Forecasting of hurricanes using weather buoy data

Ishwar Choudhary
PES1201700189
CSE Department
PES UNIVERSITY
Bangalore,India
ishwarc404@gmail.com

Shreya Banerjee
PES1201700198
CSE Department
PES UNIVERSITY
Bangalore,India
shreyabanerjee167@gmail.com

***Abstract*—This project is an attempt to predict the occurrence of a hurricane over the seas using both recent and historical data obtained from the National Oceanic and Atmospheric Administration web pages. National Oceanic and Atmospheric Administration stores and manages a very extensively detailed database of all the buoy data. It keeps a check on the condition of the Oceans, major waterways and the atmosphere. This project mainly focuses on Hurricane Irene. Hurricane Irene was a large and destructive tropical cyclone that hit the US coast in August 2011. To predict the occurrence of a hurricane, we use attributes such as wave period, wave direction, wave height, latitude, longitude, wind direction, wind direction, wind speed and pressure.**

I. Introduction

Hurricanes are a type of tropical cyclone, which form over the Atlantic Ocean and Northeastern Pacific Ocean. The creation of tropical cyclones is a subject of extensive research and is still not understood precisely. The few key factors that affect hurricanes are atmospheric pressures, warm sea surface temperatures, light winds aloft, and rotation or spin. This project deals with the prediction of hurricane formation or occurrence by evaluating such factors. Data for our project consists of buoy data which is available on the NOAA's website separately for each buoy. All buoy stations measure wind speed, direction, atmospheric pressure and air temperature. In addition, they also measure sea surface temperature and wave height and wave period. In this project, the spotlight is given to the hurricane "Irene". Using the historical data of 2011, we are developing a model, which predicts the occurrence of a hurricane.

II. Importance

The most critical reason to predict hurricanes is to minimise loss of life and property. Hurricanes are one of the most destructive natural calamities. Governments in areas prone to hurricanes develop contingency plans to help people take actions such as boarding up windows, moving in land, sheltering in safe locations etc. By predicting hurricanes, people can be directed to take action that minimises chaos and loss of life. Hurricane predictions and the probability that a hurricane will hit a specific area has a great deal of relevance to the flood risk of an area. Flooding from a hurricane can be caused by excessive quantities of rain, broken and breached levees, and storm surges from the ocean or a major lake. Many of the government operated hurricane management centres like THE NATIONAL HURRICANE CENTRE and the CENTRAL PACIFIC HURRICANE CENTRE give an accurate prediction about the formation and trajectory of hurricanes using satellite imaging up to 48 hours before they occur. In most of the cases, 2 to 5 days are not enough for entire cities to prepare themselves for a hurricane. This is why an early and accurate prediction of such calamities is very important.

III. Problem Statement

To analyse the data related to hurricane Irene in order to find out the features that can help us to predict the possibility of occurence of a hurricane in general.

IV. Dataset Description

1) The National Ocean and Atmospheric Administration (NOAA) has weather buoys deployed all over the world and in high concentration in the North Atlantic Oceans which are a hot-spot for the formation of hurricanes. NOAA maintains a very detailed archive about all the previously buoy recorded data in the National Data Buoy Center (NDBC).
2) These data tuples contain atmospheric and water-body information such as Pressure, Wind Speed, Temperature, Wind Direction and lot more. These tuples are readily available in the form of rich text files on their website.
3) Our datasets are a collection of such tuples that were scraped from the NDBC website. The tuples scraped were those of the weather buoys which were in very close proximity to the path traversed by Hurricane Irene. Hurricane Irene was a large and destructive tropical cyclone that affected much of the East Coast of the USA during late August 2011.
4) Tuples consists of detailed weather information, spread over 8 days, recorded by 8 buoys, 5 of which were in direct line of Irene and 3, a few nautical miles away.
5) We selected the buoys - 42060, 41046, 41010, 41013, 41025, 44014, 44009, 44065 for our analysis.

V. Assumptions

During the course of our entire study about the movement and predictions of hurricanes, we have taken Hurricane Irene as our sole reference point.

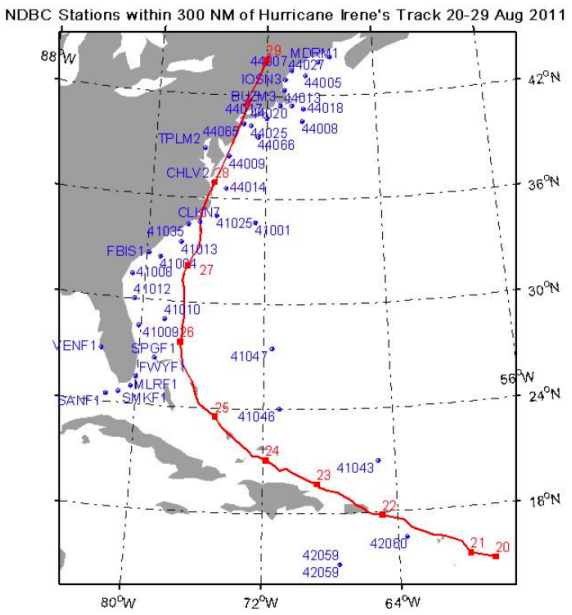

Fig. 1. Trajectory traced by Hurricane Irene in the year 2011 and the corresponding weather buoys in the path.

VI. Pipeline

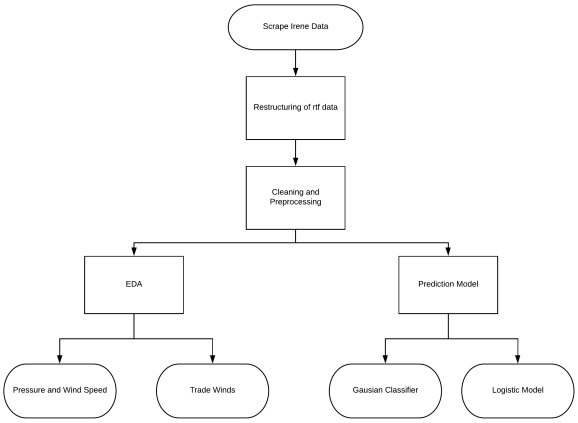

Fig. 2. Project pipeline

A. Scraping Irene Data

NOAA maintains a very extensive archive of all the weather information recorded by its buoys which are deployed all over the oceans of the world. Data scraping was the first step in our project, but it required extensive initial research to know which buoy data we had to scrape. For this, we had to analyze the trajectory of Hurricane Irene and we selected the buoys which lay directly or nearby the path of the hurricane. Further, through a Python script, we automated the scraping of these selected buoys. The result was in the form of multiple RTF files downloaded, which we then had to pre-process and convert into CSV files.We have one CSV file named "IRENECSV.csv" with data from all the buoys initially.

B. Cleaning and Pre-processing

On plotting a box plot for the columns pressure, windspeed and wave height,we realised that there are some outliers. The outliers really far away from the main distribution are replaced by the mean of each column. Some of the outliers are nearer to the main distribution and present in a significant quantity. They might be important factors in prediction of hurricane and thus, have not been removed or replaced.

C. Exploratory Data Analysis

Exploratory data analysis is a vital part of the entire project.It performs to define and refine our important features variable selection, that will be used in our model. When a tropical cyclone reaches hurricane strength, its low-pressure centre is called the "eye" of the storm. Acting like fuel that feeds more energy into the storm, moisture from the warm water is converted to heat in the bands of rain that spiral around the eye. The lower the barometric pressure at the centre of the storm, the stronger the hurricane, and vice versa. The movement of the hurricane is also very important as it is essential to help predict the trajectory. Thus it is very important to analyse these factors effecting the sole existence of the hurricane. The results obtained after analysing the scraped data for these factors are in direct agreement with the science behind Hurricanes and they are summarised as follows:

1) The scatter plots of pressure, windspeed and wave height from all the buoys showed significant variance in the readings from 25th August to 30th August, 2011.
   - Drop in pressure
   - Increase in Windspeed
   - Random variation in wave height
2) On plotting line plots for each individual buoy showing variation of pressure and windspeed together in one graph, it could be clearly seen that some of the buoys neither showed pressure drop nor windspeed increase but the other buoys which were exactly on the trajectory showed a significant pressure drop and windspeed increase.

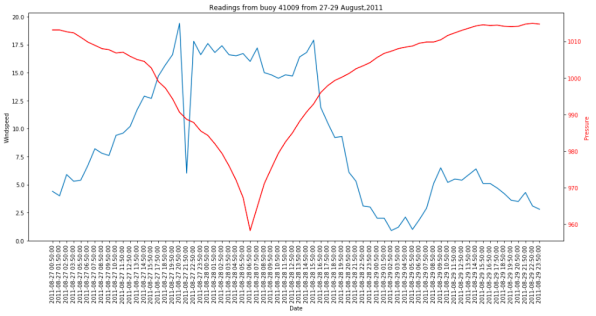

Fig. 3. The pressure reaches minimum level and then increases gradually signifying the passing of hurricane through that point. At the same time, there is a significant increase in windspeeds.

3) It can be seen in the Fig 3 that the pressure reaches minimum level and then increases gradually signifying the passing of hurricane through that point.

Since buoy number 44009 and 44065 were right on the trajectory, we decided to use the data from these buoys for accurate prediction. We created new CSVs for these buoys and created a new column , called ”possibility” (indicating the possibilty of a hurricane). The value of this column was 1 for rows with pressure lesser than 1000, otherwise it was 0.

One of the major factors affecting the movement of these tropical hurricanes is the existence of Trade Winds.Hurricanes are steered by these global winds which blow from east to west in the tropics. As the trade winds get stronger it becomes easier to predict where the storm will travel whereas when they are weak it's more difficult. After a hurricane crosses an ocean and reaches a continent, the trade winds weaken. This means that the Coriolis Effect has more of an impact on where the storm goes. In the Northern Hemisphere the Coriolis Effect can cause a tropical storm to curve northward.

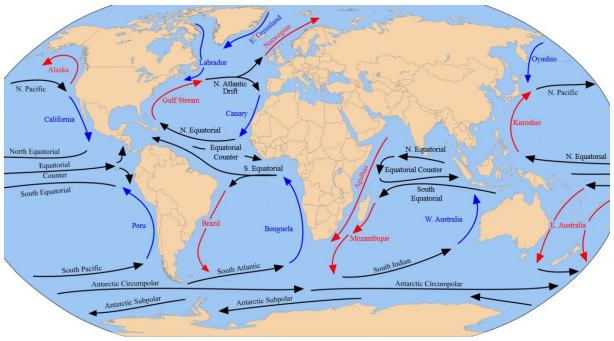

Fig. 4. Global Trade Winds

On analysing buoy data for the direction of winds, our findings were clearly in agreement with the science behind the movement of the hurricanes. On visualising the winds, we obtained the followed figure.

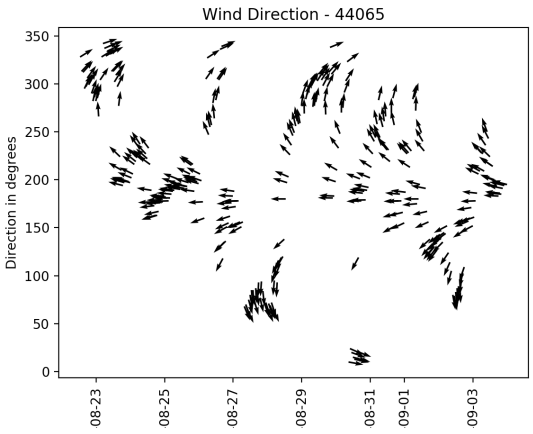

Fig. 5. Wind Directions observed by the buoy 440065

Through the wind direction plots,the change in wind direction towards NW and N is clearly noticeable which prove the existence of trade winds. Buoy numbered as 44065 is situated very close to the land of USA, and once the hurricane passes(after 8-31),we can clearly see the upward trend in the direction of the trade winds.

D. Prediction Models

1) Gaussian Naive Bayes Classifier: A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. The model was trained using ”Pressure”, ”Wind Speed” as input features and ”hurrthreat” as our output label.

2) Logistic Regression: One of the main problems with our data-set is that there are a very few instances of the hurricane actually occurring (i.e, the possibility is 1). There is a gross imbalance of data. There are very small number of instances of actual interest.This hindered in the training of the model, and provided us with unsatisfactory results. Hence, we proceeded with the up-sampling of data using SMOTE(Synthetic Minority Oversampling Technique) algorithm.
A few facts about SMOTE:

- Works by creating synthetic samples from the minor class (no-subscription) instead of creating copies.
- Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked, new observations.

We over-sampled only on the training data, because by over-sampling only on the training data, none of the information in the test data is being used to create synthetic observations, therefore, no information will bleed from test data into the model training. After up-sampling the data, we have equal number of rows for possibility 1 or 0. Now, the next question was , which features to take as predictor variables ? Irrelevant or partially relevant features can negatively impact model performance. The main problem is that factors like day, month, hour and minute obviously do not affect the occurrence of a hurricane. Features like pressure, windspeed and wave height might change drastically in such events. Our previous visualisations and conclusions support this fact. Since the changes in wave-height are extremely random, we do not take it into account. We built a model using pressure and windspeed.

VII. Experiments and Results

A. Gaussian Naive Bayes Classifier

a) Accuracy: On training our Bayesian Classifier on the combined CSVs,the accuracy score achieved was around 91% Although this is a good accuracy score,on plotting the confusion matrix, we notice that such a high score is because the model classifying the 0s more accurately as compared to the 1s, and the relatively large frequency of the 0s, leads to the biasing of the model.

1) Precision : What proportion of positive identifications was actually correct? = 0.03
2) Recall: What proportion of actual positives was identified correctly? = 0.07

Hence our model predicted the 0s correctly but not the 1s. This is downside is due to the sparse number of rows.

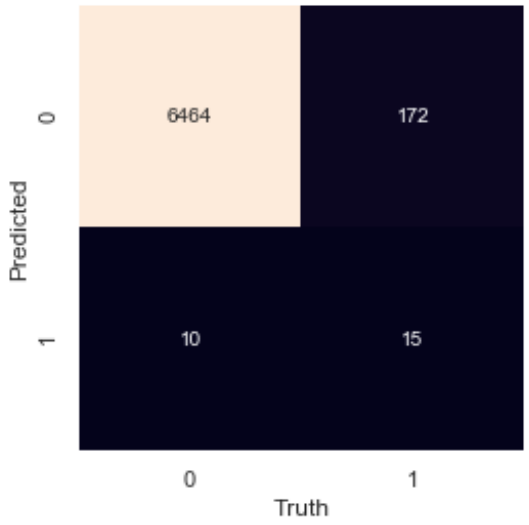

Fig. 6. Confusion Matrix for the Bayesian Classifier model

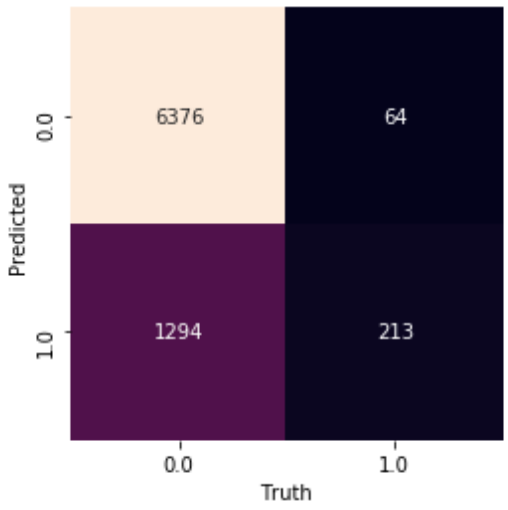

Fig. 7. Confusion Matrix for the logistic regression model

B. Logistic Regression Model

a) Accuracy: When we ran this model on the data from buoy 44065, it gave an accuracy of 0.83. In our dataset, the positive class — occurence of hurricane — is greatly outnumbered by the negative class. Even if it misclassifies the minority data, accuracy will be high. This is called "accuracy paradox". Therefore, accuracy is not a good measure for assessing model performance.

b) Confusion matrices: Looking at the confusion matrix in Fig 4, it can be clearly said that:-

- The recall for 1.0 is 0.77 i.e, out of all the positive classes (occurence of hurricane), we predicted 0.77 correctly.
- The precision for 1.0 is 0.15 i.e, out of all the positive classes we have predicted correctly, 0.15 are actually positive (i.e, a lot of false positives).

Most of the hurricane occurences are predicted correctly, but out of all the positive predictions, a lot of them are actually negative.

c) Drawbacks: The chances of raising false alarms are pretty high.

d) Potential reasons for the drawbacks:

- When we trained the model without upsampling the data, it misclassified the minority data. Standard classifier algorithms like Logistic Regression have a bias towards classes which have higher number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there was a high probability of misclassification of the minority class as compared to the majority class.
- There are a few disadvantages of upsampling the data by SMOTE algorithm. While generating synthetic examples, SMOTE does not take into consideration neighboring examples can be from other classes. This can increase the overlapping of classes and can introduce additional noise.
- We have only considered factors like pressure and windspeed to predict the occurence of a hurricane, there might be a lot of other factors that contribute to occurence of a hurricane which were not present in our dataset.

VIII. Overcoming the limitations

Detailed study of the hurricane trajectory, pressure and windspeed variation during emergence of a hurricane and existence of trade winds were some of the interesting features that we worked on in this project. They might be useful in prediction of hurricanes in future. However, the models used for the prediction had several limitations. These limitations could surely be overcome by using some other methods. More features could be brought into the picture, pressure and windspeed values could be predicted using time-series analysis and a more accurate model could be run on the data for the prediction of hurricane. Modelling of data using extreme value distribution might give better results. One must also look into alternate forms of data such as Satellite Imaging, to approach the problem with a different point of view.

IX. Contributions

- *Shreya Banerjee*
  Worked on the visualisations related to pressure and windspeed and the Logistic Regression model.
- *Ishwar Choudhary*
  Worked on scraping of data, visualisations for wind direction and Gaussian Naive Bayes Classifier.

X. Acknowledgment

We would like to express our sincere gratitude to our supervisor to Dr.Gowri Srinivas for providing her invaluable guidance and suggestions throughout the course. Furthermore we would like to extend our warmest gratitude to each of the teacher assistants Mr.Sumanth, Mr.Ninaad, Ms.Malaika and

Ms.Maanvi for taking time out of their busy schedules to help with our curriculum and projects.

REFERENCES

[1] How meteorologists predict the next big hurricane from https://theconversation.com/how-meteorologists-predict-the-next-big-hurricane-102827

[2]National Hurricane Center Track Forecasts from https://www.wunderground.com/cat6/nhc-track-forecasts-best-ever-2017-no-improvement-intensity-forecasts

[3]Predicting Hurricanes: A Not So Exact Science by Aubrey Samost from https://web.mit.edu/12.000/www/m2010/teams/neworleans1/predicting %20hurricanes.html

[4] Barometric Pressure Hurricanes from https://sciencing.com/stages-tropical-cyclone-8709867.html

[5]Operational Hurricane Track and Intensity Forecasting from https://www.gfdl.noaa.gov/operational-hurricane-forecasting/

[6]Science behind the movement and steering of hurricanes https://www.windows2universe.org/earth/Atmosphere/hurricane/movement.html

[7]Gaussian implementation of NB Classifier https://dataaspirant.com/2017/02/20/gaussian-naive-bayes-classifier-implementation-python/

[8] Wind Direction chart for mapping of the quiver plots https://uni.edu/storm/Wind%20Direction%20slide.pdf

[9] Logistic regression in python from https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8