Gregory Olson Capstone Project Proposal

Can we save lives by predicting the likelihood of a tornado occurring within a 50 mile radius of a particular weather station within 24 hours of the storm event occurring?

Across the United States, there are over 1000 tornados every single year. These extreme weather events can be very dangerous, culminating in a devastating whirlwind that can wreak havoc on anything in its path. Oftentimes, these storms can appear quite suddenly, with little time for people to prepare and take shelter for potential impact. Likewise, these storms can cause serious damage to homes and businesses, and can be deadly in some cases. As a result, it is crucial that individuals in the path of one of these storms have the appropriate amount of time to gather supplies, seek safe shelter, and give governments enough time to prepare to reduce the incoming damage. Being able to predict the likelihood of one of these storms occurring is vital to the survival of sometimes hundred of thousands of people at any given time. If we were able to use other meteorological data from days prior to determine how likely a tornado event occurred, there would be ample time for inhabitants of the at risk area to take appropriate safety precautions.

Furthermore, I would like to build a model that takes meteorological data and uses it to predict the likelihood of a tornado occurring on a given day. This involves gathering meteorological data from various weather stations, combining them into a singular data frame, and then merging attributes from another data set containing tornado occurrences and their respective locations and other attributes. This data wrangling process may be complicated being that meteorological data is very localized by nature, and two locations only ten miles apart could have very different forecasts.

There are a few limitations to completely and accurately solving this problem. First, the process of obtaining data from a large number of weather stations will be very time consuming,

yet it would be appropriate to acquire as much data from as many stations as possible. For the sake of this project though, it would make more sense to gather enough data so as to attain sufficient information for the model yet not have to spend a month on data wrangling. Another limitation is that the accuracy of predicting a storm from a given distance away from the closest weather station will be less accurate as this distance is increased by the nature of weather data. It would be nice if all of these tornadoes occurred right next to the weather stations, but that is not the way that these storms work.

This model will be very beneficial to a multitude of stakeholders. Local government will want their people to be safe from extreme weather events. Emergency services like EMTs, firefighters, police officers, and hospital workers would need to know of the likelihood of future tornado occurrences so that they can be properly staffed to perform their essential duties on the day of the storm. On a broad scale, every single person in the United States will be a stakeholder in this analysis, being that it is their safety that is directly affected by the accuracy of this model.

The data that will be used for this project will be sourced by the NOAA. I will use individual weather station datasets, as well as the most recent yearly storm event dataset, that will provide details about the location and other attributes for each tornado event.