Business Presentation

Visit with Us Case Study

Background & Context

- A **tourism company** named **"Visit with us"** is looking at **ways** to enable and establish a viable business model **to expand the customer base on the tourism sector**.
- **One** of their **ways** to expand the customer base **is to introduce a new offering of packages**.
- Currently, there are **5 types of packages** the company is offering - **Basic, Standard, Deluxe, Super Deluxe, King.**
- Looking at the data of the **last year**, we observed that **18% of the customers purchased the packages.**
- The company in the **last campaign contacted the customers at random** without looking at the available information.
- However, this time company is now planning to launch a new product i.e. **Wellness Tourism Package**.
- **Wellness Tourism** is defined as Travel that **allows the traveler to maintain, enhance or kick-start a healthy lifestyle**, and support or increase one's sense of well-being.
- The company wants to **make use the available data of existing and potential customers** to **make the marketing expenditure more efficient**.

Objective

- **Explore and visualize data**
- **Build a model**

  Model should be able to

  - To predict whether a customer will buy a Wellness tour package or not.
  - Which variables are most significant.
  - Which segment of customers should be targeted more.
- **Draw Conclusions and Business Recommendations**

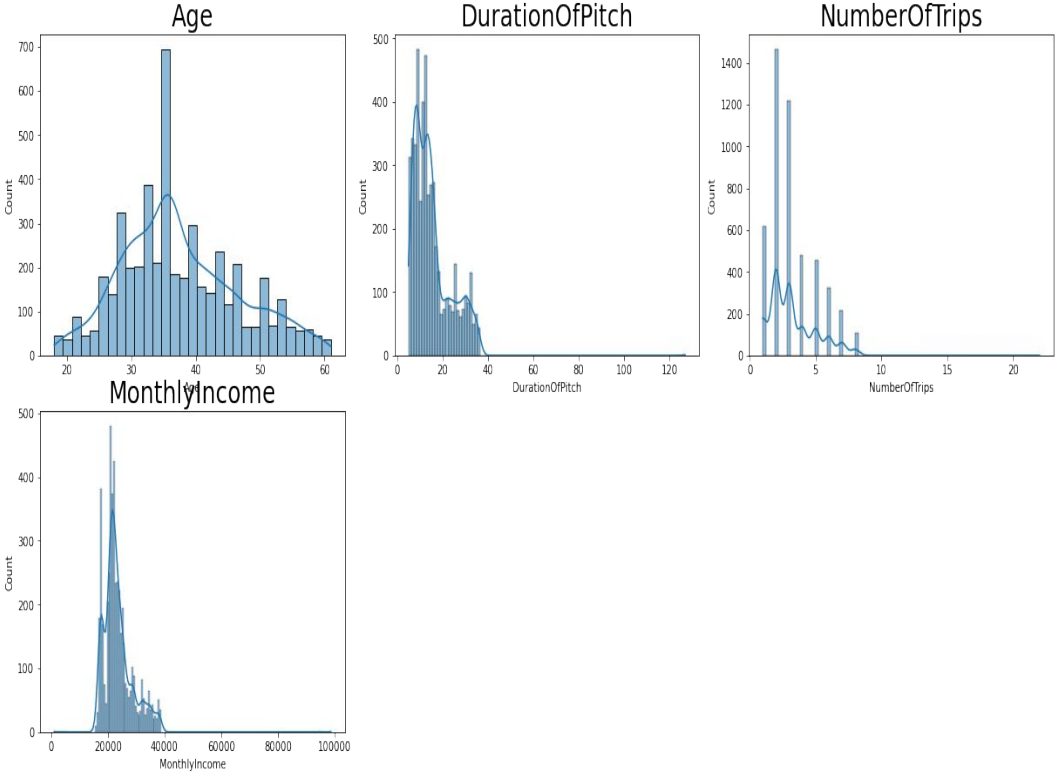
Data Overview

- Data contains 4888 (rows) customer data with 20 (columns) characteristics.
- There are null values in many columns which we will be imputing with median value for analysis.
- There are no duplicates in the dataset.
- The Customer ID column is of no significance and will be removed for analysis.
- The Gender column will be mapped to Male and Female.'Fe Male' will be grouped under Female.
- Marital Status columns has been grouped into Married and Unmarried bucket.
- Age, DurationOfPitch,NumberOfTrips,MonthlyIncome, Mortgage, CCAvg will be binned and mapped to new range columns respectively.

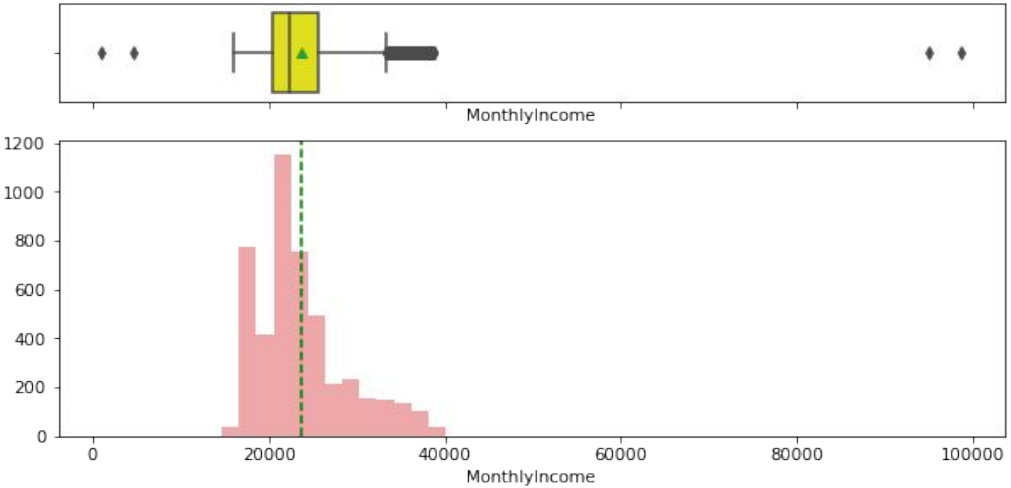
Variable	Description
CustomerID	Unique customer ID
ProdTaken	Whether the customer has purchased a package or not (0
Age	Age of customer
TypeofContact	How customer was contacted (Company Invited or Self Inquiry)
CityTier	City tier depends on the development of a city, population, facilities, and living standards. The categories are ordered i.e. Tier 1 > Tier 2 > Tier 3
Occupation	Occupation of customer
Gender	Gender of customer
NumberOfPersonVisiting	Total number of persons planning to take the trip with the customer
PreferredPropertyStar	Preferred hotel property rating by customer
MaritalStatus	Marital status of customer
NumberOfTrips	Average number of trips in a year by customer
Passport	The customer has a passport or not (0 - No , 1 - Yes)
OwnCar	Whether the customers own a car or not (0 - No , 1 - Yes)
NumberOfChildrenVisiting	Total number of children with age less than 5 planning to take the trip with the customer
Designation	Designation of the customer in the current organization
MonthlyIncome	Gross monthly income of the customer
PitchSatisfactionScore	Sales pitch satisfaction score
ProductPitched	Product pitched by the salesperson
NumberOfFollowups	Total number of follow-ups has been done by the salesperson after the sales pitch
DurationOfPitch	Duration of the pitch by a salesperson to the customer

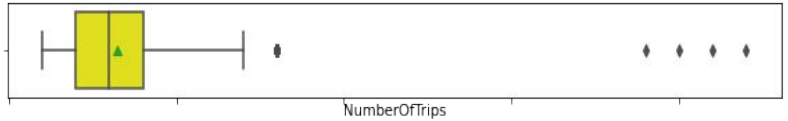
Exploratory Data Analysis - Univariate

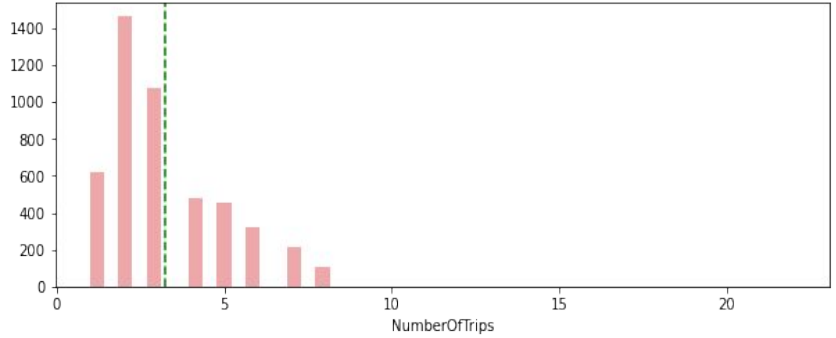
- Age: Average age of the customers in the dataset is 37 years, most of them is concentrated between 30-40 years of age has a wide range from 18 to 61 years with almost a normal distribution.
- DurationOfPitch: Most of the data is concentrated between 5 and 35 minutes.On average the duration of pitch is around 15 minutes.the data is right skewed with a long right tail indicating a lot of outliers on the right end.
- NumberOfTrips:The average number of trips for the customers is 3. the data is heavily right skewed indicating that there are outliers on the right tail.
- MonthlyIncome: MonthlyIncome data is concentrated between 20k - 40K.The average Income of the customers is around(23k USD).The data is heavily right skewed indicating the presence of lot of outliers present in the variable on the right tail. MonthlyIncome has outliers on both ends 1-20K and 40k to 210k.

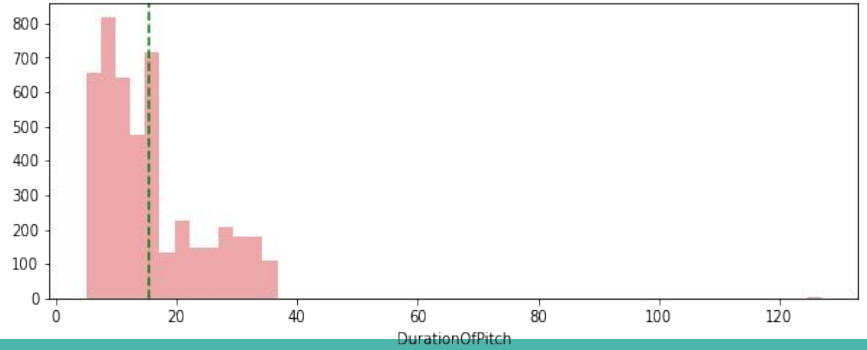


Exploratory Data Analysis - Univariate









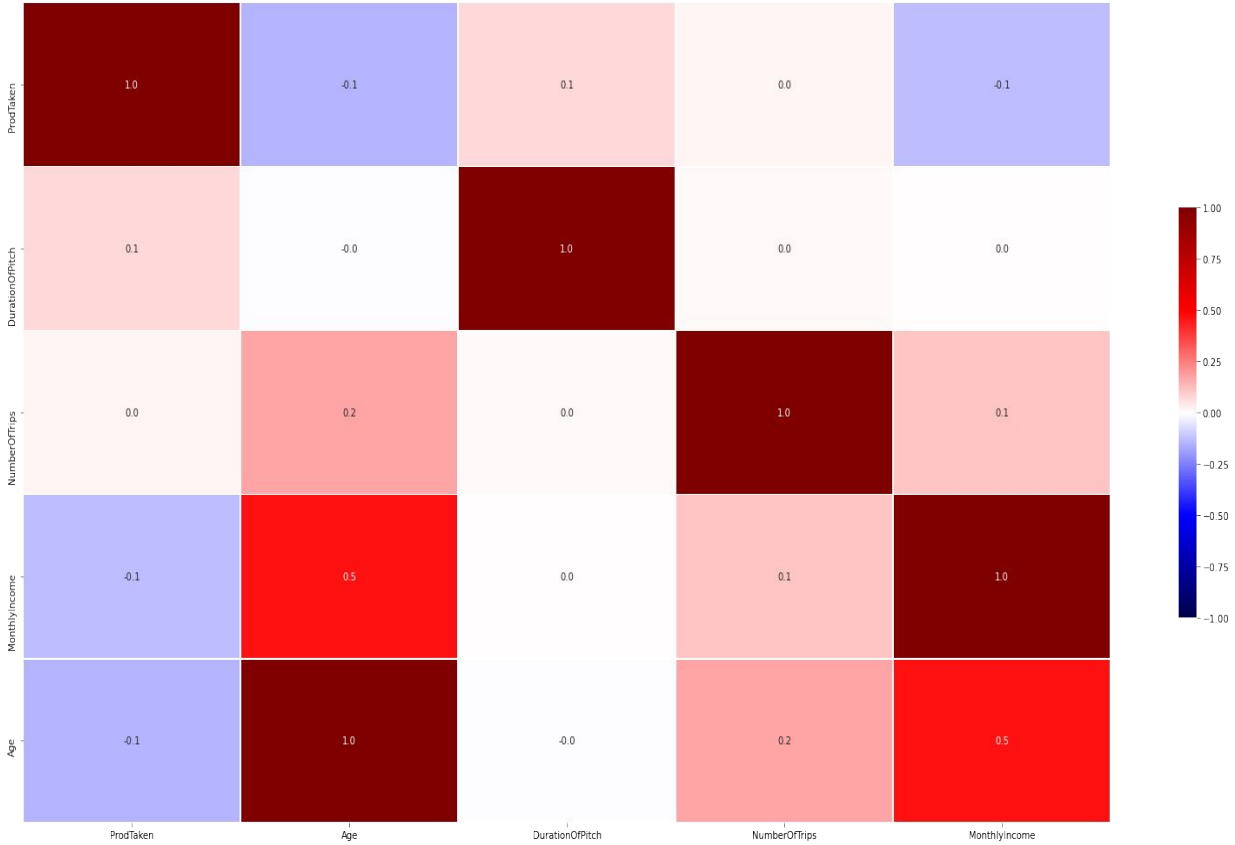
MonthlyIncome:We can see 2 extreme outlier values between 95K to 100k and 0 to 5K at both the tails of boxplot.

DurationOfPitch:We can see that there are clear 2 extreme outliers values post 120 duration that matches the boxplot and the histogram.

NumberOfTrips:There are 4 extreme outlier values between 19 and 22.

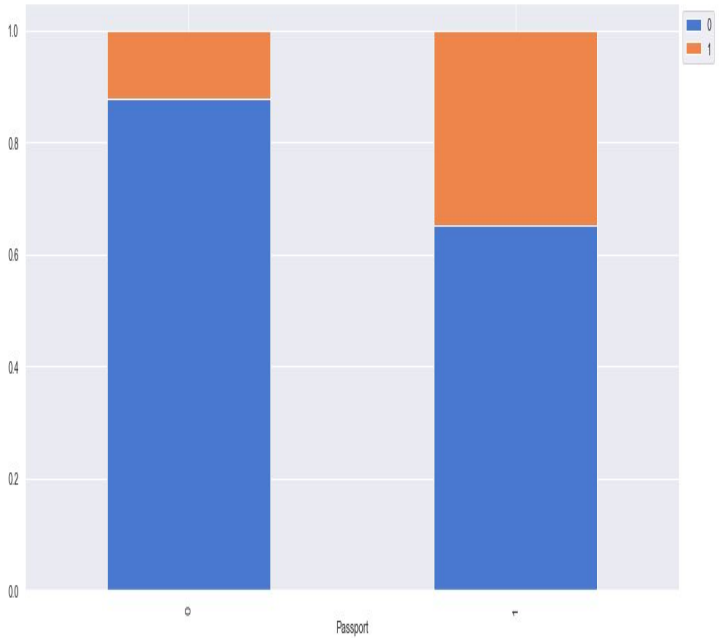
Exploratory Data Analysis - Bivariate

- Age and MonthlyIncome are positively correlated with a correlation of 0.5 which is not a surprise as the Age increases, the experience also increases and so they get higher salaries.
- We do not see any other strongly correlated variables in this data.
- ProdTaken vs Age and MonthlyIncome are negatively correlated indicating less the Age and MonthlyIncome more chances of customer buying the package but it is not a very strong correlation rather a weak one.
- DurationOfPitch has no correlation with Age, MonthlyIncome or the number of Trips and very weak correlation with Product taken.

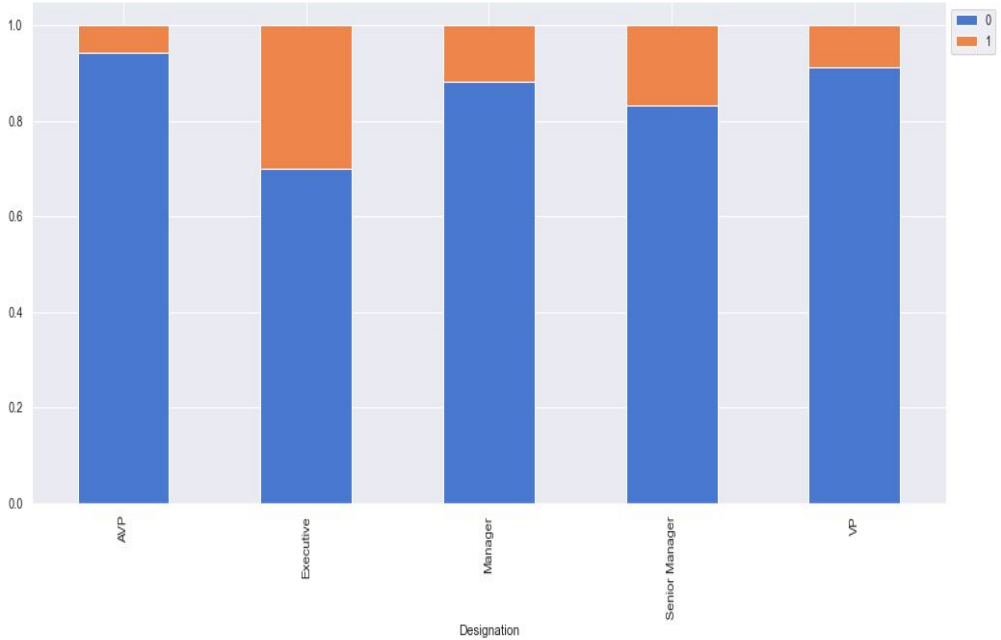


Exploratory Data Analysis - Bivariate

Passport vs Product Taken



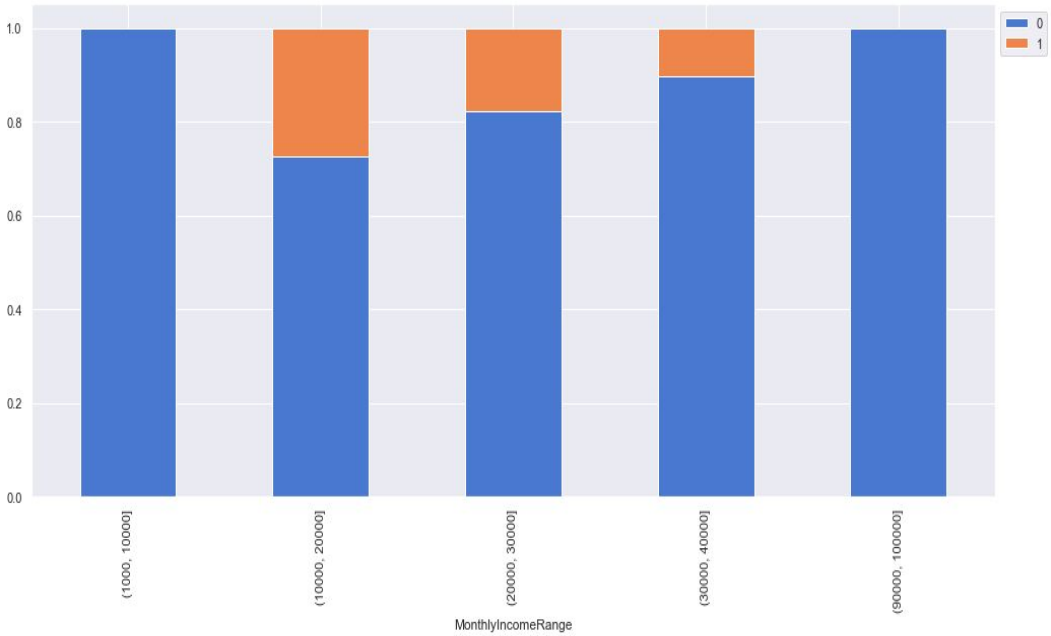
Designation vs Product Taken



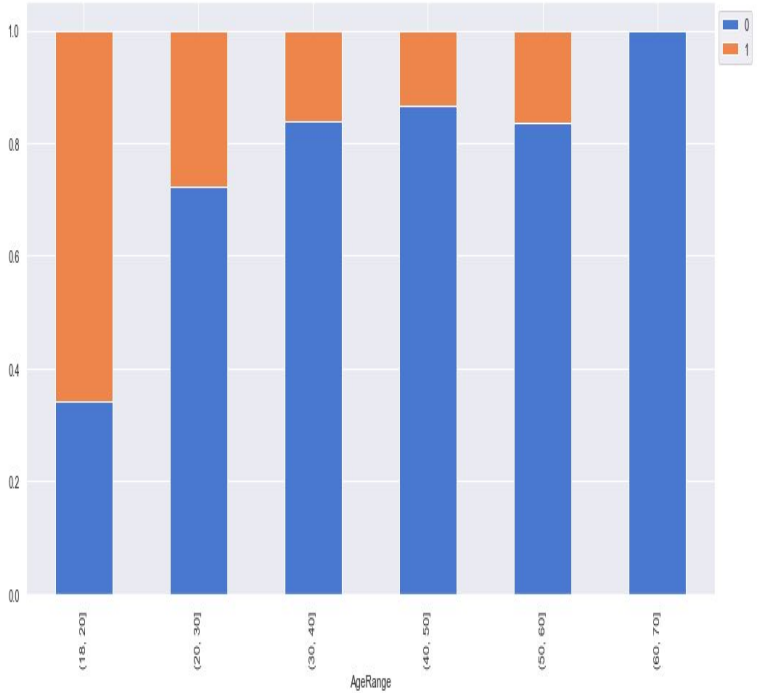
- **Customers who have passport travel more and hence buys the Tour package more.**
- **Customers whose Designation are Executives buys the Tour package more.**

Exploratory Data Analysis - Bivariate

MonthlyIncome vs Product Taken



Age vs Product Taken



- **Most of customers who has bought the tour package fall in the Income Range 10 -30K.**
- **Most customer who has bought the tour package are 18- 30 years of Age.**

Exploratory Data Analysis - Customer Profiling - Basic Package

- :
- Basic package can be targeted for customers between 20-40 years of age
- Customers whose Monthly Income is between 10-20k
- Customers with Children
- Customers who did Self Enquiry has got the Basic package the most.
- Customers who hold a passport and own a car
- Salaried customers prefer Basic package and Deluxe package
- Large Business customers prefer the Basic package.
- Male prefer Basic package.
- Unmarried customers prefer Basic package the most.
- Customers living in Tier 1 and Tier 2 cities prefer the Basic Tour package.
- Customers with PitchSatisfaction score 1,3,4,5 prefer Basic package

Exploratory Data Analysis - Customer Profiling - Deluxe package

- Deluxe Package can be targeted for customers between 30-50 years of age.
- Married customers prefer Deluxe package
- Customers with Children
- Customers living in Tier 3 cities prefer the Deluxe Tour package.
- Company Invited Customers have preferred Deluxe package
- Small Business prefer the Deluxe package.
- Female prefer Deluxe package
- Customers who travel more than 3 times a year prefer Deluxe package
- Customers who hold a passport and own a car
- Salaried customers prefer Basic package and Deluxe package
- Customers with PitchSatisfactionScore 2 prefer Deluxe package
- Duration of Pitch for 10-15 duration or 30-40 duration

Exploratory Data Analysis - Customer Profiling - Other packages

- **King Package**
  a. King Package can be targeted for customers between 40-60 years of age.
  b. Customers whose Monthly Income is between 30-40k
- **Standard Package**
  a. Customers whose Monthly Income is between 20-30k
  b. Standard Package can be targeted for customers between 30-40 years of Age
- **Super Deluxe Package**
  a. Customers whose Monthly Income is between 30-40k prefer the Super Deluxe Package
  b. Super Deluxe package can be targeted for customers between 40-60 years of age.

Model Performance Comparison- Bagging, Random Forest, Decision Tree

- In our case of predicting Tour Package Buyers ,not being able to identify a potential customer is the biggest loss Business can face.
- Recall is the right metric to check the performance of the model.
- Among Decision Tree,Bagging and Random Forest, Bagging with tuning gave us the best recall results with 79% on training and 81% on test data.

Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_f1score	Test_f1score
Decision Tree	1.00	0.88	1.00	0.68	1.00	0.67	1.00	0.68
Bagging	0.99	0.90	0.97	0.58	1.00	0.82	0.98	0.68
Bagging with class weights	0.99	0.90	0.97	0.54	1.00	0.88	0.99	0.67
Random Forest	1.00	0.90	1.00	0.50	1.00	0.94	1.00	0.65
Random Forest with weights	1.00	0.89	1.00	0.47	1.00	0.93	1.00	0.62
Decision Tree Tuned	0.77	0.77	0.69	0.69	0.43	0.44	0.53	0.54
Bagging Tuned	0.62	0.62	0.79	0.81	0.31	0.31	0.44	0.45
Random Forest Tuned	1.00	0.92	1.00	0.66	1.00	0.89	1.00	0.76

Model Performance Comparison- Boosting, Stacking

- In our case of predicting Tour Package Buyers ,not being able to identify a potential customer is the biggest loss Business can face.
- Recall is the right metric to check the performance of the model.
- Comparing all the boosting and stacking models,the tuned XGBoost Model gave the highest test recall results of 84%.
- Hyperparameter tuning helped with overfitting and improved the recall scores
-

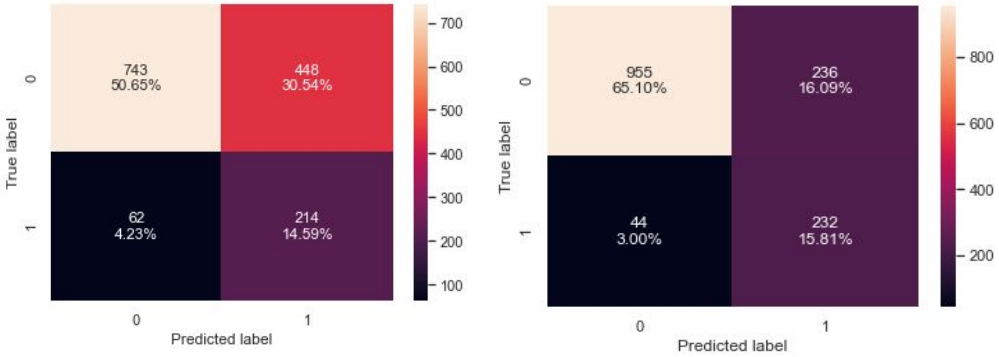
Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_f1score	Test_f1score
AdaBoost with default paramters	0.85	0.85	0.30	0.29	0.71	0.75	0.42	0.42
AdaBoost Tuned	0.99	0.89	0.94	0.68	0.99	0.72	0.96	0.70
Gradient Boosting with default parameters	0.88	0.88	0.44	0.42	0.89	0.84	0.59	0.56
Gradient Boosting with init=Adaboost	0.88	0.87	0.43	0.39	0.90	0.81	0.58	0.53
Gradient Boosting Tuned	0.92	0.88	0.59	0.47	0.94	0.81	0.72	0.60
XGBoost with default parameters	1.00	0.92	0.99	0.68	1.00	0.89	1.00	0.77
XGBoost Tuned	0.83	0.81	0.90	0.84	0.53	0.50	0.67	0.62
Stacking Estimator	1.00	0.91	1.00	0.70	0.98	0.80	0.99	0.75

Model Performance - All Models Comparison

- In our case of predicting Tour Package Buyers ,not being able to identify a potential customer is the biggest loss Business can face.
- Recall is the right metric to check the performance of the model.
- We used hyperparameter tuning technique and that helped to overcome overfitting and improved the recall scores.
- Using the tuned XGBoost Model we got the highest test recall results of 84% and we were able to predict the False Negatives at 3% and False positives at 16% which gives us a good model with low error false negative rates.

Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_f1score	Test_f1score
Decision Tree	1.00	0.88	1.00	0.68	1.00	0.67	1.00	0.68
Bagging	0.99	0.90	0.97	0.58	1.00	0.82	0.98	0.68
Bagging with class weights	0.99	0.90	0.97	0.54	1.00	0.88	0.99	0.67
Random Forest	1.00	0.90	1.00	0.50	1.00	0.94	1.00	0.65
Random Forest with weights	1.00	0.89	1.00	0.47	1.00	0.93	1.00	0.62
Decision Tree Tuned	0.77	0.77	0.69	0.69	0.43	0.44	0.53	0.54
Bagging Tuned	0.62	0.62	0.79	0.81	0.31	0.31	0.44	0.45
Random Forest Tuned	1.00	0.92	1.00	0.66	1.00	0.89	1.00	0.76
AdaBoost with default paramters	0.85	0.85	0.30	0.29	0.71	0.75	0.42	0.42
AdaBoost Tuned	0.99	0.89	0.94	0.68	0.99	0.72	0.96	0.70
Gradient Boosting with default parameters	0.88	0.88	0.44	0.42	0.89	0.84	0.59	0.56
Gradient Boosting with init=Adaboost	0.88	0.87	0.43	0.39	0.90	0.81	0.58	0.53
Gradient Boosting Tuned	0.92	0.88	0.59	0.47	0.94	0.81	0.72	0.60
XGBoost with default parameters	1.00	0.92	0.99	0.68	1.00	0.89	1.00	0.77
XGBoost Tuned	0.83	0.81	0.90	0.84	0.53	0.50	0.67	0.62
Stacking Estimator	1.00	0.91	1.00	0.70	0.98	0.80	0.99	0.75

Confusion Matrix Comparison - Top Two Models Bagging,Boosting



- **As False Negatives are opportunity cost, it is more expensive in our case than False positives which can be better planned and handled.**
- **Hence XGBoost with hyperparameter tuning gives the best results with low False negative rates and highest recall 84% on test data.**
- **Bagging with hyperparameter tuning also gives us equally good result with no overfitting but the precision scores are very low as compared to XGBoost tuned version.However the recall is very close at 81% and this performed even better in test than training data.**

**** True Positives:***
Reality: A customer buys a tour package
Model predicted: The customer will buy a tour package.
Outcome: The model is good.

**** True Negatives:***
Reality: A customer did NOT buy a tour package.
Model predicted: The customer will NOT buy a tour package.
Outcome: The business is unaffected.

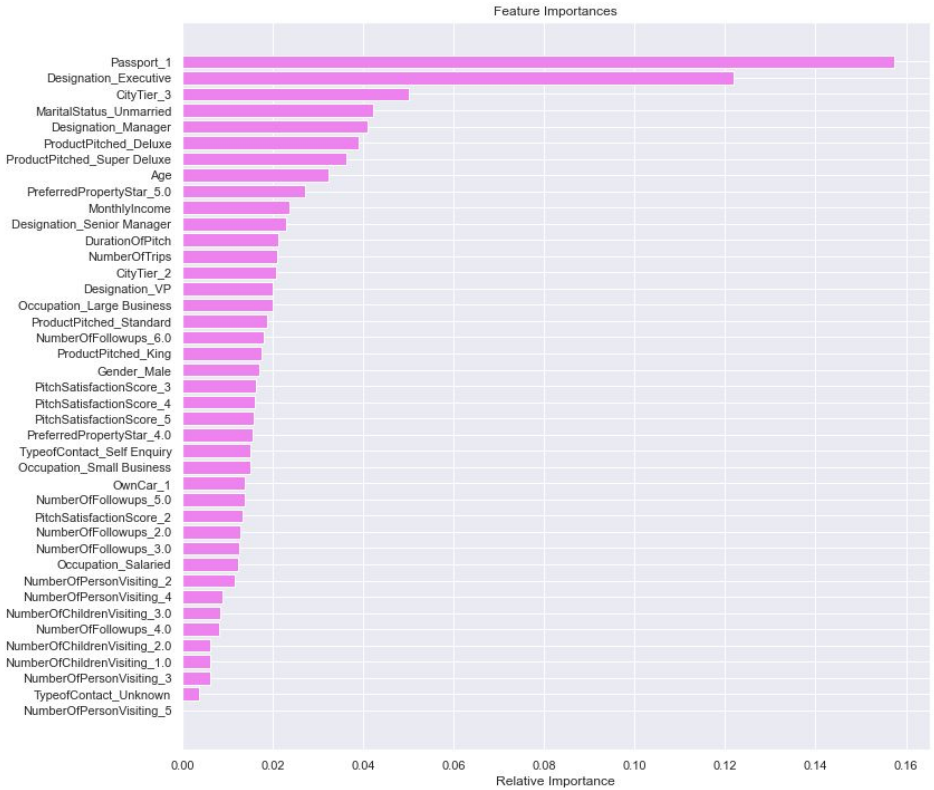
**** False Positives:***
Reality: A customer did NOT buy a tour package.
Model predicted: The customer will buy a tour package.
Outcome: The team which is targeting the potential customers will be wasting their resources on the people/customers which will not be a very big loss compared to losing a customer who will buy a tour package..

**** False Negatives:***
Reality: A customer buys a tour package.
Model predicted: The customer will NOT buy a tour package..
Outcome: The potential customer is missed by the sales/marketing team, the team could have offered the potential customer some discount or loyalty card to make them a regular customer. (Customer retention will get affected.)

Feature Importance of Top Model - XGBoost Tuned

The model indicates that the most significant predictors of Potential Tour Package buyers are

1. Customers who **hold a passport**
2. Customers whose **Designation is Executive**
3. Customers living in **City Tier 3.**
4. Marital Status as **Unmarried.**



Conclusion

- **XGBoost tuned with hyperparameter tuning technique** gave us the best model compared to all other boosting and stacking models with the **highest recall score of 90% on train and 84% on test data** and the least error of **3% False negatives** and 16% False positives.
- **Bagging tuned with hyperparameter tuning technique** also gave us a equally good model that performed even better in test data than training data **with a recall score of 76% on training data and 77% on test data a**nd the error of **4.23% False negatives** and 30.54% False positives.
- **Bagging tuned** gave the second best results as compared to Random Forest and Decision Tree Models.
- The **bivariate results of EDA clearly matched with the the XGBoost tuned Model predicted important variables**.

Recommendations

- **Basic** and **Deluxe** packages are very popular and the most preferred packages among customers.
- **Unmarried people** with Designation **Executives** living in **City Tier 3** who holds a **passport** are the potential customers to target for Sales pitching
- **Detailed Customer profiling** for every package has been listed and can be used for Sales pitching.
- Employ the predictive model to target the potential customers and market real time deals and offers only to those customers to save the advertising cost.