

# Tourism Analysis and Prediction in the Context of Nepal

Sandesh Sharan Poudel<sup>1</sup>, Ashok GM<sup>2</sup>

<sup>1</sup>Department of Computer Engineering, Lalitpur Engineering College, Lalitpur, Nepal, poudelsandesh321@gmail.com

<sup>2</sup>Associate Professor, Department of Computer Engineering, Himalaya Engineering College, Lalitpur, Nepal, ashokgm22@gmail.com

---

## Abstract

Tourism is a vital pillar of Nepal's economy, offering substantial contributions to employment, foreign exchange earnings, and regional development. This research presents a data-driven approach to analyze and forecast Tourist arrivals in Nepal using advanced statistical and machine learning models. Historical tourism data spanning from 1992 to 2017 was obtained from the Ministry of Culture, Tourism and Civil Aviation (MoCTCA), the World Travel & Tourism Council (WTTC), and other survey-based sources.

The collected data underwent rigorous preprocessing, including cleaning, scaling, transformation, and decomposition, to ensure suitability for predictive modeling. Three forecasting models were implemented: Simple Linear Regression, Seasonal ARIMA (SARIMA), and Multi-layer Perceptron (MLP). The SARIMA model captured seasonal trends in monthly Tourist arrivals, while the MLP model integrated multivariate inputs—such as accessibility, accommodation, and healthcare infrastructure—for enhanced nonlinear forecasting. Performance was evaluated using metrics like RMSE and MSE.

Among the models, the MLP achieved the highest predictive accuracy, effectively modeling the complex relationships and patterns in the data. These results offer valuable insights for policymakers and tourism stakeholders to optimize planning, marketing, and infrastructure development strategies based on robust forecasts.

**Keywords:** Tourism, Data analysis, Prediction, Machine learning, Future planning

---

## 1. Introduction

### 1.1. Background

Tourism is one of the world's largest industries, not only due to the volume of international travelers and their economic impact but also because of its profound influence on national economies and local livelihoods (World Travel & Tourism Council, 2018). In the context of Nepal, tourism stands as the largest industry and the most significant source of foreign exchange earnings (World Travel & Tourism Council, 2017). According to the World Economic Forum (2017), Nepal ranked 103rd in the Travel & Tourism Competitiveness Index (TTCI). With eight of the world's ten highest peaks and rich cultural and geographical diversity, Nepal remains a popular destination for mountaineers, adventure seekers, and cultural explorers.

In the modern era, leveraging insights from data across sectors is essential for national development. To support such data-driven efforts, the Ministry of Culture, Tourism and Civil Aviation (MoCTCA) of the Government of Nepal annually publishes comprehensive tourism statistics (MoCTCA, 2019). However, in a globally competitive, data-centric environment, traditional statistical summaries are no longer sufficient. The integration of data science—an interdisciplinary field combining statistics, machine learning, and analytical techniques—is necessary for more profound understanding and forecasting.

This study proposes a web-based application that utilizes official tourism statistics to perform analytical modeling and predictive forecasting using machine learning techniques. Accurate planning is critical in the tourism industry due to the sector's sensitivity to political, environmental, and economic factors. Effective planning depends on robust analytical methods and accurate forecasts (Frechtling, 2001). Prior studies have demonstrated that applying advanced data analytics significantly contributes to the improvement and growth of the tourism sector (Marine-Roig & Anton-Clavé, 2015).

## **1.2. Objective**

The primary objective of this study is to analyze tourism data in Nepal and develop predictive models using machine learning techniques, specifically focusing on the Multi-layer Perceptron (MLP), Simple Linear Regression, and Seasonal Autoregressive Integrated Moving Average (SARIMA) models. The study aims to provide insights into Tourist arrivals, enabling stakeholders in the tourism sector to make informed decisions for better planning and resource allocation.

## **2. Literature Review**

The tourism sector in Nepal is influenced by various factors, including technological advancements, data analytics, and external global and local events. This literature review synthesizes prior research that supports the objectives of this study, specifically focusing on time series forecasting methods, the relevance of machine learning techniques, and the use of data science tools in tourism analytics. While previously studies have addressed the effects of primary disruptions like the 2015 earthquake and the COVID-19 pandemic, this study does not include data related to these events and instead concentrates on general historical patterns of Tourist arrivals.

### **Advancements in Technology and Data Analytics**

Recent technological developments have significantly transformed how tourism data is analyzed and applied. Big data analytics, in particular, has enabled researchers and tourism planners to understand traveler behavior, preferences, and patterns more effectively. Xiang and Fesenmaier (2018) emphasized the importance of integrating data analytics into tourism to enhance user experiences and stakeholder decision-making.

### **Structure and Nature of Tourism Data**

Tourism data is typically collected in time series format, where variables like monthly Tourist arrivals, revenue from tourism, or number of visits to specific locations are recorded chronologically. Sharma et al. (2019) demonstrated the value of such data in understanding tourism trends through appropriate visualizations, such as line graphs, bar charts, and pie charts. Frechtling (2001) noted that time series analysis is essential for identifying seasonal patterns, growth trends, and irregular fluctuations, all of which are critical for strategic planning in the tourism sector.

### **Forecasting Methods in Tourism**

Accurate forecasting of Tourist arrivals helps in better planning, resource allocation, and policy formulation. Classical statistical models like ARIMA (Autoregressive Integrated Moving Average) and its seasonal variant SARIMA have long been used to capture linear patterns and seasonality in tourism data. Adhikari et al. (2021) found that ARIMA models effectively predicted Tourist arrivals in Nepal, particularly in stable historical datasets. This research incorporates SARIMA to model the seasonal characteristics of monthly Tourist arrival data.

### **Impact of External Events on Tourism**

Studies such as those by Ghimire et al. (2020) and Khatri et al. (2021) have documented how external shocks like earthquakes and pandemics can drastically affect tourism flows. These works highlight the need for adaptive forecasting models that can accommodate unexpected disruptions. However, the present study does not incorporate such event-specific data and instead focused on patterns inherent in historical trends.

### **Machine Learning Approaches in Forecasting**

Machine learning methods have increasingly been used for tourism forecasting due to their ability to model complex and nonlinear relationships. Ahmed et al. (2010) demonstrated that Multi-layer Perceptron (MLP) neural networks outperform traditional statistical models in many time series forecasting scenarios. In this study, the MLP model is applied to multivariate data—including variables such as accessibility, accommodation, and medical infrastructure—to predict regional Tourist distributions.

### **Tools and Technologies for Tourism Forecasting**

The use of data science tools, especially programming environments like Python, has become essential for tourism analytics. Python offers powerful libraries such as Pandas for data handling, Matplotlib for visualization, and

TensorFlow for building neural network models (Raschka, 2015). These tools were used in this study for data preprocessing, visualization, and implementation of forecasting models, contributing to efficient and replicable analysis.

### 3. Methodology

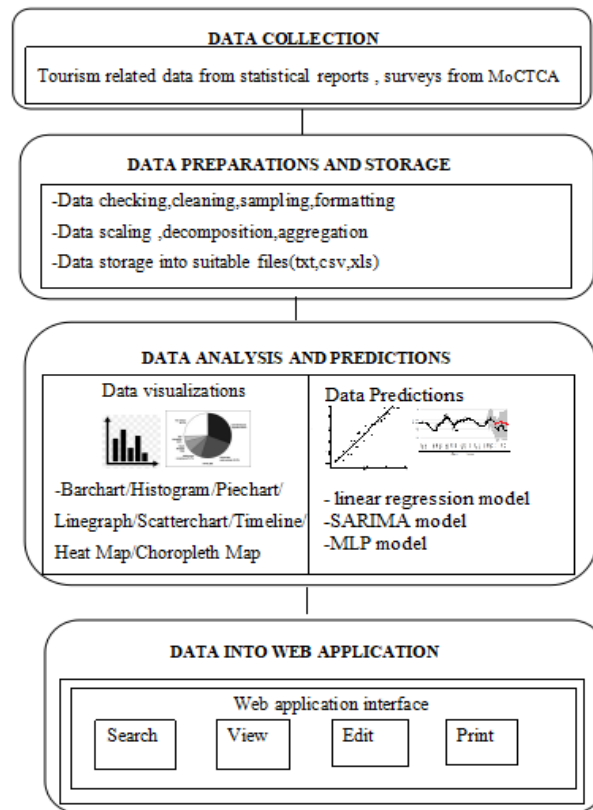


Figure 1: System Architecture

#### 3.1. System block diagram for analysis and prediction

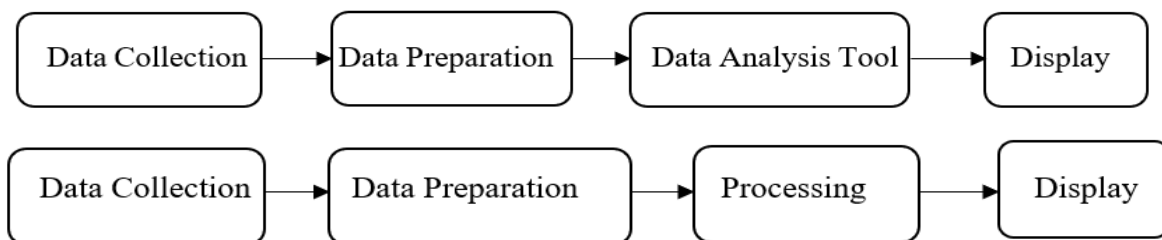


Figure 2: System block diagram

##### 3.1.1. Data Collections and Preparation

Data Sources:

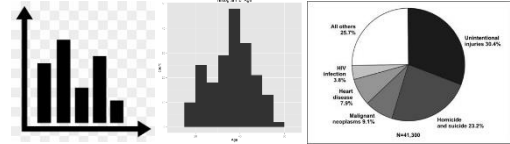
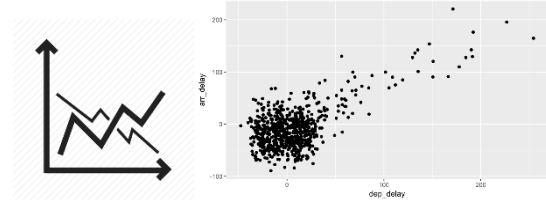
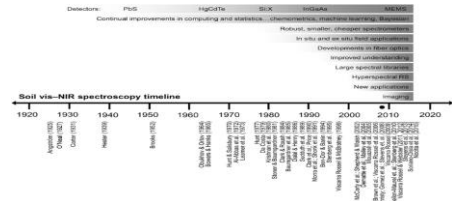
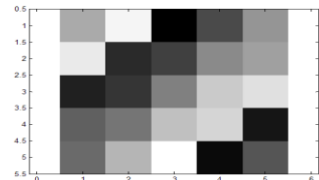
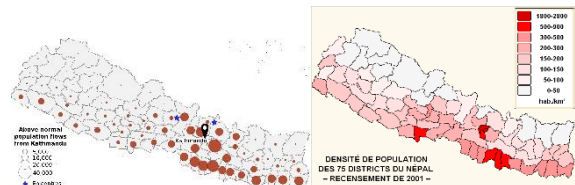
- Government of Nepal (MoCTCA), Tourism statistics report [3]
- Government of Nepal (MoCTCA), Civil Aviation report [21]
- World Travel & Tourism Council. The WTTC report: Travel and Tourism Economic Impact Nepal report's [2]
- Surveys report with Tourists about their experiences in Nepal [3]

The foremost step initiating this research is proper data collection and preparation. The preliminary data preparation tasks includes' data checking, cleaning, editing, sampling organizing, formatting into suitable forms (XML, csv), scaling, decomposition, aggregations and so on before being used in analysis and prediction model.

### 3.1.2. Data Analysis and Predictions

The first primary objective of the research to suitable data analysis of tourism data can be achieved using different statistical data presentations and visualization techniques. These can be achieved with the help of suitable data visualizations and analysis tools that use statistics through libraries in Python [17] which can be used in the proposed research. Some primary graphical techniques along with their usage to visualize data are:

Table 1: Graphical techniques

Diagrams	Example Usage
<p>Barchart/Histogram/Piechart</p>  <p>Linegraph/Scatter chart</p>  <p>Timeline</p>  <p>Heat Map</p>  <p>Scatter Map /Choropleth Map</p> 	<ul style="list-style-type: none"> <li>Tourist arrivals for different purposes <ul style="list-style-type: none"> <li>No. of visitors in different places</li> <li>No. of Tourist Tourists related accidents/incidents</li> </ul> </li> <li>No. of flights in different domestic airlines</li> </ul> <ul style="list-style-type: none"> <li>International and Domestic flight movements by month on the basis of gender/country</li> <li>Total Tourist arrival volume by month on the basis of gender/country/age group</li> <li>Gross foreign exchange earnings from tourism by fiscal year</li> </ul> <ul style="list-style-type: none"> <li>All available statistical records with highest and lowest value records on the basis of year</li> <li>Important toursim-related events/incidents</li> </ul> <ul style="list-style-type: none"> <li>Economic indicators values of hotels and restaurants on yearly</li> <li>Civil aviation indicators values on yearly basis</li> <li>No. of different types of Tourist industries</li> </ul> <ul style="list-style-type: none"> <li>No. of Tourists visit in primary places of Nepal</li> <li>Places in Nepal with different Tourist activities records</li> <li>Places in Nepal with different Tourist standard hotels/homestays</li> </ul>

### 3.1.3. Regression prediction using the simple linear regression

Usually, linear regression is an approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) . It finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variables.

A simple model function of our data is represented by equation  $y=b_0+b_1x$  which describes a line where y is the output variable we want to predict, x is the input variable we know and b0 and b1 are coefficients that we need to estimate that moves the line around and with slope b1 and y-intercept b0. In general, such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables, we call the unobserved deviations from the above equation the errors. Suppose we observe n data pairs and call them  $\{(x_i, y_i), i = 1, \dots, n\}$ . We can describe the underlying relationship between  $y_i$  and  $x_i$  involving this error term  $E_i$  by  $y_i=b_0+b_1x_i+E_i$

This relationship between the true write nothing /delete it underlying parameters  $\alpha$  and  $\beta$  and the data points is called a linear regression model. The goal is to find the best estimates for the coefficients to minimize the errors in predicting y from x. Simple regression is excellent, because rather than having to search for values by trial and error or calculate them analytically using more advanced linear algebra, we can estimate them directly from our data.

Let us introduce the following terms:

$\bar{x}$  and  $\bar{y}$  as the Average of the  $x_i$  and  $y_i$ , respectively

$Var(x)$ ,  $Cov(x, y)$  as the sample variance, sample covariance respectively.

We can start off by estimating the value for b1 as:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ or by using } b_1 = \frac{Cov(x,y)}{Var(x)} \dots\dots\dots \text{Equation (1)}$$

Finally, we can calculate b0 using b1 and some statistics from our dataset, as follows:

$$b_0 = \sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i \dots\dots\dots \text{Equation (2)}$$

Implementation of Simple linear regression:

A simple regression model approach matches the following dataset as it seemed to have linear relationships and consists of only few data.

Table 2: Data set for simple regression model approach

Start of fiscal year (A.D)	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Net foreign exchange earnings (NRs in million)	8523	9881.6	12167.8	12073.9	11717	8654.3	11747.7	18147.4	10463.8	9556	10125.3	18653.1	27959.8	28138.6	24610

Here given net foreign exchange earnings for each fiscal starting fiscal year we are interested in predicting the net foreign exchange earnings in the upcoming fiscal year start. This simple linear problem can be implemented in Python programming language with the use of the following libraries:

The following are the steps to implement and train simple linear regression models for the prediction problems.

**a. Calculation of Mean and Variance:**

The first step is to estimate the mean and the variance of both the input and output variables from the data. The mean of a list of numbers can be calculated by creating functions as:

$$\text{mean}(x) = \text{sum}(x)/\text{count}(x) \dots\dots\dots \text{Equation (3)}$$

Where, mean(x) gives the mean or Average value of x, sum(x) gives the sum of values of x and count(x) gives the no of x data values present.

Similarly, variance for a list of numbers can be calculated as:

$$\text{variance} = \text{sum}((x - \text{mean}(x))^2) \dots\dots\dots \text{Equation (4)}$$

**b. Calculation of Covariance:**

The covariance of two groups of numbers describes how those numbers change together. We can calculate the covariance between two variables as follows:

$$\text{covariance}(x, y) = \text{sum}((x(i) - \text{mean}(x)) * (y(i) - \text{mean}(y))) \dots\dots\dots \text{Equation (5)}$$

**c. Estimate Coefficients:**

We must estimate the values for two coefficients in simple linear regression. The first is b1 which can be estimated as:

$$\begin{aligned} b1 &= \text{sum}((x(i) - \text{mean}(x)) * (y(i) - \text{mean}(y))) / \text{sum}((x(i) - \text{mean}(x))^2) \quad \text{or} \\ b1 &= \text{covariance}(x, y) / \text{variance}(x) \dots\dots\dots \text{Equation (6)} \end{aligned}$$

Next, we need to estimate a value for b0, also called the intercept as it controls the starting point of the line Where it intersects the y-axis.

$$b0 = \text{mean}(y) - b1 * \text{mean}(x) \dots\dots\dots \text{Equation (7)}$$

**d. Prediction of the new values**

The simple linear regression model is a line defined by coefficients estimated from training data. Once the coefficients are estimated, we can use them to make predictions. The equation to make predictions with a simple linear regression model is as follows

$$y = b0 + b1 * x \dots\dots\dots \text{Equation (8)}$$

**3.1.4. Regression Prediction using Seasonal Autoregressive Integrated Moving Average (SARIMA) Model**

ARIMA, short for 'Auto-Regressive Integrated Moving Average' is actually a class of models that explains a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. Any non-seasonal time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models. Basically, ARIMA model is the combination of Autoregressive (AR) model, Integrated(I) model, Moving Average (MA) model. ARIMA model can be characterized by following notation below:

ARIMA (p, d, q)

Where, p is the order of the AR term, d is the number of differencing required to make the time series stationery and q is the order of the MA term.

The problem with plain ARIMA model is it does not support seasonality. Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. SARIMA model can be characterized by following notation below:

SARIMA (p, d, q) (P, D, Q) s

Where p, q, d are respective order of AR, MA and differencing of non-seasonal ARIMA model, **P** is seasonal autoregressive (AR) order, **D** is seasonal difference (I)order, **Q** is seasonal moving Average (MA)order and **S** is the number of time steps for a single seasonal period.

#### **Autoregressive Part (AR Part)**

A pure Auto-Regressive (AR only) model is one Where  $Y_t$  depends only on its own lags. That is,  $Y_t$  is a function of the 'lags of  $Y_t$ '. AR part of a time series  $Y_t$  is that the observed value depends on some linear combination of previously observed values up to a defined maximum lag (denoted p), plus a random error term  $\varepsilon_t$  and given as:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \varepsilon_t \dots \dots \dots \text{Equation (9)}$$

Where,  $Y_{t-1}$  is the lag1 of the series  $\beta_1$  is the coefficient of lag1,  $\alpha$  is the intercept term,  $\varepsilon_t$  is a random error term that the model estimates

#### **Moving Average Part (MA Part)**

A pure **Moving Average (MA only) model** is one Where  $Y_t$  depends only on the lagged forecast errors. MA part of a time series  $Y_t$  is that the observed value is a random error term plus some linear combination of previously random error terms up to a defined maximum lag (denoted q).

$$Y_t = \alpha + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \dots + \Phi_q \varepsilon_{t-q} \dots \dots \dots \text{Equation (10)}$$

Where the error terms are the errors of the autoregressive models of the respective lags. The errors  $\varepsilon_t$  and  $\varepsilon_{t-1}$  are the errors from the following equations:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_0 Y_0 + \varepsilon_t \dots \dots \dots \text{Equation (11)}$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + \dots + \beta_0 Y_0 + \varepsilon_{t-1} \dots \dots \dots \text{Equation (12)}$$

#### **The integration part (I Part)**

Time series are usually non-stationary and in order to achieve stationary the series has to be differenced. The process of differencing is known as integration part (I) and the order of differencing is denoted as d. Differencing removes the signals (the trend or seasonality) from the series so that series consists only the noise or the irregular component to be modeled. This can be expressed algebraically as:

$$\Delta^1 Y_t = Y_t - Y_{t-1} \dots \dots \dots \text{Equation (13)}$$

Using backshift operator B (Where  $By_t = y_{t-1}$ ,  $B^2 y_t = y_{t-2}$ , and so on) above equation

$$\Delta^1 Y_t = (1 - B)Y_t \dots \dots \dots \text{Equation (14)}$$

#### **ARIMA model Equation**

An ARIMA model is one Where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So, the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \dots + \Phi_q \varepsilon_{t-q} \dots \dots \dots \text{Equation (15)}$$

or it can be expressed as

$$\beta_p(B)\Delta^d Y_t = \Phi_q(B) \varepsilon_t \dots \dots \dots \text{Equation (16)}$$

Where  $\Delta^d$  is the non-seasonal difference operator, B is backshift operator (Where  $By_t = y_{t-1}$ ,  $B^2 y_t = y_{t-2}$ , and so on)

ARIMA model in words:

Predicted  $Y_t = \text{Constant} + \text{Linear combination Lags of } Y \text{ (up to } p \text{ lags)} + \text{Linear Combination of Lagged forecast errors (up to } q \text{ lags)}$

### Seasonal ARIMA Equation

SARIMA allows for the presence of seasonality in a series. This leads to the general seasonal ARIMA (p d q) s (P D Q) model, Where P, D and Q refer to the orders of the seasonal AR, I and MA parts of the model respectively and s refers to the number of periods in each season. This can be expressed algebraically as:

$$\beta_p(B)\theta_p(B^s)\Delta_s^d Y_t = \Phi_q(B)\theta_q(B^s) \varepsilon_t \dots \dots \dots \text{Equation (17)}$$

Where,  $\theta_p(B^s)$  is the seasonal AR operator,  $\Delta_s^d$  is the seasonal I operator,  $\theta_q(B^s)$  is the seasonal MA operator and s is the seasonal period.

### ACF and PACF plots

Statistical correlation summarizes the strength of the relationship between two variables. Pearson's correlation coefficient is a number between -1 and 1 that describes a negative or positive correlation respectively. A value of zero indicates no correlation. Given a pair of random variables (X, Y) the formula for Pearson's correlation coefficient ( $\rho$ ) is given by:

$$P_{(X,Y)} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \dots \dots \dots \text{Equation (18)}$$

Where,  $cov(X, Y)$  is the covariance of (X, Y),  $\sigma_X$  is the standard deviation of X and  $\sigma_Y$  is the standard deviation of Y

If we are interested in finding whether or to what extent there is a numerical relationship between two variables of interest, using their correlation coefficient will give misleading results if there is another, variable that is numerically related to both variables of interest. This misleading information can be avoided by controlling for the confounding variable, which is done by computing the partial correlation coefficient. Partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed.

Let us suppose three terms denoted by 1,2,3 (for  $y_t, y_{t+1}, y_{t+2}$ ).  $P_{13.2}$  is the correlation of  $y_t$  and  $y_{t+2}$  given (conditional on)  $y_{t+1}$ . The standard equation for partial correlation is given by:

$$P_{13.2} = \frac{\rho_{13} - \rho_{12}\rho_{32}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{32}^2}} \dots \dots \dots \text{Equation (19)}$$

Where,  $\rho_{ab}$  is correlation coefficient between two terms a and b.

We can calculate the correlation for time series observations with observations with previously time steps, called lags. Autocorrelation (ACF), also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay or lags. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies.

A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed. partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. This function plays an important role in data analysis aimed at identifying the extent of the lag in an auto-regressive models.

For example, for data below:

Table 3: No. of Tourist arrive in month

Month	No. of Tourists arrivals
1992-01	17451
1992-02	27489



Month	No. of Tourists arrivals
1992-03	31505
1992-04	30682
1992-05	29089
1992-06	22469
1992-07	20942
1992-08	27338
1992-09	24839
1992-10	42647
1992-11	32341
1992-12	27561
.....	.....
2017-12	82966

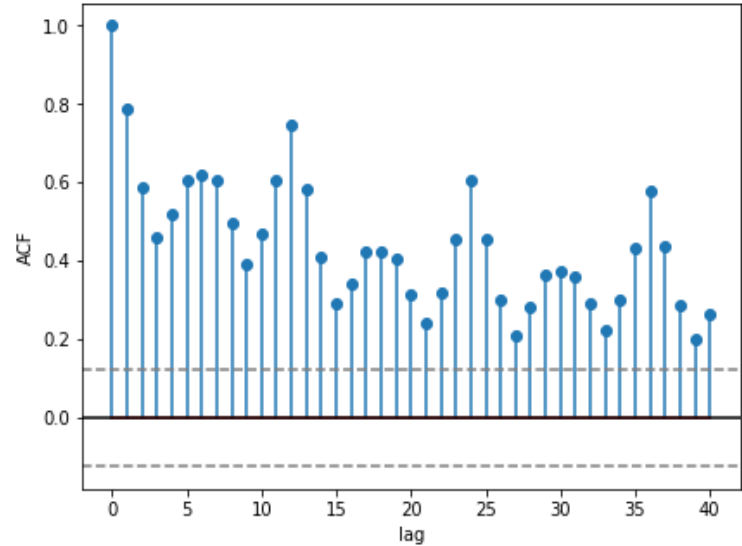


Figure 3: Plot of auto-correlation function

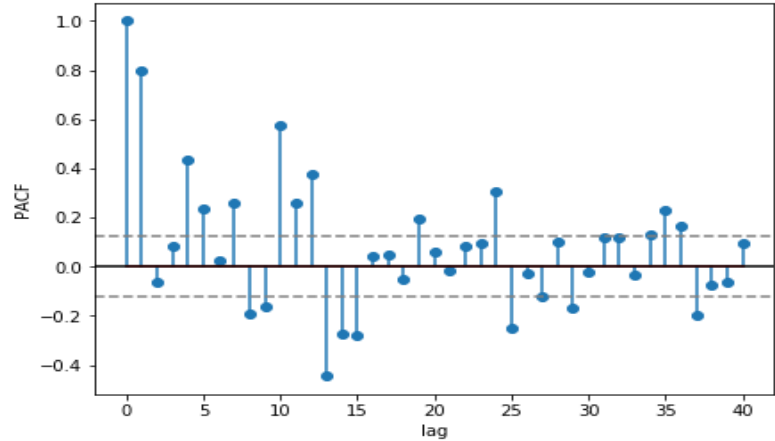


Figure 4: Plot of partial auto-correlation function

### Implementation of SARIMA Model:

A SARIMA model is appropriate for regression prediction of the following dataset as it seemed to have a seasonal pattern in the time series dataset as shown below:

Table 4: Seasonal pattern of number of tourist arrival

Month	No. of Tourists arrivals
1992-01	17451
1992-02	27489
1992-03	31505
1992-04	30682
1992-05	29089
1992-06	22469
1992-07	20942
1992-08	27338
1992-09	24839
1992-10	42647
1992-11	32341
1992-12	27561
2017-12	82966

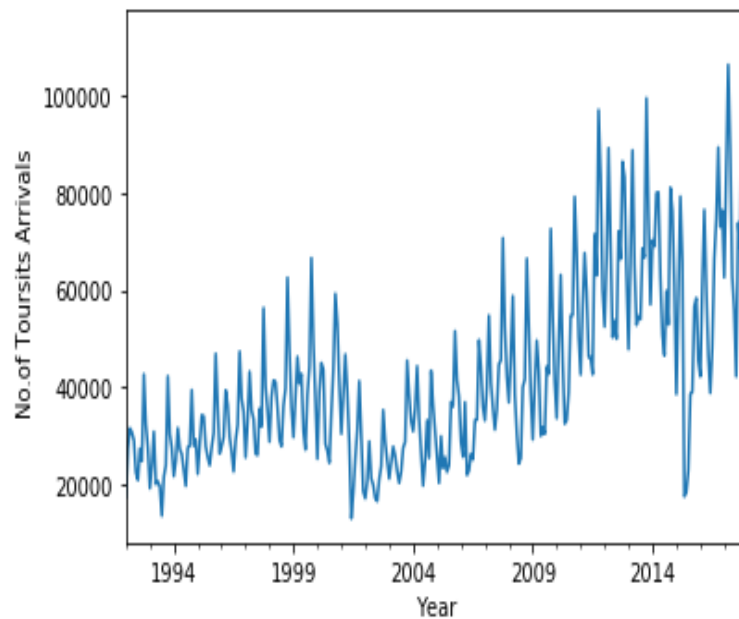


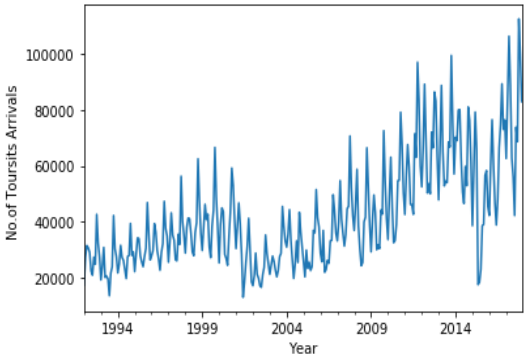
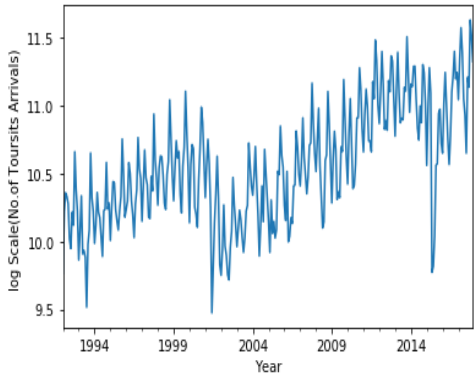
Figure 5: Plot of No. of Tourist arrive in month

These are the steps to implement SARIMA regression models for the given prediction problem:

#### i. Observation and transformation of time series data

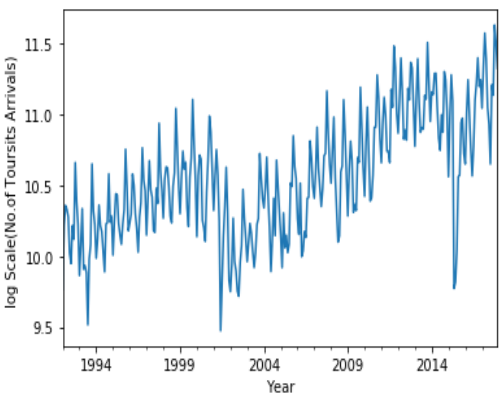
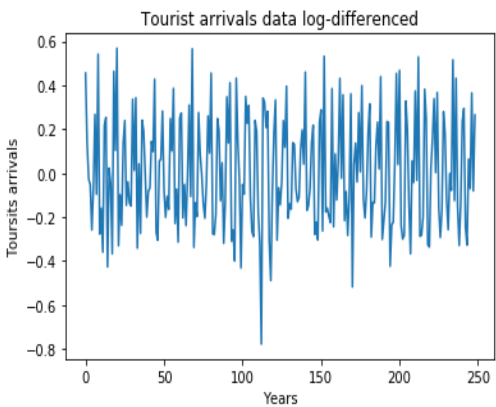
Plot the given time series in the original scale and observer its characteristics like trend, seasonality, stationarity. Also log transform the response if the seasonal variation is increasing with time.

Table 5: Transformation of time series data

Original Scale	Log Transformed Scale
	
-no stable trend	-log transformed values of no of Tourist arrivals
-non-stationary	- non-stationary
-seasonality present	-seasonality present

ii. Differencing of time series data

Since stationarity means that the statistical properties (mean, variance) of a process generating a time series do not change over time. Usually, for non-stationary time series, differencing is done to achieve stationary time series. Different order of differencing is done and it is checked whether the series is stationary or not.

Log transformed series	Log differenced series (1 <sup>st</sup> order differencing)
	

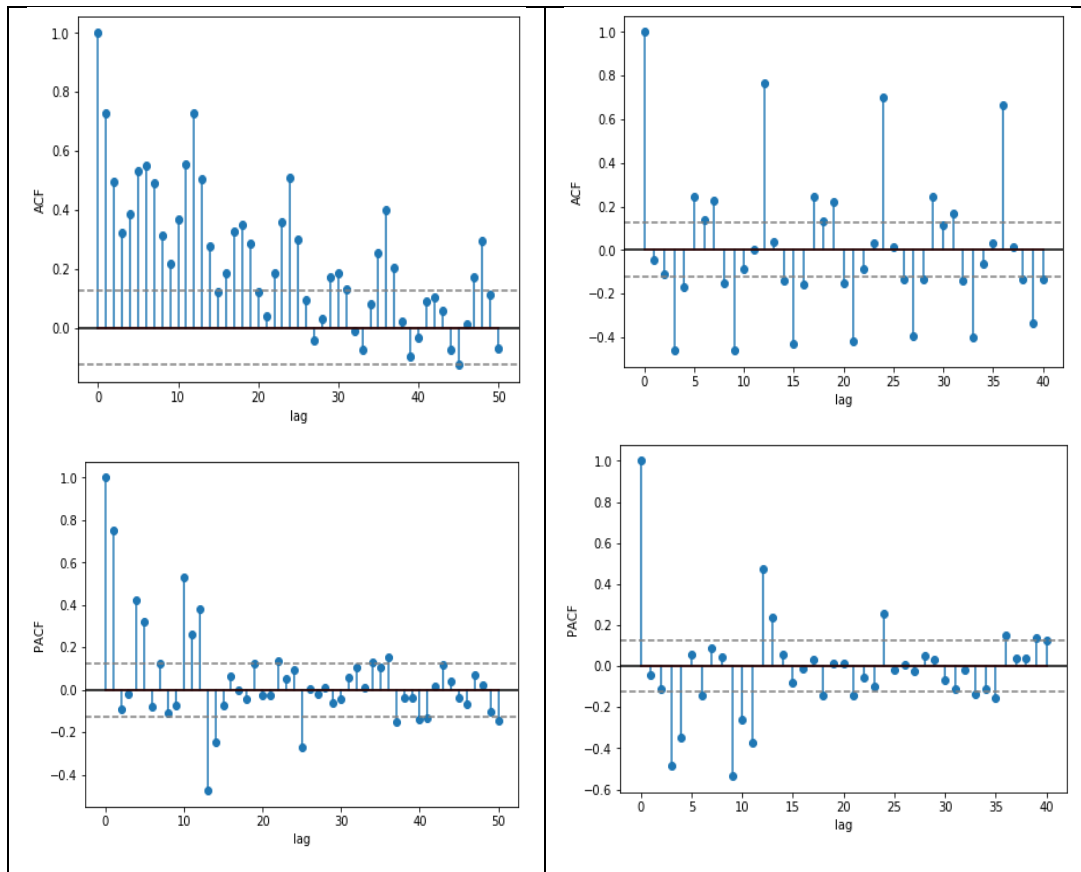


Figure 6: Plot of time series data

### iii. Split of the dataset into the Training set and Test set

This is the step Where a whole dataset is divided into test and train sets. Here for this problem, train to test set can be divided into ratio 4:1. ie 20 percent of dataset is used as a test set for model and the remaining as train set. Usually train set helps to fit or build a prediction model and test set is used to evaluate the performance of that model.

### iv. Identification of non-seasonal and seasonal level model

Examine ACF and PACF plots to tentatively identify nonseasonal level model.

#### Identifying the order of differencing

$d=0$  if the series has no visible trend or ACF at all lags is low.

$d \geq 1$  if the series has visible trend or positive ACF values out to a high number of lags.

if after applying differencing to the series and the ACF at lag 1 is  $-0.5$  or more negative the series may be over differenced.

If you find the best  $d$  to be  $d=1$  then the original series has a constant trend. A model with  $d=2$  assumes that the original series has a time-varying trend

#### Identifying the number of AR and MA terms

$p$  is equal to the first lag Where the PACF value is above the significance level.

$q$  is equal to the first lag Where the ACF value is above the significance level.

#### Identifying the seasonal part of the model:

$s$  is equal to the ACF lag with the highest value (typically at a high lag).

$D=1$  if the series has a stable seasonal pattern over time.

$D=0$  if the series has an unstable seasonal pattern over time.

Rule of thumb:  $d+D \leq 2$

$P \geq 1$  if the ACF is positive at lag  $s$ , else  $P=0$ .

$Q \geq 1$  if the ACF is negative at lag  $s$ , else  $Q=0$ .

Rule of thumb:  $P+Q \leq 2$

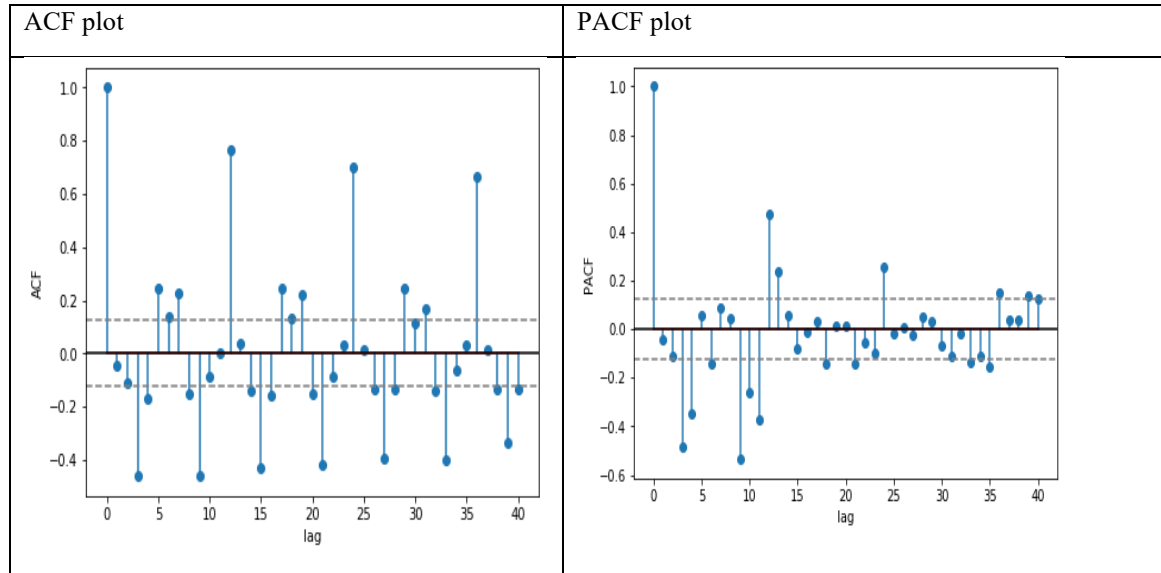


Figure 7: Plot of ACF and PACF

#### v. Creating SARIMA model

Combine models from Steps 4 to arrive at a tentative overall seasonal ARIMA model, i.e.  $ARIMA(p, d, q) \times (P, D, Q)$ . Where  $d$  &  $D$  are based on what differencing you used to achieve stationarity. Fit tentative model and look performance statistics, and the ACF/PACF of the residuals from the fit. Explore other models by changing parameters to achieve better model. During model parameters estimation using step 4 multiple SARIMA models were obtained and it was found that SARIMA (3,0,3) (2,1,0) [12] achieved better forecast accuracy.

#### vi. Evaluations of the model and further predictions

After creating the better SARIMA model for prediction, its performance is evaluated by calculating the error between the actual test set output and predicted output of the model. For this regression problem loss or error metric used is Root mean square error (RMSE) is used to evaluate the model. It was found from 80 percent trained model that 20 percentage test data were predicted with about 15172 RMSE error. Also using this model further one year no of Tourist arrivals was predicted as shown below:

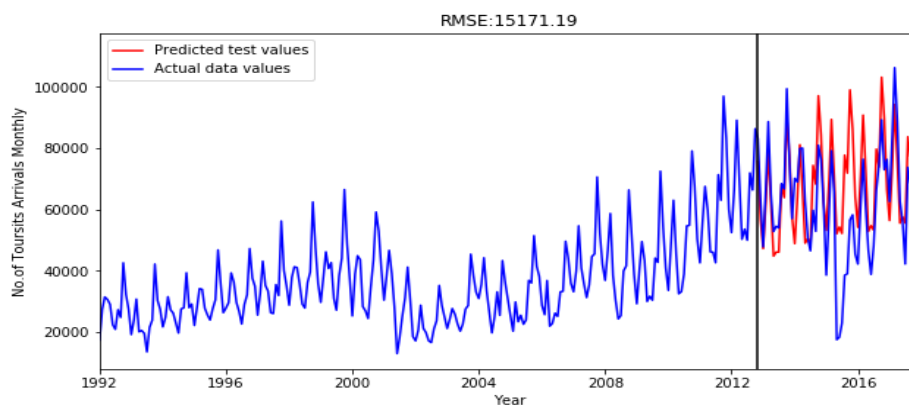
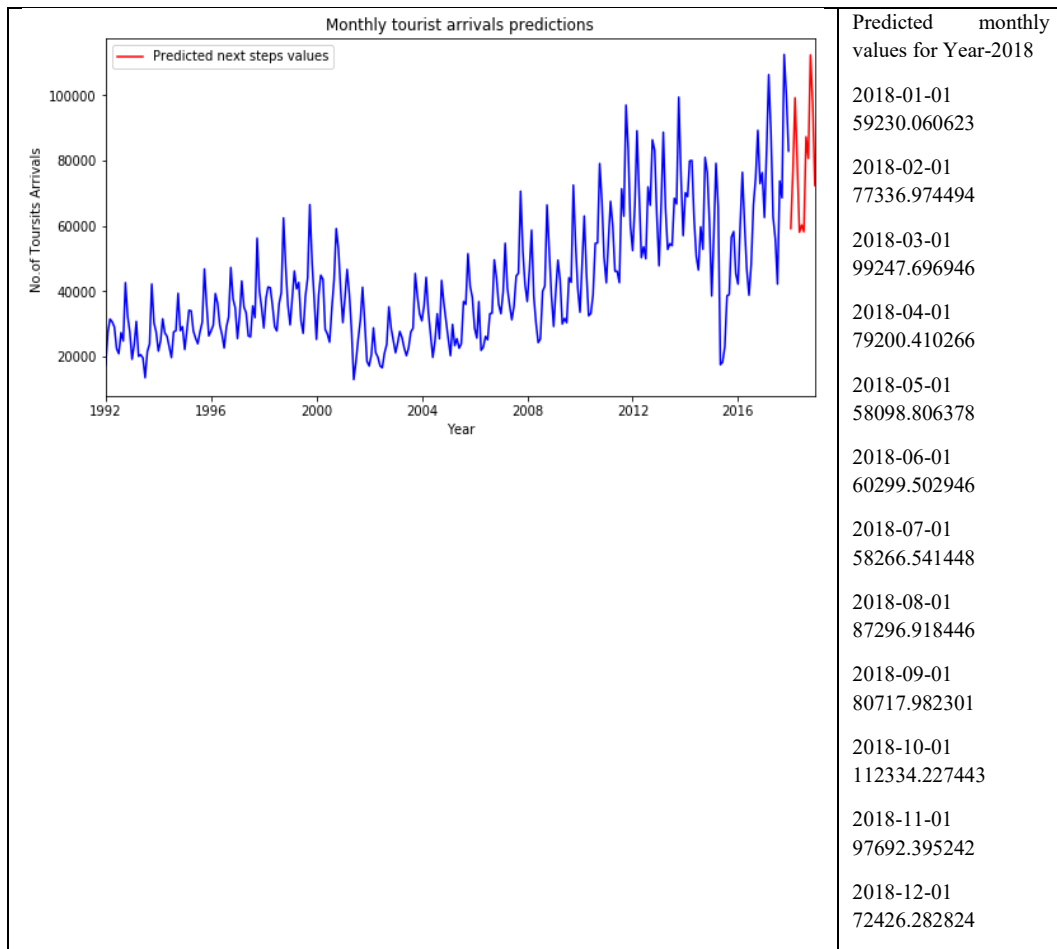


Figure 8: Plot of actual and predicted value

Table 6: Predicted monthly values for Year-2018



### 3.2. Model Selection

In this study, we selected the Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) models for forecasting Tourist arrivals in Nepal. The choice of these models is based on their proven effectiveness in handling time series data, particularly in the context of tourism forecasting.

- **Rationale for Selecting ARIMA:**

The ARIMA model is a well-established statistical method for time series forecasting. It is particularly effective for datasets that exhibit linear trends and seasonality. Previously research has demonstrated its utility in predicting Tourist arrivals, especially in stable conditions without significant external shocks. ARIMA's simplicity and interpretability make it a suitable choice for initial forecasting efforts.

- **Rationale for Selecting LSTM:**

The LSTM model, a type of recurrent neural network (RNN), is designed to capture long-term dependencies in sequential data. It is particularly effective for time series forecasting involving nonlinear relationships and complex patterns. Given the nature of tourism data, which can be influenced by various external factors (e.g., natural disasters, economic changes), LSTM is expected to provide more accurate predictions than traditional models like ARIMA.

- **Preliminary Testing:**

To evaluate the performance of the selected models, preliminary testing was conducted using historical tourism data. The dataset was divided into training and testing sets, with 80% of the data used for training

and 20% for testing. Both ARIMA and LSTM models were trained on the training set, and their predictions were compared against the actual values in the testing set.

- **Performance Metrics:**

The models were evaluated using performance metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Initial results indicated that the LSTM model outperformed the ARIMA model in terms of accuracy, particularly in capturing nonlinear trends in Tourist arrivals. This finding supports the decision to utilize LSTM for forecasting in this study.

### **3.3. Scope of Data Visualization**

In this study, data visualizations are essential tools for analyzing and interpreting tourism data in Nepal. We employ various graphical techniques—including scatter plots, bar charts, pie charts, and time-series graphs—to reveal patterns and trends in key tourism metrics such as Tourist arrivals, visitor purposes, activities, duration of stay, and foreign exchange earnings. The data are sourced from authoritative agencies, including the Ministry of Culture, Tourism and Civil Aviation (MoCTCA) and the World Travel & Tourism Council (WTTC).

These visualizations serve several important purposes:

- **Exploratory Analysis:** They provide clear and intuitive insights into the distribution and temporal dynamics of tourism data, helping to identify seasonal trends, regional differences, and shifts in Tourist behavior.
- **Model Validation:** Visual comparisons between actual data and predictions from forecasting models (ARIMA, LSTM, MLP) allow for an accessible assessment of model accuracy and effectiveness.
- **Stakeholder Communication:** Designed for diverse audiences such as policymakers, tourism officials, and researchers, the visualizations translate complex data into understandable formats that support informed decision-making.

#### **Limitations:**

The visualizations primarily reflect aggregated quantitative data over the available range of years ([insert years]). They do not capture qualitative factors like Tourist satisfaction or real-time behavioral changes, nor do they account for sudden external shocks such as natural disasters or pandemics.

By clearly defining the scope and purposes of the data visualizations, this study ensures transparent interpretation while reinforcing the validity and practical relevance of its analytical findings.

### **3.4. Regression prediction using MLP Neural Network**

Multi-layer perceptron is the classical type of neural network which comprised of one or more layers of neurons. Data is fed to the input layer, there may be one or more hidden layers providing levels of abstraction, and predictions are made on the output layer, also called the visible layer. They are suitable for regression prediction problems Where a real-valued quantity is predicted given a set of inputs. Data is often provided in a tabular format, such as you would see in a CSV file or a spreadsheet.

Another primary objective of the research to suitable quantitative forecasting can be achieved through machine learning approach using Multi-layer Perceptron (MLP) Model. The MLP model will learn a function that maps a sequence of past observations as input to an output observation. As such, the sequence of observations must be transformed into multiple examples from which the model can learn. The way MLP learns the training set is by supervised learning process Where all labeled data i.e. input and output pair is given to model and it learns from it and finally evaluates its working by testing against test sets. A simple model of MLP with one input layer, one hidden layer and an output layer can be used for prediction. To evaluate the predictive model some recent samples in time series data can be taken for testing purposes calculating root mean square error. A simple step's to training the MLP model is given below

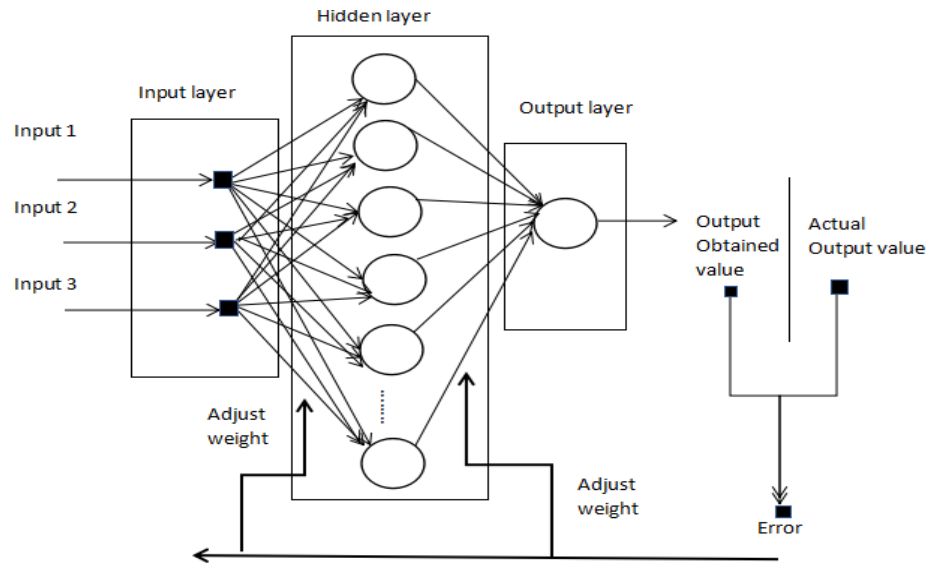


Figure 9: Multi-layer Perceptron Model

### Training the neural network model:

STEP 1: Randomly initialize the weights to small numbers close to zero.

STEP 2: Input the first observation of your dataset in the input layer, each feature in one input node.

STEP 3: Forward-Propagation: from left to right, the neurons are activated in a way that the impact of each neuron's activation is limited by weights. Propagate the activation until getting the predicted result.

STEP 4: Compare the predicted result to the actual result. Measure the generated error.

STEP 5: Back-Propagation: from right to left, the error is back-propagated. Update the weights according to how much they are responsible for the error. The learning rate decides by how much we update the weights.

STEP 6: Repeat Steps 1 to 5 and update the weights after each observation.

STEP 7: When the whole training set is passed through the network that makes an epoch. Finally redo more epochs.

### Implementation of MLP Model:

An MLP model approach matches the following dataset as it seemed to have a non-linear relationship and is multi-variate.

Table 7: Data of different Tourist place in the year 2011

Tourist Places	Mustang	Lower Dolpa	Upper Dolpa	Kanchanjunga	Manaslu	Koshi Tappu wildlife reserv	Parsa Wildlife Reserve
Year	2011	2011	2011	2011	2011	2011	2011
No. of other tourist attraction spots nearby	5	5	4	4	5	2	2
No. of available major tourist activities nearby	3	3	3	3	3	4	3
Main purpose of visit	treeking	treeking	treeking	treeking&Mountaineeri	treeking &Mountain	holiday/Pleasure	holiday/Pleasure
Accessibility status	Good	Poor	Poor	Poor	Poor	Better	Better
Accomodation status	Fair	Fair	Poor	Fair	Fair	Fair	Better
health services status	Good	Poor	Poor	Poor	Poor	Better	fair
Percentage of tourist arrival	0.400698	0.109750548	0.053924465	0.080275463	0.382089471	0.024585209	0.001901618



From the given dataset we can separate the dependent and independent variables as:

Table 8: Input and output with data of accessibility, accommodation, health and medical

Inputs (independent variables)	Output (dependent variable)
1.Year	1.Percentage out of total Tourist arrivals in that place
2.Number of Tourist attraction spots	
3.Number of Tourist activities available	
4.Main Purposes of visit	
5.Accessibility status	
6.Accommodation status	
7.Health services status	

Table 9: List of primary Tourist activities in Nepal

	POOR	FAIR	GOOD	BETTER	
Table List	<b>ACCESSIBILITY</b> References: -Department of Roads of Nepal(map) -Civil Aviation Authority of Nepal (CAAN)(map)	-Only local track, trials roads	-Only graveled or secondary roads	-Metaled primary roads or feeder road -Railways -National domestic airports	9: of
y es visit epal Table	<b>ACCOMODATION</b> References: -Hotel association Nepal(records)	-Local shops/teahouse and simple homestays	-Local hotels and lodges -Well managed or tourism-oriented homestays and guest house	-Tourist standard hotels -Tourist class lodge -Registered resorts	9: of in
	<b>HEALTH &amp; MEDICAL</b> References: -Ministry of health and population(map)	-Only simple sub-health post and health posts	-Primary health care center -community hospital	-Private clinics -Private small hospitals -District hospitals	

#### List of primary Tourist activities in Nepal

- Mountain climbing or Mountaineering
- Trekking/hiking
- Scenery, birds, animals watching /Photography
- Mountain flight
- Rock Climbing
- Rafting/kayaking/canyoning/boating
- Hot air Ballooning
- Bungy jumping
- Paragliding
- Mountain Biking
- Bicycle/Horse riding
- Jungle safari /Elephant riding/hunting
- Indoor Enjoyment
- Meditation /religious activities

Table 9: Primary purpose of visit in Nepal

#### List of primary purposes of visit in Nepal

- Holiday/Pleasure (includes: indoor enjoyment, photography, rafting, bungy jumping, camping, paragliding, biking, jungle safari etc.)

- Trekking/hiking (includes: visiting along primary trekking and hiking route's)
- Mountaineering (includes: climbing primary route allowed mountain's)
- Mountaineering and trekking (include both mountaineering and trekking)
- Official (includes visit for official or government purposes)
- Business (includes visit for business research, operations or investments)
- Conference/Conventions (includes visit during special conventions or conferences)

Here given independent variables for each year we are interested in predicting the percentage out of total Tourist arrivals in that place. This multi-variate regression problem can be solved using Python programming language with the use of following libraries:

These are the steps to implement and train MLP regression models for the given prediction problems:

#### a) Import of the dataset and separation dependent and independent variables

This is the first step Where dataset is imported as csv file and stored using panda's data frame. Now this data frame can be separated into input /output pairs ie independent and dependent variables using data frame dissects as:

Table 11: Separation dependent and independent variables

X								Y
Year	No. of spots	No. of activities	Purposes of Visit	Accessibility	Accommodation	Health & Medical	% Arrival	
2008	2	3	Holiday/Pleasure	Fair	Good	Poor	1.54291743e+01	
2008	4	3	Holiday/Pleasure	Poor	Fair	Poor	2.19195388e+00	
2008	4	3	Holiday/Pleasure	Fair	Fair	Good	3.921810000e-04	
2009	2	4	Holiday/Pleasure	Better	Fair	Better	1.49193442e+01	
2009	2	3	Holiday/Pleasure	Better	Fair	Fair	1.96088902e+03	
2009	2	4	Holiday/Pleasure	Better	Fair	Poor	1.69978489e+01	
2009	2	4	Holiday/Pleasure	Better	Good	Poor	2.69755488e-01	
2009	5	5	Holiday/Pleasure	Better	Better	Good	1.66169417e+01	
2009	3	5	Holiday/Pleasure	Better	Fair	Good	2.45852099e-02	
2009	4	3	Holiday/Pleasure	Good	Better	Fair	2.65563949e+01	

#### b) Encoding and labeling the categorical data

In this step we can see those categorical variables in X set which should be converted into proper numeric format before applying to input of MLP model. So, data labeling and encodings are done to convert the independent variables into following numeric formats that is suitable during computations:

Table 12: Encoding and labeling the categorical data

X										Y
Purposes encoded	of	visit	Year	No. of spots	No. of activities	of	Accessibility	Accommodation	Medical	Percent arrivals
0.0	1.0	0.0	2008.0	2.0	3.0		2.0	3.0	1.0	1.54291743e+01
0.0	1.0	0.0	2008.0	4.0	3.0		1.0	2.0	1.0	2.19195388e+00

0.0	1.0	0.0	2008.0	4.0	3.0	2.0	2.0	3.0	3.921810000e-04
0.0	1.0	0.0	2009.0	2.0	4.0	4.0	2.0	4.0	1.49193442e+01
0.0	1.0	0.0	2009.0	2.0	3.0	4.0	2.0	2.0	1.96088902e+03
0.0	1.0	0.0	2009.0	2.0	4.0	4.0	2.0	1.0	1.69978489e+01
0.0	1.0	0.0	2009.0	2.0	4.0	4.0	3.0	1.0	2.69755488e-01
0.0	1.0	0.0	2009.0	5.0	5.0	4.0	4.0	3.0	1.66169417e+01
0.0	1.0	0.0	2009.0	3.0	5.0	4.0	2.0	3.0	2.45852099e-02
0.0	1.0	0.0	2009.0	4.0	3.0	3.0	4.0	2.0	2.65563949e+01

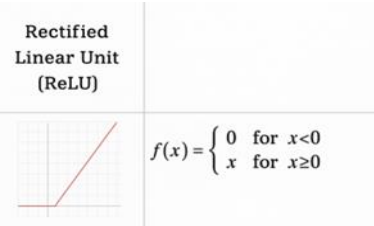
### c) Split of the dataset into the Training set and Test set

This is the step Where a whole dataset is divided into test and train sets. Here for this problem, train to test set can be divided into ratio 4:1.ie 20 percent of dataset is used as a test set for model and the remaining as train sets (X\_train and Y\_train). Usually train set helps to fit or build a prediction model and test set is used to evaluate the performance of that model.

### d) Creating regression model

This is the step Where actual model is created from the train set of the available dataset. Following provides the summary of MLP regression model that is to be trained later:

Table 13: Creating regression model

LAYERS	NUMBERS	ACTIVATION FUNCTION USED
Input layer	No. of inputs=9	-None
Hidden layer	-No. of hidden layers=2 -No. of neurons or nodes in each layer=20	-Rectifier linear unit function (Relu) if input > 0: return input else: return 0  <div style="text-align: center;">  </div>
Output layer	No. of outputs=1 No. of output node=1	-None

### e) Fitting the MLP model to the Training set

After the prediction model is created or designed it should be trained enough to make ready for predictions. The model is fitted by providing the train sets of X and y data parts. The summary of fitting the created model to training set is given below:

Table 14: Fitting the MLP model to the Training set

Train sets (X_train and y_train)	Batch Size Used	Epochs
About 80 percent of total dataset	10	300

### f) Evaluations of the model and further predictions

After fitting the regression model, its performance is evaluated by calculating the error between the actual test set output and predicted output of the model. For this regression problem loss or error metric used is mean square error (MSE) is used to evaluate the model. Lower its value better is the prediction power of the model. It was found that from multiple tests on trend model the min-square error found in the range 4-6 value.

After evaluation of model and identification of its strength of prediction, it can be used for further prediction from given inputs X or independent variables. Here's a quick prediction summary for ranking of next top Tourist destinations in Nepal:

Table 15: Prediction of next top Tourist destinations in Nepal

Location	Year	No. of Tourist spots	No. of Tourist activities	Purposes of visit	Accessibility status	Accommodation status	Health services status	Percent of Tourist arrivals predicted
Dhulikhel	2019	6	4	Holiday/Pleasure	Better	Better	Good	21.879255294799805
Bandipur	2019	4	4	Holiday/Pleasure	Good	Good	Fair	2.0926218032836914
Helambu	2019	4	3	trekking	Good	Better	Fair	1.0644960403442383
Taplejung	2019	5	3	Treeing Mountaineering	Poor	Fair	Poor	0.5248016119003296
Gokyo Valley	2019	4	3	trekking	Good	Good	Fair	0.13958019018173218

### 3.5. Model Comparison

In this section, we compare the performance of the Autoregressive Integrated Moving Average (ARIMA) model and the Long Short-Term Memory (LSTM) model with other forecasting models commonly used in tourism demand forecasting. This comparison aims to elucidate the advantages and limitations of each approach, thereby justifying the selection of ARIMA and LSTM for this study.

#### 3.5.1. ARIMA Model

The ARIMA model is a widely used statistical method for time series forecasting, particularly effective for linear data patterns. It combines autoregressive (AR) and moving Average (MA) components, making it suitable for datasets that exhibit trends and seasonality. Previously studies, such as those by Adhikari et al. (2021), have demonstrated the effectiveness of ARIMA in predicting Tourist arrivals, especially in stable conditions without significant external shocks.

#### Strengths:

- Simplicity and interpretability.
- Effective for short-term forecasting when data is stationary.
- Well-suited for linear relationships in time series data.

**Limitations:**

- Assumes linearity, which may not capture complex patterns in data.
- Requires the data to be stationary, necessitating differencing and transformation.

**3.5.2. LSTM Model**

The LSTM model, a type of recurrent neural network (RNN), is designed to capture long-term dependencies in sequential data. It is particularly effective for time series forecasting involving nonlinear relationships and complex patterns. Research by Khatri et al. (2021) has shown that LSTM models can outperform traditional statistical methods in scenarios where data exhibits volatility and nonlinearity.

**Strengths:**

- Capable of modeling complex, nonlinear relationships.
- Handles large datasets with multiple features effectively.
- Learns from historical data to make predictions, adapting to changes over time.

**Limitations:**

- Requires a larger amount of data for training compared to traditional models.
- More complex and less interpretable than ARIMA.

**3.5.3. Comparison with Other Models**

In addition to ARIMA and LSTM, other forecasting models such as Seasonal Decomposition of Time Series (STL) and Exponential Smoothing State Space Model (ETS) have been employed in tourism forecasting.

- STL is effective for decomposing time series data into seasonal, trend, and residual components, making it useful for understanding underlying patterns. However, it may not perform as well in predicting future values compared to LSTM, especially in the presence of nonlinear trends.
- ETS models are advantageous for their simplicity and effectiveness in capturing exponential trends. However, they may struggle with complex seasonal patterns that LSTM can handle more adeptly.

**3.5.4. Summary of Findings**

The comparative analysis indicates that while ARIMA is effective for linear time series data, LSTM models excel in capturing complex, nonlinear patterns, making them suitable for tourism forecasting. The choice of model should be guided by the specific characteristics of the data and the forecasting objectives. Future research could explore hybrid models that combine the strengths of both statistical and machine learning approaches to enhance forecasting accuracy.

**4. System Design**

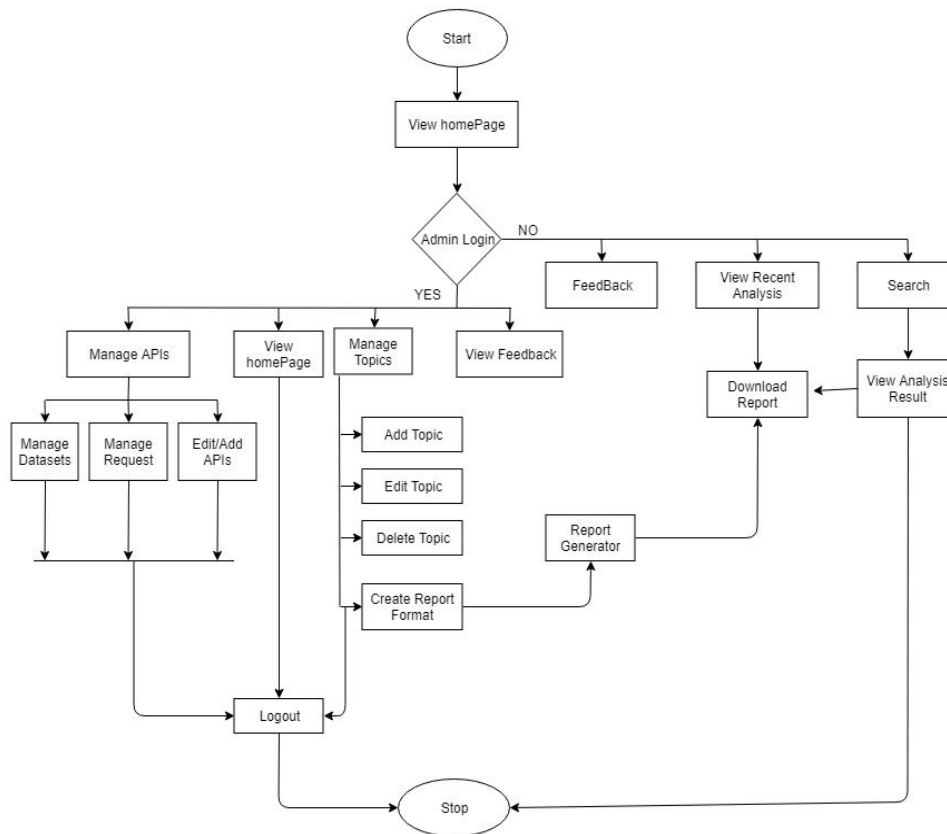


Figure 10: System flow diagram

## 5. Result and Analysis

### 5.1. Testing

#### 5.1.1. Unit Testing

Unit testing refers to the process of testing modules against the detailed design. The inputs to unit testing are the successfully compiled modules from the coding process. These are assembled during unit testing to make the largest units, i.e. the components of architectural design.

Testing has been performed in each phase of research design and coding. The module interface is tested to ensure that information properly flows into and out of the program unit under testing. The local data structure is examined to ensure that data stored temporarily maintains its integrity during all steps in an algorithm's execution. And finally, all error-handling paths are tested.

#### 5.1.2. Integration Testing

Integration testing is the phase in software testing in which individual software modules are combined and tested as a group. **Integration testing** is a level of software testing Where individual units are combined and tested as a group. The purposes of this level of testing are to expose faults in the interaction between integrated units. Test drivers and test stubs are used to assist in integration testing. Integration testing is conducted to evaluate the compliance of a system or component with specified functional requirements. It occurs after unit testing and before validation testing. Integration testing takes as its input modules that have been unit tested, groups them in larger aggregates, applies tests defined in an integration test plan to those aggregates, and delivers as its output the integrated system ready for system testing.

#### 5.1.3. System Testing

System testing is testing conducted on a complete integrated system to evaluate the system's compliance with its specified requirements. System testing takes, as its input, all of the integrated components that have passed integration testing. The purposes of integration testing are to detect any inconsistencies between the units that are

integrated together. System testing process is concerned with finding errors that results from unanticipated interactions between sub-systems and system components. Once source code has been generated, software must be tested to uncover (and correct) as many errors as possible before delivery to customers. Our goal is to design a series of test cases that have a high likelihood of finding errors. System testing seeks to detect defects both within the "inter-assemblages" and also within the system as a whole. The actual result is the behavior produced or observed when a component or system is tested.

System testing is performed on the entire system in the context of either functional requirement specifications (FRS) or system requirement specification (SRS), or both. System testing tests not only the design, but also the behavior and even the believed expectations of the customer.

Model testing:

Table 16: Testing for Simple Linear Regression Model

Case			Process	Result
<b>Dataset: YEAR (2012-2016)</b>			Identifying Linear regression equation: $Y = -4478336.394805195 + 2244.041558441558X$	Predicted
Year		Actual		36675.221
2012		34210.6		38919.263
2013		46374.9		41163.303
2014		53428.8		43407.345
2015		41765.4		45651.3870
2016		58526.9		

$$Y = -4478336.394805195 + 2244.041558441558X$$

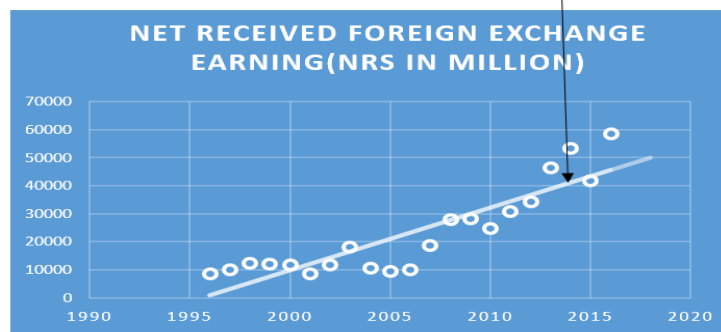


Figure 11. Net Received Foreign Exchange Earnings from Tourism (2012-2016).

As illustrated in Figure 12, the net received foreign exchange earnings from tourism in Nepal showed significant fluctuations from 2012 to 2016. There was a steady increase in earnings from 2012 to 2014, indicating a growing influx of Tourists during this period. However, a decline in earnings in 2015 suggests potential challenges faced by the tourism sector, possibly due to external factors such as natural disasters or geopolitical issues. The recovery in 2016 reflects a rebound in Tourist arrivals, highlighting the resilience of the tourism industry in Nepal.

The trends observed in Figure 12 align with the predictive models developed in this study, reinforcing the importance of accurate forecasting in understanding and responding to fluctuations in tourism demand.

Table 17. Result of Simple Linear Regression Model

Year	Actual	Predicted
2012	34210.6	36675.221

Year	Actual	Predicted
2013	46374.9	38919.263
2014	53428.8	41163.303
2015	41765.4	43407.345
2016	58526.9	45651.3870

Table 18. Forecasted Data 2017 & 2018

Year	Forecasted
2017	47895.429
2018	50139.470

Table 19: Testing for SARIMA Model

Case	Process	Result
Monthly data of year 2017		2017-01-01 56410.815408
2017-02 84061	Identifying SARIMA model parameters	2017-02-01 73766.262696
2017-03 106291	SARIMA(2,0,3)(2,1,0)[12]	2017-03-01 94300.221202
2017-04 88591		2017-04-01 75569.492869
2017-05 62773		2017-05-01 55572.991131
2017-06 55956		2017-06-01 57552.560166
2017-07 42240		2017-07-01 55608.270074
2017-08 73778		2017-08-01 83756.635899
2017-09 68634		2017-09-01 77339.780661
2017-10 112492		2017-10-01 108273.349913
2017-11 99804		2017-11-01 94110.867153
2017-12 82966		2017-12-01 69684.201630

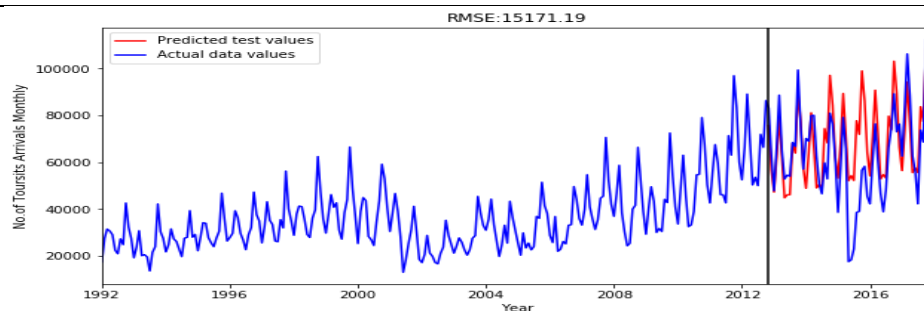


Figure 12. Result of SARIMA model I



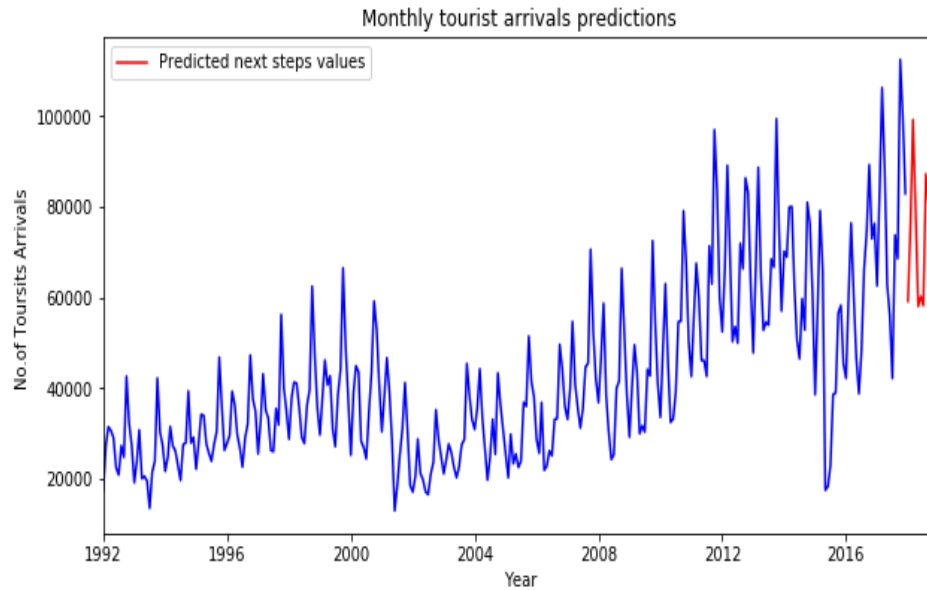


Figure 13: Result of SARIMA model II

Table 20: Testing for Multi-layer Perceptron Model

Case	Process	Result
Dhorpatan Hunting Reserve,2008,2,4, holiday/Pleasure, Fair, Fair, Poor ,0.010993909	Backpropagation learning algorithm	0.0718440961837769
Chitwan National Park,2008,5,5, holiday/Pleasure, Better, Better, Good ,16.53543937		18.303565979003906

Table 21: Result of MLP model

Places	Year	Number of Tourist attraction spots	Number of Tourist activities available	Main Purposes of visit	Accessibilit y status	Accommod ation status	Health services status	% Tourist arrival
Dhulikhel	2019	6	4	Holiday/Pleasure	Better	Better	Good	21.87925529 4799805
Helambu	2019	4	3	Trekking	Good	Better	Fair	1.064496040 3442383
Gokyo Valley	2019	4	3	Trekking	Good	Good	Fair	0.139580190 18173218
Taplejung	2019	5	3	Treeing & Mountaineering	Poor	Fair	Poor	0.524801611 9003296
Bandipur	2019	4	4	Holiday/Pleasure	Good	Good	Fair	2.092621803 2836914

## 6. Discussion

### 6.1. Interpretation of Results

The analysis conducted in this study revealed significant insights into the forecasting of Tourist arrivals in Nepal. The results indicated that the Multi-layer Perceptron (MLP) model outperformed both the Simple Linear Regression and Seasonal Autoregressive Integrated Moving Average (SARIMA) models in terms of accuracy, as evidenced by lower Root Mean Square Error (RMSE) values. This finding aligns with existing literature that suggests machine learning models, particularly neural networks, are better suited for capturing complex, nonlinear patterns in time series data. The ability of the MLP model to learn from historical data and adapt to changing trends makes it a valuable tool for predicting tourism demand.

### **6.2. Implications for Tourism Stakeholders**

The implications of these findings are profound for various stakeholders in the tourism sector. For government agencies, accurate forecasting of Tourist arrivals can facilitate better resource allocation and infrastructure development, ensuring that the needs of Tourists are met effectively. Travel agencies can leverage these predictions to tailor their marketing strategies, optimizing their outreach during peak seasons and enhancing customer satisfaction. Additionally, local businesses can use these insights to prepare for fluctuations in demand, allowing them to manage inventory and staffing levels more efficiently.

### **6.3. Limitations of the Study**

Despite the valuable insights gained, this study is not without limitations. One significant limitation is the reliance on historical data, which may not fully account for sudden changes in tourism patterns due to unforeseen events, such as natural disasters or global pandemics. Furthermore, the models used in this study primarily focused on quantitative data, potentially overlooking qualitative factors that influence Tourist behavior. The MLP model, while effective, requires a substantial amount of data for training, which may not always be available in emerging tourism markets.

### **6.4. Recommendations for Future Research**

Future research should consider exploring additional forecasting models, such as ensemble methods that combine the strengths of ARIMA and machine learning techniques to improve accuracy. Incorporating external variables, such as economic indicators, social media sentiment, and geopolitical factors, could provide a more comprehensive understanding of the dynamics influencing Tourist arrivals. Furthermore, longitudinal studies that track changes over time could enhance the robustness of predictions and provide more profound insights into trends in the tourism sector.

## **7. Conclusion**

In conclusion, this study successfully demonstrates the application of machine learning techniques, particularly the Multi-layer Perceptron (MLP), in analyzing and predicting Tourist arrivals in Nepal. The findings indicate that the MLP model outperforms traditional forecasting methods, such as Simple Linear Regression and SARIMA, in terms of accuracy. This research highlights the importance of leveraging data-driven approaches in the tourism industry, allowing stakeholders—including government agencies, travel companies, and local businesses—to enhance their strategic planning and operational efficiency. By adopting these advanced forecasting techniques, the tourism sector in Nepal can better adapt to changing market conditions and ensure sustainable growth, ultimately improving the overall Tourist experience.

## **References**

- Adhikari, T., Gautam, A., & Bhattarai, K. (2021). The impact of COVID-19 on tourism in Nepal: A case study of the tourism sector. *Journal of Tourism and Hospitality Management*, 9(1), 1–15. <https://doi.org/10.17265/2328-2169/2021.01.001>
- Ahmed, N. K., Atiya, A. F., El Gayar, N., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5–6), 594–621. <https://doi.org/10.1080/07474938.2010.481556>
- Civil Aviation Authority of Nepal. (2019). *Annual report*. <http://caanepal.gov.np/>
- Dorffner, G. (1996). Neural networks for time series processing. *Neural Network World*, 6(4), 447–468.
- Frechtling, D. C. (2001). *Forecasting tourism demand: Methods and strategies*. Butterworth-Heinemann.
- Ghimire, R., & Kafle, K. (2020). The impact of the 2015 earthquake on tourism in Nepal: A review of the literature. *Tourism Management Perspectives*, 35, 100689. <https://doi.org/10.1016/j.tmp.2020.100689>
- Khatri, D. B., & Shrestha, S. (2021). Post-pandemic recovery strategies for tourism in Nepal: A data-driven approach. *Nepalese Journal of Statistics*, 5(1), 55–70. <https://doi.org/10.3126/njs.v5i1.38003>

- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, 13(3), e0194889. <https://doi.org/10.1371/journal.pone.0194889>
- Ministry of Culture, Tourism and Civil Aviation (MoCTCA). (2023). *Tourism statistics 2000–2023*. Government of Nepal. <http://tourism.gov.np/downloads>
- Raschka, S. (2015). *Python machine learning: Unlock more profound insights into machine learning with Python*. Packt Publishing.
- Sharma, A., & Thapa, B. (2019). Time series modeling of Tourist arrivals in Nepal: An alternative approach. *Nepalese Journal of Statistics*, 3(1), 41–54. <https://doi.org/10.3126/njs.v3i1.27571>
- Subedi, A. (2017). Time series modeling on monthly data of Tourist arrivals in Nepal: An alternative approach. *Nepalese Journal of Statistics*, 1, 41–54. <https://doi.org/10.3126/njs.v1i1.12345>
- Voyant, C., Nivet, M.-L., Paoli, C., Muselli, M., & Notton, G. (2014). Meteorological time series forecasting based on MLP modeling using heterogeneous transfer functions. *Renewable Energy*, 68, 1–10. <https://doi.org/10.1016/j.renene.2014.01.001>
- Xiang, Z., & Fesenmaier, D. R. (2018). *Analytics in smart tourism design: Concepts and methods*. Springer. <https://doi.org/10.1007/978-3-319-44263-1>