Travel Package Purchase Prediction

Anisah Inua Mohammed

Description

"Visit with us" travel company wants to retain its customers for a longer time period by launching a long-term travel package. The company had launched a holiday package last year and 18% of the customers purchased that package however, the marketing cost was quite high because customers were contacted at random without looking at the available information.

Now again the company is planning to launch a new product i.e. a long term travel package, but this time company wants to utilize previously available data to reduce the marketing cost.

Objective

- Explore and visualize the dataset.
- To predict which customers will purchase the long term travel package
- Which variables are most significant.
- Build Models using Bagging & Boosting to predict whether a person will take travel package or not
- Generate a set of insights and recommendations that will help the business to understand which segment of customers should be targeted more.

Data Set

Customer details:

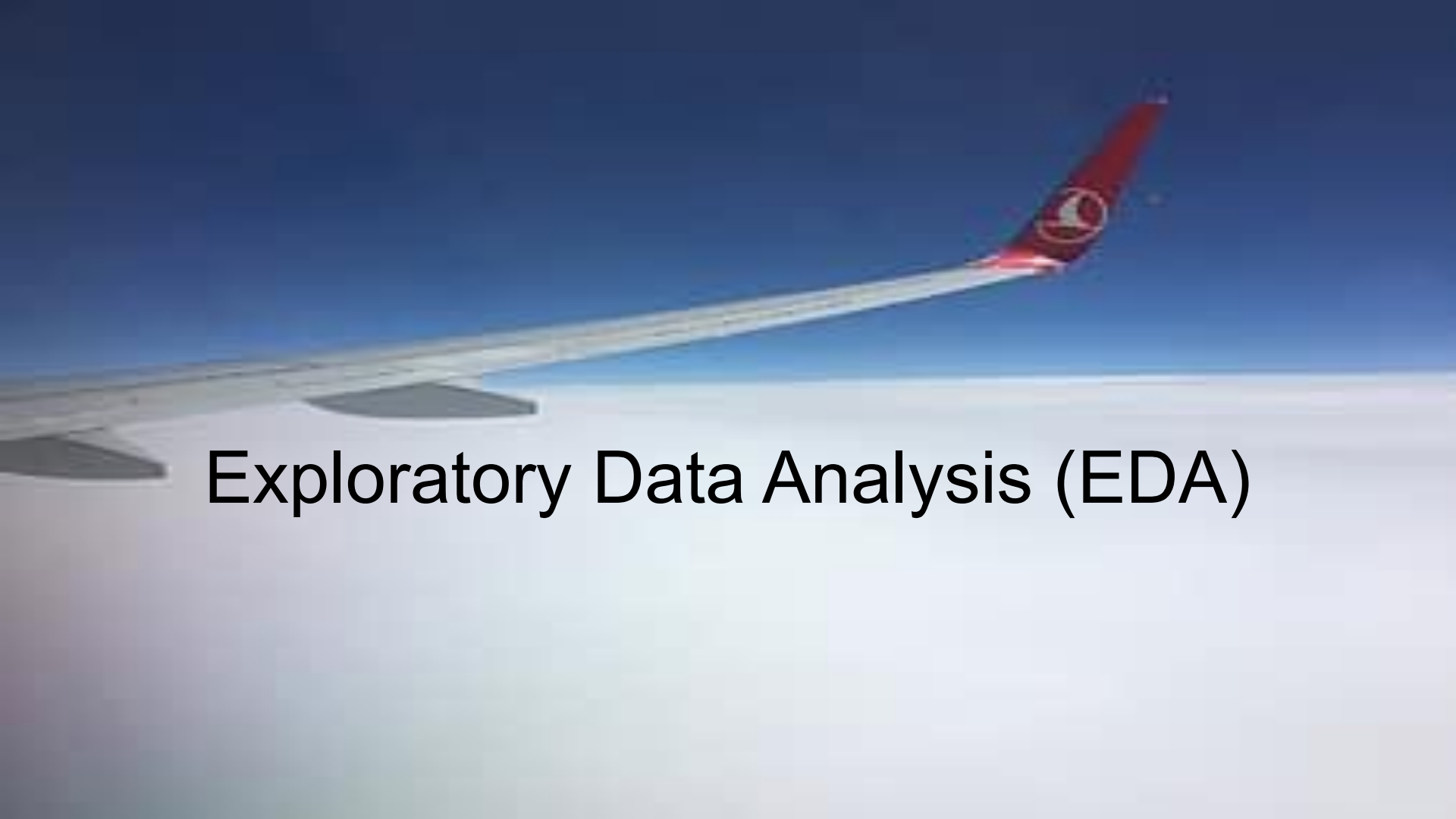
- CustomerID: Unique customer ID
- ProdTaken: Product taken flag
- Age: Age of customer
- PreferredLoginDevice: Preferred login device of the customer in last month
- CityTier: City tier
- Occupation: Occupation of customer
- Gender: Gender of customer
- NumberOfPersonVisited: Total number of person came with customer
- PreferredPropertyStar: Preferred hotel property rating by customer
- MaritalStatus: Marital status of customer
- NumberOfTrips: Average number of the trip in a year by customer
- Passport: Customer passport flag
- OwnCar: Customers owns a car flag
- NumberOfChildrenVisited: Total number of children visit with customer
- Designation: Designation of the customer in the current organization
- MonthlyIncome: Gross monthly income of the customer

Customer interaction data:

- PitchSatisfactionScore: Sales pitch satisfactory score
- ProductPitched: Product pitched by a salesperson
- NumberOfFollowups: Total number of follow up has been done by sales person after sales pitch
- DurationOfPitch: Duration of the pitch by a salesman to customer

Initial Observations

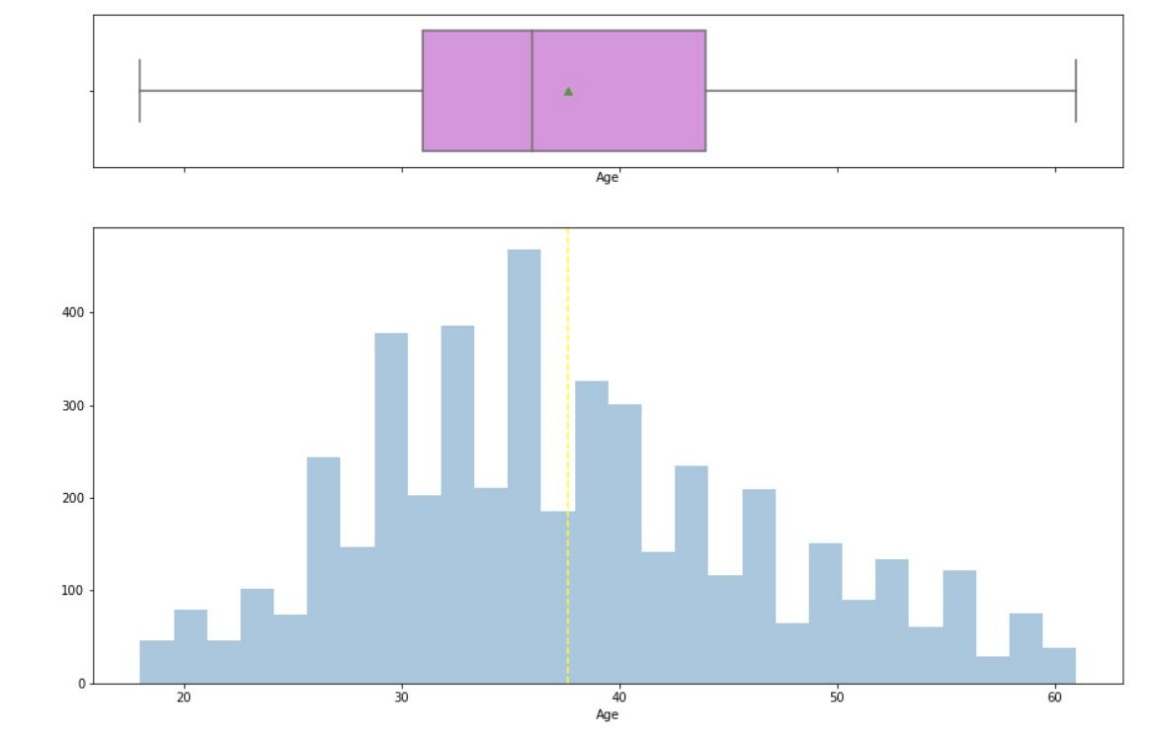
- Most of the customers used self inquiry as type of contact
- Most customers have a "salaried" occupation
- Most customers used in this dataset are Male
- Most customers are married
- The customers have not used the travel package from the old campaign
- The preferred property for most customers is 3 star properties
- The highest number of trips taken yearly by the customers is 2
- There are unique values in the dataset
- There are more customers which are Male and salaried occupation
- There is very low number of trip values for the younger ages
- In the gender category this might be an error so i would suggest combining the Fe Male with the Female category



Exploratory Data Analysis (EDA)

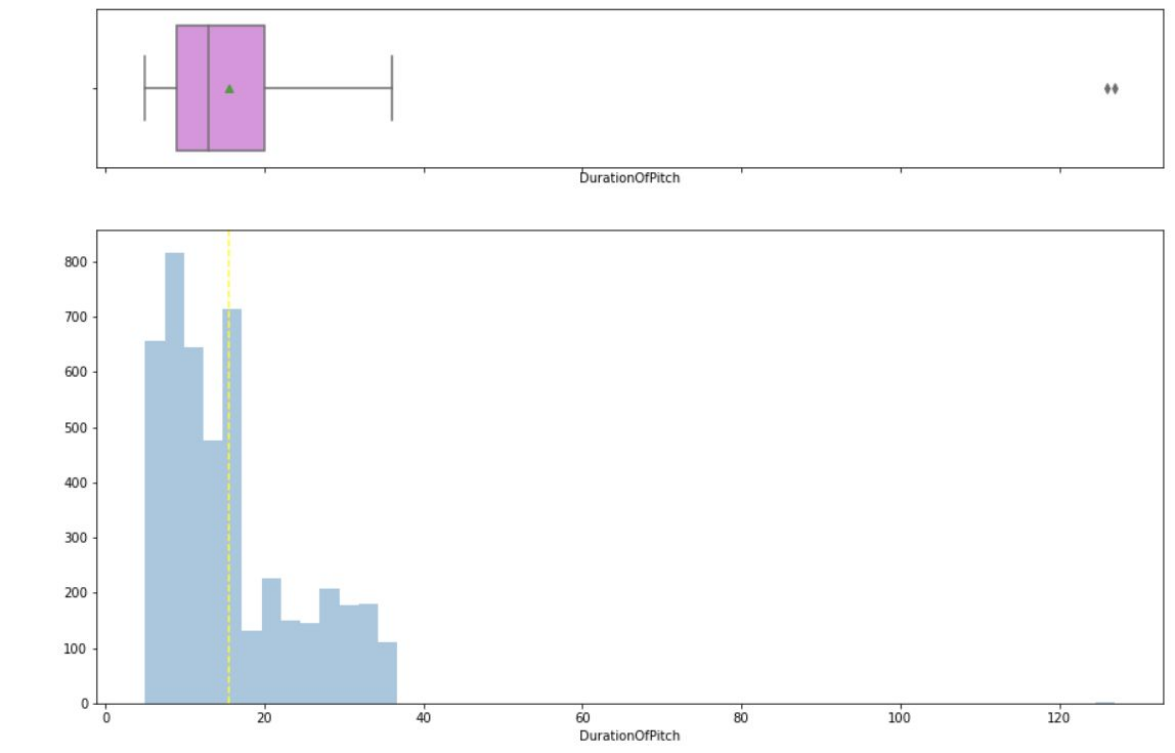
Age

From the graph it reveals that Age is normally distributed and most of the customers age is greater than 30



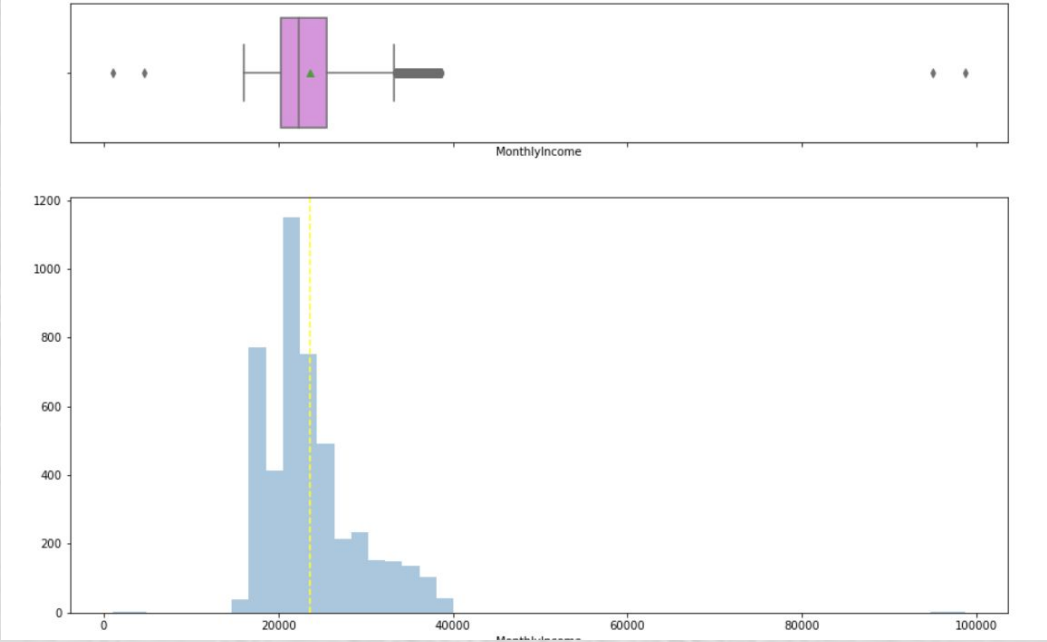
Duration of Pitch

The graph shows that the distribution is left skewed which displays presence of outliers.



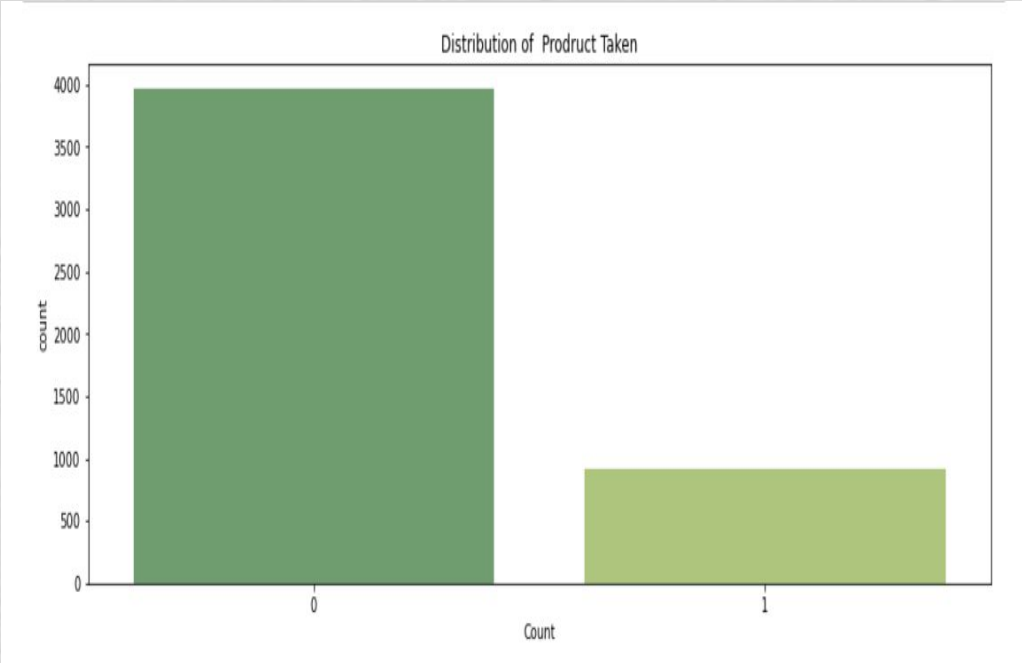
Monthly Income

This shows that most of the customers are earning 20k to 30k, where as 10% of the customers are earning around 40k per month. This also shows that there a few outliers representing very few high and low income.



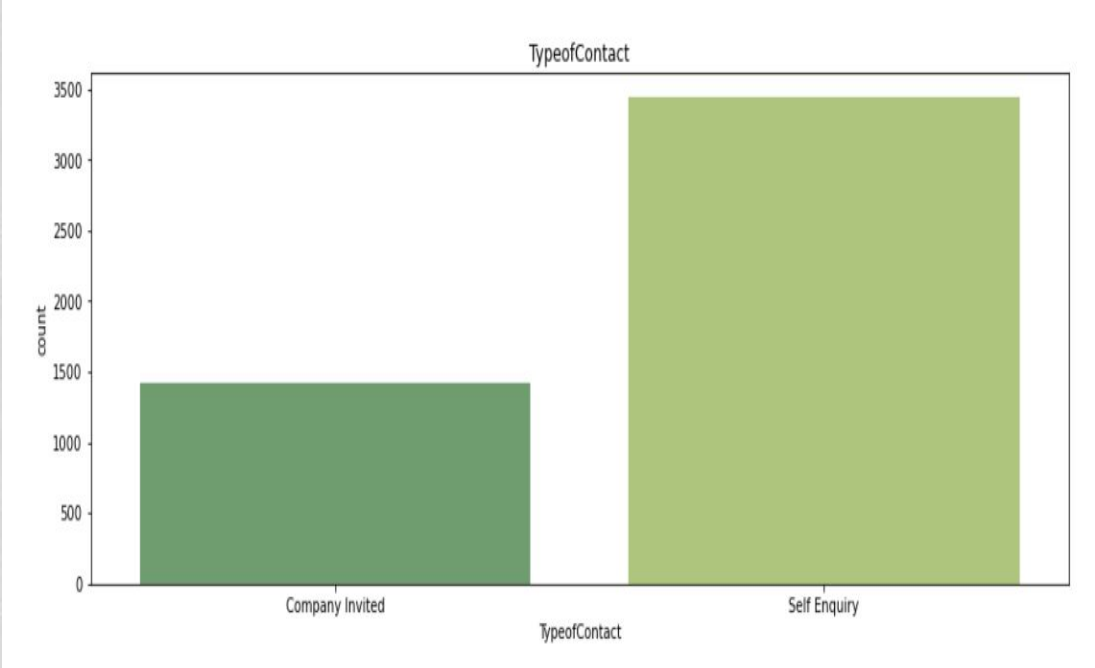
Product Taken

- 18.8% of the records have acquired the holiday package
- However the other 80% of customers did not opt for any travel package.



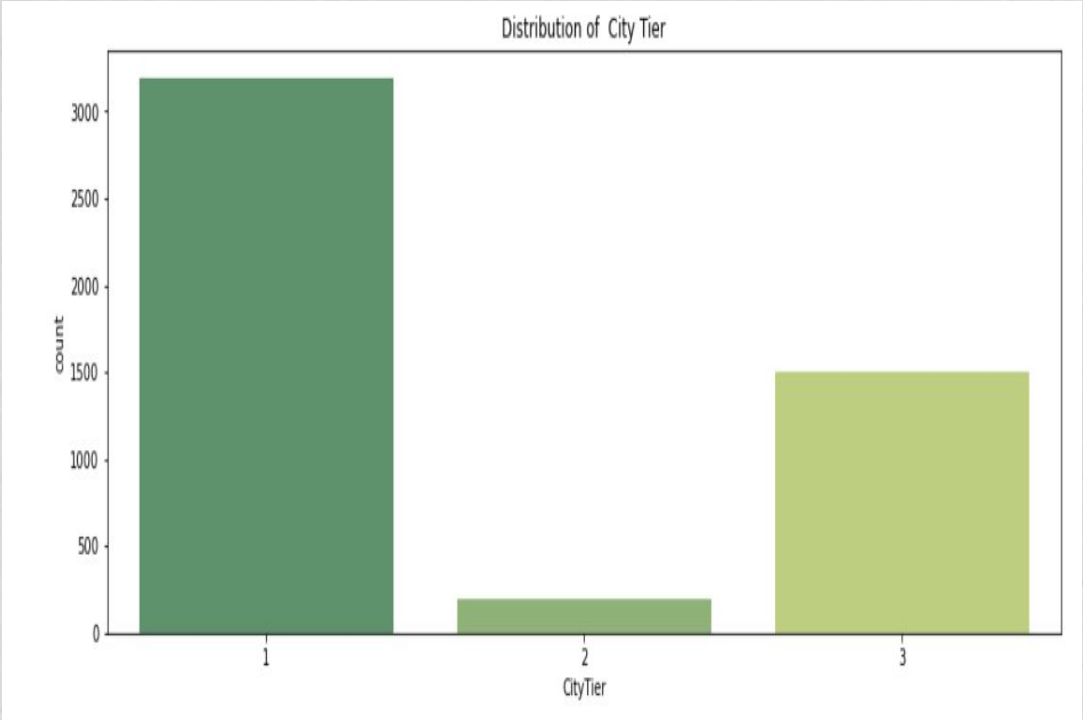
Type of Contract

- About 70% of the customers have reached the company through "Self-Enquiry"
- While the other 30% of customers have known the company through "Company Invites"



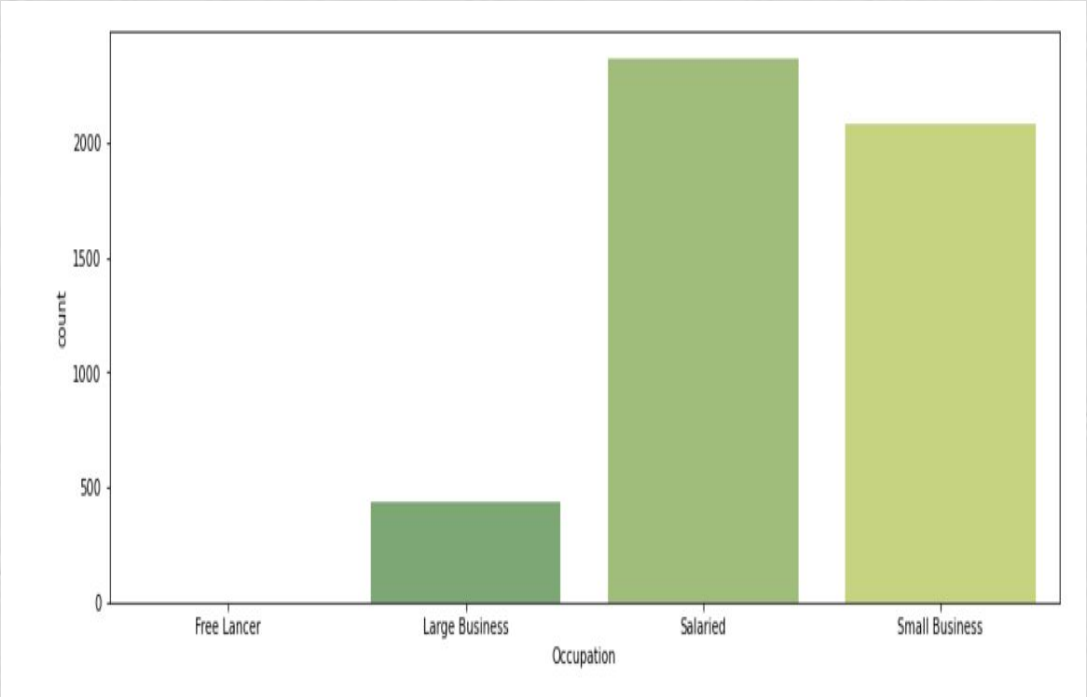
City Tier

- 65% of the customers are from Tier-1 cities and 30% of the customers from Tier 3 cities.



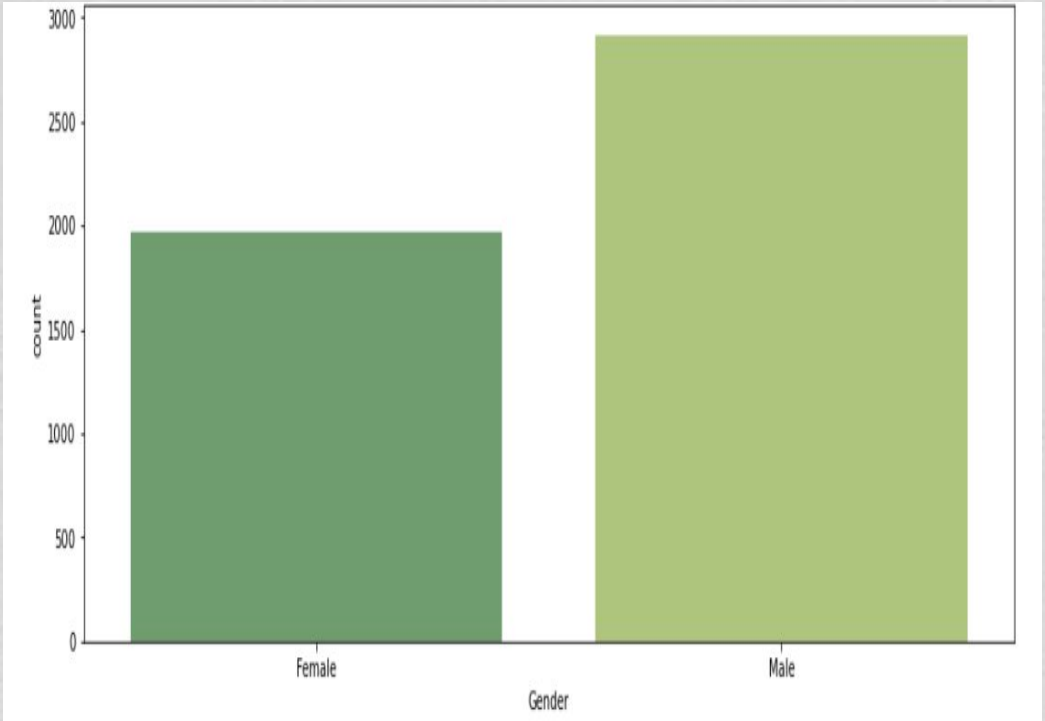
Occupation

- Almost up to half of the customers are salaried
- While the other customers are either running large or small businesses
- This can also that there is presence of outliers for Freelancer



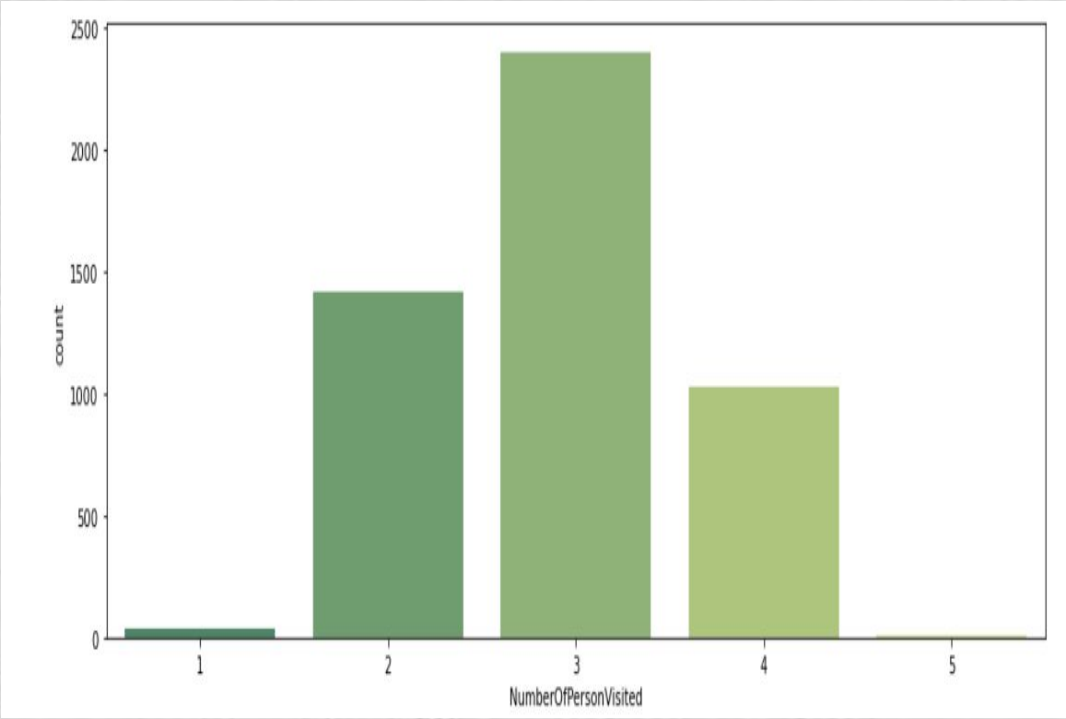
Gender

- 60% of the customers are Male and rest of the 40% are female



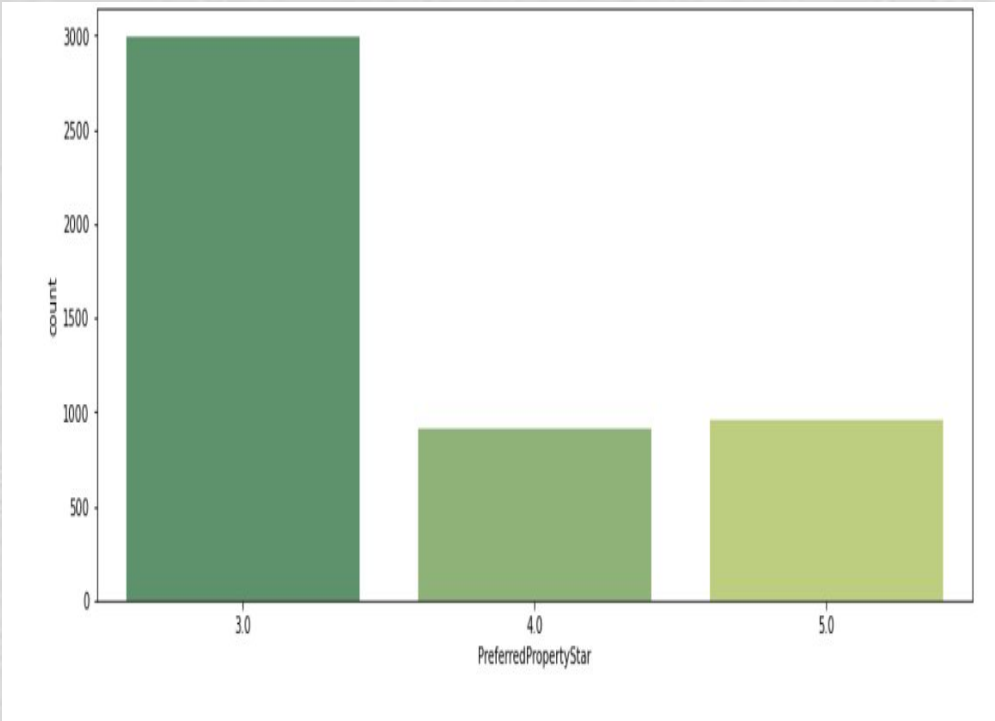
Number of Person Visited

- Most of the times 3 to 2 persons visited
- only 21% of the times 4 persons visited



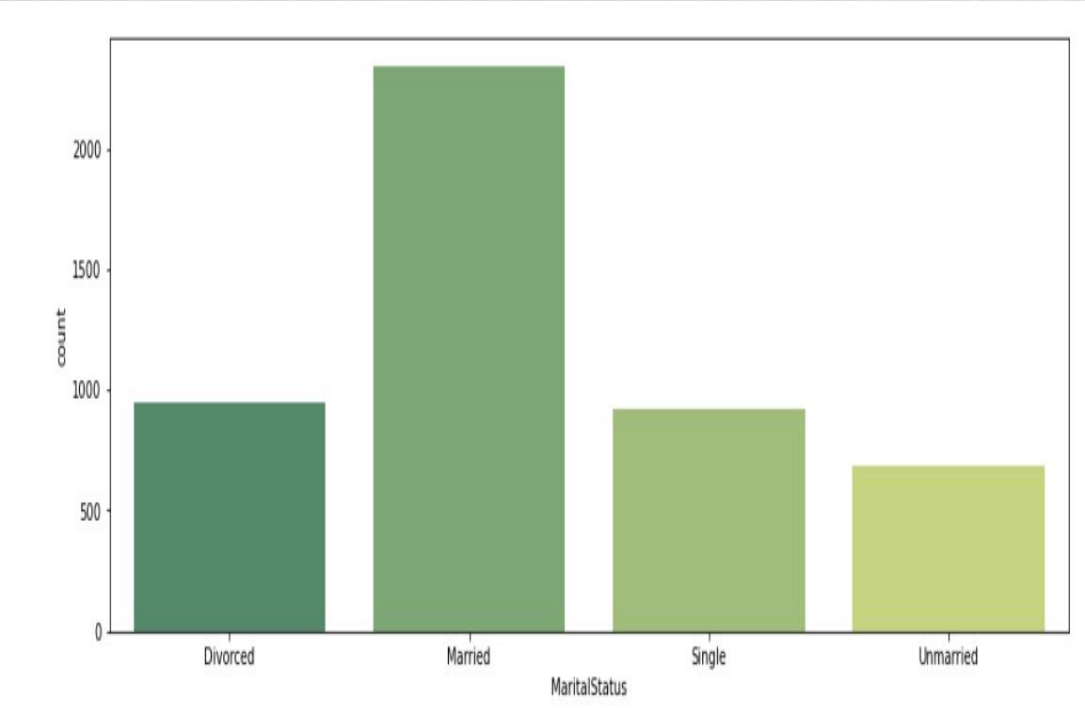
Preferred Property Star

- Most of the customers prefer a 3 star property
- This also shows that customers are selecting only the hotels which are rated above 3 stars



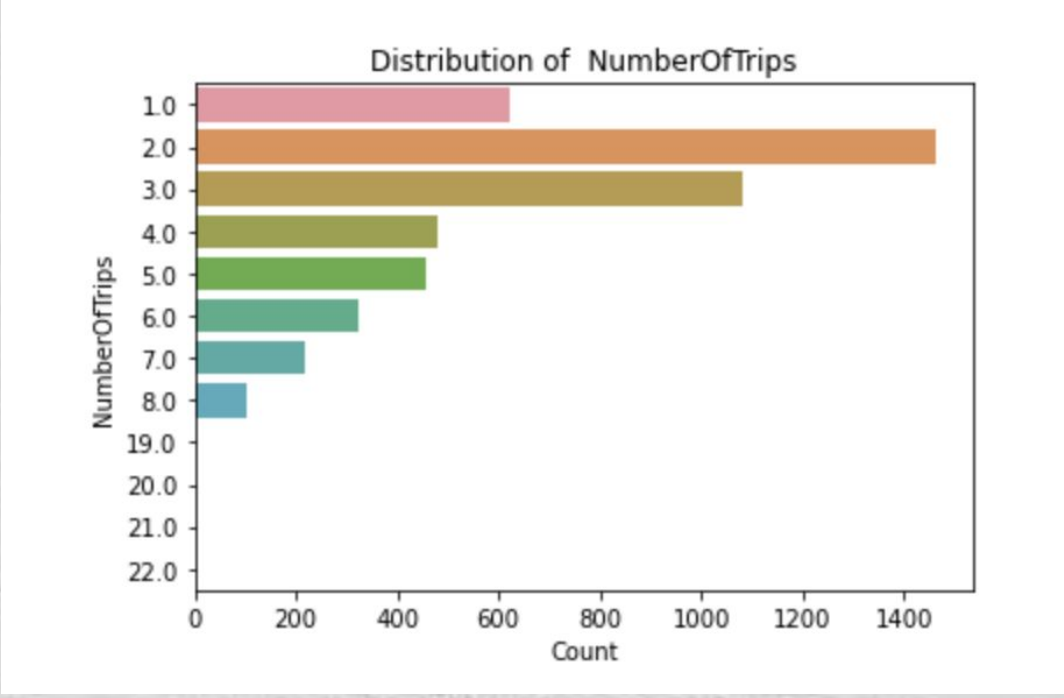
Marital Status

- Majority of the customers are married followed
- 38% of the customers are Divorced/Single
- While only 14% of the customers are Unmarried



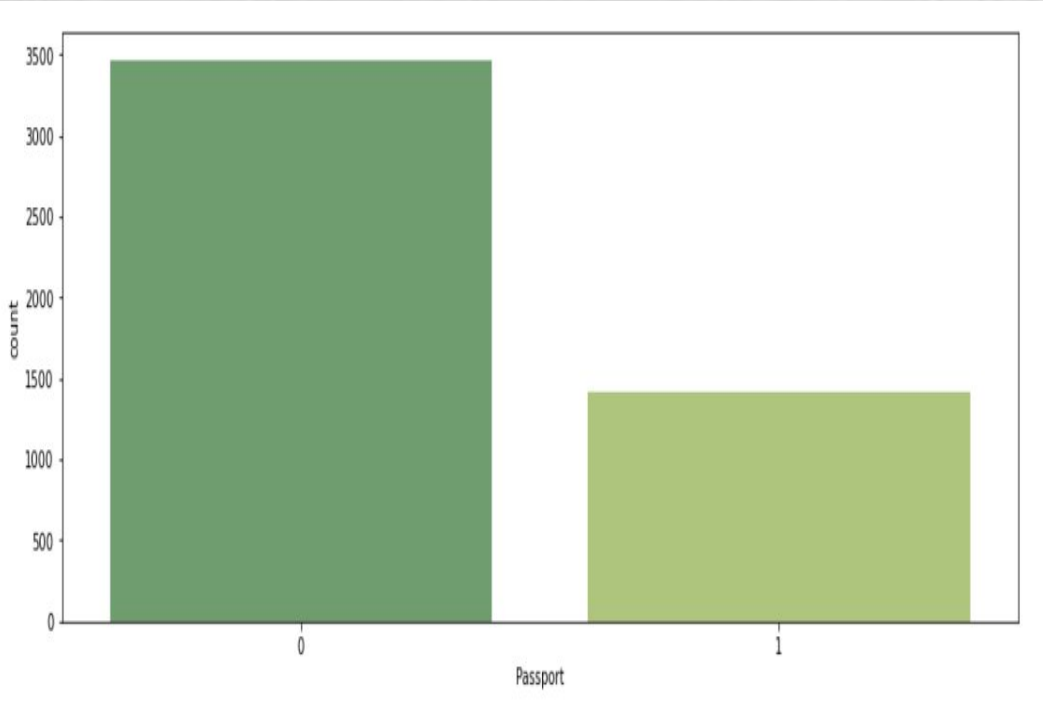
Number of Trips

- Most of the customers take 2 trips per year
- 52% of the customers have made 2 to 3 trips and there are 6% of the customers who made 7 to 8 trips



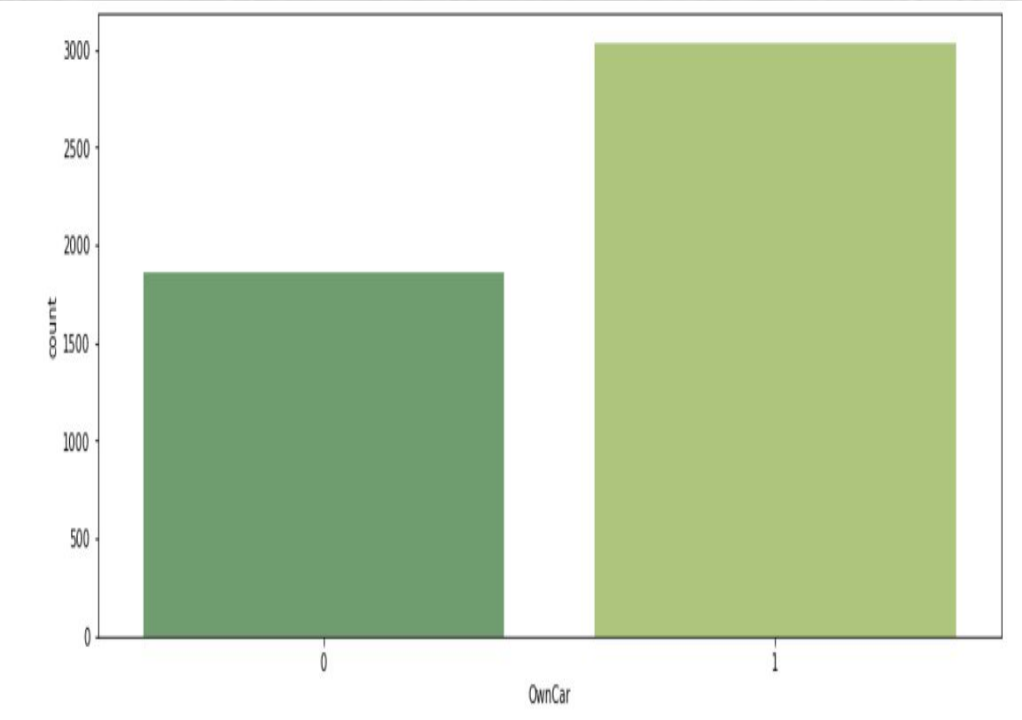
Passport

- Majority of the customers do not own passport



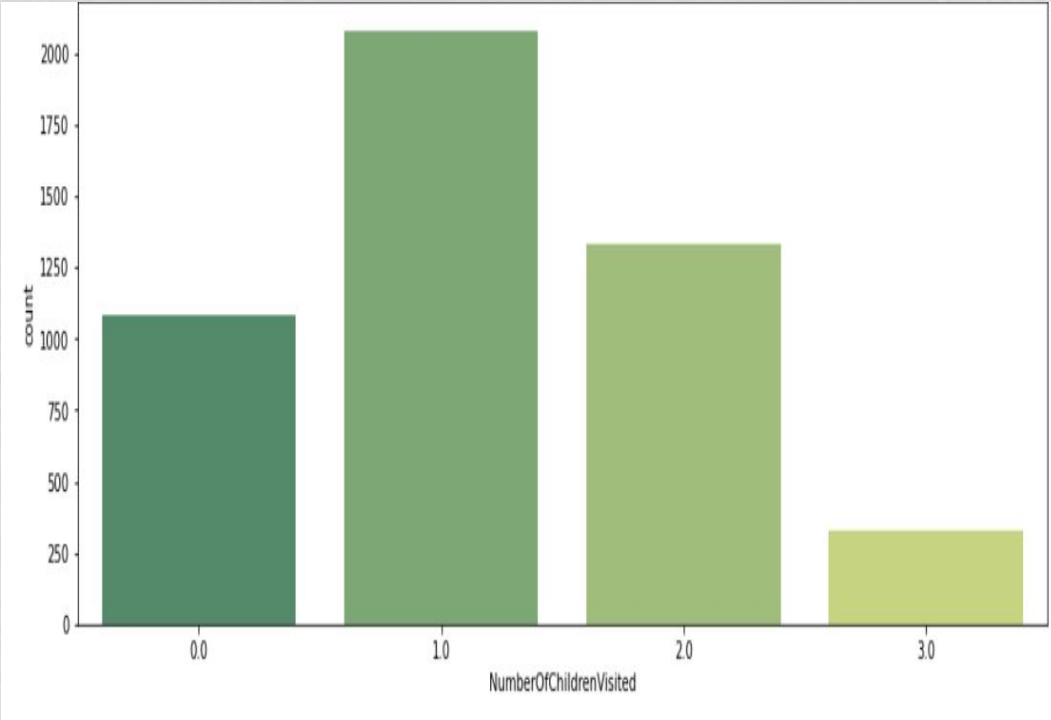
Own Car

- Most of the customers are owners of a car



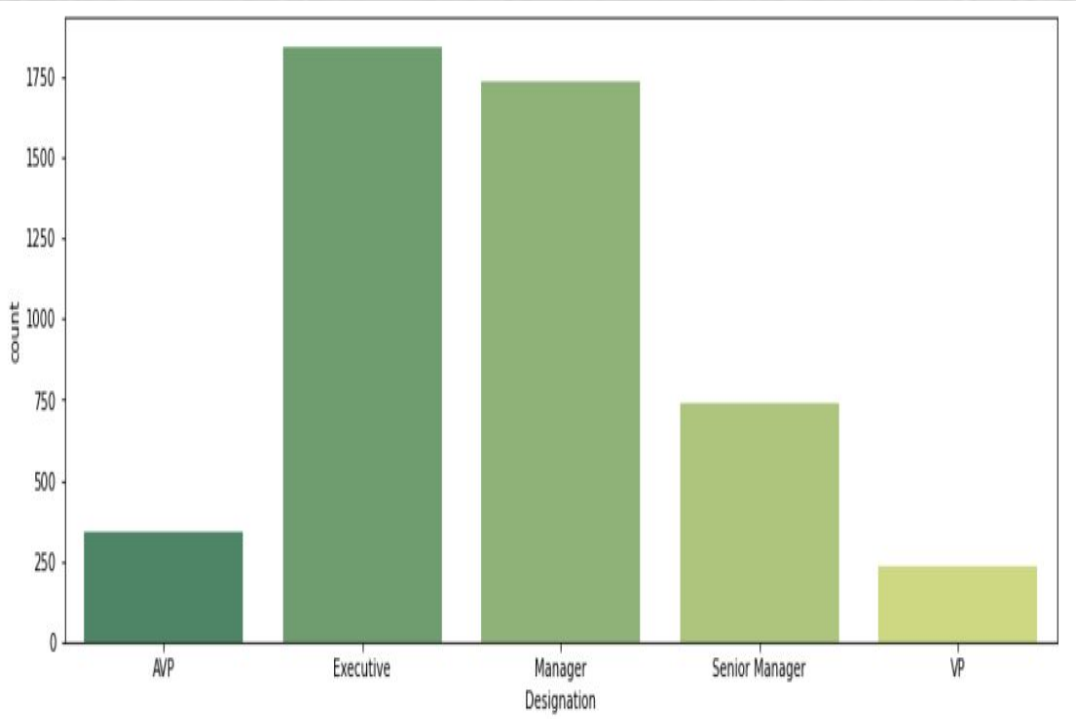
Number of Children Visited

- Most of the customers had only one child visit with them



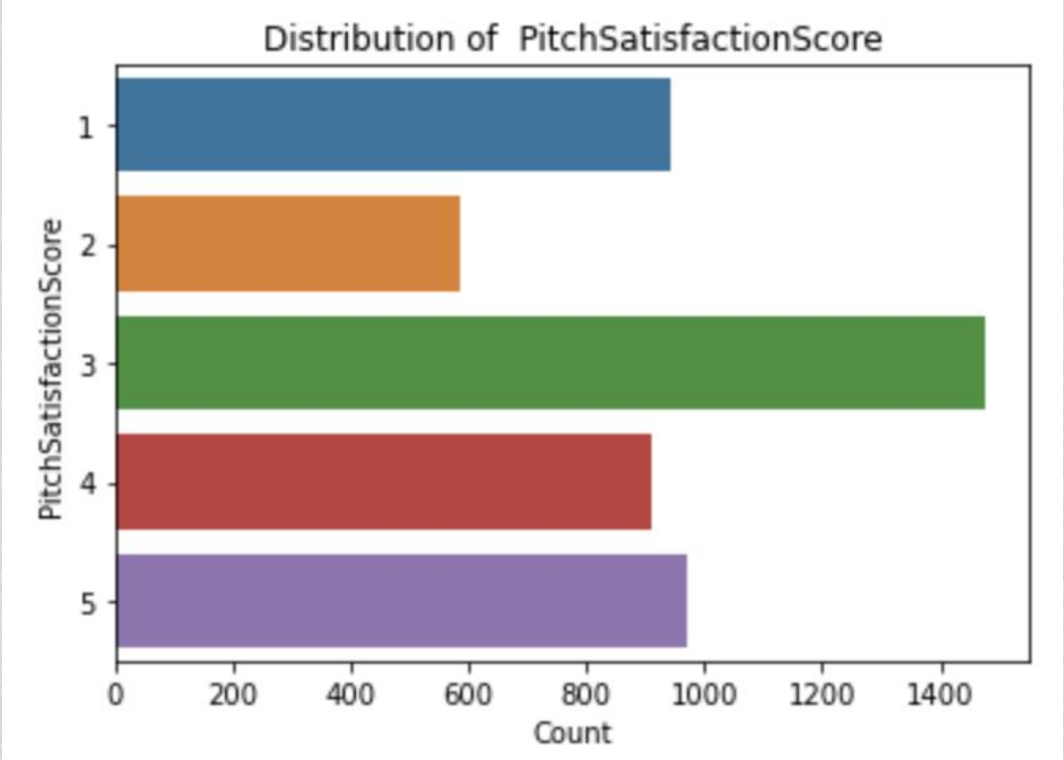
Designation

- Most of the customers at Manager/Executive level in their occupation and rest of the customers at higher positions



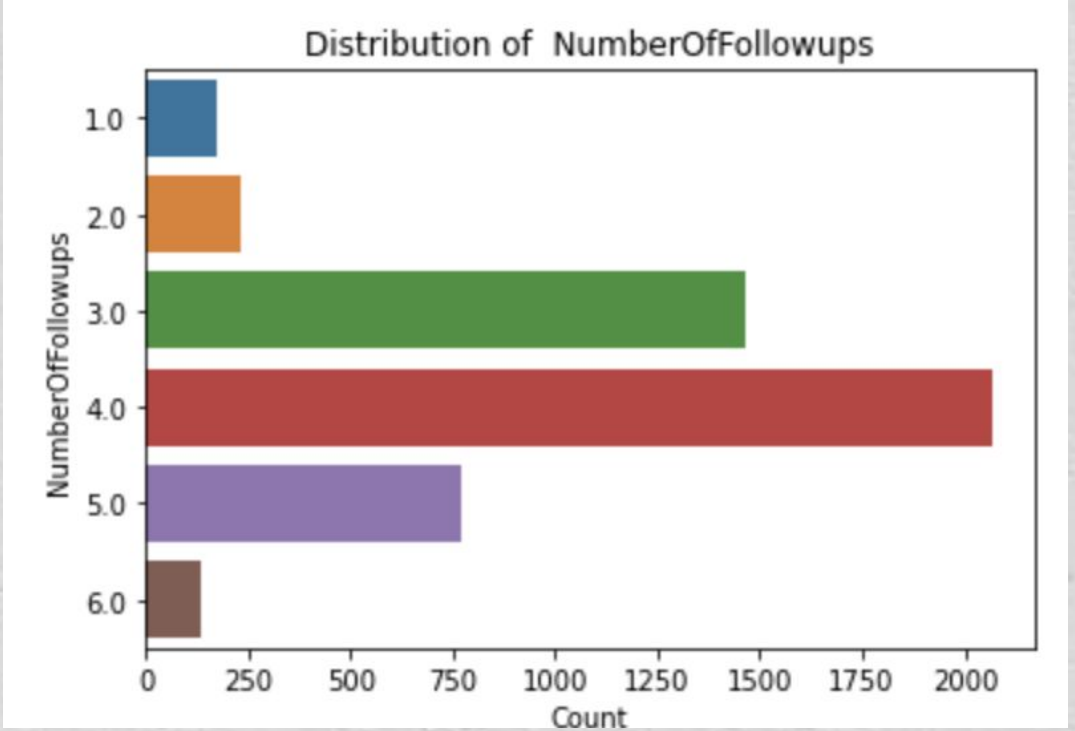
Pitch Satisfaction Score

- Most of the Pitch Satisfaction score indicate a score of 3



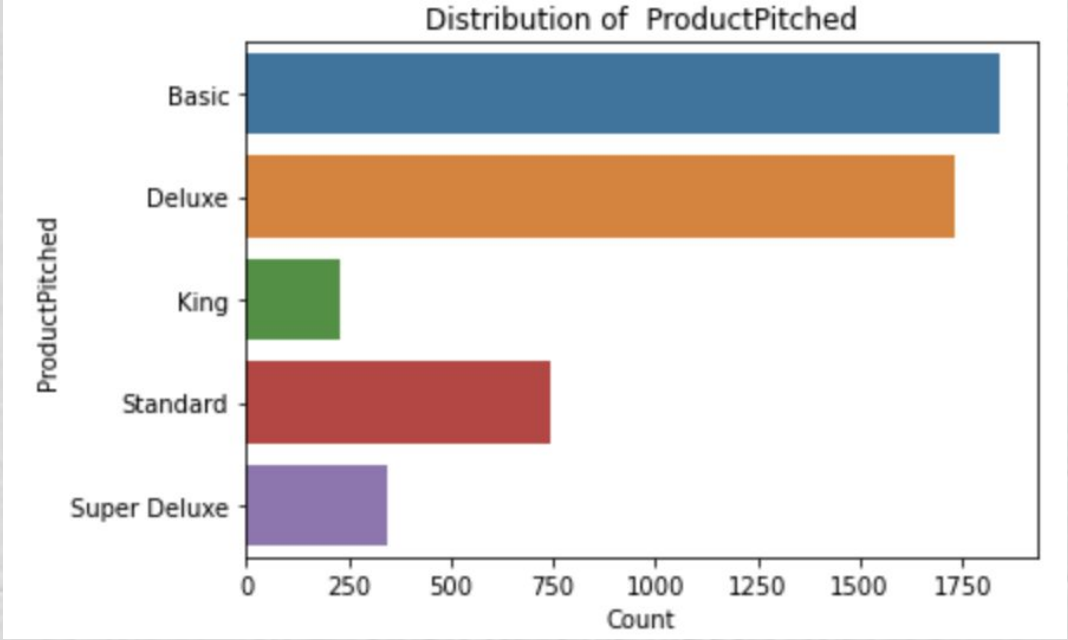
Number of Follow Ups

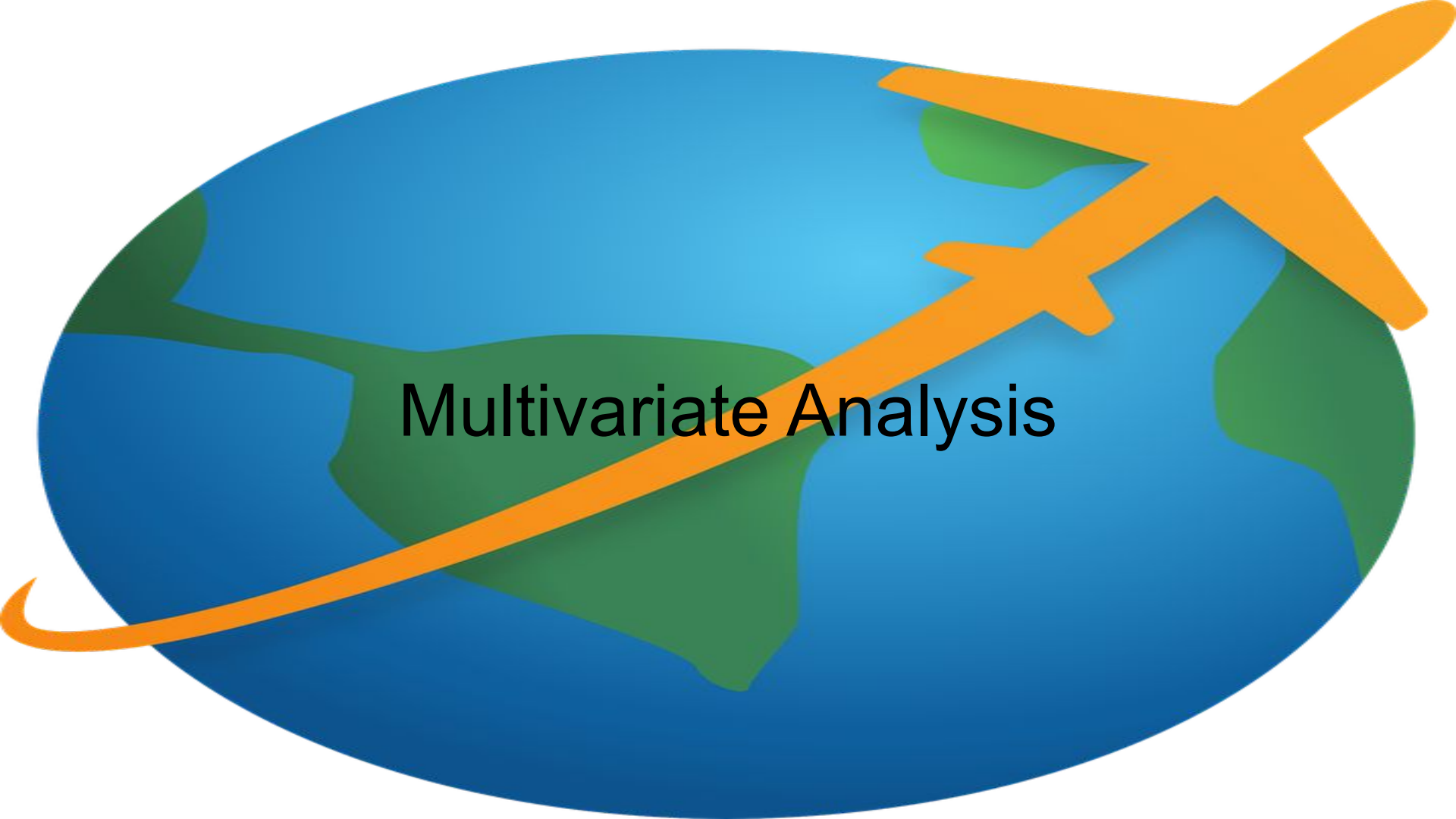
- 4 has been the most number of follow ups made having percentage of 42.3%



Product Pitched

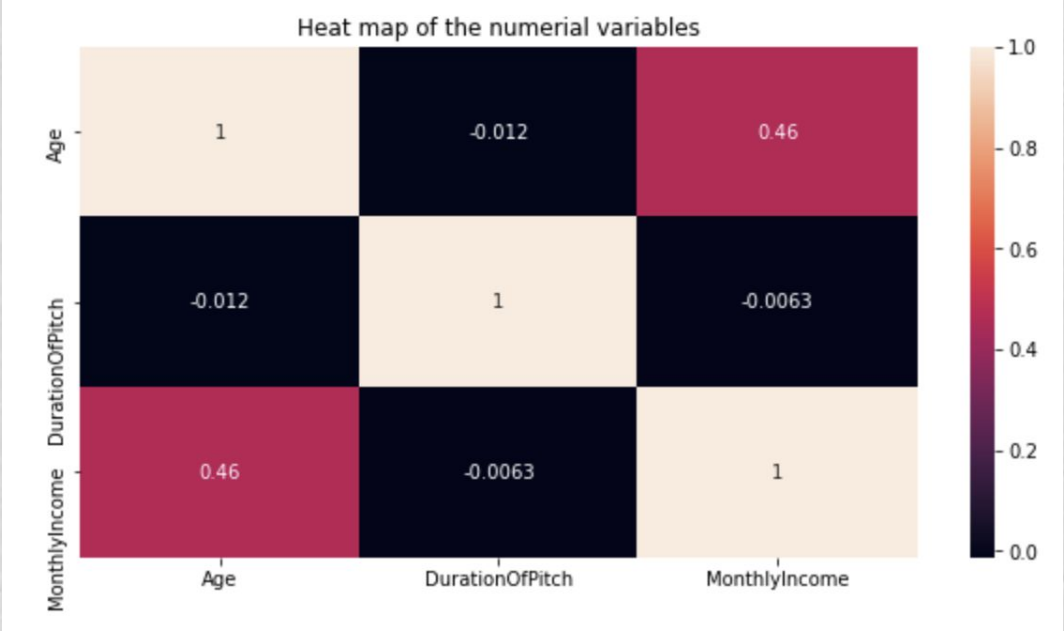
- Basic and Deluxe products were pitched most of the times




Multivariate Analysis

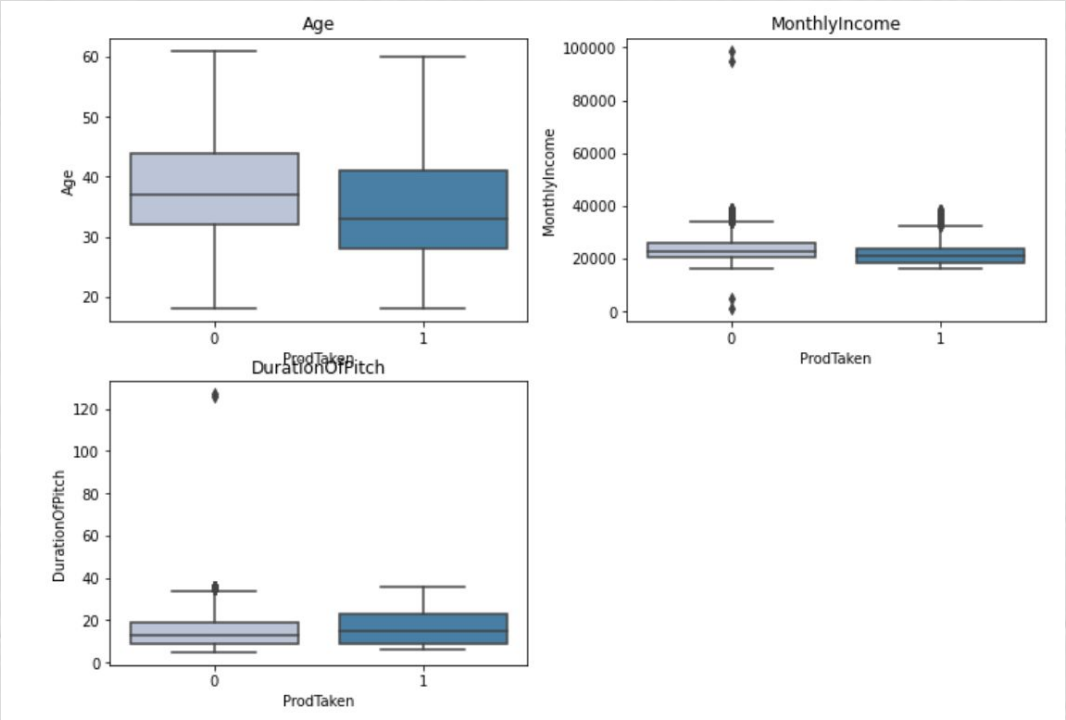
Heatmap of Numerical Values

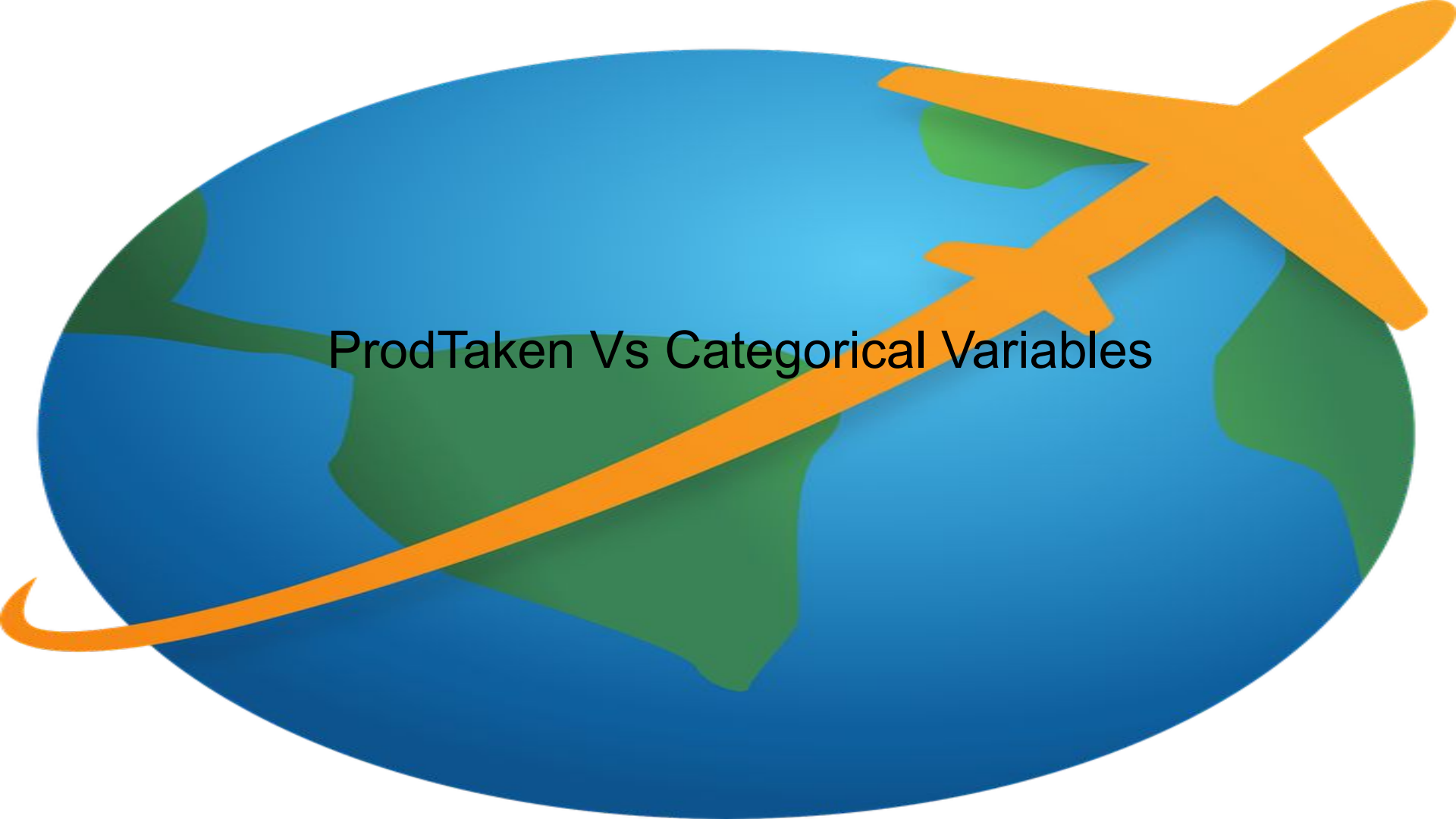
- The heatmap reveals that Age and MonthlyIncome are correlated which is makes sense.

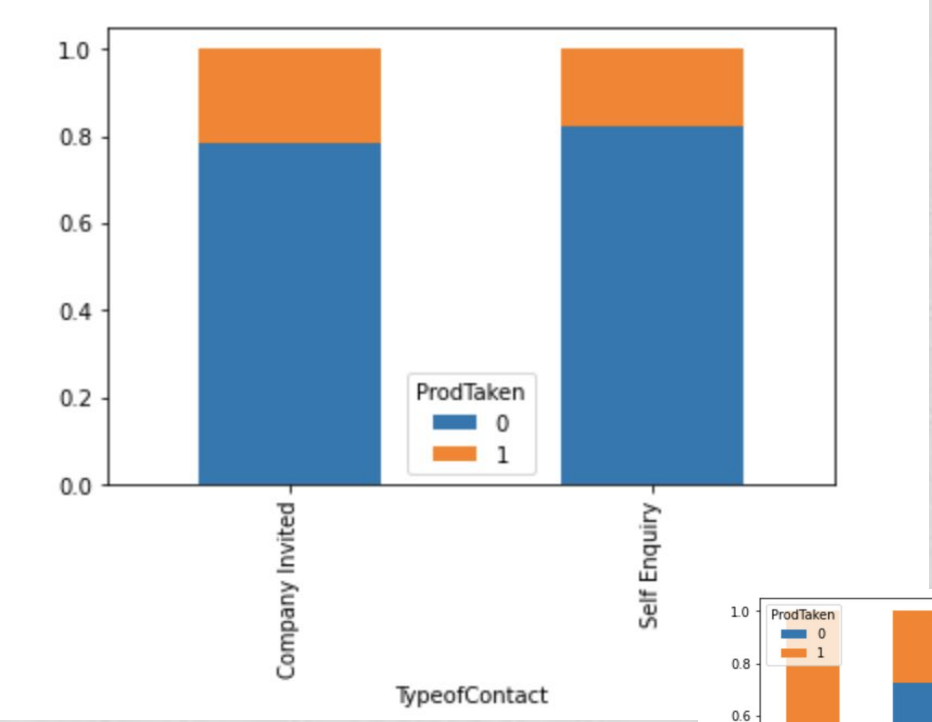


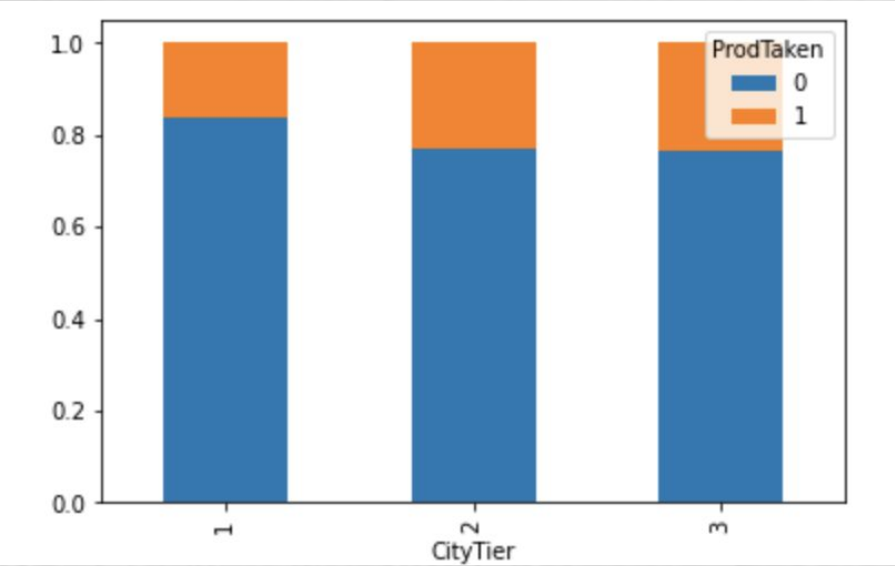
Bivariate Analysis of ProdTaken with continuous variables

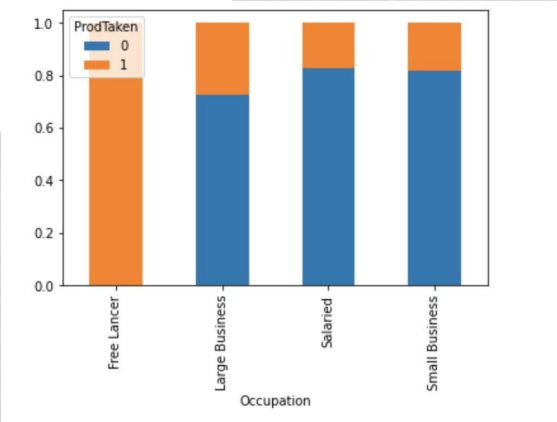
- There is not much difference in the monthly income of customers who have and have not taken the package
- The holiday package has been taken mostly by customers who are of less than 40 years of age
- The that package has been taken when there is higher duration of pitch by salesman to the customer

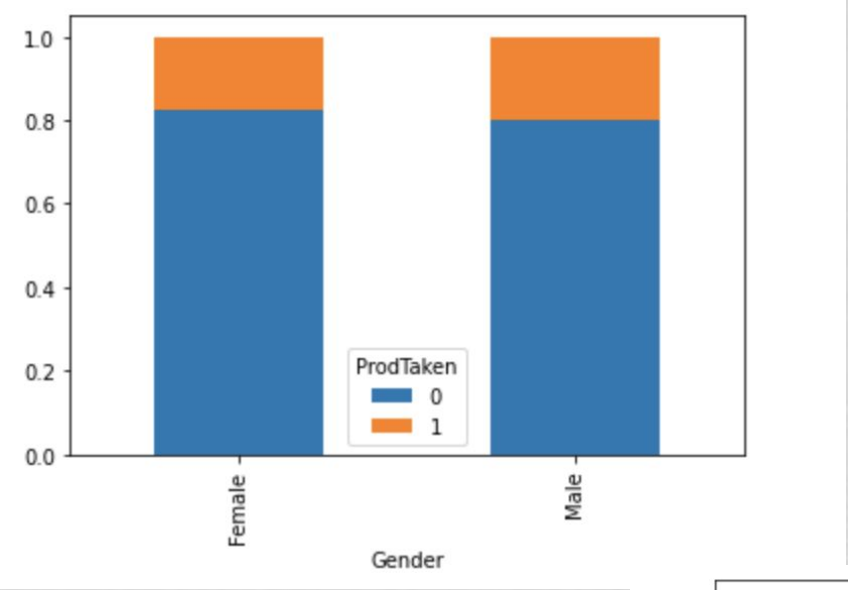


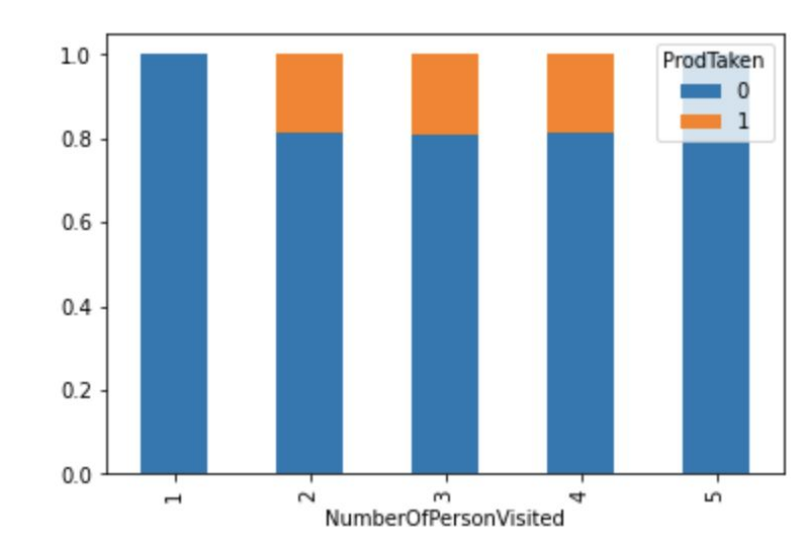

ProdTaken Vs Categorical Variables

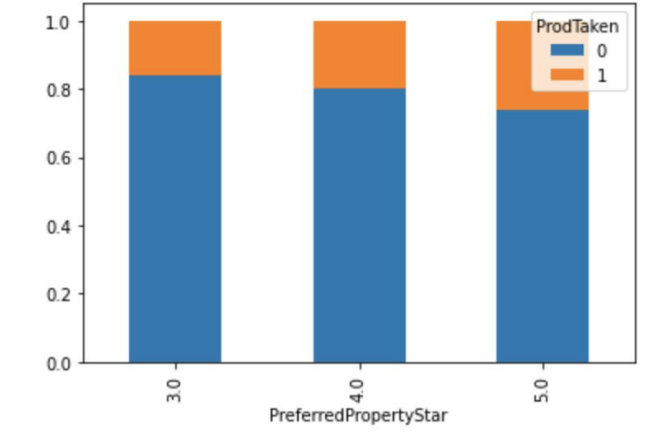

1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
Company Invited
Self Enquiry
TypeofContact

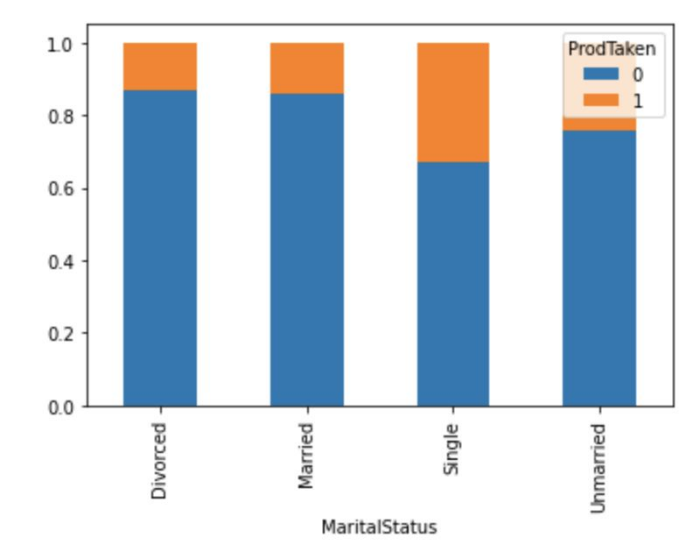

1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
1
2
3
CityTier

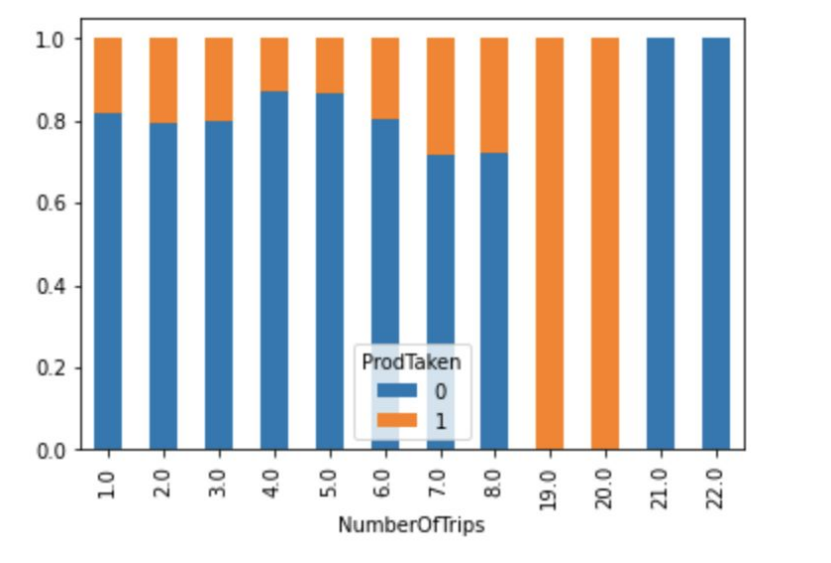

1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
Free Lancer
Large Business
Salaried
Small Business
Occupation

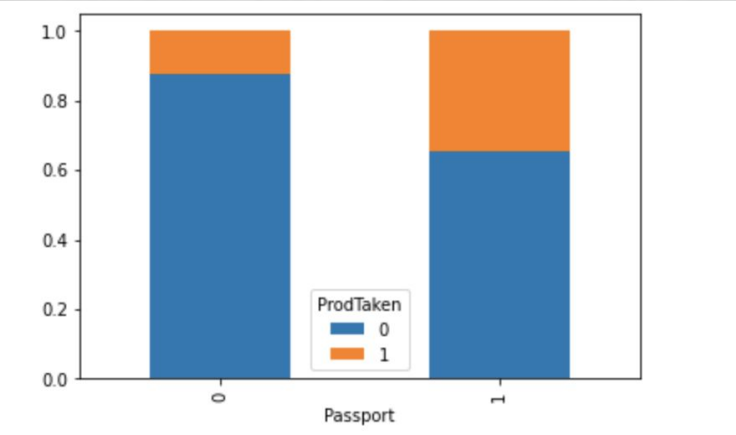

1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
Female
Male
Gender

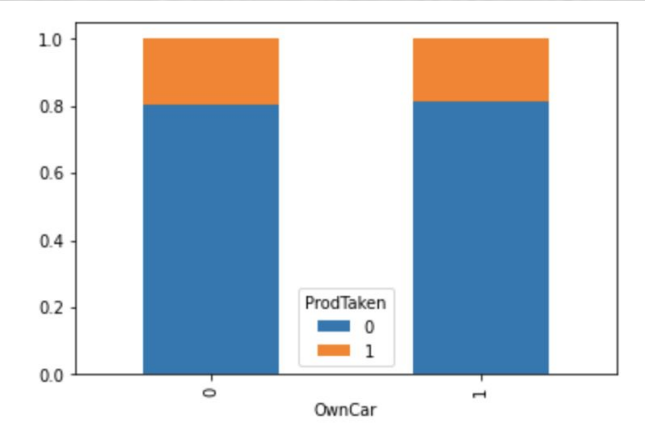

1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
1
2
3
4
5
NumberOfPersonVisited

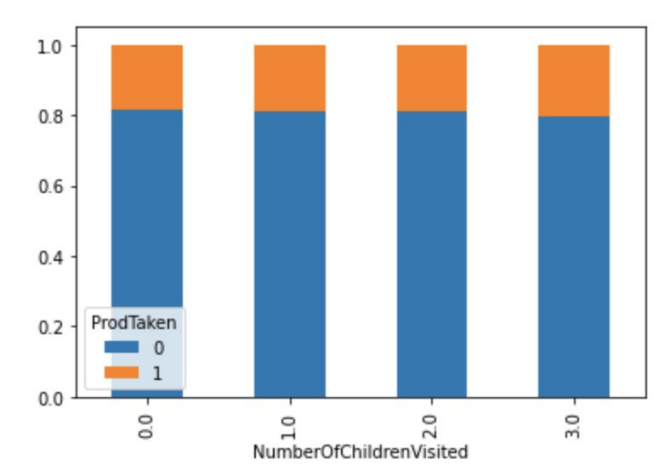

1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
3.0
4.0
5.0
PreferredPropertyStar

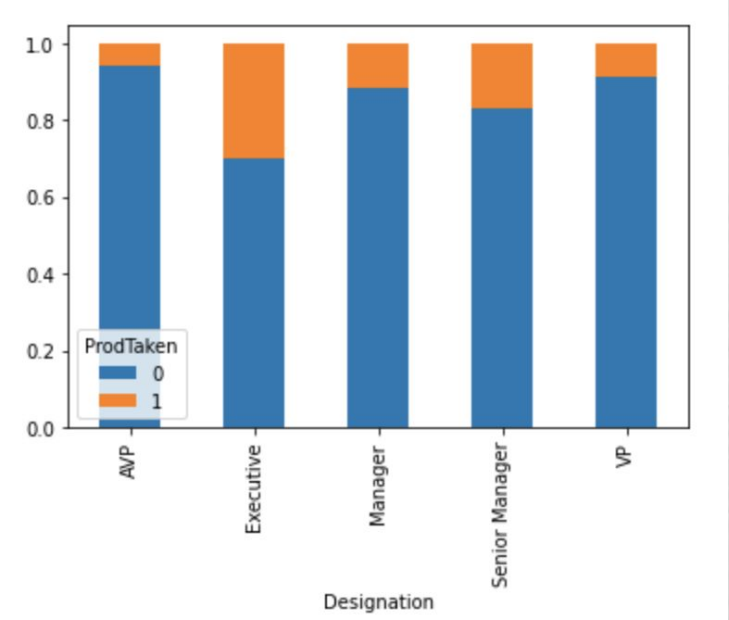

ProdTaken
0
1
1.0
0.8
0.6
0.4
0.2
0.0
Divorced
Married
Single
Unmarried
MaritalStatus

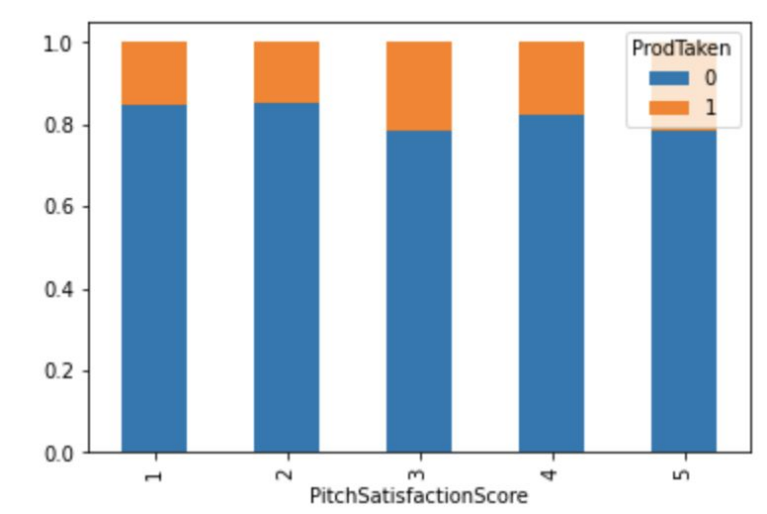

ProdTaken
0
1
1.0
0.8
0.6
0.4
0.2
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
19.0
20.0
21.0
22.0
NumberOfTrips

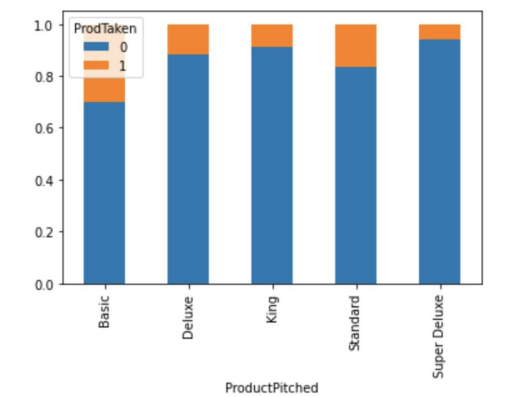

ProdTaken
0
1
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Passport

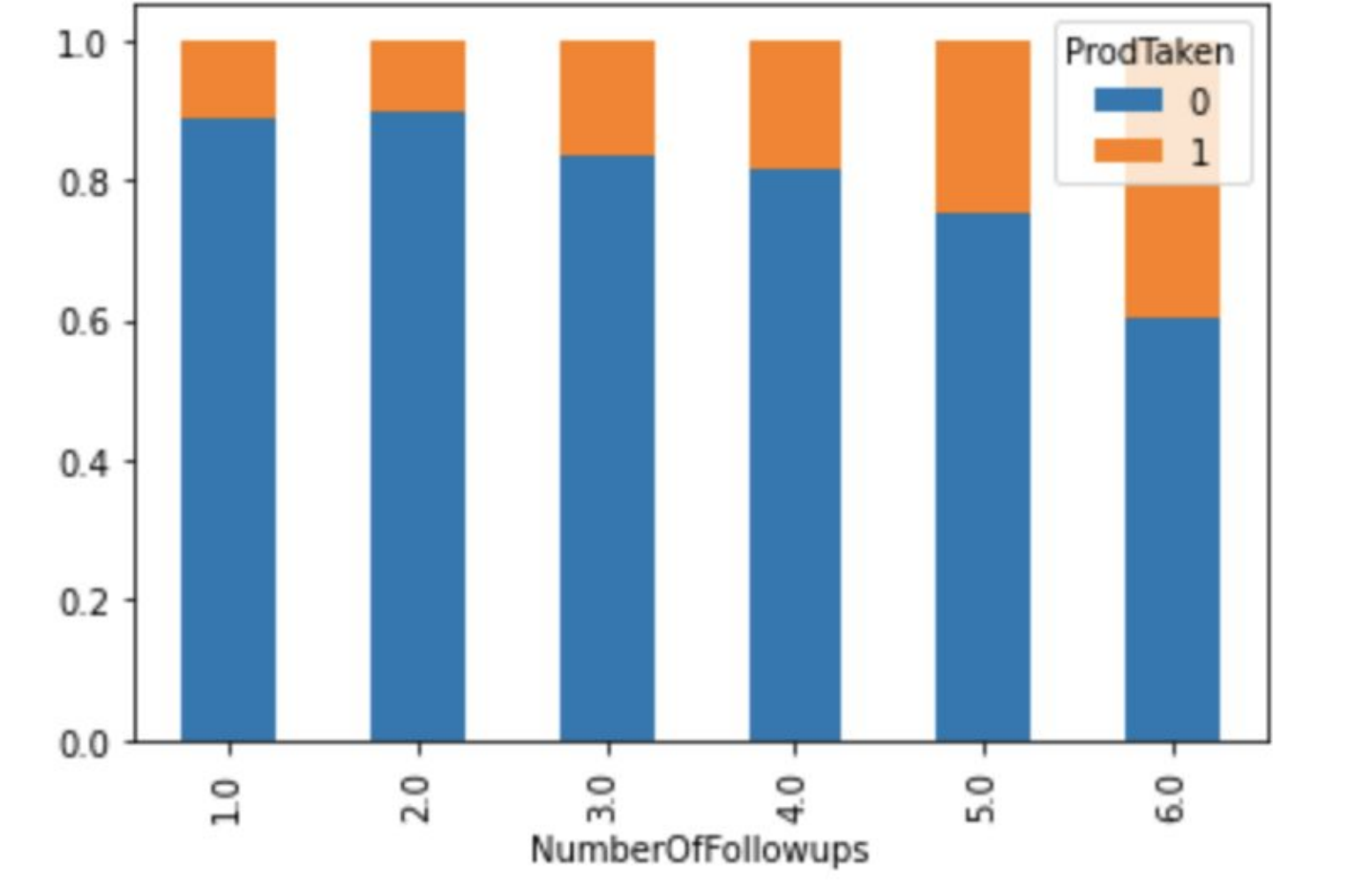

ProdTaken
0
1
1.0
0.8
0.6
0.4
0.2
0.0
0
1
OwnCar


1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
0.0
1.0
2.0
3.0
NumberOfChildrenVisited


1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
AVP
Executive
Manager
Senior Manager
VP
Designation


1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
1
2
3
4
5
PitchSatisfactionScore


1.0
0.8
0.6
0.4
0.2
0.0
ProdTaken
0
1
Basic
Deluxe
King
Standard
Super Deluxe
ProductPitched

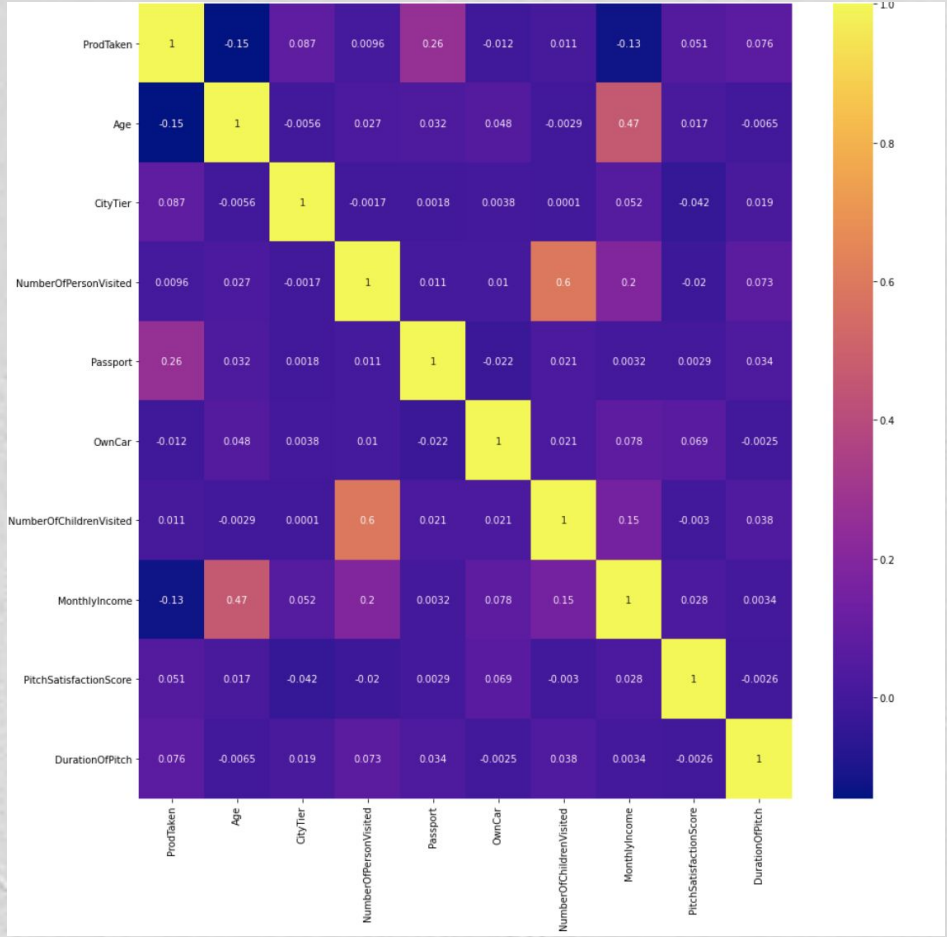

ProdTaken
0
1
1.0
0.8
0.6
0.4
0.2
0.0
1.0
2.0
3.0
4.0
5.0
6.0
NumberOfFollowups

Insights from the Bivariate Analysis

- This shows that the customers who took the travel package was those who came into company's contact as company invited
- The customers from city tier 2 and 3 have are most likely or mostly purchased the travel package
- This shows that all the customers who are "Free Lancers" have purchased the product and also customers that own or work for "Large Business" have also purchased the product.
- From the graph it reveals that there is not really much difference between the genders of the customers who have purchased the travel package
- Customers that are travelling with 2,3,4 persons have a higher chance of obtaining the travel package
- Whereas customers travelling with either 1 or 5 persons are less likely or did not acquire any travel package
- Customers who prefer 5 star rated property has brought the travel package more followed by 4 and 3 star preferred customers
- Customers who are single and unmarried people have purchased the travel package in comparison to the married or divorced customers
- There is not much difference in purchasing travel package between customers with and without a car
- There is no significant difference in the purchasing travel package for customers who had 0,1,2,3 children visiting with them.
- This shows that customers in "Executive" position are more likely or have purchased the travel package, followed by customers in the "senior manager" position.
- Those customers who have given a PitchSatisfactionScore of 3 or 5 have purchased the travel package more
- The package was purchased by those customers to whom the Basic product was pitched, followed by standard product
- The chance of purchasing increased as the number of followups went up with number of followups of 6 showing good chance of the customer purchasing the package.
- Customers that take 7 or 8 trips per year has more chace of pacquiring the travel package.Also those those travel 2 or times per year has about 20% chance of purchasing the package
- Customers who own passport are more likely to purchase the travel package compared to customers that do not own passport

Correlation Matrix

- Passport has very strong correlation
- City Tier has weak strong positive correlation towards product taken
- Age and Monthly income, No.of Persons visited & No.of Children visited have very strong positive correlations
- Owning a car and Monthly Income have positive correlations
- Product Taken and Monthly Income have negative Correlation and this can be caused due to the skewness in the dataset
- This graph shows that there is no much interaction between 'PitchSatisfactionScore', 'NumberOfFollowups' and 'DurationOfPitch' variables.



General Observation From EDA

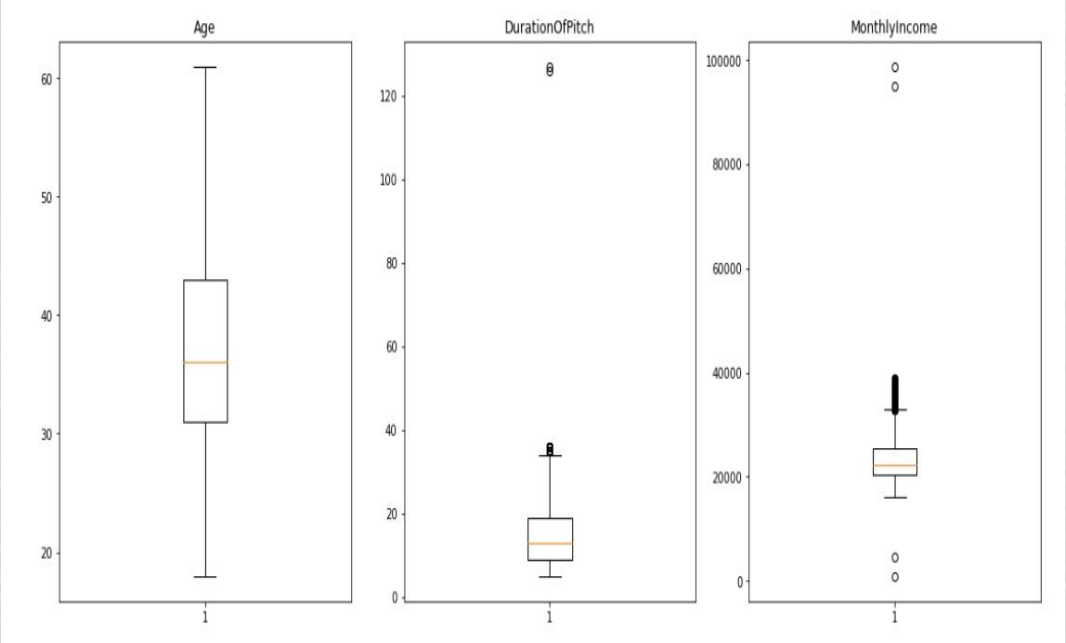
- Customers who purchased travel package preferred 5 star rated properties and were mostly from city tier 2,3 which they were mostly who came into the company's contact as company invite
- Single and unmarried customers has higher chance of purchasing the travel package.Also customers having 2,3,4 persons travelling with them has more chance of buying the travel package
- The holiday package has been taken mostly by customers who are above 30 years of age and who are in executive positions
- Customer who travel 7 or 8 trips per year have a higher chance of purchasing the travel package
- Also having a passport showed chances of purchasing the package
- For occupation the FreeLancers and Large Business owners have higher chance of purchasing the travel package
- Variables that did not have much impact travel package were Gender,number of children visiting and having a car.

While in terms of customer interaction data customers mostly purchased the travel package when

- The PitchSatisfactionScore was either 3 or 5
- The product pitched was Basic or standard product
- The longer duration of pitch by salesman to the customer
- The number of followups was high as with the number of 6 followups which displayed the higher chance of a customer acquiring the travel package

Outlier Detection

- This shows that there are extreme values such as 127 and 126 which can be possible but there are quite extreme
- From the results above, we have successfully treated missing values
- And the outliers have been treated successfully for NumberOfTrips, DurationOfPitch columns.
-



Model Building

1. Data preparation
2. Splitting the dataset into test and train
3. Build a model from the train data
4. Hypertuning the model
5. Run the data on the test dataset

Splitting the Dataset

From the above EDA it is clear that there are imbalanced distribution among the class variables. Hence, we would use stratified sampling to make sure that that the class frequencies are relatively sustained in both the training and test dataset

Since we are going to use startify sampling the parameter that would be used for this is callled train_test_split function.

Model Evaluation Criterion

ın order to prevent myself from making the wrong prediction such as :

- A customer can purchase the travel package when the customer does not actually purchase the travel package
- Also predicting that a customer will not purchase the travel package whereas the customer actually purchases the travel package

So it is important to understand the dataset further if we want to achieve better insights. Since the dataset is focused on a travel company that wants to cut the cost of marketing. Hence, it will be better to predict a customer that will purchase the travel package when the customer does not actually purchase the travel package in order to reduce the cost of marketing and places more focus on customers who would buy the travel package

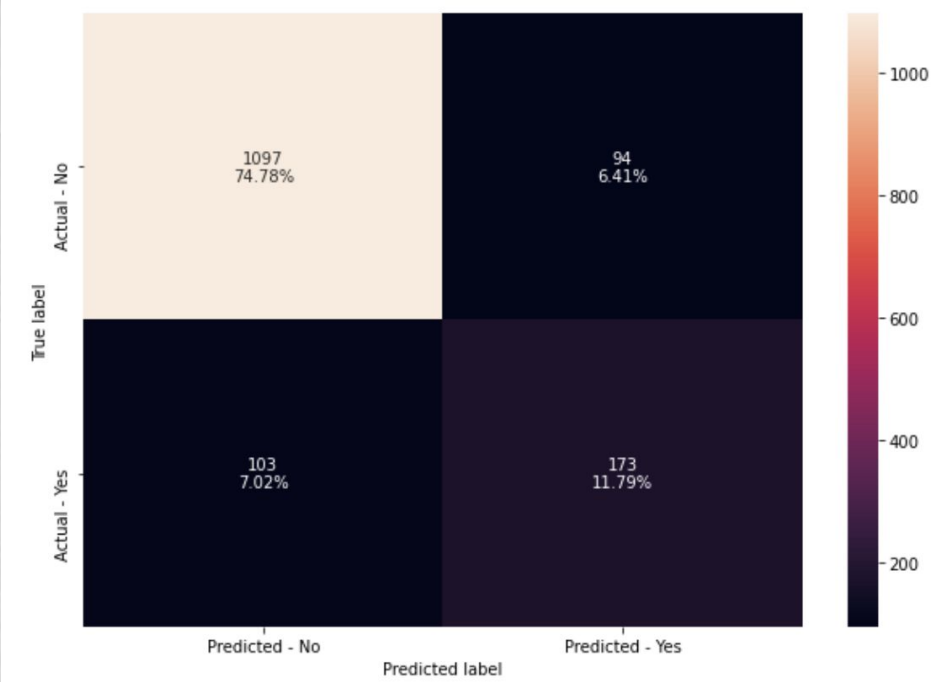
So the direction for us to take is to centre our focus on Precision to be maximized, greater the Precision higher the chances of minimizing false positives. Hence, the focus should be on increasing Precision or minimizing the false positives or in other words identifying the true positives,so that the company can focus on customers who will actually purchase the travel package and hence reduce marketing cost for the company

Decision Tree

Now we are going to build the model with use of

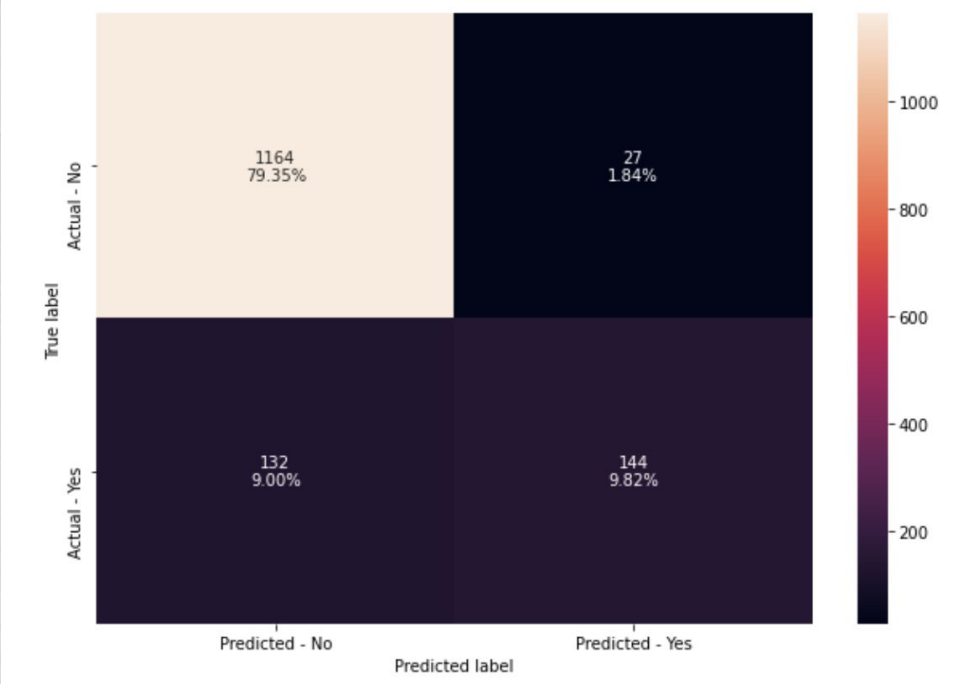
- DecisionTreeClassifier function with the default set as 'dini' criteria to split and the decision tree will give more weightage to class 1.
- Also class_weight is a hyperparameter for the decision tree classifier.

It is clear that the decision tree is overfitting the data and also the precision is also not that good.



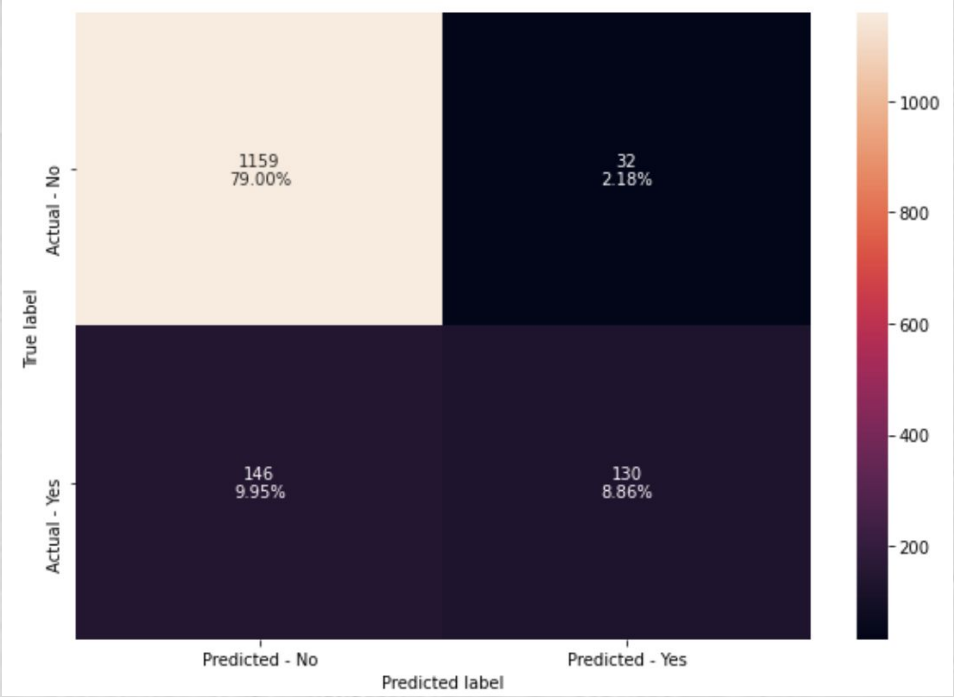
Bagging Classifier

- This shows that Precision has increased compared to the initial Decision tree
- Accuracy and recall have also increased



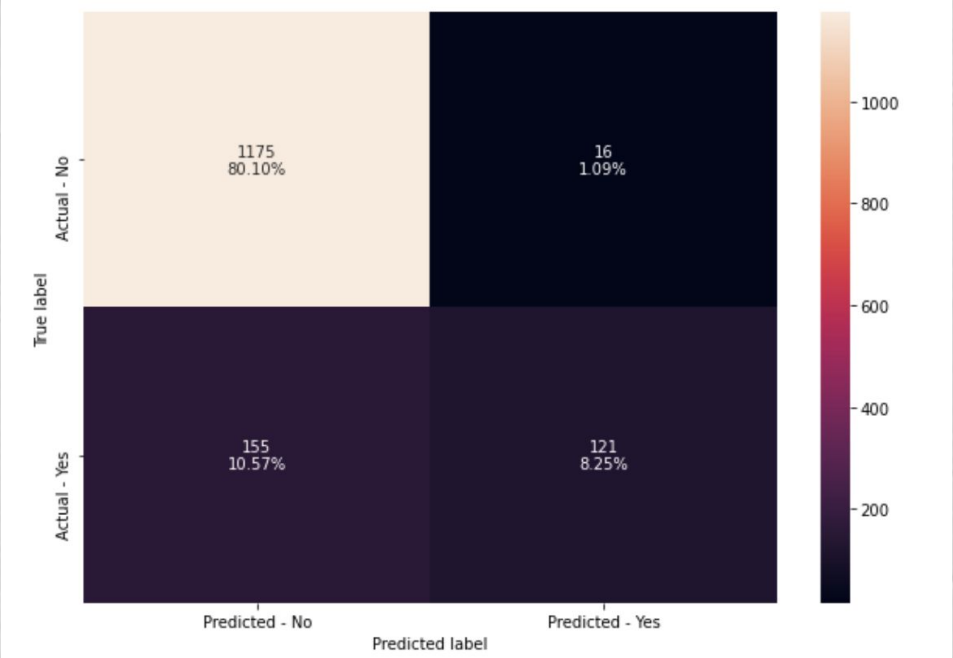
Bagging Classifier with weighted Decision Tree

- Precision has gotten better results
- While there is also a slight improvement in Accuracy
- The decision tree model is able to identify the customer which is likely to buy the travel package



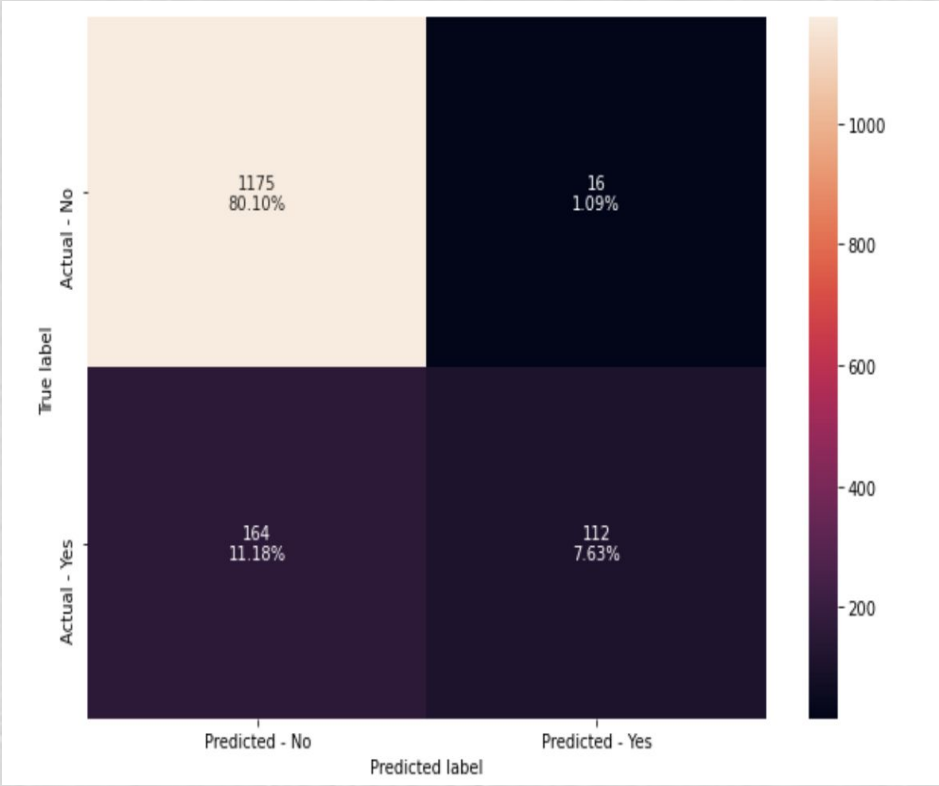
Random Forest

- Precision is looking good however Recall seems to have dropped



Random Forest with Class_weights

- There is still overfitting but there has been an improvement in Precision

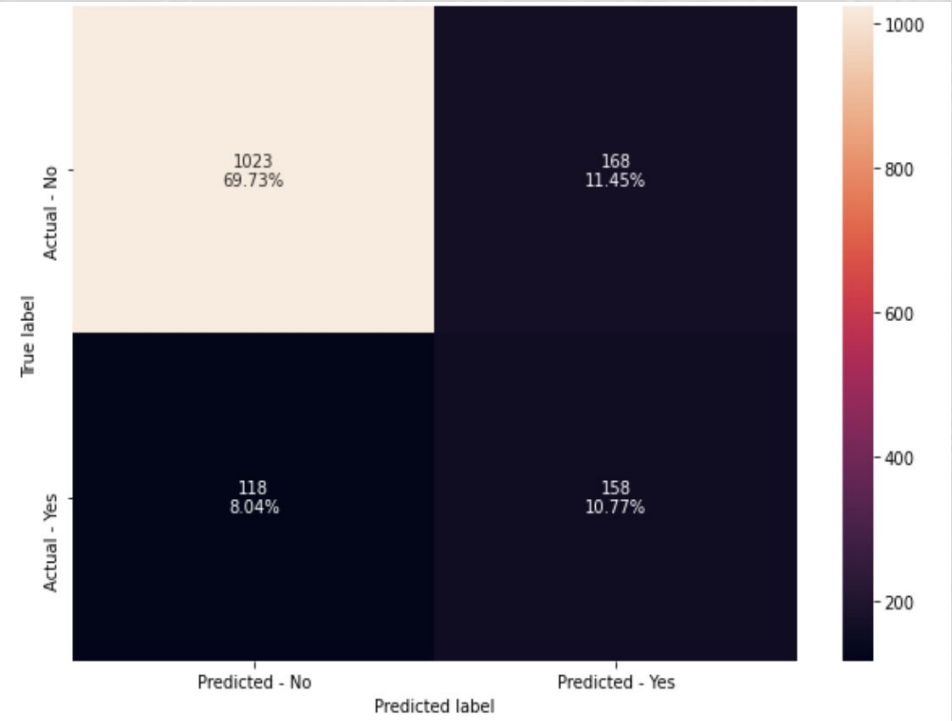


Model Performance Evaluation and Improvement - Bagging

- Customer purchased package and model predicted customer will purchase travel package : True Positive (observed=1,predicted=1)
- Customer did not purchase package and model predicted customer will purchase travel package : False Positive (observed=0,predicted=1)
- Customer did not purchase package and model predicted customer will not purchase travel package : True Negative (observed=0,predicted=0)
- Customer purchased package and model predicted customer will not purchase travel package : False Negative (observed=1,predicted=0)

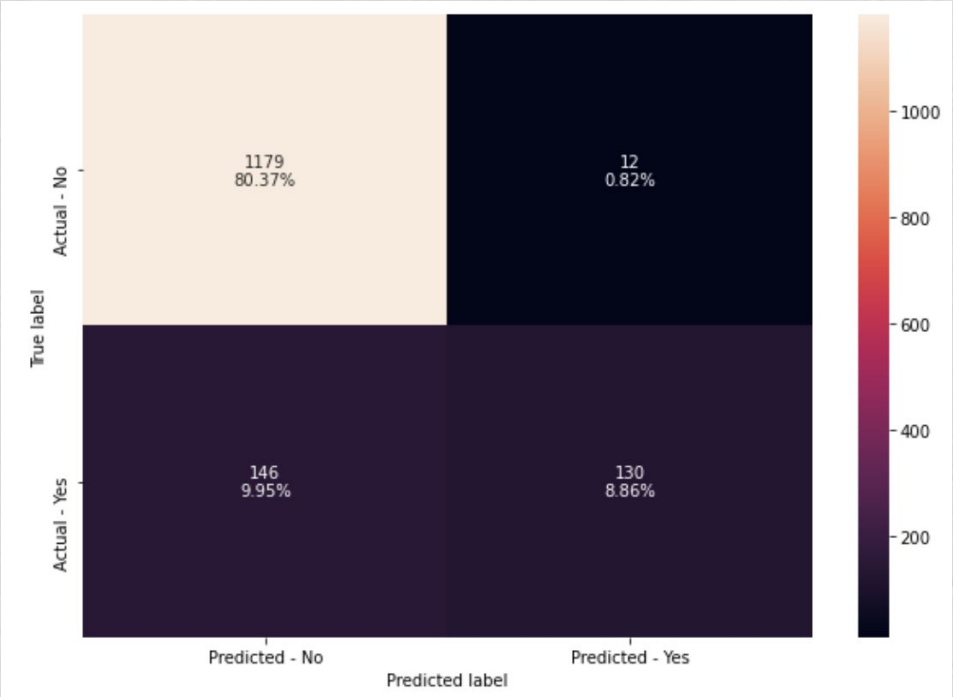
We can see a clear difference from the above analysis and the hypertuned model such as

- Recall has improved
- Precision has really decreased
- There is still overfitting however it has also reduced



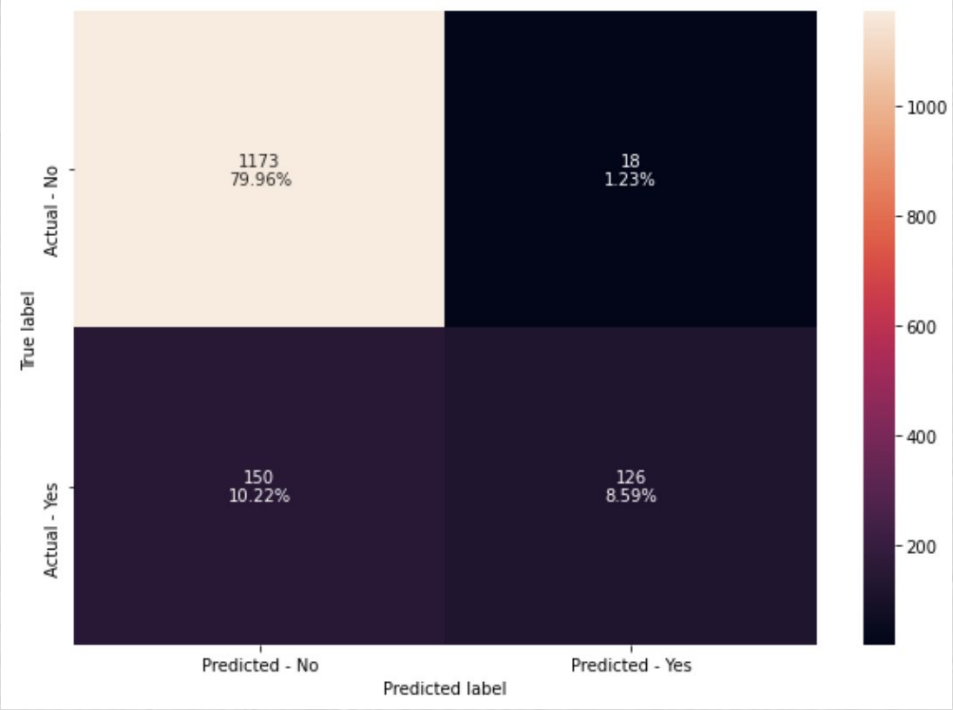
Hypertunning Bagging Classifier

- Now there is a great improvement in both Precision and also in Accuracy



Hypertuned Random Forest

- Even still after hypertunning the Random Forest the model still displays that there is still overfitting in the data



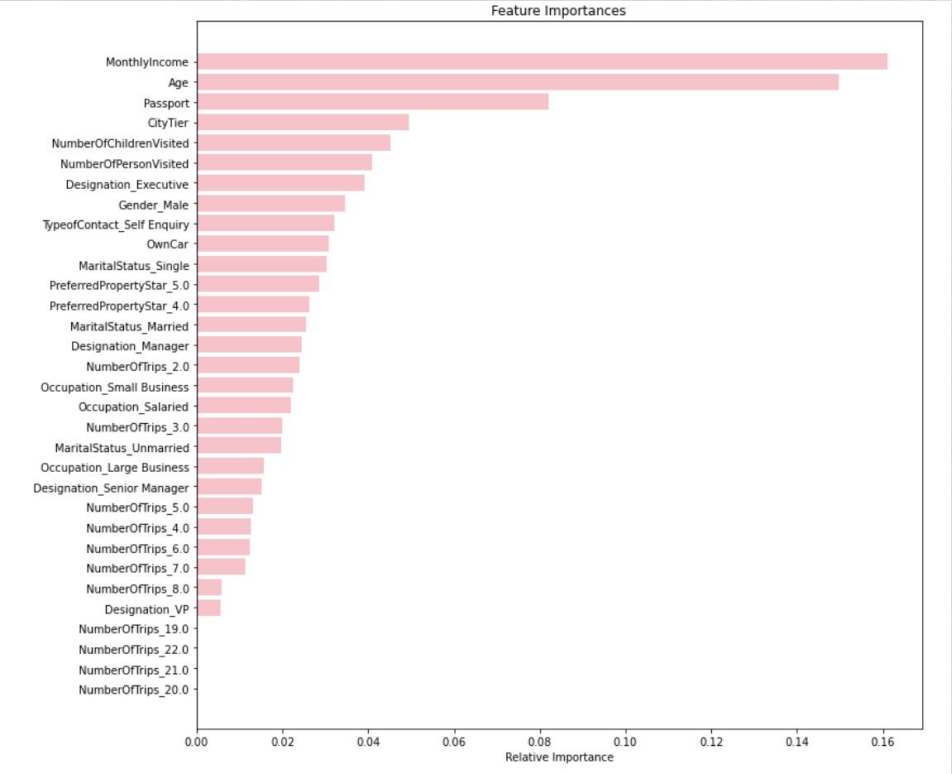
Results After comparing the models

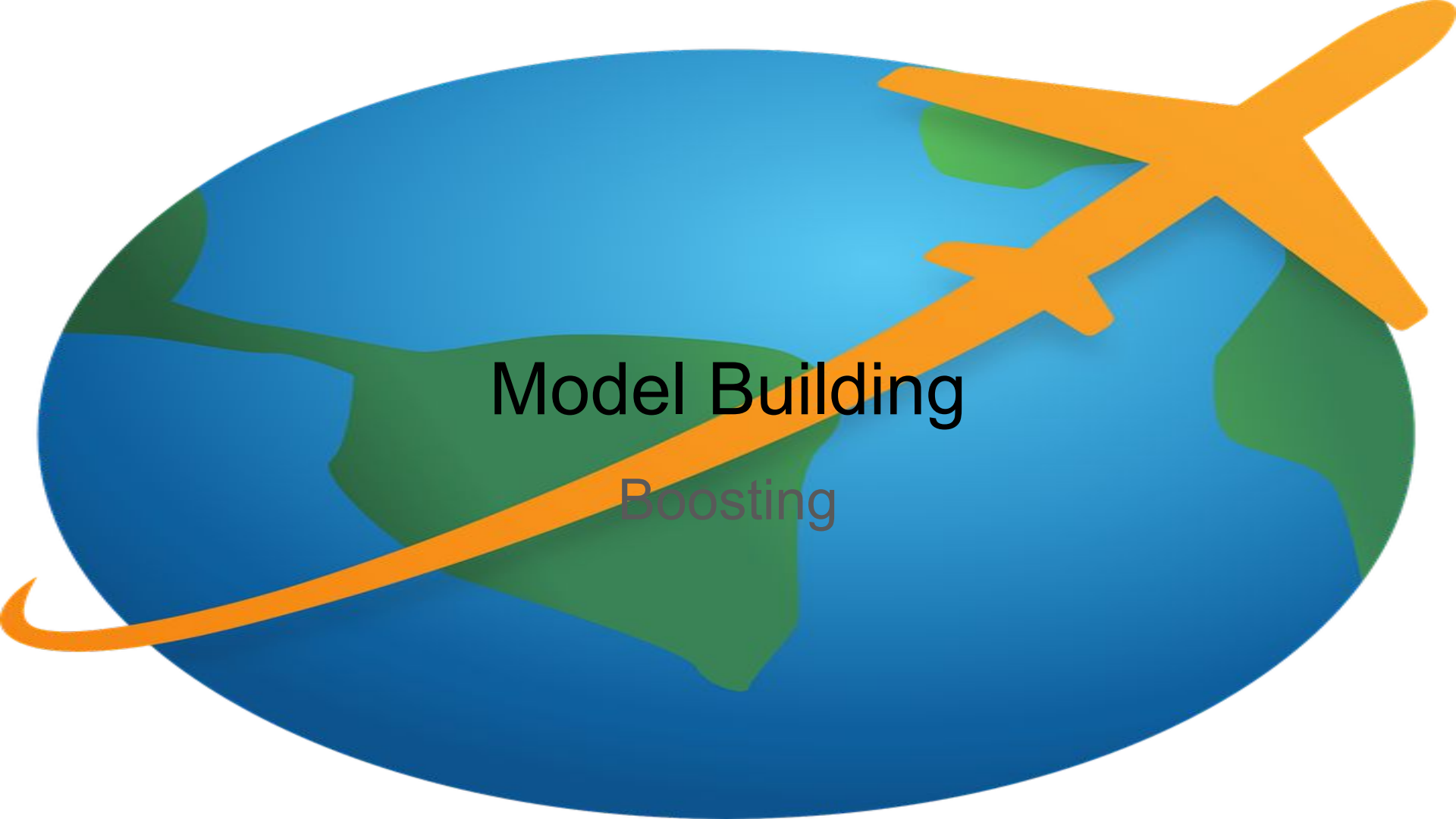
- Hypertuned Bagging classifier has the highest precision
- Hypertuned Bagging classifier also has a good accuracy
- Also RandomForest With Class weights is good to

Important Features of the Random Forest with Class_weights

Observation

- The top features which were selected by RandomForest With Class weights are Monthly Income, Age, Passport, CityTier and NumberOfChildrenVisited



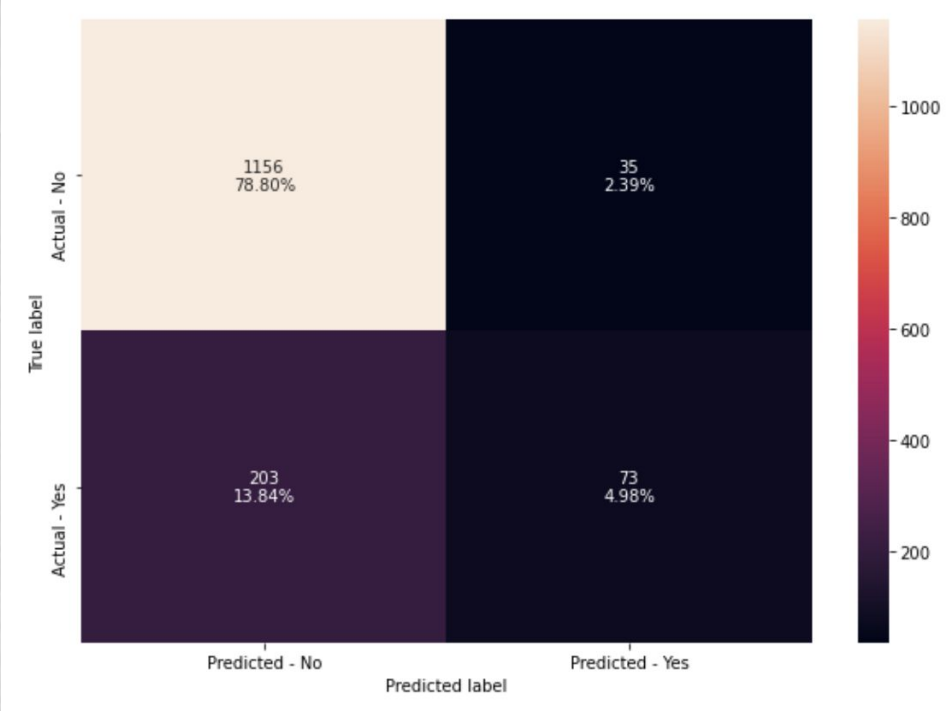


Model Building

Boosting

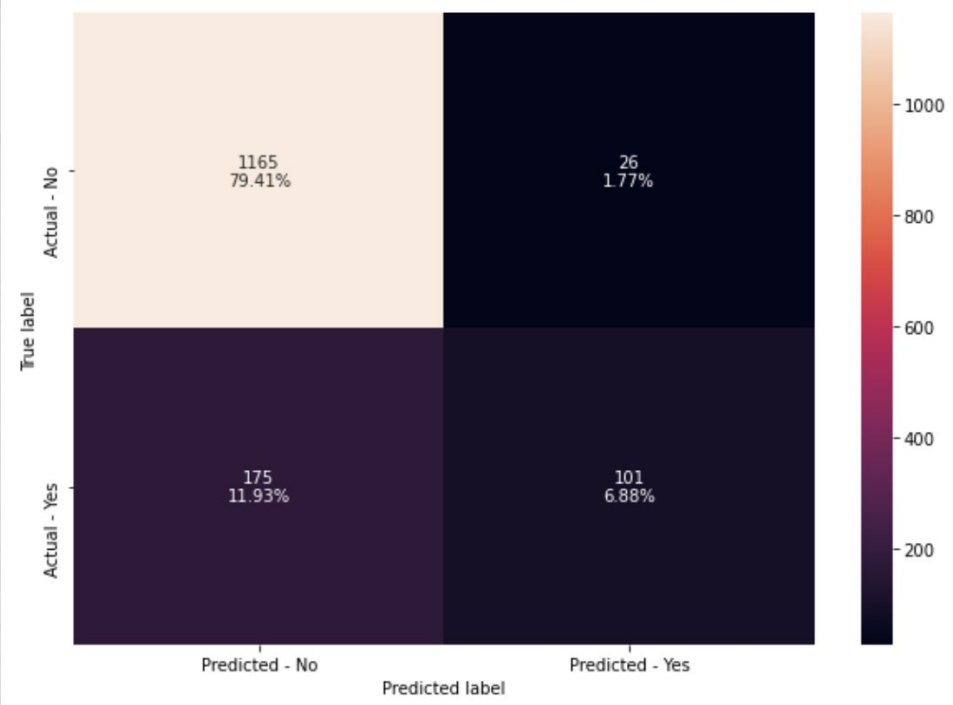
AdaBoost Classifier

- It can be deduce that both accuracy and Precision are quite good. However, recall does not good at all.



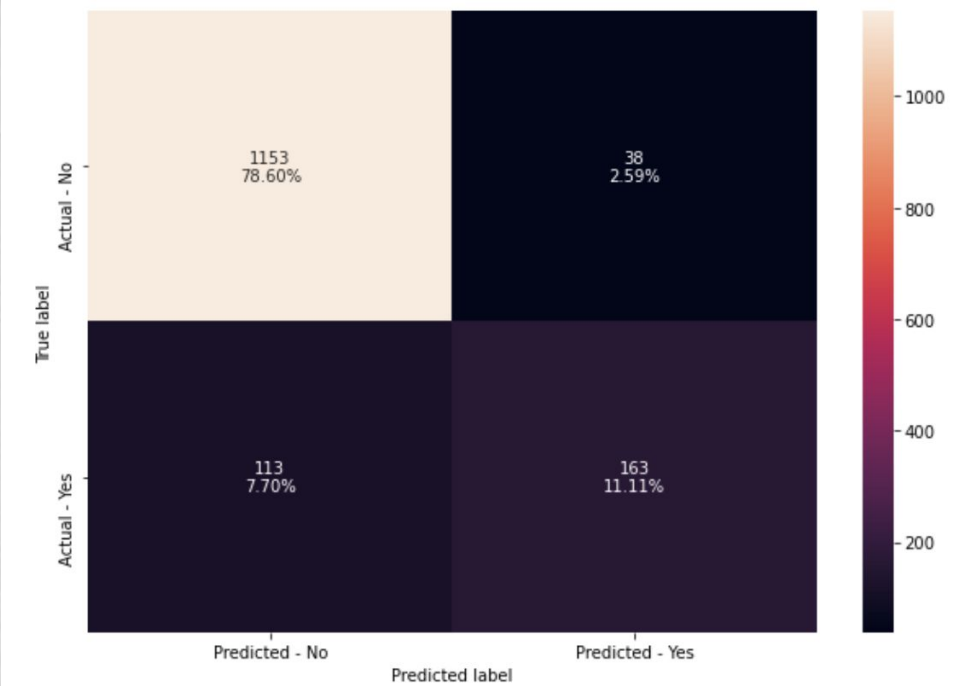
Gradient Boost Classifier

- The recall looks better compared to the AdaBoost model above
- Precision also looks better here too



XGBoost Classifier

- Compared to the rest of the other models XGBoost performs the best but also displays some overfitting in the model

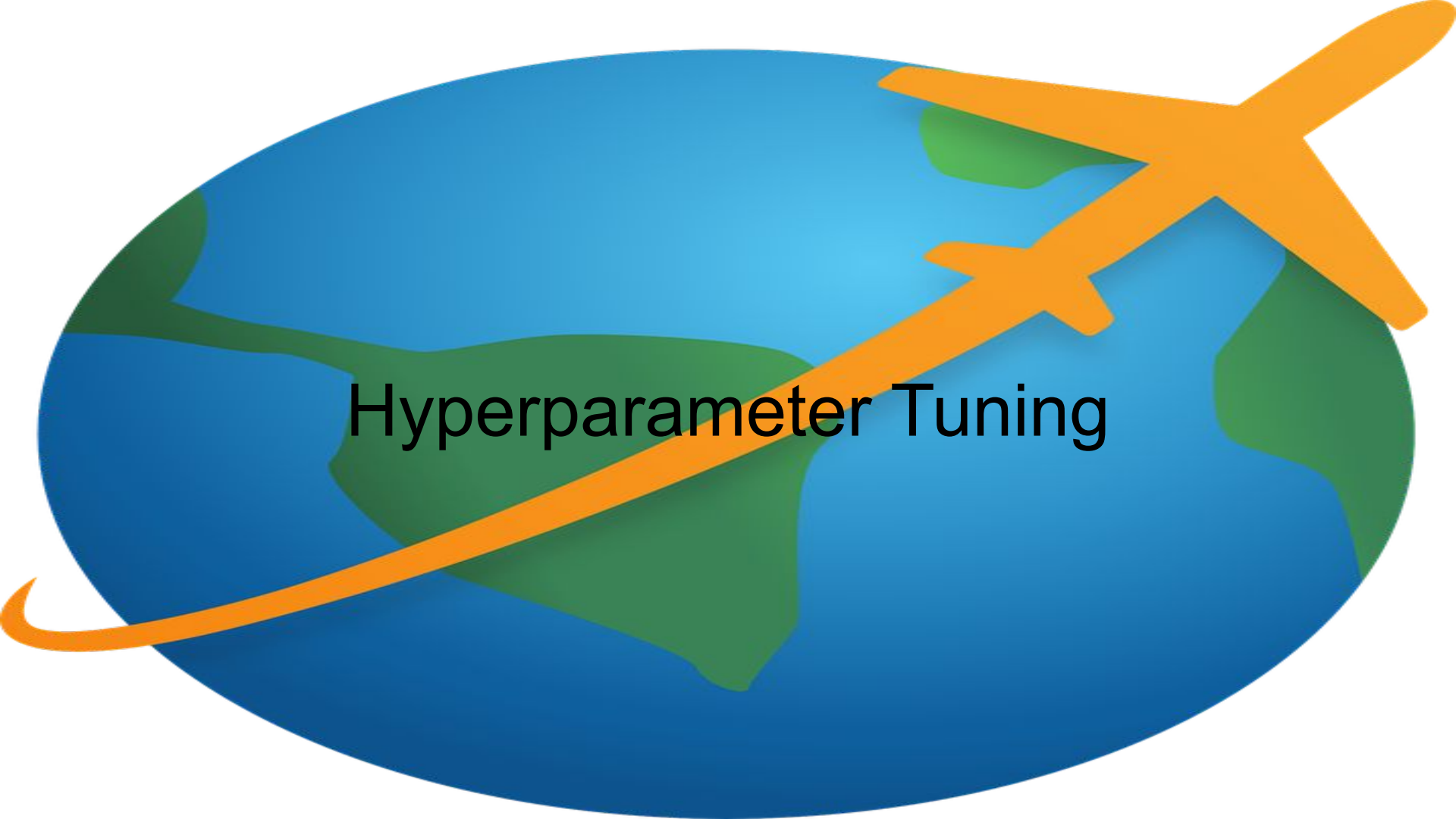


Model Performance Improvement - Boosting

- Customer purchased package and model predicted customer will purchase travel package : True Positive (observed=1,predicted=1)
- Customer did not purchase package and model predicted customer will purchase travel package : False Positive (observed=0,predicted=1)
- Customer did not purchase package and model predicted customer will not purchase travel package : True Negative (observed=0,predicted=0)
- Customer purchased package and model predicted customer will not purchase travel package : False Negative (observed=1,predicted=0)

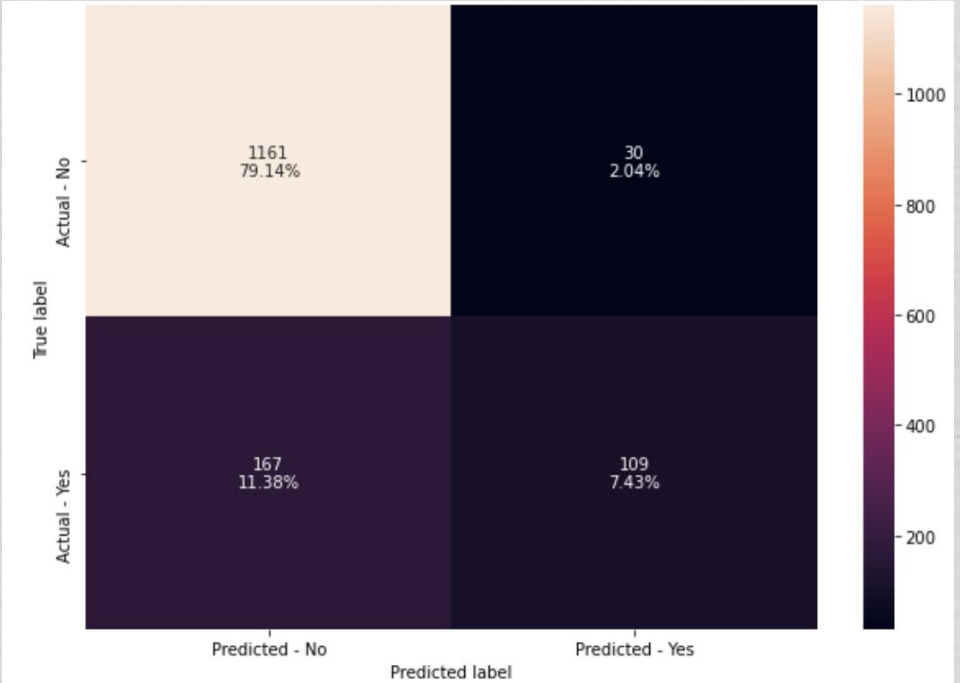
Since from the above models we had already decided to take Precision and the Travel Company would prefer if Precision is maximised which means that the higher Precision is it gives greater chances of minimising the false positives in the model. This will aid the company in centering more attention to the customers who stand a higher chance of purchasing the travel package and will reduce the company's cost of marketing.

Now we are going to hypertune the models to check if we can improve it further.


Hyperparameter Tuning

Hyperparameter Tuning - AdaBoost Classifier

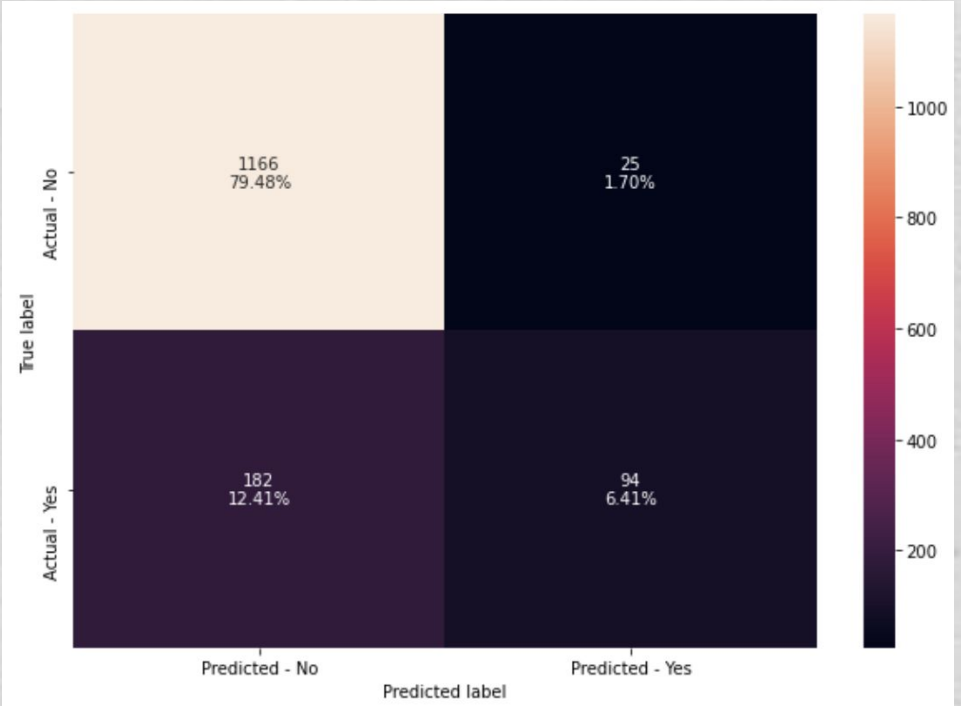
- There is a bit of overfitting after tuning but its not that bad



Hyperparameter Tuning - Gradient Boost Classifier

Using AdaBoost classifier as the estimator for initial predictions

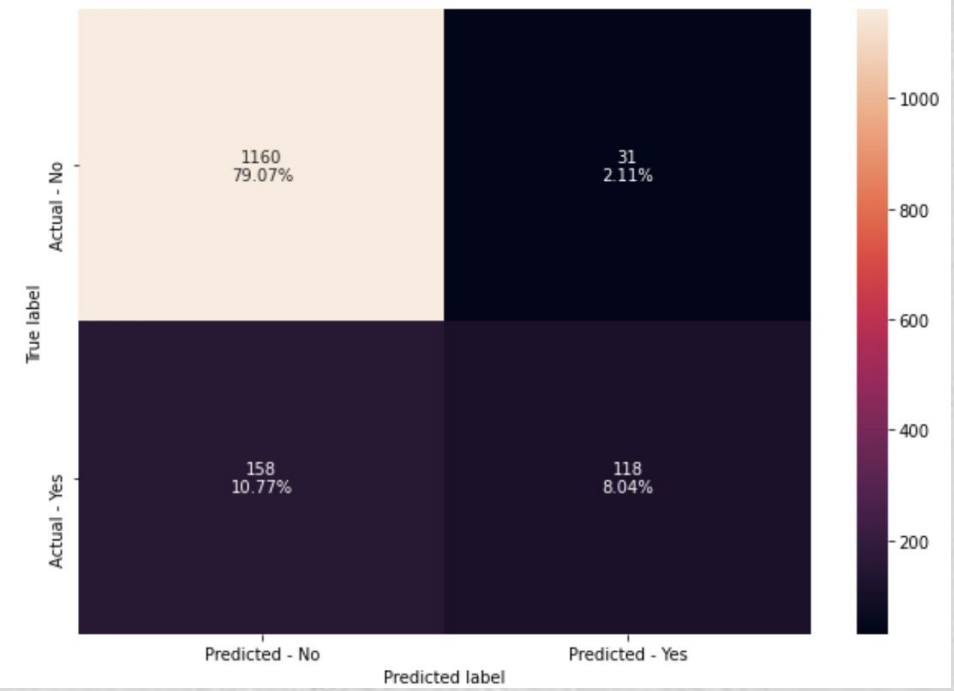
- Gradient Boosting with Adaboost as base estimator is giving good predictions and there is slight overfitting



Hyperparameter Tuning - Gradient Boost Classifier

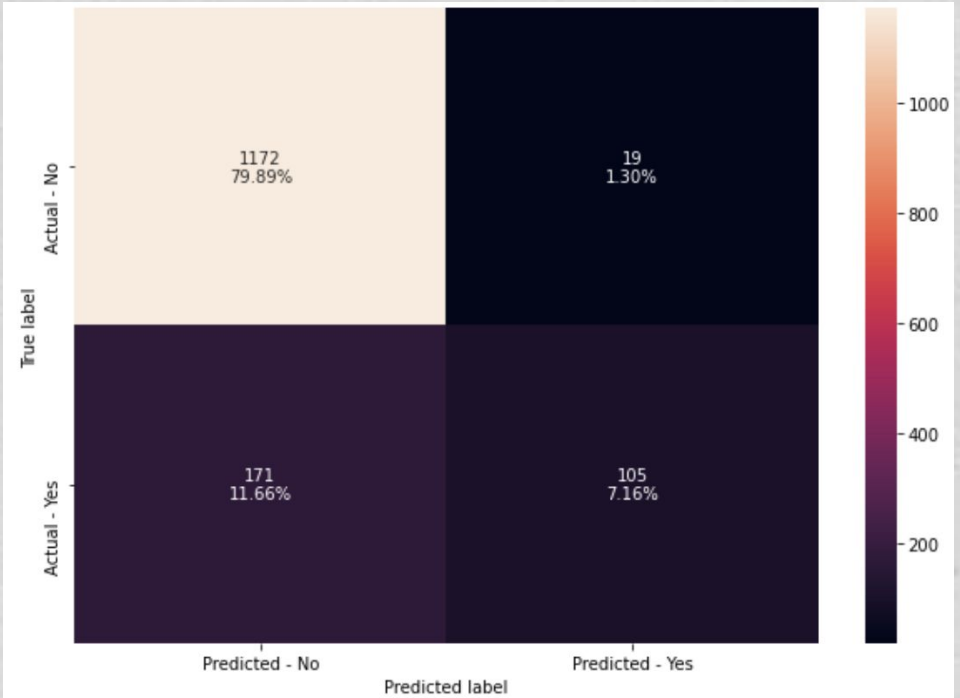
Without using AdaBoost classifier as the estimator for initial predictions:

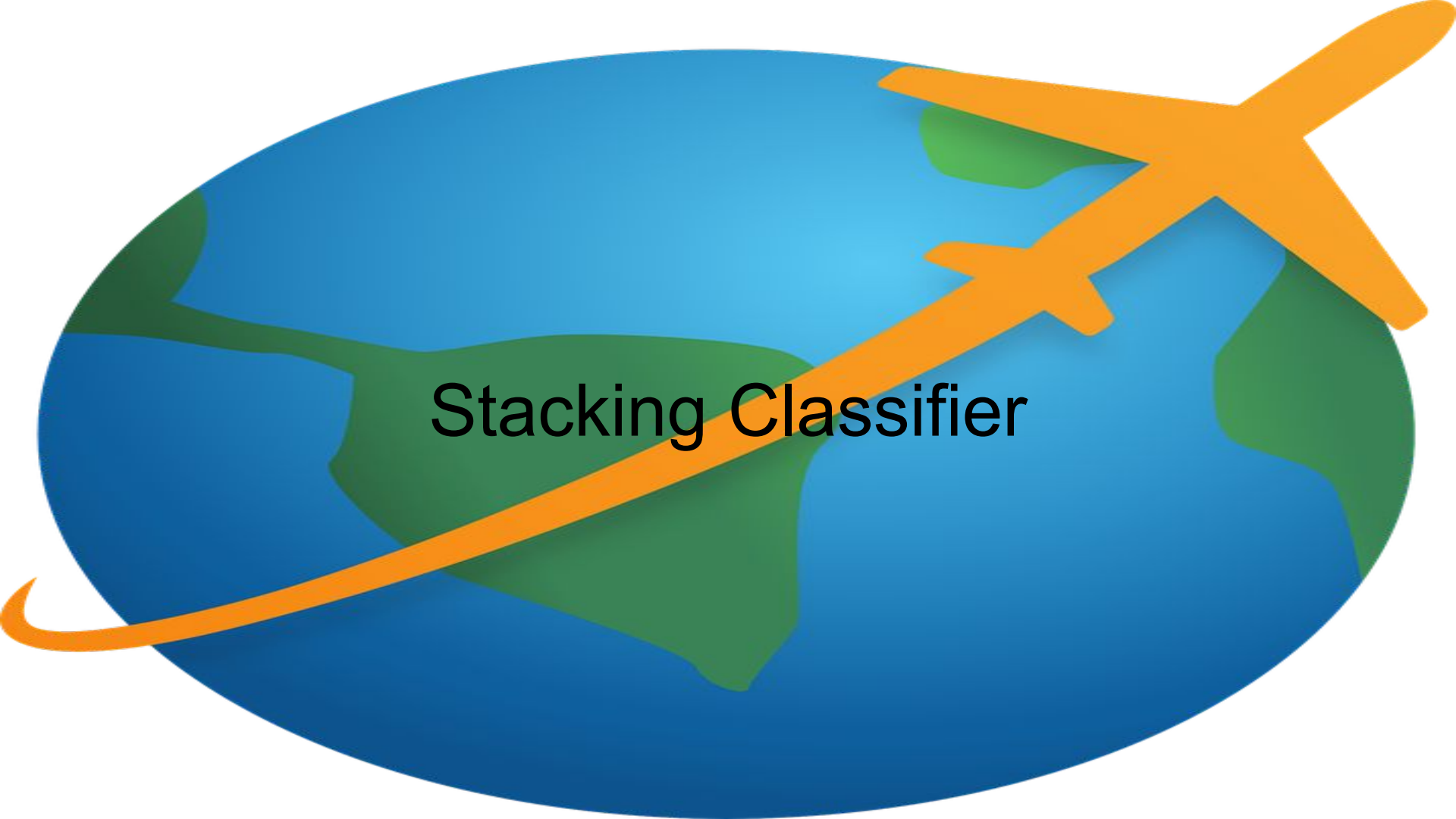
- After Tuning it shows that the model has improved.



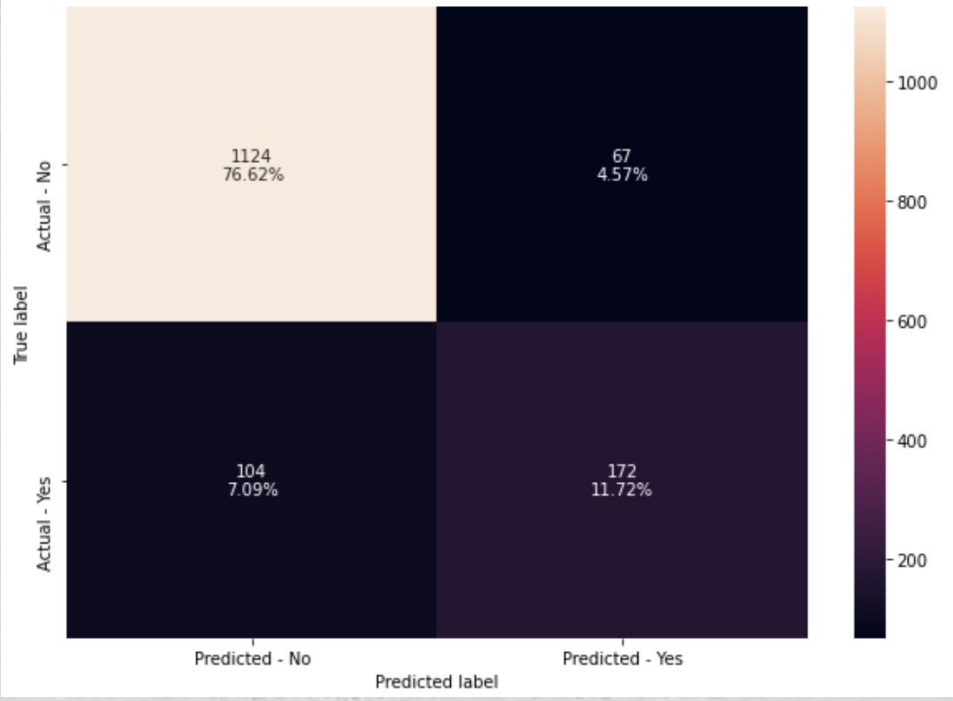
Hyperparameter Tuning - XGBoost Classifier

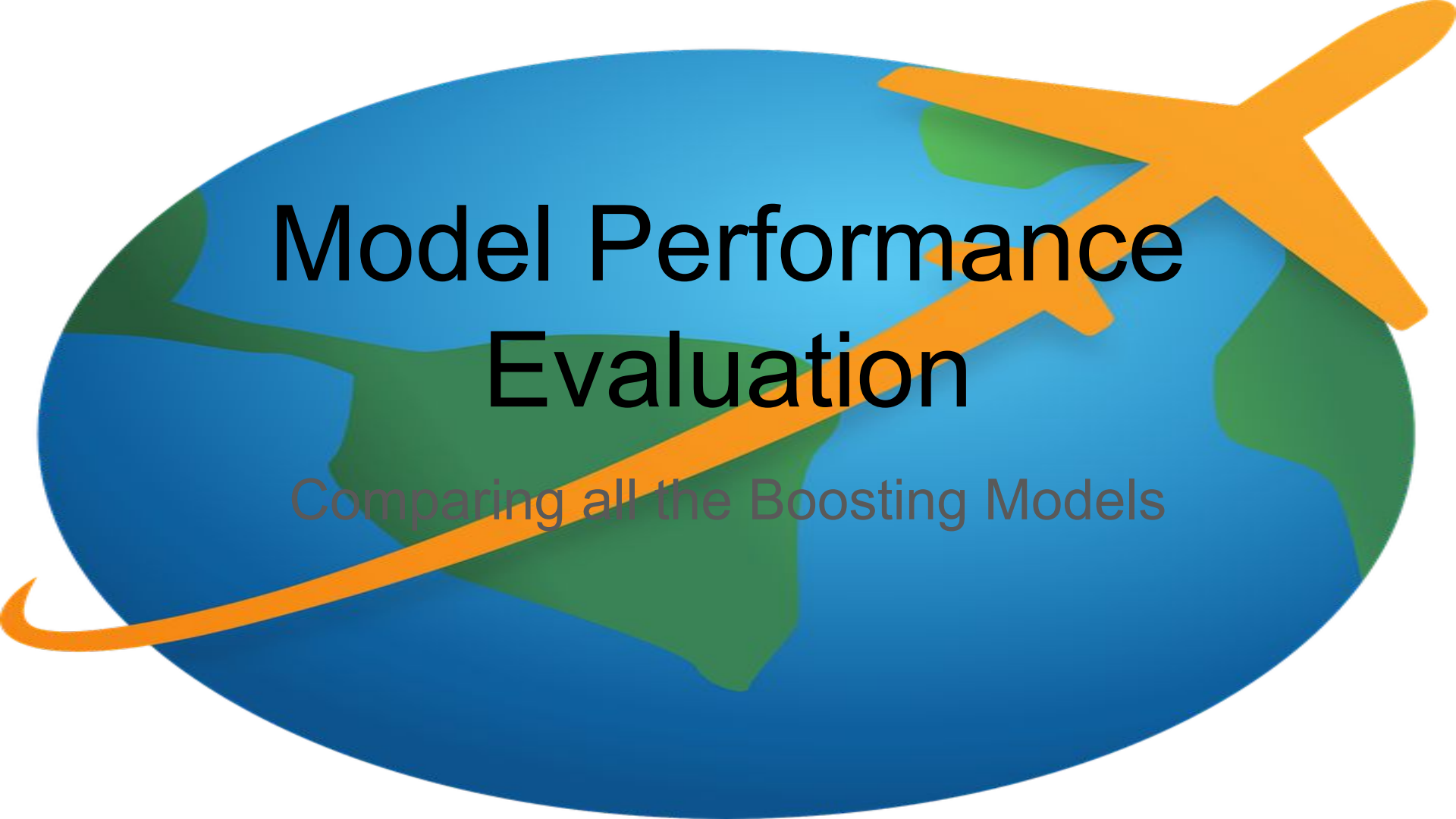
- Both in the initial XGBoost model and the hypertuned model it still displays that there is still overfitting involved in the model.




Stacking Classifier

- Even with the use of Stacking Classifier it still reveals that there is still overfitting of the data.
- Since with model performance improvement now we will evaluate and compare which model performance metrics is the best





Model Performance Evaluation

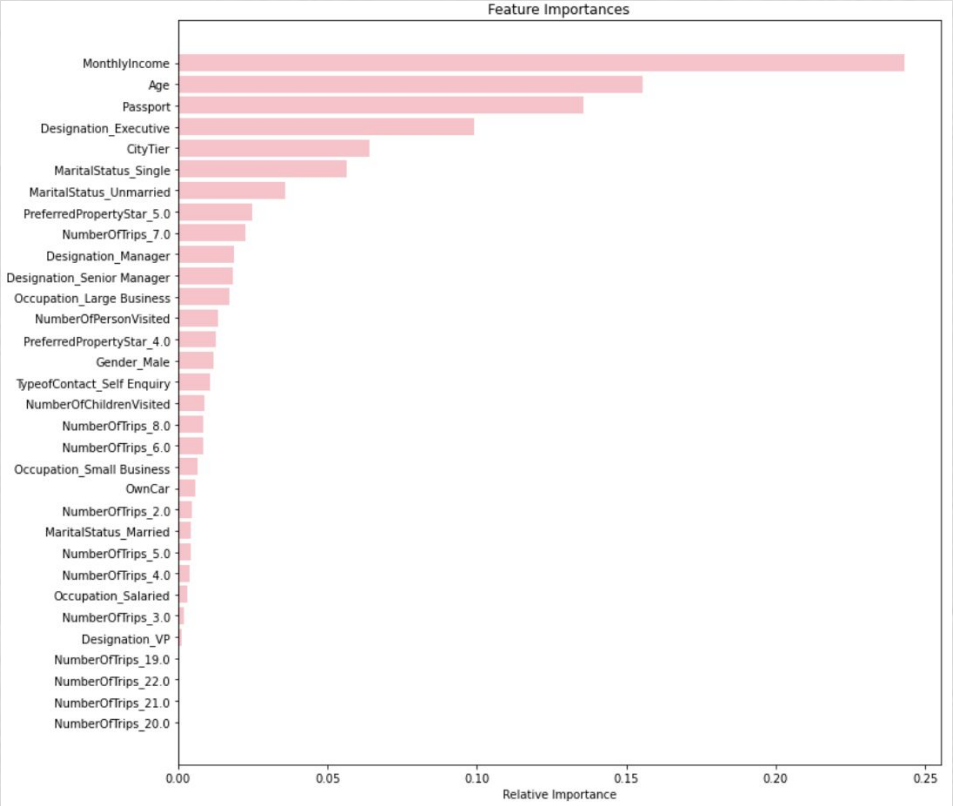
Comparing all the Boosting Models

Observation

- Random Forest gives a good performance
- Tuned Bagging Classifier & Random Forest with class weight is also giving a good performance
- Tuned Gradient Boosting Classifier has comparable accuracy ,precision and recall.

Feature Importance of Tuned Gradient Boost Classifier

- MonthlyIncome, Age, Designation_Executive, Passport and City_Tier are more positively impacting the target variable




Actionable Insights & Recommendations

Business Insights and Recommendation

Using the model to predict which customers will buy the travel package when offered, here are factors would be given importance:

- The most important factors which could be used to determine if customers that most likely to buy a package are Monthly Income, Age of the customer and if a customer owns a Passport or not and this plays very important role in customer choosing the package.
- For the new package to be picked, all of the above specified feature values should be High.
- The model would recommend the travel packages advertised to customers the age from 40 and below and with a Monthly Income of up to 40,000.
- Also try splitting Customers based on their salary, passport, owning a car.
- Customers in City Tier especially those in tier 3 can also be targeted to advertise the travel package because they also have a high likelihood of also purchasing the travel package.
- Also customers who are in Manager designation has to be target for Marketing.
- Male and also single customers display a greater chance of acquiring the travel package.
- Those customers who prefer 5 star properties can also be targeted to buy travel package.

Factors from customer interaction data which will boost the chances of the customer buying the travel package that needs to be considered are:

- Earning a PitchSatisfactionScore of either 3 or 5
- A higher duration of pitch by salesman to the customer
- The travel company doing numerous or collective followups with a customer. An example would be have the number of 6 followups with a customer can persuade a customer to purchase a travel package.


Thank You!
Anisah Inua Mohammed