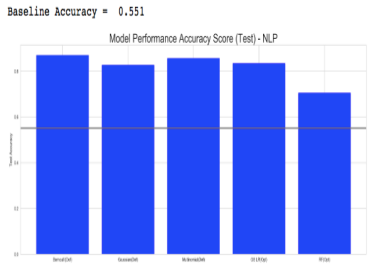
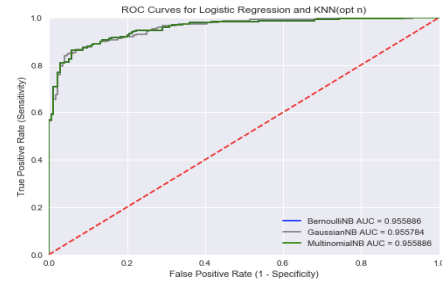
Predicting the popularity of online news

Extracted a broad set of features that are known prior to to an article publication, in order to predict its future popularity, under a binary classification task.

Naïve Bayes algorithm: Accuracy 0.87 & AUC 0.95 (Baseline 0.5)

Features extracted from the article content : Tfidf & Select K best

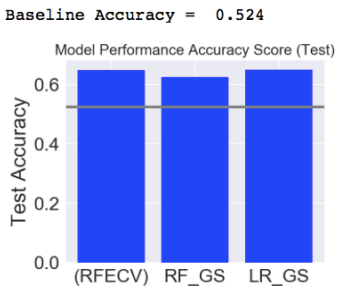




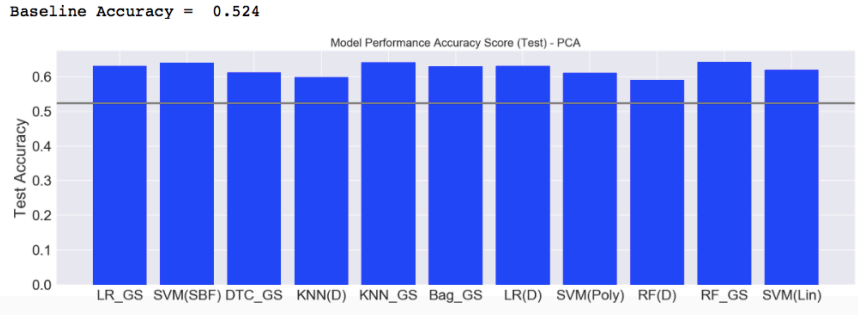
Classification models : Implemented more than 10 models

Random Forest & KNN algorithm: Accuracy 0.65 (Baseline 0.5)

Features: All variables

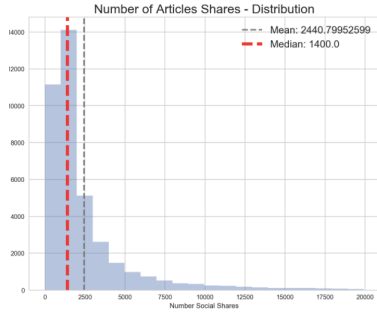


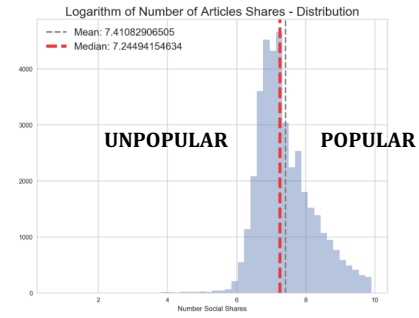
Features: PCA Components



Applied Linear Regression to get a start. But, the target shows a high variance

Target



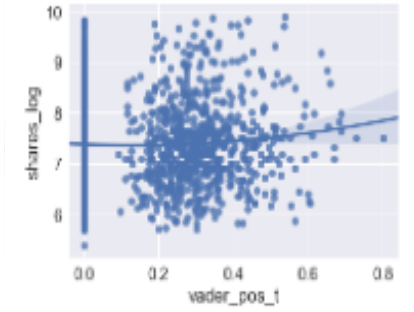


Predicting the popularity of online news

The biggest influencing factors that result in a particular 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

Correlation: score positive in the title Vs. social shares



Top 10 titles with positive score

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'

specific keywords maybe more informative and technological

What words were Hot in Titles

consideration the number of keywords

Feature importance

- Avg. keywords (avg. shares)
- Avg. keywords (max. shares) the lower is the maximum influence
- Entertainment Data Channel

NLP: Sentimental Analysis in the Title. Based on 2000 articles of the news
Ranking of features accoridng to their importance in the Random Forest model