TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

HIMALAYA COLLEGE OF ENGINEERING

A FINAL YEAR PROJECT REPORT

ON

"TOURISM ANALYSIS AND PREDICTION"

[CT 755]

SUBMITTED TO:

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

Chyasal, Lalitpur

SUBMITTED BY:

Sachin Bhattarai (39588)

Sakar Mainali (39592)

Sandesh Sharan Poudel (39595)

Utsab Pokharel (39605)

August, 2019

“TOURISM ANALYSIS AND PREDICTION”

[CT 755]

“A FINAL YEAR MOJOR PROJECT REPORT SUBMITTED FOR PARTIAL FULFILLMENT OF THE DEGREE OF BACHELORS’ IN COMPUTER ENGINEERING”

SUPERVISOR

Er. Ashok GM

SUBMITTED TO:

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

HIMALAYA COLLEGE OF ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

Chyasal, Lalitpur

SUBMITTED BY:

Sachin Bhattarai (39588)

Sakar Mainali (39592)

Sandesh Sharan Poudel (39595)

Utsab Pokharel (39605)

August, 2019

Copyright

Head of Department

Department of Electronics and Computer Engineering

Himalaya College of Engineering

ACKNOWLEDGEMENT

We are deeply grateful to our project coordinator **Er. Narayan Adhikari Chhetri** for providing the necessary guidelines and support to understand the feasibility and technical aspects of project. We would also like to offer gratitude to our project supervisor **Er. Ashok GM** whose lectures and ideas were the basis of our project research.

Group Members

Sachin Bhattarai (39588)

Sakar Mainali (39592)

Sandesh Sharan Poudel (39595)

Utsab Pokharel (39605)

ABSTRACT

In this modern data driven world, data analytics is found fruitful in various aspects like including financial services, healthcare, government, retail, e-commerce, media, manufacturing and so on. Tourism sector that affects broad aspects like economy, export, transport and so on contains a large data on respective fields also needs proper data analysis and predictions for its further improvements.

The project titled “Tourism Analysis and Prediction in Nepal” is a web application that helps generate some useful insights from tourism data collected from different sources about Nepal. Basically, the web performs analysis on the tourism data using statistical data presentations and visualization techniques like scatter plots, bar graphs, histograms, pie charts, different tourism related data- tourist’s arrivals, purpose of visits or tourist’s activities, length, places, hotels of visits, flights and passenger movements, income generation and so on. Also using suitable statistical method of simple linear regressions, SARIMA model and machine learning approach of MLP were used in regression problems including foreign exchange earnings and total percentage of tourist arrivals in particular location from the available tourism data that can be very helpful to various sectors involving in tourism- Government bodies, travel and tour agencies, guides, tourists for proper planning and decision making in future.

Keywords: *tourism, data analysis, prediction, machine learning, future planning*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ARIMA:	Auto Regressive Integrated Moving Average
CSS:	Cascading Style Sheet
HTML:	Hypertext Markup Language
MLP:	Multi-Layer Perceptron
MoCTCA:	Ministry of Culture, Tourism and Civil Aviation
TTCI:	Travel & Tourism Competitiveness Index
WEF:	World Economic Forum
SARIMA:	Seasonal Autoregressive Integrated Moving Average

1. INTRODUCTION

1.1. Background

Tourism being one of the world's largest industries represents a major area of interest, not just because of its size in terms of the enormous number of visitors or tourists, activities or the size of their consumption, but also because of its enormous impact on national economies and people's live's [1]. In context of Nepal tourism is the largest industry and its largest source of foreign exchange and revenue[2] and is ranked 103th in the Travel & Tourism Competitiveness Index (TTCI) by World Economic Forum(WEF) (2017).Possessing eight of the ten highest mountains in the world along with cultural and geographical diversity, Nepal is a hotspot destination for mountaineers, rock climbers and people seeking adventure.

In the modern world insights from data collected from various sectors is very essential tool for further improvements and development of nation. In order to analyze and visualize the data collected from different aspects of Nepalese tourism, Government of Nepal Ministry of Culture, Tourism and Civil Aviation (MoCTCA) [3] publishes a tourism statistics report yearly that has helped a lot to gain some insights about Nepalese tourism. But competing in data-driven world relying purely on statistical reports is not enough. There needs an application of Data Science that unifies statistics, data analysis, machine learning and their related methods to understand and analyze actual phenomena with data. The purposed project is a simple application of Data Science with a quest to add some insights from data in tourism sector of Nepal. The proposed project is a web application that uses the data from tourism statistics and performs suitable analysis and predictions on different aspects of tourism based on machine learning approach.

In the tourism industry planning is particularly important because of the rapid economic and political changes, the perishable nature of the tourism industry's products and services. To a large extent, planning relies heavily on accurate analysis and forecasts in order to reduce the risk in the decision making. The accurate analysis and forecasts then depend on the use of suitable methods and strategy's [4]. Research's [5] have shown that the use of modern data analysis and prediction in the field of tourism assists the improvement and growth of tourism sectors.

1.2. Problem statement

The existing statistical reports based on tourism sector which cannot produce more varied insight on different tourism data. There was lack of system that produces useful analysis and prediction specifically based on the particular topic that the user chooses to know more. There was also no proper system that can store and send tourism related data as per request of user that may be used for different other applications. So, this web application tries to solve the above mention problems.

1.3. Objectives

The general objective is to develop web application to provide useful information by analyzing and predicting from different aspects of tourism in Nepal. The specific objectives:

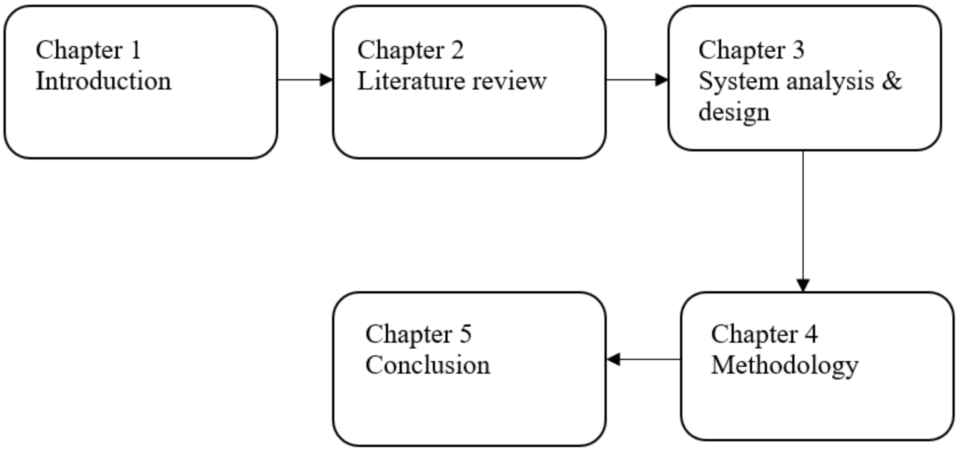
- To build a website and provide visualization on topics of tourist arrivals
- To predict the percentage of tourist arriving at particular location based on available facility
- To predict annual foreign income and monthly tourist arrival

1.4. Project scope

Since the project is a web application that performs analysis and predictions on available tourism data of Nepal, the insights on data can be useful to different sectors affected by tourism. Government bodies can use the insights to set marketing goals, explore potential markets, formulate necessary plans and policies related to travel and tourism. Managers of travel and tour agencies can use the insights to determine operational requirements such as staffing and capacity, and study project feasibility such as the viability to build a new hotel. Travel guides and agents can find the insights useful for trip plan and preparations. Tourists can find the insights useful to know more about trends in places, activities, services that can be useful during their visits.

The web application is an implementation of analytics with the aid of machine learning that is mainly search driven which analyzes data using search terms like places or locations, tourist's activities, foreign countries. It uses text search input and results to guide users to the information they are requesting for. Based on data of tourism statistic's it provides analysis with suitable visualizations on topics-tourist's arrivals, purpose of visits or tourists activities, length, places, hotels of visits, flights and passenger movements, income generation, employments or trainings and educations related to tourism, no of tourist's standard hotels, home stay, industries, guides, agencies, aviation in Nepal, tourism related accidents/incidents, feedbacks by tourists. The web application also provides a search for next coming years value prediction on topics like total international tourist's arrivals by volume, total and gross foreign exchange earnings from tourism.

1.5. Report organization



Chapter 1 includes introduction about the tourism with their background, objectives and project scope. Chapter 2 includes literature reviews which contains all the past research on topic related to tourism analysis. Chapter 3 includes system analysis and design which contain functional and non- functional requirement as well as system flow diagram and use case diagram. Chapter 4 includes methodology which is the core part of the project where we discussed about how we analyze and predict the data and what kind of tool we used to generate the result. Finally, chapter 5 includes conclusion where task accomplished as well as task remaining topic were discussed.

2. LITERATURE REVIEW

The drastic advancements in technology has transformed the tourism experience. There is huge potential in developing big data analytics in travel and tourism's[6]. Particularly, the design and development of tourism requires a profound understanding of what today's travelers need and want, how they move through and interact with physical and social spaces, and what leads to their enjoyment, happiness, and the realization of personal values.

Usually the nature of data found in tourism is a time series type where data sequence is in the order of time. Examples includes no of tourist's arrivals yearly, no of visitors in particular places in yearly basis, foreign exchange earnings in fiscal years etc. Simply these time series data may be encoded as visual objects with graphical techniques like scatter-plot, bar-graph, time-lines, pie-charts and so on which helps to analyze and reason about data and evidence. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Some of the major advantages of time series models include their systematic search capability for identification, estimation, and diagnostic checking.

For analytical tasks, such as for making comparisons use of charts of various types can be used to show patterns or relationships in the data for one or more variables. Examples of data analysis can be found on websites Google analytic[7], Destination analyst[8] where different data visualizations are used that helps to communicate information clearly and effectively through graphical means.

For forecasting time series there are different methods and approaches ranging from classical approach [9] like native approach, simple average, moving average (MA), exponential smoothing, Auto Regressive Integrated Moving average (ARIMA), Vector auto regression (VAR) models and so on to machine learning approach [10] like Artificial neural networks (ANN), Fuzzy logic, etc. for analysis and prediction of time series. The forecasting of non-linear time series can be accomplished using statistical models like ARIMA, SARIMA. There is wide use of SARIMA model for forecasting seasonal type time series data like monthly sales of items, no of tourist arrivals monthly[11]. The SARIMA model is based on the application of ARIMA models to transformed time series, where the seasonal and non-stationary behavior has been eliminated. Time series models, like the Autoregressive Integrated Moving Average (ARIMA), effectively consider serial linear correlation among observations, whereas Seasonal Autoregressive Integrated Moving Average (SARIMA) models can satisfactorily describe time series that exhibit non-stationary behaviors both within and across seasons.

In terms of techniques compared to classical models, machine learning based approaches count on several significant advantages like ability to handle complex nonlinear pattern, less restrictions and assumptions, easy automation[12]. One of the approach to forecasting time series is through artificial neural networks[13]. One simple way is using Multilayer Perceptron (MLP) which is a simple feed-forward neural network model with backpropagation learning that are most employed in time series studies[14]. They are 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. MLPs are useful in research for their ability to solve problems stochastically, which often allows approximate solutions for extremely complex problems like fitness approximation.

Typical examples that uses this approach are in market predictions, sales predictions, network traffic, meteorological forecasting[15]. The prediction of time series using neural network consists of teaching the network, the history of the variable in a selected limited time and applying the taught information to the future. Data from past are provided to the inputs of neural network and we expect data from future from the outputs of the network. Usually MLP model will learn a function that maps a sequence of past observations as input to an output observation. Regression predictive modeling is the task of approximating a mapping function from input variables to a continuous output variable. A continuous output variable is a real-value, such as an integer or floating points value. These are often quantities, such as amounts and sizes. A problem with multiple input variables is often called a multivariate regression problem. Also, if input variables are ordered by time is called a time series forecasting problem.

There are different tools and technologies used to apply data science and machine learning using different programming languages. For data science and machine learning, python is one of the most widely used programming language due to its simplicity and containing variety of data analysis libraries pandas, matplotlib, shapely, seaborn, geo-pandas along with large machine learning libraries NumPy, scikit-learn, TensorFlow where large people have made contributions and is in rapid improvements and research phase’s[16] .

3. SYSTEM ANALYSIS AND FEASIBILITY STUDY

3.1. Feasibility Analysis

A feasibility study is a preliminary study which investigates the information needs of prospective users and determines the resources requirements, costs, benefits, and feasibility of a proposed system. A feasibility study takes into account various constraints within which the system should be implemented and operated. In this stage the resource needed for implementation such as computing equipment, manpower and costs are estimated. The estimated are compared with available resources and a cost benefit analysis of the system is made. The feasibility analysis activity involves the analysis of the problem and collection of all relevant information relating to the project. The main objectives of the feasibility study are to determine whether the project would be feasible in terms of economic feasibility, technical feasibility and operational feasibility and on or not. The input data which are required for the project are available or not. Thus, we evaluated the feasibility of the system in terms of the following categories:

- Technical feasibility
- Operational feasibility
- Economic feasibility
- Schedule feasibility

3.1.1. Technical feasibility

Since the web application uses software technologies and tools which are freely available and technical skills required can be easily manageable. There are many free machine learning libraries available for data analysis and predictions with proper documentations and courses. The hardware system in the project need not be highly computing but requires a normal computing and the system server must be adequate enough and manageable in future. So, it is seen that the hardware and software meet the need of the system. So, it’s clear that the proposed project is technically feasible.

3.1.2. Operational Feasibility

Since the web application is interactive and data driven, the user need to be only a bit familiar with the software system backed with graphical explanations and that can be easily be understood faster in time with usage. This system highly focuses on design-dependent parameters like reliability, maintainability, supportability, usability, predictability, disposability, sustainability, affordability etc. that fits into the operating functions of the project. So, the project is feasible in operation.

3.1.3. Economic feasibility

Economic feasibility attempts to weigh the costs of developing and implementing a new system, against the benefits that would accrue from having the new system in place. This feasibility study gives the top management the economic justification for the new system. A simple economic analysis which gives the actual comparison of costs and benefits are much more meaningful in this case. In addition, this proves to be a useful point of reference to compare actual costs as the project progresses. There could be various types of intangible benefits on account of automation. These could include increased customer satisfaction, improvement in product quality, better decision making, and timeliness of information, expediting activities, improved accuracy of operations, better documentation and record keeping, faster retrieval of information, better employee morale.

3.1.4. Schedule feasibility

Since the project is web-based analytics, the software development time do take longer. The dateline of software system can be easily estimated if the proper team and achievable goals are formed. Roughly the proposed system can be delivered within 24 weeks on tight schedule. The Gantt chart below shows the schedule feasibility of the project.

3.2. Requirement analysis

3.2.1. Functional requirement

There are functions or features that must be included in any system to satisfy the needs and be acceptable to the users. Based on this, some basic functional need of the project is:

- The system should produce suitable data analysis and predictions on tourism dataset that can be easily downloadable for normal user and easily updatable for admin
- The system should provide API for accessing data

Hardware Requirements

Processer:	Intel Pentium 4 or higher
Memory:	1 GB (recommended)
Storage devices:	20 GB or above

Software Requirements

Operating system:	Windows/Linux with web browser installed
Web browser:	Windows Explorer/Edge/Chrome or equivalent
Database:	MySQL or MariaDB

3.2.2. Non-functional requirement

The non-functional requirements like performance, information, economy, control and security efficiency and services are very essential for successful project completion. Based on these, the non-functional requirements of the project are as follows:

- The system should be easy to use, user friendly in operation and provide useful information to user
- The system should perform with efficient throughput and response time
- The system should provide better accuracy in analysis and prediction of datasets

4. System Design

4.1. System Overview

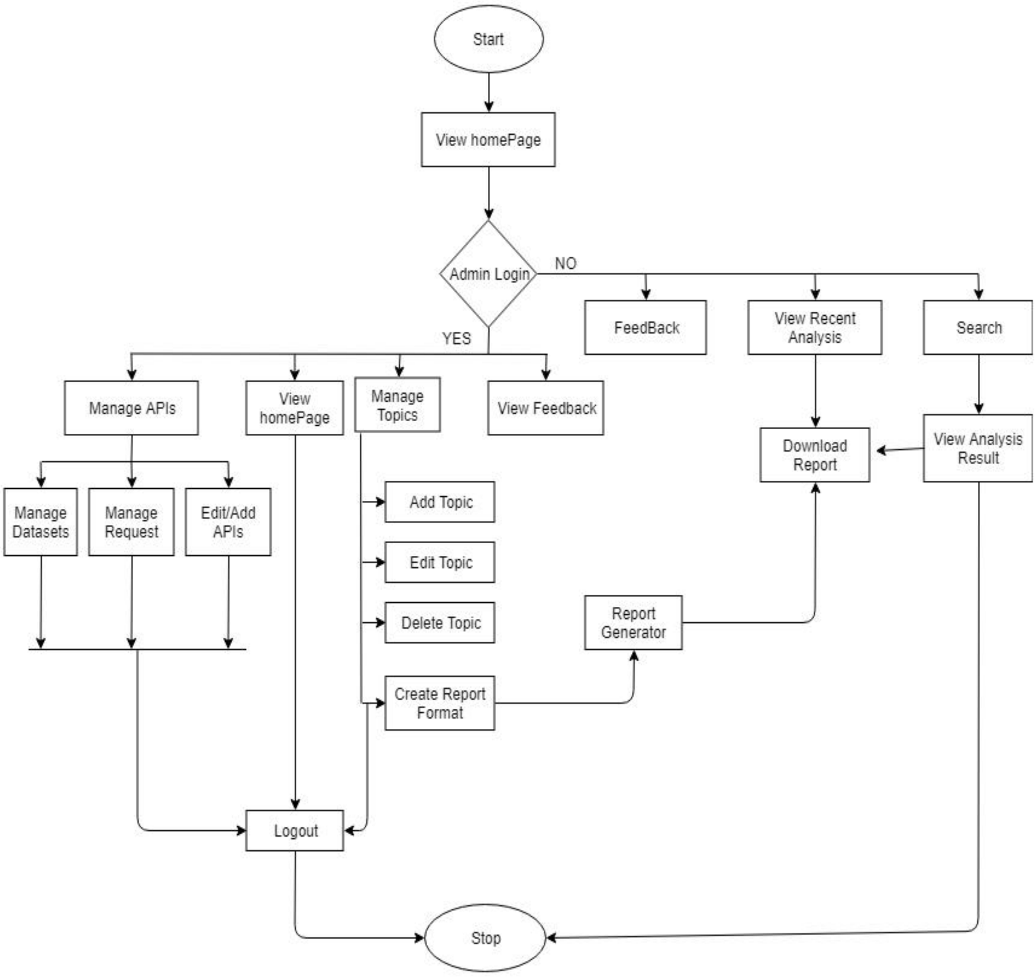


Figure 4.1: System flow diagram

4.2. USE Case Diagram

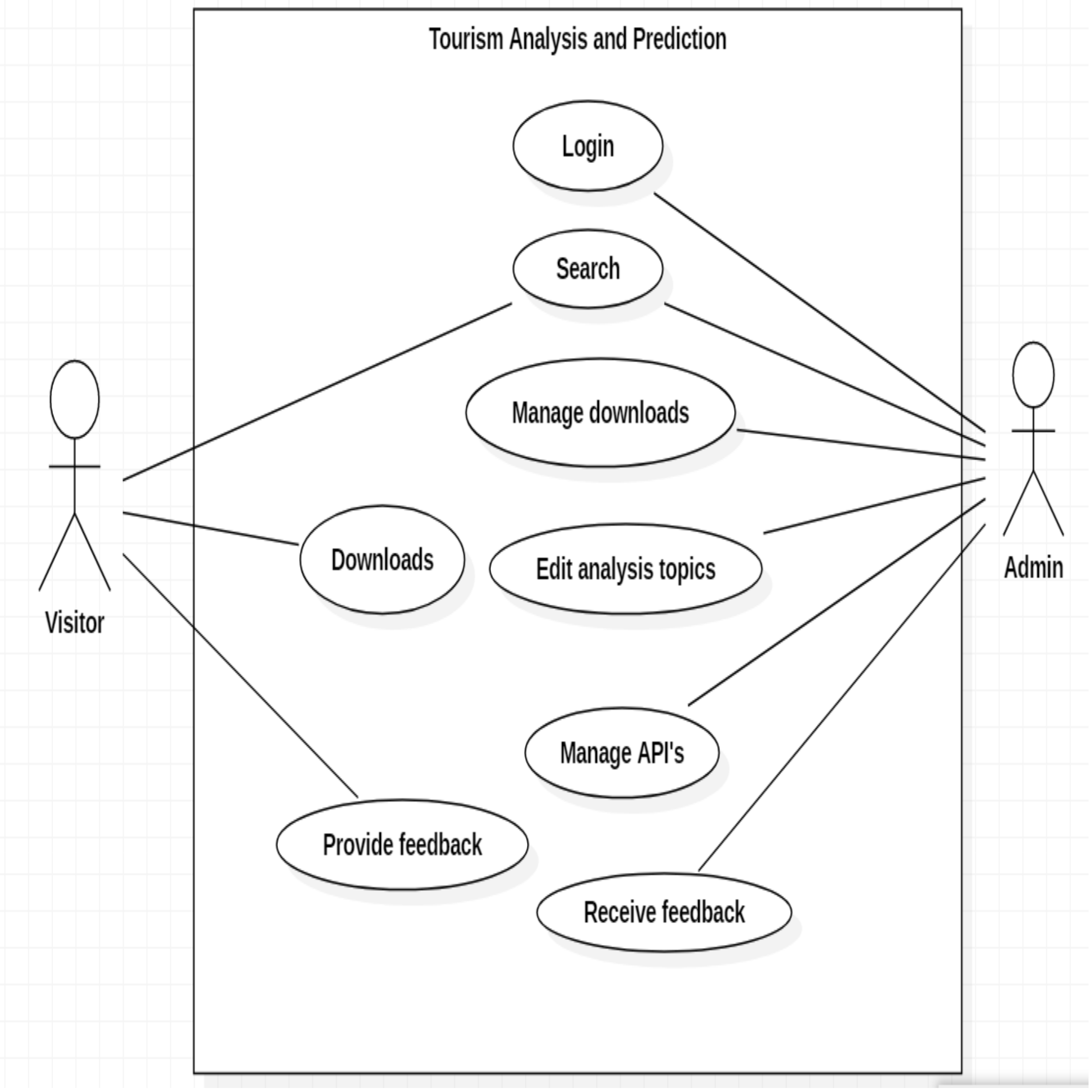


Figure 4.2: Use Case diagram

4.3. DFD

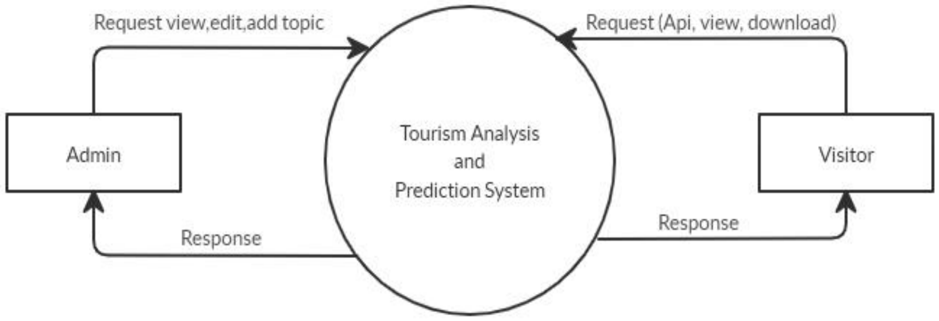


Figure 4.3: DFD level-0

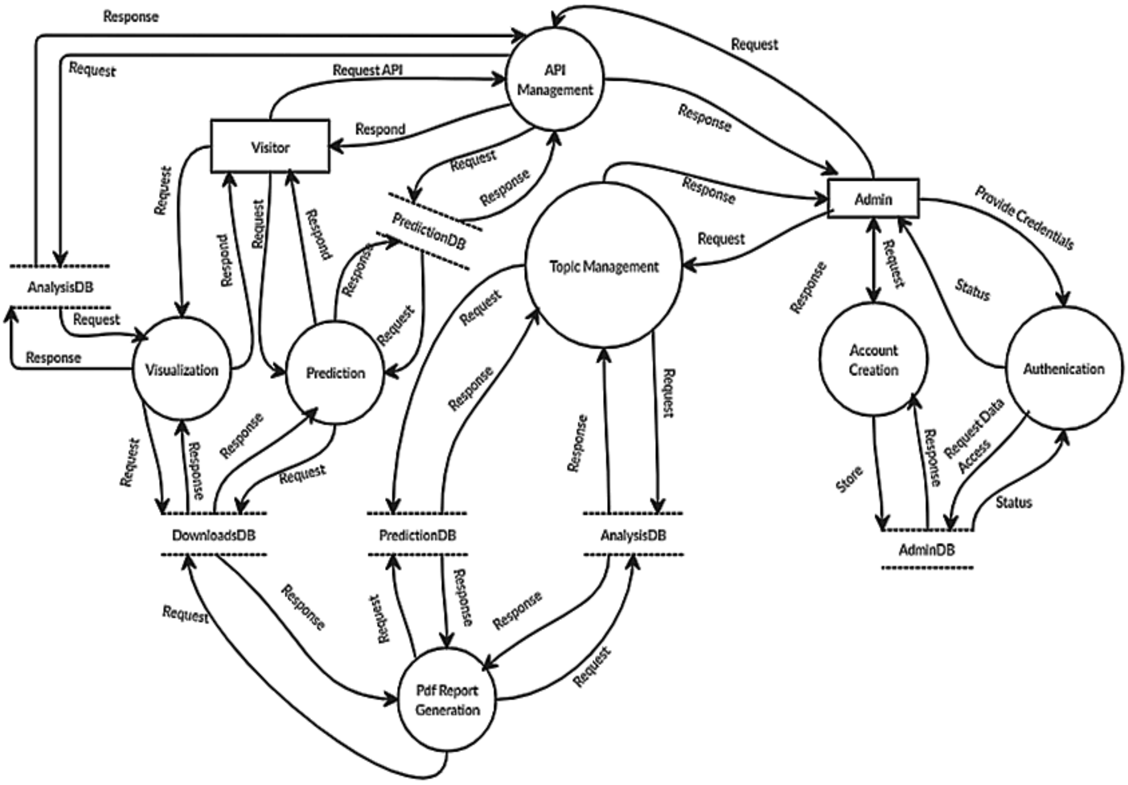


Figure 4.4: DFD level-1

4.4. Class Diagram

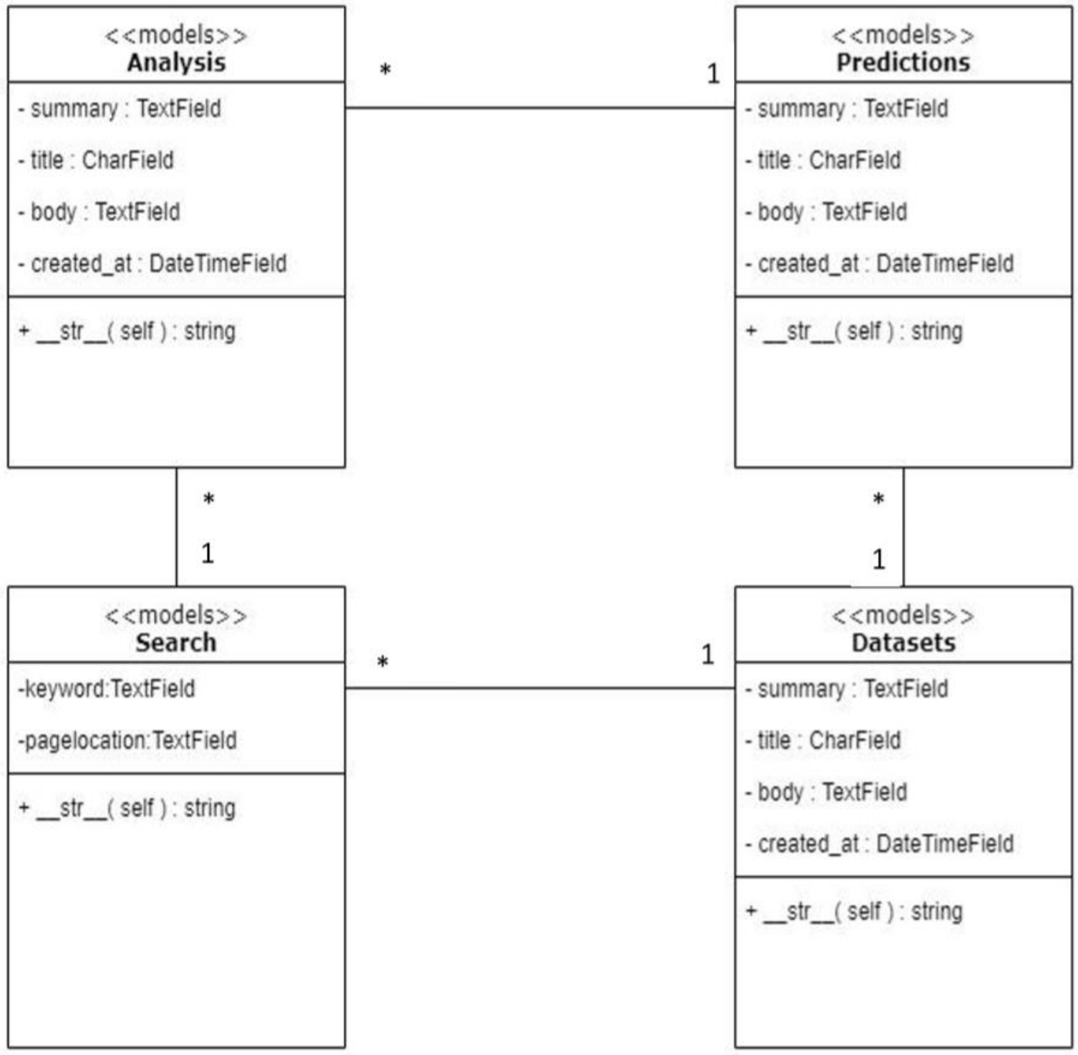


Figure 4.5: Class Diagram

5. METHODOLOGY

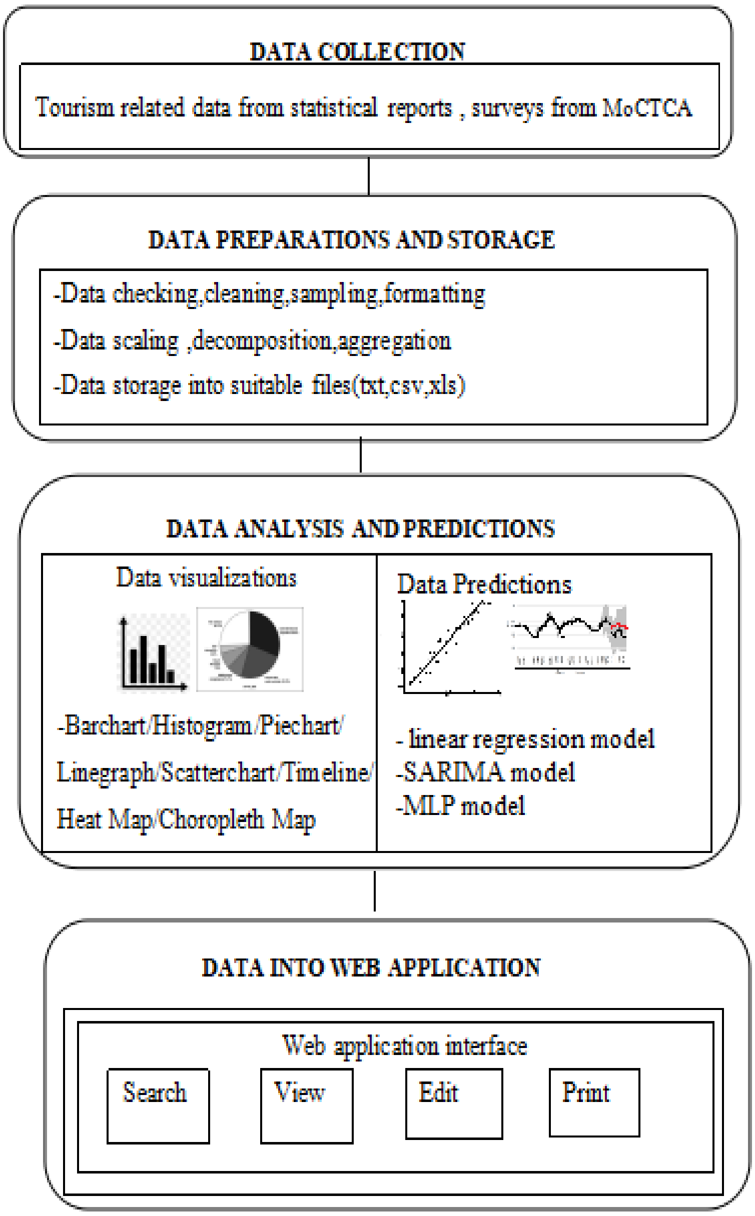

DATA COLLECTION
Tourism related data from statistical reports , surveys from MoCTCA
DATA PREPARATIONS AND STORAGE
-Data checking,cleaning,sampling,formatting
-Data scaling ,decomposition,aggregation
-Data storage into suitable files(txt,csv,xls)
DATA ANALYSIS AND PREDICTIONS
Data visualizations
-Barchart/Histogram/Piechart/
Linegraph/Scatterchart/Timeline/
Heat Map/Choropleth Map
Data Predictions
- linear regression model
-SARIMA model
-MLP model
DATA INTO WEB APPLICATION
Web application interface
Search
View
Edit
Print

Figure 5.1: System Architecture

5.1. System block diagram for analysis and prediction

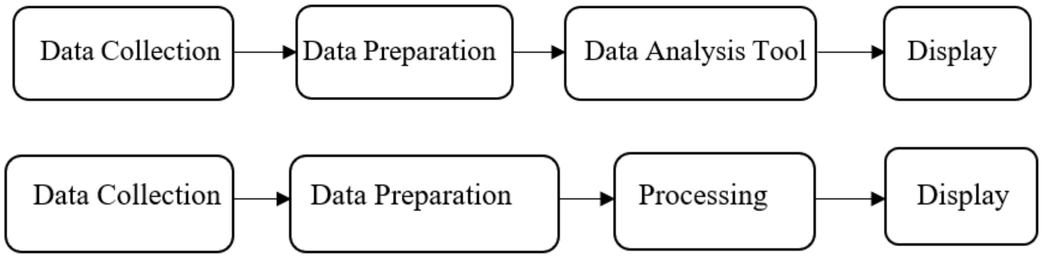


Figure 5.2: System block diagram

5.1.1. Data Collections and Preparations

Data Sources:

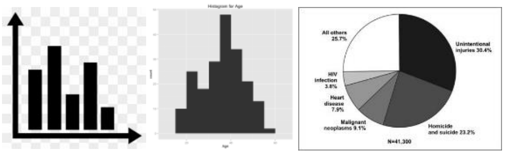
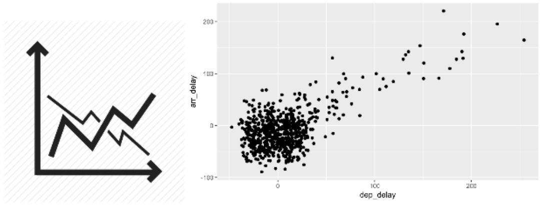
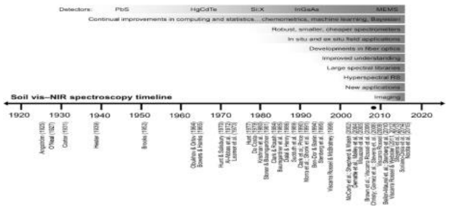
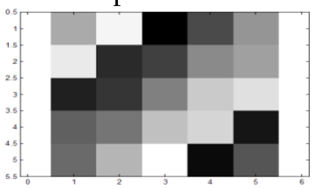
- Government of Nepal(MoCTCA), Tourism statistics reports[3]
- Government of Nepal(MoCTCA), Civil Aviation reports[16]
- World Travel & Tourism Council. The WTTC report: Travel and Tourism Economic Impact Nepal report's[2]
- Surveys reports with tourists about their experiences in Nepal[3]

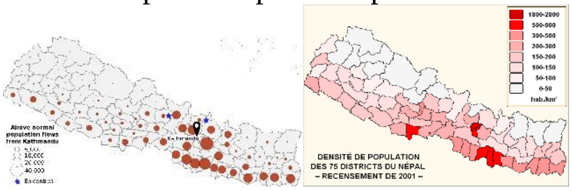
The foremost step initiating this project is proper data collection and preparations. The preliminary data preparations tasks incudes' data checking, cleaning, editing, sampling organizing, formatting into suitable forms (xml, csv), scaling, decomposition, aggregations and so on before using into analysis and prediction model.

5.1.2. Data Analysis and Predictions

The first major objective of the project 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 uses statistics through libraries in python[17] which can be used in the proposed project. Some major graphical techniques along with their usage to visualize data are:

Table 5.1: Graphical techniques

DIAGRAMS	**EXAMPLE USAGE**
Barchart/Histogram/Piechart  	▪ Tourist arrivals for different purpose ▪ No. of visitors in different places ▪ No. of tourist tourists related accident/incidents ▪ No. of flights in different domestic airlines
Linegraph/Scatterchart  	▪ International and Domestic flight movements by month on the basis of gender/country ▪ Total tourist arrival volume by month on the basis of gender/country/age group ▪ Gross foreign exchange earnings from tourism by fiscal year
Timeline  	▪ All available statistical records with highest and lowest value records on the basis of year ▪ Important tourism related events/incidents
Heat Map  	▪ Economic indicators values of hotels and restaurant on yearly ▪ Civil aviation indicators values on yearly basis ▪ No. of different types of tourist industries

Scatter Map /Choropleth Map  Above normal population flows from Kathmandu 5,000 10,000 20,000 40,000 Epicentres Kathmandu 1800-2800 500-900 300-500 200-300 150-200 100-150 50-100 0-50 hab./km² DENSITÉ DE POPULATION DES 75 DISTRICTS DU NÉPAL – RECENSEMENT DE 2001 –	▪ No. of tourists visits in major places of Nepal ▪ Places in Nepal with different tourist activities records ▪ Places in Nepal with different tourist standard hotels/home stays

5.1.2.1. 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) and 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=b0+b1x** 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 move 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 {(xi, yi), i = 1, ..., n}. We can describe the underlying relationship between yi and xi involving this error term Ei by **yi=b0+b1xi+Ei**

This relationship between the true write nothing /delete it underlying parameters α and β 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 great, 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:

x̅ and y̅   as the average of the xi and yi, respectively

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

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

$$\boldsymbol{b1} = \frac{\sum_{i=1}^{n}(\boldsymbol{xi}-\bar{\boldsymbol{x}})(\boldsymbol{yi}-\bar{\boldsymbol{y}})}{\sum_{i=1}^{n}(\boldsymbol{xi}-\bar{\boldsymbol{x}})^{2}} \quad \text{or by using} \quad \boldsymbol{b1} = \frac{\boldsymbol{Cov}(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{Var}(\boldsymbol{x})} \text{..............Equation (5.1)}$$

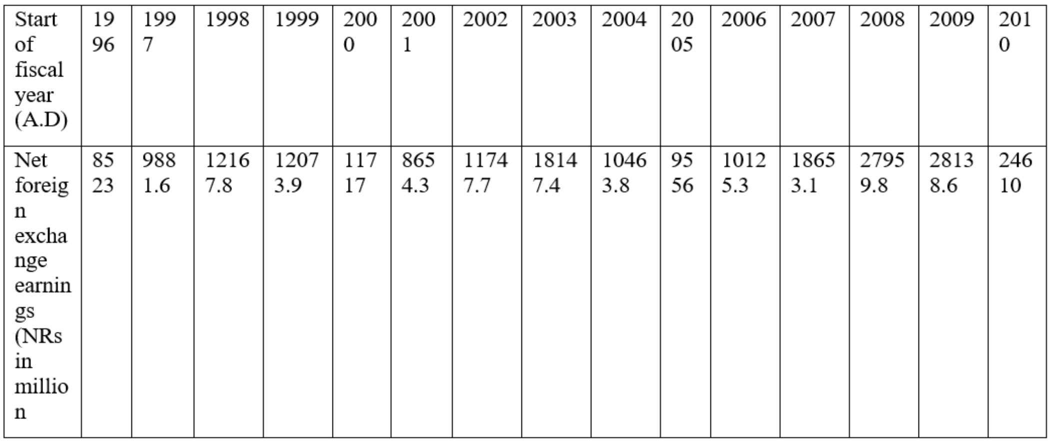
Finally, we can calculate b0 using b1 and some statistics from our dataset, as follows:

$$\boldsymbol{bo} = \sum_{i=1}^{n}\boldsymbol{yi} - \boldsymbol{b1}\sum_{i=1}^{n}\boldsymbol{xi} \text{..............Equation (5.2)}$$

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 5.2: 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:

$$\boldsymbol{mean(x) = sum(x)/count(x)} \text{ ……………Equation (5.3)}$$

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:

$$\boldsymbol{variance} = \boldsymbol{sum}((\boldsymbol{x} - \boldsymbol{mean}(\boldsymbol{x}))^{\mathbf{2}}) \text{ ....Equation (5.4)}$$

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:

$$\boldsymbol{covariance}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{sum}((\boldsymbol{x}(\boldsymbol{i}) - \boldsymbol{mean}(\boldsymbol{x})) * \big(\boldsymbol{y}(\boldsymbol{i}) - \boldsymbol{mean}(\boldsymbol{y})\big))$$

..............Equation (5.5)

c. Estimate Coefficients:

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

$$\boldsymbol{b1} = \boldsymbol{sum}((\boldsymbol{x}(\boldsymbol{i}) - \boldsymbol{mean}(\boldsymbol{x})) * (\boldsymbol{y}(\boldsymbol{i}) - \boldsymbol{mean}(\boldsymbol{y}))) / \boldsymbol{sum}(\ (\boldsymbol{x}(\boldsymbol{i}) - \boldsymbol{mean}(\boldsymbol{x}))\text{^}\mathbf{2}\ ) \quad \text{or}$$

$$\boldsymbol{b1} = \boldsymbol{covariance}\ (\boldsymbol{x},\boldsymbol{y})\ /\ \boldsymbol{variance}(\boldsymbol{x}) \text{ .... Equation (5.6)}$$

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.

$$\boldsymbol{b0} = \boldsymbol{mean}(\boldsymbol{y}) - \boldsymbol{b1} * \boldsymbol{mean}(\boldsymbol{x}) \text{ .... Equation (5.7)}$$

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

$\boldsymbol{y = b0 + b1 * x}$**……………………..** Equation (5.8)

5.1.2.2. 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 combinations 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 previous observed values up to a defined maximum lag (denoted p), plus a random error term ε_t and given as:

$$\mathrm{Y_t} = \alpha + \beta_1 \mathrm{Y_{t-1}} + \beta_2 \mathrm{Y_{t-2}} + \cdots + \beta_p \mathrm{Y_{t-p}} + \varepsilon_t \text{..............Equation (5.9)}$$

where, $\mathrm{Y_{t-1}}$ is the lag1 of the series β_1 is the coefficient of lag1, α is the intercept term, ε_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 previous random error terms up to a defined maximum lag (denoted q).

$$\mathrm{Y_t} = \alpha + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \cdots + \Phi_q \varepsilon_{t-q} \text{...........Equation (5.10)}$$

where the error terms are the errors of the autoregressive models of the respective lags. The errors ε_t and ε_{t-1} are the errors from the following equations:

$$\mathrm{Y_t} = \beta_1 \mathrm{Y_{t-1}} + \beta_2 \mathrm{Y_{t-2}} + \cdots + \beta_0 \mathrm{Y_0} + \varepsilon_t \text{..............Equation (5.11)}$$

$$\mathrm{Y_{t-1}} = \beta_1 \mathrm{Y_{t-2}} + \beta_2 \mathrm{Y_{t-3}} + \cdots + \beta_0 \mathrm{Y_0} + \varepsilon_{t-1} \text{..............Equation (5.12)}$$

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:

$$\Delta^1 \mathrm{Y_t} = \mathrm{Y_t} - \mathrm{Y_{t-1}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. \text{ Equation (5.13)}$$

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

$$\Delta^1 \mathrm{Y_t} = (1 - \mathrm{B})\mathrm{Y_t} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ Equation (5.14)}$$

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:

$$\mathrm{Y_t} = \alpha + \beta_1 \mathrm{Y_{t-1}} + \beta_2 \mathrm{Y_{t-2}} + \cdots + \beta_p \mathrm{Y_{t-p}} + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \cdots + \Phi_q \varepsilon_{t-q} \ldots\ldots\ldots\ldots\ldots \text{ Equation (5.15)}$$

or it can be expressed as

$$\beta_p(B)\Delta^d \mathrm{Y_t} = \Phi_q(B)\ \varepsilon_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ Equation (5.16)}$$

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

ARIMA model in words:

Predicted Y_t = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q 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:

$$\beta_{\mathrm{p}}(B)\theta_P(\mathrm{B}^s)\Delta^d\ \Delta_s{}^D\mathrm{Y}_{\mathrm{t}} = \Phi_{\mathrm{q}}(B)\Theta_Q(\mathrm{B}^s)\ \varepsilon_t \text{.............. Equation (5.17)}$$

Where, $\theta_P(\mathrm{B}^s)$ is the seasonal AR operator, $\Delta_s{}^D$ is the seasonal I operator, $\Theta_Q(\mathrm{B}^s)$ 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. The 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 (ρ) is given by:

$$\mathrm{P}_{(X,Y)} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} \text{.............................. Equation (5.18)}$$

Where, $cov(X,Y)$ is the covariance of (X, Y), σ_X is the standard deviation of X and σ_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:

$$P_{13.2} = \frac{\rho_{13}-\rho_{12}\rho_{32}}{\sqrt{1-\rho_{12}{}^2}\sqrt{1-\rho_{32}{}^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{Equation (5.19)}$$

Where, ρ_{ab} is correlation coefficient between two terms a and b.

We can calculate the correlation for time series observations with observations with previous 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 5.3: No. of tourist arrive in month

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

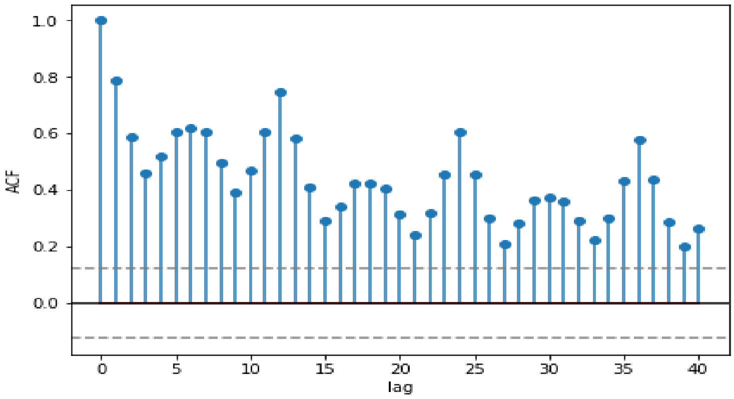


Figure 5.3: Plot of auto-correlation function

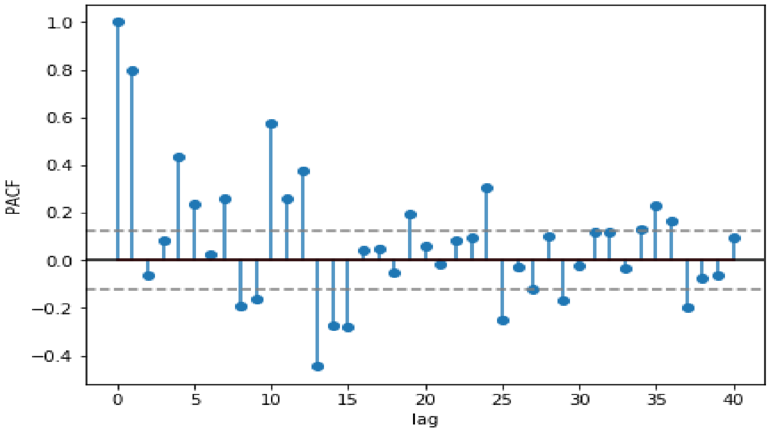


Figure 5.4: Plot of partial auto-correlation function

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:

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

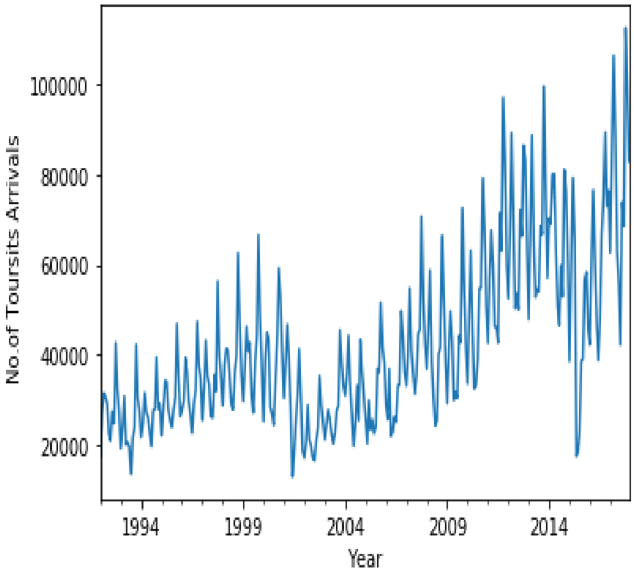


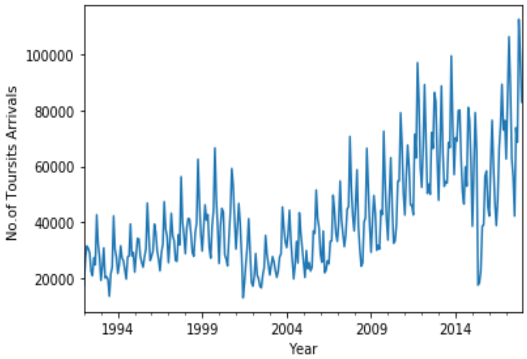
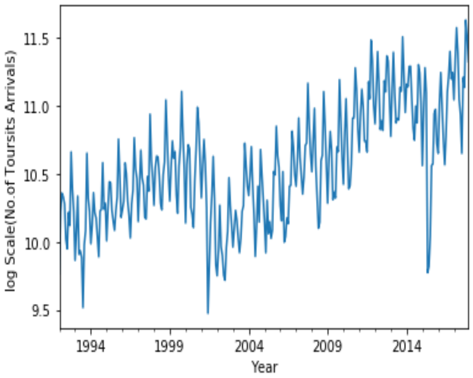
Figure 5.5: Plot of No. of tourist arrive in month

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 5.4: Transformation of time series data

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

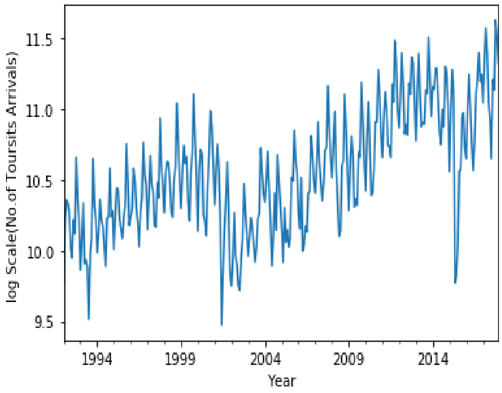
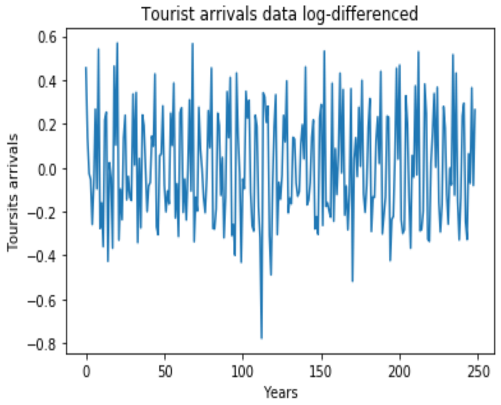
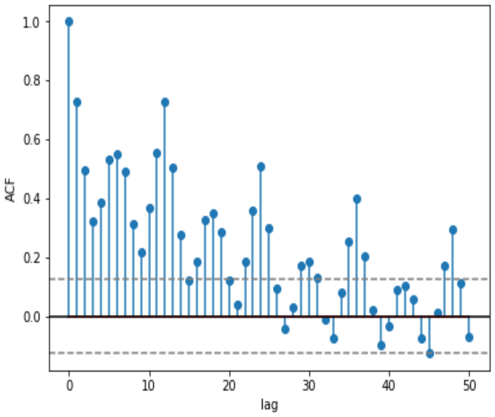
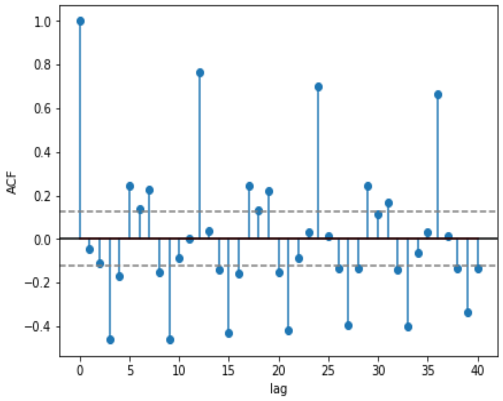
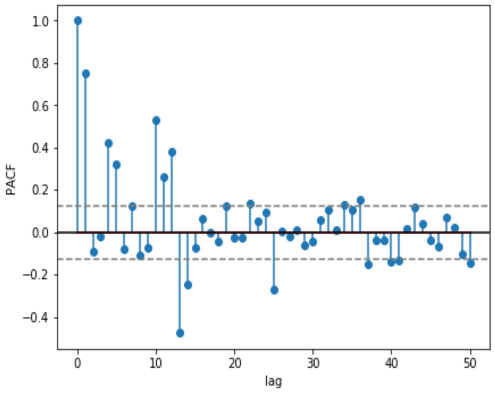
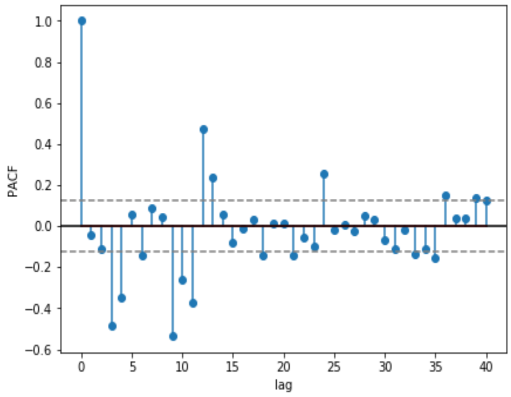
Log transformed series	Log differenced series(1st order differencing)
	
	
	

Figure 5.6: 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≥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≤2

P≥1 if the ACF is positive at lag s, else P=0.

Q≥1 if the ACF is negative at lag s, else Q=0.

Rule of thumb: P+Q≤2

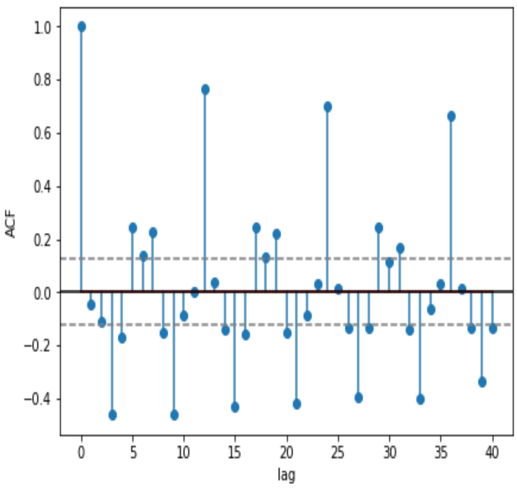
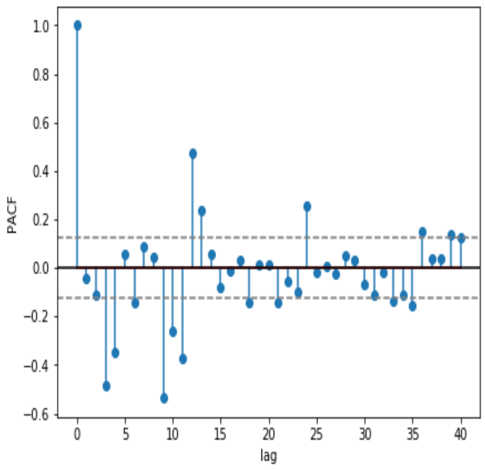
ACF plot	PACF plot
 	

Figure 5.7: 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:

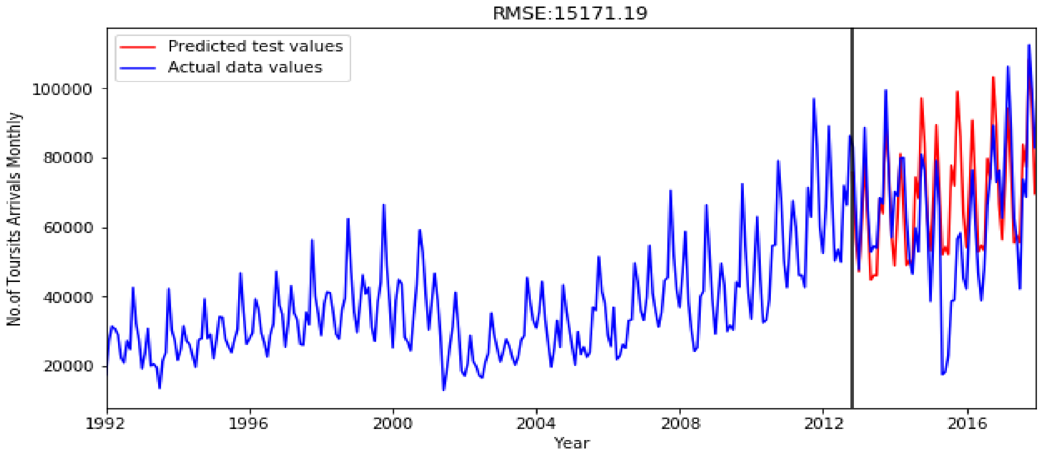
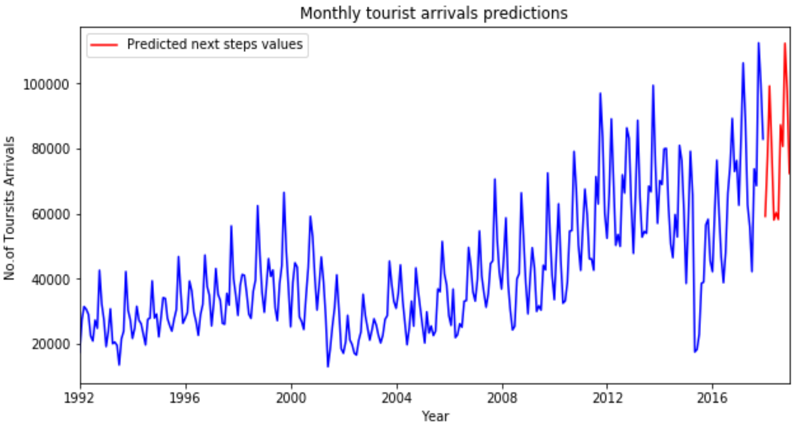


Figure 5.8: Plot of actual and predicted value

Table 5.5: Predicted monthly values for Year-2018

<table>
<tr>
<td>


</td>
<td>Predicted monthly values for Year-2018
2018-01-01
59230.060623
2018-02-01
77336.974494
2018-03-01
99247.696946
2018-04-01
79200.410266
2018-05-01
58098.806378
2018-06-01
60299.502946
2018-07-01
58266.541448
2018-08-01
87296.918446
2018-09-01
80717.982301
2018-10-01
112334.227443
2018-11-01
97692.395242</td>
</tr>
</table>

	2018-12-01 72426.282824

5.1.2.3. Regression prediction using MLP Neural Network

Multilayer 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 major objective of the project 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 purpose calculating root mean square error. A simple step's to training the MLP model is given below

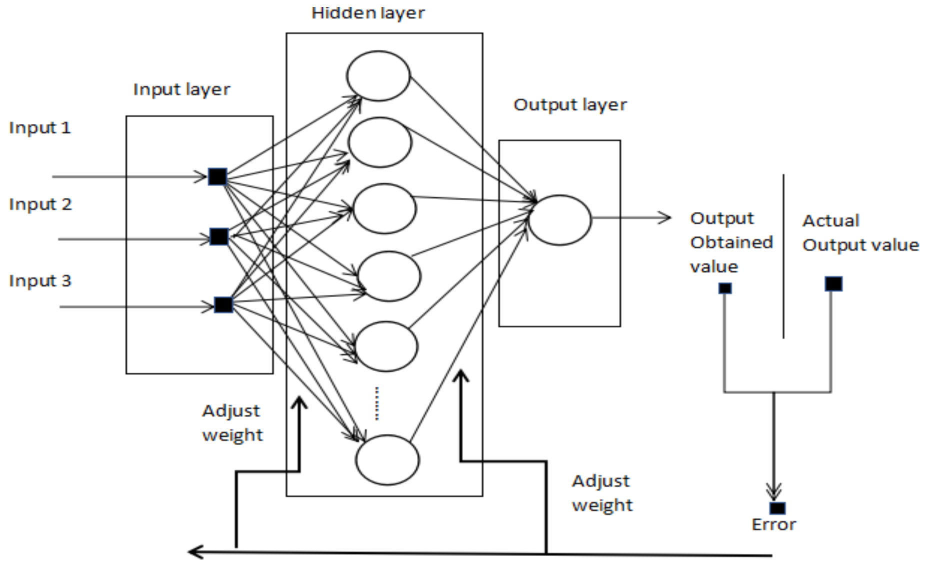


Figure 5.9: Multi-layer Perceptron Model

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.

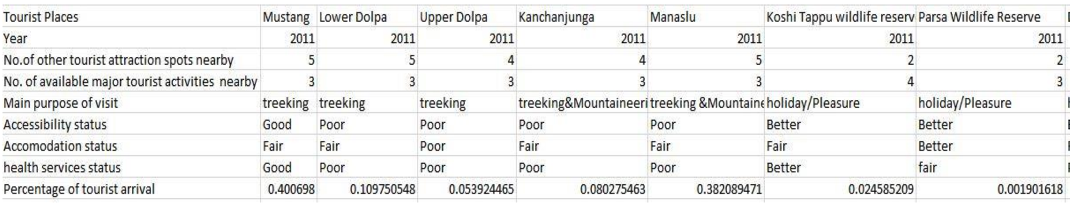
STEP 5: Back-Propagation: from right to left, the error is back-propagated. Update the weights according to how much they are responsible for the error. The learning rate decides by how much we update the weights.

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.



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 &Mountaine	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

Figure 5.10: Data of different tourist place in the year 2011

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

Table 5.6: Input and output with data of accessibility, accommodation, health and medical

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

	POOR	**FAIR**	**GOOD**	**BETTER**
ACCESSIBILITY 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 or primary roads or feeder road -Railways -National domestic airports	-Highway -International Airport
ACCOMODATION References: -Hotel association Nepal(records)	-Local shops/tea house and simple home stays	-Local hotels and lodges -Well managed or tourism-oriented homestays and guest houses	-Tourist standard hotels -Tourist class lodge -Registered resorts	-Registered star hotels and lodges -Larger hotels for tourist accommodation
HEALTH & MEDICAL 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	-Zonal, Regional Central gov hospitals

				-Private large hospital

Table 5.7: List of major tourist activities in Nepal

List of major 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
• Meditaion /religious activities

Table 5.8: List of major purpose of visit in Nepal

List of major purpose of visit in Nepal
• Holiday/Pleasure (includes: indoor enjoyment, photography, rafting, bungy jumping, camping, paragliding, biking, jungle safari etc.)
• Trekking/hiking (includes: visiting along major trekking and hiking route’s)
• Mountaineering (includes: climbing major route allowed mountain’s)
• Mountaineering and trekking (includes both mountaineering and trekking)
• Official (includes visit for official or government purposes)
• Business (includes visits for business research, operations or investments)
• Conference/Conventions (includes visit during special conventions or conferences)

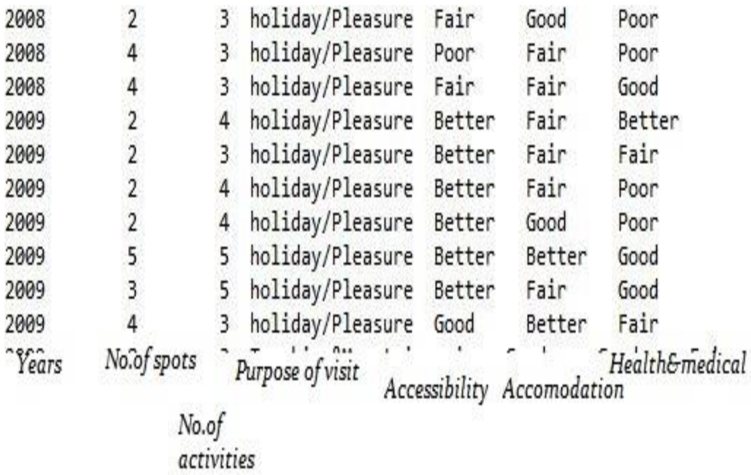
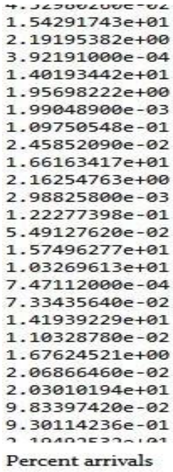
Here given independent variables for each year we are interested in predicting the percentage out of total tourist arrivals in that place. This multi-variate regression problem can be solved using python programming language with the use of following libraries:

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:

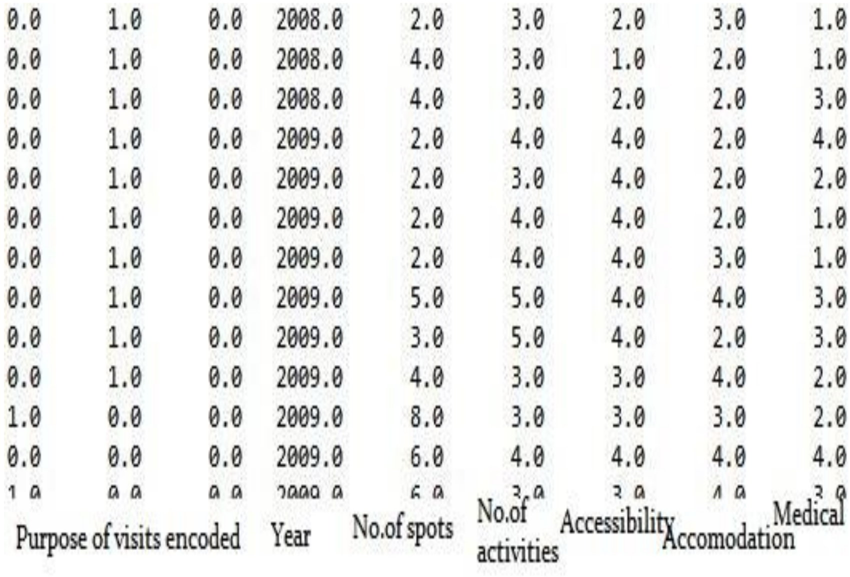
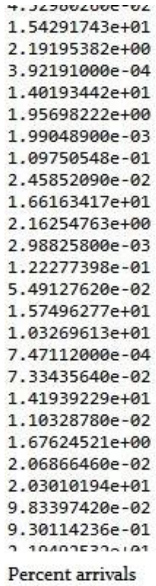
Table 5.9: Separation dependent and independent variables

X	y
 2008 2 3 holiday/Pleasure Fair Good Poor 2008 4 3 holiday/Pleasure Poor Fair Poor 2008 4 3 holiday/Pleasure Fair Fair Good 2009 2 4 holiday/Pleasure Better Fair Better 2009 2 3 holiday/Pleasure Better Fair Fair 2009 2 4 holiday/Pleasure Better Fair Poor 2009 2 4 holiday/Pleasure Better Good Poor 2009 5 5 holiday/Pleasure Better Better Good 2009 3 5 holiday/Pleasure Better Fair Good 2009 4 3 holiday/Pleasure Good Better Fair Years No.of spots No.of activities Purpose of visit Accessibility Accomodation Health&medical	 1.54291743e+01 2.19195382e+00 3.92191000e-04 1.40193442e+01 1.95698222e+00 1.99048900e-03 1.09750548e-01 2.45852090e-02 1.66163417e+01 2.16254763e+00 2.98825800e-03 1.22277398e-01 5.49127620e-02 1.57496277e+01 1.03269613e+01 7.47112000e-04 7.33435640e-02 1.41939229e+01 1.10328780e-02 1.67624521e+00 2.06866460e-02 2.03010194e+01 9.83397420e-02 9.30114236e-01 Percent arrivals

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 5.10: Encoding and labeling the categorical data

X	y
 0.0 1.0 0.0 2008.0 2.0 3.0 2.0 3.0 1.0 0.0 1.0 0.0 2008.0 4.0 3.0 1.0 2.0 1.0 0.0 1.0 0.0 2008.0 4.0 3.0 2.0 2.0 3.0 0.0 1.0 0.0 2009.0 2.0 4.0 4.0 2.0 4.0 0.0 1.0 0.0 2009.0 2.0 3.0 4.0 2.0 2.0 0.0 1.0 0.0 2009.0 2.0 4.0 4.0 2.0 1.0 0.0 1.0 0.0 2009.0 2.0 4.0 4.0 3.0 1.0 0.0 1.0 0.0 2009.0 5.0 5.0 4.0 4.0 3.0 0.0 1.0 0.0 2009.0 3.0 5.0 4.0 2.0 3.0 0.0 1.0 0.0 2009.0 4.0 3.0 3.0 4.0 2.0 1.0 0.0 0.0 2009.0 8.0 3.0 3.0 3.0 2.0 0.0 0.0 0.0 2009.0 6.0 4.0 4.0 4.0 4.0 1.0 0.0 0.0 2009.0 6.0 3.0 3.0 4.0 3.0 Purpose of visits encoded Year No.of spots No.of activities Accessibility Accomodation Medical	 1.54291743e+01 2.19195382e+00 3.92191000e-04 1.40193442e+01 1.95698222e+00 1.99048900e-03 1.09750548e-01 2.45852090e-02 1.66163417e+01 2.16254763e+00 2.98825800e-03 1.22277398e-01 5.49127620e-02 1.57496277e+01 1.03269613e+01 7.47112000e-04 7.33435640e-02 1.41939229e+01 1.10328780e-02 1.67624521e+00 2.06866460e-02 2.03010194e+01 9.83397420e-02 9.30114236e-01 Percent arrivals

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

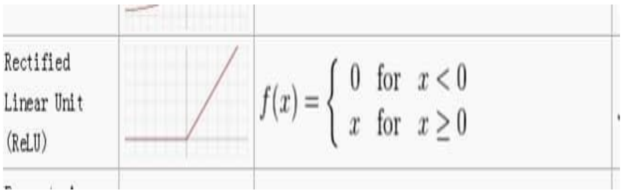
This is the step where a whole dataset is divided into test and train sets. Here for this problem, train to test set can be divided into ratio 4:1.ie 20 percent of dataset is used as a test set for model and the remaining as train sets (X_train and y_train). Usually train set helps to fit or build a 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 5.11: 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
		
		Rectified Linear Unit (ReLU)
		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$

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 5.12: 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 test 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 5.13: Prediction of next top tourist destinations in Nepal

Location	Year	No. of tourist spots	No. of tourist activities	Purpose 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	2.0926218032

								836914
Helambu	2019	4	3	treeking	Good	Better	Fair	1.0644960403442383
Taplejung	2019	5	3	Treeing &Mountaineering	Poor	Fair	Poor	0.52480161190032 96
Gokyo Valley	2019	4	3	treeking	Good	Good	Fair	0.139580190181732 18

6. RESULT AND ANALYSIS

6.1. Testing

6.1.1. 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 project 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.

6.1.2. 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 purpose of this level of testing is 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.

6.1.3. System Testing

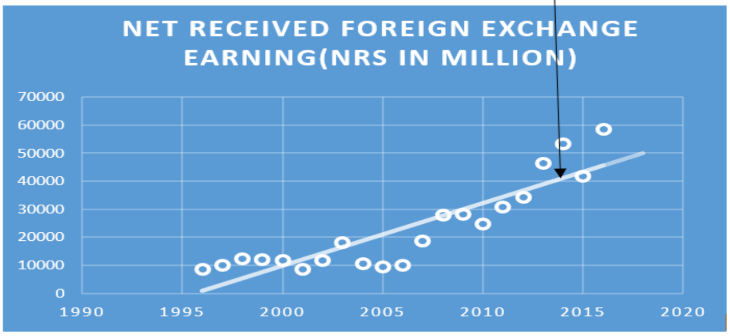
System testing is testing conducted on a complete integrated system to evaluate the system's compliance with its specified requirements. System testing takes, as its input, all of the integrated components that have passed integration testing. The purpose of integration testing is to detect any inconsistencies 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 6.1: Testing for Simple Linear Regression Model

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



Y=-4478336.394805195+2244.041558441558X

Year	Actual	Predicted
2012	34210.6	36675.221
2013	46374.9	38919.263
2014	53428.8	41163.303
2015	41765.4	43407.345
2016	58526.9	45651.3870

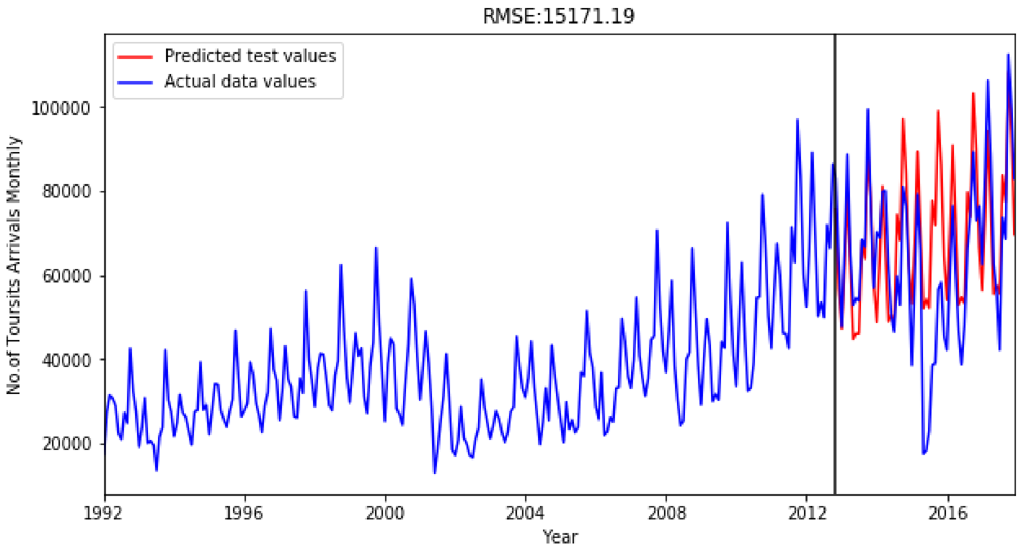
Year	Forecasted
2017	47895.429
2018	50139.470

Figure 6.1: Result of Simple Linear Regression Model

Table 6.2: Testing for SARIMA Model

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

		2017-10-01 108273.349913 2017-11-01 94110.867153 2017-12-01 69684.201630



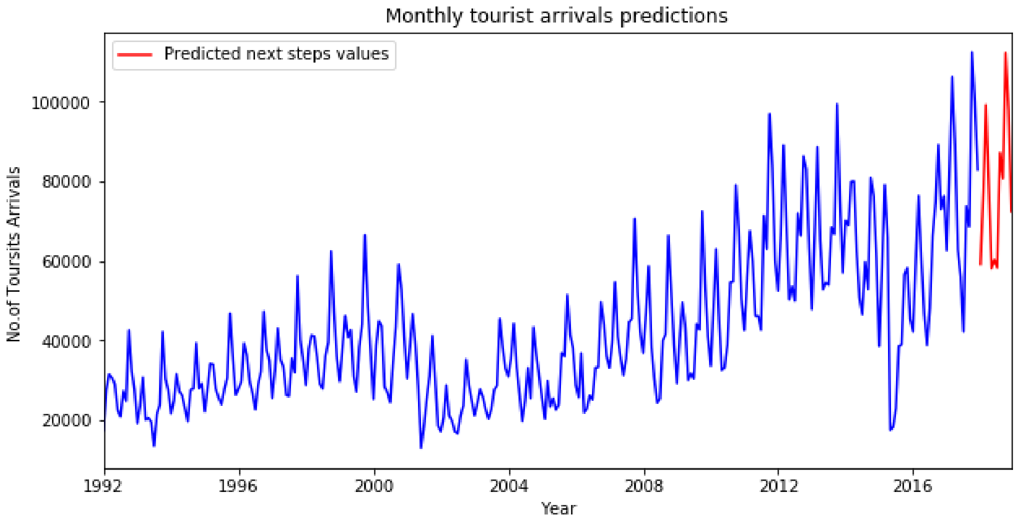
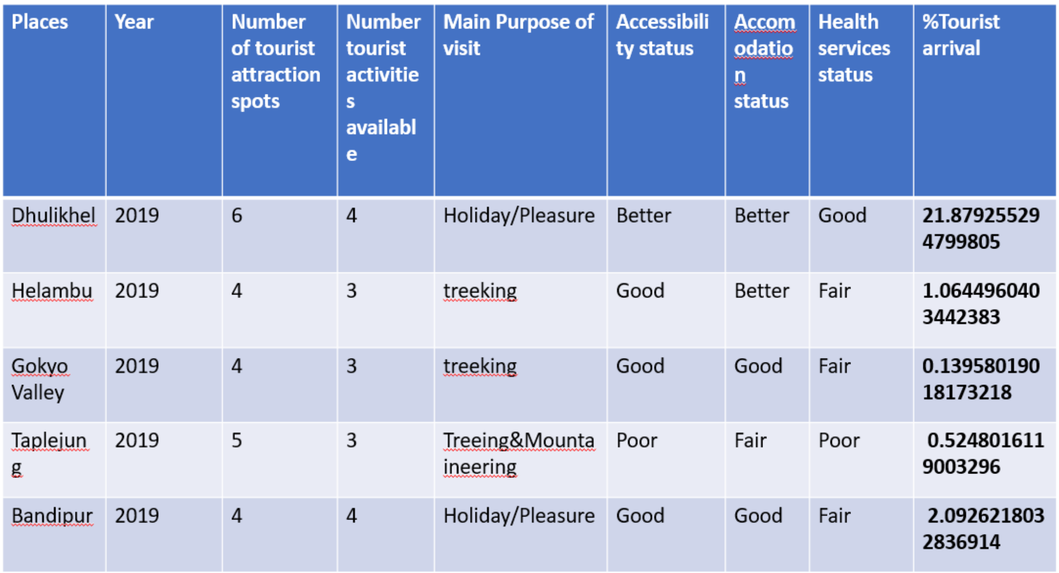


Figure 6.2: Result of SARIMA model

Table 6.3: Testing for Multilayer Perceptron Model

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



Places	Year	Number of tourist attraction spots	Number tourist activities available	Main Purpose of visit	Accessibility status	Accomodation status	Health services status	%Tourist arrival
Dhulikhel	2019	6	4	Holiday/Pleasure	Better	Better	Good	**21.879255294799805**
Helambu	2019	4	3	treeking	Good	Better	Fair	**1.0644960403442383**
Gokyo Valley	2019	4	3	treeking	Good	Good	Fair	**0.13958019018173218**
Taplejung	2019	5	3	Treeing&Mountaineering	Poor	Fair	Poor	**0.5248016119003296**
Bandipur	2019	4	4	Holiday/Pleasure	Good	Good	Fair	**2.0928218032836914**

Figure 6.3: Result of MLP model

7. CONCLUSION

The project activities were focused basically on data collection and preparations along with basic user interface design of the web application. Data collection and preparations tasks usually consumed more time as this required detail consideration of statistical reports and conversion and storage to suitable format before performing data analysis and predictions tasks. The user interface design also involved the understanding of what the application layout looks like and what functionalities should be there exactly in implementation of project goals.

7.1. Limitation

- The web application support only two-level role (i.e. Admin and Guest). There is no guest or user account and admin need to register to this system separately before actually using system
- The analysis and predictions highly depend on the accuracy of data in the data sources. Also, the latest updates depend only on update made by the data source

7.2. Future Enhancement

- Enable application to keep track of guest most visited pages, most downloaded contents
- Enhance admin functionalities like creating other admin registrations, easy admin communications, application usage analytics etc.
- Enhance API functionalities like extract latest data from direct data sources, keep track of API request, responses and report it to admin

REFERENCES

[1] World Travel & Tourism Council. The WTTC report: Travel and Tourism Economic Impact 2018 world [Accessed: 02-Jan-2019]

[2] World Travel & Tourism Council. The WTTC report: Travel and Tourism Economic Impact 2017 Nepal [Accessed: 02-Jan-2019]

[3] Ministry of Culture,Tourism & Civil Aviation,Nepal. [Online]. Available: http://tourism.gov.np/downloads. [Accessed: 02-Jan-2019]

[4] D. C. Frechtling, Forecasting Tourism Demand: Methods and Strategies, Oxford: Butterworth-Heinemann, 2001. [Accessed: 02-Jan-2019]

[5] Marine-Roig, E. & AntonClave, S. Tourism analytics with massive user generated content: A case study of Barcelona. Journal of Destination Marketing & Management (2015) [Accessed: 02-Jan-2019]

[6] Xiang, Zheng, and Daniel R. Fesenmaier. Analytics in Smart Tourism Design Concepts and Methods. Springer International Publishing, 2018. [Accessed: 02-Jan-2019]

[7] "Redirecting...",Analytics.google.com,2019.[Online].Available:https://analytics.google.com/analytics/web/. [Accessed: 02- Jan- 2019].

[8] "Destination Analysts | Tourism Market Research", Destination Analysts, 2019. [Online]. Available: http://www.destinationanalysts.com/. [Accessed: 02- Jan-2019].

[9] Shumway R.H., Stoffer D.S. (2000) Time Series Regression and ARIMA Models. In: Time Series Analysis and Its Applications. Springer Texts in Statistics. Springer, New York, NY [Accessed: 02-Jan-2019]

[10] 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 [Accessed: 02-Jan-2019]

[11] Subedi, A. (2017). Time series modeling on monthly data of tourist arrivals in Nepal: An alternative approach. Nep. J. Stat., 1, 41-54[Accessed: 02-Jan-2019]

[12] Roman Josue de las Heras Torres, “7 Ways Time Series Forecasting Differs from Machine Learning,” Oracle DataScience.com. [Online]. Available: https://www.datascience.com/blog/time-series-forecasting-machine-learning-differences. [Accessed: 02-Jan-2019].

[13] Dorffner, G. 1996, Neural Networks for Time Series Processing. Neural Network World [Accessed: 02-Jan-2019]

[14] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny. An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29(5-6):594, 2010. [Accessed: 02-Jan-2019]

[15] Voyant, Cyril & Nivet, Marie-Laure & Paoli, Christophe & Muselli, Marc & Notton, G. (2014). Meteorological time series forecasting based on MLP modelling using heterogeneous transfer functions. [Accessed: 02-Jan-2019]

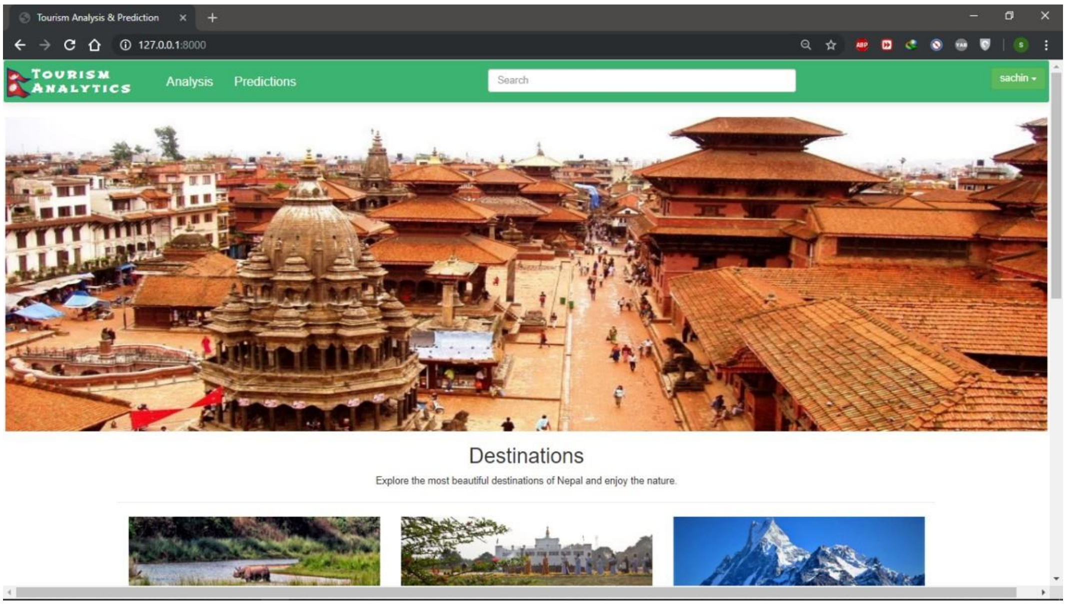
[16] C. Pelletier, A. Almalaq, and D. M. de Lachapelle, “Deep Learning for Time Series Forecasting,” Machine Learning Mastery. [Online]. Available: https://machinelearningmastery.com/deep-learning-for-time-series-forecasting. [Accessed: 02-Jan-2019].

[17] Civil Aviation Authority of Nepal. [Online]. Available: http://caanepal.gov.np/. [Accessed: 02-Jan-2019].

[18] Raschka, Sebastian. Python Machine Learning. Packt, 2015. [Accessed: 02-Jan-2019]

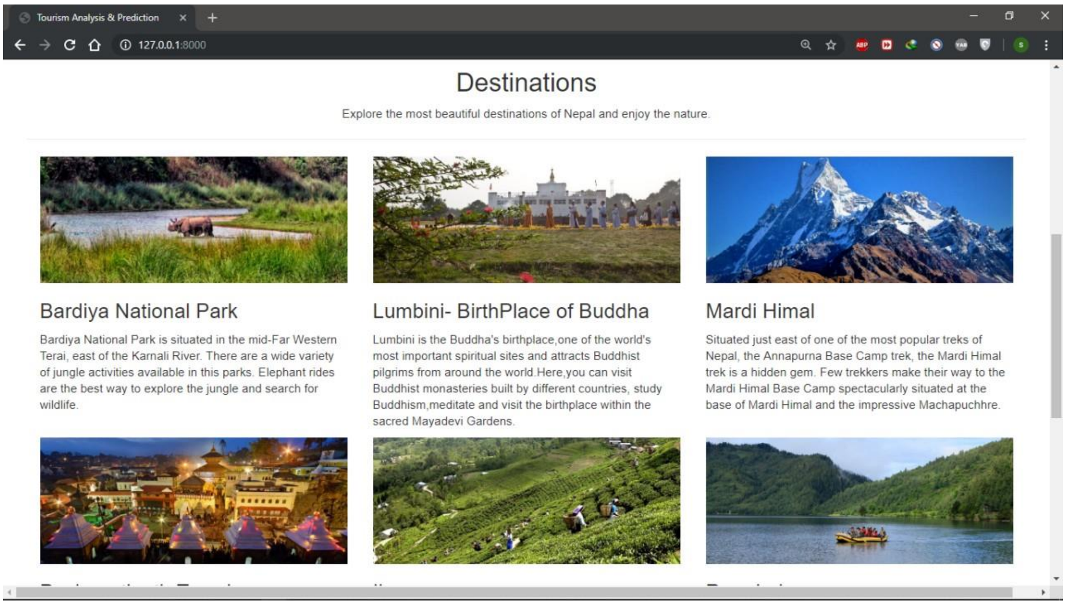
APPENDIX

Snapshot of project



Destinations

Explore the most beautiful destinations of Nepal and enjoy the nature.



Destinations

Explore the most beautiful destinations of Nepal and enjoy the nature.

Bardiya National Park

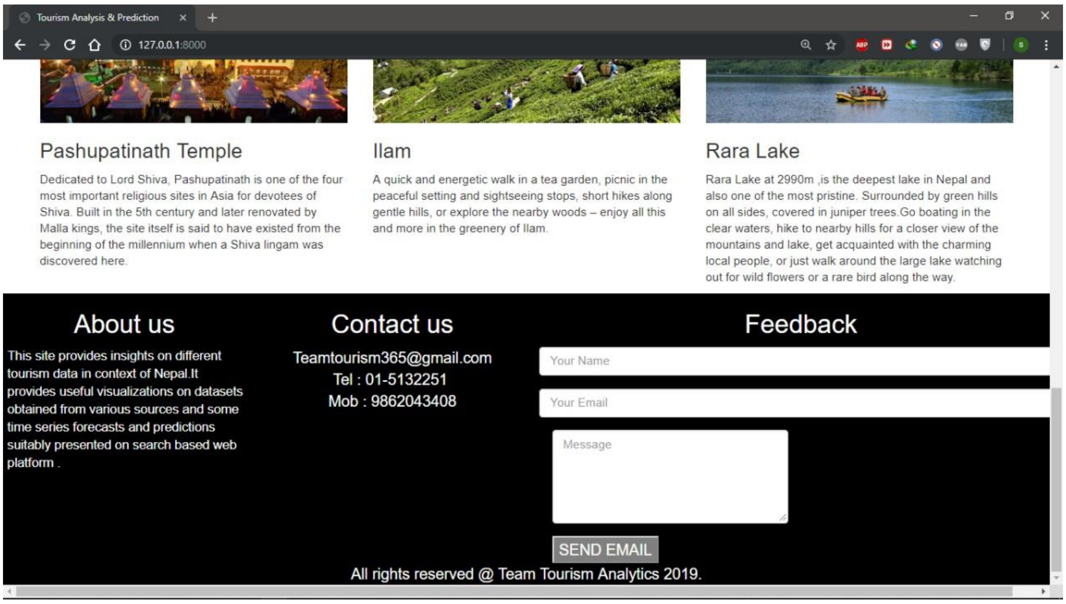
Bardiya National Park is situated in the mid-Far Western Terai, east of the Karnali River. There are a wide variety of jungle activities available in this parks. Elephant rides are the best way to explore the jungle and search for wildlife.

Lumbini- BirthPlace of Buddha

Lumbini is the Buddha's birthplace,one of the world's most important spiritual sites and attracts Buddhist pilgrims from around the world.Here,you can visit Buddhist monasteries built by different countries, study Buddhism,meditate and visit the birthplace within the sacred Mayadevi Gardens.

Mardi Himal

Situated just east of one of the most popular treks of Nepal, the Annapurna Base Camp trek, the Mardi Himal trek is a hidden gem. Few trekkers make their way to the Mardi Himal Base Camp spectacularly situated at the base of Mardi Himal and the impressive Machapuchhre.



Pashupatinath Temple

Dedicated to Lord Shiva, Pashupatinath is one of the four most important religious sites in Asia for devotees of Shiva. Built in the 5th century and later renovated by Malla kings, the site itself is said to have existed from the beginning of the millennium when a Shiva lingam was discovered here.

Ilam

A quick and energetic walk in a tea garden, picnic in the peaceful setting and sightseeing stops, short hikes along gentle hills, or explore the nearby woods – enjoy all this and more in the greenery of Ilam.

Rara Lake

Rara Lake at 2990m ,is the deepest lake in Nepal and also one of the most pristine. Surrounded by green hills on all sides, covered in juniper trees.Go boating in the clear waters, hike to nearby hills for a closer view of the mountains and lake, get acquainted with the charming local people, or just walk around the large lake watching out for wild flowers or a rare bird along the way.

About us

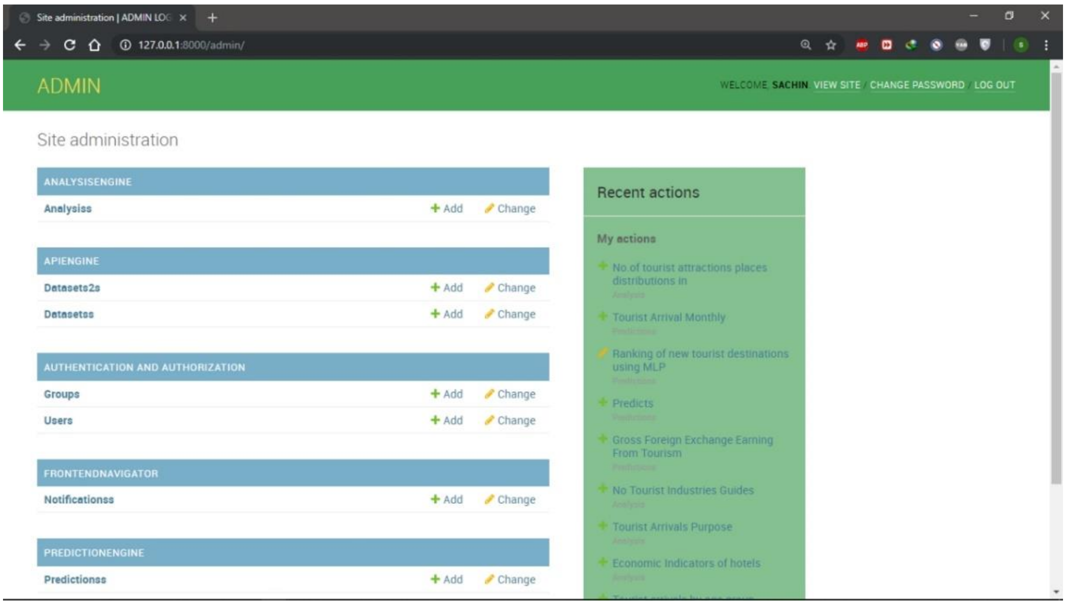
This site provides insights on different tourism data in context of Nepal.It provides useful visualizations on datasets obtained from various sources and some time series forecasts and predictions suitably presented on search based web platform .

Contact us

Teamtourism365@gmail.com
Tel : 01-5132251
Mob : 9862043408

Feedback

All rights reserved @ Team Tourism Analytics 2019.



Site administration | ADMIN LOG
127.0.0.1:8000/admin/
ADMIN
WELCOME SACHIN VIEW SITE / CHANGE PASSWORD / LOG OUT

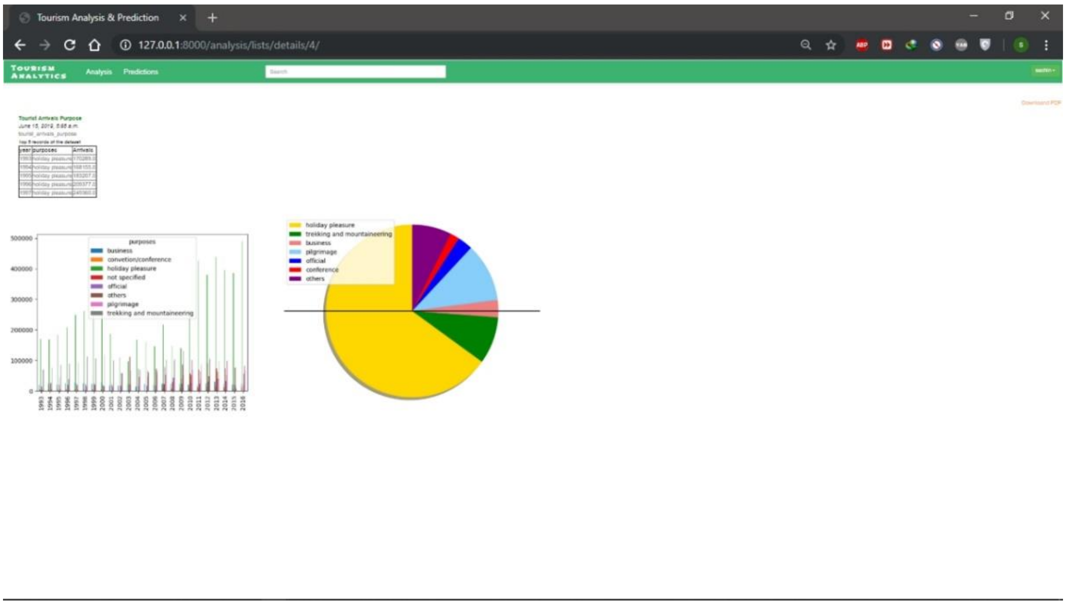
Site administration

ANALYSISENGINE		
Analysiss	+ Add	Change
APIENGINE		
Datasets2s	+ Add	Change
Datasetss	+ Add	Change
AUTHENTICATION AND AUTHORIZATION		
Groups	+ Add	Change
Users	+ Add	Change
FRONTENDNAVIGATOR		
Notificationss	+ Add	Change
PREDICTIONENGINE		
Predictionss	+ Add	Change

Recent actions

My actions

- No.of tourist attractions places distributions in
  Analysis
- Tourist Arrival Monthly
  Predictions
- Ranking of new tourist destinations using MLP
  Predictions
- Predicts
  Predictions
- Gross Foreign Exchange Earning From Tourism
  Predictions
- No Tourist Industries Guides
  Analysis
- Tourist Arrivals Purpose
  Analysis
- Economic Indicators of hotels
  Analysis



Tourism Analysis & Prediction
127.0.0.1:8000/analysis/lists/details/4/
TOURISM ANALYTICS
Analysis
Predictions
Search
Download PDF

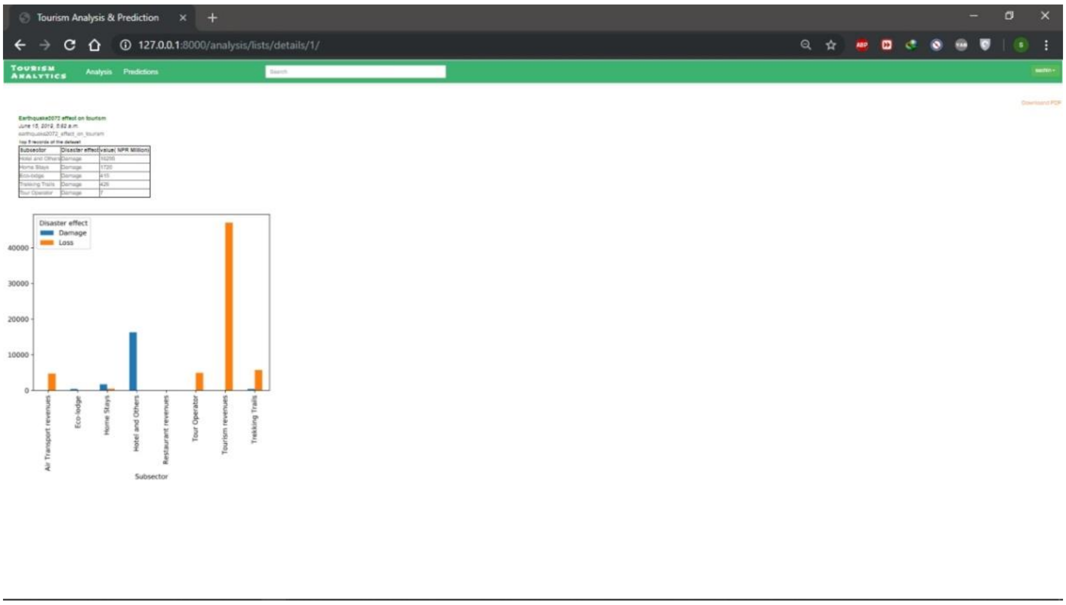
Tourist Arrivals Purpose

June 15, 2019, 5:55 a.m.

tourist_arrivals_purpose

top 5 records of the dataset

year	purposes	Arrivals
1993	holiday pleasure	170289.0
1994	holiday pleasure	168155.0
1995	holiday pleasure	183207.0
1996	holiday pleasure	209377.0
1997	holiday pleasure	249360.0



Tourism Analysis & Prediction
127.0.0.1:8000/analysis/lists/details/1/
TOURISM ANALYTICS
Analysis
Predictions
Search
Download PDF

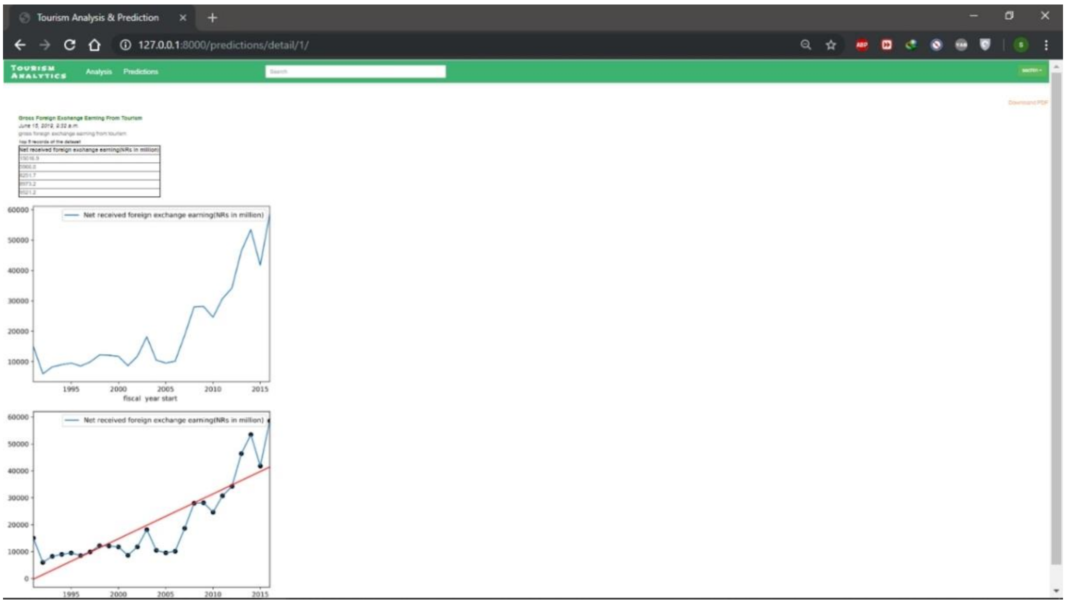
Earthquake2072 effect on tourism

June 15, 2019, 5:52 a.m.

earthquake2072_effect_on_tourism

top 5 records of the dataset

Subsector	Disaster effect	value(NPR Million)
Hotel and Others	Damage	16250
Home Stays	Damage	1720
Eco-lodge	Damage	415
Trekking Trails	Damage	426
Tour Operator	Damage	7


Tourism Analysis & Prediction
127.0.0.1:8000/predictions/detail/1/
Tourism Analytics
Analysis
Predictions
Search
sachin
Download PDF

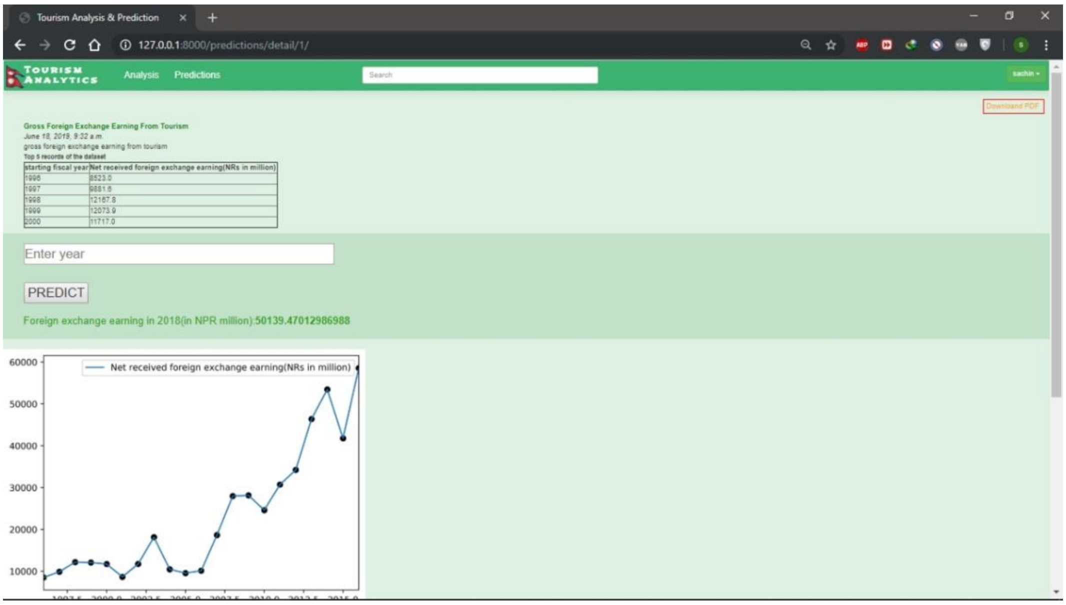
Gross Foreign Exchange Earning From Tourism

June 18, 2019, 9:32 a.m.

gross foreign exchange earning from tourism

Top 5 records of the dataset

Net received foreign exchange earning(NRs in million)
15016.9
5966.0
8251.7
8973.2
9521.2


Tourism Analysis & Prediction
127.0.0.1:8000/predictions/detail/1/
Tourism Analytics
Analysis
Predictions
Search
sachin
Download PDF

Gross Foreign Exchange Earning From Tourism

June 18, 2019, 9:32 a.m.

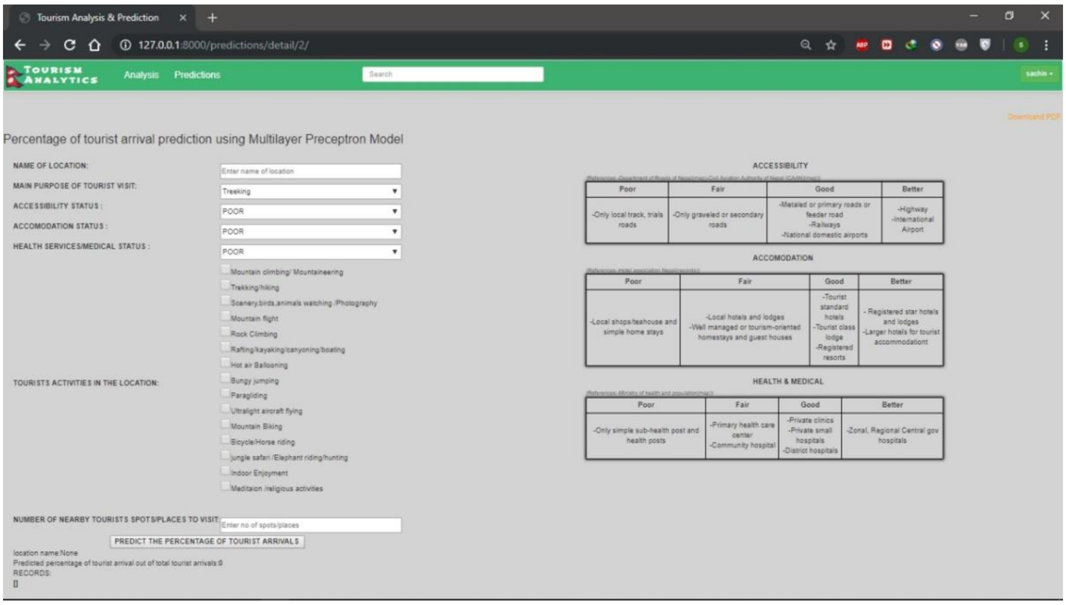
gross foreign exchange earning from tourism

Top 5 records of the dataset

starting fiscal year	Net received foreign exchange earning(NRs in million)
1996	8523.0
1997	9881.6
1998	12167.8
1999	12073.9
2000	11717.0

Foreign exchange earning in 2018(in NPR million):50139.47012986988

60000
50000
40000
30000
20000
10000
Net received foreign exchange earning(NRs in million)


Tourism Analysis & Prediction
127.0.0.1:8000/predictions/detail/2/
TOURISM ANALYTICS
Analysis
Predictions
Search
sachin
Download PDF

Percentage of tourist arrival prediction using Multilayer Preceptron Model

NAME OF LOCATION: Enter name of location

MAIN PURPOSE OF TOURIST VISIT: Treeking

ACCESSIBILITY STATUS : POOR

ACCOMODATION STATUS : POOR

HEALTH SERVICES/MEDICAL STATUS : POOR

TOURISTS ACTIVITIES IN THE LOCATION:

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

NUMBER OF NEARBY TOURISTS SPOTS/PLACES TO VISIT: Enter no of spots/places

PREDICT THE PERCENTAGE OF TOURIST ARRIVALS

location name:None
Predicted percentage of tourist arrival out of total tourist arrivals:0
RECORDS:
[]

ACCESSIBILITY

(Reference: Department of Roads of Nepal(map)-Civil Aviation Authority of Nepal (CAAN)(map))

Poor	Fair	Good	Better
-Only local track, trials roads	-Only graveled or secondary roads	-Metaled or primary roads or feeder road -Railways -National domestic airports	-Highway -International Airport

ACCOMODATION

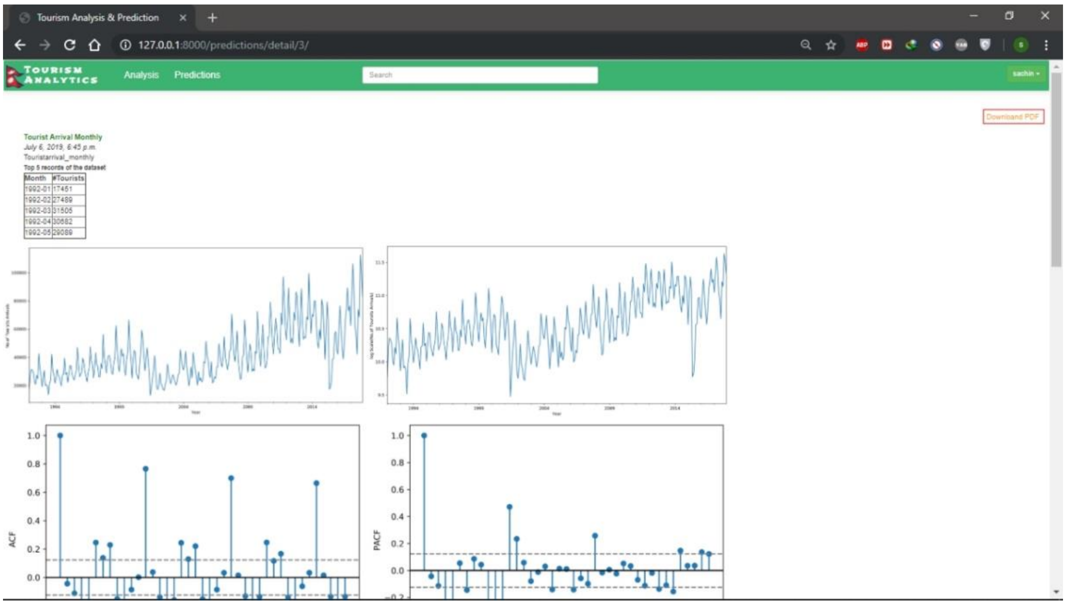
(Reference: Hotel association Nepal(reports))

Poor	Fair	Good	Better
-Local shops/teahouse and simple home stays	-Local hotels and lodges -Well managed or tourism-oriented homestays and guest houses	-Tourist standard hotels -Tourist class lodge -Registered resorts	- Registered star hotels and lodges -Larger hotels for tourist accommodationt

HEALTH & MEDICAL

(Reference: Ministry of health and population(map))

Poor	Fair	Good	Better
-Only simple sub-health post and health posts	-Primary health care center -Community hospital	-Private clinics -Private small hospitals -District hospitals	-Zonal, Regional Central gov hospitals



Tourist Arrival Monthly
July 6, 2019, 6:45 p.m.
Touristarrival_monthly
Top 5 records of the dataset

Month	#Tourists
1992-01	17451
1992-02	27489
1992-03	31505
1992-04	30682
1992-05	29089