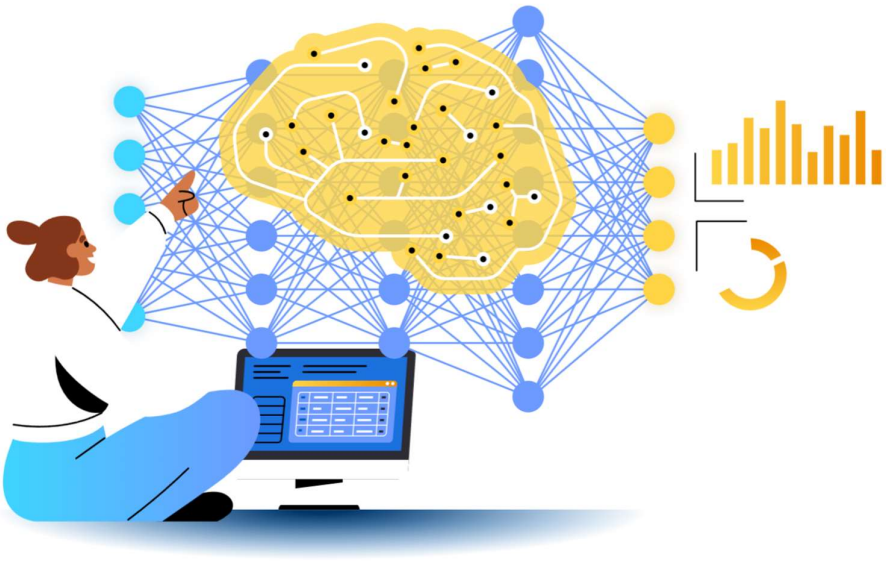
Forecasting Footfall Trends

A Data-Driven Approach



Report by-

DATA_TITANS

SERIAL NO.: D49

INDEX

PROBLEM STATEMENT:

Our goal was to design and implement a machine learning model that predicts tourist arrivals to a specific destination using Internet search index data as a key input. The goal of this competition is to develop an accurate and robust predictive model that can assist tourism authorities and businesses in forecasting tourist arrivals, thereby enabling better resource allocation, marketing strategies, and overall destination management

ABSTRACT:

The tourism industry plays a pivotal role in stimulating economic growth by contributing significantly to revenue generation and employment opportunities for local communities. To effectively support the increasing tourist influx, accurate forecasting of tourist arrivals is imperative. The inaccuracies in forecasting models can have direct and indirect consequences on the tourism industry. In this study, we propose a solution involving the development of a Machine Learning (ML) model for predicting tourist arrivals, utilizing time series data and Google search index data.

Our solution relies on time series data as the foundational component for predictions, with an added optimization element based on the search frequency of specific queries listed in the Google search index. Furthermore, we allow for interpretability and customization of the results by adjusting the parametric weights used in the calculation process. To perform our analysis, we leveraged publicly available data on tourist arrivals in Manali and integrated it with Google Trends data covering worldwide search queries since 2008.

In addition to our forecasting model, we have developed a recommendation system aimed at enhancing the overall user experience for tourists. This system provides real-time information about nearby hotels based on the user's current location, thereby offering a valuable service to tourists.

INTRODUCTION

The tourism industry in India is a vital sector, making a substantial contribution to the country's GDP, accounting for approximately 9% in 2018. Furthermore, it plays a pivotal role in providing livelihoods, with approximately 42 million jobs, constituting around 8.1% of India's total employment, directly dependent on this industry. However, tourism is subject to considerable uncertainty in terms of the number of tourist arrivals due to a multitude of interconnected factors, both dependent and independent, that influence this sector. This uncertainty places stress on existing infrastructure and resources, subsequently impacting the tourism industry and the overall economy. As a result, there is a pressing need for the accurate and dynamic forecasting of tourist arrivals, especially in regions that experience seasonal fluctuations in tourism.

Tourist footfall is influenced by a myriad of factors, including seasonality, the popularity of destinations, regulatory policies, the country's GDP, news events, special occasions, Consumer Price Index (CPI) fluctuations, and the shifting interests and preferences of tourists. Although it is challenging to quantify the exact influence of each of these parameters on tourist arrivals, one valuable dataset that offers insights into the interest in a particular location over time is the Google search index.

Our approach will include the implementation of the AutoRegressive Integrated Moving Average (ARIMA) model, which is well-suited for time series forecasting. We'll determine the appropriate values for model parameters, such as differencing orders (d), autoregressive order (p), and moving average order (q).In a nutshell, we will be using SARIMA coupled with XGBoost.

With the integration of technology and data, the mood and interests of tourists are directly reflected in their search queries, which are meticulously cataloged by Google and made available as open-source information through Google Trends. The combination of this data with historical records of tourist footfall can serve as a powerful tool for accurately forecasting tourist arrivals. Therefore, our project aims to construct a system that enables the provision of the aforementioned forecasting service, leveraging the synergy between Google Trends data and historical tourist arrival data. This system will aid in not only forecasting but also in optimizing and efficiently managing tourism resources, thereby benefiting both the industry and the economy.

METHODOLOGY:

Data Collection:

1.Monthly Tourism Data

We collected monthly total footfall data for Himachal Pradesh tourism destinations from the year 2010 to 2019. This data was obtained from a government portal, providing insights into the number of tourists visiting the state's key destinations.

2.Google Search Index Data

Additionally, we obtained data from various Google search indexes related to tourism, including queries for tourist spots, travel agencies, train and flight bookings, and tourism packages. We sourced this data from Google Trends, which offers information on the interest over time for specific locations. This search index data, combined with historical tourism footfall data, will be crucial for accurate forecasting.

Data Preprocessing:

After data collection, our next step involved data preprocessing to make it suitable for modeling. The process was as follows:

1. Data Loading: We utilized the powerful pandas library to load the CSV files containing both the monthly footfall data and the Google Trends data into dataframes. This provided a structured format for subsequent analysis.

2. Data Concatenation: To integrate the two datasets effectively, we concatenated them, utilizing the month as the primary index. This step ensured that the data was aligned and ready for analysis.

3. Data Cleaning: Fortunately, our dataset was in a relatively clean state, with no missing values requiring attention. This allowed us to proceed directly to modeling without extensive data cleaning efforts.

Notable Observation:

Throughout our data preprocessing phase and exploratory data analysis, a conspicuous observation emerged. The COVID-19 pandemic, which unfolded during 2020 and continued into 2021, had a profound impact on tourist footfall. There was a sharp and substantial decline in the number of tourists visiting these destinations during this challenging period. This observation served as a crucial backdrop for understanding the unique dynamics of the data.So we refrained us from considering the data for this time period assuming pandemics like COVID-19 are very less probable to occur in the near future.

Model Creation:

For our time series forecasting task of Shimla's tourist footfall, we have several methods at our disposal, and we will discuss our approach in detail.

ARIMA Model:

The ARIMA (AutoRegressive Integrated Moving Average) model is a classical and widely-used time series forecasting technique. We've already performed the necessary data differencing (d) and have an idea about the values for p and q. With this information, we're ready to proceed with fitting an ARIMA model to our data.

Seasonal Decomposition of Time Series (SARIMA):

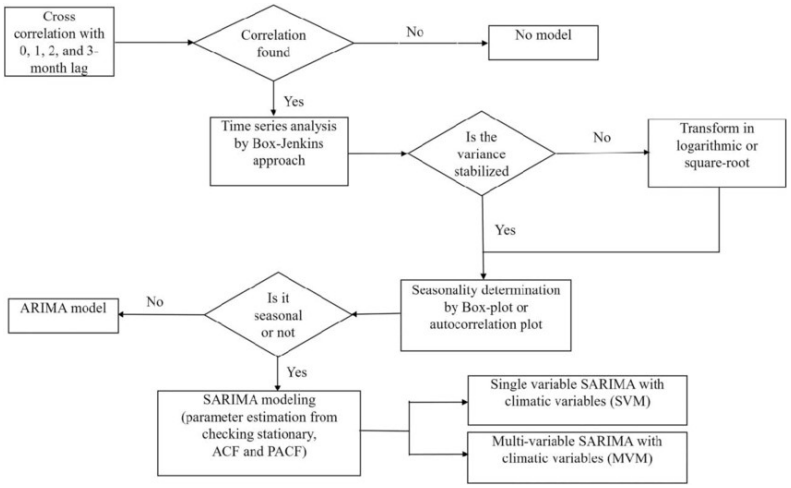
Given that tourism often exhibits seasonal patterns, we must consider the impact of seasonality on our forecasts. If the initial ARIMA model does not adequately capture the seasonality in the data, we will explore the Seasonal ARIMA (SARIMA) model. This extended version of ARIMA incorporates seasonal components to better account for recurring patterns in our tourist footfall data.

Model Validation:

Once we've developed our models, it's essential to assess their performance. To do this, we will split the dataset into a training and validation set. The training set will be used to fit the models, while the validation set will allow us to evaluate their predictive accuracy. This step ensures that we don't just have models that fit the historical data but can also make accurate predictions.

Forecasting:

With the ARIMA and, if necessary, SARIMA models in place, we'll be able to make forecasts for future time periods. These forecasts will provide valuable insights into the expected tourist footfall in the Shimla region, which can help local authorities and businesses prepare for changes in demand.



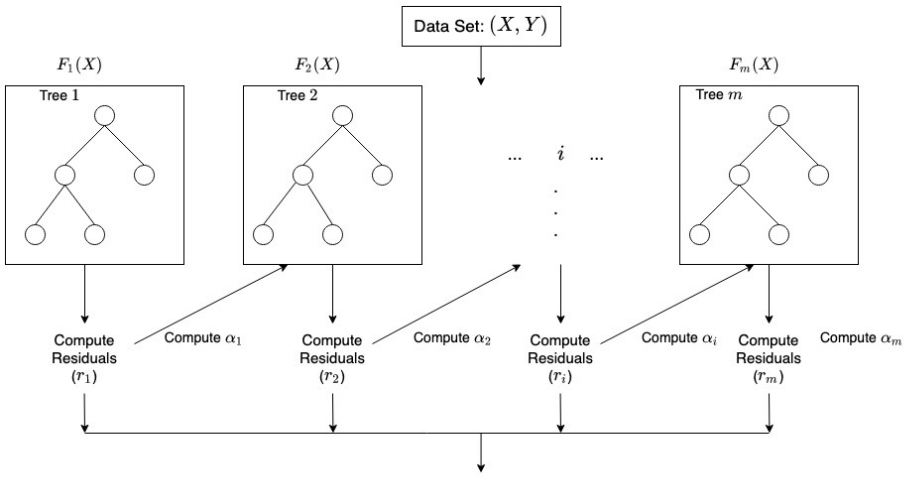
Model Evaluation:

Assessing the accuracy of our models is crucial. We will use common evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to compare our forecasts against the actual footfall data. This evaluation step will give us a clear understanding of how well our models are performing.At last we used NRMSE(nrmse = rmse / y_true_range).

Model Improvement:

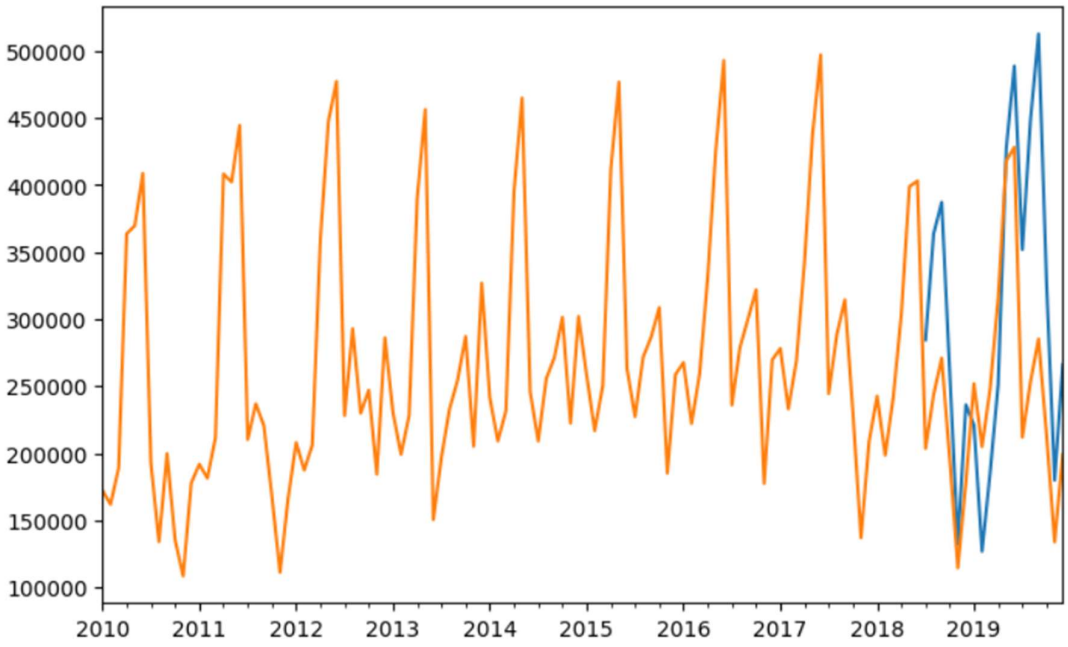
If we find that the ARIMA or SARIMA models' performance is not up to our expectations, we are prepared to take further steps to improve our forecasting accuracy. Potential actions include fine-tuning the hyperparameters of the models or exploring other time series forecasting methods. For instance, we may consider Seasonal Decomposition of Time Series (STL) or machine learning approaches such as XGBoost. The choice will depend on the specific characteristics of our data and the results of our initial model evaluations.

In conclusion, we are well-prepared to tackle the task of forecasting tourist footfall in Shimla using a combination of time-tested ARIMA models and more advanced methods, with the ultimate goal of providing accurate and actionable insights for the tourism industry in the region.

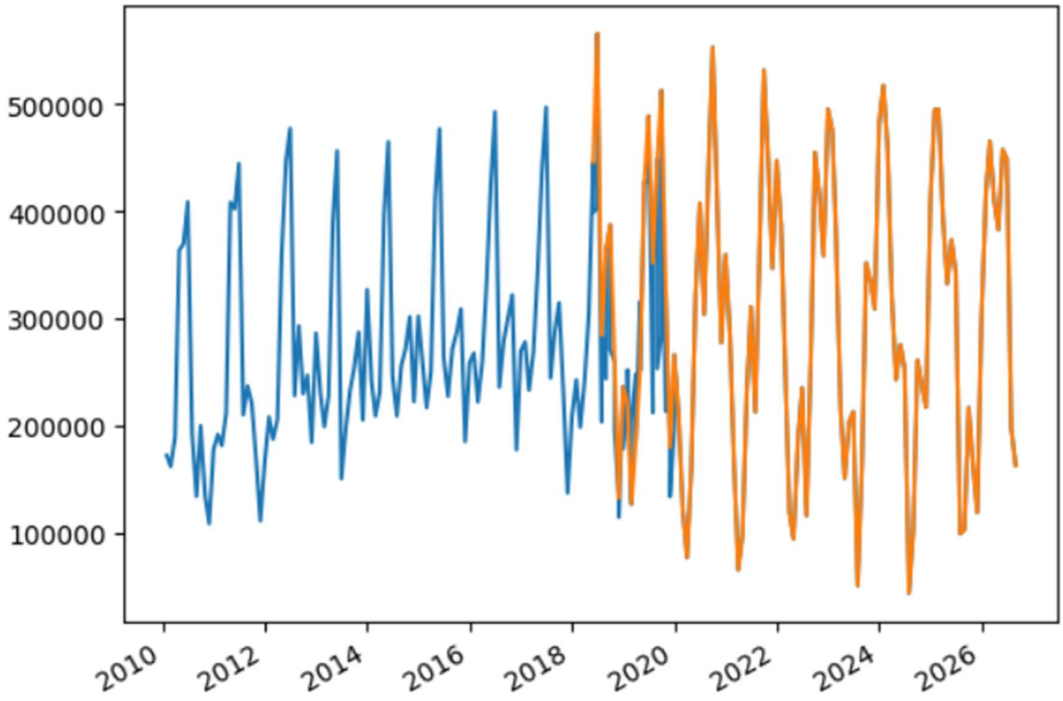


PREDICTIONS:

1. **PREDICTION WITH SARIMA:**



2. **PREDICTION WITH SARIMA COUPLED WITH XGBOOST**



ERROR ANALYSIS:

TRAINING RMSE: 4430.334481319969

TRAINING NRMSE: 0.011405042762643631

TESTING RMSE: 82897.05636564002

TESTING NRMSE: 0.21340250419776863

OUR MODEL:
https://colab.research.google.com/drive/1X21Z6CKvt9PZvrURsNYIWOqX2ezDKq1h#scrollTo=cOqJ6BfHF_t1

CONCLUSION:

In this initial phase of our project, we've laid a strong foundation for forecasting tourist footfall in Shimla, one of the most popular tourist destinations in India. We've accomplished critical data collection and preprocessing steps, setting the stage for effective time series forecasting.

Data collection involved sourcing historical tourist footfall data from 2010 to 2019, alongside other valuable datasets. These additional datasets include snowfall data for Kufri, weather conditions in Shimla and Kufri, Google search trends, and Rohtang Pass temperature. This rich and diverse dataset provides a comprehensive view of the factors influencing tourist visits to Shimla.

Data preprocessing was a crucial aspect of our work. We ensured data consistency and completeness by handling missing values and concatenating data from different sources. Furthermore, we performed differencing on the time series data to achieve stationarity, a prerequisite for many time series forecasting models.

Through exploratory data analysis, we gained significant insights into the data. Notably, we identified clear seasonality in tourist footfall, which aligns with the typical tourist seasons in Shimla. Additionally, we found evidence of autocorrelation, indicating that past footfall is a valuable predictor of future footfall.

As we move forward, the project's next phases will focus on modeling and forecasting. Our approach will include the implementation of the AutoRegressive Integrated Moving Average (ARIMA) model, which is well-suited for time series forecasting. We'll determine the appropriate values for model parameters, such as differencing orders (d), autoregressive order (p), and moving average order (q).

To assess the model's accuracy, we'll split the dataset into a training set and a validation set, allowing us to evaluate how well the model performs on unseen data. Common evaluation metrics like Root Mean Squared Error (RMSE) and NRMSE will be used to gauge model performance.

If the ARIMA model's performance doesn't meet our expectations, we'll explore alternative modeling approaches, including Seasonal Decomposition of Time Series (STL) or machine learning models like XGBoost. The ultimate goal is to provide stakeholders with accurate tourist footfall forecasts that can guide decisions related to tourism planning, resource allocation, and business strategies in Shimla.

This project plays a pivotal role in supporting the local tourism industry, as well as local authorities and businesses, as they make informed, data-driven choices that enhance the

visitor experience and boost the regional economy. Through our work, we aim to contribute to the sustainable growth and success of Shimla as a tourist destination**.**

REFERENCES:

- **IEEE International Conference on Communication information and Computing Technology (ICCICT), June 25-27, 2021,Mumbai, India**
- **Binru Zhang, Yulian Pu, Yuanyuan Wang, Jueyou Li,Forecasting Hotel Accommodation Demand Based on LSTM Model Incorporating Internet Search Index,Sustainability,11, 4708; doi:10.3390/su11174708,2019**
- **Machine Learning: A Probabilistic Perspective,Book by Kevin P. Murphy**
- **https://ieeexplore.ieee.org/document/9510074**
- **https://trends.google.com/trends/**
- **https://himachaltourism.gov.in/wp-content/uploads/2023/03/Tourist-Statistics.pdf**