Travel Package Purchase

Nibu Kuriakose

Jun 2021

Content

1. Introduction and Proposed Approach

Background

- 'Visit with us', a tourism company wants to expand its customer base by introducing a new 'Wellness Tourism package'.
- The company already offers five types of packages- Basic, Standard, Deluxe, Super Deluxe and King.
- Cost of marketing the current packages is high as customers are contacted at random, and only 18% of the current customers purchase the above five existing packages.
- Company wants to use existing data to perform effective targeted marketing.

Purpose and Benefits

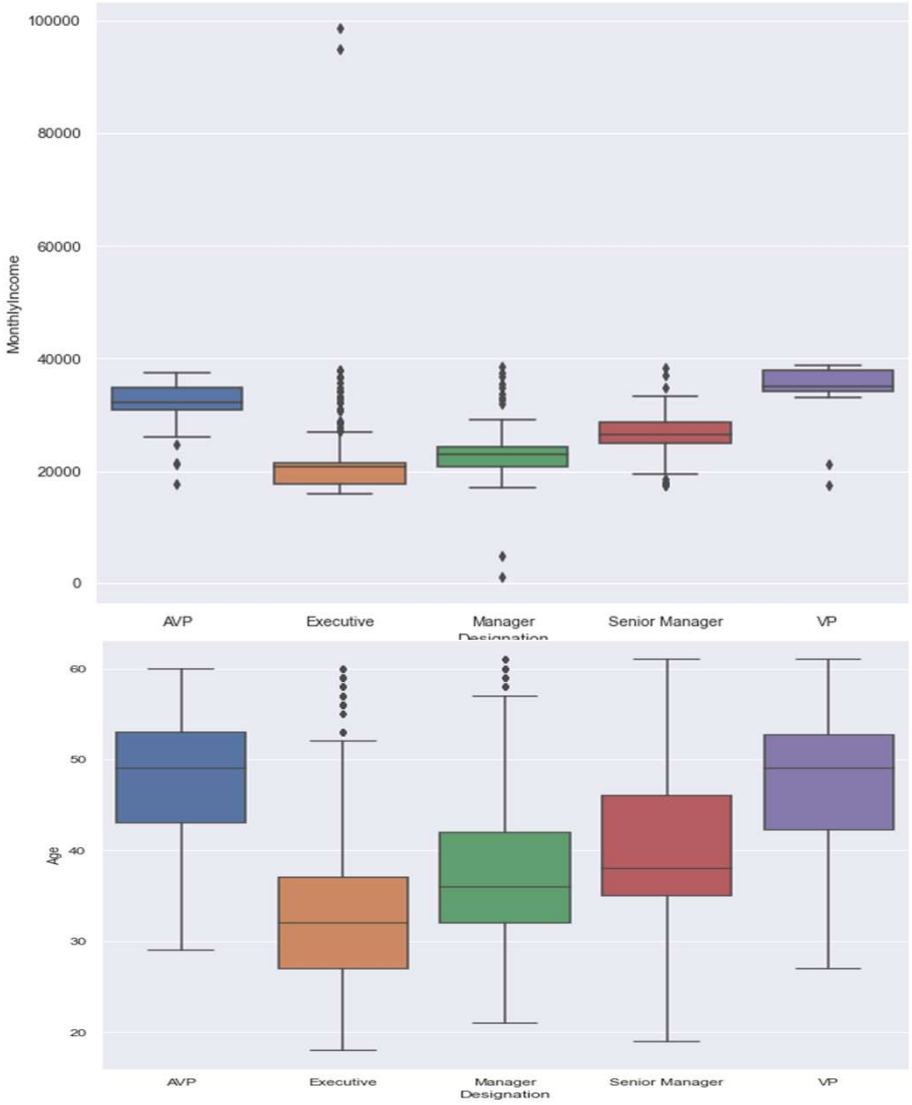
- Perform exploratory data analysis to understand the type of customers purchasing the existing five packages.
- Create a model to predict potential customers who might buy the new wellness package. This will allow for an effective targeted marketing which uses minimum resources to get maximum number of customers.
- Provide insights and recommendation based on the data.

Proposed Approach

1. Perform exploratory data analysis to find a) type of customers buying various packages and b) key factors influencing purchase of packages
2. Build prediction models using the following methods
   a) Decision Tree b) Bagging methods c) Boosting methods d) Stacking method
2. Find the appropriate performance measure to evaluate the models and perform model improvement to get the best model
3. Improve the model using additional data gathered when selling wellness package.

2. Dataset Information and Feature Engineering

- There are 4,888 samples
- Each sample has 20 attributes
- It is an imbalanced dataset as only 18% of customers have taken a package which is the target dependent variable for building the model. Hence, a class weight has been used to compensate for this.
- Data Cleaning:
  - ✓ Converted 'unmarried' to 'single'
  - ✓ Changed 'Fe Male' to 'Female'
- Feature Engineering- Converted designation into the numeric field in the order of hierarchy after reviewing the average salary and age for various designations.
  - ✓ Executive-1, Manager- 2 , Senior Manager- 3,AVP- 4 ,VP- 5

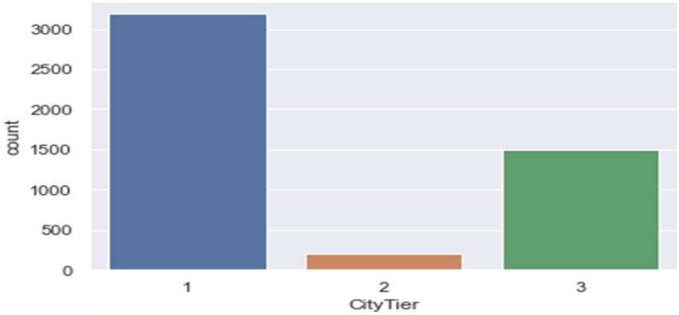


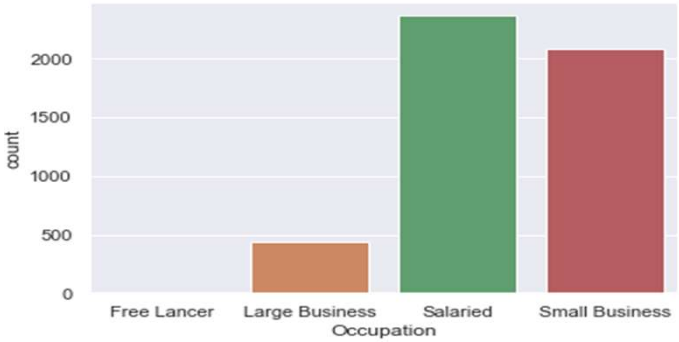
3.1 Univariate Analysis

Attribute	Description
Age	• Slightly right skewed with 50% of customers between 31 and 44 years.
Duration of pitch	• Right skewed with 50% of the customers being pitched for 9 to 20 minutes.
Number of person visiting	• Majority of customers travel in group of 3 and only 0.7% travelling alone.
Number of follow-ups	• Most of the customers are followed up 4 times with few customers even followed-up 6 times.
Preferred Property Star	• Majority of customers (62%)prefer 3 star which is the minimum preferred star too. Remaining prefer 4 star (19%) or 5 stars (20%).
Number of trips	• Most customers have done 2 trips.
Pitch Satisfaction Score	• Slightly right skewed with majority of the customers giving a score of 3, followed by the highest score of 5. 50% of customers give score of 3 or more.
Number of Children Visiting	• Majority of customers have 1 child coming for the trip, with few customers having 3 children in the trip.
Monthly Income	• 50% of the customers have a monthly incomes of $20k to $25k, with max. salary of $98k.

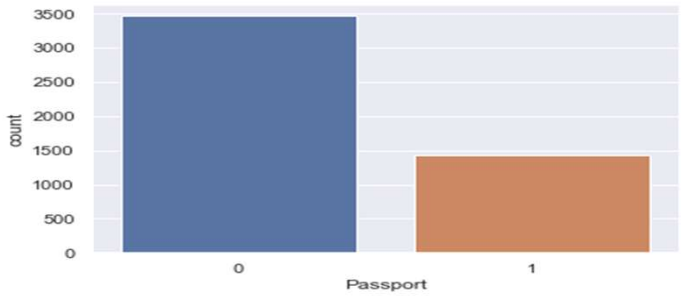
	Age	Duration of Pitch	Number of Person Visiting	Number of Follow ups	Preferred Property Star	Number of Trips	Pitch Satisfaction Score	Number of Children Visiting	Monthly Income
count	4662.0	4637.0	4888.0	4843.0	4862.0	4748.0	4888.0	4822.0	4655.0
mean	37.6	15.5	2.9	3.7	3.6	3.2	3.1	1.2	23619.9
std	9.3	8.5	0.7	1.0	0.8	1.8	1.4	0.9	5380.7
min	18.0	5.0	1.0	1.0	3.0	1.0	1.0	0.0	1000.0
25%	31.0	9.0	2.0	3.0	3.0	2.0	2.0	1.0	20346.0
50%	36.0	13.0	3.0	4.0	3.0	3.0	3.0	1.0	22347.0
75%	44.0	20.0	3.0	4.0	4.0	4.0	4.0	2.0	25571.0
max	61.0	127.0	5.0	6.0	5.0	22.0	5.0	3.0	98678.0

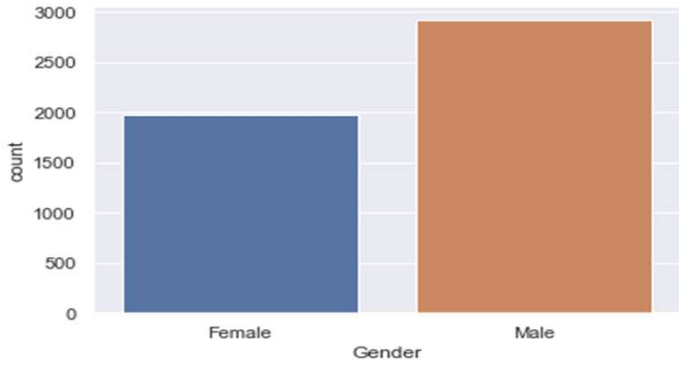
3.1 Univariate Analysis

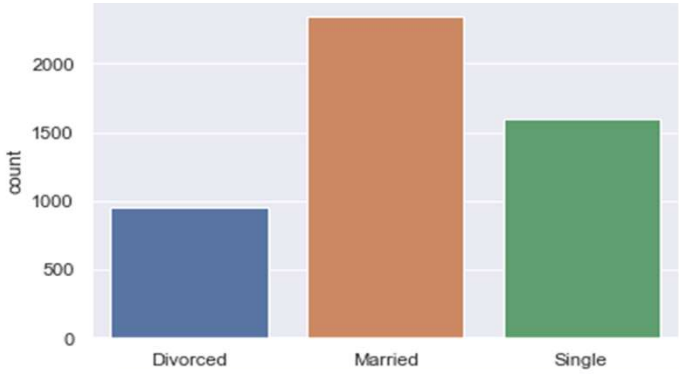


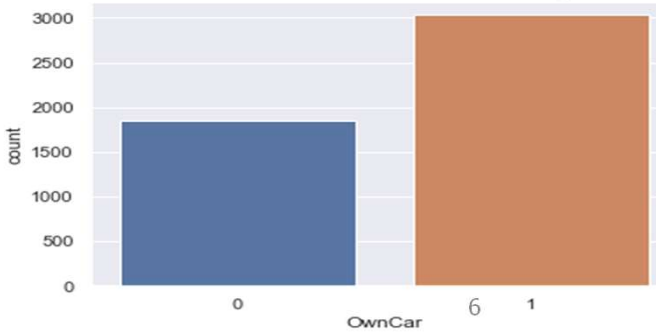


- Majority of customers are from tier-1 followed by tier-3.
- Majority of the customers are either salaried or has a small business. Majority of customers are executives or managers, with fewer customers holding senior roles such as AVP and VP.
- There are more male customers than females. Majority of customers are married.
- Majority of customers do not have a passport.
- Majority of customers own a car.

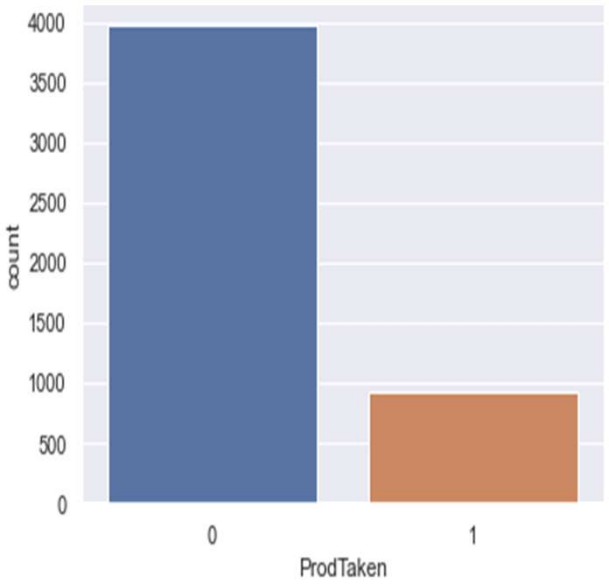




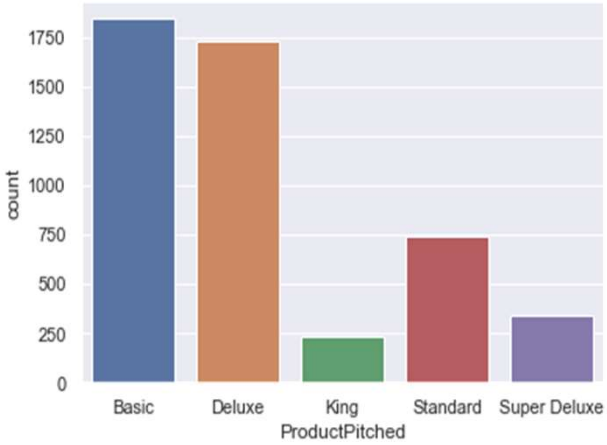


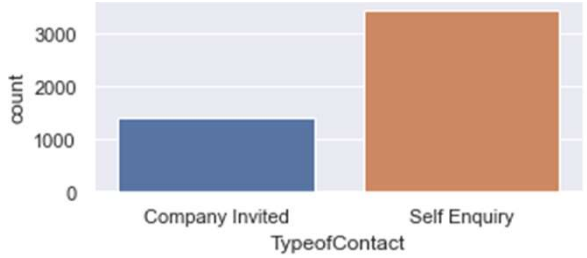


3.1 Univariate Analysis

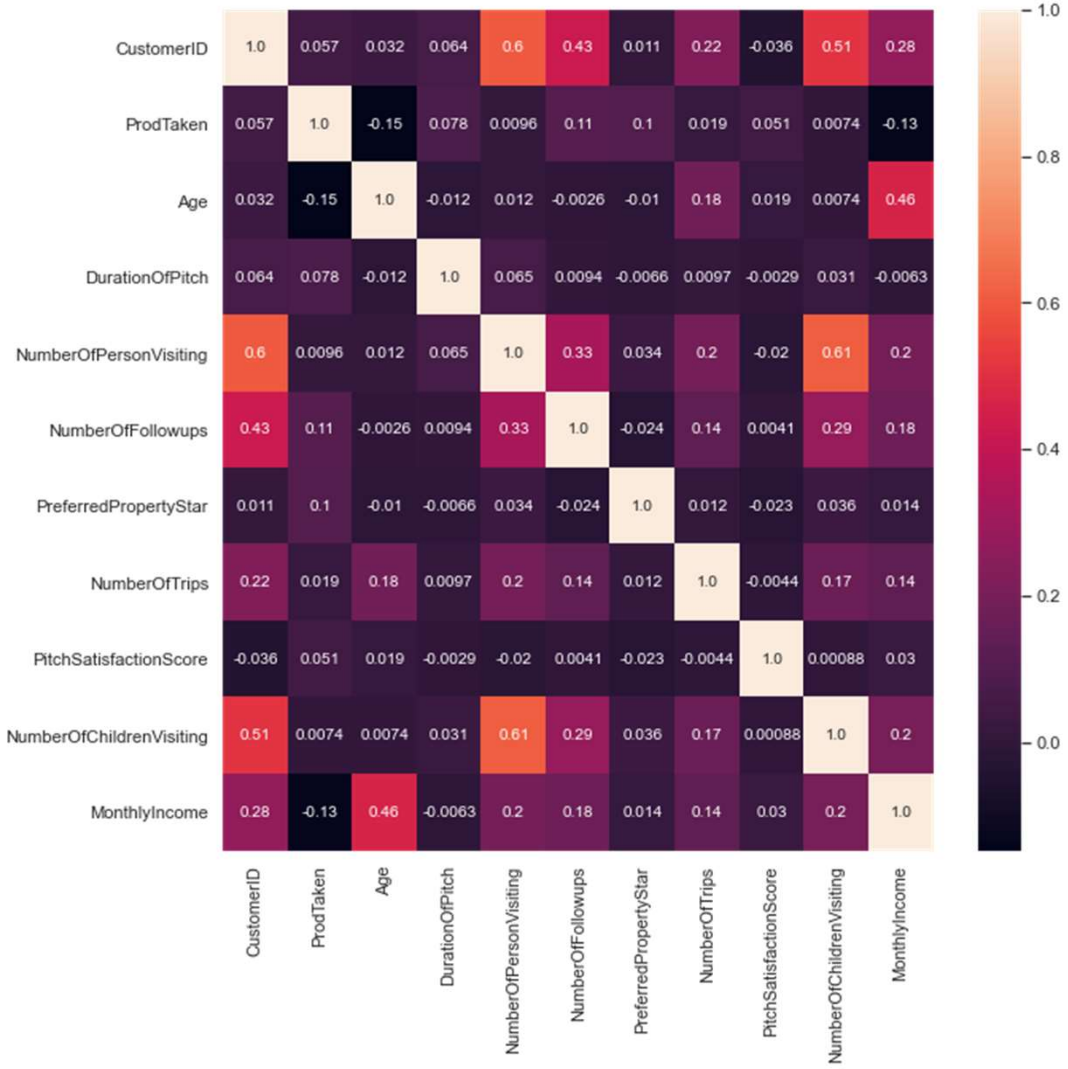


- Only 18% of the customers have taken the package.
- Majority of the contact was initiated by the customer rather than the company.
- Majority of customers were offered the basic and deluxe product, and king was pitched the least.



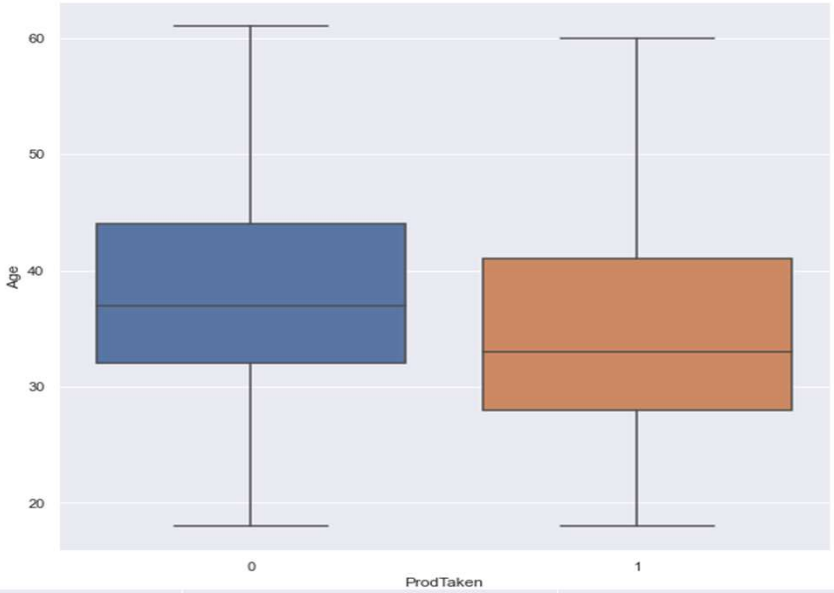


3.2 Multivariate Analysis

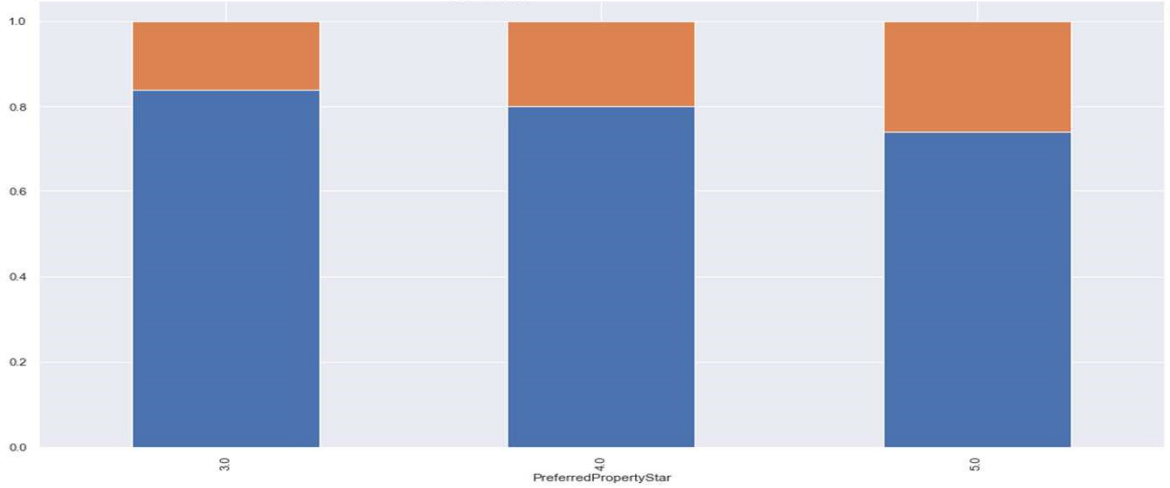


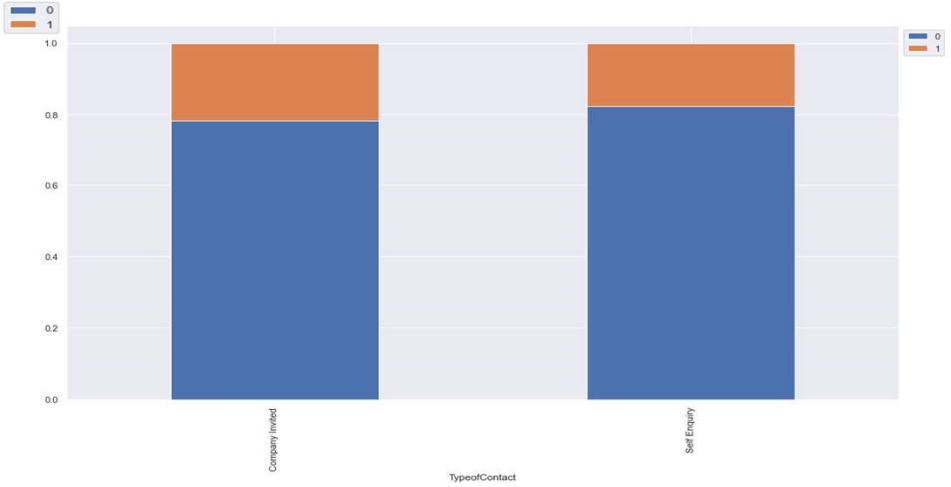
- Customers taking the package is not highly correlated to any attribute.
- Age and monthly income are positively correlated.
- Number of persons visiting is positively correlated to number of children visiting and number of follow-ups.

4. Package Purchase- Key Factors (I/III)

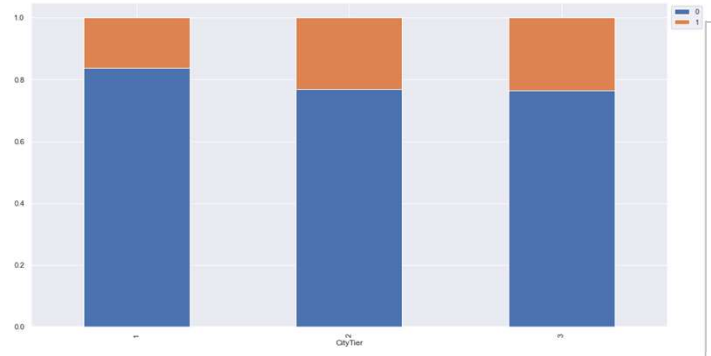


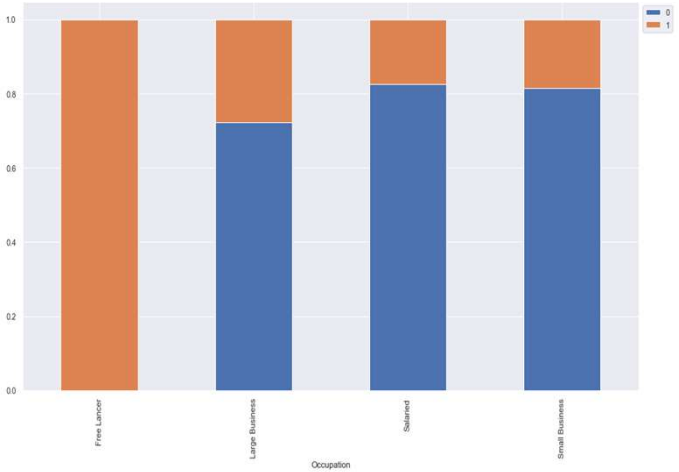
- More **younger customers** (average of early 30s) have purchased packages.
- Customers who have a **higher preferred property star** (4 or 5) have purchased more packages than ones with 3 star as their preference.
- Customers who are **contacted by company** have purchased more packages.



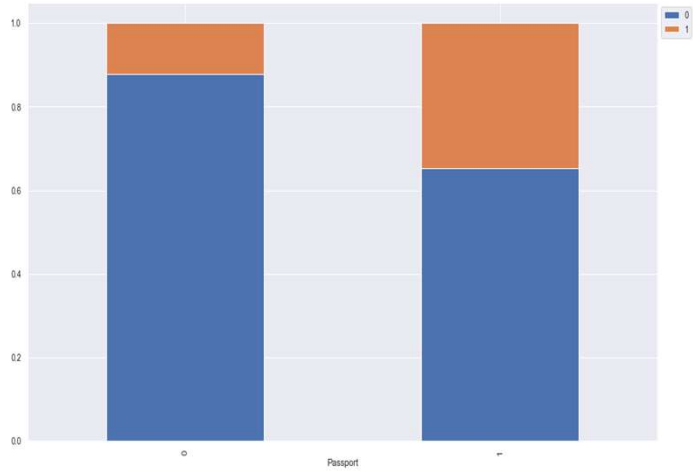


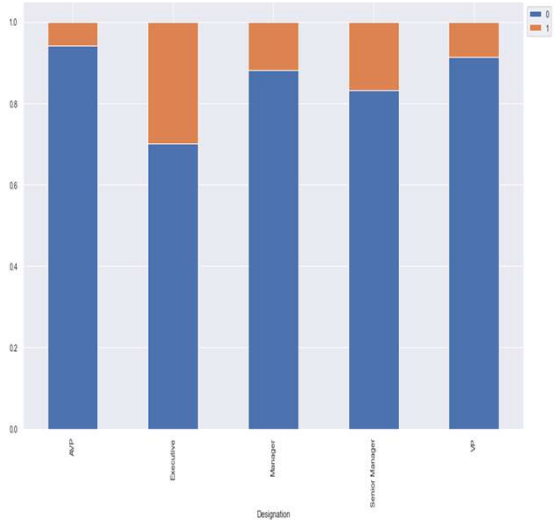
4. Package Purchase- Key Factors (II/III)



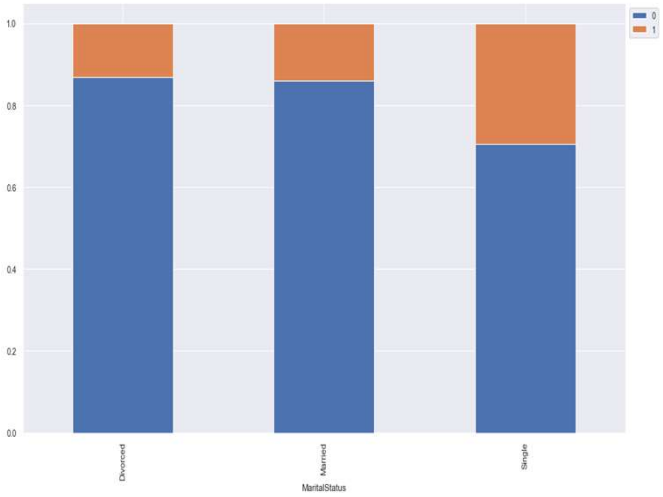


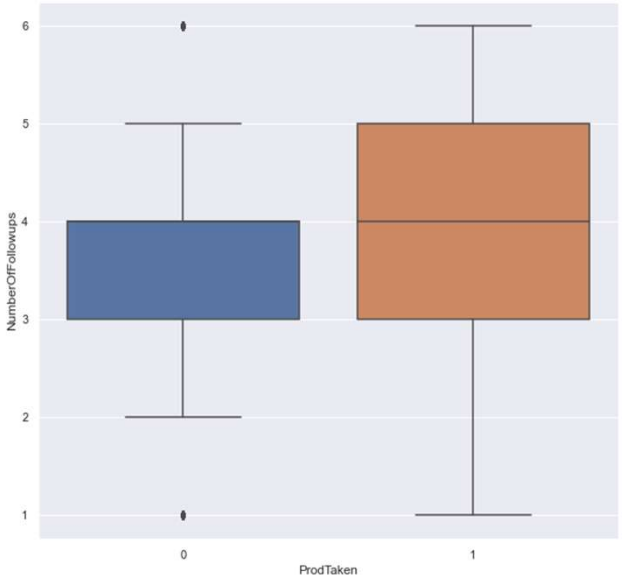
- Customers in tier-3 and tier-2 cities purchase more packages more compared to tier-1.
- Customers with large business or free lancers purchase more than salaried or customers with a small business.
- Customers with passport purchase more packages than ones without a passport.
- Customers on relatively junior roles such as executives and managers purchase packages more than ones in senior roles such as AVP and VP.



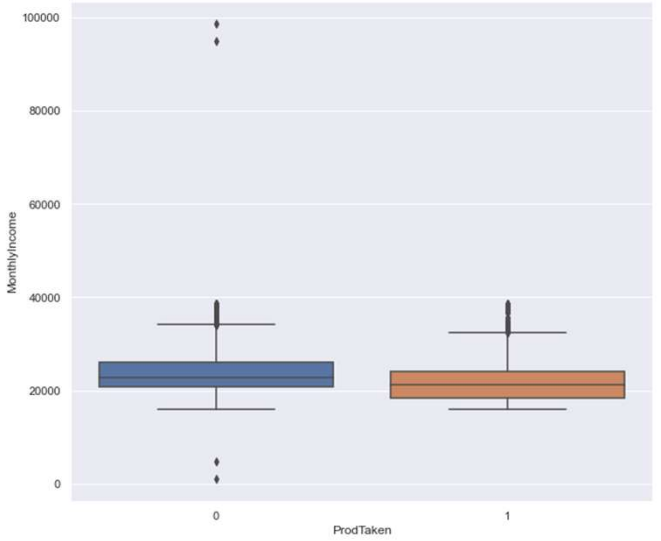


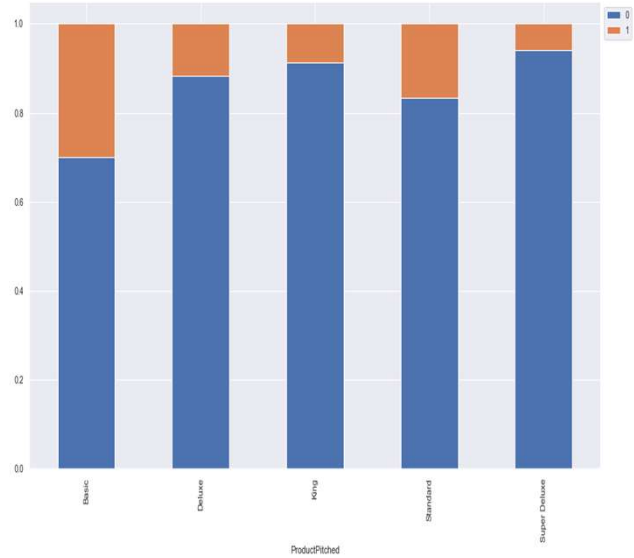
4. Package Purchase- Key Factors (III/III)





- Single customers take more packages than ones who are married or divorced.
- Average monthly incomes of customers who take packages are lower than ones who doesn't take the package.
- Number of follow-ups is higher for customers who have taken the package.
- Customers who are pitched basic and standard take the package more than the ones who are pitched other packages.

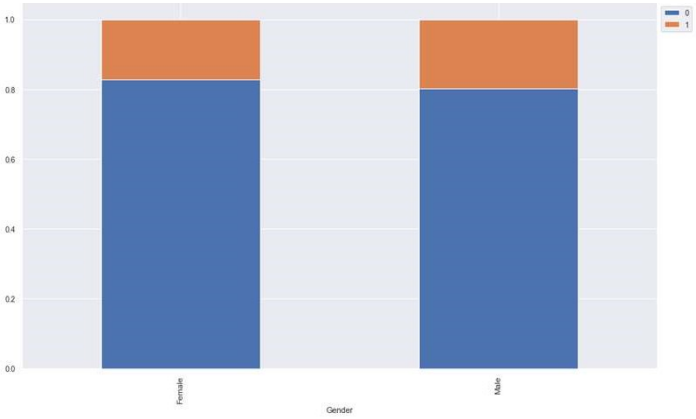


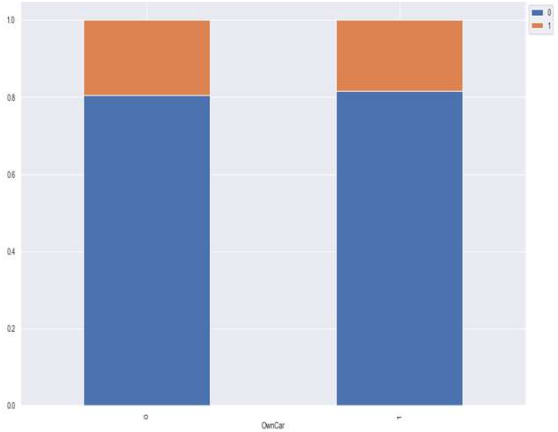


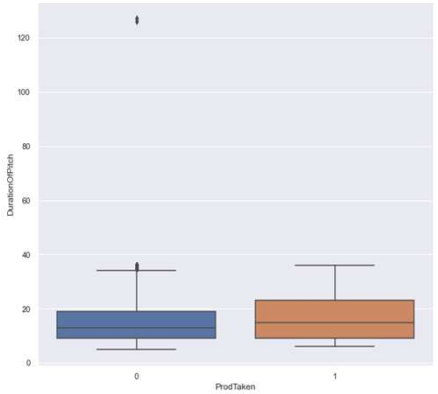
4. Package Purchase- Non-Key Factors

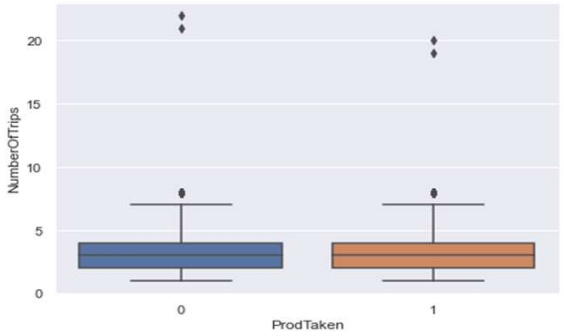
The following factors does not have notable variance for customers who have taken the package and those who have not.

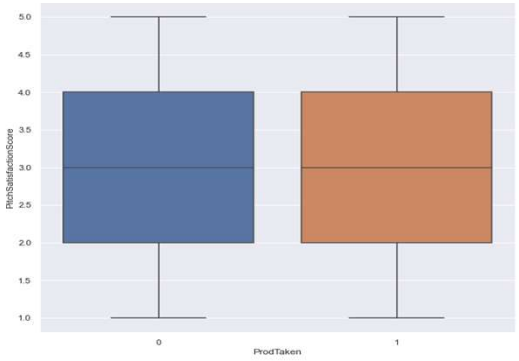
- Gender
- Number of trips
- Owning a car
- Number of persons visiting
- Duration of Pitch
- Number of children visiting

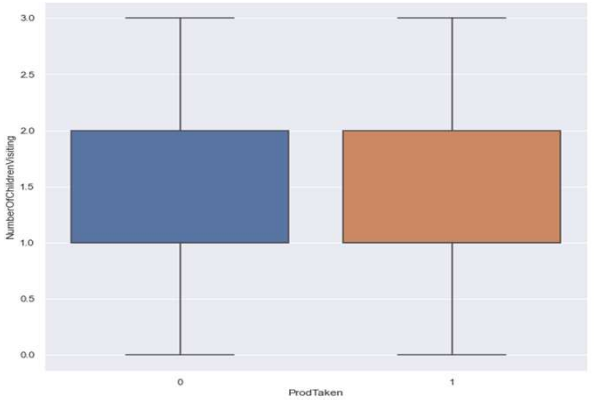






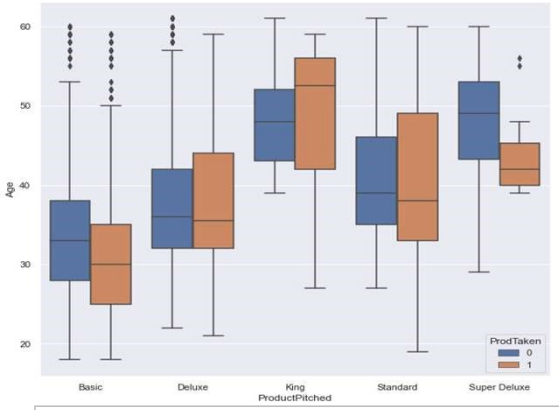


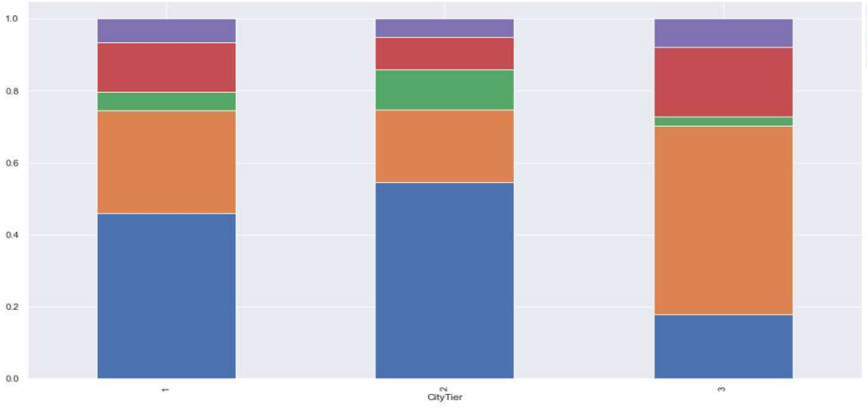


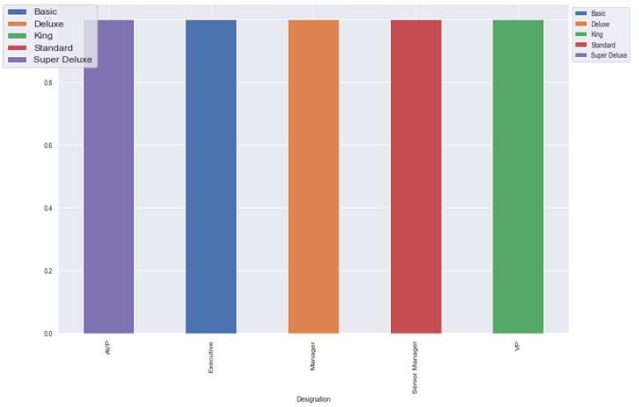


5.1 Types of Package Purchase- Key Factors

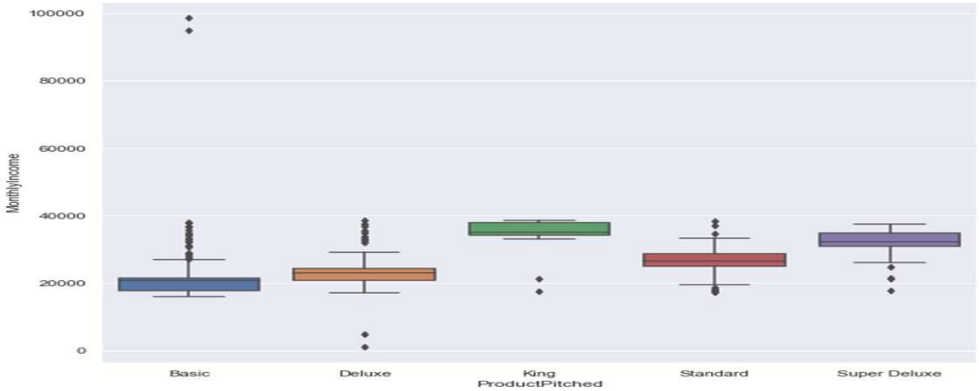
Key factors related to the choice of package by the customer are age, city tier, designation, monthly income and occupation.

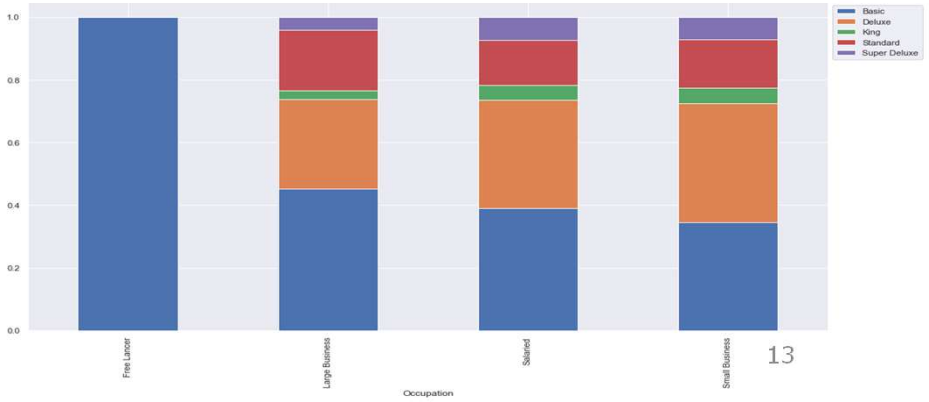






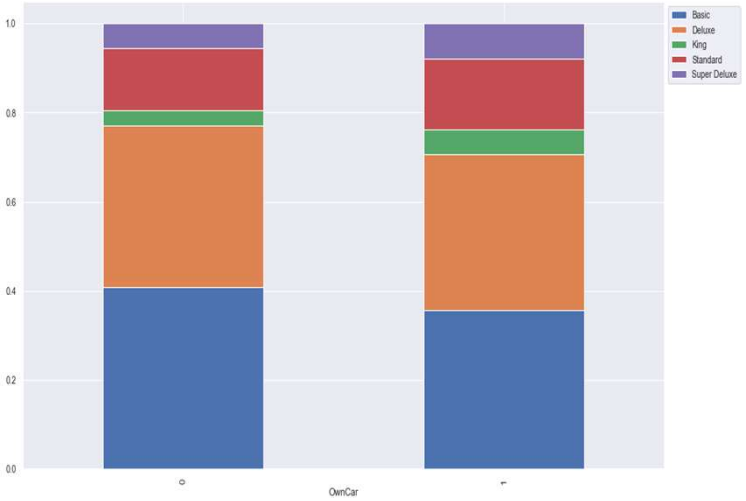
- King and Super Deluxe are purchased by older customers (between 40 and 50), but younger customers (below 40) purchase the other products
- Tier-3 city customers prefer deluxe and standard compared to tier-1 and tier-2. King is most preferred in Tier-2.
- Package selected is closely related to the designation, with senior most VP roles taking king, with executives taking basic. Similarly, the selection of package is reflected in higher monthly income customers taking king.
- Occupation does not seem to be as big an influencer as designation on the package selected.

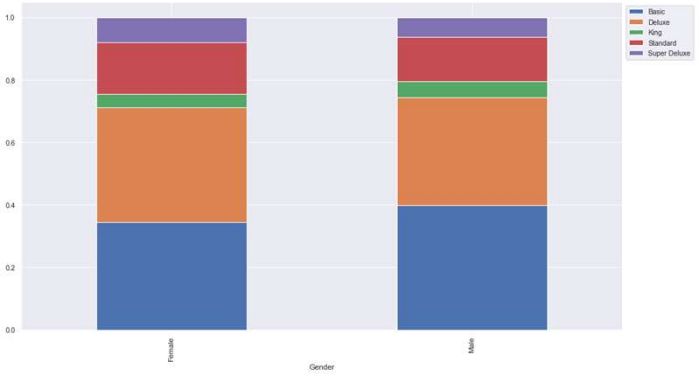


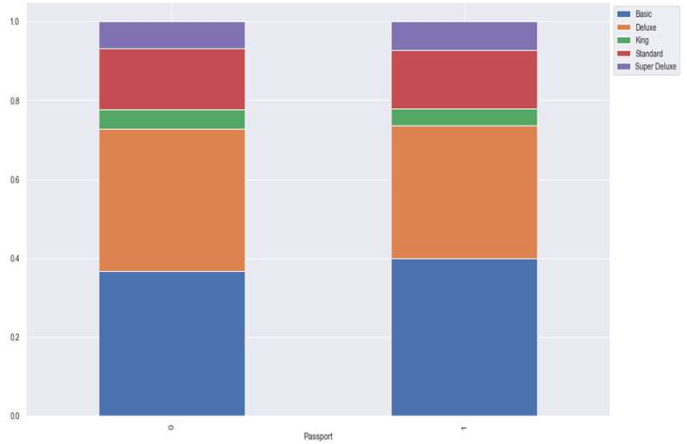


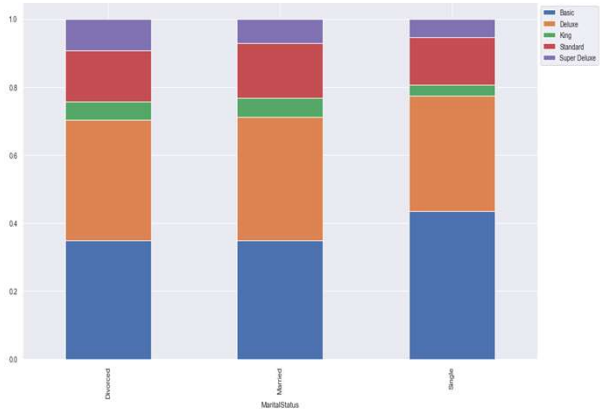
5.1 Type of Package Purchase- Non-Key Factors (I/II)

Customer gender, marital status and possessing a passport or a car does not seem to vary notably between customers who have taken different packages.



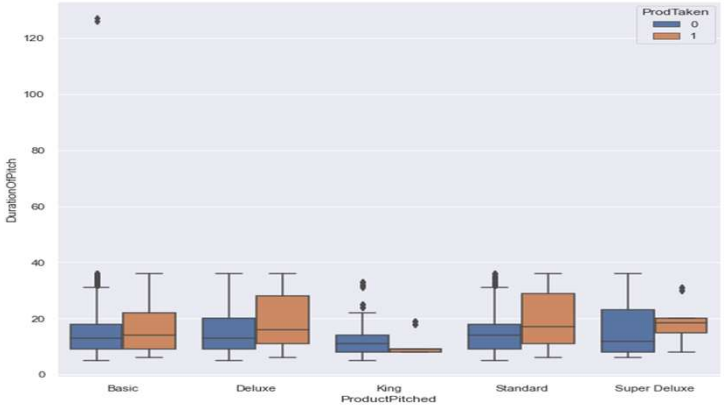


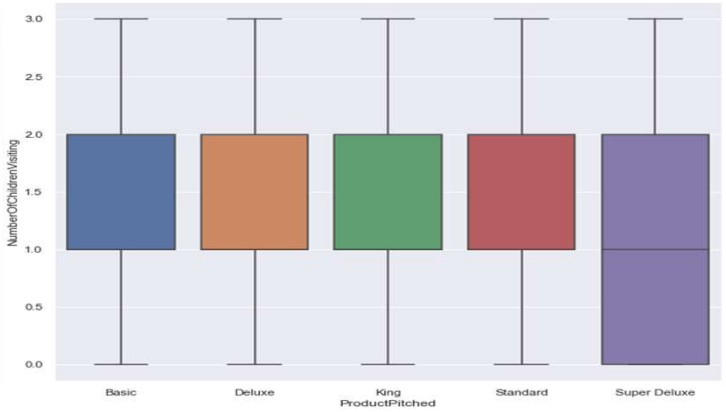


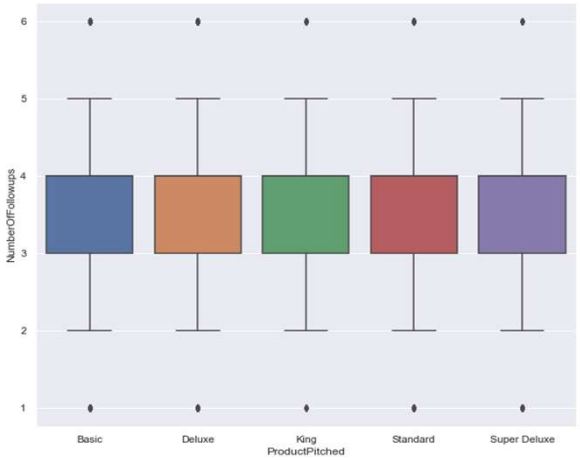


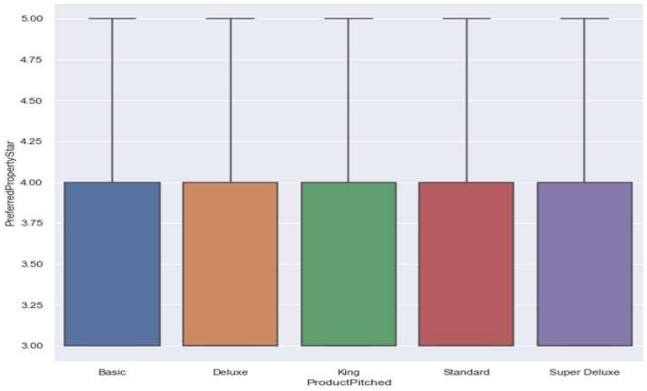
5.1 Type of Package Purchase- Non-Key Factors (II/II)

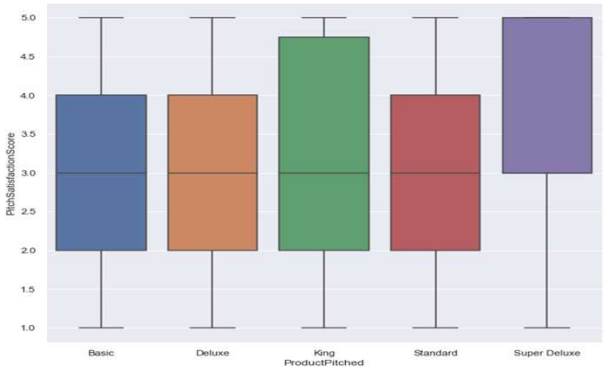
The following factors does not have notable variance between customers selecting different packages except for super deluxe showing some variance based on children visiting and preferred star selected.

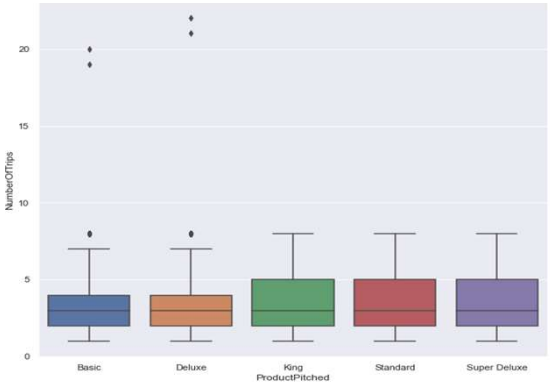


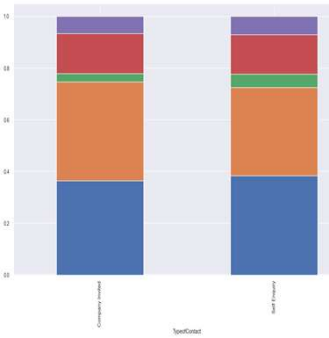












5.2 Customer Profiles

Below are profile of customers who buys various packages

Package	Customer Profile	Key customer feature
Basic	• Customers mostly between mid-20s to mid-30s • Monthly salary of below $20k, with mostly junior executive roles or free lancers • Mostly be from tier-1 and tier-2 cities, but can also be from tier-3	• Young customers who has just started their career mostly in big cities, and has low discretionary spend
Deluxe	• Customers mostly between early 30s and early 40s. • Monthly salary of just above $20k, with manager role • Mostly be from tier-3, but can also be from other cities	• Young customers who are in early stages of management, and higher discretionary income if they are from tier-3 cities.
Standard	• Customers mostly between early 40s and mid-50s. • Monthly salary of around $30k, with senior manager role • Mostly be from tier-3 and tier-1 cities, but can also be from tier-2	• Middle aged customers in middle management
Super Deluxe	• Customers mostly between 40 and 45 years • Monthly salary of around $35k, with AVP role • Mostly be from tier-3 and tier-1 cities, but can also be from tier-2	• Middle aged customers in relatively senior roles
King	• Customers mostly between early 40s and mid-50s • Monthly salary of around $40k, with VP role • Mostly be from tier-2 cities, but can also be from other cities	• Older high earner customers, with senior roles.

6.1 Model Building- Overview

- Decision Tree, Bagging methods, Boosting methods and Stacking method were used to build models to predict whether a customer would buy a package.
- Recall is the performance measure used to compare the performance of models in this scenario rather than accuracy.
  - ✓ Recall measures what proportion of the customers who take the package can be identified.
  - ✓ The cost of not being able to rightly identify a customer who buy a package is high which is measured using recall.
  - ✓ Accuracy is less important as it only the proportion of cases which are correctly identified. We will need to keep this high as possible to reduce contacting customers who will not buy a package.
- Following method will be used for the different modelling techniques:
  - ✓ Base model is built with no limits which is expected to overfit.
  - ✓ Overfitting is minimised and best recall performance obtained by using hyperparameters.
  - ✓ If there is high overfitting, attributes will be dropped to check whether overfitting can be reduced as it simplifies model.

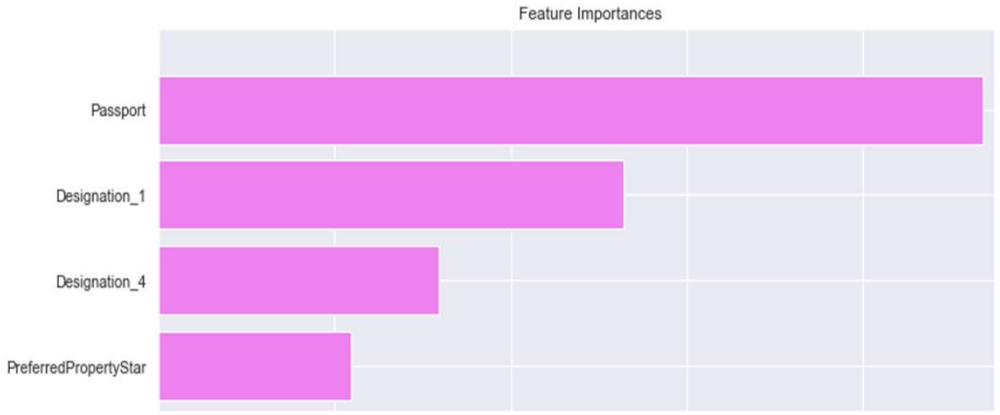
6.2 Decision Tree

- **Hyperparameter tuned model** has the highest recall, with least overfitting. However, the accuracy level is low.
- The hypertuned model will be able to get 75% of the potential customers. Low accuracy would mean that only 55.2% of the customers contacted would take the package. Hence, a better model needs to be identified.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
Decision Tree- Base model	100	72.9	100	90.3
Decision Tree- Max-depth=3	65.0	61.5	76.5	76.3
Decision Tree- Hyperparameter tuning (pre pruning)	75.5	75.0	54.2	55.2
Decision Tree- Cost complexity (post pruning)	98.1	75.7	99.6	91.5

Key Attributes

Following are key attributes by importance for the hypertuned model.



6.3 Bagging and Random Forest

- Bagging is an ensemble technique where a set of models are used together for prediction. A set of models are created in parallel, and the mode of the predictions from these models are used to get the final prediction.
- Models were created using the following methods:
  - ✓ Using all attributes
  - ✓ Selected key attributes based on eliminating the ones which either caused overfitting or reduced performance
- Models using all attributes is resulting in overfitting. Models created with selected attributes using random forest hyperparameter gave the best test recall, and higher accuracy and lower overfitting. Need to look into other models for higher performance and reduce overfitting.

All attributes

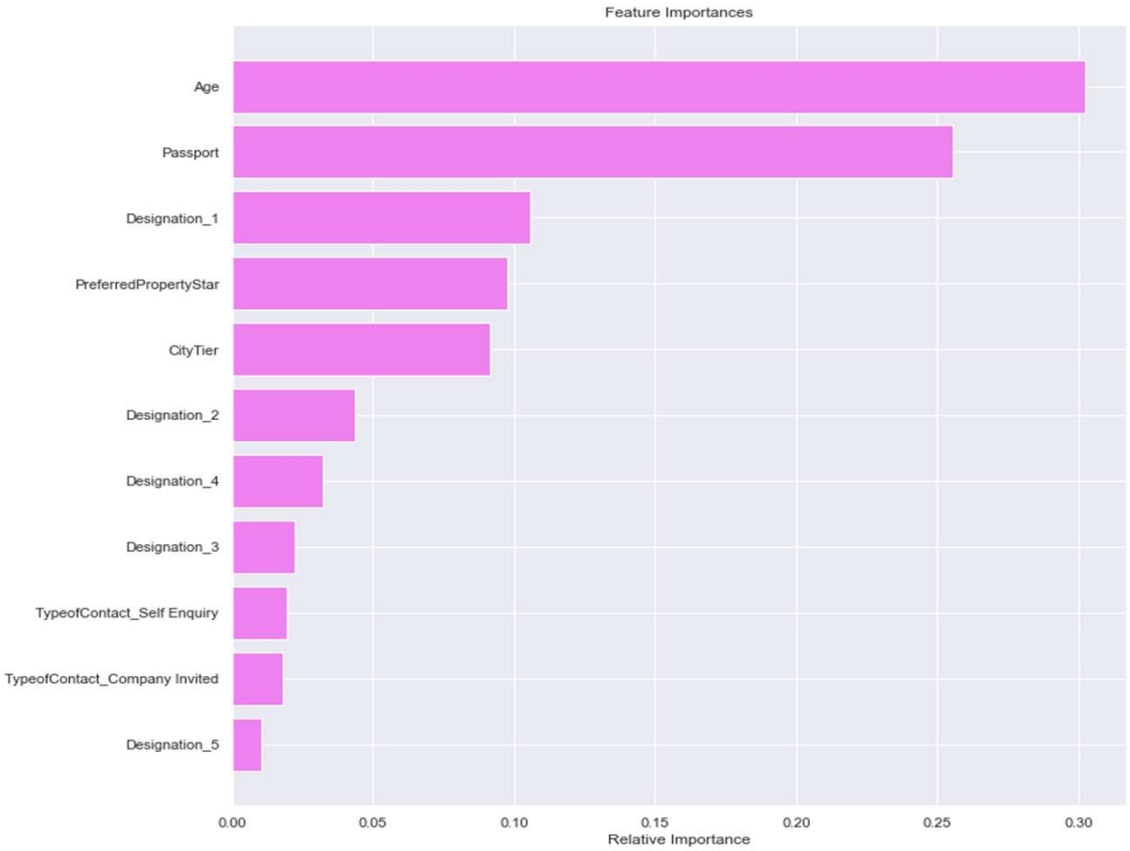
Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
Bagging estimator- base	97.8	63.2	99.6	91.2
Bagging estimator- hyperparameter	99.5	61.8	99.9	91.5
Random Forest- base	100	55.1	100	90.4
Random Forest- hyperparameter	80.3	60.8	89.6	85.1

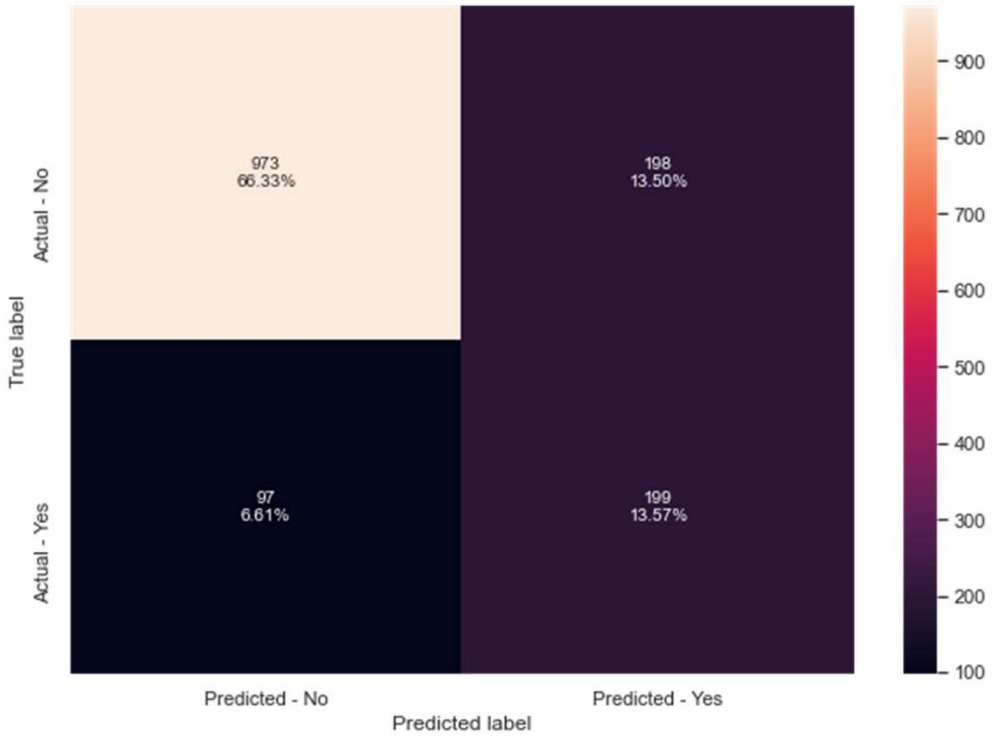
Selected key attributes

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
Bagging estimator- base	67.5	48.6	92.1	85.8
Bagging estimator- hyperparameter	57.7	39.2	91.3	85.5
Random Forest- base	69.4	51	92.7	86.8
Random Forest- hyperparameter	76.1	67.2	80.6	79.9

6.3 Bagging and Random Forest

Below are the key attributes by importance for the random forest hypertuned model, and confusion matrix.

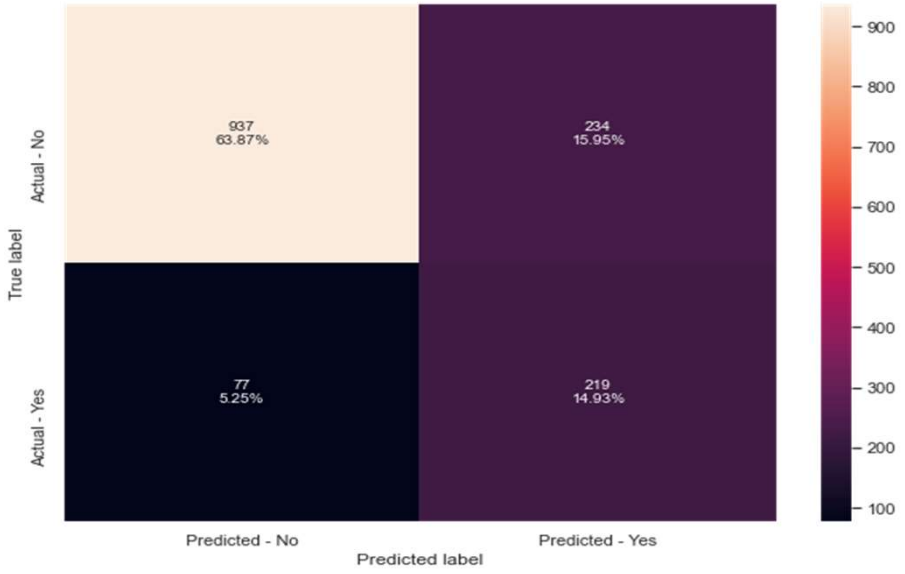




6.4 Boosting

- Boosting is another ensemble technique where models are created sequentially. The final prediction is decided by taking the weighted average or voting of models.
- **XGBoost-hyperparamter-A** gives the best recall with least overftting and reasonable accuracy. This is the **recommended model.**
- XGBoost-hyperparamter-B has higher recall than above, but is showing higher overfitting. Hence, XGBoost-hyperparamter-A is preferred over XGBoost-hyperparamter-B.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
Adaptive Boosting- base	23.4	19.9	84.0	81.7
Adaptive Boosting- hyperparamter	56.9	46.6	88.6	84.1
Gradient Boosting- base	33.9	26.3	85.5	82.5
Gradient Boosting- hyperparamter	37.5	28.7	86.3	82.9
XGBoost- base	63.6	46.3	91.4	85.6
XGBoost- hyperparamter-A	77.2	74	79.2	78.8
XGBoost- hyperparamter-B	82.5	77.0	80.2	78.7



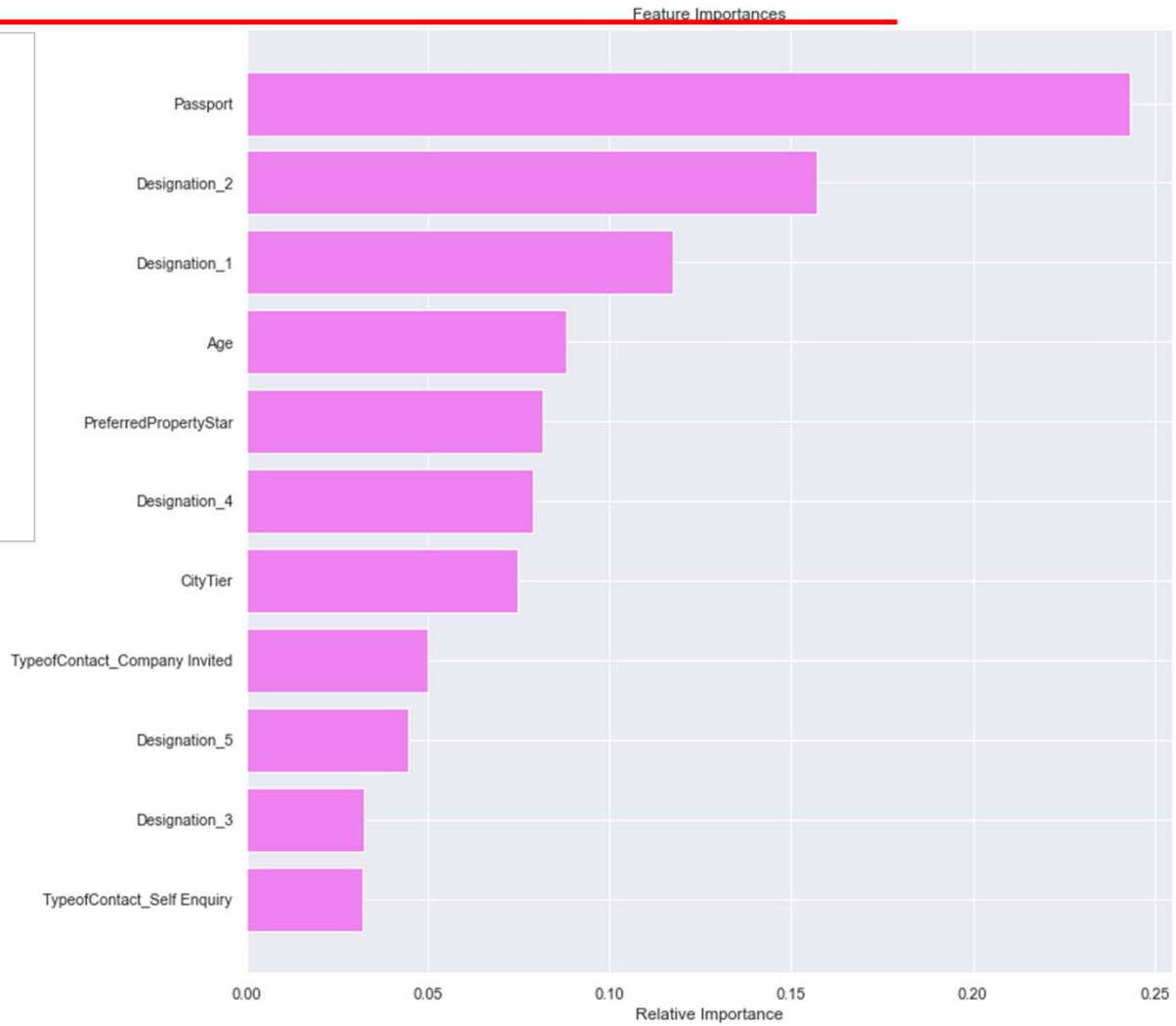
XGBoost- hyperparamter-A (recommended model)

```
base_score=0.5, booster='gbtree', colsample_bylevel=0.5, colsample_bynode=1, colsample_bytree=0.5,
eval_metric='logloss', gamma=3, gpu_id=-1, importance_type='gain', interaction_constraints='',
learning_rate=0.2, max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan, monotone_constraints='()',
n_estimators=41, n_jobs=4, num_parallel_tree=1, random_state=1, reg_alpha=0, reg_lambda=1, scale_pos_weight=5,
subsample=1, tree_method='exact', validate_parameters=1, verbosity=None
```

6.4 Boosting

- In the **XGBoost-hyperparamter-A model**, following are the key attributes in decreasing order of importance used for predicting whether a customer buys a package:

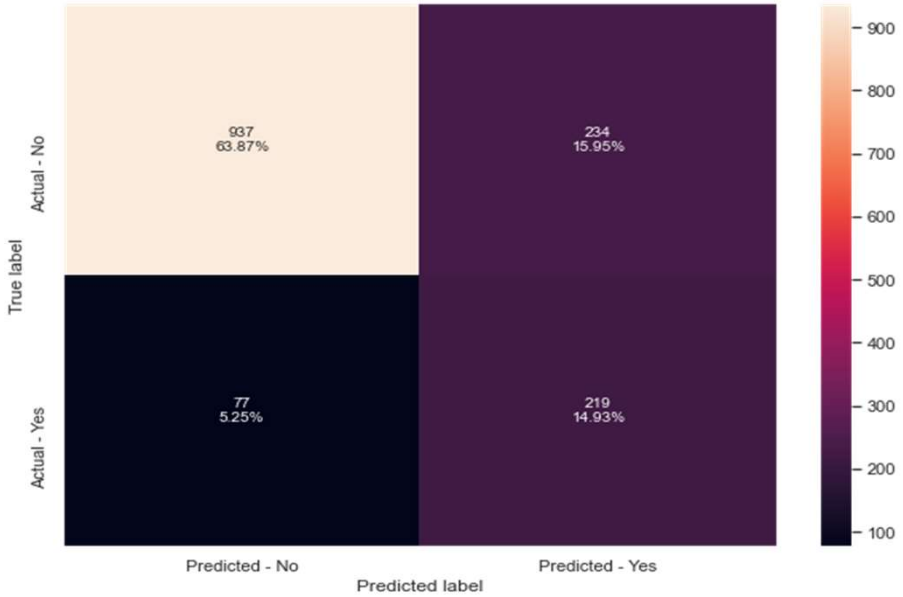
1. Possess a passport
2. Designation of customer
3. Age
4. Preferred property star
5. Tier of the customer's city
6. How the customer was contacted-initiated by company or customer.



6.5 Stacking

- Stacking is an ensemble technique where different type of models (heterogeneous) are used to create a meta model which is used for prediction. In bagging and boosting, homogeneous models are used.
- Here, the random forest, gradient boosting and decision tree were used to create a XGboost meta model which was used for predicting.
- There was more overfitting compared to XGBoost-hyperparamter-A model. Hence, this model is not preferred. The confusion matrix for test data using the stacking model is given below.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
Stacking	80.2	73.6	75.1	74.3



6.6 Model- Selection Summary

- Below is the summary of all the models created.
- Models were rejected for the following reasons:
  - ✓ models marked in amber have overfitting
  - ✓ models marked in blue had low recall or accuracy
- Model marked in green is the best model based on low overfitting and reasonable performance.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
Decision Tree- Base model	100	72.9	100	90.3
Decision Tree- Max-depth=3	65	61.5	76.5	76.3
Decision Tree- Hyperparameter tuning (pre pruning)	75.5	75	54.2	55.2
Decision Tree- Cost complexity (post pruning)	98.1	75.7	99.6	91.5
Bagging estimator- base	67.5	48.6	92.1	85.8
Bagging estimator- hyperparameter	57.7	39.2	91.3	85.5
Random Forest- base	69.4	51	92.7	86.8
Random Forest- hyperparameter	76.1	67.2	80.6	79.9
Adaptive Boosting- base	23.4	19.9	84	81.7
Adaptive Boosting- hyperparamter	56.9	46.6	88.6	84.1
Gradient Boosting- base	33.9	26.3	85.5	82.5
Gradient Boosting- hyperparamter	37.5	28.7	86.3	82.9
XGBoost- base	63.6	46.3	91.4	85.6
XGBoost- hyperparamter-A	77.2	74	79.2	78.8
XGBoost- hyperparamter-B	82.5	77	80.2	78.7
Stacking	80.2	73.6	75.1	74.3

7. Key Insights

Customer Profile

Customer characteristics

- Customer base is young with 50% of customers between 31 and 44 years
- There are more customers in junior roles, and are either salaried or has a small business.
- 75% of customer monthly income is below $25k and prefer 3 star+
- Low number of customers in tier-2 city.
- More married customers than single.

Travel characteristics

- 50% of customers goes on trips in group of 2 or 3, with 1 or 2 children.
- 75% of customers have taken at least 2 trips
- Majority of customers do not have a passport.

7. Key Insights

Purchase of Package

- Younger customers who are single and ones in junior roles with lower incomes purchase packages, than older ones who are married and are in senior roles with higher incomes.
- Customers contacted by the company purchases packages more than when the contact is initiated by the customer.
- Customers who prefer 5 star purchase the package more than the ones who prefer lower star
- Customers in lower tier (tier-2 and tier-3) takes packages more than ones in tier-1
- Customers who have passport takes packages than ones who does not have a passport.
- As number of follow-ups increases, there is a higher chance of taking a package. However, the duration of the pitch does not seem to impact the purchase of package.

Type of package purchased

Key factors impacting the type of package purchased are below:

1. Age of customer
2. Income and designation of customer
3. Occupation (Salaried, business etc.) of customer
4. City tier of customer

8. Recommendations

- Increase the customer base to have more high earners who are older in senior roles for the following reasons
  - ✓ they buy more expensive packages such as ‘king’ and super deluxe’.
  - ✓ Current customer base is of more younger lower earners
  - ✓ proposed wellness package might appeal to older customers than the current younger customer base
- Introduce service to obtain passports for customers as customers with passport buy more packages. Majority of current customer base do not have a passport.
- Increase company initiated contacts to increase package purchase, and also the follow-ups. These both increase probability of buying a package.
- Duration of pitch is not impacting package purchase. This needs to be reviewed to consider any improvements required.
- The type of package selected changes based on specific customer attributes such as age (refer slide-13 for details). Hence, the following steps needs to be followed for predicting the customers who will buy the wellness package:
  - ✓ Start using the XGBoost hypertuned model to predict which customers will buy packages to target customers who might buy the wellness package
  - ✓ After data is collected about which customers buy the wellness package, the model should be enhanced using the data gathered from the initial sales. This will help to improve the model performance which will lead to targeted marketing of the wellness package.