**GLM Results**

Our hypothesis is that generating more retweets will help emphasize the seriousness of a situation while reaching the highest level of government and decision makers. In our attempt to predict retweet count, we started with a GLM model since the response value (retweets) was not continuous and was truncated at 0.

In this first run, we found the following to be significant predictors of retweet count (see appendix):

* isretweetedTRUE (+)
* length (-)
* isvirus (+)
* isemergency (-)
* isdeaths (-)
* iswho (+)
* iscdc (-)
* isnih (-)
* isdisease (-)
* isquarantine (-)
* isban (+)
* iscoronavirus (+)
* iscovid19 (-)
* iswash (-)
* isracist (-)
* ischinese (+)
* ch.confirmed (-)
* it.confirmed (-)
* ot.confirmed (+)
* us.confirmed (-)
* ch.deaths (+)
* it.deaths (+)
* ot.deaths (-)

Looking at the list above you can see that isretweetedTRUE, isvirus, iswho, isban, iscoronavirus, ischinese, ot.confirmed, ch.deaths and it.deaths all have positive effect to retweets. This tells me that a retweet will generate more retweets. As such, the goal should be to have a serious tweet retweeted as much as possible. This is very strong since each retweeted tweet generates about 6.546 more retweets compared to the original non-retweeted tweet. Additionally, using words that directly represent what is going on seems to also increase retweets. In our current coronavirus (covid19) crisis, using words virus (1.358), world health organization (2.319), ban (1.502), coronavirus (5.704), and China (4.201) in your tweets will help increase retweet count. It is a very powerful cycle, each additional retweet generates over 6 more retweets which in turn will generate 6 more retweets.

There were two marginally significant variables, ispandemic and isrecover (p-value = .09). It was surprising that the effect of using the word pandemic in a tweet did not have a higher significant value. What was even more surprising was the word infectious and infections (grouped together as isinfectious) was not even slightly significant.

In our model, we added in confirmed cases and death counts for China, Italy, US and all other countries (others). With the exception of US death count, these seem to have a significant effect on retweet count. We need to further explore if these confirmed cases and death count impact retweet count/reachability more than the content of the tweets.

lt was a good start to understanding how tweets can help prepare countries from a crisis. We believe retweet count can explain reachability. However, we need to further check the predictors/x variables for collinearity, apply variable selection or shrink methods, and implement cross-validations.

**Appendix**

**Generalized Linear Models (GLM)**

The following predictors were significant in this first model:

* isretweetedTrue - On average, holding everything else constant, retweeted tweets (TRUE) increases retweet total by 6.546 when compared to non-retweeted tweets (FALSE = ref).
* length - On average, holding everything else constant, an increase in tweet length decreases retweets total by -5.138 (longer tweets generate less retweet and reachability).
* isvirus - On average, holding everything else constant, an increase in usage of the word “virus” increases the retweet total by 1.358.
* isemergency - On average, holding everything else constant, an increase in usage of the word “emergency” decreases the retweet total by -4.448.
* isdeaths - On average, holding everything else constant, an increase in usage of the word “dead” or “deaths decreases the retweet total by -3.387.
* iswho - On average, holding everything else constant, an increase in usage of the word “WHO” or “World Health Organization” increases the retweet total by 2.319.
* iscdc - On average, holding everything else constant, an increase in usage of the word “CDC” or “Centers for Disease Control” decreases the retweet total by -3.676.
* isnih - On average, holding everything else constant, an increase in usage of the word “NIH” or “National Institutes of Health” decreases the retweet total by -6.797.
* isdisease - On average, holding everything else constant, an increase in usage of the word “disease” decreases the retweet total by -4.519.
* isquarantine - On average, holding everything else constant, an increase in usage of the word “quarantine” decreases the retweet total by -3.845.
* isban - On average, holding everything else constant, an increase in usage of the word “ban” increases the retweet total by 1.502.
* iscoronavirus - On average, holding everything else constant, an increase in usage of the word “coronavirus” increases the retweet total by 5.704.
* iscovid19 - On average, holding everything else constant, an increase in usage of the word “covid19” decreases the retweet total by -4.556.
* iswash - On average, holding everything else constant, an increase in usage of the word “quarantine” decreases the retweet total by -2.697.
* isracist - On average, holding everything else constant, an increase in usage of the word “racist” or “racism” decreases the retweet total by -8.455.
* ischinese - On average, holding everything else constant, an increase in usage of the word “chinese” or “china” increases the retweet total by 4.201.
* ch.confirmed - On average, holding everything else constant, when confirmed cases in China increases by 1, retweet total decreases by -1.806.
* it.confirmed - On average, holding everything else constant, when confirmed cases in Italy increases by 1, retweet total decreases by -1.475.
* ot.confirmed - On average, holding everything else constant, when confirmed cases in others (all other countries not China, Italy, or US) increases by 1, retweet total increases by 1.124.
* us.confirmed - On average, holding everything else constant, when confirmed cases in US increases by 1, retweet total decreases by -5.654.
* ch.deaths - On average, holding everything else constant, when the number of deaths in China increases by 1, retweet total increases by 4.929.
* it.deaths - On average, holding everything else constant, when the number of deaths in Italy increases by 1, retweet total increases by 3.917.
* ot.deaths - On average, holding everything else constant, when the number of deaths in others (all other countries not China, Italy, or US) increases by 1, retweet total increases by -6.725.

The following predictors were marginally significant (not really significant at p-value ≤ .05 but still show effect) in this first model:

* ispandemic (p-value = 0.0952) - On average, holding everything else constant, an increase in usage of the word “ispandemic” decreases the retweet total by -1.525.
* isrecover (p-value = 0.0933) - On average, holding everything else constant, an increase in usage of the word “recover” decreases the retweet total by -3.071.

Call:

glm(formula = retweets ~ isretweeted + length + favorites + ishealth +

ispandemic + isvirus + isemergency + isdeaths + iswho + iscdc +

isnih + isdisease + isquarantine + isrecover + isban + iscoronavirus +

iscovid19 + iswash + isracist + isasian + ischinese + isinfectious +

ch.confirmed + it.confirmed + ot.confirmed + us.confirmed +

ch.deaths + it.deaths + ot.deaths + us.deaths, data = covid19current)

Deviance Residuals:

Min 1Q Median 3Q Max

-139722 -7264 -2056 2839 187104

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.299e+07 5.800e+05 22.387 < 2e-16 \*\*\*

isretweetedTRUE 6.546e+03 3.160e+02 20.713 < 2e-16 \*\*\*

length -5.138e+01 6.945e+00 -7.398 1.40e-13 \*\*\*

favorites 8.145e-01 4.204e+00 0.194 0.846379

ishealth -4.176e+02 5.016e+02 -0.833 0.405118

ispandemic -1.525e+03 9.141e+02 -1.669 0.095180 .

isvirus 1.358e+03 3.725e+02 3.646 0.000267 \*\*\*

isemergency -4.448e+03 1.301e+03 -3.419 0.000630 \*\*\*

isdeaths -3.387e+03 8.674e+02 -3.905 9.46e-05 \*\*\*

iswho 2.319e+04 5.130e+02 45.218 < 2e-16 \*\*\*

iscdc -3.676e+03 1.008e+03 -3.645 0.000268 \*\*\*

isnih -6.797e+03 2.430e+03 -2.797 0.005167 \*\*

isdisease -4.519e+03 1.164e+03 -3.883 0.000103 \*\*\*

isquarantine -3.845e+03 8.403e+02 -4.575 4.77e-06 \*\*\*

isrecover -3.071e+03 1.830e+03 -1.678 0.093328 .

isban 1.502e+03 7.469e+02 2.011 0.044285 \*

iscoronavirus 5.704e+03 4.150e+02 13.744 < 2e-16 \*\*\*

iscovid19 -4.556e+03 2.076e+02 -21.950 < 2e-16 \*\*\*

iswash -2.697e+03 8.856e+02 -3.046 0.002322 \*\*

isracist -8.455e+03 3.713e+03 -2.277 0.022795 \*

isasian -3.421e+03 2.978e+03 -1.149 0.250661

ischinese 4.201e+03 6.736e+02 6.237 4.51e-10 \*\*\*

isinfectious -2.736e+03 2.172e+03 -1.259 0.207878

ch.confirmed -1.806e+02 8.150e+00 -22.157 < 2e-16 \*\*\*

it.confirmed -1.475e+01 6.582e-01 -22.414 < 2e-16 \*\*\*

ot.confirmed 1.124e+01 3.724e-01 30.190 < 2e-16 \*\*\*

us.confirmed -5.654e+01 4.096e+00 -13.802 < 2e-16 \*\*\*

ch.deaths 4.929e+02 2.713e+01 18.167 < 2e-16 \*\*\*

it.deaths 3.917e+01 5.369e+00 7.296 3.02e-13 \*\*\*

ot.deaths -6.725e+01 1.311e+01 -5.131 2.89e-07 \*\*\*

us.deaths -2.440e+02 1.523e+02 -1.603 0.109003

---

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

(Dispersion parameter for gaussian family taken to be 458727180)

Null deviance: 2.3711e+13 on 41999 degrees of freedom

Residual deviance: 1.9252e+13 on 41969 degrees of freedom

AIC: 956870

Number of Fisher Scoring iterations: 2

**Regression Tree**

Regression tree:

tree(formula = retweets ~ isretweeted + length + favorites +

ishealth + ispandemic + isvirus + isemergency + isdeaths +

iswho + iscdc + isnih + isdisease + isquarantine + isrecover +

isban + iscoronavirus + iscovid19 + iswash + isracist + isasian +

ischinese + isinfectious + ch.confirmed + it.confirmed +

ot.confirmed + us.confirmed + ch.deaths + it.deaths + ot.deaths +

us.deaths, data = covid19current)

