	support
popular	0.54	0.51	0.52	1003
unpopular	0.53	0.55	0.54	997
accuracy			0.53	2000
macro avg	0.53	0.53	0.53	2000
weighted avg	0.53	0.53	0.53	2000

DECISION TREES (hyperparameters tuning)

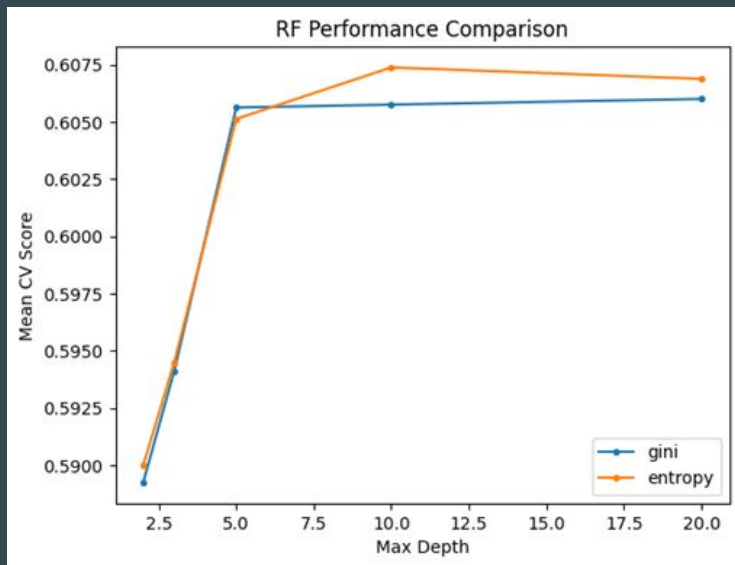


	precision	recall	f1-score	support
popular	0.58	0.56	0.57	1003
unpopular	0.57	0.59	0.58	997
accuracy			0.58	2000
macro avg	0.58	0.58	0.58	2000
weighted avg	0.58	0.58	0.58	2000

RANDOM FOREST (parameters by default)

	precision	recall	f1-score	support
popular	0.61	0.57	0.59	1003
unpopular	0.59	0.63	0.61	997
accuracy			0.60	2000
macro avg	0.60	0.60	0.60	2000
weighted avg	0.60	0.60	0.60	2000

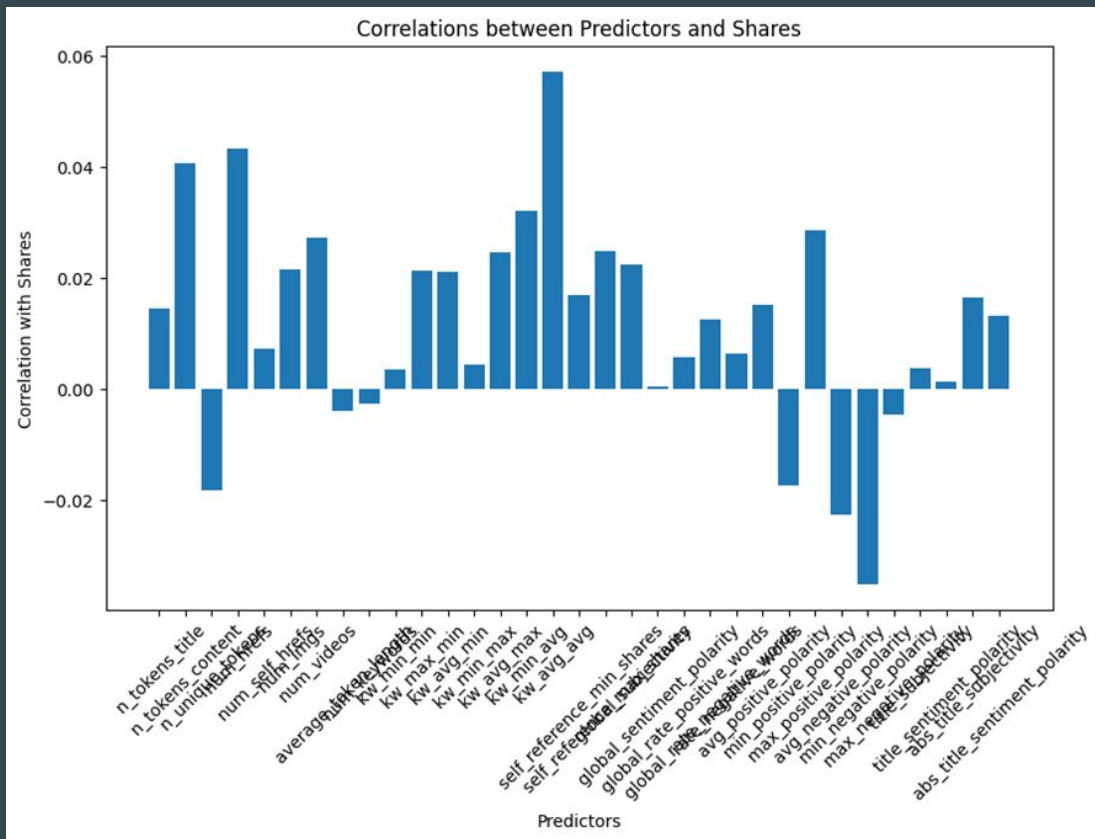
RANDOM FOREST (hyperparameters tuning)



	precision	recall	f1-score	support
popular	0.63	0.54	0.58	1003
unpopular	0.59	0.68	0.63	997
accuracy			0.61	2000
macro avg	0.61	0.61	0.61	2000
weighted avg	0.61	0.61	0.61	2000

LINEAR REGRESSION

- Simple linear regression
- Multiple linear regression



6.CONCLUSION



CONCLUSION

- The objective of the project was to predict online news popularity through data preprocessing, analysis, and modeling.
- Linear regression models did not perform well due to the lack of correlation between independent and dependent variables.
- Classification models such as decision trees and random forests provided better results with the highest accuracy of 61% obtained using random forests with maxdepth=10 and entropy.
- However, even 61% accuracy is not considered high for predicting online news popularity, and the data may not be sufficient to predict this target accurately.
- Future work:
 - Collecting additional information related to the author's reputation could improve the accuracy of the model.
 - Training more models different from the ones used in the project could lead to better results, and it is essential to check the properties of the dataset before deciding on the models to be trained.

THANK YOU