Travel Package Purchase Prediction

Contents

- Background & Problem Statement
- Data Dictionary
- Exploratory Data Analysis Highlights
- Model Performance Summary
- Recommendations

Background & Problem Statement

There is a need to expand the customer base, and hence newly introduced packages are introduced.

However, the marketing cost was quite high as customers were contacted at random without looking at the available information.

We require a model that can predict potential customers who have a higher probability of purchasing the newly introduced packages.

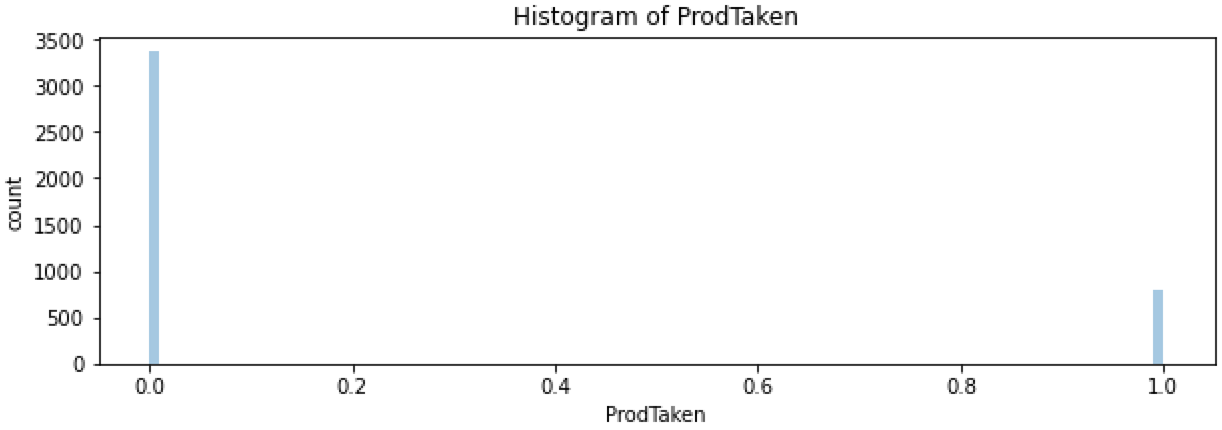
Here, several models from ensemble techniques will be used.

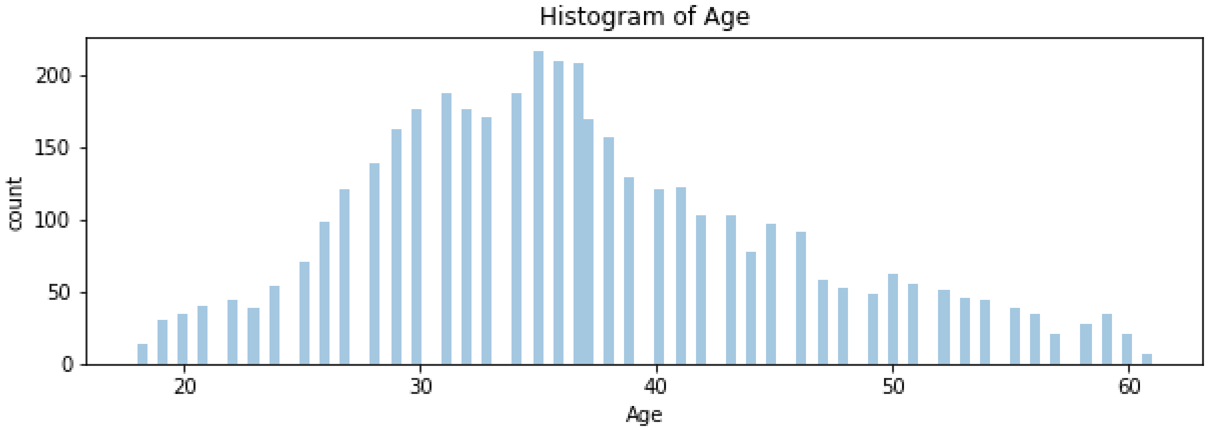
Data Dictionary

The data contains information about 4888 customers and their characteristics. Data pre-processing was done, which includes outlier treatment, conversion of ‘Fe Male’ to ‘Female’, and imputation of missing data.

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

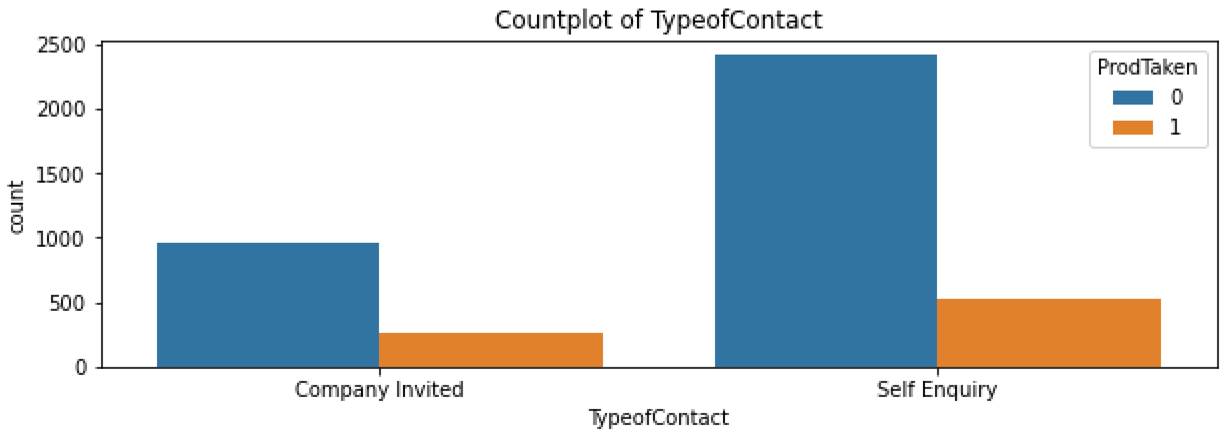
Exploratory Data Analysis Highlights

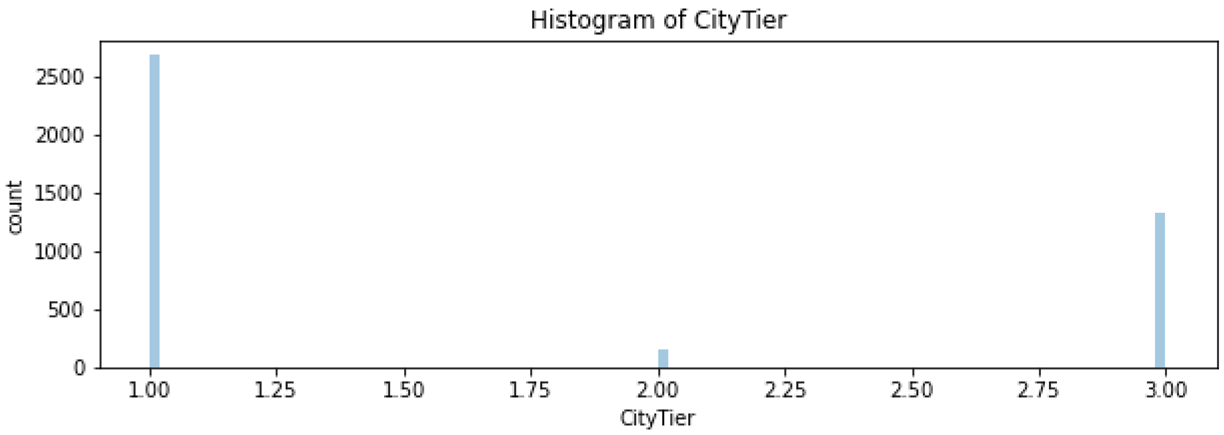




- There are more customers who did not take the package
- Age is normally distributed

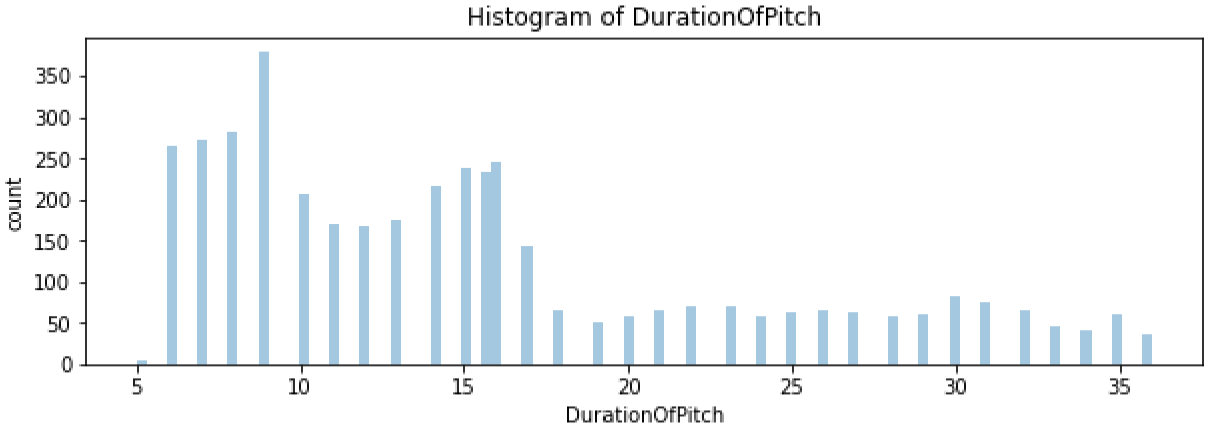
Exploratory Data Analysis Highlights

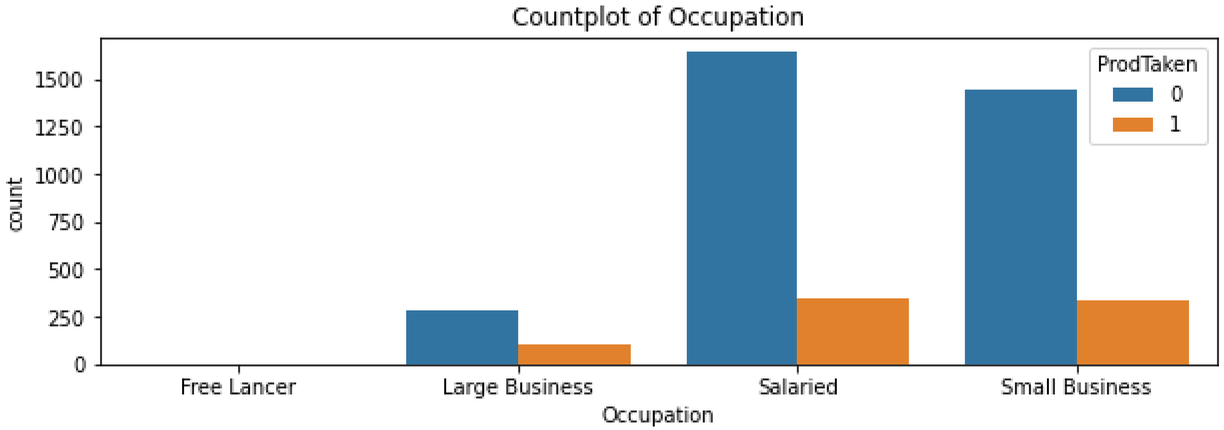




- There are more people that made a self enquiry for the product.
- Most of the people are from CityTier of 1

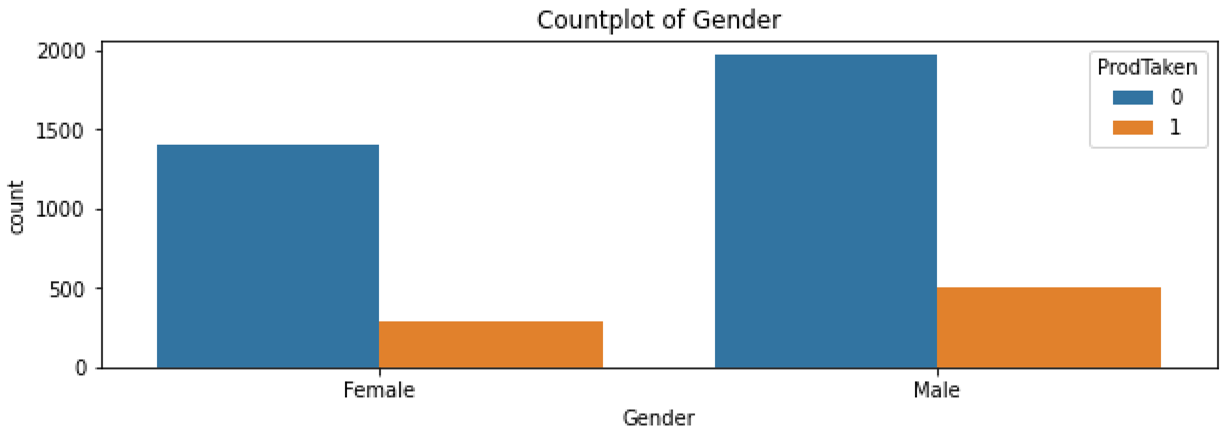
Exploratory Data Analysis Highlights

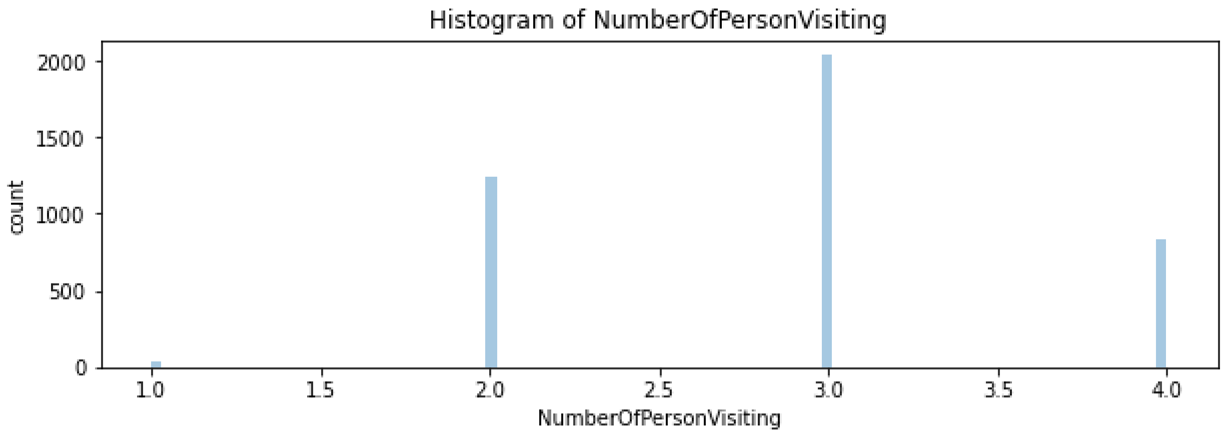




- DurationOfPitch is right skewed
- Most of the customers are salaried or have a small business

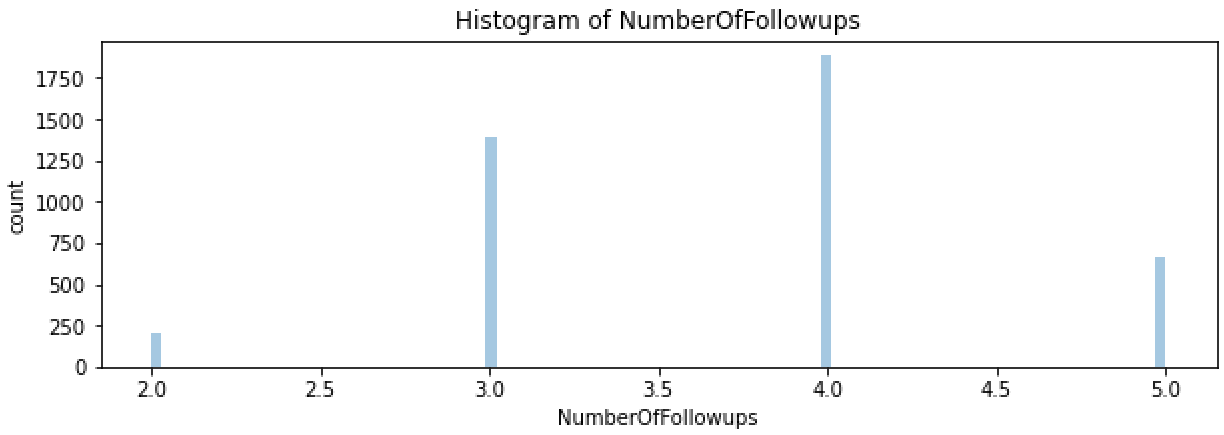
Exploratory Data Analysis Highlights

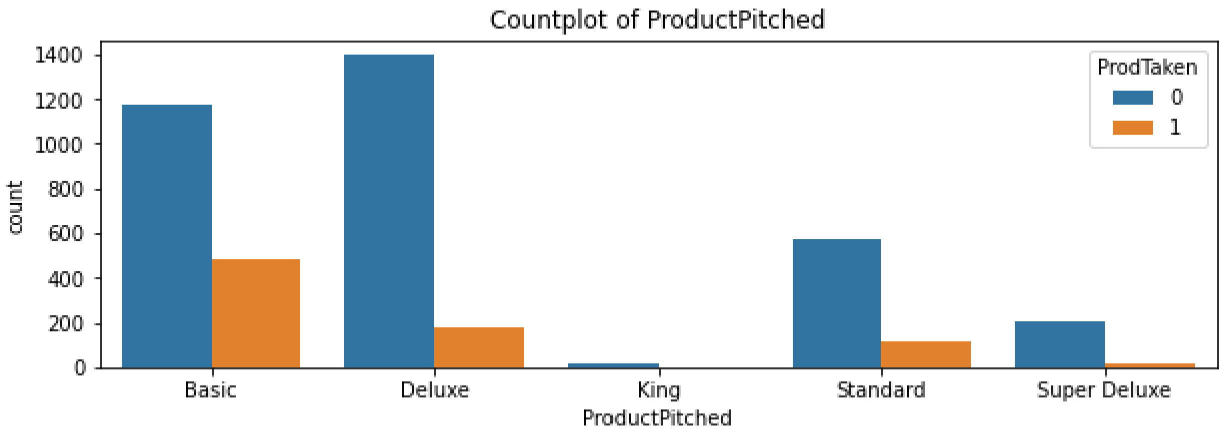




- Most of the customers are male
- On average, around 3 people join the trip with the customer

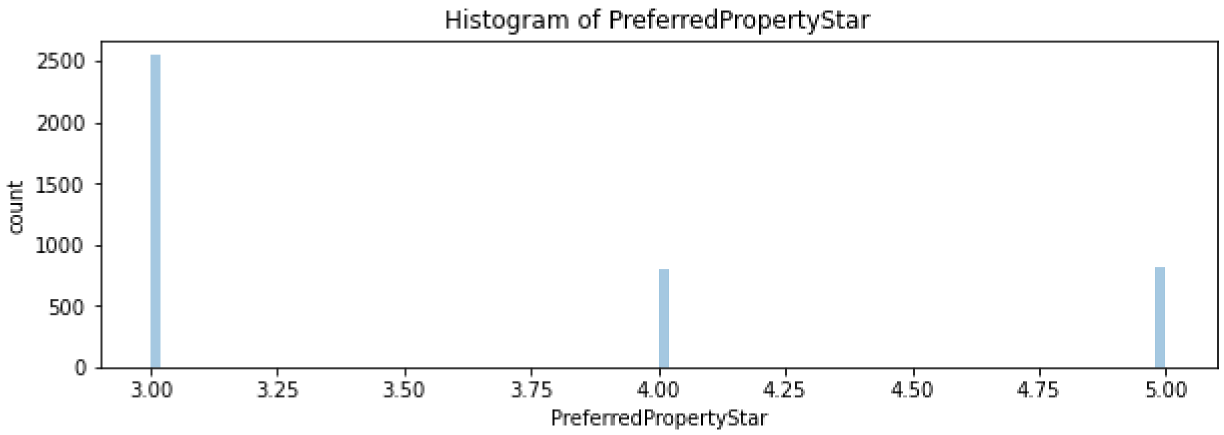
Exploratory Data Analysis Highlights

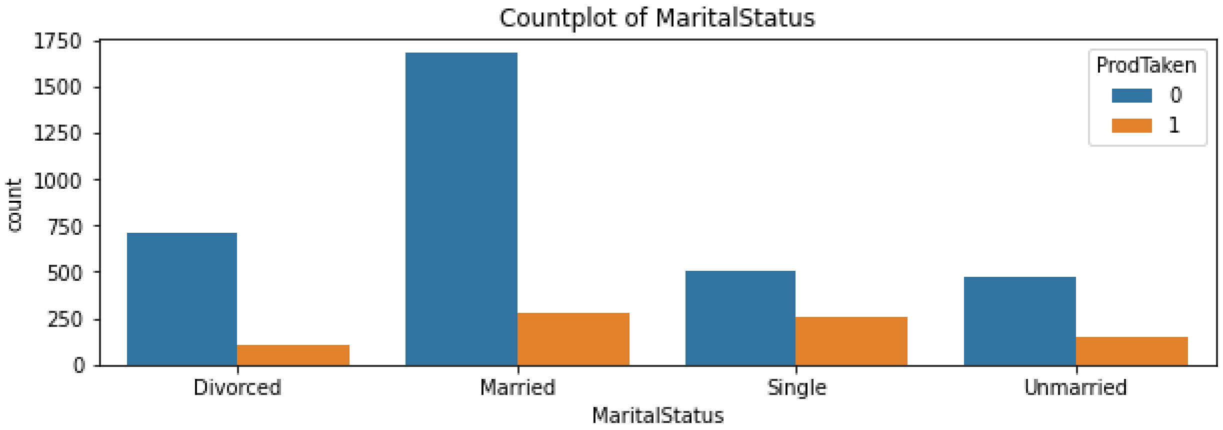




- The number of followups are around 3-4 generally
- There are more product pitches of Basic and Deluxe

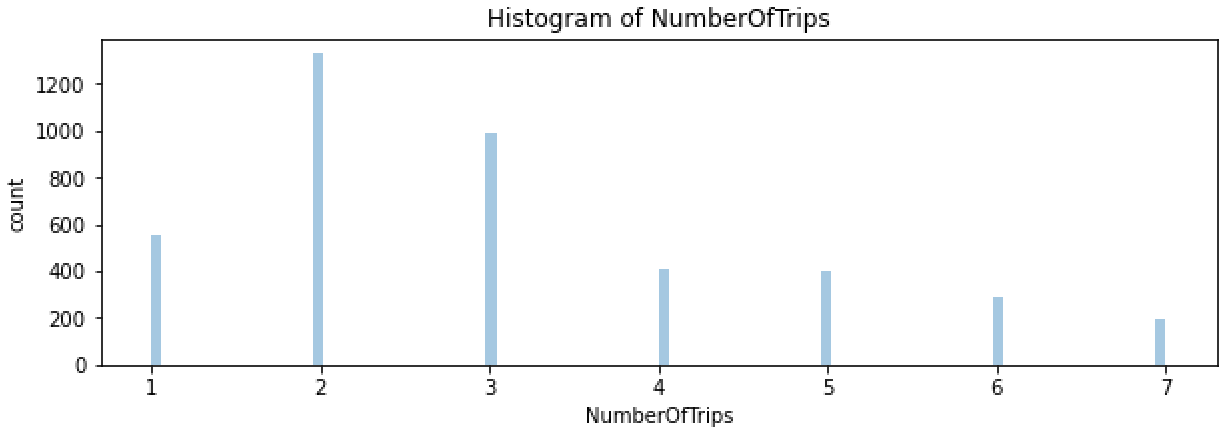
Exploratory Data Analysis Highlights

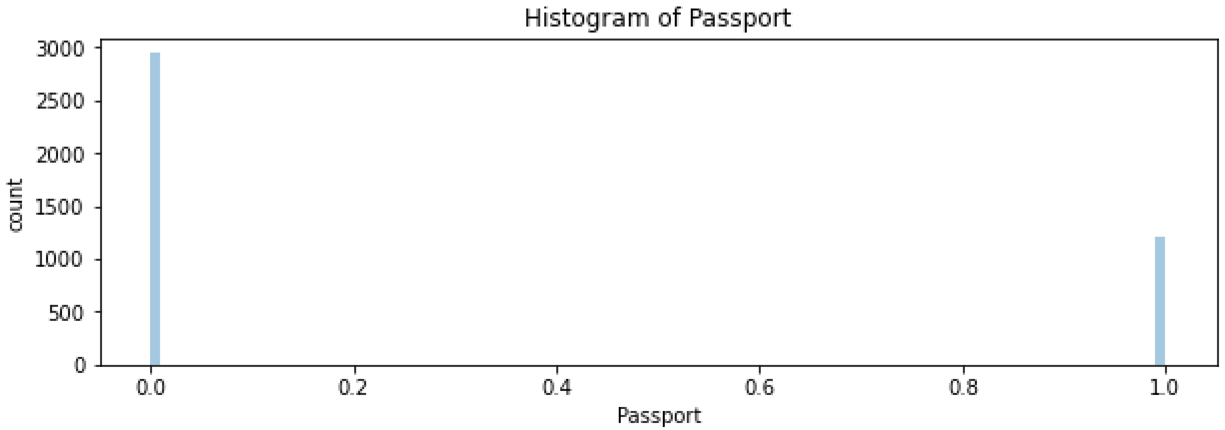




- Most of the preferred property star is 3
- Most of the customers are married. For product takers, most are either married or single.

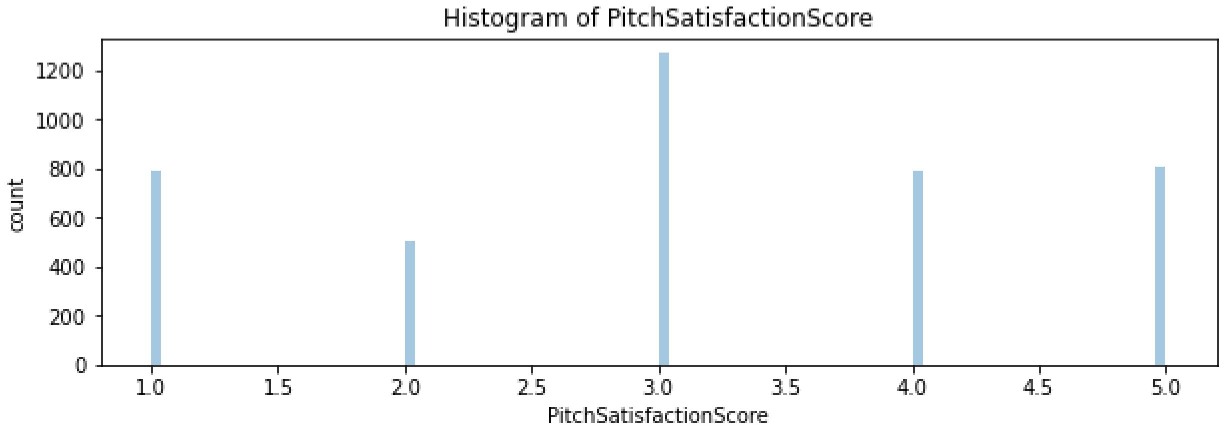
Exploratory Data Analysis Highlights

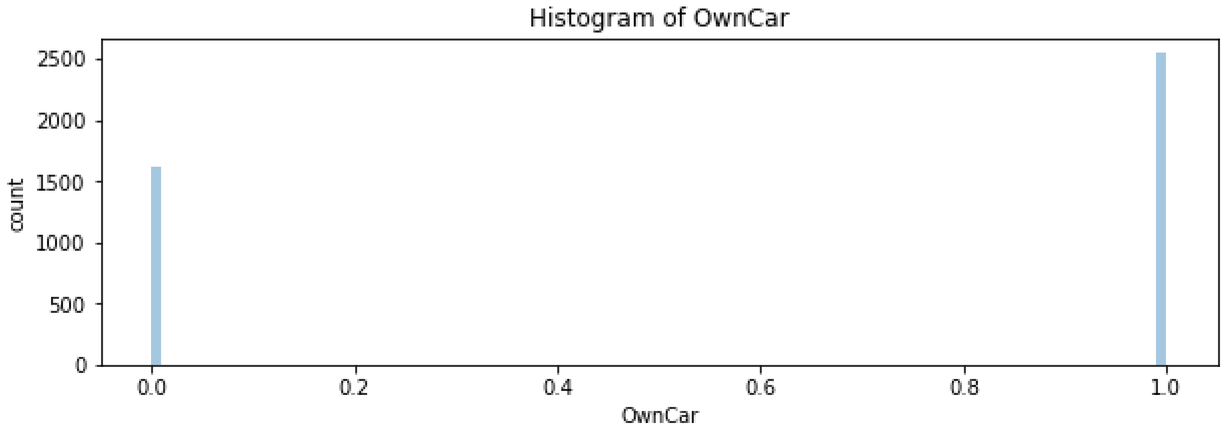




- Most of the number of trips per year of customer is 2.
- Most of the customers do not have a passport.

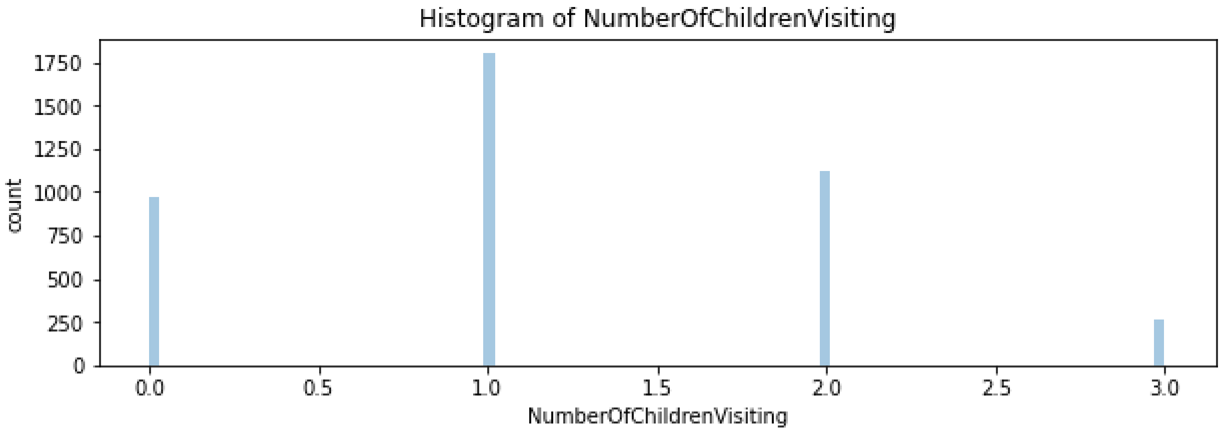
Exploratory Data Analysis Highlights

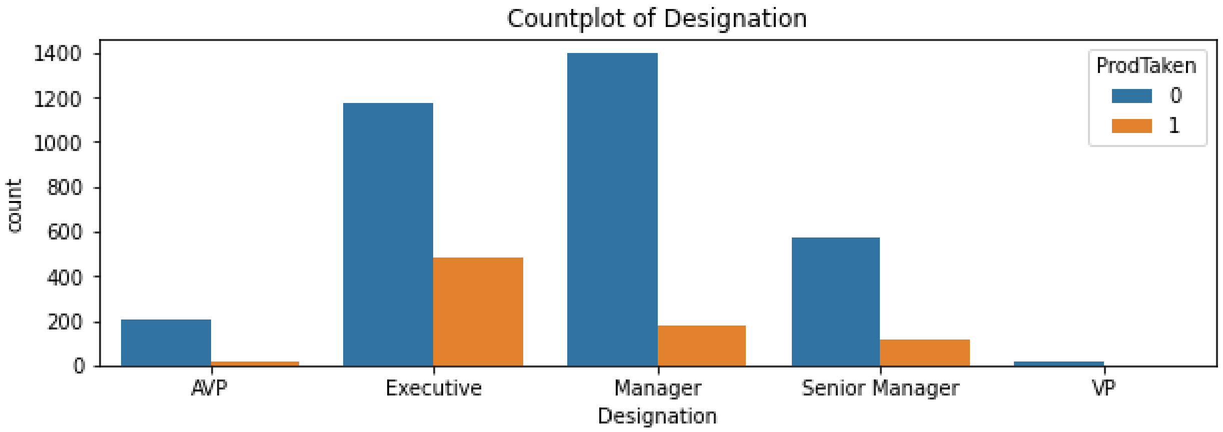




- It seems that the pitch satisfaction score is at an average of 3
- Many of the customers own a car

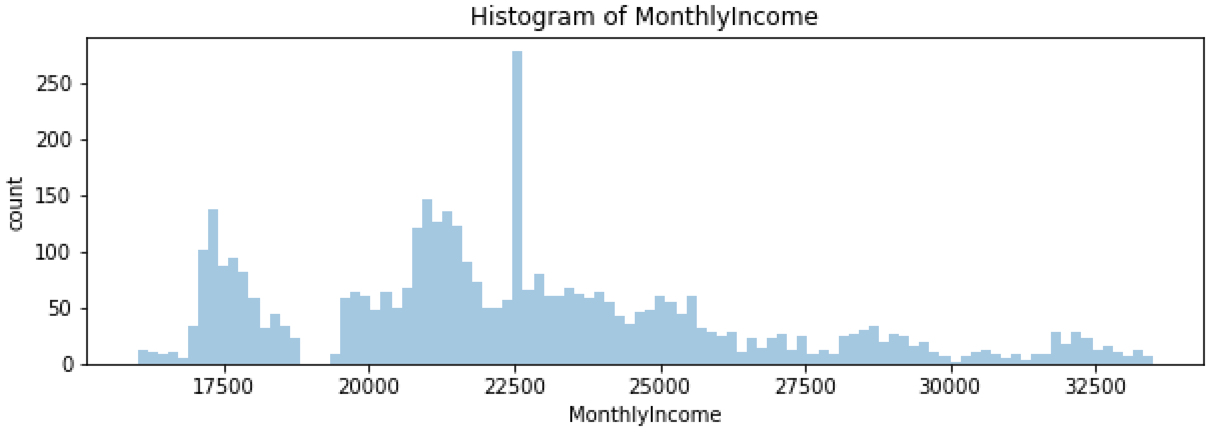
Exploratory Data Analysis Highlights





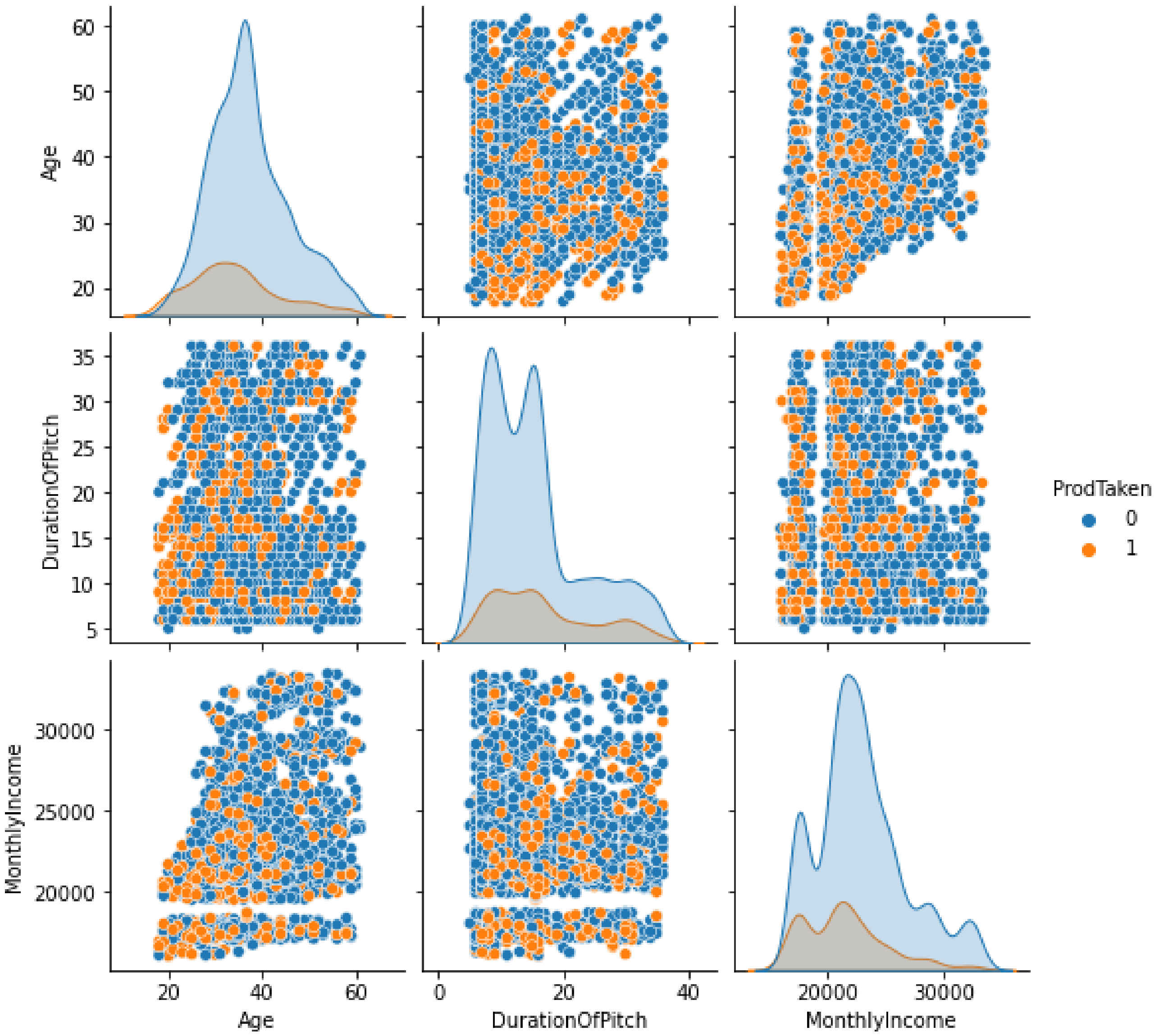
- Generally, the number of children that went with customer is 1
- The product is pitched mostly to executives and managers. However, there are more executives who ended up taking the product/package.

Exploratory Data Analysis Highlights

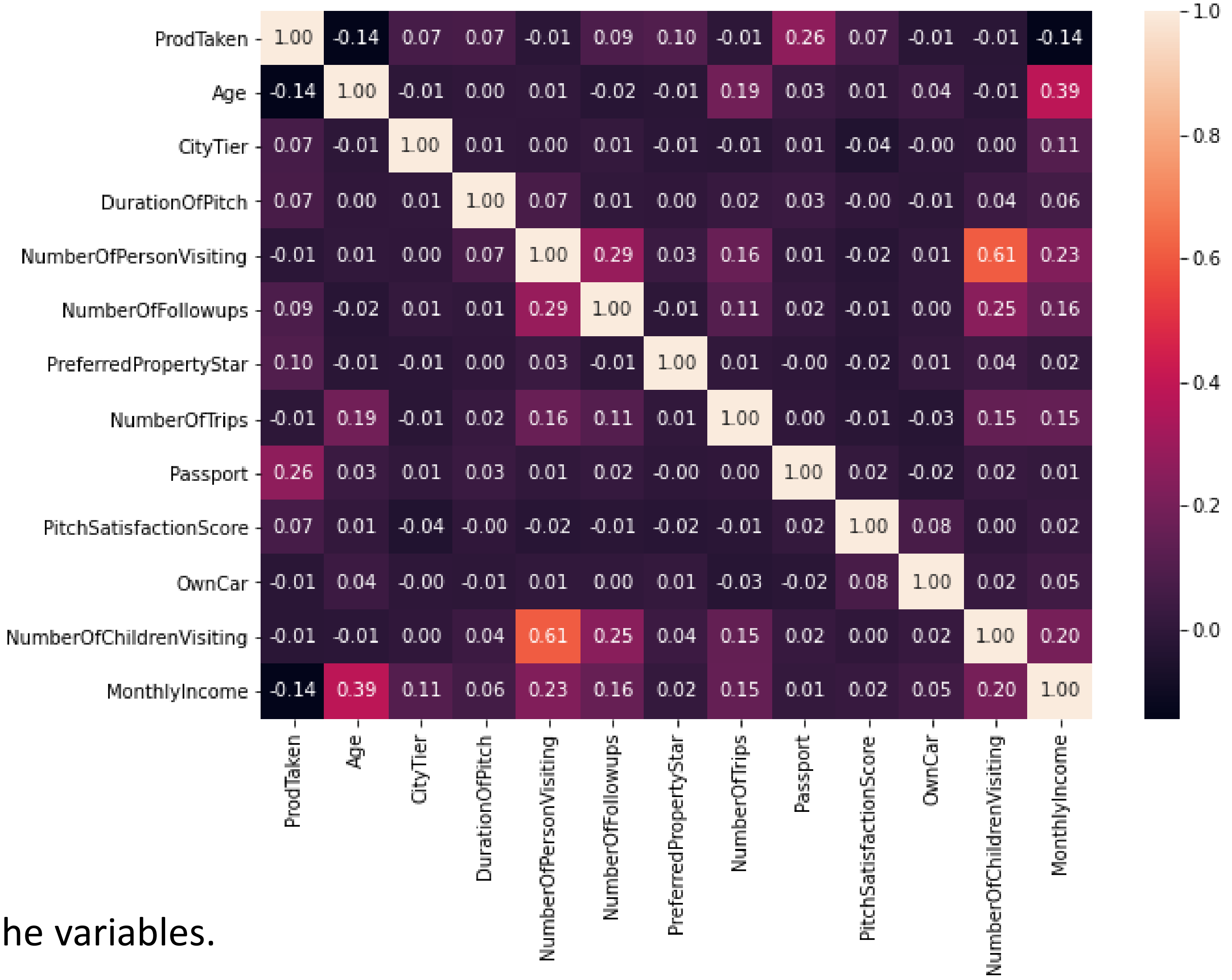


- Many of the customers have a monthly income of around 17500 to 25000.

Exploratory Data Analysis Pairplot

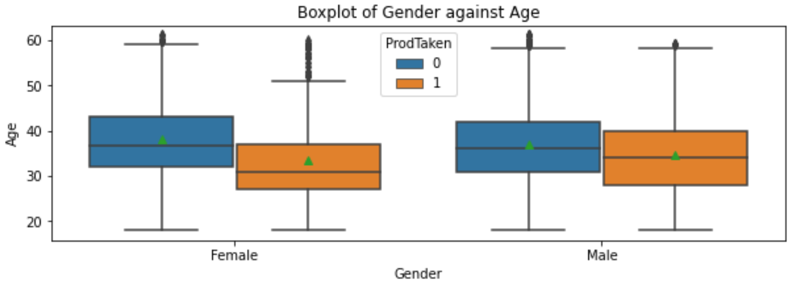


Exploratory Data Analysis Heatmap

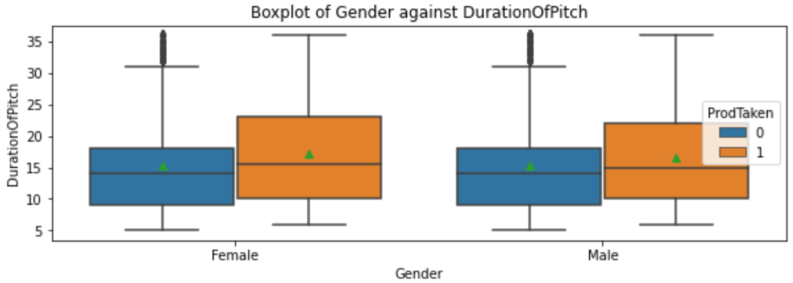


	ProdTaken	Age	CityTier	DurationOfPitch	NumberOfPersonVisiting	NumberOfFollowups	PreferredPropertyStar	NumberOfTrips	Passport	PitchSatisfactionScore	OwnCar	NumberOfChildrenVisiting	MonthlyIncome
ProdTaken	1.00	-0.14	0.07	0.07	-0.01	0.09	0.10	-0.01	0.26	0.07	-0.01	-0.01	-0.14
Age	-0.14	1.00	-0.01	0.00	0.01	-0.02	-0.01	0.19	0.03	0.01	0.04	-0.01	0.39
CityTier	0.07	-0.01	1.00	0.01	0.00	0.01	-0.01	-0.01	0.01	-0.04	-0.00	0.00	0.11
DurationOfPitch	0.07	0.00	0.01	1.00	0.07	0.01	0.00	0.02	0.03	-0.00	-0.01	0.04	0.06
NumberOfPersonVisiting	-0.01	0.01	0.00	0.07	1.00	0.29	0.03	0.16	0.01	-0.02	0.01	0.61	0.23
NumberOfFollowups	0.09	-0.02	0.01	0.01	0.29	1.00	-0.01	0.11	0.02	-0.01	0.00	0.25	0.16
PreferredPropertyStar	0.10	-0.01	-0.01	0.00	0.03	-0.01	1.00	0.01	-0.00	-0.02	0.01	0.04	0.02
NumberOfTrips	-0.01	0.19	-0.01	0.02	0.16	0.11	0.01	1.00	0.00	-0.01	-0.03	0.15	0.15
Passport	0.26	0.03	0.01	0.03	0.01	0.02	-0.00	0.00	1.00	0.02	-0.02	0.02	0.01
PitchSatisfactionScore	0.07	0.01	-0.04	-0.00	-0.02	-0.01	-0.02	-0.01	0.02	1.00	0.08	0.00	0.02
OwnCar	-0.01	0.04	-0.00	-0.01	0.01	0.00	0.01	-0.03	-0.02	0.08	1.00	0.02	0.05
NumberOfChildrenVisiting	-0.01	-0.01	0.00	0.04	0.61	0.25	0.04	0.15	0.02	0.00	0.02	1.00	0.20
MonthlyIncome	-0.14	0.39	0.11	0.06	0.23	0.16	0.02	0.15	0.01	0.02	0.05	0.20	1.00

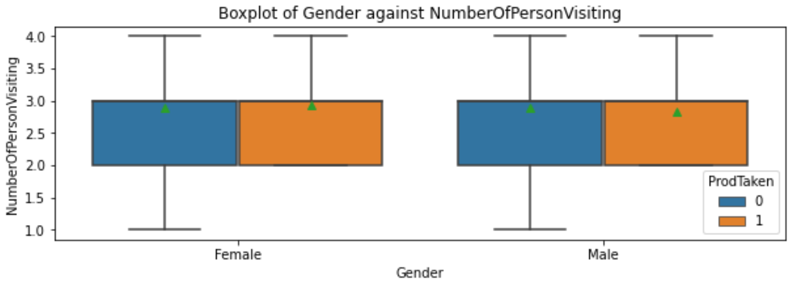
There are no notable correlation between the variables.



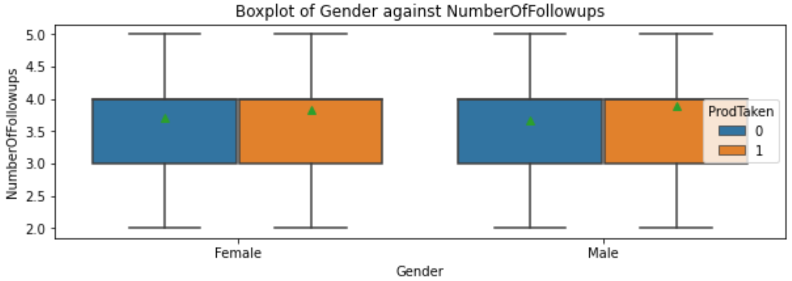
Boxplot of Gender against Age



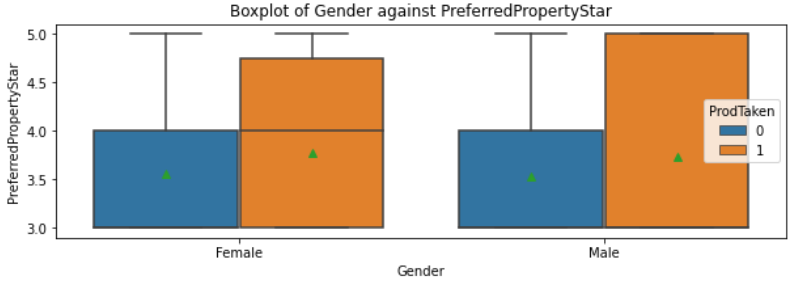
Boxplot of Gender against DurationOfPitch



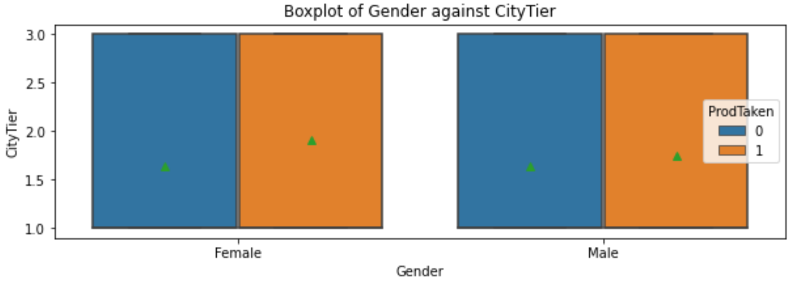
Boxplot of Gender against NumberOfPersonVisiting



Boxplot of Gender against NumberOfFollowups

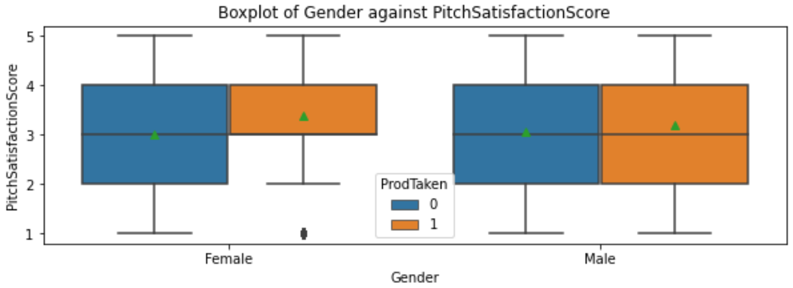


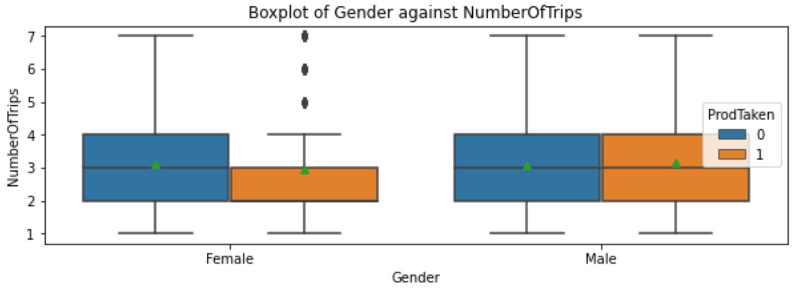
Boxplot of Gender against PreferredPropertyStar

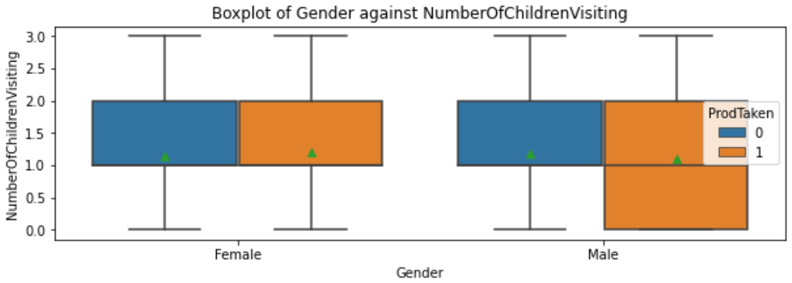


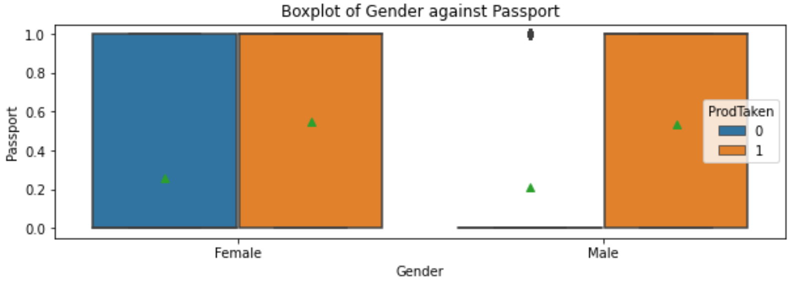
Boxplot of Gender against CityTier

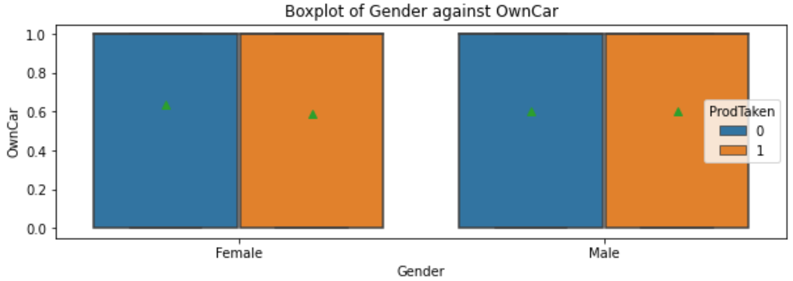
Characteristics of Product Takers

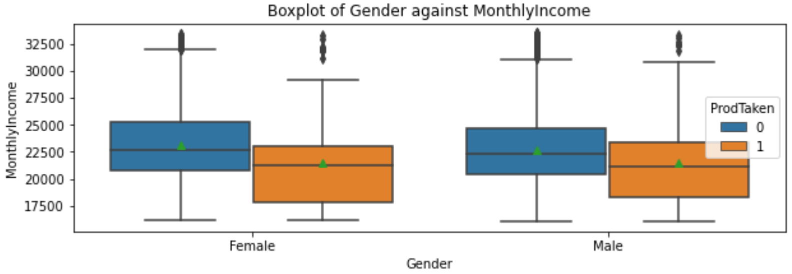








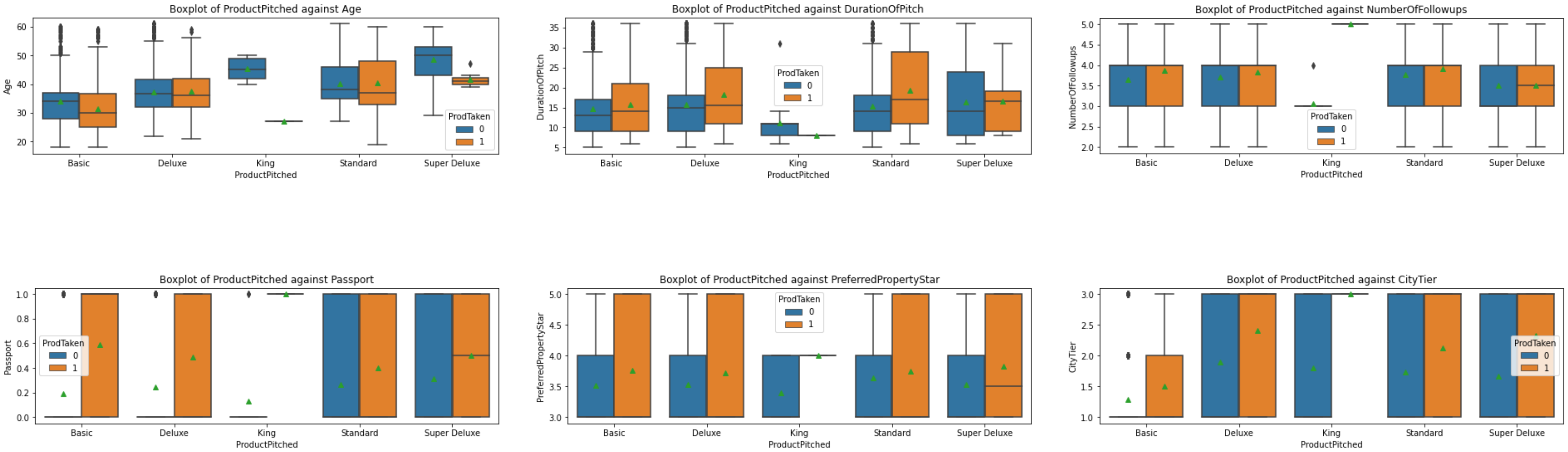




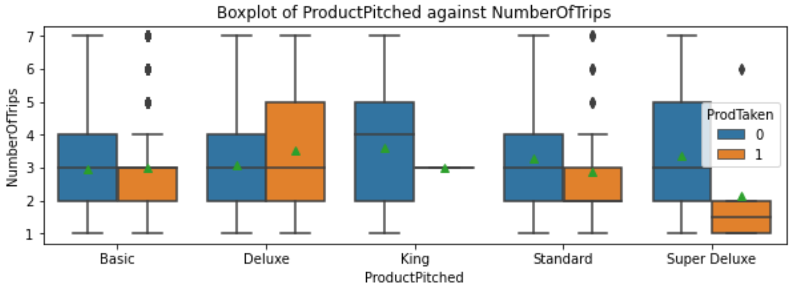
Characteristics of Product Takers

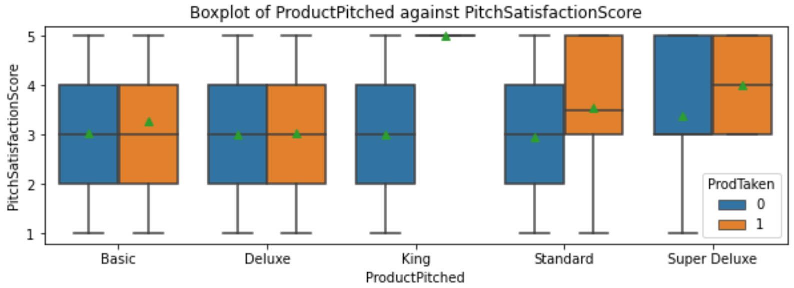
Characteristics of Product Takers

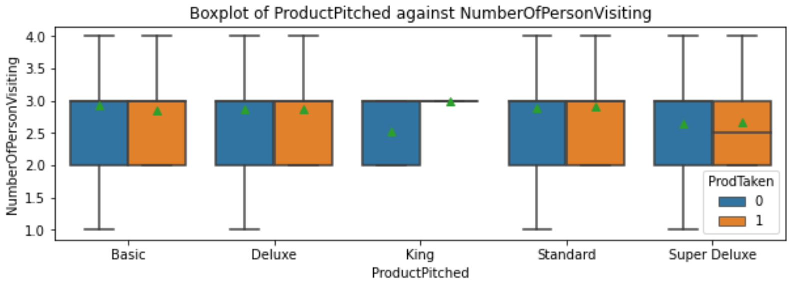
- Have lower age.
- Have longer duration of pitch.
- Have passports.
- Have somewhat higher preferred property star.
- Have somewhat higher number of followups.
- Have a higher average of city tier.
- Have higher pitch satisfaction score.
- Have less monthly income.

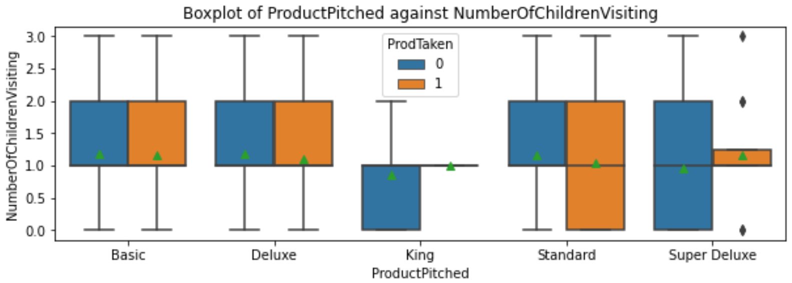


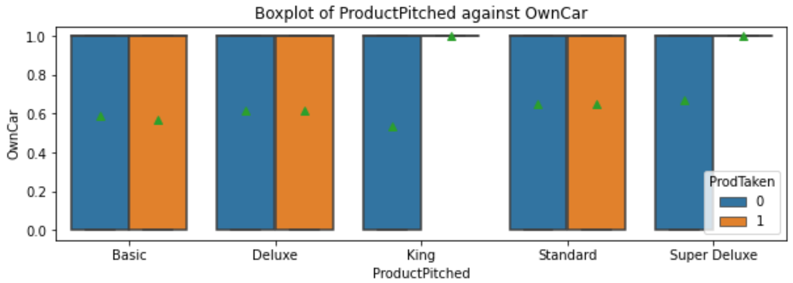
Characteristics of Customer Over Different Products

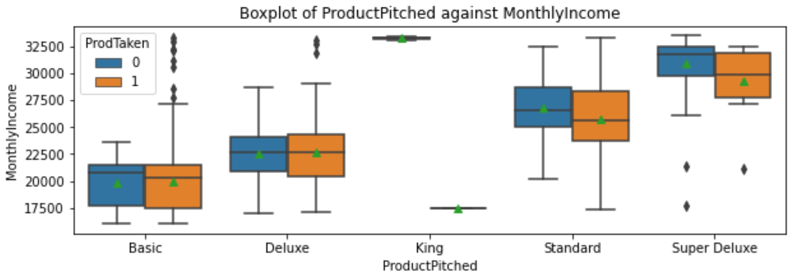












Characteristics of Customer Over Different Products

Characteristic of Basic Product Customers

a. Taken by younger people - aged around 25 to 35

b. City tier around 1-2

c. Duration of pitch around 10-20 mins

d. NumberOfPersonVisiting is around 2-3 people

e. Followed up around 3-4 times, with 4 on average

f. Average property star is around 3.8

g. Average number of trips per year is 3

h. Around 60% have passport

i. Pitch satisfaction score average is around 3.3 on average

j. Slightly less than 60% own a car

k. NumberOfChildrenVisiting is 1 on average

l. Have average salary of around 20000

Characteristic of Deluxe Product Customers

a. Taken by people aged around 30 to 40

b. City tier mostly are 2 or 3

c. Average duration of pitch is around 18 mins

d. NumberOfPersonVisiting is 2-3 people

e. Followed up around 3-4 times, with 4 on average

f. Average property star is around 3.8

g. Average number of trips per year is 3.5

h. Around 50% have passport

i. Pitch satisfaction score is around 3 on average

j. Around 60% own a car

k. NumberOfChildrenVisiting is 1 on average

l. Have average salary of around 22500

Characteristic of King Product Customers

a. The product is marketed to those with age above 40, but none took the product. The buyers are young adults slightly below 30 years old.

b. All buyers are of city tier 3.

c. Duration of pitch is slightly below 10 minutes for buyers.

d. All buyers have 3 NumberOfPersonVisiting

e. All buyers are followed up 5 times. Nonbuyers are not followed up up to 5 times.

f. All buyers have preferred property star of 4

g. All buyers have number of trips per year of 3

h. All buyers have passport.

i. All buyers have pitch satisfaction score of 5.

j. All buyers have a car.

k. All buyers have NumberOfChildrenVisiting of 1.

l. The product is marketed to those with high monthly salary (more than 32500), but buyers have low monthly salary of around 17500.

Characteristic of Standard Product Customers

a. Taken by people of age around 35 to 45 years old

b. City tier of 2 on average

c. Duration of pitch of 20 mins on average

d. NumberOfPersonVisiting is 2-3 people

e. Followed up around 3-4 times, with 4 on average

f. Average property star is around 3.8

g. Average number of trips per year of 3

h. Around 40% have passport

i. Have an average of 3.5 of pitch satisfaction score

j. More than 60% have a car

k. NumberOfChildrenVisiting is 1 on average

l. Have monthly income of around 25000 to 27500

Characteristic of Super Deluxe Product Customers

a. Product is marketed to those with age around 45-55 years old, but buyers are around 40-45 years of age.

b. City tier of 2.25 on average.

c. Average duration of pitch is around 18 mins

d. The average NumberOfPersonVisiting is around 2.75

e. Average number of followups is 3.5

f. Average number of preferred property star is around 3.9

g. The average number of trips per year is 2

h. Around 55% have passport

i. The average pitch satisfaction score is 4

j. All buyers have a car

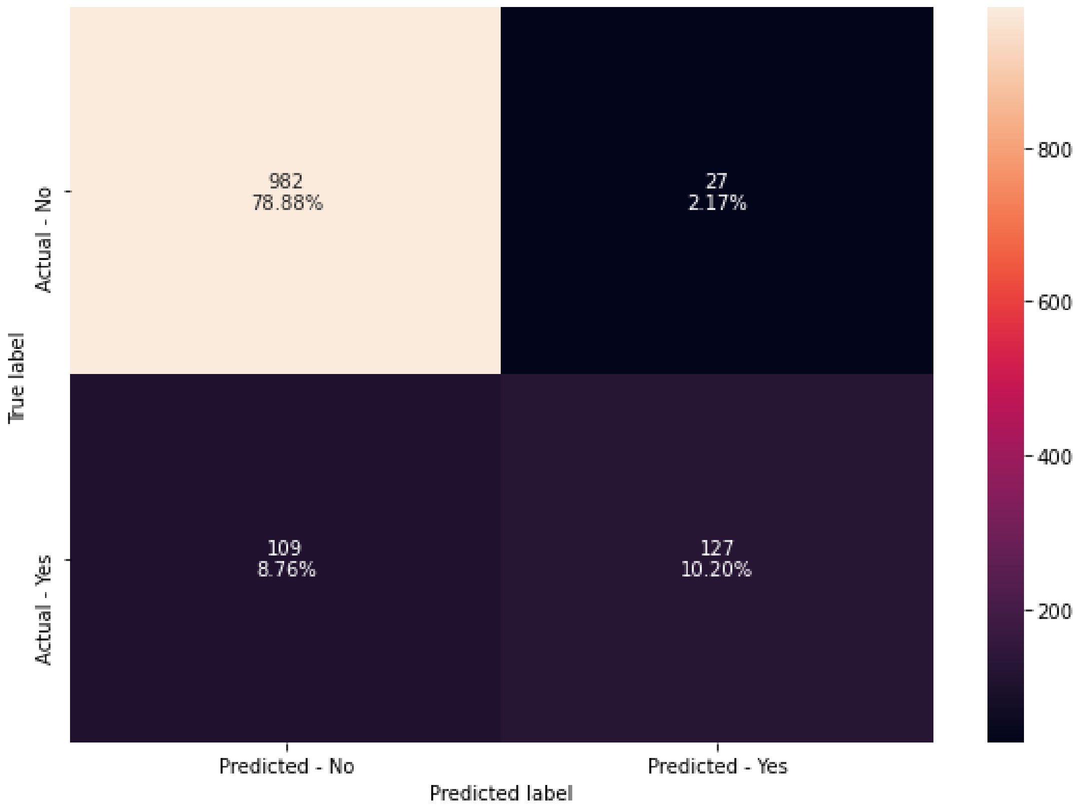
k. NumberOfChildrenVisiting is 1 on average

l. Have monthly income of around 27500 to 32500

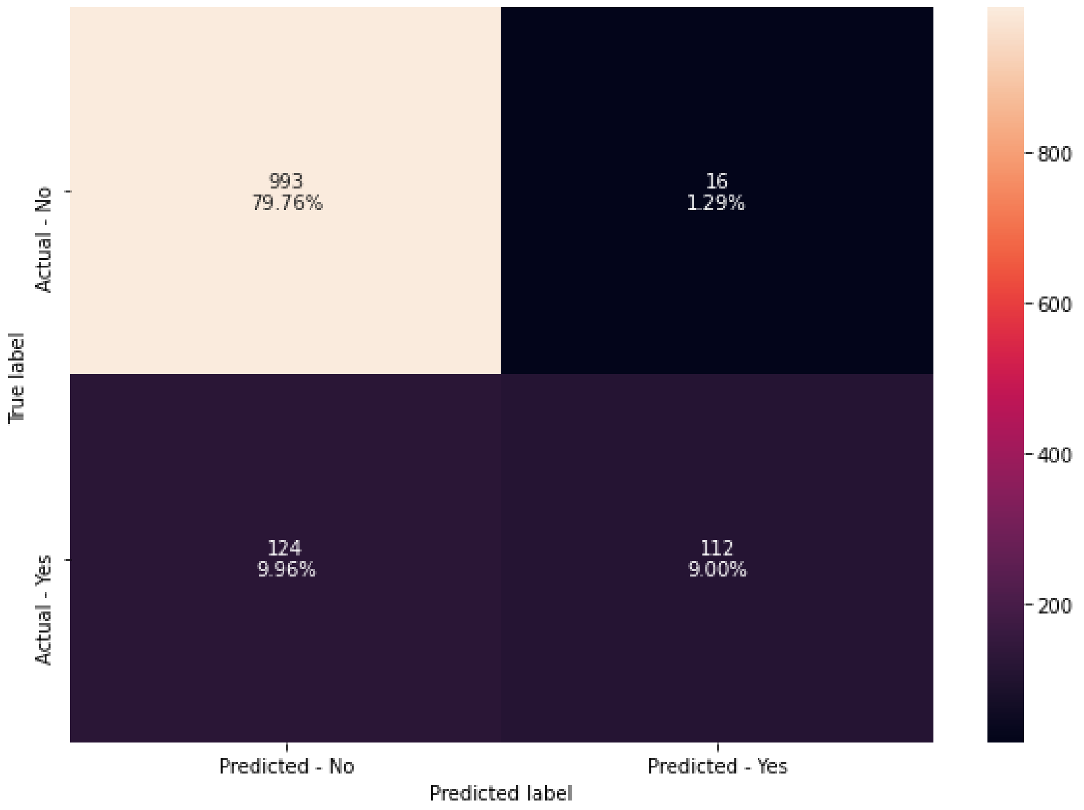
Model Performance Summary

- Both bagging and boosting methods are applied, with 70% of data used as training data and 30% of data used as testing data
- Precision is used as the main indicator of the models, as we want to avoid having false positives which could incur money loss due to waste of marketing efforts.
- We also want to maintain a good number of correct predictions.
- We also want to use F1 score as another indicator of the model.

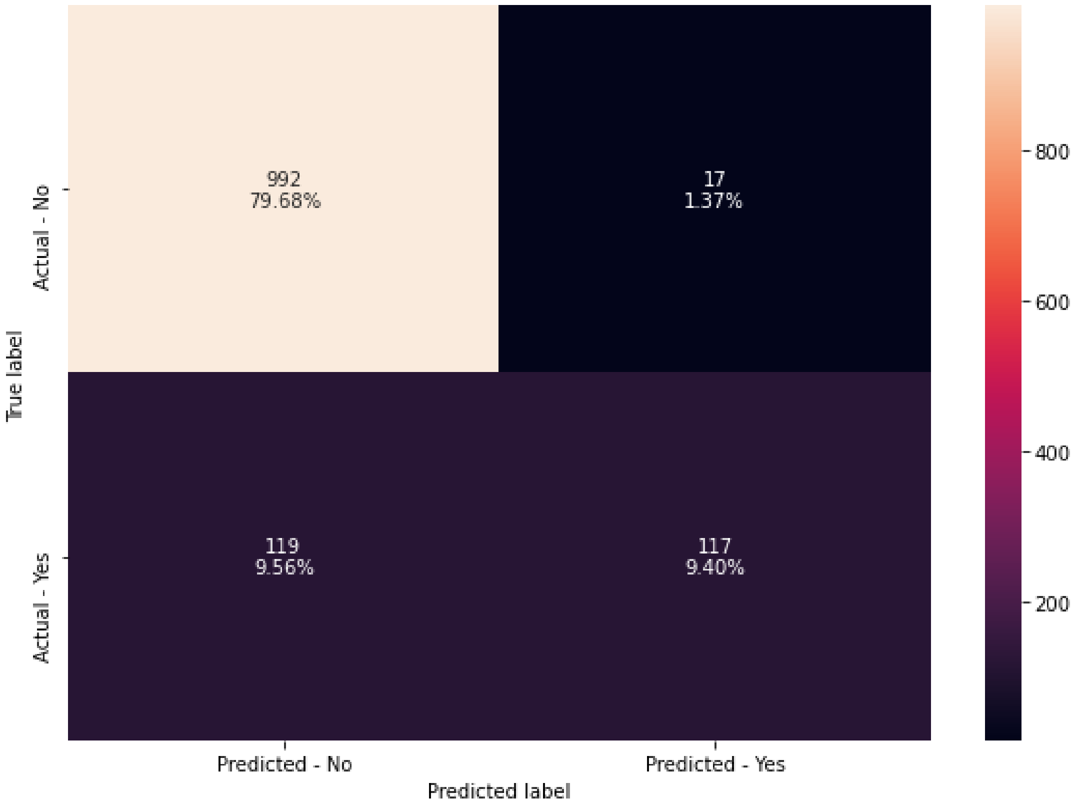
Bagging Classifier (Decision Tree as base estimator by default)



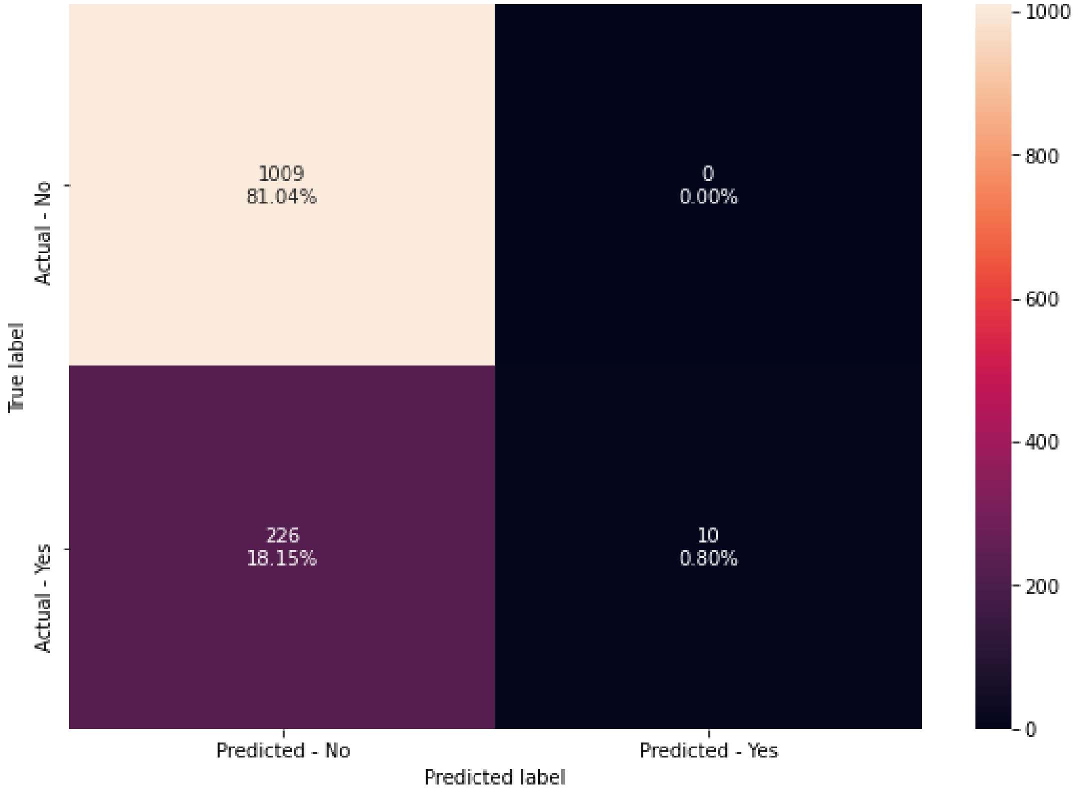
Random Forest



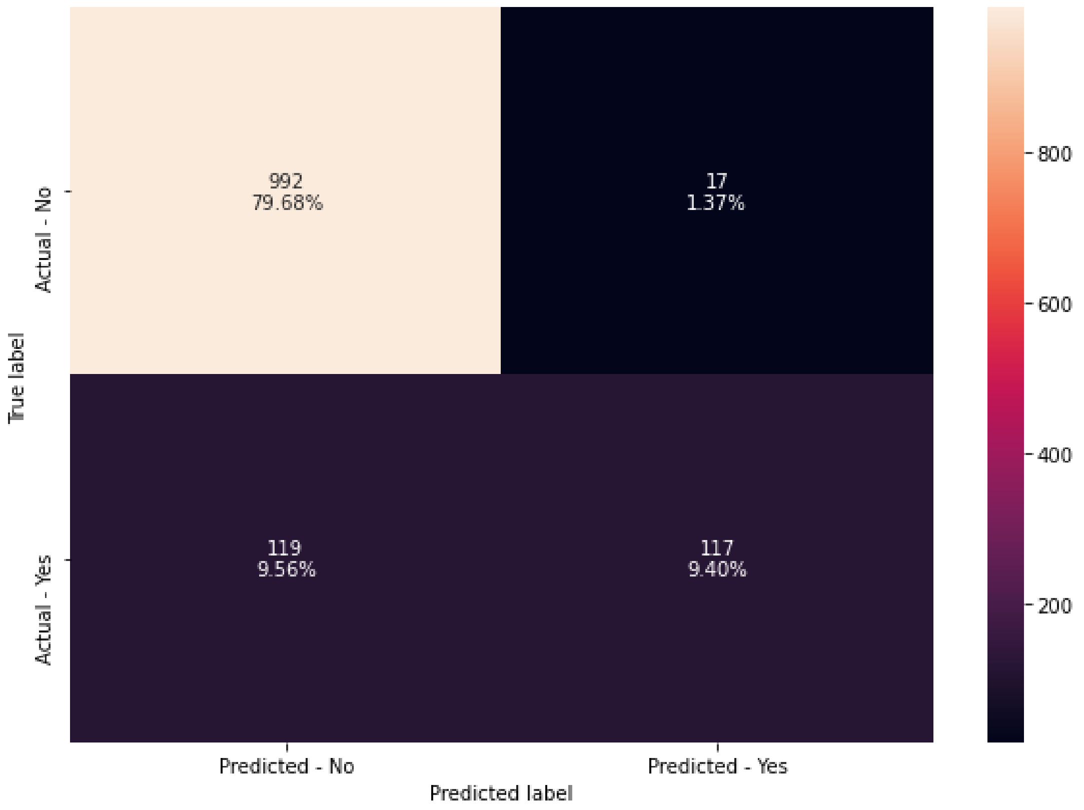
Tuned Bagging Classifier



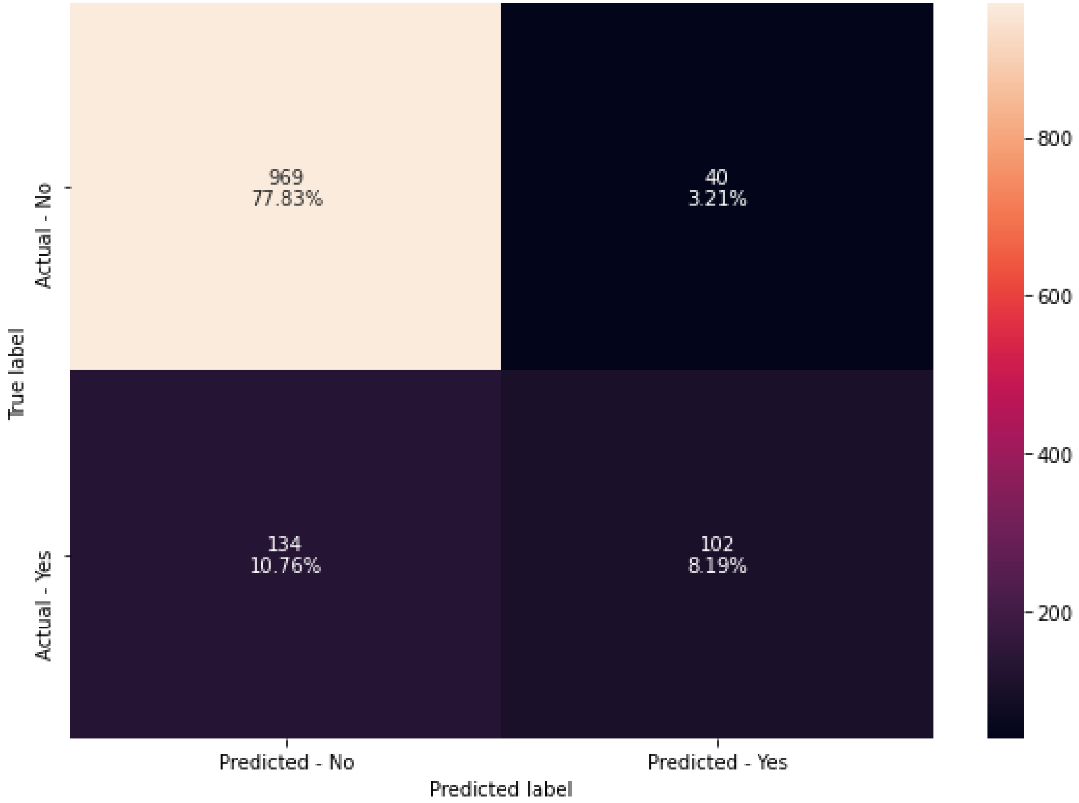
Bagging Classifier with Logistic Regression as base estimator



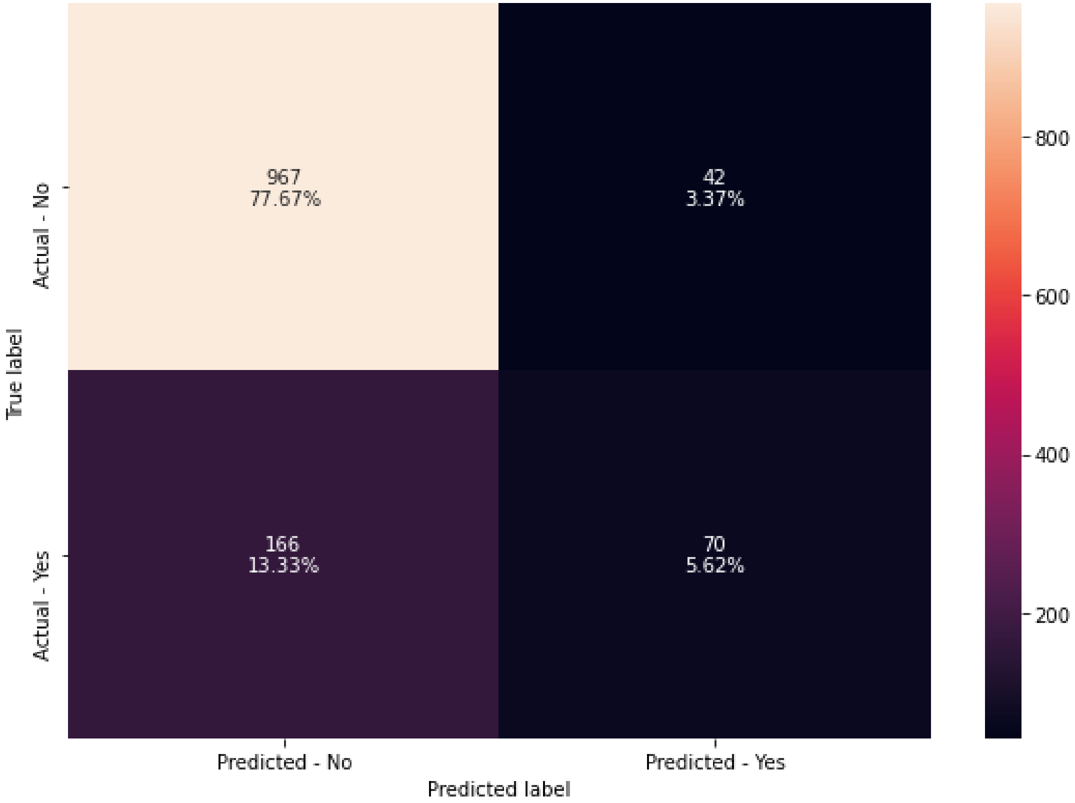
Tuned Bagging Classifier with Logistic Regression as base estimator



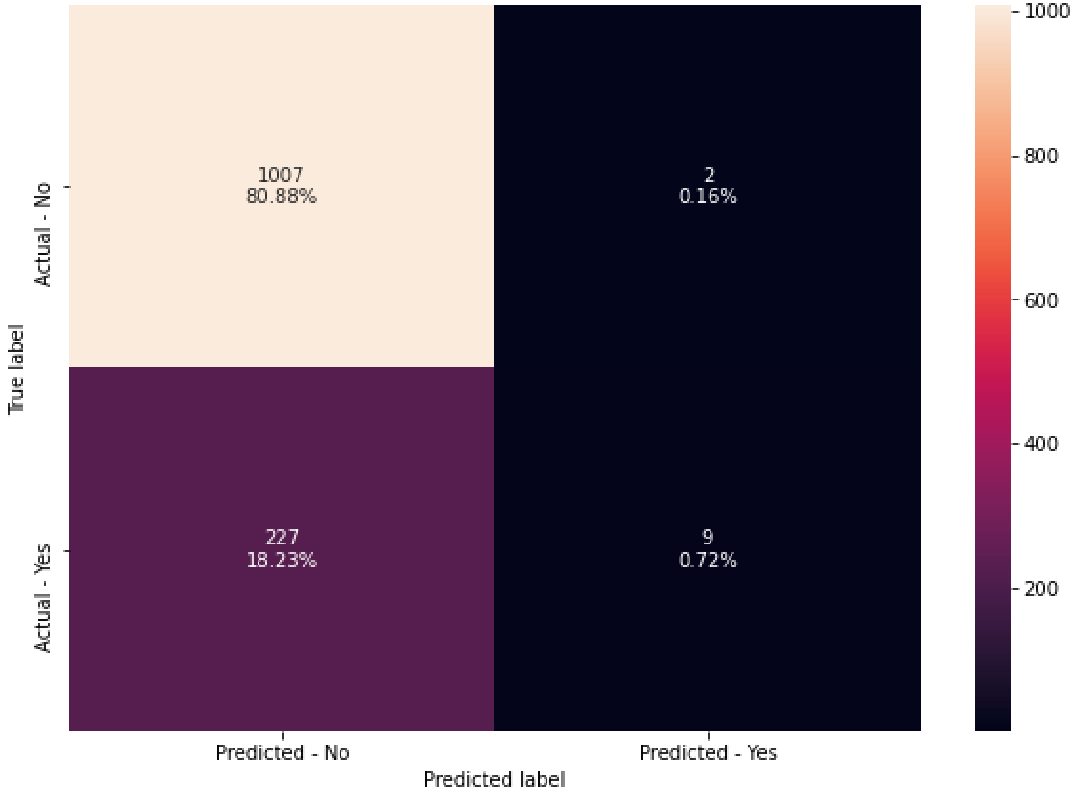
Tuned Random Forest Classifier



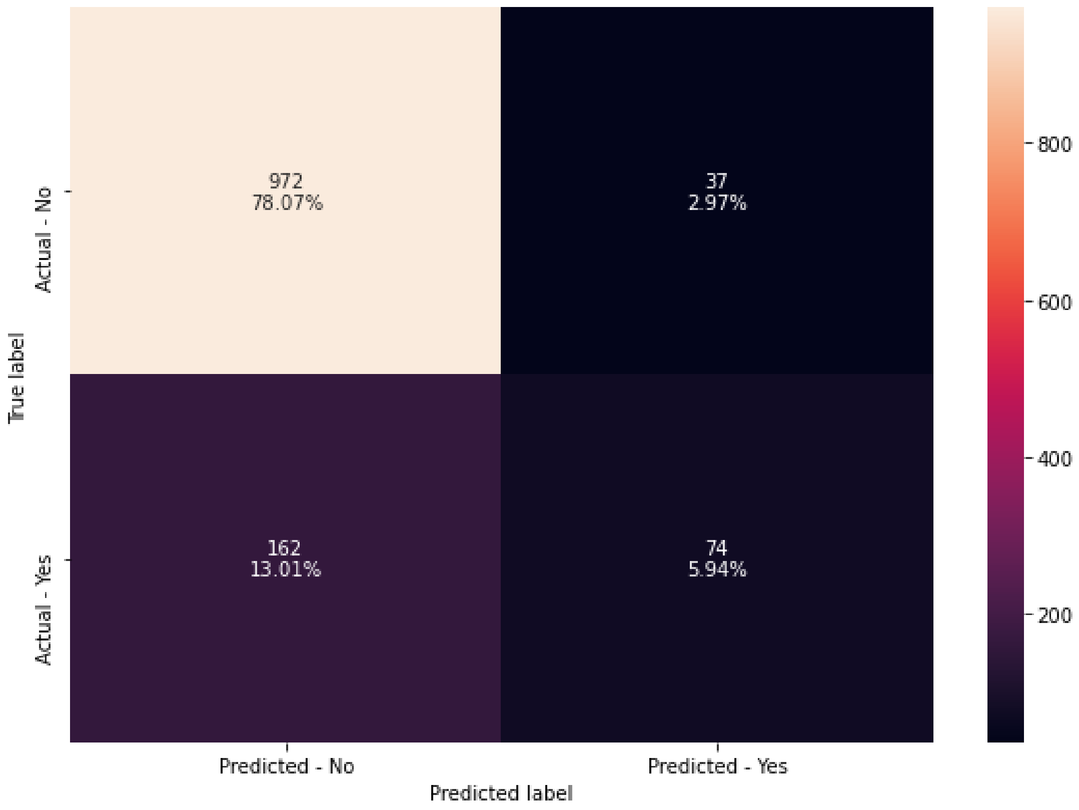
AdaBoost Classifier



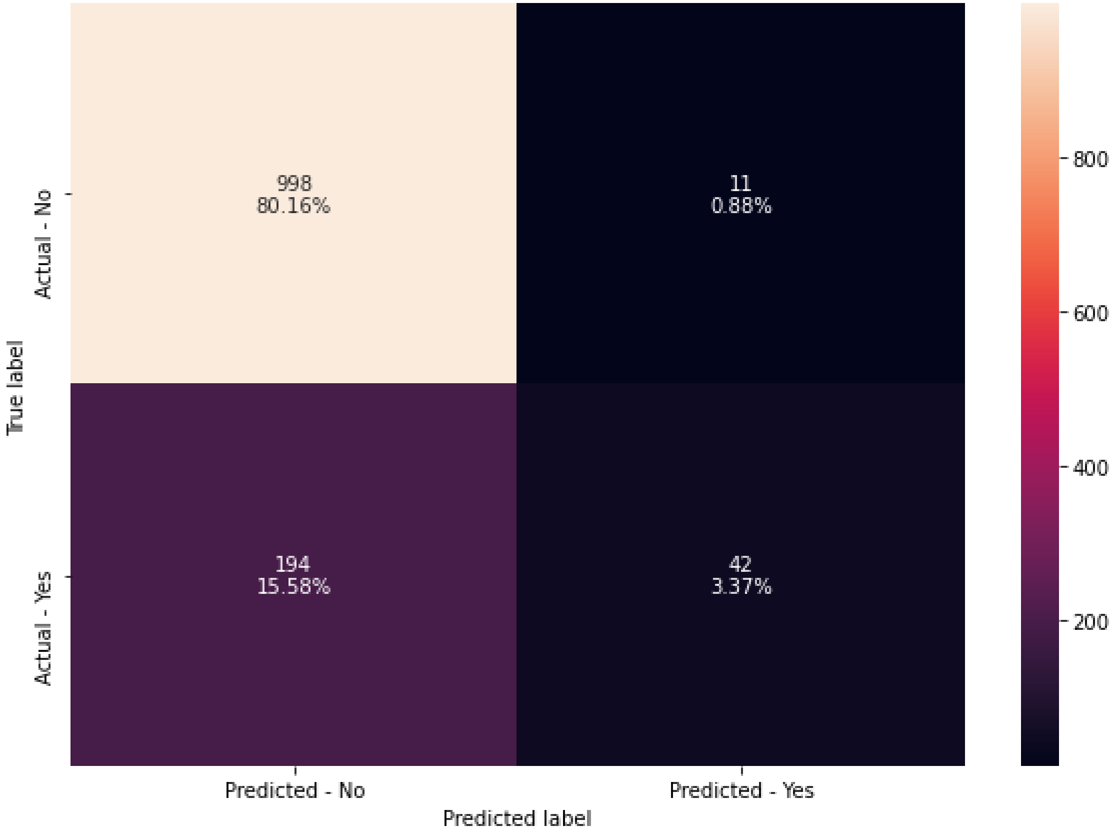
Tuned AdaBoost Classifier



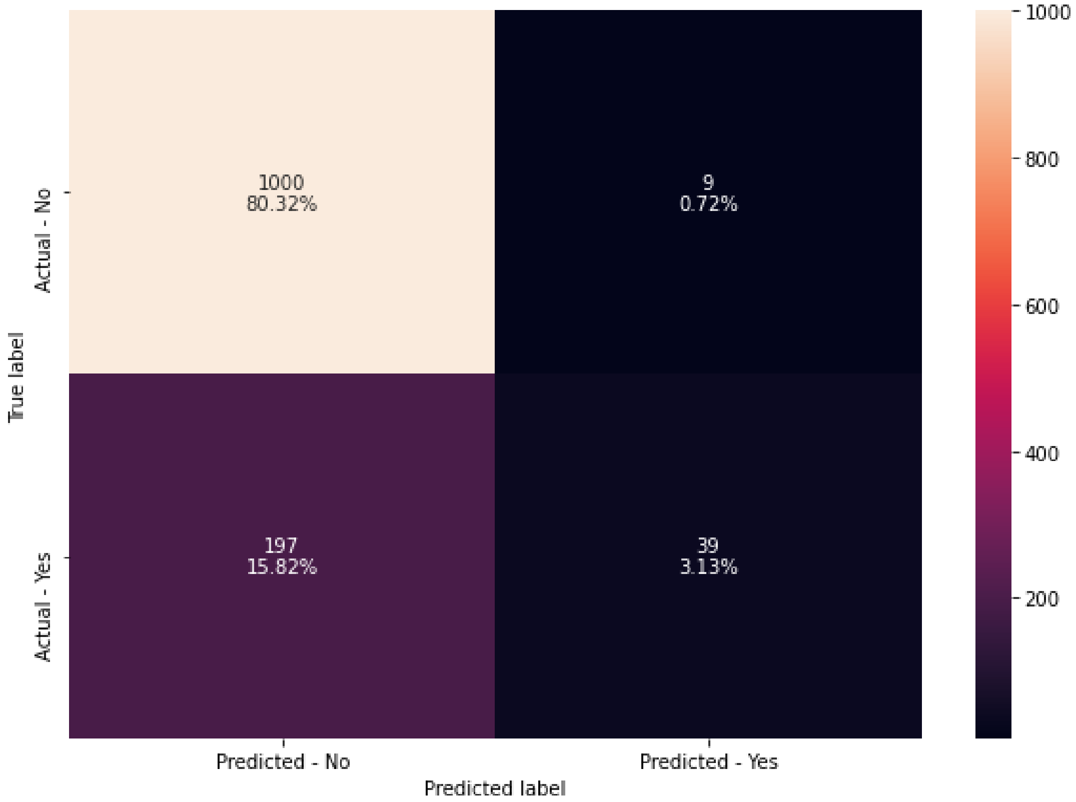
Gradient Boosting



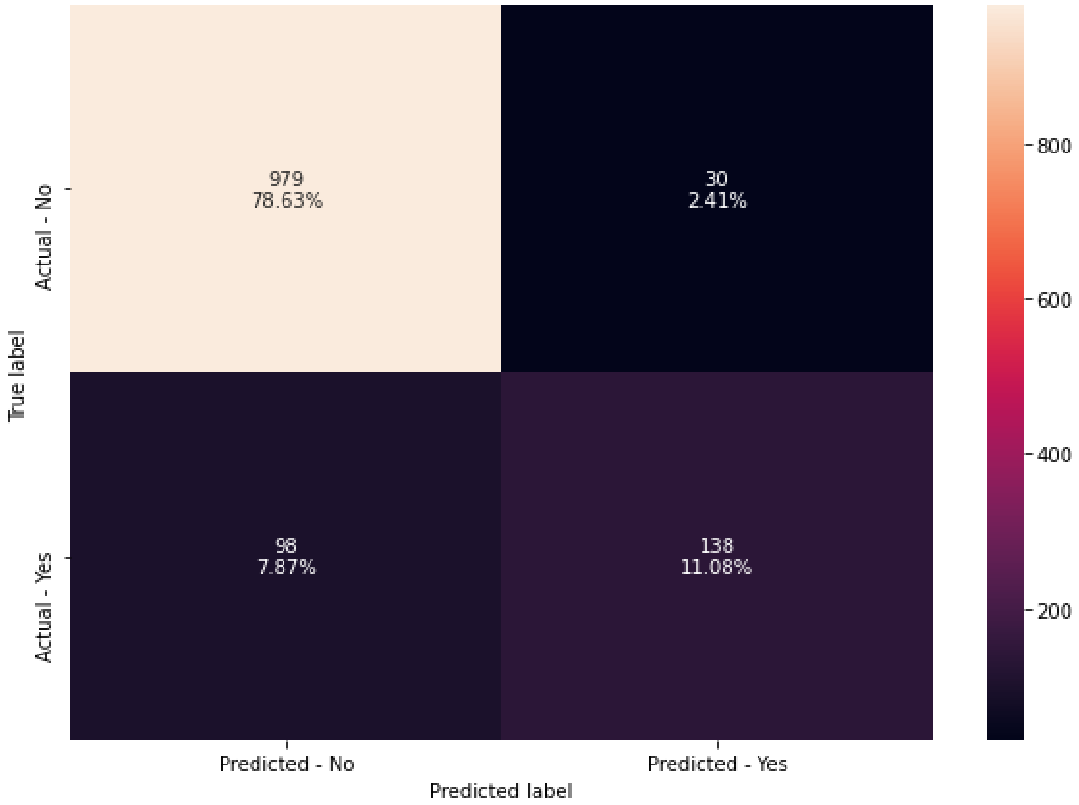
Tuned Gradient Boosting



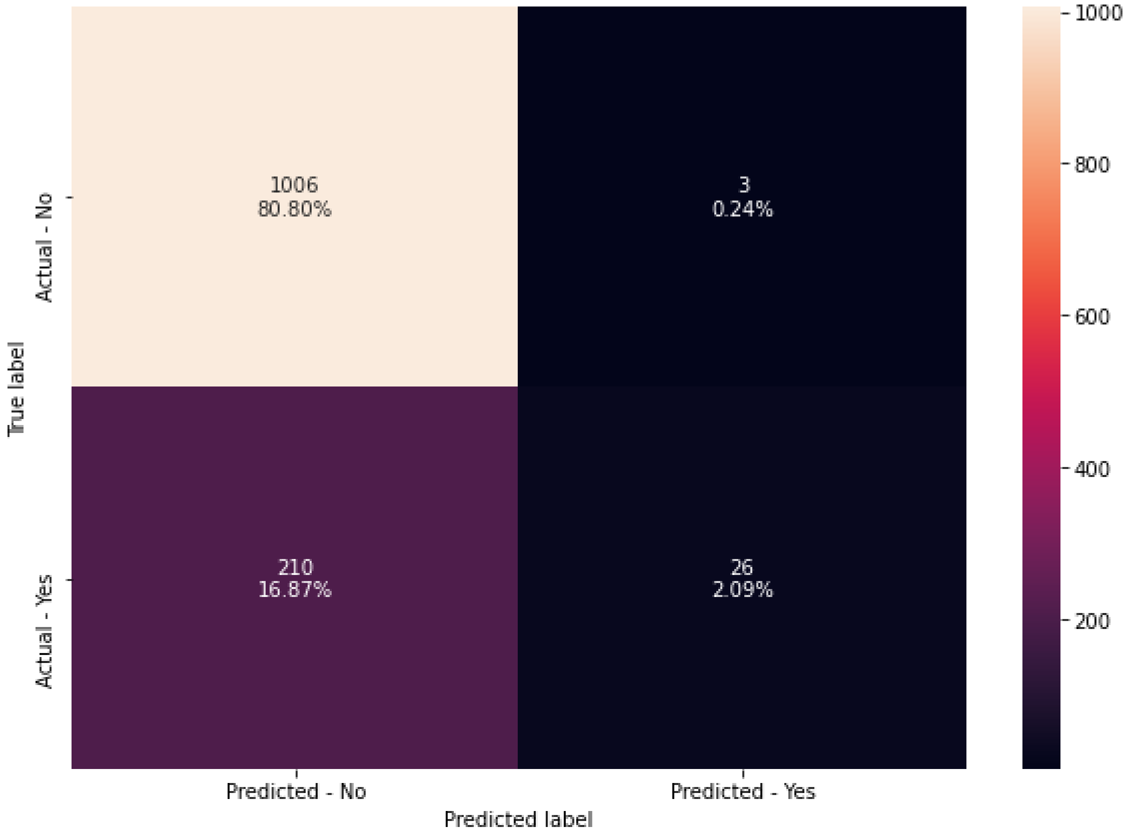
Tuned Gradient Boosting with init = Adaboost



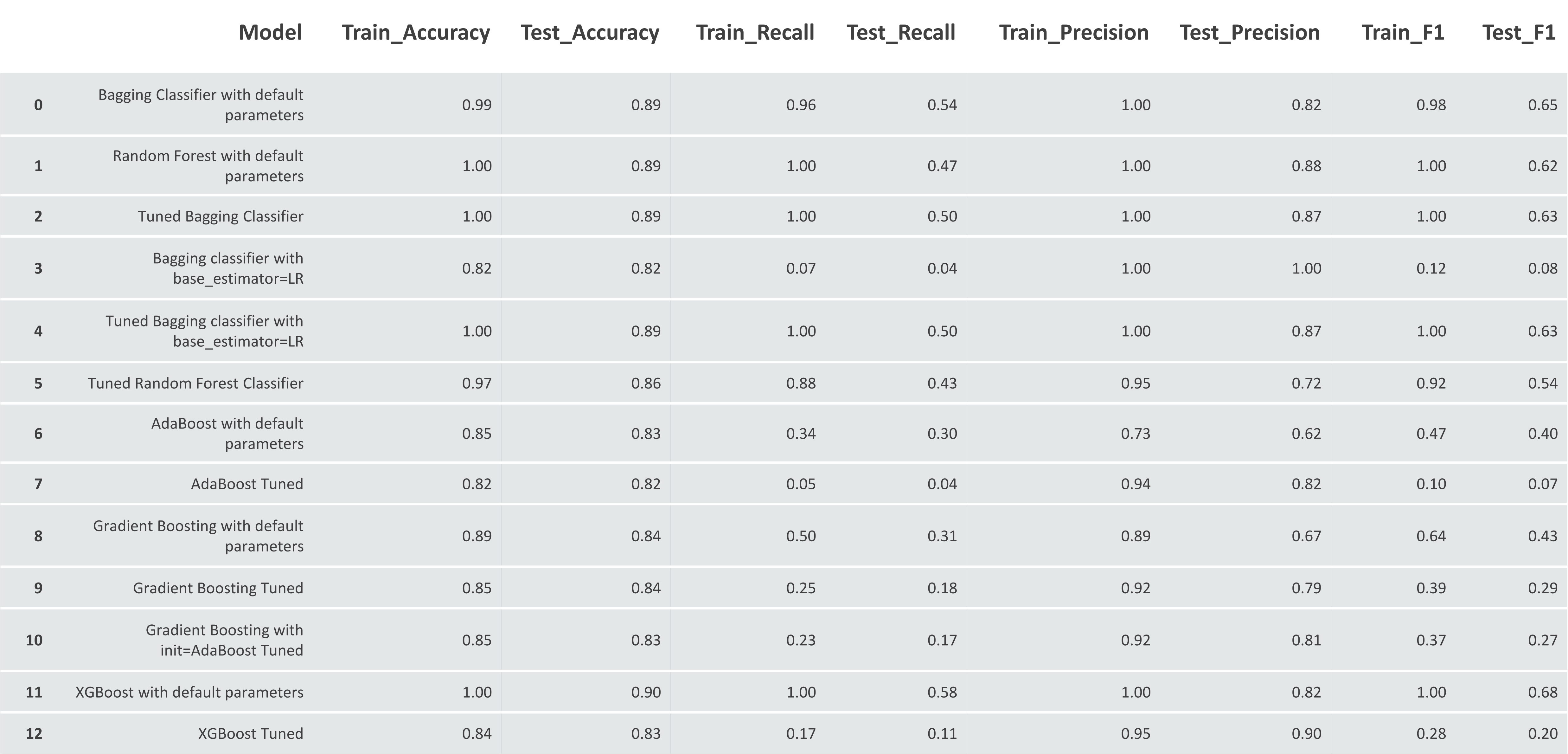
XGBoost


979
78.63%
30
2.41%
98
7.87%
138
11.08%
Actual - No
Actual - Yes
True label
Predicted - No
Predicted - Yes
Predicted label
800
600
400
200

Tuned XGBoost



Metrics of the models



	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_F1	Test_F1
0	Bagging Classifier with default parameters	0.99	0.89	0.96	0.54	1.00	0.82	0.98	0.65
1	Random Forest with default parameters	1.00	0.89	1.00	0.47	1.00	0.88	1.00	0.62
2	Tuned Bagging Classifier	1.00	0.89	1.00	0.50	1.00	0.87	1.00	0.63
3	Bagging classifier with base_estimator=LR	0.82	0.82	0.07	0.04	1.00	1.00	0.12	0.08
4	Tuned Bagging classifier with base_estimator=LR	1.00	0.89	1.00	0.50	1.00	0.87	1.00	0.63
5	Tuned Random Forest Classifier	0.97	0.86	0.88	0.43	0.95	0.72	0.92	0.54
6	AdaBoost with default parameters	0.85	0.83	0.34	0.30	0.73	0.62	0.47	0.40
7	AdaBoost Tuned	0.82	0.82	0.05	0.04	0.94	0.82	0.10	0.07
8	Gradient Boosting with default parameters	0.89	0.84	0.50	0.31	0.89	0.67	0.64	0.43
9	Gradient Boosting Tuned	0.85	0.84	0.25	0.18	0.92	0.79	0.39	0.29
10	Gradient Boosting with init=AdaBoost Tuned	0.85	0.83	0.23	0.17	0.92	0.81	0.37	0.27
11	XGBoost with default parameters	1.00	0.90	1.00	0.58	1.00	0.82	1.00	0.68
12	XGBoost Tuned	0.84	0.83	0.17	0.11	0.95	0.90	0.28	0.20

Model Performance Summary

- Based on the metrics of all the models above combined, the model with the best test precision is **bagging classifier with logistic regression as base estimator**, followed by **tuned XGBoost**.
- However, the F1 score for the two models are very low, shown by a very low number in true positives. Adding the F1 score and number of true positives as consideration, a good model for the prediction would be either **tuned bagging classifier**, **tuned bagging classifier with logistic regression as base estimator**, and **XGBoost with default parameters**.

Most Important Predictors

1st : Executive

2nd : Passport

3rd : Single

4th : Large Business

Recommendations to Marketing Team (ordered by priority)

- To stop marketing the product King to those aged above 40, as there is 0% success rate, and it would be a waste on resources
- To target more people with executive as designation
- To target more people with passports
- To target more single people
- To target more people who works in a/owns a large business
- To target customers on higher city rating
- To focus more on selling the product Basic, Deluxe and Standard as they are much more frequently bought by customers