Predicting Article Shares Within The Digital Marketing Industry

Problem Statement + Why?

- **Problem**: Content creation is expensive and time consuming.
- **Why**: Article data is segmented and hasn’t been aggregated for analysing the performance of individual articles:
  - Platforms: (backlink providers, content analytics platforms).
  - HTML: the web page itself (article).
  - Influencers: (social media posting + sharing).

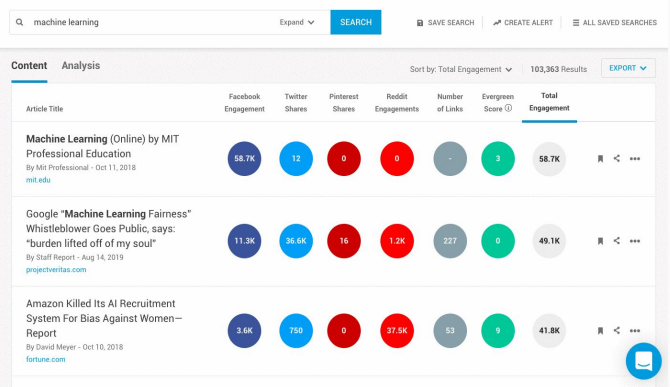
Goals + Success Metrics

- Primary goal: To predict the **number of shares an article will earn** after being published for a minimum of 1 year on the internet.
- Success Metrics:
  - To identify **3 - 5 core components** that marketers can leverage to improve the shareability of their articles.
  - To improve the score of a range of machine learning models by combining 3 unique datasets with the URL as a common key.

Three Data Sources

BuzzSumo + Web Scraping + Page Speed Data

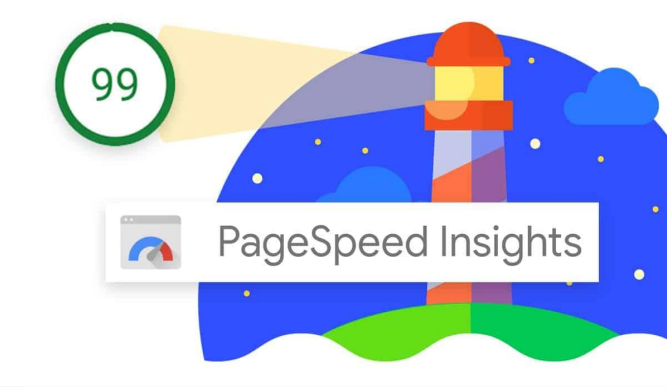
1.



2.

```
class MultiThreadScraper:


    def __init__(self, links, data_dict):
        self.pool = ThreadPoolExecutor(max_workers=20)
        self.scraped_pages = set([])
        self.to_crawl = Queue()
        self.start_url = 'https://gatheringdreams.com/affiliate-marketing-for-dummies/'
        self.to_crawl.put(self.start_url)
        self.links = ['https://www.entrepreneur.com/article/319017',
                      'https://maybethisway.com/blogging-tips/intro-affiliate-marketing/']

        self.data_dict = data_dict


    def technical_page_metrics(self, req):
        #Page_Size_In_Bytes
        page_size_in_bytes = len(req.content)

        text = fulltext(req.text)

        #Plain_text_size
        plain_text_size = len(text)

        #plain_text_rate --> plaintext rate value (plain_text_size / page_size)
        plain_text_rate = (plain_text_size / page_size_in_bytes) * 100

        #Encoding
        encoding = req.encoding

        #Detecting SSL Encryption
        if 's' in req.url:
```

3.



Buzzsumo Pro

Top Content | Influencers | Content Alerts | Reports | API | Resources | Email

Filter by Type:

- [x] Bloggers
- [x] Influencers
- [x] Companies
- [x] Journalists
- [x] Regular People

Uncheck All

- [ ] Ignore Broadcasters

Location:

E.g city or country

Filter

Reset Filters

"marketing technology"

Search! | Export

Enter a topic or username: @buzzsumo, big data. Advanced Search Options | All Links Shared | Save Search

Sort by: Number of

	PAGE AUTHORITY	DOMAIN AUTHORITY	FOLLOWERS	RETWEET RATIO
Rod Banner @rodbanner banner.net After years marketing technology; now up to my eyes in the 'technofication' of marketing. Geek, tech investor, fan of techno, humanity and overfunning.	22	10	223,786	13%
Travis Wright @teedubya linkedin.com/in/teedubya Venture Catalyst, Speaker, CMTO, StandUp Comic, Marketing Technology Entrepreneur, Data & Analytics Geek, Startup Growth Hacker. Business Journalist. Smart Ass. Influencer blogger	51	100	148,581	7%
Zain @zainb Triple agent of Marketing, Technology and Art شاعر (بحالي)!، من أبياتي: سيفتحُ الله باباً كنت تحسبهُ.. من شدة اليأس لم يخلق بمفتاح Influencer	0	0	147,689	31%
Cain Ransbottyn @ransbottyn about.me/ransbottyn A Chief Digital Officer with a broad technical background intersecting corporate #strategy, #marketing, #technology & #innovation. Fan of #GrowthHacking! Geek?	1	91	86,437	23%
Christopher Penn @cspenn cspenn.com/w VP Marketing Technology @shiftcomm, ninja, PodCamp cofounder, Marketing Over Coffee cohost, speaker, bestselling author, Buddhist. More: http://t.co/AMOsoidtNO	49	57	77,215	2%

Follow | View Links Shared

Overall Approach

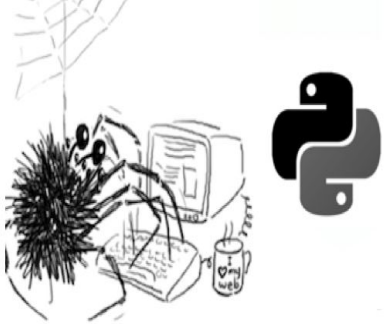
1. Data Collection - BuzzSumo Topic Data

17 topics were included:

- Affiliate Marketing, Content Marketing, Copywriting, Display Advertising, Email Marketing, Growth Hacking, Influencer Marketing, Link Building, Marketing Automation, Podcast Marketing, Search Engine Marketing, Social Media Marketing, Video Marketing, Website Design.

Overall Approach

1. Data Collection - Web Scraping



```
master_dict = {
'HTML_Content': [],
'Full_Text': [],
'Url': [],


#### Article Information
'Authors': [],
'Publish_Date': [],
'Article_Text': [],
'Article_Text_Length': [],
'Has_Top_Image': [],
'Number_of_Movies': [],
'Article_Is_Media_News': [],
'Number_Of_Images':[],
'Is_Valid_Body': [],

#### NLP Features
'Setences_Text': [],
'Number_Of_Sentences': [],
'Lexicon_Count': [],
```

```
#### Readability Scores
'Flesch_Reading_Ease_formula': [],
'Flesch_Kincaid_Grade_Level': [],
'FOG_Scale': [],
'SMOG_Index': [],
'ARI_Index': [],


#### Meta_Data
'Title_Text': [],
'Title_Tag_Length' : [],
'Meta_Description' : [],
'Meta_Description_Length': [],
```

```
#### Extract_page_features
'Body_Content_Links': [],
'Number_Of_Links': [],
'Links_To_Text_Ratio': [],


#### Technical Page Metrics
# 'Page_Size_In_Bytes': [],
# 'Plain_Text_Size': [],
# 'Plain_Text_Rate': [],
'Encoding': [],
'SSL': []
}
```

Overall Approach

1. Data Collection - PageSpeed Insights

https://searchengineland.com/

ANALYZE

MOBILE DESKTOP

75

https://searchengineland.com/

The speed score is based on the lab data analyzed by Lighthouse.

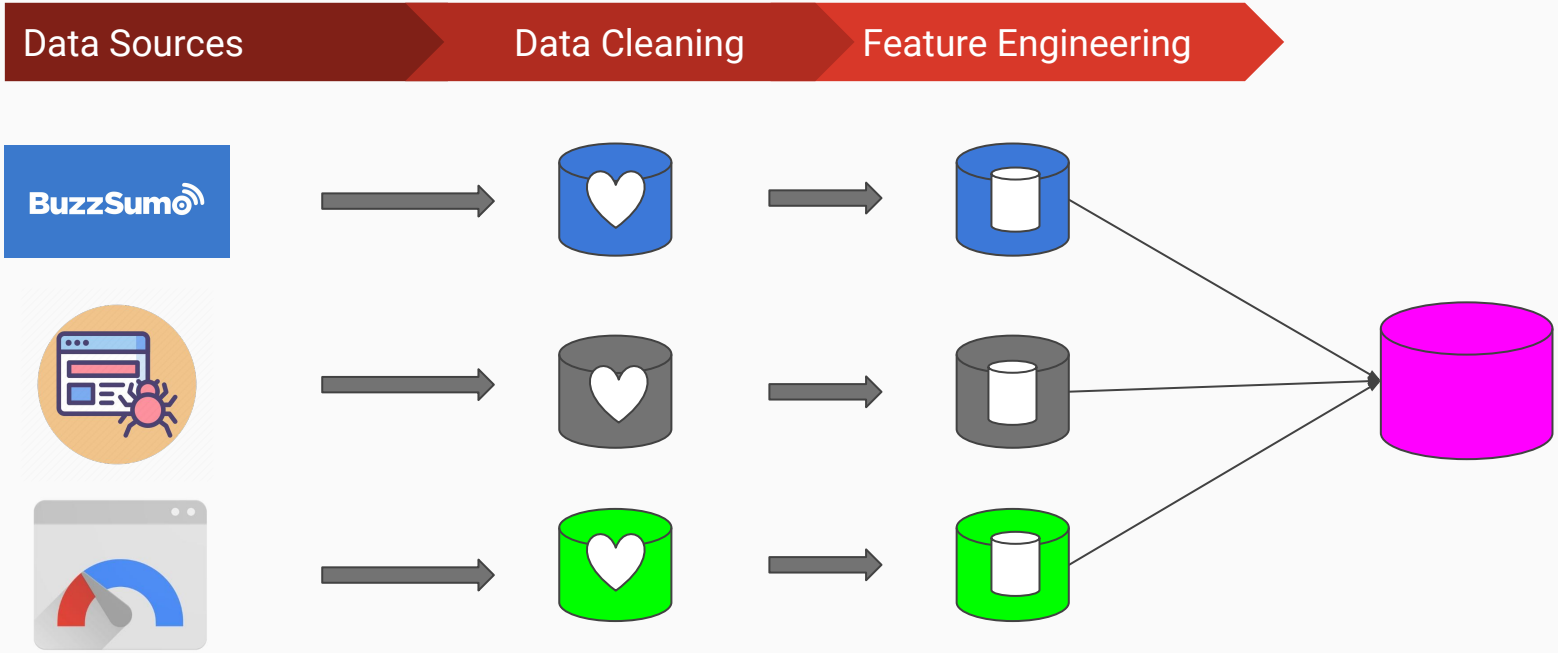
Analysis time: 11/12/2018, 8:42:11 AM

Scale: 90-100 (fast) 50-89 (average) 0-49 (slow)

Field Data

Overall Approach

2. Data Cleaning Architecture



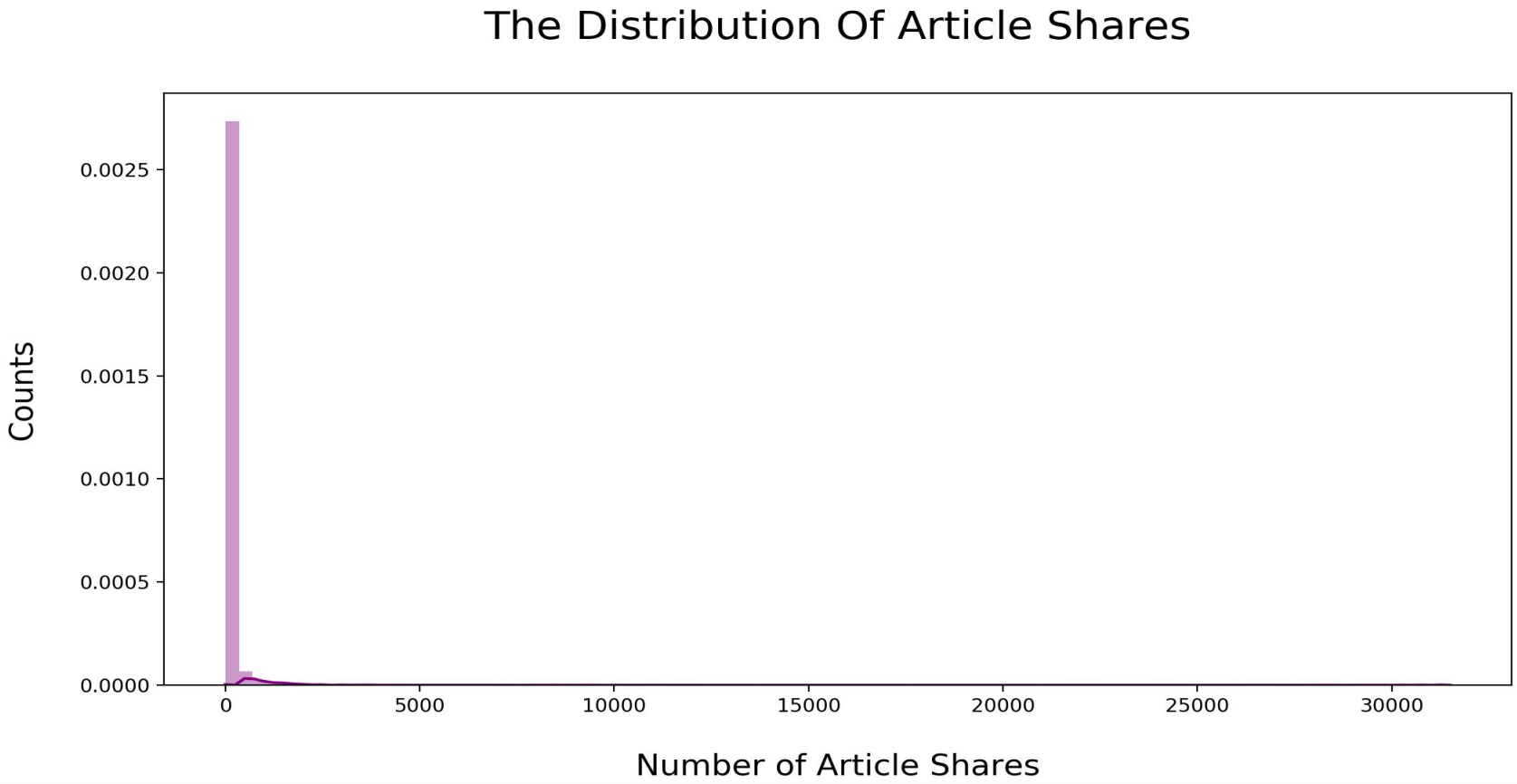
Overall Approach

	Url	Evergreen_Score	Total_Shares	Published_Date	Word_Count	num_linking_domains	Article_Text	Article_Text_Length
0	https://gatheringdreams.com/affiliate-marketin...	1.54	8021	2018-08-23	4767	1.0	Some of the links below are affiliate links, s...	27301
1	https://itsclaudiag.com/2018/09/how-to-use-aff...	1.44	2569	2018-09-16	1181	2.0	Would you like to make money while you sleep?\...	6519
2	https://www.entrepreneur.com/article/319017	5.68	844	2018-09-12	996	12.0	Learn three simple strategies to help you stac...	5916

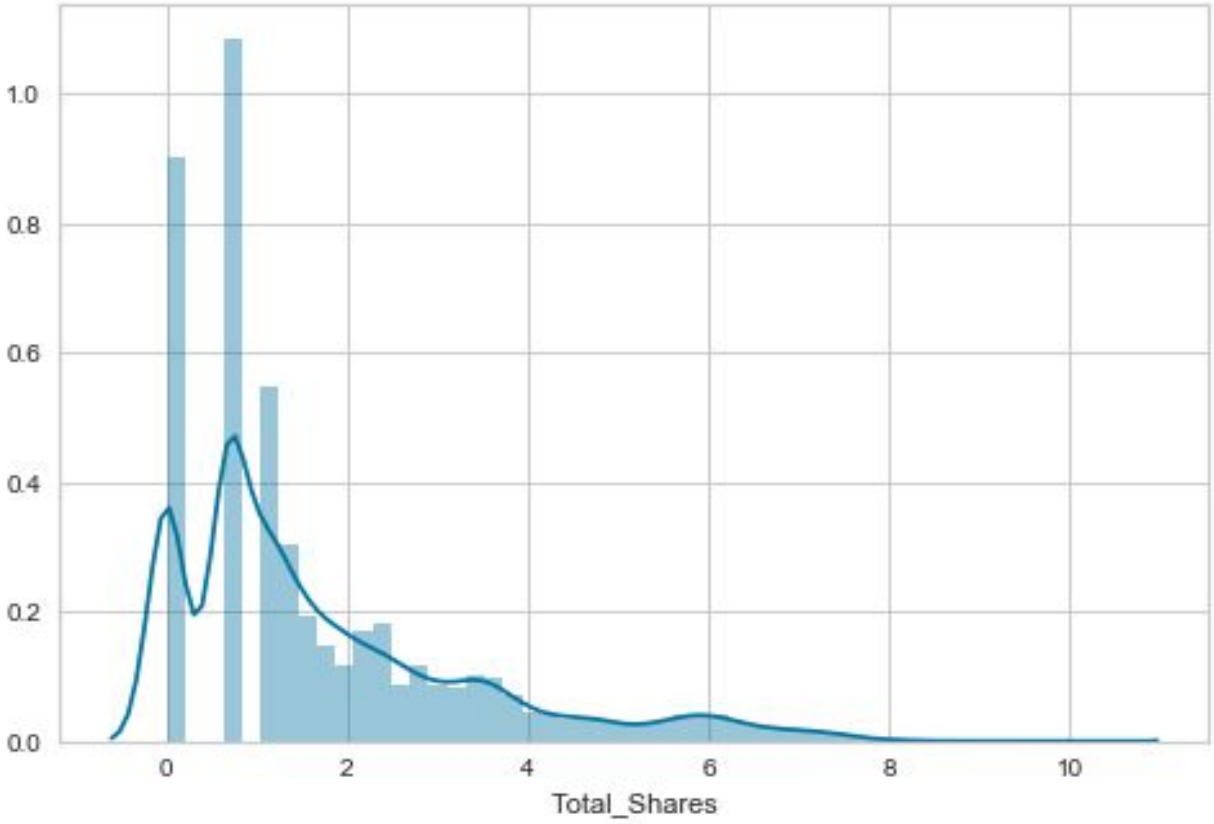
12	http://editor.ne16.com/vo/	0.7	0.7
13	http://rayhigdon.libsyn.com/how-to-build-your-...	0.7	0.7
14	http://rayhigdon.libsyn.com/social-media-in-yo...	0.7	0.7
15	http://rayhigdon.libsyn.com/two-methods-of-soc...	0.7	0.7
16	http://pages.rediff.com/responsive-website-des...	0.7	0.7
17	http://downarchive.org/e-books/video-training/...	0.7	0.7
18	http://inspiredconversation.libsyn.com/412-tyl...	0.7	0.7

0	2018-08-30	91

EDA


The Distribution Of Article Shares
Counts
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
5000
10000
15000
20000
25000
30000
Number of Article Shares

EDA - Applying a Np.log + 1 Transformation

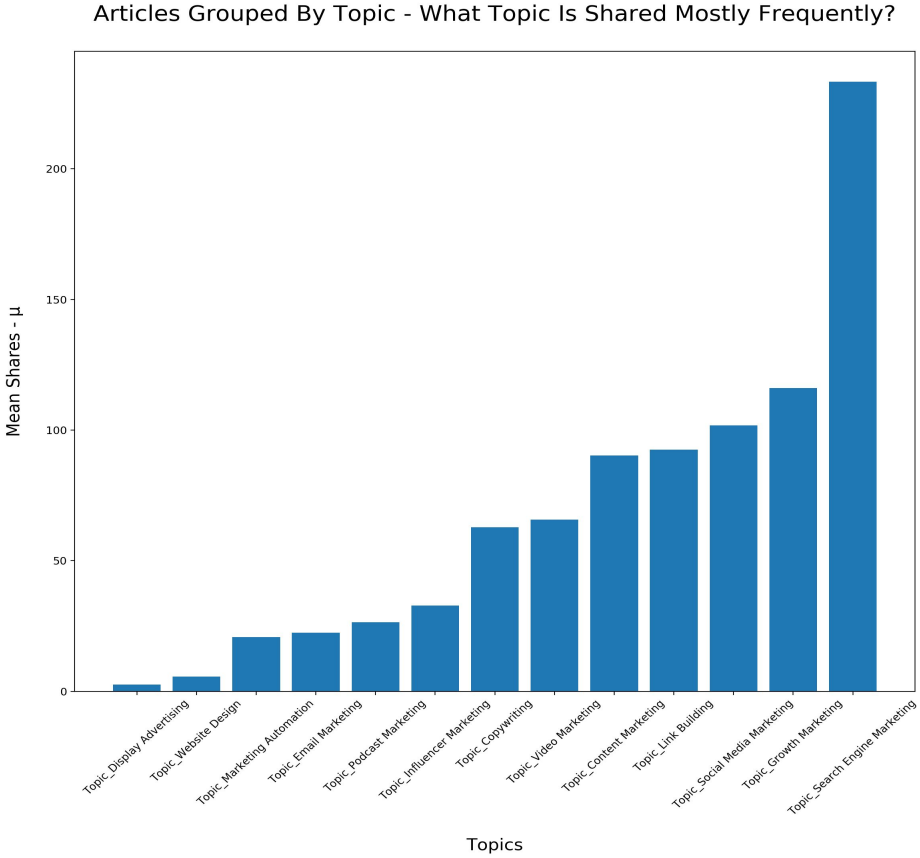


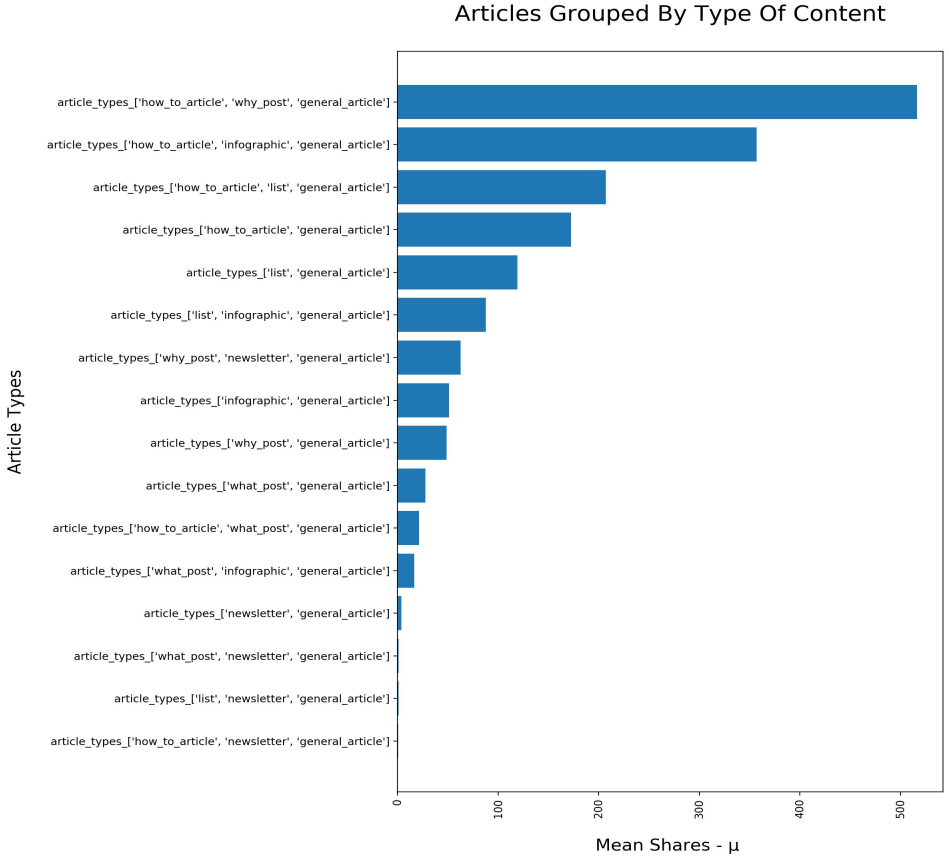
EDA

	Total_Shares
Evergreen_Score	0.529019
num_linking_domains	0.334792
Has_Article_Amplifiers	0.270627
Number_Of_Article_Amplifiers	0.240860
Has_Referring_Domains	0.240579
Number_Of_Sentences	0.115524
Lexicon_Count	0.114333
Plain_Text_Size	0.110962
Article_Text_Length	0.110816
Has_Author_Name	0.078322
Word_Count	0.075736
article_types_['how_to_article', 'general_article']	0.070345

	Total_Shares
Article_Is_Media_News	-0.006113
Topic_Display Advertising	-0.008335
article_types_['what_post', 'general_article']	-0.010957
article_types_['newsletter', 'general_article']	-0.012466
Encoding_ISO-8859-1	-0.017215
Topic_Influencer Marketing	-0.020869
Topic_Podcast Marketing	-0.021483
Topic_Marketing Automation	-0.028973
Topic_Email Marketing	-0.035100
Title_Tag_Length	-0.039640
Topic_Website Design	-0.042093
Encoding_utf-8	-0.050426

EDA


Articles Grouped By Topic - What Topic Is Shared Mostly Frequently?
Mean Shares - μ
0
50
100
150
200
Topic_Display Advertising
Topic_Website Design
Topic_Marketing Automation
Topic_Email Marketing
Topic_Podcast Marketing
Topic_Influencer Marketing
Topic_Copywriting
Topic_Video Marketing
Topic_Content Marketing
Topic_Link Building
Topic_Social Media Marketing
Topic_Growth Marketing
Topic_Search Engine Marketing
Topics


Articles Grouped By Type Of Content
Article Types
article_types_['how_to_article', 'why_post', 'general_article']
article_types_['how_to_article', 'infographic', 'general_article']
article_types_['how_to_article', 'list', 'general_article']
article_types_['how_to_article', 'general_article']
article_types_['list', 'general_article']
article_types_['list', 'infographic', 'general_article']
article_types_['why_post', 'newsletter', 'general_article']
article_types_['infographic', 'general_article']
article_types_['why_post', 'general_article']
article_types_['what_post', 'general_article']
article_types_['how_to_article', 'what_post', 'general_article']
article_types_['what_post', 'infographic', 'general_article']
article_types_['newsletter', 'general_article']
article_types_['what_post', 'newsletter', 'general_article']
article_types_['list', 'newsletter', 'general_article']
article_types_['how_to_article', 'newsletter', 'general_article']
0
100
200
300
400
500
Mean Shares - μ

Modeling

1. What Models Were Implemented?

- Linear Regression
- LassoCV, RidgeCV (modified versions of linear regression)
- Decision Trees, Random Forests
- Support Vector Machine

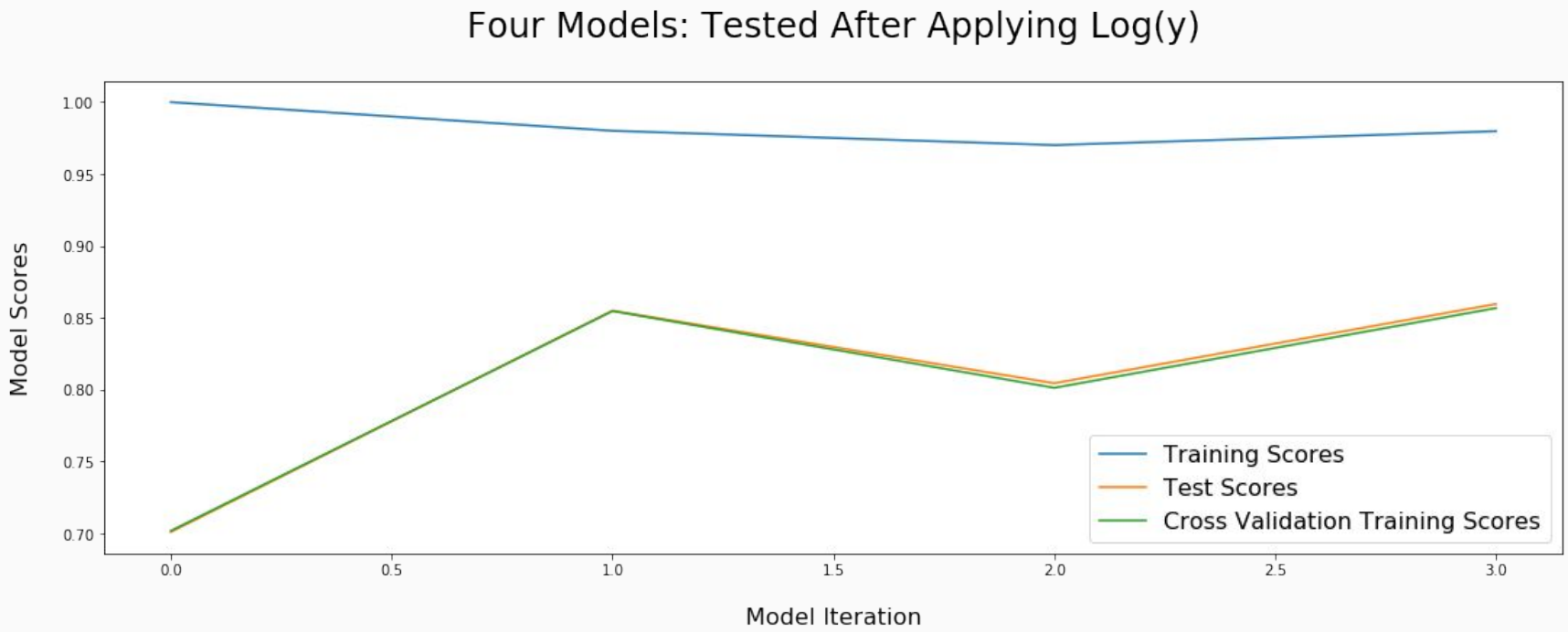
Modeling

2. Results Summary

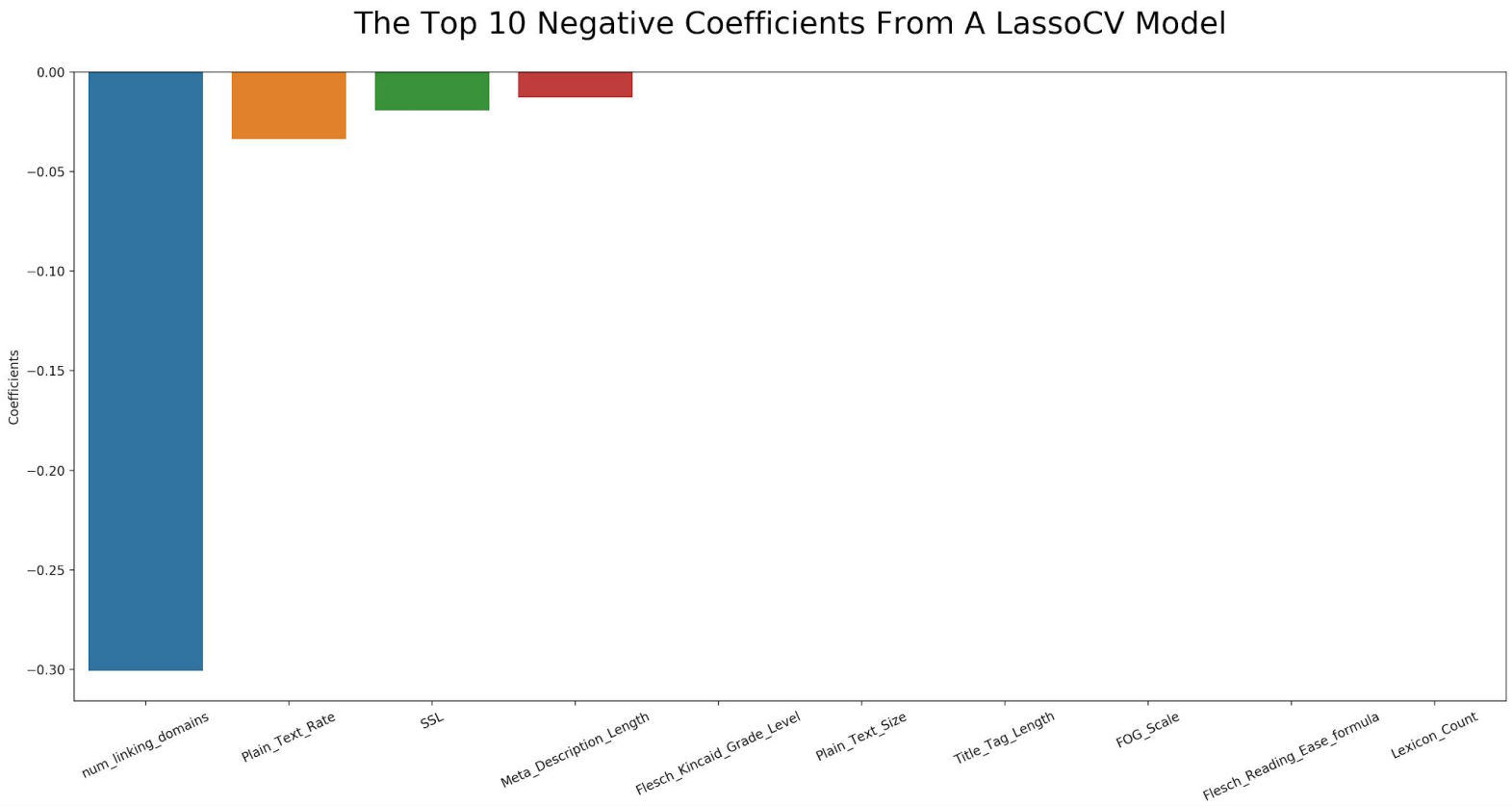
	Data_Used	Data_Type	Model_Name	Model_Training_Score	Model_Test_Score	Model_Cross_Val_Score
0	Numerical	Non-Logged Data	LinearRegression(copy_X=True, fit_intercept=Tr...	0.533955	0.155411	-3.677588e+19
1	Numerical	Non-Logged Data	RidgeCV(alphas=array([1.00000e-05, 1.26186e-05...	0.533842	0.155194	5.400575e-01
2	Numerical	Non-Logged Data	LassoCV(alphas=array([1.00000e-05, 1.26186e-05...	0.533555	0.155288	5.404661e-01
3	Numerical	Logged	DecisionTreeRegressor(criterion='mse', max_dep...	1.000000	0.701080	7.018761e-01
4	Numerical	Logged	RandomForestRegressor(bootstrap=True, criterio...	0.980080	0.854972	8.547309e-01
5	Numerical	Logged	AdaBoostRegressor(base_estimator=RandomForestR...	0.970121	0.804518	8.013276e-01
6	Numerical	Logged	AdaBoostRegressor(base_estimator=RandomForestR...	0.979795	0.859630	8.568030e-01

Modeling

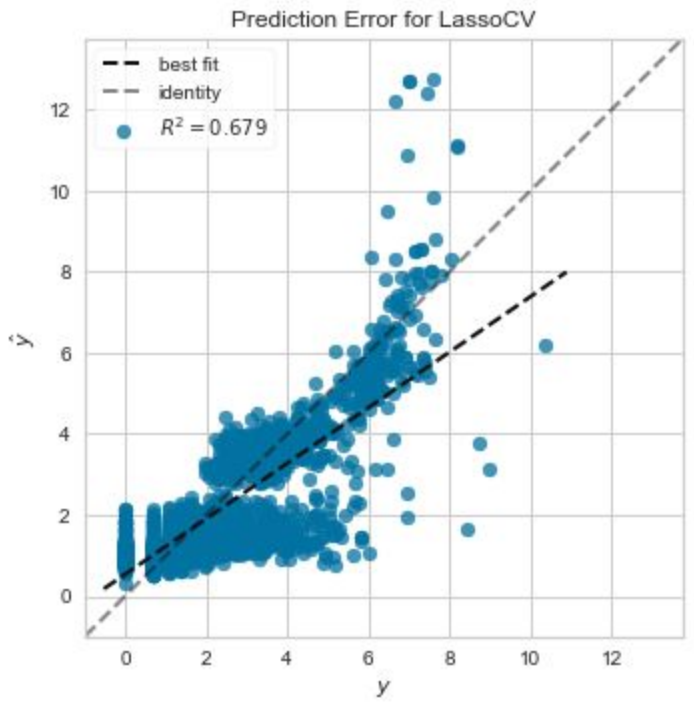
2. Results Summary

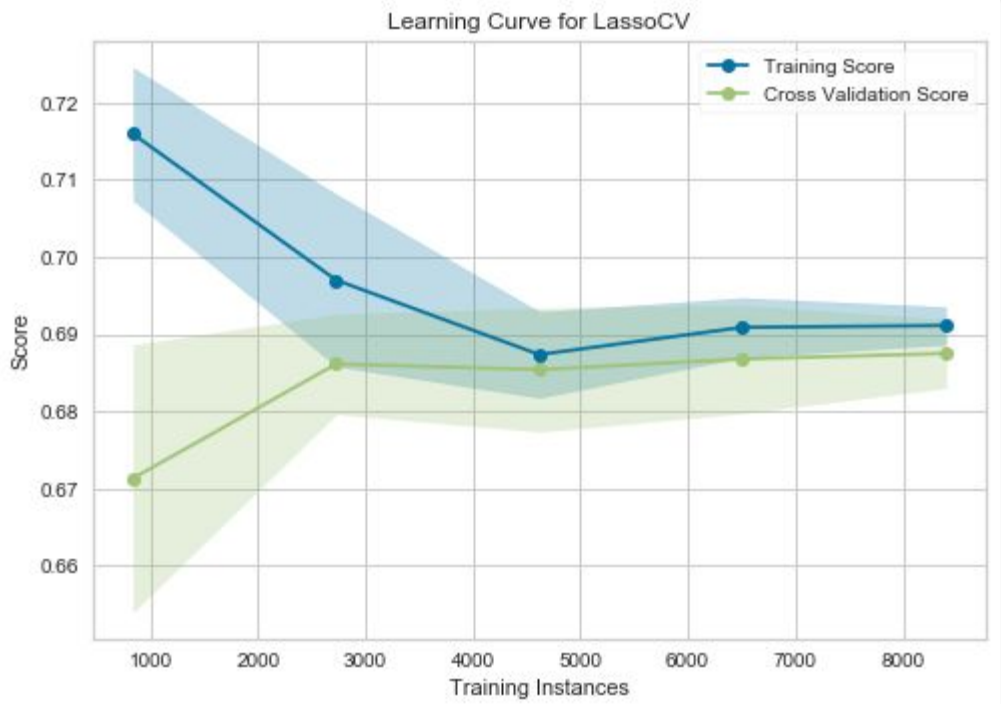


Modeling - LassoCV

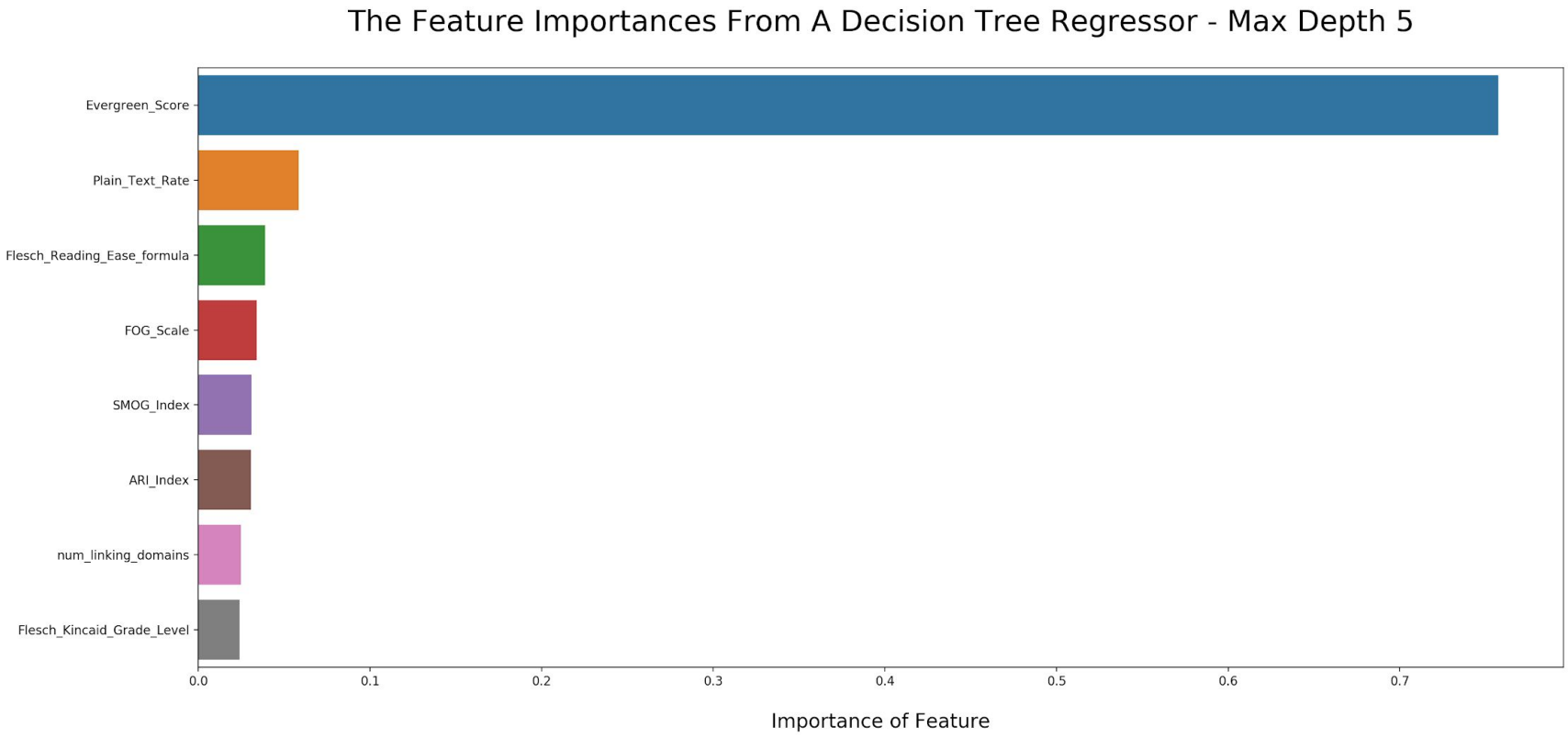


Modeling - LassoCV





Modeling - Decision Tree Regressor



Modeling - Decision Tree Regressor

Google

site:https://seo-gold.com/

Best Twitter Image Dimensions for Sharing is a 2:1 Ratio - SEO Gold
https://seo-gold.com/best-twitter-image-dimensions-for-sharing-is-a-2-1-ratio/
1 Sep 2017 - When sharing links on Twitter for best results use images with a 2:1 ratio with minimum dimensions 440px by 220px as the image shared. ... If for example I have an image with width 1800px for best Twitter results when sharing I want the height to be 900px (2:1 ratio). Twitter 2:1 Ratio ...

Optimized Images Load Faster and Consume Less Cellular Data ...
https://seo-gold.com/lighthouse-report-optimized-images-load-faster-and-consume-les...
22 Feb 2018 - How to completely wreck your WordPress sites SEO performance and speed metrics by using the Slider Revolution Responsive WordPress ...

... Indicates Truncated Title Tag, it's Too Long.

Foundem Electronics Comparison Shopping Links - SEO Gold
https://seo-gold.com/comparison.../foundem-electronics-comparison-shopping-links/
4 days ago - Foundem electronics comparison #shopping links are low on content, AKA thin affiliate content. #Google tends not to rank thin content.

Consider Modifying Truncated Title Tags.

SEO Anatomy of a Text Link - SEO Gold
https://seo-gold.com/search-engine-optimization-tips/seo-anatomy-of-a-text-link/
4 days ago - SEO Anatomy of a Text Link: Anchor Text = Very Important to Google. Title Attribute Text = Ignored by #Google.

Robotwity Collect From Followers - SEO Gold
https://seo-gold.com/robotwity-twitter-bulk.../robotwity-collect-from-followers/
24 Jan 2018 - Robotwity Collect From Followers @SEOGoldUK #Robotwity #Follow.

Content is King SEO Myth - SEO Gold

Risks / Limitations

- Sample Selection / Omission Bias.
- The Newspaper3k python library.
- The data snapshot consists of a two month segment where:
  - All of the articles are 1+ years old.
  - 15 topics were selected to cover the digital marketing industry publishing space.
- NLP (Investigating using Spacy)

Assumptions

- Is a one month sample, representative of the total article shares that would be found in the true population?
- Seasonality or consumer trends / times during the year have not been considered and have little impact on article shareability.

Next Steps

- Sentiment analysis.
- Scrape additional topics & monitor topics over time.
- RNN (LSTM) Neural Network implementation + Time Series Models
- To scrape additional link metrics from 3rd party providers including:
  - Ahrefs
  - SEMrush
  - Majestic

Conclusion

- Ada Boosted RandomForestRegressor was the best performing model.
- Taking the logarithm of the target variable (Total_‘Article_Shares’) helped to improve the model scores by creating a more linear relationship between our predictor features and ‘Total_Article_Shares’.

Key Takeaways

- Focus on producing more **evergreen content.**
- Prioritise creating **how to guides** over infographics and list posts.
- **Leverage relationships with key influencers** to increase the number of article shares.
- Focus on **long-form content** as this was a positive coefficient in the LassoCV model (higher number of sentences).