Open IIT Data Analytics
IIT KHARAGPUR
2023-24

TEAM-D21

Index

1. INTRODUCTION

Tourism is a major economic sector in many countries, and forecasting tourist arrivals is essential for effective planning and management. In an age marked by rapid technological advancement, big data, and the internet, the ability to make data-driven decisions has become a powerful tool for optimising the tourism sector. Machine learning can be used to develop predictive models that can forecast tourist arrivals with high accuracy.

This report describes the **development** and **evaluation** of a machine learning model for **predicting tourist arrivals** to a specific destination in India using Internet search index data as a key input.

2. OBJECTIVES

To build a predictive model to forecast tourist arrivals based on **historical tourist arrival** data and **Internet search index** data:

1. Collect and prepare data.
2. Choose and train a model
3. Evaluate and deploy the model.

3. DATASET

The dataset is meticulously collected to ensure its accuracy and relevance. The dataset comprises historical data on tourist arrivals. The dataset contains data for 12 months, from **Jan 2010** to **Dec 2022.**

Our scraped dataset includes **156 training examples** including **14 features.**

The dataset you provided contains the following information:

- Number of tourists
- Year
- Month
- Number of flight bookings
- Number of hotel bookings
- Tourism in Kerala
- Tourism Packages
- Tourist Places
- Food & Drink
- Flights_Kerala: Number of flights toNo of Tourists

- Trains_Kerala: Number of trains to Kerala
- Rain: Rainfall in Kerala
- T_Max: Maximum temperature in Kerala
- T_Min: Minimum temperature in Kerala

The key components of the dataset are:

Tourist Arrivals(Target variable): This is the core data on the number of tourists arriving at the destination. The dataset covers a substantial time period, from 2010 to 2022, allowing for the analysis of long-term trends.

Internet Search Indexes: Internet search data, obtained from Google Trends, plays a crucial role in this analysis. It provides insights into the popularity and interest in the destination over time.

Hotel Booking and Flight Search Data: These datasets provide information on the demand for accommodations and flights to the destination. It is a supplementary data source that can further enhance the accuracy of predictions, as it reflects the intent of tourists to visit.

4. METHODOLOGY

Forecasting the number of tourists in future time periods involves time series forecasting, as it is a sequence of data points measured at successive points in time. You can implement a model to forecast the number of tourists using various time series forecasting methods. Here's a high-level methodology to guide you through the process:

- **Data Collection and Preparation**:
  Collected historical data: Gathered the historical data on the number of tourists, including the time period and any relevant factors or features that might affect tourism, such as weather conditions, holidays, or special events.
  Cleaned the data by handling missing values, outliers, and inconsistencies.

- **Data Exploration and Visualization:**
  Visualised the data: Created time series plots and other relevant visualisations to understand the historical trends, patterns, and any seasonality in the data.
  Identified trends and seasonality: Identified long-term trends and seasonal patterns in the data that helped in model selection.

- **Time Series Decomposition**:
  Decomposed the time series into its components, which typically include trend, seasonality, and residual (error). This decomposition was done using methods like additive or multiplicative decomposition.

- **Feature Engineering :**
  Created additional features: We identified factors that affect tourism, created additional features to incorporate these factors into our model. For example, weather data, holidays, or special events can be important predictors.

- **Model Selection**:
  Choose a forecasting model: Selected an appropriate time series forecasting model based on the characteristics of our data. Common models include:
  ARIMA (AutoRegressive Integrated Moving Average):
  Suitable for stationary time series data.
  Exponential Smoothing methods: ETS (Error, Trend, Seasonal) models.
  Machine Learning models: XGBoost & LSTM (Long Short-Term Memory)

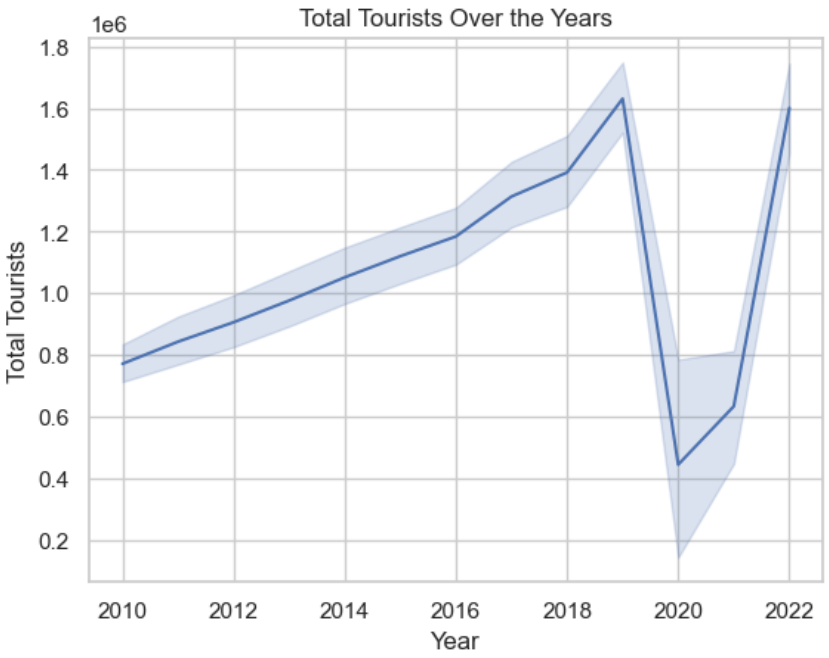
- **Model Training :**
  We trained the model with the data set for the period of **2010** to **2019** and later used **2022** as the testing data.
  Split the data: Divide the historical data into training and testing datasets. The testing dataset should be used to evaluate the model's performance.
  Trained the model: Fitted the chosen forecasting model to the training data, considering the identified seasonality and any additional features.

- **Model Evaluation:**
  Evaluated the model's performance using appropriate metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or others, depending on the nature of the data.
  Checked for overfitting: Ensured that the model is not overfitting the training data by comparing its performance on the testing dataset.

- **Model Forecasting**:
  Used the trained model to forecast the number of tourists for future time periods.

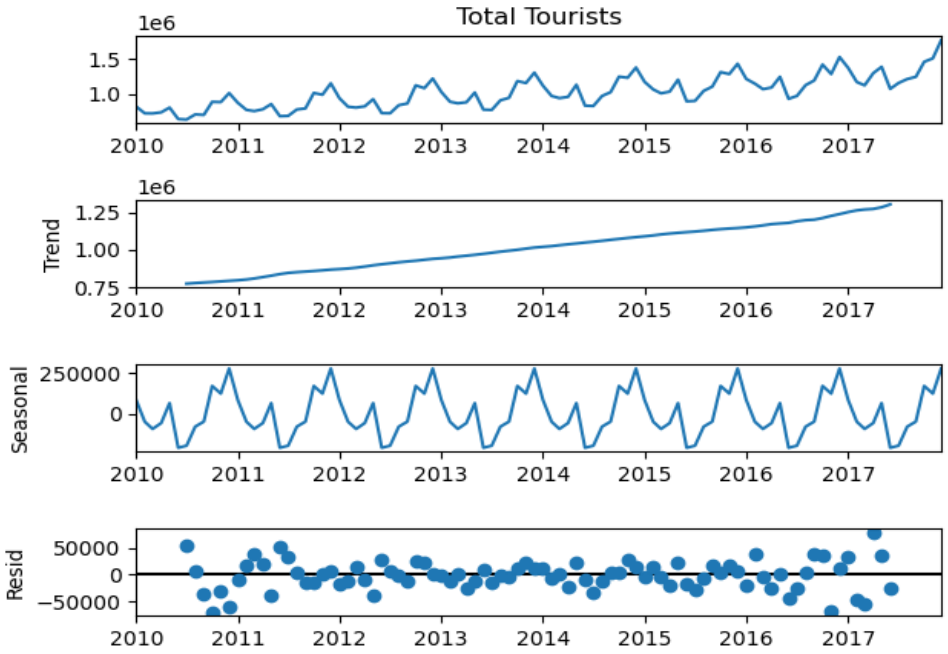
- **Visualisation and Interpretation**:
  Visualised the forecasted results along with historical data and any associated factors to gain insights and interpret the forecasts.

5. DATA VISUALISATION

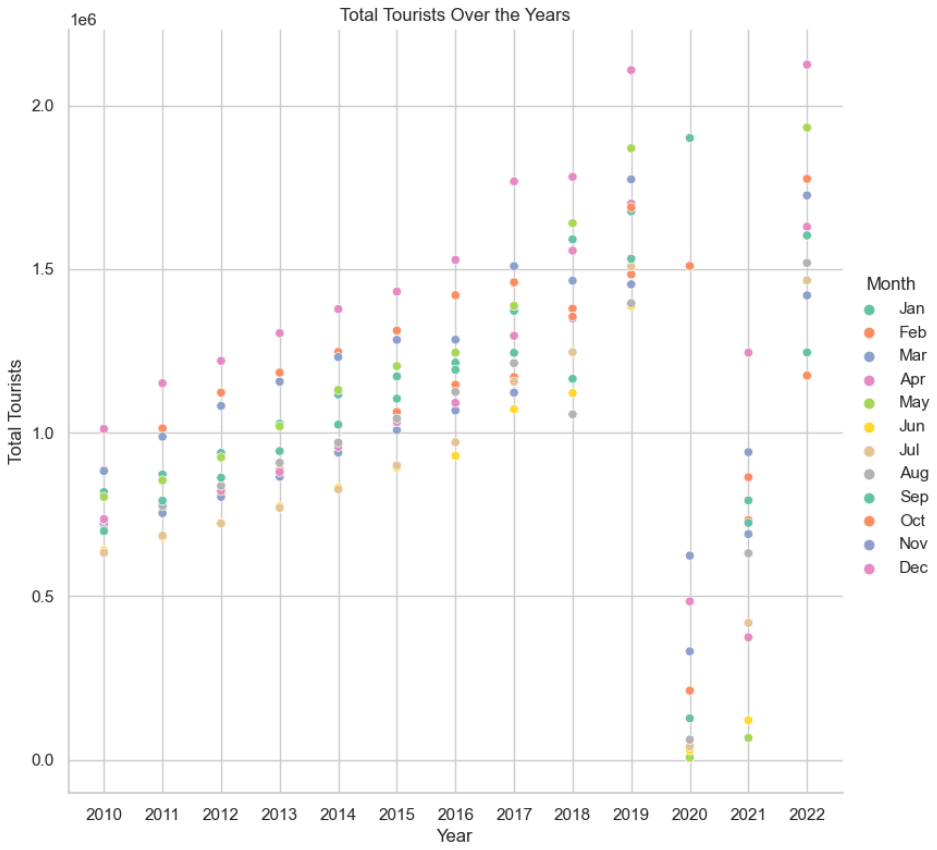


- **Tourist numbers have seen a steady increase, rising from around 1 million in 2010 to 1.8 million in 2023.**

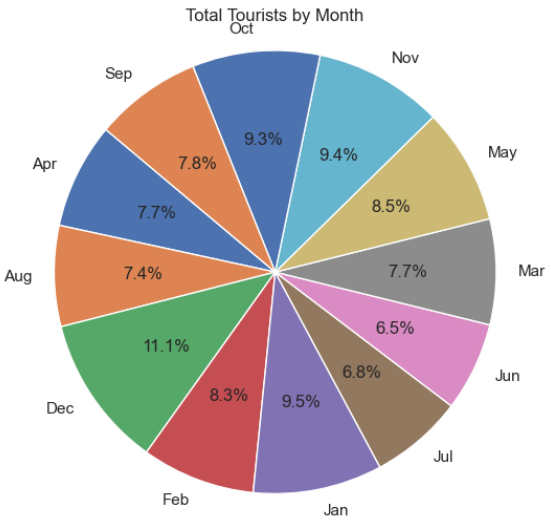
- **A significant dip occurred in 2020-21 due to the COVID-19 pandemic.**

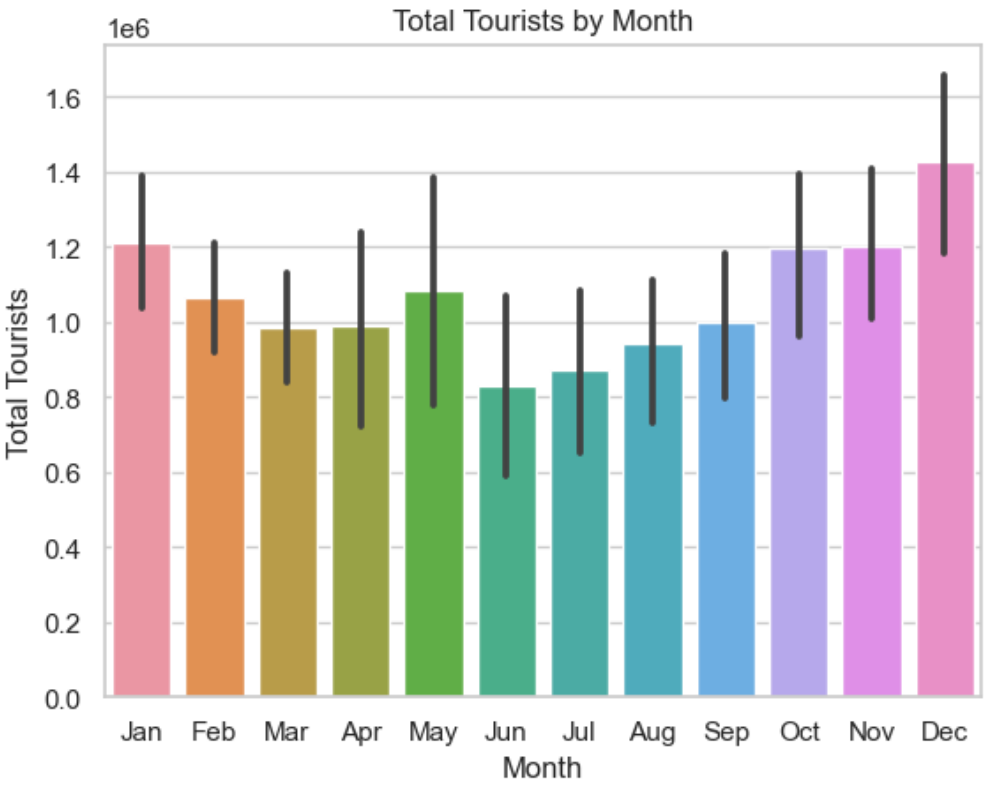


- **The winter months are the peak tourist season in Kerala, while the summer monsoon months are the off-season.**



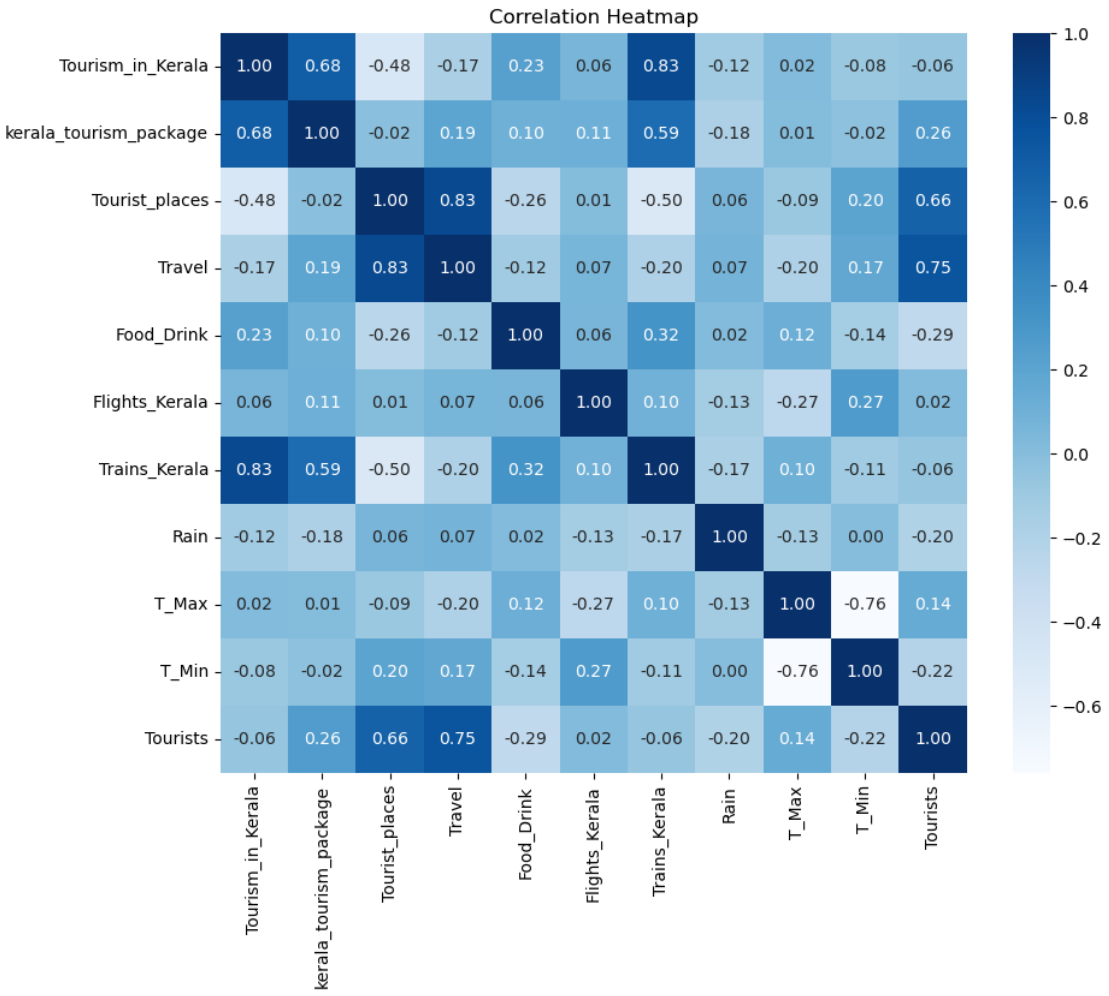
- **The number of tourists has been increasing over the years, despite some fluctuations along the way.**
- **The COVID-19 pandemic caused a significant decline in the number of tourists in 2020.**
- **There is an expectation that the number of tourists will continue to increase in the future.**





- **Lowest tourist numbers: June, July, and August (Summer months).**
- **Highest tourist numbers: December and January (Winter months).**
- **Tourist numbers rise consistently from June to October.**
- **Tourist numbers decline steadily from February to April.**

After data preprocessing we have dropped some features whose correlation with the target variable was less.



- **Number of tourists** is strongly correlated with **Tourist places** (0.66) and **Travel** (0.75).This means that the number of tourists visiting Kerala is positively correlated with the number of tourist places in Kerala and the amount of travel to Kerala.
- **Kerala tourism package** is also strongly correlated with **Number of Tourists** (0.68).
  This means that the number of tourists visiting Kerala is positively correlated with the number of tourism packages sold for Kerala.
- **Trains to Kerala** are also strongly correlated with **Number of Tourists** (0.83). This means that the number of tourists visiting Kerala is positively correlated with the number of train tickets sold to Kerala.
- **T_Max** (maximum temperature) is negatively correlated with the number of tourists (-0.76). This means that the number of tourists visiting Kerala is negatively correlated with the maximum temperature in Kerala.
- **T_Min** (minimum temperature) is also negatively correlated with the number of tourists (-0.22). This means that the number of tourists visiting Kerala is negatively correlated with the minimum temperature in Kerala.

The following observations could be drawn from the above graphs:

Tourism in Kerala is growing rapidly. The number of tourists visiting Kerala has increased by over 50% in the past decade.
There is a clear seasonal pattern to tourism in Kerala, with the peak tourist season being from October to February. This is likely due to the state's pleasant weather during these months.
The growth of tourism in Kerala has had a positive impact on the state's economy. Tourism is now one of the largest contributors to Kerala's GDP and employs a large number of people.

MODELS

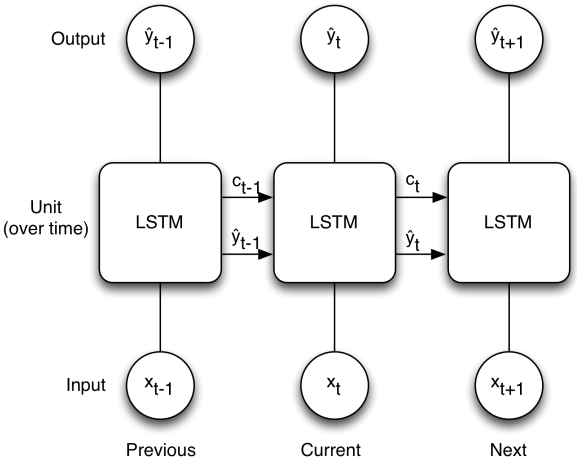
The training data was scraped from google trends and other tourism sites like Kerala Tourism from 2010 to 2019. And we are going to forecast the number of tourists for the year 2022.

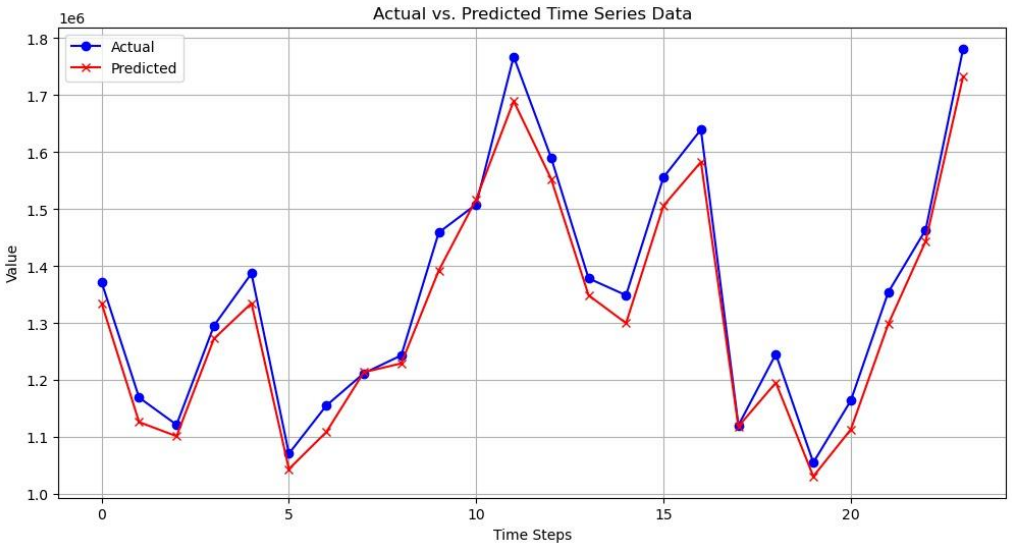
The following basic models were applied (We trained the model with the data set for the period of 2010 to 2019 and later used 2022 as the testing data):

- LSTM (Long Short-Term Memory Network)
- ARIMA
- Temporal fusion Transformer

	MAE	**RMSE**	**MAPE**
LSTM	37155.966665	42058.99635271	2.70500311
ARIMA	106445.6375	135297.6510	7.8947
Temporal fusion Transformer	33287.3280	3854.4397	2.5634

LONG SHORT TERM MEMORY (LSTM)





- LSTM models worked best on the dataset when MAPE was taken as the metric for evaluation. **Overall value of MAPE = 2.705**

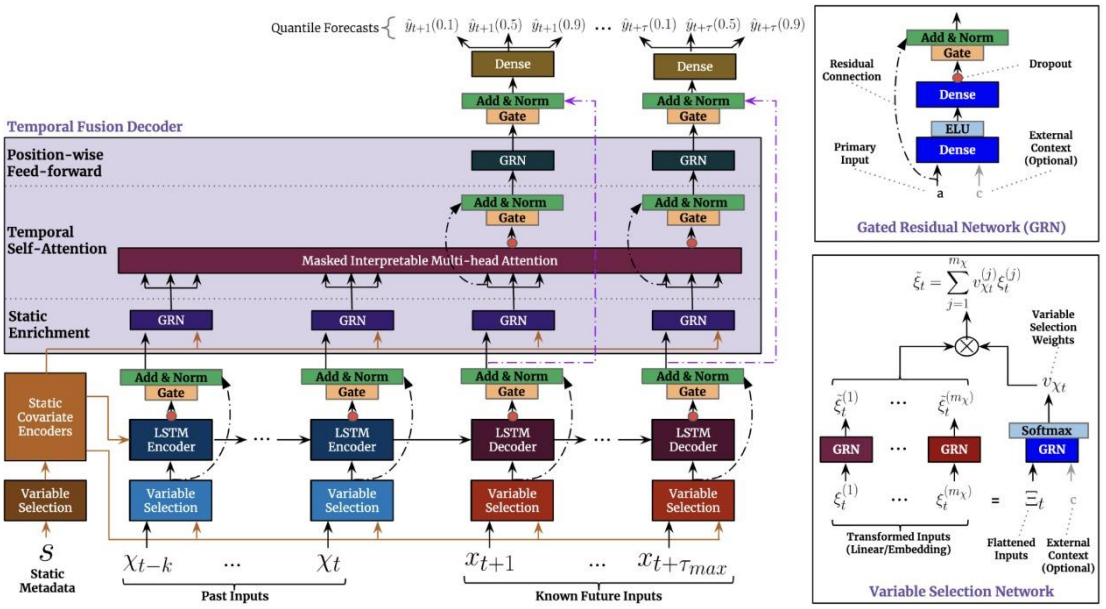
WORKFLOW OF LSTM :

We have implemented two LSTM layers, each consisting of 50 neurons.

```
model = Sequential()
model.add(LSTM(50, activation='relu', return_sequences=True, input_shape=(sequence_length, X.shape[2])))
model.add(LSTM(50, activation='relu'))
model.add(Dense(1))  # Output size is 1 for the "Total Tourists" prediction
```

TEMPORAL FUSION TRANSFORMER

MODEL ARCHITECTURE



TFT is based on the Transformer architecture, which is a type of neural network that is well-suited for modelling long-range dependencies in sequential data. TFT uses a combination of recurrent layers and self-attention layers to capture the temporal dynamics of time series data. The recurrent layers learn local temporal dependencies, while the self-attention layers learn long-term global dependencies.

TFT also uses a special type of layer called the Temporal Fusion Layer to combine the outputs of the recurrent layers and the self-attention layers.

CONCLUSION

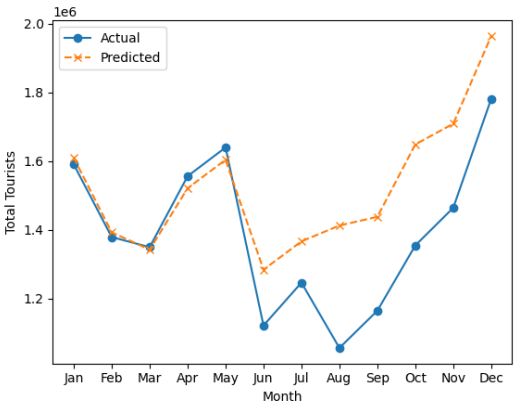
The **LSTM model** achieved the lowest MAPE, RMSE, and MAE scores indicates that it is able to predict tourist arrivals with the highest degree of accuracy.
This is likely due to the fact that LSTM models are well-suited for modelling long-range dependencies in sequential data, such as tourist arrival data.

Here are some of the reasons why LSTM models are effective for time series forecasting:

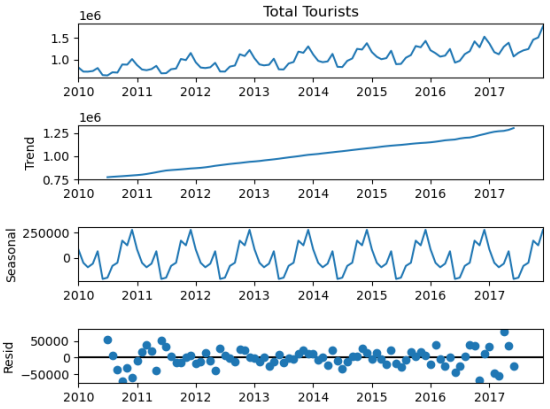
- They can learn long-term dependencies in sequential data.
- They are able to handle non-linear relationships between the features and the target variable.
- They are relatively robust to noise in the data.

Overall, the LSTM model is a powerful tool for time series forecasting. It is able to achieve state-of-the-art results on a variety of tasks, including forecasting tourist arrivals.

ANNEXURE



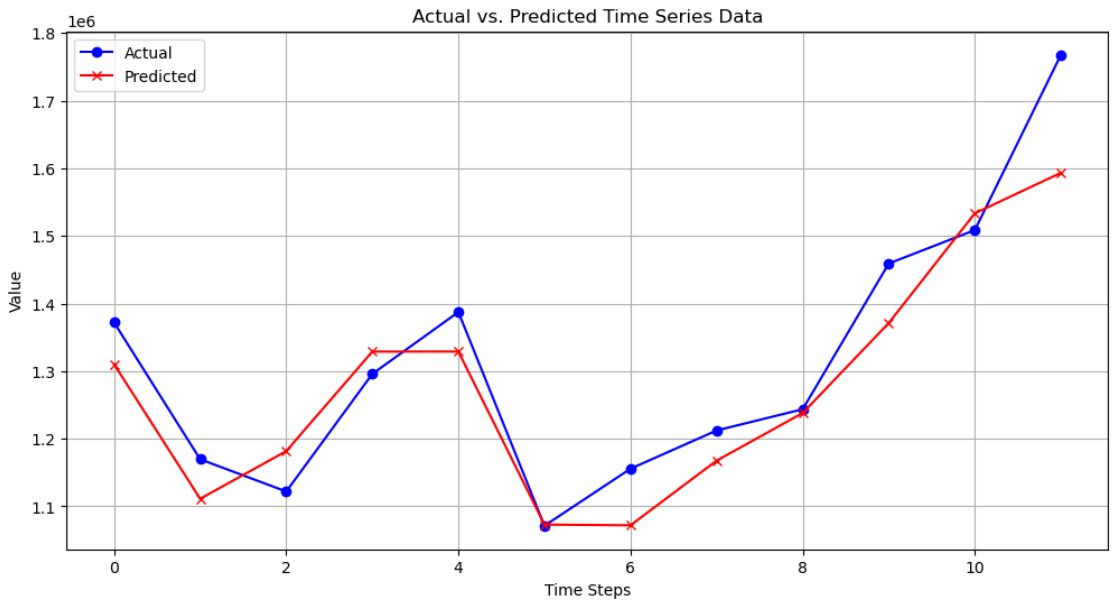
ARIMA: Actual vs Predicted



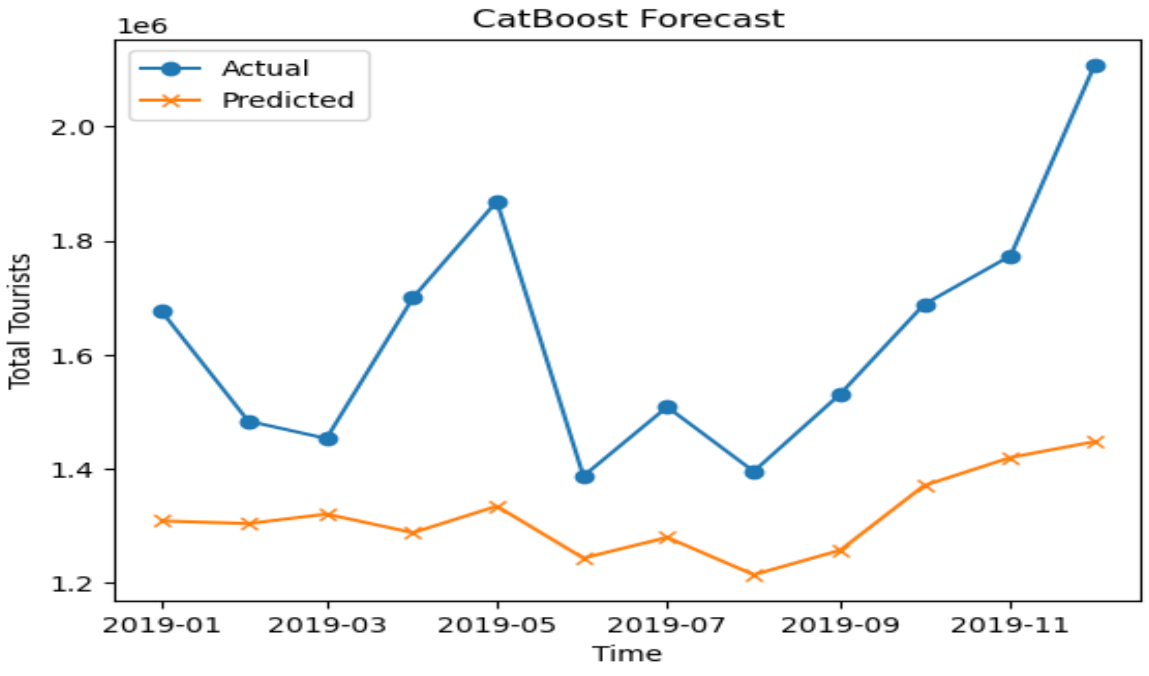
ARIMA: Seasonality and Trend

+

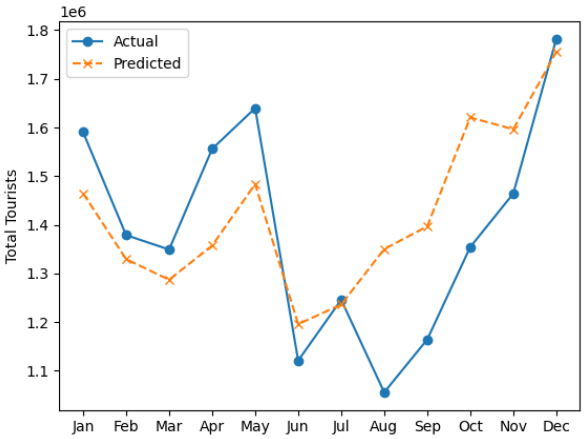
ARIMA: Training and Prediction



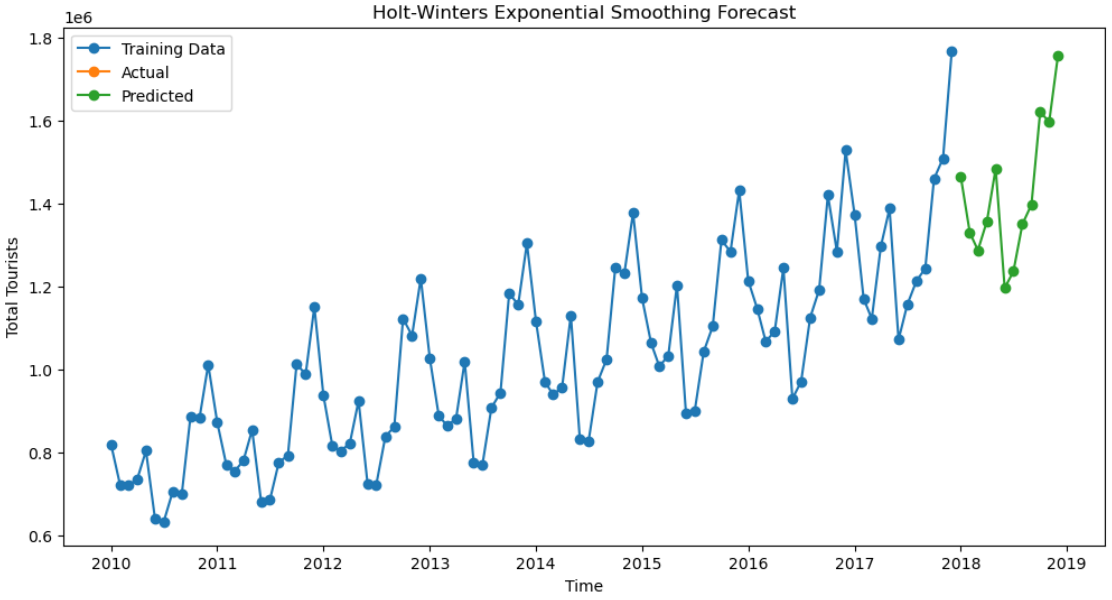
LSTM: Actual vs Predicted



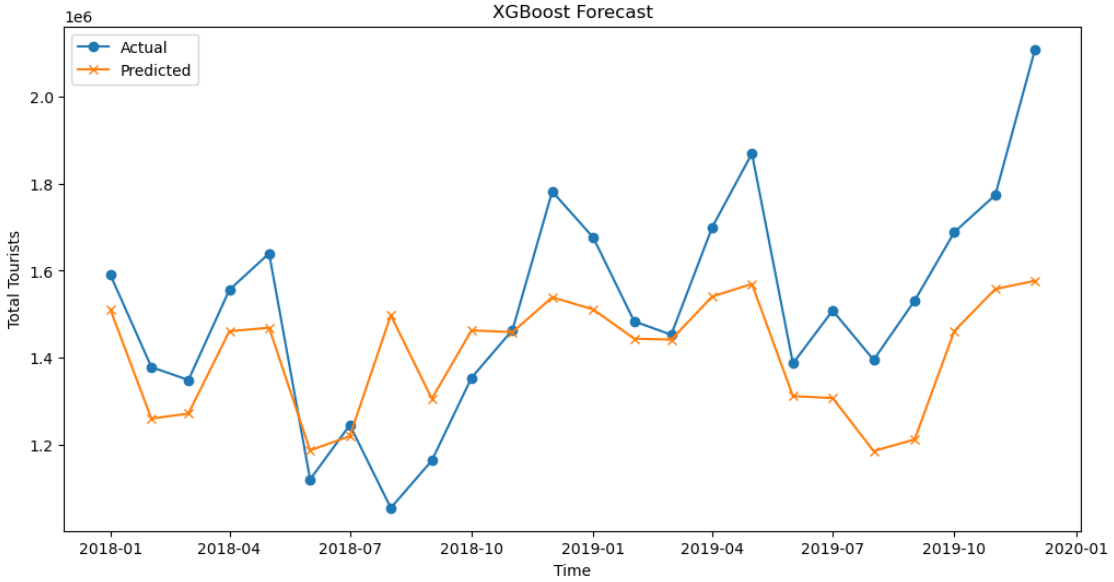
CatBoost: Actual vs Predicted



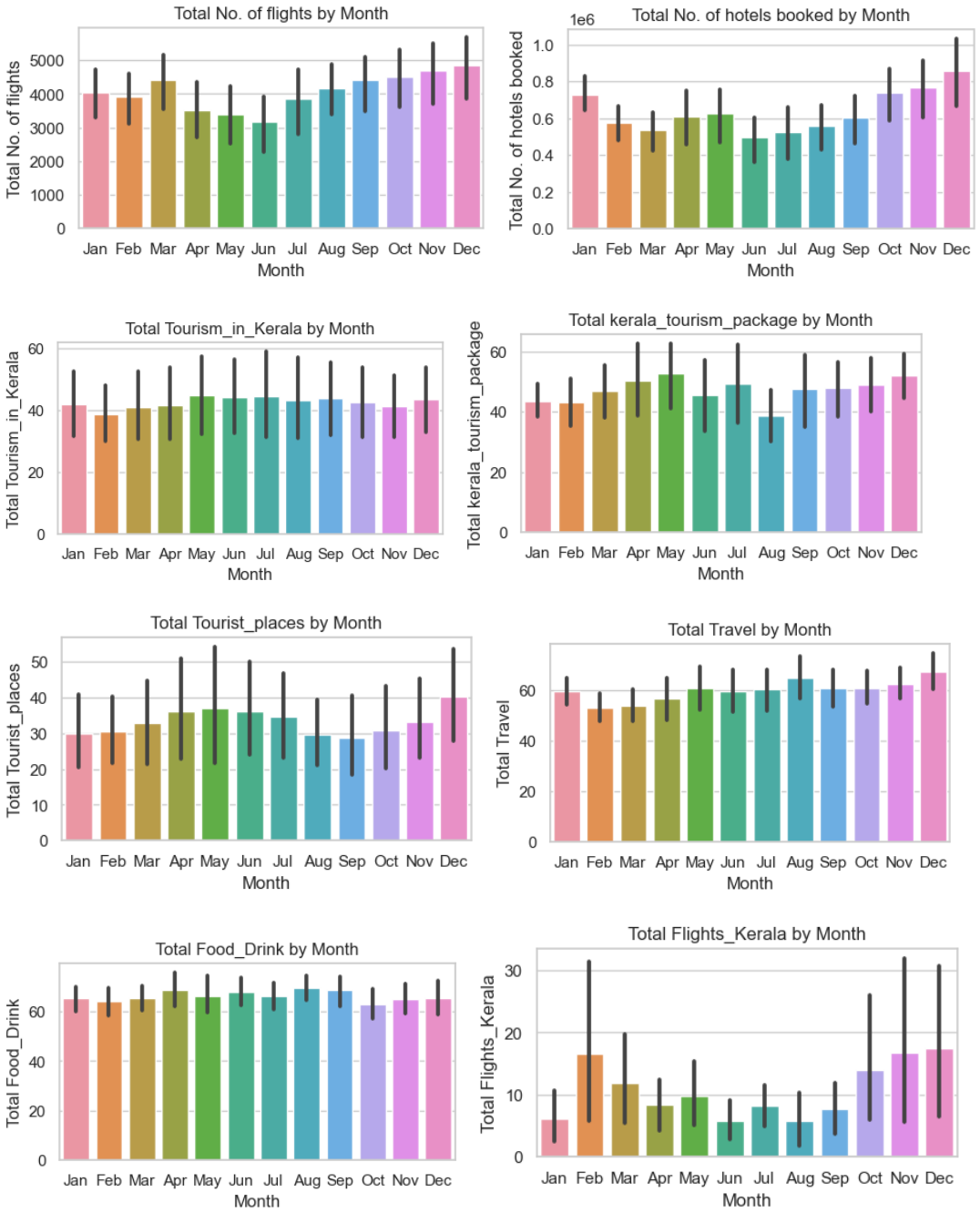
Holt-Winters Exponential Smoothing Actual Forecast: Actual vs Predicted

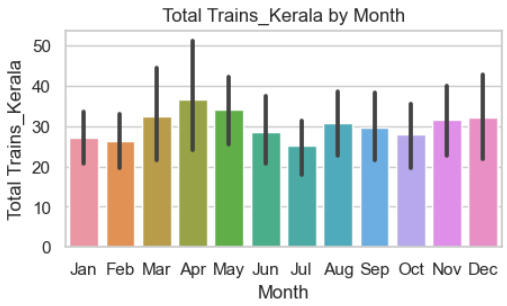


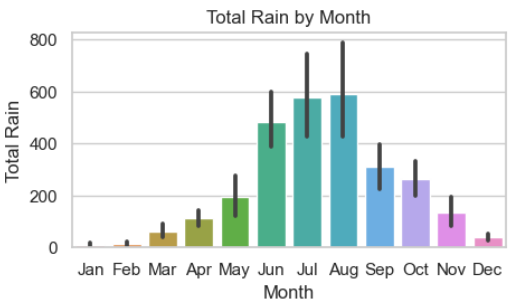
Holt-Winters Exponential Smoothing Actual Forecast: Training Data

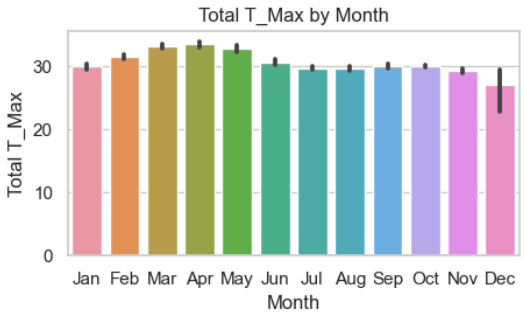


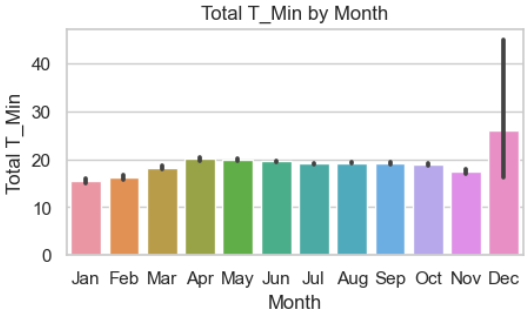
XGBoost: Actual vs Predicted

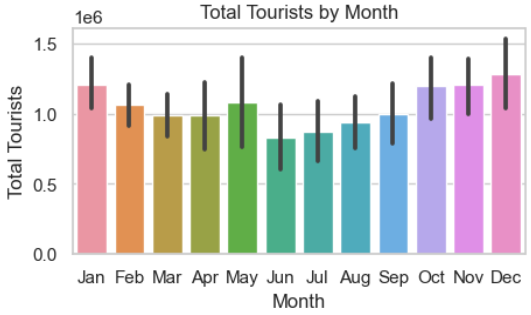


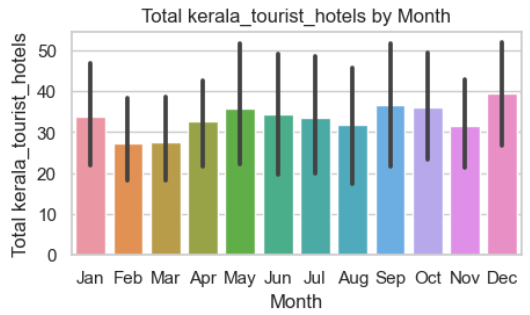

Total Trains_Kerala by Month
Total Trains_Kerala
0
10
20
30
40
50
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


Total Rain by Month
Total Rain
0
200
400
600
800
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


Total T_Max by Month
Total T_Max
0
10
20
30
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


Total T_Min by Month
Total T_Min
0
10
20
30
40
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


Total Tourists by Month
1e6
Total Tourists
0.0
0.5
1.0
1.5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


Total kerala_tourist_hotels by Month
Total kerala_tourist_hotels
0
10
20
30
40
50
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month