Identifying Major Electrical Disturbances in the U.S. Using Social Media Posts

Presented by:
Creighton Ashton
Meroe Yadollahi
Jack Wang

Problem Statement

- The traditional method to spot a power outage/electrical disturbance is to check the live feeds provided by major utility companies or the satellite data that capture the extent of light emitted at night.

- We will build a tool that identifies the major electrical disturbances using social media posts. Unlike the traditional methods, our tool will identify major electrical disturbances more timely.

Data Gathering and Cleaning

Twitter
Power Outage
Weather

“

Twitter

- Twitterscraper
  - github.com/taspinar/twitterscraper
- Scanned for keywords
  - Blackout, Power Outage, etc.
  - Every state in the U.S. for the last 5 years
  - 18,990 tweets
- Cleaned
  - Removed links
  - Tokenize tweets
  - Tried Portstemmer (poor results)
  - Formatted timestamp

“

Power Outage

- Energy.gov
  - www.oe.netl.doe.gov/OE417_annual_summary.aspx
- Combined 5 years of historical data
  - 1,325 total accounts
- Formatted date/time/location

“

Weather

- NOAA
  - https://www.ncdc.noaa.gov/data-access/severe-weather
- No shortage of data
  - 8 Key states with varied weather
    - CA, NY, OK, IL, FL, MI, NV, WA
- 53,718 entries
  - Formatted Date and location

“

Combining data

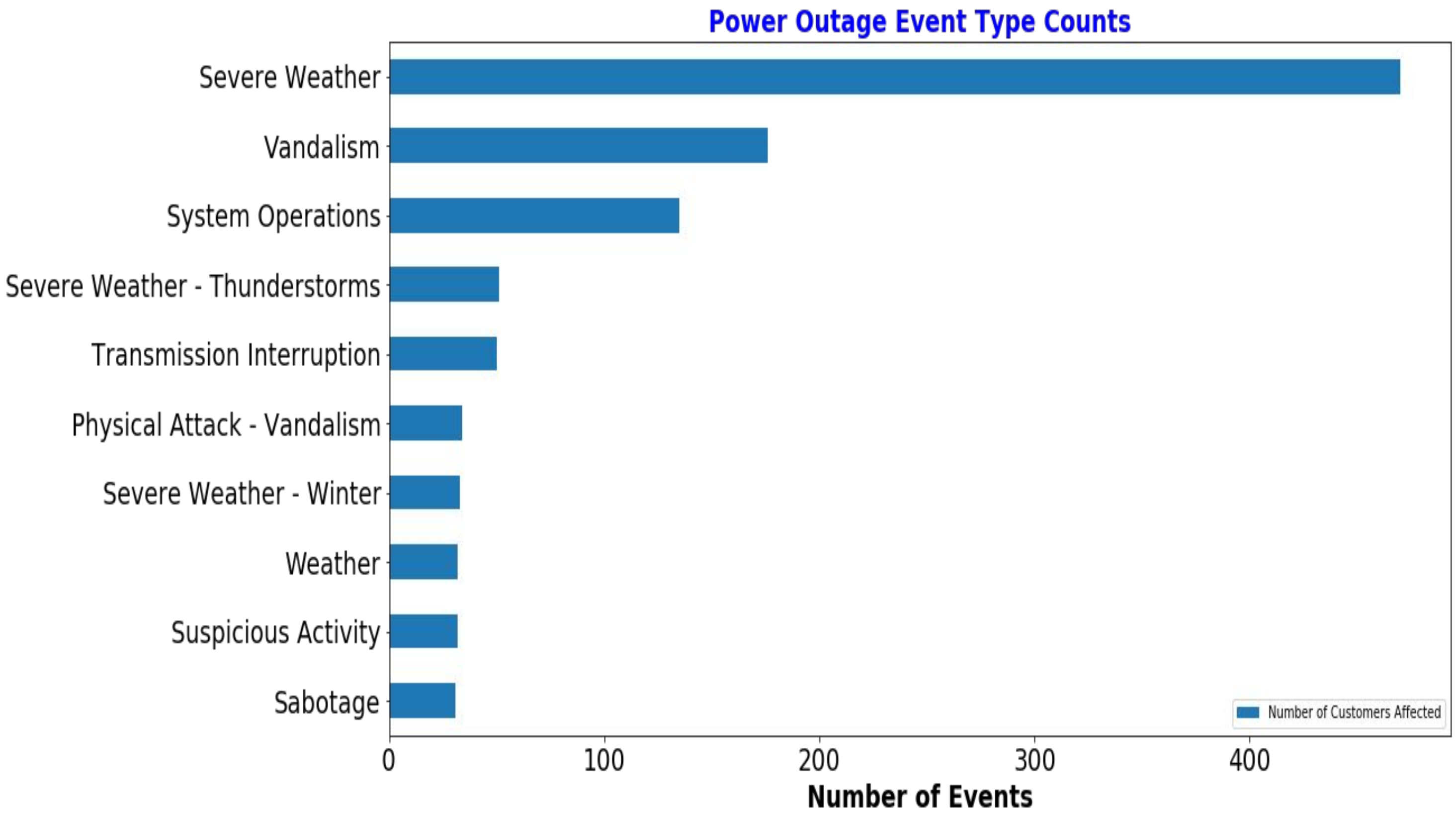
- Twitter and Power Outage
  - Created target column for twitter data
    - Checked if a tweet’s time and location was in the range of a power outage time frame in the same location
    - About 5% was the target class
- Power Outage and Weather
  - Merged tables on Date/Location
    - 5,205 entries for EDA

Exploratory Data Analysis

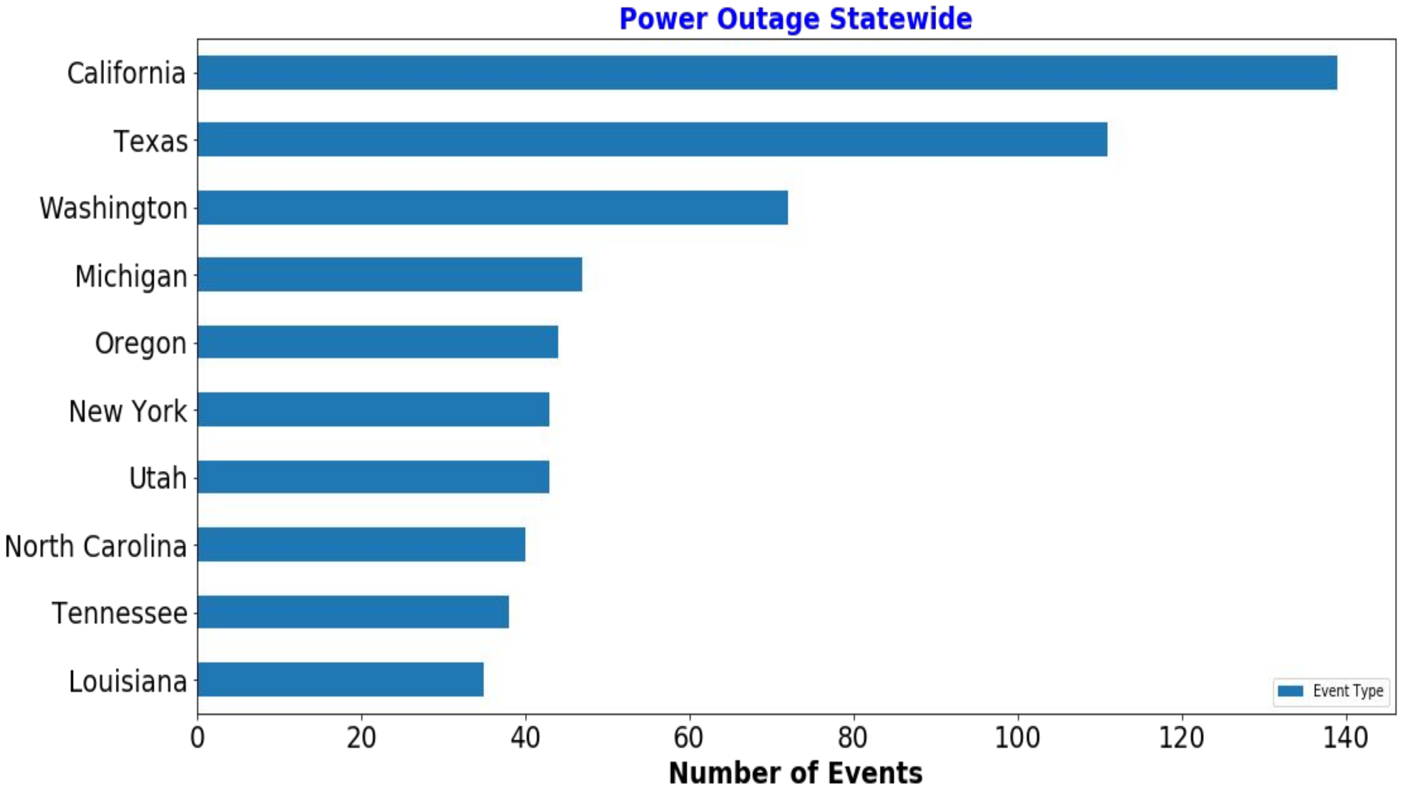
Power Outage

& Weather

Power Outage Event Type



Power Outage Events Per State



n = 139

Number of Events in California	
Vandalism	36
Severe Weather	23
System Operations	17

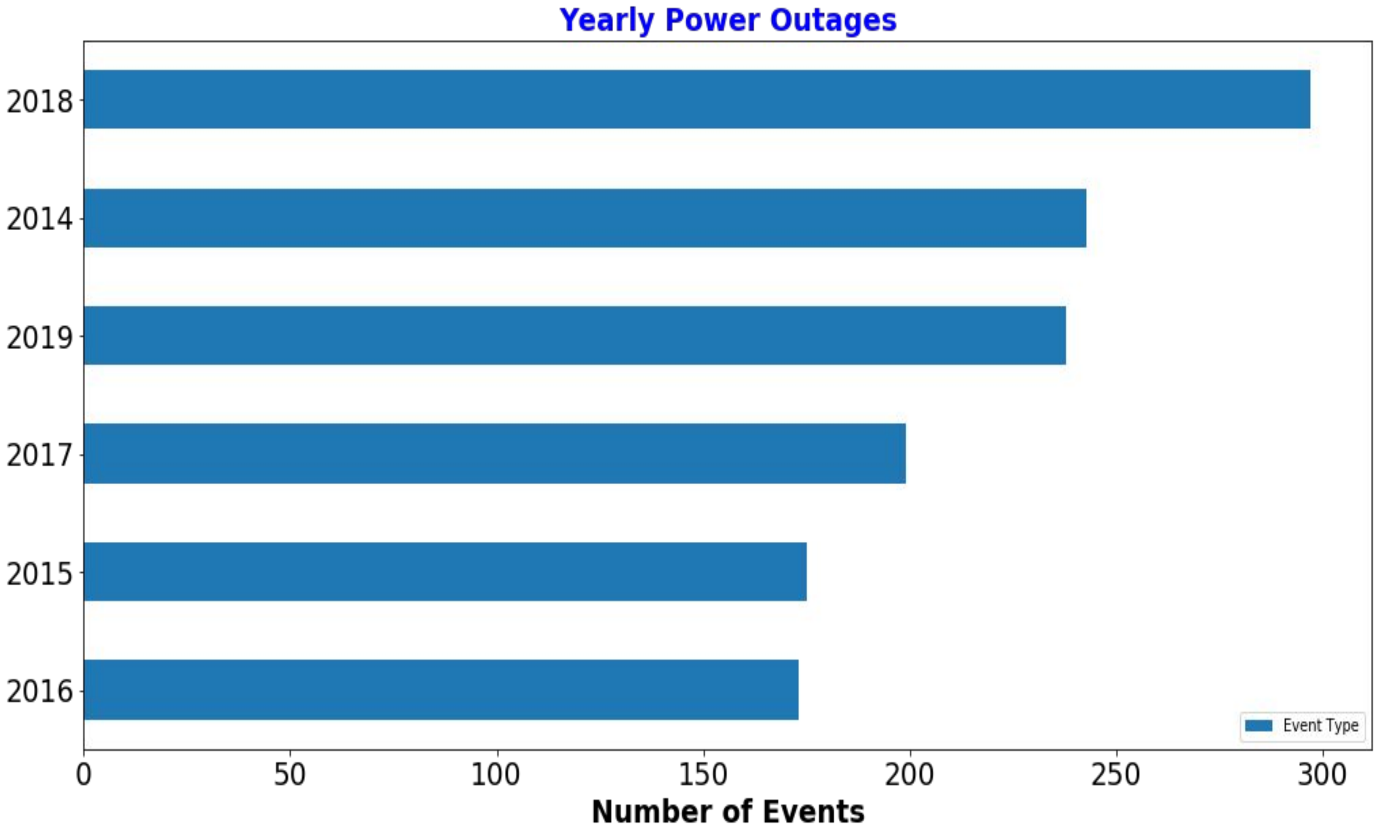
n = 111

Number of Events in Texas	
Severe Weather	48
System Operations	9
Transmission Interruption	6

n = 72

Number of Events in Washington	
Severe Weather	22
Vandalism	12
Transmission Interruption	7

Power Outage Per Year



n = 297

	Number of Events in Year 2018
Severe Weather	149
System Operations	55
Vandalism	46

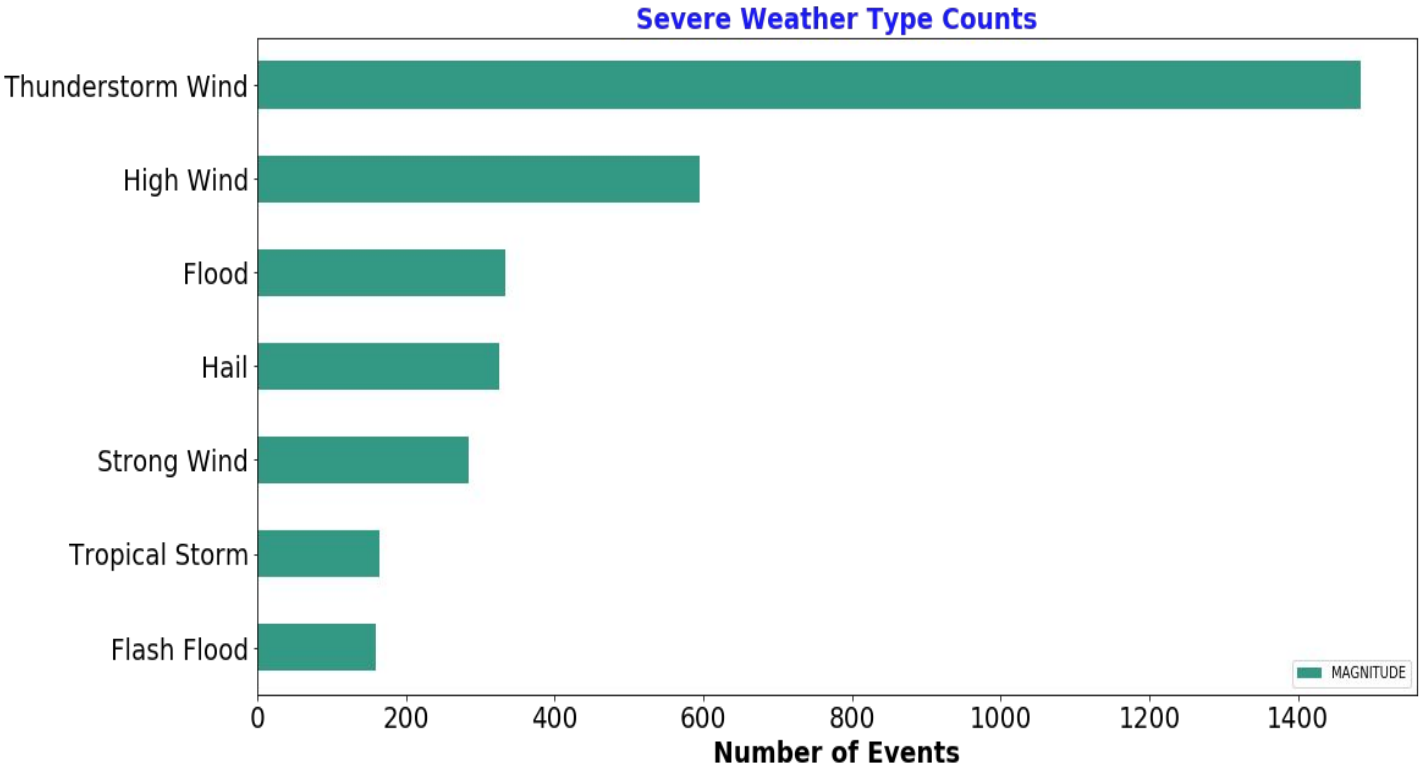
n = 243

	Number of Events in Year 2014
Severe Weather - Thunderstorms	50
Physical Attack - Vandalism	34
Fuel Supply Emergency - Coal	16

n = 238

	Number of Events in Year 2019
Severe Weather	78
Vandalism	51
System Operations	41

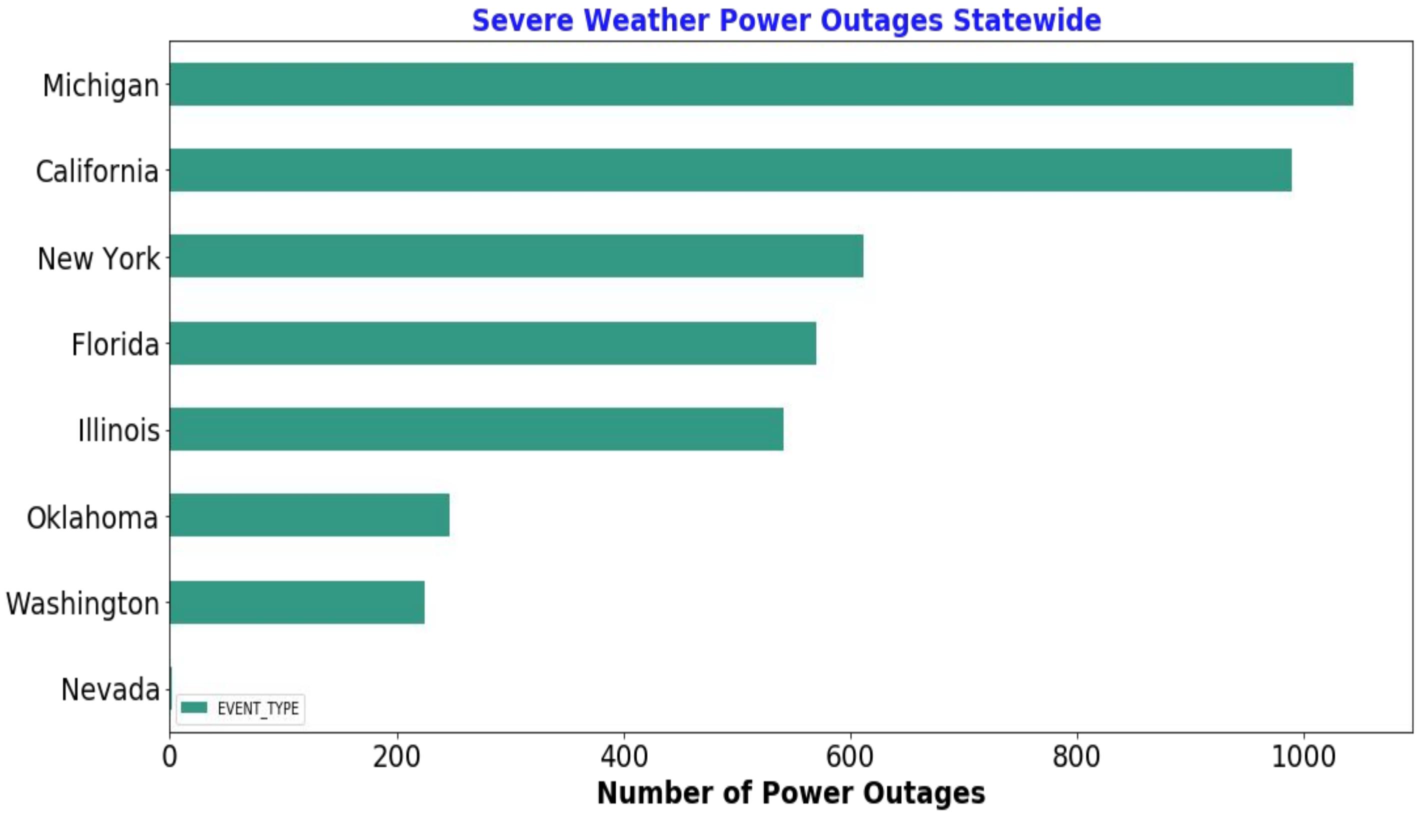
Severe Weather vs Power Outage Event Types



- Filtered the Merged Data by Severe Weather Event Type for Power Outages

n = 4229

Severe Weather vs Power Outage Statewide



n = 1043

Number of Events in Michigan	
Thunderstorm Wind	511
Hail	217
High Wind	135

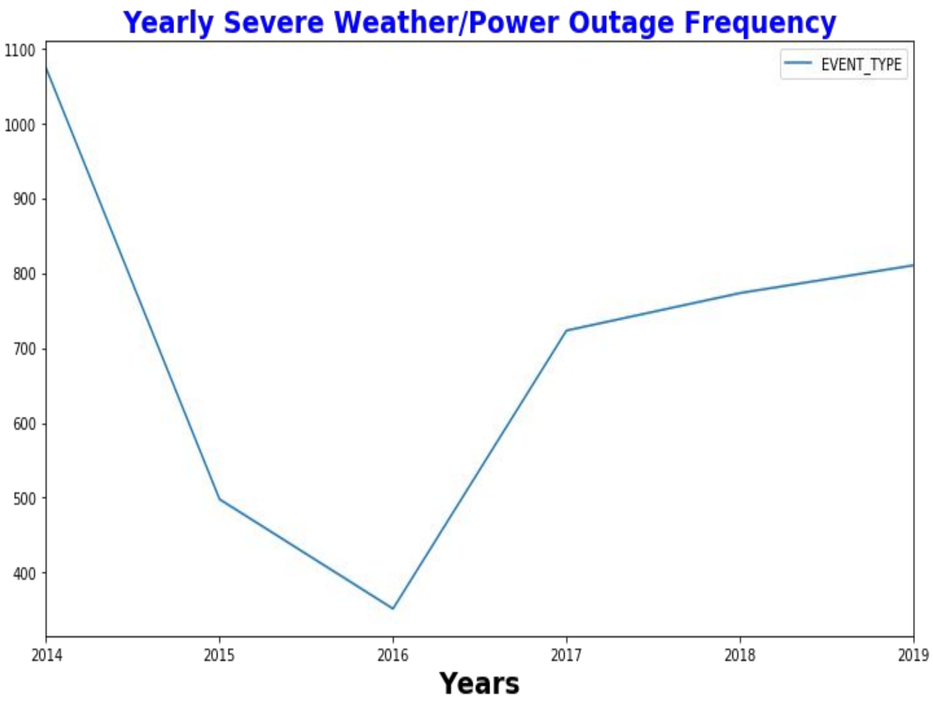
n = 989

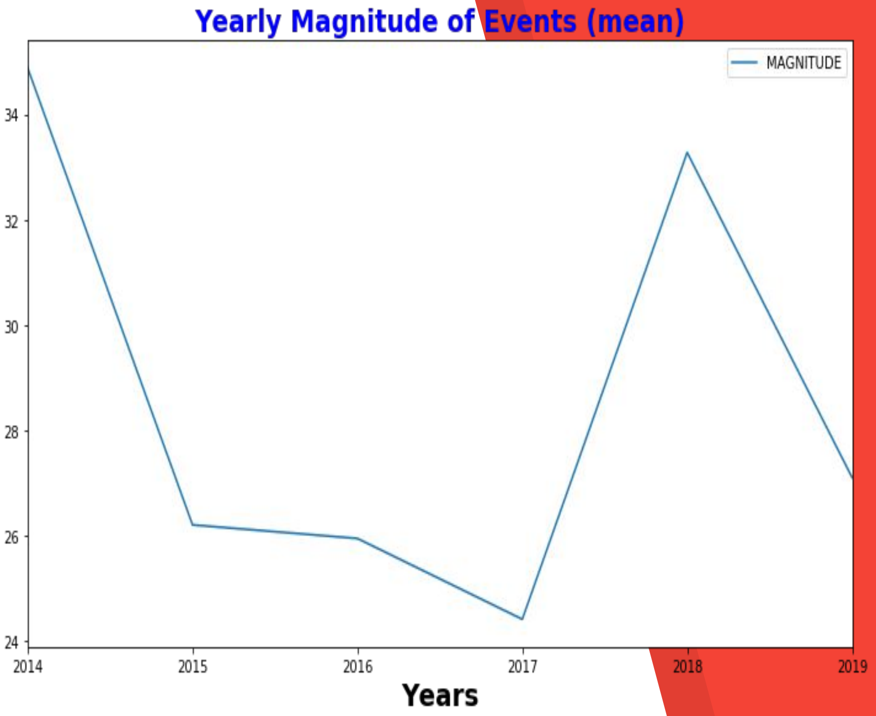
Number of Events in California	
Flood	220
Strong Wind	213
High Wind	164

n = 612

Number of Events in New York	
Thunderstorm Wind	415
High Wind	86
Hail	42

Power Outage vs Magnitude





Magnitudes Above The Average	High Wind	Thunderstorm Wind
	Strong Wind	

Magnitudes Below the Average	Heavy Rain	Lightning	Wildfire
	Heavy Snow	Hail	Excessive Heat

Modeling & Evaluation

Logistic Regression Model
Random Forest Classifier
DT Bagging Classifier

Supervised Classification Model

- **Supervised:**

  our target is major power outage events
- **Classification Model:**

  we want to predict whether or not a major power outage event is happening
- **Input (X):**

  tweets on Twitter

Classifier Models

- **Logistic Regression Model:**

  classic classification model & easy for interpretation

- **DT Bagging Classifier:**

  ensemble model

- **Random Forest Classifier:**

  lower variance

Model Evaluation

We try to minimize false negative

- ~~Accuracy Score~~
- Recall Score
- F1 Score

Final Model

Model Comparison

Model	Model Type	Sample	Vectorizer	Training Score (accuracy)	Testing Score (accuracy)	ROC Score	Recall Score	F1 Score
Model 1	Logistic Regression	Original Sample	CountVerctorizer	96.20%	95.21%	54.71%	9.75%	16.84%
Model 2	DT Bagging	Original Sample	CountVerctorizer	98.63%	94.48%	62.16%	26.27%	32.12%
Model 3	Random Forest	Original Sample	CountVerctorizer	95.04%	95.03%	50.00%	0.00%	0.00%
Model 4	Logistic Regression	Original Sample	Tfidf	95.87%	95.19%	54.50%	9.32%	16.17%
Model 5	DT Bagging	Original Sample	Tfidf	98.52%	94.79%	58.91%	19.06%	26.70%
Model 6	Random Forest	Original Sample	Tfidf	95.10%	95.03%	50.20%	0.42%	0.84%
Model 7	Logistic Regression	Decreased Sample	CountVerctorizer	92.42%	65.86%	65.22%	52.96%	59.52%
Model 8	DT Bagging	Decreased Sample	CountVerctorizer	95.57%	66.86%	66.68%	63.13%	64.36%
Model 9	Random Forest	Decreased Sample	CountVerctorizer	70.37%	62.44%	60.71%	27.54%	41.00%
Model 10	Logistic Regression	Decreased Sample	Tfidf	86.73%	67.87%	67.11%	52.54%	60.78%
Model 11	DT Bagging	Decreased Sample	Tfidf	94.57%	65.06%	64.01%	44.06%	54.45%
Model 12	Random Forest	Decreased Sample	Tfidf	71.78%	61.64%	60.33%	35.16%	46.49%
Model 13	Logistic Regression	Increased Sample	CountVerctorizer	98.64%	97.97%	97.97%	98.22%	97.96%
Model 14	Random Forest	Increased Sample	CountVerctorizer	67.34%	66.97%	66.90%	42.72%	56.33%
Model 15	Logistic Regression	Increased Sample	Tfidf	98.20%	97.42%	97.42%	98.22%	97.43%
Model 16	Random Forest	Increased Sample	Tfidf	69.77%	69.33%	69.31%	60.67%	66.36%

Model Performance

We already optimized the hyperparameters

- Unbalanced Classes
  - Dropping majority class
  - Bootstrapping minority class
  - Ensemble Models
- Similar target == 0 & target == 1
  - Both are power outages

Future Steps & Discussion

Logistic Regression Model

Random Forest Classifier

DT Bagging Classifier

Improving the Model

- **More Quality Data:**

  We are limited by our data, which only provide us major power outage events and the locations are all state-level or county-level

- **Better Keyword Choice:**

  Blackouts is causing troubles - drunk people “blackout”

- **More Features:**

  We can implement weather data to our model since they are highly correlated

Discussion

- **Data Collection:**

  Definitely a big part, there is a website selling power outage data - major utility companies do not provide historical power outage data

- **App Implementation:**

  It would be great if we can have more time to finish up the App since we already have our model pickled

Source & Reference

- **the U.S. Department of Energy:**

  https://www.oe.netl.doe.gov/OE417_annual_summary.aspx
- **National Oceanic and Atmospheric Administration:**

  https://www.ncdc.noaa.gov/data-access/severe-weather
- **Using word2vec to Analyze News Headlines and Predict Article Success:**

  https://towardsdatascience.com/using-word2vec-to-analyze-news-headlines-and-predict-article-success-cdeda5f14751
- **7 Techniques to Handle Imbalanced Data:**

  https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html
- **twitterscraper:**

  https://github.com/taspinar/twitterscraper

Question & Feedback