**Domain:Machine Learning**

**Machine learning –Based Prediction**

**on Travel Datasets**

**Abstract**

**This project presents a machine learning pipeline designed to analyze and classify travel-related data for predictive insights. The dataset, loaded from a CSV file (travel\_data.csv), undergoes extensive preprocessing including handling of categorical and numerical variables using encoding and feature scaling techniques. Exploratory data analysis (EDA) is performed through visualizations such as bar charts and heatmaps to uncover patterns and correlations within the data. This end-to-end approach demonstrates the application of supervised learning techniques in travel data analytics for business intelligence and decision-making.**

**Project Overview**

**An aviation company is shifting from tele-calling to a digital advertising strategy by partnering with a social media platform. The goal is to target customers with a high likelihood of booking domestic or international flights based on their digital and social behavior. Since user behavior differs by device, two separate predictive models will be developed—one for laptop users and another for mobile users. Given the high cost of digital ads, the models must be highly accurate to ensure effective and cost-efficient targeting.**

**Solution Overview**

**This project builds a machine learning model to classify travel data. It involves data loading, preprocessing (encoding and scaling), and exploratory analysis. Two classification models—Random Forest and XGBoost—are trained and evaluated using metrics like ROC-AUC and confusion matrix. The final model is saved using Pickle for future use. This program showcases how machine learning can generate insights and predictions from travel-related data**.

**Dataset Description**

* **User id - Unique identifier for each user**
* **Taken\_product - Target variable indicating whether the user has taken a travel product or service (Yes/No or 1/0)**
* **Yearly\_avg\_view\_on\_travel\_page- Average number of views on the travel page per year by the user**
* **preferred\_device-Device most frequently used by the user (e.g., Mobile, Desktop)**
* **total\_likes\_on\_outstation\_checkin\_given-Total number of likes the user has given to others' outstation check-ins**
* **yearly\_avg\_Outstation\_checkins-User's average number of outstation check-ins per year**
* **member\_in\_family-Number of family members associated with the user**
* **preferred\_location\_type -Preferred travel location type (e.g., Hill, Beach, Urban)**
* **Yearly\_avg\_comment\_on\_travel\_page-Average number of comments made by the user on the travel page each year**
* **total\_likes\_on\_outofstation\_checkin\_received-Total likes received by the user for their own outstation check-ins**
* **week\_since\_last\_outstation\_checkin-Number of weeks since the user’s last outstation check-in following\_company\_page Whether the user follows the company’s page (Yes/No or 1/0)**
* **montly\_avg\_comment\_on\_company\_page-Average monthly comments made by the user on the company’s page**
* **working\_flag-Whether the user is currently employed (Yes/No or 1/0)**
* **travelling\_network\_rating-A rating (likely out of 5 or 10) based on the user’s travel activity or influence in their network**
* **Adult\_flag-Whether the user is an adult (Yes/No or 1/0)**
* **Daily\_Avg\_mins\_spend\_on\_traveling\_page-Average daily time (in minutes) spent by the user on the travel page**

**Code Overview**

# **1.Importing the libraries:**

**Essential libraries for data handling (*pandas,numpy*),visualizations(*matplotlib,seaborn*),and machine learning(*scikit-learn,xgboost*)are imported**

**2.Loading the datasets:**

**The travel dataset is loaded using pd.read\_csv(“*travel\_data.csv*”)**

**3.Data preprocessing:**

**\*Handling Missing values,Duplicated values and Not applicable values**

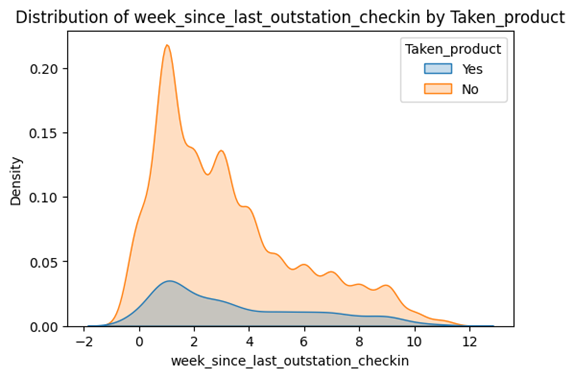
**\*Encoding Categorical Variables**

**\*Feature Scaling**

**\*Splitts training and testing datasets**

**4.Exploratory Data Analysis(EDA)** :

**It is used to understand the Structure,distribution and relationships within the travel\_data.csv dataset before training the classification models(*Random Forest and XGBoost*)**



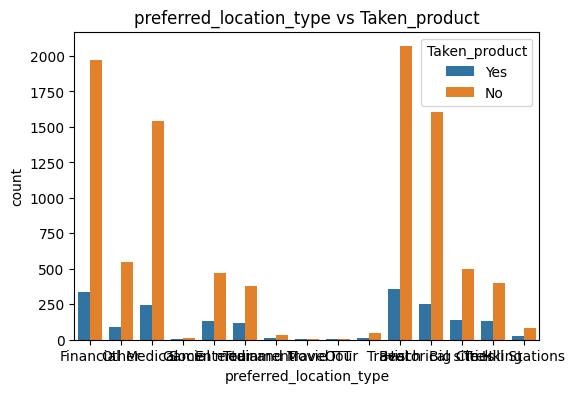
**This graph shows the** **distribution of the feature week\_since\_last\_outstation\_checkin for customers,** **separated by whether they took the product or not** (**Taken\_product = Yes or No).**

**\*X-axis: week\_since\_last\_outstation\_checkin – how many weeks ago the user last checked in at an outstation location.**

**\* Y-axis: Density – reflects the probability distribution (area under each curve = 1).**

**\* Legend:**

* **Blue (Yes): Users who took the product.**
* **Orange (No): Users who did not take the product**.



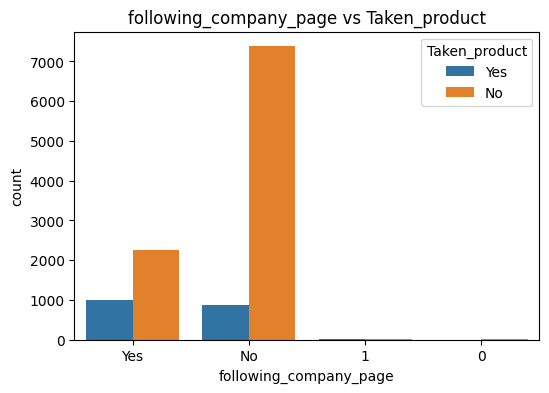
**This is a count plot showing how users' preferred location types are distributed between those who took the product (Taken\_product = Yes) and those who did not (Taken\_product = No).**

**\*X-axis: preferred\_location\_type — categories like Hill,Beach,Urban etc.**

**\* Y-axis: count — number of users in each category.**

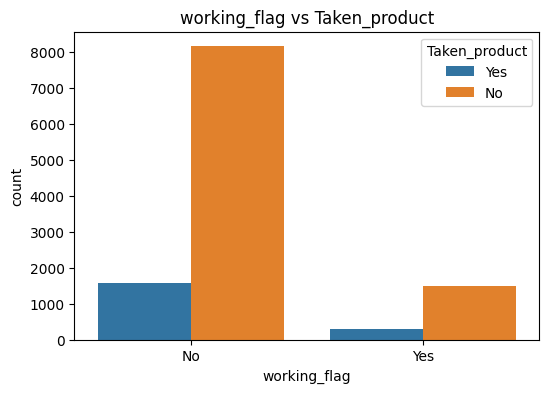
**\* Bars:**

* **Blue: Users who took the product (Taken\_product = Yes)**
* **Orange: Users who did not take the product (Taken\_product = No)**



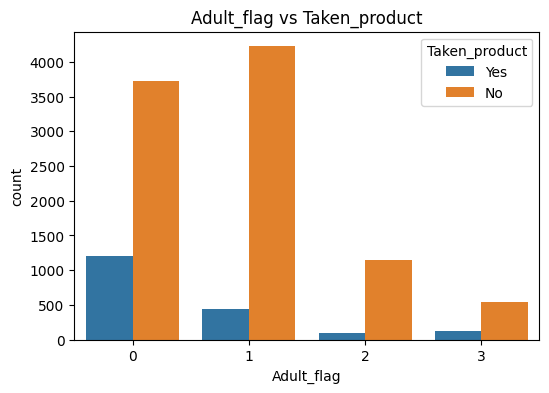
**This is a count plot showing the relationship between whether users are following the company page (following\_company\_page) and whether they took the product (Taken\_product = Yes or No).**

* **X-axis: Values of following\_company\_page (i.e., whether the user follows the company page or not).**
* **Y-axis: Count of users for each category.**
* **Bars:**
  + **Blue = Users who took the product (Yes)**
  + **Orange = Users who did not take the product (No)**



**This is a count plot showing how the variable working\_flag (i.e., whether a user is employed) relates to whether they took the product (Taken\_product = Yes or No).**

* **X-axis: working\_flag — has two categories:**
  + **No = user is not working**
  + **Yes = user is working**
* **Y-axis: Count of users**
* **Bar colors:**
  + **Blue = Users who took the product (Taken\_product = Yes)**
  + **Orange = Users who did not take the product (Taken\_product = No)**

****

**This is a count plot showing how the Adult\_flag feature relates to whether users took the product (Taken\_product = Yes or No).**

**\* X-axis: Values of the Adult\_flag column  
It appears to have values: 0, 1, 2, and 3 (likely indicating Whether the user is Adult or Not).**

**\* Y-axis: Count of users in each group.**

**\* Bars:**

* **Blue: Users who took the product (Taken\_product = Yes)**
* **Orange: Users who did not take the product (Taken\_product = No)**

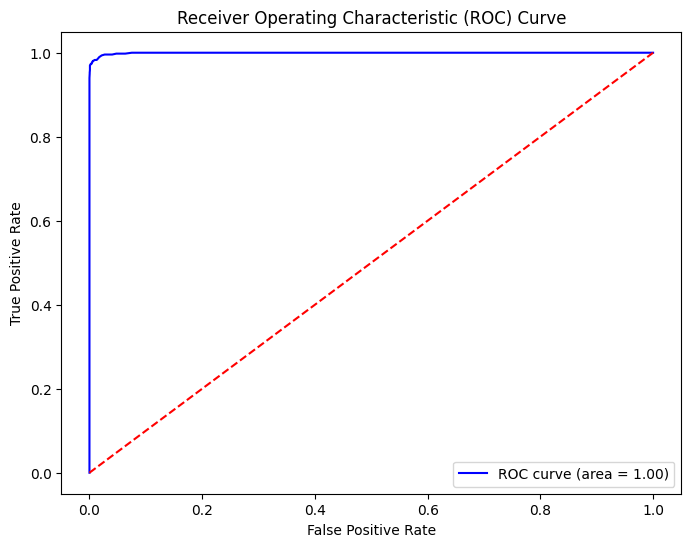
**5.Model Buliding:**

**\*Training the model**

**\*Make Predictions**

**\*Evaluating the Model**

**\*Plot ROC curve**



**This is a *Receiver Operating Characteristic (ROC) Curve*, which is used to evaluate the performance of a binary classification model. This refers to the accuracy of the project**

**\* X-axis: False Positive Rate (FPR)**

* **A lower value is better — fewer wrong positive predictions.**

**\* Y-axis: True Positive Rate (TPR)**

* **A higher value is better — more actual positives correctly predicted.**

**\* Blue Curve: This is your model’s performance.**

* **It plots the trade-off between TPR and FPR at different thresholds.**

**\*Red Dashed Line: This is the baseline for a random classifier (area = 0.5).**

* **Any useful model should perform above this line**.

Conclusion:

**This project applies a complete machine learning pipeline to a travel dataset to predict users' likelihood of purchasing a travel product. It includes data preprocessing, exploratory data analysis (EDA), feature engineering, and model training. Two ensemble models—Random Forest and XGBoost—were used and evaluated with accuracy metrics, confusion matrices, and ROC-AUC scores. The final model, optimized for performance, was serialized using Pickle for real-world deployment in targeted marketing or travel recommendation systems. The project demonstrates the practical value of machine learning in analyzing user behavior and driving data-driven decisions in the travel industry.**