

Natural Disaster Damage Prediction

“A storm is coming”

Group 5

Hemangi Kinger
Kartik Jain
Ishan Kuchroo
Siddharth Das

Problem Statement

- **Predict** the **damage** caused by any of the **storm** events in United States.

Data Overview

Description: The Storm Events Database documents the occurrence of storms and other significant weather phenomena having sufficient intensity to cause loss of life, injuries, significant property damage, and/or disruption to commerce.

- **Source:** [NOAA \(National Oceanic and Atmospheric Administration\)](#)
- **Data Availability:** January 1950 to August 2021
- **Number of Observations:** 1,710,146
- **Number of features (*before EDA*):** 51
- **Number of features (*post EDA*):** 184
- **Target Variable:** TOTAL_DAMAGE

Features

- Type of the storm
- Geographic location
- Intensity of the storm
- Storm size
- Number of Deaths
- Cause
- Month, Year



Data Preprocessing

Data Cleaning

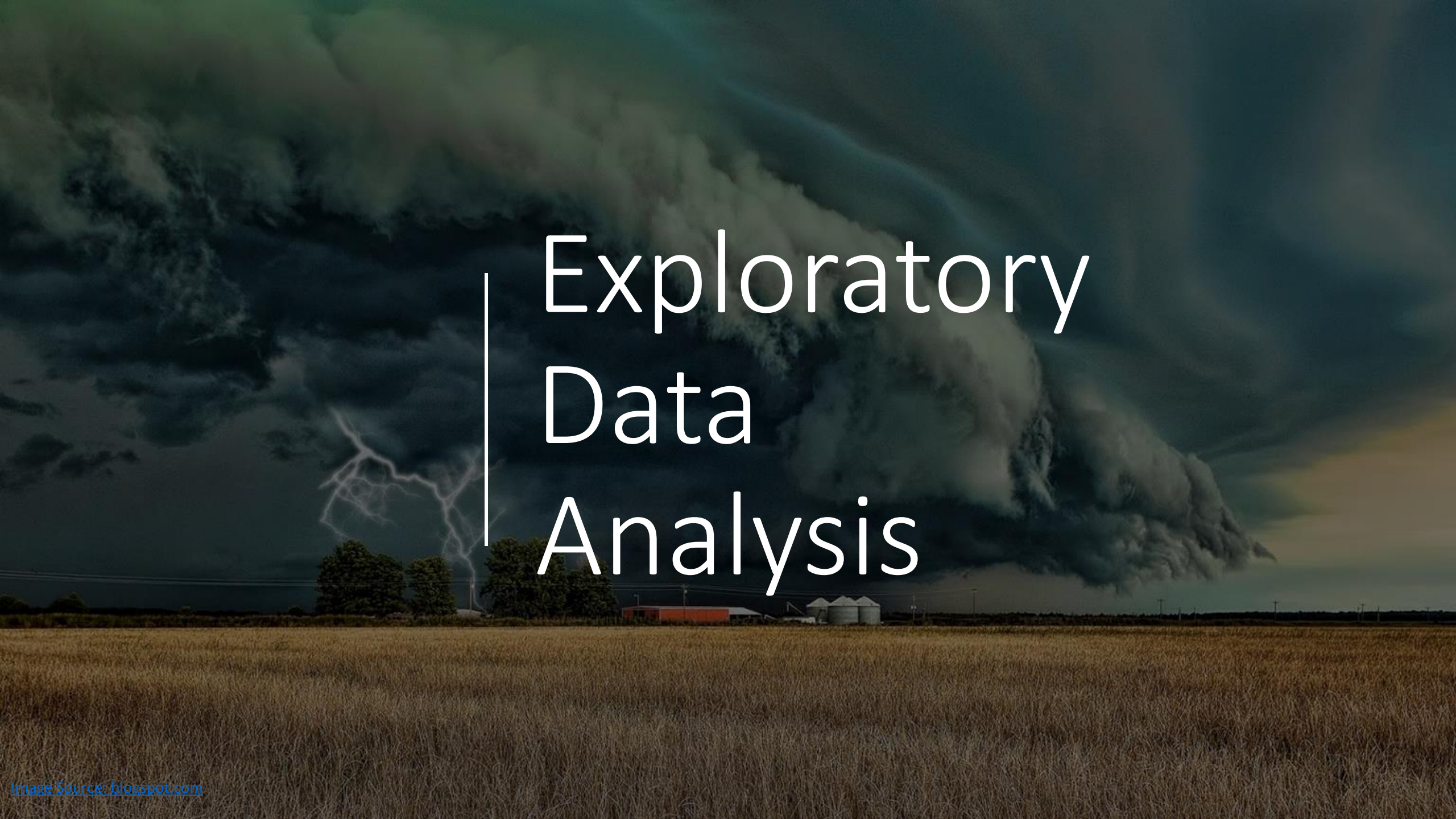
- Data Integration.
- Data from 2005 to 2020
- Removing records where the target variable(“Total_Damage”) is NULL
- Few attributes, like id columns and text columns were removed

Data Transformation

- Alphanumeric values were cleaned to get the numeric value
- Few of the classes were and merged to create consistent classes
- Converted variables to float
- Missing Data Imputation with zero or mean
- Outlier removal using IQR

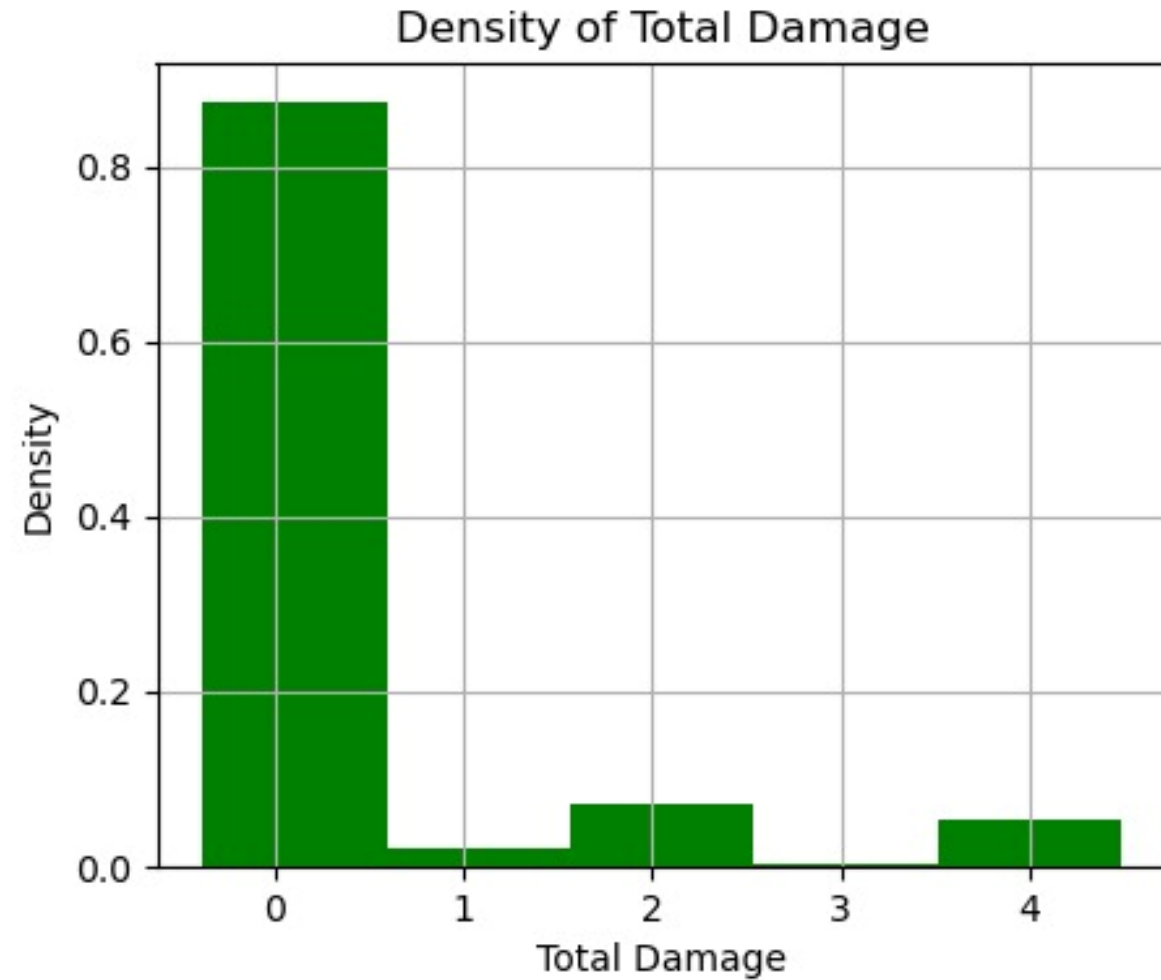
Feature Engineering

- Storm duration
- Distance covered by the storm
- Cold Weather Events
- WIND_SPEED / HAIL_SIZE
- WINDY_EVENTS
- WATER_EVENTS

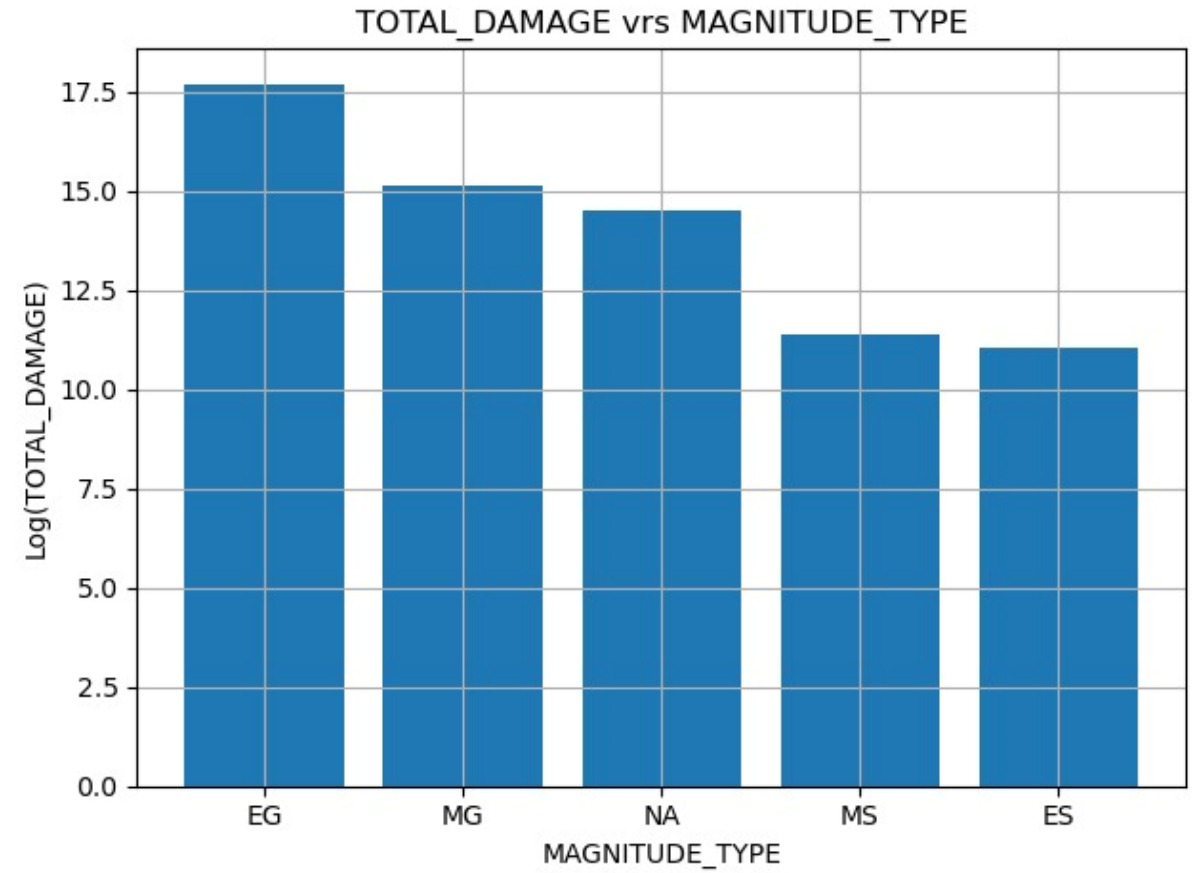
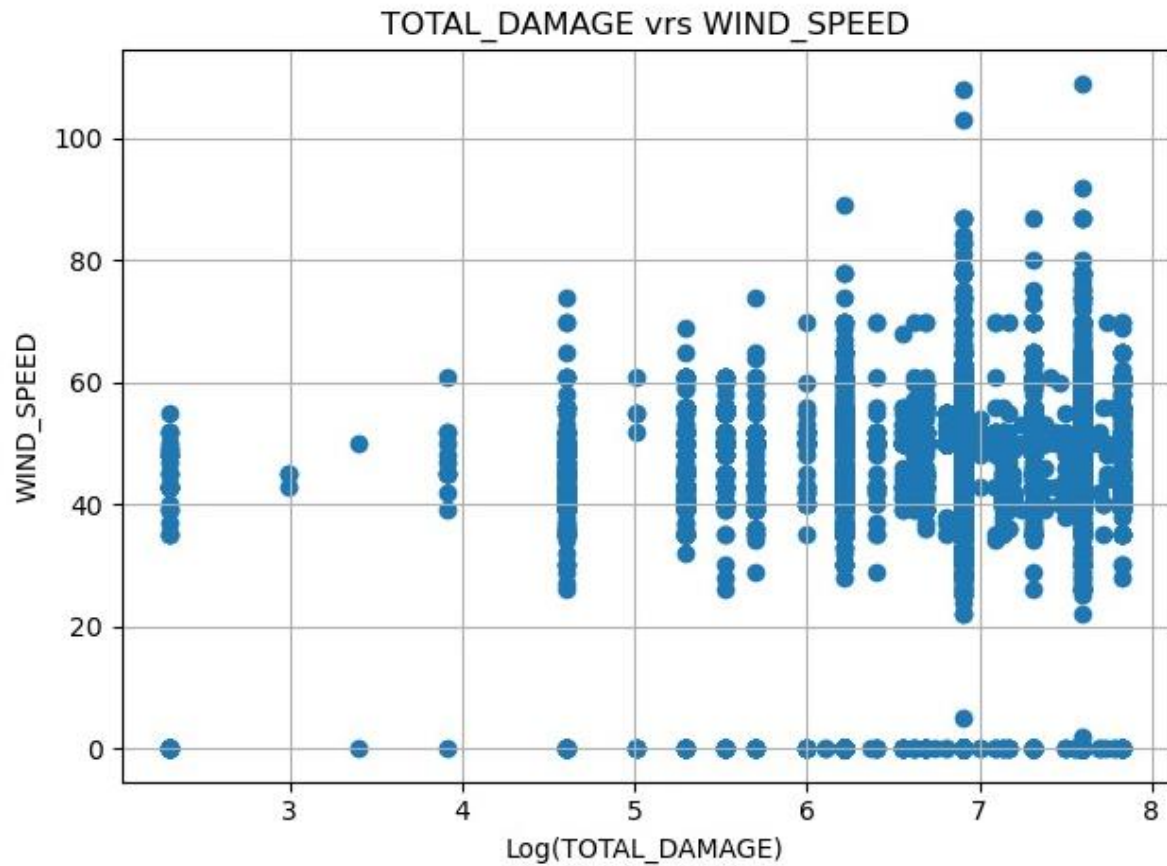
A dramatic landscape photograph featuring a vast, golden-brown field in the foreground. In the middle ground, a small farm with a red barn and two white silos is visible. The sky is dark and stormy, with a large, billowing cloud formation and a bright lightning bolt striking down on the left side. The text "Exploratory Data Analysis" is overlaid in white, with a vertical line to its left.

Exploratory Data Analysis

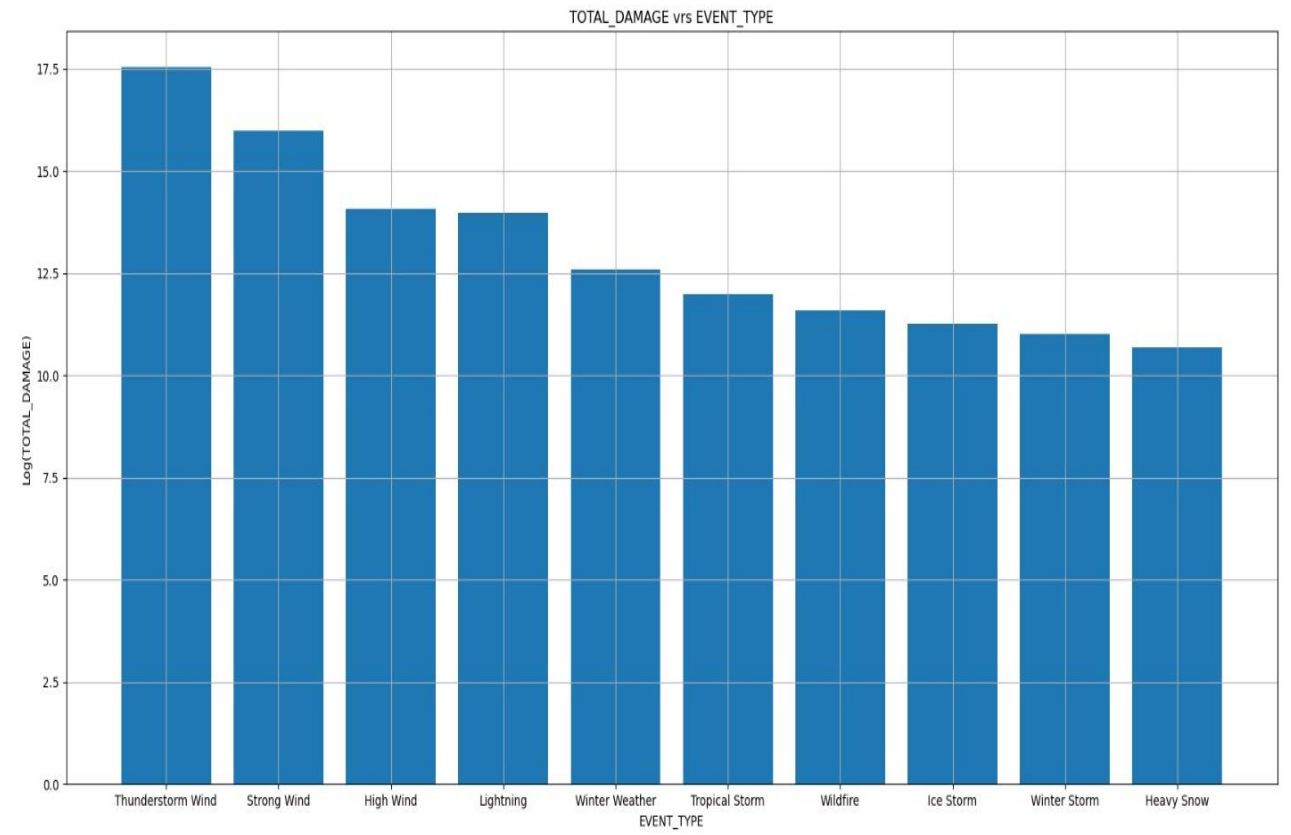
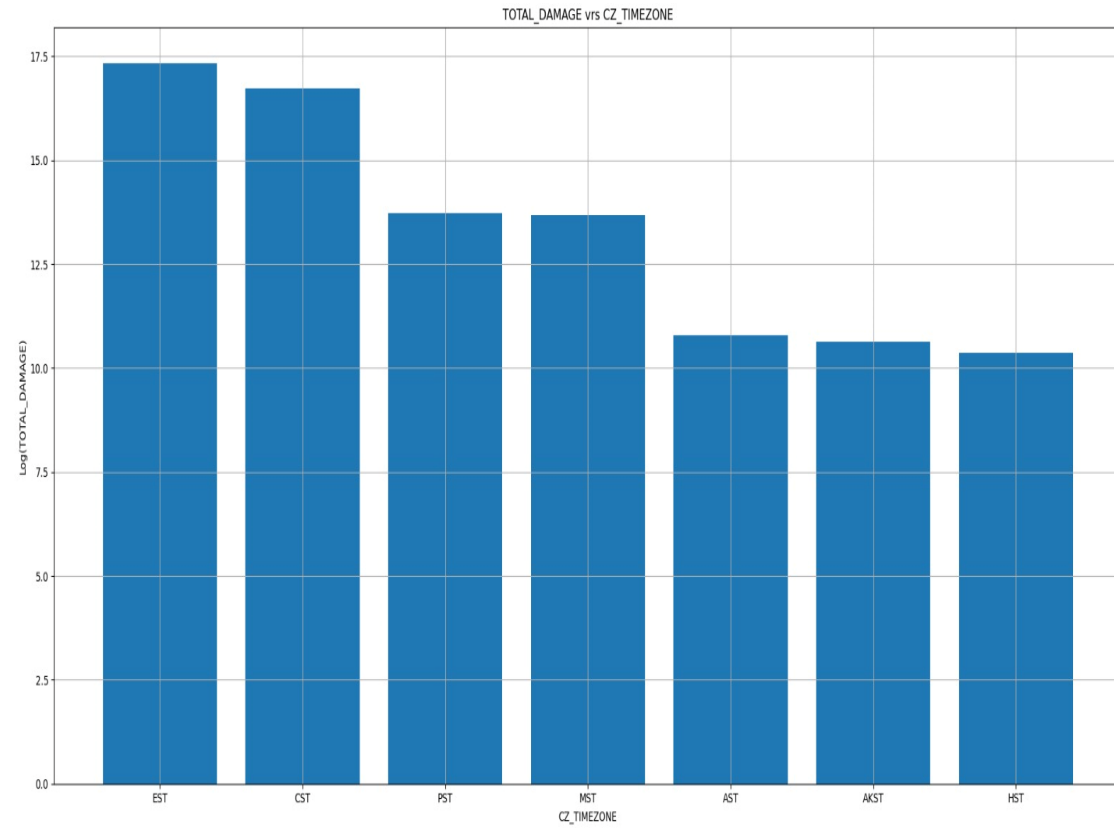
Trend for the Target Variable



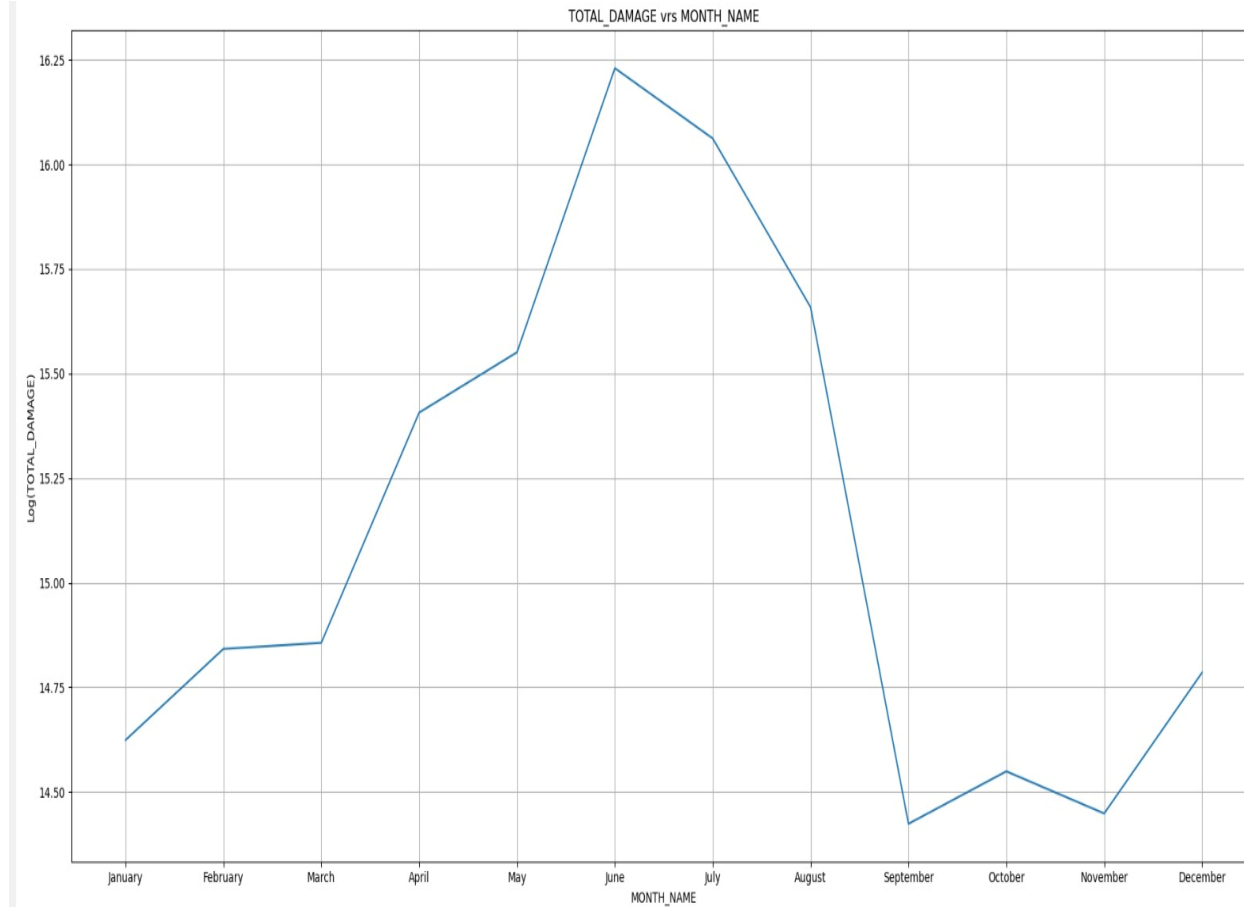
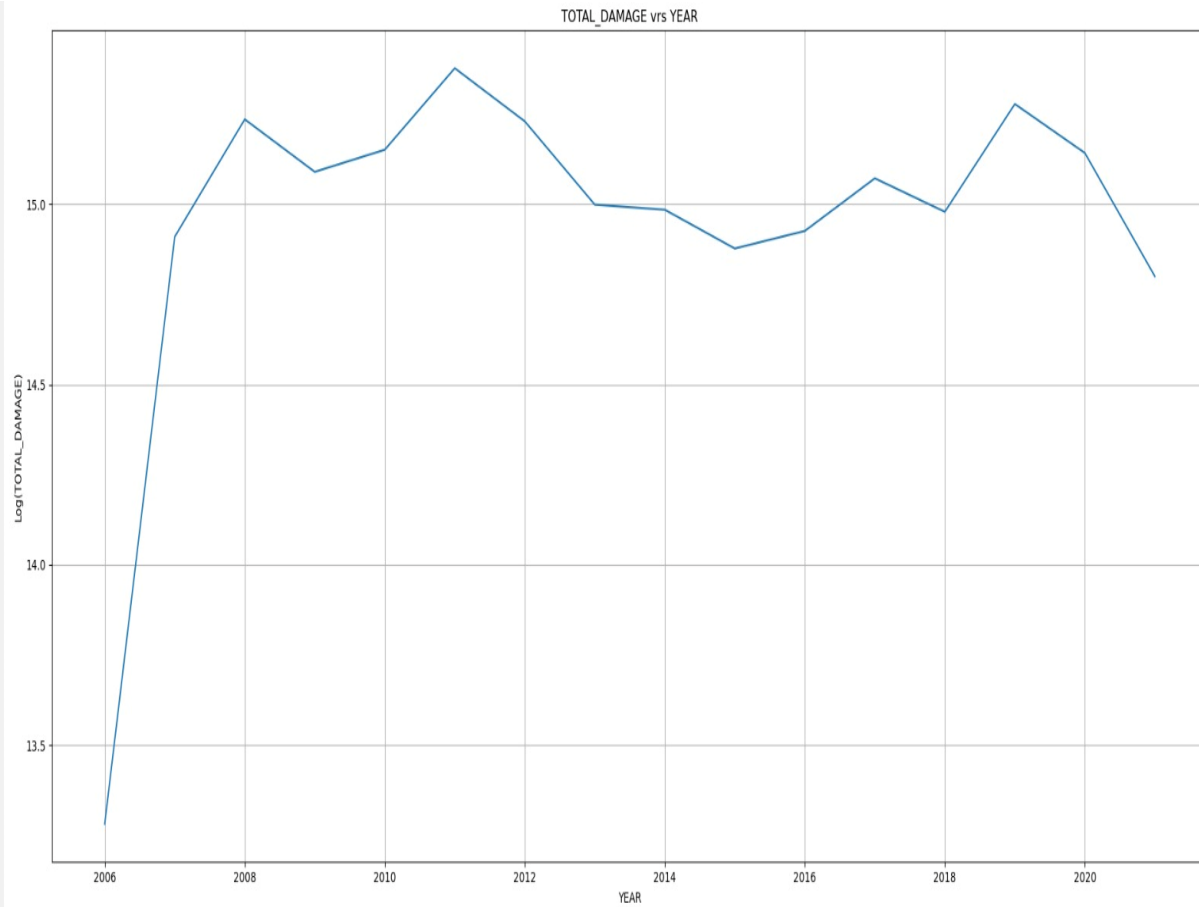
Magnitude and Wind Speed



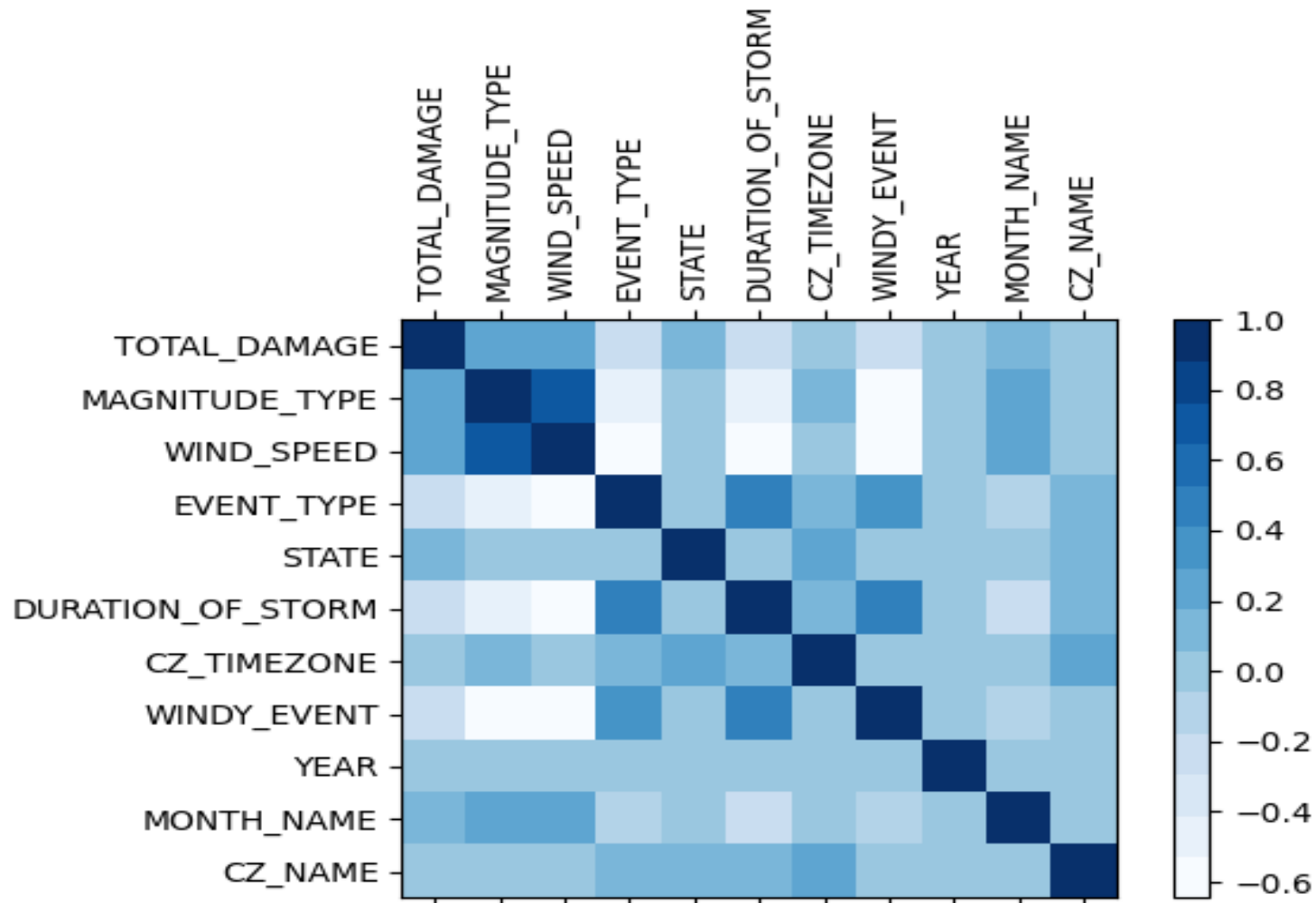
Event Type and Time Zone



Month and Year



Feature Correlation

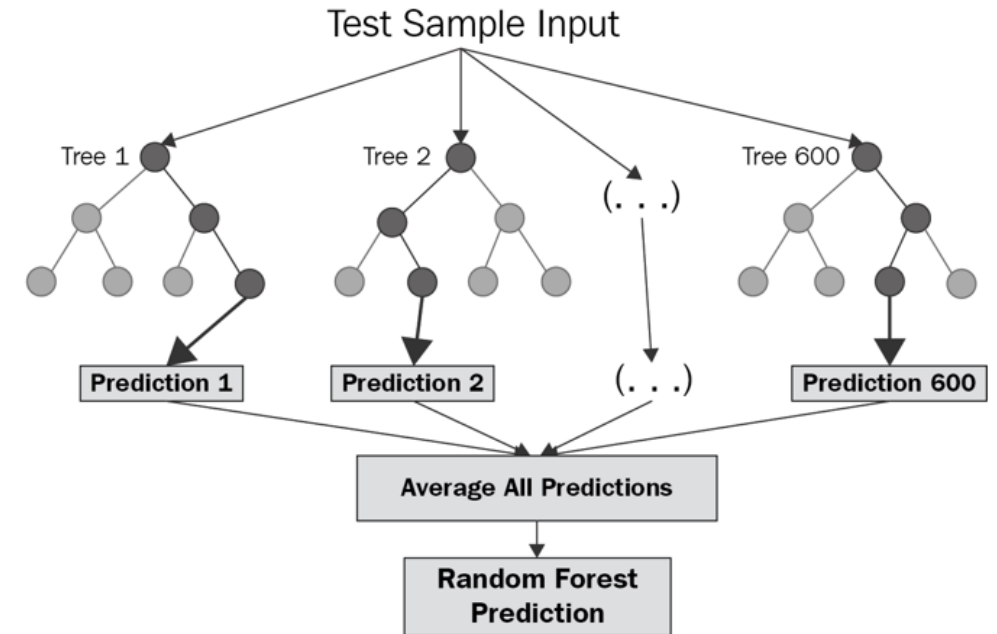




Regression Data Models

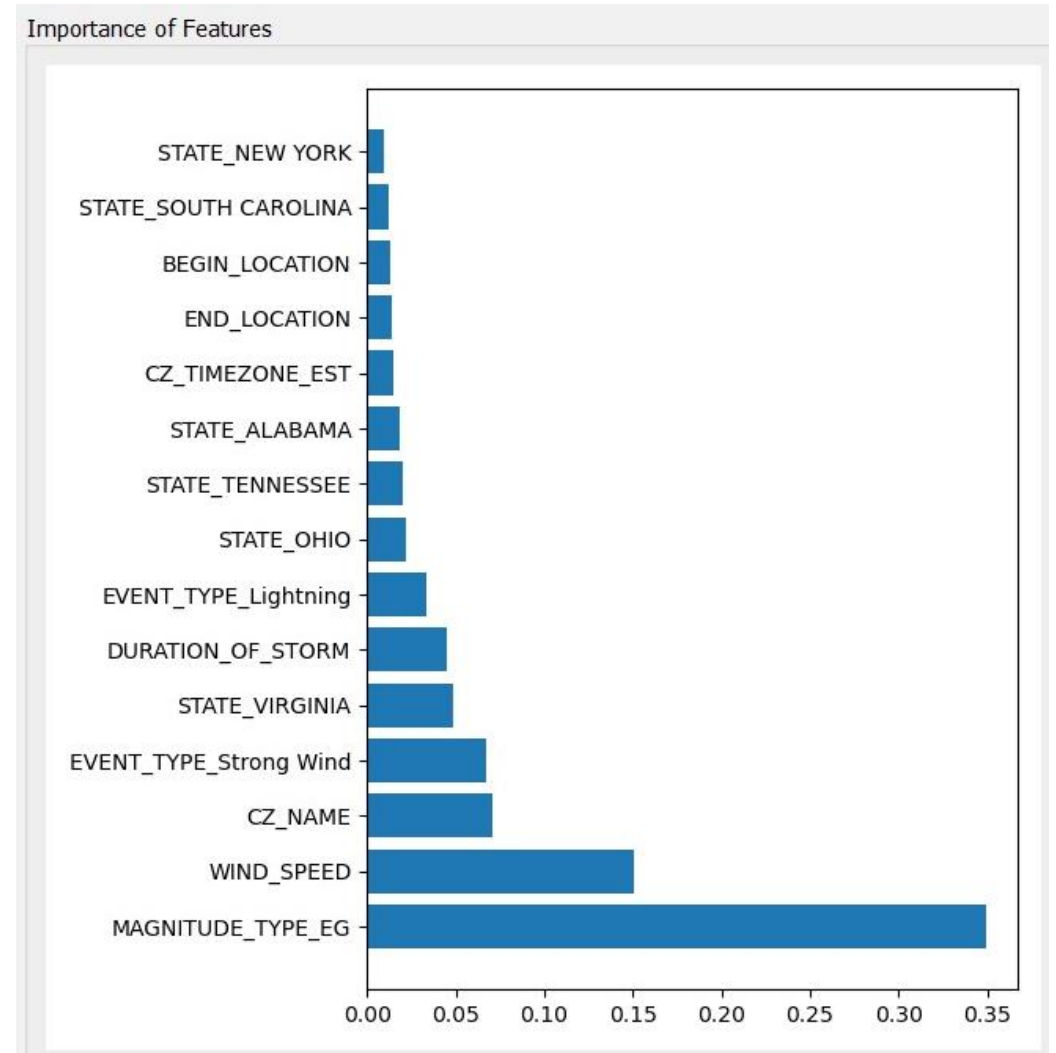
Random Forest Regressor

- A **supervised learning** algorithm that is based on the **ensemble learning** method and many Decision Trees. Random Forest uses a **Bagging technique**, so all calculations are run in parallel and there is no interaction between the Decision Trees when building them.
- We are using Random Forest Regressor to help NOAA predict a continuous value: Predict future Damage Property



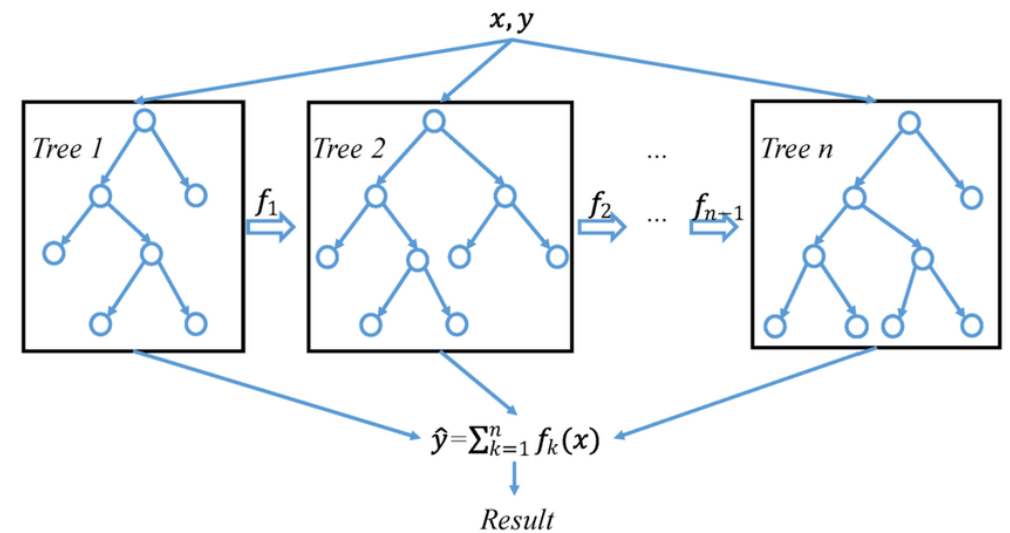
Random Forest Regressor

- Model Training Parameters:
 - `n_estimators= 100`,
 - `oob_score = 'TRUE'`,
 - `n_jobs = -1`,
 - `random_state =50`,
 - `max_features = "auto"`,
 - `min_samples_leaf = 50`
- Coefficient of determination (R^2 score):
 - Training: 53%
 - Validation: 50%



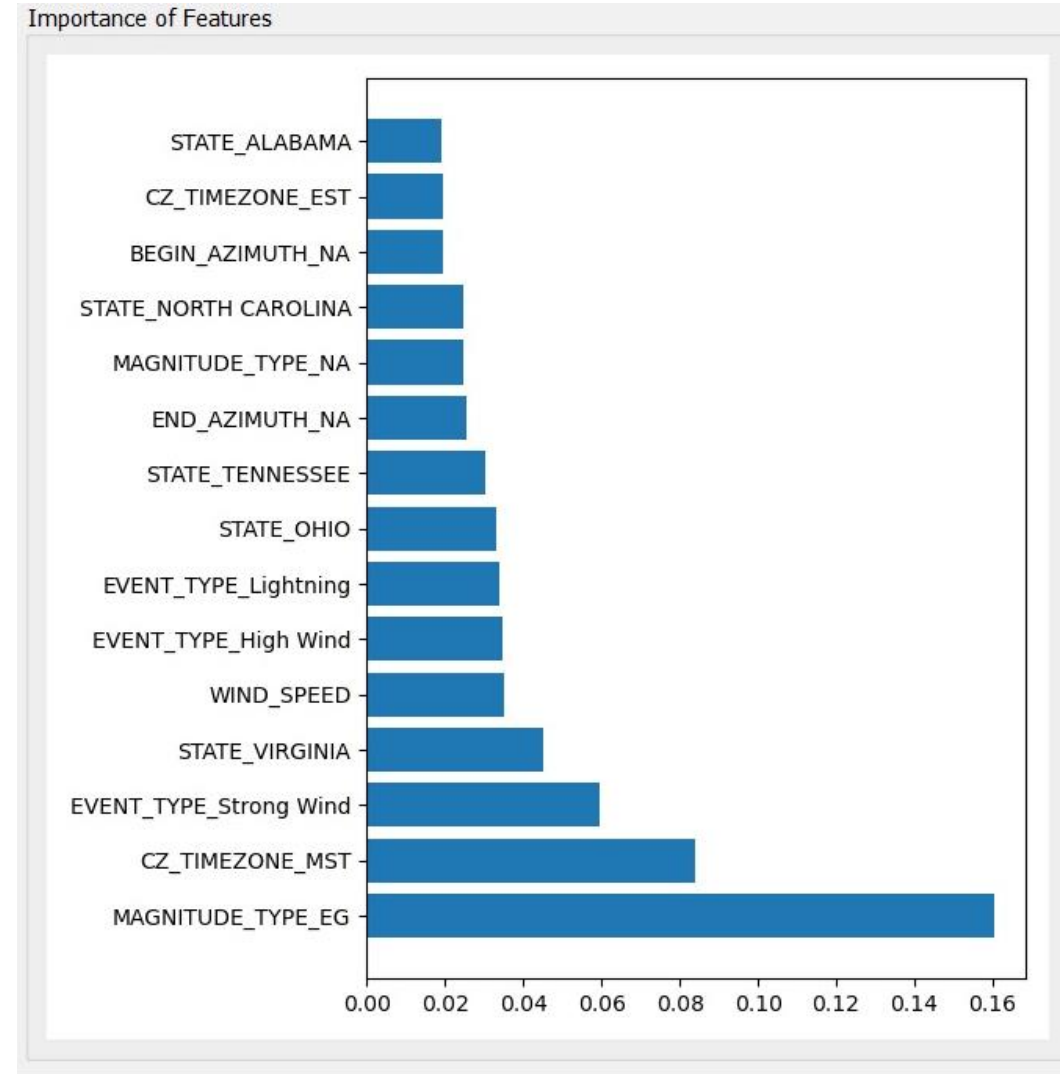
Xtreme Gradient Boosting Regressor

- Gradient boosting refers to a class of ensemble machine learning algorithms constructed from decision tree models. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, “gradient boosting,” as the loss gradient is minimized as the model is fit.
- Extreme Gradient Boosting, or XGBoost for short, is an efficient open-source implementation of the gradient boosting algorithm. XGBoost is a powerful approach for building supervised regression models.
- We are using XGBoost Regressor to help NOAA predict a continuous value: Predict future Damage Property



Xtreme Gradient Boosting Regressor

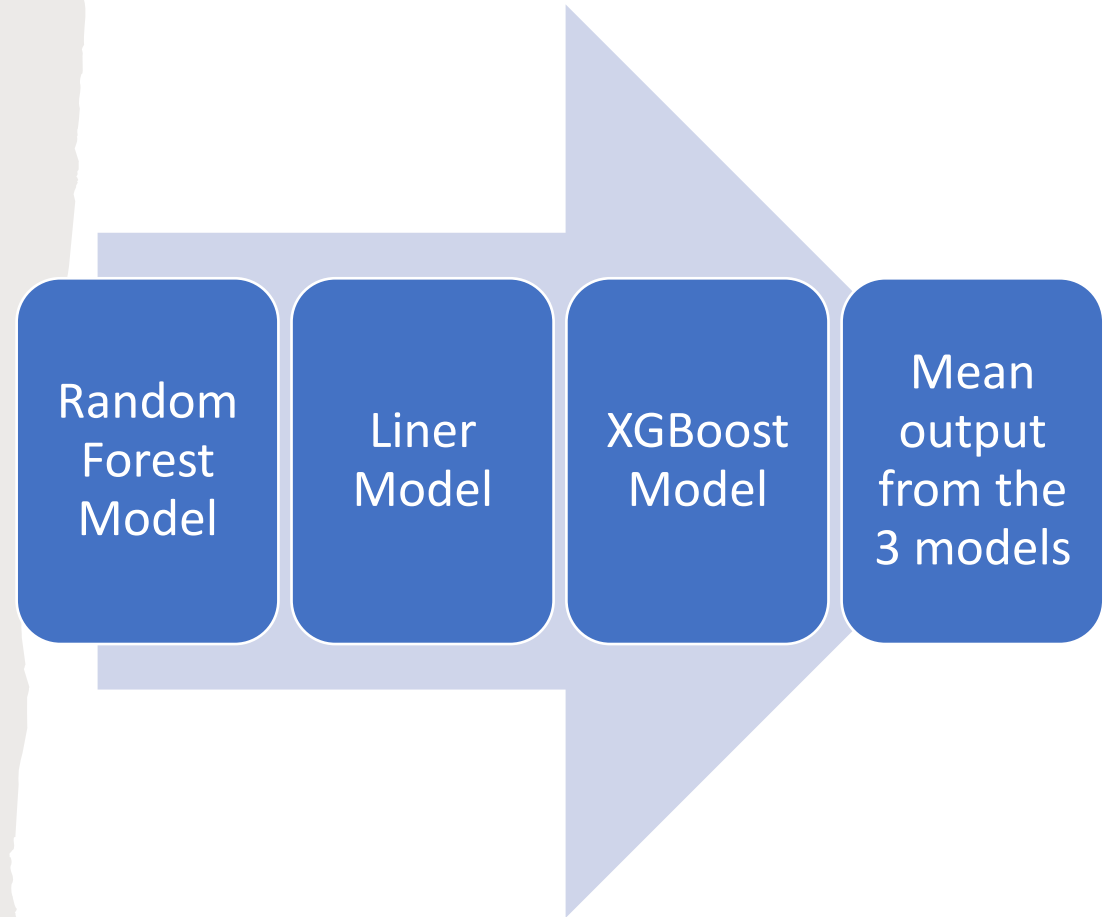
- Model Training Parameters:
 - learning_rate =0.01
 - subsample =0.7
 - max_depth=5
 - n_estimators=500
- Coefficient of determination (R^2 score):
 - Training: 0.45
 - Validation : 0.44



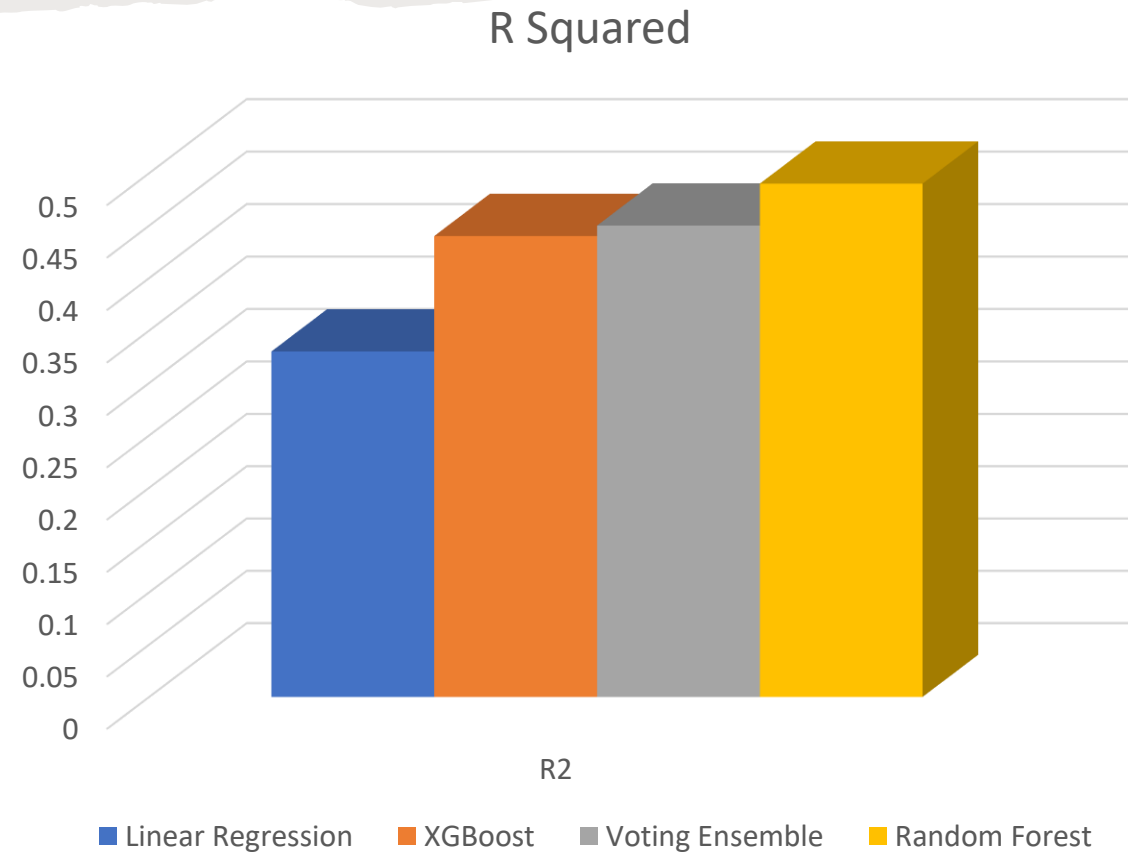
Voting Ensemble

The ensemble model uses individual model predictions and then averages them out to form a final prediction.

- Model Training Parameters:
- Coefficient of determination (R^2 score):
 - Training: 0.44
 - Validation : 0.45
- MSE: 146355



Conclusion





Questions?

