IndiaTour Insights :

Project Report

"A platform to analyze and predict tourist behavior in India using machine learning."

Course: Data Engineering - CSL4030

Group: 13

Saurav Soni - (B22AI035) | Dhruva Kumar Kaushal - (B22AI017)

All Source Code Available at Our Github:https://github.com/dhruvak001/data_engg

Demo Video Link: 🎬 **Screen Recording 2024-11-19 at 11.11.35 AM.mov**

1. INTRODUCTION

The **Tourist Behavior Analysis** project involves analyzing foreign visitors to India, including foreign tourists, overseas Indians, and crew members (excluding non-tourist arrivals such as diplomats and soldiers). The dataset used for this analysis has been constructed from Indian Government data, making it suitable for thorough analysis. This report outlines the various technologies, methodologies, and steps employed in the project, as well as the final deliverables.

The project aims to predict tourist behavior, identify trends in tourism, and provide insights into foreign exchange earnings. It leverages machine learning models, data wrangling techniques, and effective deployment strategies to analyze data and present results interactively.

2. Use Case of Technologies

The project involves a wide array of technologies that facilitate data analysis, machine learning, backend development, and deployment. The key technologies are:

3.1. Data Processing and Analysis

- **Python**:
  - Used for data wrangling, manipulation, and visualization.
  - Libraries like Pandas, NumPy, and Matplotlib were used for data cleaning, feature extraction, and plotting.
- **Machine Learning Models**:
  - **Scikit-learn**: For implementing and training predictive models.
  - **Linear Regression & Decision Trees**: Used to forecast tourism patterns based on historical data.
  - **Model Validation**: Cross-validation techniques to ensure robustness of predictions.
- **Apache Spark**:
  - **Big Data Processing**: Apache Spark was used to process large datasets efficiently and in parallel, enabling faster computations compared to traditional methods.
  - **Spark MLlib**: Leveraged for scalable machine learning algorithms, improving the performance and scalability of the prediction models.
  - **Data Handling**: Used for aggregating and transforming large datasets (e.g., foreign and domestic visitors, monument statistics) and for performing

large-scale data analysis in a distributed environment.

3.2. Web Development

- **Streamlit**:
  - Chosen for its simplicity and efficiency in building interactive web applications.
  - Used for integrating machine learning models and visualizing predictions on a real-time dashboard.
- **Frontend Design**:
  - Focus on a clean and user-friendly interface to present the analysis in an easily understandable format for both technical and non-technical users.

3.3. Backend and Database

- **MySQL**:
  - Used for storing data related to tourist arrivals, monuments, and other relevant statistics.
  - Efficient queries were designed to retrieve data and support user interactions.

Here backend serves as the engine for performing all machine learning operations, such as training models and making predictions. It manages the flow of data between the frontend and the database, processing incoming requests and returning predictions.

3.4. Containerization and Deployment

- **Docker** is employed for containerizing the frontend, backend, and database components.
- The use of Docker ensures that the entire application, from data processing to visualization, is contained in a consistent and repeatable environment.
- This simplifies the deployment and scaling of the application across different environments, such as development, staging, and production.
- With Docker, the containers are isolated, ensuring that dependencies do not conflict between the different parts of the system. This also allows for easier version control and updates without affecting other components.

4. Key Features of the System

4.1. Data Integration and Storage

- Data from different sources (government databases and online surveys) was integrated into a unified MySQL database.
- The database stores information on:
  - Number of foreign and domestic visitors to various monuments.
  - Growth percentages, comparison metrics, and foreign exchange earnings.

4.2. Machine Learning Predictions

- Developed predictive models to forecast the number of visitors based on historical data and key variables.
- **Foreign Visitor Predictions**: Predicted trends for foreign visitors, considering factors like global tourism trends and regional events.
- **Domestic Visitor Predictions**: Focused on domestic tourism within India, factoring in local preferences and regional events.

4.3. Web Interface (Streamlit App)

- The Streamlit web interface was designed to:
  - Allow users to input specific criteria (like region or monument) to view predictions and trends.
  - Display charts and graphs for better visualization of data trends.
  - Allow comparison between actual and predicted data, providing insights into the accuracy and potential gaps in the data.

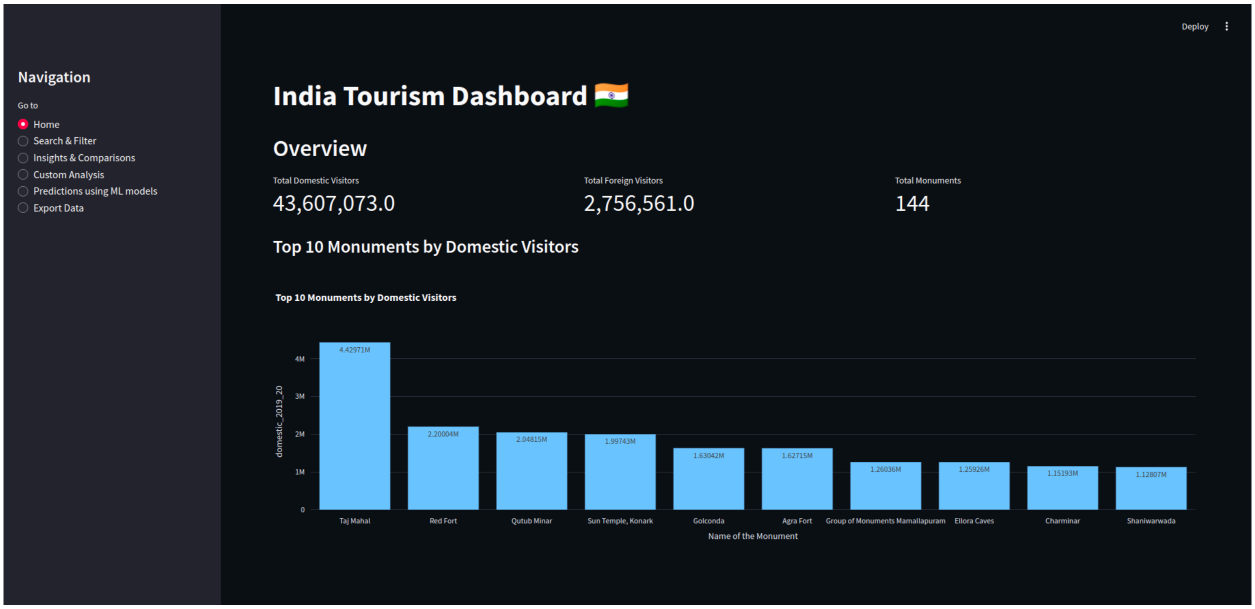
4.4. Dashboard Features

- **Interactive Charts**: Users can explore actual vs predicted trends for domestic and foreign visitors.
- **Real-Time Data**: The dashboard pulls real-time data from the MySQL database to display the most up-to-date information.
- **User-Friendly Interface**: Simple navigation that makes it easy for non-technical users to interact with the system and obtain relevant insights.

5. Key Platform Features:

- Page Descriptions

Home Page:



India Tourism Dashboard 🇮🇳

Overview

Total Domestic Visitors: 43,607,073.0

Total Foreign Visitors: 2,756,561.0

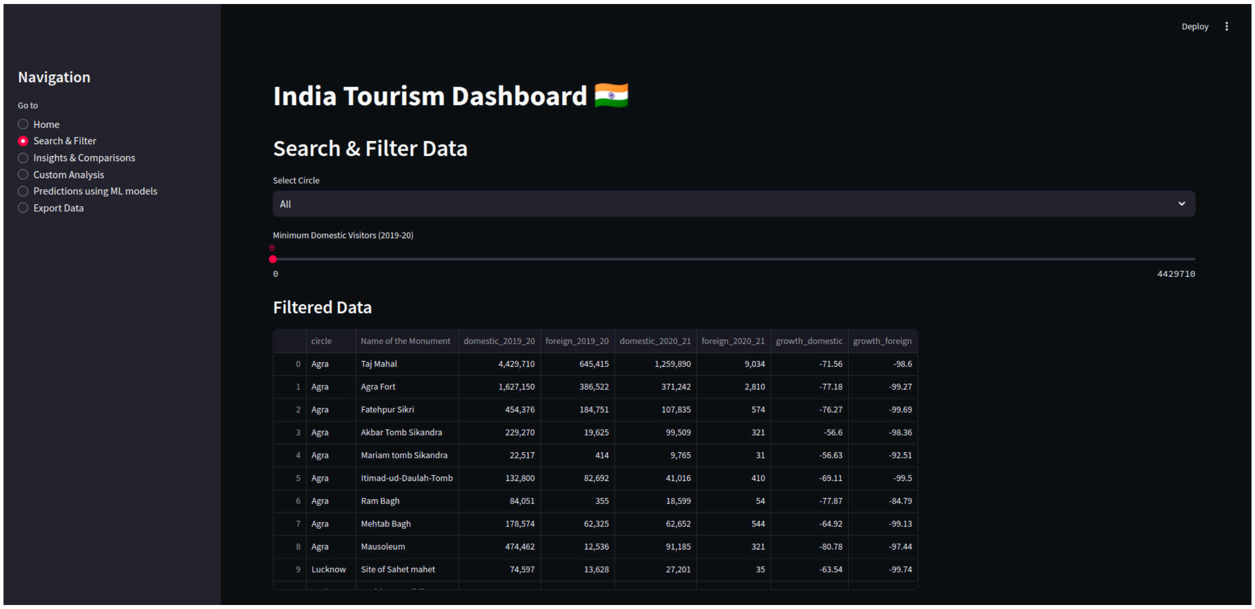
Total Monuments: 144

Top 10 Monuments by Domestic Visitors

The Home page provides a quick overview of key metrics and insights related to the tourism dataset:

- **Key Metrics Display**:
  - **Total Domestic Visitors**: The cumulative number of domestic tourists during 2019-20.
  - **Total Foreign Visitors**: The total number of foreign tourists for the same period.
  - **Total Monuments**: The number of unique monuments included in the dataset.
- **Top 10 Monuments**:
  - A bar chart visualization ranks the top 10 monuments by the number of domestic visitors in 2019-20. This helps identify the most popular monuments among domestic tourists.

Search & Filter:



India Tourism Dashboard 🇮🇳

Search & Filter Data

Select Circle: All

Minimum Domestic Visitors (2019-20): 0 (0 – 4429710)

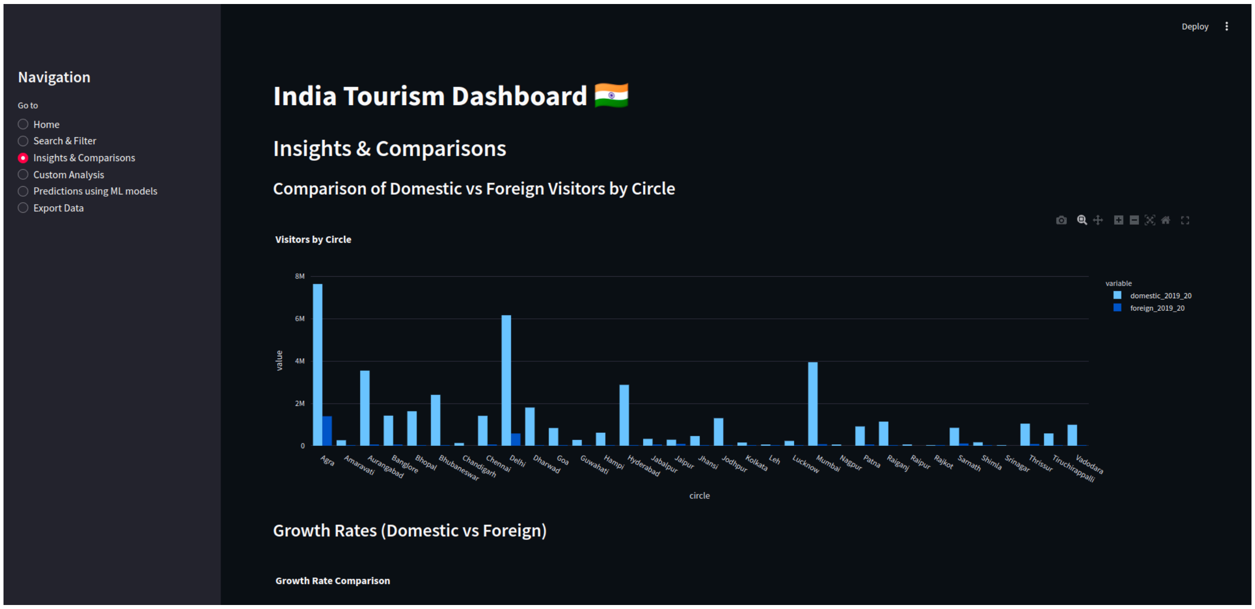
Filtered Data

	circle	Name of the Monument	domestic_2019_20	foreign_2019_20	domestic_2020_21	foreign_2020_21	growth_domestic	growth_foreign
0	Agra	Taj Mahal	4,429,710	645,415	1,259,890	9,034	-71.56	-98.6
1	Agra	Agra Fort	1,627,150	386,522	371,242	2,810	-77.18	-99.27
2	Agra	Fatehpur Sikri	454,376	184,751	107,835	574	-76.27	-99.69
3	Agra	Akbar Tomb Sikandra	229,270	19,625	99,509	321	-56.6	-98.36
4	Agra	Mariam tomb Sikandra	22,517	414	9,765	31	-56.63	-92.51
5	Agra	Itimad-ud-Daulah-Tomb	132,800	82,692	41,016	410	-69.11	-99.5
6	Agra	Ram Bagh	84,051	355	18,599	54	-77.87	-84.79
7	Agra	Mehtab Bagh	178,574	62,325	62,652	544	-64.92	-99.13
8	Agra	Mausoleum	474,462	12,536	91,185	321	-80.78	-97.44
9	Lucknow	Site of Sahet mahet	74,597	13,628	27,201	35	-63.54	-99.74

This page enables users to search and filter the dataset based on specific criteria:

- **Circle-Based Filter**:
  - Users can filter the dataset by selecting a specific archaeological circle (administrative divisions of monuments).
- **Visitor Count Filter**:
  - A slider allows users to set a minimum threshold for the number of domestic visitors (2019-20).
- **Filtered Data**:
  - The filtered results are displayed in a tabular format.
- **Statistics**:
  - Descriptive statistics of the filtered data (e.g., mean, median, and range) are provided to summarize the subset.

Insights & Comparisons:



Deploy

Navigation

Go to

○ Home
○ Search & Filter
● Insights & Comparisons
○ Custom Analysis
○ Predictions using ML models
○ Export Data

India Tourism Dashboard 🇮🇳

Insights & Comparisons

Comparison of Domestic vs Foreign Visitors by Circle

Visitors by Circle

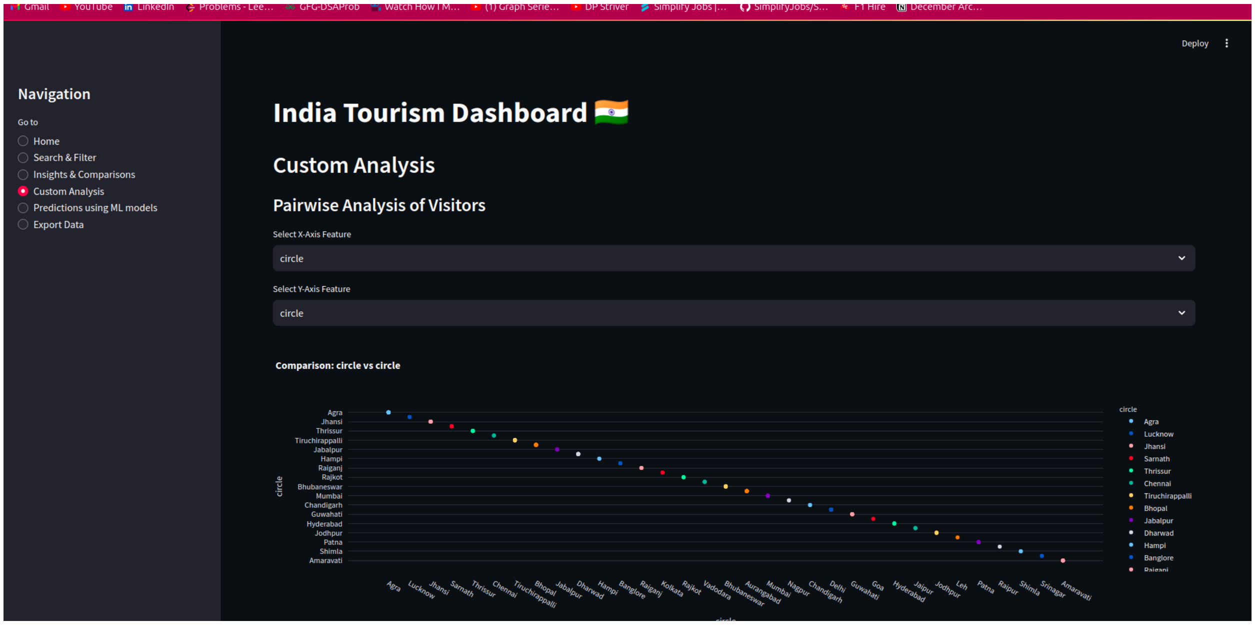
Growth Rates (Domestic vs Foreign)

Growth Rate Comparison

This page focuses on visual comparisons and growth rate analysis:

- **Visitors by Circle**:
  - A grouped bar chart compares domestic and foreign visitor counts for each archaeological circle.
- **Growth Rate Analysis**:
  - A scatter plot visualizes the growth rates of domestic vs. foreign visitors for each monument. This highlights trends and anomalies in visitor growth.

Custom Analysis:



India Tourism Dashboard 🇮🇳

Custom Analysis

Pairwise Analysis of Visitors

Select X-Axis Feature: circle

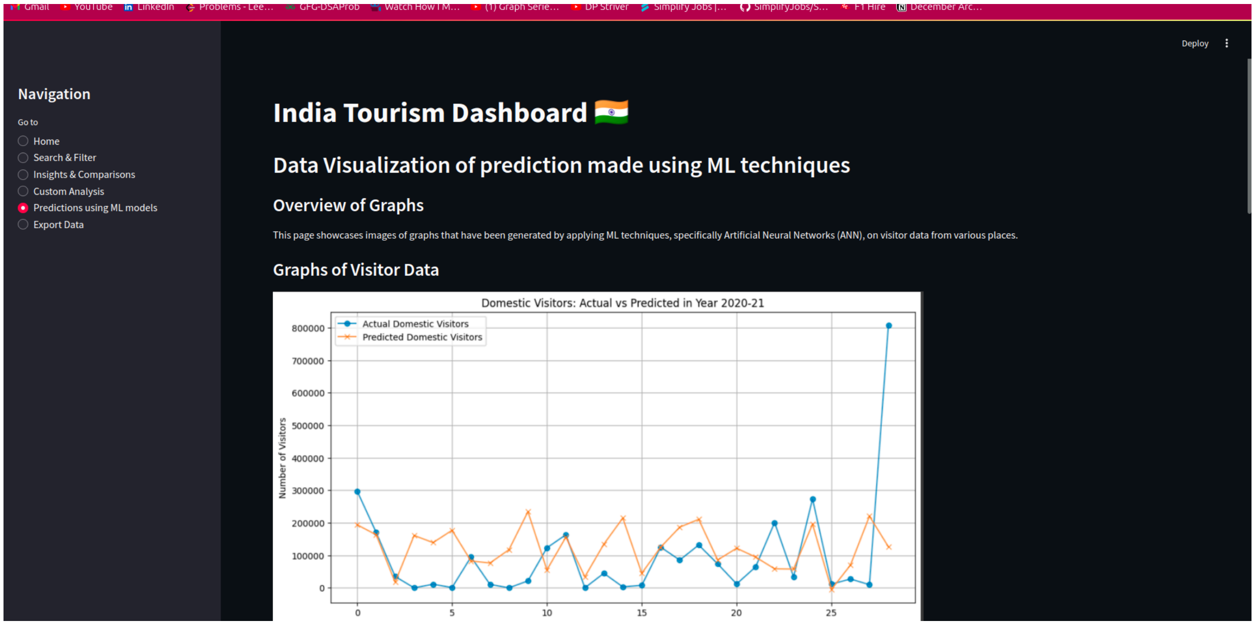
Select Y-Axis Feature: circle

Comparison: circle vs circle

The Custom Analysis page allows for advanced, user-defined data exploration:

- **Pairwise Comparisons**:
  - Users can select any two features from the dataset for a scatter plot comparison. This supports hypothesis testing and deeper insights into feature relationships.
- **Correlation Heatmap**:
  - A heatmap displays correlations between numerical features, aiding in identifying strong relationships or dependencies among variables.

Predictions Using ML Models:



Navigation

Go to

○ Home
○ Search & Filter
○ Insights & Comparisons
○ Custom Analysis
● Predictions using ML models
○ Export Data

India Tourism Dashboard 🇮🇳

Data Visualization of prediction made using ML techniques

Overview of Graphs

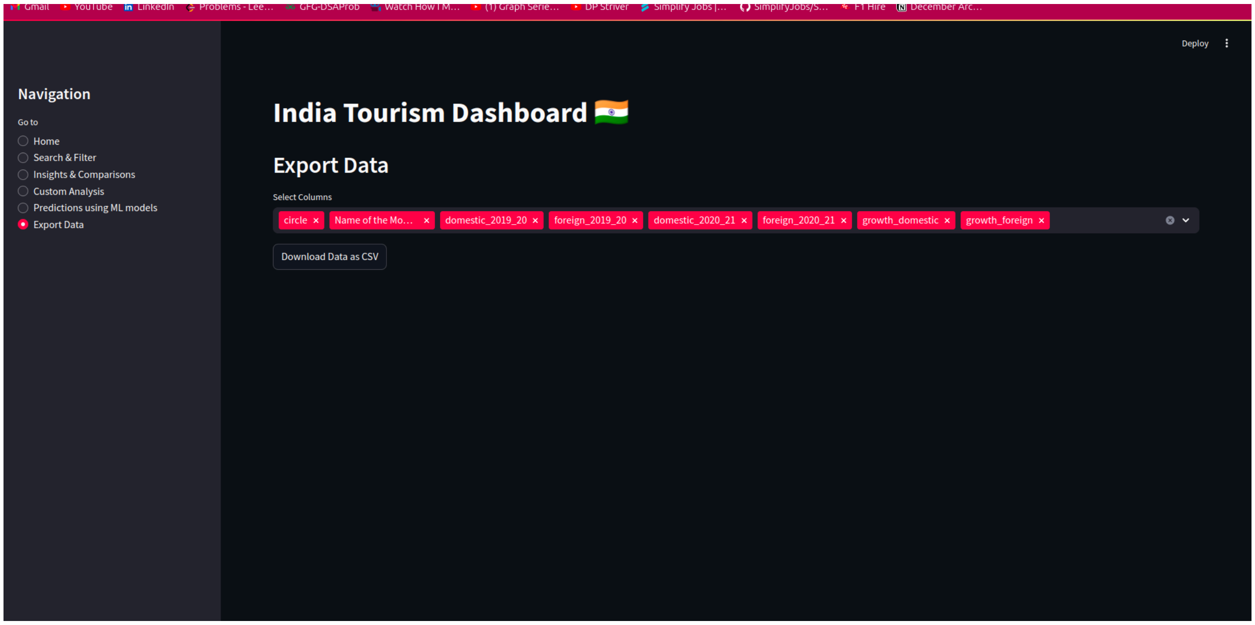
This page showcases images of graphs that have been generated by applying ML techniques, specifically Artificial Neural Networks (ANN), on visitor data from various places.

Graphs of Visitor Data

This page showcases predictions made using machine learning techniques, specifically we have used Artificial Neural Networks (ANNs):

- **Visualization of Predictions**:
  - Pre-generated graphs illustrate trends in visitor predictions for various monuments.
  - Graphs include both domestic and foreign visitor trends predicted by ANN models.
- **Explanation**:
  - Descriptions accompany each graph, explaining the insights derived from the predictions and the ANN model's objectives.

Export Data:



India Tourism Dashboard 🇮🇳

Export Data

Select Columns: circle ×, Name of the Mo... ×, domestic_2019_20 ×, foreign_2019_20 ×, domestic_2020_21 ×, foreign_2020_21 ×, growth_domestic ×, growth_foreign ×

This page provides functionality to export the dataset:

- **Custom Column Selection**:
  - Users can select specific columns of interest from the dataset to include in the export.
- **Download Option**:
  - The filtered data is made available for download in CSV format, facilitating offline analysis or sharing.

6. Challenges and Solutions

5.1. Data Quality and Integrity

- **Challenge**: The dataset contained missing or inconsistent data.
- **Solution**: Used data wrangling techniques in Python to clean and fill missing values, ensuring accurate predictions.

5.2. Model Accuracy

- **Challenge**: Achieving high accuracy in predicting visitor trends.

- **Solution**: Multiple models were tested, with the final choice being based on performance metrics such as Mean Squared Error (MSE). Cross-validation ensured that the models were robust and not overfitting.

5.3. Containerization and Deployment

- **Challenge**: Managing the different components of the project (frontend, backend, database) and ensuring smooth deployment.
- **Solution**: Docker and Docker Compose were used to containerize each component, ensuring a seamless and consistent deployment process across environments.

5.4. Big Data Processing

- **Challenge**: The large volume of data made it difficult to process efficiently.
- **Solution**: Apache Spark was employed to handle and process large datasets in parallel, significantly improving the performance of data aggregation and analysis tasks.

Data Analysis :

Google Colab LINK: **Data_Analysis.ipynb**

1. **Domestic Visitors Analysis :**

Observation of Prediction Accuracy:

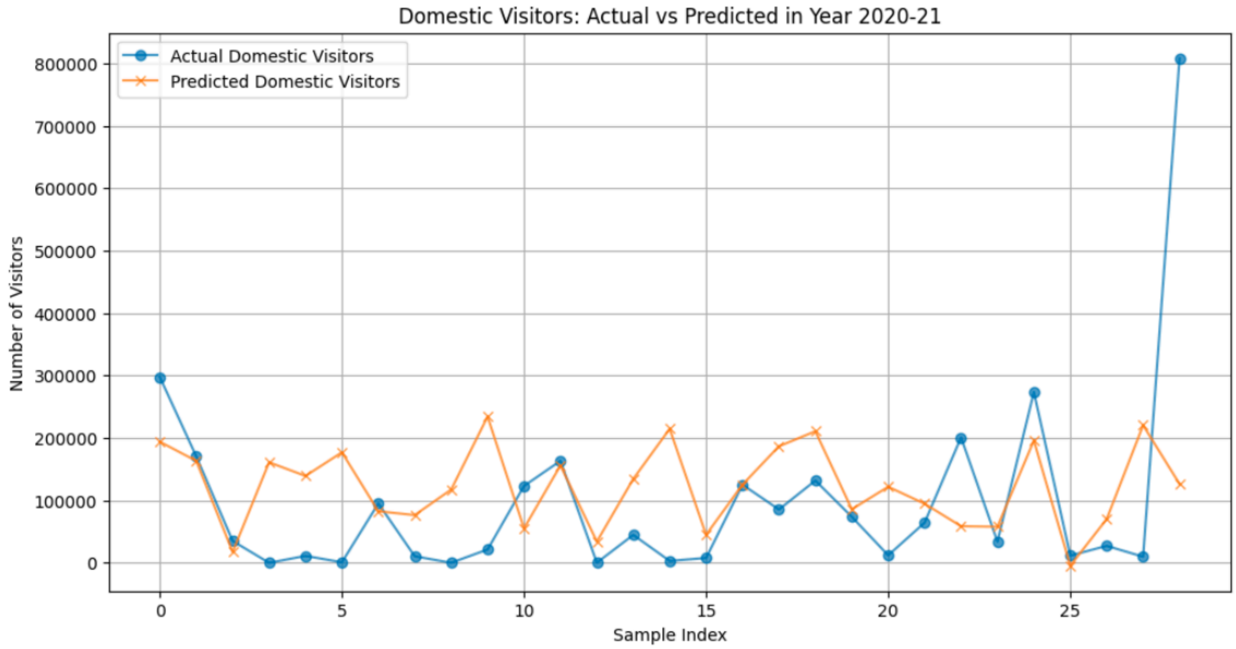
- The predictions closely follow the actual visitor trends for most samples.
- However, there are some deviations for monuments with higher visitor counts (e.g., sample 0 and 26), where the model underpredicts the actual visitor numbers.

General Trends:

- A significant dip in the number of domestic visitors is noticeable for most samples. This aligns with the overall decline in tourism during 2020-21 due to the pandemic.
- Some monuments (e.g., sample 26) show a large actual visitor count despite the general downward trend, potentially due to local tourism still being active.

Model Performance:

- The prediction model captures the general patterns effectively but struggles with extreme values, particularly for monuments with very high visitor counts.



2. **Foreign Visitors Analysis**

Observation of Prediction Accuracy:

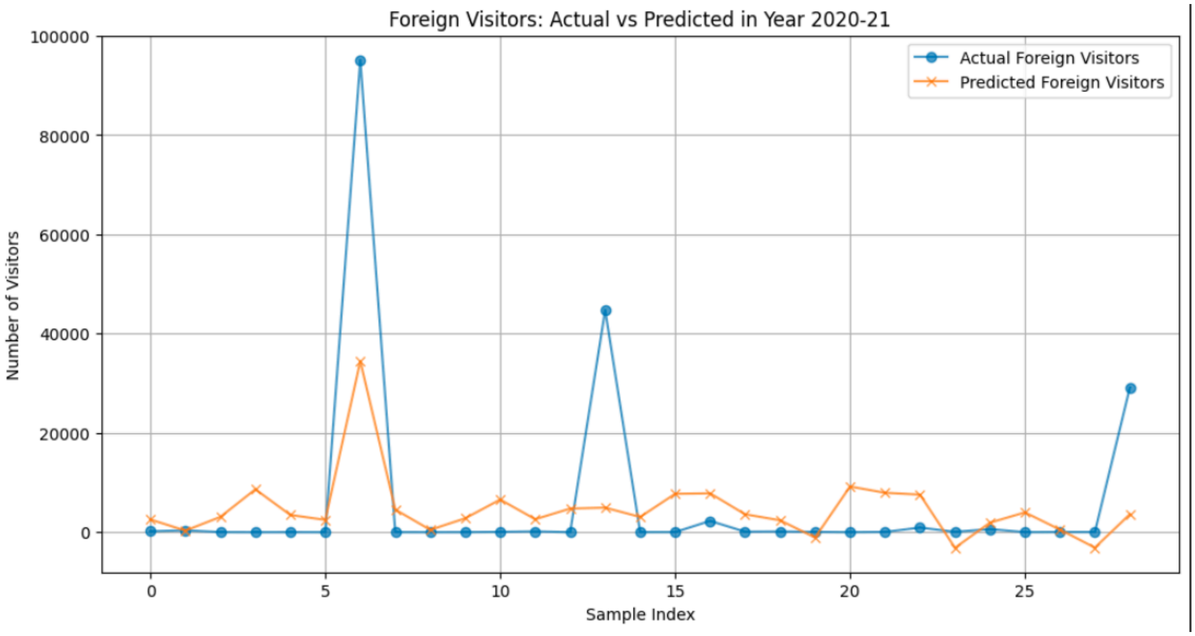
- The predictions are fairly close to actual values for lower visitor counts (e.g., most samples).
- However, significant deviations occur for outliers, such as sample 5, where actual visitors greatly exceeded the prediction.

General Trends:

- Foreign tourism faced a far steeper decline compared to domestic tourism, as evident from the much lower visitor counts for most samples.
- The few outliers (e.g., sample 5 and 20) suggest that a handful of monuments might still have attracted foreign tourists, possibly due to their prominence or delayed international travel bans.

Model Performance:

- The model accurately predicts trends for smaller visitor counts but does not generalize well to outliers (e.g., monuments with unexpectedly high foreign visitor numbers).



3. Key Insights

Pandemic Impact:

- Both domestic and foreign tourism saw sharp declines, but the impact was more pronounced for foreign visitors.
- Domestic tourism had pockets of resilience, potentially due to localized or regional travel being less restricted.
- The dataset tracks a significant decline in tourism during the **2020-21 period**, due to the COVID-19 pandemic.

Outliers:

- Some monuments deviate significantly from general trends, indicating unique factors influencing their visitor counts (e.g., cultural significance, regional policies, or marketing efforts).

- The Taj Mahal experienced a drastic reduction in both domestic and foreign visitors: **-71.56%** domestic and **-98.60%** foreign.

:CONCLUSION

This project successfully implemented a predictive system for analyzing tourist behavior, integrating machine learning with a user-friendly web interface. By leveraging technologies like Streamlit, Docker, MySQL, and Apache Spark, we created a scalable and efficient platform that can be extended to include additional features, such as real-time data collection and more granular predictions.