Predicting Online News Popularity

Mo Hannan & Aroa Gomez

Objectives

- Prediction of news articles popularity prior its publication
- Explore what features could help to improve popularity:
  - Day of publication
  - Article length
  - Topic sentiment
  - Channel

News UK

THE VOICE

Mashable

Data & Methodology

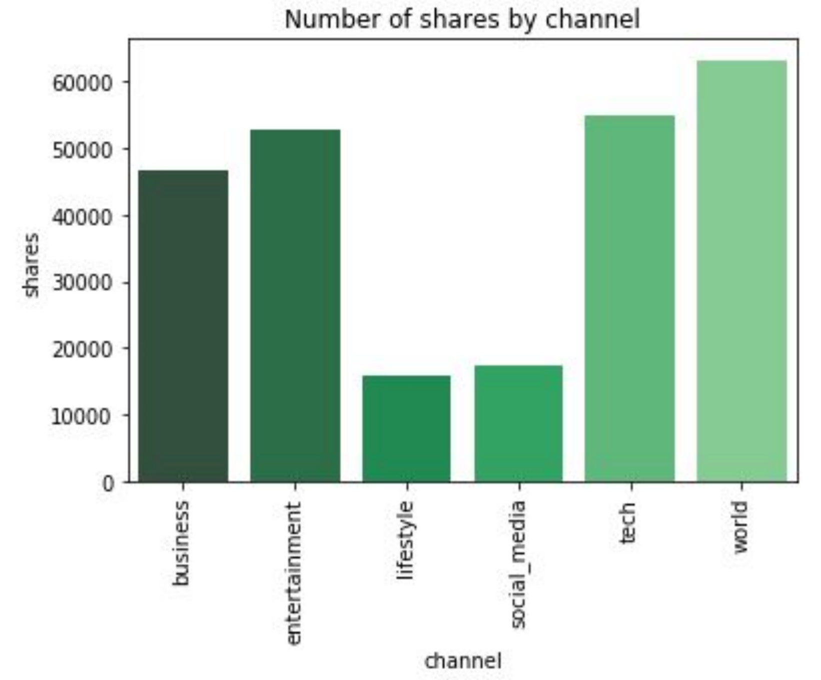
Data Set

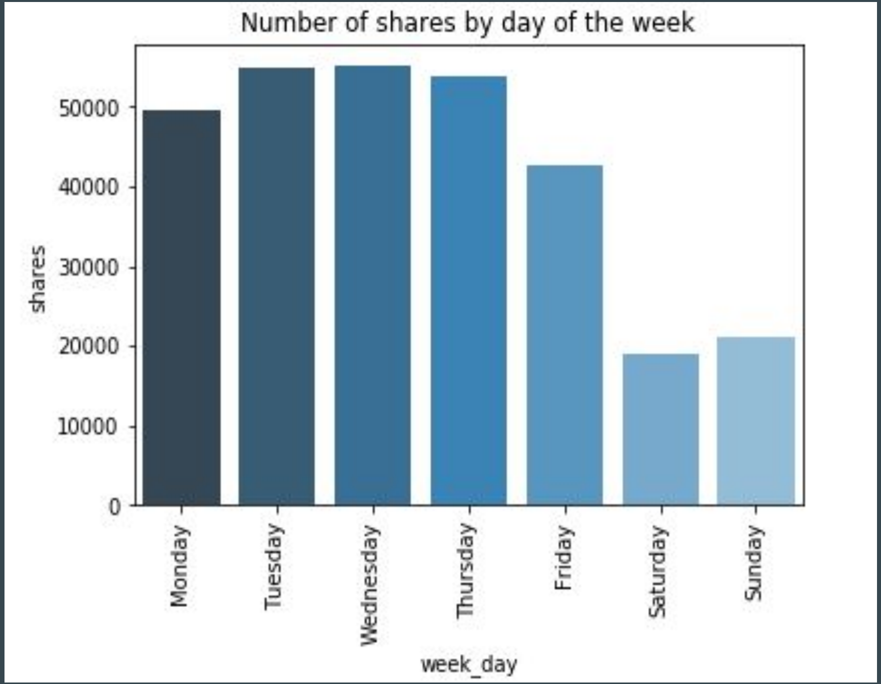
★ Articles published by Mashable in 2015
★ 58 predictive features
★ About 40,000 articles

Methodology

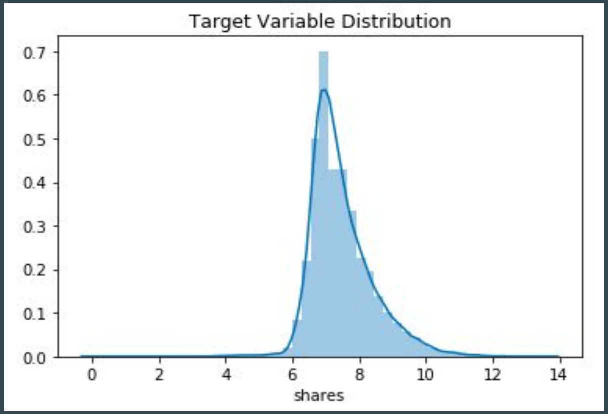
★ Popularity: It’s defined by the number of times an article gets shared
★ Target: estimate number of shares
★ Analysis: Linear and Polynomial Regression

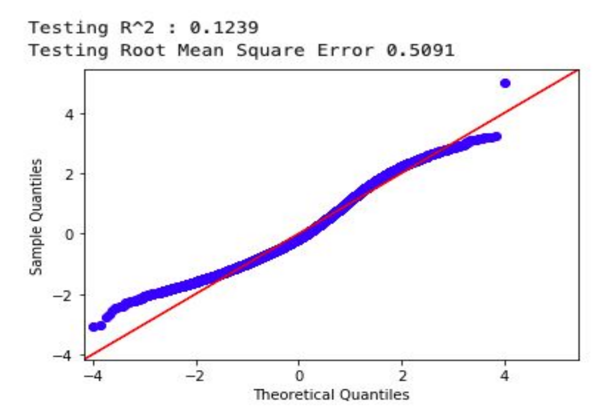
Popularity by Channel and Week Day





Regression Analysis





```
Testing R^2 : 0.1239
Testing Root Mean Square Error 0.5091
```

Linear Regression

- We start with 30 predictive variables
- **R**^{2} : 0.10
- MSE: 0.51

Polynomial Regression - Quadratic

- After removing non significant features (p-value>0.05) we reduced the model to 20 variables
- **R**^{2}: 0.12 (20% increase vs initial model)
- MSE: 0.50

Findings and Next Steps

Recommendations

- The features gathered by UCI on Mashable.com news articles have little predictive power
- News published on the weekends are less popular than those published during the week
- Lifestyle and Social Media articles are the least popular

Next Steps

- Other type of regressions could be more suitable for this data set. For instance logistic regression to predict popular vs non-popular articles
- Increasing the data set and looking at different news providers could improve the predictability of the model