**PROPERTY DAMAGE PREDICTION**

**Individual Final Report**

**Author**: Kartik Jain

**Email**: [karjain@gwu.edu](mailto:karjain@gwu.edu)

**Overview of Project**

NOAA (National Oceanic and Atmospheric Administration) records 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.

**What is a Storm?**

According to Wikipedia, a storm is any disturbed state of an environment or in an astronomical body's atmosphere especially affecting its surface, and strongly implying severe weather. It may be marked by significant disruptions to normal conditions such as strong wind, tornadoes, hail, thunder, and lightning (a thunderstorm), heavy precipitation (snowstorm, rainstorm), heavy freezing rain (ice storm), strong winds (tropical cyclone, windstorm), or wind transporting some substance through the atmosphere as in a dust storm, blizzard, sandstorm, etc.

**What are we predicting?**

NOAA stores the observations of storm events in a database of csv files (<https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/>). We are using the features and observations from this data to predict the property damage caused by any of the storm events in United States

**Roles and Responsibility**

|  |  |  |
| --- | --- | --- |
| **Team Member** | **Area of Work** | **Shared Responsibility** |
| Siddharth Das | Preprocessing | EDA |
| Kartik Jain | Extraction | Modelling |
| Hemangi Kinger | PyQt5 and Visualization | EDA |
| Ishan Kuchroo | Modelling | Preprocessing |

**What is my responsibility?**

Created the initial web scrapping and preprocessing flow which was later refined. Web scraping to extract the data from the website and creating a single dataframe by combining multiple CSVs. Cleaning the data to remove the alpha numeric values to numeric values. Additionally, build and tune the XGBoost model once the preprocessing is executed. Suggested feature creations to increase the include in the such as geo distance, duration of the storm. Experimented with the unsupervised K-Means clustering feature creation to increase the performance of the mode. Created the power point presentation with Ishan. Created the map plot using the geo coordinates.

**Extraction**

1. Web scrap to get the data files to unzip and read the file into a data frame and concat it to a single data frame.

**Graphical user interface, text, application

Description automatically generated**

1. Extracted and merged the EPA climate data to create a single dataframe for further preprocessing

**Text

Description automatically generated**

1. Preprocessing: Created a method to clean the target variable and the damage crop variable and get a consistent numeric variable

Text

Description automatically generated

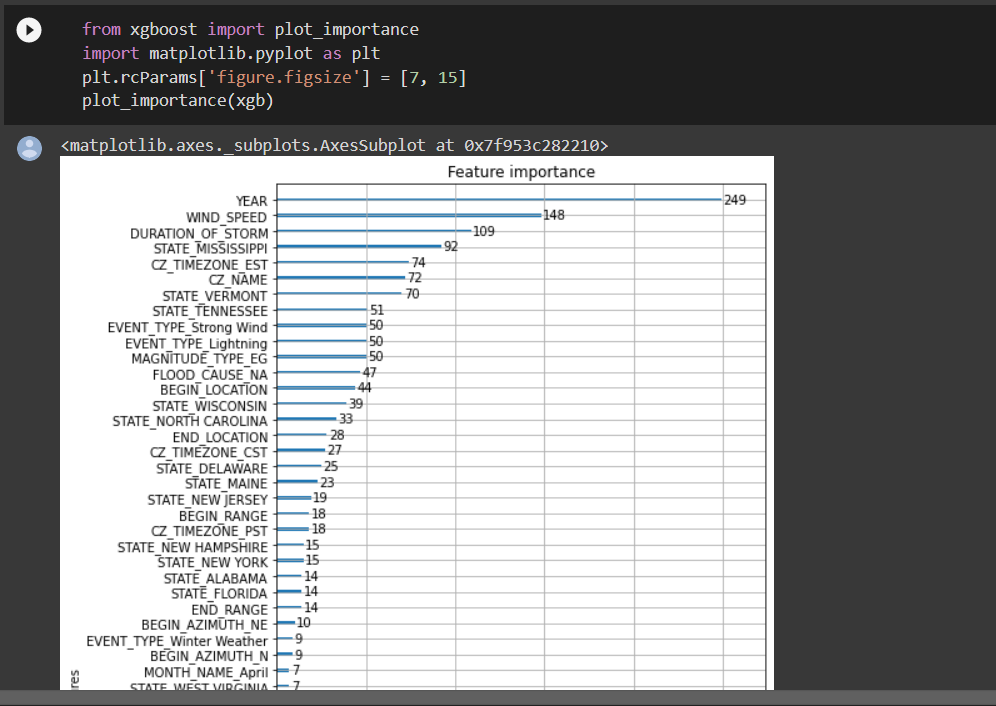
1. Checked for null values per year to understand the data and created plots for the number of missing valuesGraphical user interface, text, application

   Description automatically generated
2. Created the XGBoost model with tuning hyperparameters

Text

Description automatically generated

1. Created the feature importance for the XGBoost regressor



1. Advance feature creation using the unsupervised Kmean clustering with the scaling, sum of squared distance and silhouette score

Text

Description automatically generated

Text

Description automatically generated

**Referenced Code %**

(70 − 50) /(70 + 20) × 100 = 22.22

**References**

* <https://medium.com/@juniormiranda_23768/ensemble-methods-tuning-a-xgboost-model-with-scikit-learn-54ff669f988a>
* <https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html>
* <https://docs.python.org/3/library/pickle.html>
* <https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/#:~:text=In%20k%2Dmeans%20clustering%2C%20the,tree%20shape%20form%20(dendrogram)>.
* <https://link.springer.com/article/10.1007/s10115-021-01572-6#Sec33>
* <https://medium.com/greyatom/using-clustering-for-feature-engineering-on-the-iris-dataset-f438366d0b4b>
* <https://www.kaggle.com/learn/feature-engineering>