Online News Popularity Prediction

Data Scientist : Chun Liu

Data Description

- **39,644** articles published by Mashable from 2013 to 2015

- **58** predictive features:
    - number of words, links, images, videos, day of the week, article category, etc.

- **Target variable:** number of article shares

- **Goal**: build a prediction model to help publishers to maximize popularity of their articles and sell advertisement

```
Data Cleaning  →  Label Classes  →  Normalization
```
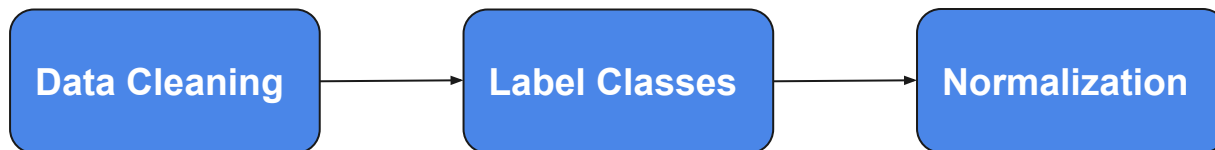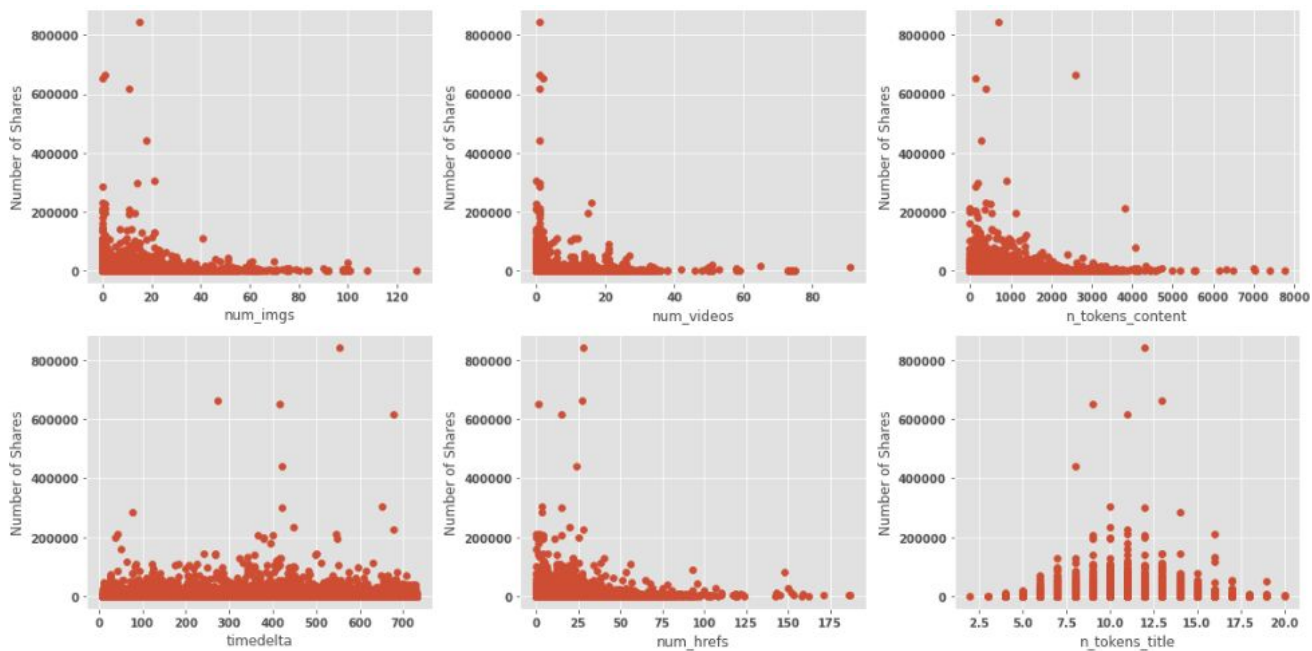
Feature Importance

Some important features
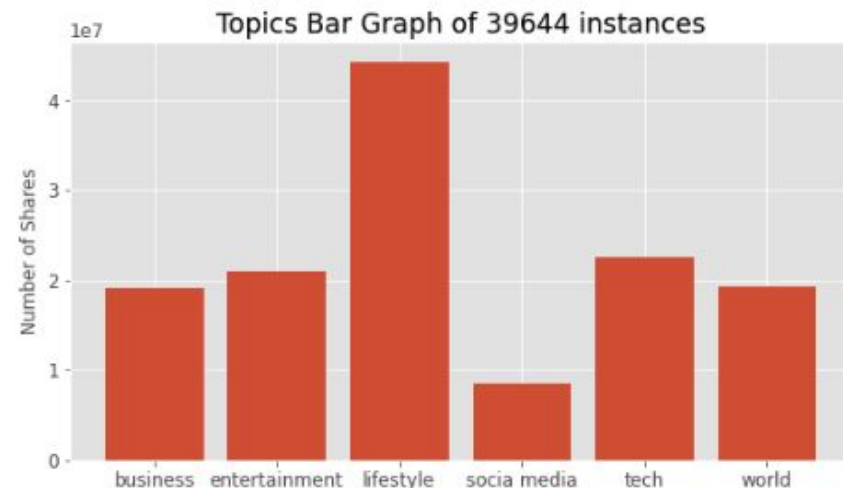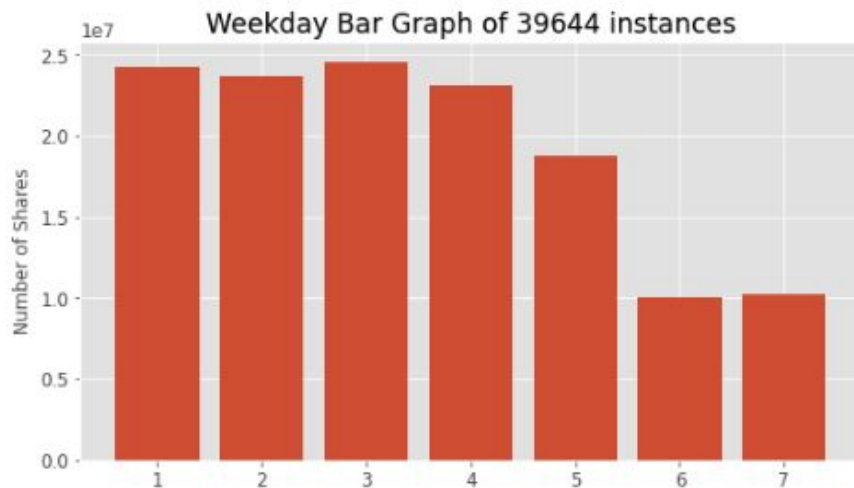
- average number of keywords
- published on a weekend?
- minimum number of shares of Mashable links
- article category
- etc.

Mean Decrease Accuracy
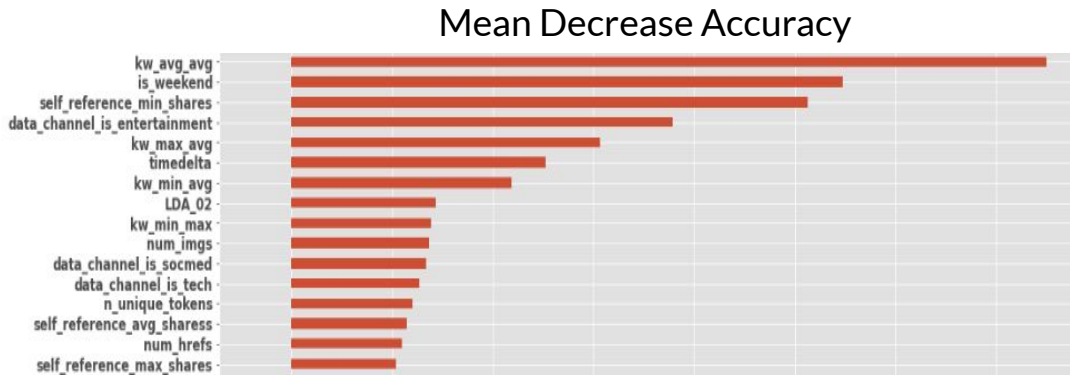
Exploratory Data Analysis

Exploratory Data Analysis

Weekday Bar Graph of 39644 instances

Topics Bar Graph of 39644 instances
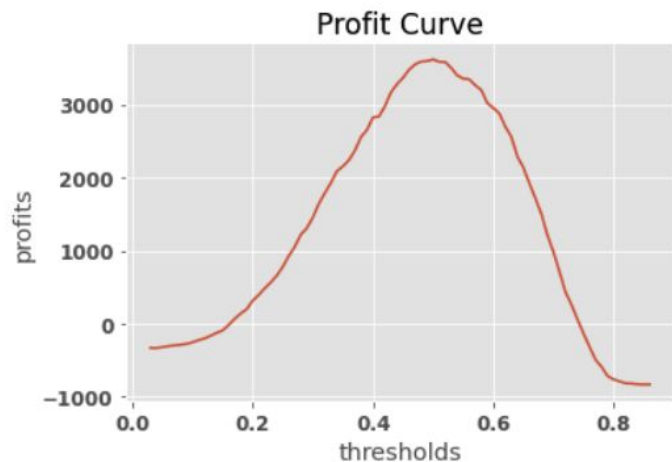
Multi-Classes Classification

Models trained: **Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier**

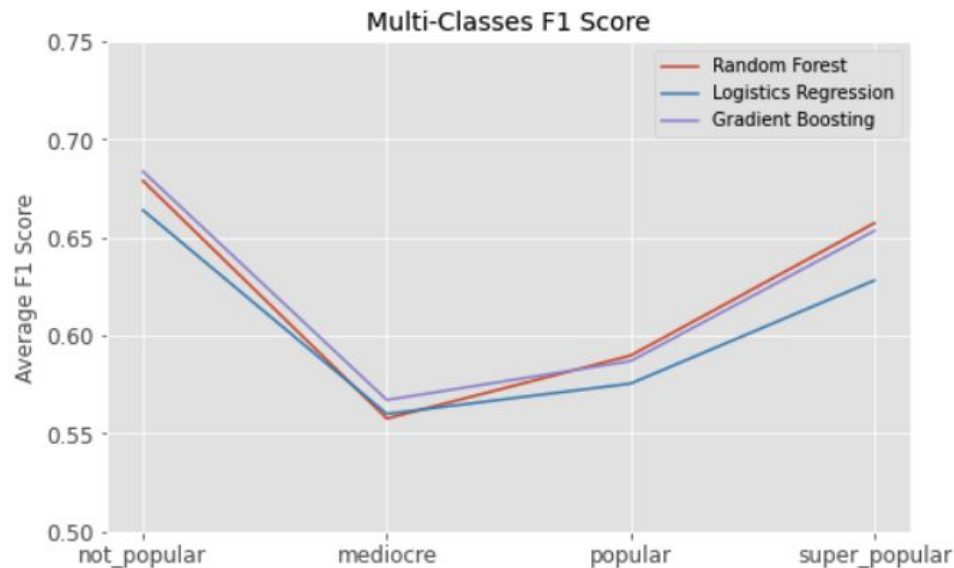Baseline model F1 score: 0.49
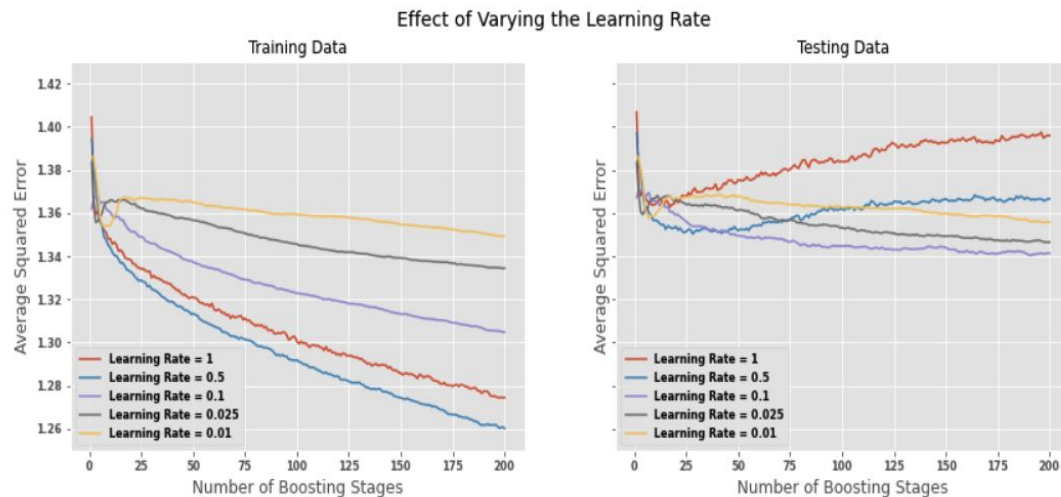
Model Performance

- Performance measure: **F1 score**

$$F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{1}{\frac{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}{2}}$$

- The models were better at predicting "not popular" and "super popular" categories with F1 score greater than 0.65

Business Insights

Recommendations to improve popularity:

- increase the embedded links to articles with high popularity
- increase amount of subjectivity in title
- increase number of positive/trending words in the content
- decrease number of longer words in the content

Next Steps

- web scraping 39,644 articles based on their URLs
- NLP analysis on article content
- identify trending words in popular articles

Any questions?

Appendix. Cost matrix and profit curve

Assumptions:

- a popular article will bring $5 in ads revenue in average
- a not popular article will bring $-2 in ads revenue
- it costs $3 to do improvement on not popular articles
- the opportunity cost of a popular article which predicted as not popular is $3

Optimal threshold: 0.5

predicted/actual	not popular	popular
not popular	-$2	-$5
popular	$2	$5

Profit Curve