Predicting popularity of online news articles

Aditya Bugalya

Ashwin Tomer

Rupin Oberoi

Sarthak Dixit

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation – 1

- Despite thousands of news articles published everyday by several news organizations, there are some which gain popularity but others not so much. How can we explain it?
- Can we identify characteristics which influence the popularity of articles?
- We were curious to know if we can predict the popularity even before the publication using only the intrinsic features of the article such as its language

Motivation – 2

- A reliable predictor can prove to be helpful for a lot of stakeholders: news organizations, independent journalists and more importantly the readers who would get more engaging content
- The online news industry is growing at a rapid pace and hence approaching this problem is quite relevant

Literature Review

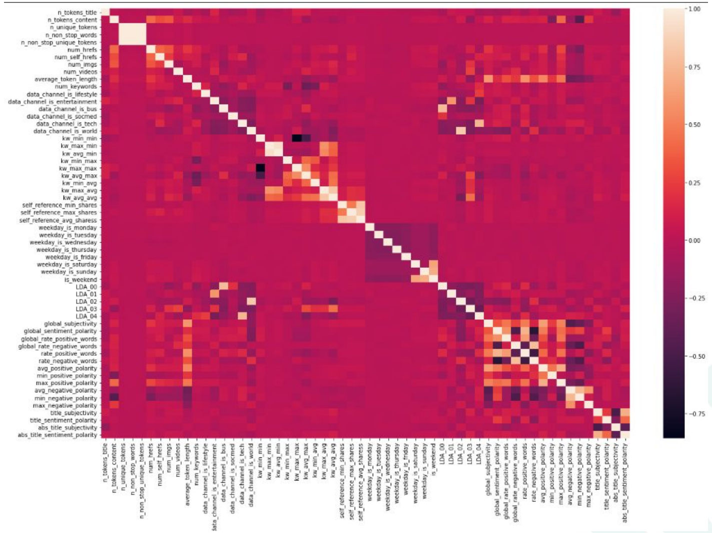
- Roja Bandari et al. in their paper ’The Pulse of News in Social Media: Forecasting Popularity’ have addressed the problem of predicting popularity of news items on social media prior to its release.
- They used features like no. of times a tweet related to the article was shared, news source, article category etc. and assign a score to each news source based on articles shared and t-density(avg no. of shares per article).
- Found that the reputation of the news source has almost no correlation with its popularity.

Literature Review – 2

- Carlos Castillo et al. in their paper “Characterizing the Life Cycle of Online News Stories Using Social Media Reactions” present the study of life cycle of a news article.
- Studied article views in uneventful weeks to avoid unnecessary spike of views towards a particular set of articles.
- Articles are of two types breaking news and in-depth and the way people arrive at them are divided into 4 classes internal, external, direct links and through searches.
- Several variables were computed such as number of Facebook shares/minute, number of tweets/minute, number of followers of the people who tweeted them etc to figure out the trends.
- 3 patterns emerged decreasing(78%), steady or increasing(12%) and rebounding(10%).
- Seeing which in-depth articles are gaining popularity also helps traditional TV news in figuring out what to show to keep up the viewership in case of no breaking news.

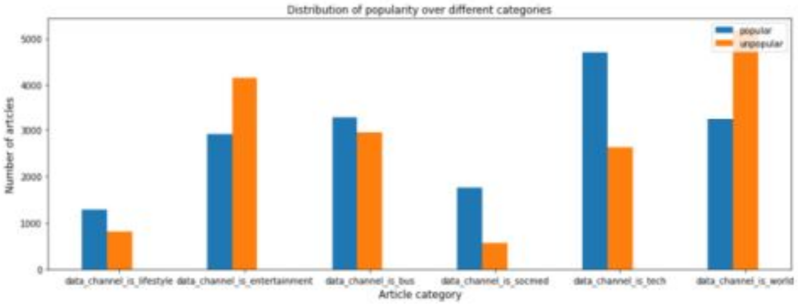
Dataset Attributes and Visualization

- 61 attributes -> 58 predictive, 2 non- predictive and a target variable - number of shares
- 39644 samples consisting of information about articles published by Mashable
- Attributes cover a wide range of information and are very detailed: sentiment oriented, number of stop words, number of images etc.



Dataset attributes and visualisation

- Very few attributes show high correlation between them



app
new york
samsung galaxy
google glass
selfie
china
iphone
startup
video
tweet
fact
dog
social media
super bowl
youtube
obama
australia
ad
twitter
apple
amazon
google
future
space
fan
drone
year
3d printed
netflix
review
star wars
comic
instagram
snapchat
internet
throne
smartphone
hand
viral video
game
facebook
million
online
trailer
song
job
new
photo
christmas
kid watch
best
world cup
gif
tumblr
microsoft
vine
tip
tech
brief
first
robot
nfl

Preprocessing

- **Fundamental Preprocessing:** No null values or invalid values found and articles not older than 2 weeks were removed.
- **Handling Outliers:** Significant no. of outliers were found so they were capped at respective bounds of non-outliers to avoid loss of much data.
- **Converting URLs to article titles:** Extracted information from the URLs using regex and nltk library to maybe explore in the future as URL is not a predictive feature.
- **Feature Selection:** Plotted pairwise heatmaps of features that helped in identifying highly correlated features. Two features were deemed not necessary using this.

Methodology

- Tried **Linear Regression** to fit the data but the results (bias and variance) showed that the model underfit.
- Classified articles with >= median no. of shares as popular and others as unpopular to study this as a classification problem.

- **Naive Bayes Classifier -** One of the fastest classifiers there is, suited to use when no. of features is huge.

- **Logistic Regression -** Performs best on linearly separable data, but difficult to fit complex models on it, may overfit when no. of features is large.

Methodology

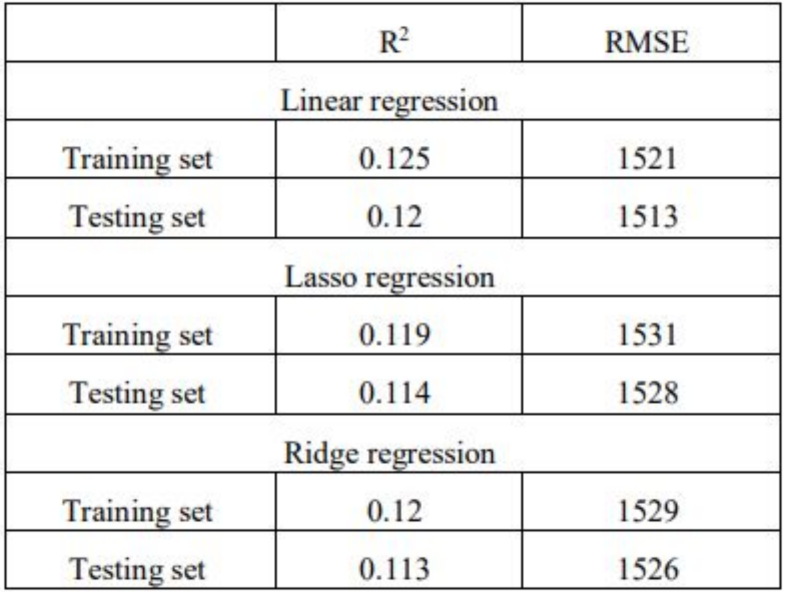
- **Random Forest -** Based on decision trees and bagging it is a very robust classifiers best at handling outliers.

- **Support Vector Machine(SVM):** We tried different models, degrees (in case of poly model), the penalty parameter of the error term.

- **Artificial Neural Network(ANN):** Here, we tried different activation functions like the relu, logistic, identity, tanh, keeping overfitting in check.

Results and Analysis

Regression: High bias, low variance -> Likely underfitting, high RMSE is also explained by the high variance in the target variable (shares)

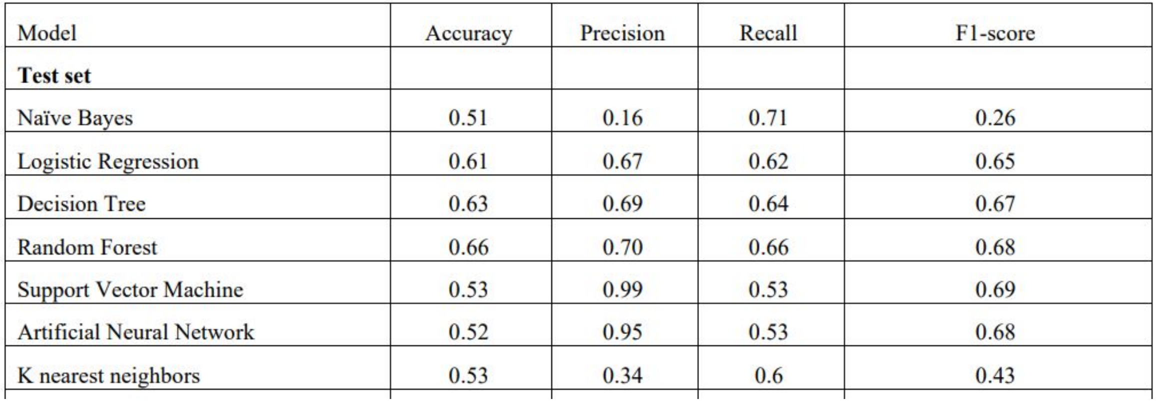


	R^2	RMSE
	Linear regression	
Training set	0.125	1521
Testing set	0.12	1513
	Lasso regression	
Training set	0.119	1531
Testing set	0.114	1528
	Ridge regression	
Training set	0.12	1529
Testing set	0.113	1526

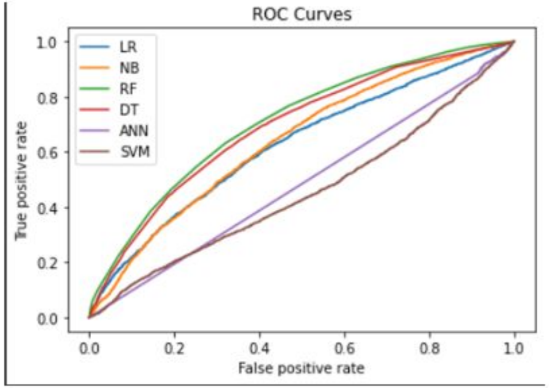
Results and Analysis

Classification:

- Naive Bayes is used as baseline classifier and all other models have significantly better metrics.
- Best performing hyperparameters found using grid search.

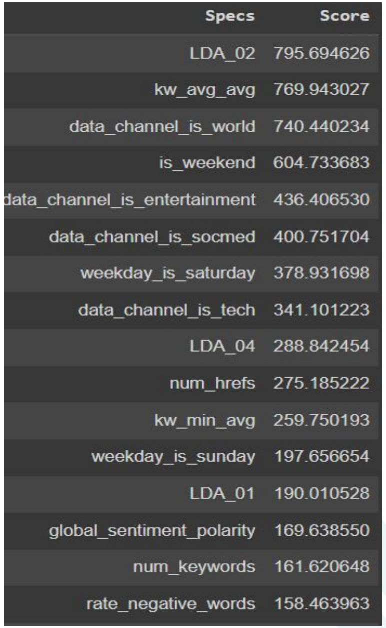


Model	Accuracy	Precision	Recall	F1-score
Test set				
Naïve Bayes	0.51	0.16	0.71	0.26
Logistic Regression	0.61	0.67	0.62	0.65
Decision Tree	0.63	0.69	0.64	0.67
Random Forest	0.66	0.70	0.66	0.68
Support Vector Machine	0.53	0.99	0.53	0.69
Artificial Neural Network	0.52	0.95	0.53	0.68
K nearest neighbors	0.53	0.34	0.6	0.43



Results and Analysis

- Using Sklearn’s SelectKBest method, feature importance was determined.
- Article category, day of the week of publication are among the ones with highest scores
- We tried using different subsets of features by selecting top 20, top 30 etc. features but the results were not significantly different than using all features together.



Specs	Score
LDA_02	795.694626
kw_avg_avg	769.943027
data_channel_is_world	740.440234
is_weekend	604.733683
data_channel_is_entertainment	436.406530
data_channel_is_socmed	400.751704
weekday_is_saturday	378.931698
data_channel_is_tech	341.101223
LDA_04	288.842454
num_hrefs	275.185222
kw_min_avg	259.750193
weekday_is_sunday	197.656654
LDA_01	190.010528
global_sentiment_polarity	169.638550
num_keywords	161.620648
rate_negative_words	158.463963

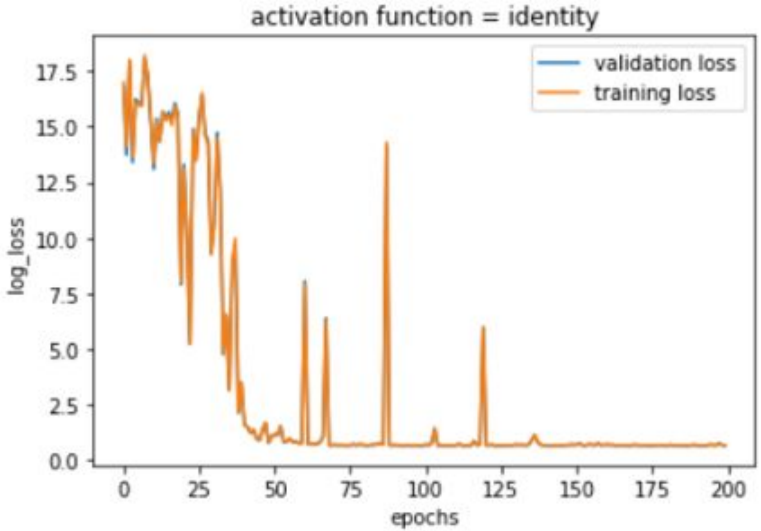
Results and Analysis

For Support Vector Machine(SVM), on using different models, degrees(in case of poly mod), altering the regularization parameter of the error term. We got the highest accuracy on RBF of 58.40% using the penalty parameter of the error term as 1.0 .

Model	Accuracy
Sigmoid	52.74
RBF	58.40
Poly	52.82
Linear	52.36

Results and Analysis

For **Artificial Neural Network(ANN)**, on using different activation functions like the relu, logistic, identity, tanh. We got the highest accuracy, on Identity activation function, of 58.40% at learning rate of 0.001 and hidden layer sizes of 40, 32, 20,12, 8 one after the other.



Conclusion

- We were able to identify the top 25-30 characteristics which influence the popularity of articles. The popularity of a news article depends mostly on just these characteristics.

- We also could predict the popularity of a news article even before it was published saving a lot of time and money of the publisher.

- The readers would also get more engaging articles

Individual contributions

- Aditya Bugalya
  - EDA, Model Training
- Ashwin Tomer
  - Literature Review, Model Training
- Rupin Oberoi
  - Data Preprocessing, Literature Review
- Sarthak Dixit
  - Feature Selection, Data Cleaning, model training