Master Modeler Competition
Finalist Round

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.”

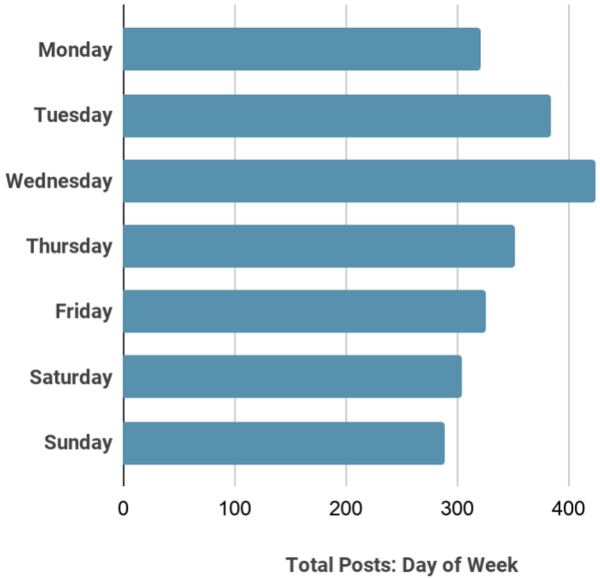
A. Determine the types of posts that are most successful
B. Estimate the value of each type of engagement (likes, etc.)
C. Evaluate post attributes (use of celebrities, etc.) that tend to drive more positive post results

Agenda

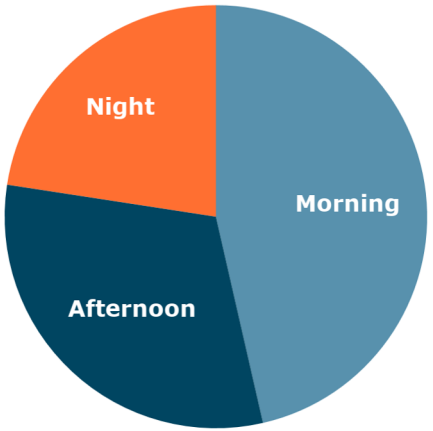
- Introduction
- Data Exploration
- Data Cleaning
- Features and Post Scoring
- Model
- Appendix

Post Frequency and Timing

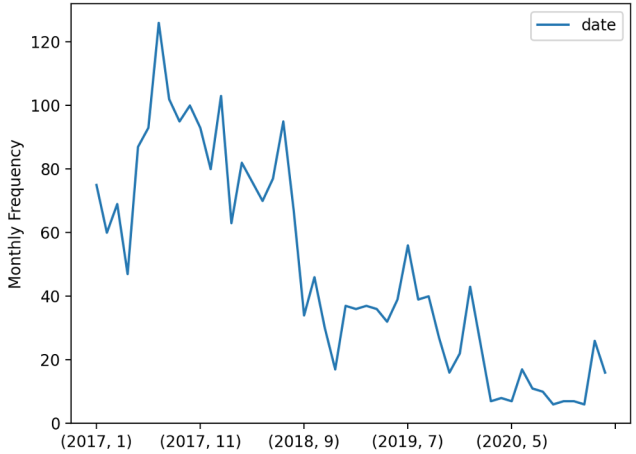
Day of Week



Time of Day

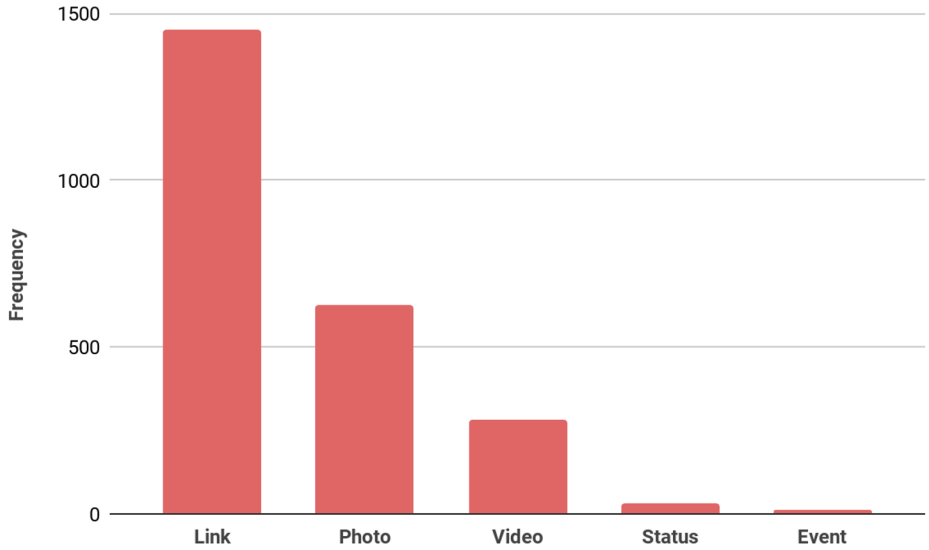


Monthly Post Frequency



Media Type and Body Text

Total Posts: Media Type



Body Text/Hashtag Summary

Text Length (characters):

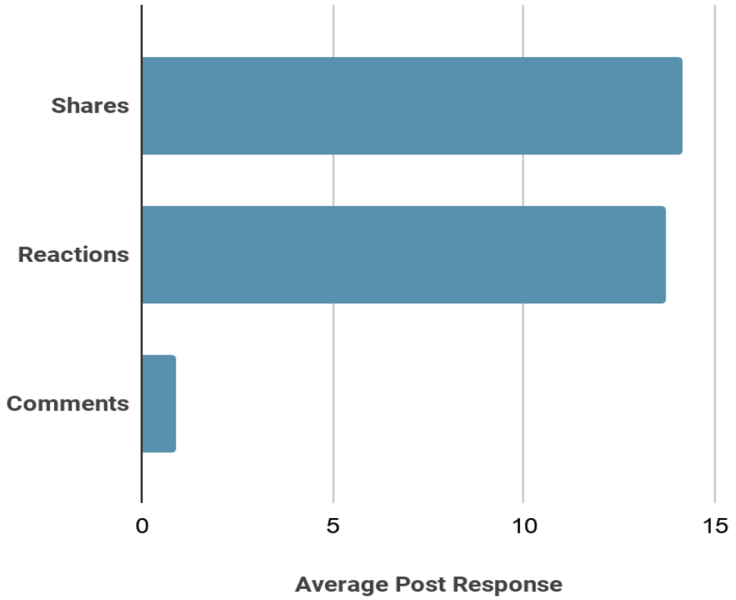
- 192 - mean ~ 30 words
- 182 - standard deviation

Hashtag Usage:

- 0.85 - mean
- 1.80 - standard deviations

Post Outcome Breakdown

Average Response to Posts



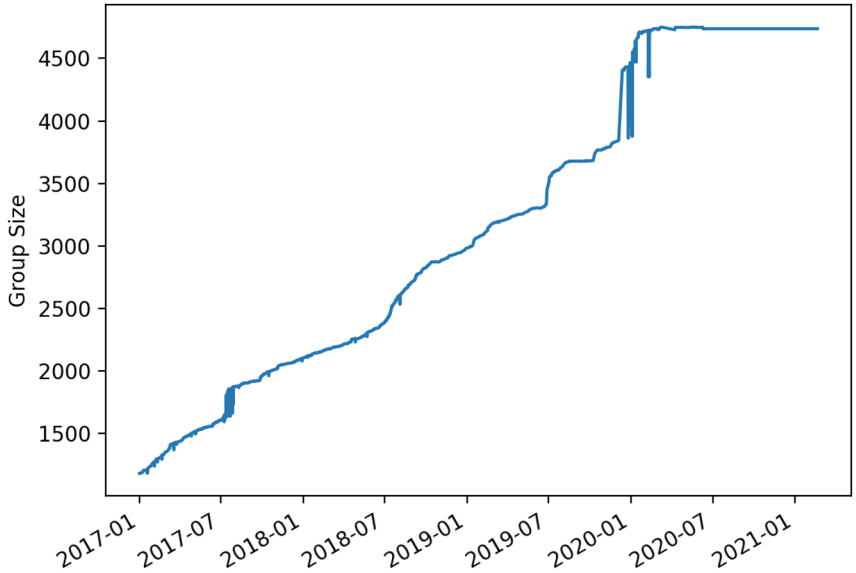
Engagement Summary

Total Engagement:
(Shares+Reactions+Comments)

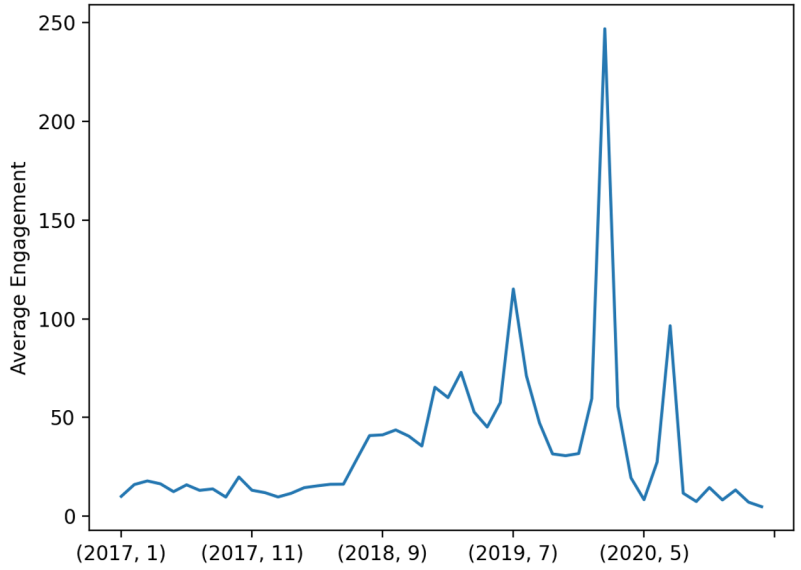
- 29 - mean
- 100 - standard deviation
- 12 - median
- 4057 - best post

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

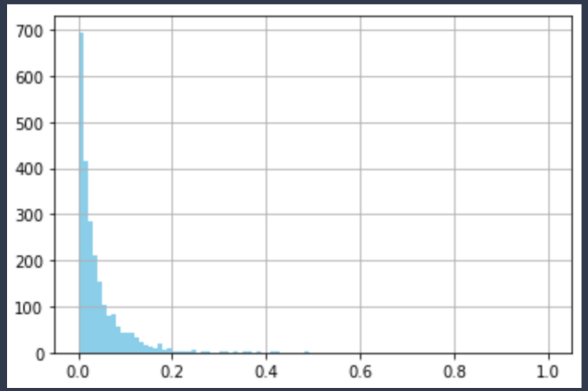
(1)
Weight engagement

(2)
Divide by Group Size

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

Reactions	×	**1**
Comments	×	**10**
Shares	×	**50**

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



Agenda

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

Information Flow

Train Models

1

Input features & post score into model

Extract Features

2

Mathematically derive significant features

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	Negative 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 matching

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 counting with REGEX
    b. Extract variable Length counting characters
    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
2. Post Scoring: We value a share at 50x that of a reaction and a comment at 10x that of a reaction. This is based on how many people on average it reaches. This is based on Facebook statistics. https://www.brandwatch.com/blog/facebook-statistics/#:~:text=Worldwide%2C%2026.3%25%20of%20the%20online,photos%20on%20that%20person's%20account.

	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	2400.000000
mean	0.089467	0.026523	0.549440	0.044679
std	0.085092	0.056298	0.272513	0.070369
min	0.000000	0.000000	0.000000	0.000000
25%	0.036662	0.000000	0.351188	0.007426
50%	0.078678	0.000000	0.500201	0.023578
75%	0.119181	0.031250	0.801513	0.053947
max	1.000000	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|)	
(Intercept)	0.0267729	0.0075510	3.546	0.000399	***
spring	-0.0010880	0.0041033	-0.265	0.790919	
summer	0.0083927	0.0038705	2.168	0.030228	*
fall	-0.0068928	0.0042195	-1.634	0.102482	
winter	NA	NA	NA	NA	
sunday	-0.0021187	0.0057359	-0.369	0.711878	
morning	0.0046292	0.0039665	1.167	0.243305	
afternoon	-0.0004717	0.0042852	-0.110	0.912358	
night	NA	NA	NA	NA	
monday	0.0074337	0.0055984	1.328	0.184358	
tuesday	0.0056253	0.0053971	1.042	0.297381	
wednesday	0.0047575	0.0053075	0.896	0.370151	
thursday	0.0028226	0.0054666	0.516	0.605676	
friday	0.0090065	0.0055825	1.613	0.106804	
saturday	NA	NA	NA	NA	
norm.length	0.0144721	0.0182279	0.794	0.427300	
norm.sentiment	-0.0054779	0.0054602	-1.003	0.315847	
norm.hashtags	0.1034904	0.0286971	3.606	0.000317	***
event	-0.0293940	0.0224639	-1.308	0.190830	
link	0.0160778	0.0050464	3.186	0.001461	**
photo	0.0015604	0.0053675	0.291	0.771300	
status	-0.0085057	0.0132957	-0.640	0.522408	
video	NA	NA	NA	NA	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0697 on 2381 degrees of freedom
Multiple R-squared: 0.02641, Adjusted R-squared: 0.01905
F-statistic: 3.589 on 18 and 2381 DF, p-value: 4.633e-07

Appendix 2: Linear Regression

Appendix Contents

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

Our second model (Model 2) 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. RF in general also benefits from handling unbalanced data as well as having a low bias and moderate variance. This compliments our linear regression results well.

- 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.

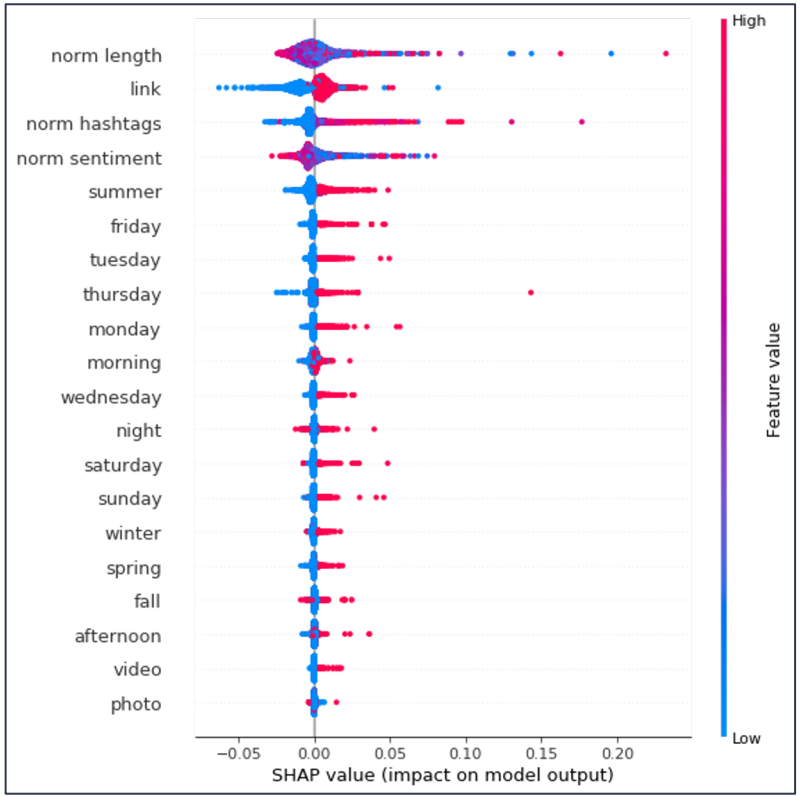
Our third and fourth models aimed to look at the problem from a different lens in order to:

- 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 (Model 3) 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 (Model 4). 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.

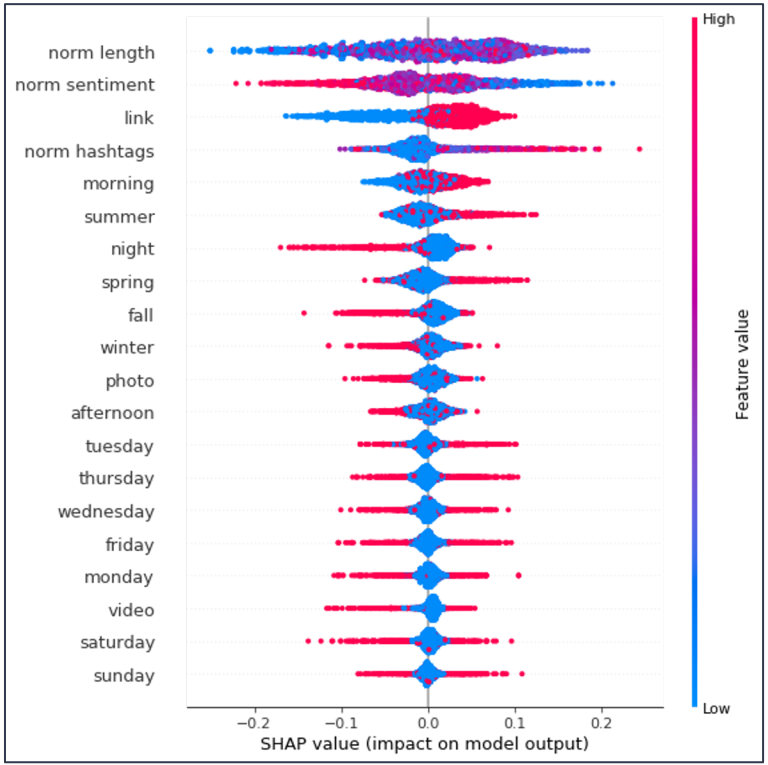
RFR Model 2: Feature Importance for High Post Scores

- 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

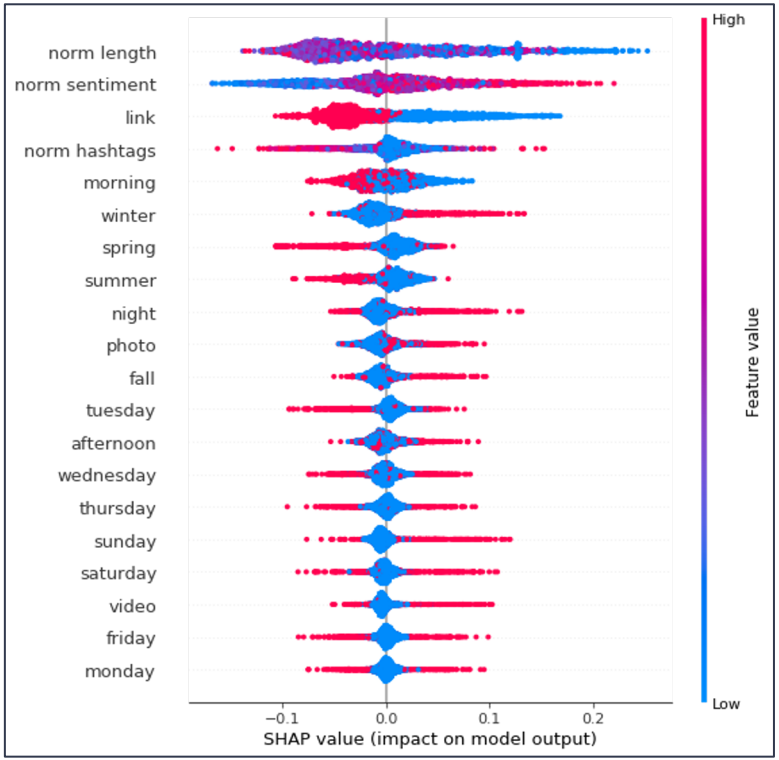


Note: This is feature importance for the classification Model 3 labeling “bad” and “good”. We are showing the features that are impactful for a “good” classification. Ex: Link is 3rd in importance and having a high(red) link aka (having a link) positively impacts the classification of “good”. In our next slide which is the classifier for “good” “medium” and “bad” you will see we’ve included feature importance for “bad” post and for “good” posts, their feature values look like opposites due to the fact that a feature that is harmful for predicting a “bad” post is beneficial when predicting a “good” post.

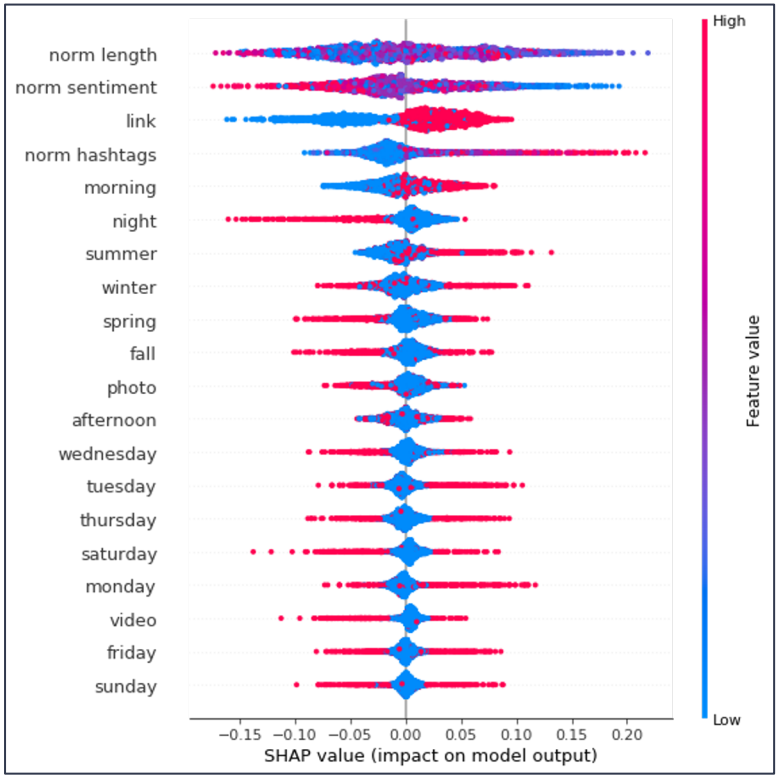
RFC Model 3: Feature importance for classifying “Good” posts



RFC Model 4: Feature importance for classifying "Bad" posts



RFC Model 4: Feature importance for classifying "Good" posts



Appendix 3d: Random Forest Classifier - Model 4

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: Our sentiment scoring is limited by VADER’s implementation of positive, neutral, and negative. Besides the obvious sentiment, when people read posts our language contains various complexities that are hard to capture in our feature set.

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 it's best to limit the biases listed above but again is never a perfect answer allowing us to say X causes Y.