Predicting Article Popularity

Siboney Cardoso • Adithya Murali • Andrew White • Ramzi Kattan • Carissa Ing

Agenda

01 About Our Data

02 EDA & Data Wrangling

03 Modeling

04 Solution & Insights

About Our Data

- Mashable articles posted from 2013-2014
- 58 predictive variables
  - Num links, images, videos
  - Avg len of words, num unique words
  - What day article was published
  - Sentiment and polarity of words
- ~40,000 observations
- Target: How many times each article was shared over social networks (aka popularity)

Business Problem

How to format newsletters/articles/blogs/ads to gain more consumer attention

Why It Matters

- Maximize use of limited company resources (especially for small businesses)
- Fulfill revenue goals through advertisements and sales

Our Team's Goal

- Build a model to predict popularity of Mashable articles
- Identify what makes an article popular to apply to small businesses

EDA & Wrangling

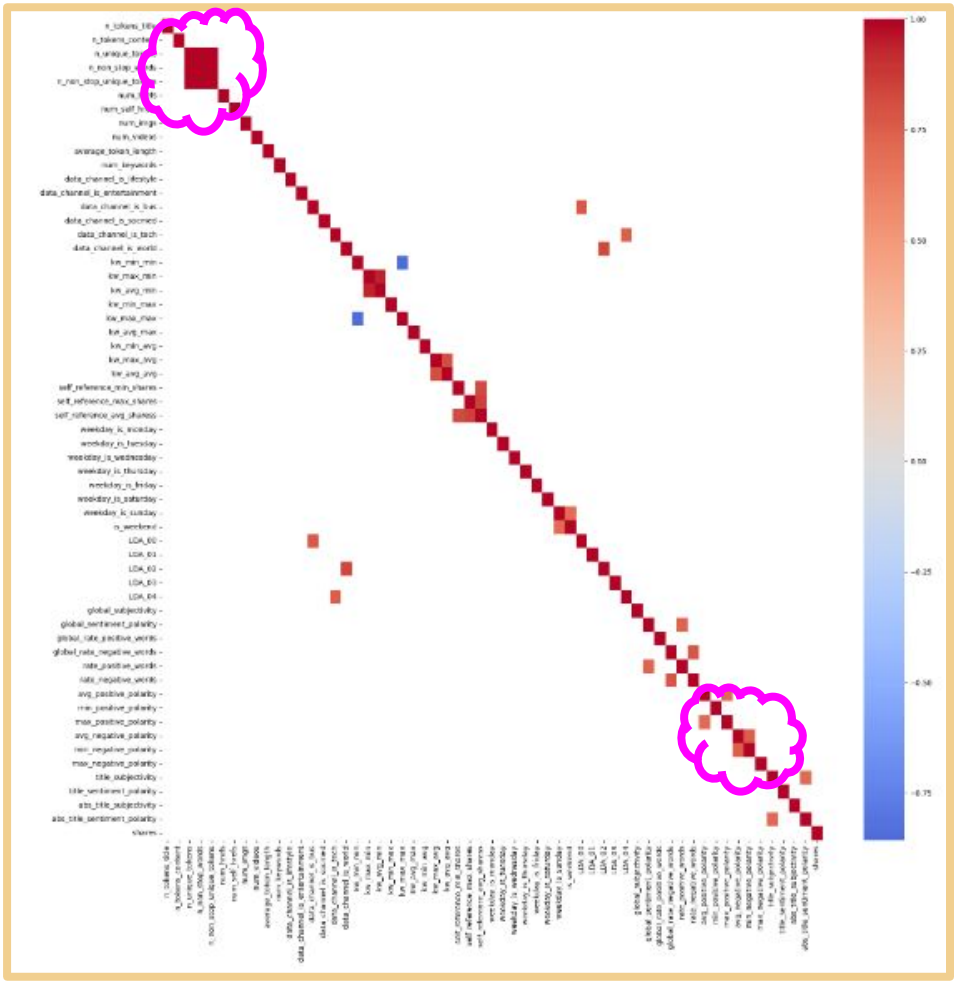
Checked for collinearity

- Correlation >= 0.7

Of correlated vars, drop features with less effect on target

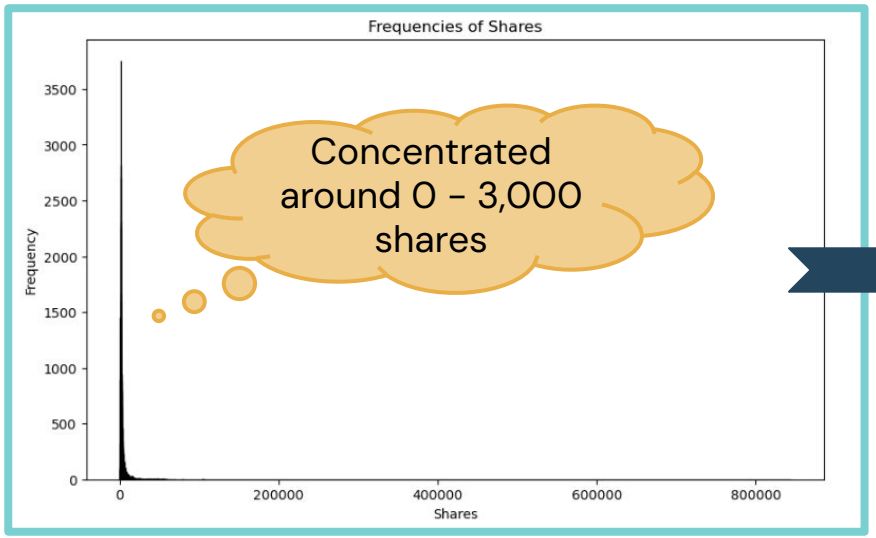
- n_non_stop_words
- n_non_stop_unique_tokens

- max_positive_polarity



EDA & Wrangling

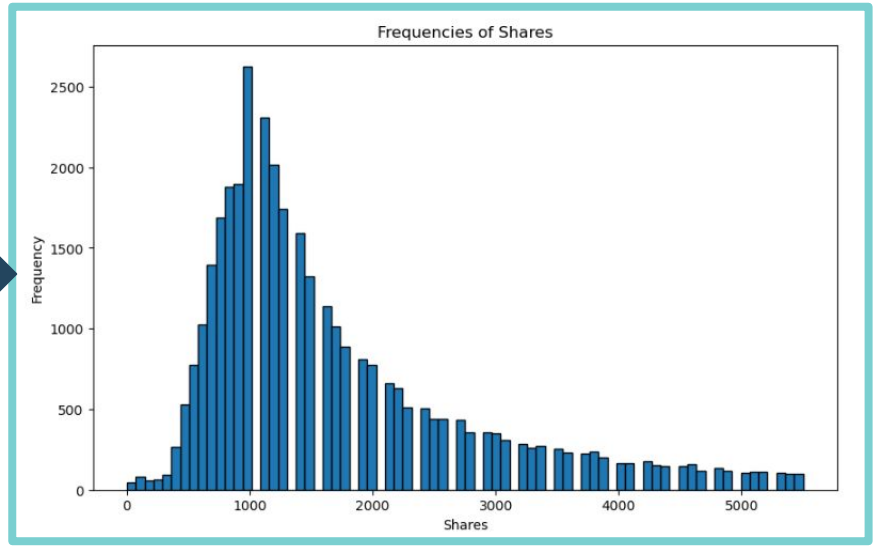
Checked distribution of target var



Removed outliers to better predict the avg article

Outliers = Shares > 3 std deviations from mean

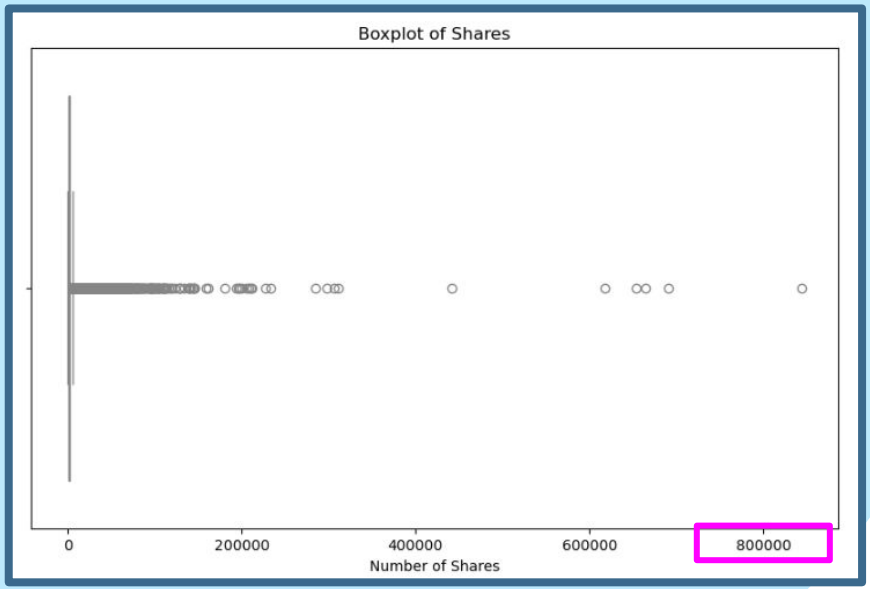
Note: Did not log(shares) bc logging is for ease of interpretability & does not increase predictability



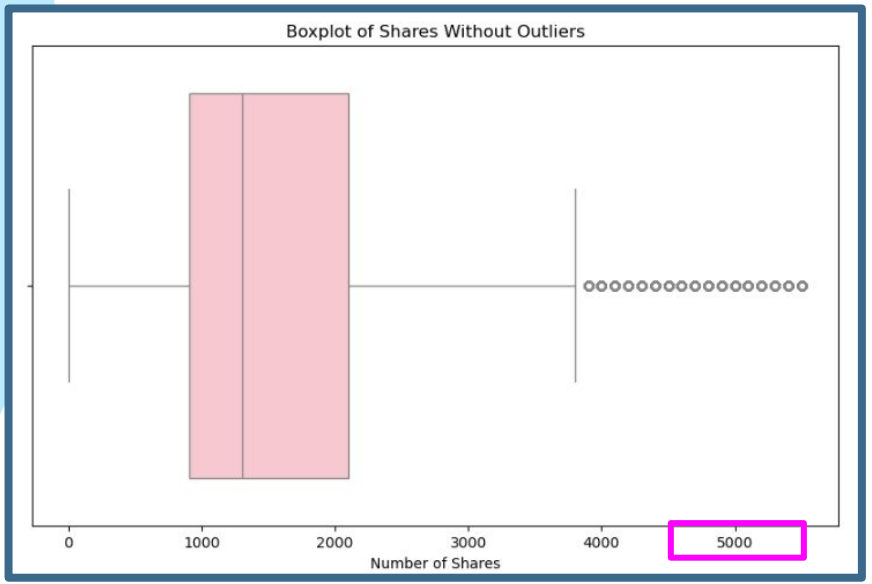
EDA & Wrangling

Another view of the *shares* distribution before and after removing outliers

Before



After

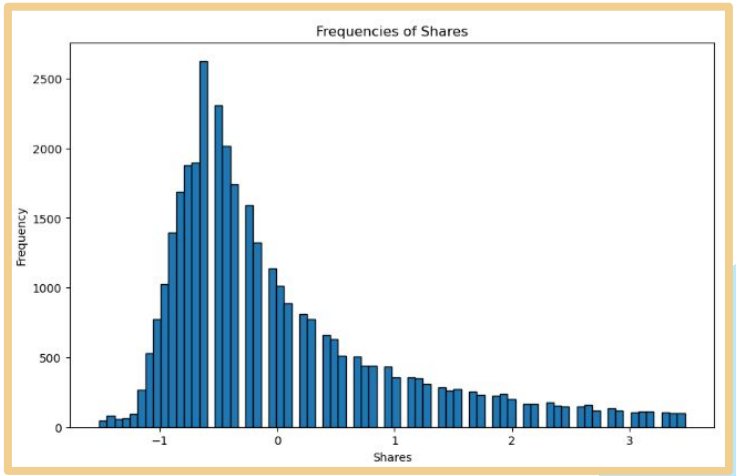


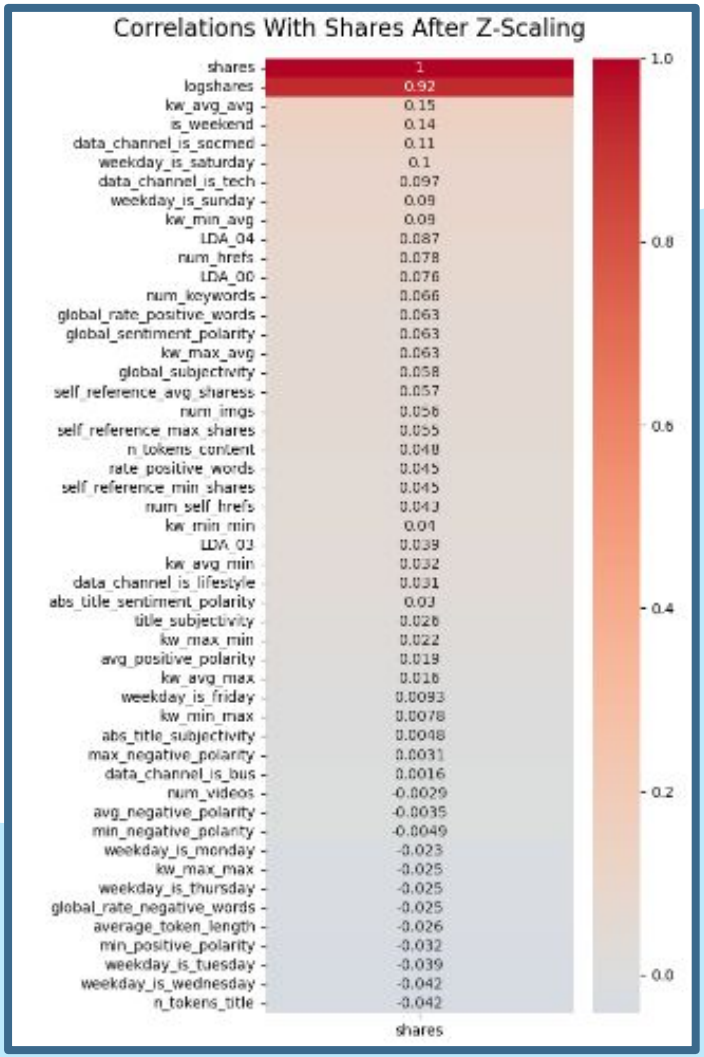
EDA & Wrangling

Have a lot of variables with different scales, i.e.:

- *number* of images
- *avg polarity* of positive words

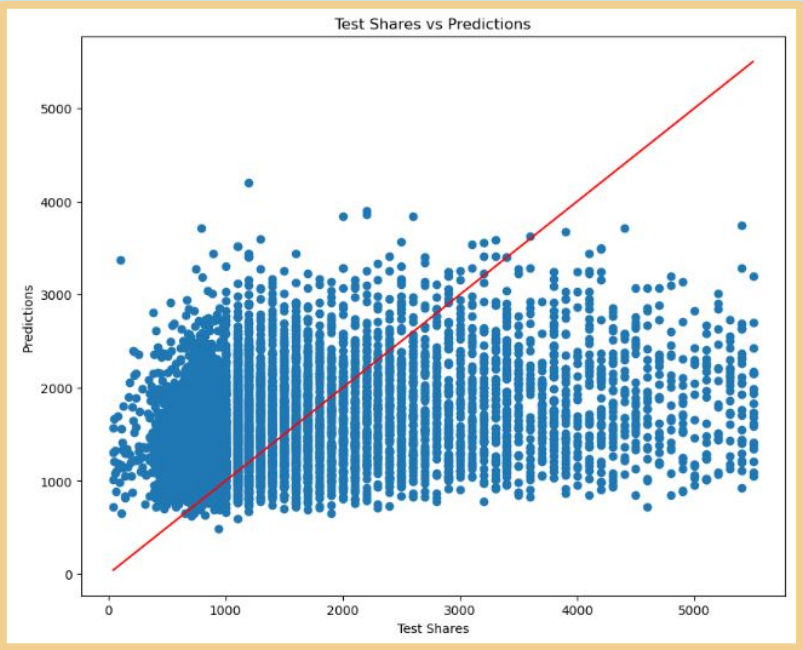
Standardized our variables by z-scoring

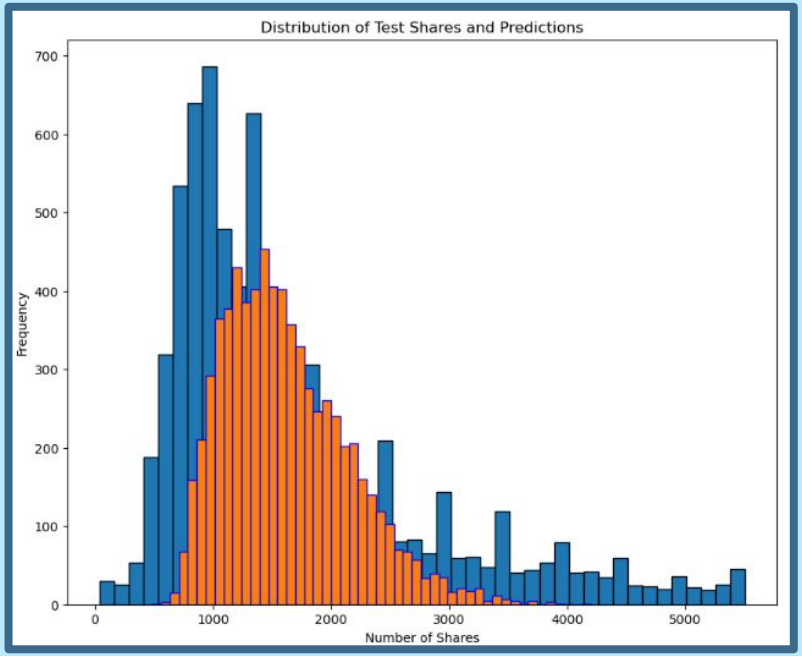




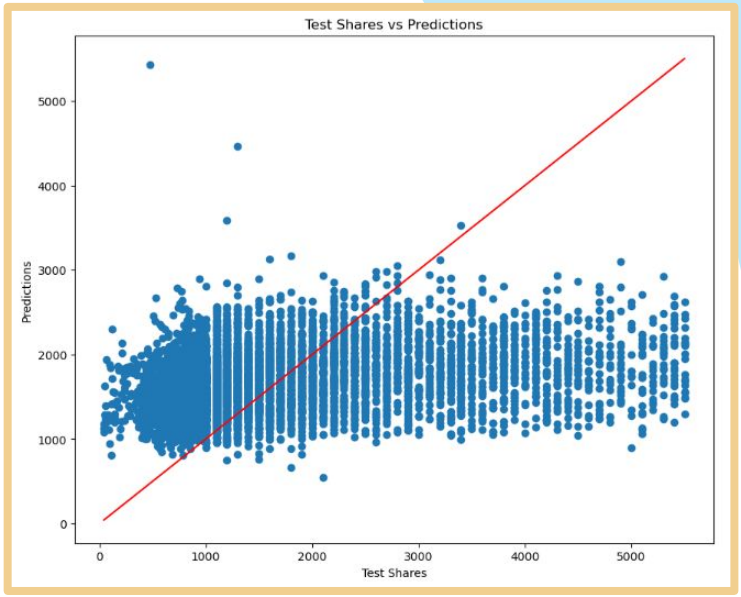
Model 1: KNN

Test RMSE	1,084.20
Normalized RMSE	19.86%
Coefficient of Variation RMSE	64.69%

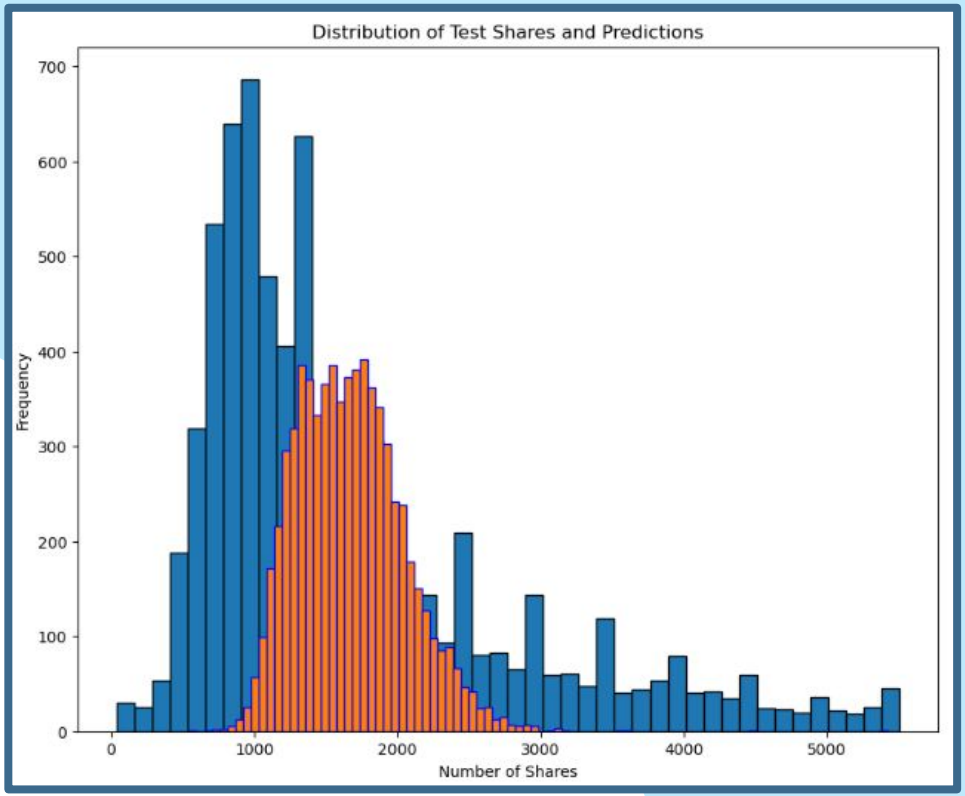




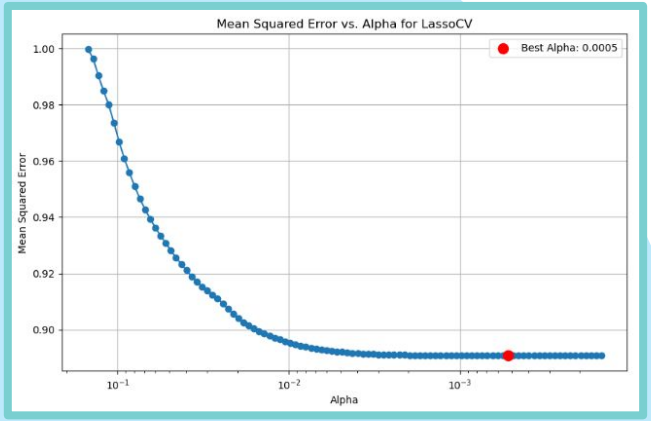
Model 2: Linear



Test RMSE	1,044.41
Normalized RMSE	19.14%
Coefficient of Variation RMSE	62.31%

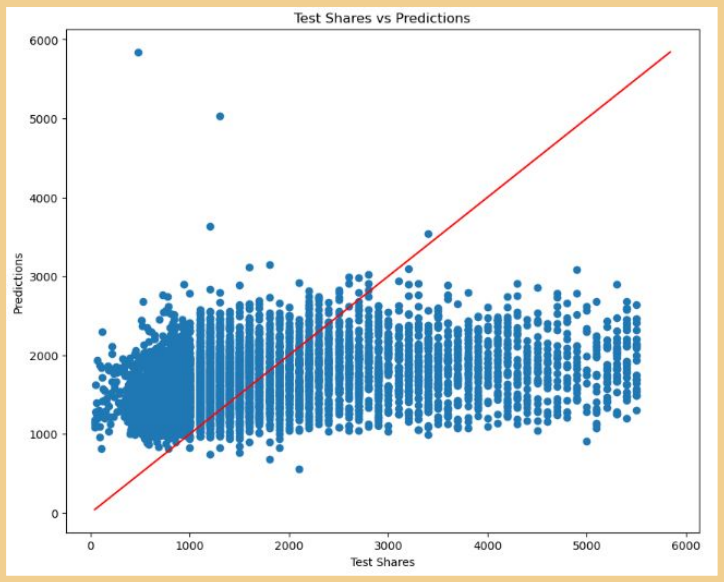


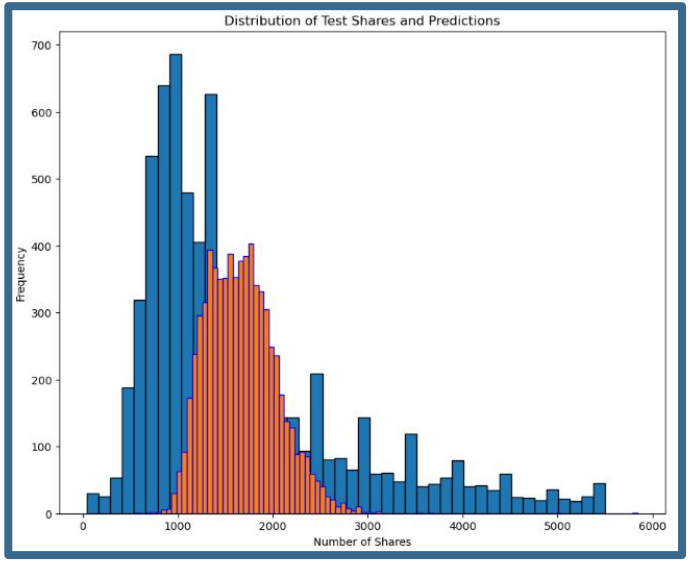
CV alpha:
~ 0.005



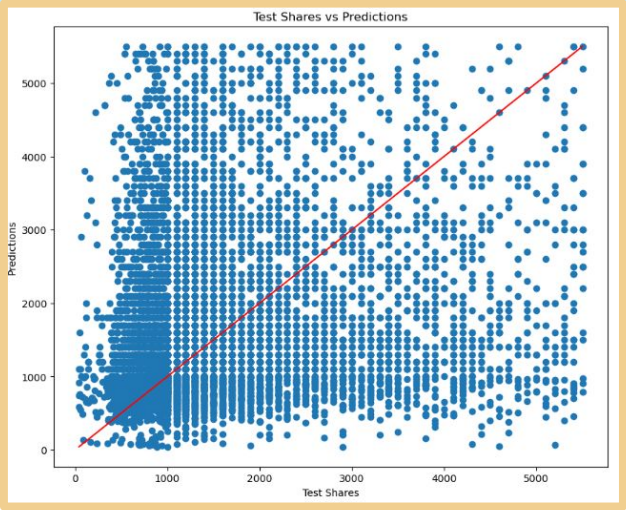
Model 3: Lasso

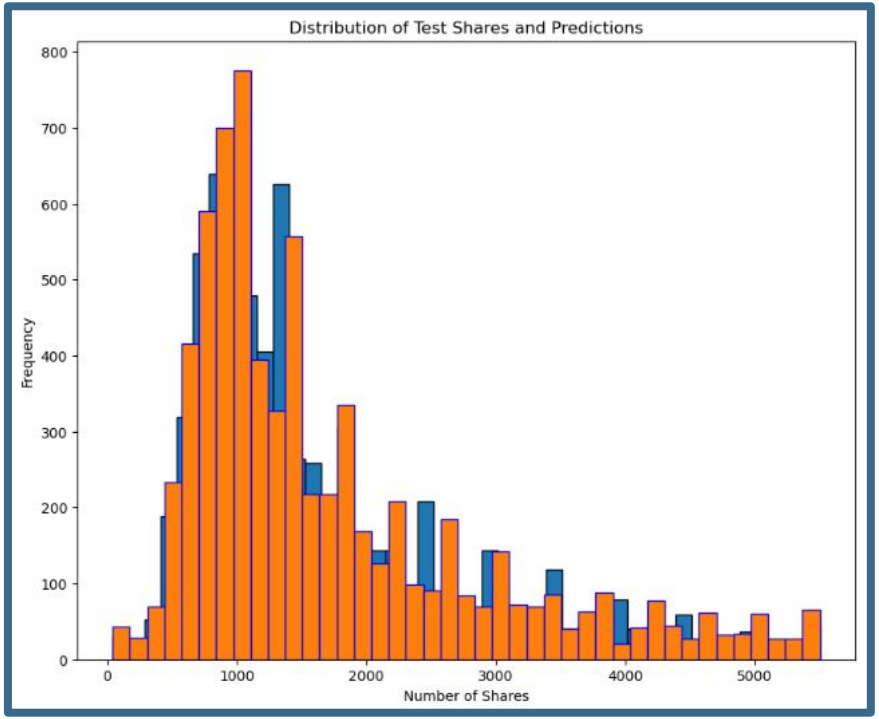
Test RMSE	1,044.66
Normalized RMSE	19.14%
Coefficient of Variation RMSE	62.33%





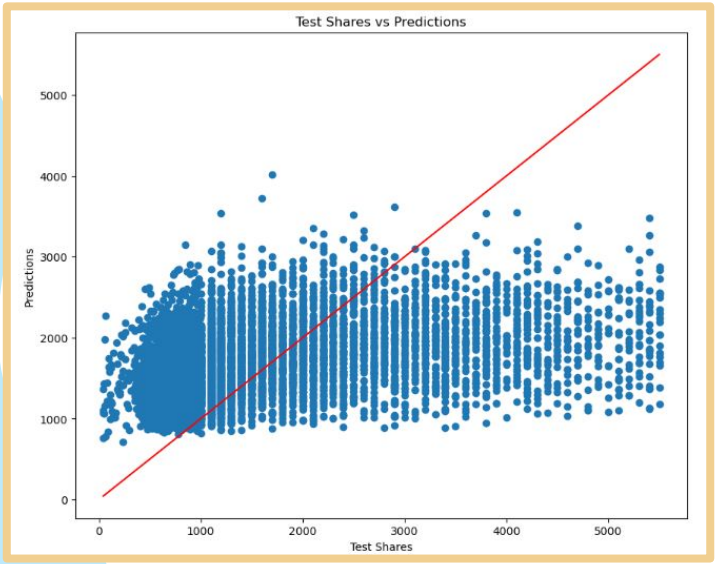
Model 4: Decision Tree

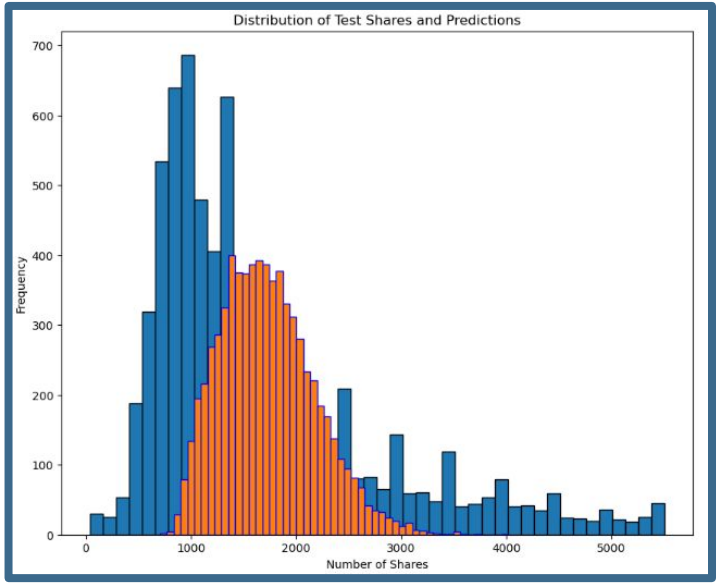




Test RMSE	1,495.22
Normalized RMSE	27.33%
Coefficient of Variation RMSE	89.21%

Model 5: Random Forest

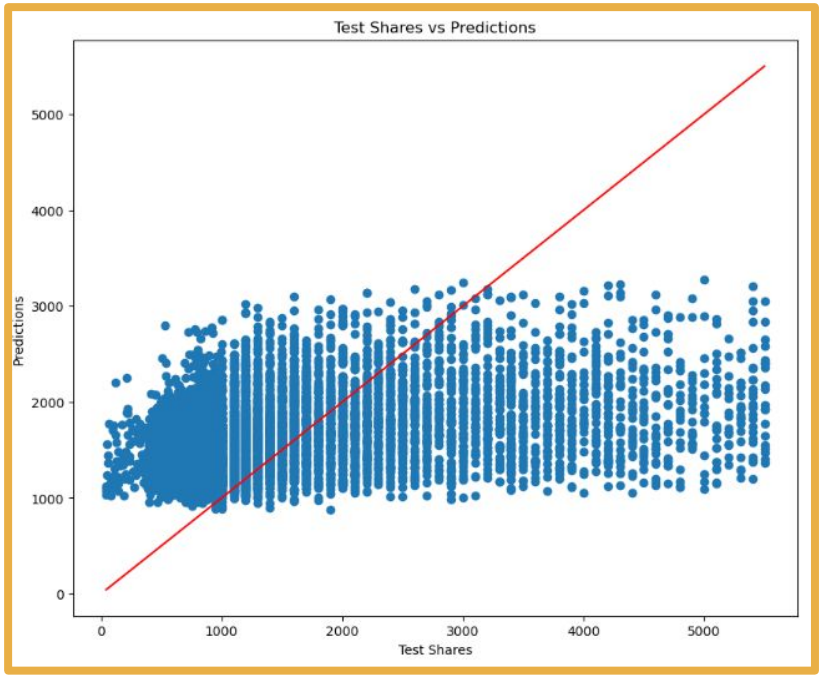


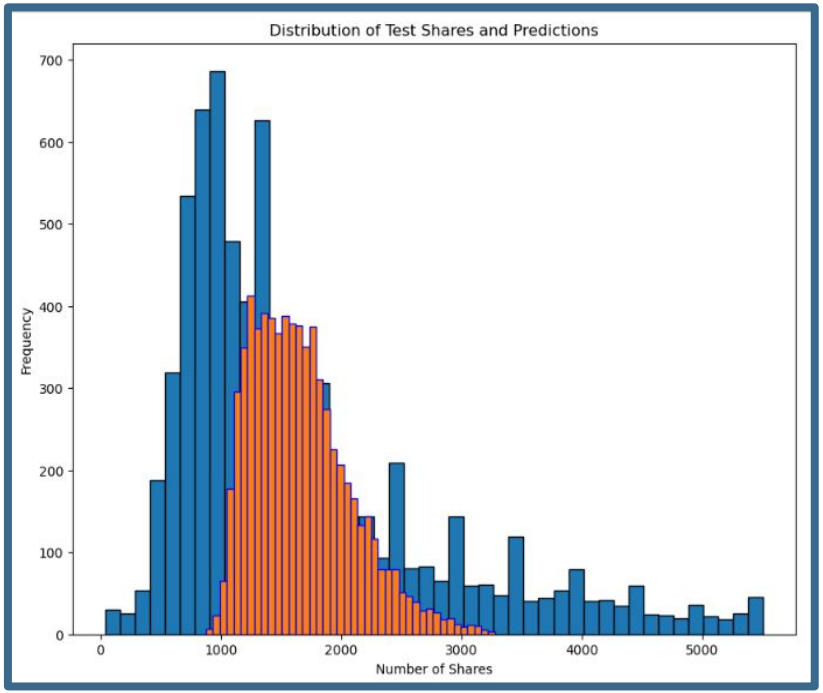


Test RMSE	1,035.02
Normalized RMSE	18.96%
Coefficient of Variation RMSE	61.75%

Model 6: Neural Network

Test RMSE	1,027.35
Normalized RMSE	18.82%
Coefficient of Variation RMSE	61.30%





Additional Model: Gaussian Naïve Bayes

Important Variables

```
data_channel_is_socmed: 0.016 +/- 0.002
weekday_is_saturday: 0.007 +/- 0.002
is_weekend: 0.006 +/- 0.002
weekday_is_sunday: 0.006 +/- 0.002
kw_avg_avg: 0.005 +/- 0.001
data_channel_is_tech: 0.005 +/- 0.002
LDA_04: 0.004 +/- 0.002
data_channel_is_lifestyle: 0.003 +/- 0.001
num_hrefs: 0.003 +/- 0.001
kw_min_avg: 0.003 +/- 0.001
```

Accuracy	42.94%
Avg Cross-Validated Accuracy	42.46%
Calibrated Accuracy	47.76%

Not very good so we disregarded

Solution & Insights

	Model	RMSE	NRMSE	CVRMSE
0	KNN	1084.203492	19.864483	61.296118
1	Ordinary Least Squares	1044.410331	19.135404	62.313746
2	Lasso Linear Regression	1044.655823	19.139901	62.328393
3	Decision Tree Classifier	1491.535674	27.327513	88.991053
4	Random Forest	1035.024315	18.963436	61.753738
5	Neural Network	1027.354370	18.822909	61.296118

Neural Network
performed best

BUT

Random Forest is
more interpretable
(for variable importance)

Difference of ~8 shares is
negligible

RF Top 5 Important Predictors

kw_avg_avg	The average performance of the keywords in terms of shares.	Keyword Performance
kw_avg_max	The average performance of top-performing keywords.	
kw_max_avg	The best possible average performance among all keyword.	
LDA_00	The strength of an article with particular topics.	Topic Relevance
n_unique_tokens	The richness of vocabulary in the articles.	Vocab Diversity

Conclusion: By focusing on these key predicting metrics, we can better predict and produce articles that not only attract readers but also compel them to share the content, increasing advertisement and sales revenue.

Future Steps

- Combine more correlated variables
- Consider popularity of articles from wider range of publishers
- Extend the search to social media posts to determine what types of posts from other businesses gain the most traction

Thank you!

Questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.