Spatio-Temporal Extreme Event Prediction over Indo Gangatic Plain

Name:	**Abhishek Kuriyal**
Registration No./Roll No.:	19349
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	February 02, 2022
Date of Submission:	April 24, 2022
Team Members:	Vishv Jeet, Piyush Verma

1 Introduction

1.1 Dataset Description

The training dataset provides 122 features vectors spanning over 934807 training instances. They comprise geographical coordinates, date time of recording and certain weather metrics recorded via multiple sensing devices. The testing dataset provides 103868 instances for prediction testing. The target csv file provides a relative score for the event. All geographical coordinates corresponds to total 29 unique districts. The frequency of data distribution among these districts can be seen in fig-1. Out of 934807 training instances provided in dataset, only 654903 feature vectors are relevant rest were mostly with missing values. The reason why these missing values weren't handled because the they were missing continuously for "four" decades for almost all districts as shown in fig-2, for district "Aligarh".

As per the requirement of the model we are using, the dataset must be in uniform time distribution, which isn't the case here so instead of handling continouous missing values, they were dropped.

This raise a very valid question, "don't these values play any role"? Yes, they do but they aren't required, as from fig-2 the most significant property this data has is periodicity which started near year 2016 and continued with similar trend till end. And this is the key for any regression model to learn this data distribution.

Since its the periodic trend what matters here so this entire data set is clipped from year 2016 till year 2021. The reason we've chosen this range is that the distribution is both continuous, with each day observations except for few gaps which are replaced with "mean" of adjacent values, and

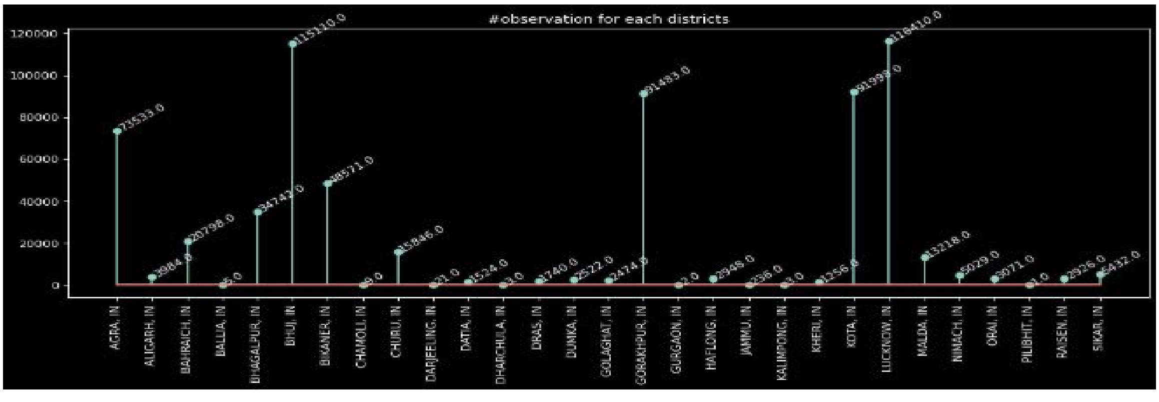


Figure 1: Data distribution frequency for each district in dataset.

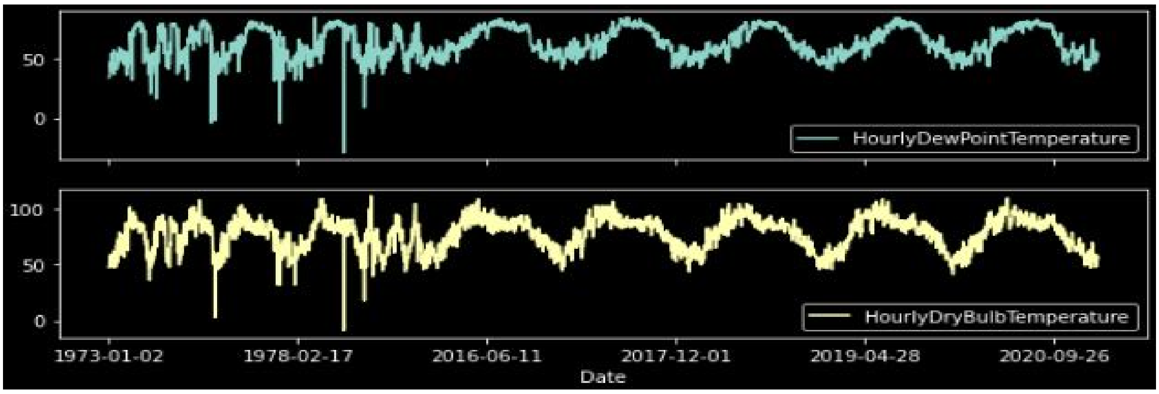


Figure 2: Continous Missing Values from year 1973 to 2016 [“Aligarh”]

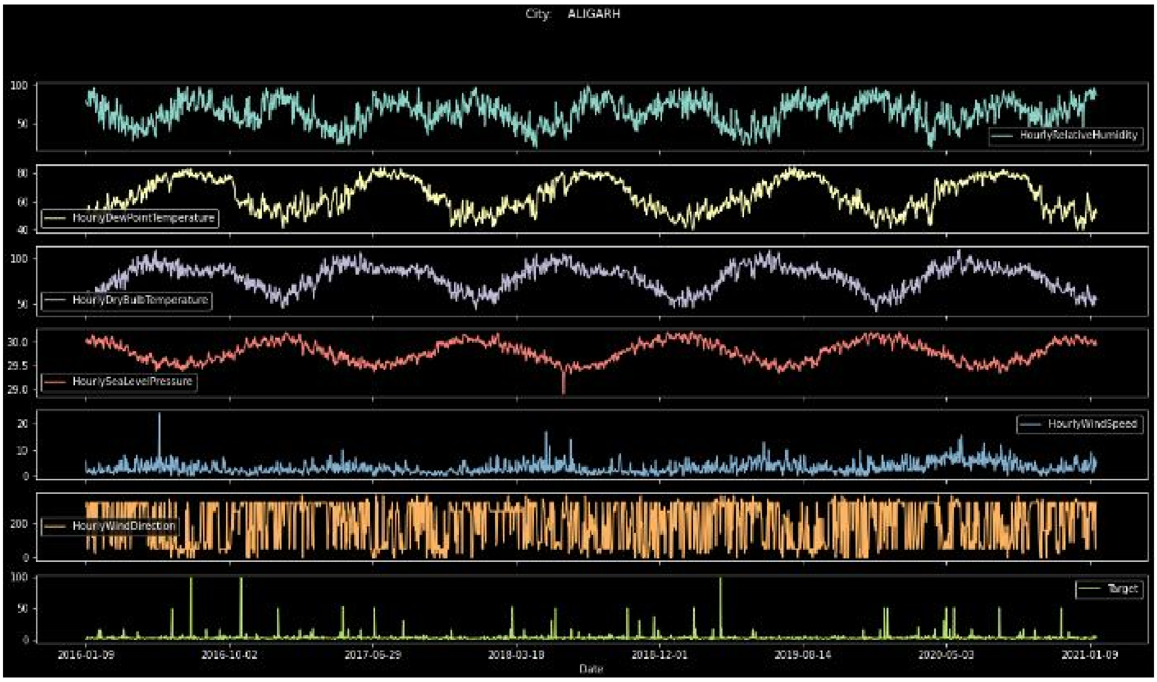


Figure 3: Data distribution for city “Aligarh” used in training regression model.

seasonally periodic, so even after dropping such large number of data points, we haven’t lost most crucial information about this distribution. As shown in figure-3, is the distribution range, for district “Aligarh”, which will be passed into the training model. The distribution of other cities is handled with same approach.

2 Methodology

2.1 Custom made LSTM model

We construct a Regression model for our task. Considering the time-series nature of our dataset, we propose a custom neural network with LSTM cell embeddings. The neural network takes data features as inputs and produces predictions for the Target variable.
The figure-4 below shows the input and outputs of an LSTM for a single timestep. This is one timestep input, output and the equations for a time unrolled representation. The LSTM has an input x(t) which can be the output of a CNN or the input sequence directly. h(t-1)and c(t-1) are the inputs from the previous timestep LSTM. o(t) is the output of the LSTM for this timestep. The LSTM also generates the c(t) and h(t) for the consumption of the next time step LSTM.

An end to end visual architecture can be observed in fig5.

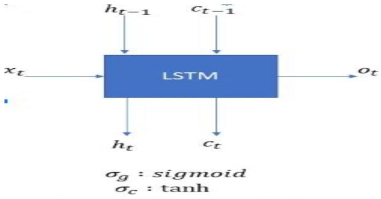


Figure 4: Input and Outputs of a LSTM for single time step

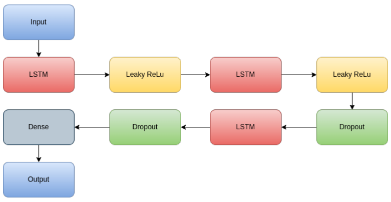


Figure 5: Our LSTM based neural network architecture.

3 Experimental analysis

3.1 MSE scores

For all the cities provided in the dataset, we calculate MSE scores for corresponding Target variable. MSE is a standard evaluation function for Regression based task. Below table summarizes the MSE scores for different cities after training the model for 10 epochs.

We have also attached a visual prediction graph for a continuous 350 days in fig6.

Github: `https://github.com/Abhishek19009/DSML_Project.git`

4 Discussion

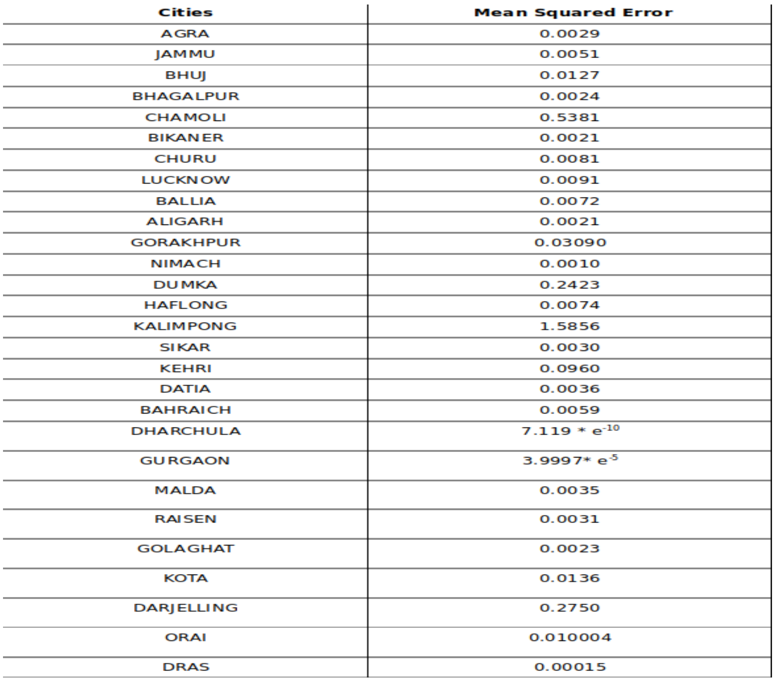
4.1 Limitations and Challenges

Our observation showcased that the Target values are stationary over a range for most of days. Thus for most part, no extreme event was found. There are certain outliers that were minimal in number and hence hard to determine by model as the model generalizes rather than overfits.

The challenges are lack of seasonality in most of the provided feature vectors and the target vector itself. Research needs to be done on imputing the missing values in the dataset, some of the techniques being mean imputation between successive intervals, calculating fourier transformation of feature vectors and performing inverse of it.

5 Contribution

The very first and most essential part in entire project pipeline is the feature engineering, if done carefully, provides not only strong and ideally simple relationships between new input features and the output feature but also more clean and structured data for the learning algorithm to model. However in case of spatial-temporal data, this task becomes more complex due to data's unique structure with focus on both, time and space. My work mainly includes Exploratory Data Anaslysis (EDA), handling missing values, mainly using interpolation methods, handling imbalanced dataset, outliers, feature scaling and converting categorical feature such as location into numerical feature.



Cities	Mean Squared Error
AGRA	0.0029
JAMMU	0.0051
BHUJ	0.0127
BHAGALPUR	0.0024
CHAMOLI	0.5381
BIKANER	0.0021
CHURU	0.0081
LUCKNOW	0.0091
BALLIA	0.0072
ALIGARH	0.0021
GORAKHPUR	0.03090
NIMACH	0.0010
DUMKA	0.2423
HAFLONG	0.0074
KALIMPONG	1.5856
SIKAR	0.0030
KEHRI	0.0960
DATIA	0.0036
BAHRAICH	0.0059
DHARCHULA	$7.119 * e^{-10}$
GURGAON	$3.9997* e^{-5}$
MALDA	0.0035
RAISEN	0.0031
GOLAGHAT	0.0023
KOTA	0.0136
DARJELLING	0.2750
ORAI	0.010004
DRAS	0.00015

Figure 6: Table of MSE scores for different cities.

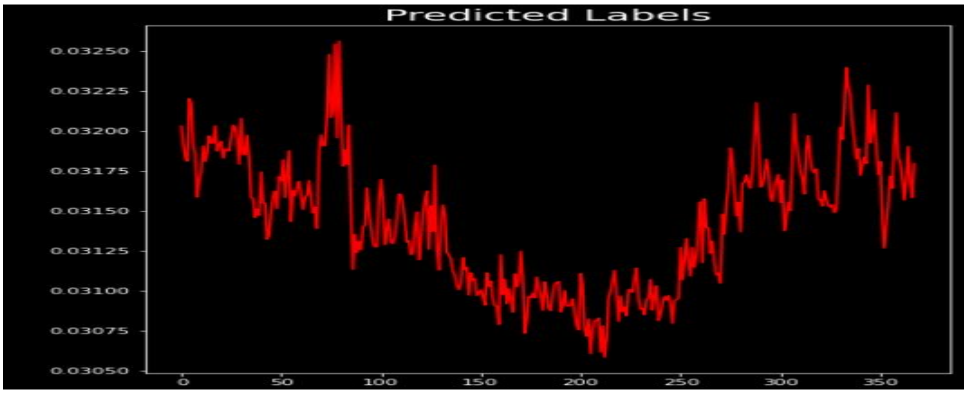


Figure 7: Visual prediction for 350 days.