EXL

IIT Kharagpur

EQ 2023 Round 2 Submission

Ritwik Jain
Ninad Sahu

EXL

- Understanding the Statement
- Identifying Key Patterns
- Selection of Modelling Methodologies
- Solution Design
- Model Predictions and Recommendations
- Evaluation of Strengths and Weaknesses

UNDERSTANDING OF THE PROBLEM STATEMENT

WHAT IS PM 2.5 PREDICTION? →

PM 2.5 pollution comes from a variety of sources, both natural and human-made. Natural sources include dust storms, wildfires, and volcanic eruptions. Human-made sources include transportation, power generation, industrial processes, and agriculture. Many people's lives are harmed due to pollution.

WHY IS IT NECESSARY? →

It is necessary to predict PM 2.5 because this information can be used for developing and implementing effective pollution control strategies and regulations

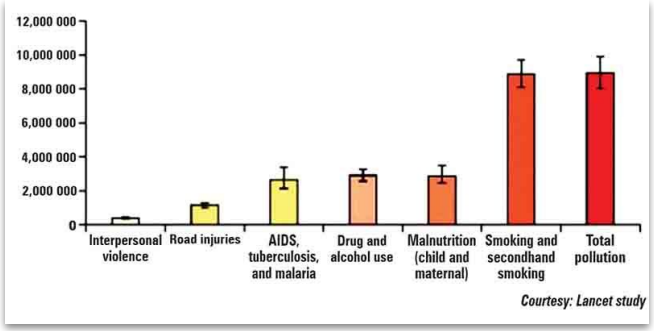
OUR UNDERSTANDING AND STRATEGY

Our understanding is that the PM 2.5 has a seasonal trend as well as it is dependent upon the various feature values given we model it as a multivariate time series prediction

CHALLENGES IN PREDICTION

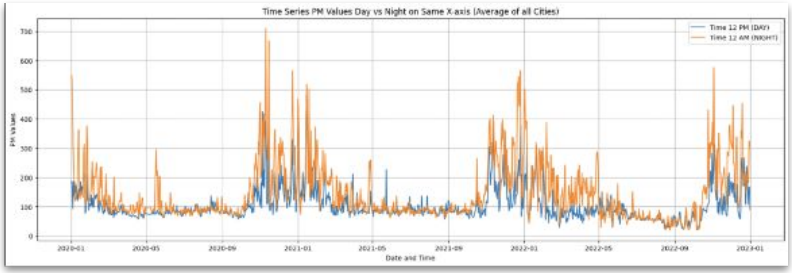
Selection of the right model and its hyperparameters is a difficult task

Too many features for prediction would add unnecessary noise hence the optimal set of features need to be selected.



Total pollution related deaths in World

IDENTIFYING KEY PATTERNS FROM THE DATA

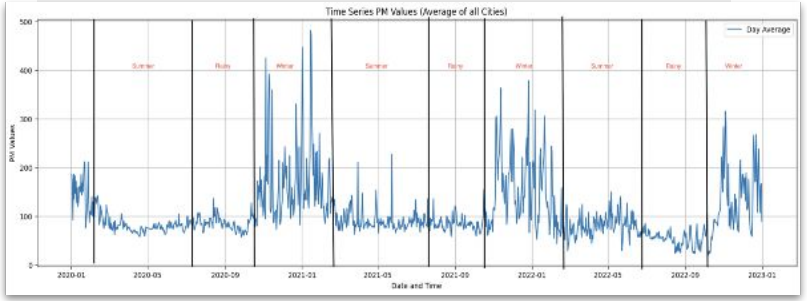


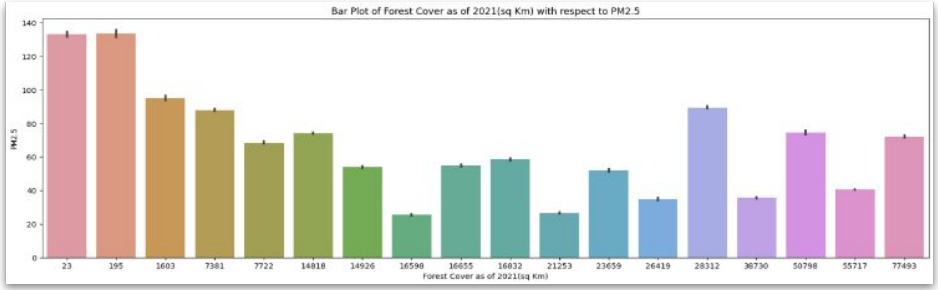
A comprehensive analysis was conducted to study the variations in PM2.5 levels, incorporating both day and night readings.

PM2.5 levels were consistently higher during the night compared to the levels observed during the day.

Comprehensive analysis conducted: Daily average PM values plotted against seasons and seasonality test performed.

Winter months showed notable increase in PM levels due to decreased temperature and reduced particle entropy.





Relationship between Forest Cover and PM2.5 values analysed in various cities using dataset.

Trend observed: Increase in Forest Cover area correlates with a decrease in PM2.5 values.

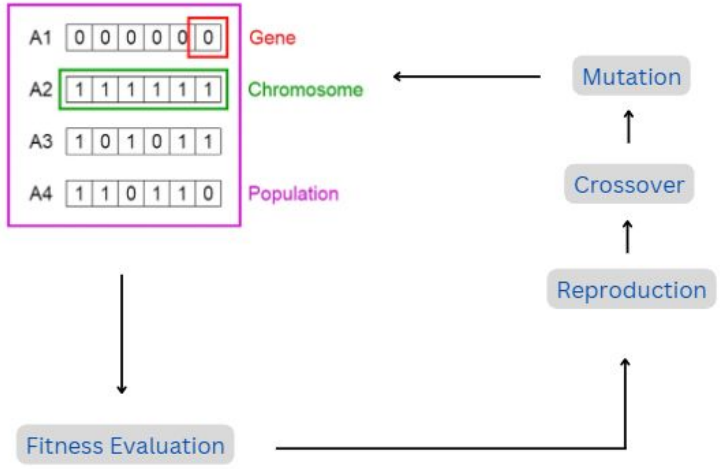
SUMMARY OF SELECTION OF MODELLING METHODOLOGIES

- Understanding the Statement
- Identifying Key Patterns
- **Selection of Modelling Methodologies**
- Solution Design
- Model Predictions and Recommendations
- Evaluation of Strengths and Weaknesses

DATASET (ALL FEATURES AS INPUT) → GENETIC SELECTION OF MODEL AND HYPERPARAMETERS → BEST MODEL ACCORDING TO GENETIC SELECTION IS CHOSEN → SELECTION OF FEATURES BY GENETIC ALGORITHM

In a genetic algorithm, chromosomes are made up of each hyperparameter where the decimal of that parameter value is each gene and , and during feature selection it represents the feature vector with 1 meaning the feature is selected whereas 0 means it is not

Various chromosomes make up a population



Mutations also occur on a probabilistic basis where genes of an offspring are mutated

Crossovers happen on a probabilistic basis where genes of the parents chosen for reproduction are interchanged

Two of the fittest individuals are chosen for reproduction

Fitness is evaluated based on rolling cross validation accuracy on the training set.

SOLUTION DESIGN

Search Space

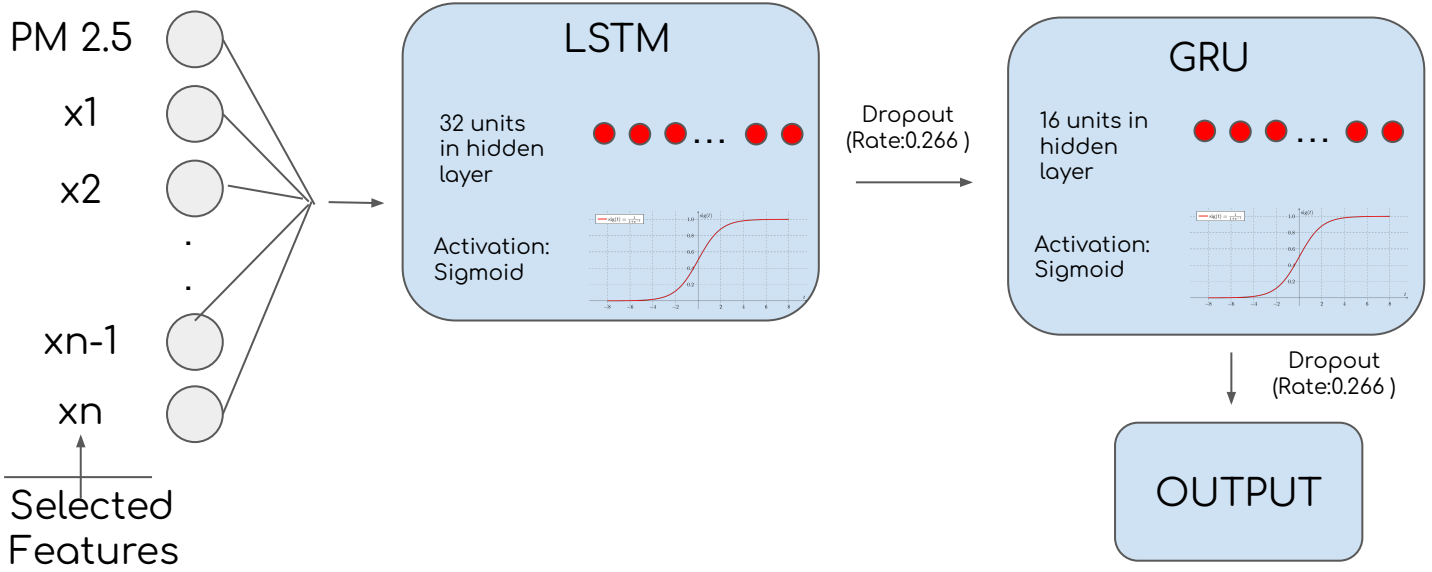
'learning rate' – np.linspace(start=0, end=1, num=35);

'dropout rate' - np.linspace(start=0, end=0.5, num=25); 'layer size' - [1,2,4]; 'unit size' – [8, 16, 32];

'activations' - ['relu', 'sigmoid', 'softmax', 'softplus', 'softsign', 'tanh', 'selu', 'elu'];

'models' – ['GRU', 'LSTM', 'RNN'];

The search space for the genetic algorithm is shown above, out of the various combinations the best model pipeline as determined by the genetic algorithm is shown below



MODEL PREDICTIONS AND POSSIBLE RECOMMENDATIONS

Learning Rate	Dropout Rate	Layer Size	Hidden units	Activation	Models	MAPE
0.08912	**0.266**	**2**	**[32, 16]**	**Sigmoid**	**[LSTM, GRU]**	**0.221011**
0.017318	0.349	1	[8]	Relu	[LSTM]	0.309722
0.089739	0.149	2	[16, 16]	Softsign	[LSTM, RNN]	0.335088
0.045603	0.399	2	[32, 32]	Sigmoid	[GRU, RNN]	0.405112
0.076634	0.358	1	[32]	Elu	[GRU]	0.527352

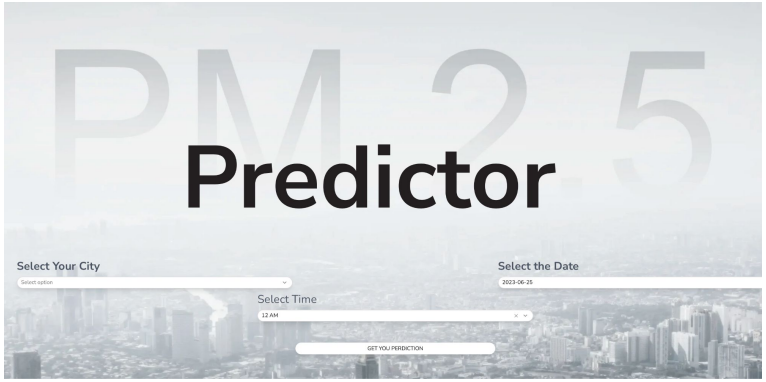
The MAPE of the predictions for the top five models as calculated by genetic algorithm of the five fold rolling cross validation on the training dataset

1. Muzaffarpur
2. Gaya
3. Patna
4. Gwalior
5. Delhi

The top five most polluted cities according to our model

EVALUATION OF STRENGTHS AND WEAKNESSES OF OVERALL FRAMEWORK

Using the trained weights we have created an application so that the predictions are available to both governments and corporates as well as individuals.



Our App Link

Advantages: The use of Genetic Algorithms allow us to overcome the problem of trial and error for the selection of the model pipeline and the hyperparameters, as far as feature selection is concerned it is a great tool when we have many features as in our case as it can give us an optimal set of features which gives high accuracy as well as reduces the prediction time.

Drawbacks: Genetic algorithms can be computationally expensive, especially when dealing with large and complex problem spaces. As the number of variables, constraints, and fitness evaluations increases, the time required to find a solution also increases.