SCHOOL OF COMPUTER SCIENCE

GROUP ASSIGNMENT TASK 1 (30%)

September 2024 – SEMESTER 6

Module Code:	**ITS 69504**
Module Name:	**Data Analytics and Machine Learning**
Deadline:	**30th October 9th October 2024, 23:59 PM (NPT) via MyTIMeS Portal**

Laboratory 3 Group 6 **Module Leader: Mr. Anmol Adhikari**

Student Declaration: We Declare That –

- ✓ **I confirm my awareness about the university's regulations, governing cheating in tests and assignments, and form the guidance issued by the school of computing and it concerning plagiarism and proper academic practices, and the assessed work now submitted is in accordance with this regulation and guidance.**
- ✓ **I understand that, unless already agreed with the school of computing and it, that the assessed work has not been previously submitted, either in whole or in part, in this or any other institution.**
- ✓ **I recognize that should evidence emerge that my work fails to comply with either of the above declarations, then i may be liable to proceeding under regulation.**

S. No.	Student Full Name	University ID	Signatures	Scores
1.	Amogh Shakya	036 2073		_____ / 100
2.	Mamata Shrestha	036 2047		_____ / 100
3.	Monalisha Thapa Magar	036 2948		_____ / 100
4.	Rahul Wasti	036 2512		_____ / 100
5.	Sujal Ratna Tuladhar	036 2483		_____ / 100

DECLARATION

- ✓ I pledge to be respectful and supportive of my team members.
- ✓ I pledge to abide by the deadline set by my lecturer and team members.

S. NO.	STUDENT NAME & ID	WORK BREAKDOWN	SIGNATURE
1.	**Amogh Shakya | 036 2073**	2.3. Comparative and Critical Analysis, 3.5. Evaluation Matrices, 3.5.1. Linear Regression, 3.5.2. Random Forest, 5.4. Impact of Visualizations on Decision-Making	
2.	**Mamata Shrestha | 036 2047**	2.3.2. Existing Tourism Management, 3.1.1. Import Libraries, 3.1.2. Data Loading, 3.1.3. Data Structure, 4.1. Processing Techniques, 4.2. Technology Stack Justification, 5.2. Visualization Tools for Complex Tourism Data	
3.	**Monalisha Thapa Magar | 036 2948**	2.2. Dataset Relevance to Scope, 3.3.1. Dimensionality Reduction, 3.3.2. Target and Features, 5.2. Visualization Tools for Complex Tourism Data	
4.	**Rahul Wasti | 036 2512**	2.3.1. State-of-the-Art Solutions, 3.3.3. Standard Scaler, 3.4. Selection Algorithms, 3.4.1. Linear regression, 3.4.2. Random Forest Regressor, 3.4.3. XGBoost, 5.3. Insights for Data Analysis	
5.	**Sujal Ratna Tuladhar | 036 2483**	2.1. Problem Identification, 3.1.4. Data Preprocessing, 3.2. Feature Engineering, 3.3. Data Preparation, 5.1. Reporting Frameworks for Key Performance Indicators	

- Provide A Clear Work Breakdown Structure To Describe What Each Member Is Doing.

Table of Contents

Table of Figures

Table of Tables

Table of Codes

CASE STUDY ON TRAVEL AND TOURISM: TREKKING IN NEPAL

Abstract:

This study investigates challenges and opportunities within the tourism sector in Nepal, focusing on the trekking industry as a vital economic pillar. Utilizing a dataset covering diverse aspects of visitor demographics, costs, and preferences. We try to address several barriers related to sustainable tourism management, including lack of consistency, increasing operational costs, health and safety concerns, and the need for sustainable growth practices. The report proposes a data-driven solution making use of machine learning and regression models, for predictive insights and trend analysis. Evaluation metrics such as ereor and score demonstrate the efficacy of these models in forecasting costs and visitor preferences with high accuracy, particularly through the XGBoost model. Furthermore, visualization tools are leveraged to present complex tourism data in an actionable format for stakeholders. Key recommendations include the integration of real-time monitoring, enhancing safety protocols, and data-based marketing strategies to support informed decision making and promote sustainable development.

Keywords:

Travel and Tourism, Trekking, Machine Learning, Artificial Intelligence, Regression Model, Linear Regression, Data Analysis, Random Forest Regression, XGBoost, Mean Absolute Error, Mean Square Error, R^2

1. Introduction to Case Study

1.1. Background Overview

1.1.1. Travel and Tourism Sector

Nepal is regarded as the tourist destination hub, with most complete and unmatched natural beauties, geographical attributes glittered with plethora of cultural, traditional and religious heritage. Nepal is host to numerous mountain peaks “ceiling of the worlds”, recognized world heritage sites. Drawing in millions of tourists every year for activities like recreations, trekking, mountaineering, adventurous games and religion. Tourism sector has made a considerable impact in the economy of Nepal, being one of the primary drivers of socio-economic development of the country.

Developing a sustainable and responsible infrastructure for tourism while displaying and establishing the country as a safe, attractive, exciting, unique location helps in surging influx of visitors and their activities is a high priority policy set by the ministry. also directly helping in creating employment opportunities in rural areas, equitable distribution of benefits of tourism at grass-root level and establishing the tourism sector as the foundation of prosperous Nepal. Due to covid-19 global pandemic, but with prudent plans and policies formulated by the ministry in coordination with stakeholders, the tourism sector is rejuvenated showing great resilience.

The ministry is also working on economic mainstreaming of tourism by expanding and diversifying the product offering and destination marketing in the national and international arena focusing on quality-focused and purpose driven tourism. In order to promote the sector, it calls for an integrated database with details about tourists, their activities, average duration of stay, revenue, and expenditures, foreign capital gains and other information related to tourists.

1.1.2. Trekking Industry

Nepal's situation is unique as it is provided with rich geographical attractions, incomparable spectrum location altitude from highest of mountains to lowest of fields. placed between two fastest growing giant countries: China and India. The tourism sector may experience a boost due to some competitive advantage, which also instruments in providing economic prosperity for the citizens by earning foreign currency.

Beginning of year 2022 till now when covid restrictions are lifted and the world is moving on, Nepal has shown signs of recovery as the government is focused on developing airports, roads, trails, and other infrastructural facilities. such actions are strategized by the ministry, committed to uplift the living standard of people and improve daily livelihood.

1.2. Purpose

- **Understanding Demographics of Travelers:** Insights into trekkers' age, nationality, and health profiles will help tourism providers better tailor services and marketing efforts to meet the needs of specific traveler groups.
- **Analyzing Costs and Expenses:** The breakdown of the costs of the trekking-undoing permit and expenses on accommodation-will be helpful for travelers with regards to budgeting. Businesses also need this data as it helps them ensure that they have competitively-priced services for the affordability of the services offered.
- **Deriving the Trek Condition Information:** This helps the recommendation of package provided to be matched closer to the level of experience of the traveler by judging the difficulty of the trek, duration, and travel timings. It also keeps insights into resource allocation and infrastructure planning informed.
- **Informed Decision Making for Tourism Management:** Data-driven insights will allow tourism managers to align services with traveler demand, supporting strategic planning and resource distribution.
- **Informing Policy Improvement:** The findings shall enlighten the policy makers on the areas of improvement in safety, regulations, and environmental practices. Improvement in policy will surely guarantee further sustainable growth within the trekking industry in Nepal.

- **Allowing Business Optimization and Growth:** Understanding the trends and pattern of traveler/visitors in tourism and trekking will definitely allow business people to optimize such operations and improve their services correspondingly. These insights further encourage sustainable growth and also ensure there is a great improvement in tourist experience.

2. Problem Definition

2.1. Problem Identification

Nepal is lacking behind in such tourism related fields although the country has abundance of resources, natural beauty heritage and scenery to take advantage off "Nepal is also considered a paradise for trekkers" they have 8 of the 10 tallest mountains in the world due to tens of millions of years where the Indian subcontinental plate has crashed and collided with the Eurasian tectonic plate. Although these peaks are shared between Nepal and Tibet (autonomous region of China), due to harsh conditions form the Tibetan side Nepal is considered as the starting place for the adventure.

The tourism sector in Nepal, while holding a lot of potential, faces a multitude of challenges, from inconsistent service quality to rising operational costs. Nepal's tourism sector, particularly trekking, is crucial to the country's economy, yet sustainability and efficient management is obstructed by the following issues:

- **Inconsistent Services:** Visitors experience varied service quality, impacting their overall satisfaction.
- **Rising Costs:** Remote areas with limited infrastructure drive costs higher and contribute to the expense of visiting Nepal, potentially discouraging tourists.
- **Health and Safety Concerns:** Trekking brings unique concerns; lack of consistent health and safety measures can make it challenging to ensure safety and maintain tourist confidence.
- **Sustainable Growth:** Rapid increase in tourism strain resources, while sustainable practices are essential to preserve Nepal's natural and cultural heritage.

To address these challenges, a data-driven approach that combines machine learning, predictive analysis, and real-time data monitoring can offer solutions. This could involve utilizing tourism data analytics to improve service consistency, cost efficiency, safety, and sustainability efforts.

2.2.Dataset Relevance to Scope

The available dataset provides valuable insights into various aspects of Nepalese tourism, particularly in trekking routes, travel costs, and trekker demographics. The dataset provided has been scrapped from online sites related to trekking and specially focus on Nepali locations. This dataset requires a lot of cleaning. The dataset provided is has multiple key columns that tells a specific information at that instance.

the columns provided in the dataset contains 383 different types of record related to Nepal Trekking.

Table 1. Column Descriptions

Column	Description
Trek	The name or route of the trek.
Cost	Estimated cost of the trek
Time	The duration of the trek in days
Trip Grade	The difficulty of the trek
Max Altitude	The max altitude reached during the trek, in meters
Accomodation	Type of accommodation the trekkers stayed in during the trek
Best Travel Time	Optimal months for the trek
Date of Travel	The actual date the trek was undertaken
Sex	Gender of the trekker (Male, Female, Transgender, Non-Binary)
Age	The age of the trekker
Employment Type	The sector where the trekker is employed (Government or Private)
Graduate or Not	The graduation status of the trekker
Annual Income	The annual income of the trekker in USD
Family Members	Number of family members of the trekker
Chronic Diseases	Status indicating whether the trekker has any chronic diseases or not
Frequent Flyer	Status indicating if the traveler is a frequent flier (Yes/No)
Ever Travelled Abroad	Indicates if the trekker has traveled abroad before (Yes/No)
Travel Insurance	Indicates whether the trekker has travel insurance (1 for Yes, 0 for No)
Insurance Year	indicates when the insurance was filled
Regional Code	Regional code for the trekker's place of origin
Country	Country of origin of the trekker

2.3.Comparative and Critical Analysis

Few innovative approaches in tourism management have been rising and have pointed out the effectiveness and adaptability of machine learning in addressing the existing problems in tourism management.

2.3.1. State-of-the-Art Solutions

2.3.1.1. Data-driven Personalized Tourist Recommender System

A recent article, Personalized Tourist Recommender System: A Data-Driven and Machine-Learning Approach, by (Shrestha et al., 2024), introduces an advanced, data-driven recommendation system tailored specifically for tourism in Pokhara, Nepal. The objective of this research is to create a Tourist Recommender System that leverages machine learning to deliver highly personalized travel recommendations based on extensive tourist data. Its approach revolves around four crucial data categories: demographics, travel planning behavior, spending habits, preferences, and satisfaction indicators—each adopted in providing a holistic profile for every tourist.

For machine learning, the following seven algorithms were studied: k-Nearest Neighbors, Decision Trees, Support Vector Machines, Random Forest, Gradient Boosting, Naïve Bayes, and Neural Networks. After much testing and evaluation, it appeared that both Decision Trees and Gradient Boosting performed with really high accuracy, precision, and reliability during the training and testing phases.

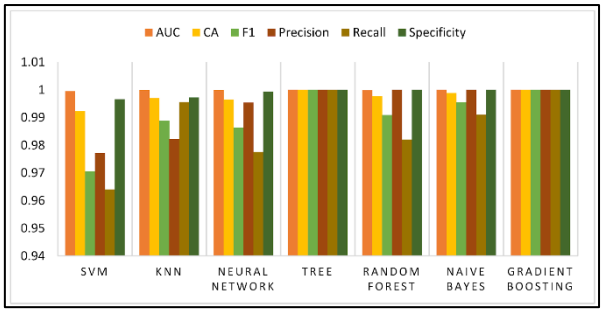


Figure 1. Performance Eval of the TRS

For validation, their TRS was benchmarked against some of the widely used platforms like TripAdvisor and Google Maps. It was then found that the proposed system performed not only more accurately but also in a way that produced recommendations more contextually relevant to a particular region of Pokhara, filling a gap in existing recommendation platforms.

This study highlights the potential of tailored, data-driven solutions in the tourism industry, especially for regions with unique geographic and demographic characteristics.

2.3.2. Existing Tourism Management

The Nepal Tourism Board , established in collaboration with the private sector, is the primary governing body for tourism promotion. NTB organizes campaigns such as “Visit Nepal” to boost tourism while working closely with private stakeholders to develop targeted initiatives. However, gaps remain in data-driven decision making and targeted service improvements. Integrating analytics and predictive insights could address these gaps fostering more resilient operations.

2.3.2.1. Trekking Route Management and Safety Initiatives

Given that trekking is central to Nepal's tourism, NTB, local government units, and private trekking agencies manage and maintain popular trails such as the Annapurna Circuit and Everest Base Camp. This may include trail maintenance, provision of guide and porter services, and setting up checkpoints to monitor and track activities for safety. High-altitude trekking, however, is not without risks. While some health-care facilities are found along the major routes, access to advanced medical services is still limited in most remote regions.

2.3.2.2. Sustainable Tourism and Conservation Initiatives

Sustainable tourism is a critical focus within the trekking industry, with initiatives aimed at balancing tourist arrivals with environmental preservation. This includes conservation efforts by organizations like the Annapurna Conservation Area Project (ACAP) and Sagarmatha National Park (SNP), which are responsible for maintaining biodiversity and regulating trekking activities while encouraging local businesses to minimize environmental impact. These organizations emphasize responsible tourism, encourage visitors to respect local culture and nature, and actively engage in community-based programs to distribute tourism benefits to local residents (National Trust for Natural Conservation, n.d.).

2.3.2.3. Marketing and Promotion Gaps

Digital transformation in Nepal's tourism seems to be in the early stages. While NTB has started some digital marketing activities to reach audiences globally, embedding advanced analytics and real-time monitoring systems still seems to be very restricted. Social media platforms are used to attract tourists, yet there is potential for further development in data analytics to understand traveler behavior, predict tourism demand, and customize marketing based on demographic insights.

2.3.2.4. Gaps in Health and Safety Measures

Lack of standardized health and safety measures is a pressing issue in Nepal's tourism management. While some trekking areas are well-regulated, safety protocols, emergency response, and healthcare access are inconsistent, especially in high-altitude regions. Trekkers' safety could be significantly improved by adopting GPS/GPRS tracking systems, such as those used in Mahabir Pun's 'eTag' initiative, which would allow for real-time monitoring and emergency response (Pandey & Dhakal, 2019, 12).

3. Data Modelling

Data modeling is the approach of creating a mathematical representation of a system based on real world data. Simply put, it is the process of creating an abstract representation of data and its relationship to solve specific problems. (IBM, n.d.)

3.1. Outline for Approach

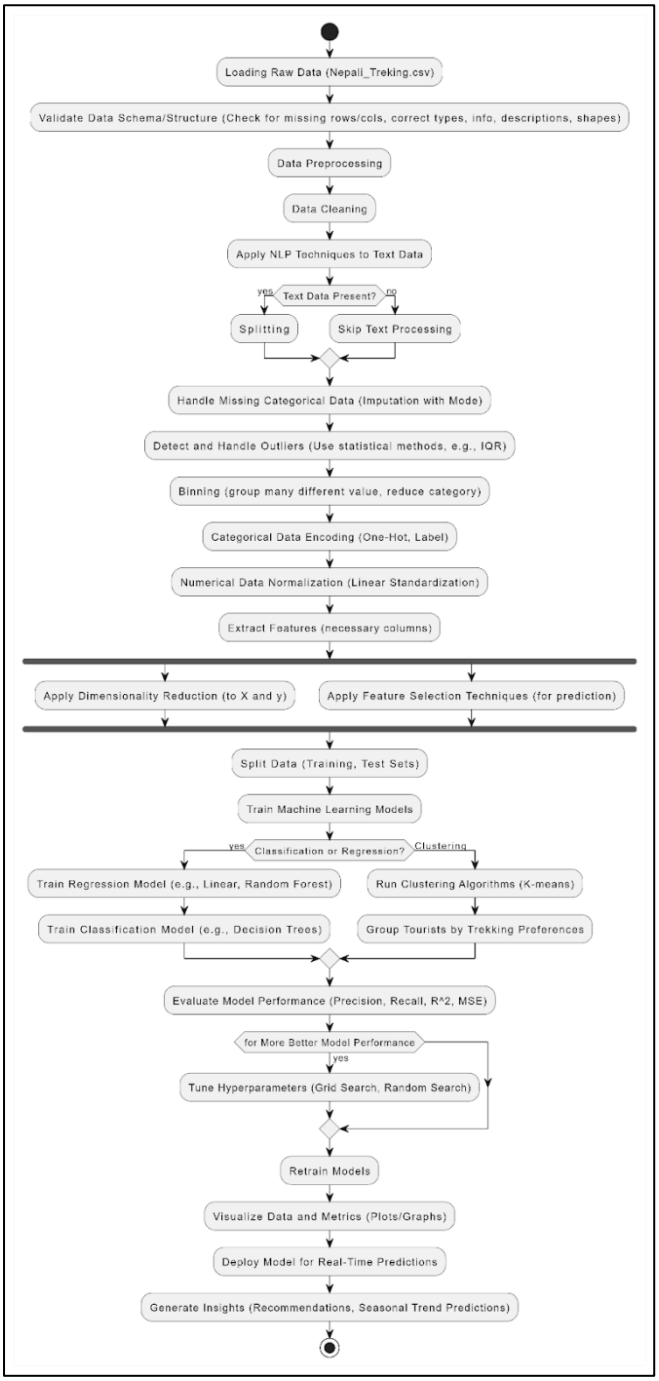


Figure 2. Flow Diagram of Approach

3.1.1. Import Libraries

```python
import pandas as pd
import numpy as np

import re

import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LinearRegression
from sklearn.tree import plot_tree
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

import xgboost as xgb

import warnings
warnings.filterwarnings("ignore")

from google.colab import drive
drive.mount('/content/drive')
```

Code 1, Libraries and Modules Imported

this step is necessary for understanding what libraries are used in overall project

3.1.2. Data Loading

Data loading refers to importing data from external sources, such as CSV files or other formats, so that it can be analyzed, processed, or visualized.

```python
df = pd.read_csv("Nepali_Treking.csv")
```

Code 2. Loading Dataset

3.1.3. Data Structure

The Nepali Trekking dataset is provided as a comma separated values (CSV) file containing different trek details in 21 columns. The columns in the provided dataset contains 383 different types of records related to Nepal Trekking. we can use different functions and properties that are in-built or provided by the respective libraries.

- pd.set_option()

```
[5]: pd.set_option("display.max_columns", None)
     # pd.set_option("display.max_rows", None)
     # pd.set_option('display.max_colwidth', None)
```

Code 3. to view max columns

the code above helps to view all present cloumns.

- .iloc()

```
[8]: df.iloc[:5]
```

[8]:

	Unnamed: 0	Trek	Cost	Time	Trip Grade	M Altitu
0	0	Everest Base Camp Trek	\n$1,420 USD	16 Days	Moderate	5545

Code 4. using iloc

by using iloc property we can view row with the interger location

- .drop

```
[11]: df = df.drop(columns={'Unnamed: 0'}, axis=1)
      df.head(5)
```

[11]:

	Trek	Cost	Time	Trip Grade	Max Altitude	Accomodation	B Tra T
0	Everest Base Camp Trek	\n$1,420 USD	16 Days	Moderate	5545 m	Hotel/Guesthouse	Ma - M Se
1	Everest Base Camp Short Trek	\n$1,295 USD	14 Days	Moderate	5545 m	Hotel/Guesthouse	Ma - M Se

Code 5. dropping existing index

drop the unnamed column that gives extra index column which the pandas dataframe already provide

```
[14]: df.drop_duplicates(inplace=True)
```

Code 6. dropping duplicates if any

drop duplicate values just in case to avoid inaccuracies with their non random sample, prevent bias and reliable for modelling and avoid overfitting.

- .info()

```
[6]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 383 entries, 0 to 382
Data columns (total 22 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Unnamed: 0          383 non-null    int64
 1   Trek                383 non-null    object
 2   Cost                383 non-null    object
 3   Time                383 non-null    object
 4   Trip Grade          383 non-null    object
 5   Max Altitude        383 non-null    object
 6   Accomodation        383 non-null    object
 7   Best Travel Time    383 non-null    object
 8   Date of Travel      383 non-null    object
 9   Sex                 383 non-null    object
 10  Age                 383 non-null    int64
 11  Employment Type     383 non-null    object
 12  GraduateOrNot       383 non-null    object
 13  AnnualIncome        383 non-null    int64
 14  FamilyMembers       383 non-null    int64
 15  ChronicDiseases     383 non-null    int64
 16  FrequentFlyer       383 non-null    object
 17  EverTravelledAbroad 383 non-null    object
 18  TravelInsurance     383 non-null    int64
 19  Year                383 non-null    int64
 20  Regional code       383 non-null    int64
 21  Country             383 non-null    object
dtypes: int64(8), object(14)
memory usage: 66.0+ KB
```

Code 7. dataframe info

used for summarizing dataframe in use by showing range of entries, total number of column, column names, count of nun-null values, data type of each column, and memory usage.

- .describe()

```
[11]: df.describe()
```

[11]:

	Age	AnnualIncome	FamilyMembers	ChronicDiseases	TravelInsurance	Year	Regional code
count	383.000000	3.830000e+02	383.000000	383.000000	383.000000	383.000000	383.000000
mean	29.673629	9.492167e+05	4.843342	0.268930	0.344648	2025.519582	4081.302872
std	2.868042	3.720953e+05	1.667836	0.443983	0.475875	175.953867	2432.427473
min	25.000000	3.000000e+05	2.000000	0.000000	0.000000	2016.000000	0.000000
25%	28.000000	6.500000e+05	4.000000	0.000000	0.000000	2016.000000	2030.000000
50%	29.000000	9.000000e+05	5.000000	0.000000	0.000000	2017.000000	4140.000000
75%	32.000000	1.250000e+06	6.000000	1.000000	1.000000	2017.000000	6080.000000
max	35.000000	1.800000e+06	9.000000	1.000000	1.000000	5460.000000	9990.000000

Code 8. dataset describe

usually generate description of statistics of numeric values giving dispersion, distributions and central tendency

3.1.4. Data Preprocessing

In order to perform statistical analysis on the data, the dataset has to be purged of various different anomalies such as invalid data types or missing values.

```
[16]: def format_column(col):
          col = re.sub(r'\s+', '', col)
          col = re.sub('(?<!^)(?=[A-Z])', '_', col).lower()
          return col

      df.columns = [format_column(col) for col in df.columns]
      print(df.columns)

Index(['trek', 'cost', 'time', 'trip_grade', 'max_altitude', 'accomodation',
       'best_travel_time', 'dateof_travel', 'sex', 'age', 'employment_type',
       'graduate_or_not', 'annual_income', 'family_members',
       'chronic_diseases', 'frequent_flyer', 'ever_travelled_abroad',
       'travel_insurance', 'year', 'regionalcode', 'country'],
      dtype='object')
```

Code 9. coulmns to snake case

using regular expression to rename all columns to a standardized, easy to read, consistent. for now, it is in snake case format

3.1.4.1. Data Cleaning

each and every column and entries were analyzed and cleaned

- Trek

there were extra characters [\xa0, which is defined as non-breaking space in different encode] that were present on trek which were removed and turned to lower case

```
df = df.sort_values(by='trek', ascending=True)
# remove non breakable space
df['trek'] = df['trek'].str.replace('\xa0', '').lower()
```

Code 10. sort, clean trek

- Cost

The currency, unit and additional keyword [\n; new line break] can be stripped away from the values before converting the data into a floating type.

```
df['cost'] = df['cost'].str.replace(r"\W|\D|\s", "", regex=True).astype('float32')
exchange_rate = 134.20  # 1 USD to NPR as of Oct 27, 2023
df['cost_in_npr'] = df['cost_in_usd'] * exchange_rate
```

- Time

Since this column represents duration in days, the term “days” were removed from the values.

```
df['time'] = df['time'].apply(lambda x: re.sub(r'\b[Dd]ays?\b', '', x)).astype(int)
df.rename(columns={'time': 'time_in_days'}, inplace=True)
```

Code 11. regex, clean cost, rename

- Max Altitude

Similarly, the unit “m” (meters) can be removed from “Max Altitude”.

```
df['time'] = df['time'].apply(lambda x: re.sub(r'\b[Dd]ays?\b', '', x)).astype(int)
df.rename(columns={'time': 'time_in_days'}, inplace=True)
```

Code 12. regex altitude, rename

- Accommodation

accommodation is fixed and replaced with correct spelling

```
accomodation_mapping = {"Teahouses": "Teahouse",
                        "Guest Houses": "Guesthouse",
                        "Guesthouses": "Guesthouse",
                        "Teahouses/Lodges": "Lodges",
                        "Luxury Lodges": "Luxury Lodge"}
df['accomodation'] = df['accomodation'].replace(accomodation_mapping)
```

Code 13. accomodation

- Best Travel Time

the travel time were cleaned where all months were turn to it first 3 characters and unnecessary character were replaced and regex was used to match certain format style

```python
def clean_travel_time(value):
    value = value.replace(".", "")
    value = value.replace("March", "Mar")
    value = value.replace("April", "Apr")
    value = value.replace("Setpt", "Sep")
    value = value.replace("Sept", "Sep")
    value = value.replace(" ", "")
    value = re.sub(r'(\w+)-(\w+)', r'\1 - \2', value)
    value = re.sub(r'&(\w+)', r' & \1', value)
    return value


df['best_travel_time'] = df['best_travel_time'].apply(clean_travel_time)
```

Code 14. cleaning best travel time, replace month with season

- Date of Travel

Date of Travel is split to individual entity for future visualizations

```python
In [ ]: df['date_of_travel'] = pd.to_datetime(df['date_of_travel'], errors='coerce')
        df['date_of_travel'] = df['date_of_travel'].dt.strftime('%m-%d-%y')
        df['date_of_travel'] = pd.to_datetime(df['date_of_travel'], format='%m-%d-%y')
```

Code 15. change date of travel format

- Country

```python
country_corrections = {
        "Total": "Puerto Rico",
        "Hong Kong Di": "Hong Kong",
        "Brush": "Burundi",
        "Carbo Verde": "Cabo Verde",
        "Hong Kong Sar": "Hong Kong",
        "Jibuti": "Djibouti",
        "Kiribass": "Kiribati",
        "Komoro": "Comoros",
        "Lesot": "Lesotho",
        "Macau Travel Certificate": "Macau",
        "Mari": "Mali",
        "Moldoba": "Moldova",
        "Naure": "Nauru",
        "Nigail": "Niger",
        "Others": "Virgin Islands",
        "Palao": "Palau",
        "Saechel": "Seychelles",
        "Santa Principa": "Sao Tome and Principe",
        "St. Christopher Navis": "Saint Kitts and Nevis",
        "Sun Marino": "San Marino",
        "Swaji Land": "Eswatini",
        "Torque Menistan": "Turkmenistan",
        "Zaire": "Democratic Republic of the Congo",
        "Israel": "Palestine" # isntreal
    }
df['country'] = df['country'].replace(country_corrections)
```

Code 16. replace and add country name based on UN ISO

Countries and dependencies were replaced with their current name and spelling were matched according to how it is defined by UN and ISO 3166-1, which represent them based on number codes

- Regional Code

same over here extra 0 was present in units place of regional codes columns which were removed and to match 3-digit UN code and some were replaced

```python
df['regional_code'] = df['regional_code'].astype(str).apply(
    lambda x: int(x[:-1]) if x[-1] == '0' and x[:-1] else int(x) if x else 0)
```

Code 17. remove extra zero to match UN code

```python
regional_code_corrections = {
    "Macau": 446,
    "Democratic Republic of the Congo": 178,
    "Eswatini": 748,
    "Hong Kong": 344,
    "Palestine": 275,
    "Virgin Islands": 92,
    "Puerto Rico": 630
}

# Apply regional code corrections based on country
for country, regional_code in regional_code_corrections.items():
    df.loc[df['country'] == country, 'regional_code'] = regional_code
```

Code 18. map code with new country

3.1.4.2. Data Validation

replace is used/applied to these columns giving them appropriate name and representation for easy recognition and value count is used on each to see both unique options with count

- Gender

```python
[ ]: df.rename(columns={'sex': 'gender'}, inplace=True)
```

Code 19. rename gender column

- Employment Type

```
df.rename(columns={'employment_type': 'employment_sector'}, inplace=True)
df['employment_sector'] = df['employment_sector'].replace({'Government Sector': 'Government',
                                                           'Private Sector/Self Employed': 'Private'})
```

```
df['employment_sector'].value_counts()
```

employment_sector	count
Private	277
Government	106

dtype: int64

Code 20. rename employment column and values

- Age

```
df.rename(columns={'age': 'age_in_years'}, inplace=True)
```

Code 21. rename age column

- Year

```
mode_year = df['year'].mode()[0]
df['year'] = df['year'].replace(5460, mode_year)

df.rename(columns={'year': 'insurance_year'}, inplace=True)
```

Code 22. replace with most frequently occurring value

3.1.4.3. Remove Outliers

Box plot was used to identify the outliers in the "cost" column and removed using the Interquartile Range (IQR) method. Q1 - 1.5 * IQR determined the bottom and Q3 + 1.5 * IQR determined as upper bounds; any data points outside these limits were seen as outliers and removed. This improved the dataset so that an analysis might be more exact.

```
# Calculate the first and third quartiles for the 'Cost' column
Q1 = data['cost_in_npr'].quantile(0.25)
Q3 = data['cost_in_npr'].quantile(0.75)

# Calculate the Interquartile Range (IQR)
IQR = Q3 - Q1

# Define lower and upper bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filter the dataset to remove outliers
new_data = data[(data['cost_in_npr'] >=
                 lower_bound) & (data['cost_in_npr']
                                 <= upper_bound)]
```

Code 22. apply interquartile range

$$Q1 = 25th\ Percentile,$$

$$Q2 = 50th\ Percentile,$$

$$Q3 = 75th\ Percentile$$

$$IQR = Q3 - Q1$$

$$Lower\ Outlier = Q1 - 1.5 * IQR,$$

$$Upper\ Outlier = Q3 - 1.5 * IQR$$

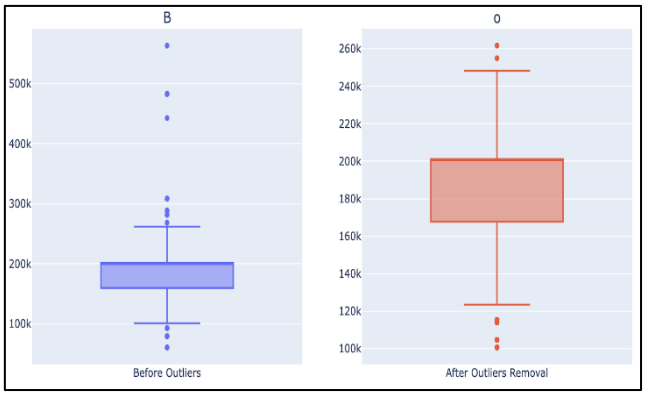


Figure 3.Before and After Removing Outlier form cost in npr

3.2. Feature Engineering

We applied feature engineering techniques to transform the "Date Of Travel" column into a standard Date and Time format. This made sure that the data was consistent across the whole set. Additionally, we changed the "Time Duration" column to show duration in days, simplifying the analysis of time-based data.

- Treks Peaks, Routes, and Activity

New features were extracted from the "Trek" column to create new columns such as "trek_peaks," "trek_routes," and "trek_activity," which offer more detailed information about each trek. We used mode to fill in null values for missing data, ensuring the data remained full and reliable.

```python
peak_list = ['kanchenjunga', 'makalu', 'everest', 'helambu',
             'langtang' , 'manaslu', 'mustang', 'annapurna']

def find_treks(text):
  found_treks = []
  for trek in peak_list:
    if re.search(r'\b' + trek + r'\b', text.lower()):
      found_treks.append(trek)
  return ', '.join(found_treks)

df['trek_peaks'] = df['trek'].apply(find_treks)
```

Code 23. separate trek peak from trek

```python
route_list = ['base camp', 'tilicho lake', 'poon hill', 'tibet', 'high',
              'gokyo ri', 'gokyo lake', 'cho la pass', 'renjo la pass',
              'kalapathar', 'kangshung', 'ghorepani', 'khopra ridge',
              'gosaikunda', 'ganja la pass', 'mardi himal', 'tamang heritage',
              'nar phu valley', 'rara lake', 'tsum valley', 'upper', 'yara', 'nepal', 'royal']

def find_type(text):
  found_type = []
  for trek in route_list:
    if re.search(r'\b' + trek + r'\b', text.lower()):
      found_type.append(trek)
  return ', '.join(found_type)

df['trek_routes'] = df['trek'].apply(find_type)
```

Code 24. separate trek route from trek

```python
activity_list = ['heli', 'trek', 'short', 'circle', 'circuit', 'luxery',
                 'panaroma', 'sanctuary', 'sunrise', 'view', 'classic',
                 'advanced', 'shuttle', 'lodge', 'seniors', 'youths',
                 'trekking', 'face', 'instant', 'community', 'experience',
                 'hiking', 'culture', 'passes', 'tenzing hillary', 'marathon',
                 'trial', 'tiji festival']

def find_activity(text):
  found_type = []
  for trek in activity_list:
    if re.search(r'\b' + trek + r'\b', text.lower()):
      found_type.append(trek)
  return ', '.join(found_type)

df['trek_activity'] = df['trek'].apply(find_activity)
```

Code 25. separate trek activities from trek

```python
routes_peaks = {
    'everest': ['gokyo lake','gokyo ri', 'gokyo lake, renjo la pass'],
    'langtang': ['tamang heritage'],
    'manaslu': ['tsum valley'],
    'annapurna': ['khopra ridge', 'nar phu valley', 'mardi himal', 'poon hill, ghorepani','royal'],
    'jumla':['rara lake'],
    "everest, annapurna": ['nepal']
    }

def fill_empty_trek_peaks(row):
        if pd.isnull(row['trek_peaks']) or row['trek_peaks'] == '':
            for peak, routes in routes_peaks.items():
                if row['trek_routes'] in routes:
                    return peak
        return row['trek_peaks']

df['trek_peaks'] = df.apply(fill_empty_trek_peaks, axis=1)
```

Code 26. fill remaining trek routes

3.2.1.1. Data Imputation

```python
# Replace empty strings with NaN
df['trek_routes'] = df['trek_routes'].replace('', np.nan)

# Find the most frequent value, excluding NaN
most_common_route = df['trek_routes'].mode()[0]

# Replace empty strings with most frequent value
df['trek_routes'].fillna(most_common_route, inplace=True)

# Check the result
print(df['trek_routes'].value_counts())
```

```
trek_routes
base camp                195
upper                     20
mardi himal               20
tilicho lake              15
tsum valley               15
tamang heritage           15
nepal                     15
high                       6
khopra ridge               5
```

Code 27. fill remaining trek routes with most frequent value

```python
# Replace empty strings with NaN
df['trek_activity'] = df['trek_activity'].replace('', np.nan)

# Find the most frequent value, excluding NaN
most_common_activity = df['trek_activity'].mode()[0]

# Replace empty strings with the most frequent value
df['trek_activity'].fillna(most_common_activity, inplace=True)

# Check the result
print(df['trek_activity'].value_counts())
```

```
trek_activity
trek                     200
trek, circuit             40
trek, short               17
trekking                  15
heli, trek                12
trek, view                 6
```

Code 28. fill remaining trek activities with most frequent activities

- Accommodation Types

Additionally, we also splitted the "accommodation" column into two independent columns: "accommodation_1" for hotels and "accommodation_2" for tea, guest, and lodge.

Finally, we used label encoding to transform the category columns like "Trek Peaks" and "Trip Grade" into numerical values, thereby preparing them for quantitative analysis.

```python
df.split_accomodation = df.accomodation.str.split('\n')

df[['accomodation_1','accomodation_2']] = df['accomodation'].str.split('/',n=1,expand=True)

print(df['accomodation_1'].unique())
print(df['accomodation_2'].unique())

print("*"*24)

df['accomodation_1'] = df['accomodation_1'].str.rstrip('s')
df['accomodation_2'] = df['accomodation_2'].str.rstrip('s')

def refine_accomodation(value):
  if isinstance(value, str):
    value = value.replace("Teahouse", "Tea")
    value = value.replace("Guesthouse", "Guest")
    value = value.replace("Luxury Lodge", "Lodge")
    value = value.replace("Guest House", "Guest")
  return value

df['accomodation_1'] = df['accomodation_1'].apply(refine_accomodation)
df['accomodation_2'] = df['accomodation_2'].apply(refine_accomodation)

print(df['accomodation_1'].unique())
print(df['accomodation_2'].unique())
```

Code 29. split accommodation, replace values, apply function

```
['Hotel' 'Teahouses']
['Teahouses' 'Guesthouse' 'Luxury Lodges' 'Teahouse' 'Guest Houses'
 'Guesthouses' 'Lodges']
************************
['Hotel' 'Tea']
['Tea' 'Guest' 'Lodge']
```

Code 30. output before and after accommodation split

- Converting Best Travel Months to Seasons

```python
def season_to_month(season):
  season_month = {
      'Mar - May': 'Spring',
      'Jun - Aug': 'Summer',
      'Sep - Nov': 'Autumn',
      'Dec - Feb': 'Winter',
      'Sep - Dec': 'Autumn-Winter',
      'Jan - May': 'Spring-Winter',
      'Apr - May': 'Spring',
      'Mar - Nov': 'Non-Winter'
  }
  return season_month.get(season)

df['best_travel_time_2'] = df['best_travel_time_2'].apply(season_to_month)
df['best_travel_time_1'] = df['best_travel_time_1'].apply(season_to_month)
```

Code 31. replace months with season mapping

- Date of Travel

```python
df['traveled_day'] = df['date_of_travel'].dt.day
df['traveled_month'] = df['date_of_travel'].dt.month
df['traveled_year'] = df['date_of_travel'].dt.year
```

Code 32. split date into year month day

3.3. Data Preparation

Data preparation was done to make sure that only the most important feature for the model was used and that any data that wasn't needed was removed.

3.3.1. Dimensionality Reduction

Dimensionality reduction was used to lower the number of features while preserving important ones. The correlation method was used to find and remove unnecessary features. This strategy simplified the dataset by removing characteristics with weak correlation.

```python
data = df[['trek_peaks', 'cost_in_npr', 'time_in_days', 'trip_grade', 'max_altitude']]
```

Code 33. Selecting only required columns

```
In [172]:
row = feature.iloc[50]
row

Out[172]:
```

	93
Encoded_Trek	0.0
time_in_days	18.0
Encoded_Trip_Grade	0.0
max_altitude	5416.0

Code 34. feature check

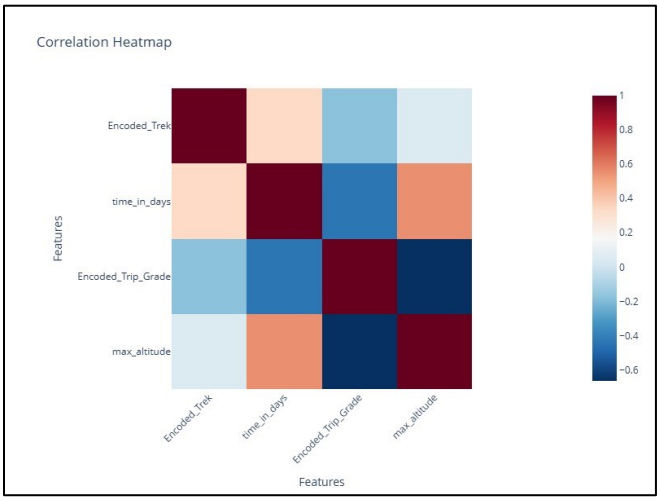


Figure 4. Coorelation matrix

3.3.2. Target and Features

The feature selection method separated the target and features. The features were "Encoded_Trek", "time_in_days", "Encoded_Trip_Grade", and "max_altitude." The target was "cost_in_npr" This separation guaranteed that the model could be trained efficiently by splitting it into 80% training data and 20% testing data using a train-test split, which could be predicted from those used for prediction.

```
feature = new_data.drop(columns=['cost_in_npr'])
target = new_data['cost_in_npr']
```

Code 35. feature and target

```
feature.corr()
```

	Encoded_Trek	time_in_days	Encoded_Trip_Grade	max_altitude
Encoded_Trek	1.000000	0.333292	-0.176267	0.050786
time_in_days	0.333292	1.000000	-0.437575	0.553098
Encoded_Trip_Grade	-0.176267	-0.437575	1.000000	-0.663754
max_altitude	0.050786	0.553098	-0.663754	1.000000

Code 36. feature coordination

```
feature.shape
```

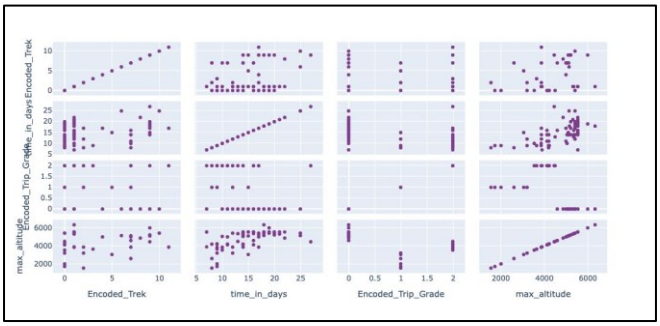
```
(335, 4)
```

```
X_train, X_test, Y_train, Y_test = train_test_split(feature, target, test_size=0.2, random_state=42)
```

```
X_train.shape
```

```
(268, 4)
```

Code 37. shaping and train text split



Code 38. scatter plot matrix of feature

3.3.3. Standard Scaler

The data was normalized by using a standard scaler on features. This made sure that each feature had a mean of 0 and a standard deviation of 1. By making all the features the same size, this scaling process helped the model perform better, which is especially helpful for methods that depend on feature sizes.

```
scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
X_train_scaled
```

```
array([[-0.81635607,  0.48103063, -0.72325553,  0.71653075],
       [-0.23985165,  0.48103063,  1.59809157, -0.80653362],
       [-0.52810386,  0.7044556 , -0.72325553,  0.84361697],
       ...,
       [-0.81635607,  0.92788057, -0.72325553,  0.71653075],
       [ 1.20140939, -0.8595192 , -0.72325553,  0.3067023 ],
       [-0.81635607,  0.48103063, -0.72325553,  0.71653075]])
```

Code 39. standard scalar

3.4.Selection Algorithms

3.4.1. Linear regression:

Simple, interpretable algorithm that searches in the feature space a linear relationship towards the target variable and predicts the outcome by minimizing an error between the predicted and actual values.

$$\Upsilon_i = \beta_0 + \beta_1 \chi + \epsilon_i$$

Figure 5. linear regression actual vs predicted value.

3.4.2. Random Forest Regressor:

Ensemble methods that combine multiple decision trees in order to generally have better generalization of results, with reduced overfitting risks. Thus, this model is suitable for capturing the complex relationships in data.

$$Final\ Prediction = \frac{1}{n}\sum_{i=1}^{n} Prediction\ from\ Tree_i$$

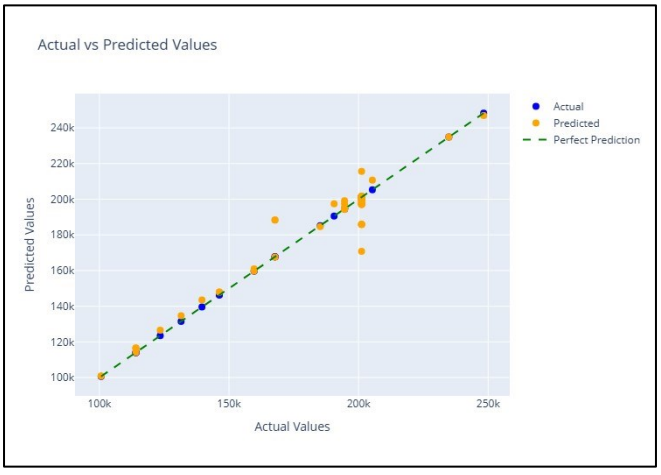


Figure 6. random forest regressor actual vs predicted value

3.4.3. XGBoost:

A powerful gradient boosting algorithm widely used because of its high performance and efficiency. It is efficient in handling big datasets by iteratively building models to correct errors from previous iterations.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

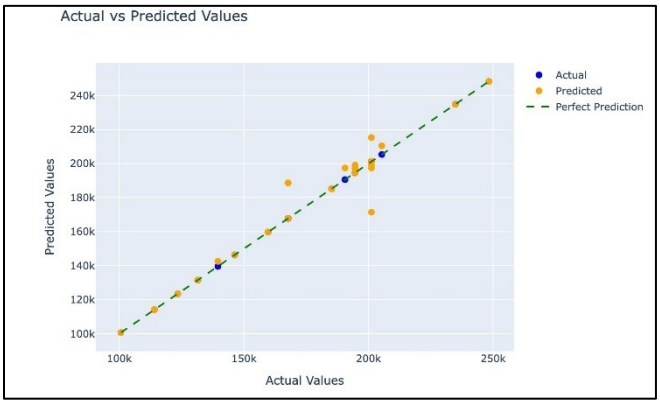


Figure 7. xgboost actual vs predcited value

3.5. Evaluation Matrics

3.5.1. Linear Regression:

Table 2. linear regression metrics

Mean Absolute Error (MAE):	22,098.48
Mean Squared Error (MSE):	753,175,881.57
R-squared (R^2):	14.44

This model showed a high error and low R^2, indicating a poor fit for the data.

3.5.2. Random Forest Regressor

(Before Hyperparameter Tuning and Cross-Validation)

Table 3. random forest before hyperparameter tuning and cross validation

Mean Absolute Error (MAE):	3,612.80
Mean Squared Error (MSE):	47,847,137.76
R-squared (R^2):	94.18

Initial performance was strong, with a low error and high R^2. (After Hyperparameter Tuning and Cross-Validation)

Table 4. . random forest after hyperparameter tuning and c

Mean Absolute Error (MAE):	3,612.80
Mean Squared Error (MSE):	48,758,676.63
R-squared (R^2):	94.05

Results remained similar, showing robust model performance with minimal improvement from tuning.

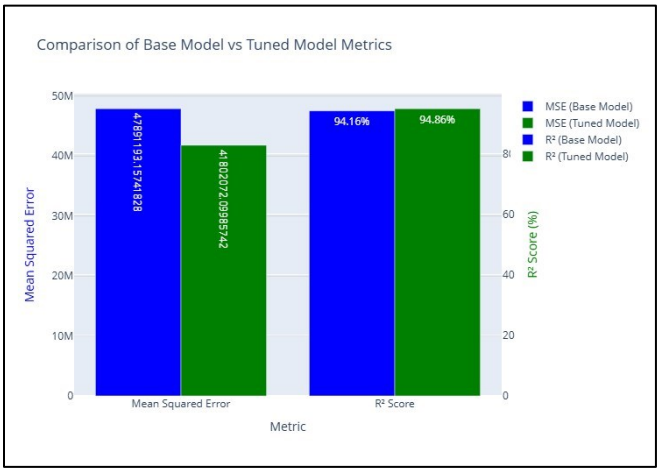


Figure 8. compararision base and tuned model

XG Boost

Table 5. . xgboost metrics

Mean Absolute Error (MAE):	1,761.11
Mean Squared Error (MSE):	31,330,042.00
R-squared (R^2):	96.43

XGBoost outperformed other models, achieving the lowest error and highest R^2, indicating an excellent fit for the data.

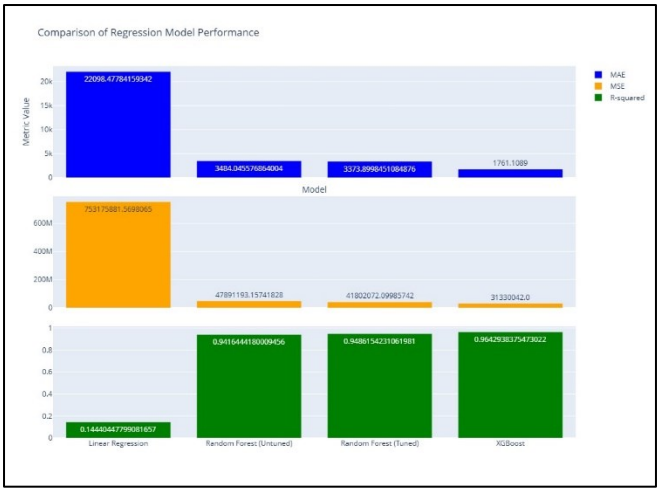


Figure 9. comparision of all 4 models

4. Processing and Technologies

4.1. Processing Techniques

4.2. Technology Stack Justification

We used **PYTHON** language for this project, as it is projected to being one of the most popular languages in Artificial Intelligence, Data Science communities in Data Analytics and Machine Learning tasks. It is easily accessible, and extensible comprehensive open-source libraries via modules. (Welcome to Python.org, n.d.)

Interactive Python Notebooks: are those platforms which maybe on-premise like 'Jupyter Project' or based on cloud like 'Google Colaboratory'. were the important packages being present in both distributions. (Anaconda Navigator — Anaconda Documentation, 2024). This versatile interactive interface allow us users to write and excute interactive python codes, which we can visualize the insights based on exploratory analysis mathematical computations, statistical simulation and explanatory text all looks like sharable and readable documents

Pandas is one of the most in use library supported by communities all over the world which also the reason it is popular. here we can read DF object form multiple formats like csv, xlsx. manipulate by writing over them, transform and reshape them. making it super functional and practical in analysis workflows. (Pandas, 2024)

NumPy provides analysts of any platforms with very fast, versatile, comprehensive functions for mathematics and scientific computing. it is useful for manipulating and transforming on vast multi-dimensional matrices at once with vectorized approach. (NumPy, 2022)

RegEx is important package/module built for matching operations, froming a search pattern for specific set of string sequence, meta-characters which can be used for substitution, replacement or removal, of ordinary and special characters

Plotly is one of the visualization libraries, we were able used to create super interactive, top quality and unique graphical figures like basics (scatter, line, bar, pie, dot, sunburst, Sankey, tree-maps), statistics (histograms, box plots, dist.-plots, scatter matrix, parallel categories, violins, trendlines), scientific (heatmaps, wind rose and polar bear, radar, network), financial (time series), geographic maps (bubble,), ai/ml (ROC curves, KNN classifications, regression, PCA 3D axes). (Plotly, 2023)

Seaborn is also a visualization library based on **Matplotlib** which provide attractive visuals for statistical data, and also highly informative interface, with plots like relational, distributions, categorical, regression, multi gird. these library are very comprehensive making anytype of static,

animated or ineeractive visuals it integrates well with pandas data structure applying transformations and abstractions, with functions like axes or figure level

Scikit-learn is a machine learning python module, consisting which is indispensable as a multi-functional ML toolkit, widely in use in state-of-the-art models/tasks due it its straight forward API, simple yet powerful with detailed quality documentations. It is fantastic to apply in our academics and suits us students for the straight forward mathematical operations. Excellent for iterative process of (exploring, selecting, training, evaluating) (Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.6.Dev0 Documentation, 2024)

5. Reporting, Visualization, and Insights

5.1. Reporting Frameworks for Key Performance Indicators

Key Performance Indicators are the measurements that will allow the company or different departments in a sector of an industry to, learn from the data they have collected through different sources. these are viewed by the by stakeholder and other companies, sub directories for improving their longievity and performance. these are important indicators that will help any industry to make sound decisions on different areas and apply stratigic approach, while juding overall performance. these historical data can be use to judge progress and set bench marks based on past performances. types like statigics, operational, functional, leading are the board categories

the major KPI in tourisn and travel sector and trekking industries whould be

Table 6. major KPIs in travel, tourism and trekking

Arrival by: (from national statistics office)	Land or Air medium public or private rentals planes, jeeps, buses, taxi
Average Length of Stay (formr Department of Immigration)	duration spent on destination country for valuable economic increase, boost local economy
Vistor Demographics (from nepal rastra bank)	Gender and Age Group Behaviour preference, policy maker, create job
Purpose of Visit (from local heritage development trusts, national parks and wildlife conservation reserves)	holiday, pleasure, pilgrimage, trekking, mountaineering, expedition, business, official, conventions and conference
airlines and land transports (fom civil aviation authority)	operational and non operational airport, domestic and international,

incidents (reports from tourist police)	Lost/Missing Items, documents and belongings, Stolen/theft, Fraud/Cheating, Robbery, Pickpocket, Harassment/dispute, Accident, Damage, Attack/ Assault, Snatching, Threat, Missing Person, Rape, Misbehaviour, Kidnap, Against Foreigner
education and business (to national academy of tourism and hotel management)	institutions, academic courses, programmes, professional training,
tourism enterprise (from nepal tourism board, department of tourism and office)	hotels with stars (up to tourist standards) or no stars like lodges, stay in houses, agencies and guides for travel overall, activities and trekking specific while providing services, operations

- trek popularity

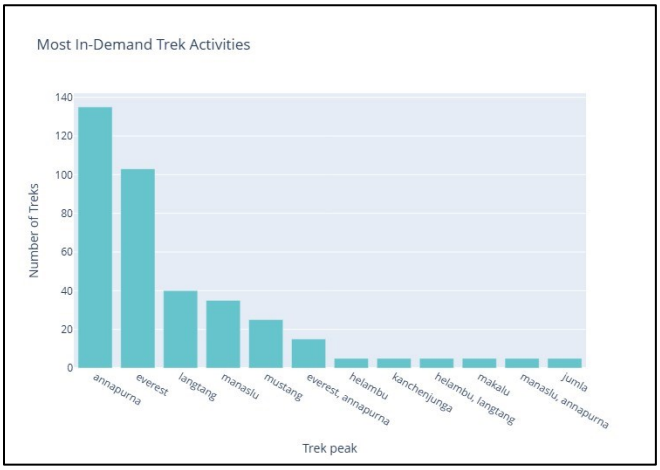


Figure 10. trek peak and trek count

here we can see distribution of trek and see which is more popular among the data collected. this illustrate the number of treks taken per peaks. so the ministry can focus on marketing effort and other organizations can allocate resources, and improve overall services.

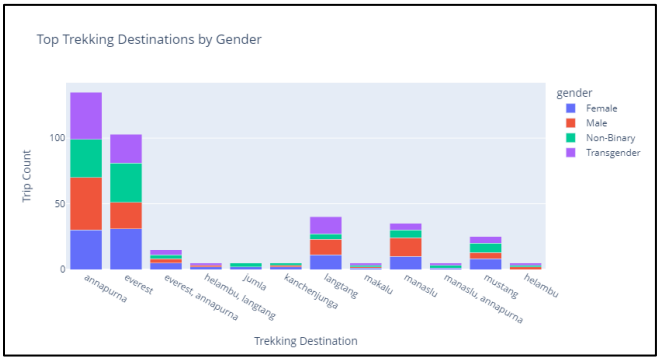


Figure 11. trip destination by gender

we can further distribute popular trekking destination based on gender Age-wise preference for popular trekking destinations within a particular age group (Bar chart) To understand in which age group what is the distribution over destinations, so that they can segmentise their marketing into Age divided trekking preferences.

- travel demographic

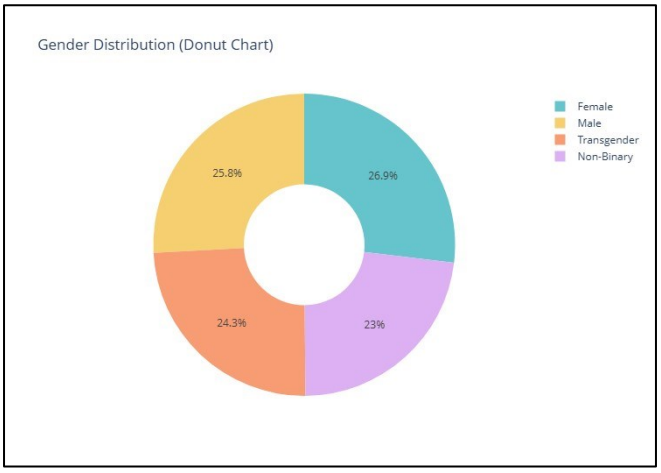


Figure 12. gender distribution

this chart show the diversity that is present among the trekker. with this we can help stakeholders to target differnent genders. while promoting taylor made experience while being inclusive to minorities. participation in trekking can also be calculated to visualize a clear picture of the community.

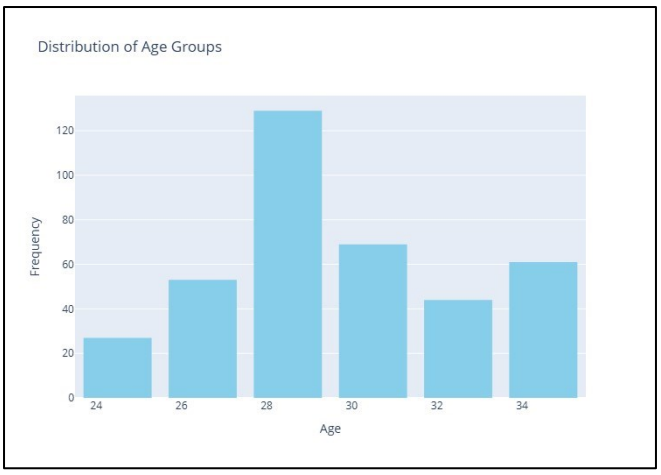


Figure 13. age group distribution

this chart shows the range of different age groups, predominant praticipants. we can target different group with different products and increase local marketing and maybe increase global export overall. increasing the demographic trend insights.

- health risk

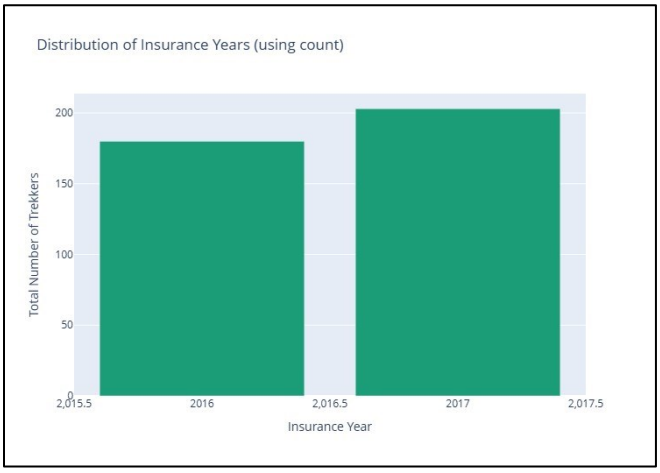


Figure 14. insurance count distribution

we can see the increase in insurance per year

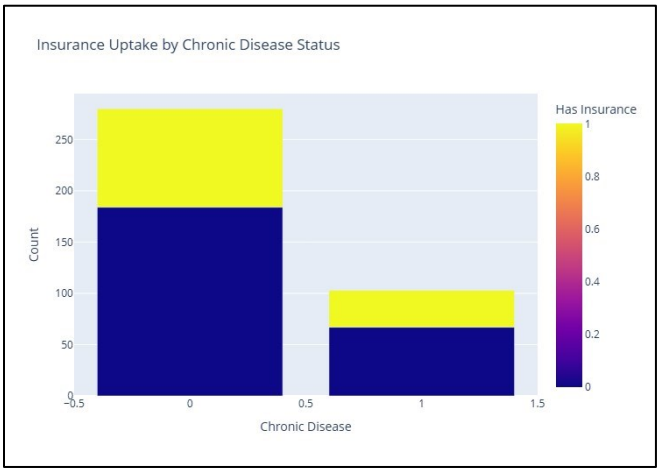


Figure 15. insurance uptake by chronic disease status

how chronic health conditions affect the likelihood of purchasing travel insurance. A stacked bar format breaks down the count of individuals with and without insurance by chronic disease status. The KPI here is the insurance uptake rate by health condition—higher rates among those with chronic conditions indicate a successful targeting strategy for insurance providers.

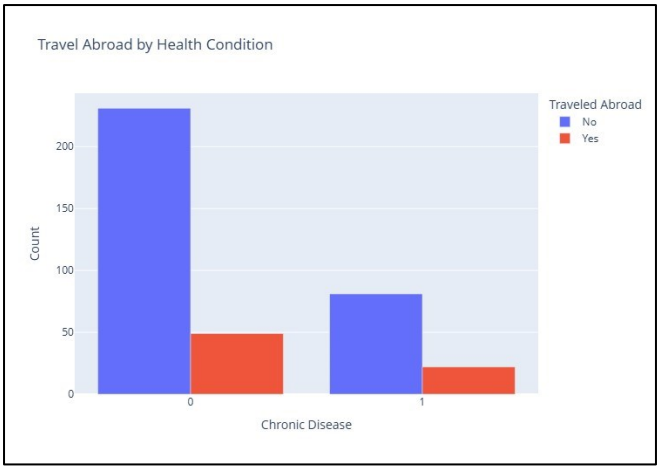


Figure 16. travel abroad by health condition

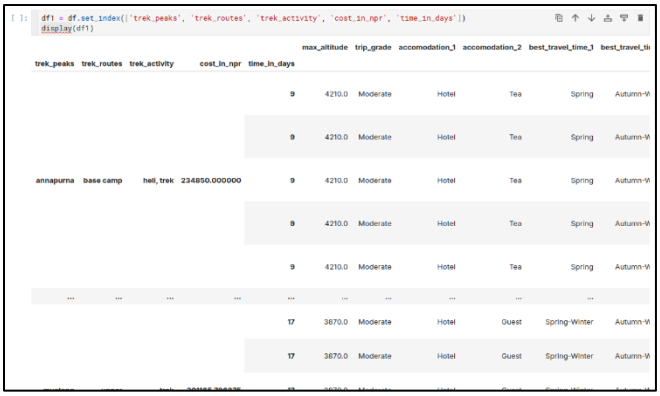
- Basis of Report for Stakeholder

Figure 17. distribution of travel month

we can check for most busy months for trekking and per season to properly plan different activities, planning human resources and marketing promotions further aiding in seasonal plans.

5.2. Visualization Tools for Complex Tourism Data

- how it presrnt complex data in understanding manner



```
df1 = df.set_index(['trek_peaks', 'trek_routes', 'trek_activity', 'cost_in_npr', 'time_in_days'])
display(df1)
```

trek_peaks	trek_routes	trek_activity	cost_in_npr	time_in_days	max_altitude	trip_grade	accomodation_1	accomodation_2	best_travel_time_1	best_travel_ti
				9	4210.0	Moderate	Hotel	Tea	Spring	Autumn-W
				9	4210.0	Moderate	Hotel	Tea	Spring	Autumn-W
annapurna	base camp	heli, trek	234850.000000	9	4210.0	Moderate	Hotel	Tea	Spring	Autumn-W
				9	4210.0	Moderate	Hotel	Tea	Spring	Autumn-W
				9	4210.0	Moderate	Hotel	Tea	Spring	Autumn-W
...	...	...	...	...	...	...	...	...	...	...
				17	3870.0	Moderate	Hotel	Guest	Spring-Winter	Autumn-W
				17	3870.0	Moderate	Hotel	Guest	Spring-Winter	Autumn-W

Figure 18. set_index method

```
df.query('country == "Palestine"')
```

	trek	cost_in_usd	time_in_days	trip_grade	max_altitude	accomodation	best_travel_time	date_of_travel
65	annapurna circuit trek with tilicho lake and p...	1499.0	20	Difficult	5416.0	Hotel/Guesthouse	Jan - May & Sep - Dec	2021-10-29
268	gokyo lakes and gokyo ri trek	1499.0	14	Difficult	5357.0	Hotel/Guesthouse	Jan - May & Sep - Dec	2020-03-24
289	langtang valley trek with ganja la pass	1499.0	14	Difficult	5122.0	Hotel/Guesthouse	Jan - May & Sep - Dec	2021-05-25
86	manaslu circuit trek	1190.0	17	Difficult	5220.0	Hotel/Guesthouse	Jan - May & Sep - Dec	2023-01-11

Figure 19. query fucntion

- Interactive Format for Decision Maker

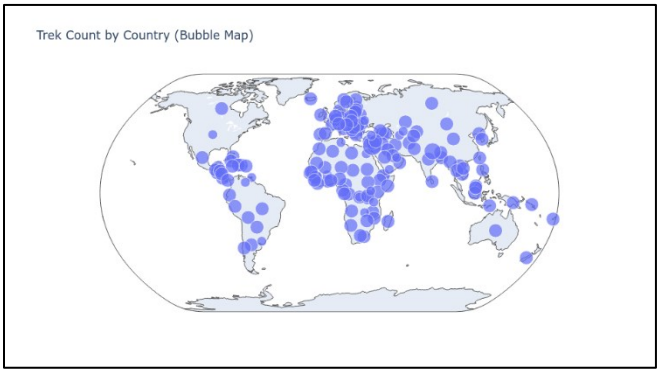


Figure 20. bubble map for trek count by country

this map provide a geographic overview of traveller all over the world, those who have visitied. this will help stakeholder to identify foreign market and build stronger bonds and have potential partnership.

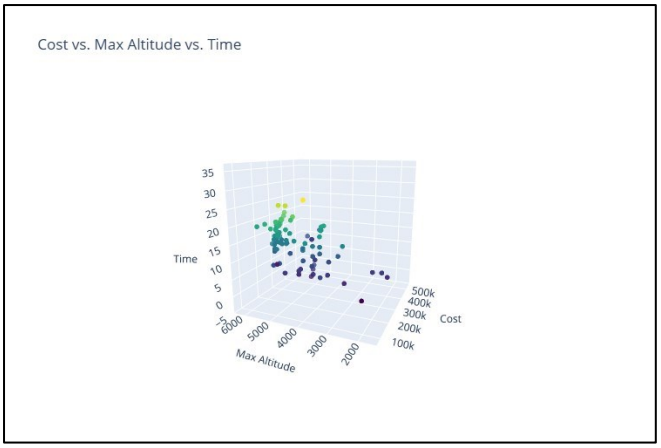


Figure 21. 3d axes cost altitude time

stakeholders are able to view multidimentional interactive view and understand the how points vary with different comparing columns. like trek option, cost, duration, altitude, to help in further adjusting the pricing make it profittable for business affordable for users.

5.3. Insights for Data Analysis

- Trends, durations

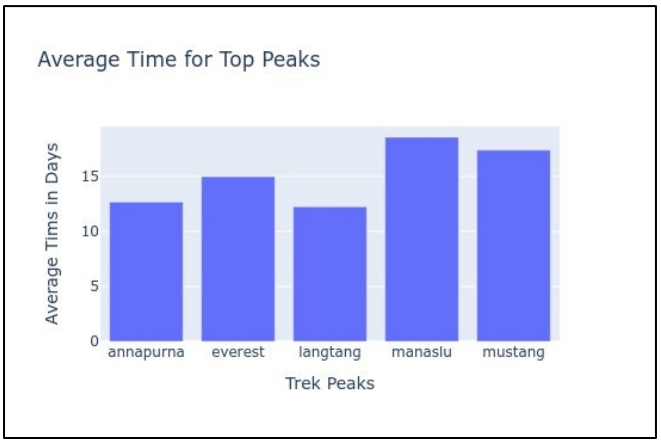


Figure 22. average time for top peaks

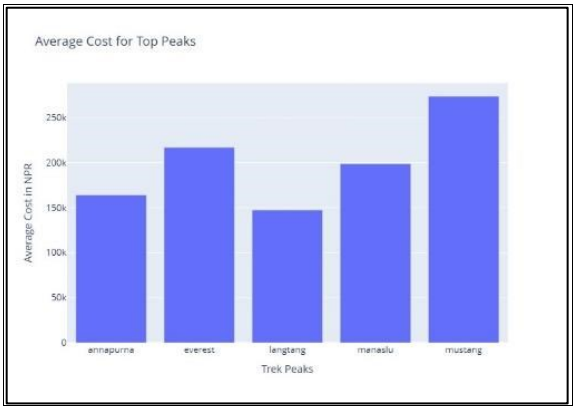


Figure 23. distribution of trek peak and average cost

- round figure

```
display(df.groupby(['trek_peaks', 'trek_routes', 'trek_activity'])['cost_in_usd'].aggregate(['min', 'max', 'mean','median']))
```

trek_peaks	trek_routes	trek_activity	min	max	mean	median
		heli, trek	1750.0	1750.0	1750.000000	1750.0
		trek	860.0	1499.0	1197.800049	1090.0
		trek, circle	1499.0	1499.0	1499.000000	1499.0
		trek, circuit	1100.0	1499.0	1283.000000	1250.0
	base camp	trek, circuit, classic	1350.0	1350.0	1350.000000	1350.0
		trek, instant	980.0	980.0	980.000000	980.0
		trek, sanctuary	1499.0	1499.0	1499.000000	1499.0
		trek, short	590.0	1090.0	840.000000	840.0
annapurna		trek, short, circuit	1250.0	1250.0	1250.000000	1250.0
	khopra ridge	trek, community	1499.0	1499.0	1499.000000	1499.0
	mardi himal	trek	590.0	1499.0	979.666687	850.0
		trekking	850.0	850.0	850.000000	850.0
	nar phu valley	trek	1499.0	1499.0	1499.000000	1499.0
	poon hill, ghorepani	trek	450.0	450.0	450.000000	450.0
	royal	trek	1499.0	1499.0	1499.000000	1499.0
	tilicho lake	trek	1450.0	1499.0	1474.500000	1474.5
	tilicho lake, poon hill	trek, circuit	1499.0	1499.0	1499.000000	1499.0

Figure 24. group by aggregate fucntions for mean median min max

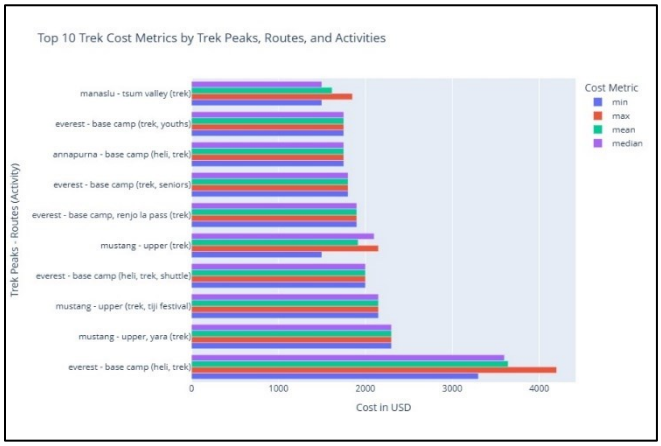


Figure 25. trek cost metric by peaks routes and activities

- Preferences

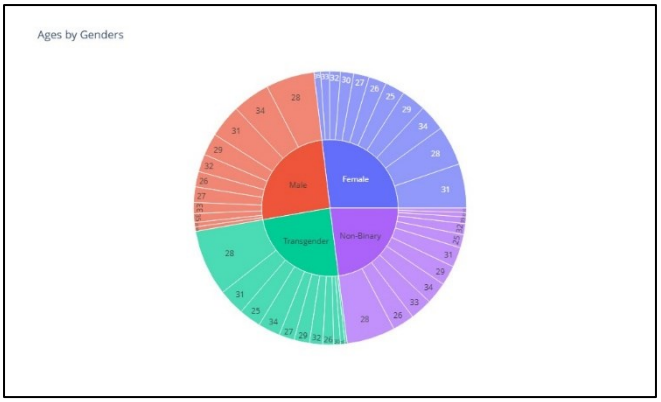


Figure 26. age and gender sunburst chart

with sunburst chart, users can view into multiple layers of informations, demographic intersections, patterns to guide, plan trek and activities. pairing common groups

- Season Propularity

the diagram represent the which solactice and equinox or quarter attract the most visitors are attracted to. instead of just promoting on season, shareholder can adjust price of packages and add extra services to optimize and distribute, allocate resources year long.

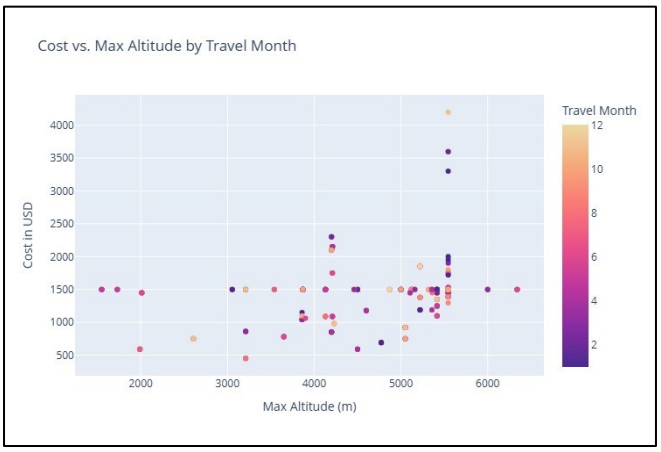


Figure 27. distribution of cost, altitude, by month

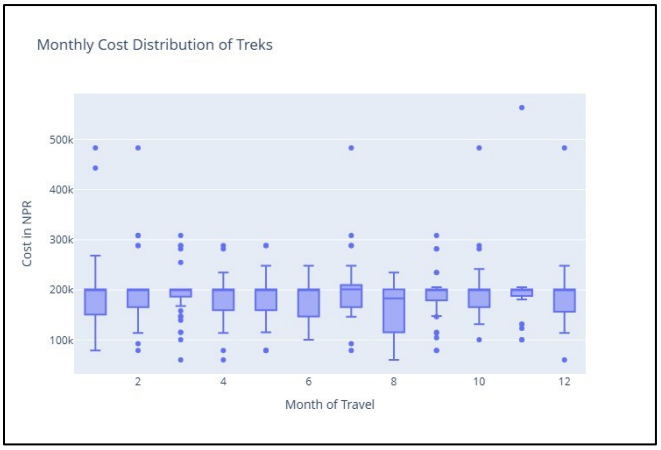


Figure 28. monthly cost distribution

5.4. Impact of Visualizations on Decision-Making

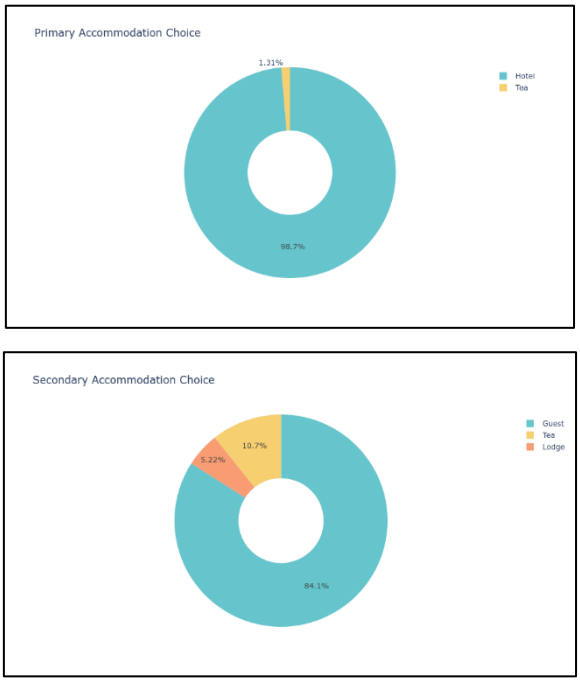


Figure 29. accomodation choices

Primary and Secondary Accommodation Choices display the preference distribution, the popularity of preferred housing the trekking sample would choice and spend money on if they had option to. pririotizing comfort when not in harsh environment and understanding the limits of rural area. the trekker would also show interest in tea houses, camping. in future it helps in resource planning like making a hotel that accommodate and bring money

explain how it help in stakeholder to quickly understand and make informed decisions

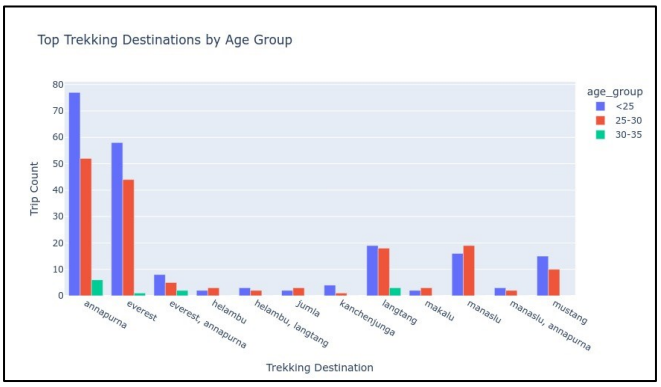


Figure 30. popular destination by age group

Figure 31. average cost for different altitudes across a year

attempted understand with average trip count by income group and a scatter plot of income vs. trip count, covering this subplot, since it covers which part has trekked, KPI by Income GroupAverage trip count average per income groupServes as a guiding metric in developing different packages targeting various income levels.

Figure 32. income analysis

6. MEMORANDUM OF UNDERSTANDING (MOU)

between

RAHUL WASTI, AMOGH SHAKYA, SUJAL RATNA TULADHAR,

MAMATA SHRESTHA, MONALISHA THAPA MAGAR

and

MR. ANMOL ADHIKARI

This is an agreement between "Party A", hereinafter called **TUTORIAL 3 GROUP 6** and "Party B", hereinafter called **Consultant Anmol Adhikari**

I. PURPOSE & SCOPE

The purpose of this MOU is to clearly identify the roles and responsibilities of each party as they relate to collaborative vase study project within the Artificial Intelligence specialization, with a specific focus on Data Analytics and Machine Learning for the Travel, Trekking, Tourism industry objective of this partnership.

In particular, this MOU is intended to:

- Enhance the quality and delivery of educational resources and support, enhance the knowledge and application of machine learning and data analysis technique for Party A
- Increase collaboration and knowledge sharing between the parties, especially for Party A's understanding pf AI-drive solutions tailored to tourism and trekking.
- Reduce operational costs and improve efficiency for project completion for stakeholders, and apply real-world insights into how data analytics, science and machine learning can be applied within the tourism sector to improve decision making in various level of industry
- Establish clear guidelines and expectations for structured project-framework related tasks from Party B with guidance and feedback

II. BACKGROUND

Laboratory 3 Group 6 (Party A) consists of student specializing in Artificial Intelligence, with aa particular focus on Data Analytics and Machine Learning. as part of our coursework, Party A is conducting a case study that applies advanced data analytics techniques and machine learning techniques to tourism and trekking in Nepal, aiming to derive insights and optimize processes within the sector.

Lecturer Anmol Adhikari (Party B) is an experienced instructor in data analytics machine learning and artificial intelligence, providing mentorship and expertise to Party A and helping them to understand theoretical aspect while apply practical knowledge training for scenarios that are relevant to industries in Nepal.

III. [PARTY A] RESPONSIBILITIES UNDER THIS MOU

- Development of data models, use of machine learning algorithms, and use of relevant analytical frameworks that would seem as best fit based on the case study scenario and task requirement adhering to the rubrics
- Deliver periodic updates of the project to Party B when asked to, showing transparency and open communications and success of project, deliver draft and comply to feedbacks
- Sharing of all relevant datasets, process of finding better analysis, maintain document methodologies, organized and accurate result for review, refollowing the rules and upholding ethical standard outlined by Party B

IV. [PARTY B] RESPONSIBILITIES UNDER THIS MOU

- provide expert guidance that are applicable for the case study, evaluate and promote ethical pactices
- review the work provided by Party A and provide feedback to further enhance the understanding of applications and principles
- providing any teaching training helping resources supplies to Party A to navigate safely through complex task when required

VII. EFFECTIVE DATE AND SIGNATURE

This MOU shall be effective upon the signature of Parties A and B authorized officials. It shall be in force from (date)_____ to (date) _____.

Parties A and B indicate agreement with this MOU by their signatures.

VIII. SIGNATURES AND DATES

PARTY A] [PARTY B]

_____ _____

DATE: 30TH OCTOBER 2024 DATE: _____ _____ 2024

7. PROJECT CHARTER

Table 7. charter form

General Project Information	
Project Name:	TREK TOUR TRAV | AI POWERED INSIGHTS FOR NEPAL
Executive Sponsors:	PHILANTHROPIST, BOARD AND DIRECTORS, MR. ANMOL ADHIKARI
Department Sponsor:	SCHOOL OF COMPUTING SCIENCE ARITFICIAL INTELLEGENCE SPECIALIZATION
Impact of project:	IMPACT ON TRAVEL AND TOURISM SECTOR OF NEPAL

Project Team			
	Name	**Department**	**E-mail**
Project Leader:	RAHUL WASTI	Bachelor of Computer Science (Honors)	rahul22013279@iimscollege.edu.np
Team Members:	AMOGH SHAKYA	Bachelor of Computer Science (Honors)	amogh22013021@iimscollege.edu.np
	MAMATA SHRESTHA	Bachelor of Computer Science (Honors)	mamata22013229@iimscollege.edu.np
	MONALISHA THAPA MAGAR	Bachelor of Computer Science (Honors)	monalisha22013291@iimscollege.edu.np
	SUJAL RATNA TULADHAR	Bachelor of Computer Science (Honors)	sujal22013230@iimscollege.edu.np

Stakeholders (e.g., those with a significant interest in or who will be significantly affected by this project)
Volunteer and Community: who is part of the initiative, activists for globalization and climate change
Organizations: trekking and tourism agencies, governmental and private enterprise
Individuals Families: who experienced as in tour guides and operators as it is primary or secondary source
Educational Institutions: colleges, schools, universities for student led project in hotel management
Agencies: local, regional, national, international, social services,

Scope Statement
Project Purpose / Business Justification Describe the business need this project addresses
The project addresses the need to optimize the trekking experience in Nepal by leveraging AI and advanced data analytics. It aims to support decision-making and route planning, enhance customer satisfaction, and enable trekking business to deliver a superior
Objectives (in business terms) Describe the measurable outcomes of the project, e.g., reduce cost by xxxx or increase quality to yyyy
• develop predictive model to analyze tourist preferences and trends, enable real-time recommendation for trekking agency for operational efficiency • implement route optimization algorithms to improve trekking safety and reduce route planning time
Deliverables List the high-level "products" to be created (e.g., improved xxxx process, employee manual on yyyy)
• a comprehensive AI model for trend analysis and route optimization • codes for real-time insights and behavior analysis • report documenting findings, predictive insights, and recommendations • interactive visualizations for businesses and stakeholders
Scope:

List what the project will and will not address (e.g., this project addresses units that report into the Office of Executive Vice President. Units that report into the provosts Office are not included)

- data collection, validation, cleaning, feature engineering and analysis specific to Nepal trekking industry
- application of machine learning and data analytics for customer insights
- development of advanced data aggregation visualization for key performance indicators

Project Milestones Propose start and end dates for Project Phases (e.g., Inception, Planning, Construction, Delivery) and other major milestones

Inception:		**Planning:**	
Start Date:	September 5, 2024	**Start Date:**	September 13, 2024
End Date:	September 13, 2024	**End Date:**	September 26, 2024
Milestones:	Meetings, Ideate, Research, Define, Document.	**Milestones:**	Meetings, Planning, Feedback, Technical Research
Construction:		**Delivery:**	
Start Date:	September 26, 2024	**Start Date:**	October 19, 2024
End Date:	October 19, 2024	**End Date:**	October 29, 2024
Milestones:	Meetings, Design, Development, Integration of Features and Functionality, Testing, Report	**Milestones:**	Testing, Feeback, Evaluations, Report

Major Known Risks (including significant Assumptions) Identify obstacles that may cause the project to fail.

Risk	Risk Rating	Risk	Raing
Regular Monitoring	Medium	Lacking Expertise	Medium
Technology Combability	Low	Collaboration	Medium

Constraints List any conditions that may limit the project team's options with respect to resources, personnel, or schedule (e.g., predetermined budget or project end date, limit on number of staff that may be assigned to the project).

Fixed and short deadline, personal problems (sickness), learning curve, stakeholder expectations

External Dependencies Will project success depend on coordination of efforts between the project team and one or more other individuals or groups? Has everyone involved agreed to this interaction?

analyzing websites related to tourism, trekking, travelling, agencies, governmental, private, institution feedback and guidance from consultant to ensure relevence

1. Communication Strategy (specify how the project manager will communicate to the Executive Sponsor, Project Team members and Stakeholders, e.g., frequency of status reports, frequency of Project Team meetings, etc.

V. Team: Weekly briefing on progress updates, Daily meetings for assignment update on distributed task and feedback recommendation inputs and solutions, engagement and participations,

2. Sign-off

Role	Name	Signature	Date (MM/DD/YYYY)
Project Stakeholder	**Mr. Anmol Adhikari**		__/__/2024
Project Leader	**Mr. Rahul Wasti**		10/30/2024

3. Notes

- Use of google meet for project discussion between team.
- communication with messaging app like WhatsApp.
- Documentation and report through platform like google docs,
- Immediate communication through direct messages and emails.

8. Appendix

8.1. References

1.1. Linear Models — scikit-learn 0.24.0 documentation. (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/linear_model.html

3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.23.2 documentation. (n.d.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor

Adventure Himalaya Circuit Treks. (2024). Adventure Himalaya Circuit Treks. https://www.adventurehimalayacircuit.com/

Anaconda Navigator — Anaconda documentation. (2024). Anaconda.com. https://docs.anaconda.com/navigator/

Himalayan Glacier Adventure and Travel Company. (n.d.). Himalayan Glacier. https://www.himalayanglacier.com/

NumPy. (2022). *Overview — NumPy v1.19 Manual*. Numpy.org. https://numpy.org/doc/stable/

Pandas. (2024). *pandas documentation — pandas 1.0.1 documentation*. Pandas.pydata.org. https://pandas.pydata.org/docs/

Plotly. (2023). *Plotly Python Graphing Library*. Plotly.com. https://plotly.com/python/

scikit-learn: machine learning in Python — scikit-learn 1.6.dev0 documentation. (2024). Scikit-Learn.org. https://scikit-learn.org/dev/index.html

sklearn.feature_extraction. (2024). Scikit-Learn. https://scikit-learn.org/stable/api/sklearn.feature_extraction.html

sklearn.feature_selection. (2024). Scikit-Learn. https://scikit-learn.org/stable/api/sklearn.feature_selection.html

sklearn.preprocessing. (2024). Scikit-Learn. https://scikit-learn.org/stable/api/sklearn.preprocessing.html

Trekking in Nepal: Nepal Trekking Agency. (n.d.). Nepal Hiking Team. https://www.nepalhikingteam.com/

IBM. (n.d.). *What Is Data Modeling?* IBM. Retrieved October 25, 2024, from https://www.ibm.com/topics/data-modeling

National Trust for Natural Conservation. (n.d.). *Annapurna Conservation Area Project (ACAP) | The National Trust for Nature Conservation (NTNC)*. National Trust for Nature Conservation. Retrieved October 26, 2024, from https://ntnc.org.np/project/annapurna-conservation-area-project-acap

Pandey, R., & Dhakal, R. K. (2019). Visit Nepal Year 2020: Some Imperatives. *Social Inquiry: Journal of Social Science Research*, *1*(1). https://doi.org/10.3126/sijssr.v1i1.26919

Ratner, B. (2017). *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, Third Edition* (Third ed.). CRC Press LLC. https://doi.org/10.1201/9781315156316

Shrestha, D., Wenan, T., Shrestha, D., Rajkarnikar, N., & Jeong, S.-R. (2024, March 18). Personalized Tourist Recommender System: A Data-Driven and Machine-Learning Approach. *Computation*. https://doi.org/10.3390/computation12030059

8.2.Table of Abbveration and Acronymns

AI	Artificial Intelligence	**NB**	Naïve Bayes
ML	Machine Learning	**NTB**	Nepal Tourism Board
MSE	Mean Square Error	**ACAP**	Annapurna Conservation Area Project
NP	NumPy	**SNP**	Sagarmatha National Park
PD	Pandas	**GPS**	Global Positioning System
PY	Python	**GPRS**	General Purpose Radio Service
R^2	Coefficient of Determination	**CSV**	Comma Separated Value
SK	Scikit-learn	**UN**	United Nations
XG	Extreme Gradient	**ISO**	International Organization Standarization
TSR	Tourist Recommender System	**IQR**	Inter Quartile Range
KNN	k-Nearest Neighbour	**RE**	Regular Expression
RSVM	Support Vector Machine	**NSO**	National Statistic Office
RF	Random Forest	**DOI**	Department of Immigration
GB	Gradient Boosting	**HTF**	Heritage Trust Fund
NP	National Park	**CA**	Conservation Area
WR	Wildlife Reserve	**HR**	Hunting Reserve
KPI	Key Performance Index	**DOT**	Department of Tourism
NTB	Nepal Tourism Board	**TO**	Tourism Office
CAA	Civil Aviation Authority	**NHT**	Nepal Hiking Team
TAAN	Trekking Agencies Association Nepal	**TIMS**	Trekkers Information Management Systems
MoCTCA	Ministry of Culture, Tourism and Civil Aviation	**NRB**	Nepal Rastra Abnk
CBS	Central Bureau of Statistics		