D45

Open IIT - Data Analytics

TEAM Data_Ninjas

AGENDA

Data Extraction

Exploratory Data Analysis

Data pre-processing & Feature Engineering

Forecasting Models

Results

Extraction Sources

Government datasets

Tourism Annual Reports

IndiaStat

Google Trends

Kaggle

Yahoo WebScope

Extraction Techniques

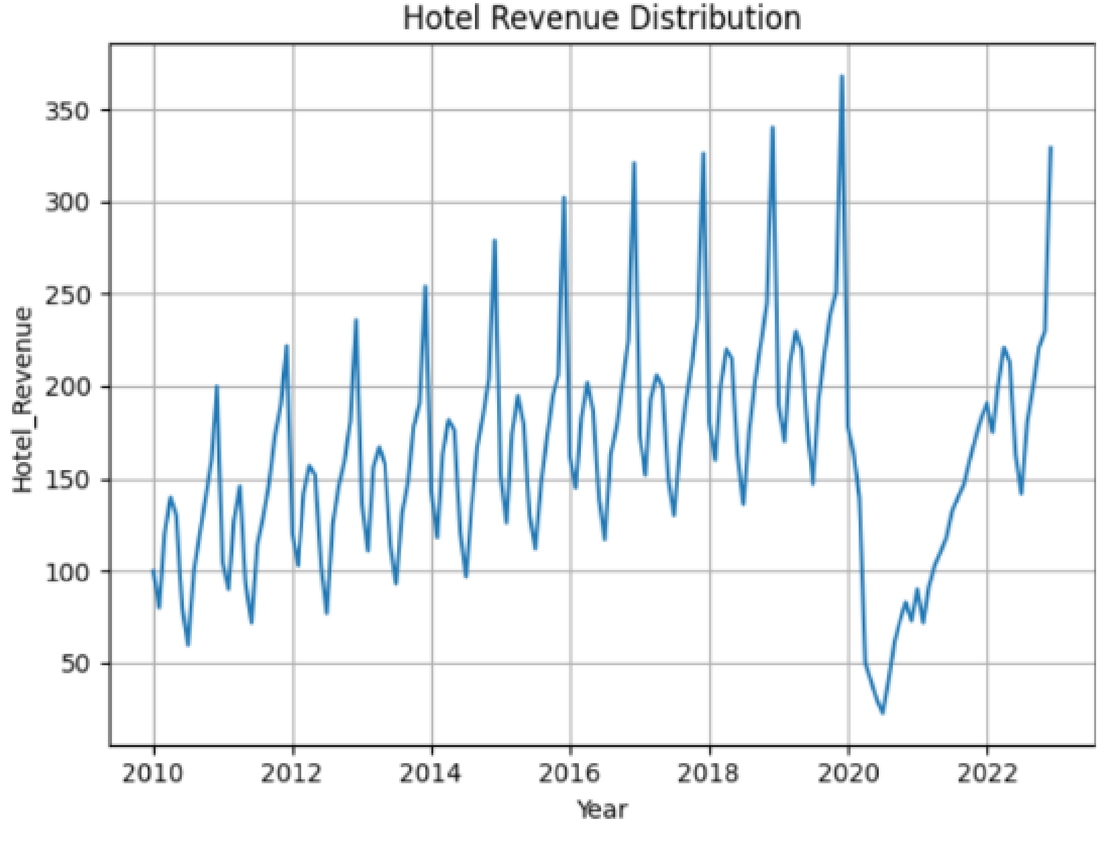
- Monthly and weekly data corresponding to preselected features like month and year, search index, hotel revenue and their corresponding tourist arrival data were obtained from several sources. Strings were converted to CSVs and eventually read into dataframes.
- CSVs for each year were compiled to form a dataset spanning from 2010 to 2022

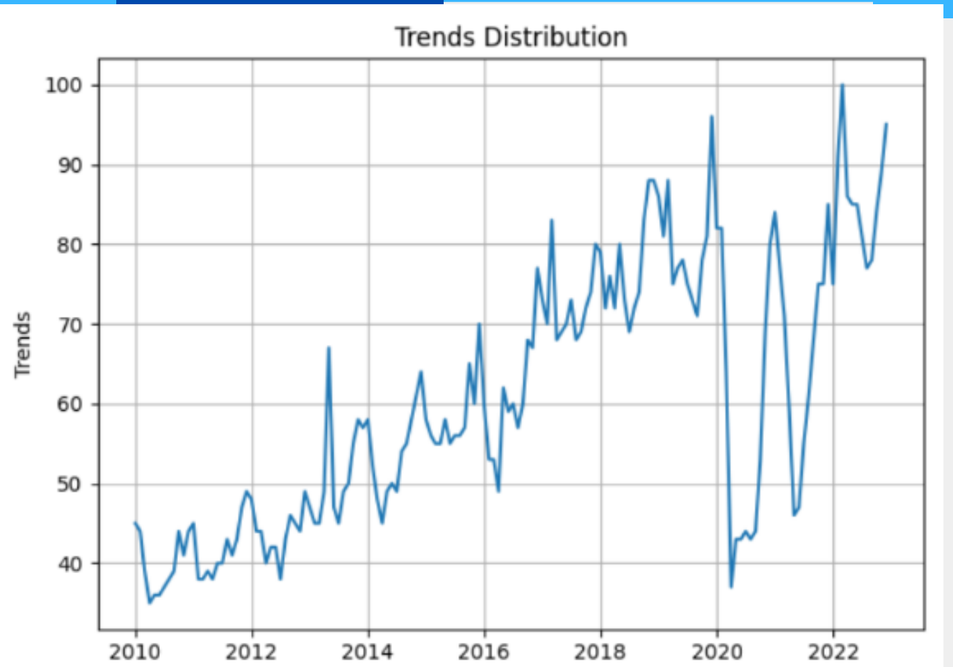
Feature selection

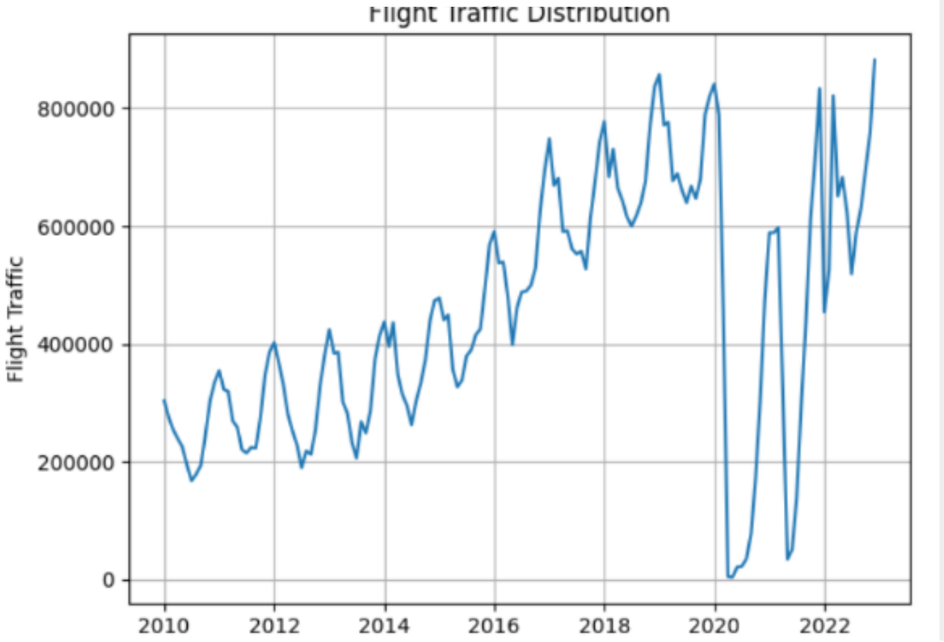
- Time Series forecasting model neccesisates month,year and number of tourists as a feature.
- Flight bookings to the destination have a high correlation with tourist data
- Hotel revenue is directly correlated with the number of tourists utilising lodgings
- Parameters like crime rate have a high negative correlation with number of tourists, indicating an inverse relationship

EXPLORATORY DATA ANALYSIS

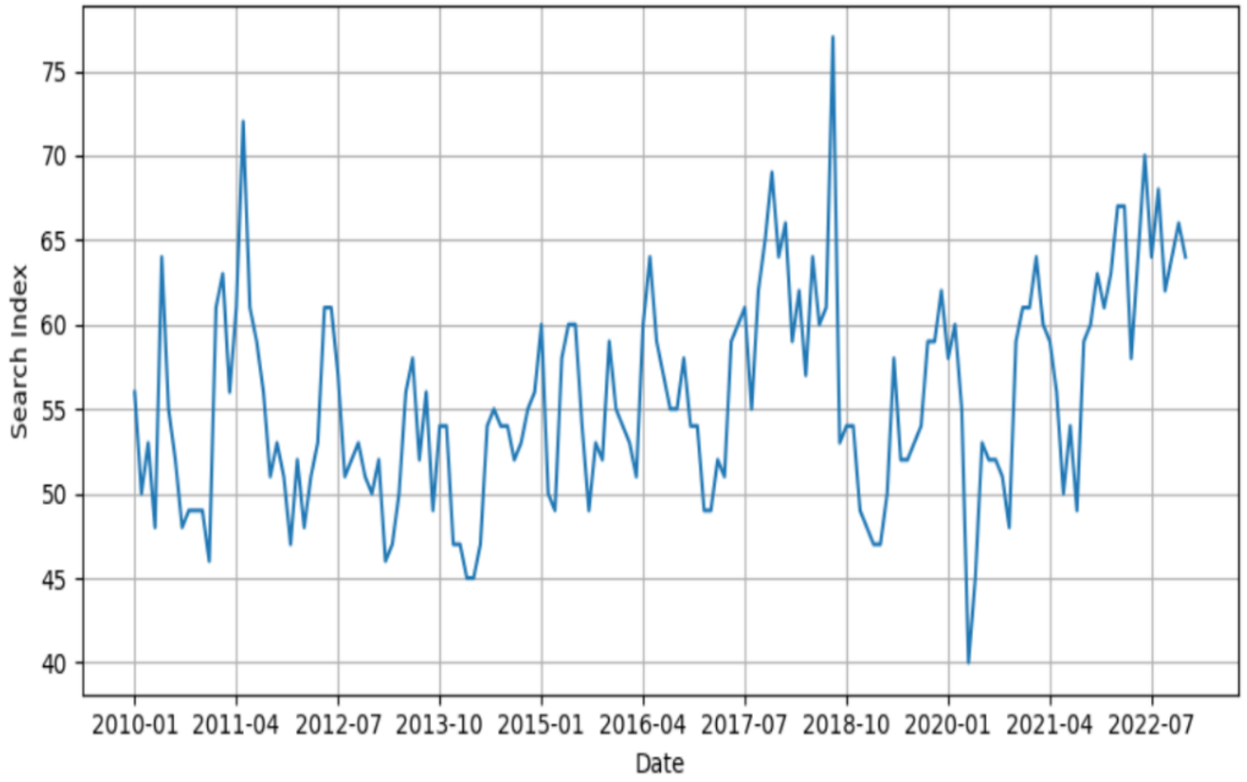
Goa

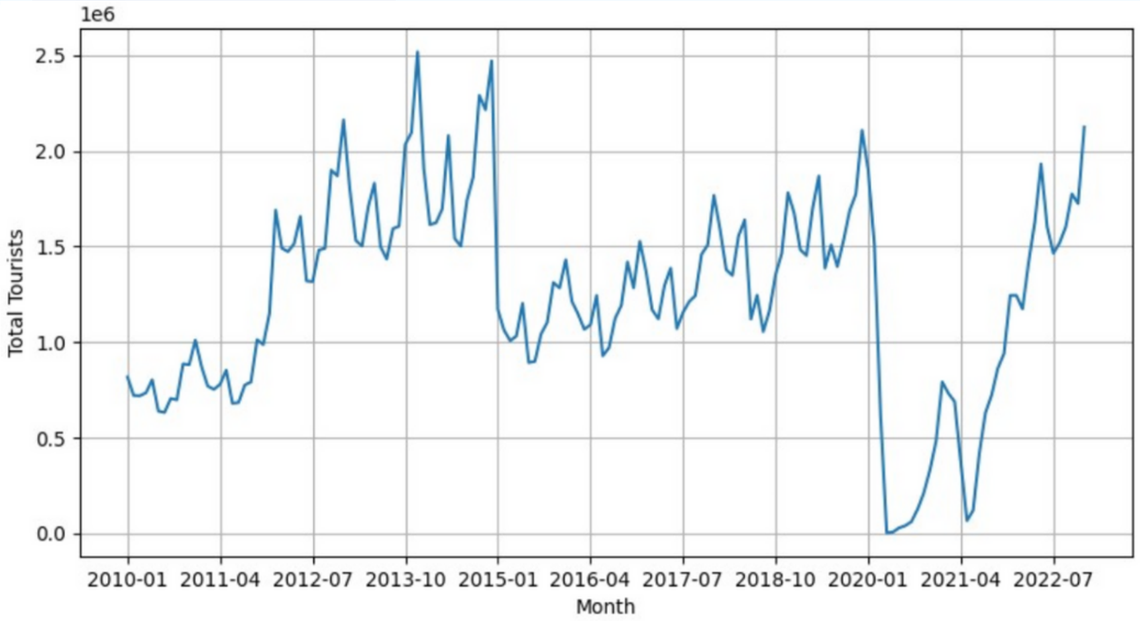


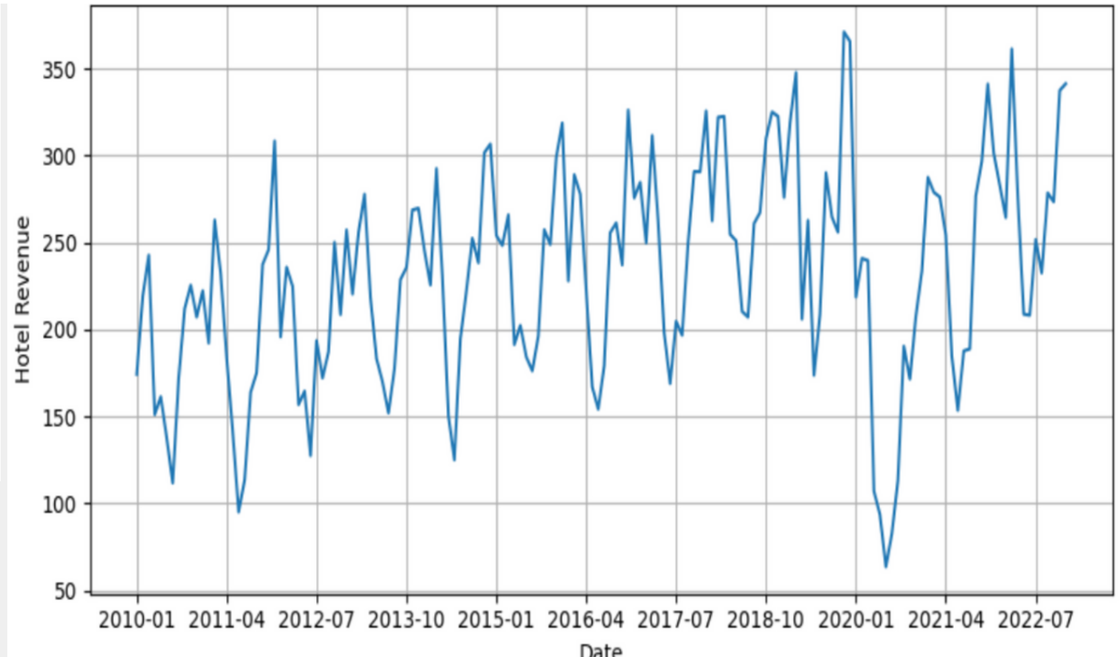




Kerala

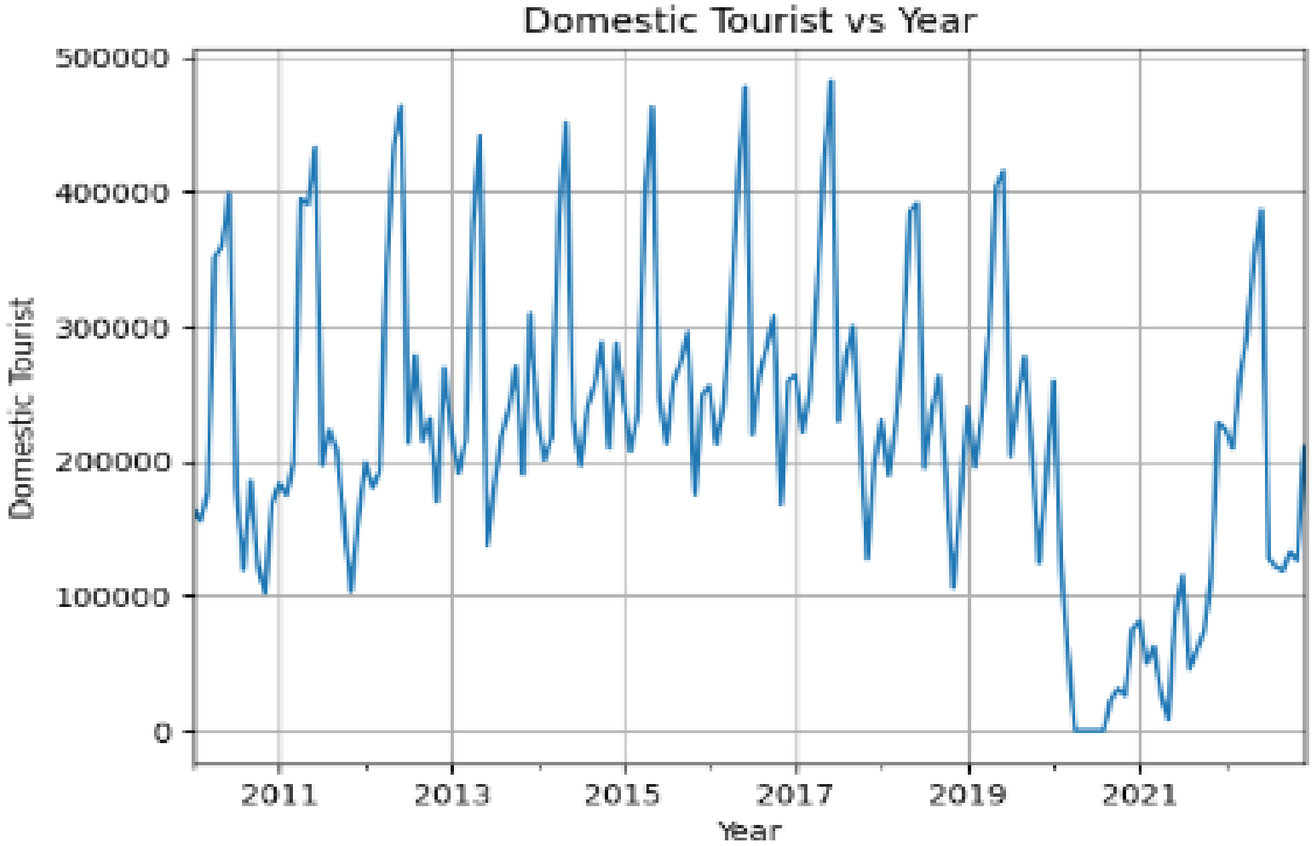

Search Index
75
70
65
60
55
50
45
40
2010-01
2011-04
2012-07
2013-10
2015-01
2016-04
2017-07
2018-10
2020-01
2021-04
2022-07
Date

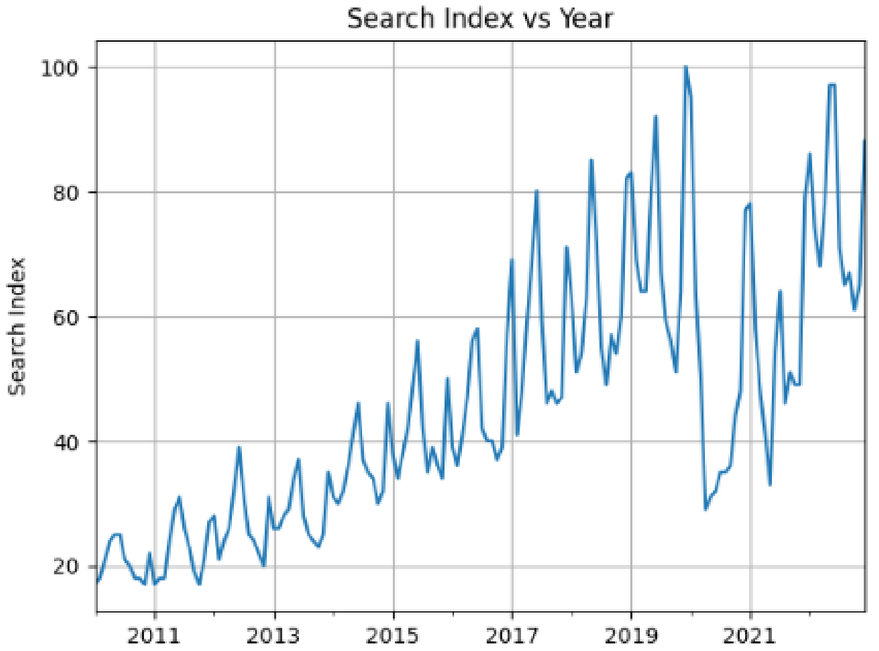

1e6
Total Tourists
2.5
2.0
1.5
1.0
0.5
0.0
2010-01
2011-04
2012-07
2013-10
2015-01
2016-04
2017-07
2018-10
2020-01
2021-04
2022-07
Month

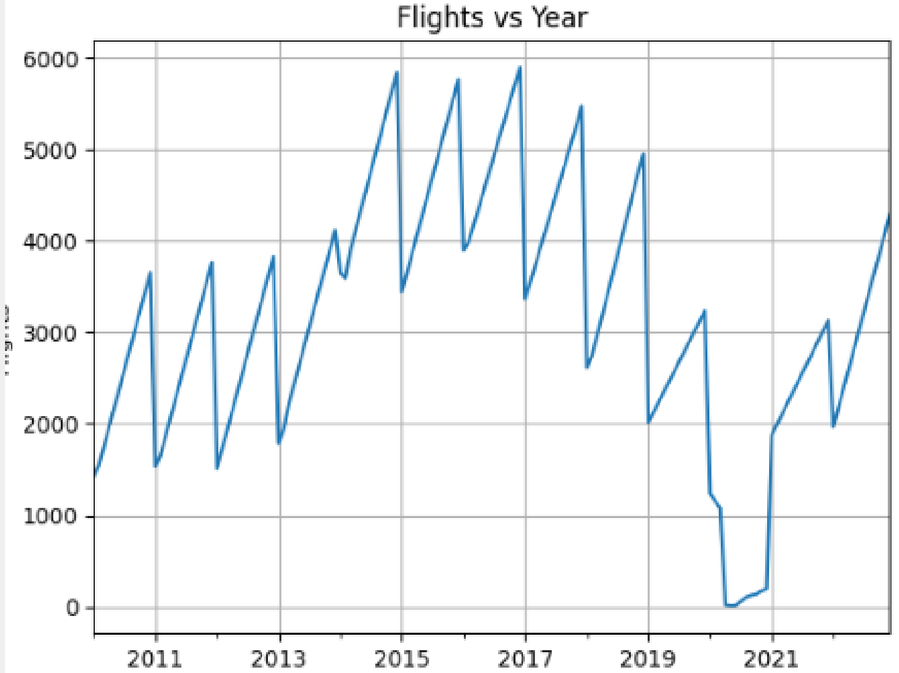

Hotel Revenue
350
300
250
200
150
100
50
2010-01
2011-04
2012-07
2013-10
2015-01
2016-04
2017-07
2018-10
2020-01
2021-04
2022-07
Date

EXPLORATORY DATA ANALYSIS

Shimla

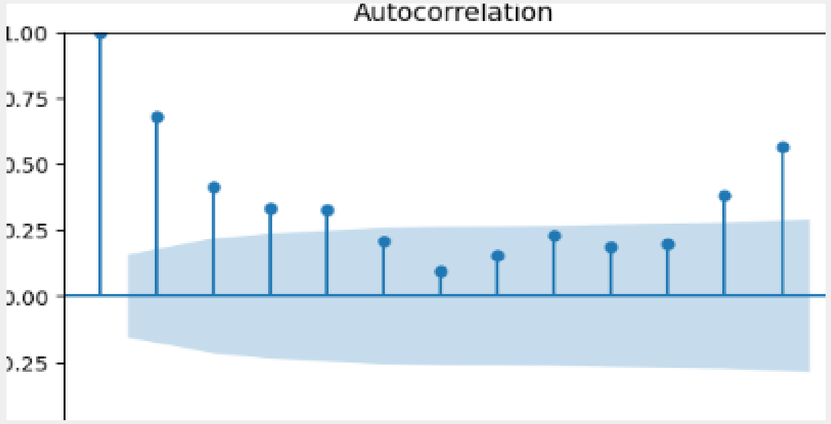




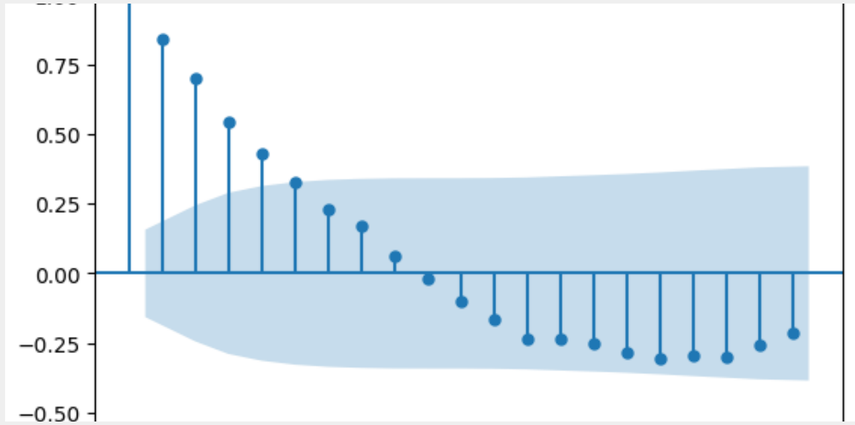


Auto Correlation plots for different states

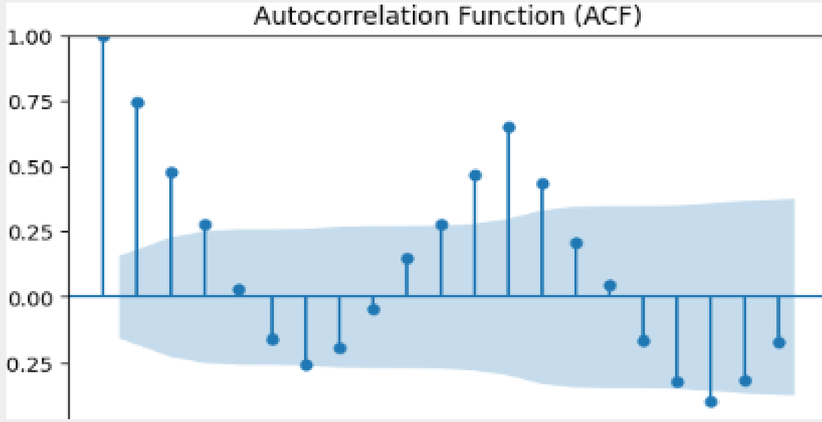
Autocorrelation helps to identify patterns in the time series data, which can provide insights into the behavior of the variable over time. This information can be useful for understanding the underlying factors that affect the variable and for making informed decisions.



ACF plot for Shimla



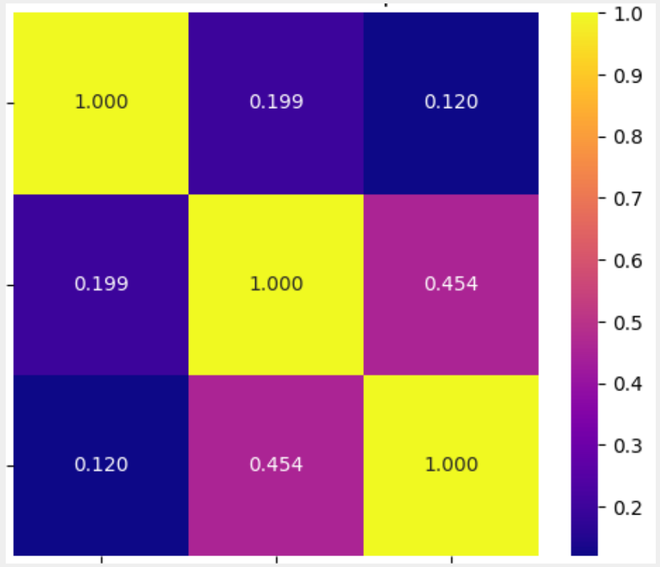
ACF plot for Kerala



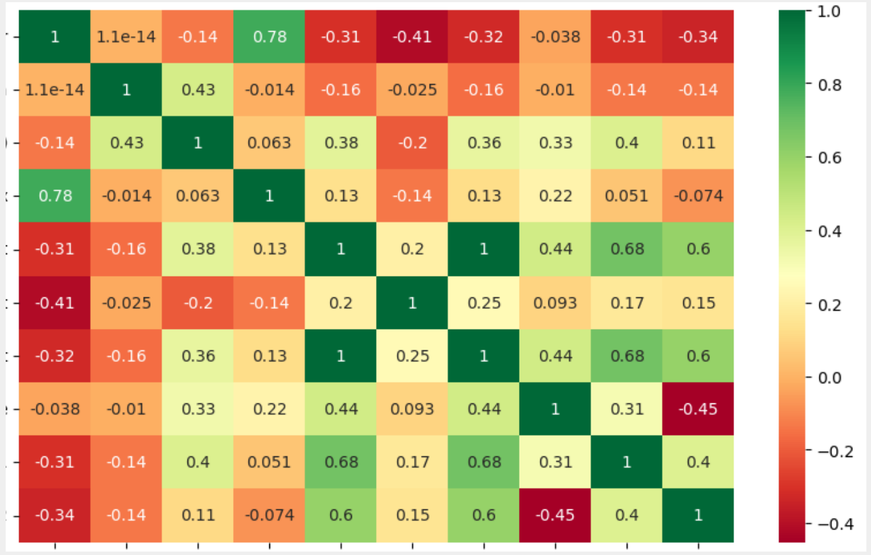
ACF plot for Goa

Correlation Heatmap for Different States

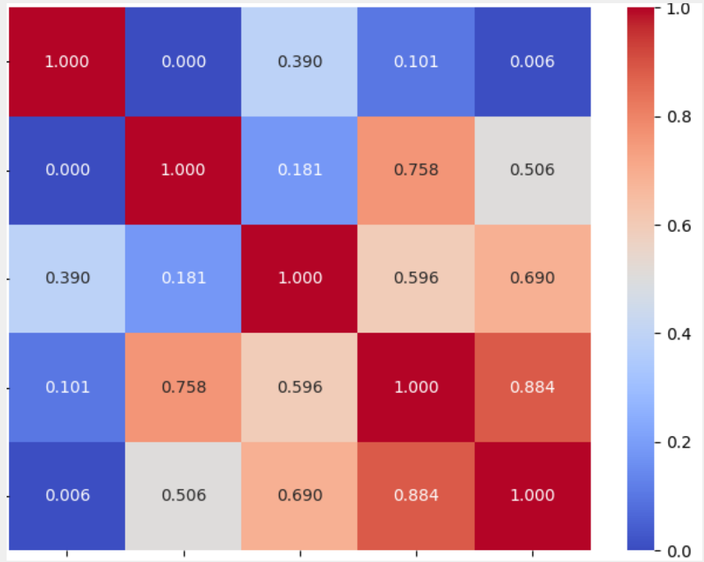
A correlation heatmap visually displays relationships between variables, using colors to represent the strength and direction of correlations in data.



Correlation for Kerala



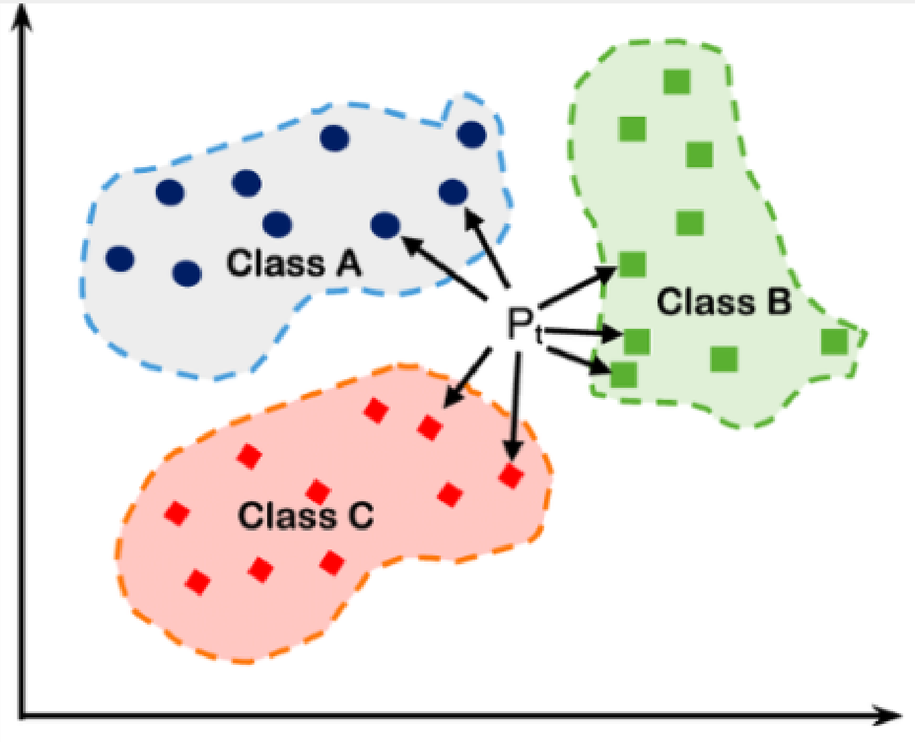
Correlation for Shimla



Correlation for Goa

Imputation using KNN

- A non-parametric, instance-based supervised learning algorithm for classification and regression.
- Predicts based on proximity to 'K' closest training examples in feature space
- Defers training until prediction, storing data for fast retrieval.
- Determines the number of neighbors considered for prediction, influencing model sensitivity.



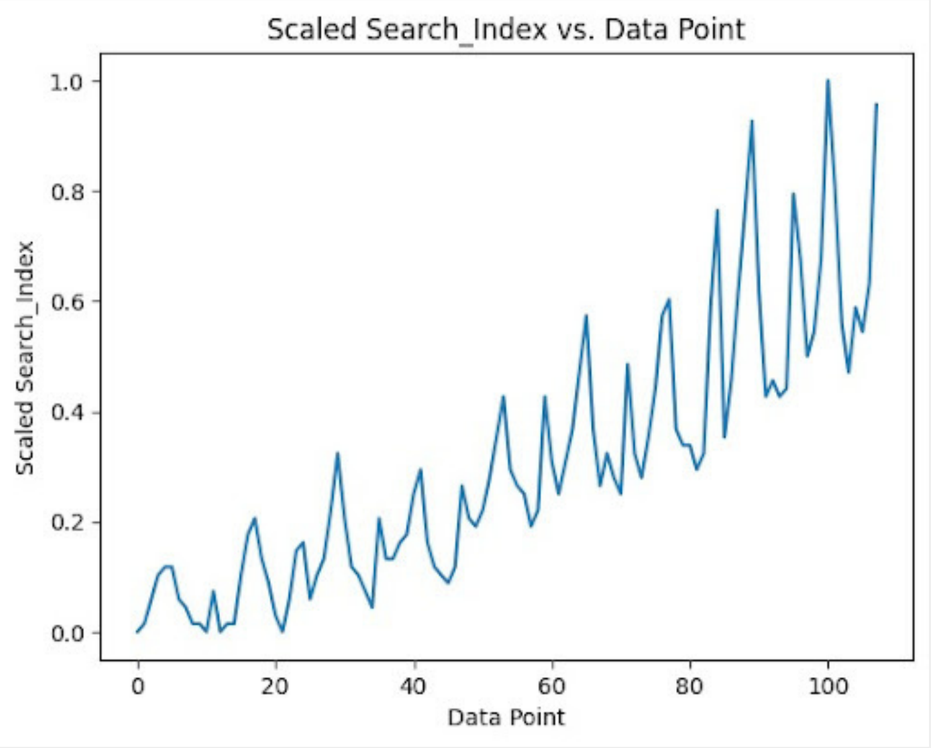
Scaling and Normalization

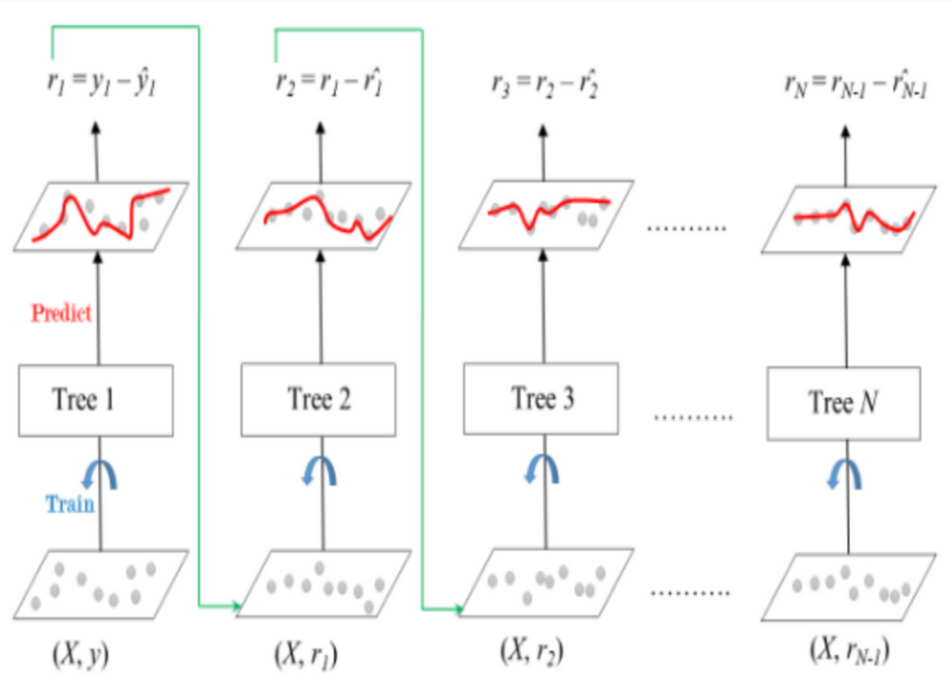
Scaling for Feature Consistency:

- Scaling techniques, such as Min-Max scaling or Standardization (Z-score scaling), are used to bring features to a consistent scale.
- This ensures that no single feature dominates the learning process, helping algorithms converge faster and improving model performance.

Normalization for Data Distribution:

- Normalization helps address skewed data distributions, making the data more suitable for algorithms that assume a normal (Gaussian) distribution.





Best model

- **Gradient Boosting** R2 Score **0.97**

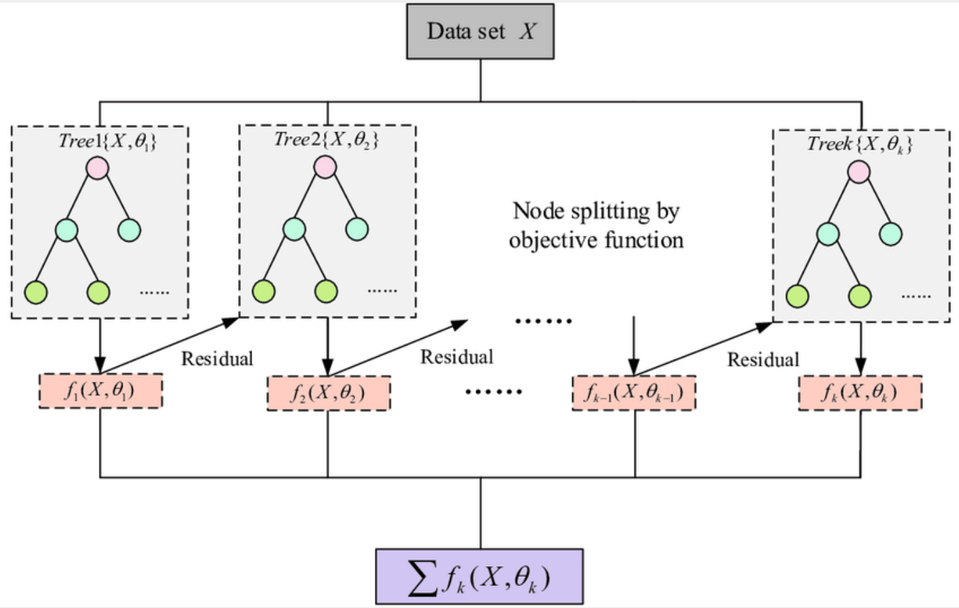
GOA Gradient Boosting

Gradient Boosting combines weak learners to create a strong learner. It iteratively corrects errors by calculating gradients, training weak models to approximate these gradients, and adding their predictions. This process repeats until a stopping criterion is met.

Why Gradient Boost?

Gradient Boosting performs here better for its technical advantages. It's an ensemble learning method that minimizes loss by iteratively adding decision trees, optimizing through gradient descent.

GOA XGBoost



Best model

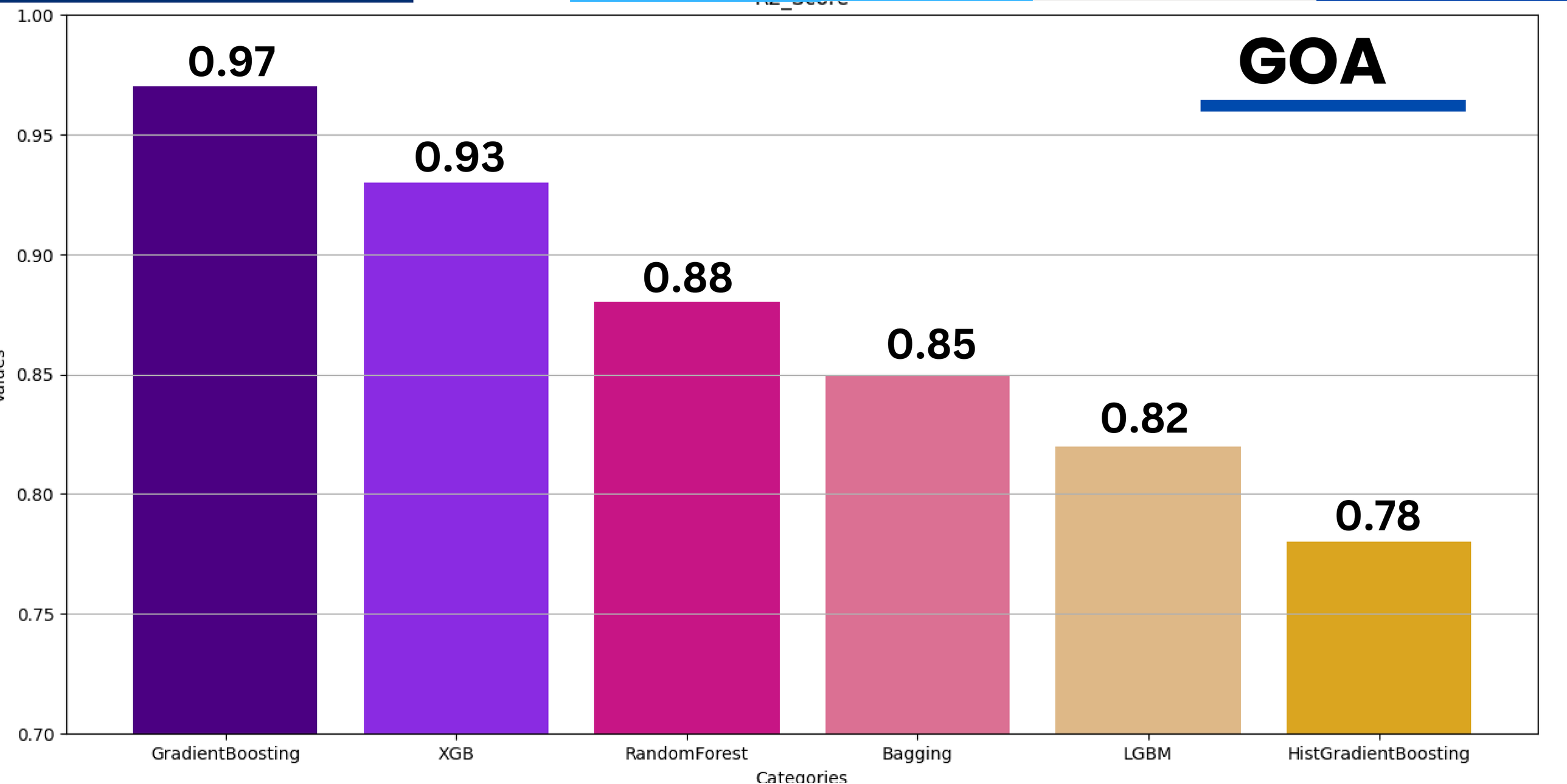
- **XGBoost** R2 Score **0.93**

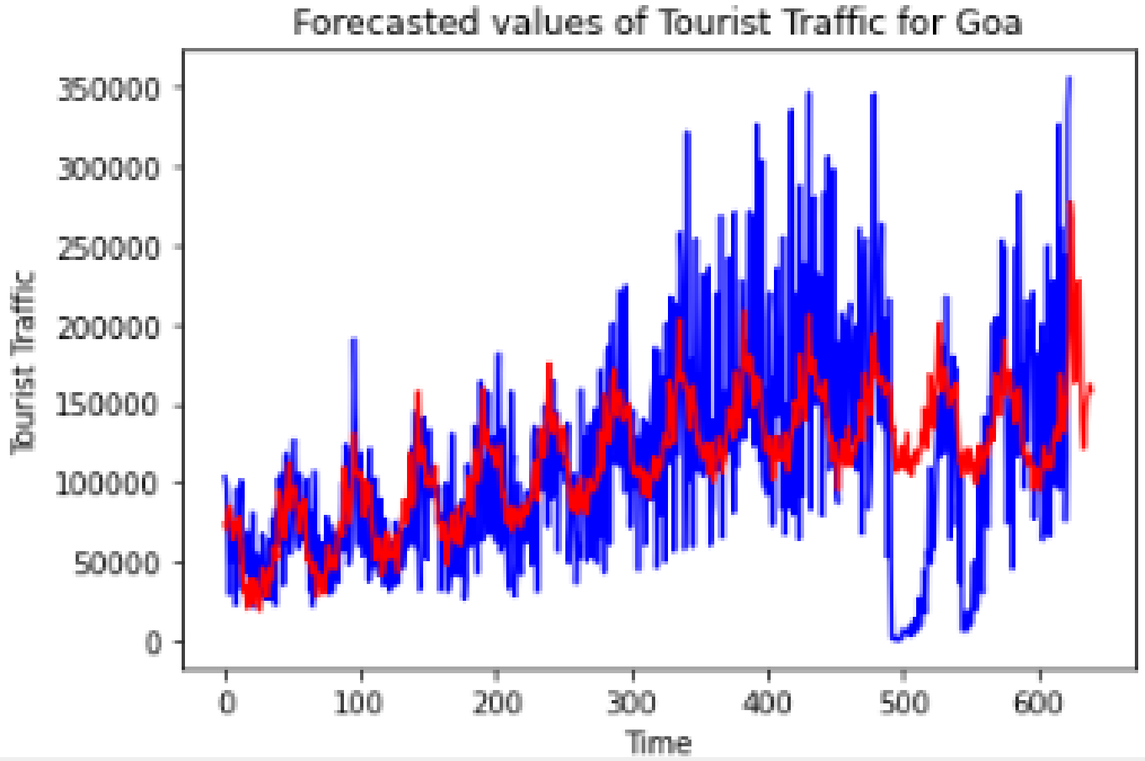
XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

Why XG Boost?

It can capture the nuances of the relationships between features, especially when you have observed seasonality and trends over time. Its ability to handle complex, non-linear patterns in data and missing values.

FORECASTING MODELS





GOA PROPHET

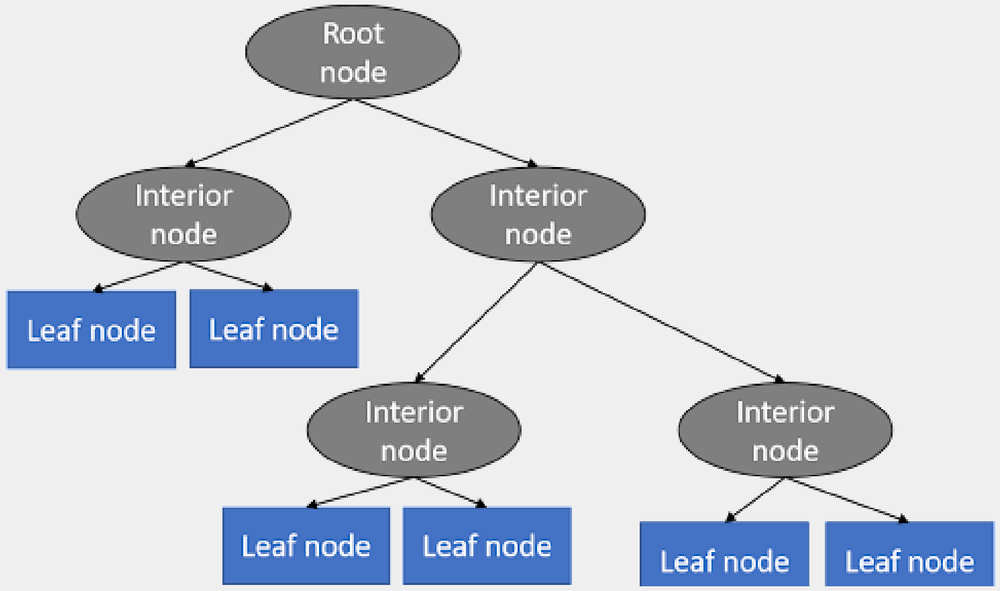
Prophet is a time series forecasting model developed by Meta that offers several compelling reasons for its use in various applications. It excels in handling data with strong seasonal patterns, holiday effects, and missing values.

Why Prophet?

Prophet is designed to work well with data that has strong seasonal patterns and multiple sources of uncertainty, which is common in our case. It includes the ability to account for holidays and special events, which is crucial when forecasting tourist

Kerala

Decision Tree Regressor



Best model

- **Decision Tree Regressor** R2 Score: **0.93**

A Decision Tree is a fundamental machine learning algorithm that builds a tree-like structure to make decisions. It partitions data into subsets at each node by selecting the feature that minimizes a splitting criterion. The tree continues to split, recursively refining decision boundaries, until a predefined stopping criterion is met.

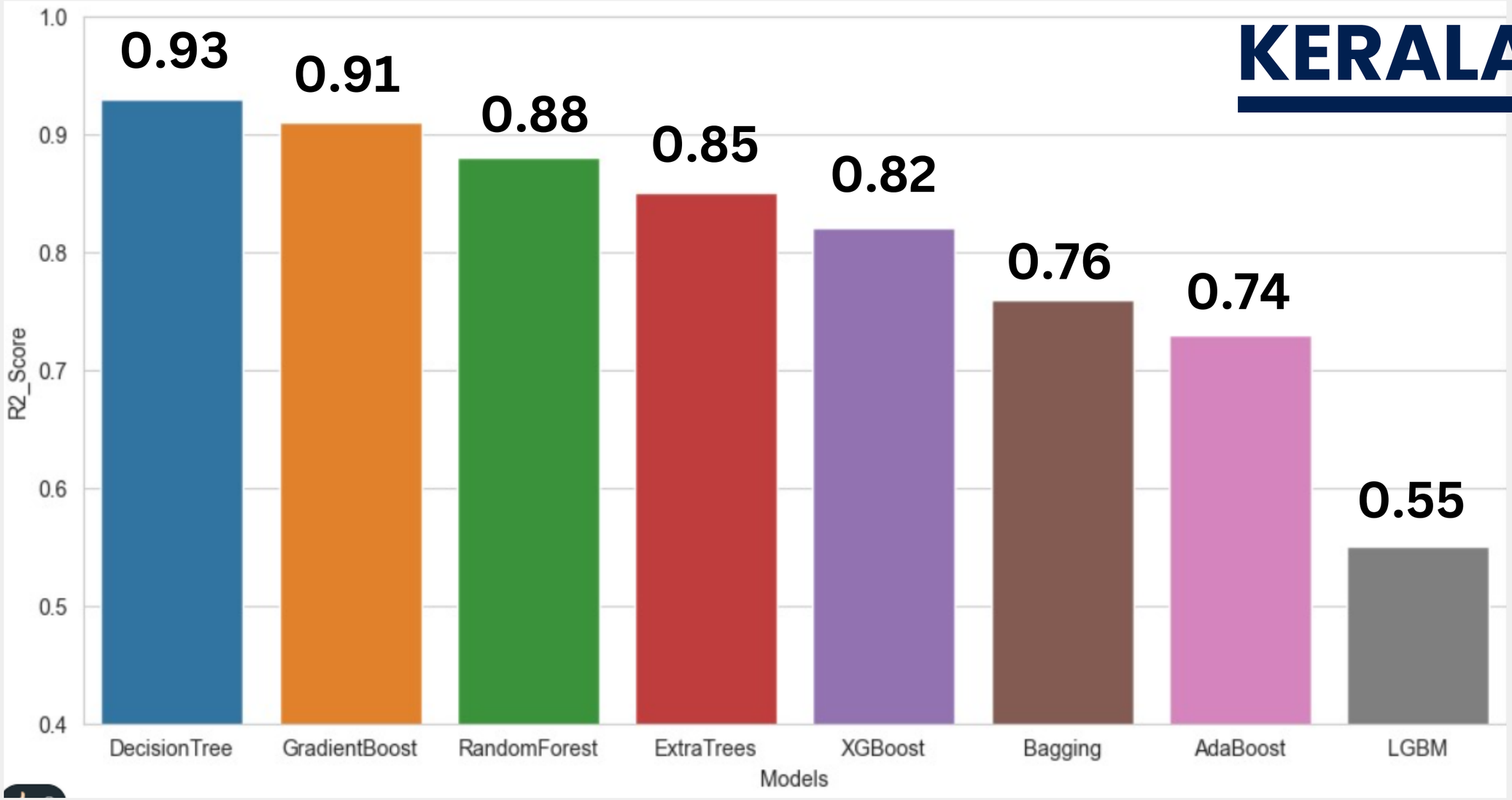
Why Decision Tree Regressor?

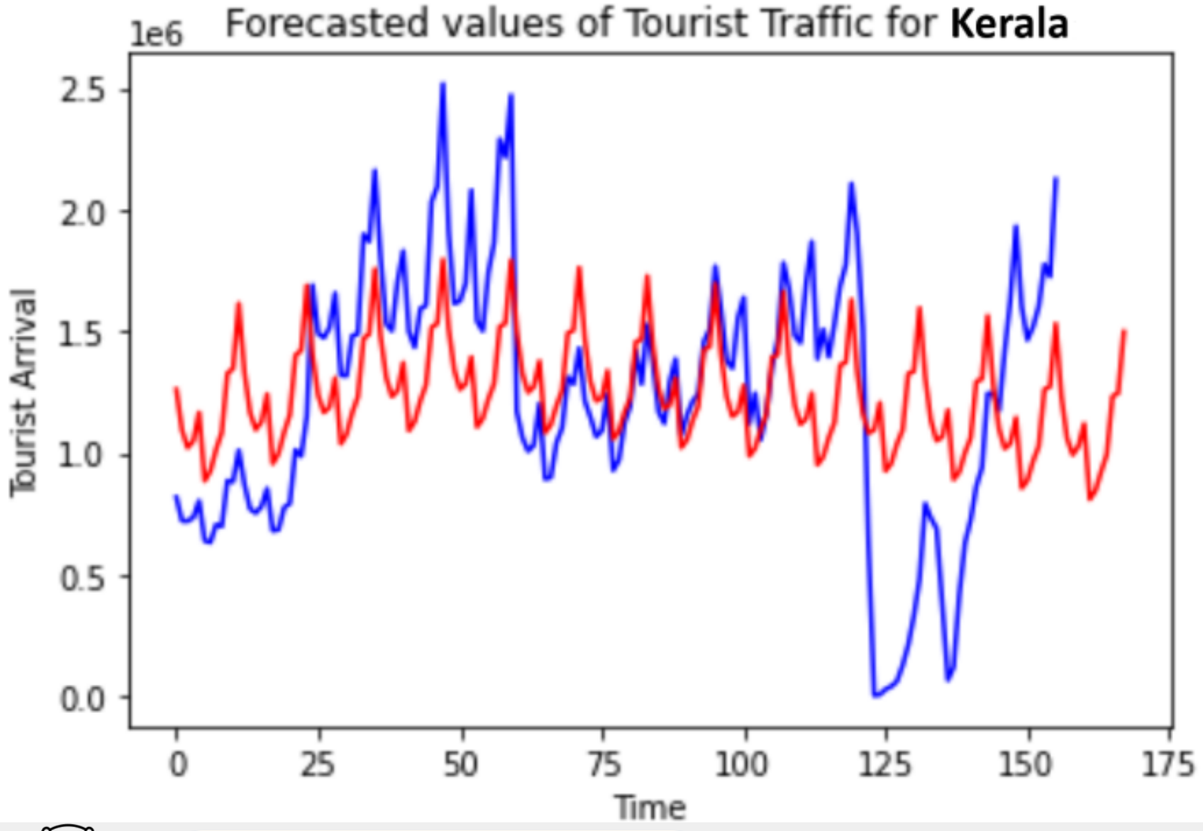
Decision Tree Regressor gives the best results for our dataset because it can handle both categorical and numerical data, which are present in our dataset. It can also handle non-linear relationships between the input features and the target variable, which may be present in our dataset.

FORECASTING MODELS

DATA EXTRACTION | EDA | PRE-PROCESSING | FORECASTING MODEL

KERALA



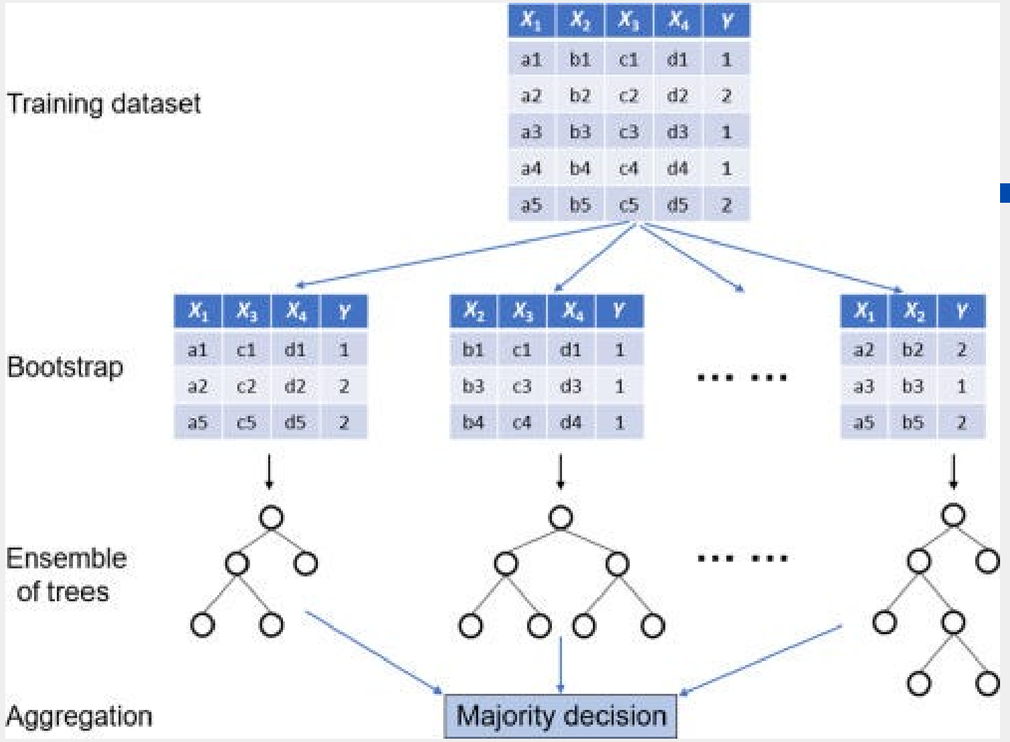


KERALA PROPHET

Prophet is a time series forecasting model developed by Meta that offers several compelling reasons for its use in various applications. It excels in handling data with strong seasonal patterns, holiday effects, and missing values.

Why Prophet?

Prophet is designed to work well with data that has strong seasonal patterns and multiple sources of uncertainty, which is common in our case. It includes the ability to account for holidays and special events, which is crucial when forecasting tourist



X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

Best model

- **Extra Trees Regressor** R2 Score **0.95**

Shimla

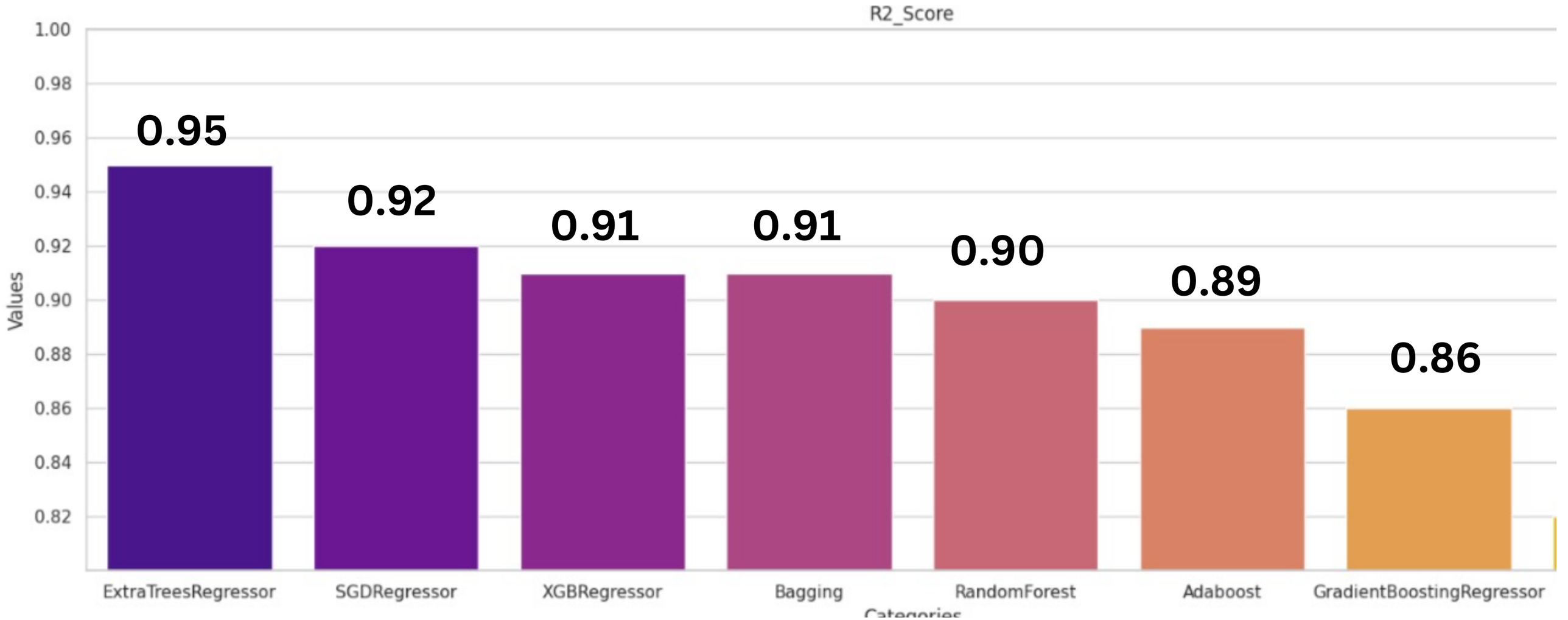
Extra Trees Regressor

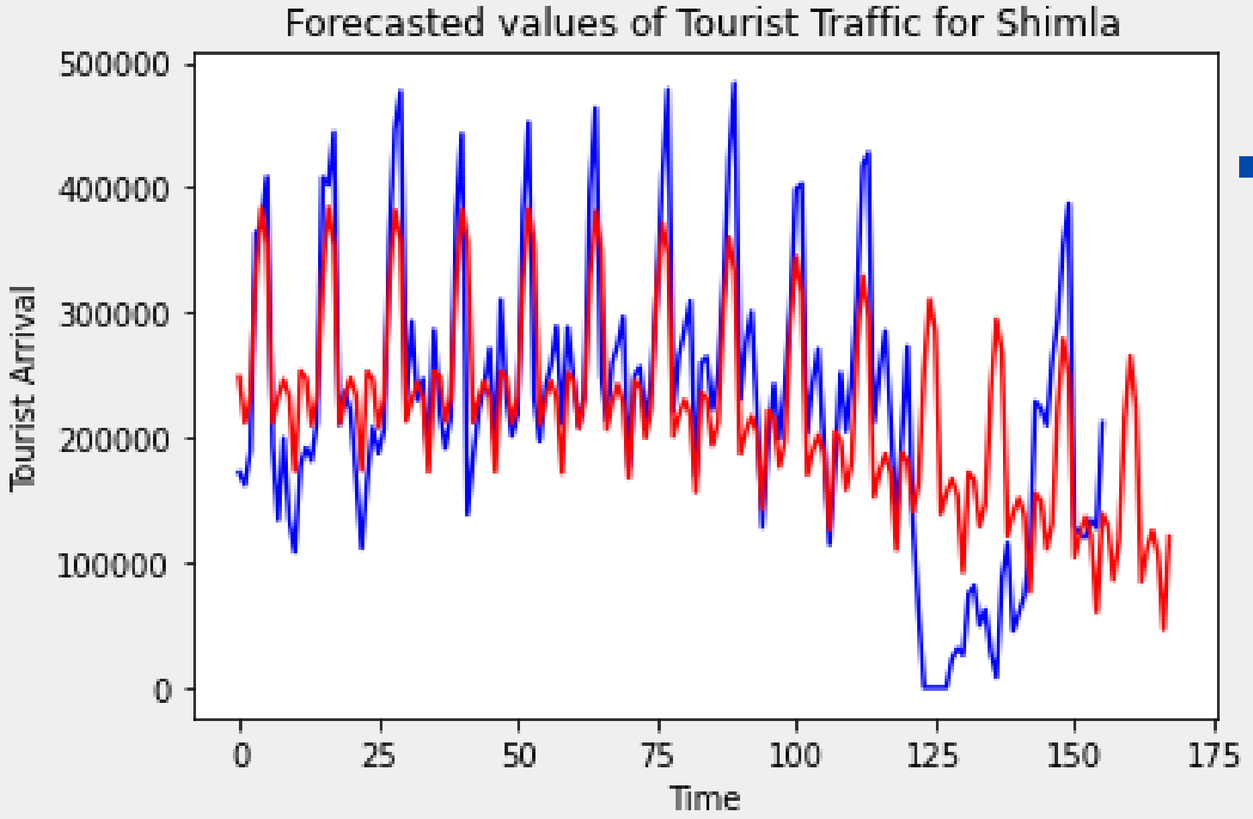
Extra trees regressor is a type of ensemble learning technique that uses randomized decision trees to improve predictive accuracy and control over-fitting. The algorithm works by creating a large number of fitted trees from the training dataset and then averaging the predictions of these decision trees. It is faster than Random Forest since it chooses the splitting node randomly and not the optimal one.

Why Extra Trees Regressor?

Extra Trees Regressor works with a lot of randomization at each step. This makes it robust to noise and outliers in the data. Also extra trees has the ability to capture non-linearity in data due to its ability to create deep and diverse decision trees and ensemble them to get the final prediction.

SHIMLA


R2_Score
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
Values
0.95
0.92
0.91
0.91
0.90
0.89
0.86
ExtraTreesRegressor
SGDRegressor
XGBRegressor
Bagging
RandomForest
Adaboost
GradientBoostingRegressor
Categories



SHIMLA PROPHET

Prophet is a time series forecasting model developed by Meta that offers several compelling reasons for its use in various applications. It excels in handling data with strong seasonal patterns, holiday effects, and missing values.

Why Prophet?

Prophet is designed to work well with data that has strong seasonal patterns and multiple sources of uncertainty, which is common in our case. It includes the ability to account for holidays and special events, which is crucial when forecasting tourist

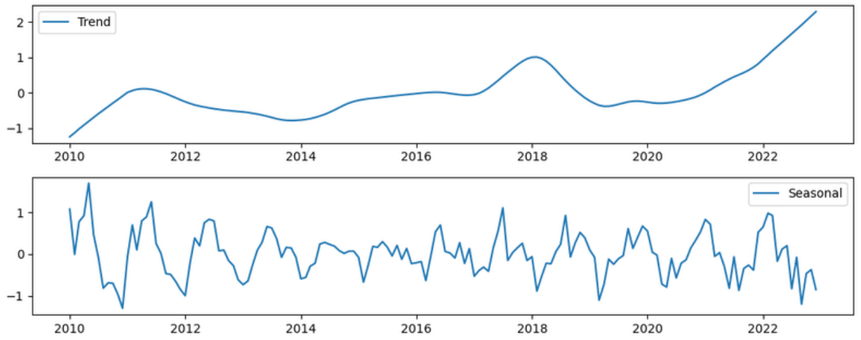
RESULTS

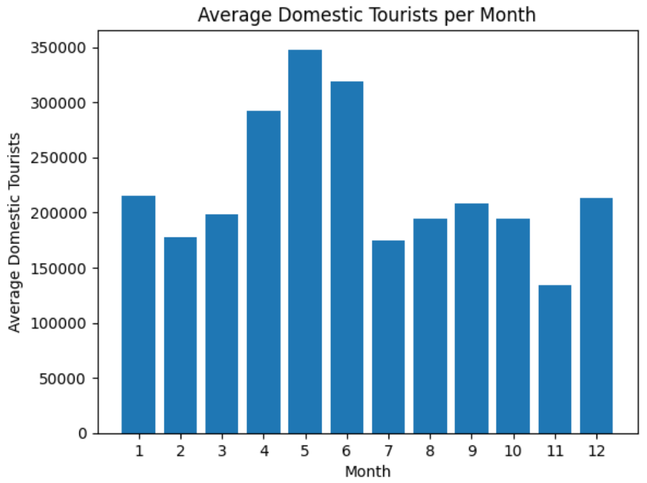
REGION	Best Performing Models	R2 Score
GOA	Gradient Boosting	0.97
SHIMLA	Extra Trees Regressor	0.95
KERALA	Decision Tree Regressor	0.93

THANK YOU!

ANNEXURE

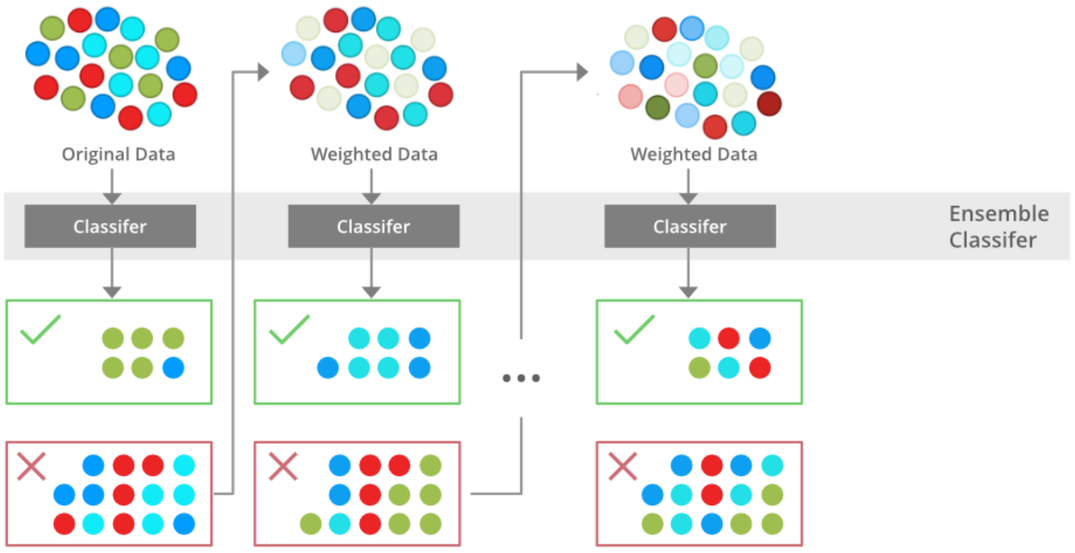
EXPLORATORY DATA ANALYSIS





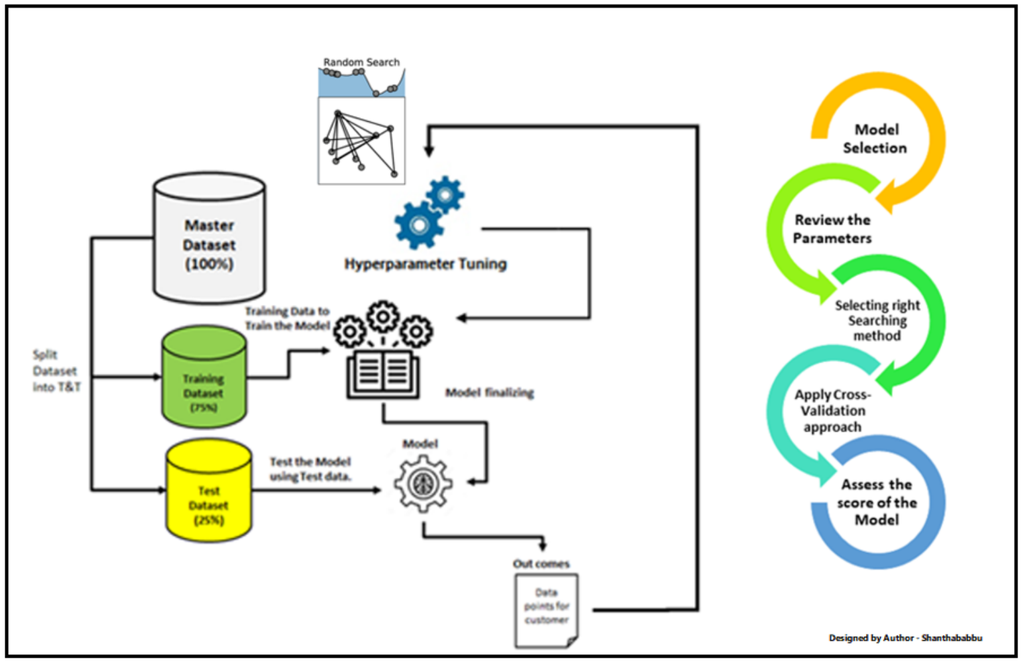
On applying EDA to various dataset, we have observed that there is a seasonal variation in the data for the tourist destinations. For example in Shimla, it records its highest visitors in the months of April to June.

BOOSTING



Boosting is an ensemble machine learning technique that combines the predictions of multiple weak learners (usually simple models) to create a strong learner. It works by sequentially training a series of models, with each subsequent model focusing on the mistakes made by the previous ones.

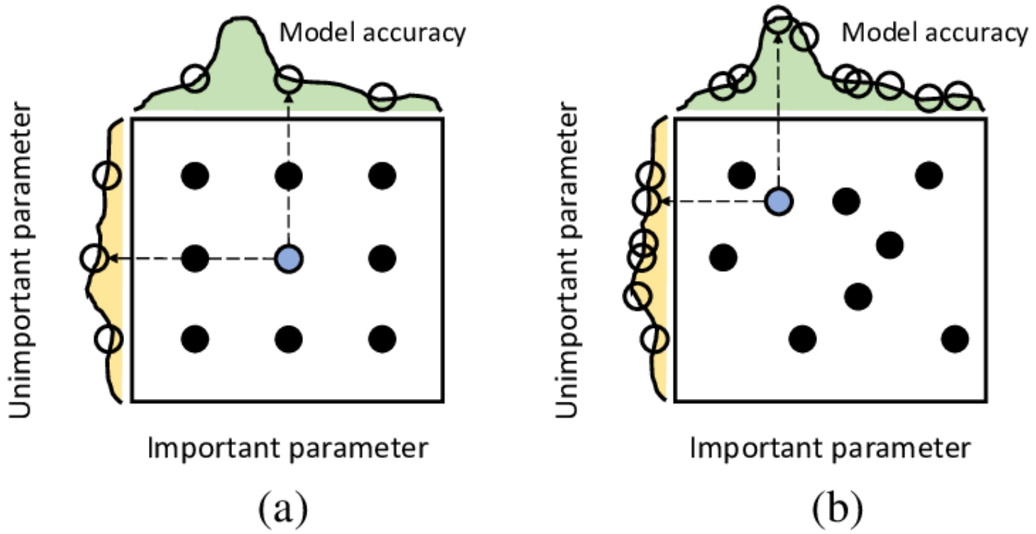
Hyperparameter Tuning



Hyperparameter tuning is the process of finding the best set of hyperparameters for a machine learning model in order to optimize its performance. Hyperparameters are configuration settings that are not learned from the data during training but are set prior to training. They control aspects like the complexity of the model, the learning rate, regularization strength, and more.

GridSearch

GridSearchCV is a technique to search through the best parameter values from the given set of the grid of parameters. GridSearchCV is a useful tool to fine tune the parameters of your model.



MODEL	PARAMETERS
Bidirectional LSTM	layers.Bidirectional(layers.LSTM(15,dropout=0.2))(x) # LSTM layer; layers.Dropout(0.2, name='dropout')(x); layers.Dense(64, activation='relu', name='dense')(x)
CatBoost	'learning_rate': [0.1, 0.01, 0.02,0.04], 'depth': list(range(4,11))
Support Vector Machines	'C': [0.1, 1, 10, 100, 1000], 'gamma': [1, 0.1, 0.01, 0.001, 0.0001], 'kernel': ['rbf', 'linear', 'poly', 'sigmoid']