Project Milestone

Seunghyuk Baek email: sbaek44@vt.edu
Jeevan Thapa email: gforest5@vt.edu

Introduction

Every year, many tropical cyclones or hurricanes are developed in Atlantic Ocean and some of them approach to the US soil. When a tropical cyclone falls onto US soil, it often causes massive damage to human life and property. The development of a tropical cyclone depends highly on the environmental temperature. Normally, a tropical cyclone loses its strength when it moves within mainland. However, sometimes a cyclone can be strengthened even though it is hovering over ground especially when it contacted warm fresh water. Therefore, It is plausible to correlate temperatures with development of a hurricane.

Project Problem Statement

Preventing the hurricane damage is a hard task since Saffir-Simpson Wind Scale (often denoted by 'category') is hard to estimate at the beginning of cyclone development. If there is a correlation between earlier year weather (roughly from January to July) and the hurricane strength, people can prepare in advance for the damage that future cyclone will cause. Moreover, in government's perspective, it is easier to allocate aid. Therefore, the task of this project is to find any correlation between the weather condition of certain period and the strength of a tropical cyclone.

Data Set

Data set contains location specific data and is collected from NOAA (National Ocean and Atmospheric Administration). Since, hurricanes are developed from ocean, we picked random cities and their weather observation stations located near US east coasts. Each station provides different data points and features. In average, there are 60 different features for each data set. The number of data points depends on requested range of period. The data we collected from Jan 1, 2009 to Dec 31, 2017 has 1800 data points. Another dataset that acquired from NOAA is historic data of landfall hurricanes. This is table from NOAA's html page with 345 data points and 7 features.

Preprocessing steps

The climate dataset is a csv file which contains string and numeric values. The first 5 features of this dataset are: station code, station name, latitude, longitude and recording date. Because this is monthly data, there is only year and month for the date. After these 5 columns, actual measurement data is provided with data attributes. Data attributes describe the characteristic of this collected data in a given month. For example, TAVG is average air temperature of the month and data attribute describe if there is missing value for calculating the average. For this analysis, all data attributes are deleted with first 4 columns (station id, name of station, altitude and longitude). In climate dataset, there are values that can be represented by average. For example, all other temperature related data such as extreme heat or minimum temperature will

be eliminated if there is average temperature value. Some data from the dataset seem irrelevant for the analysis. Data such as snowfall and snow depth are eliminated. There are many empty values (or null) present in the climate data. To eliminate the eliminate the empty values, entire instance with empty values are removed. Because there are data collected on same day but in different location, there are multiple values for each month. We average these values from same year and month to combine as one.

The list of hurricane data is trimmed so that only data from 2009 to 2017 is available. The all the data from Jan to the month before a hurricane occur set as training input and the wind speed of the hurricane is set as training output.

Methods and Models

1. Linear Regression
2. *k*-NN Regression
3. Neural Network
4. Logistic Regression