Storm Damage Prediction

“A storm is coming”

Image Source: blogspot.com

Problem Statement

- **Predict** the **total 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

Data Preprocessing

Image Source: blogspot.com

Data Cleaning

- Data Integration.
- Data from 2015 to 2020
- Removing records where the target variable(“Damage Property”) 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

Exploratory Data Analysis

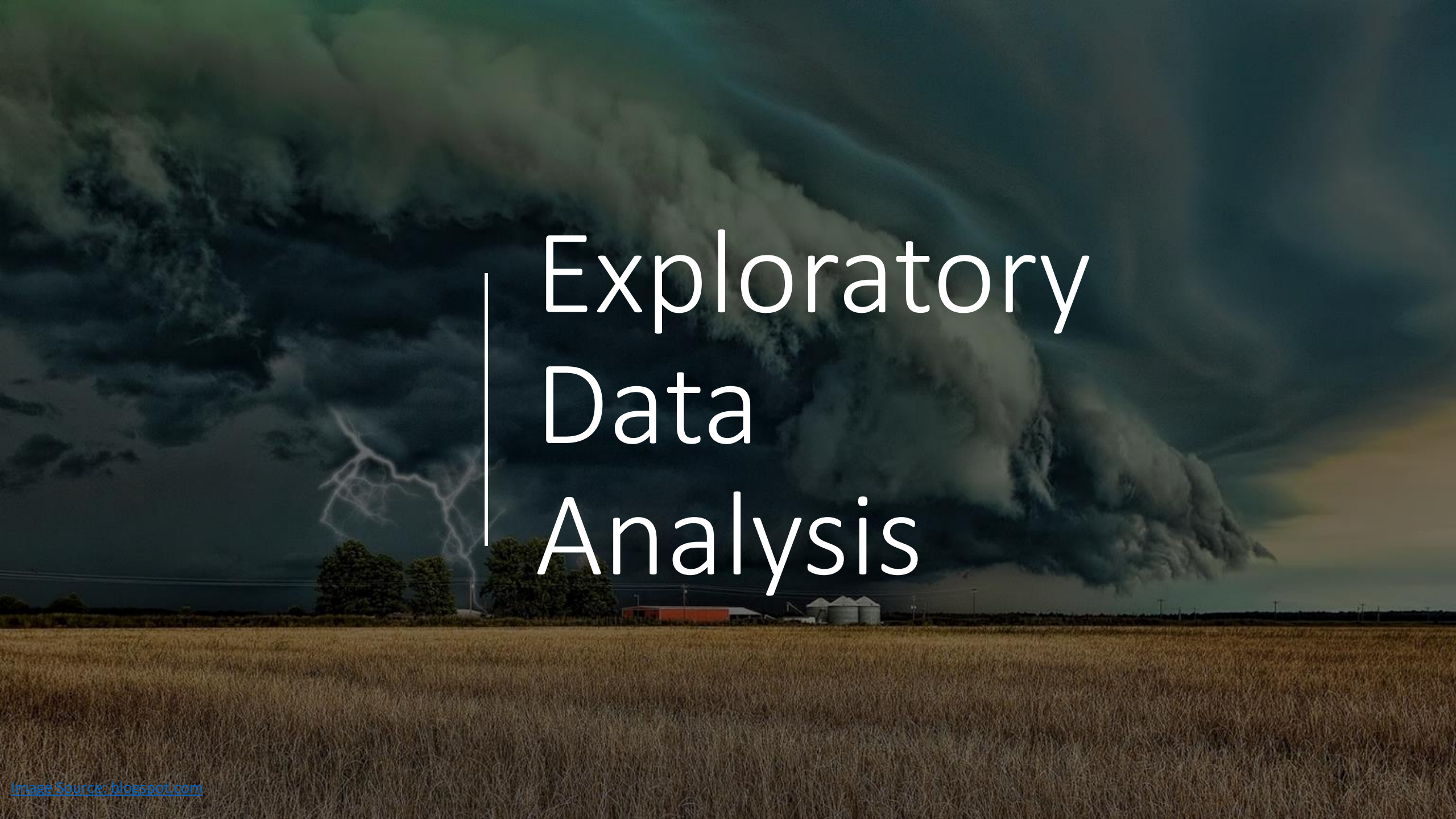
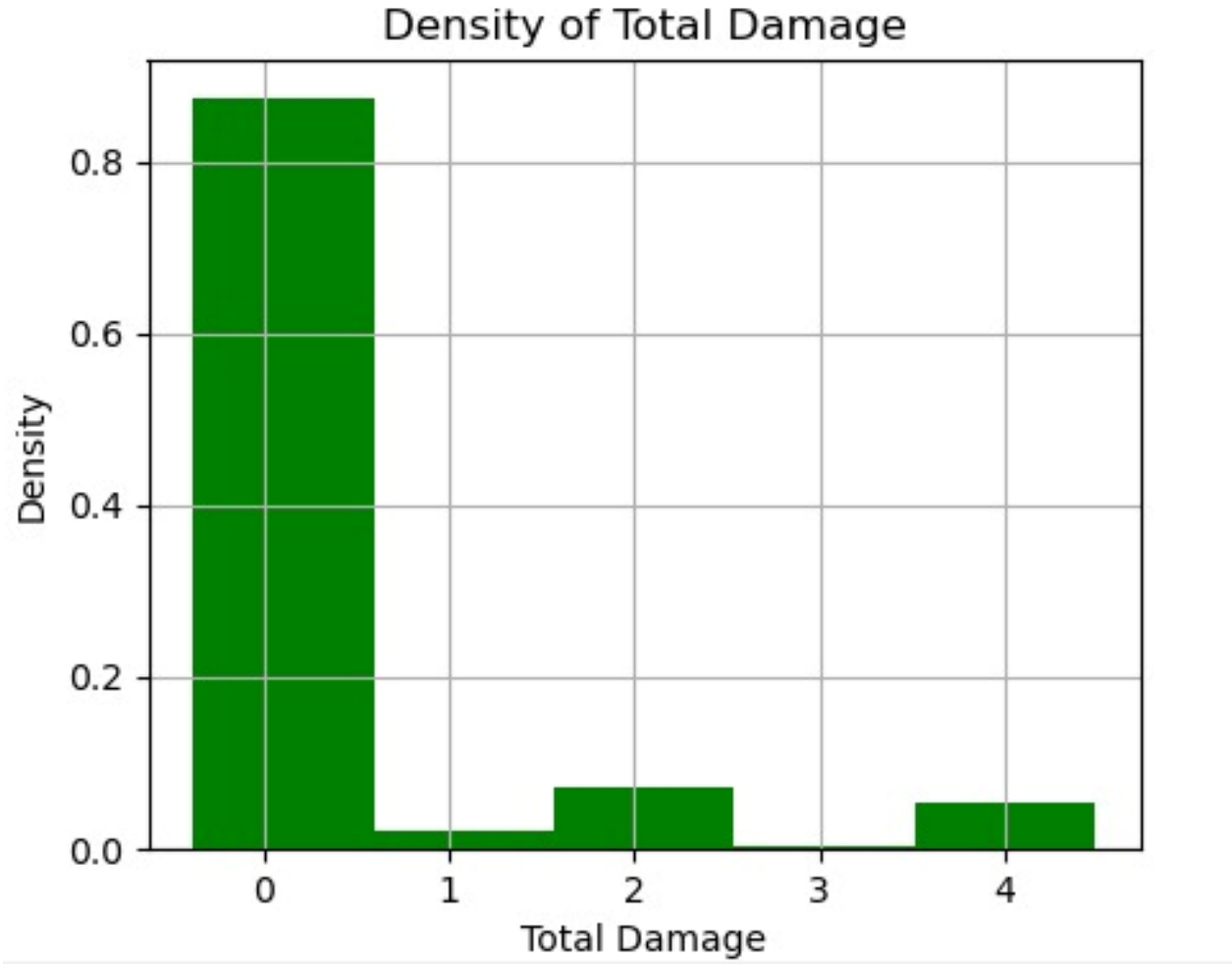
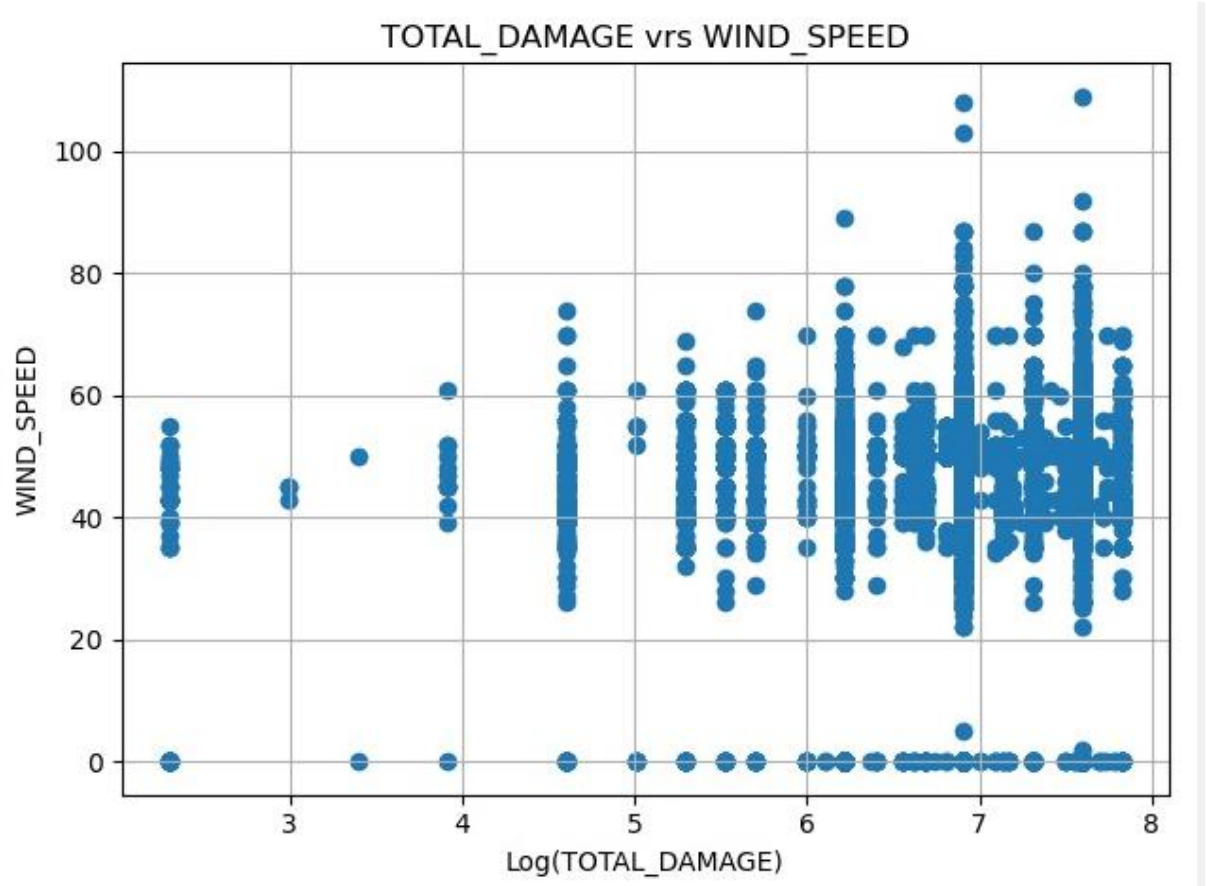


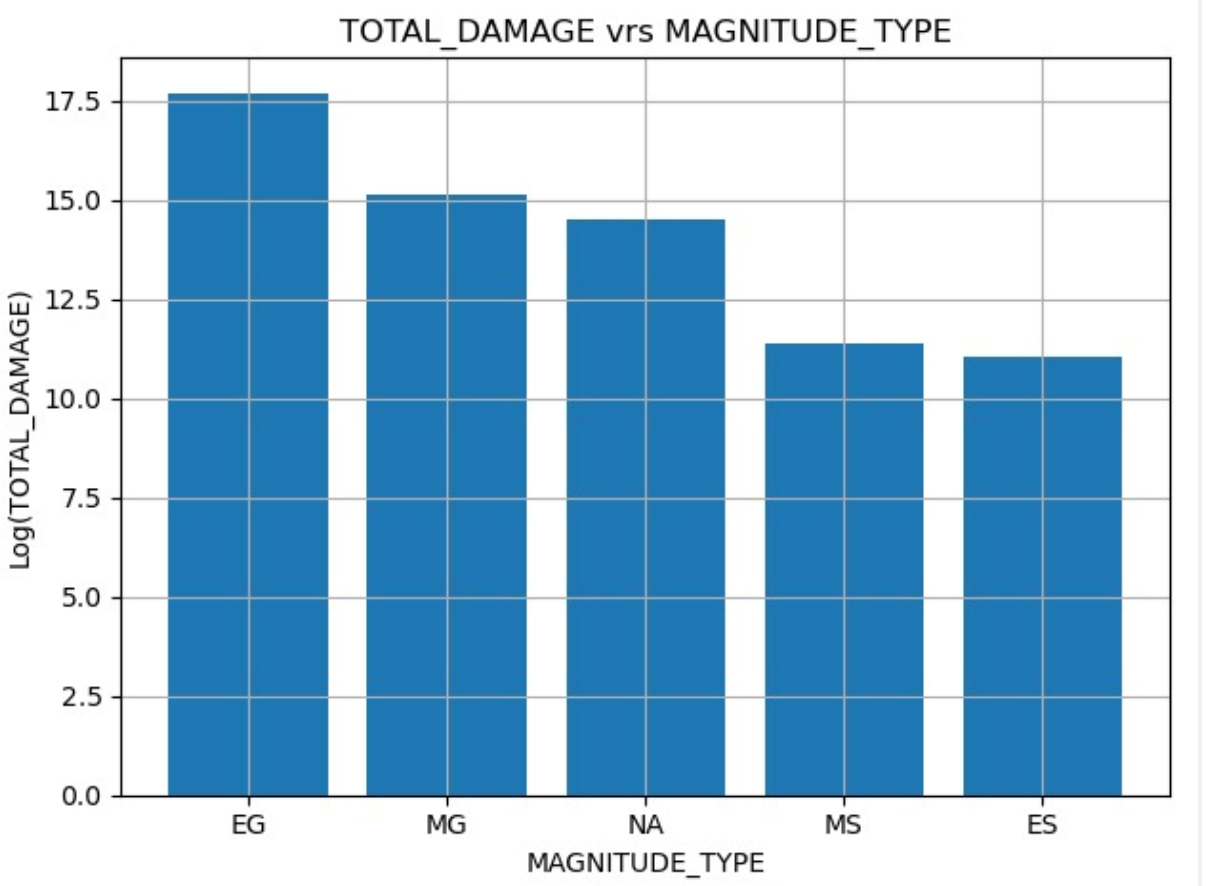
Image Source: blogspot.com

Trend for the Target Variable

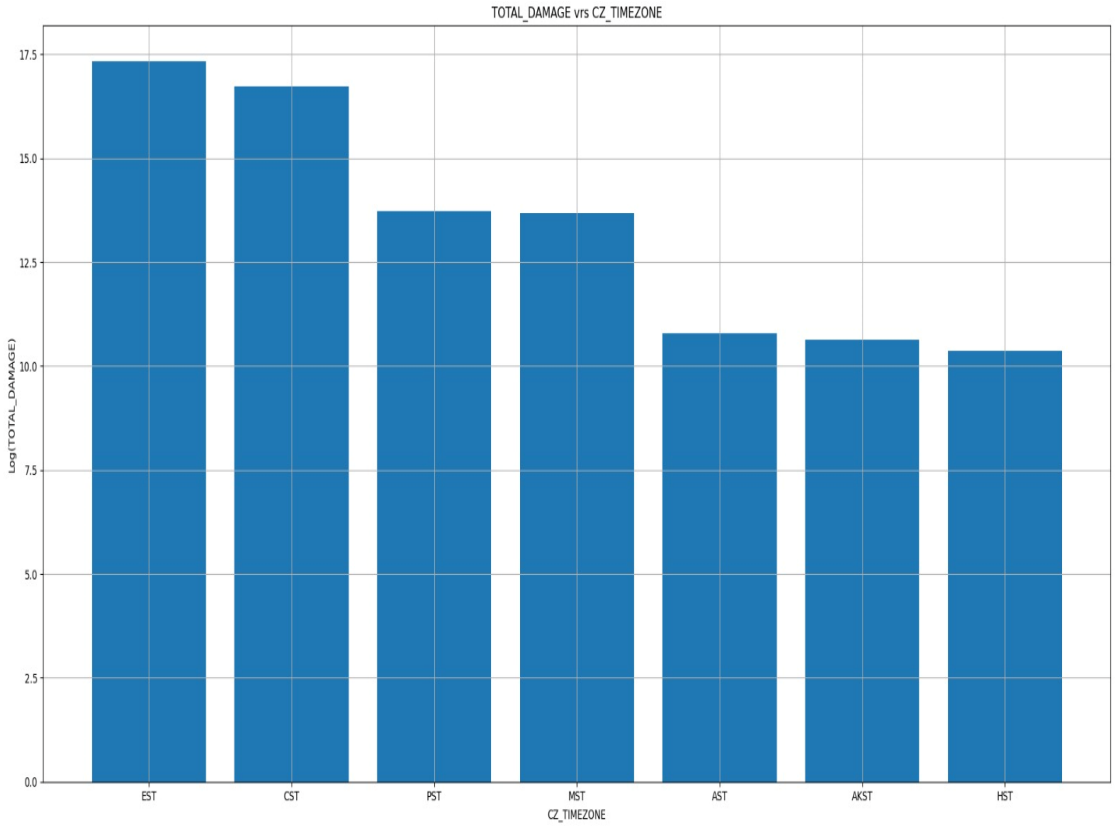


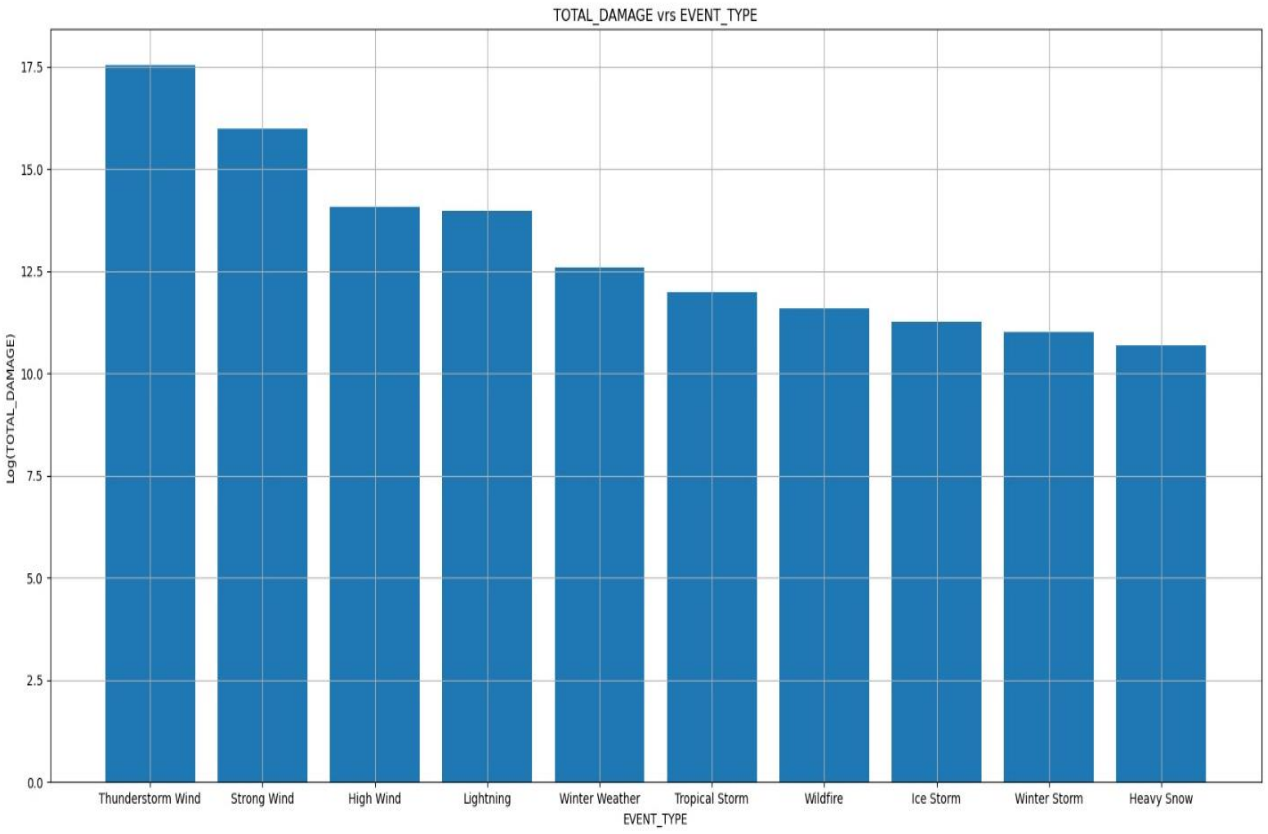
Magnitude and Wind Speed



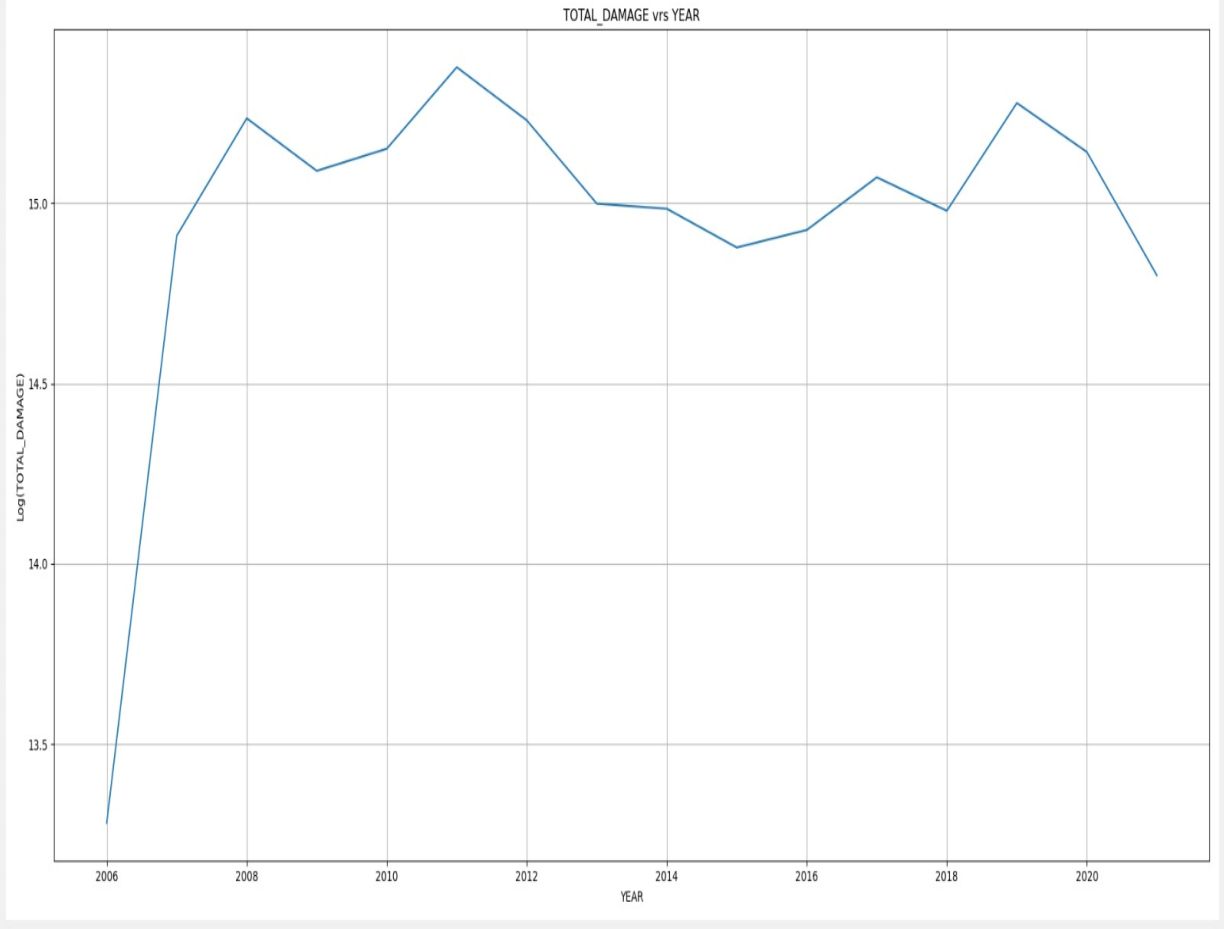


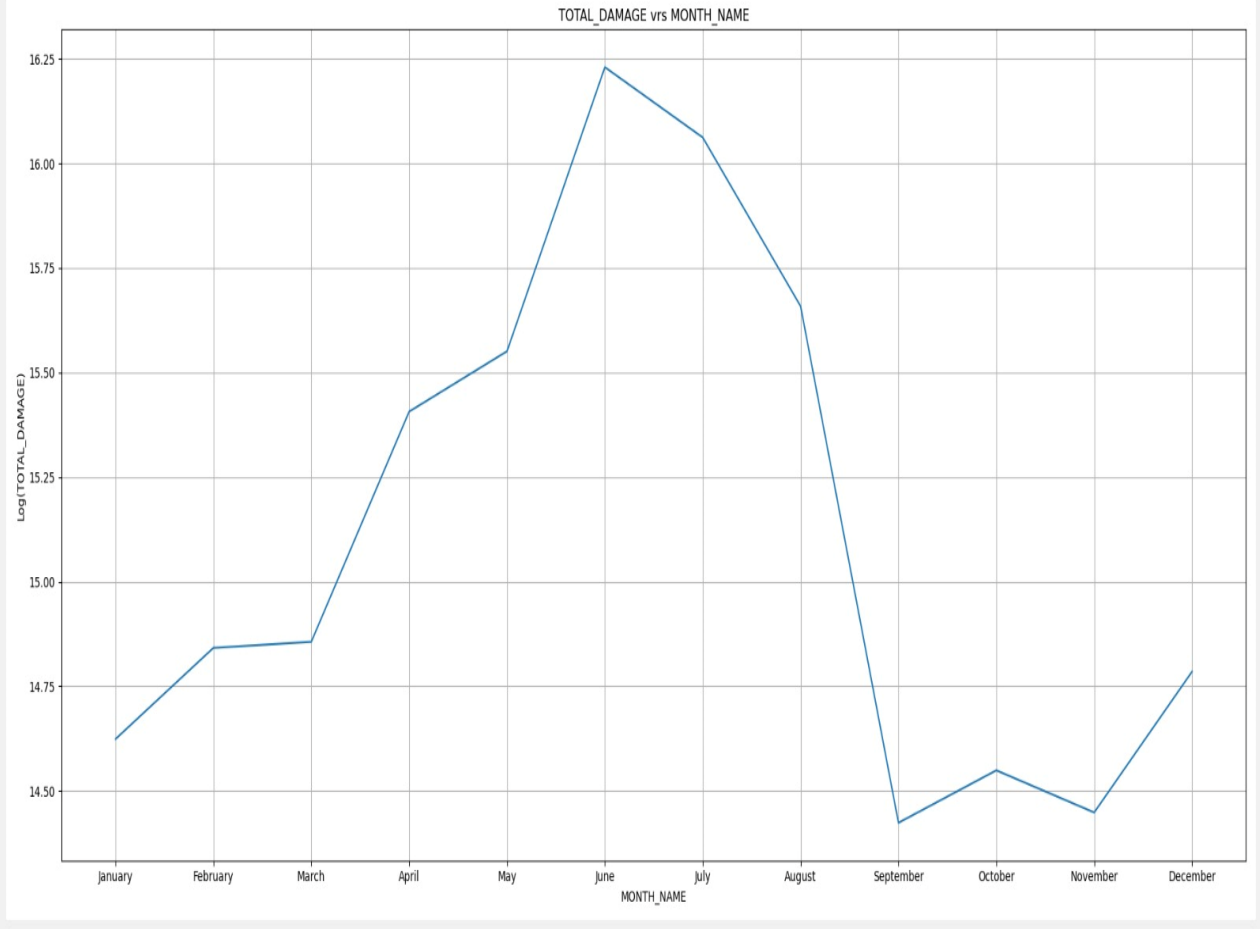
Event Type and Time Zone





Month and Year



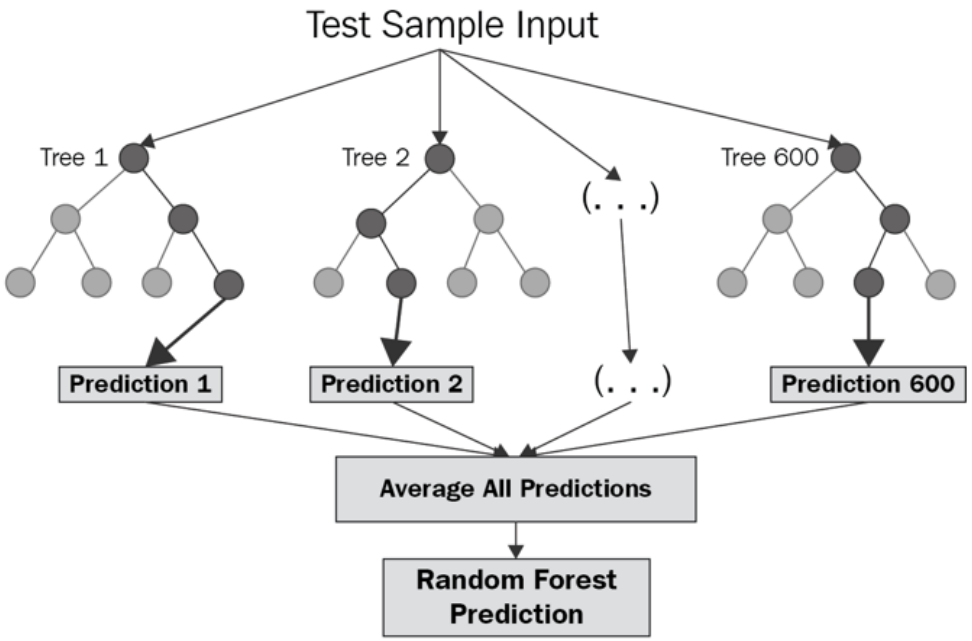


Regression Data Models

Image Source: blogspot.com

Random Forest Regressor

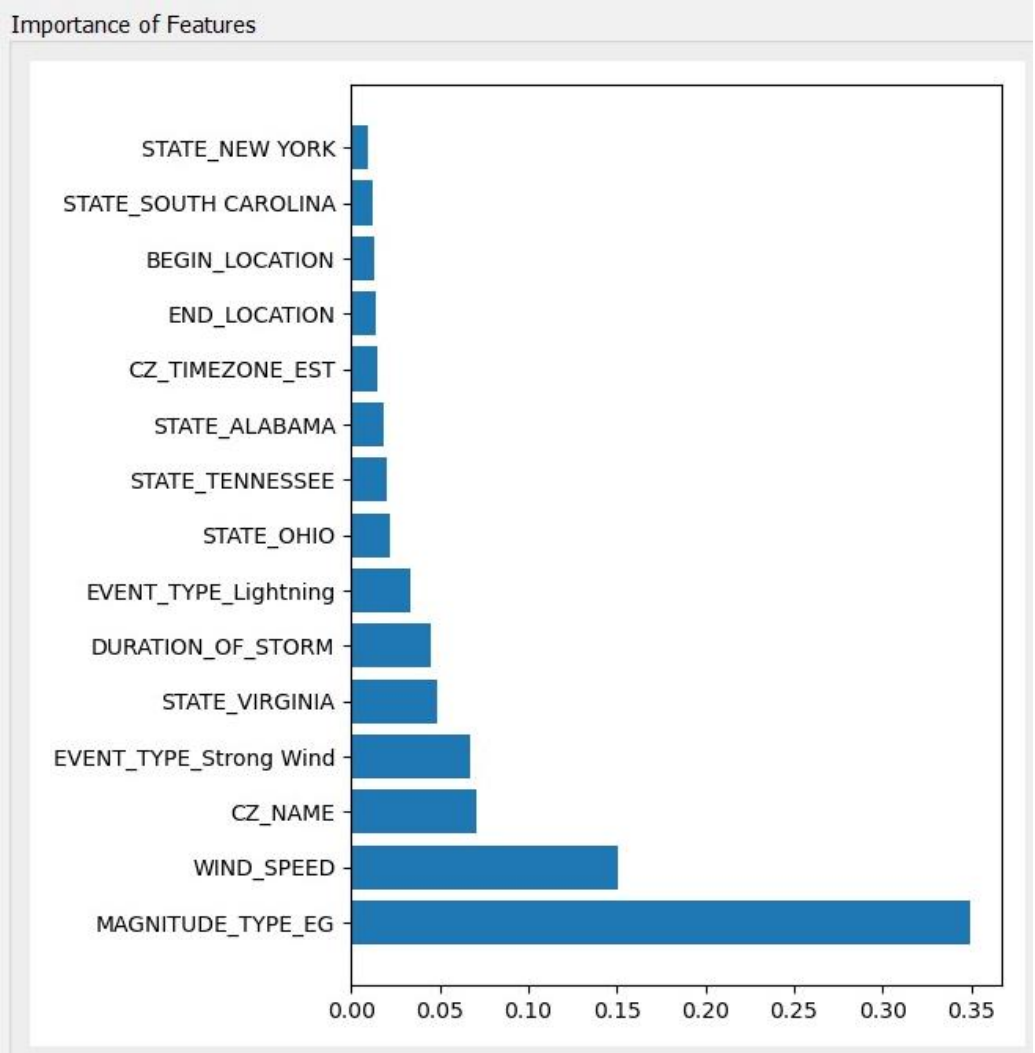
- A **supervised learning** algorithm that is based on the **ensemble learning** method and many Decision Trees. Random Forest uses a **Bagging technique**, so all calculations are run in parallel and there is no interaction between the Decision Trees when building them.
- We are using Random Forest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



Random Forest Regressor

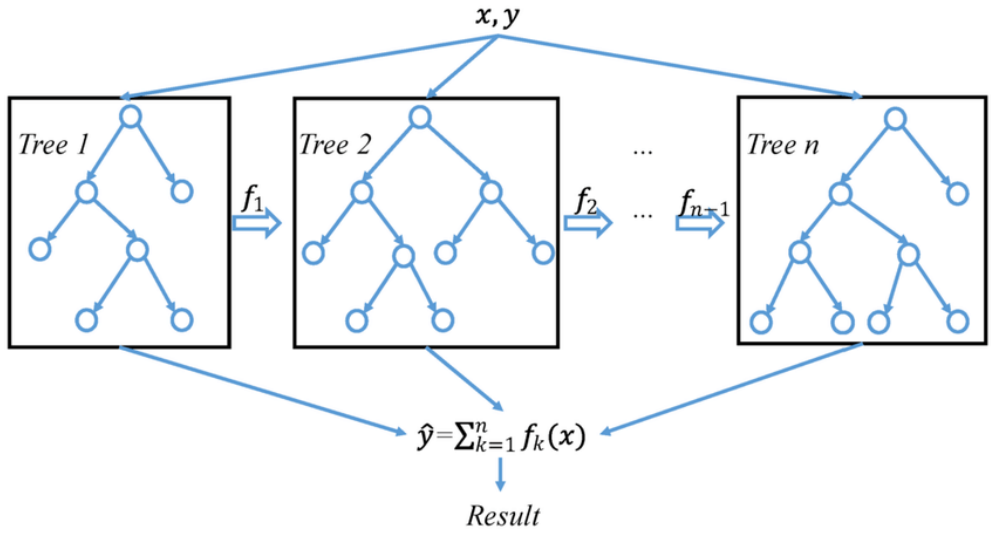
- Model Training Parameters:
  - n_estimators= 100,
  - oob_score = 'TRUE’,
  - n_jobs = -1,
  - random_state =50,
  - max_features = "auto",
  - min_samples_leaf = 50

- Coefficient of determination (R^2 score):
  - Training: 0.53
  - Validation : 0.50
- MSE: 132421



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

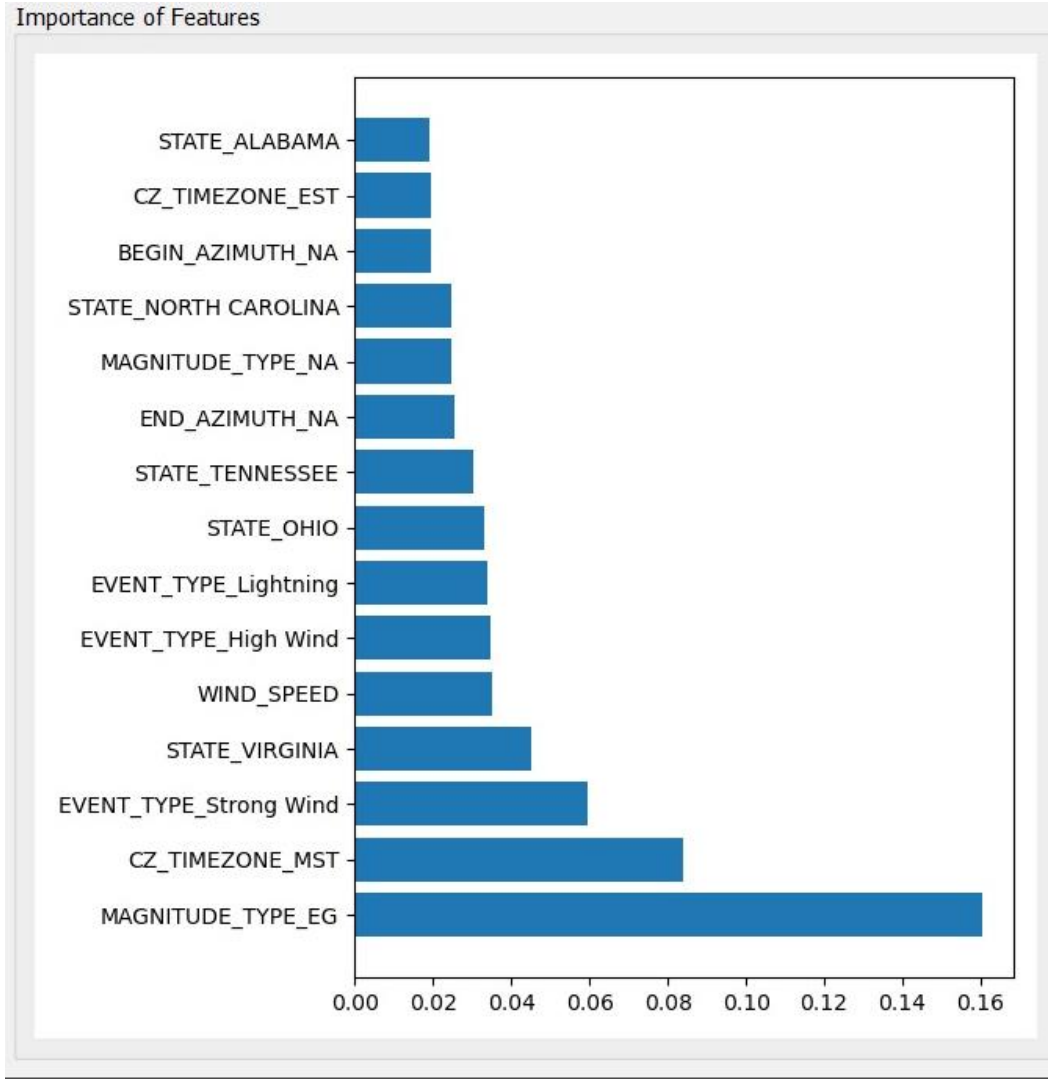


Xtreme **G**radient **B**oosting Regressor

- 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

- MSE: 146355



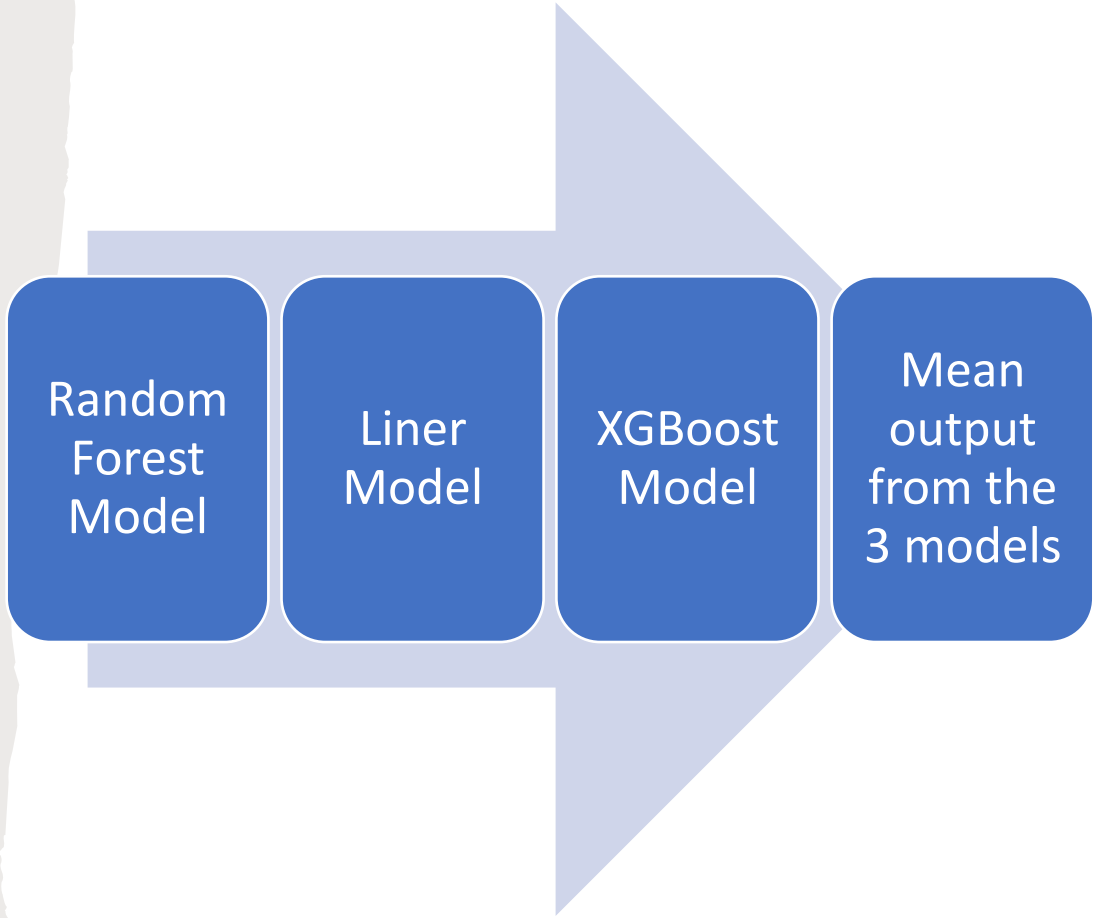
Voting Ensemble \

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

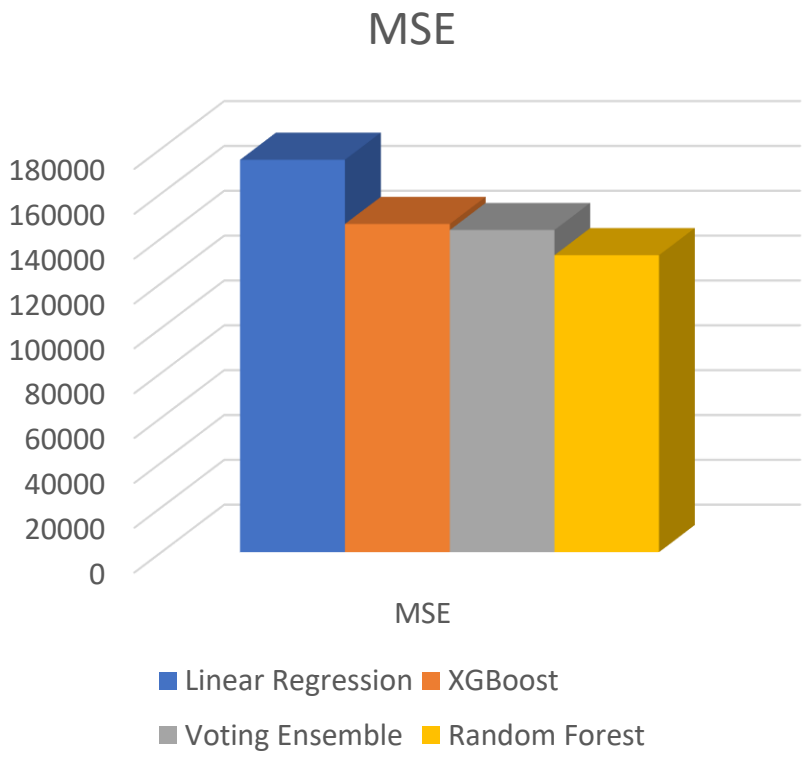
- Model Training Parameters:

- Coefficient of determination (R^2 score):
  - Training: 0.44
  - Validation : 0.45

- MSE: 146355



Conclusion


MSE
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
MSE
Linear Regression
XGBoost
Voting Ensemble
Random Forest

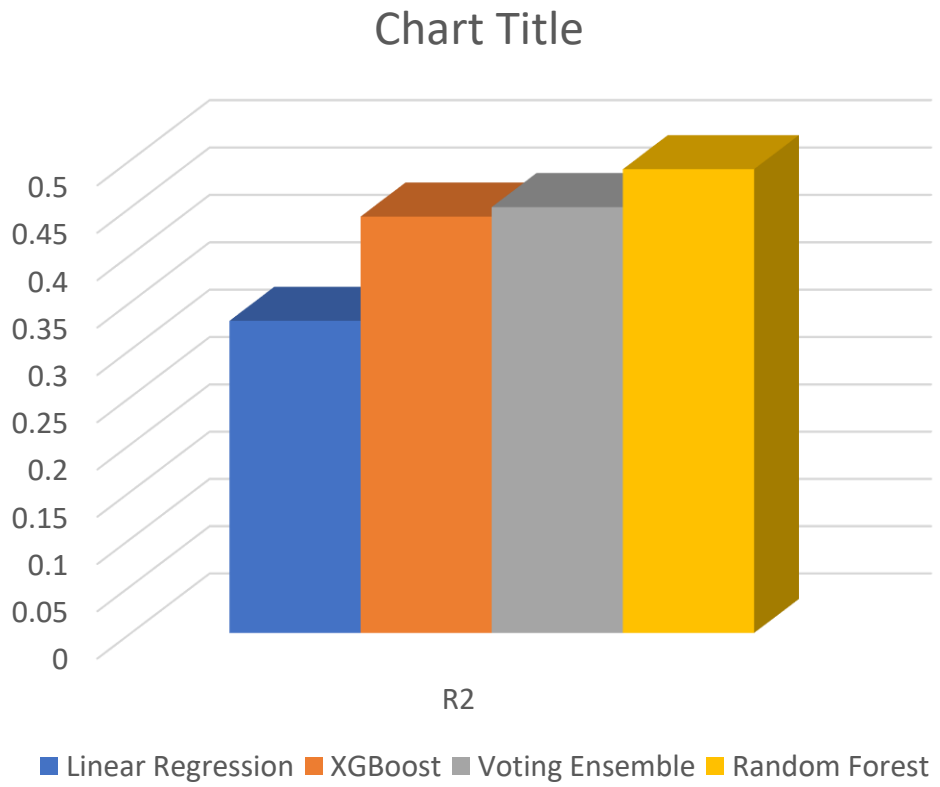


Chart Title
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
R2
Linear Regression
XGBoost
Voting Ensemble
Random Forest

Questions?