CUNY DATA 698: Master’s Research Project

Comparing Time Series Intelligence Methods and Supervised Learning Models Using Social & Political Indicators to Predict NYC Hate Crimes

Introduction

Throughout the pandemic, there has been a focus on New York City hate crime incidents that world watched on nightly news channels. These incidents gave light to the division in the city known as the "melting pot". In particular, the Asian and Jewish communities saw their neighbors being targeted unprovokedly.

In this study, the author will capture New York City social and political indicators by using proxy data to find relationships between the New York City environment and hate crimes. Most of the data used in the study are aggregated numbers using accessible data; however, one independent variable is produced using sentiment analysis. The author uses the NRC Word-Emotion Association Lexicon to capture the rolling seven-day sentiment of Donald Trump's tweets.

After gathering the dataset, the study will perform basic analysis of the independent variables and dependent variables. Next, the dataset will be pushed into supervised learning models to find variable importance. Lastly, the study will create forecasts using time series intelligence methods to predict the number of weekly New York City hate crimes incidents. The study will conclude by comparing the supervised learning models and the time series intelligence methods by looking at the predicted errors.

Hypothesis

The author uses New York City social and political indicators to create a supervised learning model used to predicting New York City hate crimes. Moreover, the author believes that quantifying Donald Trump's tweets as a political indicator will be the most significant independent variable when explaining the variance of New York City hate crimes. Next, the author believes that the independent variable gathered will perform with a higher accuracy of predicting New York City hate crimes than the best performing time series intelligence method.

The models' Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Lastly, the author believes that a either a supervised model or a time series intelligence model can be used to predict New York City hate crimes.

Data & Variables

This study exercises data from the NYPD Hate Crimes report, Donald Trump Tweets, NYC Labor Market, Department of Homeless Services Daily report, S&P stock price, and NYC Weather.

The NYPD Hate Crime report gathers information on confirmed hate crime incidents in New York City beginning January 2019. On January 8, 2021, Twitter permanently suspends Donald Trump's twitter account following the Capitol Insurrection due to the risk of further incitement of violence. These two datasets set the lower and upper bound of the timeframe of the complete dataset used.

The dependent variable in the dataset is the daily NYC hate crime count. The independent variables used for supervised modelling are the daily count of Donald Trump's tweets, rolling seven-day NRC sentiment score, daily temperature, the daily total of individuals in the NYC shelter system, daily closing S&P stock price, and monthly unemployment rate. These variables are an attempt to capture the Social and Political indicators during January 2019 to January 2021.

Literature Review

Hate Crimes

Hate Crimes are the committed criminal offense which are motivated, in whole or in part, by the offender's bias(es) against a race, religion, disability, sexual orientation, ethnicity, gender, and/or gender identity.[1] Hate crimes can be separated into four categories: thrill-seeking, defensive, retaliatory, and mission offenders.[6] Expanding into the defensive hate crime category, these occurrences include the offender group perceiving certain groups taking away their economic, social, or political power.[2]

The author uses indicators to simulate the change in the economic, social, and political environment. These indicators are transformed into a dataset and applied to three regression supervised models to predict the daily hate crime count. The results are compared to forecasting models to discover the best performing model.

NRC Emotion Lexicon

NRC Emotion Lexicon annotates words with Plutchik's eight basic emotions: joy, sadness, anger, fear, disgust, surprise, trust, and anticipation. The lexicon uses *Roget's Thesarus* as the source for terms and annotates words that occurred more than 120,000 times in the Google n-grams corpus. Each term is annotated by five different people. For 74.4% of instances, the annotators agreed on the emotion of the term. The remaining instances were decided on majority vote.[3]

The author of this study chooses the NRC Emotion Lexicon as its popularity and availability. Additionally, the lexicon allows the author to view Donald Trump's tweets into eight emotion buckets for the two-year period of this analysis.

A criticism of the NRC Emotion Lexicon is that words that should in most contexts be emotionally neutral are associated with emotional labels that are inaccurate. Examples of these scenarios are the words "lesbian", labeled as disgust and sadness, and "mountain" as anticipation.[4]

Principal Component Analysis

Principal component analysis (PCA) reduces the dimensionality of a dataset by linearly transforming the data into a new coordinate system where the variation in data can be explained in fewer dimensions.[5]

The author uses PCA as an exploratory tool to visually identify the clusters of nearby data points that is not seen otherwise.

Regression Models

Regression models are statistical processes for estimating relationships between the dependent and independent variables. These models are used to infer causal relationships and make prediction of the dependent variable.[7] The author uses Elastic Net and Neural Network models in addition to Random Forest model to predict the total hate crime count in New York City.

Statistical Methods

The author uses a combination of government and social data from January 2019 – January 2021 as these are the limits of the combined datasets used for this study. The complete list of variables is the following:

- Month
- Day
- Year
- Donald Trump tweet count – daily
- NRC Valence of Donald Trump tweet – daily
- Log of Total number of individuals in NYC shelters – daily
- New York City hate crime count – daily
- Log of S&P stock price – daily
- Unemployment Rate – monthly
- New York City average temperature – daily

Imputing missing data was done by using a naïve method, "downup". The approach sorts the data in chronological order then fills in the missing value by first looking "down' then "up". The author believes that the missing daily records can be most reflected by using the information surrounding the missing record.

All but one variable used is available using various government datasets. The rolling seven-day sentiment of Donald Trump tweet is done by capturing the tweet and performing several transformations to remove unwanted data. The preprocessing includes removing twitter handles, links, punctuations, and numbers. Furthermore, the author uses the stemming method to reduce words to its root form. Fig 1 shows a word cloud which visualizes the frequency of cleaned up dataset. Finally, the author uses the NRC Word-Emotion Association Lexicon to calculate the rolling seven-day average valence. Fig 2 highlights the of emotions by word in the

Donald Trump's tweets. Fig 3 shows the rolling seven-day mean of the tweet's valence. Valence is calculated by the sum of negative and positive emotions.

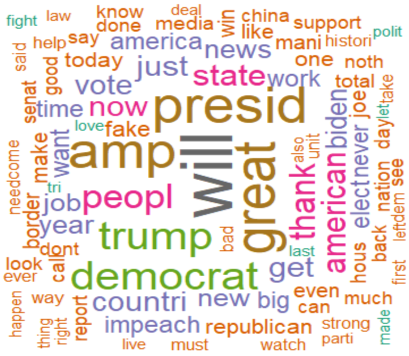

Fig 1

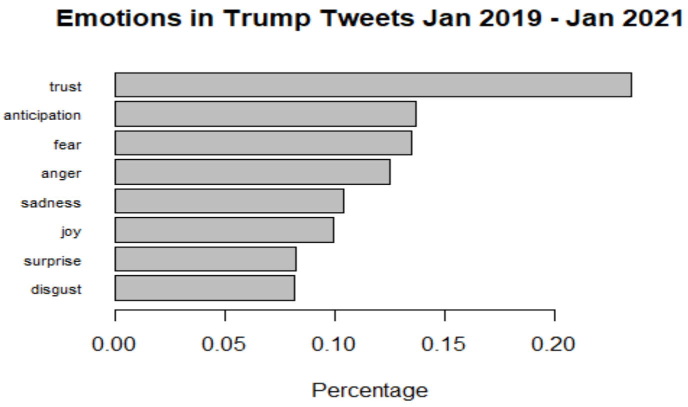

Fig 2

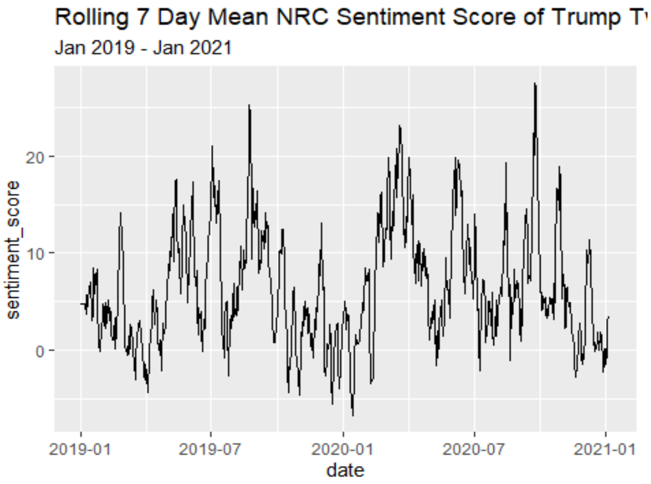

Fig 3

After the gathering of the desired data, we look at the distribution of the variables that will be used for creating a supervised learning model. Fig 4 shows the variables with a clear lower bound of zero are skewed right with no upper bound limit. An additional point to note is the bimodal distribution of the total number of individuals in New York City shelter (log).

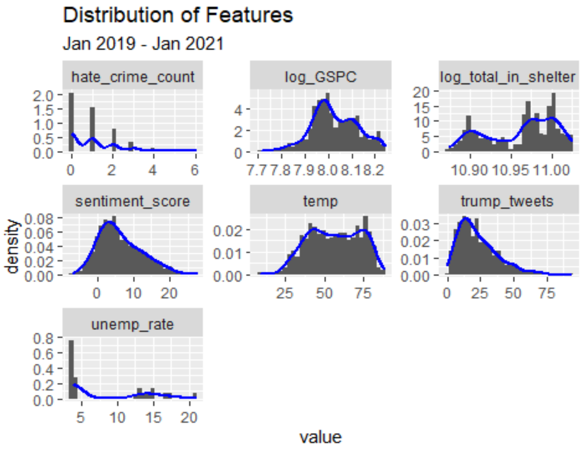

Fig 4

Next, the author creates a correlation plot to analyze the between variables and more importantly if there are any noticeably correlated variables to the dependent variable, hate crime count. Fig 5 shows that the highest correlated variables to the dependent are the total number of individuals and the unemployment rate with an absolute value of 0.2.

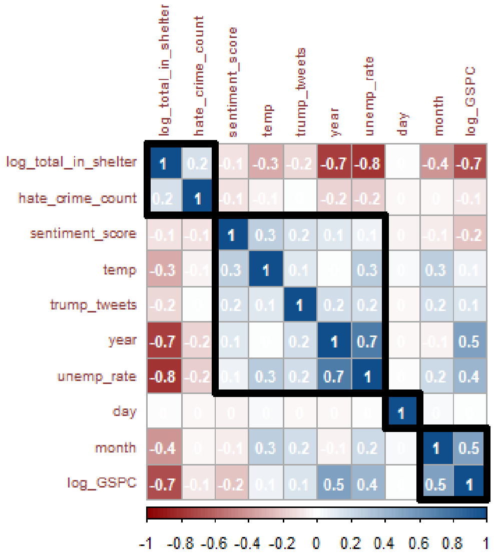

Fig 5

In final analysis of the dataset, the author uses the unsupervised learning model. Fig 6 shows the PCA analysis that identifies the principal components explaining 76.4% of the variance in the dataset. Note that the count of Donald Trump tweets is on the outskirts of the PCA variables.

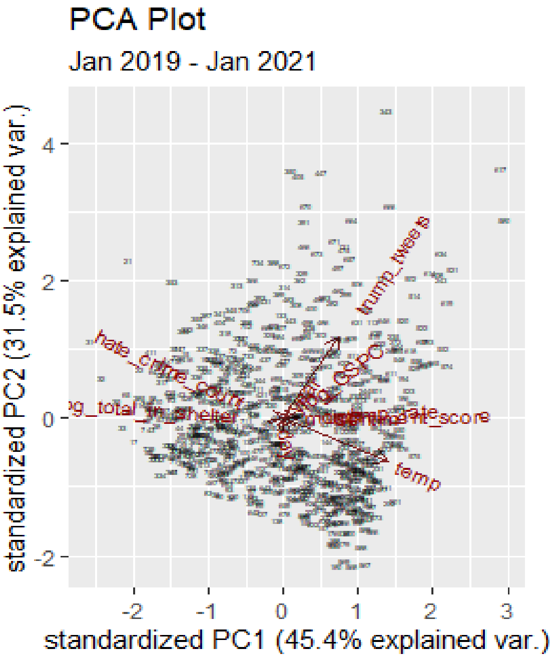

Fig 6

Supervised Models

The author splits the training and test data into a 75/25 split. Next, the independent and dependent will be identified as X and y, respectively. The datasets are then used on three regression models to predict daily New York City hate crimes. Lastly, the variable importance, R-squared, RMSE, and MAE is examined for each supervised model.

The first model used is the Elastic Net model. The variables are preprocessed to center and scale the variables used in the model, below shows the best performing tuned model using cross validation. The top three important variables are the total number of individuals in New York City shelters, New York City Unemployment Rate, and the year of the hate crime incident.

```
## Elasticnet
##
## 555 samples
##   9 predictor
##
## Pre-processing: centered (9), scaled (9)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 500, 499, 499, 499, 500, 499, ...
## Resampling results across tuning parameters:
##
##   lambda  fraction  RMSE      Rsquared    MAE
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were fraction = 0.6 and lambda = 0.1.
```

```
## Variable Importance
```

```
##                         Overall
## log_total_in_shelter 100.000000
## unemp_rate            83.088354
## year                  64.387450
## sentiment_score       43.676777
## log_GSPC              23.391869
## temp                  17.462245
## day                    2.806218
## month                  1.027059
## trump_tweets           0.000000
```

The second model used is the Random Forest model. The variables are preprocessed to center and scale the variables used in the model, below shows the best performing tuned model using cross validation. The top three important variables are the total number of individuals in New York City shelters, New York City temperature, and the rolling seven-day sentiment of Donald Trump's tweets.

```
## Random Forest
##
## 555 samples
##   9 predictor
##
## Pre-processing: centered (9), scaled (9)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 499, 500, 499, 499, 501, 500, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared    MAE
##   2     1.187267  0.04035153  0.9228649
##   5     1.194886  0.04144920  0.9340909
##   9     1.201943  0.04090473  0.9421778
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
```

```
## Variable Importance
##                         Overall
## log_total_in_shelter 100.00000
## temp                  92.73580
## sentiment_score       81.95835
## log_GSPC              80.06047
## trump_tweets          73.80812
## day                   69.75111
## month                 23.31624
## unemp_rate            19.99008
## year                   0.00000
```

The third model used is the Neural Network model. The variables are preprocessed to center and scale the variables used in the model, below shows the best performing tuned model using bootstrapping. The top three important variables are the total number of individuals in New York City shelters, New York City Unemployment Rate, and year.

```
## Neural Network

## Model Averaged Neural Network
##
## 555 samples
##   9 predictor
## Pre-processing: centered (9), scaled (9)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 555, 555, 555, 555, 555, 555, ...
## Resampling results across tuning parameters:
##
##   0.01    1    1.144454  0.029370638  0.8871494
## The final values used for the model were size = 1, decay = 0.01 and bag = FALSE.

##                          Overall
## log_total_in_shelter 100.000000
## unemp_rate            83.088354
## year                  64.387450
## sentiment_score       43.676777
## log_GSPC              23.391869
## temp                  17.462245
## day                    2.806218
## month                  1.027059
## trump_tweets           0.000000
```

A couple of points to note in using the supervised models is that the social indicators top the list of variable importance in each models used. Another point to make is that Donald Trump tweets, political indicator, play a much lesser role in each model. More specifically, the count of Donald Trump tweets plays the least amount of importance in two of the three models. The last point to make is that the R-squared values are very low, highlighting that the independent variables do not explain the variance of the dependent variable.

Forecasting

Fig 7 shows the weekly count of New York City hate crimes. The ACF plot shows that there is a trend as the plot shows positive values that slowly decrease as the lags increase. Additionally, there seasonal plot shows that there is potentially seasonal trends in hate crimes.

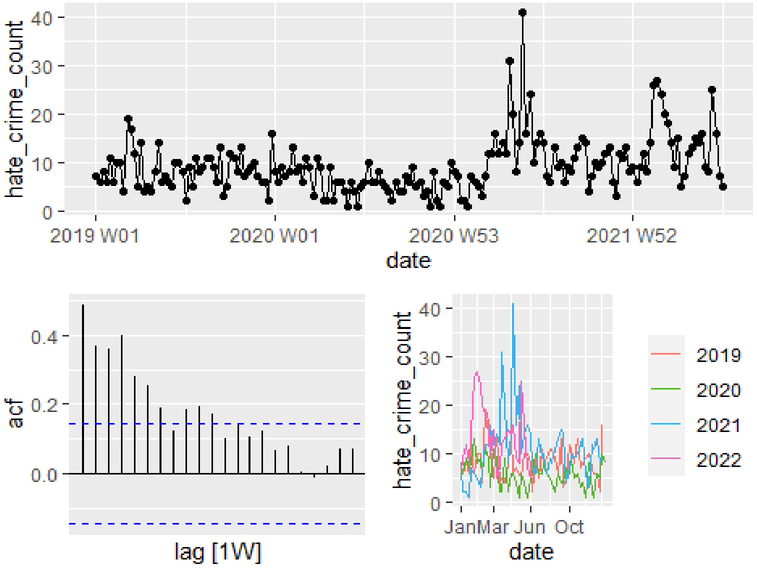

Fig 7

The author uses Box Cox transformation of the weekly count of New York City hate crimes before using the forecasting methods. Once done, the author uses the following forecasting methods to find the best performing model: ARIMA, Exponential Smoothing, Neural Network, and Stochastic. Fig 8 shows the six-month weekly forecast of the four models.

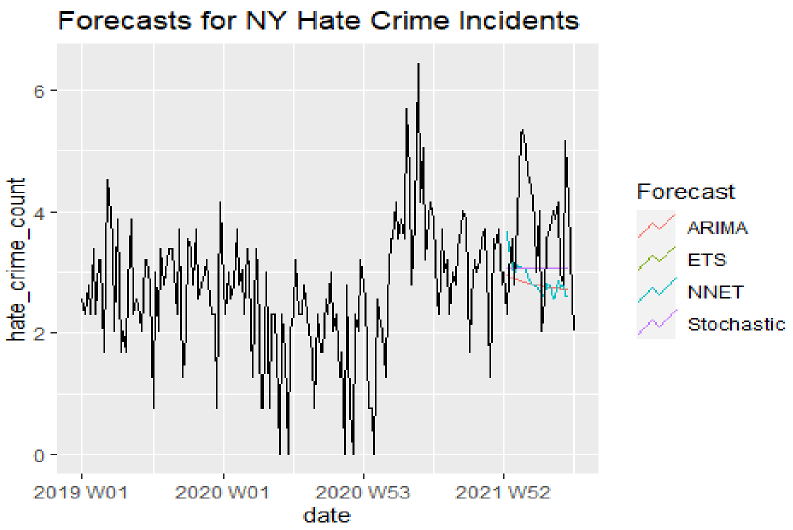

Fig 8

After the creating the models, the author reviews the accuracy of models on unused data.

```
## # A tibble: 4 x 10
##   .model     .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>      <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 ARIMA      Test  0.950  1.33 1.09   20.1  26.5 0.832 0.795 0.443
## 2 ETS        Test  0.698  1.15 0.943  13.0  23.5 0.718 0.690 0.446
## 3 NNET       Test  0.875  1.31 1.09   17.8  26.9 0.828 0.784 0.421
## 4 Stochastic Test  0.697  1.15 0.943  13.0  23.5 0.718 0.690 0.446
```

The Stochastic model performs the best when reviewing the RMSE and MAE. The author then examines the residuals and autocorrelation of the Stochastic model. Fig 9 shows that the residuals and autocorrelation are evenly distributed.

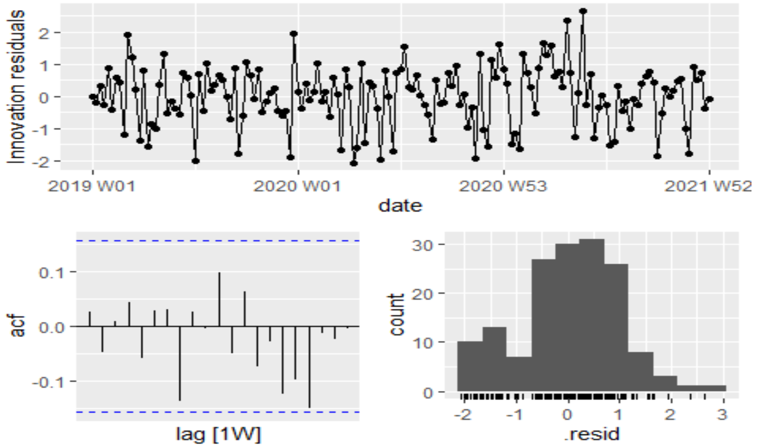

Fig 9

Lastly, the Stochastic model is applied to the future data.

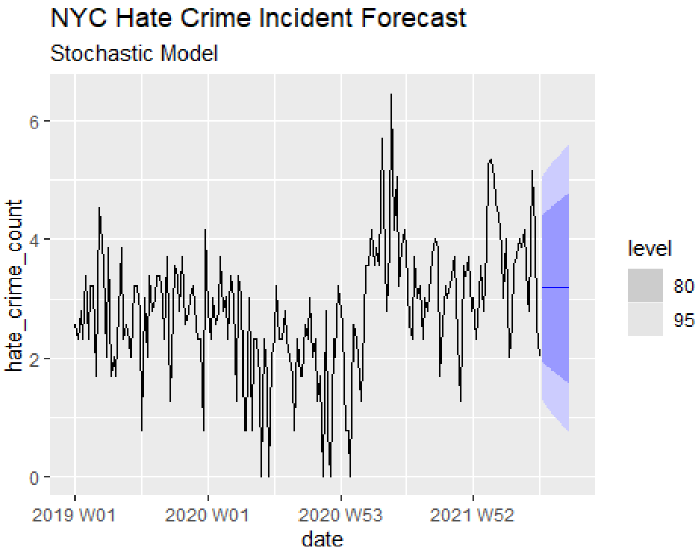

Fig 10

References

1. Federal Bureau of Investigation. Hate crimes. Washington, D.C.: U. S. Department of Justice. Retrieved from https://www.fbi.gov/about-us/investigate/civilrights/hate_crimes/overview

2. Gerstenfeld, P.B. *Hate crimes: Causes, controls, and controversies*. Thousand Oaks, CA: Sage.

3. Mohhammad, Sarif M.;Turney, Peter D. *NRC Emotion Lexicon*, National Research Council Canada.

4. Zad, Samira; Jimenez, Joshuan; Finlayson, Mark, A. *Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon*. Knight Foundation School of Computing and Information Sciences

5. Jolliffe, Ian T.; Cadima, Jorge. *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences

6. Burke, Daniel. *The four reasons people commit hate crimes*. CNN

7. Freedman , David A. *Statistical Models: Theory and Practice.* Cambridge University Press.