SPATIOTEMPORAL EXTREME EVENT PREDICTION OVER THE INDO-GANGETIC PLAIN USING MACHINE LEARNING

— PROJECT 5 - PHASE II —

Name:	**Jayesh Kumpawat**
Registration No./Roll No.:	19148
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	February 02, 2022
Date of Submission:	April 24, 2022
Team Member(s):	Shubhajit Dey - 19299

INTRODUCTION

The primary objective of the problem is to find out the spatiotemporal variations in predicting 'visibility' using machine learning techniques. The dataset comprises of meteorological dataset collected over various Indian observation stations.

DATASET

After a preliminary descriptive overview of the dataset it was found that the dataset had 122 features in total and each feature contained 934807 data points. Amongst these 122 attributes there were some which revealed **location coordinates** and there were also some which contained **time coordinates**, thus providing the dataset both **spatial** and **temporal** properties respectively. On further exploration, many relevant information about the dataset were discovered. Moreover, a statistical overview of various features can be found in the code.

PRELIMINARY CHALLENGES AND DATA CLEANING

- There were many features which contained way too many missing values to be considered. To be precise there are almost 100 features which contains more than 930000 missing values. Thus all features which contained more than 900000 missing values were dropped.
- Next, some literature survey was done for some specific features and based on the domain knowledge collected few attributes like 'ShortDurationPrecipitationValue150', 'REM', 'SOURCE' etc. were found highly irrelevant. Those features were dropped.
- Next in the queue comes the target. In the training set, it was found that the target vector contained many noises. Those noises were cleaned using appropriate ways after thorough analysis of the target vector. Similar noises were found features like 'HourlyWindDirection', 'HourlyAltimeterSetting' etc. were found and thus were cleaned in a similar manner.
- The feature 'HourlyPressureTendency' appeared to be interesting in beginning. The graph - code associated with it was manipulatively executed in order to understand the noise levels and their patterns in this feature.

DEALING WITH CATEGORICAL DATA

The dataset contained some feature which comprised categorical data (non-numeric data). Some domain knowledge were used to individually encode this features using separate methods.

- The 'HourlyPresentWeatherType' feature was encoded using 'Base N Encoder'. **N** (base) was chosen to be **6** and as result this one feature got replaced with three encode feature vectors.
- After proper literature survey it was found that in feature 'HourlyWindDirection', the data points 'VRB' meant that the wind speed were changing repeatedly in very less time intervals. Thus such data point were converted to numeric types by using the mean values.
- Encoding of 'HourlySkyConditions' took efforts and significant time. Basically this feature tells us about the the type of cloud/pollution cover in sky by level wise. We cleaned the non relevant and noisy data points and replaced the relevant ones using mean of cloud/pollution cover values of all levels.

DEALING MISSING VALUES

All missing values of the dataset at this point were imputed using the **Iterative Imputer**.

TOWARDS FINAL DATASET

After imputing the missing values, the dataset was almost ready and thus required some minute but crucial touches. One of them being a proper usage of time series attribute. The time series feature was put to work by introducing a new feature 'Lag-Feature' with its help. Basically the fact that visibility in a certain location cannot differ in an unbounded manner was used. This feature is basically the cumulative lag of visibility in various locations that were formed using their chronological stamps.

METHODS

- Classical ML Algorithms like Decision Tree Regressor, Random Forest Regressor and Linear Regression were used on the training dataset but the results obtained on the test sets were not up to the mark according to various performance matrix.
- Now, as it was a spatiotemporal dataset, so it was planned to use the time series data in a hope of better accuracy. As a result we plot the auto-correlation plot in terms of 100 lags, i.e, the previous 100 visibility values when arranged chronologically.

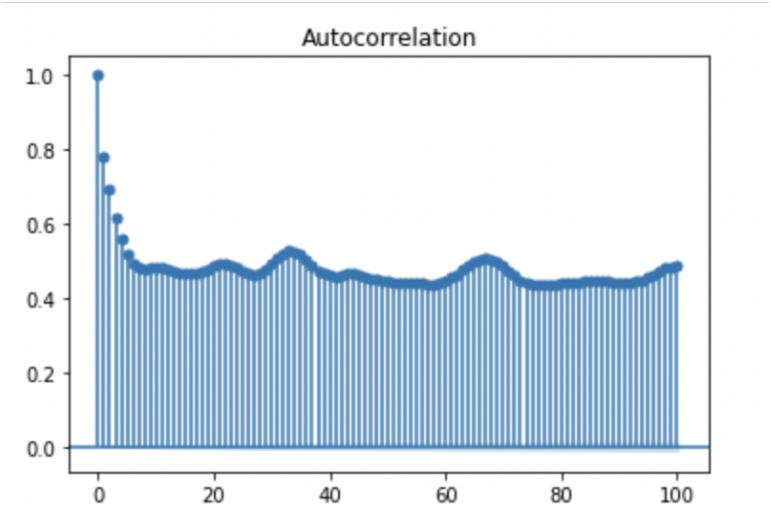

Figure 1: Autocorrelation Plot.

- Shifting by step 1, step 2 we introduced lag 1, lag2 which eventually gave significantly better result. But there was a loophole in the algorithm that those lags were also being taken under consideration which differed in location. This does not make sense because visibility of different locations are not correlated. So, we introduced a novel lag feature which uses previous visibility data from same location.
- Out of the mentioned algorithms, Random Forest Regression gave the best result.

`GitHub` : https://github.com/Jayesh-Kumpawat/SpatioTemporal-Visibility-Prediction

EXPERIMENTAL ANALYSIS

The performance metrics for the regression algorithms used are as follows:

```
MAE = 0.9038537835292219
MSE = 31.307688503630253
R2 = 0.6575345431755273
```

Figure 2: Linear Regression

```
MAE = 0.9038537835292219
MSE = 31.307688503630253
R2 = 0.6575345431755273
```

Figure 3: Decision Tree Regressor

```
MAE = 0.8617355583634753
MSE = 17.041567984255177
R2 = 0.8064704350203773
```

Figure 4: Basic Random Forest Regressor

DISCUSSIONS

- A pipeline comprising feature selection methods and several ML Algorithms was designed but due to several hardware constrains, those codes took humongous amount of time to run. The personal devices present were inefficient to handle such codes with high demands.
- The pipeline also included Grid Search CV to facilitate the desired methodologies but again hardware constraints prove to be a major hindrance.
- If continued then several other techniques like usage of pipelines and methodology of hyper-parameter tuning can be used in the project. This would greatly amplify the scopes of this project.