Predict Online News Popularity

The News Squad

📸 **Jingwen Yu, Zhentao Hou, Xiao Chu**

Outline📝

1. Ask - Why we care?
2. Acquire - Introduce the dataset
3. Process - EDA and feature engineering
4. Model - Choose the model based on the evaluation metric
5. Deliver - Takeaway

🧐Ask - Why we care about news popularity

News is an important channel for us to learn what is happening in the world.

A piece of popular news could lead to a successful advertising or public relation activity, contributing great business value to companies.

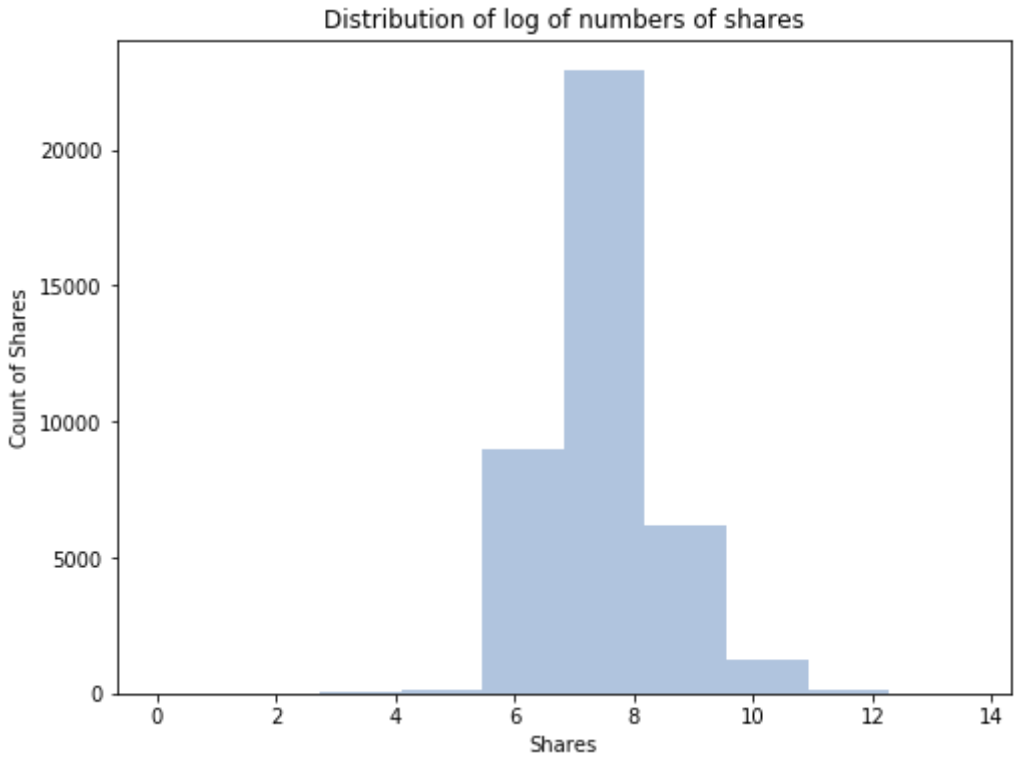
📩Acquire - Introduce the Dataset

1. Number of Features: 58
2. Target Column: Number of Shares
3. Number of Instances: 39644
4. Source: UCI Machine Learning Repository

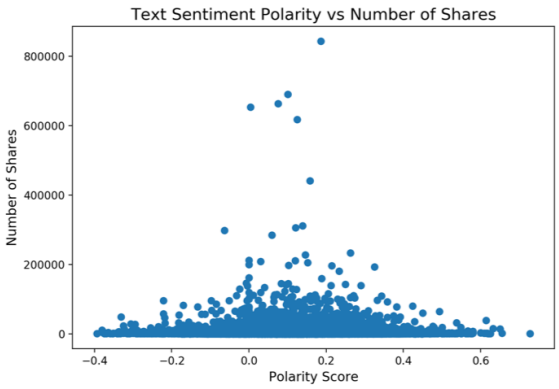
📊Process - EDA and Feature Engineering

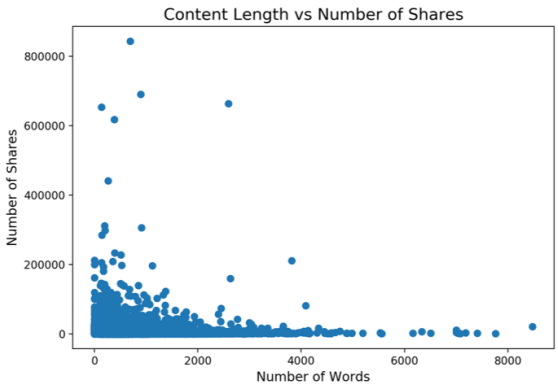
Target Exploration - Raw Dataset

Target Range from 1 to 843300



Feature Exploration





💻Model - Evaluation Metric and Model Selection

Regression Model

- Predict Number of Shares ➡️ **Median Absolute Error (MedAE)**

1. Create Pipelines
   - scaler - StandardScaler()
   - regressor - Lasso, Ridge, and Random Forest Regressor
2. Fit models
3. Evaluation

Model	R^2	Mean Absolute Error	Median Absolute Error
LassoCV	0.011630	3189.03	1678.44
RidgeCV	0.011259	3203.89	1681.21
RandomForestRegressor	-0.052303	3529.62	1597.50

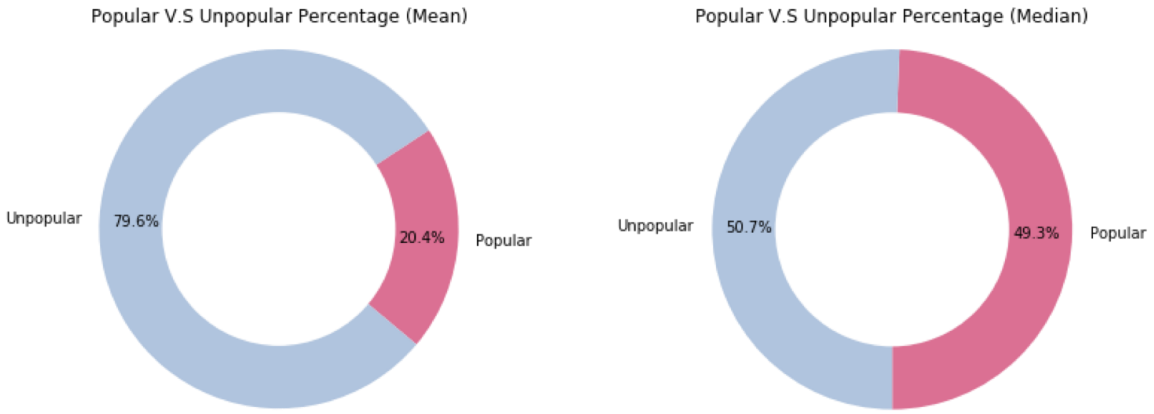
Summary

- 3 regression algorithms, MedAE as the North Star metric
- Large MedAE (> 1600)
- Classification might make more sense

Classification Model

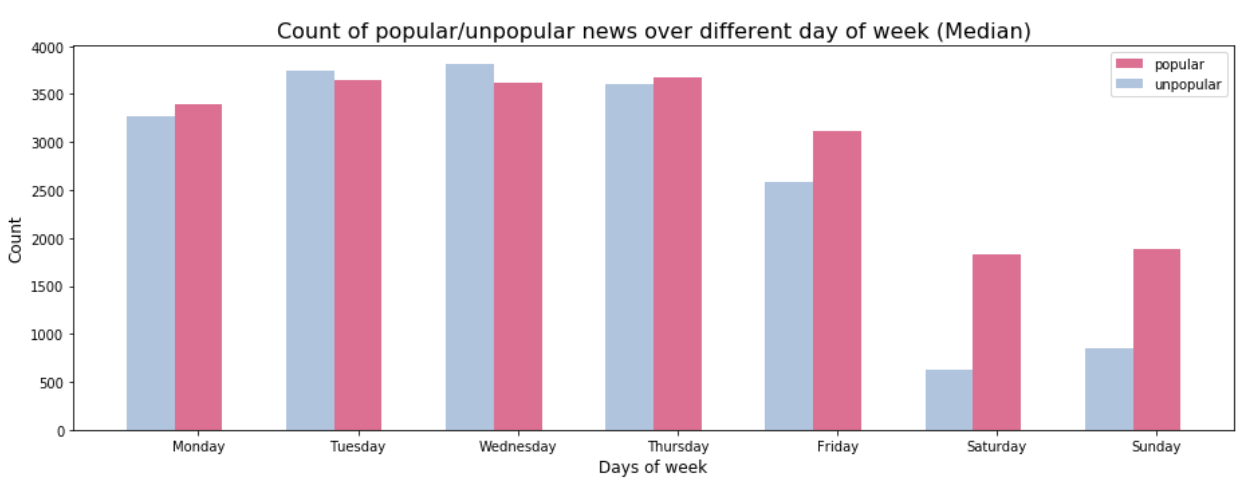
- Binary Classification: Predict Popular or Unpopular ➡️ **F1 Score**

Choosing the threshold -- Mean (3395) or Median (1400)

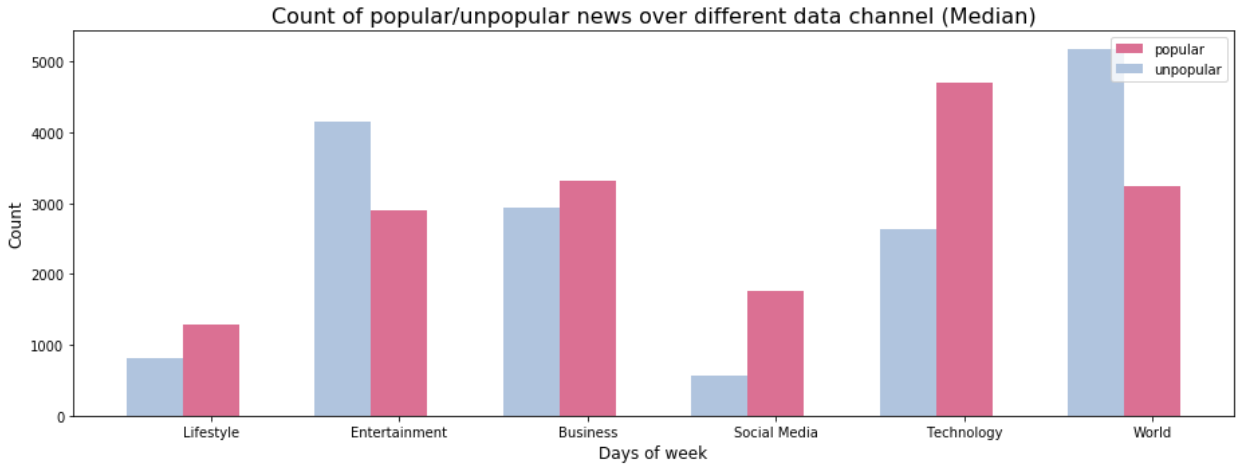


Redo the EDA Again 😹

Weekend or Weekday?



Which Channel?



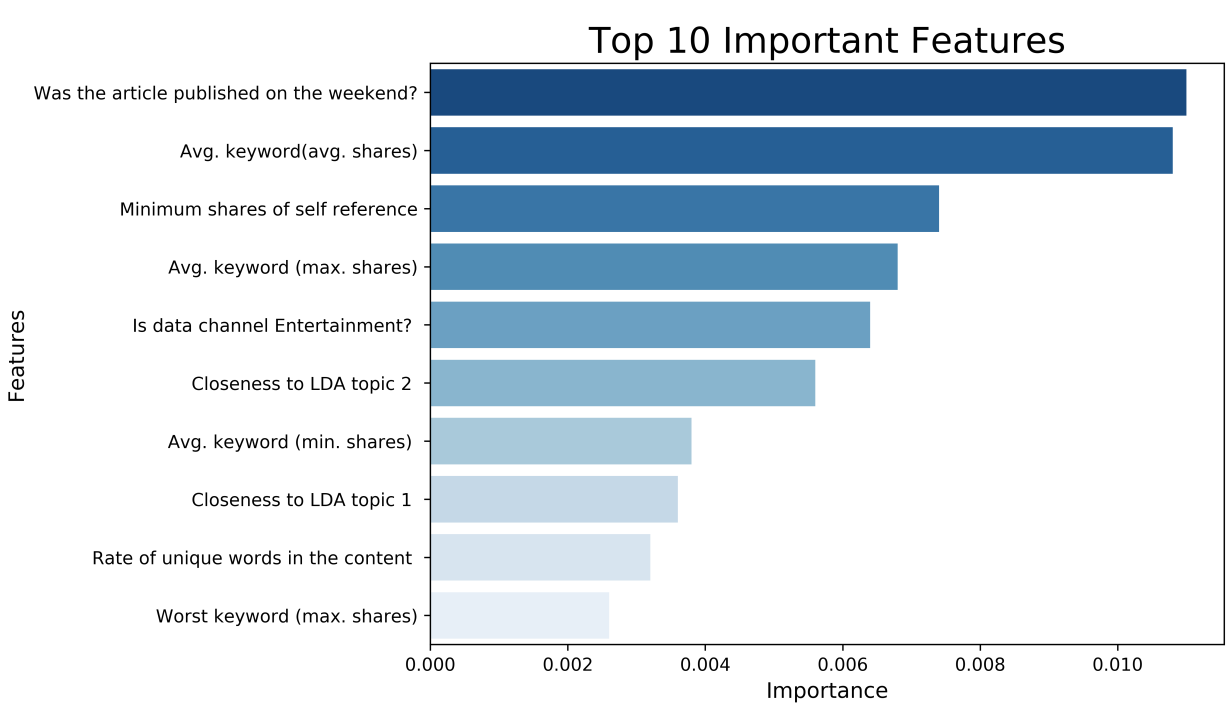
Build Pipelines

- scaler - StandardScaler()
- classifier -
  LogisticRegressionCV
  KNeighborsClassifier
  GaussianNB
  RandomForestClassifier

Fit Models & Model Evaluation

- F1-score = $\frac{2*precision*recall}{precision+recall}$
- F1-score is used when False Negatives and False Positives are crucial
- F1-score is a better metric when there are imbalanced classes

	model	accuracy	precision	recall	f1 score
0	LogisticRegressionCV	0.653424	0.656876	0.627071	0.653168
1	KNeighborsClassifier	0.599823	0.608570	0.535814	0.598168
2	GaussianNB	0.528062	0.642970	0.103747	0.426353
3	RandomForestClassifier	0.660865	0.664009	0.636758	0.660653



✍🏻Deliver - Takeaway

- Using F1-score as the North Star Metric, Random Forest Classifier is the best model.
- Recommendations for reporters and business entities:
  (1) Keywords are important.
  (2) Publication time matters.
  (3) Reference articles with high popularity would help.