Introduction

A new research field emerged in climate science in the early 2000s that wanted to explore the increasing prevalence of extreme weather events like floods, storms, cyclones, etc. The field is known as "extreme event attribution" and has gained momentum in recent years in media in addition to the scientific world. There is mounting evidence that human activity is to blame for the increased risk of these extreme weather-type events. Researchers have also given importance to analyzing the economic costs linked to the human contribution to weather events. A study in 2020 approximated that nearly $67bn of damages caused by Hurricane Harvey in 2017 could attribute to human influences on climate. There are numerous methods to carry out attribution analysis. One way is to record instances of an extreme weather event and see their frequencies change with changes in environmental factors. We aim to build a model that accurately predicts the estimated damage to property while considering various event-related factors, in addition to external factors that might be influencing the extent of the damage.

Dataset

For this project, we have used publicly available data from the National Oceanic and Atmospheric Administration (NOAA) that contains event details on disaster incidents occurring in the US ranging from 1950 to August 2021. Some of the variables that we use from this dataset are as follows:

- begin and end date-time of event
- state where the event occurred
- the type of event (Hail, Storm, Drought, etc.)
- number of injuries and deaths
- starting and ending latitudes and longitudes of the event

The complete data dictionary for reference is accessible through this link.

We have also pulled in environmental indicators from yearly data collected by the United States Environmental Protection Agency (EPA). We have joined this data as additional information against the event year. The datasets that we have considered from the EPA source are as follows:

- emissions of greenhouse gases from 1990 to 2019
- events of heavy precipitation by land area percentage
- yearly earth surface temperature
- CSIRO and NOAA data for yearly sea-level changes
- variations in average seasonal temperature for fall, winter, summer, and spring
- artic ice coverage in March (yearly high) and September (yearly low)
- Glacier mass balance and number of observed glaciers

Additional dataset information is available at this source.

Algorithms Used:

Linear Regression:

It is a method to model a relationship between one or more independent variables and a response variable by fitting a linear equation on the observed data. Regression tells us the value of the response variable for an arbitrary explanatory variable value. The regression equation is:

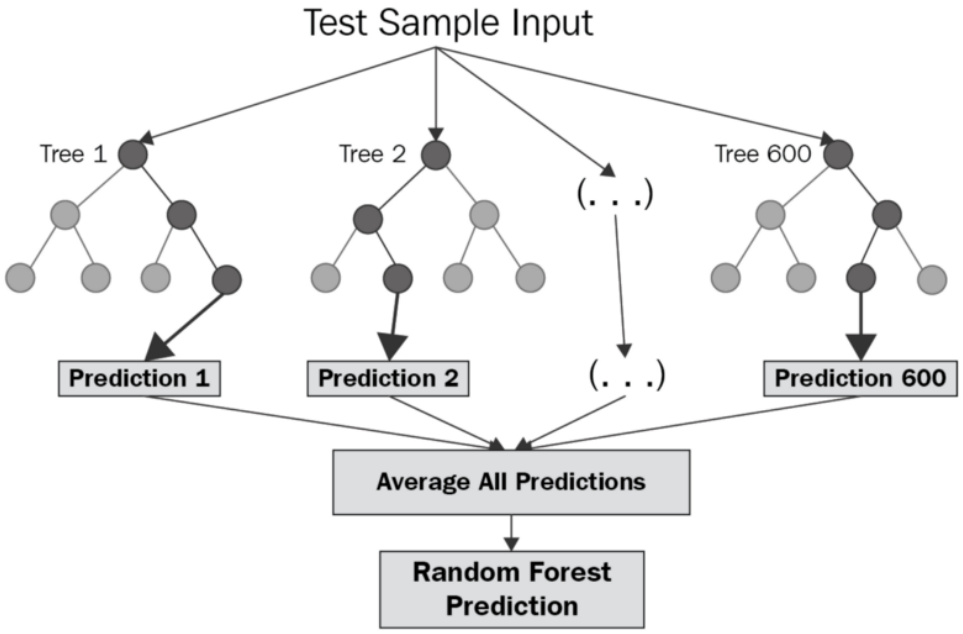
$\hat{y} = b0 + b1\,x1 + b2\,x2 + \ldots$ where
b_0: intercept
b_i: slope/rate of change

Bootstrap aggregation:

Bootstrapping is a sampling technique to create subsets of observations from the original data and is also known as bagging. In this technique, a generalized result combines the results of various predictive models. The subset size for bagging may be smaller than the original dataset.

Random Forest Regression:

It is a supervised machine learning algorithm that uses bagging to solve regression and classification problems. The algorithm works by training multiple decision tree estimators concurrently and outputting the mean or mode of all the individual predictions. It helps against individual trees overfitting the data and getting stuck in locally optimal solutions.



Random Forest prediction working [3]

Extreme Gradient Boosting Regression:

Gradient boosting is a class of ensemble machine learning algorithms constructed from decision tree models. It fits the model using any arbitrary differentiable loss function and gradient descent optimization algorithm. This technique is known as gradient boosting as we minimize the loss gradient while training the model.

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.

Experimental Setup:

Extraction:

The NOAA data files are extracted from this link and have the following naming structure: StormEvents_details-ftp_v1.0_d1950_c20210803.csv.gz.

The files are then concatenated into one to form our source data frame. We save this as a pickle file.

Preprocessing:

- The replace_str2num() function cleans up the DAMAGE_CROPS and DAMAGE_PROPERTY variables to convert them into numeric values.
- The winds() and hail() functions split the MAGNITUDE variable based on the values of MAGNITUDE_TYPE into WIND_SPEED and HAIL_SIZE.
- The missing_swap() function imputes missing values for variables where the counterpart has valid values. For example, if the BEGIN_LAT is present and the END_LAT is not present, we fill it with the BEGIN_LAT value.
- The calc_duration() function calculates the time difference between the event start and end.
- The geo_distance() function uses the Haversine formula to calculate the geographical distance covered by the event (Tornado, etc.)
- The dict_mapping() function replaces junk values from CZ_TIMEZONE, BEGIN_AZIMUTH, and END_AZIMUTH with appropriate values.
- We use the EVENT_TYPE variable to derive three variables: COLD_WEATHER_EVENT, WINDY_EVENT, and WATER_EVENT, based on keywords like Snow, Storm, Hurricane, etc.
- Tor_scale() converts the values of F_Scale for the tornado strength into numeric values.
- Then, we fill the missing values in continuous variables with 0 and the categorical variables by N/A.
- We use Pandas to read the various EPA data CSV files and collate them into one. We interpolate the missing data ranging back to the year 1950 by using the impute_EPA_data() function. We use the interp1d method from SciPy to get the extrapolated variable values.
- Finally, we join the entire data into one data frame and remove the outliers from all the numerical variables.

Modeling:

We encode the categorical variables with many unique values using the mapping() function on the range of distinct values. We use the get_dummies() method to split the other categorical variables. Our response variable TOTAL_DAMAGE is the sum of the DAMAGE_PROPERTY and DAMAGE_CROPS.
We train and compare the efficiencies of four different models. Here is the information about these models, along with their training parameters.

1. Linear Regression with default parameters
2. Random Forest Regressor with parameters as follows:
   i) n_estimators=100
   ii) oob_score='TRUE'
   iii) n_jobs=-1
   iv) random_state=50
   v) max_features="auto"
   vi) min_samples_leaf=50
3. Extreme Gradient Boosting Regressor with parameters as follows:
   i) learning_rate=0.01
   ii) subsample=0.7
   iii) max_depth=5
   iv) n_estimators=100
   v) colsample_bytree=0.8
4. Ensemble model which is a combination of above three models using VotingRegressor

We compare the performance of the different models by making use of the following metrics:

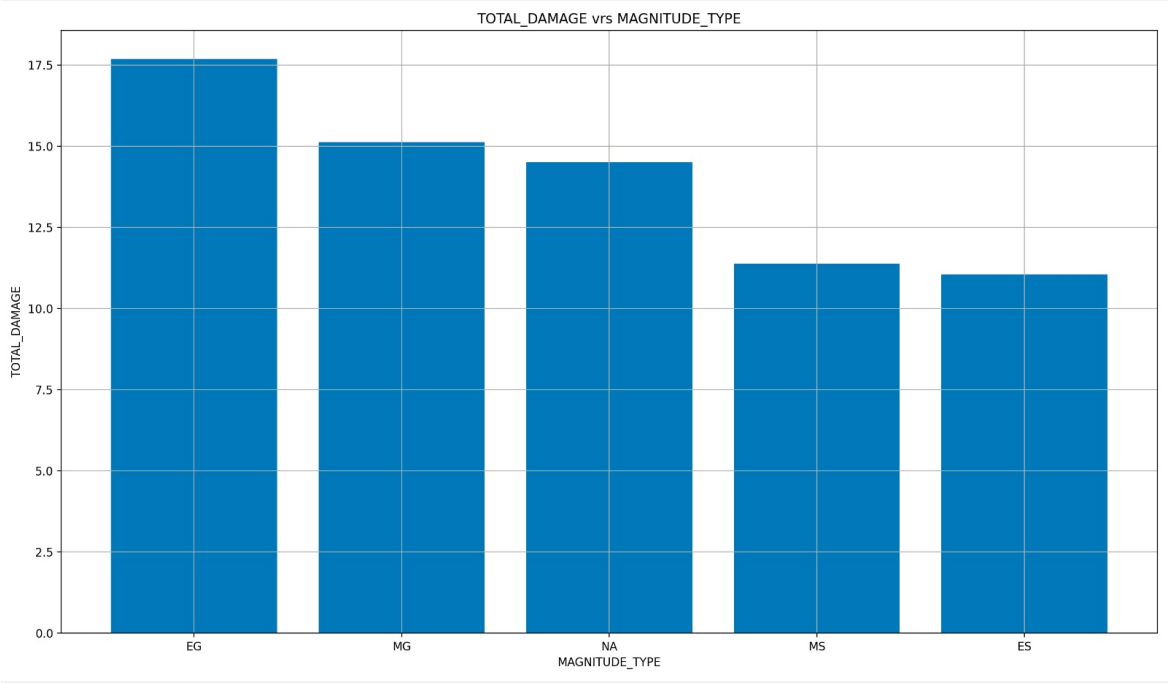
1. Mean squared error
2. Train R-squared value
3. Test R-squared value

Results:

Exploratory Data Analysis:

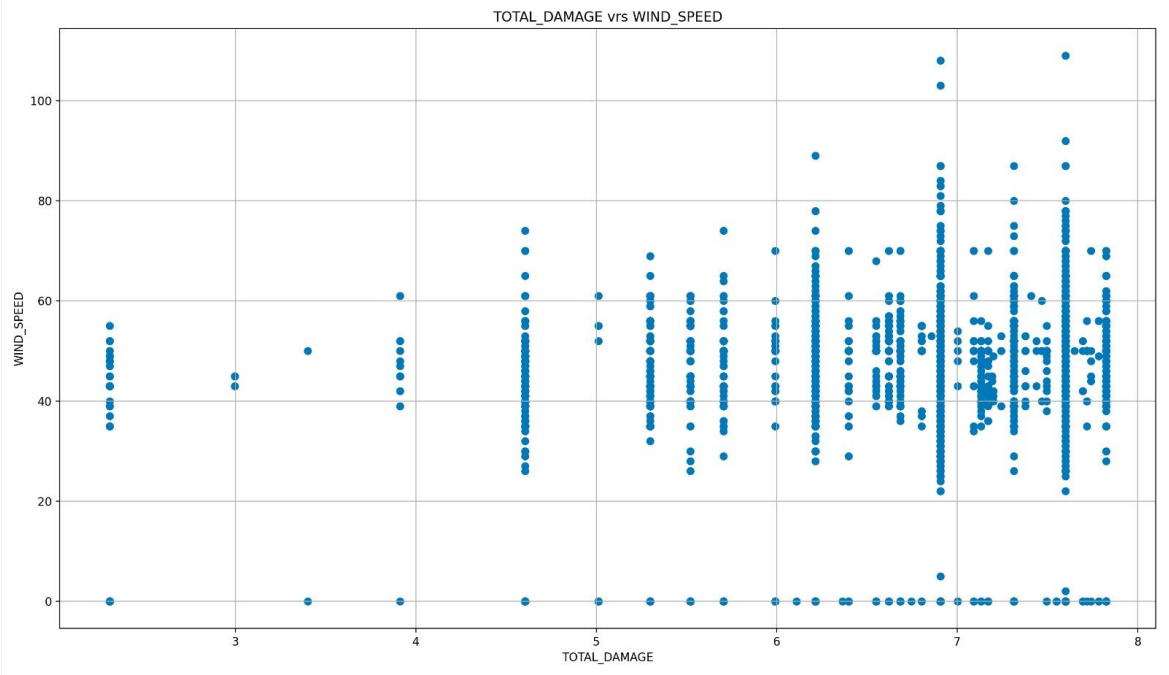
After obtaining the cleaned dataset, our objective was to get better insights about our data so that we can fix any data inconsistencies and get a clearer picture of event attributes that are explaining the variance in our target variable TOTAL_DAMAGE. We start by plotting the distributions of our target variable against various features to gauge their overall importance in our final model. For our plots, we take the logarithm of the total damage sum across different groups to plot our graphs.

Plot 1:



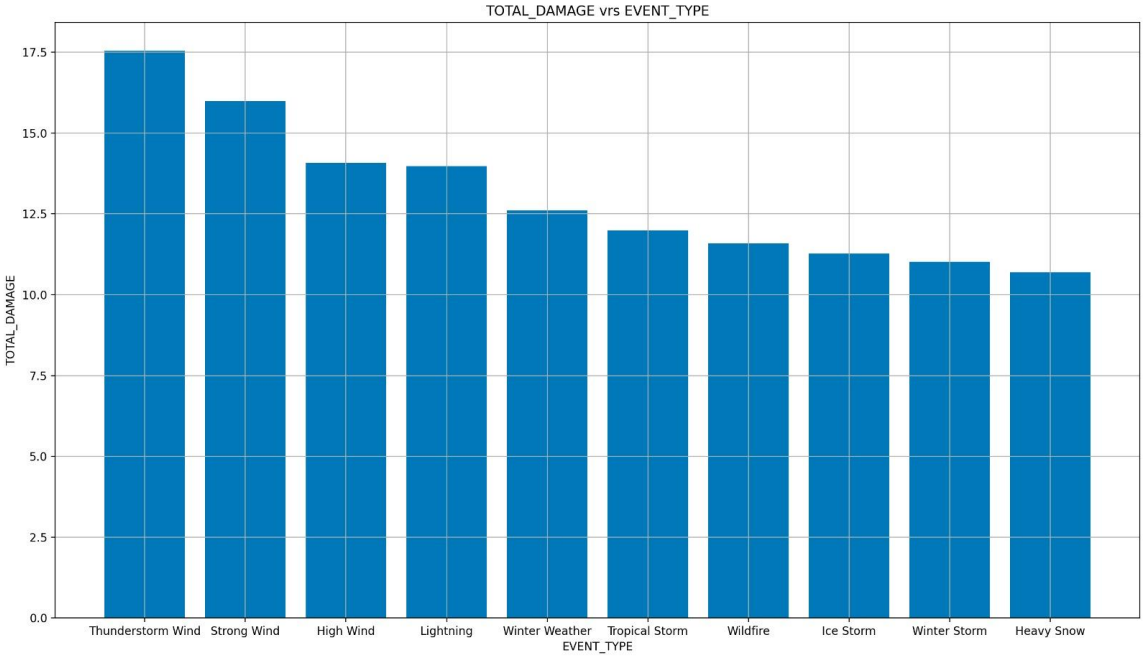
We see that Wind estimated gust (EG) has the highest total damage and Estimated Sustained Wind (ES) has the least.

Plot 2:



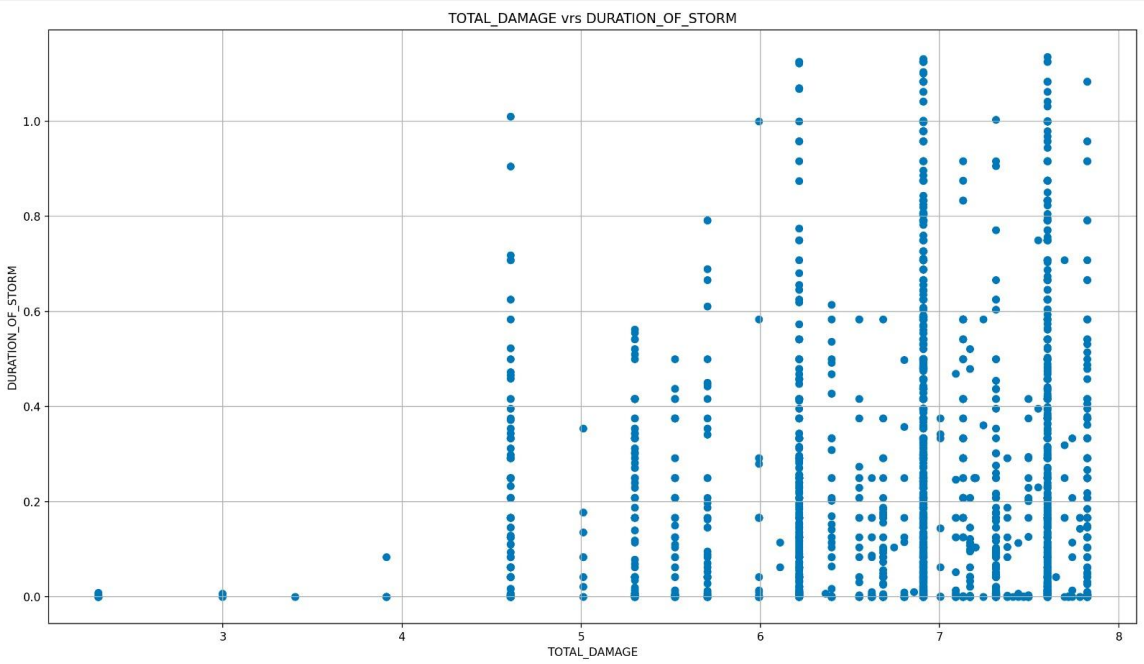
We see a slight positive correlation between wind speed and total damage.

Plot 3:



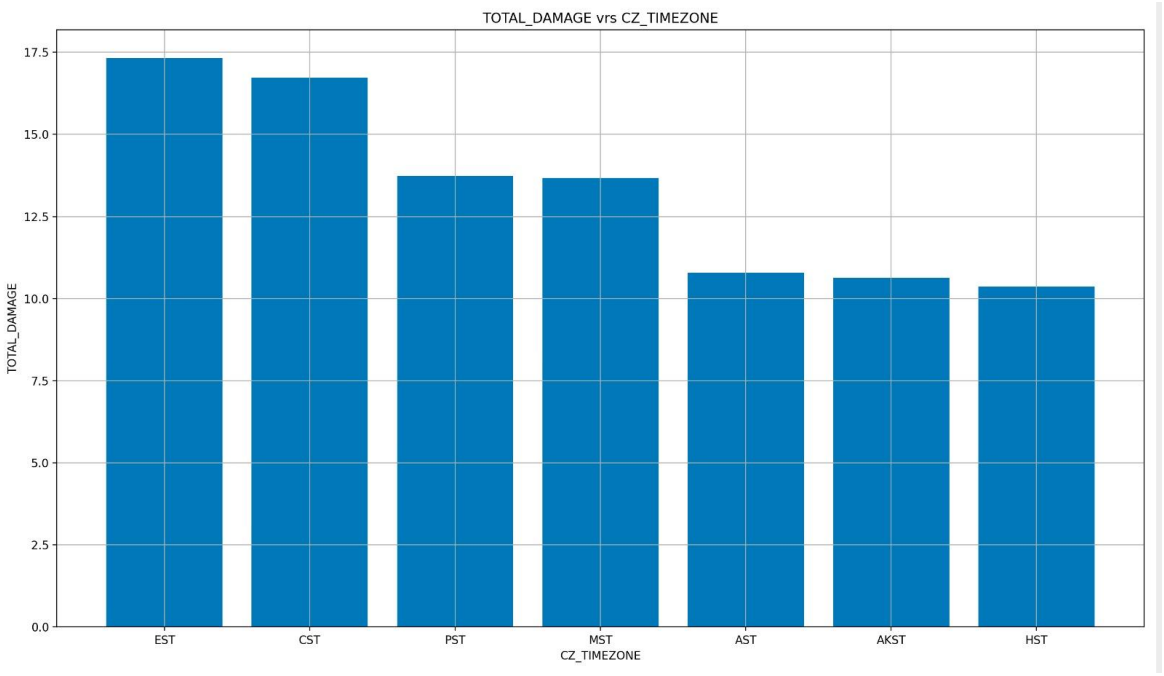
This is a list of events that have the highest TOTAL_DAMAGE. Wind related events seem to have very high damage along with winter-related events.

Plot 4:



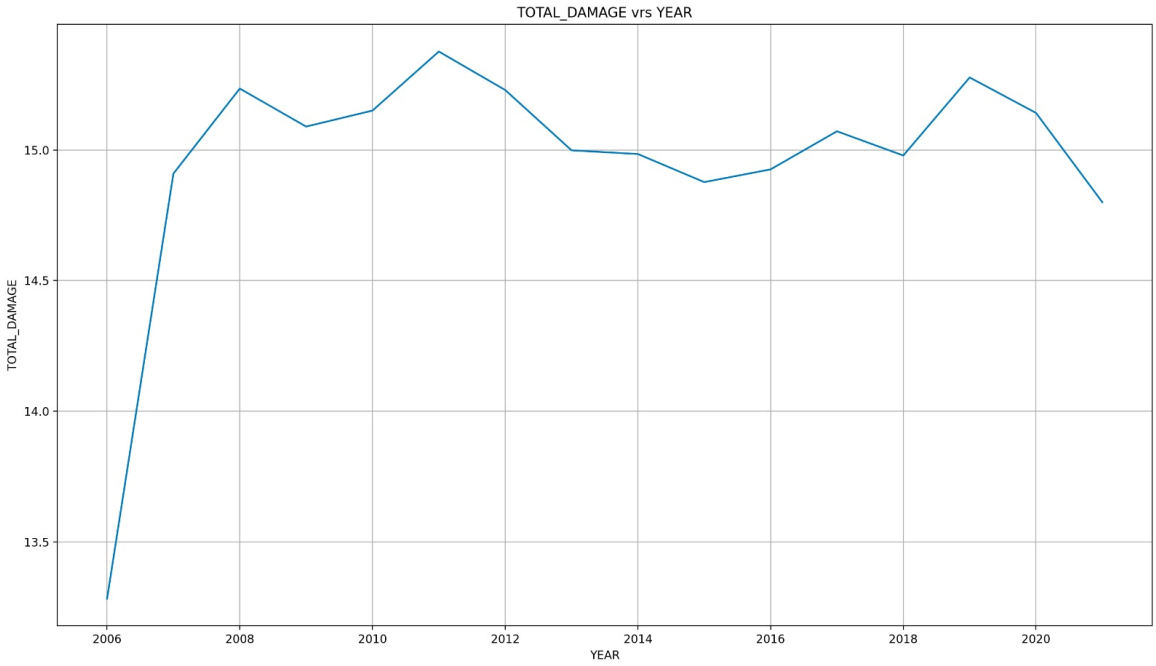
Again, we see a positive correlation between the duration of storm and the total damage caused by it.

Plot 5:



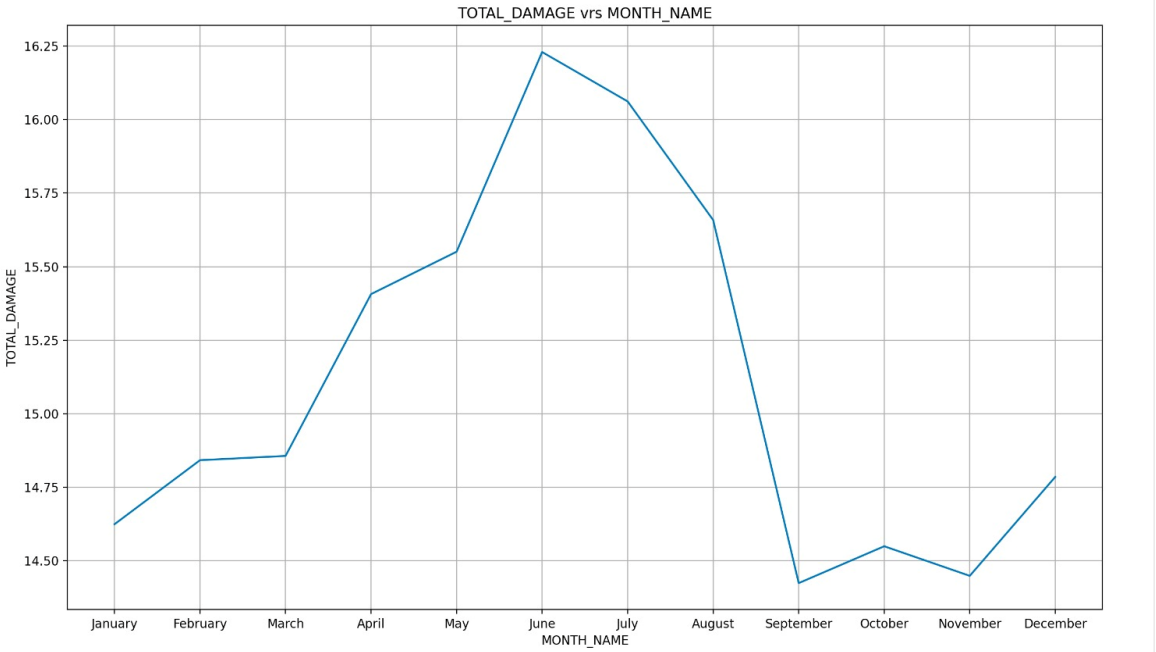
We see from this graph that the Eastern Coast has significantly higher total damage as compared to the Western Coast.

Plot 6:



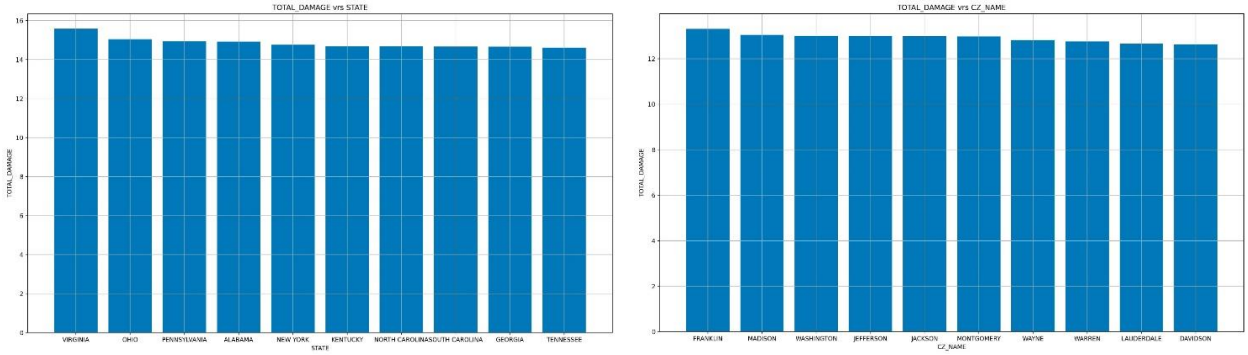
There is a generally uniform amount of total damage every year since 2007.

Plot 7:



The monthly distribution of total damage shows us that the warmer months from April to August have very high total damage as compared to the cooler months.

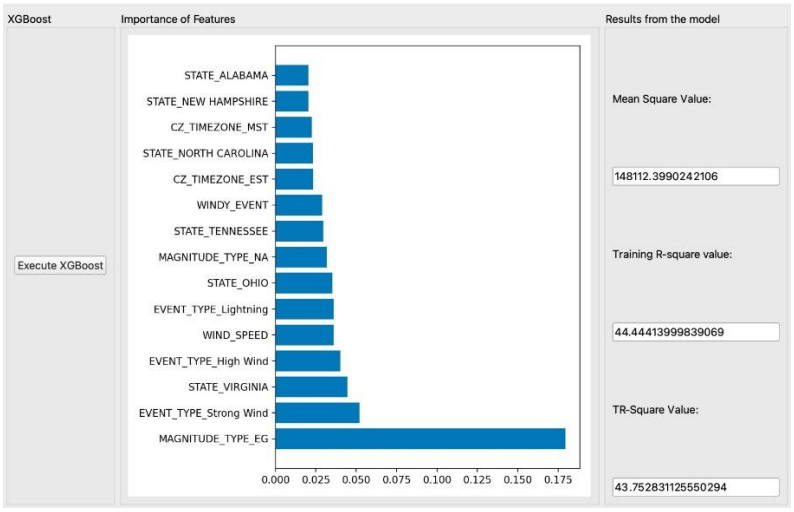
Plot 8:



These are the states and county/zone names with the highest total damage.

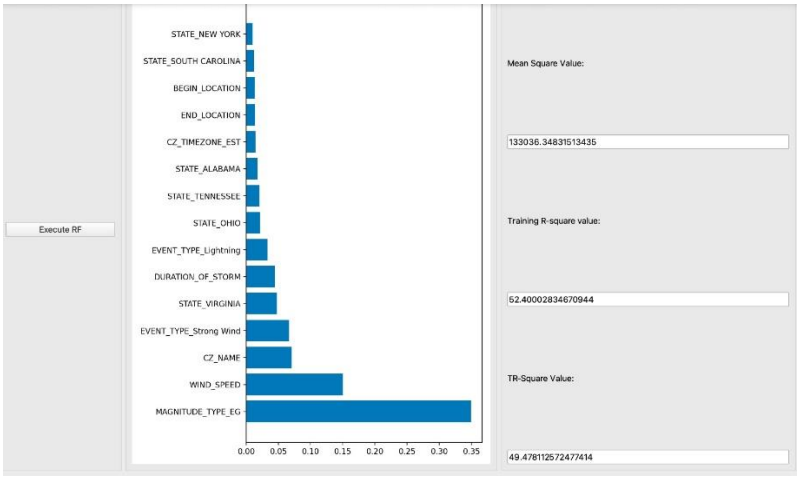
Modeling:

For XGBoost Regressor, our feature importances after training the model are given below. The RMSE values and train and test R-squared values are also included in the output.



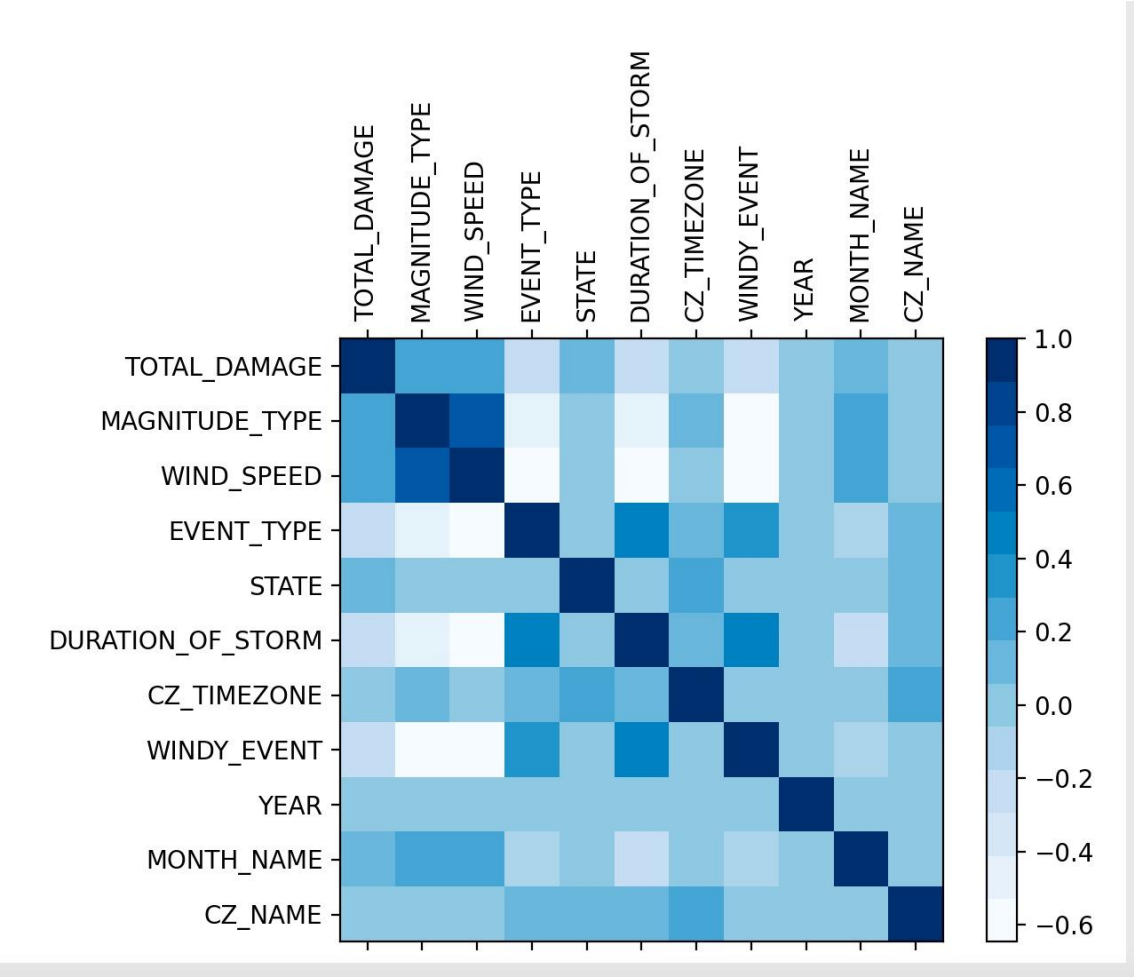
We see that some of our important features from the model come out to be Magnitude_Type, Event_Type, State, CZ_Timezone, and Wind_Speed. Additionally, we get an R-squared value of around 44% which is not that great.

For Random Forest Regressor, our feature importances after training the model are given below. The RMSE values and train and test R-squared values are also included in the output.



Similarly, we see that the important features are Magnitude_Type, Wind_Speed, State, Event_Type, Duration_of_Storm, and CZ_Timezone. We get better performance with R-squared value of about 50%.

A corrplot of our most important features against the TOTAL_DAMAGE variable returns the following result.



Summary and Conclusions

As seen from the results section, we have tried to build a disaster damage predictor by analyzing the attributes of the event and modeling them using regression. We have identified the variables that are having the most effect on the target variable. We have used a variety of models with the Random Forest Regressor giving the best results. We have also used Grid Search for hyperparameter tuning to obtain the best scores. We also see that the environmental indicators like CO2 levels, Arctic Ice coverage, Earth Surface Temperature, etc. do not have any significant impact on our final predictions. This could be attributed to a mismatch in the granularity of the environmental data and the total damage that is caused by these disaster events. Future work will focus on further refining our feature engineering process to improve the performance of our model.

References

1. https://www.carbonbrief.org/mapped-how-climate-change-affects-extreme-weather-around-the-world
2. https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8
3. https://www.oreilly.com/library/view/tensorflow-machine-learning/9781789132212/d3d388ea-3e0b-4095-b01e-a0fe8cb3e575.xhtml