Spatiotemporal extreme event prediction over the Indogangatic plane using Machine Learning

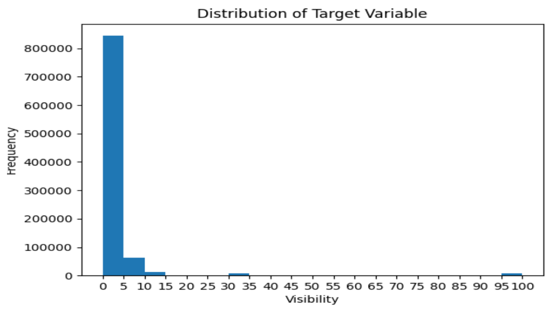
Name:	**Toraskar Saurabh Mahesh**
Registration No./Roll No.:	21290
Institute/University Name:	IISER Bhopal
Program/Stream:	EES(Earth and Environmental Sciences)
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

1 Introduction

The project aims to develop a robust model to forecast visibility conditions for Aviation and Maritime based on a set of weather parameters. The dataset used contains time-series data for various meteorological station across Gangatic plain.

1.1 Challenges

- Large amount of Feature Vectors :- The raw data for the project the training set has 934807 feature vectors, while the test set has 103868 feature vectors.Due to this, the computational complexity is elevated which will result in prolonged training and testing times.
- Highly Skewed Data:- The Target variable is visibility which is very highly skewed with 90% values in 0-5 range as shown in fig1.1



- Mixed type of features :- The given data set contains Numeric features(e.g.- Pressure,Temp) as well as Non-Numeric features (e.g.-Weather type ,Report-Type)
- Large number of Missing Values :- After extracting the useful features, 715815 feature vectors had at least one NaN Value,which is around 76% of the total feature vectors
- Noisy Data:- In the feature vectors an extra character ‘s’ is present in the numeric variables, whereas in target variable ‘V’ is present

2 Methods

Link to github repository
We tackle the Challenges mentioned in 1.1 in a step by step manner.The data handling and analysis of the project were conducted using the **Pandas library**, a tool in the Python ecosystem designed for

Table 1: Performance on various Imputing Techniques on Numeric Features

Evaluation Parameter	Error Metric	Linear Regression	Decision Tree Regressor
Mean	MSE	59.13	55.21
	RMSE	7.68	7.43
	R-2 Score	0.342	0.386
Median*	**MSE**	**59.06**	**52.85**
	RMSE	**7.68**	**7.27**
	R-2 Score	**0.339**	**0.408**
Back-fill	MSE	59.90	55.05
	RMSE	7.73	7.41
	R-2 Score	0.342	0.395

Table 2: Imputing Techniques on Categorical Features(Decision Tree Regressor)

Error Metric	**None***	Back-fill	Mode
MSE	**34.78**	42.85	43
RMSE	**5.89**	6.54	6.58
R-2 Score	**0.62**	0.53	0.53

efficient data handling, manipulation and analysis. Pandas provided a comprehensive set of functions for various tasks, from initial data exploration to preprocessing and feature engineering. The methods used for this are discussed below in detail

2.1 Feature Engineering

Only features with less than 2% NaN values are selected with exception of 'HourlySeaLevelPressure' and 'HourlyPresentWeatherType'. The other features rejected have more than 30% NaN values.Columns:- [STATION,LATITUDE,LONGITUDE,ELEVATION, NAME] give essentially the same info,thus only 'ELEVATION' is kept.
From "Date" feature, which has DateTime datatype, 'hour' which gives implied information on the intensity of sunlight was extracted, also 'month' was extracted which provides valuable insights into the seasonal variations,both of this features can contribute key insights for subsequent prediction of visibility.
The Character Anomaly was removed after isolating relevant features.Imputing NaN values is discussed in detail in 2.1.1

2.1.1 Imputing NaN Values

The data set set has mixed features i.e. Numerical features and also categorical features.We tackle both of them individually. First the model in trained on only numerical features using different Imputation techniques as shown in Table table1.Imputation for categorical features is done after imputing numerical features with 'Median' on Decision Tree Regressor.'None' refers to techinque where the NaN values are not imputed so that in encoding it becomes a new value .The results are given in Table 2 The methods chosen are written in bold and marked by a asterik

3 Experimental Setup

We have utilized the **scikit-learn library**, a popular machine learning library in Python, to implement various algorithms and functionalities.For scaling:- StandardScaler, MinMaxScaler and RobustScaler were considered and the scaler with the best performance was chosen which was RobustScaler. Since the problem is of Regression, the metrics used are MSE,RMSE,R2-Score to evaluate our models.The lower the values of MSE and RMSE and closer the value of R2-Score, the better the model is performing.
The different parameters explored for regressors are given below:-
AdaBoost Regression:- estimator,
DecisionTreeRegressor:- criterion,max_features,max_depth,ccp_alpha
RidgeRegression:- solver
LinearRegression:- positive
RandomForestRegressor:-n_estimators,max_depth,max_features
SVM Regressor:- *The model was trained only on default parameters*
KNeighbors Regressor:- n_neighbors

Grid search was implemented to get to get the best parameters and evaluation metrics are calculated after training and testing the model on best parameters.

4 Results and Discussion

Some or the models experienced prolonged execution times, e.g.
SVM Regressor on only default parameters took around 2.5 days,so Grid search implementation was unfeasible,also RandomForest Grid Search took around 2 days
The results of the best parameters and also models with good results on default parameters are given in Table 3.Scaling showed significant improvemnet in Ridge and Kneighbors Regression whereas almost no improvement is seen in Linear Regression which was not expected.From Table 3, we can see that Tree algorithms perform better on weather data set with Random Forest(which is Ensemble Tree Algorithm) on default parameters giving the best results.

5 Future Scope

- In Aviation and Maritime sector, particularly in metropolitan areas, pollution emerges as a prominent influencing factor, significantly impacting visibility conditions. Hence, pollution can be incorporated by introducing while developing a model for visibility prediction .
- To implement this model of a global scale, we would also have to consider real-time GPS data, to account for different time zones and also hemispheres.

6 Conclusion

We conclude that Random Forest Regression is the most fitting regressor for visibility prediction. Due to the huge computational complexity and long training and testing times and also lack of knowledge the models could not be cross-validated which could be done in the future. The most significant limitation in visibility prediction lies in the substantial variability exhibited by weather conditions, posing a significant challenge to accurate forecasting models

Table 3: Performance Of Different Classifiers Using All Features

Regressor	MSE	RMSE	R2-Score
Random Forest *max_depth=50* *max_features=1*	30.64	5.53	0.664
Random Forest(default)	29.22	5.4	0.678
Decision Tree *max_depth=12* *ccp_alpha=0.02*	35.30	5.9	0.61
Linear Regression (*positive=True*)			
–without scaling	59.9	7.74	0.343
–with scaling	59.8	7.73	0.344
Ridge Regression (*positive=True*)			
–without scaling	89.6	9.46	0.012
–with scaling	60.03	7.74	0.342
AdaBoost (*random_state=0*)			
base_estimator: LinearRegression	80.53	8.9	0.11
base_estimator: DecisionTreeRegressor	32.29	5.7	0.642
KNeighbors (*n_neighbors=5*)			
–without scaling	62.46	7.90	0.31
–with scaling	34.90	5.90	0.62
Support Vector Machine(default)	90.19	9.49	0.012