

Master Modeler Competition

Round 1

Joey Hodson and Laird Stewart
Emory University
March 1, 2021



Agenda

- ❖ Introduction
- ❖ Data
- ❖ Model
- ❖ Conclusion

**ERASE
ERASE
CHILD TRAFFICKING**

Introduction



Joey Hodson



Laird Stewart

Objectives

“Leverage social media to reach individuals in need of assistance for leaving a human trafficking situation.”

- A. Determine the types of posts that are most successful
- B. Estimate the value of each type of engagement (likes, etc.)
- C. Evaluate post attributes (use of celebrities, etc.) that tend to drive more positive post results

Agenda

- ❖ Introduction
- ❖ Data
- ❖ Model
- ❖ Conclusion

**ERASE
ENHANCE
CHILD TRAFFICKING**

Data Cleaning

01

Cleaning - missing values, data types, duplicates, etc..

02

Filtering - subsetting data to highlight information

Data Manipulation

01

Hot Encoding - qualitative variables to binary

02

Time Matching - utilizing time as a variable of data

03

Custom Features - combining content for additional value

04

Aggregation - merging disparate data sets

05

Dimensionality Reduction - for model training purposes

Post Attribute Analysis

- ❖ Sentiment Analysis
- ❖ Length
- ❖ Endorsements
- ❖ Media Form (video/image/etc.)
- ❖ Post Type (informational, personal, etc.)
- ❖ Time
- ❖ Hashtag Use
- ❖ Etc...

 ERASE Child Trafficking February 2 at 4:10 PM 

We love the work that Genesee Human Oppression Strike Team (GHOST) is doing to keep their community safe. In addition to their recoveries and investigations, they are educating the public about online safety.

"Kids often record these videos late at night when parents are asleep. They can be encouraged by older people on these apps to do certain things in videos that could be explicit."

#tiktok #discord



TCTIMES.COM
Sheriff holds human trafficking town hall
Please purchase a subscription to read our premium content. If ...

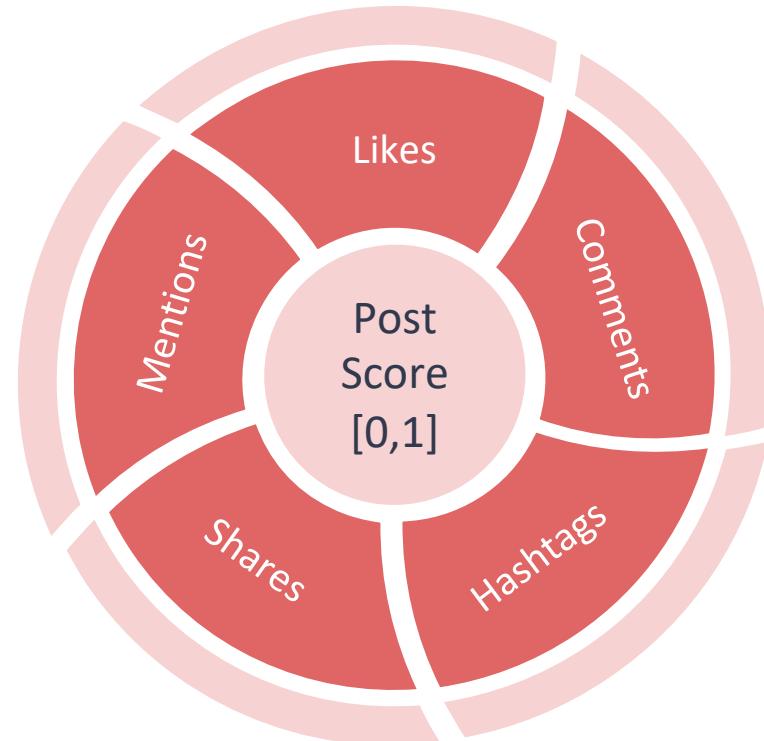
Agenda

- ❖ Introduction
- ❖ Data
- ❖ Model
- ❖ Conclusion

**ERASE
ENHANCE
CHILD TRAFFICKING**

Metric Weighting

1. Engagement Rate by Reach (ERR)
2. Engagement Rate by Post (ERP)
3. Engagement Rate by Impressions (ERI)
4. Daily Engagement Rate (DER)
5. Engagement Rate by Views (ERV)
6. Factored Engagement Rate (FER)



Machine Learning Model Development

Model objective:

Determine which attributes have the greatest impact on Post Score [0,1]

Model inputs:

Post attributes, Post Score [0,1]

Model output:

Coefficients of attributes

Resources:

Python 3.9, Scikit-learn, etc...



Agenda

- ❖ Introduction
- ❖ Data
- ❖ Model
- ❖ Conclusion

**ERASE
ENHANCE
CHILD TRAFFICKING**

Conclusion

With this analysis and modeling we hope to improve ERASE Child Trafficking's ability to increase awareness of human and child trafficking, reduce the level of trafficking, and reach individuals in need of assistance for leaving a human trafficking situation.

Master Modeler Competition

Round 2

Joey Hodson and Laird Stewart
Emory University
March 23, 2021



Agenda

- ❖ Introduction
- ❖ Data Exploration
- ❖ Data Cleaning
- ❖ Features and Post Scoring
- ❖ Model
- ❖ Appendix



(Re)Introduction



Joey Hodson



Laird Stewart

Objectives

“Leverage social media to reach individuals in need of assistance for leaving a human trafficking situation.”

- A. Determine the types of posts that are most successful
- B. Estimate the value of each type of engagement (likes, etc.)
- C. Evaluate post attributes (use of celebrities, etc.) that tend to drive more positive post results

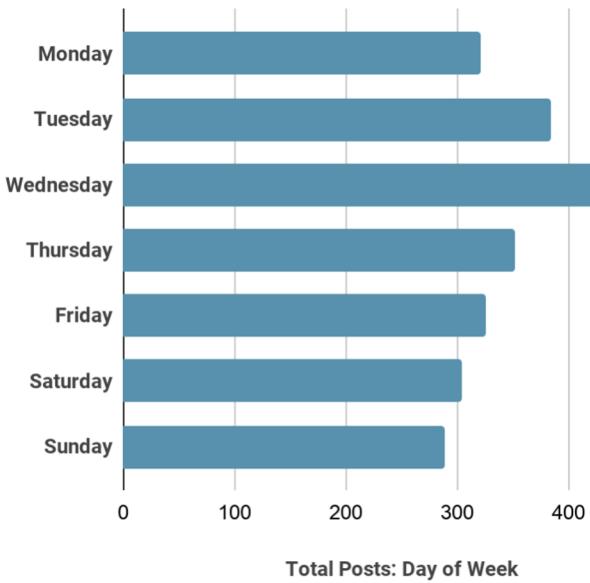
Agenda

- ❖ Introduction
- ❖ Data Exploration
- ❖ Data Cleaning
- ❖ Features and Post Scoring
- ❖ Model
- ❖ Appendix

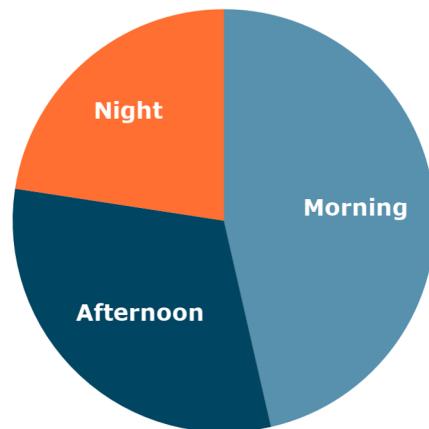


Post Frequency and Timing

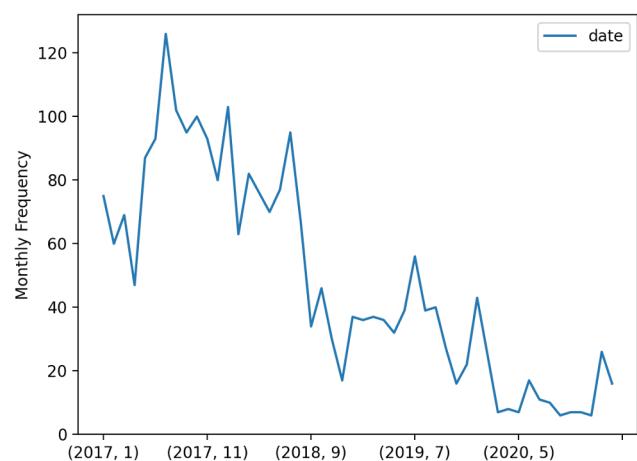
Day of Week



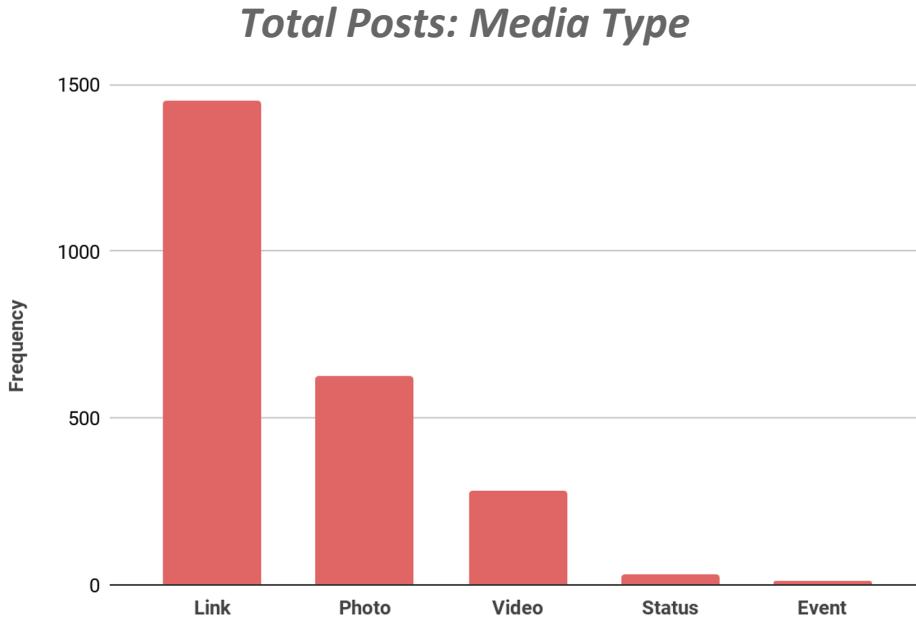
Time of Day



Monthly Post Frequency



Media Type and Body Text



Body Text/Hashtag Summary

Text Length (characters):

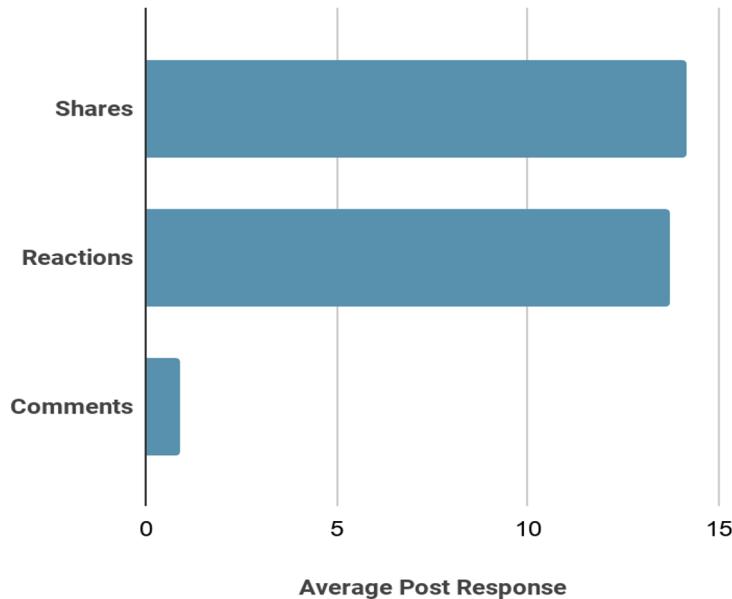
- 192 - mean ~ 30 words
- 182 - standard deviation

Hashtag Usage:

- 0.85 - mean
- 1.80 - standard deviations

Post Outcome Breakdown

Average Response to Posts



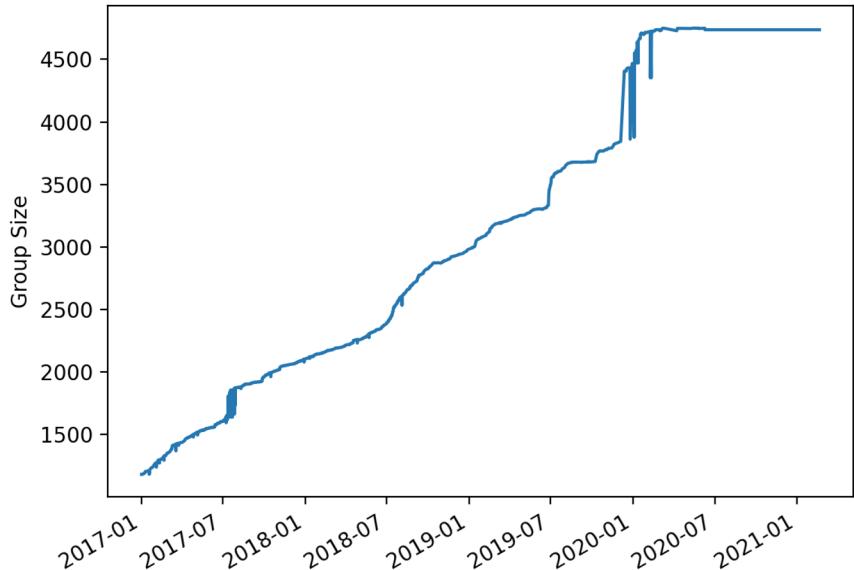
Engagement Summary

Total Engagement:
(Shares+Reactions+Comments)

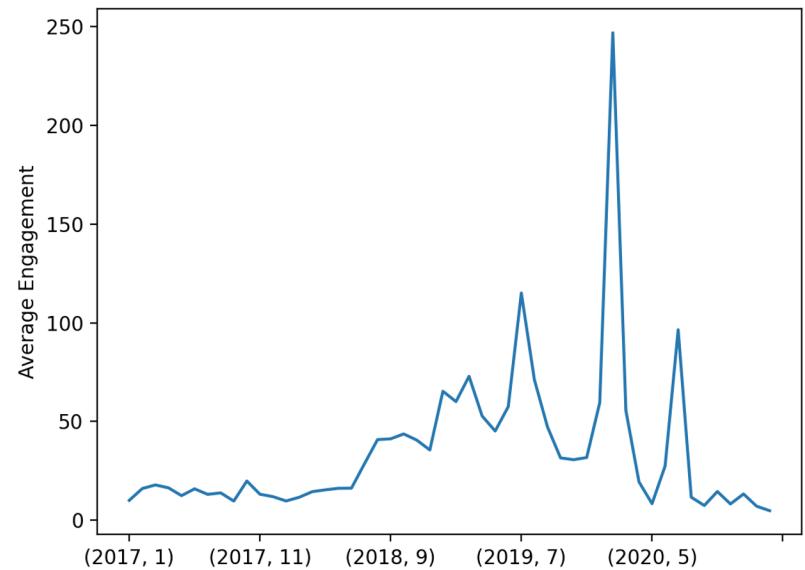
- 29 - mean
- 100 - standard deviation
- 12 - median
- 4057 - best post

Page Engagement

Total “Group Size”



Average Post Engagement (Monthly)

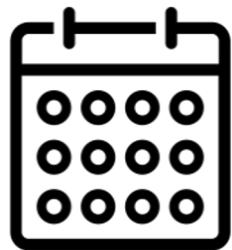


Agenda

- ❖ Introduction
- ❖ Data Exploration
- ❖ Data Cleaning
- ❖ Features and Post Scoring
- ❖ Model
- ❖ Appendix



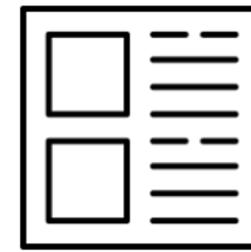
Data Cleaning



Post
Date/Time

NaN

Missing
Values



Text
Processing

Agenda

- ❖ Introduction
- ❖ Data Exploration
- ❖ Data Cleaning
- ❖ Features and Post Scoring
- ❖ Model
- ❖ Appendix



Feature Extraction (i.e. post attributes)

- ❖ Time of Day (Morning, Afternoon, Night)
- ❖ Day of Week
- ❖ Season
- ❖ Text Length
- ❖ Text Sentiment*
- ❖ Number of Hashtags
- ❖ Media Type (Photo, Video, Link, Event, Status)

* Found using VADER sentiment analysis

Post Scoring

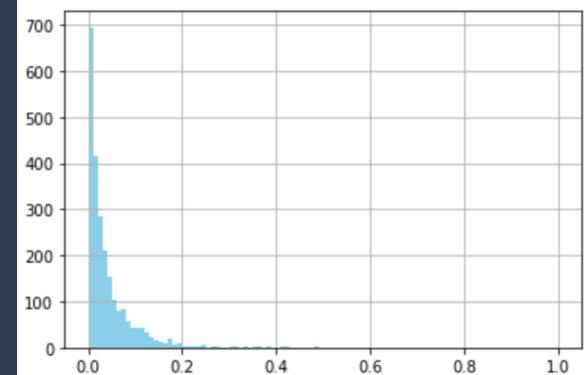
(1)
Weight
responses

(2)
Divide by
Group Size

(3)
Log Scale/
[0,1] Transform

Reactions $\times 1$
Comments $\times 10$
Shares $\times 50$

$$\frac{\text{Weighted Score}}{\text{Group Size}}$$



Agenda

- ❖ Introduction
- ❖ Data Exploration
- ❖ Data Cleaning
- ❖ Features and Post Scoring
- ❖ Model
- ❖ Appendix



Model Selection

We used a combination of Regression and Classification models to verify our results

Linear Regression

Linear Regression:

- Maps linear relations
- Derives statistical significance

Random Forest Regression:

- Maps nonlinear relations
- Versatile
- Impurity - variance

Classification

Random Forest Classification:

- Maps nonlinear relations
- Versatile
- Impurity - gini/entropy

Significant Post Features

Features	Significance	Interpretation
Length	Very Strong	Longer length associates with successful posts
Link	Very Strong	Link usage associates with successful posts
Hashtags	Strong	Hashtag usage associates with successful posts
Sentiment	Strong	<u>Negative</u> sentiment associates with successful posts
Summer	Medium	Summer posting associates with successful posts
Morning	Medium	Morning posting associates with successful posts

Takeaways

1. Write longer length posts (~50+ words)
2. Keep sharing relevant links and articles
3. Utilize multiple hashtags that describe your post
4. Posts with negative sentiment draw more attention
5. Promote events/fundraisers in the summer to naturally gain more traction
6. Schedule daily posts before 12:00 pm

Agenda

- ❖ Introduction
- ❖ Data Exploration
- ❖ Data Cleaning
- ❖ Features and Post Scoring
- ❖ Model
- ❖ Appendix



Appendix Contents

- Data Cleaning/Wrangling
- OLS Linear Regression
- Random Forest Regression/Classification
- Limitations



Data Cleaning Steps:

1. Created datetime objects for easier manipulation
2. Fill NA values with empty strings in text column
3. Clean text data using REGEX

Data Wrangling Steps:

1. Convert non-binary feature to dummy variables (EX: link to 1/0)
2. Create dummy variables from datetime date (EX: season)
3. Feature creation from text data
 - a. Extract variable Hashtags
 - b. Extract variable Length
 - c. Extract variable Sentiment using VADER Sentiment analysis. We choose VADER for our sentiment analysis because it is designed to perform well on social media text, does not require training data, and does not suffer from a speed-performance tradeoff.
4. Feature scale all variables to [0,1] using min-max scaler so that regression model isn't biased
 - * Note: Random Forest models do not require scaling but testing both scaled and unscaled data showed negligible differences in feature importance.
1. Log transform Post Score to help normalize the score

	total engagement	engagement rate	reactions	shares	comments	length	hashtags	sentiment	score
count	2400.000000	2400.000000	2400.000000	2400.000000	2400.000000	2400.000000	2400.000000	2400.000000	2400.000000
mean	28.837083	1.062614	13.742500	14.171667	0.922917	192.175417	0.848750	0.097808	0.258786
std	100.173733	2.657102	19.916801	88.641005	2.940761	182.776642	1.801544	0.541320	1.169741
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.993600	0.000000
25%	6.000000	0.272346	4.000000	1.000000	0.000000	78.750000	0.000000	-0.296000	0.027919
50%	12.000000	0.549539	8.000000	4.000000	0.000000	169.000000	0.000000	0.000000	0.091365
75%	27.000000	1.079793	15.000000	10.000000	1.000000	256.000000	1.000000	0.598525	0.221454
max	4057.000000	85.953390	292.000000	3745.000000	67.000000	2148.000000	32.000000	0.992800	39.775847
						norm length	norm hashtags	norm sentiment	norm log score
						count	2400.000000	2400.000000	2400.000000
						mean	0.089467	0.026523	0.549440
						std	0.085092	0.056298	0.272513
						min	0.000000	0.000000	0.000000
						25%	0.036662	0.000000	0.351188
						50%	0.078678	0.000000	0.500201
						75%	0.119181	0.031250	0.801513
						max	1.000000	1.000000	1.000000

Appendix 1b: Summary Statistics

Appendix Contents

- Data Cleaning/Wrangling
- OLS Linear Regression
- Random Forest Regression/Classification
- Limitations



As our first model, we chose to use an ordinary least squares linear regression. Our variables had very little multicollinearity problems making it a good candidate for testing linear relationships. From our regression coefficients, we then ran a statistical significance test to find which features had a ceteris paribus effect on the output score.

	Estimate	Std. Error	t value	Pr(> t)	norm.length	0.0144721	0.0182279	0.794	0.427300
(Intercept)	0.0267729	0.0075510	3.546	0.000399 ***	norm.sentiment	-0.0054779	0.0054602	-1.003	0.315847
spring	-0.0010880	0.0041033	-0.265	0.790919	norm.hashtags	0.1034904	0.0286971	3.606	0.000317 ***
summer	0.0083927	0.0038705	2.168	0.030228 *	event	-0.0293940	0.0224639	-1.308	0.190830
fall	-0.0068928	0.0042195	-1.634	0.102482	link	0.0160778	0.0050464	3.186	0.001461 **
winter	NA	NA	NA	NA	photo	0.0015604	0.0053675	0.291	0.771300
sunday	-0.0021187	0.0057359	-0.369	0.711878	status	-0.0085057	0.0132957	-0.640	0.522408
morning	0.0046292	0.0039665	1.167	0.243305	video	NA	NA	NA	NA
afternoon	-0.0004717	0.0042852	-0.110	0.912358	---				
night	NA	NA	NA	NA	Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
monday	0.0074337	0.0055984	1.328	0.184358	
tuesday	0.0056253	0.0053971	1.042	0.297381	Residual standard error:	0.0697	on 2381 degrees of freedom		
wednesday	0.0047575	0.0053075	0.896	0.370151	Multiple R-squared:	0.02641,	Adjusted R-squared:	0.01905	
thursday	0.0028226	0.0054666	0.516	0.605676	F-statistic:	3.589	on 18 and 2381 DF,	p-value:	4.633e-07
friday	0.0090065	0.0055825	1.613	0.106804					
saturday	NA	NA	NA	NA					

Appendix Contents

- Data Cleaning/Wrangling
- OLS Linear Regression
- Random Forest Regression/Classification
- Limitations



Our second model aimed to capture nonlinear relationships in a regression format. We needed a model that could still extract important features in this nonlinear fashion. After reviewing and testing multitudes of regression outputs we found that the Random Forest Regressor (RFR) gave us the capabilities for proper feature extraction and interpretation.

- With feature importance we decided to use Shapely values. These are the average of all marginal contributions to possible coalitions. The runtime of this increases exponentially but with a manageable feature set this was possible. Additionally, tree SHAP benefits from being consistent and accurate, the proof of this is beyond the scope of our project.

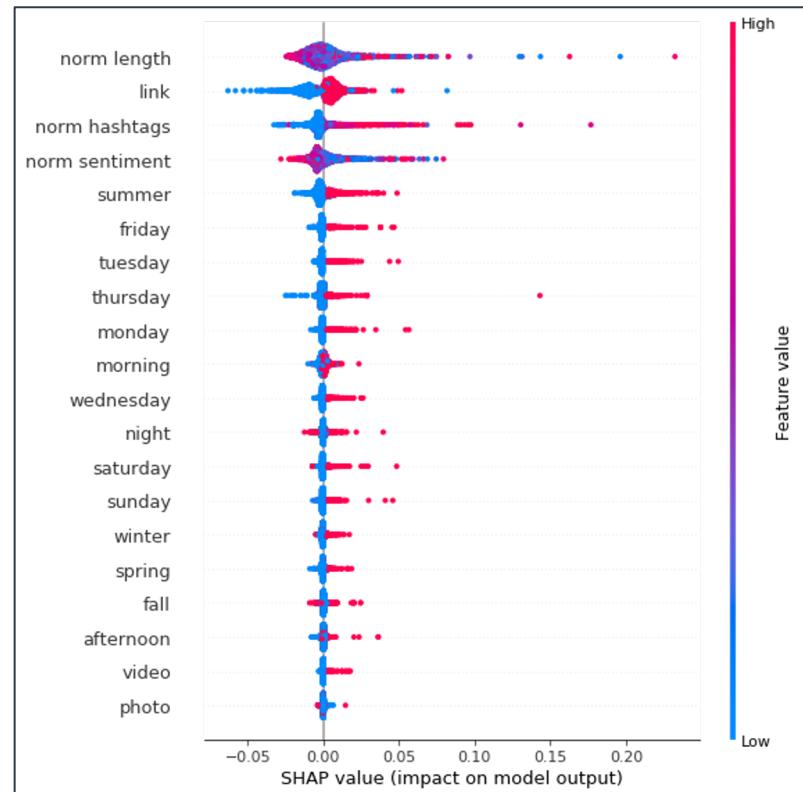
Our third and fourth models aimed to look at the problem from a different lens in order to:

- Verify the regression results
- Continue to find significant features

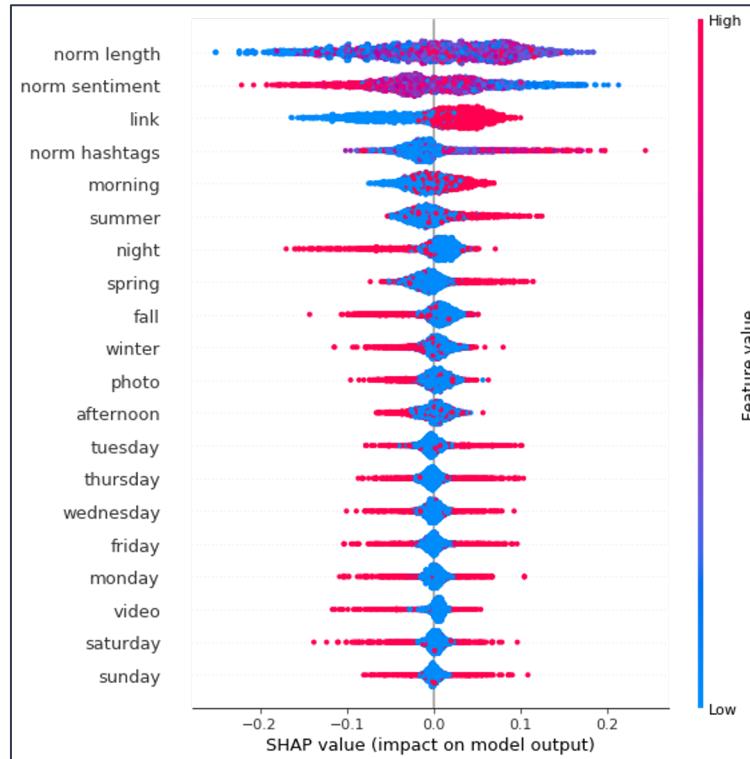
We decided to approach this problem as a multi-class classification problem. With our first model we split the output score from a continuous variable into a “bad”, and “good” classification problem. The splits were not arbitrary but rather at percentiles <50 and >50 for our third model and “bad”, “medium”, “good” at percentiles at the <33, 33-66, >66 for the fourth. We again had to keep in mind “can we extract feature importance” when choosing from a plethora of classification models. We then chose a Random Forest Classifier (RFC) on our data. We again evaluated and interpreted these using SHAP.

RFR Model 2: Feature importance for classifying “Good” posts

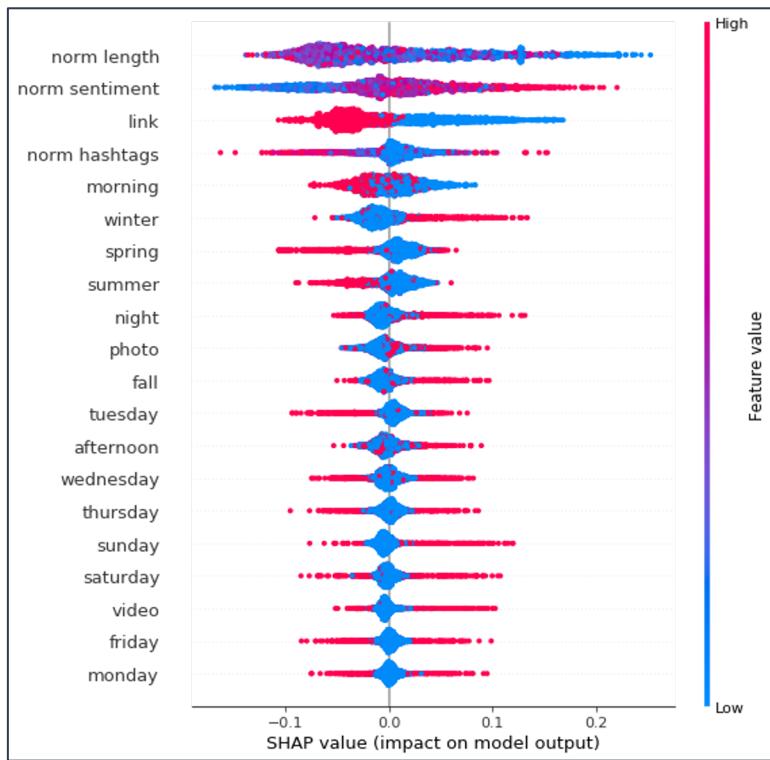
- The y-axis indicates the variable name, in order of importance from top to bottom.
- The x-axis is the SHAP value. Indicates how much is the change in log-odds. From this number we can extract the probability of success.
- Gradient color indicates the original value for that variable. In booleans, it will take two colors, but in number it can contain the whole spectrum.
- Each point represents a row from the original dataset.
- Example: Having a high value for a link (AKA a link being present) has a positive impact on SHAP value and thus, our output. The data becomes less useful when the bar/color begins to look symmetrical.
- Our method for determining the significant features was by cross referencing the top values of the SHAP charts and OLS



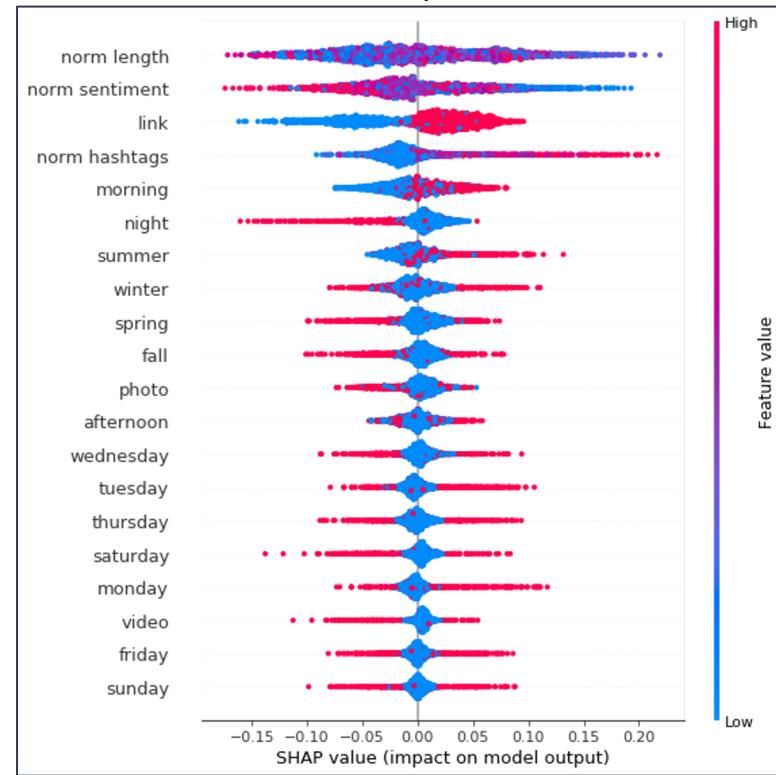
RFC Model 3: Feature importance for classifying “Good” posts



RFC Model 4: Feature importance for classifying
“Bad” posts



RFC Model 4: Feature importance for classifying
“Good” posts



Appendix Contents

- Data Cleaning/Wrangling
- OLS Linear Regression
- Random Forest Regression/Classification
- Limitations



Data Wrangling Limitations

- Omitted variable bias: We attempted to create as many variables from our given data that we felt was important. There are infinite variables that could have effects on our defined “post score” and this is by no means a comprehensive list.
- Sentiment Scoring: Although VADER was developed to analyze social media text, like any other sentiment analyzer it can struggle with more complex sentiments like irony or sarcasm. Further language and text is much more complicated than a simple positive/negative continuum.

Linear Regression Limitations

- Assumptions of Linear Regression: Using OLS requires a set of assumptions regarding linear regression such as feature independence, uncorrelated error terms, constant variance etc... It is extremely difficult to find perfect data that satisfies all these conditions but given what we know we can do our best to hedge the potential biases when implementing these techniques.

Random Forest Limitations:

- Continuous data: Random forest models sometimes tend to value continuous data more than binary, we helped hedge against this bias by implementing the SHAP feature importance vs. the built in parameter.
- Model choice: Our model choice was limited by needing to be able to accurately extract the feature importances. There are hundreds of classification and regression techniques that tinker with loss functions, parameters, and strategy that may have been a better fit for predicting the most important features. We narrowed the field with research on model types, trial and error, and intuition.
- Statistical significance: Unlike OLS where pure statistical significance can be easily derived using p-values, feature extraction from RFR and RFC can dip into a qualitative realm. We again tried to hedge against this by using the SHAP feature which helps limit the bias by being a consistent and accurate ranking of features and their impact on the post score.

SHAP:

- Although SHAP is a popular feature extraction tool for RF models it is not a golden ticket. SHAP will do its best to limit the biases listed above but again is never a perfect answer allowing us to say X causes Y.