Open IIT Data Analytics

Team Data Warriors

Problem Statement

To Implement a Machine Learning Model that predicts tourist arrivals to a specific destination using Internet Search Index Data as a key input

Index:

Introduction:

- The objective was to analyse and score a dataset that encapsulates the dynamics of tourism in the picturesque city of Shimla, as well as select destinations in the mesmerising state of Kerala, spanning the years from 2010 to 2022.

- Tourism plays a pivotal role in the socio-economic development of India, contributing significantly to the nation's GDP and employment. Understanding the trends and patterns in tourist arrivals is essential for policy formulation and resource allocation and for ensuring these tourism-centric regions' growth and sustainability. Therefore, the importance of a robust dataset that accurately reflects these trends cannot be overstated.

- To curate such a dataset, we embarked on a comprehensive data scraping journey, amalgamating data from various reliable sources. These sources included government websites, Google Trends, and other relevant resources. Our approach involved mining diverse datasets, encompassing information on the number of tourists, seasonal variations, historical trends, and factors that influence tourism in these regions. These insights will provide valuable information for the competition and have broader applications in the field of tourism research and policy development.

- With this introduction, we set the stage for the unveiling of a dataset meticulously crafted to reveal the secrets of Shimla and Kerala's tourism industry, encapsulating the ebb and flow of visitors throughout the years, and we eagerly anticipate the opportunity to present our findings to the judging panel and fellow participants at the Open-IIT Data Science Hackathon.

Dataset:

- The tourist arrivals (Both Domestic and International) in Shimla city have been obtained from the data released by the government of Himachal Pradesh on its website. We collected the tourist arrivals data from January 2010 to December 2022 for each month.

- Apart from that, For predicting tourist arrivals in Shimla city, we have collected various features from Google Trends to help our model in the prediction of Tourist arrivals.

- To transform the original monthly tourist arrival dataset into a more granular and manageable form, we employed a crucial step in data preprocessing—converting the data into weekly intervals.

- The conversion of the original monthly dataset into a weekly format was a pivotal step in data preparation. It allowed us to extract deeper insights, reduce noise, and ensure compatibility with various time series analysis models. This transformed dataset served as the foundation for our analysis, enabling us to derive meaningful conclusions and predictions regarding tourist arrivals in Shimla and Kerala.

Features Used:

The variables/features that have been collected and finally used in our dataset were:-

1. Tourist inflow features: Historic Tourist arrival data on weekly basis from taken from 2010 to 2023:

   1.1 Date - Data from January 2010 to December 2022:

   1.2 Domestic Visitors - Local : Domestic Tourist arrival data from 2010 to 2022.

   1.3 Foreign Visitors : Foreign Tourist arrival data from 2010 to 2022.

2. Transportation Data: Google trends data on search topic popularity for keywords related to transportation:

   2.1 Shimla flights

   2.2 Shimla buses

   2.3 buses - shimla

   2.4 cars - shimla

   2.5 trains - shimla

   2.6 shimla trains

   2.7 shimla cars

   2.8 shimla bus booking

   2.9 Shimla bus timings

   2.10 Shimla train booking

   2.11 Shimla train timings

   2.12 Shimla flight booking

   2.13 Flights in Shimla

   2.14 Flights near Shimla

   2.15 Total transport

3. Weather data: Historic weather data during the same time period and frequency.

3.1 temperature_2m (Â°C)

3.2 windspeed_10m (km/h)

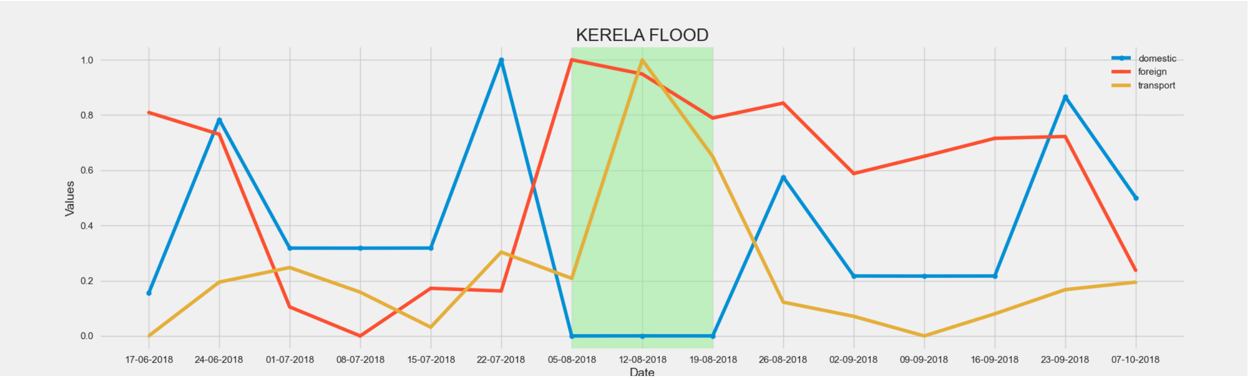
3.3 relativehumidity_2m (%)

3.4 rain (mm)_sum

3.5 snowfall (cm)_sum

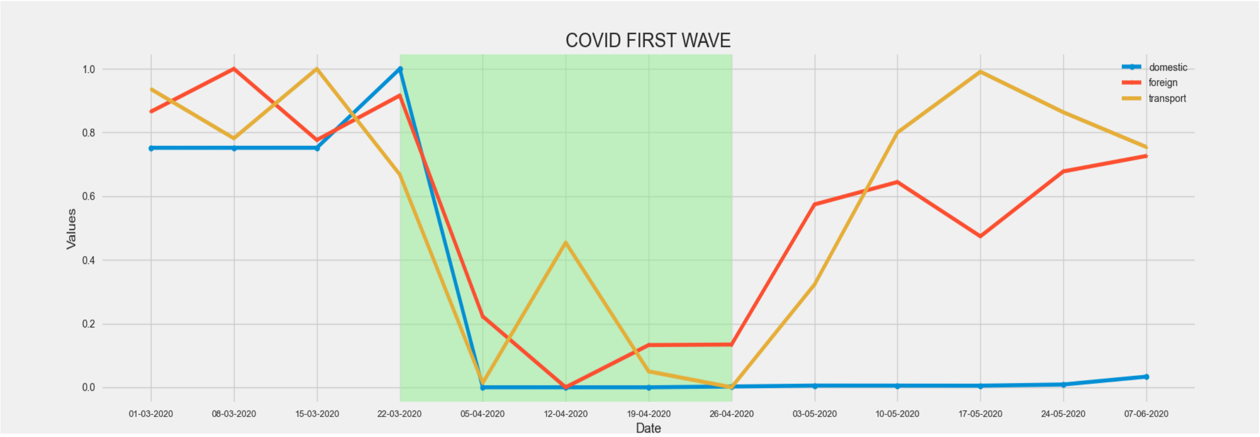
Feature Trends:-

Kerala Flood



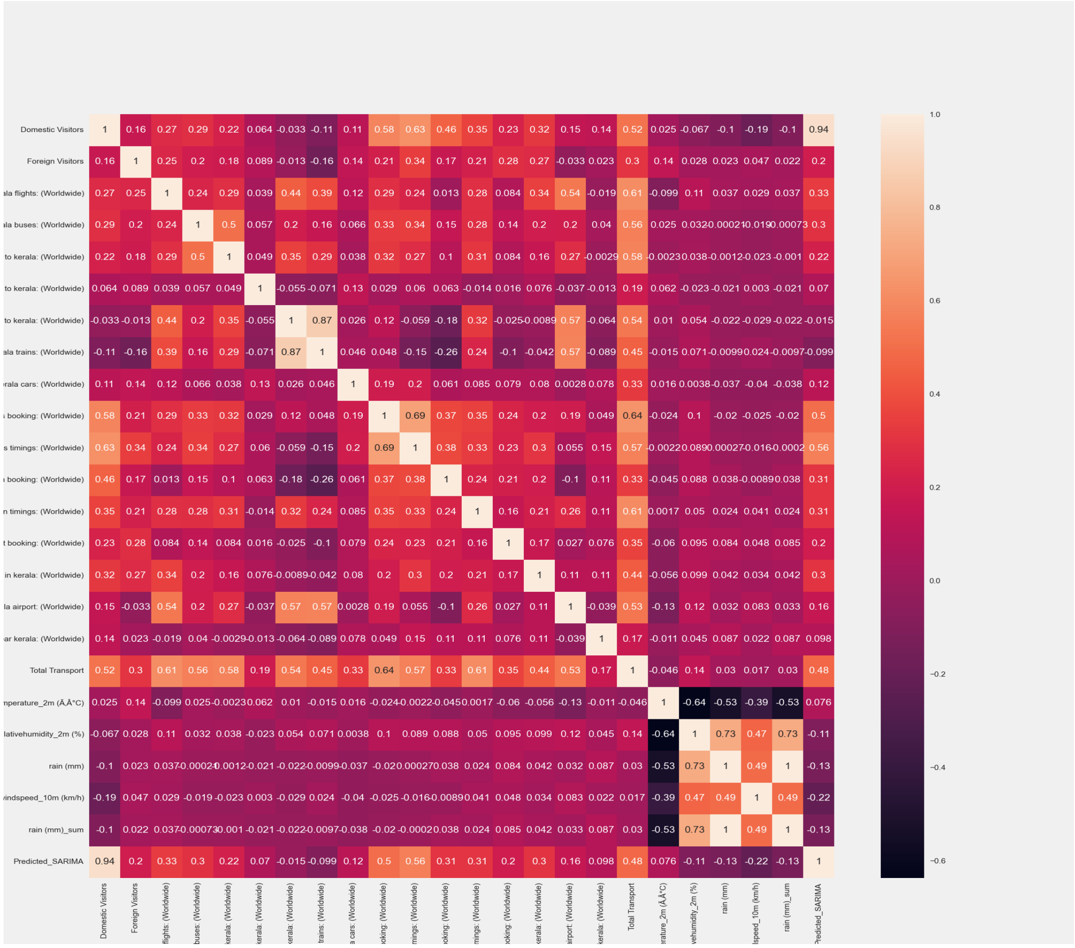
In the highlighted green area, it is apparent that there is a significant decline in domestic tourist numbers. Additionally, for foreign tourists, there is a noticeable downward trend, as many have started to depart from the Kerala city to avoid potential disruptions. It is also worth noting a peak in search interest related to transportation on Google, indicative of heightened curiosity regarding the availability of transportation services.

Covid First Wave



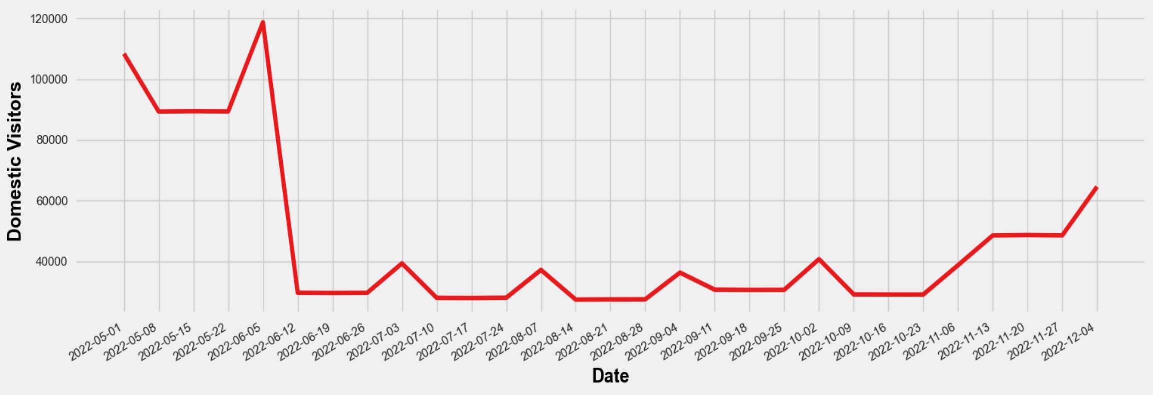
In the graph, it is evident that there was a significant decrease in tourist counts during the initial COVID-19 wave. This decline is particularly pronounced within the green-shaded area. Additionally, there was a noticeable reduction in the volume of transportation-related data searches on Google during the same period, which gradually recovered after a few weeks.

Correlation among features



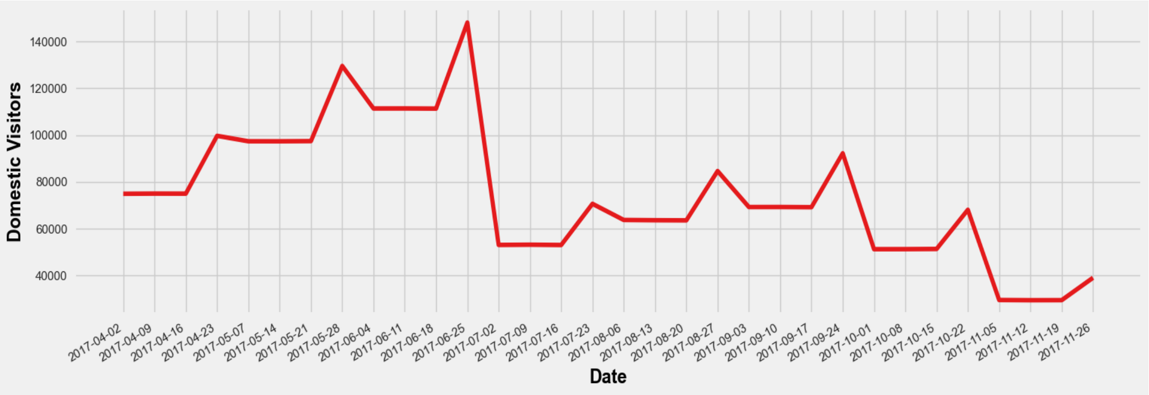
In the heatmap depicted above, it is evident that there is a notable correlation between bus timings and tourist arrivals, denoted by a correlation coefficient of 0.63. This observation suggests a substantial relationship, as tourists often show a keen interest in acquiring information about bus schedules and arrivals.

Cloud Brust in Shimla



The presented plot distinctly illustrates a sharp decline in tourist numbers, attributable to a significant cloud burst incident. The enduring effect of this disaster becomes apparent when analyzing the data over an extended timeframe.

Water Crisis



The water crisis in Shimla has notably diminished tourist arrivals, leaving a discernible impact over an extended duration.

Models tried:-

We turned to a suite of powerful time series analysis models to extract valuable insights, understand temporal patterns, and make informed predictions. This report outlines our approach and findings, detailing the utilisation of models such as ARIMA, SARIMAX, LSTM-based models, and the innovative KELM algorithm. Each model offers unique advantages in dissecting and harnessing time-dependent data, and in the following sections, we discuss their significance and the results they yield.

1. ARIMA (AutoRegressive Integrated Moving Average):

Importance: ARIMA is a classic and widely used time series forecasting model. It combines autoregressive (AR) and moving average (MA) components with differencing to make a time series stationary. This model is particularly valuable when dealing with univariate time series data, as it can capture both short-term and long-term patterns.

$$\underset{\text{differenced time series}}{y'_t} = \overset{\text{intercept}}{c} + \underbrace{\varphi_1 y'_{t-1} + \dots + \varphi_p y'_{t-p}}_{\text{lagged values}} + \underbrace{\theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t}_{\text{lagged errors}}$$

Score obtained: RMSE was 10 to 15 times worse than sarimax.

2. SARIMAX(Seasonal AutoRegressive Integrated Moving Average with Exogenous Variables):

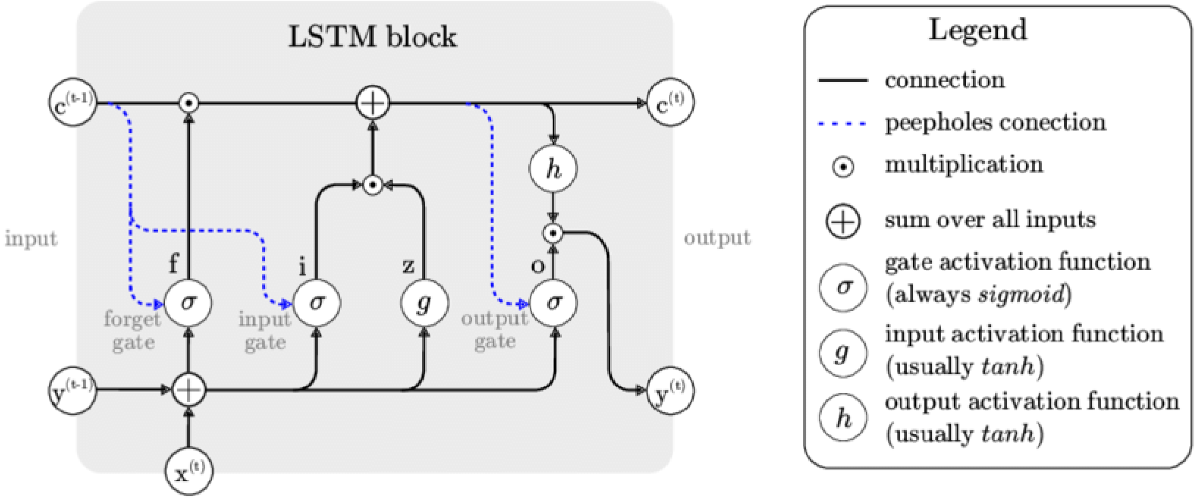
Importance: SARIMAX extends ARIMA by incorporating external factors or exogenous variables that can impact the time series. This feature is valuable for accounting for additional factors such as government policies, weather, or special events that may influence tourist arrivals.

$$d_t = c + \sum_{n=1}^{p} \alpha_n d_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{r} \beta_n x_{n_t} + \sum_{n=1}^{P} \phi_n d_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

RMSE: 0.01

3. LSTM (Long Short-Term Memory):

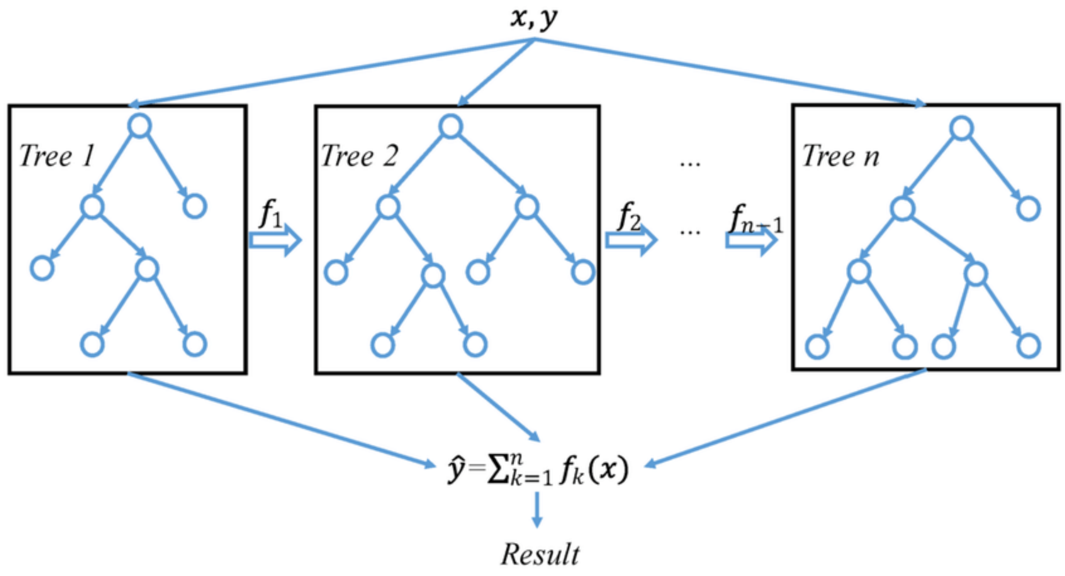
Importance: LSTM is a deep learning-based model used for time series analysis and forecasting. Unlike traditional statistical methods, LSTM can capture complex, non-linear relationships in the data. Its ability to handle sequences of data, remember long-term dependencies, and adapt to changing patterns makes it suitable for a wide range of time series applications, including natural language processing, speech recognition, and financial predictions.



4. XG-BOOST:

XGBoost works as Newton-Raphson in function space, unlike gradient boosting, which works as gradient descent in function space, a second-order Taylor approximation is used in the loss function to connect to the Newton-Raphson method.

A generic unregularised XGBoost architecture is:



Results and Discussions:

Our study focused on predicting tourist outcomes by integrating weather data, transport and other trends. Rainfall and transportation-related searches were found to have a significant impact. Surprisingly, rainfall positively influenced tourist activity, potentially due to indoor activities. Increased train and transport searches correlated with higher tourist arrivals, underscoring the importance of accessible transportation.

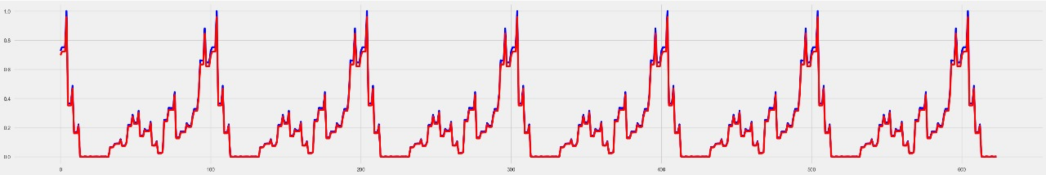
We tested ARIMA, LSTM, and SARIMAX models, with SARIMAX emerging as the most accurate predictor. This model's superior performance emphasises its utility in understanding the complex interplay between weather, transportation, and tourist behaviour, aiding policymakers and industry stakeholders in effective decision-making.

Among these models, **SARIMAX** consistently yielded the best Root Mean Squared Error (RMSE) scores, indicating its potential for generalisation to new datasets for future tourist arrival predictions.

Our findings stress the need to consider weather and transportation in tourism strategies. Future research can explore the nuanced relationship between specific weather attributes and tourist preferences. Additionally, a deeper analysis of factors driving transportation trends could optimise travel systems for tourists. Overall, this study contributes to informed and sustainable tourism development.

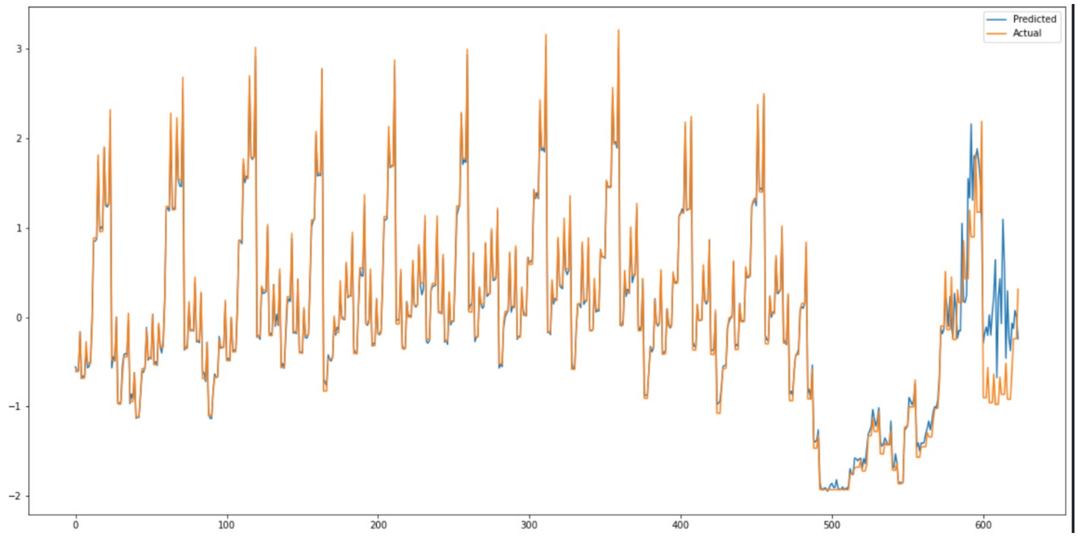
Best Model: <u>SARIMAX(Seasonal AutoRegressive Integrated Moving Average with Exogenous Variables):</u>

Forecast:



<u>RMSE: 0.0131</u>

<u>XG-BOOST:</u>



Model	**RMSE**
SARIMAX	0.0131
ARIMA	0.1
LSTM	0.51
XG Boost	0.588

The superior performance of SARIMAX in predicting tourist arrivals on our dataset suggests its potential for generalisation to new datasets with similar features. Given that we have relevant historical data and exogenous variables, we can confidently apply the SARIMAX model to forecast future tourist arrivals. The model's ability to capture seasonality and external factors makes it an attractive choice for tourism industry stakeholders.

SARIMAX has proven to be a robust and accurate model for forecasting tourist arrivals, outperforming other models such as ARIMA, LSTM, and XGBoost.

It is important to note that, despite its promising performance, regular model retraining and validation are necessary to ensure continued accuracy, especially as external factors influencing tourism can change over time.

References:

Shimla Weather

•https://open-meteo.com/en/docs/historical-weather-api

Shimla Tourism

https://himachaltourism.gov.in/types/district/

Kerala Weather

•https://open-meteo.com/en/docs/historical-weather-api

Kerala Tourism

•https://www.keralatourism.org/touriststatistics/

Google Trends

•https://trends.google.com/trends/