Predictive Analysis of Weather in Trinidad and Tobago

Jordhan Couteau   
*Department of Computing and Information Technology  
University of the West Indies, St. Augustine*The Republic of Trinidad and Tobago, West Indies  
jordhan.couteau@my.uwi.edu

Lillian Charles  
*Department of Computing and Information Technology  
University of the West Indies, St. Augustine*The Republic of Trinidad and Tobago, West Indies  
lillian.charles@my.uwi.edu

Caleb Ramdath  
*Department of Computing and Information Technology  
University of the West Indies, St. Augustine*The Republic of Trinidad and Toabgo, West Indies  
caleb.ramdath@my.uwi.edu

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—Weather Forecasting, Pre-Processing, Data, Weather Model, (key words)

# Introduction

Weather forecasting is an essential aspect of daily life, shaping decision-making across critical sectors such as agriculture, transportation, disaster management, and energy planning. The ability to predict weather conditions accurately not only influences individual activities but also drives economic efficiency and safeguards communities against adverse weather events. However, despite advancements in technology, traditional weather prediction models often fall short in providing precise forecasts especially in Trinidad and Tobago. These limitations hinder proactive planning and preparedness, especially in regions prone to severe weather phenomena.

In response to these challenges, this paper seeks to enhance the accuracy and granularity of weather predictions through the application of big data analytics and exploratory analysis of historical weather datasets. By identifying specific parameters—such as temperature thresholds, humidity levels, wind speed patterns, and atmospheric pressure changes—we aim to detect critical conditions that precede severe weather events. Through this approach, our analysis strives to bridge the gaps in traditional forecasting methods by offering data-driven insights tailored to localized needs.

Accurate and reliable weather predictions are essential for mitigating the devastating effects of natural disasters, optimizing operations within weather-dependent industries, and safeguarding public safety. Improved forecasting capabilities can empower farmers with actionable data for planting and harvesting cycles, enable more efficient and cost-effective transportation logistics, and bolster the preparedness of island communities for extreme weather events. Beyond disaster prevention, our efforts also focus on enhancing the reliability of everyday weather forecasts, making them an indispensable tool for routine decision-making.

# Methodology

## Background

In this section, the preliminary approach that involves the leveraging of machine learning algorithms to develop a predictive model for weather forecasting will be highlighted. Further, factors such as data considerations, feature engineering and challenges faced are detailed.

## Approach

Our approach consisted of gathering relevant data sources, discerning which dataset would yield the most information, selecting which variables would be the most beneficial to Trinidad and Tobago and deciding which models would be best suited to train the selected dataset.

* Data Collection

Three data sources were identified to be best suited to be utilized. Meteostat was considered as it provides historical data for Piarco International Airport, facilitating pattern and trend analysis, this dataset also offered visualization tools that could enhance data exploration and interpretation. On the other hand, Climate Change Model Portal offered an extensive historical climate data for Trinidad and Tobago, including temperature, precipitation and extreme weather events. This dataset aids in trend analysis and future climate predictions using various models and scenarios. Finally, Visual Crossing Weather API was selected as the dataset that would utilized as it provided real-time and historical data for regions across Trinidad and Tobago with metrics such as temperature, humidity, wind speed and rainfall.

* Machine Learning Models

Ridge Regression Model was not selected as it assumes a linear relationship between features in the dataset and the target variable, which would affect predictions as weather patterns are non-linear, influenced by connections between for example, temperature, pressure and humidity.

AutoRegressive Integrated Moving Average (ARIMA) Model was not chosen due to its limitation of being an univariate model [1]. This means that the model forecasts based on a singular variable, this with weather forecasting involving multiple variable, the ARIMA would not be able to handle the complex data manipulation needed for the dataset.

Extreme Gradient Boosting (XGBoost) was selected as the main model for this project. This model was selected as it handles non-linear patterns, processes large datasets, resistant to overfitting and it works well for both short-term and long-term weather forecasting [2].

# Analysis

Data from Visual Crossing API pertaining to regions in Trinidad and Tobago between the years 2000 and 2025 were downloaded, cleaned and merged to create a combined dataset with weather information pertaining to Trinidad and Tobago.

Observations were made to ensure that the data being utilized would provide valuable insights. Features were then selected and engineered to enhance the predictability of chosen model. Visualizations were also made to help identify weather patterns as well as hidden relationships between the data to make more informed decisions.

## Data Collection and Preprocessing

Figure 2 depicts the records that includes observations of the regions downloaded based on Figure 1, showing precipitation, humidity, average temperature (avgtemp c), maximum temperature (tempmax c) and minimum temperature (tempmin c).

Fig. 1. Picture showcasing regions in Trinidad and Tobago

Fig. 2. Weather Dataset showcasing its features

## Data Loading and Cleaning

This step involved loading the raw dataset into an human-readable structured format utilizing the pandas library in Python. Preliminary data exploration was conducted to identify missing values, inconsistencies, outliers and to remove any irrelevant variables such as snow that has no correlation to the region as showcased in Figure 3 and 4.

Fig. 3. Preprocessing of data

Fig. 4. Preprocessing of data

## Feature Selection and Engineering

To enhance the predictive capabilities of the XGBoost model, various key features were engineered.

* Heat Index

Average temperature and Humidity were used to calculate the Heat Index. Also known as the apparent temperature, heat index is a measurement of how hot it feels when air temperature and relative humidity are combined [3]. Calculations made were compared against the National Oceanic and Atmospheric Administration (NOAA) chart and their heat index calculator to ensure accuracy of results.

* Dew Point

Equation (1) depicts the formula needed to calculate dew point, with Ts depicting dew point, T for temperature and RH representing relative humidity. In this case, average temperature and humidity was used to engineer this feature.

Ts = (b × a (T,RH)) / (a - α(T,RH)) 

Fig. 5. Feature Engineering

## Visual Analysis

Visual analysis allowed for the data to be analysed more thoroughly to derive trends, outliers and relationships. It also plays a crucial role in both the development and validation of the weather forecasting model in various ways such as feature engineering as highlighted in Figure 5 and 6, parameter tuning and understanding the model’s limitations.

Fig. 7. Name of Screenshot

Fig. 8. Name of Screenshot

Fig. 9. Indexing of Data

# Model Training

To accomplish our task of training reliable weather forecasts, we trained machine learning models on historical weather data. This section outlines the key decisions made for modelling as well as the evaluation of the generated model.

## Model Selection

XGBoost was selected as the modelling approach due to its effective nature of handling structured data as well as its strong performance with minimal tuning. It has the ability to model nonlinear relationships and handle missing data. Therefore, this made it suitable for 20 years of historical data with daily measurements of approximately 120,000 records.

The initial approach involved creating one model to predict all the features. Generating predictions for multiple variables simultaneously. However, the results were inadequate as accuracy rates were low, resulting in a model that struggled to capture trends in the features.

Therefore, rather than building a singular model, the approach pivoted to adopt a Chained-Model Structure. A predictive model was developed for each weather variable such as temperature, humidity and windspeed. Each model predicted one after the other allowing for predictions from earlier models in to chain to be used as features in the subsequent models. This greatly improved accuracy and consistency across the various variables.

## Feature Scaling

XGBoost is a tree-based model, therefore, it does not require traditional scaling. However, certain variables such as Location were encoded numerically using label encoding for model compatibility. Time based features such as year, month and day were also engineered to enable recognition of time based trends.

Other features used include:

* dayofweek, dayofyear which reflected the periodic patterns
* location\_encoded which distinguished the various regions for which data were collected.

*Lag Features*

Lag features were considered to all the model to predict based on previous trends. However, this posed a dilemma especially in terms of future forecasting. Lag features tend to look at either the previous day or the past seven days in order to detect trends within the weekly forecast. When predicting the weather for future days, the model would need data for the lag features which may not necessarily be available. Therefore, this model was developed with the exception of Lag features.

## Model Evaluation Metrics

The Mean Absolute Error (MAE) and accuracy from sklearn was utilized to calculate the accuracy and test the model. MAE was called on each model and provided the average error per prediction. Generally, the models returned an MAE of < 1°C with the exception of the cloudcover, precip, humidity and windspeed models. However, these still returned decent accuracy scores. The calculated accuracy percentages gave a general sense of the correctness of the predictions across all the models.

# Model Results

The models demonstrated exceptional performance particularly on the variables such as temperature and windspeed with accuracy rates above 97%. Features which tend to have high levels of variability also performed well with the lowest rate 93.7% approximately.

The chained structure allowed the predictions to leverage each other creating inter-variable dependencies. The helped maintain coherence through out the model predictions and the datasets while also reducing the likelihood of propagation.

Utilizing these features, the deployed system will be able to generate forecasts for any user specified date as well as location. Ultimately returning predictions for the selected day as well as the following few days creating a 7-day query and based on those predictions, it returns risk indicators such as “Potential Flooding” to help users with daily planning and risk management.

TABLE 1. MODEL PERFORMANCE METRICS

|  |  |  |
| --- | --- | --- |
| Variable | Mean Absolute Error (MAE) | Accuracy (%) |
| cloudcover  precip  humidity  windspeed  feelslikemax c  tempmax c  tempmin c  avgtemp c  feelslikemin c  avgfeelsliketemp c  dewpoint c  visibility | 5.629  3.250  1.917  1.616  0.769  0.382  0.408  0.216  0.298  0.306  0.210  0.261 | 93.66  98.62  95.45  99.00  97.79  97.21  96.86  97.37  98.47  98.17  97.90  98.58 |

Average Accuracy: 97.88%

After training the models, they were exported as Python Pickle File (.pkl) to be utilized in the web application development process to generate predictions for queries.

# Web Application

The web application developed provides a user-friendly interface for accessing and visualizing the output of the advanced weather model. It integrates historical data, model predictions and interactive visualizations, enabling a comprehensive forecasting of the weather time period selected.

Fig. ?. Screenshot

# Discussion

This study demonstrates the practicality of utilizing machine learning modest for forecasting weather conditions and identifying the severity or risks associated with those forecasts in countries such as Trinidad and Tobago. Trained on 20 years of historical weather data, the XGBoost regression model yielded high accuracy with variables generally retaining accuracy rates of greater than 95%.

The design decision to use a chained modelling structure improved the contextual awareness of the models. Model tuning and feature designing among other things resulted in steady outcomes across the forecast horizon.

The simple and accessible nature of the final system is also noteworthy. The web application runs locally without installation requirements, allowing users to interact with a user-friendly interface. They have the ability to obtain a 7-day forecast within one query. However, it should be noted that these models were trained on historical data and as such have developed pattern-based predictions which rely on seasonal trends. Therefore, although it is effective for general trend recognition, the model may fall short of replacing real-time meteorological data sources providing short term precision forecasting.

While the current system provides reliable forecasting based on historical data, there are several areas which can be improved further for a more efficient and accurate model. The incorporation of real-time data can improve accuracy and precision of the models. Another potential enhancement involves refining the model to handle severity in a predictive sense, instead of generation based on thresholds. A more robust approach involving the use of a classification model for training to predict the severity of the weather in a day directly. In this scenario, the model will incorporate the severity label into the training dataset allowing the system to learn the patterns associated with sever weather events over time in Trinidad and Tobago. Essentially making severity a learned outcome instead of manually set rules.

# Conclusion

In this project, a fully functional weather application was developed. One which applies machine learning to provide predictions tailored specifically towards the regions of Trinidad and Tobago. By creating a Chained XGBoost model, the system provides fairly accurate 7-day forecasts for over 10 features. It provides easily interpretable forecasts with a 97.88% rate of accuracy, which demonstrates its reliability.

##### References

1. “ARIMA Modeling,” CORP-MIDS1 (MDS). https://www.mastersindatascience.org/learning/statistics-data-science/what-is-arima-modeling/
2. Analytics Vidhya, “Understanding the Math behind the XGBoost Algorithm,” Analytics Vidhya, Sep. 06, 2018. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
3. N. US Department of Commerce, “What is the heat index?,” www.weather.gov, 2024. https://www.weather.gov/ama/heatindex
4. “What is Dew Point? Camlab,” www.camlab.co.uk. https://www.camlab.co.uk/blog/what-is-dew-point
5. K. Elissa, “Title of paper if known,” unpublished.
6. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
7. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
8. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
9. K. Eves and J. Valasek, “Adaptive control for singularly perturbed systems examples,” Code Ocean, Aug. 2023. [Online]. Available: <https://codeocean.com/capsule/4989235/tree>
10. D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114. [Online]. Available: <https://arxiv.org/abs/1312.6114>
11. S. Liu, “Wi-Fi Energy Detection Testbed (12MTC),” 2023, gitHub repository. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC
12. “Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009.” U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2