Spatiotemporal Extreme Event Prediction Over the Indo-Gangetic Plane using Machine Learning

Name:	**Krishnendu J**
Registration No./Roll No.:	20341
Institute/University Name:	IISER Bhopal
Program/Stream:	BS Economic Sciences
Problem Release date:	August 17, 2023
Date of Submission:	November 16, 2023

1 Introduction

In recent years, the Indian climate system has experienced a surge in extreme weather events, including heatwaves, dense haze, droughts, and flash floods. In addition to being life-threatening, these events also have far-reaching socio-economic impacts, such as water availability, food security, employment and public health. A critical step in addressing these challenges is predicting the spatiotemporal variations of these extreme events. Visibility, a key meteorological parameter, is pivotal in assessing and responding to extreme weather. This project leverages machine learning to forecast visibility over the Indo-Gangetic plane in India. We analyse datasets collected from Indian observation stations dating from 1950 to the present, primarily focusing on predicting mean visibility., e.g., Figure 1.

2 Methods

In this study, we first perform data preprocessing and feature engineering before applying machine learning algorithms to predict visibility. The focus is to clean and condition the spatiotemporal dataset to ensure quality inputs for the model. All codes can be accessed here.

2.1 Data Preprocessing

The initial approach involved handling missing data by removing features with more than 50% missing values and rows with any missing values or placeholders indicative of erroneous data (value 9999.9), as reported in the data description. Furthermore, columns with mixed data types were coerced to numeric, facilitating subsequent operations. Outliers were identified and removed using a defined standard deviation threshold (2*SD), improving the robustness of the dataset against extreme values that could skew the model training.

2.2 Feature Engineering

As part of our method, certain redundant or non-informative features were dropped. Features likely to impact the predictive model significantly were retained and cleaned. Data type coercion and handling of outliers were performed to ensure the integrity of the input data.

2.3 Tools

The Python library Pandas was used for data manipulation, NumPy for numerical operations, and Matplotlib along with Seaborn for data visualization, providing insights into the data distribution post-cleanup. Scikit-learn library was used to develop and evaluate the machine learning models used. We develop four models using linear regression, LASSO [1] and multilayer perceptron [2] algorithms.

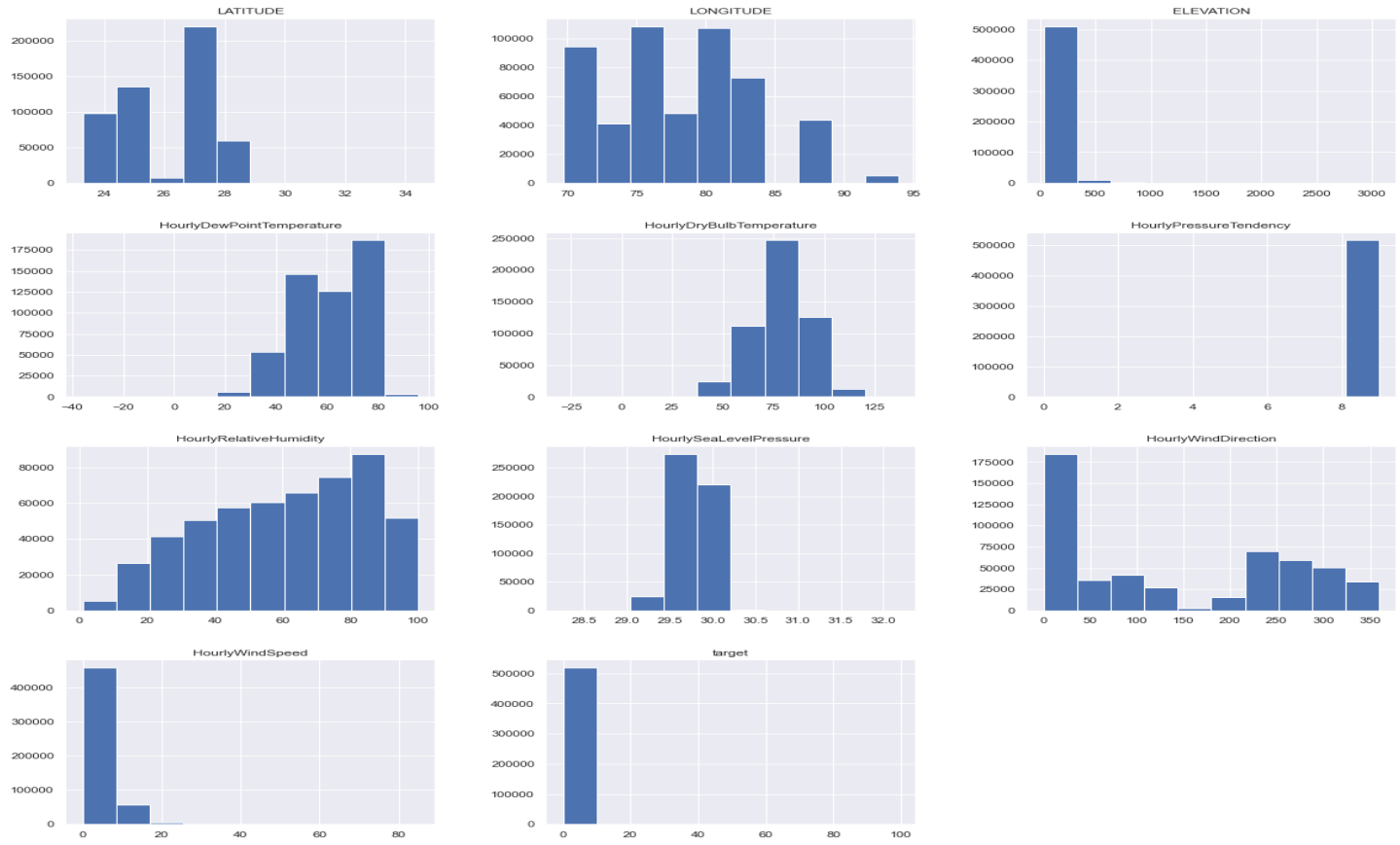

Figure 1: Distribution of features in the dataset

3 Experimental Setup

The evaluation of various models is based on standard metrics such as mean squared error (MSE), mean absolute error (MAE) and R-squared. These metrics were chosen for their ability to provide a holistic view of the model performance. The performance of the regression methods performed is reported in Table 1.

Table 1: Performance of Different Regression Methods

Technique	**MSE**	**MAE**	**R-squared**
Linear regression	0.75	0.57	0.27
k-NN regression	0.63	0.45	0.38
LASSO	1.02	0.67	-0.00006
Multilayer perceptron	0.56	0.45	0.45

4 Results and Discussion

Based on the evaluation metrics, the Multilayer Perceptron (MLP) appears to be the best-performing model for this particular problem, as it has the lowest errors and the highest R-squared value. However, the R-squared value is still below 0.5, suggesting that there is room for improvement.
The k-NN model also shows relatively better performance than Linear Regression and LASSO, which could mean that the dataset has non-linear patterns that k-NN and MLP are better able to capture.

5 Conclusion

We observed a spectrum of performance metrics after applying various regression techniques to predict visibility using spatiotemporal data of the Indo-Gangetic plain, using Linear Regression, k-NN Regression, LASSO, and Multilayer Perceptron (MLP). The MLP exhibited the most promising results, indicating a better fit and stronger predictive capabilities relative to the other models.
Despite these efforts, the performance across all models remained suboptimal, with R-squared values suggesting that a significant portion of the variance in the target variable remained unexplained. This outcome has highlighted several potential pathways for enhancement.
For future work, we recommend delving deeper into feature engineering to unearth more predictive signals. Sky condition feature may be parsed and used for predictive modelling.

References

[1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[2] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197, 1991.