PARTNERSHIPS for STRATEGIC IMPACT

Caption Analysis Capstone Project

Ningxin Zhang
Huixin Yang
Shiqi Wang

Agenda

01

Background

-Motivations
-Goals

02

Research Questions

03

Data

-Data Collection & Cleaning
-Exploratory Data Analysis

04

Models

-Linear Regression
-Regression Trees
-Random Forest

05

Conclusion

-Recommendation
-Limitations

06

Reference

Project Background

01 Dominance of social media in this digital age; Interested in the power of captions on popular social media platforms

02 Help our client explore what features of a caption could increase interaction and engagement on social media posts

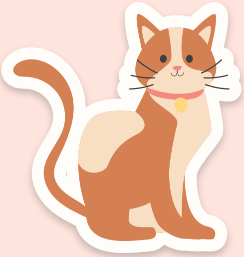
03 Chose to focus on three major social media platforms: **LinkedIn, Twitter and Facebook;** Found accounts similar to the profile of our client: small-scale, non-profit companies

04 Discarded the previous dataset and collected our own data using **web scraping**

Research Questions

As for companies similar to PSI, with the **control of number of followers**, what **factors of caption** would have the strongest association with the **engagement** of their social media posts on Facebook, Twitter and Linkedin?

DATA



Data Collection

- Python
- Scraped data with GitHub open source packages
- Post from accounts that PSI designated
- Up-to-date till Oct 19, 2022
- Initial information collected:
    - Name of the company
    - Text of the caption
    - Likes
    - Comments
    - Followers

LinkedIn

149 observations from **5** Companies

Twitter

105 observations from **8** Companies

Facebook

150 observations from **2** Companies

Data Cleaning

Special Characters

- Count the number of emojis, hashtags, and question marks in each post
- Packages used (Python): advertools, re

Word Count and Length

- Transform text into corpus
- Remove common stop words and special characters
- Calculate the number of words and average length of words in each post
- Packages used (Python): nltk, re

Sentiment

- Compute sentiment for each post
- Packages used (R): SentimentAnalysis

Variables of Interest

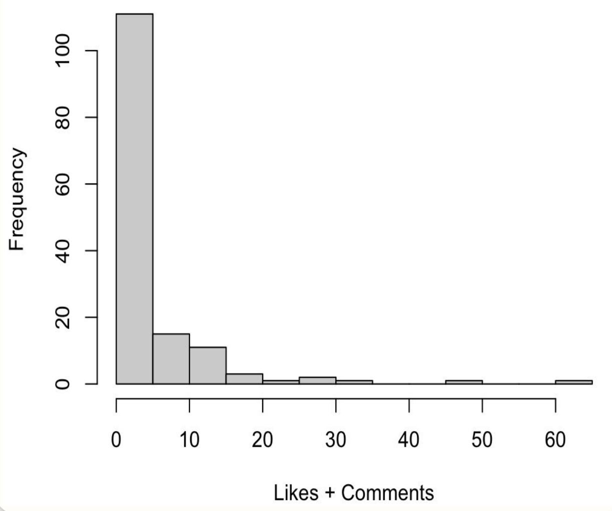
SentimentGI	The sentiment of the post text; Quantitative; Ranging from -1 to 1
Hashtag	The number of hashtags included in the post; Quantitative
Question mark	The number of question marks included in the post; Quantitative
Emoji	The number of emojis included in the post; Quantitative
Average length	The average length (number of letters) of the words in the post; Quantitative
Word count	The number of words in the post; Quantitative
c	The number of followers that the account has; Control variable, not of interest

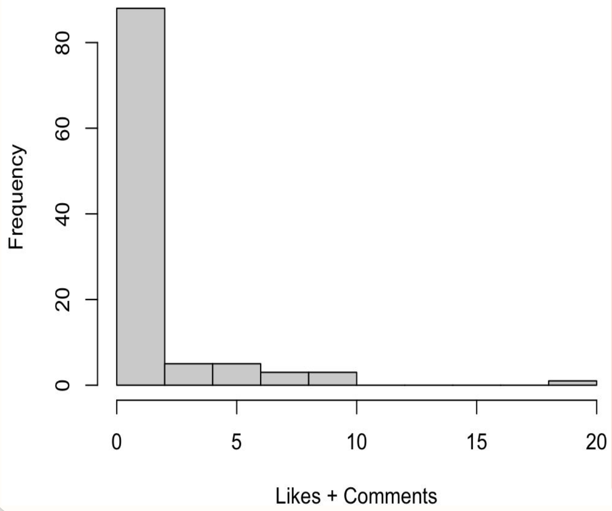
Response	The responses that a post receives; Comments + Likes; Quantitative

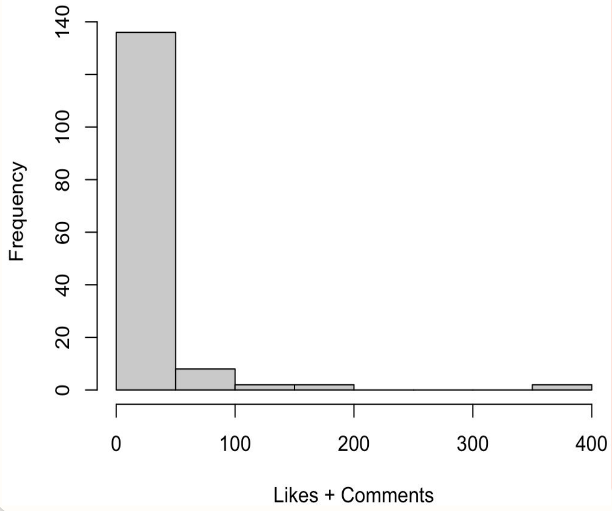
SentimentGI	The sentiment of the post text; Quantitative; Ranging from -1 to 1
Hashtag	The number of hashtags included in the post; Quantitative
Question mark	The number of question marks included in the post; Quantitative
Emoji	The number of emojis included in the post; Quantitative
Average length	The average length (number of letters) of the words in the post; Quantitative
Word count	The number of words in the post; Quantitative
Followers*	The number of followers that the account has; Control variable, not of interest
Response	The responses that a post receives; Comments + Likes; Quantitative

Response Variable Transformation

- Plotted histograms of response variable across all three platforms
- Highly right-skewed distribution, nearly 90% low values (includes 0)
- Need to perform log transformation

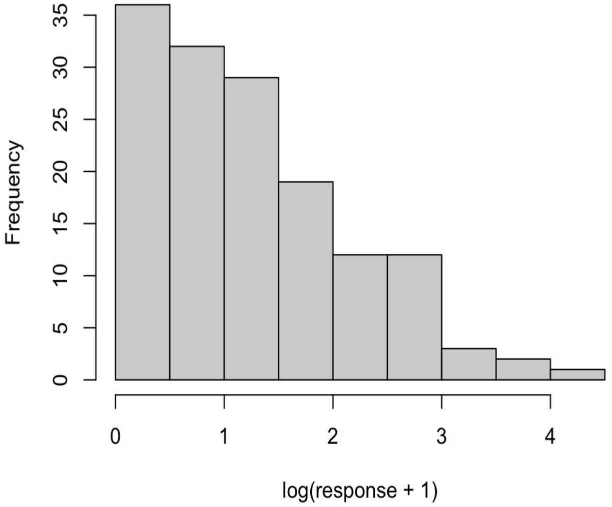


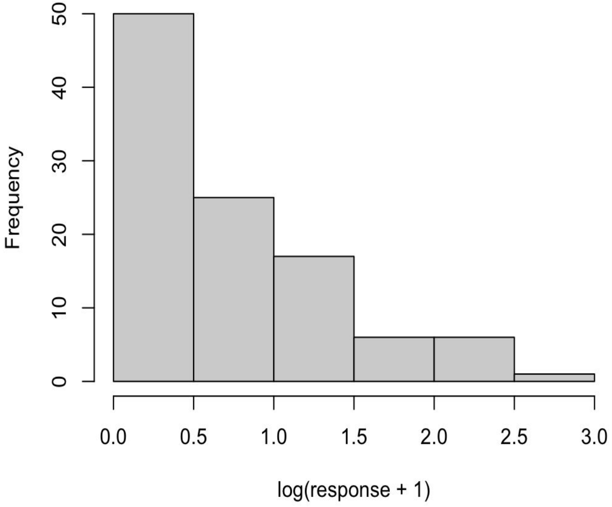


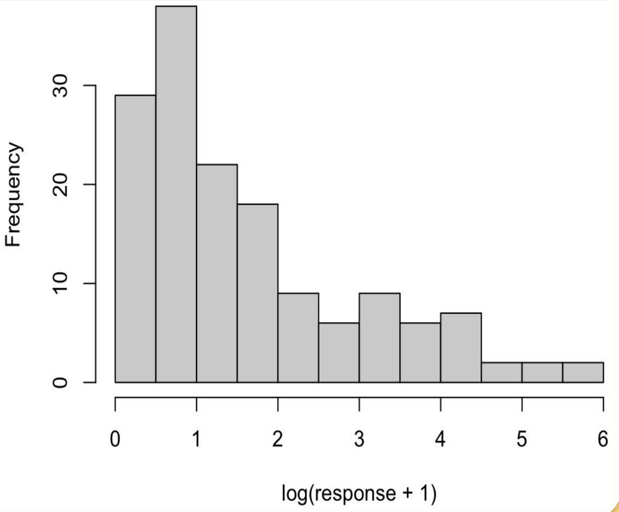


After Log Transformation

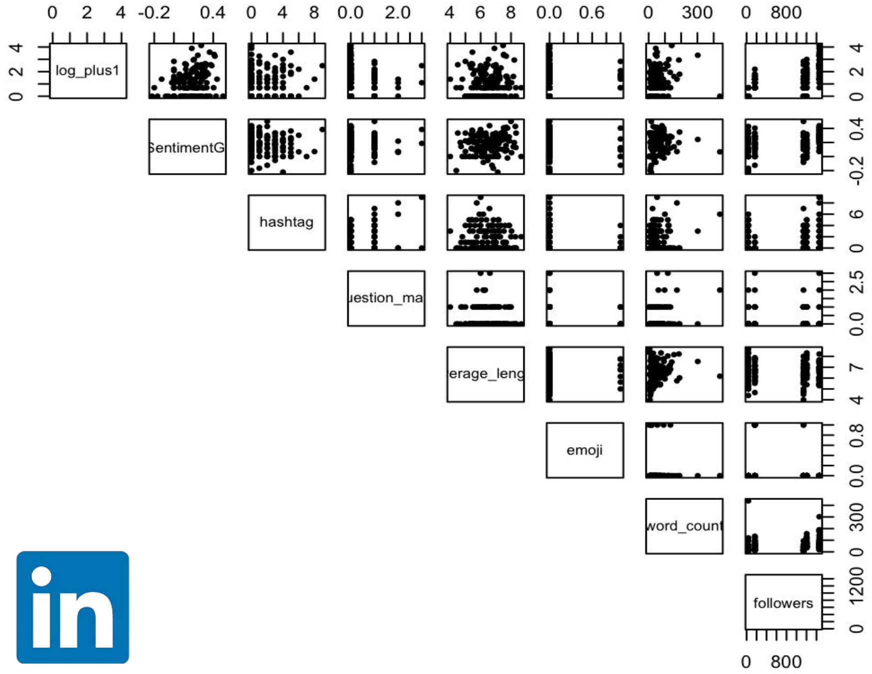
- Applied log(y+1) on all response variables
- log(0) = undefined → error, log(1) = 0
- Impossible normal distribution, no longer heavily skewed

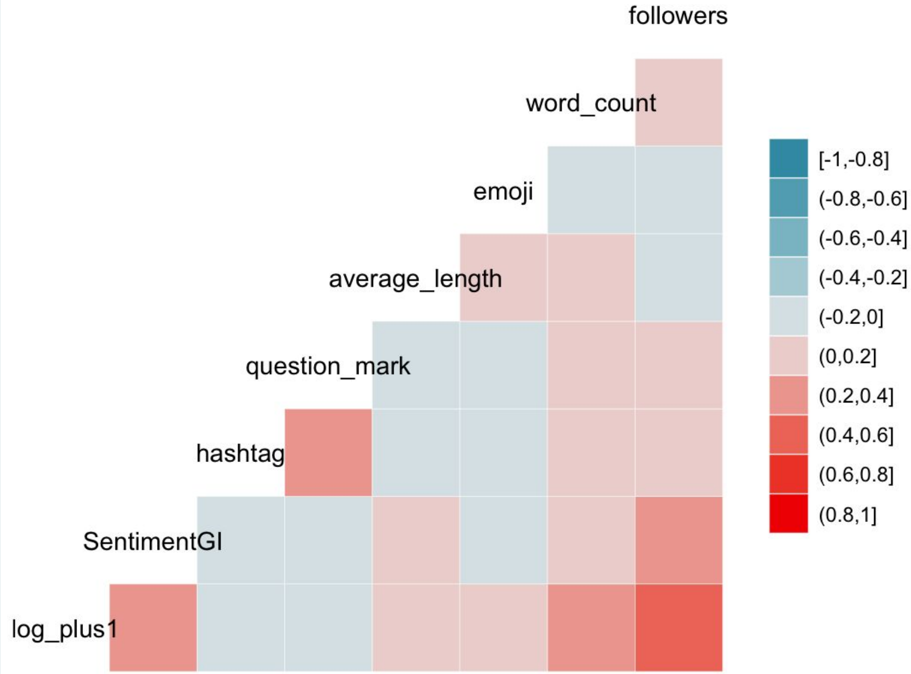




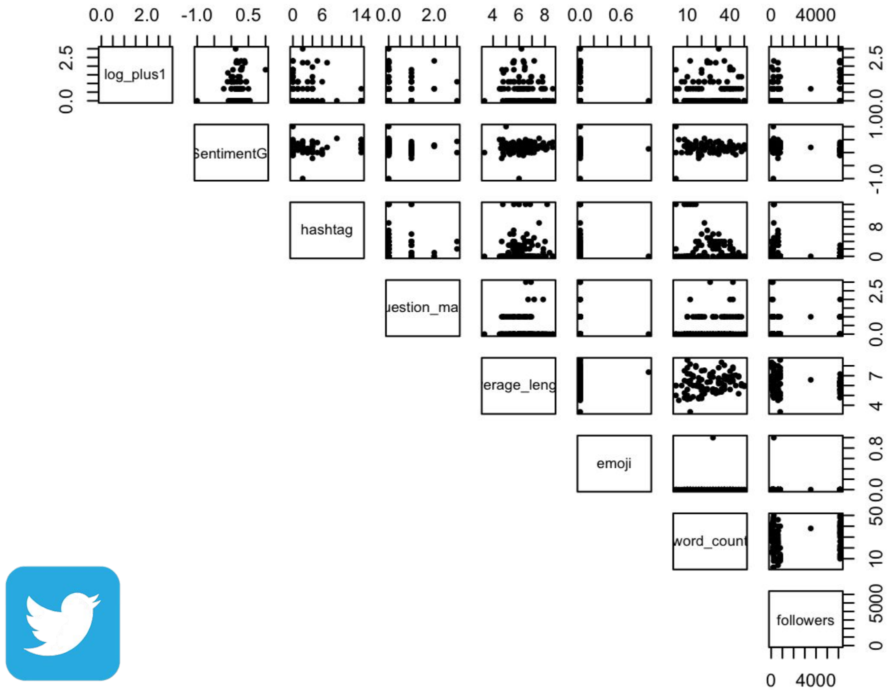


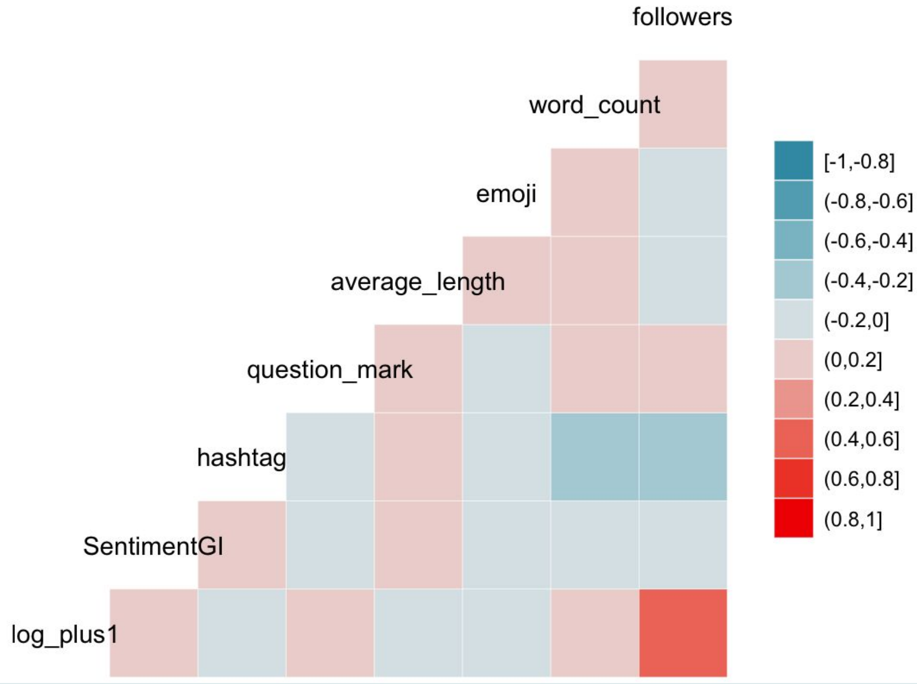
Exploratory Data Analysis



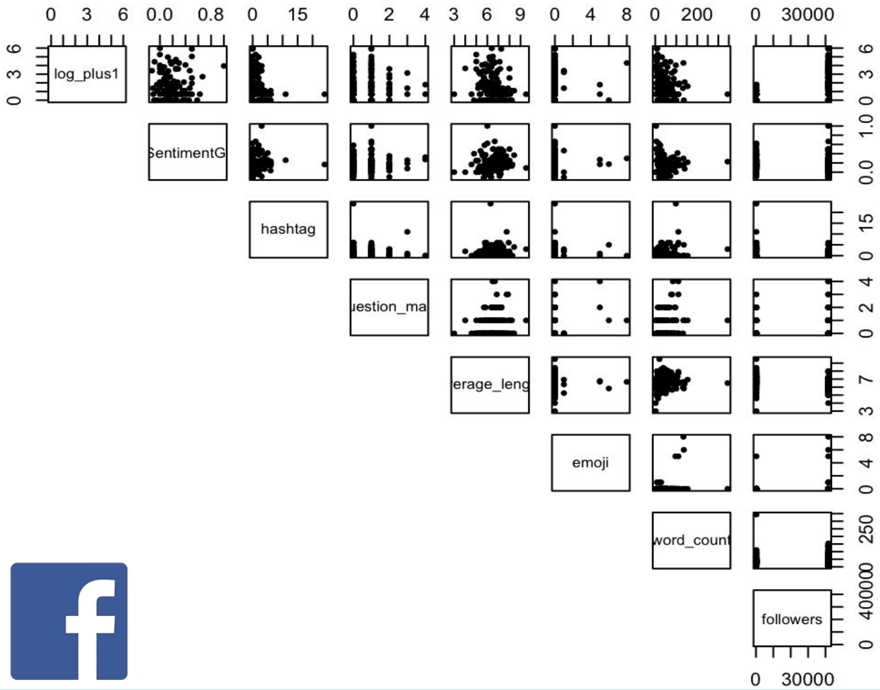


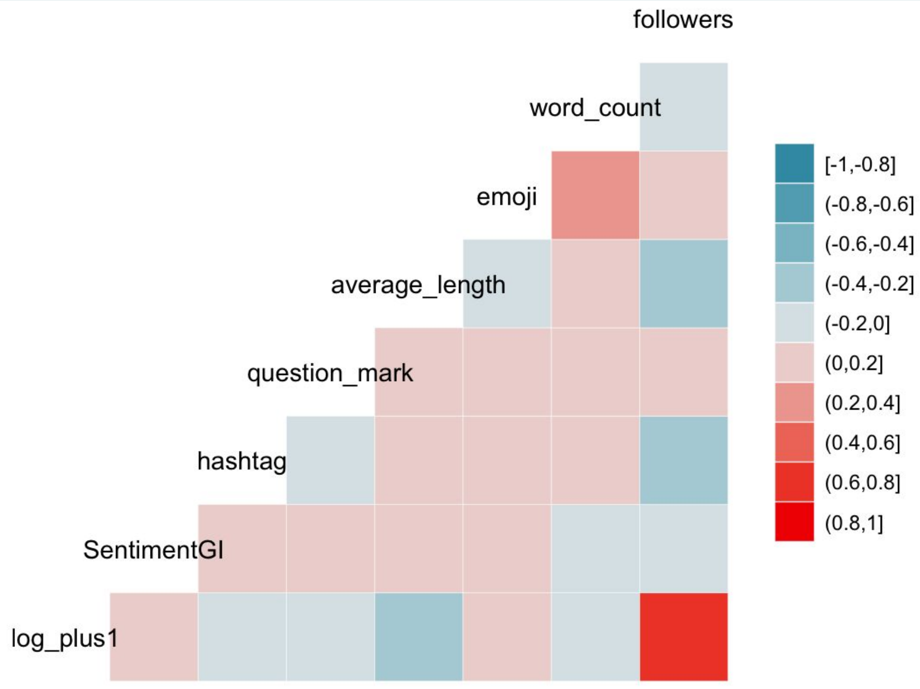
Exploratory Data Analysis





Exploratory Data Analysis

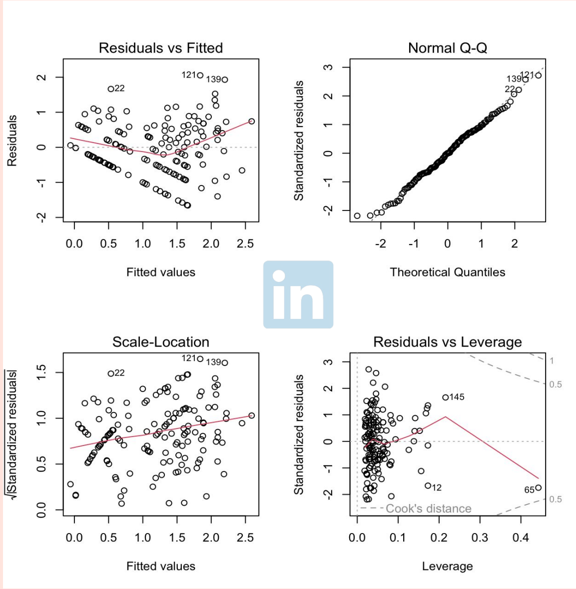


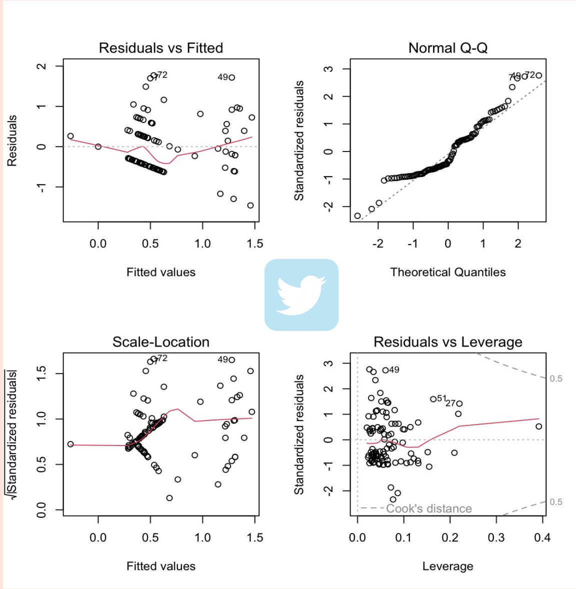


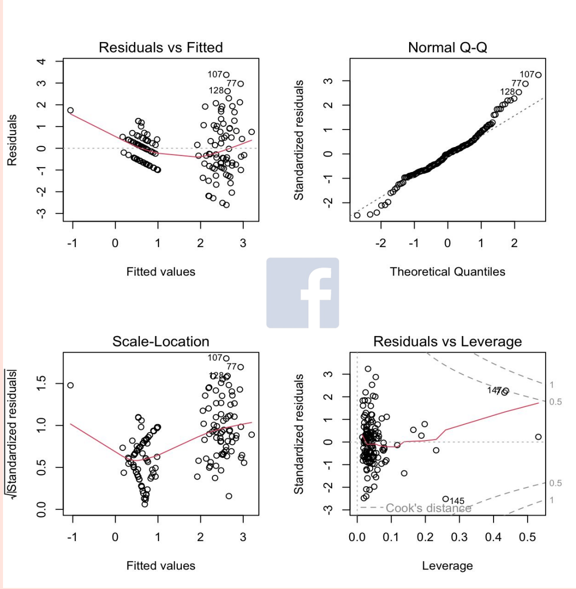
Linear Regression - Assumptions check

The errors, for each fixed value of x,

1) have mean 0
2) have constant variance
3) are independent
4) follow a normal distribution.







LinkedIn

Model Building

Linear Regression - LinkedIn

Significant Predictors

SentimentGI (p = .0343)

Question_mark (p = .040)

Emoji (p = .002)

Word_count (p = .024)

Followers* (p < .001)

Adjusted R-squared

0.3998

F-statistic

F(138) = 14.8 (p < .001)

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.65726 -0.51372  0.00378  0.53597  2.05347

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.3338731  0.4783368   0.698  0.48636
SentimentGI     1.0437475  0.4881051   2.138  0.03425 *
hashtag        -0.0072577  0.0344716  -0.211  0.83356
question_mark  -0.2247075  0.1084438  -2.072  0.04012 *
average_length -0.0293008  0.0719516  -0.407  0.68447
emoji           0.9098001  0.2852374   3.190  0.00176 **
followers       0.0009567  0.0001154   8.293 8.84e-14 ***
word_count      0.0028220  0.0012367   2.282  0.02402 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7667 on 138 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.4288,    Adjusted R-squared:  0.3998
F-statistic:  14.8 on 7 and 138 DF,  p-value: 2.638e-14
```

Improved Linear Regression - LinkedIn

Likelihood Ratio Test

p = 0.897

Significant Predictors

SentimentGI (p = .032)

Question_mark (p = .030)

Emoji (p = .002)

Word_count (p = .026)

Followers* (p < .001)

Adjusted R-squared

0.4075 (originally 0.3998)

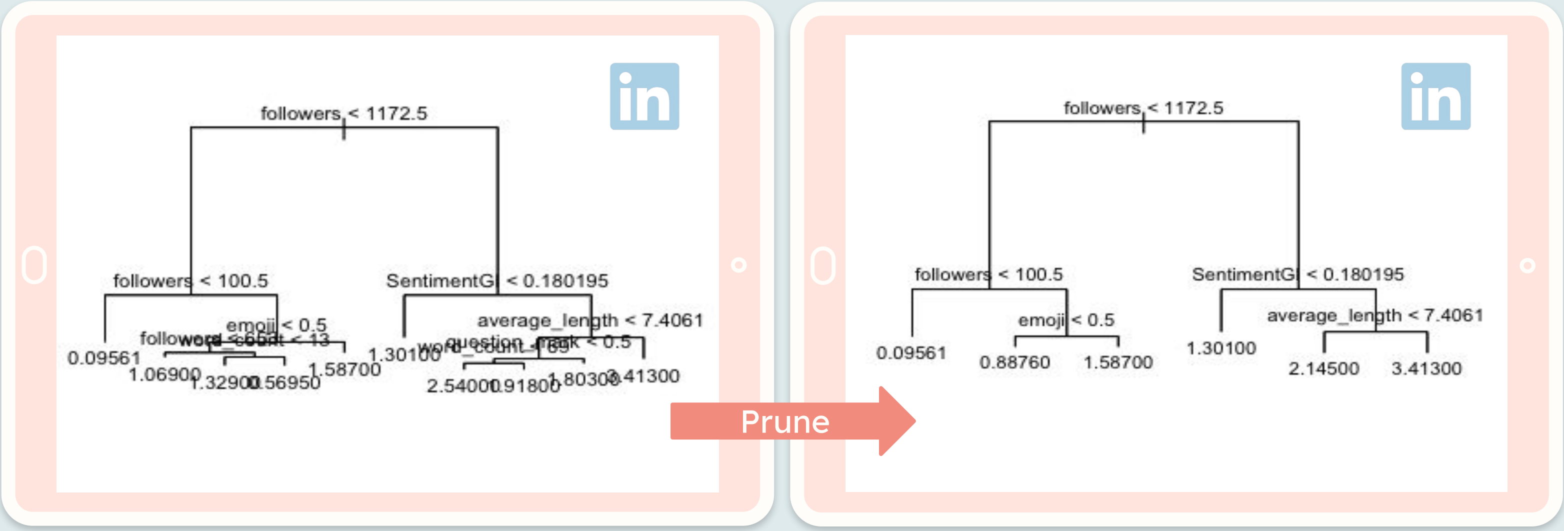
```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.6302 -0.5241 -0.0087  0.5201  2.0494 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.1393971  0.1423708   0.979  0.32921    
SentimentGI   1.0311503  0.4754367   2.169  0.03178 *  
question_mark -0.2270002  0.1038036  -2.187  0.03042 *  
emoji         0.9101241  0.2816855   3.231  0.00154 ** 
word_count    0.0027064  0.0012012   2.253  0.02580 *  
followers     0.0009591  0.0001136   8.443 3.52e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7618 on 140 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.428,	Adjusted R-squared:  0.4075 
F-statistic: 20.95 on 5 and 140 DF,  p-value: 1.362e-15
```

Regression Tree - LinkedIn

- Important predictors: **Sentiment**, **Emoji Counts**, **Average Length**, Followers*
- Pruning: **6 terminal nodes** selected from 10-fold Cross Validation
- Interpretation example:
    - For accounts with more than **1173 followers**, post with **slightly positive sentiment** (greater than 0.18) and average word length greater than **7 letters** are predicted to receive **3 to 4** likes and comments.



Random Forest - LinkedIn

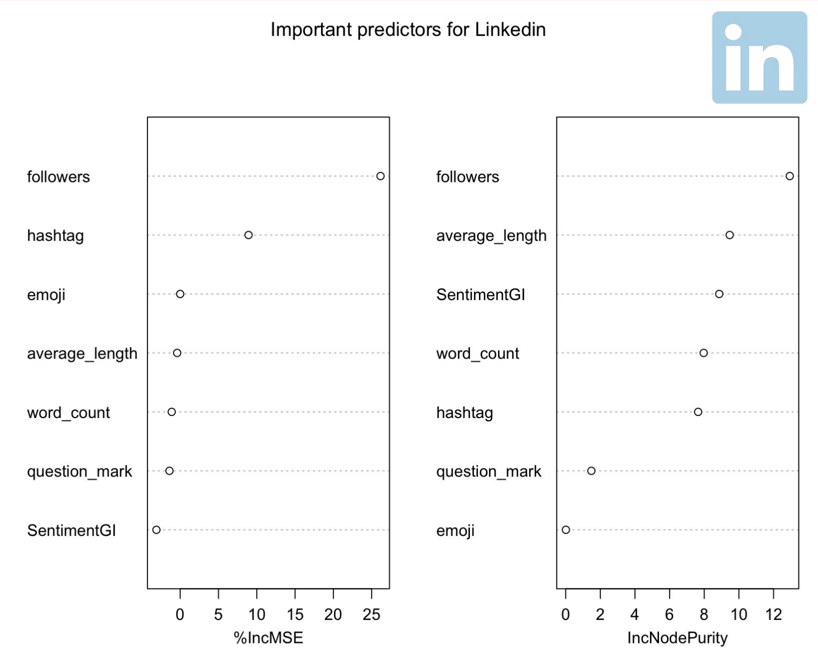
Number of trees: 500

No. of variables tried at each split: 4

Mean of squared residuals: 0.50

% Var explained: 4.55

	%IncMSE	IncNode Purity
SentimentGI	-3.10	8.87
hashtag	8.93	7.65
question_mark	-1.40	1.48
average_length	-0.40	9.47
emoji	0	0.01
followers	26.16	12.94
word_count	-1.11	7.97



Twitter

Model Building

Linear Regression - Twitter

Significant Predictors

SentimentGI (p = .061)

Followers* (p < .001)

Adjusted R-squared

0.1943

F-statistic

F(97) = 4.584 (p < .001)

```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.4585 -0.4509 -0.2127  0.3136  1.7716 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     5.602e-01  4.287e-01   1.307   0.1944    
SentimentGI     5.965e-01  3.141e-01   1.899   0.0605 .  
hashtag        -4.646e-03  1.843e-02  -0.252   0.8015    
question_mark   1.673e-03  1.039e-01   0.016   0.9872    
average_length -4.118e-02  6.621e-02  -0.622   0.5354    
emoji          -3.663e-01  6.601e-01  -0.555   0.5802    
followers       1.499e-04  2.996e-05   5.002 2.52e-06 ***
word_count     -3.848e-04  5.288e-03  -0.073   0.9421    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6498 on 97 degrees of freedom
Multiple R-squared:  0.2486,	Adjusted R-squared:  0.1943 
F-statistic: 4.584 on 7 and 97 DF,  p-value: 0.0001852
```

Linear Regression - Twitter

Likelihood Ratio Test

p = 0.9953

Significant Predictors

SentimentGI (p = .057)

Followers* (p < .001)

Adjusted R-squared

0.218 (originally 0.1943)

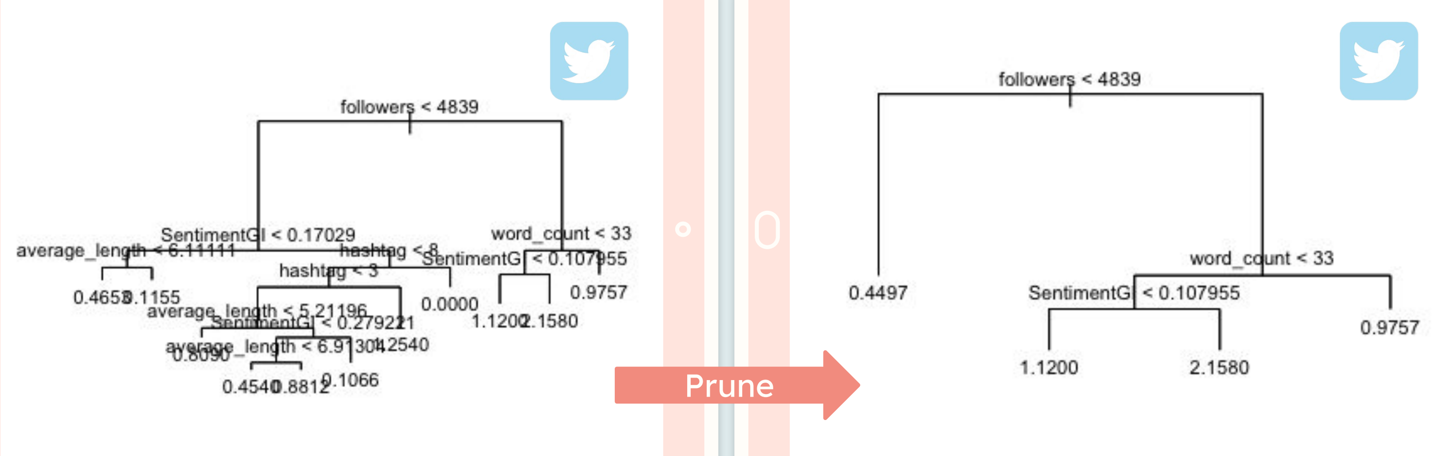
```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4595 -0.4692 -0.2118  0.3192  1.7613

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.406e-01  4.135e-01   1.307   0.1940
SentimentGI     5.959e-01  3.089e-01   1.929   0.0566 .
average_length -4.162e-02  6.393e-02  -0.651   0.5165
emoji          -3.546e-01  6.479e-01  -0.547   0.5854
followers       1.515e-04  2.807e-05   5.397 4.57e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6402 on 100 degrees of freedom
Multiple R-squared:  0.2481,    Adjusted R-squared:  0.218
F-statistic: 8.247 on 4 and 100 DF,  p-value: 8.636e-06
```

Regression Tree - Twitter

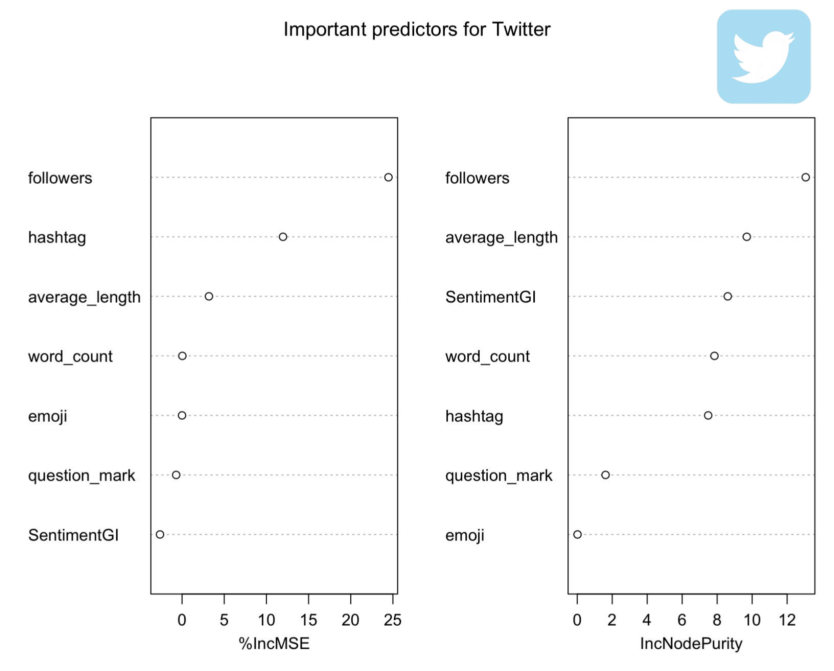
- Important predictors: **Word Counts**, **Sentiments**, Followers*
- Pruning: **4 terminal nodes** selected from 10-fold Cross Validation
- Interpretation example:
  - For accounts with more than **4839 followers**, post with less than **33 words** and a **slightly positive sentiment** (greater than 0.1) are predicted to receive **2** likes and comments.



Random Forest - Twitter

Number of trees: 500
No. of variables tried at each split: 4
Mean of squared residuals: 0.48
% Var explained: 7.16

	%IncMSE	IncNode Purity
SentimentGI	-2.62	8.61
hashtag	11.97	7.49
question_mark	-0.70	1.62
average_length	3.19	9.69
emoji	0	0.02
followers	24.47	13.06
word_count	0.04	7.84



FaceBook

Model Building

Improved Linear Regression - FaceBook

Significant Predictors

Followers* (p < .001)

Word_count (p = .018)

Adjusted R-squared

0.4522

F-statistic

F(141) = 18.45 (p < .001)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.6064 -0.6889 -0.0135  0.4867  3.3782

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.541e-01  7.828e-01   0.580   0.5628
SentimentGI     5.614e-01  5.454e-01   1.029   0.3051
hashtag         7.350e-03  3.491e-02   0.211   0.8335
question_mark  -1.084e-01  1.123e-01  -0.965   0.3360
average_length  5.355e-02  1.157e-01   0.463   0.6440
emoji           4.826e-02  9.287e-02   0.520   0.6041
followers       4.732e-05  4.739e-06   9.985   <2e-16 ***
word_count     -5.552e-03  2.326e-03  -2.387   0.0183 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.061 on 141 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4781,	Adjusted R-squared:  0.4522
F-statistic: 18.45 on 7 and 141 DF,  p-value: < 2.2e-16
```

Improved Linear Regression - FaceBook

Likelihood Ratio Test

p = .768

Significant Predictors

Followers* (p < .001)

Word_count (p = .009)

Adjusted R-squared

0.4618 (originally 0.4522)

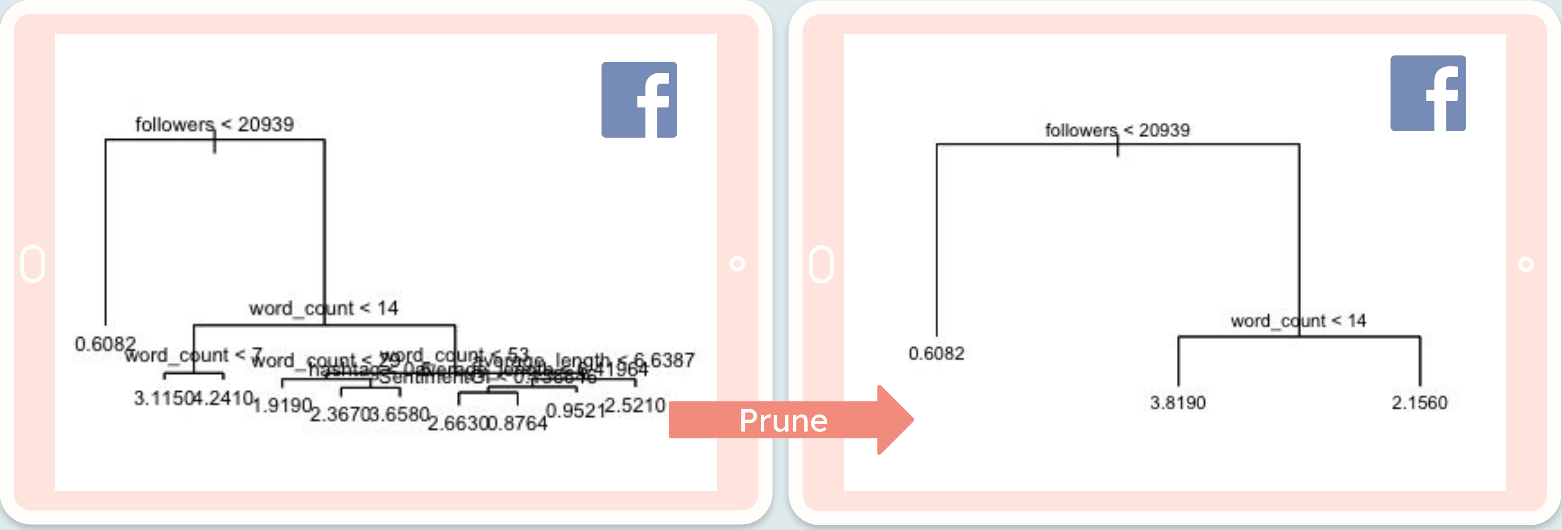
```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.6409 -0.6837  0.0012  0.5081  3.2503 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.840e-01  1.640e-01   5.389 2.78e-07 ***
word_count  -5.639e-03  2.135e-03  -2.642  0.00915 ** 
followers    4.572e-05  4.163e-06  10.983  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.051 on 146 degrees of freedom
Multiple R-squared:  0.4691,	Adjusted R-squared:  0.4618 
F-statistic:  64.5 on 2 and 146 DF,  p-value: < 2.2e-16
```

Regression Tree - FaceBook

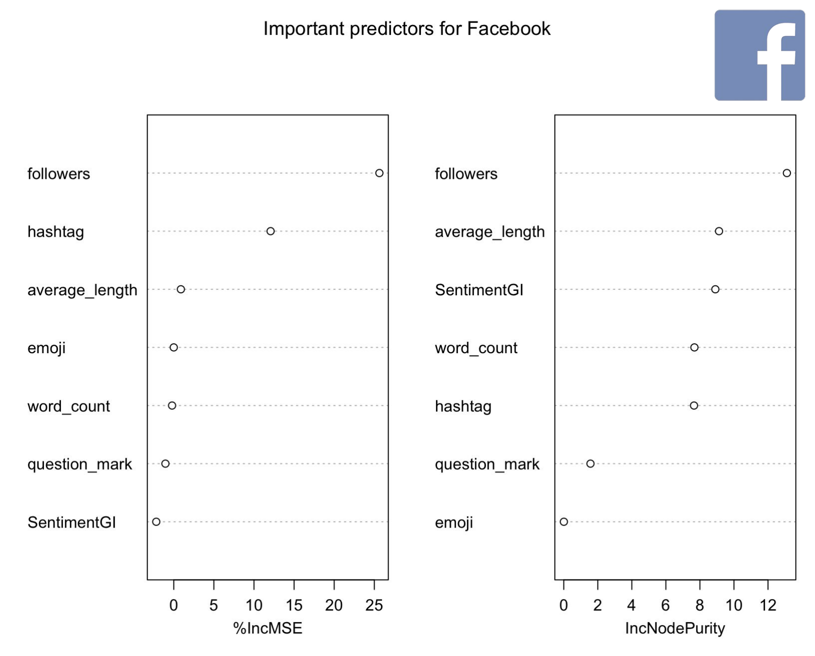
- Important predictor: **Word Counts**, Followers*
- Pruning: **3 terminal nodes** selected from 10-fold Cross Validation
- Interpretation example:
  - For accounts with more than **20939 followers**, post with less than **14 words** are predicted to receive **3.8** likes and comments.



Random Forest - FaceBook

Number of trees: 500
No. of variables tried at each split: 4
Mean of squared residuals: 0.49
% Var explained: 6.14

	%IncMSE	IncNode Purity
SentimentGI	-2.18	8.91
hashtag	12.07	7.65
question_mark	-1.04	1.57
average_length	0.89	9.12
emoji	0	0.01
followers	25.63	13.11
word_count	-0.21	7.68



Conclusion

Conclusion & Recommendation

LinkedIn

- **Sentiment, Emoji, Average_Length** are the most important predictors
- Choose a more positive tone, use more hashtags and longer words

Twitter

- **Sentiment** is the most important predictor
- Choose a more positive tone

FaceBook

- **Word_count** is the most important predictor
- Be concise and avoid tedious long sentences

Limitation - Data

01 Unbalanced Data

- Due to limitation of the open-source packages, some accounts cannot be scrapped
- A limit to the number of posts that can be scrapped from each platform
  - Log in requirement
  - Scraping restriction that results in disabling the package
- Small company size leads to few followers and low responses

02 Transformed response variable to log(y+1)

- Highly right-skewed response variable
- Presence of 0 in response variable of some observations
- Unable to simply apply log transformation
- Unusual but valid transformation

Limitation - Model Building

01 Small sample size: 100-150 posts for every platform

- Unable to split training data and test data
- Less powerful models
- High variance

02 Future Improvement

- Collect more data
- Consider using Ridge Regression to reduce variance when sample size is small

Reference

Packages:

- https://github.com/shaikhsajid1111/twitter-scraper-selenium
- https://github.com/kevinzg/facebook-scraper
- https://github.com/daniyalrmb/LinkedIn-Scraping-Sentiment
- https://advertools.readthedocs.io/en/master/
- https://www.nltk.org/howto.html
- https://cran.r-project.org/web/packages/SentimentAnalysis/SentimentAnalysis.pdf

Thank you for listening!