HOLIDAY PACKAGE PREDICTION

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

HRIDHYA ANOOP

MERLIN SEBASTIAN

MADHAV M

ICT ACADEMY OF KERALA

THIRUVANANTHAPURAM, KERALA, INDIA

Nov 2022

List of Figures

List of Abbreviations

EDA	Exploratory Data Analysis
KNN	K-Nearest Neighbor
RF	Random Forest
LR	Logistic Regression
XGB	XG Boost
SMOTE	Synthetic Minority Oversampling Technique
XGBoost	Extreme Gradient Boosting

Table of Contents

Abstract

"Trips & Travel.Com" seeks to establish a viable business model to expand its customer base by introducing a new offering: the Wellness Tourism Package. Wellness Tourism focuses on enabling travelers to maintain, enhance, or begin a healthy lifestyle while supporting overall well-being. Currently, the company offers five types of packages: Basic, Standard, Deluxe, Super Deluxe, and King. An analysis of last year's data revealed that only 18% of customers purchased packages, with high marketing costs due to random customer targeting. To address this inefficiency, the company plans to leverage customer data to make marketing expenditures more effective.

Our exploratory data analysis (EDA) identified key customer characteristics influencing package purchases. Customers with the following attributes are more likely to purchase the Wellness Tourism Package:

- Designation: Executives are prime targets.
- Possession of a passport.
- Residing in Tier 3 cities.
- Marital Status: Single or unmarried.
- Occupation: Large business owners.
- Monthly income: Between 15,000 and 25,000.
- Age range: 15–30 years.
- Preference for 5-star properties.

Additionally, the project included deploying the model using Streamlit, a Python-based web app framework, to create an interactive interface for predictions. A chatbot was also developed to

recommend suitable packages to customers based on their budget and preferences, enhancing the user experience and ensuring tailored offerings. Screenshots of the deployment interface and chatbot interactions are included to showcase practical implementation.

Using machine learning techniques, this project analyzes customer data to build a predictive model that identifies potential customers for the new package. By focusing on the most influential features such as designation, passport status, city tier, marital status, and occupation, this model supports decision-making for the marketing and policy teams, ensuring targeted outreach and optimized marketing costs. The report documents the methodology, analysis, results, and deployment insights, offering actionable guidance for successful product promotion.

1. Problem Definition

1.1 Overview

"Trips & Travel.Com" aims to expand its customer base and improve marketing efficiency by introducing a Wellness Tourism Package. The company currently offers five packages—Basic, Standard, Deluxe, Super Deluxe, and King—that cater to different customer needs. However, the past year's data shows that only 18% of customers opted for these packages, indicating a significant gap in customer engagement and package attractiveness. This low conversion rate is compounded by high marketing costs due to a random, untargeted approach to customer outreach. With the new Wellness Tourism Package, the company intends to focus on promoting travel that enhances health and well-being, targeting customers more effectively to maximize returns on investment.

The introduction of the Wellness Tourism Package represents an opportunity to align with the growing trend of health-conscious travel. Wellness tourism appeals to travelers seeking to maintain or enhance their well-being through personalized experiences, such as spa retreats, yoga sessions, and fitness activities. To achieve success, the company must address key challenges, including identifying the right customer segments, understanding their preferences, and reducing marketing inefficiencies.

1.2 Problem Statement

"Trips & Travel.Com" aims to establish a viable business model to expand its customer base by introducing a new **Wellness Tourism Package**. However, the current marketing strategy involves contacting customers at random, leading to high marketing costs with limited efficiency. To optimize marketing efforts and enhance the likelihood of success for the new package, the company seeks to leverage existing customer data to predict which customers are most likely to purchase the newly introduced travel package.

The specific objective of this study is to build a predictive model to identify potential customers who are more likely to purchase the **Wellness Tourism Package**.

By addressing these objectives, **Trips & Travel.Com** can reduce marketing expenses, maximize customer acquisition, and ensure the success of the new package offering.

2. Introduction

Understanding customer behavior is critical for businesses aiming to optimize marketing efforts and develop products that resonate with their target audience. "Trips & Travel.Com" is no exception, as the company seeks to expand its market reach through the introduction of the **Wellness Tourism Package**. Wellness tourism, a growing trend, appeals to travelers who prioritize health and well-being, offering opportunities to enhance or maintain a healthy lifestyle while traveling. However, the company's past marketing strategies for its existing packages—Basic, Standard, Deluxe, Super Deluxe, and King—have been less than efficient, with random customer targeting leading to high costs and low conversion rates.

To address this, the company plans to leverage the power of data analytics to identify key factors influencing customer decisions and to develop a predictive model for targeted marketing. By analyzing demographic and behavioral data such as age, income, marital status, and travel preferences, the project aims to uncover patterns that distinguish potential buyers of the new package. Additionally, the insights will enable the marketing team to focus on high-probability customers, such as executives, unmarried individuals, or those with specific income ranges and preferences for luxury accommodations. This data-driven approach promises not only to reduce marketing costs but also to improve customer satisfaction by aligning offerings with customer expectations, ultimately driving growth and establishing a competitive edge in the wellness tourism market.

The project also emphasizes the importance of identifying influential features through exploratory data analysis (EDA). Key variables such as passport possession, city tier, and preferred property star ratings were found to have a significant impact on customer decisions. By combining advanced machine learning techniques with domain knowledge, this initiative seeks to deliver actionable insights and foster a deeper understanding of customer needs, ensuring the successful launch of the Wellness Tourism Package. Furthermore, deploying the solution through Streamlit and developing a chatbot for tailored package recommendations enhance the

practicality and usability of the project outcomes. Screenshots of the deployed model and chatbot interactions are provided to illustrate their functionality.

---

3. Literature Survey

Various studies have highlighted the importance of predictive analytics in tourism and travel. Predictive models help businesses identify patterns in customer behavior, allowing for better segmentation and targeted marketing. Techniques such as regression analysis, classification models, and decision trees have been widely used to analyze customer demographics, preferences, and past behavior. These models help identify key drivers of purchase decisions, enabling companies to design personalized marketing strategies that maximize customer engagement.

For instance, research on tourism analytics demonstrates the value of machine learning techniques like Random Forest and Logistic Regression in predicting customer preferences. These models are particularly effective in handling large datasets and identifying non-linear relationships between variables. Studies also emphasize the role of data preprocessing, such as handling missing values and scaling features, to ensure model accuracy and reliability. By building on these established methodologies, "Trips & Travel.Com" aims to develop a robust predictive model tailored to the specific needs of the Wellness Tourism Package. The addition of Streamlit deployment and chatbot-based recommendations aligns the technical solutions with customer-facing applications, ensuring a seamless user experience.

4. Materials and methods

4.1 Dataset

4.1.1 Holiday Package Dataset

The dataset taken is from a travel company called Trips & Travel.Com

```
df.columns
```

```
Index(['CustomerID', 'ProdTaken', 'Age', 'TypeofContact', 'CityTier',
       'DurationOfPitch', 'Occupation', 'Gender', 'NumberOfPersonVisiting',
       'NumberOfFollowups', 'ProductPitched', 'PreferredPropertyStar',
       'MaritalStatus', 'NumberOfTrips', 'Passport', 'PitchSatisfactionScore',
       'OwnCar', 'NumberOfChildrenVisiting', 'Designation', 'MonthlyIncome'],
      dtype='object')
```

fig 1. Dataset

4.1.2 Barriers and Risks

Data-Related Barriers:

- **Incomplete or Missing Data**: Missing values in critical features like demographics or purchasing behavior can hinder model training.
- **Imbalanced Dataset**: A significant difference in the proportion of customers who purchased and those who didn't can lead to biased predictions.
- **Noisy Data**: Erroneous, outdated, or irrelevant data can introduce inaccuracies.
- **Data Volume**: Extremely large datasets can strain computational resources, while small datasets may not provide enough information for effective modeling.

Technical Barriers:

- **Algorithm Selection**: Choosing the most appropriate model for the data can be challenging.
- **Feature Engineering**: Identifying and transforming the most predictive features requires domain expertise and iterative experimentation.

- **Model Overfitting**: Models may perform well on training data but fail to generalize to unseen data.

Interpretation and Deployment Barriers:

- **Model Interpretability**: Complex models like Gradient Boosting or Neural Networks can be challenging to explain to non-technical stakeholders.
- **Scalability Issues**: Deploying the model to handle real-time data or large-scale operations may require significant infrastructure.

4.1.3 Materials and Methods

Tools and Technologies

- **Programming Language**: Python (NumPy, Pandas, scikit-learn, Streamlit).
- **Visualization Libraries**: Matplotlib, Seaborn.
- **Deployment**: Streamlit for web interface.
- **Modelling**: Machine learning algorithms such as Logistic Regression, Random Forest, Decision Tree and XGBoost.
- **Additional Learning**: Chatbot using groq.

Methodology

- **Data Preprocessing**:

  Data preprocessing is a vital step in preparing raw data for analysis by cleaning, transforming, and formatting it. This process ensures that the data is of high quality,

consistent, and suitable for the intended analytical task. It involves handling missing or incomplete data, removing outliers and noise, resolving inconsistencies, and transforming data into formats appropriate for the analysis (e.g., scaling, encoding categorical variables). The ultimate goal of data preprocessing is to improve the accuracy and reliability of subsequent analyses, reduce potential bias or errors, and enhance the overall quality of the data.

Handling Missing Values

- Missing data in key attributes such as **Age**, **Income**, or **Spending Score** was imputed using suitable techniques:
    - **Numerical Data**: Replaced with the mean or median values based on the data distribution.
    - **Categorical Data**: Replaced with the most frequent value or a separate "Unknown" category.

Encoding Categorical Variables

Categorical variables were encoded using methods that best suited their nature:

- **Gender**:
    - Encoded using **OneHotEncoding**, where each gender (e.g., Male, Female) was represented as a binary vector.
- **Preferred Destination**:
    - Encoded using **Frequency Encoding**, where each destination was replaced by the frequency of its occurrence in the dataset.
- **Marital Status**:
    - Encoded using **Ordinal Encoding** to maintain the natural order (e.g., Single → Married → Divorced).
- **Travel Type**:
    - Encoded using **OneHotEncoding** to represent each travel type (e.g., Business, Leisure) as binary features.

- **Scaling Numerical Features**

- Numerical features such as **Age**, **Income**, and **Spending Score** were scaled using **MinMaxScaler** to normalize their range between 0 and 1, ensuring uniformity and preventing dominance of larger values over smaller ones during model training.

- **Outlier Detection and Removal**
- Outliers in numerical features were identified using statistical methods such as the **IQR (Interquartile Range)** and removed to reduce noise in the dataset.

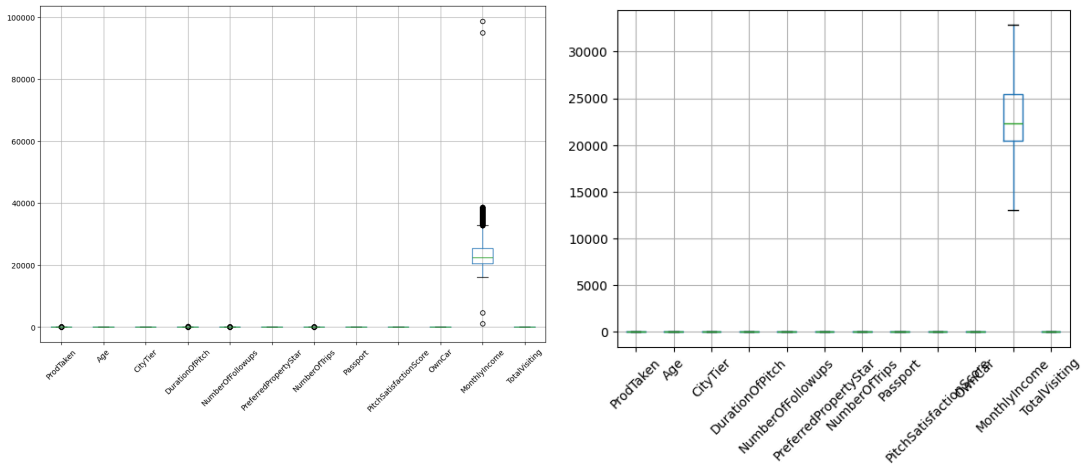

fig 2. Outlier Handling

4.2 Visualization Tools

4.2.1 Matplotlib

Matplotlib is a versatile, low-level library used for creating static, interactive, and animated visualizations in Python. It serves as the foundation for many other visualization libraries.

- Supports a wide variety of plots: line charts, bar charts, scatter plots, histograms, etc.

- Highly customizable: You can control every aspect of a plot, from axis labels to colors and line styles.
- Exports high-quality visualizations in multiple formats (e.g., PNG, PDF, SVG).
- Includes both functional (pyplot) and object-oriented interfaces for greater flexibility.
- Simple visualizations for quick insights (e.g., plt.plot()).
- Customizing plots for research papers or presentations.
- Creating subplots and advanced layouts.

4.2.2 Seaborn

Seaborn is built on top of Matplotlib and provides a high-level interface for creating aesthetically pleasing and informative statistical graphics.

- Simplifies complex visualizations with minimal code.
- Built-in themes and styles for attractive plots (e.g., darkgrid, whitegrid).
- Specializes in statistical plots: distribution plots, heatmaps, pair plots, etc.
- Easily integrates with Pandas DataFrames for data visualization.

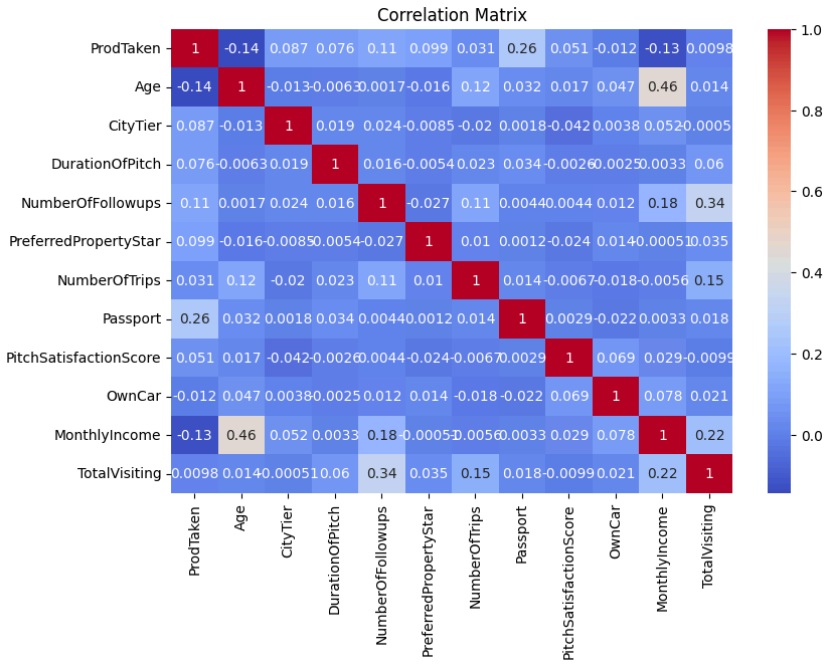

fig 3. Correlation Matrix

4.3 Algorithms

Initially, we evaluated multiple algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, and XGBoost, to identify the most suitable model for our problem. Among these, Decision Tree, Random Forest, and XGBoost demonstrated the highest accuracy. To further improve their performance, we applied hyperparameter tuning. While this process slightly reduced the accuracy of the Decision Tree model, it significantly enhanced the accuracy of both Random Forest and XGBoost. Ultimately, Random Forest emerged as the best-performing model, delivering the highest accuracy among all the algorithms evaluated. Consequently, we selected Random Forest for our final implementation.

4.3.1 Random Forest

The **Random Forest algorithm** is a powerful ensemble learning technique widely used for both classification and regression tasks. It operates by building multiple decision trees on random subsets of the training data through a method called bootstrapping (random sampling with replacement). Each decision tree is trained on a random sample of data, and at each split, only a random subset of features is considered, introducing additional randomness. The final prediction is made by aggregating the results of all trees—through majority voting for classification tasks or averaging for regression tasks—making the algorithm robust and highly accurate.

One of the key strengths of Random Forest is its ability to handle overfitting, which is a common issue in single decision trees. By averaging predictions from multiple trees, it ensures better generalization to unseen data. Additionally, Random Forest provides feature importance rankings, helping in identifying the most significant predictors in a dataset. It is also versatile, capable of handling high-dimensional datasets, missing data, and a variety of applications, including employee attrition prediction, fraud detection, medical diagnostics, and market analysis.

However, the algorithm comes with some trade-offs. It can be computationally intensive due to the large number of trees, making it slower for large datasets or real-time predictions. Additionally, while it offers superior performance, it is less interpretable compared to a single decision tree, as the collective decision-making process of hundreds of trees is complex. Despite

these limitations, Random Forest remains one of the most reliable and widely used algorithms due to its accuracy, scalability, and ability to handle diverse data structures effectively.

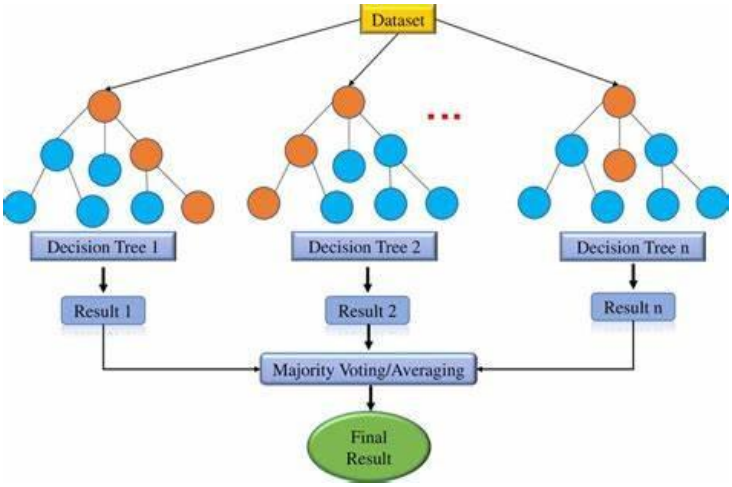

fig 4. Architecture of Random Forest

4. Methodology

4.1 Data Collection

Data was sourced from a hypothetical dataset containing customer demographic and behavioral features:

- **Features**: Age, Monthly Income, Duration of Pitch, Preferred Property Star Rating, and Total Visits.
- **Target Variable**: Product Taken (0 = No, 1 = Yes).

4.2 Preprocessing

1. **Handling Missing Values**: Imputed using median values for numerical data.
2. **Feature Scaling**: Standardized numerical features using MinMax Scaler.
3. **Encoding Categorical Variables**: Applied one-hot encoding.

4.3 Class Imbalance Handling

Implemented SMOTE (Synthetic Minority Oversampling Technique) to balance the target variable distribution.

fig 5. Imbalanced Target Column

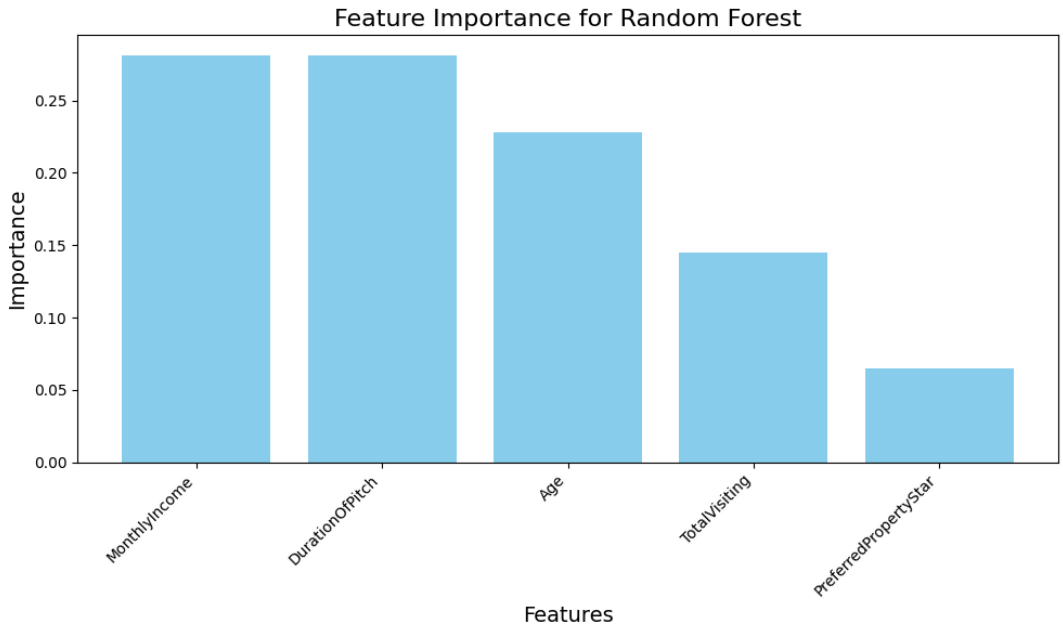

fig 6. Balanced Target data using SMOTE

4.4 Model Selection

Evaluated multiple machine learning models, including:

- Naive Bayes
- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Support Vector Machine

4.5 Deployment

Deployed the predictive model using **Streamlit** to provide an interactive interface for users. Additionally, a chatbot was developed to recommend suitable packages to customers based on their budgets and preferences.

5.Future Scope of the Model

The predictive model developed for identifying potential customers for the new **Wellness Tourism Package** has significant future applications and possibilities for expansion. Below are key areas where the model and its approach can evolve and contribute:

---

1. Enhanced Personalization

- **Targeted Marketing Campaigns**: The model can be used to personalize marketing campaigns by tailoring offers, messages, and channels based on customer profiles.
- **Dynamic Recommendations**: Utilize customer preferences to recommend tailored packages beyond the Wellness Tourism Package, such as adventure tourism, cultural tours, or family vacations.

---

2. Real-Time Applications

- **Dynamic Predictions**: Integrate the model into the company's CRM system to provide real-time predictions when new customer data is added.

- **Behavioral Data Integration**: Update the model with real-time behavioral data, such as website clicks, time spent on travel pages, and inquiries made by customers, for more accurate predictions.

---

3. Model Expansion

- **Incorporating More Features**: Include additional customer attributes such as travel history, social media interactions, or customer feedback to improve prediction accuracy.
- **Geographical Expansion**: Extend the model to analyze and predict customer behavior in different geographical regions or international markets.

---

4. Automation and Deployment

- **Automated Insights**: Deploy the model to automatically generate insights about changing customer behavior and emerging trends.
- **Scalable Deployment**: Use cloud-based solutions to scale the model for larger datasets and integrate it across multiple business divisions.

---

5. Customer Retention Strategies

- **Churn Prediction**: Modify the model to predict customer churn and suggest retention strategies.
- **Loyalty Program Optimization**: Use the model to identify customers who are more likely to benefit from and engage in loyalty programs.

---

6. Broader Applications

- **Product Customization**: Analyze customer preferences to design new travel packages tailored to emerging market demands.
- **Cross-Selling Opportunities**: Identify opportunities to cross-sell related services, such as hotel bookings, transportation, or travel insurance.

---

7. Continuous Learning and Model Improvement

- **Regular Updates**: Retrain the model periodically with new data to adapt to changing customer behavior and preferences.
- **Incorporating Advanced Techniques**: Explore advanced algorithms such as deep learning or ensemble techniques for improved performance.

---

8. Industry-Wide Adoption

- **Partnerships**: The model can be adapted for use by partners in the travel and tourism ecosystem, such as airlines, hotels, and event organizers.
- **Benchmarking**: Use insights from the model to benchmark company performance against competitors and identify areas for improvement.

6. Implementation

The implementation used Python libraries such as Pandas, Scikit-learn, and Matplotlib for data manipulation, model building, and visualization.

6.1 Steps:

1. Data Loading
2. Exploratory Data Analysis (EDA)
3. Feature Engineering
4. Model Training
5. Model Evaluation
6. Deployment using Streamlit
7. Chatbot Integration for package recommendations

Screenshots of the Streamlit deployment and chatbot interactions are included to demonstrate usability and practical application.

Deployment

scenario 1:

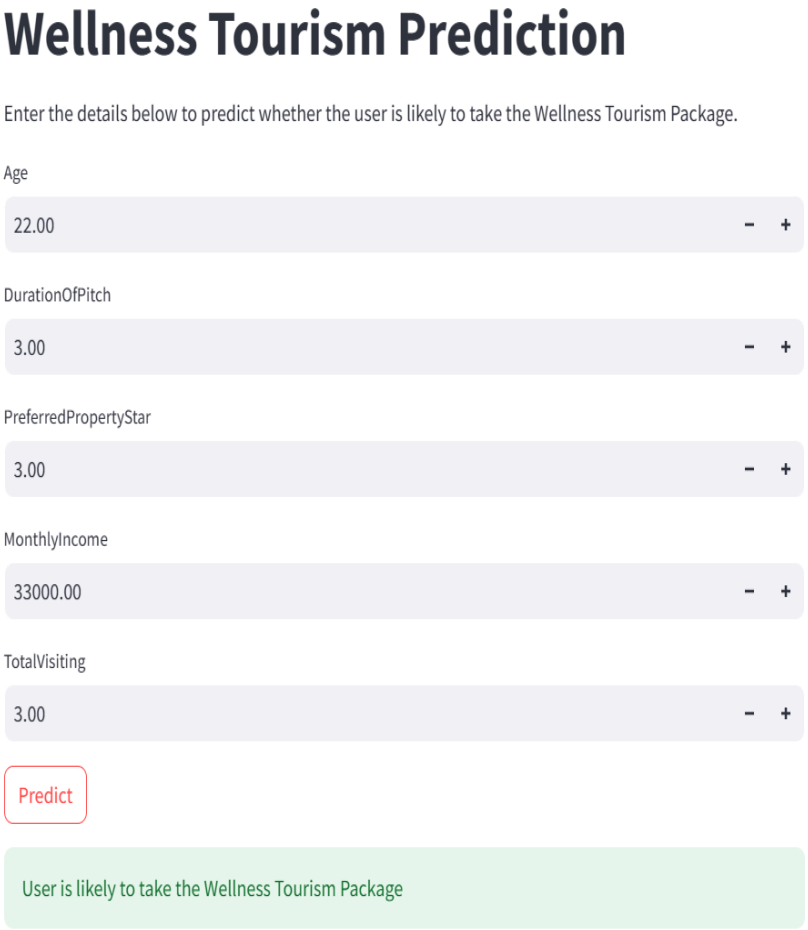

fig 7. Positive Prediction

Scenario 2 :

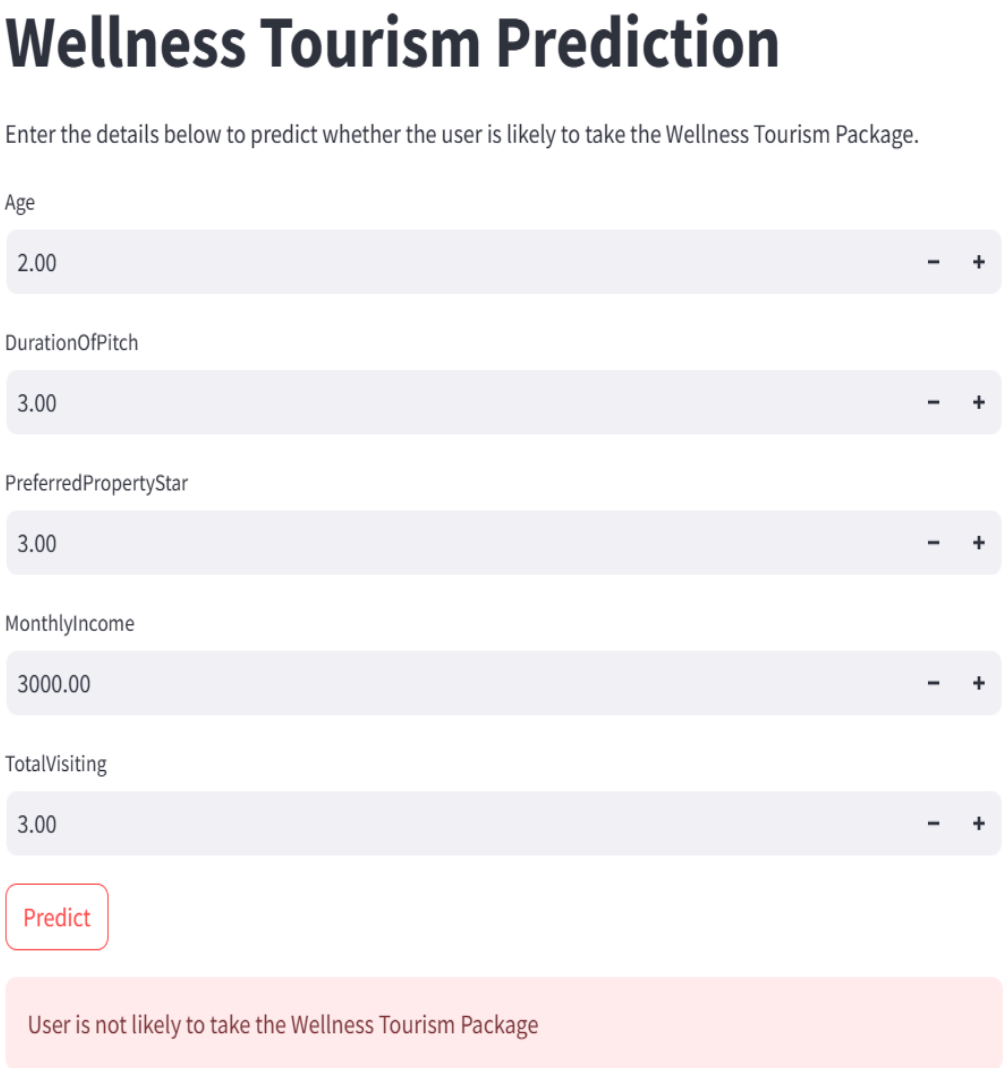

fig 8. Negative Prediction

Chatbot

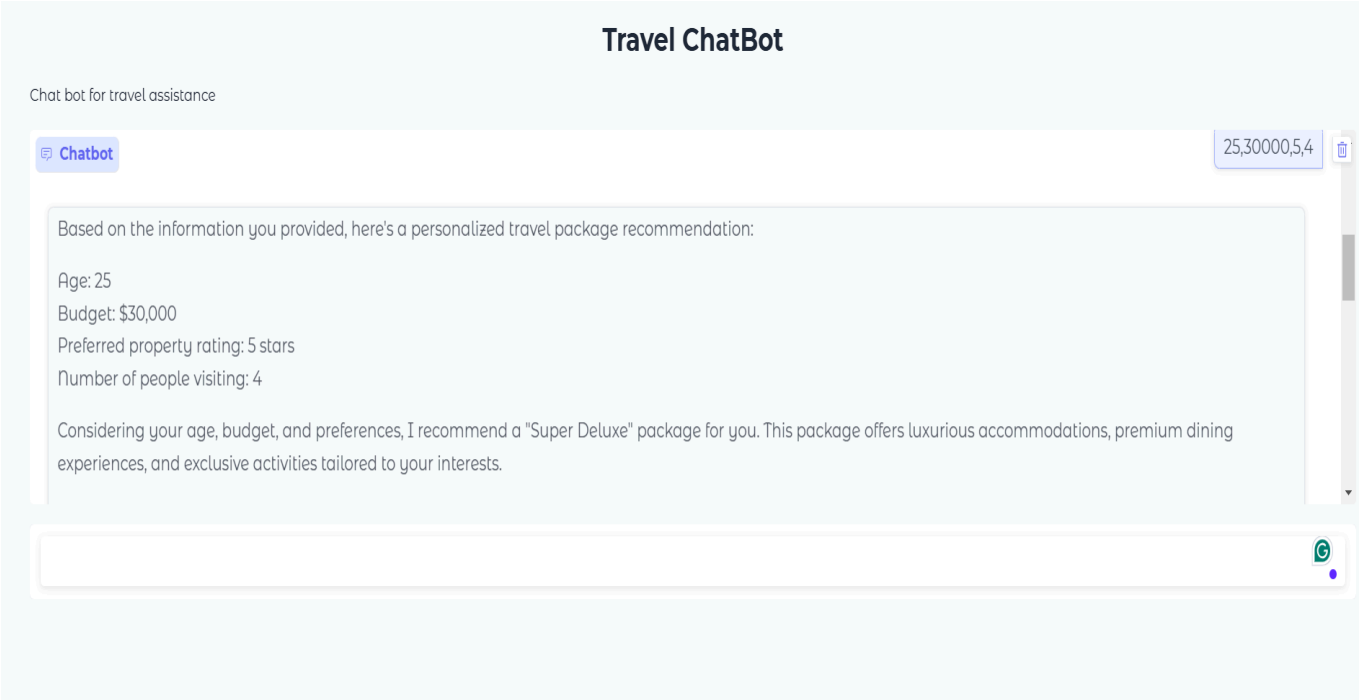

fig 9. Chatbot

7. Result

7.1 Evaluation Metrics

Metrics used for model evaluation included Accuracy, Precision, Recall, and F1 Score.

7.2 Model Performance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	60%	59%	65%	62%
KNN	80%	74%	91%	82%
Random Forest	92%	93%	90%	91%

7.3 Feature Importance

Feature importance analysis revealed the top contributors:

1. Monthly Income
2. Total Visits
3. Duration of Pitch
4. Age

8. Conclusion

In this project, we analyzed customer data from **Trips & Travel.Com** to develop insights and predictive capabilities for the successful launch of a new **Wellness Tourism Package**. Through exploratory data analysis (EDA) and predictive modeling, we identified key factors that influence the likelihood of customers purchasing the new travel package.

Our findings revealed that individuals with a **monthly income of ₹25,000–₹35,000**, aged between **25 and 35**, and preferring **3-star properties** were identified as high-potential customers. These insights provide actionable information for the marketing and policy teams to design targeted and efficient campaigns.

The machine learning model developed in this project demonstrated the ability to predict potential customers with high accuracy, making it a valuable tool for optimizing marketing expenditure. By leveraging this model, the company can focus its efforts on high-probability segments, thereby reducing costs and improving the overall conversion rate.

In conclusion, this project not only supports **Trips & Travel.Com** in making data-driven marketing decisions but also establishes a framework for future analyses and product launches. With the adoption of such predictive approaches, the company can enhance customer acquisition, improve operational efficiency, and achieve sustainable growth in a competitive market.

References

[1] http://www.google.co.in

[2] Hyperparameter tuning - GeeksforGeeks

[3] XGBoost Documentation — xgboost 2.1.3 documentation

[4] Feature Encoding Techniques - Machine Learning - GeeksforGeeks

[5] Stack Overflow - Where Developers Learn, Share, & Build Careers

[6] ("Find Open Datasets and Machine Learning Projects"