SEVERE WEATHER PREDICTION

Using Weather Data by Location to Predict Severe Weather Events

BREA BEALS, JULIAN KLEINDIEK, AND MARK ROBERTS

Agenda

Business Problem

Data Sources and Insights

Feature Engineering

Data Modeling

Future Work

Business Problem

Severe weather events caused >**1,300** deaths and >**$100B** in damages from 2015 to 2020 in the US alone.

For **communities** to better prepare for those events, our goal is to predict if a severe weather event is likely to happen in the near future and if so which type of event.

For **insurances** to be able to allocate funds early and respond to those events quickly, our second goals is to predict the expected damage from severe weather events.

Data Sources

Weather

- 16 unique weather attributes
- Data available on a daily basis including latitude and longitude
- Selected data at a 500hPa pressure level
- Approx. 3.5M rows of data accessed via the Copernicus API

Joined sources by date and a 58x58 miles grid of the US

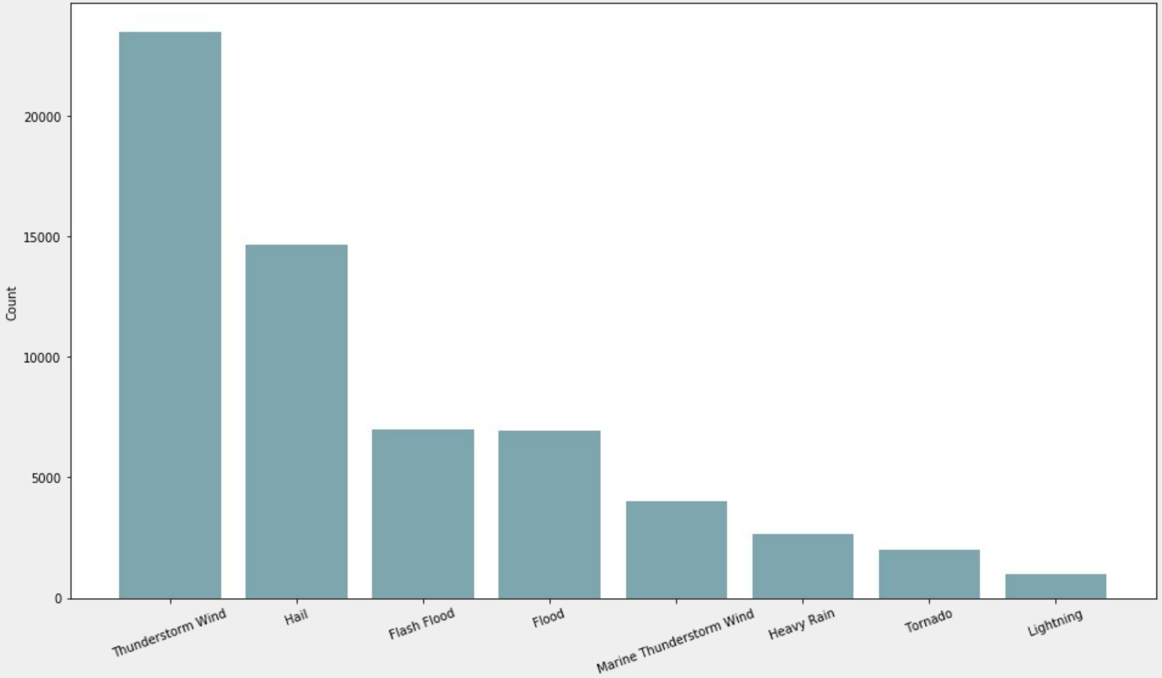
Severe Events

- Data on event type and the related damages, deaths, and injuries
- Data available by date of the event including latitude and longitude
- Approx. 60K rows of data scraped from NOAA website with BeautifulSoup

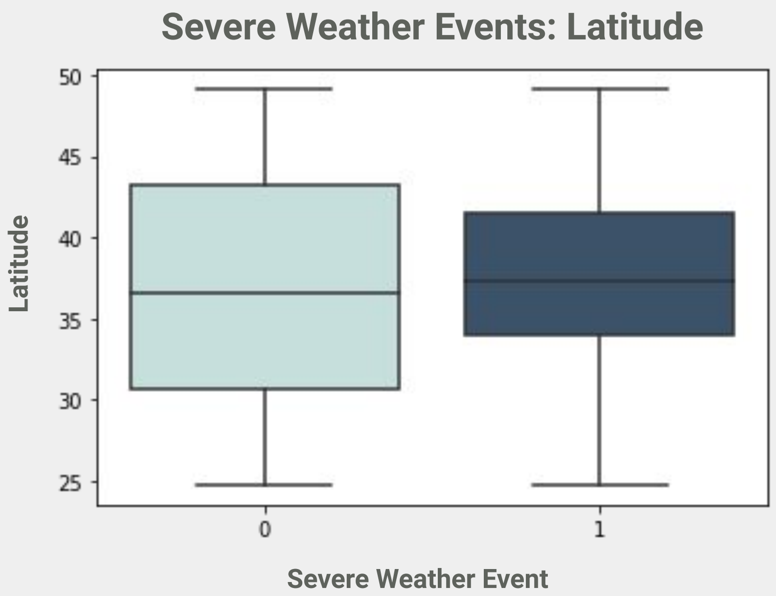
Exploratory Data Analysis

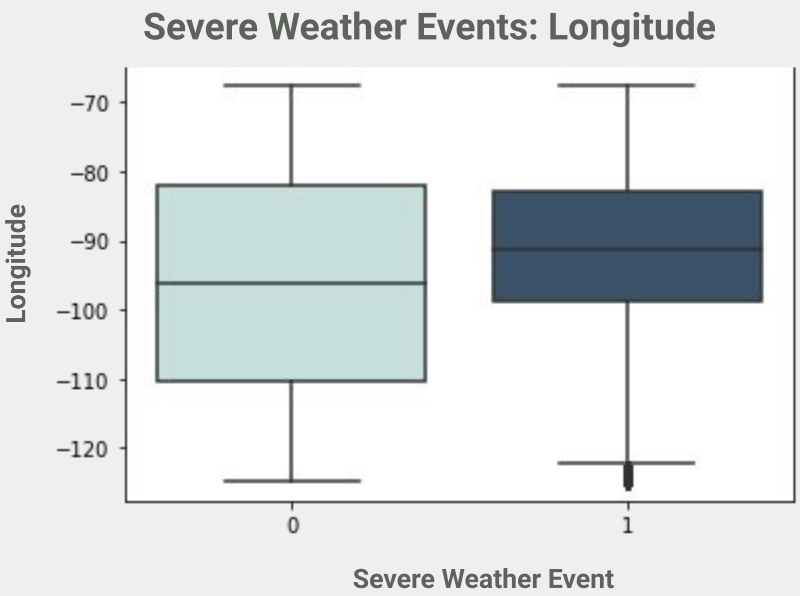
WEATHER EVENT	COUNT
Thunderstorm Wind	23,471
Hail	14,619
Flash Flood	6,885
Flood	6,807
Marine Thunderstorm Wind	4,027
Heavy Rain	2,276
Tornado	2,015
Lightning	963

Distribution of Severe Weather Events

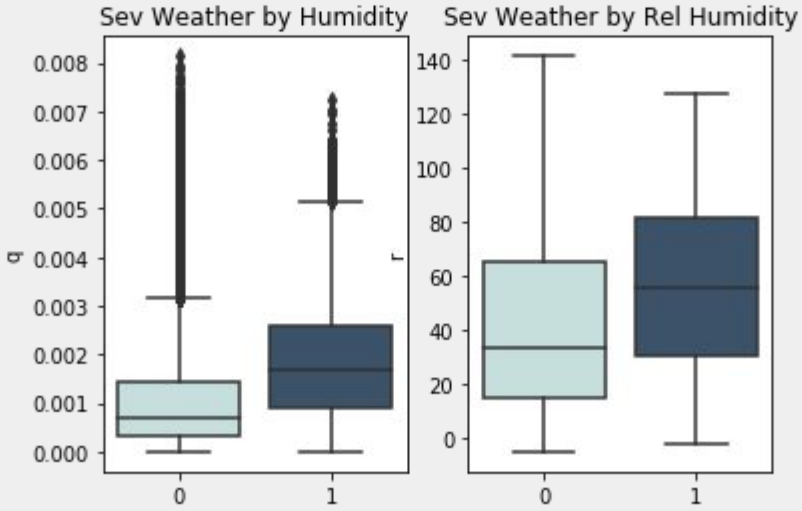


Exploratory Data Analysis: Location Features

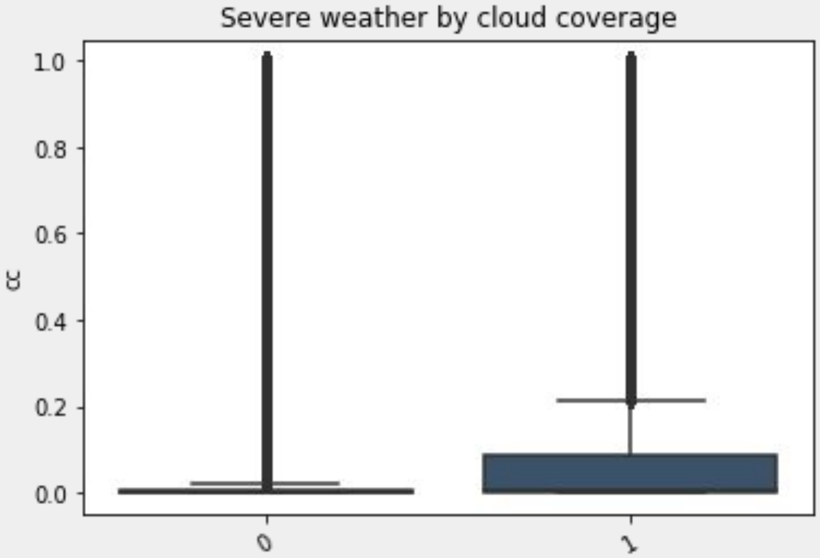




Exploratory Data Analysis: Weather Features

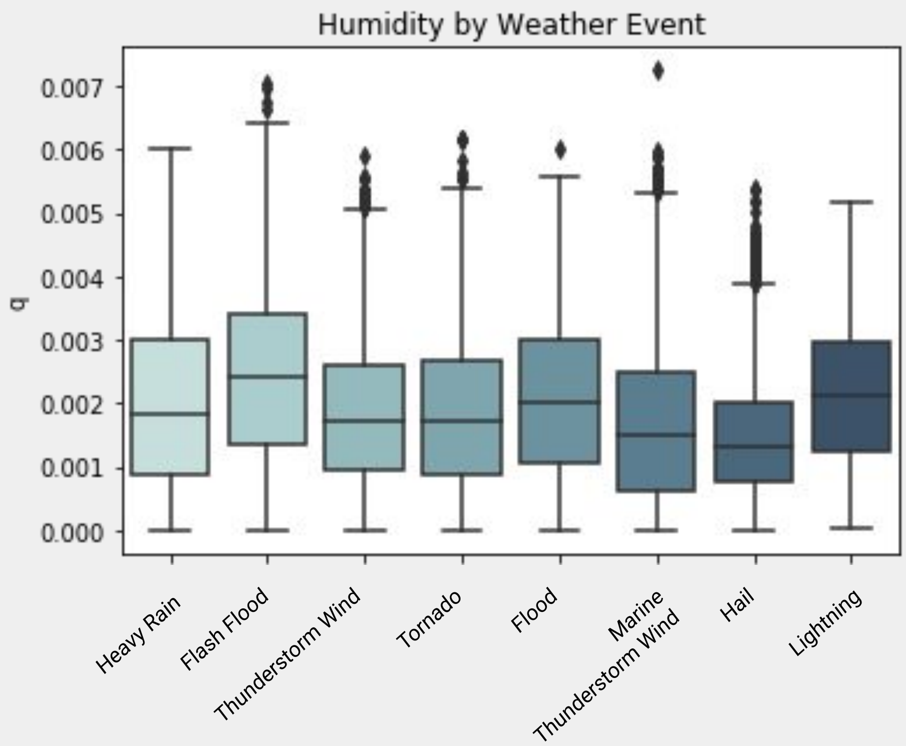


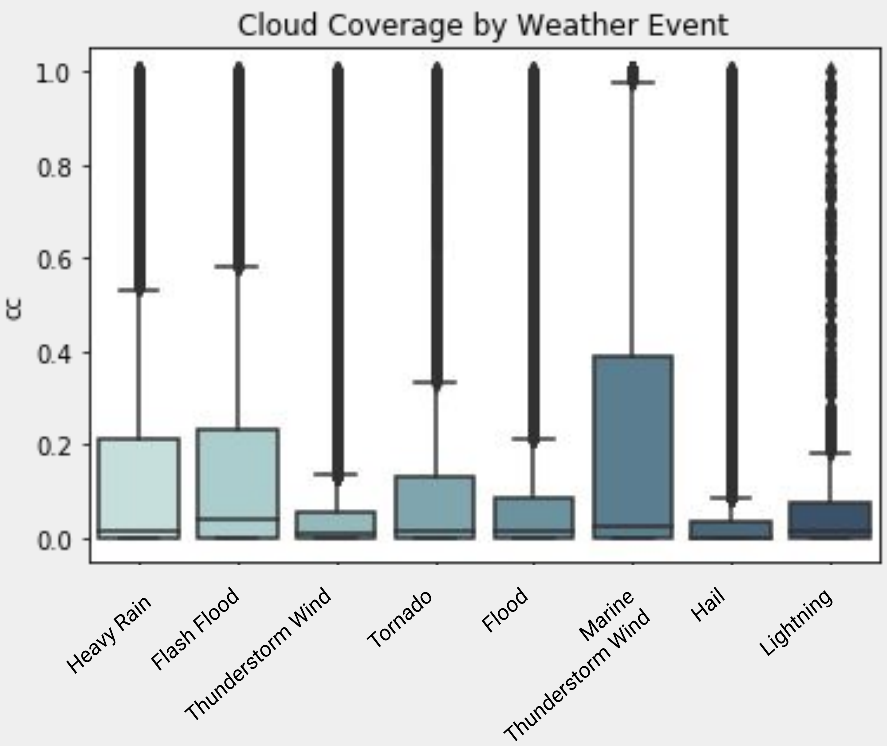
Severe Weather Event



Severe Weather Event

Exploratory Data Analysis: Weather Features





Feature Engineering

Relevant steps:

- Applied variance threshold to drop low variance features
- Dropped highly correlated features
- Lagged data by one day
- Created 10-day rolling averages, standard deviations, minima and maxima
- Created 4 geo clusters based on latitude and longitude

#	Column	Dtype
0	latitude	float64
1	longitude	float64
2	EVENT_TYPE	object
3	INJURIES_DIRECT	float64
4	INJURIES_INDIRECT	float64
5	DEATHS_DIRECT	float64
6	DEATHS_INDIRECT	float64
7	DAMAGE_PROPERTY	float64
8	DAMAGE_CROPS	float64
9	fraction_cloud_cover	float64
10	relative_humidity	float64
11	temperature	float64
12	u_component_wind	float64
13	v_component_wind	float64
14	vertical_velocity	float64
15	fraction_cloud_cover_10_day_mean	float64
16	relative_humidity_10_day_mean	float64
17	temperature_10_day_mean	float64
18	u_component_wind_10_day_mean	float64
19	v_component_wind_10_day_mean	float64
20	vertical_velocity_10_day_mean	float64
21	fraction_cloud_cover_10_day_std	float64
22	relative_humidity_10_day_std	float64
23	temperature_10_day_std	float64
24	u_component_wind_10_day_std	float64
25	v_component_wind_10_day_std	float64
26	vertical_velocity_10_day_std	float64
27	fraction_cloud_cover_10_day_max	float64
28	relative_humidity_10_day_max	float64
29	temperature_10_day_max	float64
30	u_component_wind_10_day_max	float64
31	v_component_wind_10_day_max	float64
32	vertical_velocity_10_day_max	float64
33	fraction_cloud_cover_10_day_min	float64
34	relative_humidity_10_day_min	float64
35	temperature_10_day_min	float64
36	u_component_wind_10_day_min	float64
37	v_component_wind_10_day_min	float64
38	vertical_velocity_10_day_min	float64
39	geo_cluster	int64
40	year	int64
41	month	int64
42	day	int64


MODELING

Binary Classification

Target: Severe Weather Y/N

1. Logistic Regression
2. Random Forest
3. AdaBoost
4. Gradient Boost

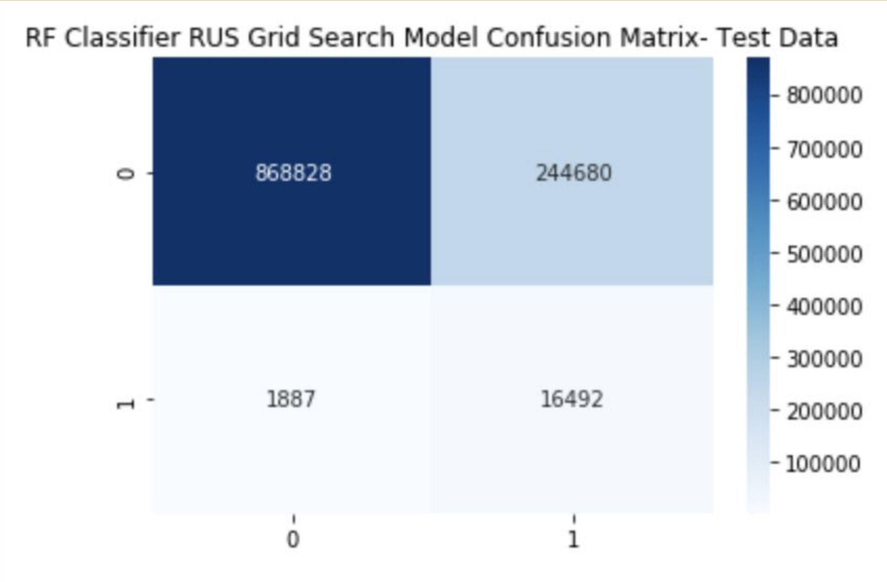
RAW DATA

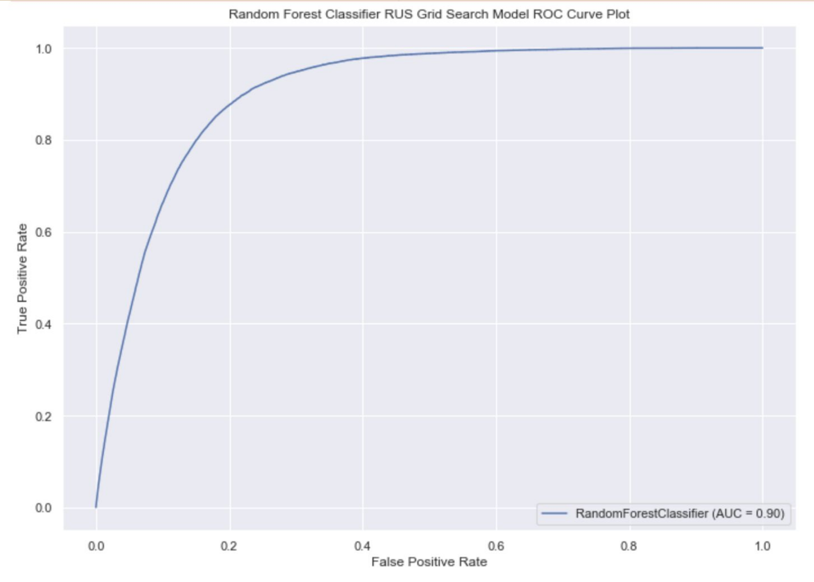
- Grid search for best parameters
- Optimize Prediction Threshold
- Ensemble Results

UNDERSAMPLED DATA

- Grid search for best parameters
- Optimize Prediction Threshold
- Ensemble Results

Random Under-Sampled Data Results



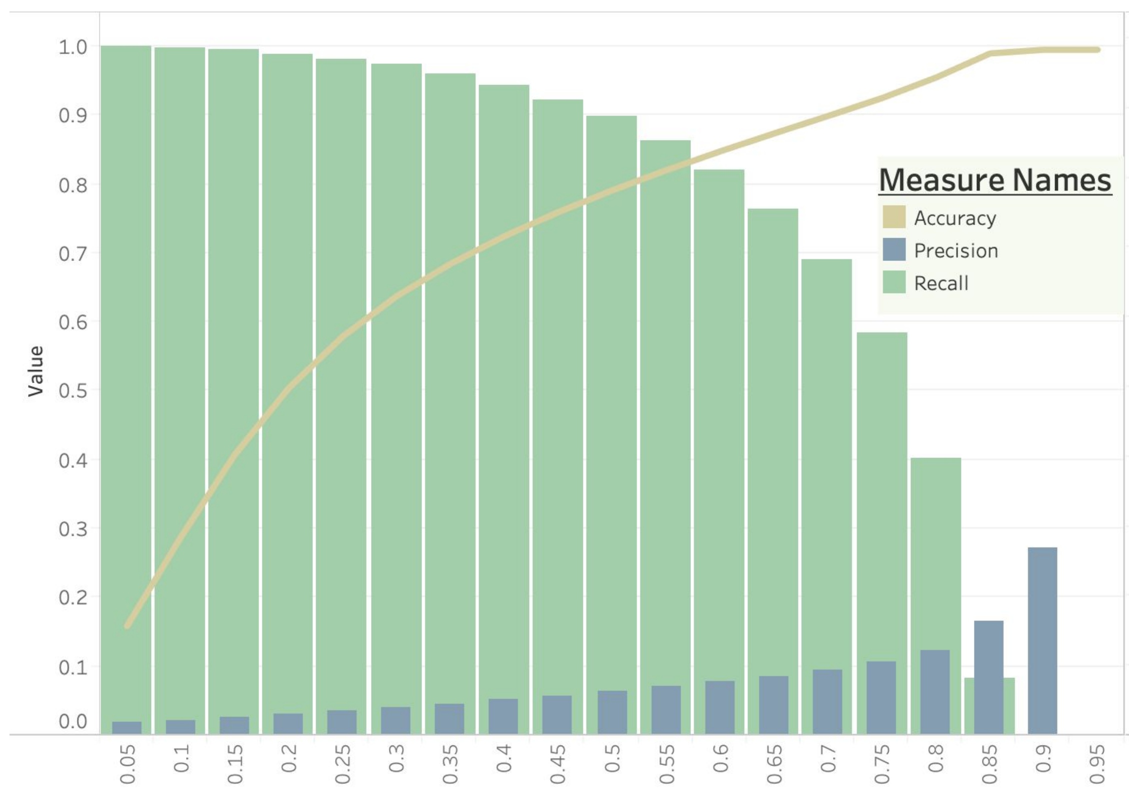


Accuracy score: 0.98
Precision/Recall Score (class 1): 0

RUS Data Prediction Threshold

Threshold: 0.75

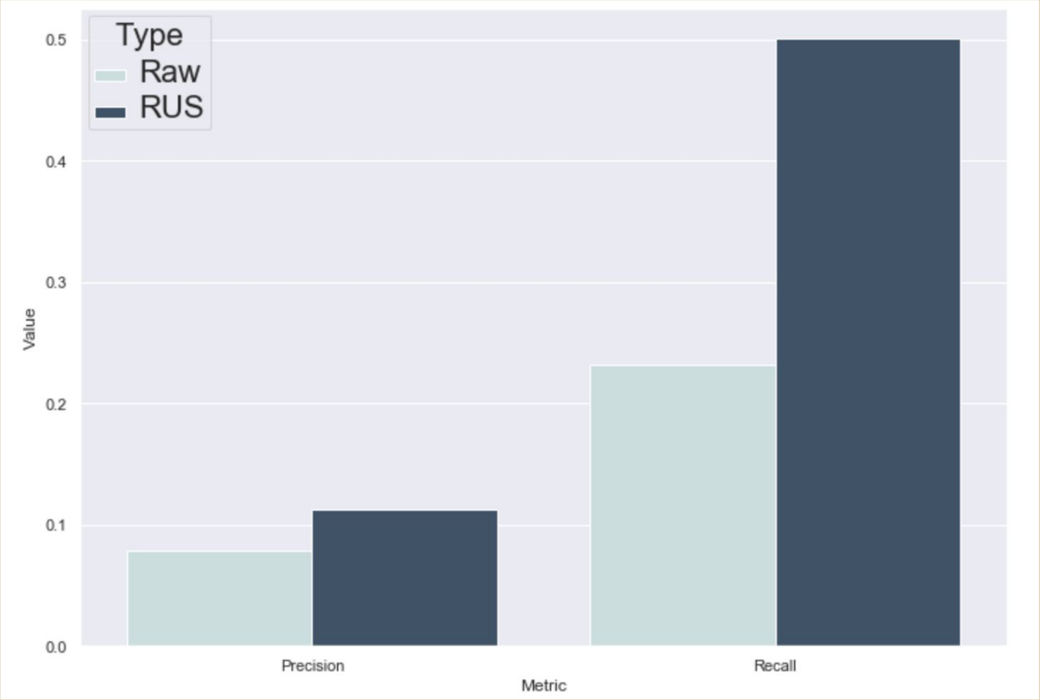
- **Precision:** 0.107
- **Recall:** 0.58
- **Accuracy:** 0.91



Ensemble Model Performance Comparison

<u>**RUS vs Raw Data**</u>

- **Precision: 37.5% lift**
- **Recall: 117% lift**



Weather Event Classification

Classification of Top 3 Weather Events by Frequency

Weather Event	Count
Wind Related Events	27,498
Hail	14,619
Flood Related Events	13,692

MODELING

- Logistic Regression, SVM, Random Forest, AdaBoosting, KNeighbors Classifier, XGBoost
- Examined accuracy, weighted average accuracy, and generalization gap for model selection
- Optimized the model hyperparameters using Randomized Cross Validation

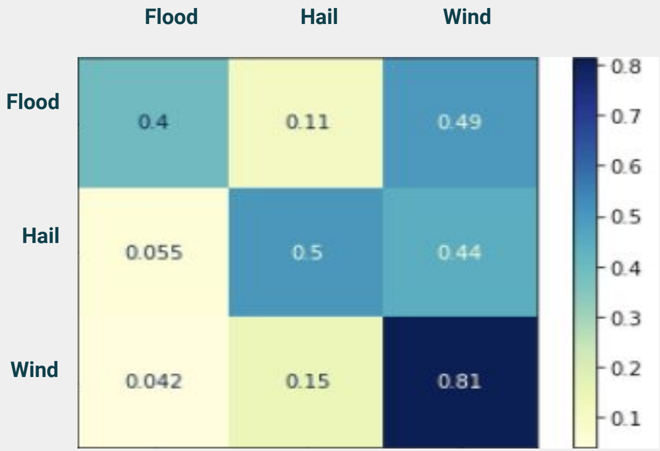
VARIOUS TECHNIQUES

- Applied various techniques to see if performance could be improved
    - PCA
    - Random Undersampling
    - SMOTE Oversampling

Weather Event Classification

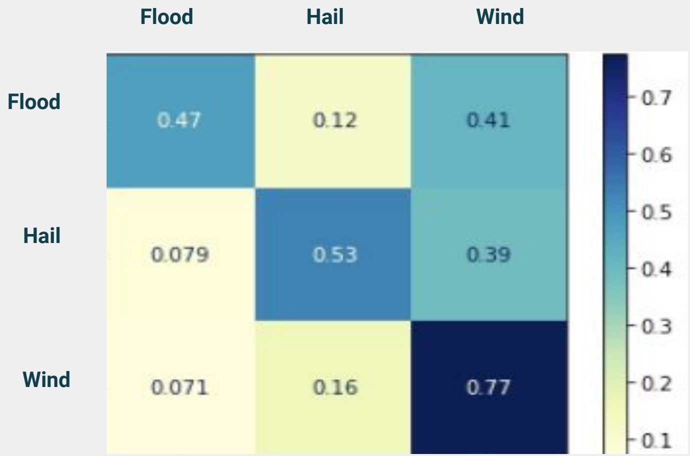
Random Forest Classifier:

- Training accuracy: 66%
- Test accuracy: 63%
- Test Weighted Accuracy: 62%



XGBoost Classifier:

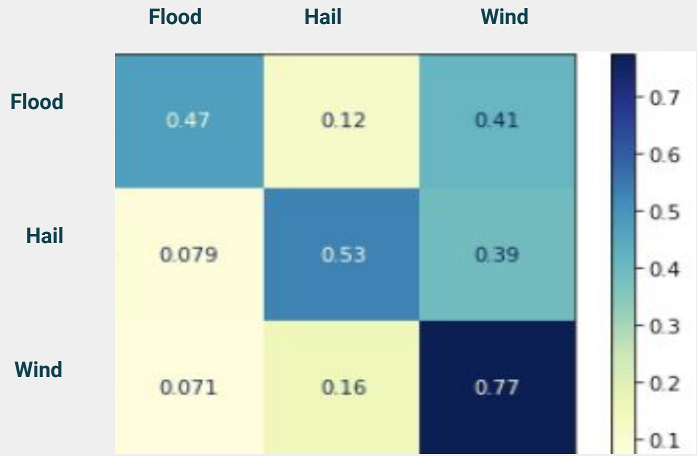
- Training accuracy: 67%
- Test accuracy: 64%
- Test Weighted Average: 63%



Randomly Undersampled Data

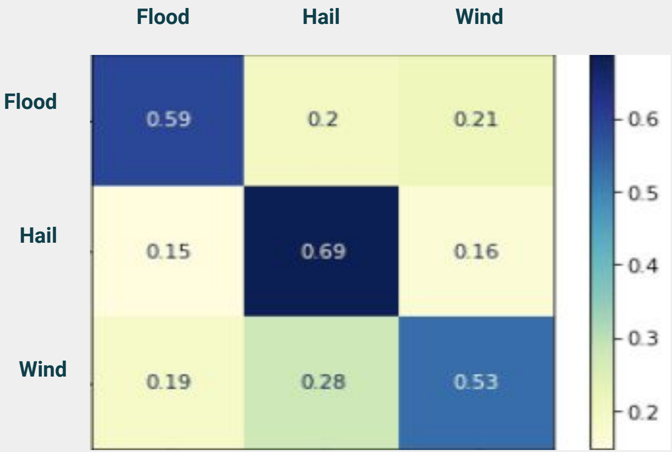
Original XGBoost Model

- Training accuracy: 67%
- Test accuracy: 64%
- Test Weighted Average: 63%



Randomly Undersampled Data

- Training accuracy: 62%
- Test accuracy: 60%
- Test Weighted Average: 60%



Damage Model

- Experimented with different target variables (regression, multiclass classification)
- Settled on a binary classification of the **target variable "damage above $100K"**
- Training on a mix of over- and undersampled data (50:50 split among classes)
- Tested SVMs, Decision Trees, Random Forests, Gradient Boosting, Ada Boosting and Artificial Neural Networks

Ada Boosting:

- Optimized the model hyperparameters using Randomized Cross Validation
- Training accuracy: 80%
- Test accuracy: 78%

Neural Network:

- 3 layers with a total of 20 neurons using relu and sigmoid as activation functions
- Training accuracy: 71%
- Test accuracy: 61%

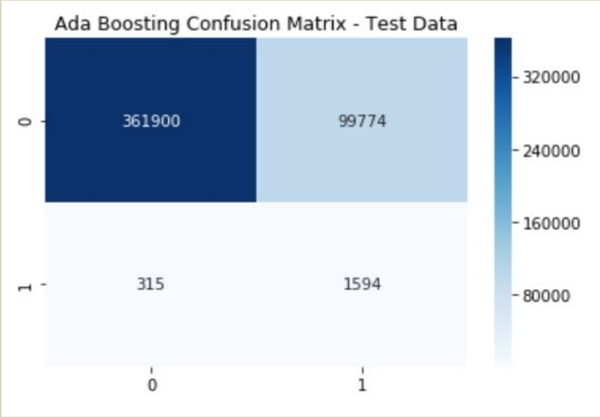
Selected Ada Boosting as final model due to superior accuracy, precision, and stability when applied to unseen data

Randomly Under- and Oversampled Data

Ada Boosting

Measures for damage >100$:

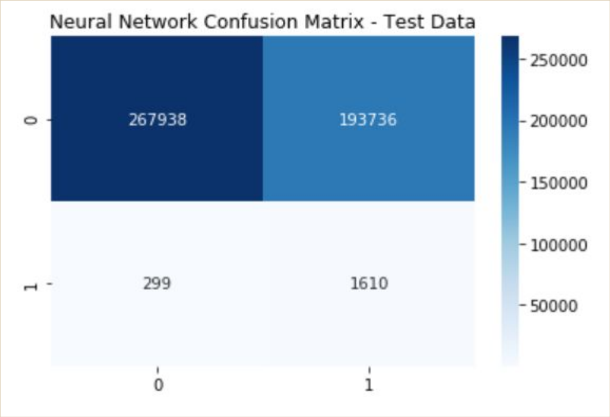
- Train / test precision: 79% / 2%
- Train / test recall: 83% / 83%

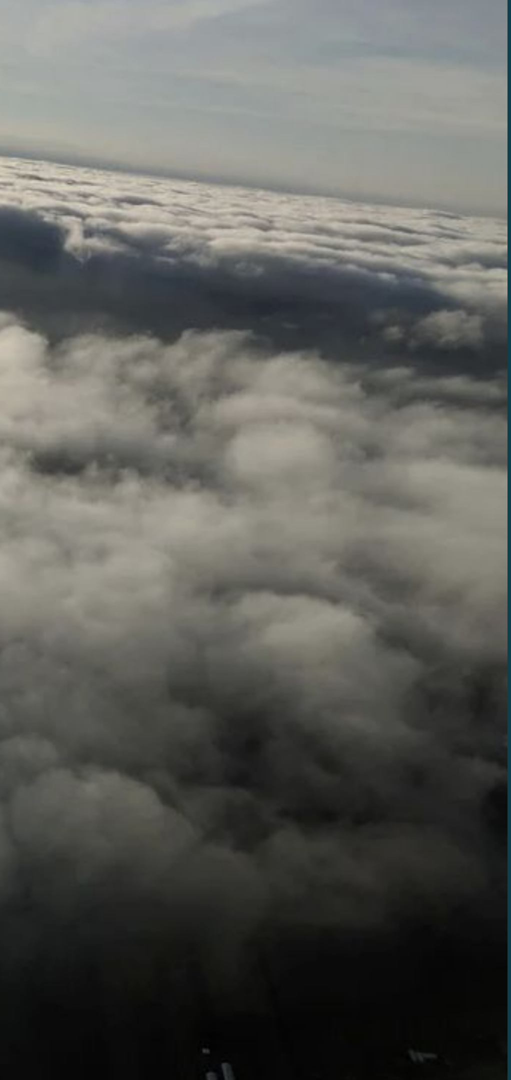


Neural Network

Measures for damage >100$:

- Train / test precision: 66% / 1%
- Train / test recall: 84% / 84%





Future Work

- Increased computing power
    - Limited original dataset to only 5 years - could be better with all 10 years of data
    - Including hourly data
    - Decreasing grid size for our locations
- Further investigation of features
    - Regions of the US
    - More relevant features to predict damage (such as infrastructure given lat/long)
    - Image Data → Google Nowcasting
- Talk with domain experts to verify/expand on assumptions

THANK YOU