Sovunh Voeu, George Zmudzien

Professor Poonam

COMP 470-001

11/6/2024

Topic: Natural Disasters Predictions Using Weather Data

The 5 different EDA operations I have done are:

* **Structure Analysis** to identify outliers or any anomalies that may create problems for the data.



* **Missing Value Analysis** to remove affected rows or chooses a suitable imputation that will target missing values that interfere.



* **Univariate Analysis** to understand and normalize the distribution if needed through using histograms and box plots.



* **Outlier Detection and Treatment Analysis** to avoid outliers that may skew the model and create a poor graph.



* **Bivariate Analysis** to create scatter plots or heatmaps to further observe the relationship between pairs of variables.



The algorithm that we chose to use from machine learning that we did not discuss in class was the **Random Forest Algorithm**.



The Random Forest algorithm tends to handle non linearity very well, and since our datasets have a lot of different parameters(cloud cover, temperature and humidity for example) it makes analysus of all those features easy to analyze. It also works incredibly well with regression analysis especially with the logistic regression algorithms we did later.

And of the two algorithms that we had learned about in class we picked the **Naive Bayes Algorithm** and **Logistic Regression**

There were a lot of consideratiuons as to why we would choose a logistic regression model with the main one being the ability for a logistic regression model to be efficiently used on large datasets. Our dataset was incredibly large and the fact that it can preform well on such a huge dataset and preform reliably made it a major contender for what we would use. Another reason to use logistic regression would be how easy it is to interpret the results of the algorithm, by having something easy to interpret it makes the time we need to focus on the results less and grants us greater focus on other areas of the project.

A lot of these points can be made about Naive Bayes, its functionally incredibly similar to logistic regression models in the fact that it is also incredibly efficient but also it is generally effective for our dataset. Since we are dealing with a lot of categorical variables it ends up being incredibly efficient when compared to some other machine learning algorithms we learned about.

Works Cited:

GeeksForGeeks. 12 July 2024, [www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/](http://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/). Accessed 7 Nov. 2024.