Tracking Hurricanes with Gradient Boosted Regressors And Multi-Output Regressors

Grant Hunter
CIS377 - Final Project
grant.hunter@students.fhu.edu

Abstract

This paper explores the benefits and challenges of using simplistic models for predicting hurricane trajectories. To accomplish this, I used two models: one that predicted the direction of the hurricane and one that predicted the distance to be traveled. Putting both models together allowed me to compute a trajectory path to reasonably low error.

1. Introduction/Background/Motivation

1.1. Hurricane Tracking

Hurricanes are becoming increasingly catastrophic in consequence of global climate changes. Tracking these hurricanes becomes ever important, and doing so in a computationally time effective matter is important since people must act fast to prepare for the worst.

People cannot always rely on a centralized agency for weather forecasts, so it is valuable to have open-source models that people with basic setups can utilize.

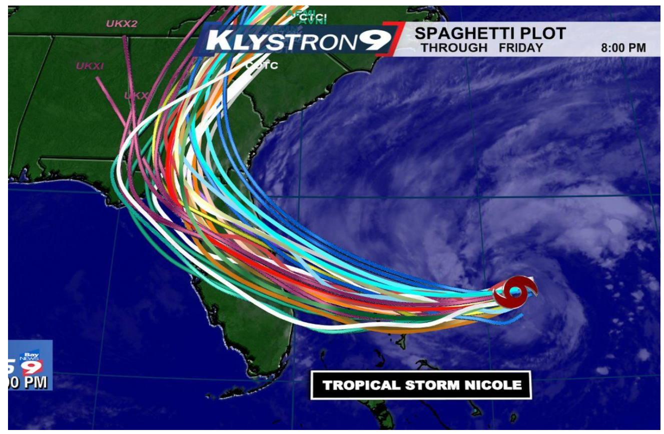
Another important characteristic to focus on is that since hurricanes are quite large (~300 miles wide), so minor imprecisions can be acceptable.

1.2. Commonly Used Approaches

The most prominent models for predicting hurricanes involve a few kinds of structures: GANs, LSTMs, and Ensembles. These can be relatively computationally expensive, so I sought to achieve the best error possible with more simple time-effective regressors.

1.3. Possible Applications

I believe that my model could be used as an additional model to be averaged into ensemble methods computed by the National Hurricane Center (NHC), or at least alongside the many other models of a spaghetti plot. An integral part of hurricane tracking is the existence of many unique models that can interact together to give us a greater sense of possibilities.



While predicting hurricane trajectories has become manageable, they are still chaotic systems with enough randomness to demand adjusting for uncertainty.

1.4. Data

The dataset used for this project was HURDAT2 retrieved from Kaggle, which was shared by NOAA. The data contains 1814 distinct hurricanes from the years 1851-2015.

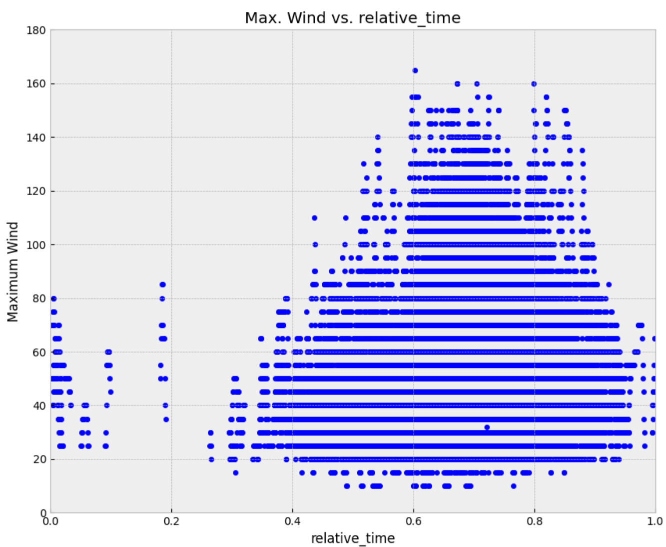
The HURDAT2 dataset contains a total of 22 features on 49,105 instances. I only utilized 7 of those original features within my models. I also computed features which will be discussed in section 2 of the paper.

2. Approach

2.1. Determining Relevant Features

An important first step in regression problems is determining the important features for informing the output that is desired. Obvious candidates were the *Latitude* and *Longitude* features, which were used in both the independent variable(s) and dependent variable(s).

To simplify the regression task, *Latitude* and *Longitude* values were converted to an xy-coordinate plane using a geographic transformation. The location of the storm on the globe can contain information about typical behaviors of storms at a given coordinate. In order to best estimate the trajectory of storms, I used the vectors between points as a calculated field. This allowed for the building of two models: one for predicting the direction of the vector and one for predicting the length of the vector.



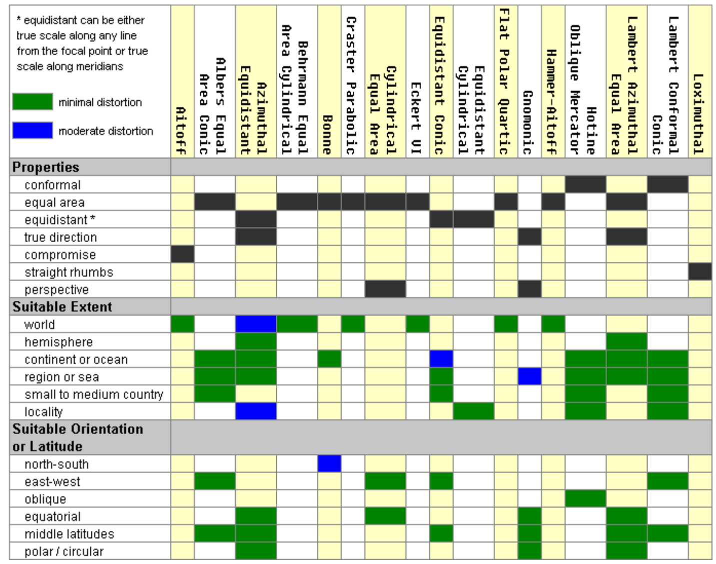
In examining the *Maximum Wind* feature against *Relative Time* (decimal value in range (0,1) representing completion of the year), it became apparent that there was indeed a "hurricane season" as is colloquially accepted. This informed my decision to include discretized datetime features in the regression models.

Maximum Wind was another feature that was used in the regression models, as I believe it contains information on the intensity of the storm, which could inform how far the storm may travel.

The final features used in the regression models were lagged copies of the *Direction* and *Lengths* of the vectors. This will be further discussed in section 2.3.

2.2. Geographic Features and Map Transformations

Meteorologists take geographic transformations of coordinates to be an important metric to have in consideration. Every transformation from a spherical globe onto an xy-coordinate plane will have distortions. The goal in choosing a transformation is to pick what distortions you want to allow.



Some are better at preserving direction, and some are good at preserving shape. It is not possible to be ideal at both. However, there do exist transformations that are good enough over a specific-sized region. Since the dataset contained points from nearly the entire Northern Atlantic, I decided to go with the Equidistant Sphere transformation, which was good up to hemisphere-sized regions and preserved both direction and shape sufficiently well.

```
Name: World Equidistant Cylindrical (Sphere)
Axis Info [cartesian]:
- E[east]: Easting (metre)
- N[north]: Northing (metre)
Area of Use:
- name: World.
- bounds: (-180.0, -90.0, 180.0, 90.0)
Coordinate Operation:
- name: World Equidistant Cylindrical (Sphere)
- method: Equidistant Cylindrical (Spherical)
Datum: Not specified (based on GRS 1980 Authalic Sphere)
- Ellipsoid: GRS 1980 Authalic Sphere
- Prime Meridian: Greenwich
```

It takes the shortest distance between two points on the plane to be equivalent to the shortest distance on a sphere: a great circle. This important characteristic would help inform a more realistic path for hurricanes, as they do not take what a Mercator map projection would show to be the shortest distance.

2.3. Converting from Time-Series Data to Supervised

An important step in taking time-series data and preparing it for regression was the addition of lagged vectors and multi-output predictions. Lagged vectors helped to communicate to the model some level of change over time in the vectors, which I believe would help to better classify what was to come.

Multi-output predictions are highly valuable in hurricane prediction, because having an online model that can only predict one step (6 hours in this case) ahead would be unhelpful for those that need to shelter or evacuate. I experimented with a 4x4 grid of input and output sizes of vectors in the development of my models, which could therefore predict up to a day in advance.

2.3.1. Chained Regressors

Most regression models do not natively adapt to predicting on multiple outputs (with the exception of Decision Trees). It became necessary to use *sklearn*'s *multioutput* sub-module to wrap around the Gradient Boosted Regressors. The wrapper simply tells the models to predict one output at a time, much like an iterator.

2.4. Error Function

2.5. Anticipated Problems

2.6. Encountered Problems

3. Experiments and Results

3.1. Grid of Input and Output Sizes

3.2. Decision Tree Regressors vs. Gradient Boosted Regressors

4. Future Work for Improvement

5. Sources