NYC Covid-19 Cases and Twitter

Eric Cusick

Business Case

Using Twint to filter through Covid related tweets from New York City to see if the message of a verified users affects the Covid infection/cases rate in NYC.

The goal is to use NLP and Random Forest Regressor modeling to see any words that have a higher correlation with new Covid cases.

Twint API

Twint API is used to gather all of the tweets relating to Covid in NYC.

Input parameters to obtain the desire tweets:

- Location : New York City
- Date : 01/01/2020 - 02/14/2021
- Verified Users
- Likes Count greater than 50

All desire tweets were compile into a csv file named ‘covidtweets.csv’

The follow search words were used to obtain the tweets:

- Covid
- Corona
- Coronavirus
- Mask
- Vaccine
- Quarantine

NLP

- Approximate 3000 tweets were obtained from using Twint API
- Removed hashtag, url links, @user, and digits
- Tokenization
- Stopwords
- Lemmatization
- Count Vectorizer

Before Cleaning

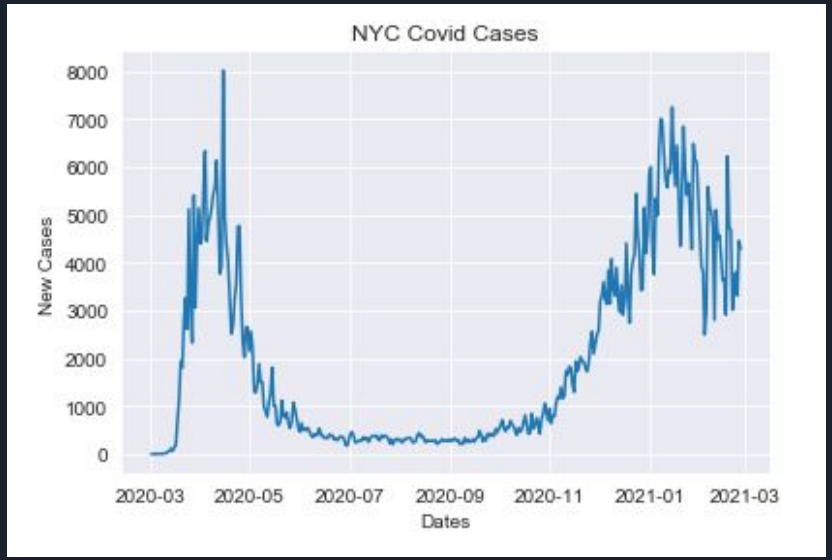
```
'Here’s a fun video by Newark’s own @DJLILMAN973 Ft. our Mayor @rasjbaraka reminding all of Newark to Mask Up. #MaskUpNewark  h
ttps://t.co/2BuHyG7KCD'
```

After Cleaning

```
'Heres fun video Newarks Ft Mayor remind Newark Mask Up MaskUpNewark'
```

NYC Cases EDA

NYC Covid Dataset were obtained and modify from the NYTime's Github page where there is ongoing reports of Covid.



Modeling Process Random Forest Regressor

Created two dataset for modeling, original and likes count.

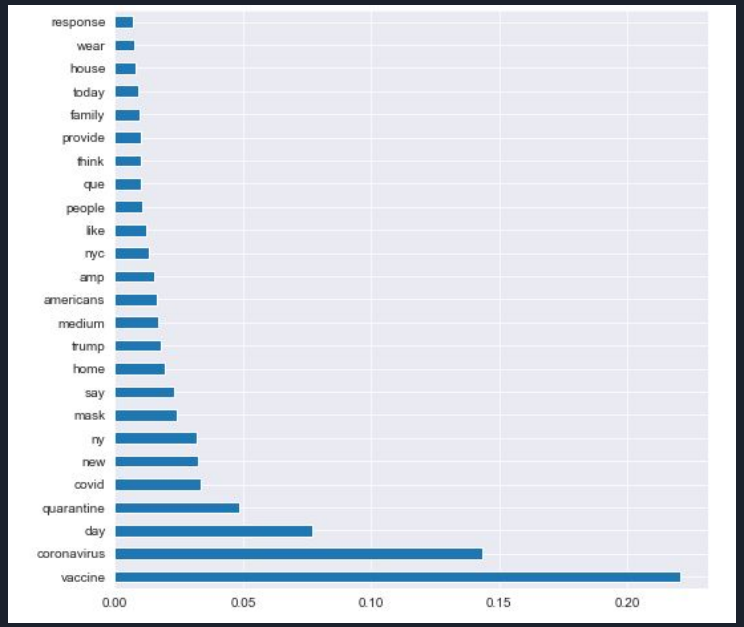
- Original dataset is there as a control model taking in a combine tweets for each individual days to correlate with NYC Covid cases

- The likes count dataset is the same as original but with one exception. The number of likes that a tweet has is the value of those words.

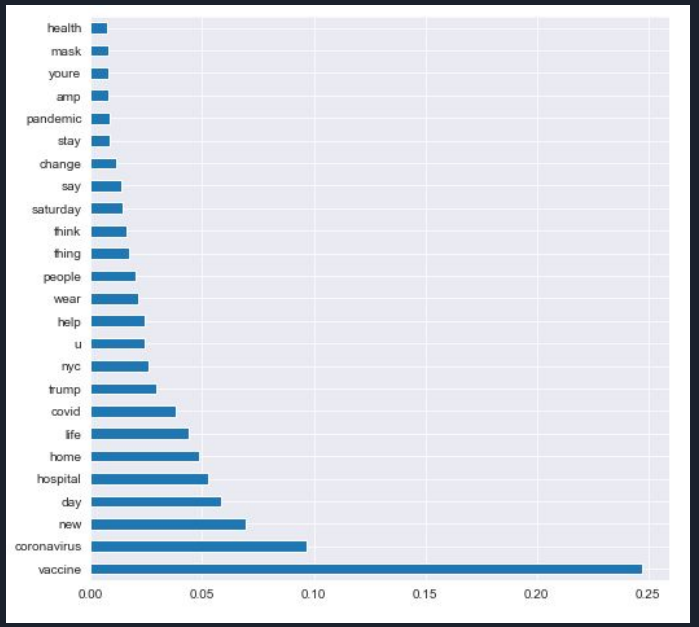
Final Model

Feature Importances of both Random Forest Regressor Models

Original Dataset



Likes Count Dataset



Final Model Cont.

Original:

- GridSearch Training: 30.13%
- Training accuracy: 0.63
- Test accuracy: 0.31
- RMSE: 2426200

Likes:

- GridSearch Training: 22.90%
- Training accuracy: 0.61
- Testing accuracy: 0.24
- RMSE: 2679992

These are the results for the Models.

The Original Dataset perform better than the likes count dataset, however both will need further work to improve the overall accuracy.

Recommendation

After looking at the features importance graphs, there is a lot of health based words and some political words as well that correlate with new covid cases. The verified user should be careful when it comes to writing and phrasing a tweets. As they could influence the behaviors of their followers.

Future Work

Some further work may include:

- Obtaining more tweets
- Further cleaning of NLP
  - Including more stopwords
  - Getting rid of Keywords used for gathering tweets
  - Placing language filter
- Create a bi-gram or n-gram datasets

Thank You

Thank for your time and attention to the presentation today.

Github link: https://github.com/Ericusick/Twint_Covid_NYC