greatlearning
Power Ahead

Travel Package Purchase Prediction Project Business Presentation

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Data Preparation
- Model Performance Summary
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

- The Policy Maker of a tourism company named "Visit with us" wants to enable and establish a viable business model to expand the customer base. A viable business model is a central concept that helps in understanding the existing ways of doing the business and how to change the ways for the benefit of the tourism sector.
- One of the 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.

Business Problem Overview and Solution Approach

- However, the marketing cost was quite high because customers were contacted at random without looking at the available information.
- The 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.
- This time company wants to harness the available data of existing and potential customers to make the marketing expenditure more efficient.

Business Problem Overview and Solution Approach

- There is a need to analyze the customers' data and information to provide recommendations to the Policy Maker and Marketing Team as well as build a model to predict the potential customer who is going to purchase the newly introduced travel package.
- The objective of the model is:
  - To predict which customer is more likely to purchase the newly introduced travel package.

Data Overview

Variable	Description
ID	Unique customer ID
ProdTaken	Whether the customer has purchased a package or not (0: No, 1: Yes)
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

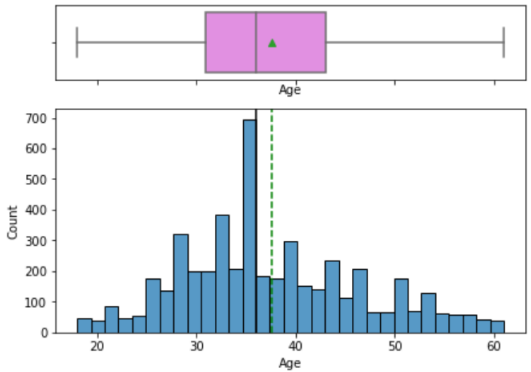
Observations	Variables
4888	20

Note:

- ID column is removed.
- Column value formatting is done for Gender (Fe Male to Female) and MaritalStatus (Single and Unmarried merged to Single).
- Missing values are filled with median values or rows removed where impact is small (~0.5%).
- Duplicated rows are removed where impact is small (< 3%).
- Outliers are capped at lower and upper whiskers of IQR.
- Age and MonthlyIncome variables are binned into categorical groups and converted to Age_bin and Income_bin variables.
- All columns except DurationOfPitch and NumberOfTrips have been converted to category.

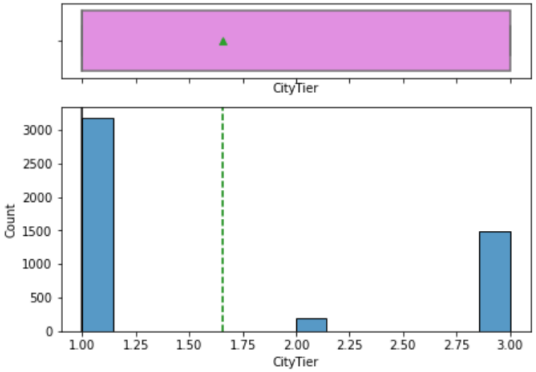
EDA – Age, CityTier & DurationOfPitch

Age



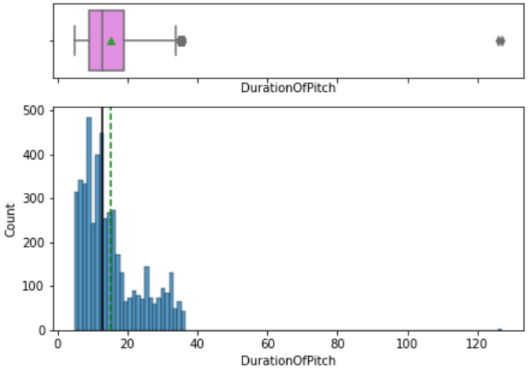
- The distribution of age is normal and no outliers.
- A significant number of customers are aged around 36.

CityTier



- A vast majority of customers come from firstly Tier 1 (3173 / 65.25%) then Tier 3 (1492 / 30.68%) cities.

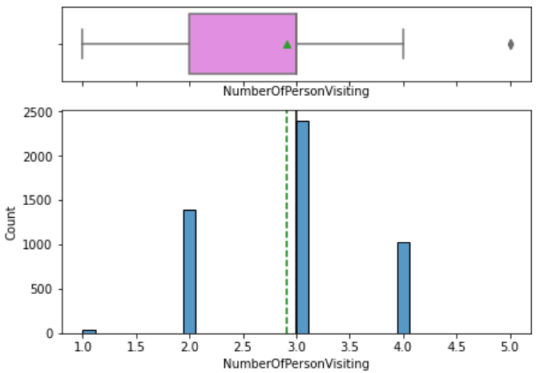
DurationOfPitch



- A vast majority of customers pitch duration are below 15 with low number of outliers indicating extremely long pitch duration.

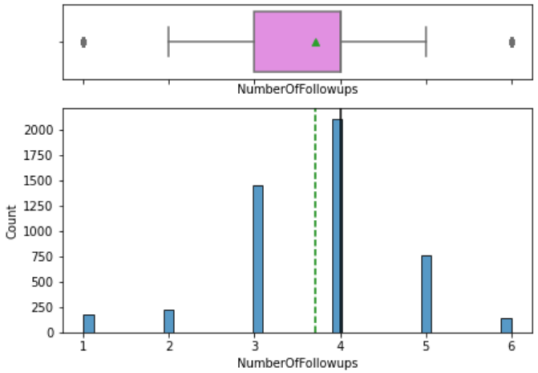
EDA – NumberOfPersonVisiting, NumberOfFollowups & PreferredPropertyStar

NumberOfPersonVisiting



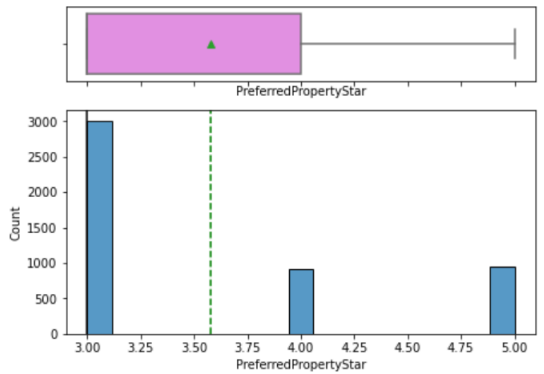
- A majority of customers have 3 people on the trip with them followed by 2 then 4 with some outliers at 5.

NumberOfFollowups



- A majority of follow ups are 4 times followed by 3 times with outliers at 1 follow-up and 6 follow ups.

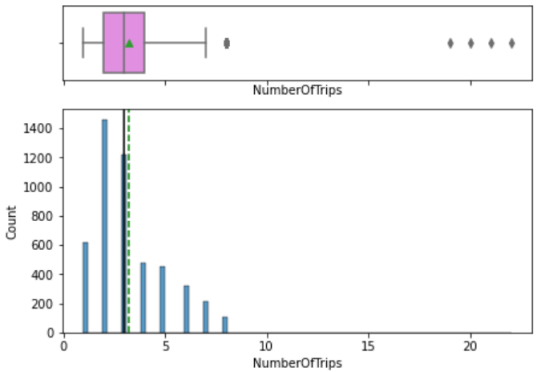
PreferredPropertyStar



- A majority of customers preferred a 3 star to 4 or 5 star ratings hotel property.

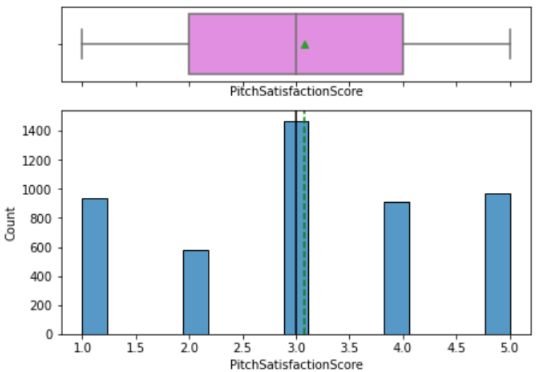
EDA – NumberOfTrips, PitchSatisfactionScore & NumberOfChildrenVisiting

NumberOfTrips



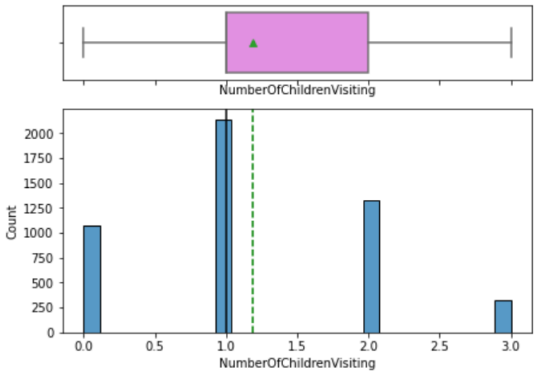
- A majority of trip numbers are 3 and below with a small number of outliers indicating a significantly higher number of trips.

PitchSatisfactionScore



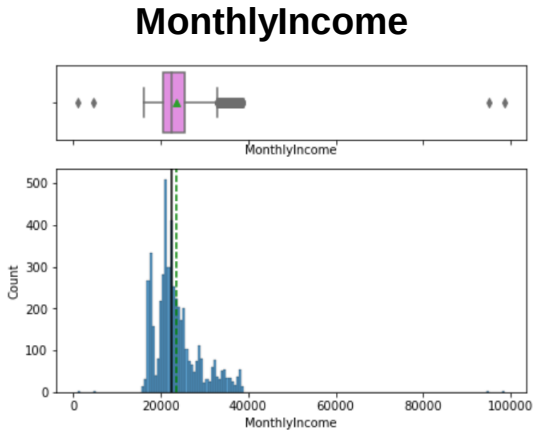
- Most pitch satisfaction ratings are 3.0 / 5.0.

NumberOfChildrenVisiting



- Most customers have up to 2 children on the trip with 1 child being the most common.

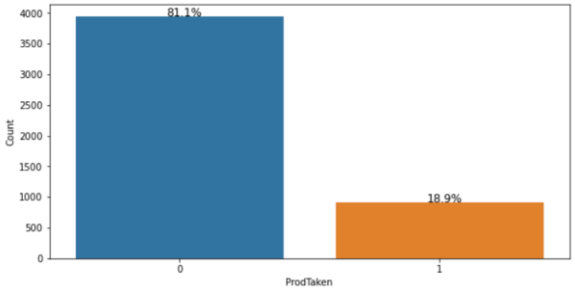
EDA – MonthlyIncome



- Most customers have monthly income around 24K with a small number of outliers of significantly higher income customers.

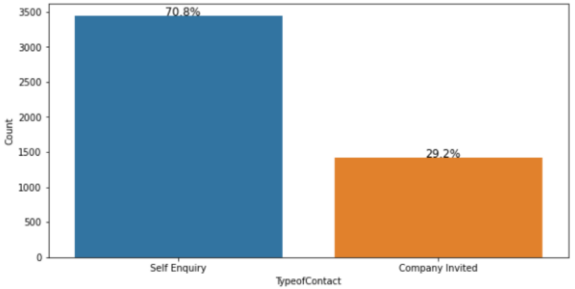
EDA – ProdTaken, TypeofContact & Occupation

ProdTaken



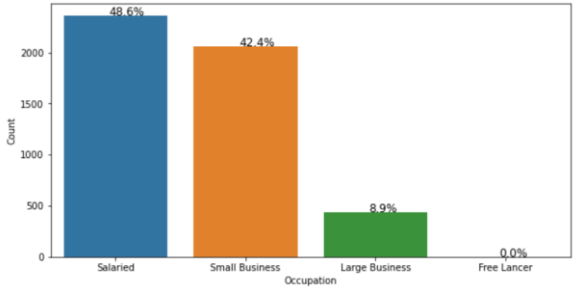
- 18.9% of customers in dataset took the package while vast majority at 81.1% did not.

TypeofContact



- Most customers at 70.8% self enquired to contact the company.

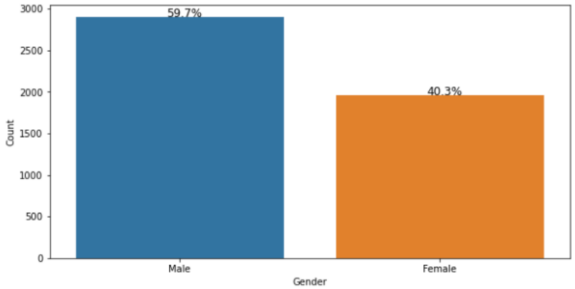
Occupation



- Most customers are salaried employees followed by small business.

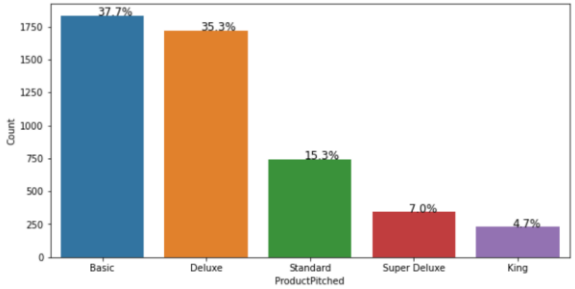
EDA – Gender, ProductPitched & MaritalStatus

Gender



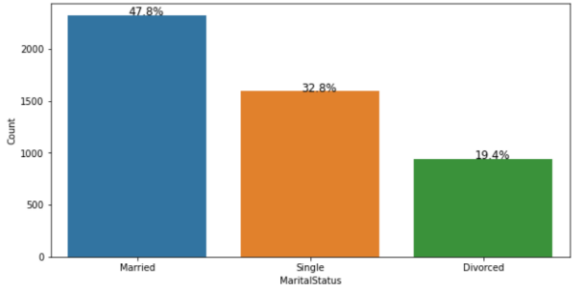
- More customers are male at 59.7%.

ProductPitched



- Most product packages pitched were Basic and Deluxe packages.

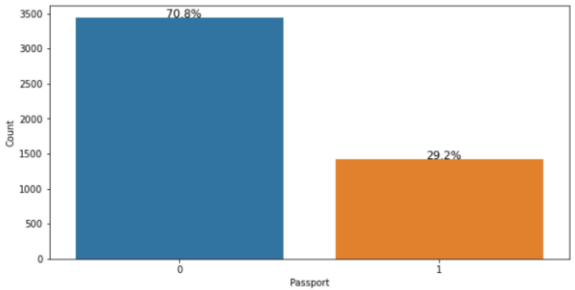
MaritalStatus



- Most customers are married followed by singles.

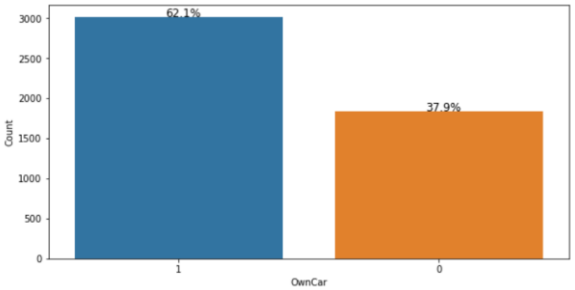
EDA – Passport, OwnCar & Designation

Passport



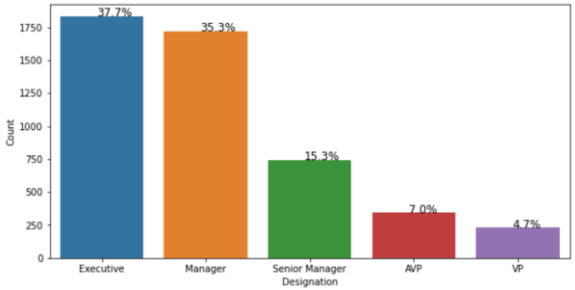
- Most customers at 70.8% do not have a passport.

OwnCar



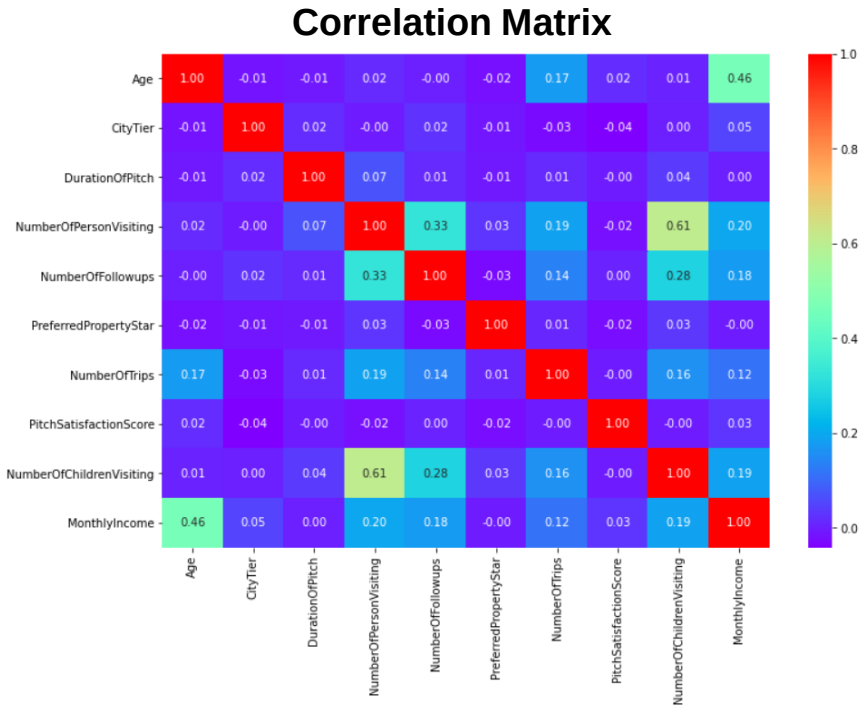
- Most customers at 62.1% own a car.

Designation



- Most customers are executives or managers to senior managers.

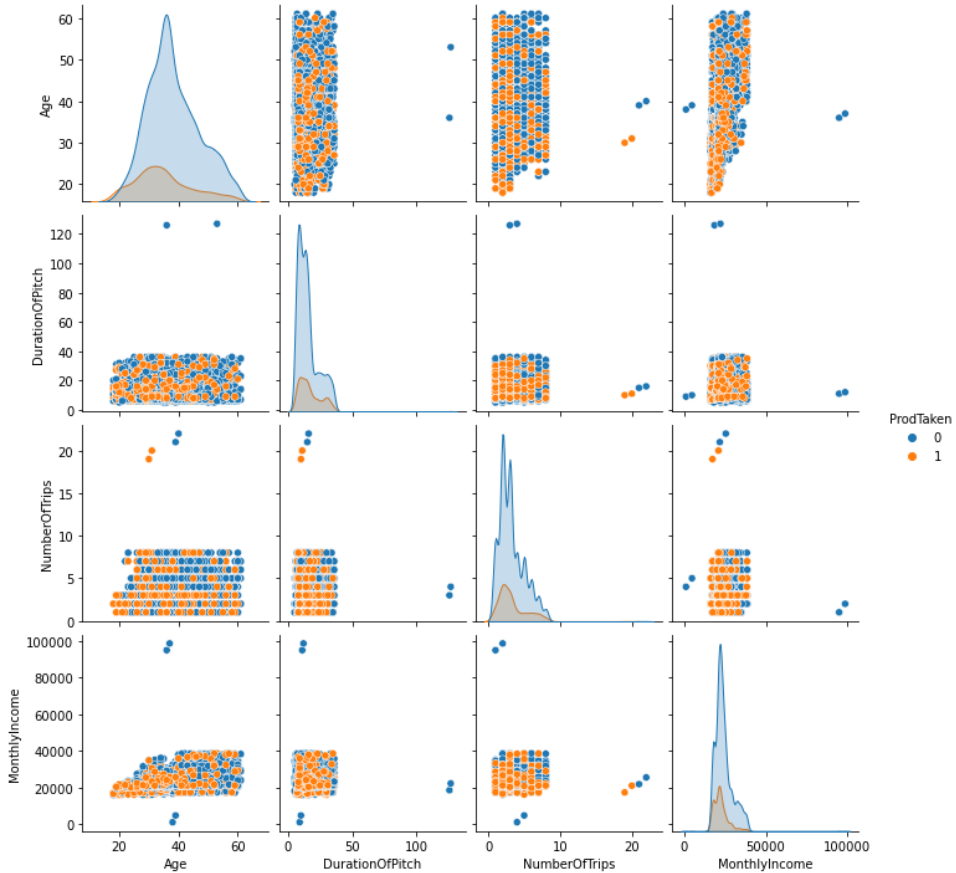
EDA – Correlation Matrix



Observations

- NumberOfPersonVisiting and NumberOfChildrenVisiting seemed to be mildly positively correlated as both indicate the number travelling with the customer.
- There does not seem to be any other correlation among numeric variables thus some variables can be converted to categorical: 'CityTier', 'NumberOfPersonVisiting', 'NumberOfFollowups', 'PreferredPropertyStar', 'PitchSatisfactionScore', 'NumberOfChildrenVisiting'

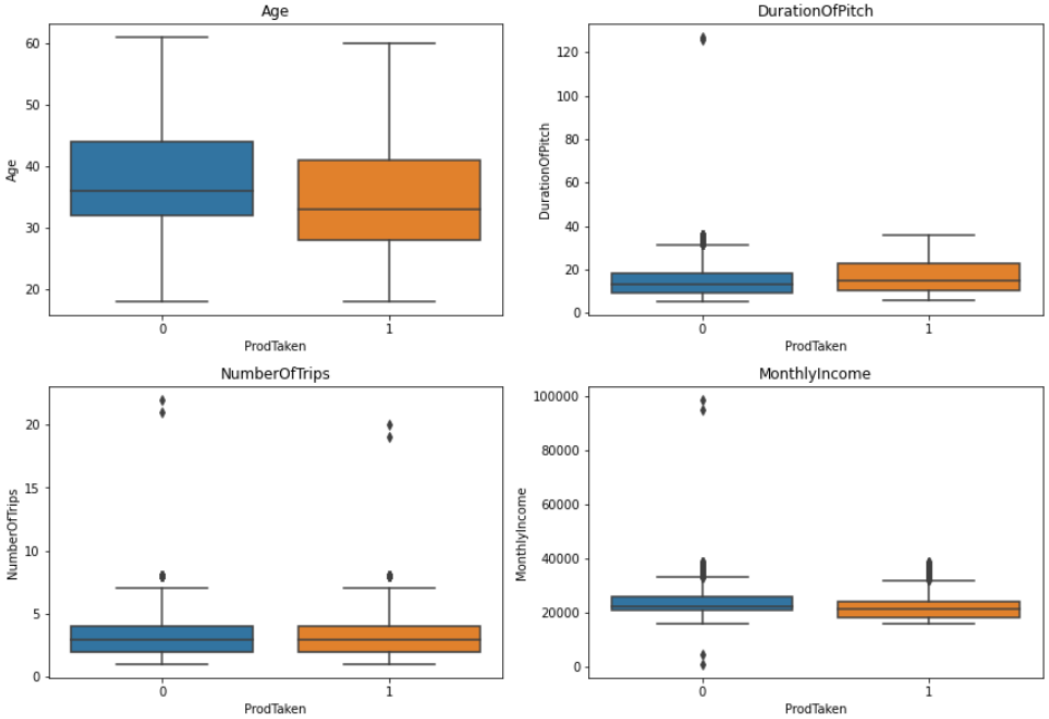
EDA – Pairplot



Observations

- There does not seem to be any discernible pattern in the remaining numeric variables.

EDA – ProdTaken with Numeric Values



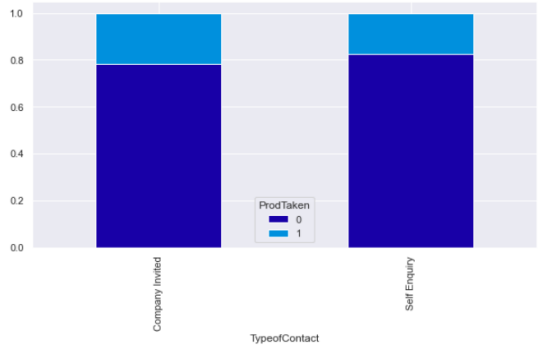
Observations

- Age median and distribution from 25th to 75th percentile of package taking customers are lower than non-package taking customers. The age range of customers taking up packages is between ~ 28 to just above 40.
- DurationOfPitch median and distribution are rather similar between package and non package takers with non package takers registering a number of outliers with higher pitch durations. Package takers have a marginally higher median and distribution.
- NumberOfTrips, for customer number of trips in a year, median and distribution are rather similar between package and non package takers with both registering a number of outliers on the higher end of number of trips.
- Monthly Income median and distribution are rather similar between package and non package takers with both registering a number of outliers. Non Package takers have a marginally higher median and distribution as well as lower end and much higher end outliers in monthly income.

EDA – ProdTaken with TypeofContact, CityTier & Occupation

ProdTaken Vs TypeofContact

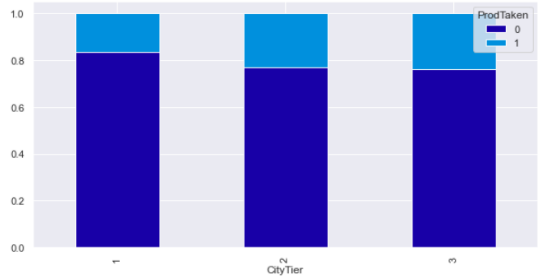
ProdTaken	0	1	All
TypeofContact			
Company Invited	1109	310	1419
Self Enquiry	2837	607	3444
All	3946	917	4863



- Company invited customers are slightly more likely to take up of a package.

ProdTaken Vs CityTier

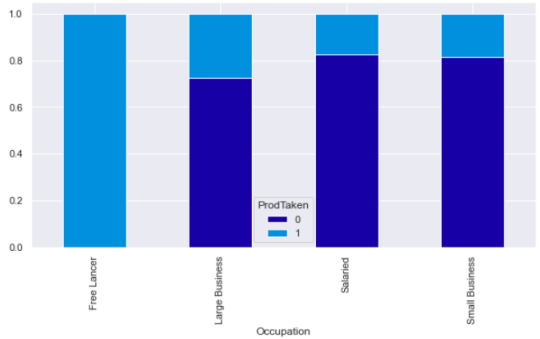
ProdTaken	0	1	All
CityTier			
1	2655	518	3173
2	152	46	198
3	1139	353	1492
All	3946	917	4863



- Tier 2 and 3 resident customers are slightly more likely to take up a package than Tier 1 residents.

ProdTaken Vs Occupation

ProdTaken	0	1	All
Occupation			
Free Lancer	0	2	2
Large Business	314	120	434
Salaried	1950	413	2363
Small Business	1682	382	2064
All	3946	917	4863

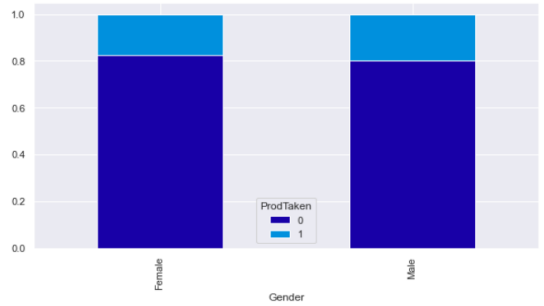


- Free Lancers are very much likely to take up a package followed by Large Business customers. However this might not be statistically significant given that there are only 2 freelancers in the data set.

EDA – ProdTaken with Gender, NumberOfPersonVisiting & NumberOfFollowups

ProdTaken Vs Gender

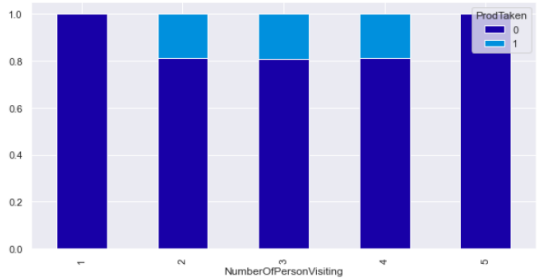
ProdTaken	0	1	All
Gender			
Female	1620	342	1962
Male	2326	575	2901
All	3946	917	4863



- There is no discernible difference in either gender of customers to determine a higher take up of a package.

ProdTaken Vs NumberOfPersonVisiting

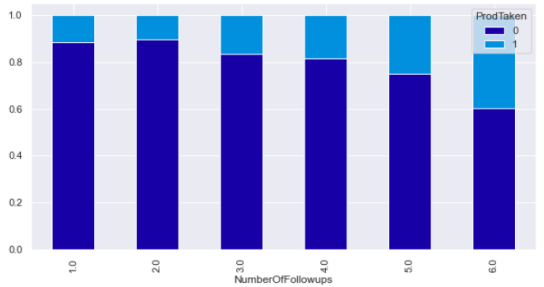
ProdTaken	0	1	All
NumberOfPersonVisiting			
1	39	0	39
2	1136	265	1401
3	1935	459	2394
4	833	193	1026
5	3	0	3
All	3946	917	4863



- 2 to 4 travel companions with the customer are more likely to take up a package.

ProdTaken Vs NumberOfFollowups

ProdTaken	0	1	All
NumberOfFollowups			
1.0	152	20	172
2.0	205	24	229
3.0	1214	242	1456
4.0	1720	386	2106
5.0	573	191	764
6.0	82	54	136
All	3946	917	4863

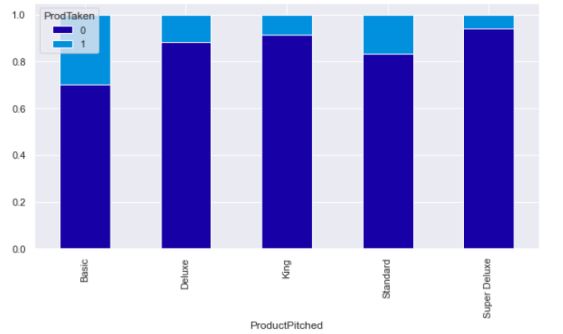


- Among customers, a higher number of follow ups to customers will lead to higher success rate to take up a package.

EDA – ProdTaken with ProductPitched, PreferredPropertyStar & MaritalStatus

ProdTaken Vs ProductPitched

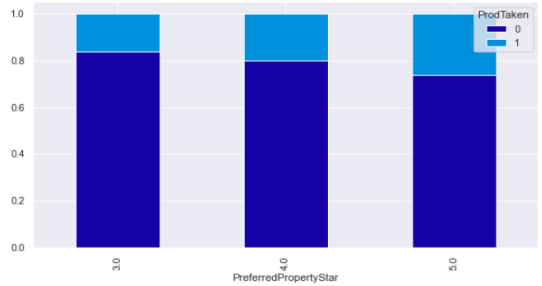
ProdTaken	0	1	All
ProductPitched			
Basic	1283	549	1832
Deluxe	1513	204	1717
King	210	20	230
Standard	618	124	742
Super Deluxe	322	20	342
All	3946	917	4863



- Basic followed by Standard then Deluxe packages have higher success rate of take up when pitched to customers.

ProdTaken Vs PreferredPropertyStar

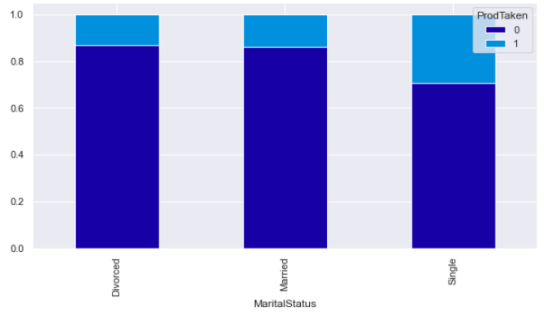
ProdTaken	0	1	All
PreferredPropertyStar			
3.0	2514	486	3000
4.0	730	182	912
5.0	702	249	951
All	3946	917	4863



- 5 star hotel preferred customers are more likely to take up a package.

ProdTaken Vs MaritalStatus

ProdTaken	0	1	All
MaritalStatus			
Divorced	821	123	944
Married	1999	326	2325
Single	1126	468	1594
All	3946	917	4863

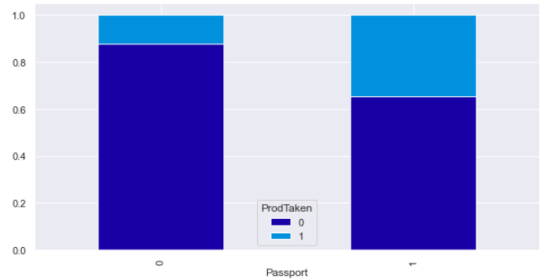


- Singles are much more likely to pick up a package.

EDA – ProdTaken with Passport, PitchSatisfactionScore & OwnCar

ProdTaken Vs Passport

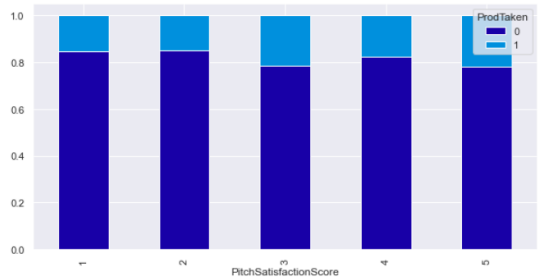
```
ProdTaken     0    1   All
Passport
0          3018  423  3441
1           928  494  1422
All        3946  917  4863
```



- Customers with passports are much more likely to pick up a package.

ProdTaken Vs PitchSatisfactionScore

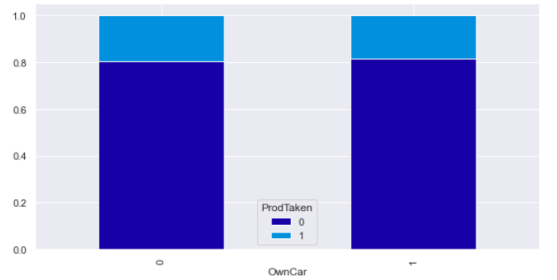
```
ProdTaken                0    1   All
PitchSatisfactionScore
1                      794  143   937
2                      495   88   583
3                     1152  314  1466
4                      748  162   910
5                      757  210   967
All                   3946  917  4863
```



- Customers who rated 3.0 or more for pitch satisfaction are more likely to pick up a package.

ProdTaken Vs OwnCar

```
ProdTaken     0    1   All
OwnCar
0          1486  359  1845
1          2460  558  3018
All        3946  917  4863
```

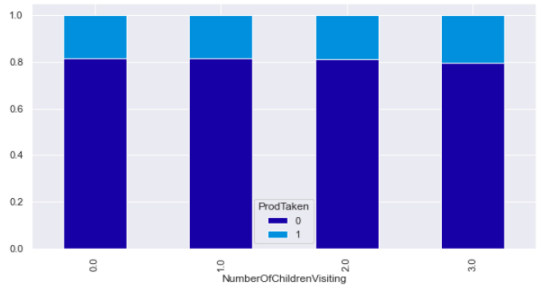


- There is no discernible difference in either owning or not owning a car to determine a higher take up of a package by a customer.

EDA – ProdTaken with NumberOfChildrenVisiting & Designation

ProdTaken Vs NumberOfChildrenVisiting

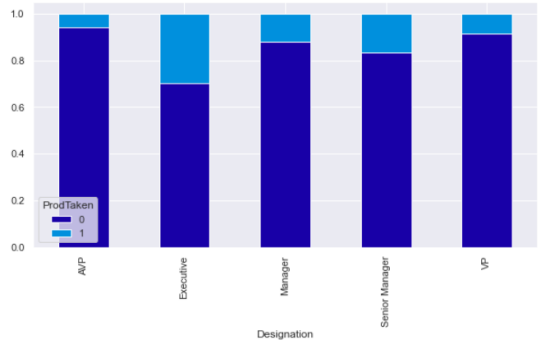
ProdTaken	0	1	All
NumberOfChildrenVisiting			
0.0	870	200	1070
1.0	1737	398	2135
2.0	1080	253	1333
3.0	259	66	325
All	3946	917	4863



- There is no discernible difference in number of children travelling with customer to determine a higher take up of a package by a customer.

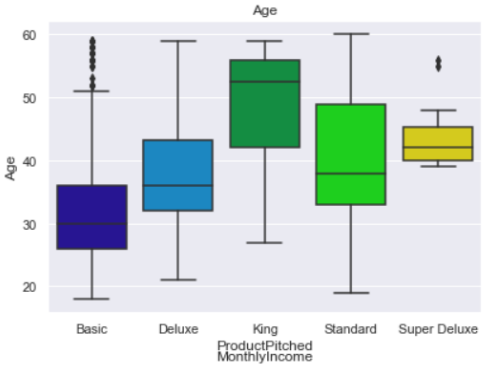
ProdTaken Vs Designation

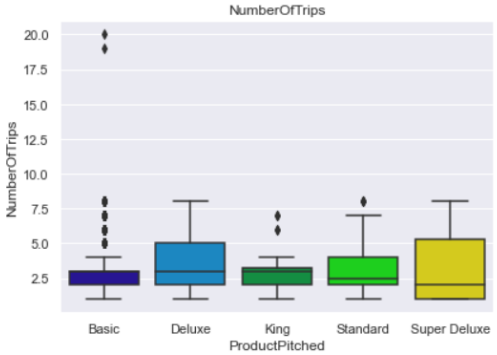
ProdTaken	0	1	All
Designation			
AVP	322	20	342
Executive	1283	549	1832
Manager	1513	204	1717
Senior Manager	618	124	742
VP	210	20	230
All	3946	917	4863

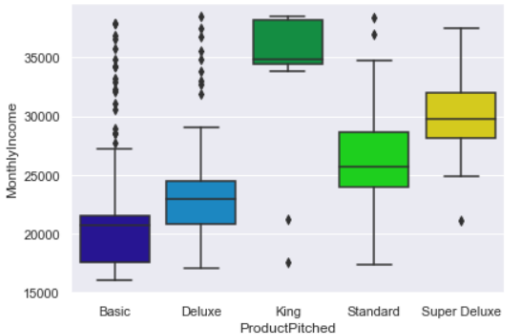


- Executives are much more likely to take up a package.

EDA – Build Customer Profile: ProductPitched with Numeric Values







Observations

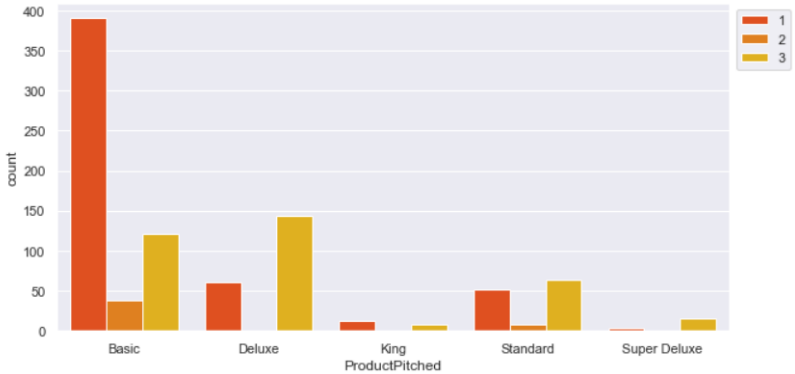
- Basic packages are mainly taken by the younger groups from late 20s to late 30s.
- Deluxe package takers mainly age range from early 30s to early 40s.
- Standard package takers mainly age range from mid 30s to late 40s.
- Super Deluxe package takers mainly age range from early 40s to mid 40s.
- King package takers mainly age range from early 40s to mid 50s.

- Basic package takers mainly travel between 2 to 3 times a year with several outliers.
- Deluxe package takers mainly travel between 2 to 5 times a year.
- King package takers mainly travel between 2 to 3 times a year with some outliers.
- Standard package takers mainly travel between 2 to 4 times a year with a outlier.
- Super Deluxe package takers travel between 1 to 5 times a year.

- Basic package takers income mostly range between 17.5K to ~ 21K.
- Deluxe package takers income mostly range between 21K to just shy of 25K.
- Standard package takers income mostly range between 24K to ~ 29K.
- Super Deluxe takers income mostly range from just below 28K to ~ 32K.
- King package takers income ranges from 35K and above.

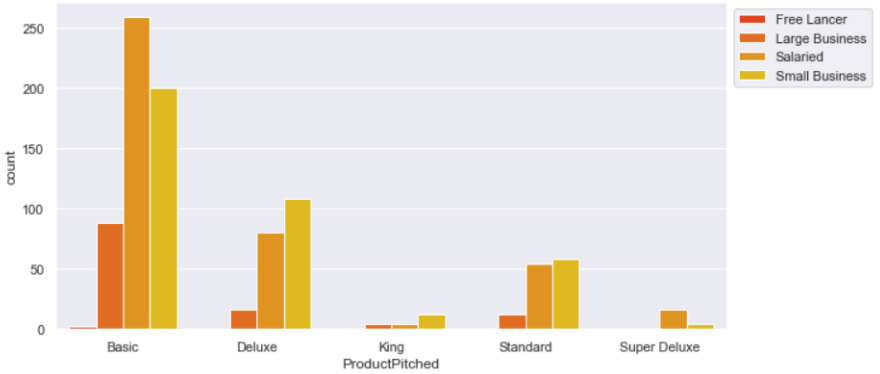
EDA – Build Customer Profile: ProductPitched with CityTier & Occupation

ProductPitched Vs CityTier



- Basic packages are most favored by Tier 1 city residents.
- Deluxe packages are most favored by Tier 3 followed by Tier 1 city residents.
- Standard to King packages are most favored by Tier 1 and 3 city residents.

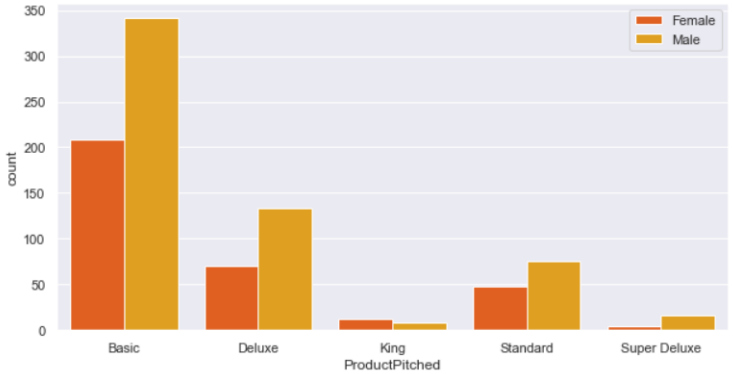
ProductPitched Vs Occupation



- Basic packages are most favored by salaried people then small business.
- Deluxe packages are most favored by small business then salaried people.
- Standard packages are most favored by both salaried people and small business almost equally.
- Super Deluxe packages are picked up most by salaried people.
- King packages are picked up most by small business.

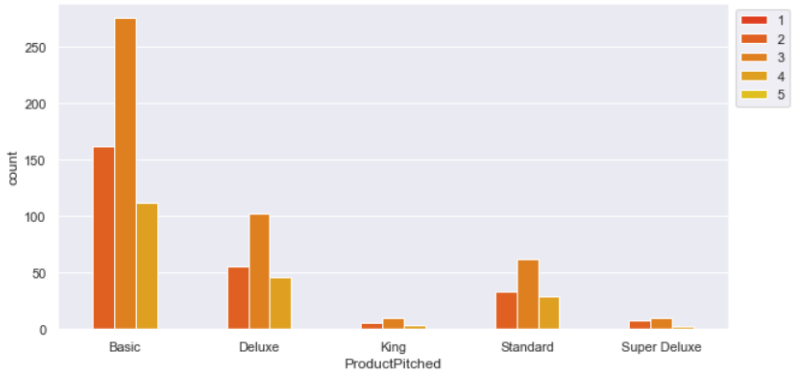
EDA – Build Customer Profile: ProductPitched with Gender & NumberOfPersonVisiting

ProductPitched Vs Gender



- Basic packages are more favored by males but remains the most popular among genders.
- Deluxe packages are more favored by males but remains 2nd most popular among genders.
- Standard packages are more favored by males.
- Super Deluxe packages are picked up more by males.
- King packages are picked slightly more by females.

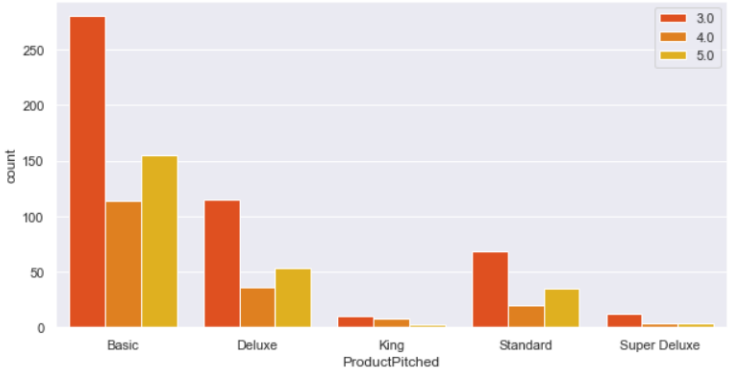
ProductPitched Vs NumberOfPersonVisiting



- Basic, Deluxe and Standard packages are picked up by customers with 2 to 4 travelling companions with 3 being the most popular.
- Super Deluxe and King packages mainly attract 2 or 3 companions groups of customers.

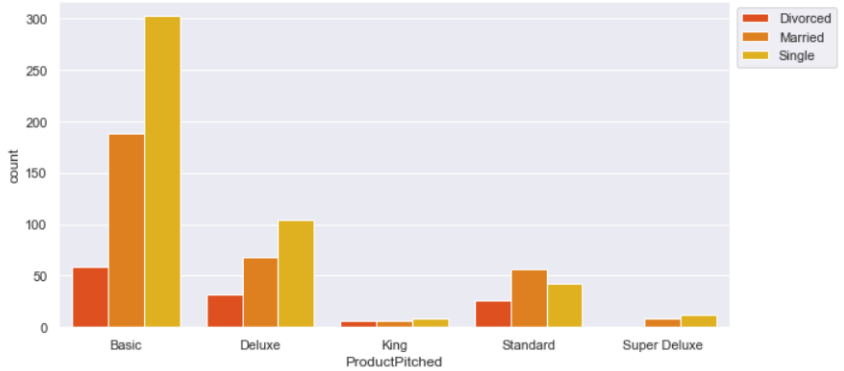
EDA – Build Customer Profile: ProductPitched with PreferredPropertyStar & MaritalStatus

ProductPitched Vs PreferredPropertyStar



- Across Basic, Deluxe and Standard packages, most customers who took up packages preferred 3 star hotels followed by 5 stars.
- In Super Deluxe and King packages, customers who took up packages tend to prefer 3-4 stars hotels.

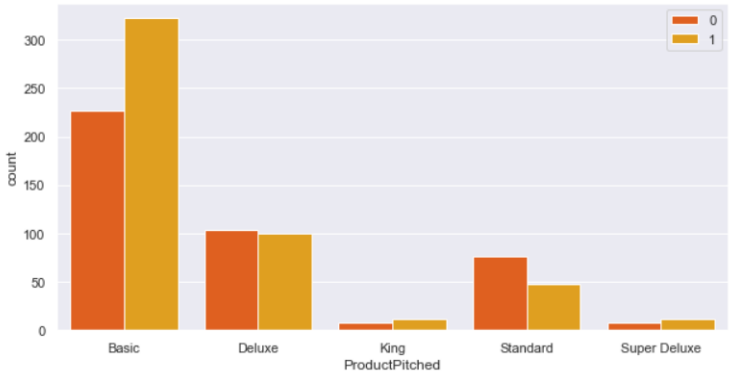
ProductPitched Vs MaritalStatus



- Basic, Deluxe packages appealed most to Single customers who took a package.
- Standard packages appealed most to married customers.
- Super Deluxe and King packages appealed to the same amount of each type of customer except no divorcees picked up a Super Deluxe package.

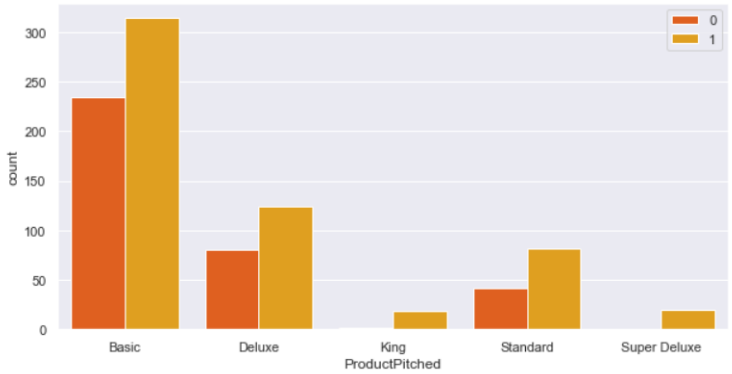
EDA – Build Customer Profile: ProductPitched with Passport & OwnCar

ProductPitched Vs Passport



- Basic packages are most likely picked up by a customer with a passport.
- Standard package is more picked up by those without passport.
- Deluxe, Super Deluxe and King customers have equal numbers having or not having a passport.

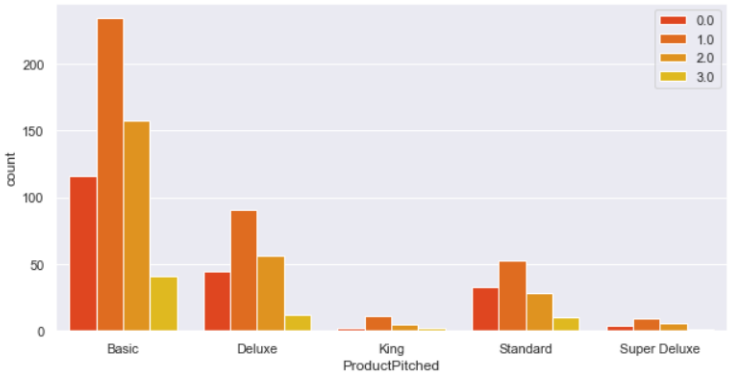
ProductPitched Vs OwnCar



- Those who own cars are more likely to pick up a package then those who don't, the disparity is more so for the Basic and King packages.

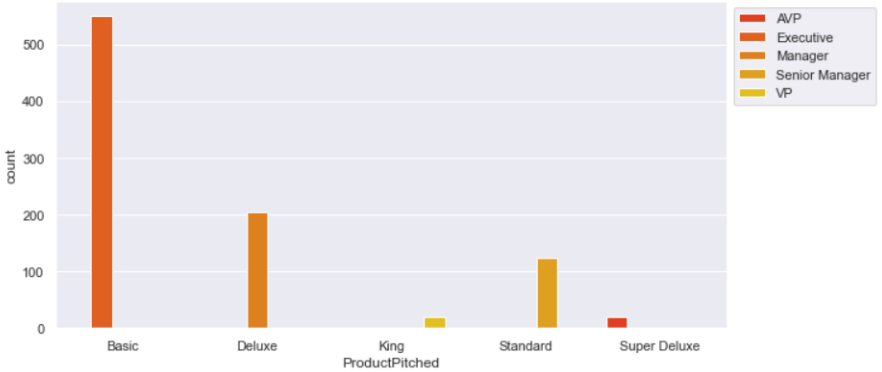
EDA – Build Customer Profile: ProductPitched with NumberOfChildrenVisiting & Designation

ProductPitched Vs NumberOfChildrenVisiting



- Basic, Deluxe and Super Deluxe packages are most likely picked up by customers who travels with 1-2 kids.
- Standard packages are most likely picked up by customers with 0-1 kids.
- King packages are most likely picked up by customers with 1 child.

ProductPitched Vs Designation



- Basic Package attracts Executive level customers.
- Deluxe package attracts Manager level customers.
- Standard package attracts Senior Manager level customers.
- Super Deluxe and King packages are picked up by AVPs and VPs respectively.

EDA – Build Customer Profile: Packages Insights

Basic Package Customer Profile

- Age ranges from late 20s to 30s.
- Mainly travel between 2 to 3 times a year.
- Monthly Income mostly range from 17.5K to ~ 21K.
- Executive level designation.
- Travels with 1-2 kids.
- Much more likely to own car.
- Most likely owns a passport.
- Most likely Single.
- Preferred 3 star hotels followed by 5 stars hotels.
- With 2 to 4 travelling companions.
- More favored by Males.
- Most favored by salaried people then small business.
- Most favored by Tier 1 city residents.

Deluxe Package Customer Profile

- Age range from early 30s to early 40s.
- Mainly travel between 2 to 5 times a year.
- Monthly income mostly range between 21K to just shy of 25K.
- Manager level designation.
- Travels with 1-2 kids.
- Likely to own car.
- Having either a passport or none.
- Most likely Single.
- Preferred 3 star hotels followed by 5 stars hotels.
- With 2 to 4 travelling companions.
- More favored by Males.
- Most favored by small business then salaried people.
- Most favored by Tier 3 followed by Tier 1 city residents.

Standard Package Customer Profile

- Age range from mid 30s to late 40s.
- Mainly travel between 2 to 4 times a year.
- Monthly income mostly range between 24K to ~ 29K.
- Senior Manager level designation.
- Travels with 0-1 kids.
- Likely to own car.
- More likely not owning a passport.
- More likely Married.
- Preferred 3 star hotels followed by 5 stars hotels.
- With 2 to 4 travelling companions.
- More favored by Males.
- Favored by both salaried people and small business almost equally.
- Most favored by Tier 1 and 3 city residents.

EDA – Build Customer Profile: Packages Insights

Super Deluxe Package Customer Profile

- Age range from early 40s to mid 40s.
- Travel between 1 to 5 times a year.
- Monthly income mostly range from just below 28K to ~ 32K.
- AVP level designation.
- Travels with 1-2 kids.
- Likely to own car.
- Having either a passport or none.
- Single or Married.
- Prefer 3-4 stars hotels.
- With 2 or 3 travelling companions.
- More favored by Males.
- Favored most by salaried people.
- Most favored by Tier 1 and 3 city residents.

King Package Customer Profile

- Age range from early 40s to mid 50s.
- Mainly travel between 2 to 3 times a year.
- Monthly income ranges from 35K and above.
- VP level designation.
- Travels with 1 kid.
- Likely to own car.
- Having either a passport or none.
- Single, Married or Divorced.
- Prefer 3-4 stars hotels.
- With 2 or 3 travelling companions.
- More favored by Females.
- Favored most by small business.
- Most favored by Tier 1 and 3 city residents.

EDA – Business Insights

- It has been shown that the package acceptance is largely among the customers ages 28 to 40. This has made our Basic, Standard and Deluxe packages successful based on marketing the right packages to our customer profiles.

- Tier 2 and Tier 3 city residents are more accepting of packages then Tier 1 city residents. However, Tier 2 residents only constituted of 4.07% of our customer pool. More expansion is needed among Tier 2 residents as well as Tier 3 customer base now at 30.68% only.

- Large Business were also found to be more accepting of packages but only constituted 8.9% of our customer base. More expansion is needed on Large business customers.

- Higher number of follow-ups by our sales teams to the potential customers lead to greater success of the customer accepting a package. This is so for 3 follow-ups and above with each successive follow-up leading to a greater acceptance level. Sales teams should be briefed to at least have 3 follow-ups with potential customers.

EDA – Business Insights

- Basic followed by Standard and Deluxe packages have the most success rates among our customers. The new Wellness package has to have similar features to them and marketed in 3 star or 5 star hotel combinations. This is because although most customers of Basic, Standard or Premium packages preferred 3 star, a number of them also picked 5 star over 4 star within these packages. 5 star preferred customers are also more accepting of a package.

- Singles are much more likely then married or divorcees to take up packages but our current customer base only has 32.8% of singles vs. 47.8% married. Perhaps a marketing campaign to attract singles to be our customers can be done.

- Passport holders are much more likely to accept a package than none passport holders but 70.8% of our current customer base do not have passports. Incentives can be launched to attract passport holders to be our customers or encourage our current customers to make a passport.

EDA – Business Insights

- Customer rating of Pitch satisfaction score of 3.0 and above are more likely to accept a package. However numbers show that ratings 1.0 is as high as ratings 4.0 as well as 5.0. This has to be further investigated on why the sales team gets such an unusual higher number of rating 1.0 to help improve sales conversion.

- Analysis has shown that least acceptance rate among packages is the Super Deluxe package which ties in with its customer profile of people designated as AVPs. AVPs number the least likely to accept a package and Super Deluxe was marketed at them. The package therefore needs a relook and if needed, it can be discontinued. The new Wellness package should not have features similar to the Super Deluxe package.

Data Preparation – Data Pre-Processing

- Data Pre-Processing
  - Missing Values
    - Median values are filled into missing value rows of numerical variables.
    - Rows are removed for missing value rows of categorical variables where impact is small (~ 0.5% of rows).
  - Duplicated Values
    - Duplicated rows are removed where impact is small (< 3% of rows).
  - Outliers
    - Outliers are capped at lower and upper whiskers of IQR.
    - Variables affected: ‘DurationofPitch’, ‘NumberOfTrips’,’ MonthlyIncome’
  - Other Formatting
    - Column value formatting is done for Gender (Fe Male to Female).
    - Column value formatting is done for MaritalStatus (Single and Unmarried merged to Single).

Data Preparation – Feature Engineering

- Data Pre-Processing
  - Age Variable Binning
    - Age has been put into age groups in Age_bin and converted to categorical variable.
    - Age Groups are: "15-25" , "25-35" , "35-45" , "45-55" , "55-65"
  - MonthlyIncome Variable Binning
    - MonthlyIncome has been put into income groups in Income_bin and converted to categorical variable.
    - Income groups are: "10k-20k" , "20k-30k" , "30k-40k"

Data Preparation – Train & Test Sets Split

- The data set is split into 70% for training and 30% for testing
- Dummy variables were prepared for categorical variables
- Stratifying is set to Y in order to ensure dependent variable train and test set has the same ratio of 1’s and 0’s as the original dataset

Data Preparation – Train & Test Sets Split

- The list of variables/features used for all the models are as below:
  - 'DurationOfPitch', 'NumberOfTrips', 'TypeofContact_Self Enquiry', 'CityTier_2', 'CityTier_3', 'Occupation_Large Business', 'Occupation_Salaried', 'Occupation_Small Business', 'Gender_Male', 'NumberOfPersonVisiting_2', 'NumberOfPersonVisiting_3', 'NumberOfPersonVisiting_4', 'NumberOfPersonVisiting_5', 'NumberOfFollowups_2.0', 'NumberOfFollowups_3.0', 'NumberOfFollowups_4.0', 'NumberOfFollowups_5.0', 'NumberOfFollowups_6.0', 'ProductPitched_Deluxe', 'ProductPitched_King', 'ProductPitched_Standard', 'ProductPitched_Super Deluxe', 'PreferredPropertyStar_4.0', 'PreferredPropertyStar_5.0', 'MaritalStatus_Married', 'MaritalStatus_Single', 'Passport_1', 'PitchSatisfactionScore_2', 'PitchSatisfactionScore_3', 'PitchSatisfactionScore_4', 'PitchSatisfactionScore_5', 'OwnCar_1', 'NumberOfChildrenVisiting_1.0', 'NumberOfChildrenVisiting_2.0', 'NumberOfChildrenVisiting_3.0', 'Designation_Executive', 'Designation_Manager', 'Designation_Senior Manager', 'Designation_VP', 'Age_bin_25-35', 'Age_bin_35-45', 'Age_bin_45-55', 'Age_bin_55-65', 'Income_bin_20k-30k', 'Income_bin_30k-40k'

Model Performance Summary – Evaluation Criterion

- Model Evaluation Criterion
  - Model can make wrong predictions as:
    - False Positive: Predicting a customer is a travel package convertable but actually not convertable.
    - False Negative: Predicting a customer is a travel package non-convertable but actually convertable.
- Which case is more important?
  - Both the cases are important as:
  - If we predict a customer is a travel package convertable but actually not convertable then a wrong person will be getting the targeted marketing effort wasting resources.
  - If we predict a customer is a travel package non-convertable but actually convertable, that person will not be able to receive targeted marketing effort and hence may not be aware of the travel package and thus a loss of business.

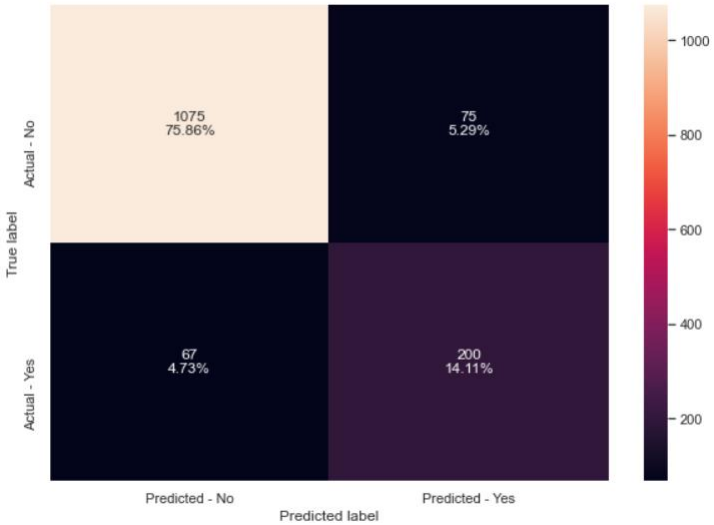
Model Performance Summary – Evaluation Criterion

- How to reduce losses?
  - We can use accuracy but since the data is imbalanced it would not be the right metric to check the model performance.
  - Therefore, f1_score should be maximized, the greater the f1_score higher the chances of identifying both the classes correctly.

- **Note: Class Weight {0:0.19,1:0.81} is applied to untuned and tuned models as a default.**

Model Performance Summary – Decision Tree Model

Test data Confusion Matrix




```
Accuracy on training set :  1.0
Accuracy on test set :  0.899788285109386
Recall on training set :  1.0
Recall on test set :  0.7490636704119851
Precision on training set :  1.0
Precision on test set :  0.7272727272727273
f1 score on training set :  1.0
f1 score on test set :  0.7380073800738007
```

Build Decision Tree Model

- We will build our model using the DecisionTreeClassifier function. Using default 'gini' criteria to split.
- If the frequency of class A is 10% and the frequency of class B is 90%, then class B will become the dominant class and the decision tree will become biased toward the dominant classes.
- In this case, we can pass a dictionary {0:0.19,1:0.81} to the model to specify the weight of each class and the decision tree will give more weightage to class 1 based on the dataset distribution.
- class_weight is a hyperparameter for the decision tree classifier and will be applied to all models here tuned or untuned as a default.

Observations

- Decision tree is working well on the training data but is not able to generalize well on the test data.
- This is so as well for the f1 score (train data: 100% ; test data: 73.8%).
- In fact this is a sign of overfitting.

Model Performance Summary – Decision Tree Model Tuned

Test data Confusion Matrix


```
Accuracy on training set :  0.7963691376701967
Accuracy on test set :  0.7692307692307693
Recall on training set :  0.7115384615384616
Recall on test set :  0.6142322097378277
Precision on training set :  0.47385272145144075
Precision on test set :  0.422680412371134
f1 score on training set :  0.5688661114670083
f1 score on test set :  0.5007633587786259
```

Hyperparameters Applied:

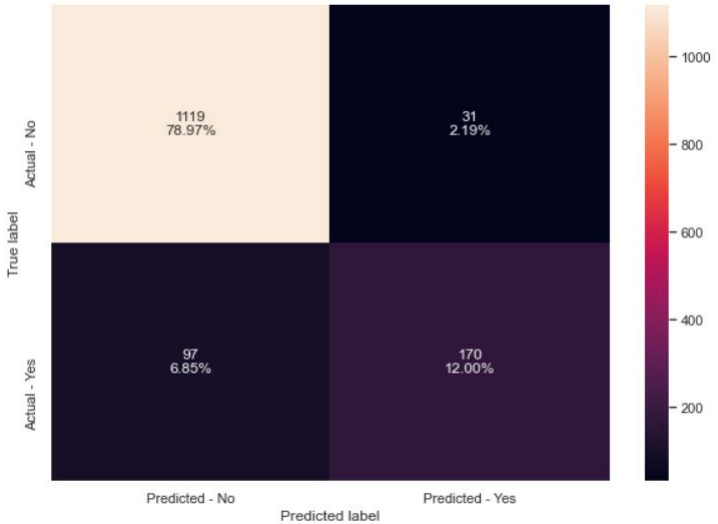
{'max_depth': np.arange(2,30),
'min_samples_leaf': [1, 2, 5, 7, 10],
'max_leaf_nodes' : [2, 3, 5, 10,15],
'min_impurity_decrease': [0.0001,0.001,0.01,0.1]}

Observations

- Overfitting in decision tree tuned has reduced but scores have also reduced.
- This is so as well for the f1 score (train data: 56.9% ; test data: 50.1%).

Model Performance Summary – Bagging Classifier Model

Test data Confusion Matrix




```
Accuracy on training set :  0.9957639939485627
Accuracy on test set :  0.9096683133380381
Recall on training set :  0.9775641025641025
Recall on test set :  0.6367041198501873
Precision on training set :  1.0
Precision on test set :  0.845771144278607
f1 score on training set :  0.9886547811993517
f1 score on test set :  0.7264957264957265
```

Observations

- Bagging classifier is still overfitting on the training set and is not generalizing well on the test data.
- This is so as well for the f1 score (train data: 98.9% ; test data: 72.6%).
- It is a slight improvement over the initial lone decision tree where the scores gap between train and test data is smaller.

Model Performance Summary – Bagging Classifier Model Tuned

Test data Confusion Matrix


```
Accuracy on training set :  0.9993948562783661
Accuracy on test set :  0.9146083274523642
Recall on training set :  0.9967948717948718
Recall on test set :  0.6179775280898876
Precision on training set :  1.0
Precision on test set :  0.8967391304347826
f1 score on training set :  0.9983948635634029
f1 score on test set :  0.7317073170731706
```

Hyperparameters Applied:

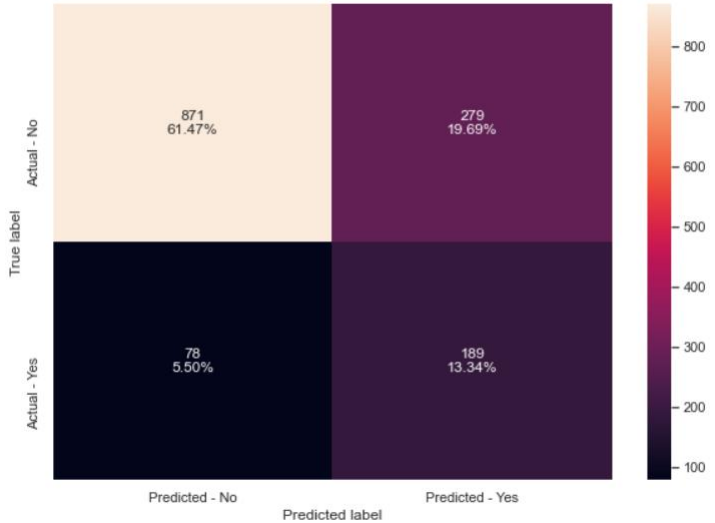
{'max_samples': [0.7,0.8,0.9,1],
'max_features': [0.7,0.8,0.9,1],
'n_estimators' : [10,20,30,40,50],}

Observations

- Bagging classifier tuned is still overfitting but less so on the training set and is generalizing a little better on the test data for precision.
- f1 score's score gap between training and test set remains (train data: 99.8% ; test data: 73.2%).
- There is no improvement on other scores but precision score gap between training and test set is smaller by 5% points.

Model Performance Summary – Bagging Classifier Log

Test data Confusion Matrix



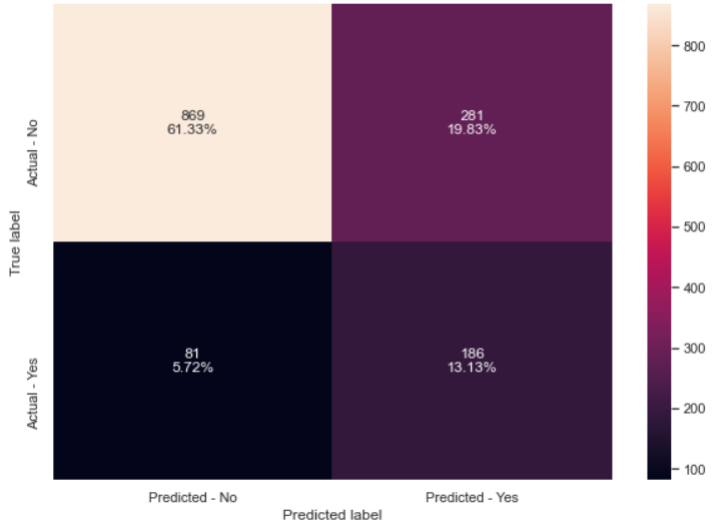

```
Accuracy on training set :  0.7655068078668684
Accuracy on test set :  0.748059280169372
Recall on training set :  0.7419871794871795
Recall on test set :  0.7078651685393258
Precision on training set :  0.4298978644382544
Precision on test set :  0.40384615384615385
f1 score on training set :  0.5443856554967665
f1 score on test set :  0.5142857142857142
```

Observations

- Bagging classifier with logistic regression as base_estimator is not overfitting the data but the scores are low.
- This is so as well for the f1 score (train data: 54.4% ; test data: 51.4%).
- Ensemble models are less interpretable than decision tree but bagging classifier is even less interpretable than random forest. It does not even have a feature importance attribute.

Model Performance Summary – Bagging Classifier Log Tuned

Test data Confusion Matrix




```
Accuracy on training set :  0.7679273827534039
Accuracy on test set :  0.744530698659139
Recall on training set :  0.7435897435897436
Recall on test set :  0.6966292134831461
Precision on training set :  0.4332399626517274
Precision on test set :  0.39828693790149894
f1 score on training set :  0.5474926253687316
f1 score on test set :  0.5068119891008175
```

Hyperparameters Applied:

{'max_samples': [0.7,0.8,0.9,1],
'max_features': [0.7,0.8,0.9,1],
'n_estimators' : [10,20,30,40,50],}

Observations

- Bagging classifier tuned with logistic regression as base_estimator is not overfitting the data but the scores performed similarly low to before tuning.
- This is so as well for the f1 score (train data: 54.7% ; test data: 50.7%).

Model Performance Summary – Random Forest Model

Test data Confusion Matrix


```
Accuracy on training set :  1.0
Accuracy on test set :  0.8892025405786874
Recall on training set :  1.0
Recall on test set :  0.449438202247191
Precision on training set :  1.0
Precision on test set :  0.9230769230769231
f1 score on training set :  1.0
f1 score on test set :  0.6045340050377833
```

Observations

- Random Forest classifier is still overfitting on the training set and is not generalizing well on the test data.
- This is so as well for the f1 score (train data: 100% ; test data: 60.5%).
- It does not do as well as the lone decision tree or the bagging classifier.

Model Performance Summary – Random Forest Model Tuned

Test data Confusion Matrix


```
Accuracy on training set :  0.9282904689863842
Accuracy on test set :  0.8623853211009175
Recall on training set :  0.875
Recall on test set :  0.599250936329588
Precision on training set :  0.774468085106383
Precision on test set :  0.6451612903225806
f1 score on training set :  0.8216704288939052
f1 score on test set :  0.6213592233009709
```

Hyperparameters Applied:

{ "n_estimators": [50,100,150],
"min_samples_leaf": np.arange(5, 10),
"max_features": np.arange(0.2, 0.7, 0.1),
"max_samples": np.arange(0.3, 0.7, 0.1),}

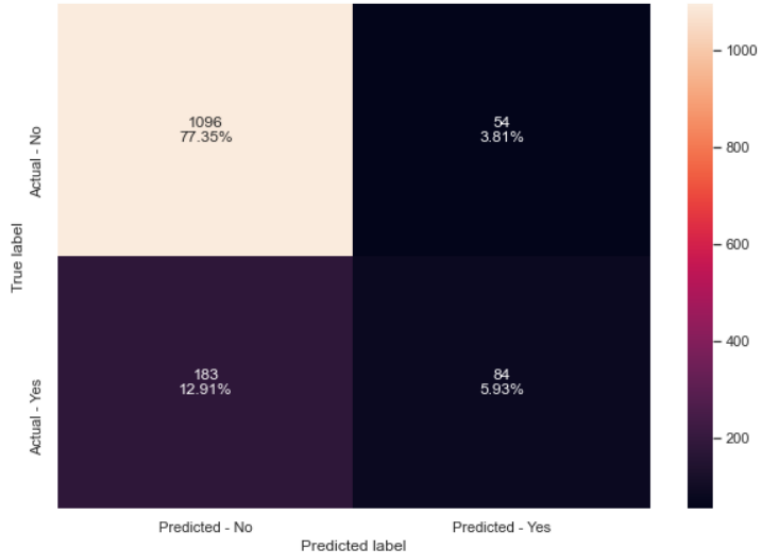
Observations

- Random Forest classifier tuned has improved much of the overfitting issue on the training set and is generalizing better on the test data.
- This is so as well for the f1 score (train data: 82.2% ; test data: 62.1%), an improvement of 20% points in the score gap but the scores are low.

Model Performance Summary – AdaBoost Classifier Model

Test data Confusion Matrix

```
Accuracy on training set :  0.848411497730711
Accuracy on test set :  0.8327452364149612
Recall on training set :  0.36538461538461536
Recall on test set :  0.3146067415730337
Precision on training set :  0.6846846846846847
Precision on test set :  0.6086956521739131
f1 score on training set :  0.47648902821316613
f1 score on test set :  0.4148148148148148
```



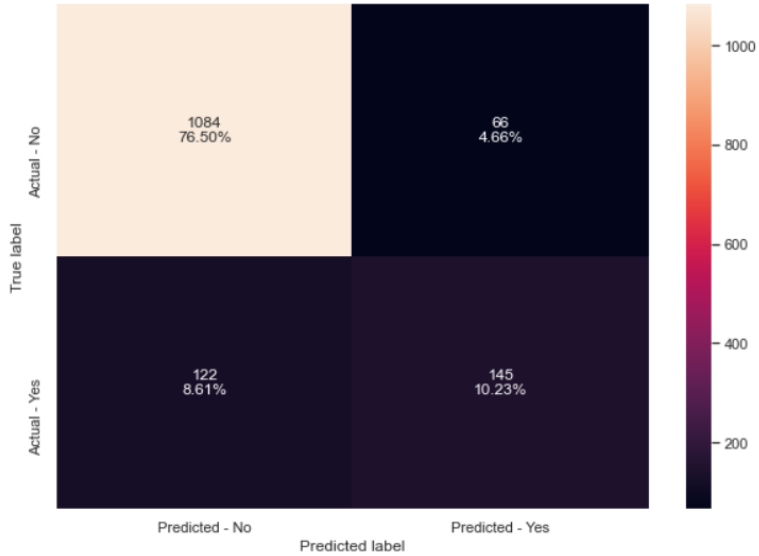
Observations

- Adaboost is giving more generalized performance than previous models but the test f1-score is too low.
- F1 score is at (train data: 47.6% ; test data: 41.5%).

Model Performance Summary – AdaBoost Classifier Model Tuned

Test data Confusion Matrix

```
Accuracy on training set :  0.9721633888048411
Accuracy on test set :  0.8673253352152435
Recall on training set :  0.8894230769230769
Recall on test set :  0.5430711610486891
Precision on training set :  0.9602076124567474
Precision on test set :  0.6872037914691943
f1 score on training set :  0.9234608985024958
f1 score on test set :  0.6066945606694562
```



Hyperparameters Applied:

{"base_estimator":[DecisionTreeClassifier(max_depth=1),
DecisionTreeClassifier(max_depth=2),
DecisionTreeClassifier(max_depth=3)],
"n_estimators": np.arange(10,110,10),
"learning_rate":np.arange(0.1,2,0.1)}

Observations

- The tuned AdaBoost model performance has increased but the model has started to overfit the training data among several metrics.
- F1 score is at (train data: 92.3% ; test data: 60.7%).

Model Performance Summary – Gradient Boosting Classifier Model

Test data Confusion Matrix

```
Accuracy on training set :  0.8886535552193646
Accuracy on test set :  0.8567395906845449
Recall on training set :  0.5
Recall on test set :  0.38202247191011235
Precision on training set :  0.8478260869565217
Precision on test set :  0.7285714285714285
f1 score on training set :  0.6290322580645161
f1 score on test set :  0.5012285012285013
```

Observations

- The Gradient Boosting model does not overfit as much but some of the test data metrics are low.
- F1 score is at (train data: 62.9% ; test data: 50.1%).

Model Performance Summary – Gradient Boosting Classifier Model Tuned

Test data Confusion Matrix

```
Accuracy on training set :  0.9234493192133132
Accuracy on test set :  0.8729710656316161
Recall on training set :  0.655448717948718
Recall on test set :  0.47191011235955055
Precision on training set :  0.9149888143176734
Precision on test set :  0.7636363636363637
f1 score on training set :  0.7637721755368814
f1 score on test set :  0.5833333333333333
```

Hyperparameters Applied:

{"n_estimators": [100,150,200,250],
"subsample":[0.8,0.9,1],
"max_features":[0.7,0.8,0.9,1]}

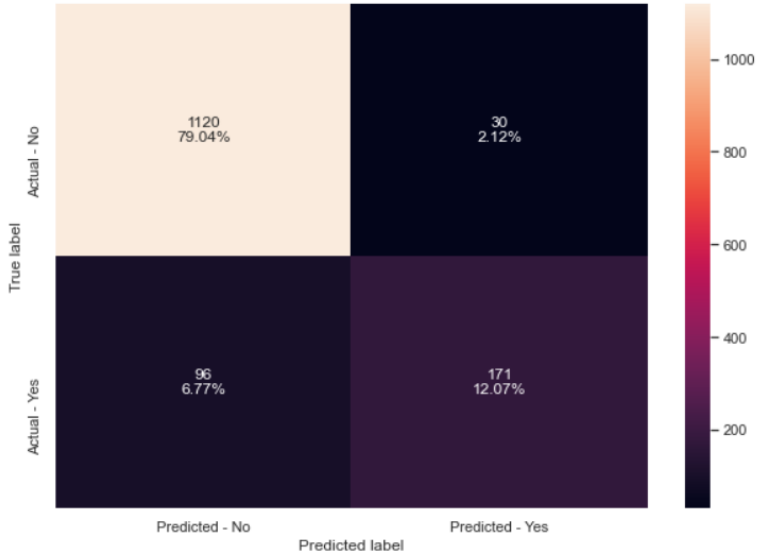
Observations

- The tuned Gradient Boosting model does not overfit yet and all test data metrics improved but there is more room for improvement.
- F1 score is at (train data: 76.4% ; test data: 58.3%).

Model Performance Summary – XGBoost Classifier Model

Test data Confusion Matrix

```
Accuracy on training set :  0.9990922844175492
Accuracy on test set :  0.9110797459421313
Recall on training set :  0.9951923076923077
Recall on test set :  0.6404494382022472
Precision on training set :  1.0
Precision on test set :  0.8507462686567164
f1 score on training set :  0.9975903614457832
f1 score on test set :  0.7307692307692307
```



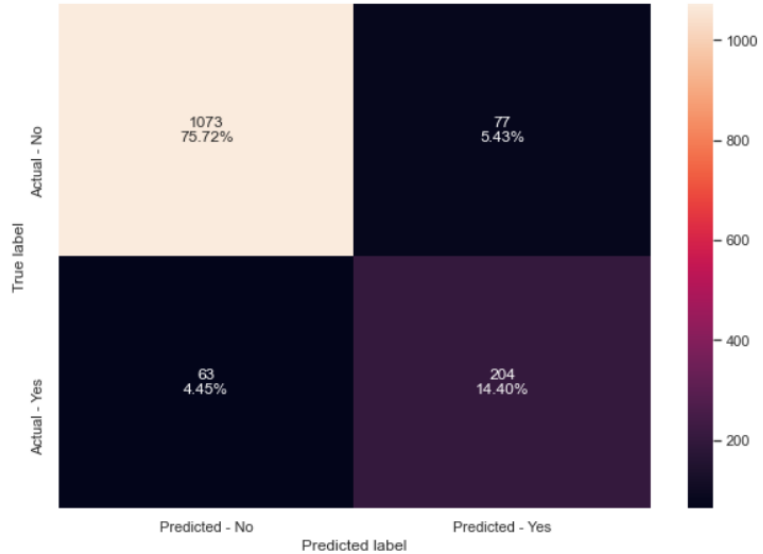
Observations

- The XGB model is starting to overfit and all test data metrics have improved significantly.
- F1 score is at (train data: 99.8% ; test data: 73.1%).

Model Performance Summary – XGBoost Classifier Model Tuned

Test data Confusion Matrix

```
Accuracy on training set :  0.9712556732223904
Accuracy on test set :  0.9011997177134792
Recall on training set :  0.9951923076923077
Recall on test set :  0.7640449438202247
Precision on training set :  0.8709677419354839
Precision on test set :  0.7259786476868327
f1 score on training set :  0.9289454001495887
f1 score on test set :  0.7445255474452555
```



Hyperparameters Applied:

{"n_estimators": [10,30,50],
"scale_pos_weight":[1,2,5],
"subsample":[0.7,0.9,1],
"learning_rate":[0.05, 0.1,0.2],
"colsample_bytree":[0.7,0.9,1],
"colsample_bylevel":[0.5,0.7,1]}

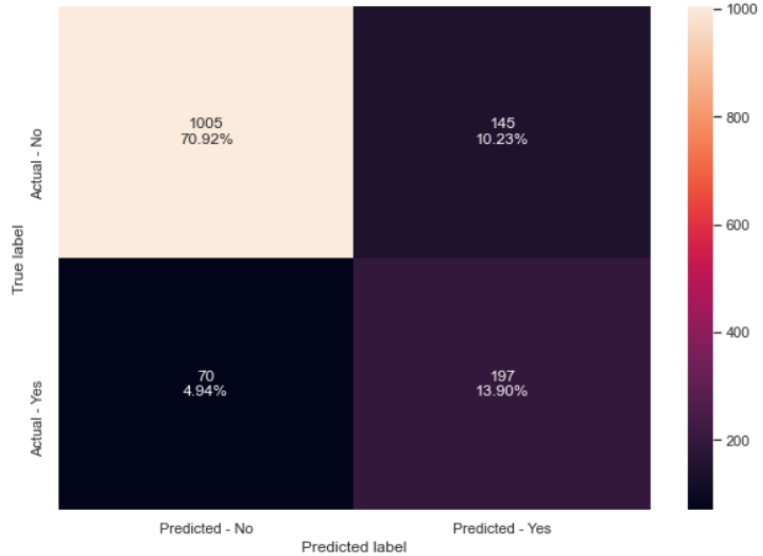
Observations

- The tuned XGB model has generalized the test data better and keeping the improved metrics scores.
- F1 score is at (train data: 92.9% ; test data: 74.5%).

Model Performance Summary – Stacking Classifier Model

Test data Confusion Matrix

```
Accuracy on training set :  0.9288956127080181
Accuracy on test set :  0.8482709950599859
Recall on training set :  0.9423076923076923
Recall on test set :  0.7378277153558053
Precision on training set :  0.747141041931385
Precision on test set :  0.5760233918128655
f1 score on training set :  0.8334514528703048
f1 score on test set :  0.6469622331691297
```



Estimators:

- Estimator 1 = Random Forest Model Tuned
- Estimator 2 = Gradient Boosting Model Tuned
- Estimator 3 = Decision Tree Model Tuned
- Final Estimator = XGBoost Model Tuned

Observations

- The stacker classifier has generalized the test data similar to XGB tuned model but scored lower on the metrics.
- F1 score is at (train data: 83.3% ; test data: 64.7%).

Model Performance Summary – Performance Metrics

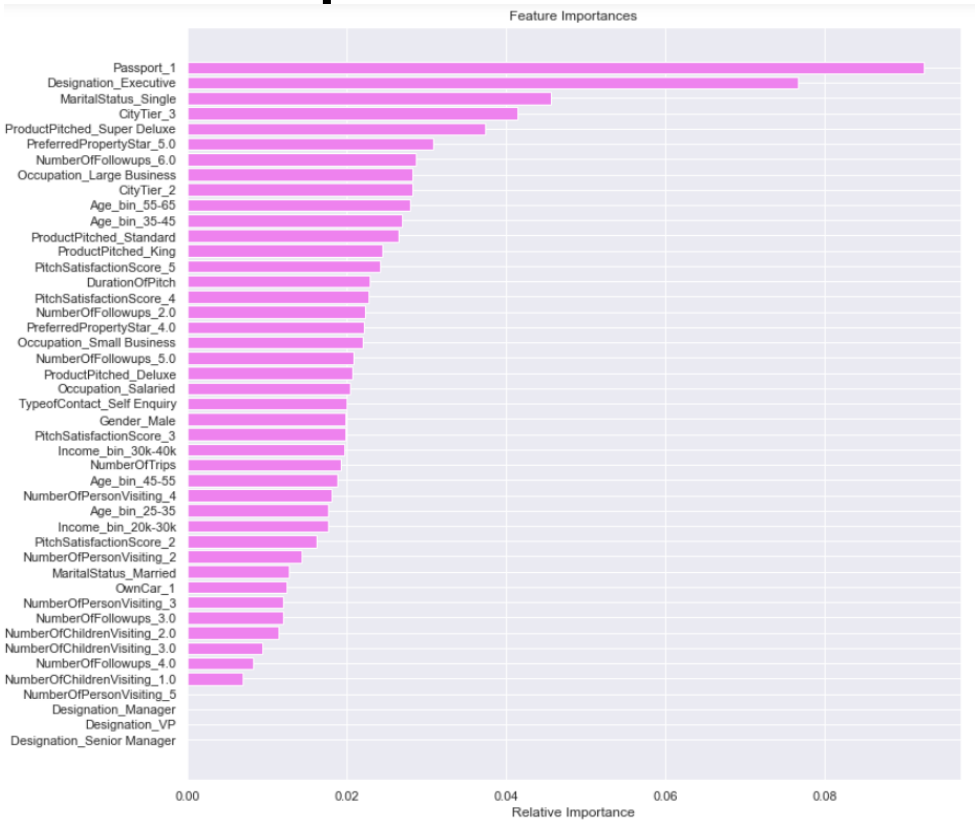
	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_F1-Score	Test_F1-Score
13	Tuned XGBoost Classifier	0.971256	0.901200	0.995192	0.764045	0.870968	0.725979	0.928945	0.744526
0	Decision Tree	1.000000	0.899788	1.000000	0.749064	1.000000	0.727273	1.000000	0.738007
5	Tuned Bagging Classifier	0.999395	0.914608	0.996795	0.617978	1.000000	0.896739	0.998395	0.731707
12	XGBoost Classifier	0.999092	0.911080	0.995192	0.640449	1.000000	0.850746	0.997590	0.730769
4	Bagging Classifier	0.995764	0.909668	0.977564	0.636704	1.000000	0.845771	0.988655	0.726496
14	Stacking Classifier	0.928896	0.848271	0.942308	0.737828	0.747141	0.576023	0.833451	0.646962
3	Tuned Random Forest	0.928290	0.862385	0.875000	0.599251	0.774468	0.645161	0.821670	0.621359
9	Tuned AdaBoost Classifier	0.972163	0.867325	0.889423	0.543071	0.960208	0.687204	0.923461	0.606695
2	Random Forest	1.000000	0.889203	1.000000	0.449438	1.000000	0.923077	1.000000	0.604534
11	Tuned Gradient Boosting Classifier	0.923449	0.872971	0.655449	0.471910	0.914989	0.763636	0.763772	0.583333
6	Bagging Classifier Logistic Regression	0.765507	0.748059	0.741987	0.707865	0.429898	0.403846	0.544386	0.514286
7	Tuned Bagging Classifier Logistic Regression	0.767927	0.744531	0.743590	0.696629	0.433240	0.398287	0.547493	0.506812
10	Gradient Boosting Classifier	0.888654	0.856740	0.500000	0.382022	0.847826	0.728571	0.629032	0.501229
1	Tuned Decision Tree	0.796369	0.769231	0.711538	0.614232	0.473853	0.422680	0.568866	0.500763
8	AdaBoost Classifier	0.848411	0.832745	0.365385	0.314607	0.684685	0.608696	0.476489	0.414815

Observations

- The models are either tending towards overfitting or score poorly in terms of f1-score.
- Tuned XGBoost Classifier should be the model to press ahead as it scored the highest test f1-score.
- Tuned XGBoost Classifier has also one of the lowest gap between training and test f1-score of the dataset among all models.
- There may well be other combinations of hyperparameters not attempted yet to improve the metrics of the models. This will however require much more time to find out more comprehensively a better model.

Model Performance Summary – XGBoost Classifier Model Tuned

Feature Importance



Observations

- In the tuned XGBoost model, Passport(Yes) is the most important feature followed by features - Designation(Executive) and Marital Status(Single).

Business Insights and Recommendations

- Based on the performances of the different models, Tuned XGBoost Classifier performed the best using f1_score as the deciding factor due to the unbalanced data as well as a lower score gap between training and testing data
  - Significant variables include ‘Passport’, ‘Designation’ and ‘MaritalStatus’.
  - Coupled with EDA insights, Passport holders, Executive level and Single customers are more likely to accept a package deal.
  - However, the current customer base has 70.8% Non-Passport holders and only 32.8% of customers are single compared to 47.8% are married.
  - Therefore more targeted marketing effort at attracting Passport holders and Singles to expand the customer base with these customer segments. Incentives can also be given to encourage existing customer base to obtain a passport or attract new passport holding customers.
  - Currently, the Executive level customers dominated the share of customers at 37.7% and more can be done to recruit them and bolster the Executive level numbers.

Business Insights and Recommendations

- Comments on additional data sources for model improvement
  - Additional data can be obtained from measured feedback of initial targeted marketing efforts to the public in order to strengthen the model.
  - Feedback can be gathered from non-package convertible customers for further analysis.

- Model implementation in real world and potential business benefits from model
  - The model implemented in the real world will help to raise more successful targeted marketing converts of its campaign and reduce the costs of marketing to potential non-converts or miss target marketing to potential converts. This will increase revenue and reduce both variable marketing costs and opportunity costs.

Business Insights and Recommendations

- Other Recommendations – From EDA
  - More expansion is needed among Tier 2 residents as well as Tier 3 customer base now at 30.68% only.
  - More expansion is needed on Large business customers now at 8.9% of our customer base.
  - Sales teams should be briefed to at least have 3 follow-ups with potential customers.
  - The new Wellness package has to have similar features to Basic, Standard and Deluxe packages and marketed in 3 star or 5 star hotel combinations.
  - Further investigation on unusual high number of rating 1.0 to help improve sales conversion.
  - Super Deluxe package performs poorly and it should require a relook at its features and discontinued it needed. The new Wellness package should not have similar features to Super Deluxe package.

greatlearning
Power Ahead

Happy Learning !