Predicting Online News Popularity

Pearly Ang
Darren Lu
Aries Li
Xiaosu Qi
Xavier Sallent

1. Objectives

1. **PREDICT** the number of shares in social networks of specific pieces of news

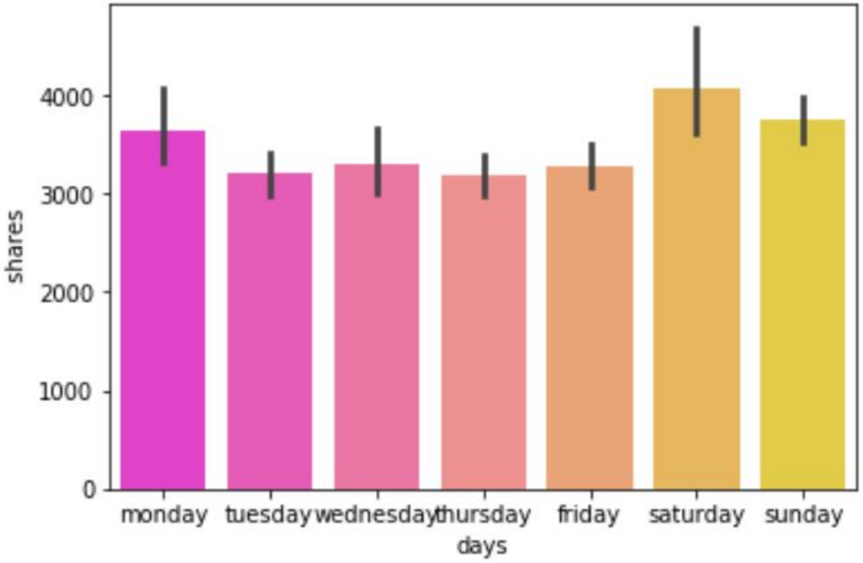
2. **RANK** variables based on contribution to number of shares

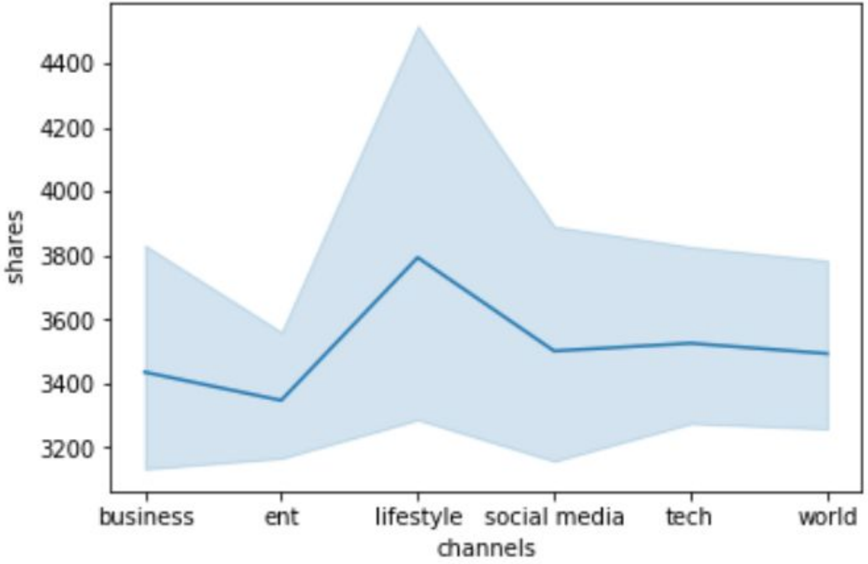
3. **GIVE** insights to online news editors to maximize number of shares, clearly pointing out to potential limitations

2. Exploratory Data Analysis (1/2)

- **Overview**: 59 numerical attributes, a total of 39,644 articles
- **Scaling**: MinMaxScaling (translates continuous feature such that it is between zero and one)
- **Outlier treatment**: removed a variable that distorted the distribution
- **Correlation matrix**: created a correlation matrix to gauge the rough correlations between feature variables

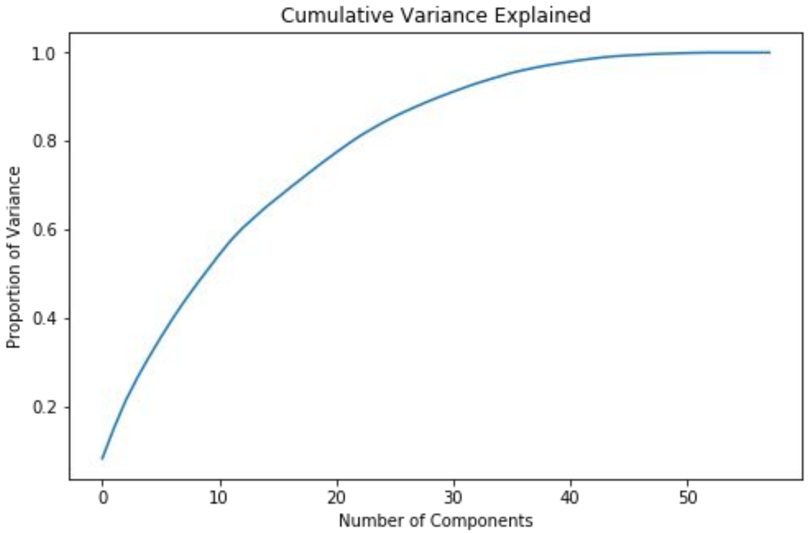
2. Exploratory Data Analysis (2/2)





3. Principal Component Analysis

We need around 35 variables to explain 95% of the variance



- Our original dataset contains 58 features
- Our aim is to reduce the number of variables running a PCA
- The plot shows an *almost* linear relationship between the number of components and the cumulative variance explained
- These are not good news; to achieve a 95% explained variance in the independent set, we still need around 35 variables

4. Models & Comparison (1/2)

Regression Models:

- Scaled continuous data
- Split dataset: Training (75%) & Testing (25%)

Algorithm	**RMSE** **k fold cross-validation: k = 10**
Linear Regression $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i,$	11760.99
Lasso Regression $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$	11656.98
Ridge Regression $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j}^{m} \beta_j^2$	**11024.96**

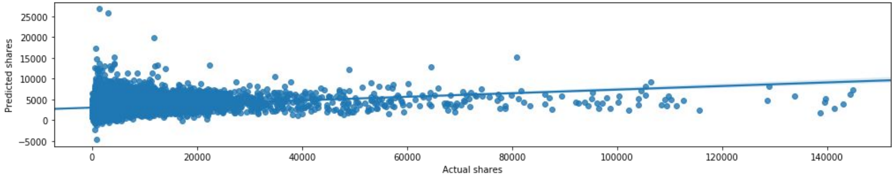


Figure 1: Linear Regression

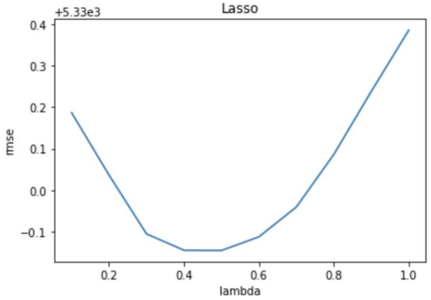


Figure 2: Lambda for Lasso
best Lambda = 0.5

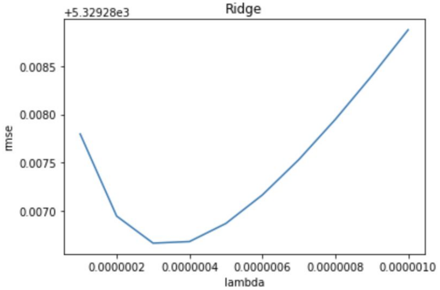


Figure 3: Lambda for Ridge
Lambda can be really low

Therefore, we choose **Ridge Regression** for our final test dataset, the final RMSE is **10659.10**

4. Models & Comparison (2/2)

Classification: Given the features of an article, predict whether the article will be popular or not.

Algorithm	Accuracy
Logistic Regression	0.599
Decision Tree	0.637
Random Forest	**0.661**
K-Nearest Neighbor (k=35)	0.588

Threshold: 1400 shares (median)

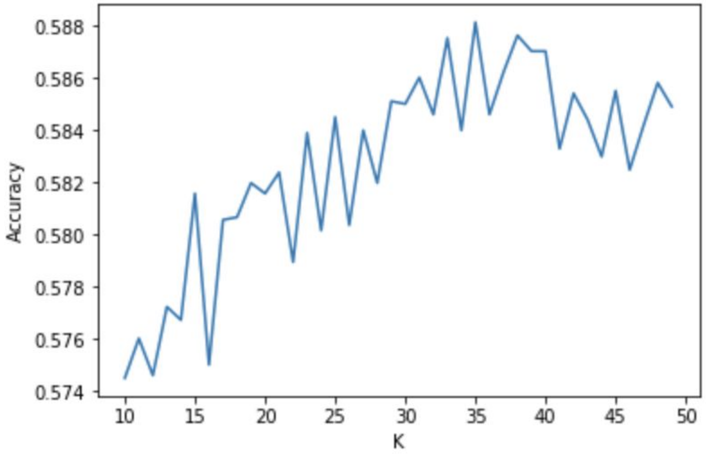
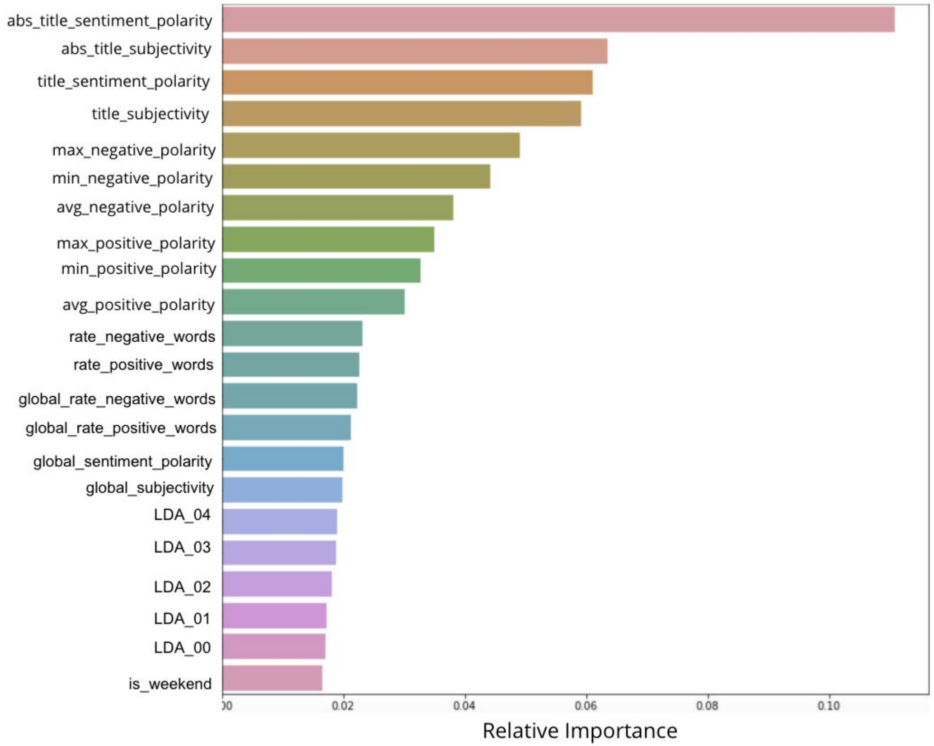


Figure1: KNN

5. Feature Importance (1/2)



Title Sentiment Polarity (Positively Correlated),

Title Subjectivity (Positively Correlated),

Positive Polarity (Positively Correlated),

Negative Polarity (Negatively Correlated),

Published on Weekend (Positively Correlated),

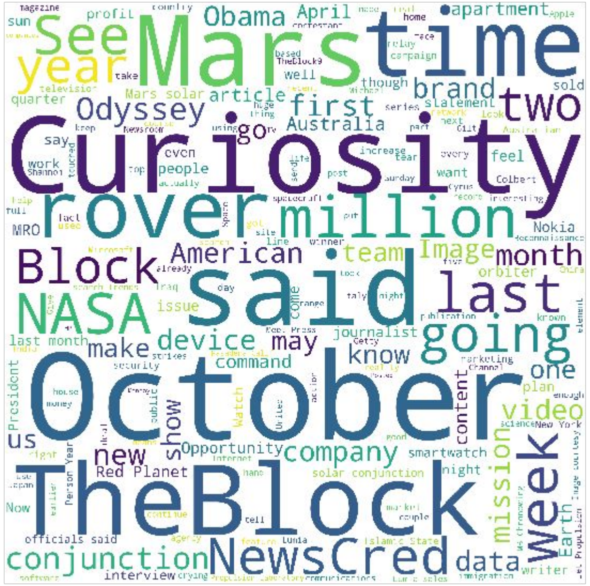
LDA_0,1,2 (Negatively Correlated),

LDA_3 (Positively Correlated)

5. Feature Importance (2/2)

LDA Topic 3: Illustrative Word Cloud

LDA Topic 2: Illustrative Word Cloud



6. The Perfect Piece of News - Conclusions

News editors should:

1. Talk about Sports events
2. Include many links
3. Include many images and videos
4. Write in a subjective way

News editors should not:

1. Talk about Science
2. Categorize articles under the “World” category
3. Write excessively long articles
4. Include negative words

7. Limitations and Further Exploration Ideas

1. **News have been analyzed independently**. Some news may rank high in our model but could be “eclipsed” by others published in the same day.
2. **Readers could face saturation**, especially if a website only focuses on a specific topic which is supposed to attract more attention.
3. **More popular is not always better.** In the long term, news agencies could incur in reputational costs if they only produce content to be shared rather than taking a more responsible and objective approach.