PREDICTING THE POPULARITY OF ONLINE NEWS STORIES

Abstract

In recent years, the style and nature of the Press Releases has changed. Sharing news is clearly impacting on how people find news today. In this analysis, I addressed the problem of predicting online news popularity based on the number of shares of articles. I needed to find a model to explain why some news stories have more share than others and identify a group of variables to predict popularity.

Key Findings

Natural Language Processing techniques were implemented to create features extracted from the content of the article and machine learning algorithms such as Naive Bayes substantially improved the accuracy of prediction by 0.87 and the ROC Curve to 0.95.

Some of the biggest influencing factors that result in a news article being popular and experiencing the highest volume of media sharing are ones that: convey positive messaging, have the placement of a single number in the title and contain strong visual imagery. This is important in considering the number of keywords in the article.

1.Introduction

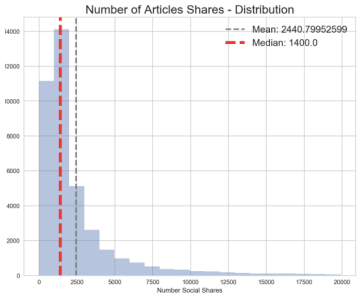
My recent professional experience in studying new ways of consuming media and entertainment included different forms from linear TV, streaming, video on demand, apps, … which aroused my curiosity about understanding how people find online articles today.

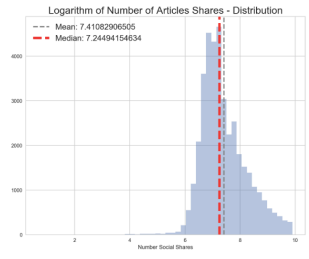
In recent years, the style and nature of Press Releases has changed. Sharing news is clearly impacting on how people find news today. In this project, I addressed the problem of predicting online news popularity based on the number of shares of articles. I needed to find a model to explain why some news stories have more shares than others and identify a group of factors to predict popularity. I was using the dataset is by Mashable and scraping in its web. Mashable is a well-known global multi-platform media and entertainment company.

Initially, I formulated the prediction task as a regression linear where the target is measured at the continuous level. However, the target showed a high variance and I observed that transforming the target into a two-class label with a partition at the median improved the accuracy of prediction and provided an acceptable model. Therefore, I extracted features from the news items and applied Bayesian algorithms. As a result, popularity was correctly classified in 87% of cases. In addition, I was using Sentimental Analysis to find the biggest influencing factors that result in a news item being popular.

1.1. Target

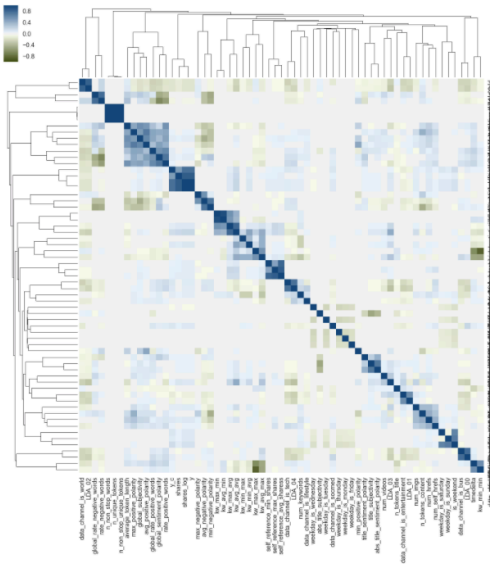
The target variable shows a high variance. I discretized the target value to binary category. To do this, I used median as measure of central tendency. The median is the value which occupies the middle position. The article considers being popular if the number of shares exceeded 1,400 social shares else it's classified as unpopular.





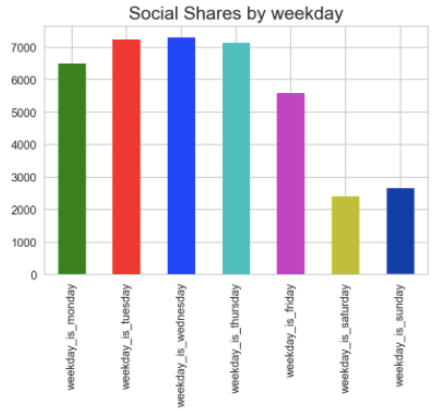
1.2. Features Engineering

This is a collection of almost 40,000 news with 58 variables. Some of them are relate to meta-data such as channel type or day the new was published. Also, the dataset contains natural language processing features (The Latent Direchlet Allocation was applied to all texts). To start, I applied a cluster map to examine the correlation among variables. This seems that some of them have high correlation. However, correlated information between features is limited.

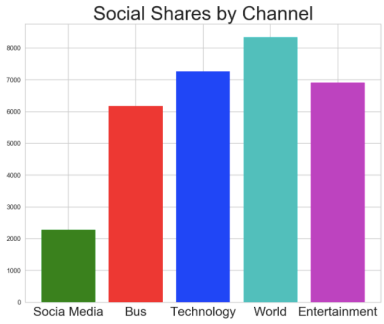


I collected all the past data and then searching for the pattern in this data.

I identified some facts in the dataset. Distribution of weekdays showed that the number of shares are highest during the middle of the week and decreased during the weekends; on Tuesday, Wednesdays and Thursday are higher than the rest of days. After some analysis, I founded that a weekend article had less opportunity to be shared as referenced article.



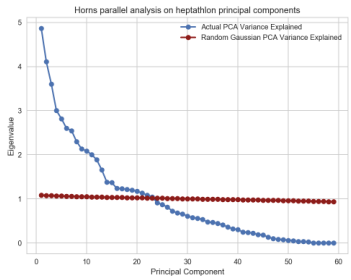
By channels, the number of shares articles in World channel is the highest. It is followed by Tech News and Entertainment. We do not have information if any channel could be recorded for more than one category.

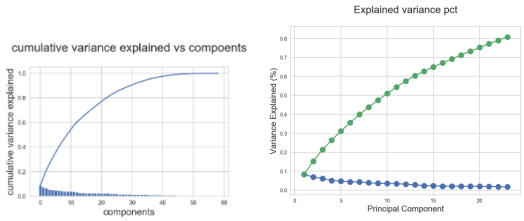


I found some limitations in the Dataset such as there is no information about the relationship between the number of time an article is shared vs the amount time the article was online.

1.3. Dimensionality reduction

Then, given the large number of variables, I considered it important to apply Principal Component Analysis to identify patterns and reduce the dimensions of the dataset with minimal loss of information. Also, I applied Horn's Parallel Analysis to determine the accurate number of components. This is the gold standard in determining which components are not noise.





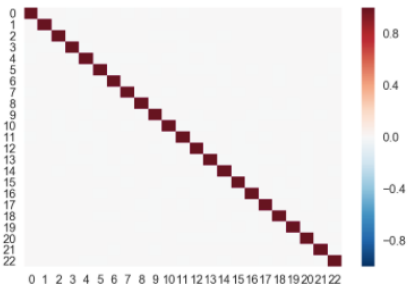
Almost 80% of the explained variance in the data can be explained by just 23 of the principal components. The dimensionality has been reduced by half.

As the number of variables in the dataset is large, the analysis shows similar values in many of the coefficients.

Component interpretation:

- PC0 data_channel_is_world
- PC1 rate_negative_words
- PC2 kw_avg_avg
- PC3 n_unique_tokens
- PC4 kw_avg_min
- PC5 self_reference_avg_sharess
- PC6 LDA_00
- PC7 kw_max_min
- PC8 LDA_00
- PC9 abs_title_sentiment_polarity
- PC10 n_tokens_content
- PC11 data_channel_is_entertainment
- PC12 LDA_01
- PC13 max_negative_polarity
- PC14 kw_min_max
- PC15 weekday_is_wednesday
- PC16 weekday_is_tuesday
- PC17 weekday_is_monday
- PC18 weekday_is_monday
- PC19 weekday_is_monday
- PC20 num_videos
- PC21 data_channel_is_socmed
- PC22 data_channel_is_socmed

Each column in the transformed data is no longer correlated.



1.4. Clustering Analysis: K-Means and Hierarchical clustering

I used the PCA method as a step-in dimensionality reduction in the dataset before clustering algorithm as K-Means and Hierarchical methods. The Clustering models determine the distance of articles based on a few parameters from the centroid of clusters.

Number of clusters used is 3 (K=3).

K-Means

```
KMeans clustering:
Silhouette score: 0.207201788072
Homogeneity score: 0.195510046802
Completness: 0.735793987952
```

Hierchical clustering

```
Hierarchical clustering:
Silhouette score: 0.204787168484
Homogeneity score: 0.181526740564
Completness: 0.696952074899
```

Note: There is a total of 38818 rows in the data due to lack of computer resources to run of them. The analysis is based on 100 rows.

Similar scores values for both methods.

High Completeness score (0.735 with Kmeans) close than 1, stands for perfectly complete labelling.

Although there seems to be some difference between the three segments, after running both the Kmeans and Hierarchical clustering test numerous times it becomes clear that the distance between each segment is not that large. Then, it does not make much sense to segment the articles based on these factors.

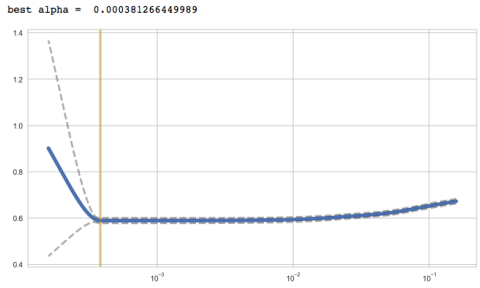
2. Machine learning algorithms

2.1. Linear Regression

Since I had applied PCA to reduce the dimensionality I divided the analysis in both: with all features and with the components from PCA.

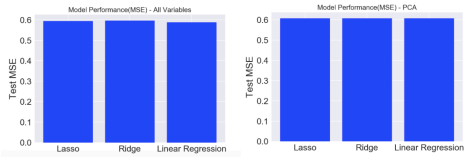
I applied Linear Regression both all principal features and PCA and I used Ridge looking to mitigate multicollinearity and Lasso to penalize the absolute size of the regression coefficients. It can reduce the variability and improving the accuracy of linear regression models. Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

I plotted of the mean CV MSE, and some bands for the standard deviation of MSE along the alphas.



Model Performance comparison – Mean Squared Error

The Mean Absolute Error (or MAE) is the sum of the absolute difference between predictions and actual values. It gives an idea of how wrong the predictions were. And, the Mean Squared Error provide a gross idea of the magnitude of error. Taking the square root of the mean squared error converts the unit back to the original units of the output variable.

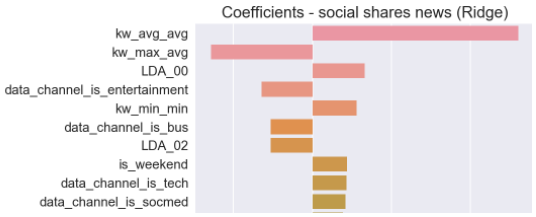


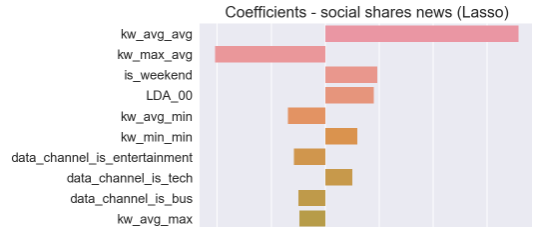
In both cases I applied Lasso and Ridge. Lasso does slightly better than linear regression but the difference is not substantial.

Initial variables do not suffer multicollinearity, independent variables are not highly correlated, as Ridge and Lasso Regression have not reduced the standard errors.

Due to the high variance of the target, linear Regression does not seem the best model.

Most importance/Predictive Features





The top features with highest score are: kw_ave_max, kw_max_max (the lower is the maximum influence) coinciding with Lasso coefficients.

The three highest predictors are related to content. This seems that the current content of an article its more predictive than the rest of metadata variables.

2.1. Classification models

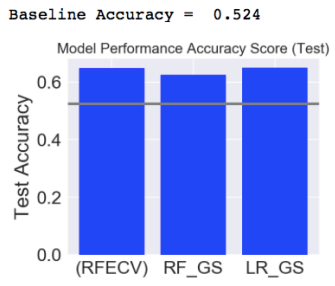
As I mentioned above the alternative approach is to transform the data into two categories (unpopular or popular news articles) to try to improve accuracy.

As in the linear regression analysis, I divided the analysis into both: with all features and with the components from PCA.

2.1.1. Features: All Initial variables

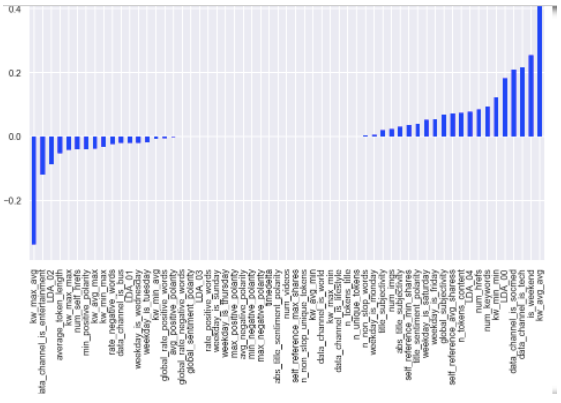
Sklearn also offers recursive feature elimination as a class named `RFECV`. Used this technique in combination with a logistic regression model to see what features would be kept with this method. Also, Logistic Regression with optima parameters, Random Forest Classifier and feature elimination using the lasso penalty.

In this case, I am using Accuracy to compare models. Classification Accuracy is the number of correct predictions made as a ratio of all predictions made. This metric is suitable when there are an equal number of observations in each class and that all predictions and prediction errors are equally important.



Using feature selection techniques, like RFECV, do not seem to improve information compared to using all the features.

From the 53 selected features, the top variables make the most sense in trying to estimate the number of shares:



Surprisingly, the TOP3 predictor are the same than linear regression.

It seems that the editorial content of an article its more predictive than the rest of features.

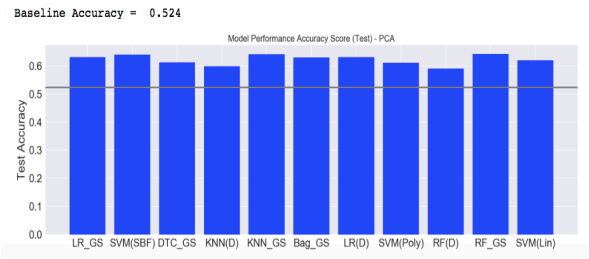
2.1.2. Features: PCA Components

I explored whether PCA components could improve the model. I examined different classifiers starting with Logistic Regression, KNN, Random Forest, Decision Tree, Bagging Classifier and Support Vector Machine with the linear and poly basis function kernels. Each model was optimised using techniques such as Gridsearch for parameter tuning to ensure optimal results.

KNN can be used for both classification and regression predictive problems. In classification problems, KNN determines the class of each testing sample by taking the majority vote from its k nearest training samples. The strength of this model is that it is robust to fluctuations in feature values, taking an aggregate similarity measurement across features (L2-Norm). A weakness of this model is that it assumes which distance metric to use.

A Random Forest model is a meta estimator that fits many decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Also, since it uses decision trees it can learn non-linear hypothesis and the top two classifiers in terms of performance.

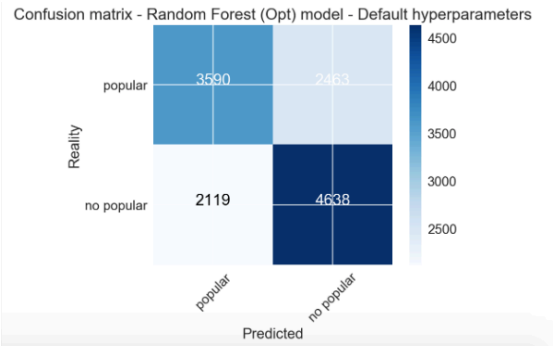
The Support Vector Machine (SVM) algorithm is a different approach to classification. SVM still fits a decision boundary like a logistic regression, but uses a different loss function called the "hinge loss" (as opposed to the log loss in logistic regression). The SVM classifier was trained with linear, polynomial, and radial kernels. An advantage of SVM is that it is guaranteed to converge to a global minimum. However, the computation time required is comparatively higher and it does not return probabilistic confidence.

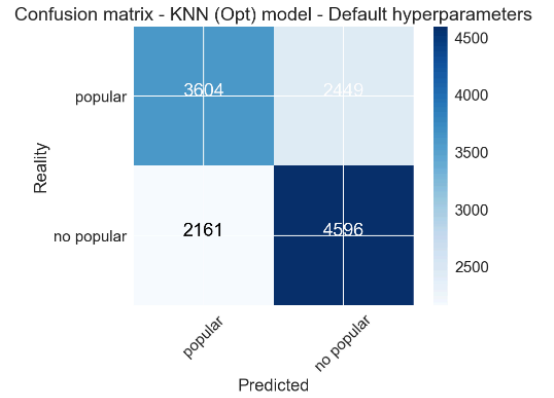


I found that Random Forest classifier, KNN and SVM can be considered as best models across different variations of the data set.

Analysing best model - The confusion matrix is a handy presentation of the accuracy of a model with two or more classes.

The table presents predictions on the x-axis and accuracy outcomes on the y-axis. The cells of the table are the number of predictions made by a machine learning algorithm. This allows more detailed analysis than mere proportion of correct classifications (accuracy).





The two final tables of confusion contain the average values for all classes combined.

We can observe that the confusion matrix is very similar but KNN (Opt) performs slightly better in the outcome True positives (3604 Vs 3590).

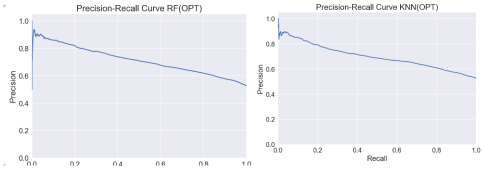
Scikit-learn does provide a convenient report when working on classification problem to give you a quick idea of accuracy of a model using a number of measures.

The classification report() function display the precision, recall, f1-socore and support for each class. The F1-Score is robust because is shows an unweighted, fair measure for accuracy by taking the harmonic mean of precision and recall.

```
RF(600)classification report:
              precision    recall  f1-score   support

           0       0.63      0.59      0.61      6053
           1       0.65      0.69      0.67      6757

avg / total       0.64      0.64      0.64     12810

SVM(rbf)classification report:
              precision    recall  f1-score   support

           0       0.63      0.60      0.61      6053
           1       0.65      0.68      0.67      6757

avg / total       0.64      0.64      0.64     12810
```

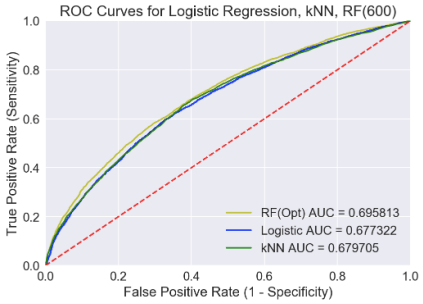
We can see that the classification report is very similar too.

Plot the Precision-Recall curve



Precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

The area ROC curve (or AUC for short) is a performance metric for classification. The AUC represents a model's ability to discriminate between positive and negatives classes. An area of 1.0 represents a model that made all prediction perfectly. An area of 0.5 represents a model as good as random.



ROC curve Random Forest with opt parameters shows the highest value, 0.695.

In this analysis, I implemented more than 10 different machine learning models and each model was optimised using techniques such as Gridsearch. Nearly two thirds of the almost 40000 observations were used for training, and the remaining, about one third were used for model validation. Most models achieve an accuracy score of at least 65%.

3. NLP: Predicting with the article content

Given that the editorial content of an article its more predictive than the rest its features. I added a new analysis, predicting the online new items popularity with the article content.

I extracted the information from the website using Web scrape and the URLs. Due to the difficulty of computer resources for running many rows, the analysis is based on 2000 rows only.

3.1. Data transform 1. TF-IDF

I tried to find words that were characteristic of certain documents using the TfidfVectorizer in scikit-learn. This would create a tfidf matrix with all the words and their scores in all the documents.

TF-IDF stands for "Term Frequency, Inverse Document Frequency". It is a way to score the importance of words (or "terms") in a document based on how frequently they appear across multiple documents. If a word appears frequently in a document, it is important. Give the word a high score. But if a word appears in many documents, it is not a unique identifier. Give the word a low score. Therefore, common words like "the" and "for", which appear in many documents, will be scaled down. Words that appear frequently in a single document will be scaled up.

Also, I filtered the data using Select K best according to the K equal to 500 highest scores. I applied score function parameter chi2 given that this function weeds out the features that are the most likely to be independent of class and therefore irrelevant for classification.

I divided the search engines into two groups: those belonging to the body and those to the title. Using the words in the title and the words in the body of an article as additional features.

3.2. Machine Learning

I implemented machine learning algorithms like Naïve Bayes. In addition, I applied Logistic Regression and Random Forest algorithms. As has been seen before, the benefits of Random Forest algorithm and is easy to learn and use.

Multinomial, Bernoulli and Gaussian models improved significantly the accuracy of prediction combined with the previous work.

Multinomial is a Bayes' theorem based on conditional probability. The conditional probability helps to calculate the probability that something will happen, given that something else has already happened.

Multinomial Naive Bayes is a specialized version of Naive Bayes that is designed more for text documents. Whereas simple Naive Bayes would model a document as the presence and absence of particular words, Multinomial Naive Bayes explicitly models the word count and adjusts the underlying calculations to deal with it.

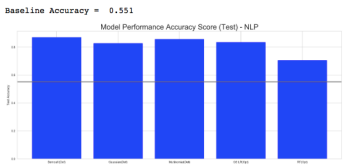
The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word count for text classification).

Bernoulli, like Multinomial, as a classifier is suitable for discrete data. The difference is that while Multinomial works with occurrence counts, Bernoulli is designed for binary/Boolean features.

A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It is specifically used when the features have continuous values. It is also assumed that all the features are following a Gaussian distribution i.e, normal distribution.

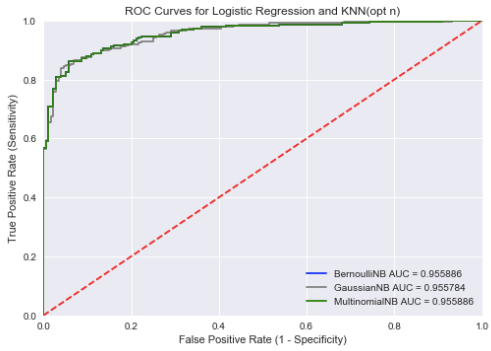
As mentioned above, the choice of metric to evaluate the models depends on the type of model and the implementation plan of the model.

I started with accuracy metric for all the models. This measure computes the proportion of correct predictions.



Naïve Bayes models achieved better results over the other models, substantially improved the accuracy of prediction by 0.87 Vs 0.65 in the previous analysis.

Then, I focussed on the models with better results. The biggest advantage of using the ROC curve is that it is independent of the change in proportion of responders. The ROC curve compares the relationship between True Positive Rate and False Positive Rate. The point along the curve that maximizes True Positive while minimizing False Positive is essential.



ROC is 0.95 which means the area under the curve is 0.95. (1 is equivalent to perfect prediction)

Similarly ROC curves resultd in the three models.

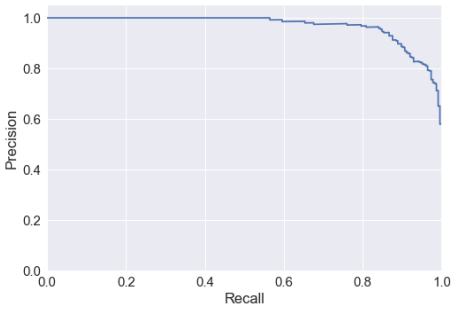
The classification report() function

	precision	recall	f1-score	support
0	0.92	0.74	0.82	175
1	0.83	0.95	0.88	225
avg / total	0.86	0.86	0.85	400

The F1-Score is robust because it shows an unweighted, fair measure for accuracy by taking the harmonic mean of precision and recall, reaching 0.85 with Bernoulli model. The scores corresponding to every class will tell you the accuracy of the classifier in classifying the data points in that particular class compared to all other classes.

The support is the number of samples of the true response that lie in that class.

The precision-recall plot is a model-wide measure for evaluating binary classifiers and closely related to the ROC plot.



Above the precision-recall curve of an almost perfect classifier. A classifier with the perfect performance level shows a combination of two straight lines – from the top left corner (0.0, 1.0) to the top right corner (1.0,1.0) and further down to the end point (1.0, P/(P+N)).

Extracting directly all the words in an article as additional features, and then applying machine learning algorithms such as Naive Bayes and Random Forest substantially improved the accuracy of prediction (0.87 Vs 0.65), the ROC Curve (0.95 Vs.) and F1 score.

3.3. Data Transform. Sentimental Analysis TITLE

One thing that caught my attention was the news headlines. I considered it important to know what characteristic in the title results in an article being popular and getting the highest volume of media sharing.

Sentimental analysis can prove a major breakthrough for the title of the article. The key to running a successful article with the sentimental data is the ability to exploit the unstructured data for actionable insights.

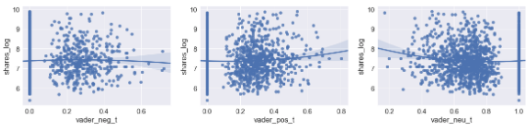
Sentimental Analysis is the use of Natural Language Processing techniques to extract subjective information from a piece of text. i.e. whether an author is being subjective or objective or even positive or negative. This can also be referred to as Opinion Mining.

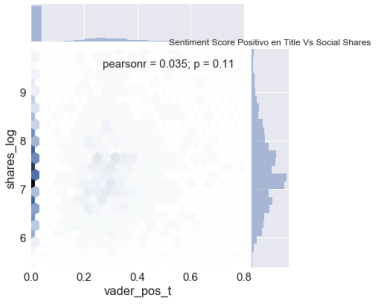
I used the VADER library to get better sentiment scores. VADER Sentiment Analysis. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.

I parsed the text with `SentimentIntensityAnalyzer`.

Through Sentimental Analysis, new variables would be added to represent how positive or negative the text is.

I analysed the correlation between the social shares news and the score in the title.





I saw there was a positive correlation between the social shares news and the positive score in the title.

Then, I decided to determine which characteristic showed the top most positive features in the title (top10).

```
Happy Superb Owl Sunday!
Puppies Adorably Predict Super Bowl Winner [VIDEO]
Introducing the Smartest Cat Alive
Journey + Memes = Faithfully Amazing [VIDEO]
9 Fresh YouTube Shows You'll Love
SAG Awards Recap: Best Moments and Acceptance Speeches
8 Romantic Gifts for Space Lovers
10 Best YouTube Channels for Free Fitness Videos
Twitter, You Won the Super Bowl
Government Wants to Create Free Public 'Super Wi-Fi'
```

I noticed that there were 3 titles which included a number and two of them had video. It seems that strong visual and including numbers in the title creates impact to share.

What were Hot in titles

The word cloud shows that specific words should be more informative, descriptive and technological.

This suggests that the influencing factors that result in the title of an article being popular and experiencing the highest volume of media sharing are ones that: have the placement of a single number in the title, contain strong visual imagery such as videos and convoy positive messaging.

References

>- Data Science central website. http://www.datasciencecentral.com

>- The Elements of Statistical Learning. Authors: Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome

>- He Ren and Quan Yan - Predicting and Evaluating the popularity of Online News. https://pdfs.semanticscholar.org/9e91/6a3469e9e2fc5f0c8f927d7d1d05f5575729.pdf

http://stevenloria.com/finding-important-words-in-a-document-using-tf-idf/