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.

Work Description

Extraction:

My role in the project was to bring in the data from various sources and make the final dataframe ready for modeling. The code to extract the zip files from the NOAA database was done by Kartik initially. I built upon his code to serialize the process so that it picks up all the zip file data from 1950. Building functions to automate the process and save the pickle file in the data folder was done by me.

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. I used 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, I filled the missing values in continuous variables with 0 and the categorical variables by N/A. I used Pandas to read the various EPA data CSV files and collate them into one. I interpolated the missing data ranging back to the year 1950 by using the impute_EPA_data() function. I used the interp1d method from SciPy to get the extrapolated

variable values. Finally, I joined the entire data into one data frame and remove the outliers from all the numerical variables.

GitHub upload and Report Creation:

I was responsible for collecting all the work and combining it into one workable code on the main branch and completing the final report.

Percentage of code:

Written by myself = 130 lines

Copied from the internet = 16 lines

Modified from the internet= 5

Percentage

- ⇨ (16-5) * 100 / (16 + 130)
- ⇨ 7.53%