

```
In [1]: #importing necessary libraries
import warnings
warnings.filterwarnings("ignore")

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn import metrics
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import BaggingClassifier, GradientBoostingClassifier, A
from xgboost import XGBClassifier
```

```
In [2]: #Loading dataset
data = pd.read_excel('Tourism.xlsx', sheet_name = 'Tourism' )
```

Overview of dataset

viewing first 10 rows of dataset

```
In [3]: data.head(10)
```

```
Out[3]:
```

	CustomerID	ProdTaken	Age	PreferredLoginDevice	CityTier	DurationOfPitch	Occupation
0	200000	1	41.0	Self Enquiry	3	6.0	Salaried
1	200001	0	49.0	Company Invited	1	14.0	Salaried
2	200002	1	37.0	Self Enquiry	1	8.0	Free Lancer
3	200003	0	33.0	Company Invited	1	9.0	Salaried
4	200004	0	NaN	Self Enquiry	1	8.0	Small Business
5	200005	0	32.0	Company Invited	1	8.0	Salaried
6	200006	0	59.0	Self Enquiry	1	9.0	Small Business
7	200007	0	30.0	Self Enquiry	1	30.0	Salaried
8	200008	0	38.0	Company Invited	1	29.0	Salaried
9	200009	0	36.0	Self Enquiry	1	33.0	Small Business

checking data types and if there is any null values under each feature

```
In [4]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4888 entries, 0 to 4887
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   CustomerID                           4888 non-null   int64
1   ProdTaken                            4888 non-null   int64
2   Age                                  4662 non-null   float64
3   PreferredLoginDevice                 4863 non-null   object
4   CityTier                             4888 non-null   int64
5   DurationOfPitch                      4637 non-null   float64
```

```

6  Occupation          4888 non-null object
7  Gender              4888 non-null object
8  NumberOfPersonVisited 4888 non-null int64
9  NumberOfFollowups    4843 non-null float64
10 ProductPitched      4888 non-null object
11 PreferredPropertyStar 4862 non-null float64
12 MaritalStatus       4888 non-null object
13 NumberOfTrips       4748 non-null float64
14 Passport            4888 non-null int64
15 PitchSatisfactionScore 4888 non-null int64
16 OwnCar              4888 non-null int64
17 NumberOfChildrenVisited 4822 non-null float64
18 Designation         4888 non-null object
19 MonthlyIncome       4655 non-null float64
dtypes: float64(7), int64(7), object(6)
memory usage: 763.9+ KB

```

- looking at the data information suggests that null values need to be imputed and 'object' data types need to be converted to category so that our models would be fed with proper data

```
In [5]: data.nunique()
```

```

Out[5]: CustomerID          4888
ProdTaken                2
Age                     44
PreferredLoginDevice      2
CityTier                 3
DurationOfPitch          34
Occupation               4
Gender                   3
NumberOfPersonVisited     5
NumberOfFollowups         6
ProductPitched            5
PreferredPropertyStar      3
MaritalStatus             4
NumberOfTrips            12
Passport                  2
PitchSatisfactionScore     5
OwnCar                    2
NumberOfChildrenVisited    4
Designation                5
MonthlyIncome            2475
dtype: int64

```

- Object data types are expected to have fewer unique values, but 'Gender' has 3 and needs to be fixed

```
In [170]: data.isna().sum() #checking number of null values in each feature
```

```

Out[170]: CustomerID          0
ProdTaken                0
Age                     226
PreferredLoginDevice      25
CityTier                 0
DurationOfPitch          251
Occupation               0
Gender                   0
NumberOfPersonVisited     0
NumberOfFollowups         45
ProductPitched            0
PreferredPropertyStar      26
MaritalStatus             0
NumberOfTrips            140
Passport                  0
PitchSatisfactionScore     0

```

```
OwnCar          0
NumberOfChildrenVisited 66
Designation     0
MonthlyIncome   233
dtype: int64
```

Summary of dataset

In [7]: `data.describe().T`

Out[7]:

		count	mean	std	min	25%	50%
	CustomerID	4888.0	202443.500000	1411.188388	200000.0	201221.75	202443.5
	ProdTaken	4888.0	0.188216	0.390925	0.0	0.00	0.0
	Age	4662.0	37.622265	9.316387	18.0	31.00	36.0
	CityTier	4888.0	1.654255	0.916583	1.0	1.00	1.0
	DurationOfPitch	4637.0	15.490835	8.519643	5.0	9.00	13.0
	NumberOfPersonVisited	4888.0	2.905074	0.724891	1.0	2.00	3.0
	NumberOfFollowups	4843.0	3.708445	1.002509	1.0	3.00	4.0
	PreferredPropertyStar	4862.0	3.581037	0.798009	3.0	3.00	3.0
	NumberOfTrips	4748.0	3.236521	1.849019	1.0	2.00	3.0
	Passport	4888.0	0.290917	0.454232	0.0	0.00	0.0
	PitchSatisfactionScore	4888.0	3.078151	1.365792	1.0	2.00	3.0
	OwnCar	4888.0	0.620295	0.485363	0.0	0.00	1.0
	NumberOfChildrenVisited	4822.0	1.187267	0.857861	0.0	1.00	1.0
	MonthlyIncome	4655.0	23619.853491	5380.698361	1000.0	20346.00	22347.0

- Duration Of Pitch, and Number of Trips seem to have outliers
- Monthly Income looks rightly skewed
- Mean value for Age is 37, Mean for Duration Of Pitch is 15

Data Cleaning:

In [8]: `tourism = data.copy() #creating a copy of our data for precautions`

In [9]: `#converting 'object' datatypes to categoricals`
`for feature in tourism.columns:`
 `if tourism[feature].dtype == 'object':`
 `tourism[feature] = pd.Categorical(tourism[feature])`
`tourism.head()`

Out[9]:

	CustomerID	ProdTaken	Age	PreferredLoginDevice	CityTier	DurationOfPitch	Occupation
0	200000	1	41.0	Self Enquiry	3	6.0	Salaried
1	200001	0	49.0	Company Invited	1	14.0	Salaried
2	200002	1	37.0	Self Enquiry	1	8.0	Free Lancer
3	200003	0	33.0	Company Invited	1	9.0	Salaried
4	200004	0	NaN	Self Enquiry	1	8.0	Small Business

```
In [10]: tourism['PreferredLoginDevice'].value_counts()
```

```
Out[10]: Self Enquiry      3444
Company Invited    1419
Name: PreferredLoginDevice, dtype: int64
```

```
In [12]: tourism['Gender'].value_counts() #checking the third incorrect value in 'Gender'
```

```
Out[12]: Male      2916
Female    1817
Fe Male    155
Name: Gender, dtype: int64
```

```
In [13]: #replacing typo with correct 'Female' name
tourism['Gender'] = tourism['Gender'].replace('Fe Male', 'Female')
```

```
In [14]: tourism['Gender'].value_counts()
```

```
Out[14]: Male      2916
Female    1972
Name: Gender, dtype: int64
```

Imputing Null Values with either Mean or Mode based on variable

```
In [15]: tourism['Age'].fillna(value=tourism['Age'].mean(), inplace=True)
```

```
In [16]: tourism['DurationOfPitch'].fillna(value=tourism['DurationOfPitch'].mean(), inplace=True)
```

```
In [17]: tourism['MonthlyIncome'].fillna(value=tourism['MonthlyIncome'].mean(), inplace=True)
```

```
In [18]: tourism['PreferredLoginDevice'].fillna(value=tourism['PreferredLoginDevice'].mode()[0], inplace=True)
```

```
In [19]: tourism['NumberOfFollowups'].fillna(value=tourism['NumberOfFollowups'].mode()[0], inplace=True)
```

```
In [20]: tourism['PreferredPropertyStar'].fillna(value=tourism['PreferredPropertyStar'].mode()[0], inplace=True)
```

```
In [21]: tourism['NumberOfTrips'].fillna(value=tourism['NumberOfTrips'].mode()[0], inplace=True)
```

```
In [22]: tourism['NumberOfChildrenVisited'].fillna(value=tourism['NumberOfChildrenVisited'].mode()[0], inplace=True)
```

```
In [23]: tourism.isna().sum() #checking now cleaned data
```

```
Out[23]: CustomerID      0
ProdTaken      0
Age            0
PreferredLoginDevice  0
CityTier      0
DurationOfPitch  0
Occupation     0
Gender         0
NumberOfPersonVisited  0
NumberOfFollowups  0
ProductPitched  0
PreferredPropertyStar  0
MaritalStatus  0
NumberOfTrips  0
Passport       0
PitchSatisfactionScore  0
OwnCar         0
NumberOfChildrenVisited  0
Designation    0
```

MonthlyIncome
dtype: int64

0

In [24]: `tourism.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4888 entries, 0 to 4887
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   CustomerID                           4888 non-null   int64
1   ProdTaken                            4888 non-null   int64
2   Age                                  4888 non-null   float64
3   PreferredLoginDevice                 4888 non-null   category
4   CityTier                             4888 non-null   int64
5   DurationOfPitch                      4888 non-null   float64
6   Occupation                           4888 non-null   category
7   Gender                               4888 non-null   category
8   NumberOfPersonVisited               4888 non-null   int64
9   NumberOfFollowups                   4888 non-null   float64
10  ProductPitched                      4888 non-null   category
11  PreferredPropertyStar                4888 non-null   float64
12  MaritalStatus                       4888 non-null   category
13  NumberOfTrips                       4888 non-null   float64
14  Passport                             4888 non-null   int64
15  PitchSatisfactionScore               4888 non-null   int64
16  OwnCar                               4888 non-null   int64
17  NumberOfChildrenVisited              4888 non-null   float64
18  Designation                          4888 non-null   category
19  MonthlyIncome                       4888 non-null   float64
dtypes: category(6), float64(7), int64(7)
memory usage: 564.3 KB
```

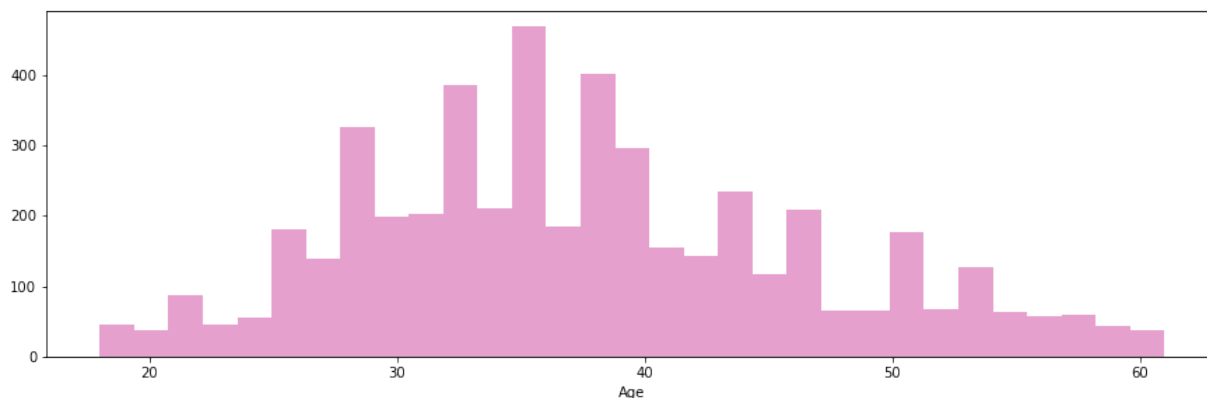
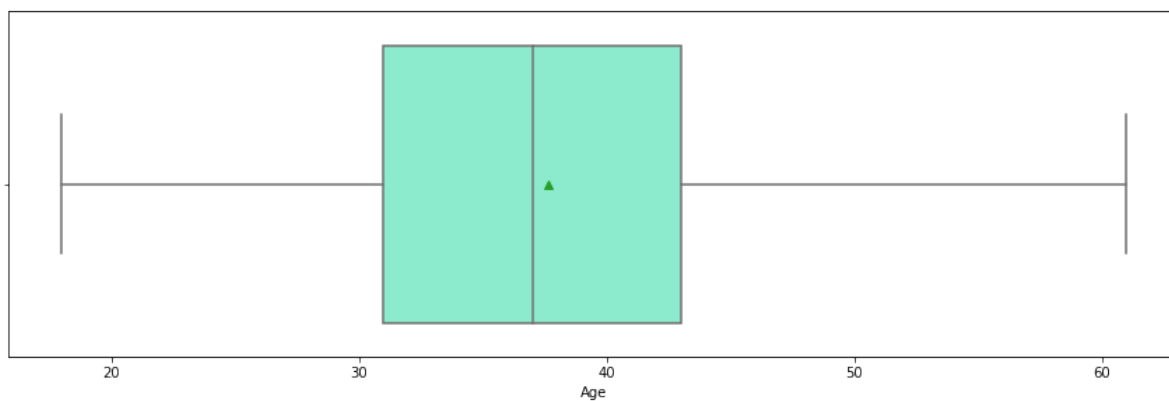
Visualization:

Univariate

In [25]: `tourism.drop(columns="CustomerID",inplace=True)`

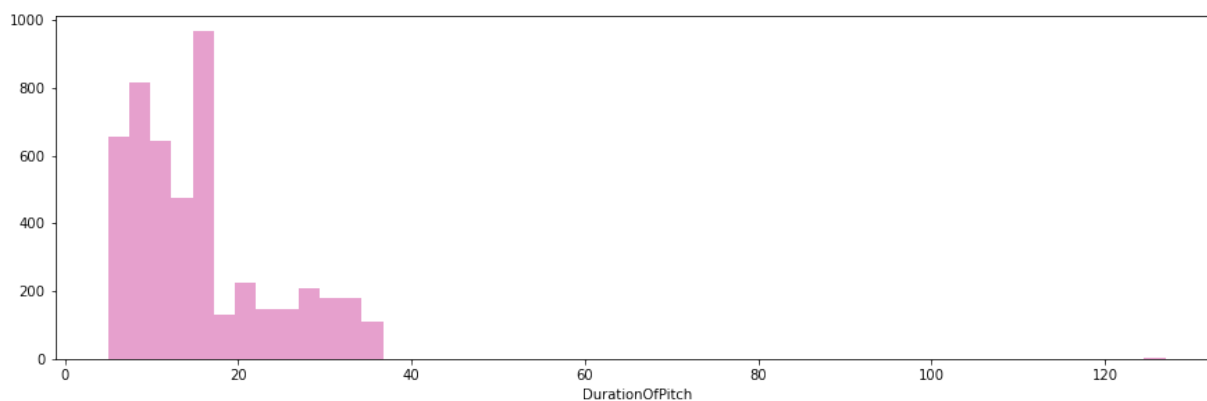
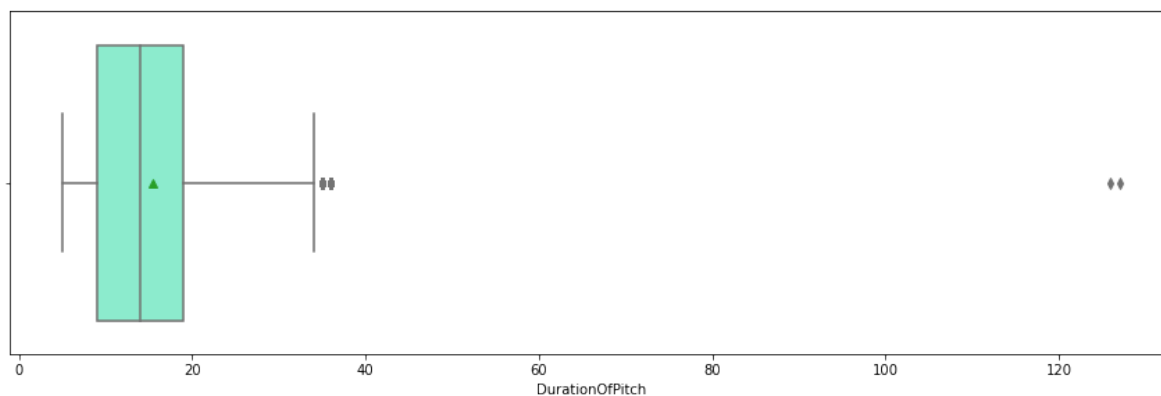
In [26]: `def histogram_boxplot(feature, figsize=(15,10)):`
 `f2, (ax_box2, ax_hist2) = plt.subplots(nrows = 2, # Number of rows of the`
 `figsize = figsize`
 `) # creating the 2 subplots`
 `sns.boxplot(feature, showmeans=True,color='Aquamarine',ax=ax_box2)`
 `sns.distplot(feature,kde=False,color="MediumVioletRed",ax=ax_hist2)`

In [27]: `histogram_boxplot(tourism['Age'])`



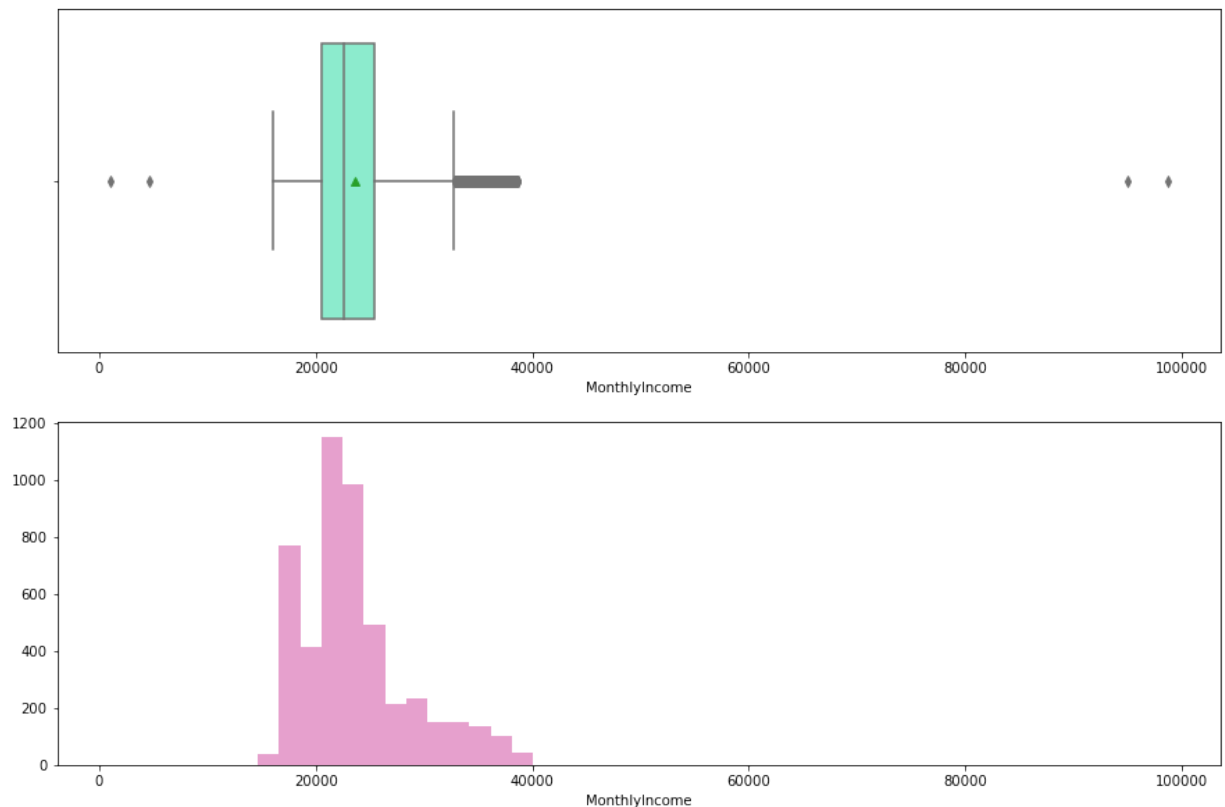
- It seems the Age is approximately symmetrically distributed
- No outliers in the distribution

In [28]: `histogram_boxplot(tourism['DurationOfPitch'])`



- Duration of sales pitches to customers are right skewed, about '18' being the highest count among durations
- There are a few outliers, as some customers may have been more interested or needed more insights during their pitch, hence a few outliers is fine

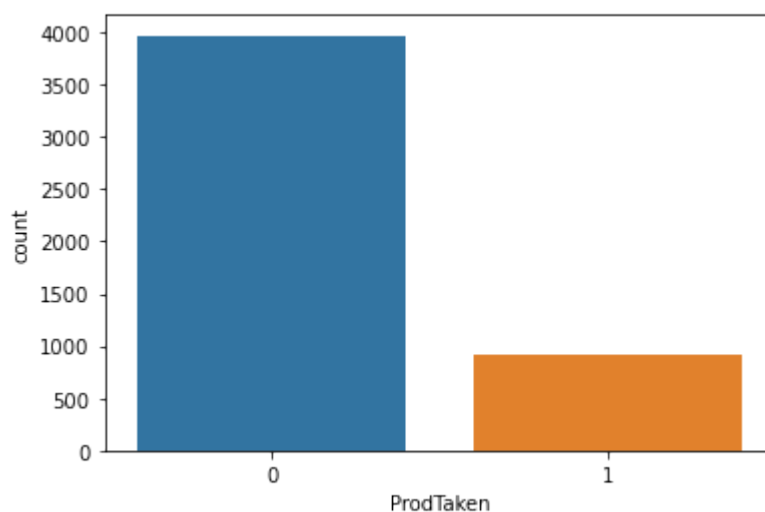
```
In [29]: histogram_boxplot(tourism['MonthlyIncome'])
```



- Distribution of Monthly Income is right skewed, that is to be expected among usual populations
- Non-Symmetric with about 22,000 having the highest count
- A good number of outliers are there and are expected with a distribution of a diversified customer pool

```
In [30]: sns.countplot(tourism['ProdTaken'])
```

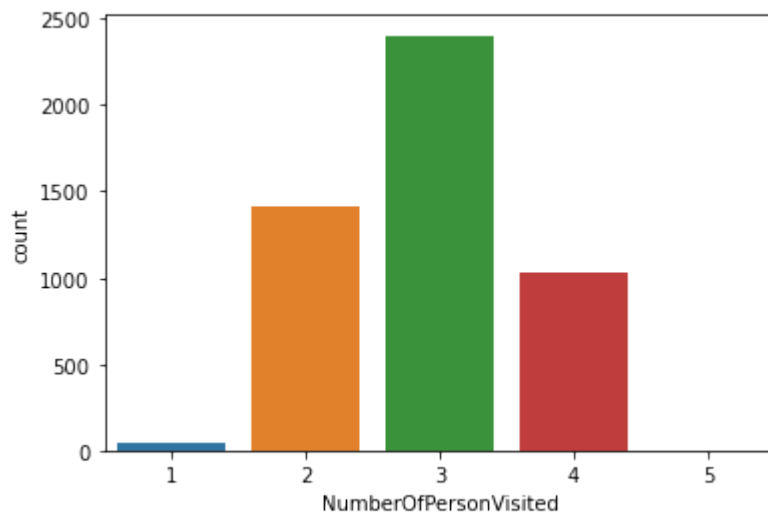
```
Out[30]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



- As just 18% customers purchased the packages, the countplot shows the 82% that have not purchased to be fairly around 3900 customers

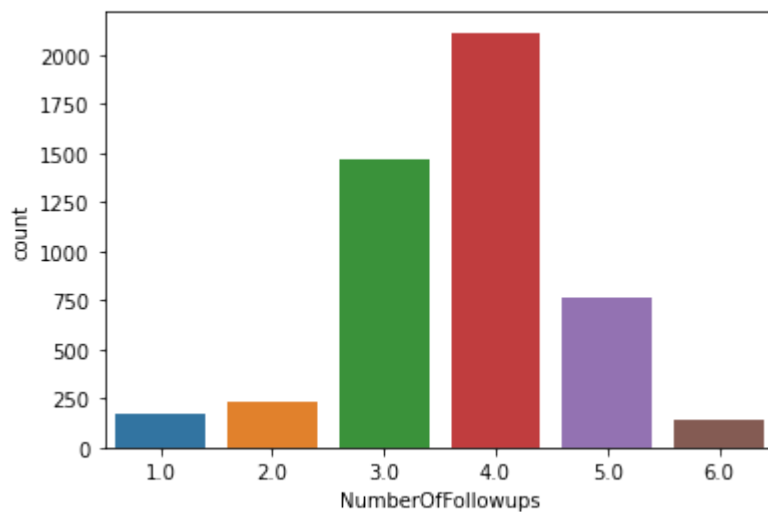
```
In [31]: sns.countplot(tourism['NumberOfPersonVisited'])
```

```
Out[31]: <AxesSubplot:xlabel='NumberOfPersonVisited', ylabel='count'>
```



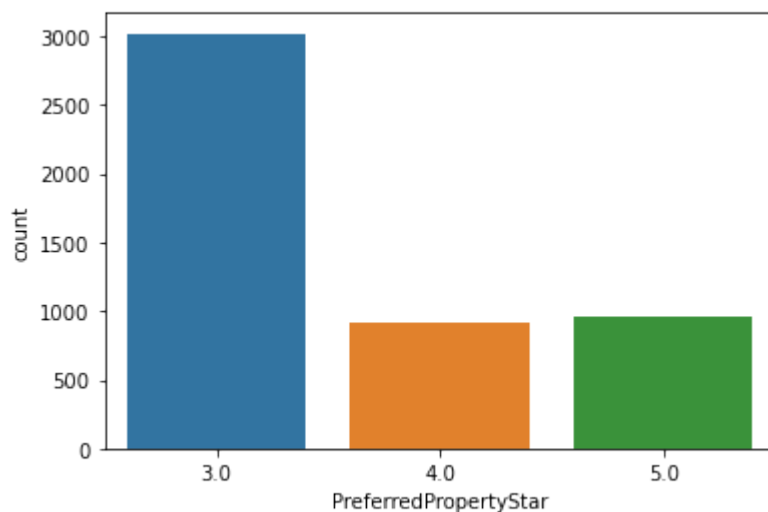
```
In [32]: sns.countplot(tourism['NumberOfFollowups'])
```

```
Out[32]: <AxesSubplot:xlabel='NumberOfFollowups', ylabel='count'>
```



```
In [33]: sns.countplot(tourism['PreferredPropertyStar'])
```

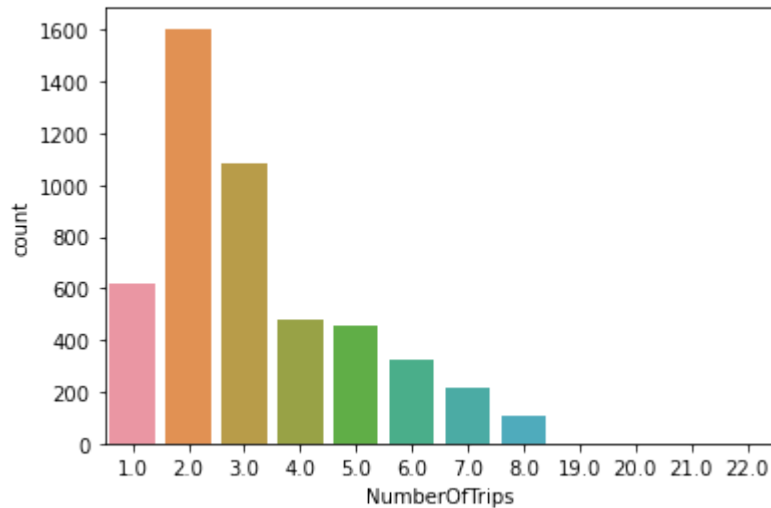
```
Out[33]: <AxesSubplot:xlabel='PreferredPropertyStar', ylabel='count'>
```



- About 75% of customers preferred 3.0 starred properties, with 4.0 and 5.0 having fairly the same amount of distribution for the remaining 15%


```
In [34]: sns.countplot(tourism['NumberOfTrips'])
```

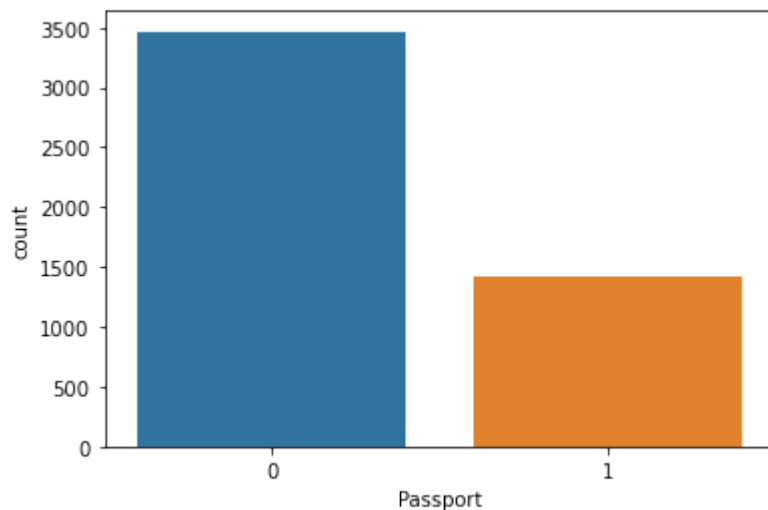
```
Out[34]: <AxesSubplot:xlabel='NumberOfTrips', ylabel='count'>
```



- Average amount per year for customers were mostly 2 or 3 trips

```
In [35]: sns.countplot(tourism['Passport'])
```

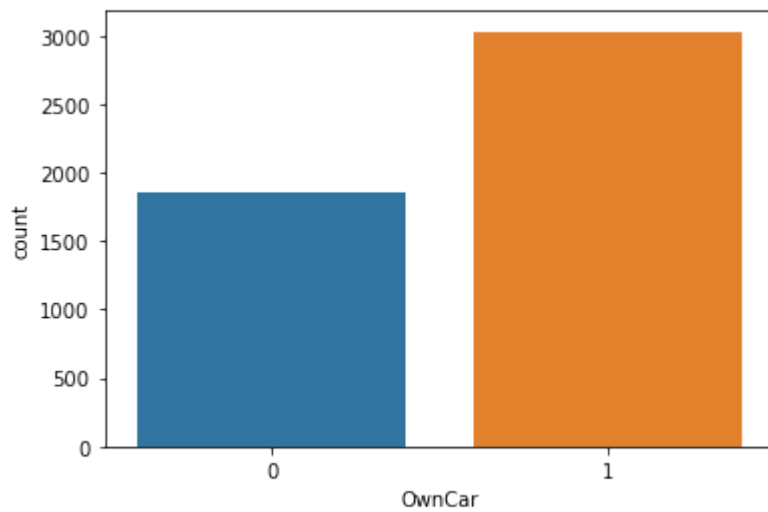
```
Out[35]: <AxesSubplot:xlabel='Passport', ylabel='count'>
```



- Most customers did not have a passport

```
In [36]: sns.countplot(tourism['OwnCar'])
```

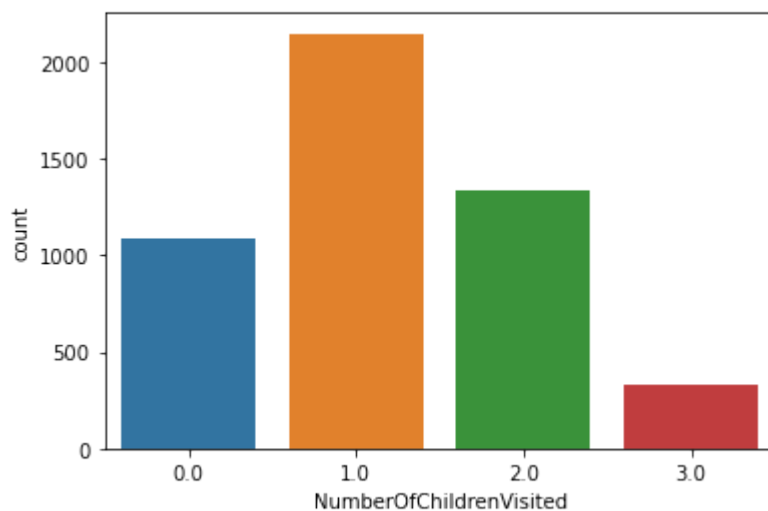
```
Out[36]: <AxesSubplot:xlabel='OwnCar', ylabel='count'>
```



- About just have the customers do not own a car, this can be useful information in sales pitches where the company can offer airport transportation to specific customers

```
In [37]: sns.countplot(tourism['NumberOfChildrenVisited'])
```

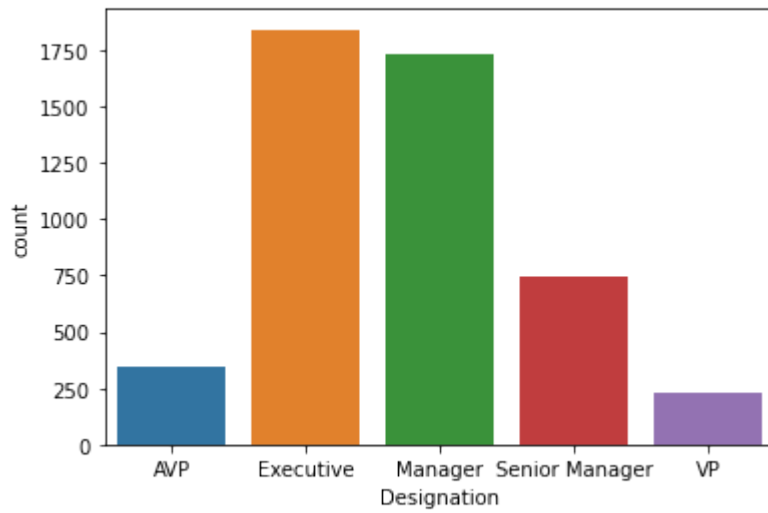
```
Out[37]: <AxesSubplot:xlabel='NumberOfChildrenVisited', ylabel='count'>
```



- Only about 22% of customers were not a family
- Majority had one child

```
In [38]: sns.countplot(tourism['Designation'])
```

```
Out[38]: <AxesSubplot:xlabel='Designation', ylabel='count'>
```

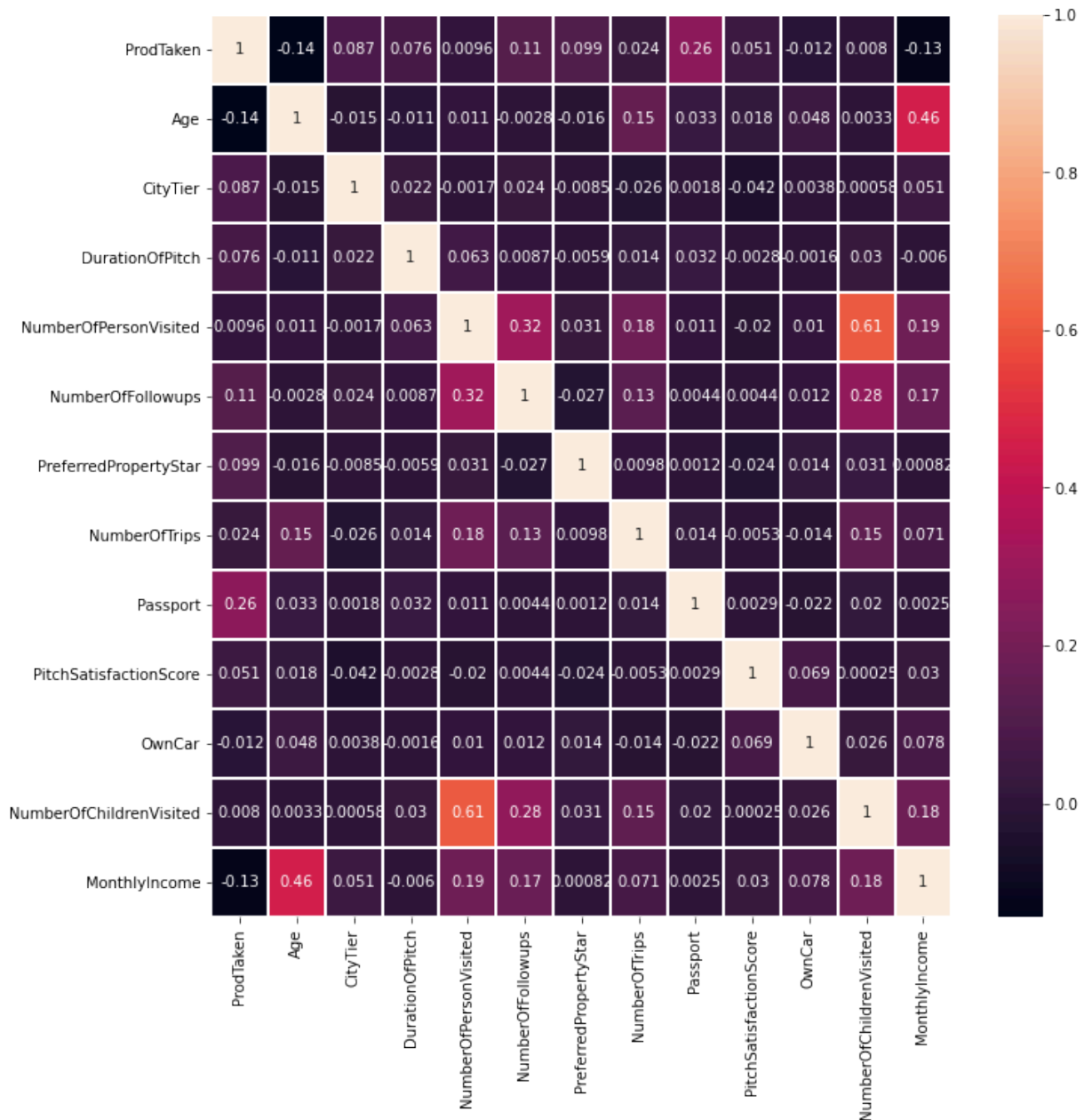


- Interestingly, most customers were either executives or managers, this should be useful for business insights

Bivariate

```
In [39]: fig, ax = plt.subplots(figsize=(11, 11))
sns.heatmap(tourism.corr(),annot=True,linewidth=1)
```

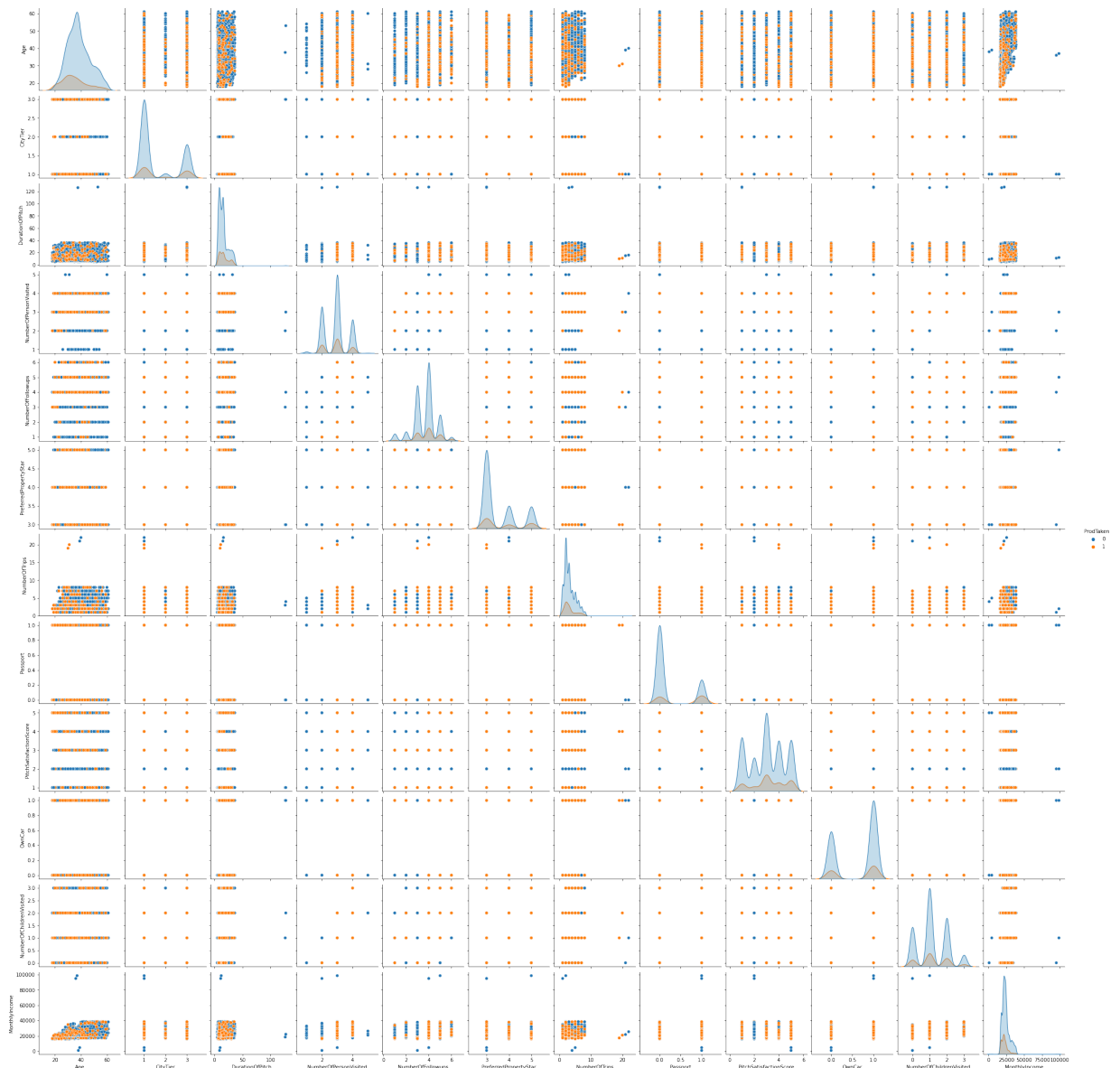
```
Out[39]: <AxesSubplot:>
```



- Number of person visited and Number of children visited have a very good correlation as it is a similar metric
- Age and Monthly Income has a good positive correlation

In [41]: `sns.pairplot(data=tourism,hue="ProdTaken")` #pairplot to further investigate c

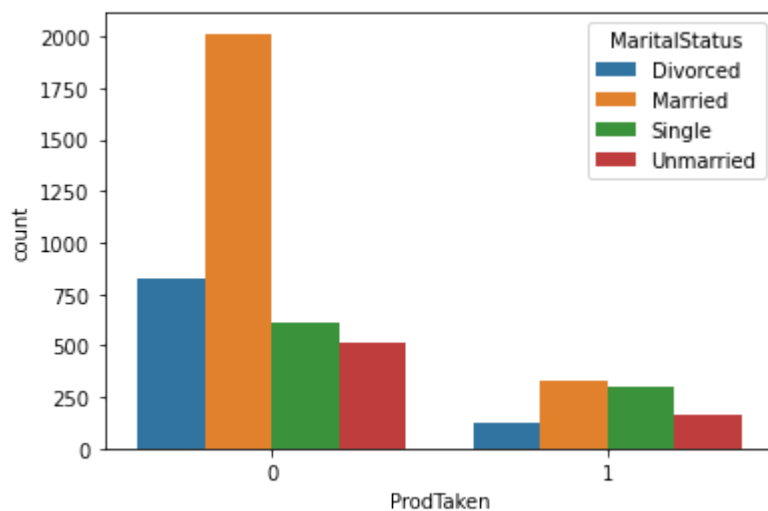
Out[41]: <seaborn.axisgrid.PairGrid at 0x1c520e396a0>



- Since the focus is on getting customers to purchase a package, in these counplots we will mainly focus on the distribution in which products have been taken:

```
In [64]: sns.countplot(tourism["ProdTaken"], hue=tourism["MaritalStatus"])
```

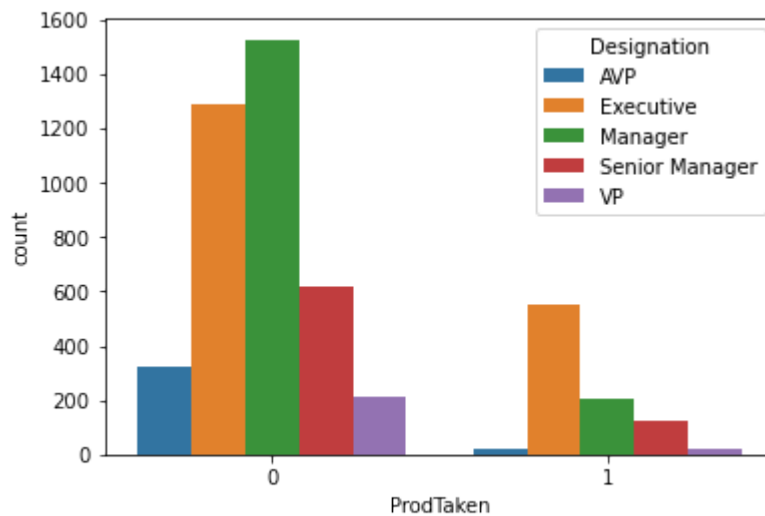
```
Out[64]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



- Since single people are considered unmarried as well, there are mostly that have purchased a package

```
In [55]: sns.countplot(tourism["ProdTaken"], hue=tourism["Designation"])
```

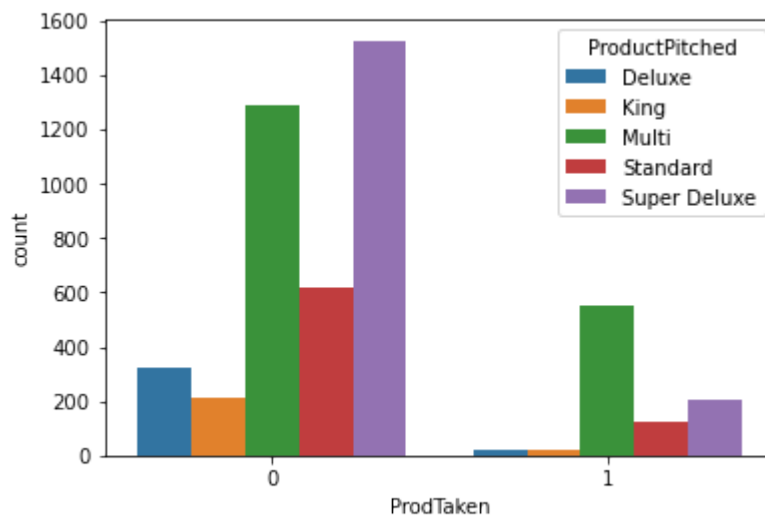
```
Out[55]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



- Executives are mainly the customers who took a package, as most customers were either managers or executive

```
In [56]: sns.countplot(tourism["ProdTaken"], hue=tourism["ProductPitched"])
```

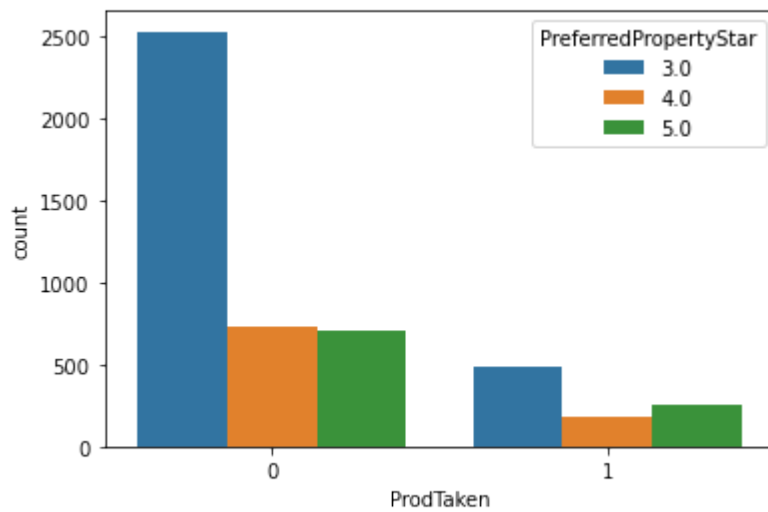
```
Out[56]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



- Multi seems to be the popular choice among the pitched products by sales men, this tells us that products such as King and Deluxe need to have a better attention on being pitched to customers

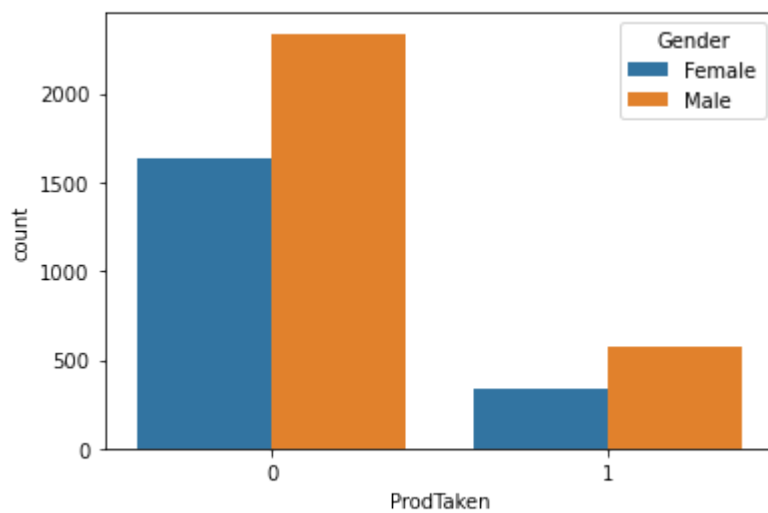
```
In [57]: sns.countplot(tourism["ProdTaken"], hue=tourism["PreferredPropertyStar"])
```

```
Out[57]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



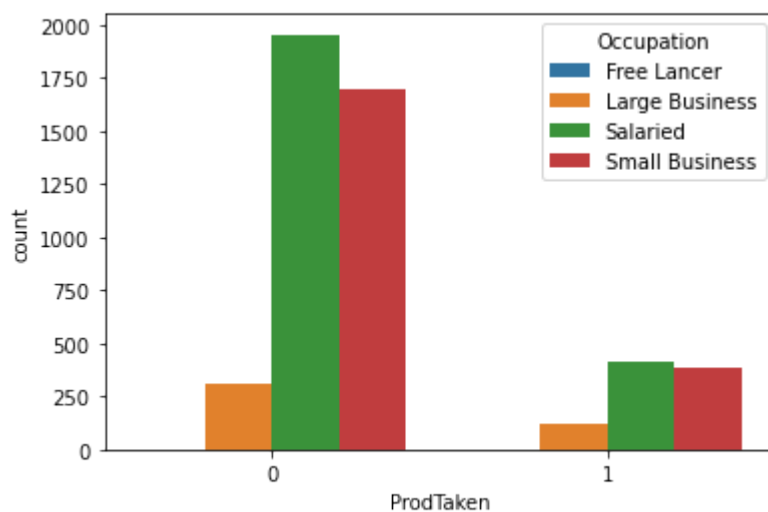
```
In [58]: sns.countplot(tourism["ProdTaken"],hue=tourism["Gender"])
```

```
Out[58]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



```
In [59]: sns.countplot(tourism["ProdTaken"],hue=tourism["Occupation"])
```

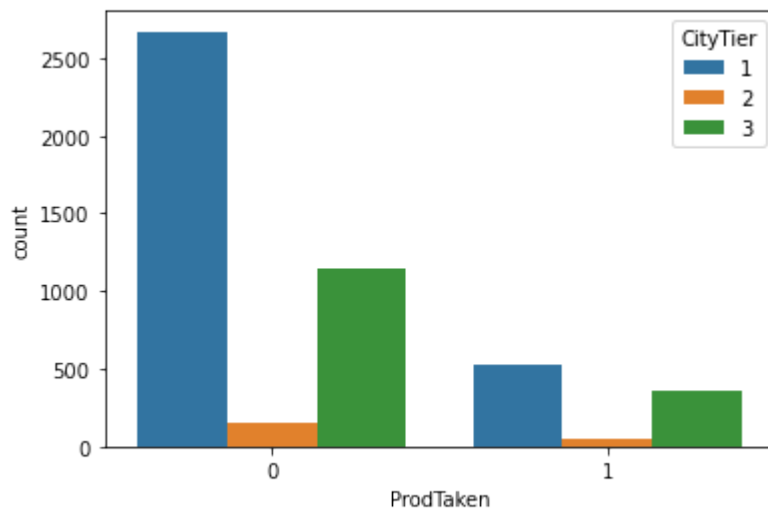
```
Out[59]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



- Occupation of customers are mainly 'Small Business' or 'Salaried'

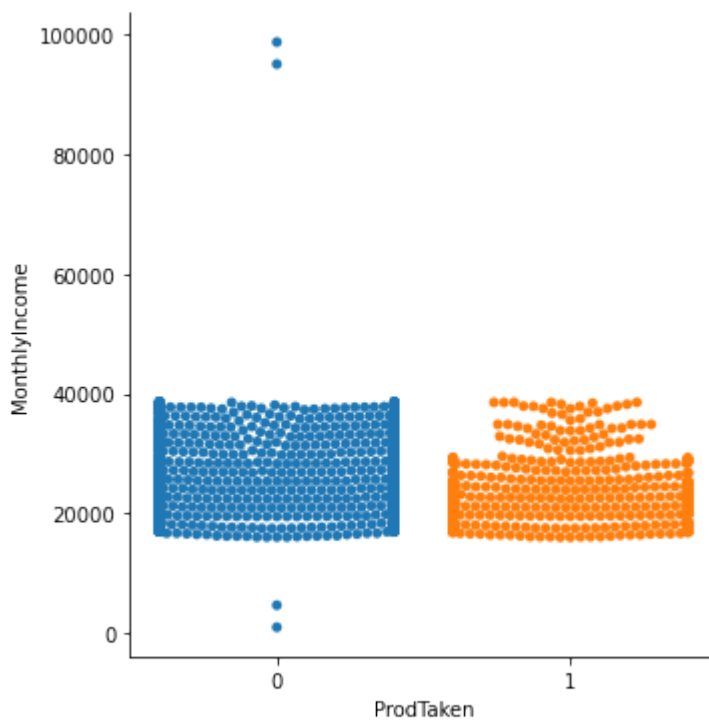
```
In [60]: sns.countplot(tourism["ProdTaken"],hue=tourism["CityTier"])
```

Out[60]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>



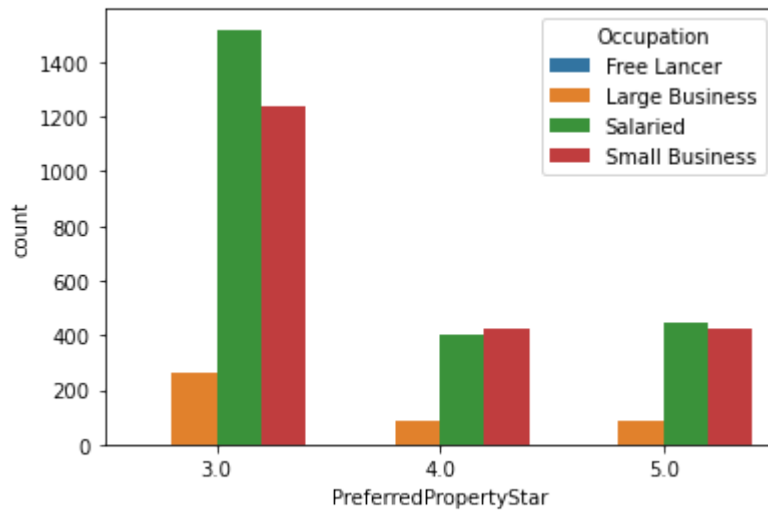
In [50]: `sns.catplot(x="ProdTaken",y="MonthlyIncome",data=tourism, kind="swarm")`

Out[50]: <seaborn.axisgrid.FacetGrid at 0x1c52dcfe8b0>



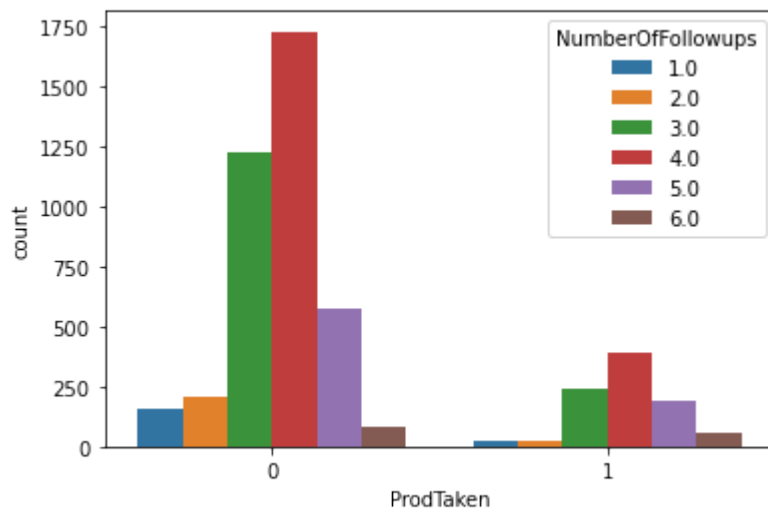
In [52]: `sns.countplot(tourism["PreferredPropertyStar"],hue=tourism["Occupation"])`

Out[52]: <AxesSubplot:xlabel='PreferredPropertyStar', ylabel='count'>



```
In [65]: sns.countplot(tourism["ProdTaken"],hue=tourism["NumberOfFollowups"])
```

```
Out[65]: <AxesSubplot:xlabel='ProdTaken', ylabel='count'>
```



Preparing Data

- There was quite a number of categorical feature, we turn them into dummy variable so our models can perform as we want them to

```
In [83]: dummy_data = pd.get_dummies(tourism, columns=['PreferredLoginDevice', 'CityTie',
                                                    'Designation', 'MaritalStatus'],
dummy_data.head()
```

```
Out[83]:
```

	ProdTaken	Age	DurationOfPitch	NumberOfPersonVisited	NumberOfFollowups	Numk
0	1	41.000000	6.0	3	3.0	
1	0	49.000000	14.0	3	4.0	
2	1	37.000000	8.0	3	4.0	
3	0	33.000000	9.0	2	3.0	
4	0	37.622265	8.0	2	3.0	

5 rows x 31 columns

Splitting the dataset

```
In [84]: x = dummy_data.drop('ProdTaken',axis=1)
         y = dummy_data['ProdTaken']
```

```
In [85]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, rand
         print(x_train.shape, x_test.shape)

(3421, 30) (1467, 30)
```

Building our models

- On each model, the train data will be fitted, analysing scores and performance
- By using hyperparameter tuning, model will be improved everytime using this technique
- With GridSearch cross validation and recall score for model optimization
- Since this is a classification solution, and we want people to purchase our packages, Recall will be used as the metric of evaluation: Recall gives the ratio of True Positives to Actual Positives, where the higher our recall the lower are the False Negatives, chances of predicting product customers as non product customers are lower
- We start by creating functions to get model scores, confusion matrices, importance of features based on the model. This is so we dont use more lines of codes for each model:

```
In [94]: def draw_cm( model, y_actual):
         y_predict = model.predict(x_test)
         cm = metrics.confusion_matrix(y_actual, y_predict)
         sns.heatmap(cm, annot=True)
         plt.ylabel('observed')
         plt.xlabel('predicted')
         plt.show()
```

```
In [92]: def get_scores(model):

         pred_train = model.predict(x_train)
         pred_test = model.predict(x_test)

         print("Accuracy on training set : ",model.score(x_train,y_train))
         print("Accuracy on test set : ",model.score(x_test,y_test))
         print("Recall on training set : ",metrics.recall_score(y_train,pred_train)
         print("Recall on test set : ",metrics.recall_score(y_test,pred_test))
         print("Precision on training set : ",metrics.precision_score(y_train,pred_train)
         print("Precision on test set : ",metrics.precision_score(y_test,pred_test)
```

```
In [132... def feature_importance(model):
            print(pd.DataFrame(model.feature_importances_, columns=["Imp"], index = x
```

Bagging Classifier

```
In [171... bagging=BaggingClassifier(random_state=1)
         bagging.fit(x_train,y_train)
```

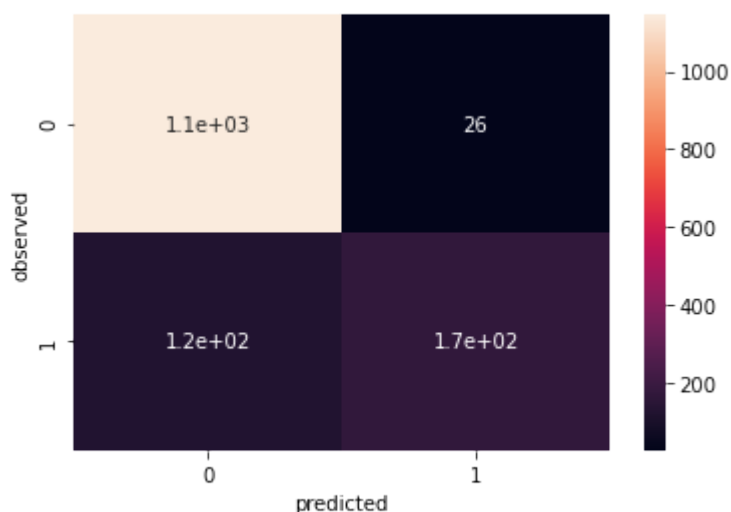
```
Out[171... BaggingClassifier(random_state=1)
```

```
In [172... get_scores(bagging)
```

Accuracy on training set : 0.9935691318327974
 Accuracy on test set : 0.8977505112474438
 Recall on training set : 0.9647435897435898
 Recall on test set : 0.581081081081081
 Precision on training set : 1.0
 Precision on test set : 0.8686868686868687

- On the training data, there is a overfitting of the data, we will see if this could be reduced with Hyperparameter tuning. Accuracy(89%) and Precision(86%) were great, but we will see if recall(58%) could be improved

In [173... draw_cm(bagging, y_test)



Hyperparameter Tuning

```

In [176... bagging_tuned = BaggingClassifier(random_state=1)

parameters = {'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],
              }

acc_scorer = metrics.make_scorer(metrics.recall_score)
grid_obj = GridSearchCV(bagging_tuned, parameters, scoring=acc_scorer,cv=5)
grid_obj = grid_obj.fit(x_train, y_train)

bagging_tuned = grid_obj.best_estimator_
tunedtree.fit(x_train,y_train)
  
```

Out[176... DecisionTreeClassifier(max_depth=2, max_leaf_nodes=3, min_impurity_decrease=0.001, random_state=1)

In [177... get_scores(bagging_tuned)

Accuracy on training set : 0.9991230634317451
 Accuracy on test set : 0.9195637355146558
 Recall on training set : 0.9951923076923077
 Recall on test set : 0.6283783783783784
 Precision on training set : 1.0
 Precision on test set : 0.9587628865979382

- Significant changes was improvement on Precision(95%) and Recall(62%)

Decision Tree

```
In [86]: treemodel = DecisionTreeClassifier(criterion='gini',random_state=1)
```

```
In [87]: treemodel.fit(x_train,y_train)
```

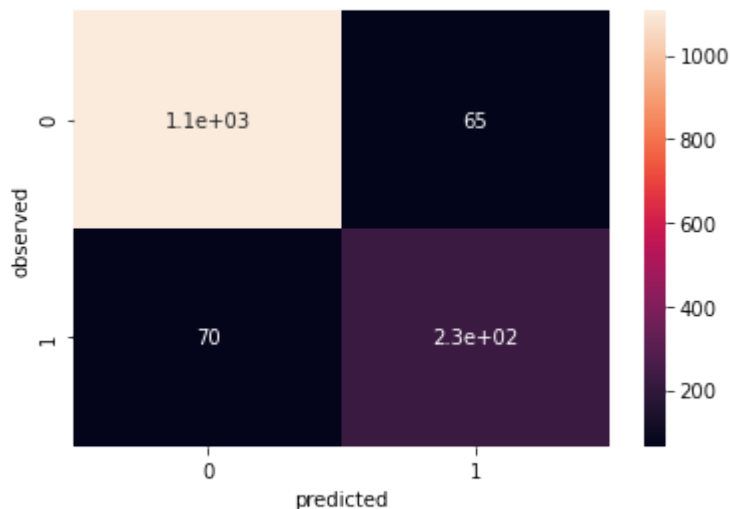
```
Out[87]: DecisionTreeClassifier(random_state=1)
```

```
In [93]: get_scores(treemodel)
```

```
Accuracy on training set : 1.0
Accuracy on test set : 0.9079754601226994
Recall on training set : 1.0
Recall on test set : 0.7635135135135135
Precision on training set : 1.0
Precision on test set : 0.7766323024054983
```

- Decision Tree with default parameters is overfitting the train data, but that is to be expected with ensemble techniques
- We will check if tuning hyperparameters will reduce the overfit and improve overall performance

```
In [95]: draw_cm(treemodel, y_test)
```



Feature Importance of each variable based on Decision Tree

```
In [133]: feature_importance(treemodel)
```

	Imp
Age	0.130905
DurationOfPitch	0.127166
MonthlyIncome	0.108246
PitchSatisfactionScore	0.081616
Passport	0.078662
Designation_Executive	0.071617
NumberOfTrips	0.065822
NumberOfFollowups	0.049526
MaritalStatus_Single	0.036244
CityTier_3	0.033803
PreferredPropertyStar_5.0	0.025536
NumberOfPersonVisited	0.020398
MaritalStatus_Married	0.018592
Gender_Male	0.018575
MaritalStatus_Unmarried	0.017877
PreferredLoginDevice_Self Enquiry	0.016409
NumberOfChildrenVisited	0.014977
OwnCar	0.014573
Occupation_Large Business	0.014132

CityTier_2	0.012213
Occupation_Small Business	0.010508
PreferredPropertyStar_4.0	0.006511
ProductPitched_Multi	0.006170
Designation_Manager	0.006143
Designation_Senior Manager	0.004737
Occupation_Salaried	0.003795
ProductPitched_Super Deluxe	0.003213
Designation_VP	0.001568
ProductPitched_King	0.000465
ProductPitched_Standard	0.000000

Hyperparameter Tuning

In [96]: `from sklearn.model_selection import GridSearchCV`

In [98]:

```
tunedtree = DecisionTreeClassifier(random_state=1)

parameters = {'max_depth': np.arange(1,10),
              'min_samples_leaf': [1, 2, 5, 7, 10,15,20],
              'max_leaf_nodes' : [2, 3, 5, 10],
              'min_impurity_decrease': [0.001,0.01,0.1]}

acc_scorer = metrics.make_scorer(metrics.recall_score)
grid_obj = GridSearchCV(tunedtree, parameters, scoring=acc_scorer,cv=5)
grid_obj = grid_obj.fit(x_train, y_train)

tunedtree = grid_obj.best_estimator_
tunedtree.fit(x_train,y_train)
```

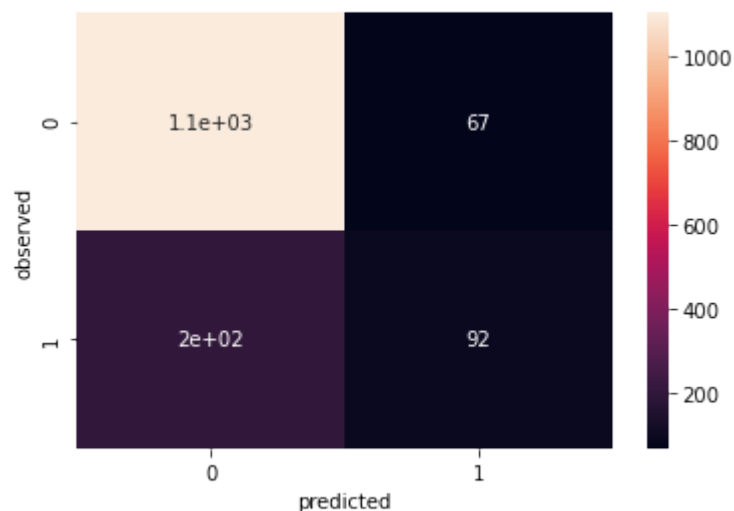
Out[98]: DecisionTreeClassifier(max_depth=2, max_leaf_nodes=3,
min_impurity_decrease=0.001, random_state=1)

In [100... `get_scores(tunedtree)`

Accuracy on training set : 0.8325051154633148
Accuracy on test set : 0.8152692569870484
Recall on training set : 0.3685897435897436
Recall on test set : 0.3108108108108108
Precision on training set : 0.5623471882640587
Precision on test set : 0.5786163522012578

- Attempting on tuning the Decision Tree underfit the train data, with a lower performance as well

In [99]: `draw_cm(tunedtree, y_test)`



Feature Importance of each variable based on Tuned Decision Tree

In [134... `feature_importance(tunedtree)`

	Imp
Passport	0.565147
ProductPitched_Multi	0.434853
Age	0.000000
Gender_Male	0.000000
MaritalStatus_Single	0.000000
MaritalStatus_Married	0.000000
Designation_VP	0.000000
Designation_Senior Manager	0.000000
Designation_Manager	0.000000
Designation_Executive	0.000000
ProductPitched_Super Deluxe	0.000000
ProductPitched_Standard	0.000000
ProductPitched_King	0.000000
PreferredPropertyStar_5.0	0.000000
PreferredPropertyStar_4.0	0.000000
Occupation_Small Business	0.000000
DurationOfPitch	0.000000
Occupation_Salaried	0.000000
Occupation_Large Business	0.000000
CityTier_3	0.000000
CityTier_2	0.000000
PreferredLoginDevice_Self Enquiry	0.000000
MonthlyIncome	0.000000
NumberOfChildrenVisited	0.000000
OwnCar	0.000000
PitchSatisfactionScore	0.000000
NumberOfTrips	0.000000
NumberOfFollowups	0.000000
NumberOfPersonVisited	0.000000
MaritalStatus_Unmarried	0.000000

Random Forest

In [102... `rfmodel = RandomForestClassifier(random_state=1)`
`rfmodel.fit(x_train,y_train)`

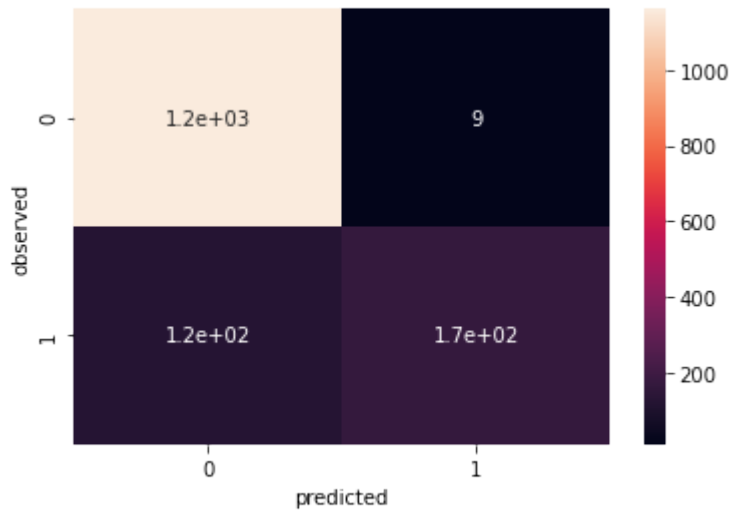
Out[102... `RandomForestClassifier(random_state=1)`

In [103... `get_scores(rfmodel)`

```
Accuracy on training set : 1.0
Accuracy on test set : 0.9086571233810498
Recall on training set : 1.0
Recall on test set : 0.5777027027027027
Precision on training set : 1.0
Precision on test set : 0.95
```

- The Random Forest has a high overfit with a recall that could be improved by hyperparameter tuning

In [150... `draw_cm(rfmodel,y_test)`



Feature Importance of each variable based on Random Forest

In [165... `feature_importance(rfmodel)`

	Imp
MonthlyIncome	0.125300
Age	0.116939
DurationOfPitch	0.102473
Passport	0.087527
PitchSatisfactionScore	0.061371
NumberOfTrips	0.059425
NumberOfFollowups	0.054855
NumberOfChildrenVisited	0.031297
CityTier_3	0.028129
NumberOfPersonVisited	0.027777
PreferredPropertyStar_5.0	0.026889
Designation_Executive	0.026225
MaritalStatus_Single	0.024092
Gender_Male	0.023962
PreferredLoginDevice_Self Enquiry	0.023547
PreferredPropertyStar_4.0	0.021003
OwnCar	0.020028
Occupation_Salaried	0.019390
Occupation_Small Business	0.019021
MaritalStatus_Married	0.016094
ProductPitched_Multi	0.016023
MaritalStatus_Unmarried	0.015771
Occupation_Large Business	0.014698
CityTier_2	0.008182
Designation_Manager	0.008004
ProductPitched_Super Deluxe	0.007556
ProductPitched_Standard	0.005842
Designation_Senior Manager	0.005290
Designation_VP	0.001786
ProductPitched_King	0.001503

Hyperparameter Tuning

```
In [104... rftuned = RandomForestClassifier(random_state=1)

parameters = {'max_depth':[4, 6, 8, 10, None],
              'max_features': ['sqrt', 'log2', None],
              'n_estimators': [80, 90, 100, 110, 120]}

acc_scorer = metrics.make_scorer(metrics.recall_score)
grid_obj = GridSearchCV(rftuned, parameters, scoring=acc_scorer, cv=5)
grid_obj = grid_obj.fit(x_train, y_train)
```

```
rftuned = grid_obj.best_estimator_  
rftuned.fit(x_train,y_train)
```

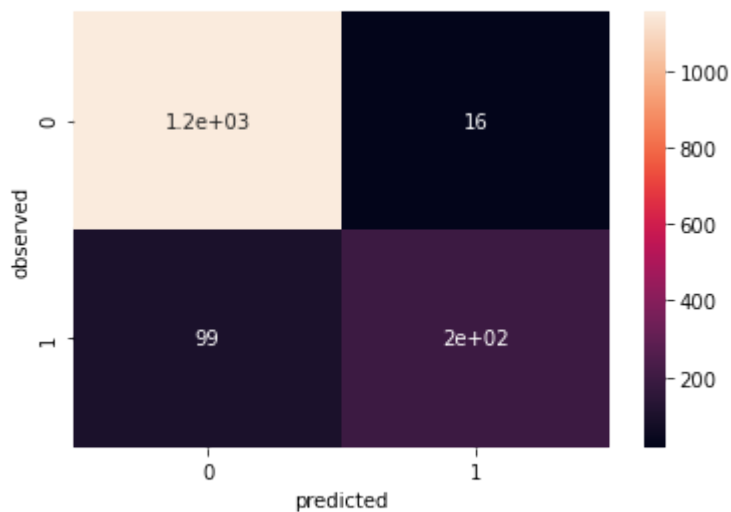
Out[104... RandomForestClassifier(max_features=None, n_estimators=120, random_state=1)

```
In [105... get_scores(rftuned)
```

```
Accuracy on training set : 1.0  
Accuracy on test set : 0.9216087252897068  
Recall on training set : 1.0  
Recall on test set : 0.6655405405405406  
Precision on training set : 1.0  
Precision on test set : 0.9248826291079812
```

-fitting of the data stayed the same, Although we have a much better recall score of 66% compared to 57%

```
In [106... draw_cm(rftuned, y_test)
```



Feature Importance of each variable based on Tuned Random Forest

```
In [135... feature_importance(rftuned)
```

	Imp
Age	0.127792
DurationOfPitch	0.125646
MonthlyIncome	0.123922
Passport	0.080834
PitchSatisfactionScore	0.059524
NumberOfTrips	0.058700
NumberOfFollowups	0.050915
CityTier_3	0.038191
Designation_Executive	0.037720
ProductPitched_Multi	0.036432
MaritalStatus_Single	0.022430
PreferredPropertyStar_5.0	0.021607
NumberOfChildrenVisited	0.020984
Gender_Male	0.019946
PreferredLoginDevice_Self Enquiry	0.019231
MaritalStatus_Unmarried	0.017303
Occupation_Large Business	0.016663
NumberOfPersonVisited	0.016658
MaritalStatus_Married	0.016134
PreferredPropertyStar_4.0	0.015565
OwnCar	0.015492
Occupation_Small Business	0.014251
Occupation_Salaried	0.010748
CityTier_2	0.008628
ProductPitched_Super Deluxe	0.007077
Designation_Manager	0.006127

Designation_Senior Manager	0.005042
ProductPitched_Standard	0.004537
Designation_VP	0.001027
ProductPitched_King	0.000874

Adaptive Boost

```
In [107... ab_classifier = AdaBoostClassifier(random_state=1)
```

```
In [108... ab_classifier.fit(x_train,y_train)
```

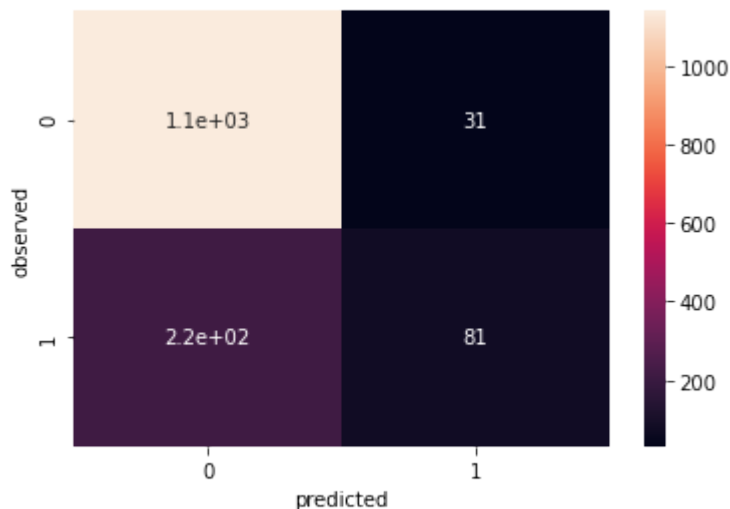
```
Out[108... AdaBoostClassifier(random_state=1)
```

```
In [109... get_scores(ab_classifier)
```

```
Accuracy on training set : 0.8520900321543409
Accuracy on test set : 0.8323108384458078
Recall on training set : 0.3269230769230769
Recall on test set : 0.27364864864864863
Precision on training set : 0.7034482758620689
Precision on test set : 0.7232142857142857
```

- There is a good fitting of the data on the AdaBoost, but a very poor score on Recall of 27%
- Hyperparameter tuning should improve this score

```
In [136... draw_cm(ab_classifier,y_test)
```



Feature Importance of each variable based on Adaptive Boost

```
In [137... feature_importance(ab_classifier)
```

	Imp
MonthlyIncome	0.36
Age	0.14
DurationOfPitch	0.06
NumberOfFollowups	0.06
NumberOfPersonVisited	0.04
Passport	0.04
PreferredPropertyStar_5.0	0.04
MaritalStatus_Single	0.02
Designation_VP	0.02
Designation_Senior Manager	0.02
Designation_Executive	0.02
PreferredPropertyStar_4.0	0.02
Gender_Male	0.02

MaritalStatus_Unmarried	0.02
Occupation_Large Business	0.02
CityTier_3	0.02
CityTier_2	0.02
PreferredLoginDevice_Self Enquiry	0.02
PitchSatisfactionScore	0.02
NumberOfTrips	0.02
Occupation_Salaried	0.00
ProductPitched_King	0.00
ProductPitched_Multi	0.00
ProductPitched_Standard	0.00
ProductPitched_Super Deluxe	0.00
NumberOfChildrenVisited	0.00
Designation_Manager	0.00
OwnCar	0.00
MaritalStatus_Married	0.00
Occupation_Small Business	0.00

Hyperparameter Tuning

```
In [110... ab_tuned = AdaBoostClassifier(random_state=1)

parameters = {'n_estimators': np.arange(10,100,10),
              'learning_rate': [1, 0.1, 0.5, 0.01],
              }

acc_scorer = metrics.make_scorer(metrics.recall_score)
grid_obj = GridSearchCV(ab_tuned, parameters, scoring=acc_scorer,cv=5)
grid_obj = grid_obj.fit(x_train, y_train)

ab_tuned = grid_obj.best_estimator_
ab_tuned.fit(x_train,y_train)
```

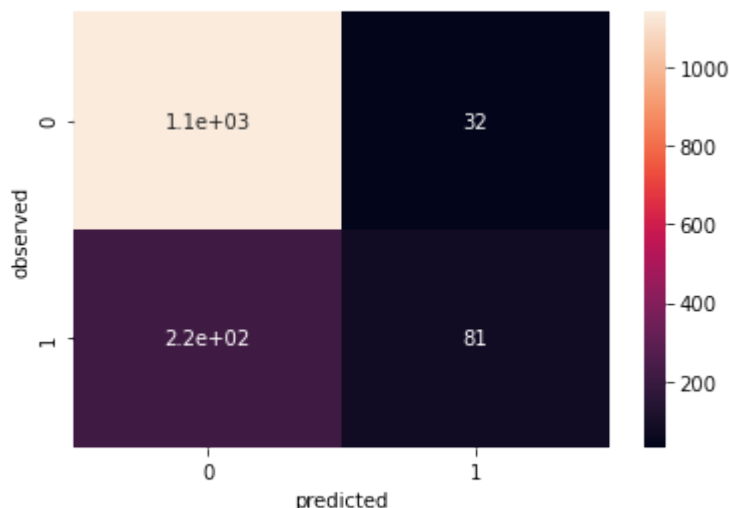
```
Out[110... AdaBoostClassifier(learning_rate=1, n_estimators=40, random_state=1)
```

```
In [111... get_scores(ab_tuned)
```

```
Accuracy on training set : 0.8488745980707395
Accuracy on test set : 0.8316291751874574
Recall on training set : 0.3173076923076923
Recall on test set : 0.27364864864864863
Precision on training set : 0.6851211072664359
Precision on test set : 0.7168141592920354
```

- Attempting to tune the AdaptiveBoost model did not return a drastic difference on improvement, continuing to see how other models can vary in performance

```
In [112... draw_cm(ab_tuned,y_test)
```



Feature Importance of each variable based on Tuned Adaptive Boost

In [138... `feature_importance(ab_tuned)`

	Imp
MonthlyIncome	0.225
Age	0.175
NumberOfFollowups	0.075
DurationOfPitch	0.050
NumberOfPersonVisited	0.050
Passport	0.050
PreferredPropertyStar_5.0	0.050
MaritalStatus_Single	0.025
Designation_VP	0.025
Designation_Senior Manager	0.025
Designation_Executive	0.025
PreferredPropertyStar_4.0	0.025
Gender_Male	0.025
MaritalStatus_Unmarried	0.025
Occupation_Large Business	0.025
CityTier_3	0.025
CityTier_2	0.025
PreferredLoginDevice_Self Enquiry	0.025
PitchSatisfactionScore	0.025
NumberOfTrips	0.025
Occupation_Salaried	0.000
ProductPitched_King	0.000
ProductPitched_Multi	0.000
ProductPitched_Standard	0.000
ProductPitched_Super Deluxe	0.000
NumberOfChildrenVisited	0.000
Designation_Manager	0.000
OwnCar	0.000
MaritalStatus_Married	0.000
Occupation_Small Business	0.000

Gradient Boost

In [114... `gb_model = GradientBoostingClassifier(random_state=1)`
`gb_model.fit(x_train,y_train)`

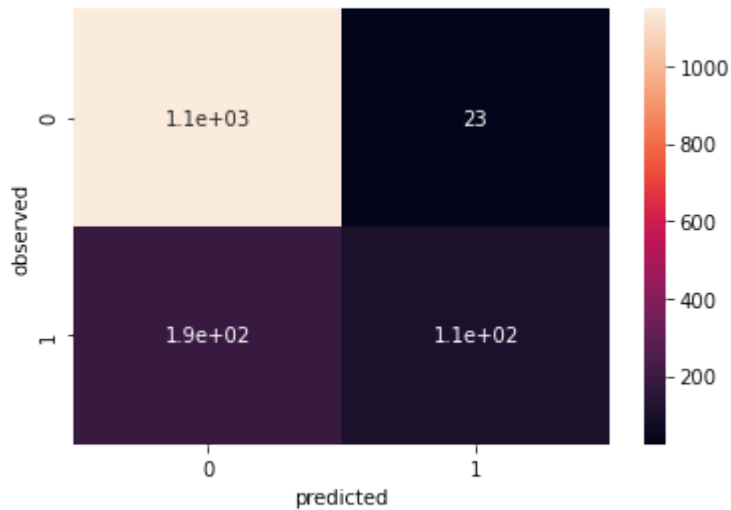
Out[114... `GradientBoostingClassifier(random_state=1)`

In [115... `get_scores(gb_model)`

Accuracy on training set : 0.8918444899152295
 Accuracy on test set : 0.8548057259713702
 Recall on training set : 0.4951923076923077
 Recall on test set : 0.3581081081081081
 Precision on training set : 0.8489010989010989
 Precision on test set : 0.8217054263565892

- Overall, the performance of this Gradient Boost is very good, but lacks on fitting the train data on recall(49%) and has a poor recall test score (35%)
- We will try to improve the fitting of the data with hyperparameter tuning

In [118... `draw_cm(gb_model,y_test)`



Feature Importance of each variable based on Gradient Boost

In [139]... `feature_importance(gb_model)`

	Imp
Passport	0.197952
MonthlyIncome	0.118461
Age	0.118287
Designation_Executive	0.087993
DurationOfPitch	0.081403
NumberOfFollowups	0.060805
ProductPitched_Multi	0.060702
CityTier_3	0.039383
MaritalStatus_Single	0.038736
PitchSatisfactionScore	0.030816
PreferredPropertyStar_5.0	0.024755
NumberOfTrips	0.018136
MaritalStatus_Unmarried	0.017813
Gender_Male	0.017259
PreferredLoginDevice_Self Enquiry	0.014413
Occupation_Large Business	0.011653
Designation_Manager	0.010871
MaritalStatus_Married	0.009972
PreferredPropertyStar_4.0	0.009452
ProductPitched_Standard	0.008074
CityTier_2	0.005142
NumberOfPersonVisited	0.004946
Designation_Senior Manager	0.003809
ProductPitched_Super Deluxe	0.003414
OwnCar	0.002666
Occupation_Small Business	0.001519
Occupation_Salaried	0.001258
NumberOfChildrenVisited	0.000308
ProductPitched_King	0.000000
Designation_VP	0.000000

Hyperparameter Tuning

```
In [121]... gb_tuned = GradientBoostingClassifier(random_state=1)

parameters = {'n_estimators': np.arange(50,200,25),
              'subsample': [0.7,0.8,1],
              'max_features': [0.7,0.8,1],
              'max_depth': [3,5,7,10]}

acc_scorer = metrics.make_scorer(metrics.recall_score)
grid_obj = GridSearchCV(gb_tuned, parameters, scoring=acc_scorer,cv=5)
grid_obj = grid_obj.fit(x_train, y_train)
```

```
gb_tuned = grid_obj.best_estimator_  
gb_tuned.fit(x_train,y_train)
```

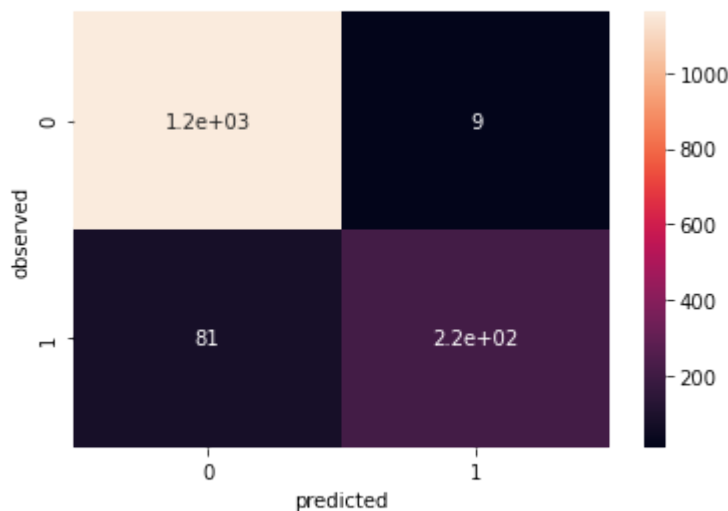
```
Out[121...] GradientBoostingClassifier(max_depth=10, max_features=0.8, n_estimators=150,  
                                random_state=1, subsample=1)
```

```
In [122...] get_scores(gb_tuned)
```

```
Accuracy on training set : 1.0  
Accuracy on test set : 0.9386503067484663  
Recall on training set : 1.0  
Recall on test set : 0.7263513513513513  
Precision on training set : 1.0  
Precision on test set : 0.9598214285714286
```

- It seems that tuning the hyper parameters of the Gradient Boost improved both Recall and Precision scores (72%) and (95%) respectively, with what seems to be a very competitive model up until now

```
In [141...] draw_cm(gb_tuned,y_test)
```



Feature Importance of each variable based on Tuned Gradient Boost

```
In [142...] feature_importance(gb_tuned)
```

	Imp
DurationOfPitch	0.133598
Age	0.128715
MonthlyIncome	0.124058
Passport	0.090284
PitchSatisfactionScore	0.063222
NumberOfFollowups	0.054898
NumberOfTrips	0.050829
Designation_Executive	0.033348
CityTier_3	0.032178
ProductPitched_Multi	0.028557
PreferredPropertyStar_5.0	0.026828
MaritalStatus_Single	0.025194
Gender_Male	0.023027
PreferredLoginDevice_Self Enquiry	0.019890
PreferredPropertyStar_4.0	0.018133
Occupation_Large Business	0.017771
MaritalStatus_Unmarried	0.016185
NumberOfChildrenVisited	0.014994
NumberOfPersonVisited	0.014680
Occupation_Small Business	0.014537
MaritalStatus_Married	0.014343
OwnCar	0.012150
CityTier_2	0.009967

ProductPitched_Super Deluxe	0.008231
Occupation_Salaried	0.007656
Designation_Manager	0.006329
Designation_Senior Manager	0.005149
ProductPitched_Standard	0.003844
ProductPitched_King	0.000848
Designation_VP	0.000557

Extreme Gradient Boost

```
In [123...] xgb = XGBClassifier(random_state=1)
xgb.fit(x_train,y_train)
```

[13:47:11] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

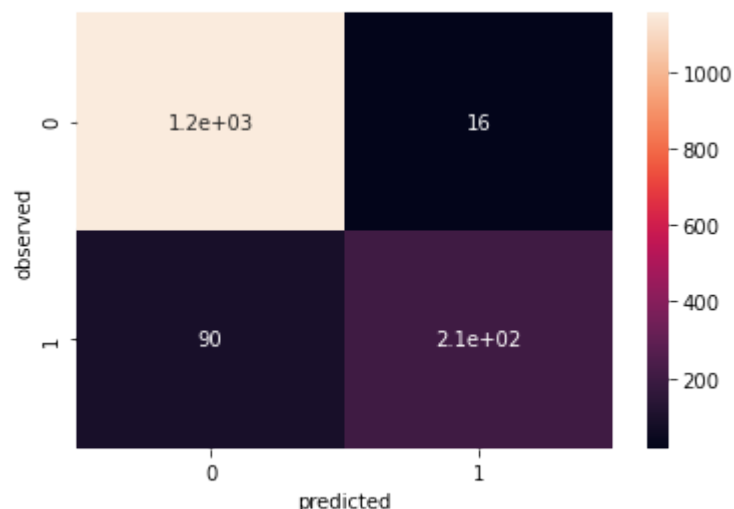
```
Out[123...] XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
importance_type='gain', interaction_constraints='',
learning_rate=0.300000012, max_delta_step=0, max_depth=6,
min_child_weight=1, missing=nan, monotone_constraints=(),
n_estimators=100, n_jobs=16, num_parallel_tree=1, random_state=
1,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
tree_method='exact', validate_parameters=1, verbosity=None)
```

```
In [124...] get_scores(xgb)
```

Accuracy on training set : 0.9994153756211634
Accuracy on test set : 0.9277436946148603
Recall on training set : 0.9967948717948718
Recall on test set : 0.6959459459459459
Precision on training set : 1.0
Precision on test set : 0.9279279279279279

- Overall, Extreme Gradient Boost returned very good scores on Recall and Precision, and as seen throughout most models, there is an overfit on training data which Ensemble models tend to do but clearly does not affect the performance

```
In [143...] draw_cm(xgb,y_test)
```



Feature Importance of each variable based on Extreme Gradient Boost

```
In [144...] feature_importance(xgb)
```

	Imp
ProductPitched_Multi	0.142975

Passport	0.117376
CityTier_3	0.047344
MaritalStatus_Single	0.045267
ProductPitched_Super Deluxe	0.040600
Occupation_Large Business	0.040528
CityTier_2	0.039117
ProductPitched_Standard	0.036149
PreferredPropertyStar_5.0	0.035984
NumberOfFollowups	0.035962
PreferredPropertyStar_4.0	0.033011
PreferredLoginDevice_Self Enquiry	0.030515
NumberOfTrips	0.030308
PitchSatisfactionScore	0.030265
Age	0.030189
MaritalStatus_Unmarried	0.029631
DurationOfPitch	0.029480
MaritalStatus_Married	0.028988
Occupation_Small Business	0.028085
MonthlyIncome	0.024513
Gender_Male	0.023719
ProductPitched_King	0.023197
OwnCar	0.021106
Occupation_Salaried	0.020260
NumberOfPersonVisited	0.019474
NumberOfChildrenVisited	0.015956
Designation_Executive	0.000000
Designation_Manager	0.000000
Designation_Senior Manager	0.000000
Designation_VP	0.000000

Hyperparameter Tuning

In [125...

```
xgb_tuned = XGBClassifier(random_state=1)

parameters = {'n_estimators': [75,100,125,150],
              'subsample':[0.7, 0.8, 0.9, 1],
              'gamma':[0, 1, 3, 5],
              'colsample_bytree':[0.7, 0.8, 0.9, 1],
              'colsample_bylevel':[0.7, 0.8, 0.9, 1]
            }

acc_scorer = metrics.make_scorer(metrics.recall_score)
grid_obj = GridSearchCV(xgb_tuned, parameters, scoring=acc_scorer,cv=5)
grid_obj = grid_obj.fit(x_train, y_train)

xgb_tuned = grid_obj.best_estimator_
xgb_tuned.fit(x_train,y_train)
```

[13:50:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[13:50:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[13:50:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[13:50:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[13:50:01] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[13:50:01] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the de

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGha... 46/332

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGha... 53/332

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGha... 54/332

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGha... 57/332

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
[13:51:50] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the de
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:52:06] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:06] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:06] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:07] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:07] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:07] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:07] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:08] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:08] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:08] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:09] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:52:29] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:52:47] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:47] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:47] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:48] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:48] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:48] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:48] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[13:52:49] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

```
[13:53:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:53:56] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
[13:54:26] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
[13:54:47] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

```
[13:54:48] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

```
[13:54:49] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

```
[13:54:50] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:55:51] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:56:07] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

```
[13:56:18] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:56:59] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
[13:57:51] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```


[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGh... 218/332

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGh... 221/332

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
[13:58:30] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed
```

[illegible]

```
[13:58:36] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[13:59:15] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGh... 278/332

```
[14:00:47] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the de
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGh... 293/332

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGh... 295/332

[illegible]

[illegible]

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGh... 301/332

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[14:02:05] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[14:02:06] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[14:02:06] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[14:02:07] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[14:02:08] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.  
[14:02:09] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[14:02:36] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the de
```


[illegible]

[illegible]

[illegible]

[illegible]

file:///Users/murtadaalghanim/Desktop/Desktop - Murtada's MacBook Air/sec/Texas Projects/Travel Package Purchase Prediction - Murtadha Assad BinGh... 327/332

fault evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:02:58] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:02:59] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:02:59] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:02:59] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:02:59] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:02:59] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:03:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:03:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:03:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

[14:03:00] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

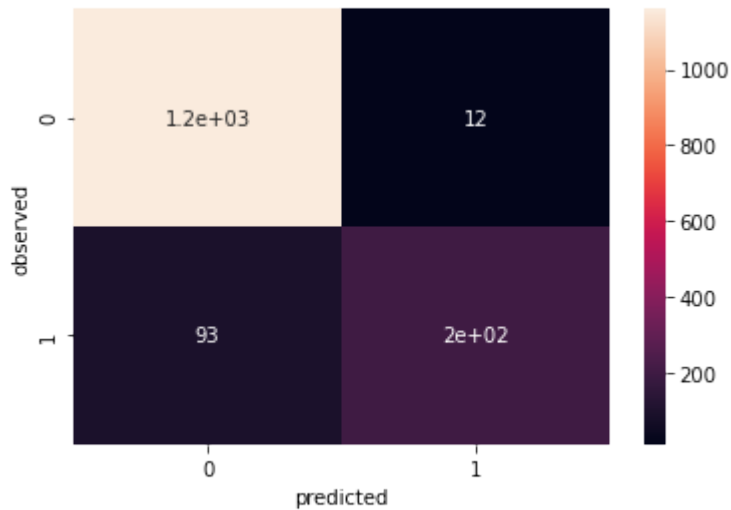
```
Out[125...] XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=0.9,
                    colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
                    importance_type='gain', interaction_constraints='',
                    learning_rate=0.300000012, max_delta_step=0, max_depth=6,
                    min_child_weight=1, missing=nan, monotone_constraints='()',
                    n_estimators=150, n_jobs=16, num_parallel_tree=1, random_state=
1,
                    reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=0.9,
                    tree_method='exact', validate_parameters=1, verbosity=None)
```

```
In [126...] get_scores(xgb_tuned)
```

```
Accuracy on training set : 1.0
Accuracy on test set : 0.9284253578732107
Recall on training set : 1.0
Recall on test set : 0.6858108108108109
Precision on training set : 1.0
Precision on test set : 0.9441860465116279
```

- Tuning the Extreme Gradient Boosting returned no significant change

```
In [127...] draw_cm(xgb_tuned, y_test)
```

Feature Importance of each variable based on Tuned Extreme Gradient Boost

In [145... `feature_importance(xgb_tuned)`

	Imp
ProductPitched_Multi	0.129620
Passport	0.091227
MaritalStatus_Single	0.044484
CityTier_2	0.042286
PreferredPropertyStar_5.0	0.042036
CityTier_3	0.041816
Occupation_Large Business	0.037933
MaritalStatus_Unmarried	0.037709
Designation_Senior Manager	0.035529
PreferredPropertyStar_4.0	0.034384
PitchSatisfactionScore	0.032002
MaritalStatus_Married	0.031690
NumberOfFollowups	0.030682
ProductPitched_Standard	0.029541
DurationOfPitch	0.028437
NumberOfTrips	0.028137
Age	0.027924
Gender_Male	0.027470
PreferredLoginDevice_Self Enquiry	0.027193
Occupation_Small Business	0.024723
MonthlyIncome	0.022910
ProductPitched_Super Deluxe	0.021302
NumberOfPersonVisited	0.020375
ProductPitched_King	0.020103
Designation_Executive	0.019425
Occupation_Salaried	0.018521
OwnCar	0.016944
NumberOfChildrenVisited	0.013741
Designation_Manager	0.012672
Designation_VP	0.009187

Stacking

In [128... `estimators=[('Decision Tree', tunedtree), ('Random Forest', rftuned), ('Gradient Boosting', gbtuned)]`
`final_estimator=XGBClassifier(random_state=1)`

In [129... `stacking_estimator=StackingClassifier(estimators=estimators, final_estimator=final_estimator)`
`stacking_estimator.fit(x_train,y_train)`

[14:14:46] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

```

Out[129... StackingClassifier(cv=5,
                                estimators=[('Decision Tree',
                                              DecisionTreeClassifier(max_depth=2,
                                                                      max_leaf_nodes=3,
                                                                      min_impurity_decrease=
0.001,
                                                                      random_state=1)),
                                              ('Random Forest',
                                              RandomForestClassifier(max_features=None,
                                                                      n_estimators=120,
                                                                      random_state=1)),
                                              ('Gradient Boosting',
                                              GradientBoostingClassifier(max_depth=10,
                                                                              max_features=0.8,
                                                                              n_estimators=150,
                                                                              random_state=1,
                                                                              S...
                                                                              importance_type='gain',
                                                                              interaction_constraints=None
e,
                                                                              learning_rate=None,
                                                                              max_delta_step=None,
                                                                              max_depth=None,
                                                                              min_child_weight=None,
                                                                              missing=nan,
                                                                              monotone_constraints=None,
                                                                              n_estimators=100, n_jobs=None
e,
                                                                              num_parallel_tree=None,
                                                                              random_state=1, reg_alpha=None
ne,
                                                                              reg_lambda=None,
                                                                              scale_pos_weight=None,
                                                                              subsample=None,
                                                                              tree_method=None,
                                                                              validate_parameters=None,
                                                                              verbosity=None))

```

```
In [130... get_scores(stacking_estimator)
```

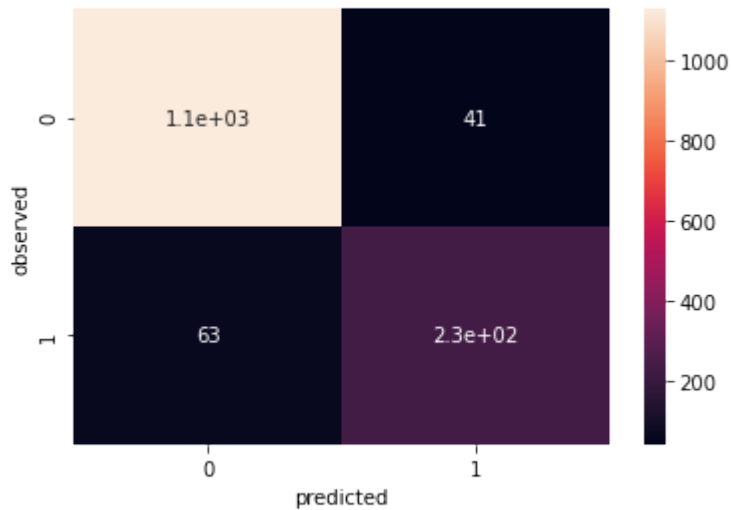
```

Accuracy on training set : 0.9997076878105817
Accuracy on test set : 0.929107021131561
Recall on training set : 1.0
Recall on test set : 0.7871621621621622
Precision on training set : 0.9984
Precision on test set : 0.8503649635036497

```

- Stacking the models returned a reasonable performance on Recall(78%) and Precision(85%)

```
In [146... draw_cm(stacking_estimator,y_test)
```



Comparing all models

In [178...

```
comparison_frame = pd.DataFrame({'Model': ['bagging', 'bagging_tuned', 'treemodel', 'xgb_tuned', 'stacking_estimator'],
                                'Train_Recall': [0.96, 0.99, 1, 0.36, 1,
comparison_frame
```

Out [178...

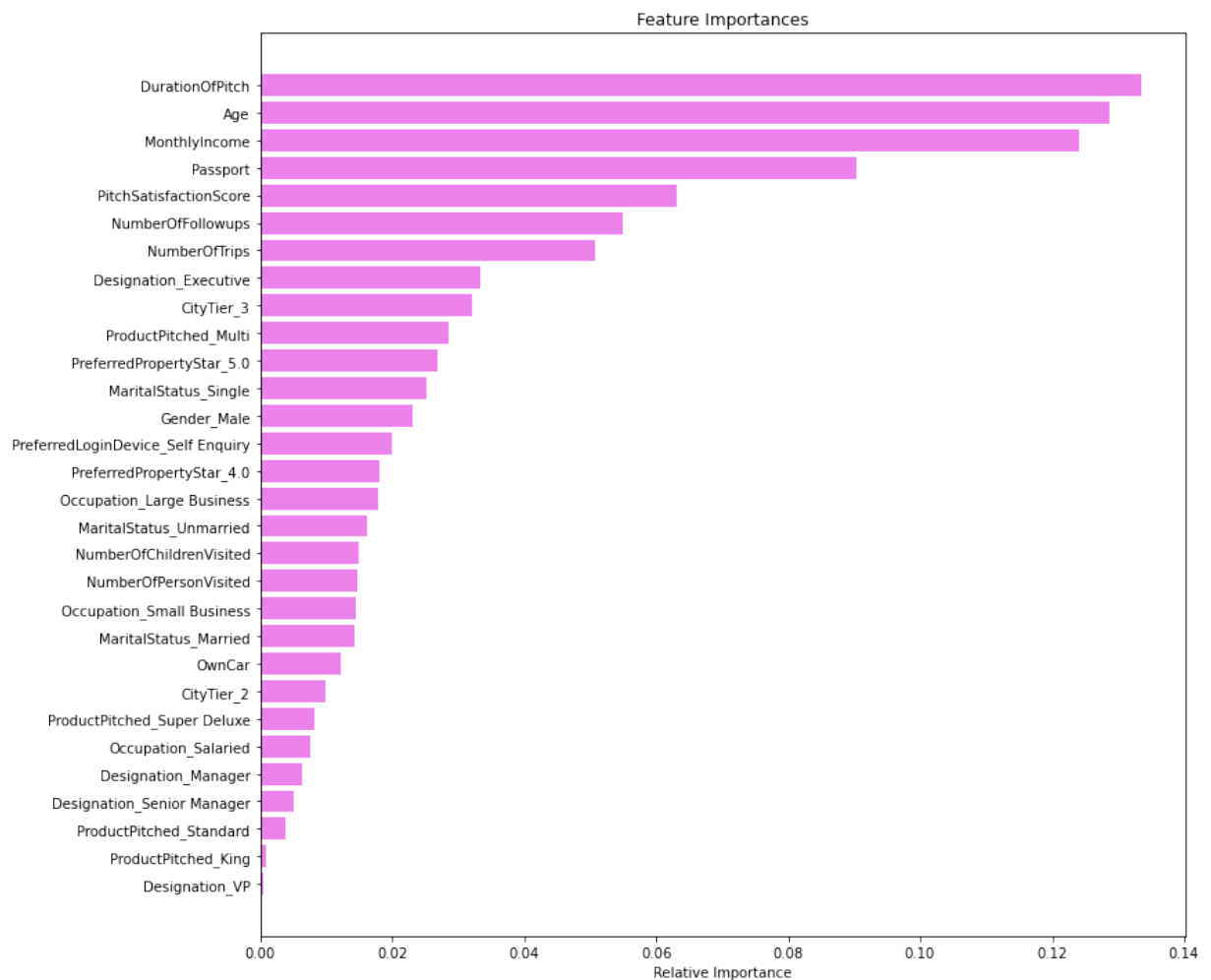
	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precisic
0	bagging	0.99	0.89	0.96	0.58	1.0
1	bagging_tuned	0.99	0.91	0.99	0.62	1.0
2	treemodel	1.00	0.90	1.00	0.76	1.0
3	tunedtree	0.83	0.81	0.36	0.31	0.5
4	rfmodel	1.00	0.90	1.00	0.57	1.0
5	rftuned	1.00	0.92	1.00	0.66	1.0
6	ab_classifier	0.85	0.83	0.32	0.27	0.7
7	ab_tuned	0.84	0.83	0.31	0.27	0.6
8	gb_model	0.89	0.85	0.49	0.35	0.8
9	gb_tuned	1.00	0.93	1.00	0.72	1.0
10	xgb	0.99	0.92	0.99	0.69	1.0
11	xgb_tuned	1.00	0.92	1.00	0.68	1.0
12	stacking_estimator	0.99	0.92	1.00	0.78	0.9

- Tuned Gradient Boost had the best overall performance, with the highest test accuracy of (93%), third best Test Recall of (72%), and the best Precision of (95%)
- Top 3 models were tuned random forest, tuned gradient boost, and stacking estimator with a similar overall performance
- Age and Monthly Income were common in models' strongest feature importance
- One improvement on overall models would be to try and see how combining Singles and Unmarried together would improve performance, as they both have the same meaning

In [169...

```
importances = gb_tuned.feature_importances_
indices = np.argsort(importances)
feature_names = list(x.columns)
```

```
plt.figure(figsize=(12,12))
plt.title('Feature Importances')
plt.barh(range(len(indices)), importances[indices], color='violet', align='center')
plt.yticks(range(len(indices)), [feature_names[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
```



Business Insights

- Further insights into our best model suggests that the duration of the pitch given by the sales team is a key factor that the company should focus on to drive in more purchases
- It is also important to consider the age and monthly income to help decide which product would be ideal to drive the interest
- Since the duration of pitch is considered the main factor, the company should also focus feedback and pitch satisfaction scores so that they can further improve pitching the products, this is proved by our model
- It was shown from EDA analysis, that preferred star ratings of 4.0 & 5.0 were not popularly taken as properties, and our model suggests that these two are significant features to be considered. Hence, the marketing team could study on pitching these properties better and could bring more sales withing those ratings