TRAVEL PACKAGE PREDICTION
VISIT WITH US

Project 5 - Ensemble Techniques - Classification

Authored by C. Kamakani Kahunahana

BUSINESS OVERVIEW
PROBLEM AND SOLUTION APPROACH

PROBLEM

Last year, only 18% of Visit with Us' customers purchased travel packages. The company wants to enable and establish a viable business model to expand the customer base.

One of the ways to expand the customer base is to introduce a new offering of packages.

Currently, there are 5 types of packages the company is offering - Basic, Standard, Deluxe, Super Deluxe, King. Looking at the data of the last year, we observed that 18% of the customers purchased the packages. The company is now planning to launch a new product.

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

SOLUTION APPROACH

Build a predictive model to help the company better target customers more likely to purchase the new packages. We will build a model using ensemble techniques including decision trees, random forests, boosting and stacking.

Using data provided, we will identify key variables from their customer database to identify which characteristics are best at predicting which customers are more likely to purchase travel packages.

Objectives:

1. To predict which customers are more likely to purchase new products

2. Identify customer characteristics which will help improve the company's marketing effectiveness and drive spending efficiency

BASE DATASET OVERVIEW

Details	Amount
Observations	4888
Variables	20
Missing Values	DurationOfPitch (251), MonthlyIncome (233), Age (226), NumberOfTrips (140), NumberOfChildrenVisiting (66), NumberOfFollowups (45), PreferredPropertyStar (26), TypeofContact (25)

This is a description of the base dataset prior to any pre-processing

Variable	Type	Description
CustomerID	Integer	Unique Customer ID
ProdTaken	Integer	Purchased package in past
Age	Float	Age of customer
TypepofContact	Object	How customer contacted
CityTier	Integer	Tier 1 > Tier 2 > Tier 3
DurationOfPitch	Float	Duration of the pitch by a salesperson to the customer
Occupation	Object	Occupation of customer
Gender	Object	Gender of the customer
NumberofPersonVisitng	Integer	Total number of persons planning to take the trip with the customer
NumberOfFollowups	Float	Total number of follow-ups has been done by the salesperson after the sales pitch
ProductPitched	Object	Product pitched by the salesperson
PreferredPropertyStar	Float	Preferred hotel property rating by customer
MaritalStatus	Object	Marital status of customer
NumberOfTrips	Float	Average number of trips in a year by customer
Passport	Integer	Customer has a passport or not (0: No, 1: Yes)
PitchSatisfactionScore	Integer	Sales pitch satisfaction score
OwnCar	Integer	Whether the customer owns a car (0: No, 1: Yes)
NumberOfChildrenVisiting	Float	Total number of children with age less than 5 planning to take the trip with the customer
Designation	Object	Designation of the customer in the current organization
MonthlyIncome	Float	Gross monthly income of the customer

01

DATA PREPROCESSING

DATA PRE-PROCESSING

PROCESSING VARIABLES AND UNNECESSARY DATA

Variable	Pre-Processing Strategy
CustomerID	Dropped during model building
Age	Fill nulls with median
CityTier	Convert to category and onehotencode
DurationOfPitch	Nulls = 0
NumberOfFollowups	Fill nulls with median
PreferredPropertyStar	Fill nulls with median
NumberOfChildrenVisiting	Fill nulls with median
MonthlyIncome	Fill nulls with median grouped by Designation

- Dropped CustomerID
- Nulls are filled with Medians
- CityTier converted to category and onehotencoded

DATASET OVERVIEW
NUMERICAL VALUES

NUMERICAL VARIABLE DESCRIPTION

	Count	Mean	Standard Deviation	Minimum	25%	50%	75%	Maximum
ProdTaken	4888	0.18	0.39	0	0	0	0	1
Age	4888	37.54	9.11	18	31	36	43	61
DurationOfPitch	4888	15.36	8.32	5	9	13	19	127
NumberOfPersonVisiting	4888	2.91	0.72	1	2	3	3	5
NumberOfFollowups	4888	3.71	1	1	3	4	4	6
PreferredPropertyStar	4888	3.58	0.80	3	3	3	4	5
NumberOfTrips	4888	3.23	1.82	1	2	3	4	22
Passport	4888	0.29	0.45	0	0	0	1	1
PitchSatisfactionScore	4888	3.08	1.37	1	2	3	4	5
OwnCar	4888	0.62	0.49	0	0	0	1	1
NumberOfChildrenVisiting	4888	1.18	0.85	0	1	1	2	3
MonthlyIncome	4888	23546.84	526628	1000	20485	22413.50	25424.75	98678

- ProdTaken is our dependent variable

NEW DATASET OVERVIEW
CATEGORICAL VALUES

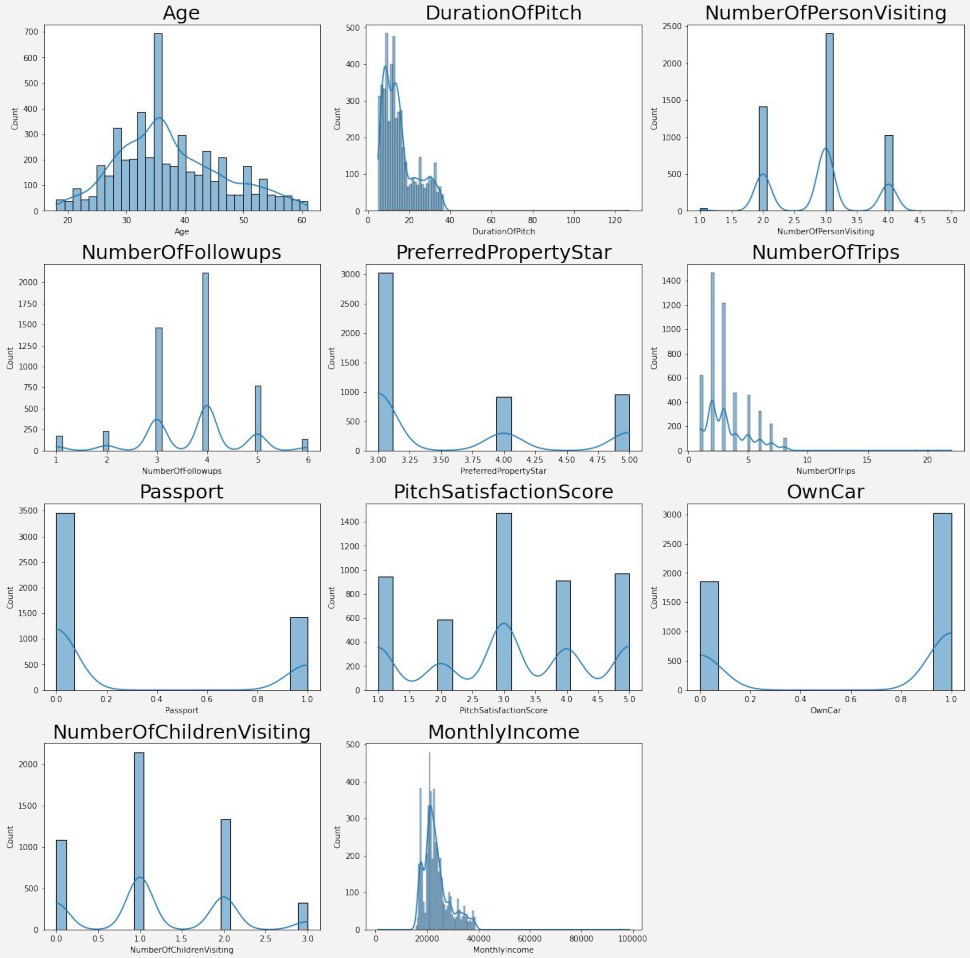
CATEGORICAL VARIABLE DESCRIPTION

	Count	Unique	Top	Frequent
TypeofContact	4888	3	Self Enquiry	3444
CityTier	4888	3	1	3190
Occupation	4888	4	Salaried	2368
Gender	4888	2	Male	2916
ProductPitched	4888	5	Basic	1842
MaritalStatus	4888	4	Married	2340
Designation	4888	5	Executive	1842

- TypeOfContact most common is Self Enquiry, Company Invited is second, followed unknown (25)
- CityTier_1 has more than 2x more than CityTier_3 (1500) followed by CityTier_2 (198)
- Occupation is led by Salaried (2368) followed by Small Business with (2084)
- ProductPitched second place is Deluxe (1732) followed by Standard (742)
- Married is most common MaritalSatus followed by Divorced (950), Single (916) and Unmarried (682)
- Most customers identify as Executives followed closely by Manager (1732)

02

EXPLORATORY DATA ANALYSIS

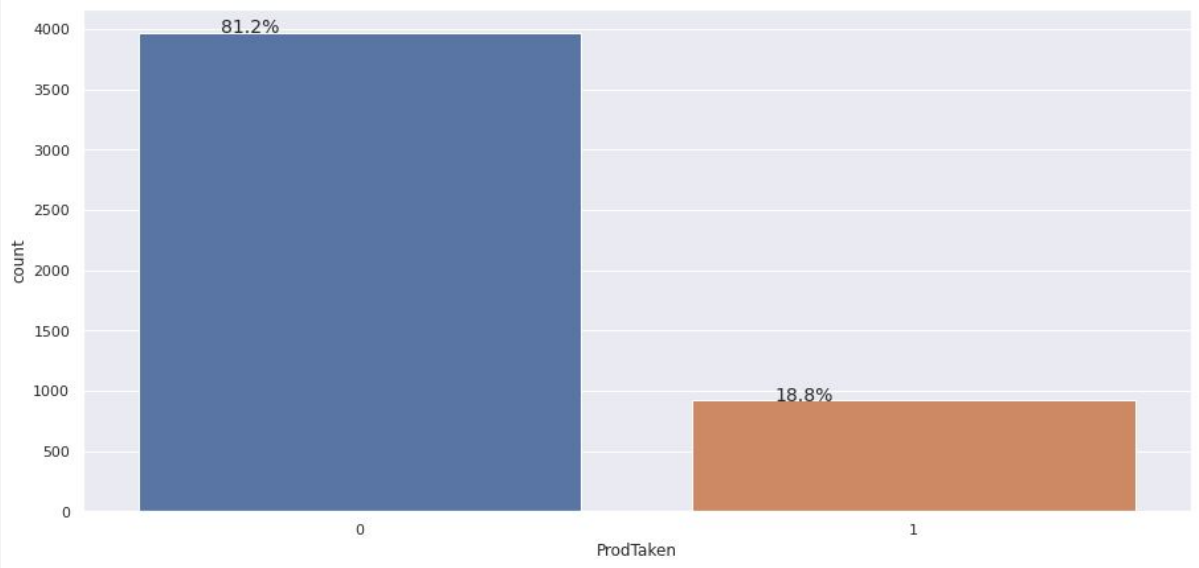


DISTRIBUTION ANALYSIS

- Age, DurationOfPitch, NumberOfTrips, and MonthlyIncome are continuous variables
- We will look at these variables in closer detail on the following pages

CUSTOMER ANALYSIS
PURCHASED PACKAGE

PURCHASED PACKAGE (Y/N)

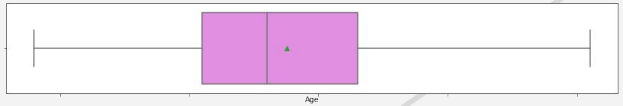


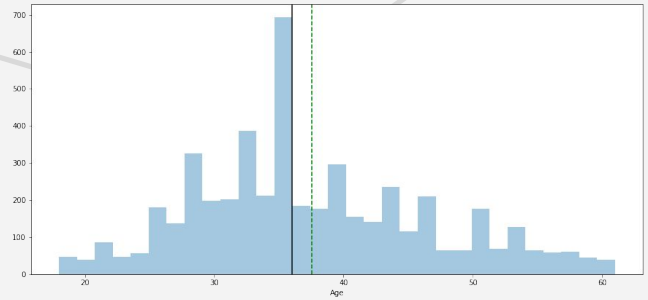
- 81.2% of customers did not purchase a travel package

CUSTOMER ANALYSIS

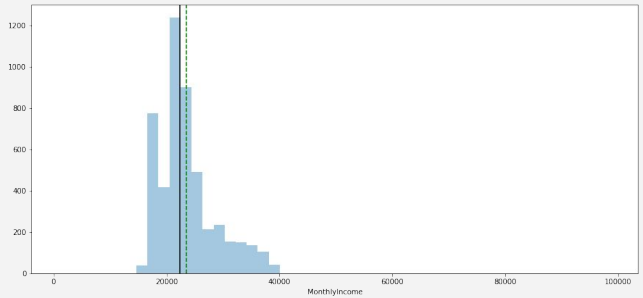
AGE, MONTHLY INCOME, NUMBER OF TRIPS TAKEN

CUSTOMER AGE

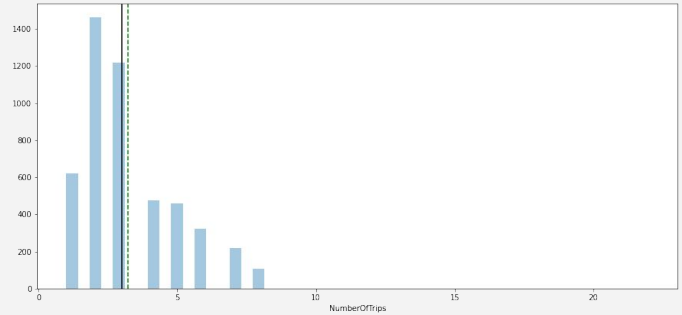




MONTHLY INCOME



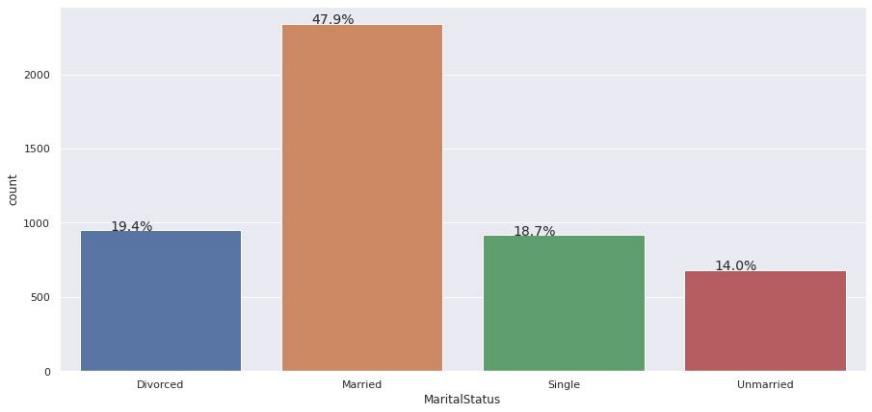
NUMBER OF TRIPS TAKEN



- Age is fairly well distributed with a mean of 37.6 and median of 36
- Monthly Income is right skewed with several outliers at the top and bottom; min 1000, median 22347, max 98678
- Number Of Trips is right skewed with several major outliers; min 1.0, median 3.0, max 22.0

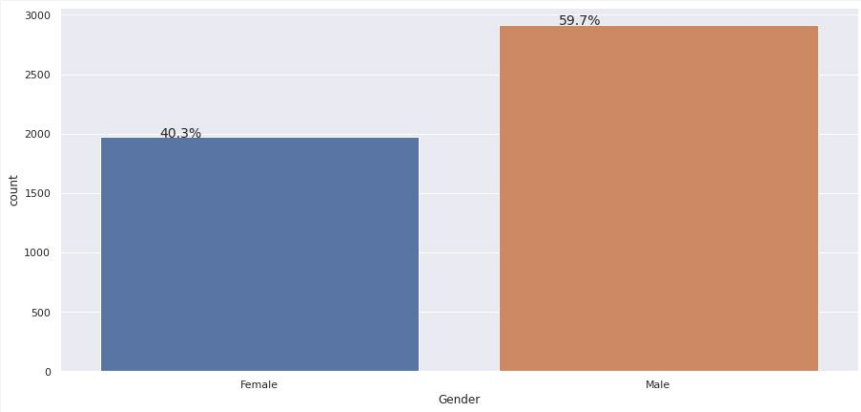
CUSTOMER ANALYSIS
MARITAL STATUS AND GENDER

MARITAL STATUS



- 47.9% of customers are Married

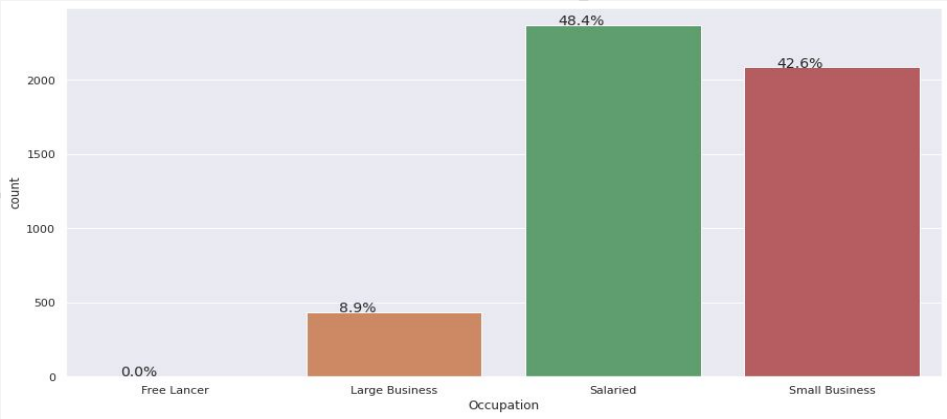
GENDER



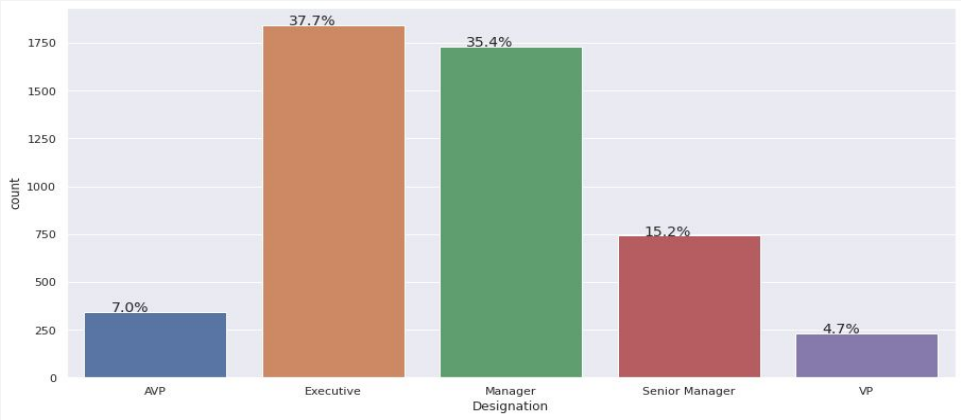
- 50.7% of customers are Male

CUSTOMER ANALYSIS
OCCUPATION AND DESIGNATION

OCCUPATION



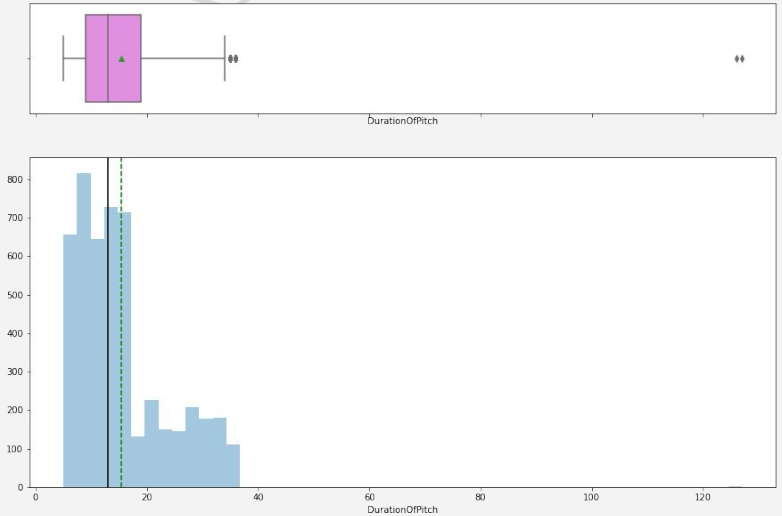
DESIGNATION



- The majority of customers identify as Salaried Employees or Small Business
- Customers tend to be either Executives (37.7%) or Managers (35.4%)

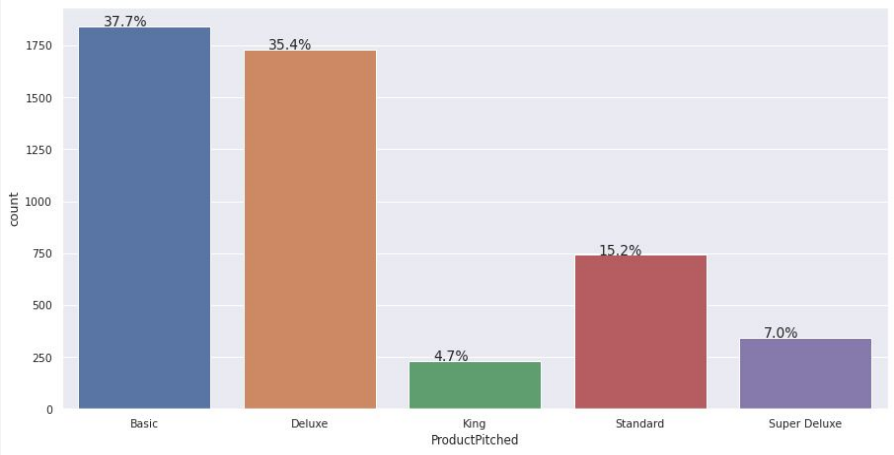
SALES & MARKETING ANALYSIS
PITCH DURATION AND PACKAGE PITCHED

DURATION OF PITCH



- Pitch Duration is right skewed with major outliers. However, we did not find that these outliers affected our predictive results

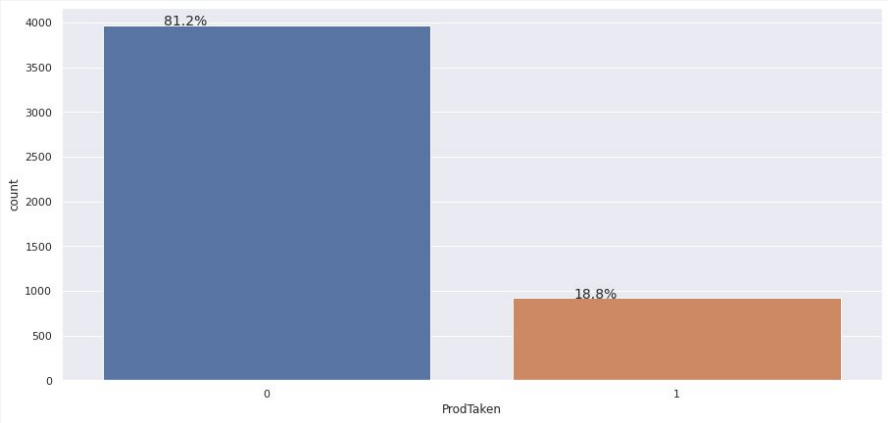
PACKAGE PITCHED



- Employees tend to focus on pitching Basic and Deluxe packages to customers

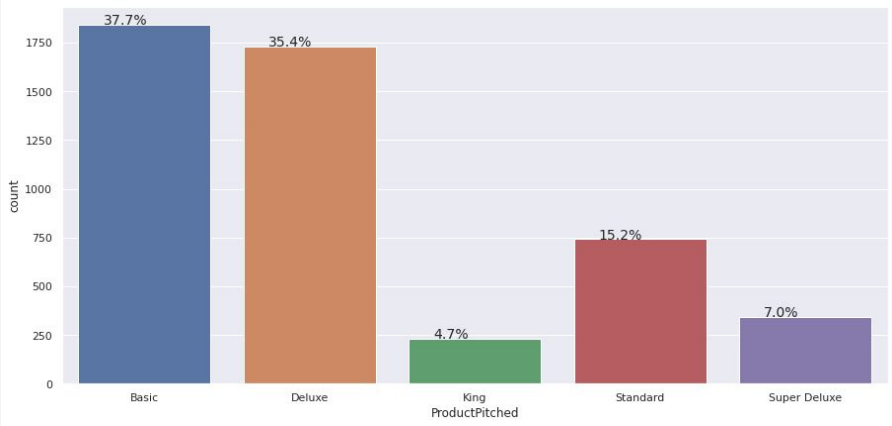
CUSTOMER ANALYSIS
HOW CONTACTED

PURCHASED PACKAGE (Y/N)



- 81.2% of customers did not purchase a travel package

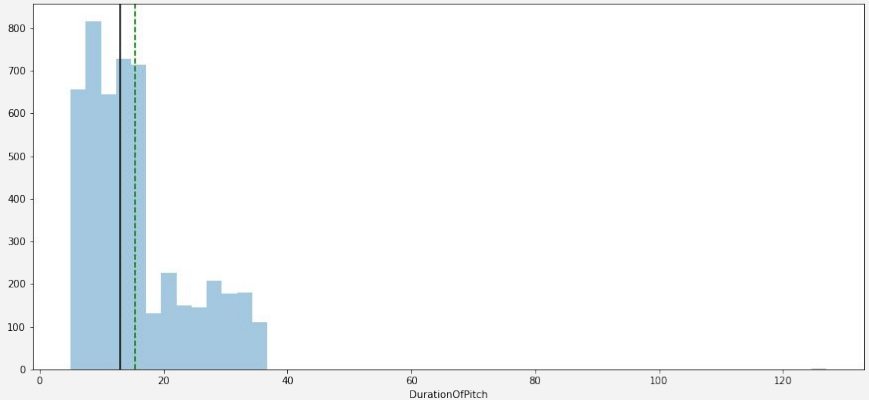
PACKAGE PITCHED



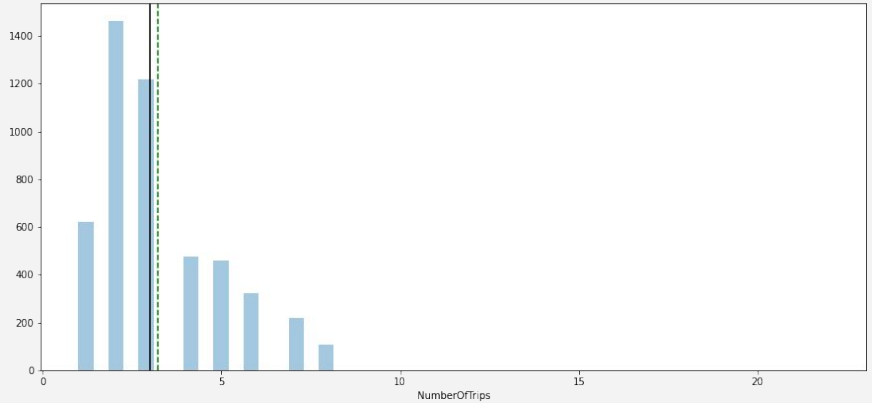
- 50.7% of customers are Male

SALES & MARKETING ANALYSIS
DURATION OF PITCH AND PRODUCTS PITCHED

DURATION OF PITCH



PRODUCTS PITCHED



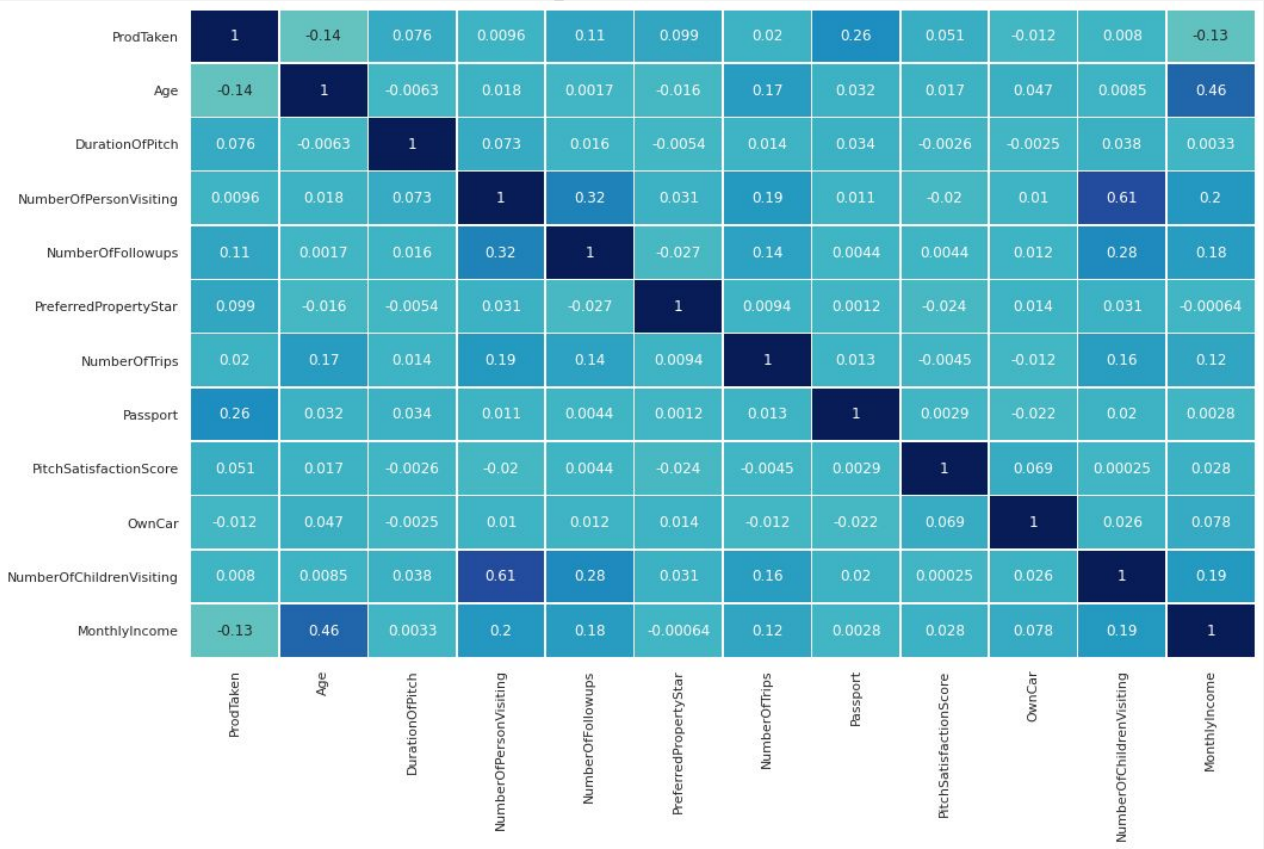
- Product
- Monthly Income is right skewed with several outliers at the top and bottom; min 1000, median 22347, max 98678

03

BIVARIATE & MULTIVARIATE ANALYSIS

VARIABLE ANALYSIS

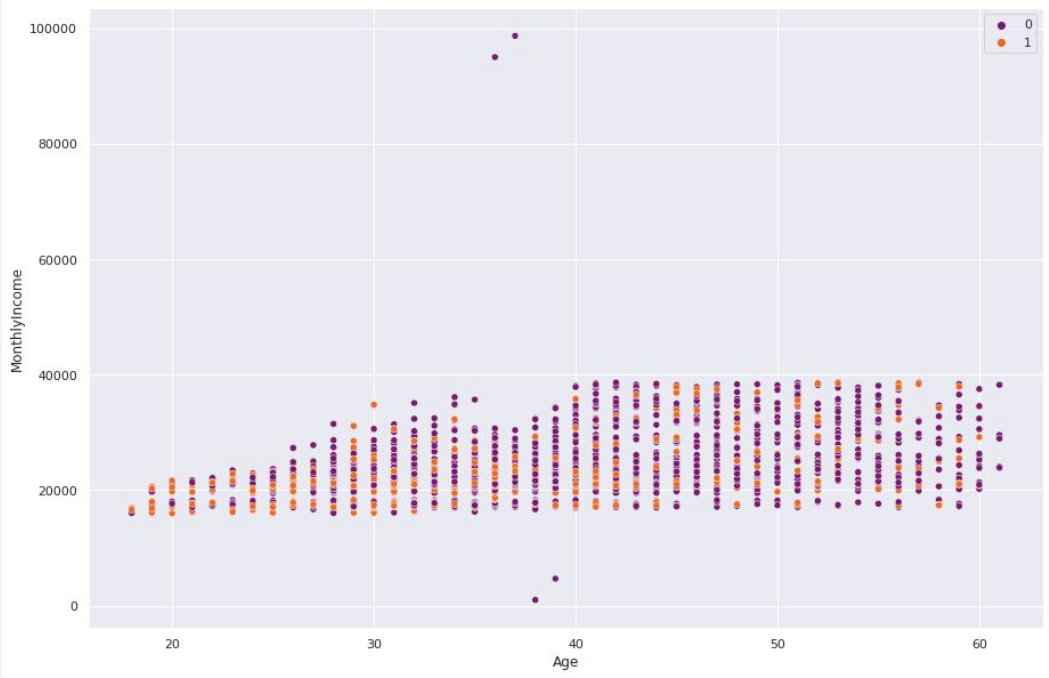
CORRELATION MATRIX



- There does not seem to much strong correlation in the dataset which is surprising
- Age has a loose correlation with MonthlyIncome
- NumberOfPersonVisiting has a positive association with NumberofChildrenVisiting which makes sense.

PACKAGE PURCHASED ANALYSIS
AGE AND MONTHLY INCOME

AGE & MONTHLY INCOME VS PACKAGE PURCHASED

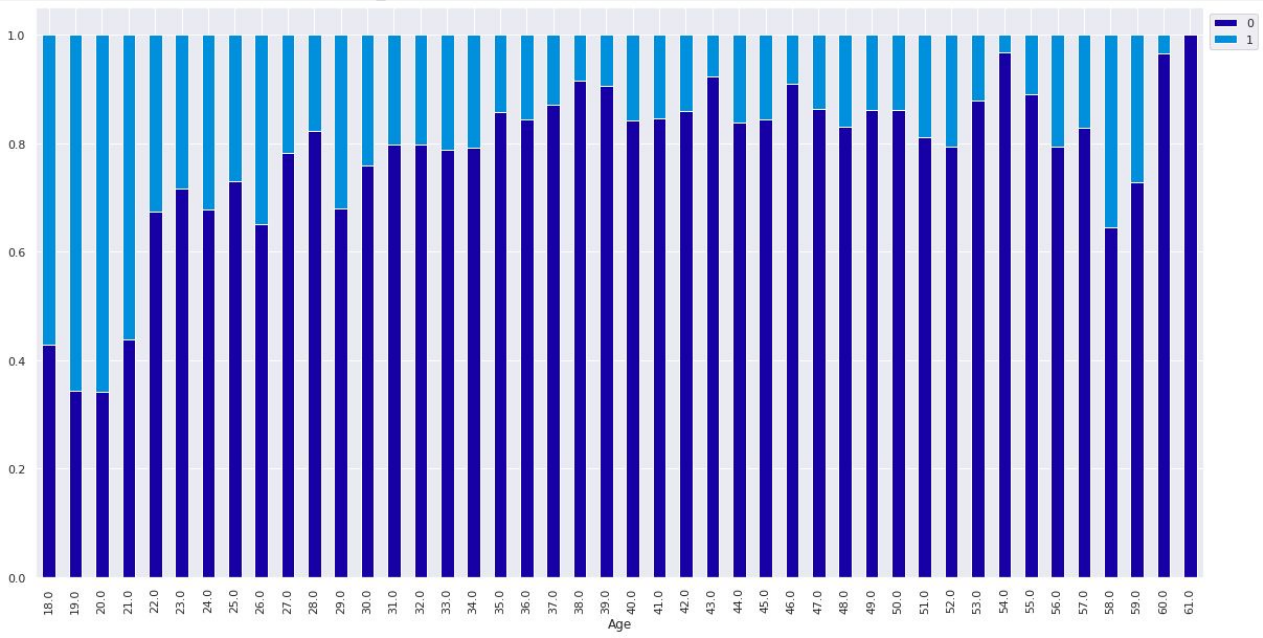


- Younger customers tend to purchase more travel packages than older customers
- There is not a strong relationship between Monthly Income and purchasing of packages
- There is a positive relationship between income and age which is not unexpected

PACKAGE PURCHASED ANALYSIS

AGE

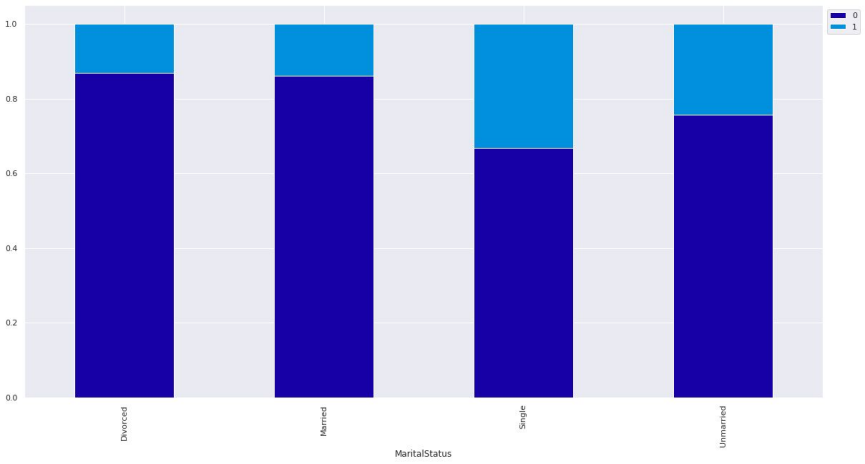
AGE VS PACKAGE PURCHASED



- Younger customers tend to be more likely to purchase packages than older customers

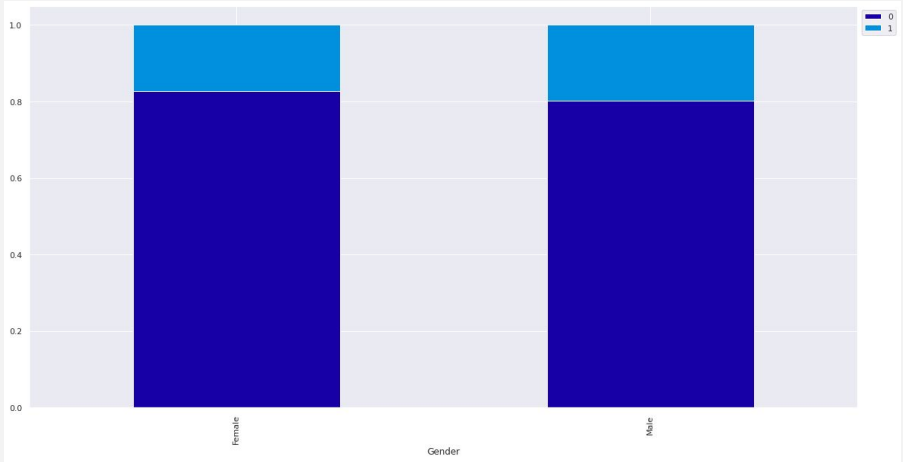
PACKAGE PURCHASED ANALYSIS
MARITAL STATUS AND GENDER

MARITAL STATUS VS PACKAGE PURCHASED



- Single and Unmarried customers are relatively more likely to make a purchase than Divorced or Married customers

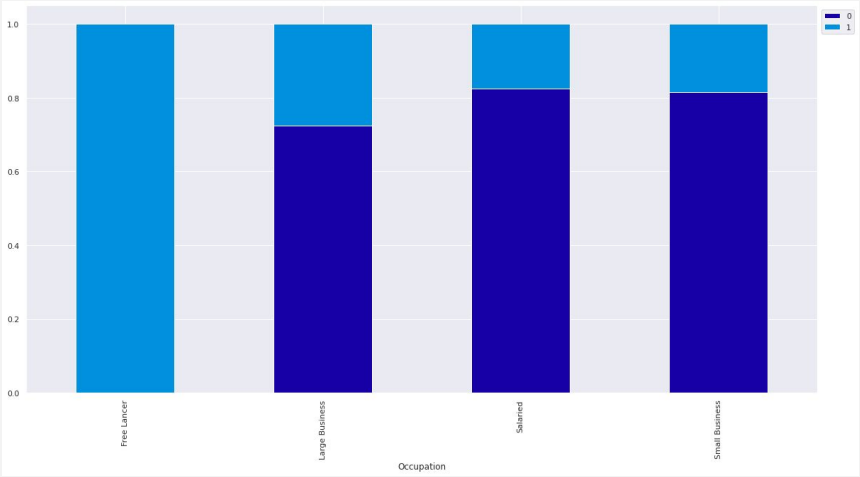
GENDER VS PACKAGE PURCHASED



- Men are relatively more likely to make a purchase

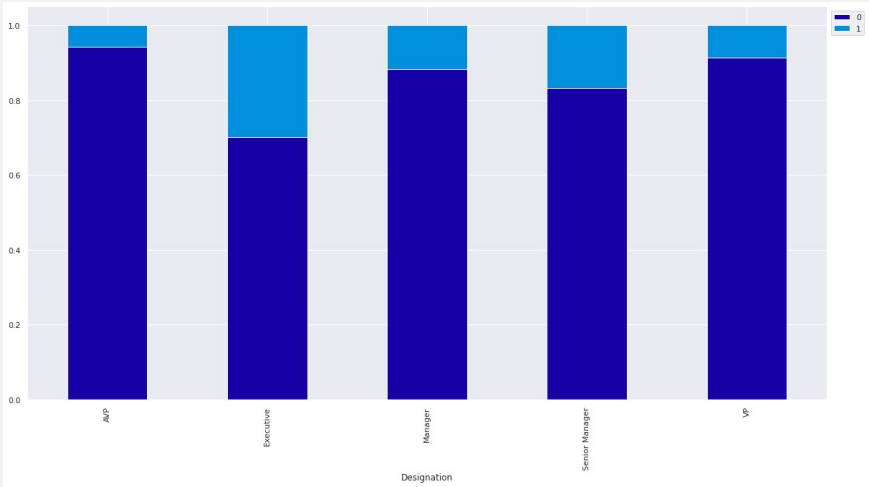
PACKAGE PURCHASED ANALYSIS
OCCUPATION AND DESIGNATION

OCCUPATION VS PACKAGE PURCHASED



- Freelancer have a very small sample size but are likely to purchase
- Employees at large business are relatively frequent purchasers

DESIGNATION VS PACKAGE PURCHASED

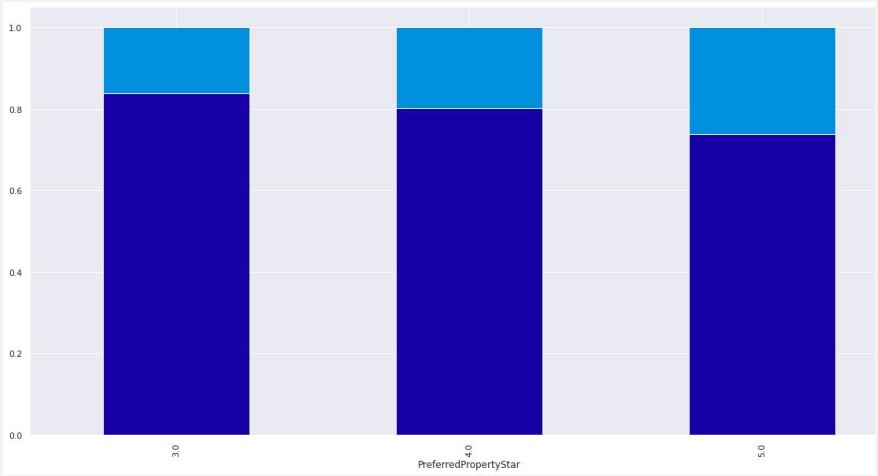


- Executives seem to be the best target market

PACKAGE PURCHASED ANALYSIS

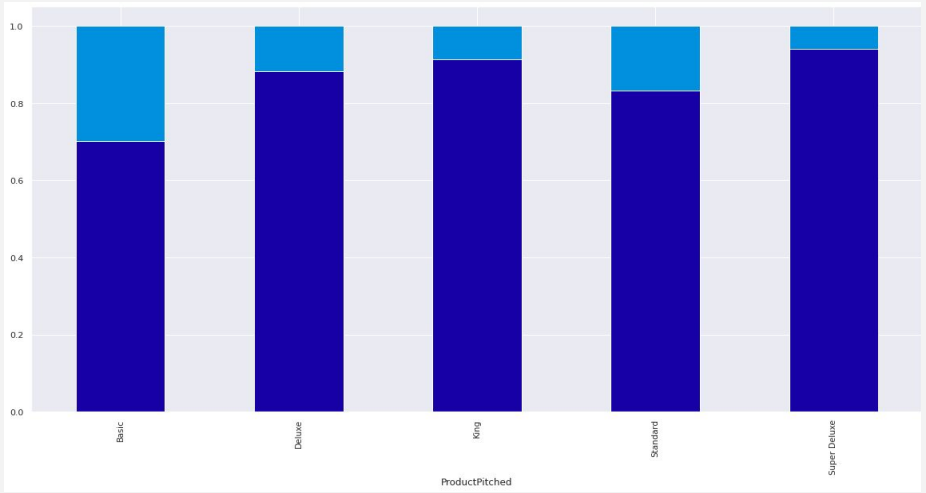
PREFERRED PROPERTY SCORE AND PRODUCT PITCHED

PREFERRED PROPERTY VS PACKAGE PURCHASED



- Customers who prefer higher rated properties are relatively more likely to purchase packages

PRODUCT PITCHED VS PACKAGE PURCHASED

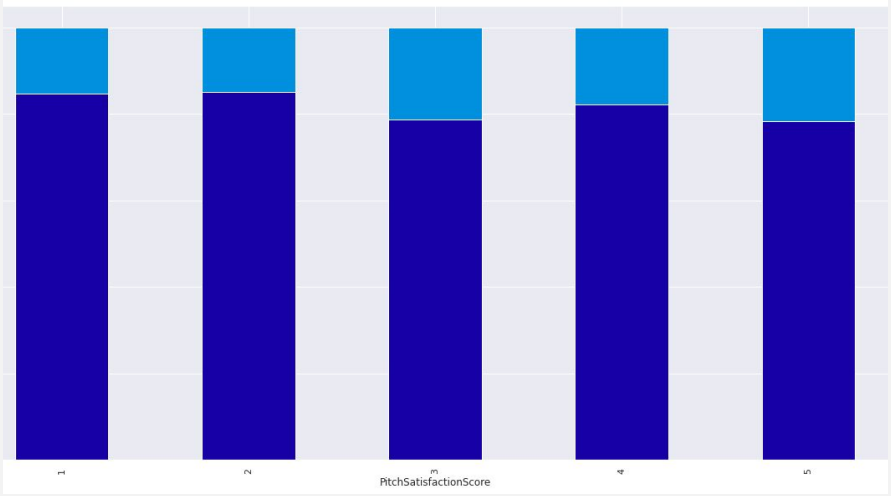


- Customers who are pitched basic or standard packages are relatively more likely to purchase

PACKAGE PURCHASED ANALYSIS

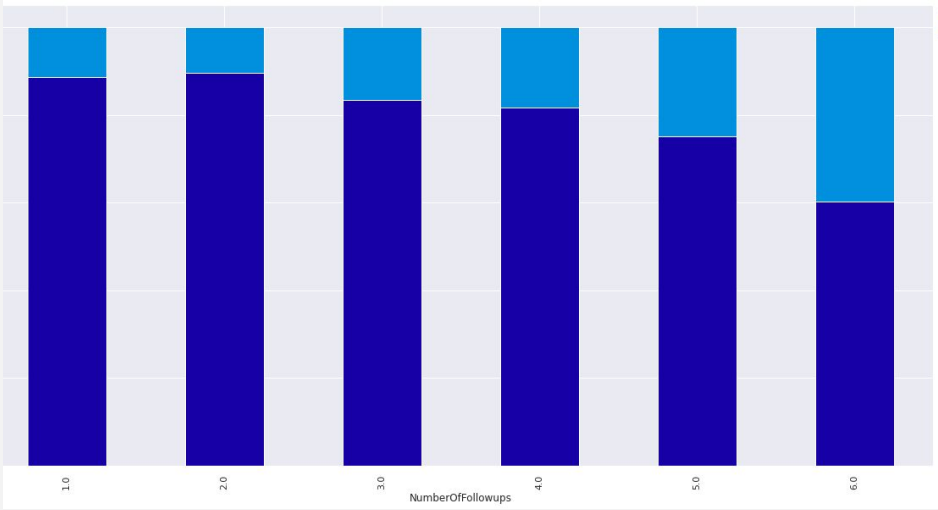
PITCH SATISFACTION AND NUMBER OF FOLLOW UPS

PITCH SATISFACTION VS PACKAGE PURCHASED



- Customers who reported satisfaction scores of 3 and 5 are relatively more likely to purchase packages

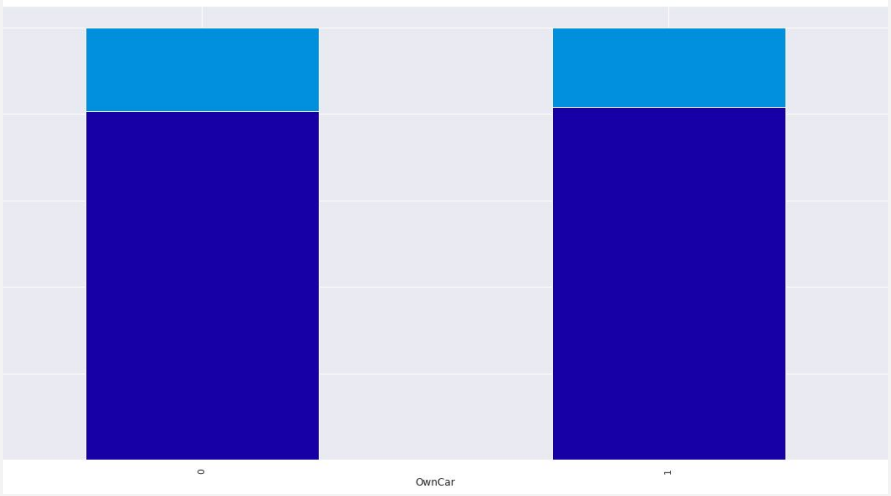
NUMBER OF FOLLOW UPS VS PACKAGE PURCHASED



- The more times employees follow up with customers the more likely they are to make a purchase

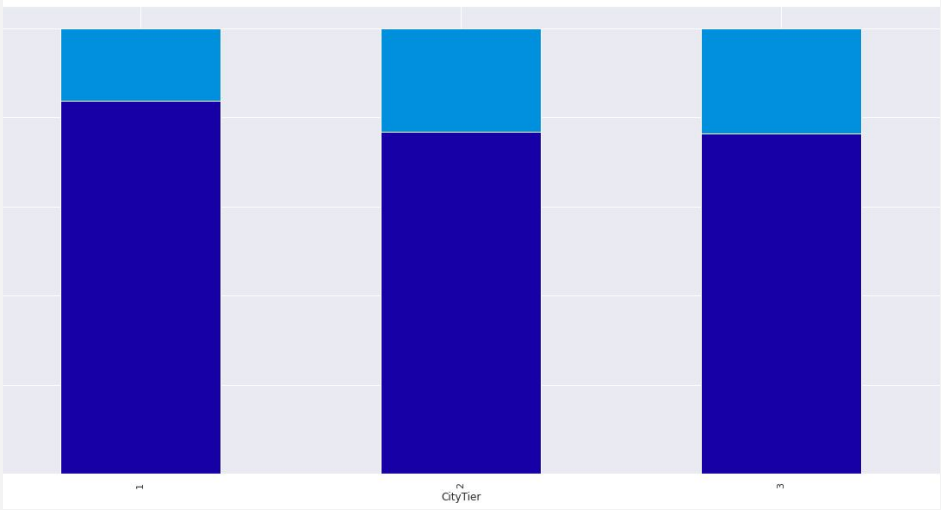
PACKAGE PURCHASED ANALYSIS
CAR OWNERSHIP AND CITY TIER

CAR OWNERSHIP VS PACKAGE PURCHASED



- Whether or not a customer own a car does not seem to be a good indicator of purchase intent

CITY TIER VS PACKAGE PURCHASED



- Customers who live in third tier cities are more relatively more likely to make a purchase

EXPLORATORY DATA ANALYSIS
SUMMARY CONCLUSION AND OBSERVATIONS

GENERAL CUSTOMER DEMOGRAPHICS

- 36 years old male
- Executive or Manager
- Salaried Employee
- $22K monthly income
- Travels 3x per year
- Lives in Tier 1 Cities
- Married

PACKAGE PURCHASE STATISTICS

- 18% of customers have purchased a package
- Younger male customers tend to purchase more packages
- Executives who prefer 5 star properties tend to purchase more packages
- Basic and Standard packages are the most purchased kind of packages
- Customer in Tier 3 cities purchase the most packages

RECOMMENDATIONS

- Company should target their marketing toward younger unmarried male customers with higher incomes, work and large business, and live in Tier 3 cities.
- Secondary market are young married couples
- Company should train their salespeople to take their time during presentations and follow-up at least 6 times.

04

MODEL BUILDING

DECISION TREES, BAGGING & RANDOM FORESTS

BUILDING THE PREDICTIVE MODEL

Approach

1. Data preparation
2. Partition the data into train and test set.
3. Build model on the train data.
4. Tune the model if required.
5. Test the data on test set.
6. Measure performance using
    a. Recall and Accuracy

Objectives

- To predict which customers are more likely to purchase new products
- Identify customer characteristics which will help improve the company's marketing effectiveness and drive spending efficiency

Model can make wrong predictions as:

- Predicting an customer will purchase ProdTaken and the customer doesn't purchase (increase marketing costs; False Positive)
- Predicting an customer will not purchase ProdTaken and the customer would have made a purchase (lose customers; False Negative)

Which case is more important?

Predicting that customer will not purchase ProdTaken and the customer would have made a purchase (lose customers). We want to reduce False Negatives and increase True Positives.

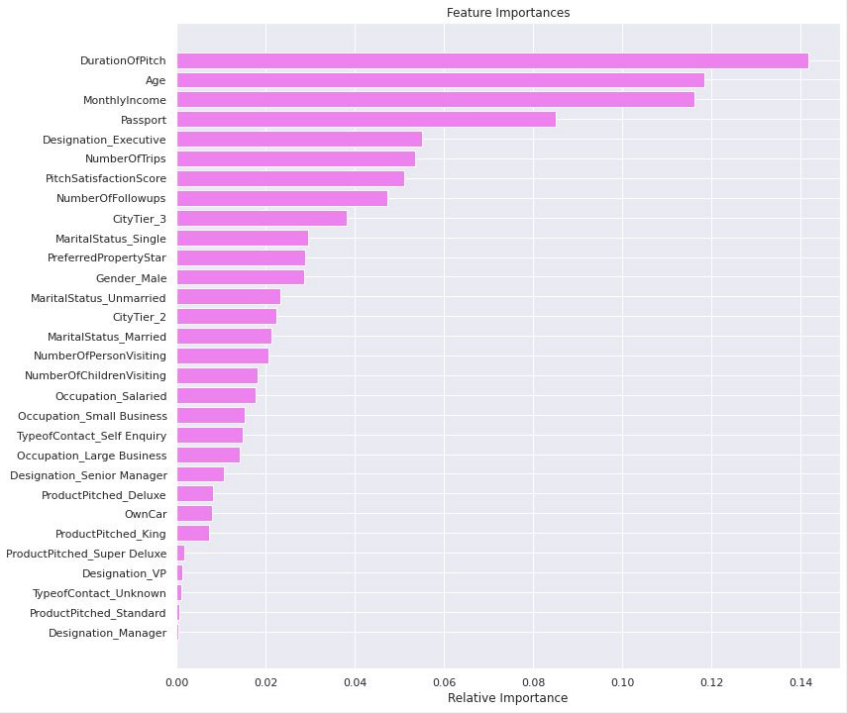
How to reduce this loss i.e need to reduce False Negatives?

Company wants Recall to be maximized, greater the Recall higher the chances of minimizing false negatives. Hence, the focus should be on increasing Recall or minimizing the false negatives or in other words identifying the true positives(i.e. Class 1) so that the company can expand its customer base and increase profits.

Accuracy is also a good score as it can help the company better refine its marketing to be more capital efficient. However, Recall is our primary metric.

DECISION TREE
MODEL PERFORMANCE SUMMARY

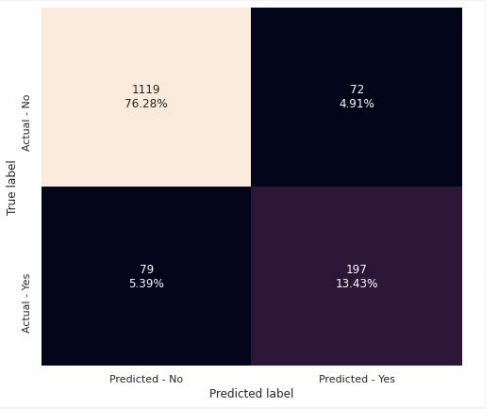
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	1.000	0.897
Recall	1.000	0.714
Precision	1.000	0.732

CONFUSION MATRIX



- Accuracy score of 89.7% exceeds the base accuracy of 81% indicating the model has some predictive value
- Using the Recall measure, the model can label customers who purchased travel packages 71.4% of the time.
- Scores of 100% on the training data indicate a classic overfitting which means this model will not generalize well.
- DurationOfPitch, Age and MonthlyIncome were key features
- True Positive = 197 (higher better)
- False Negative = 79 (lower better)

BAGGING
MODEL PERFORMANCE SUMMARY

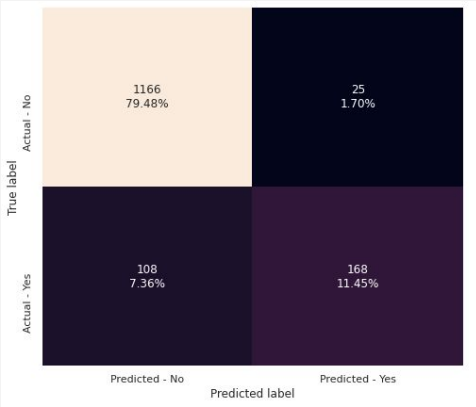
FEATURE IMPORTANCE

Note: there is no feature importance in bagging classification

METRICS

Metric	Train	Test
Accuracy	0.994	0.909
Recall	0.972	0.608
Precision	0.996	0.870

CONFUSION MATRIX



- Recall score on the test data of 60.8% is less than the Decision Tree model
- There is still some overfitting which means this model will not generalize well.
- True Positive = 168 (higher better)
- False Negative = 108 (lower better)

BAGGING with Class Weights
MODEL PERFORMANCE SUMMARY

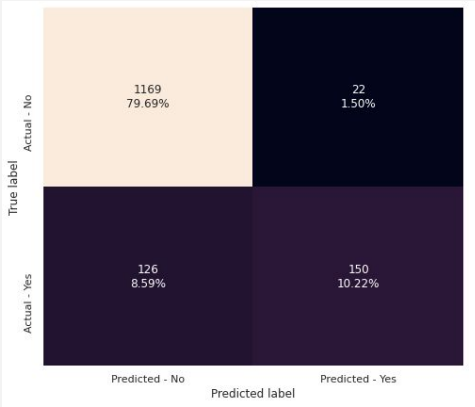
FEATURE IMPORTANCE

Note: there is no feature importance in bagging classification

METRICS

Metric	Train	Test
Accuracy	0.995	0.899
Recall	0.977	0.543
Precision	0.995	0.872

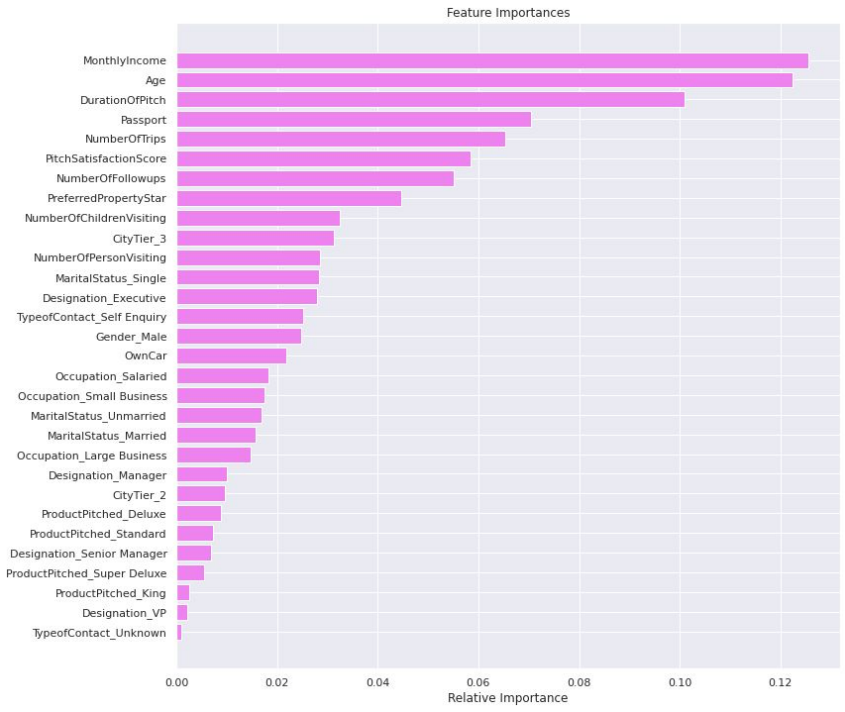
CONFUSION MATRIX



- Class weights were 19% and 81% to mirror the original dataset
- This weighted model underperformed the non-weighted model
- True Positive = 150 (higher better)
- False Negative = 126 (lower better)

RANDOM FOREST
MODEL PERFORMANCE SUMMARY

FEATURE IMPORTANCE



METRICS

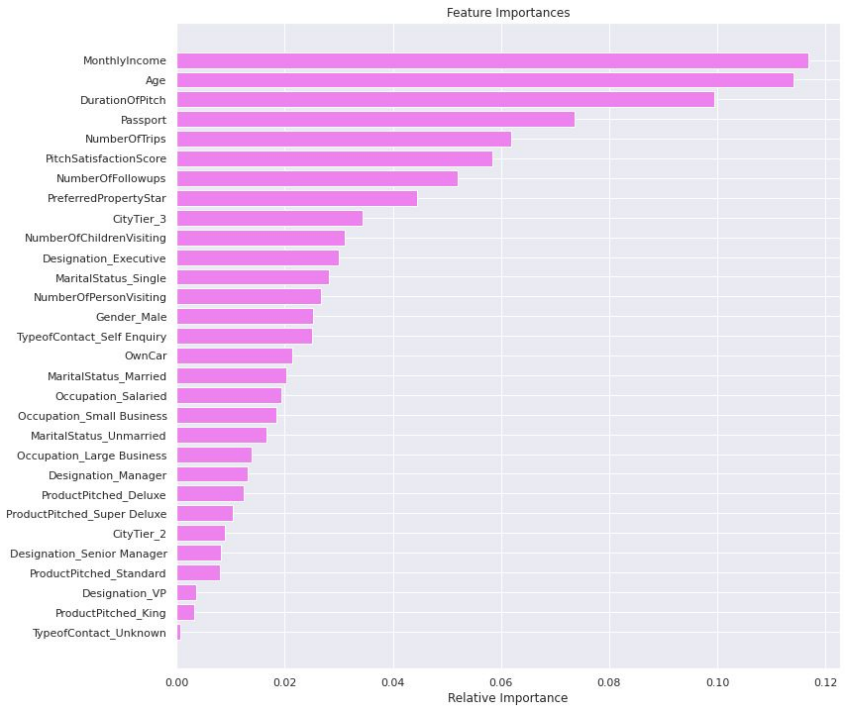
Metric	Train	Test
Accuracy	1.000	0.919
Recall	1.000	0.605
Precision	1.000	0.944

CONFUSION MATRIX

- Accuracy scores of 91.9% is the highest thus far but Recall, our key metric, declined significantly
- Scores of 100% on the training data indicate a classic overfitting which means this model will not generalize well.
- MonthlyIncome and Age were key features
- True Positive = 167 (higher better)
- False Negative = 109 (lower better)

RANDOM FOREST with Class Weights
MODEL PERFORMANCE SUMMARY

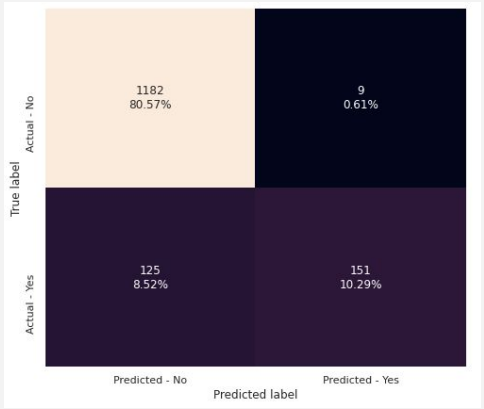
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	1.000	0.909
Recall	1.000	0.547
Precision	1.000	0.943

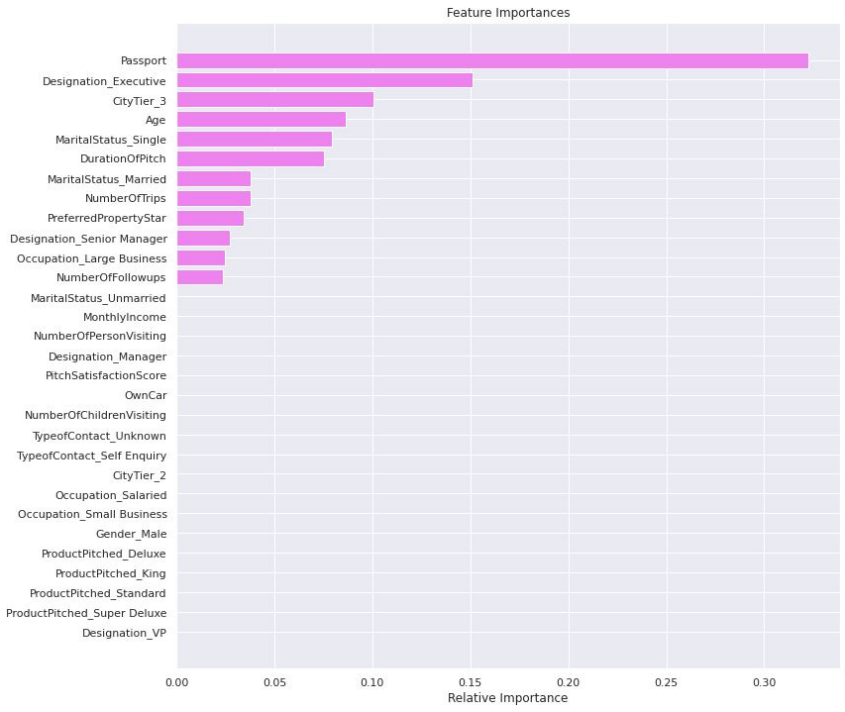
CONFUSION MATRIX



- Class weights were 19% and 81% to mirror the original dataset
- Recall and Accuracy scores declined versus previous models
- Scores of 100% on the training data indicate a classic overfitting which means this model will not generalize well.
- MonthlyIncome and Age were again key features
- True Positive = 151 (higher better)
- False Negative = 125 (lower better)

DECISION TREE - Tuned
MODEL PERFORMANCE SUMMARY

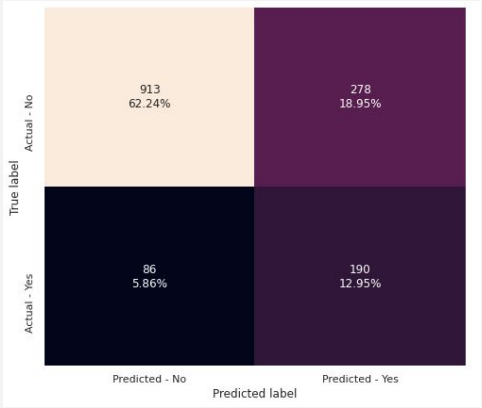
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	0.750	0.752
Recall	0.682	0.688
Precision	0.403	0.406

CONFUSION MATRIX



- Recall and Accuracy scores are not the highest but the model appears to be generalizing well which is key
- Feature Importance was trimmed significantly by the tuning.
- Passport, and Designation_Executive emerged as key features
- True Positive = 190 (higher better)
- False Negative = 86 (lower better)

BAGGING - Tuned
MODEL PERFORMANCE SUMMARY

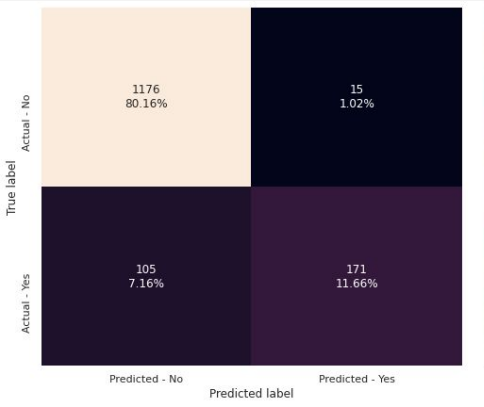
FEATURE IMPORTANCE

Note: there is no feature importance in bagging classification

METRICS

Metric	Train	Test
Accuracy	1.000	0.918
Recall	1.000	0.620
Precision	1.000	0.919

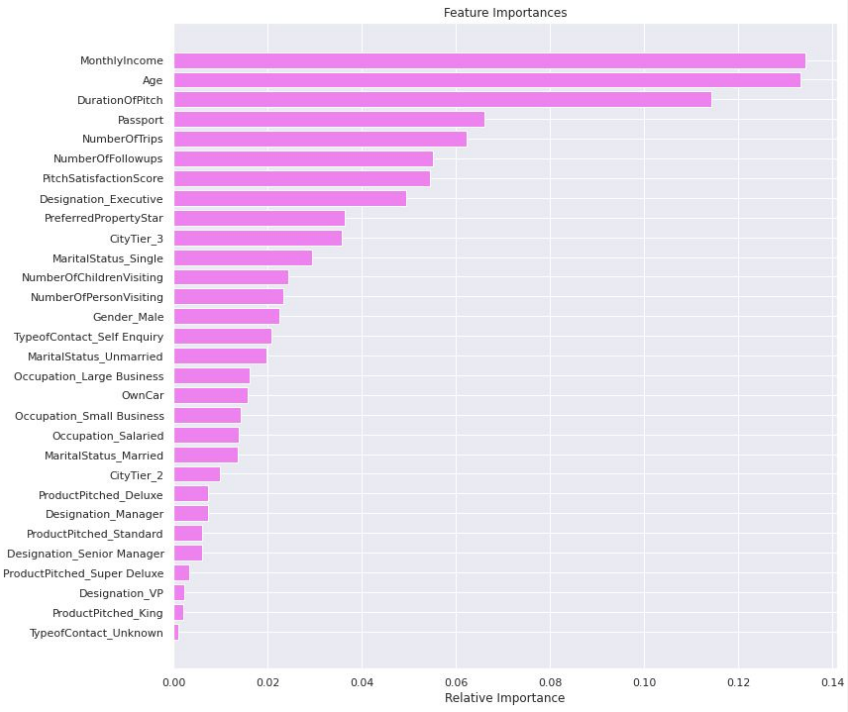
CONFUSION MATRIX



- Bagging continues to underperform on recall relative to Decision Trees and Random Forests
- We continue to see overfitting
- True Positive = 171 (higher better)
- False Negative = 105 (lower better)

RANDOM FOREST - Tuned
MODEL PERFORMANCE SUMMARY

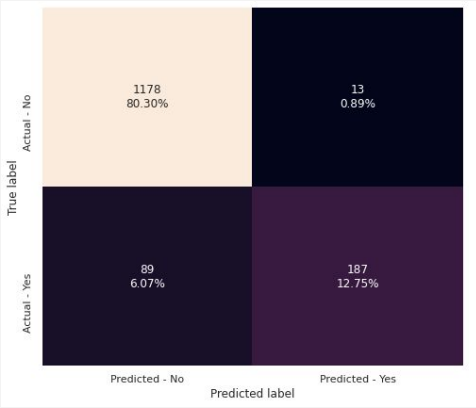
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	1.000	0.930
Recall	1.000	0.678
Precision	1.000	0.935

CONFUSION MATRIX



- Accuracy and Precision are high but Recall underperform the other models.
- There continues to be overfitting which is not unusual for Random Forest models
- MonthlyIncome, Age, and DurationOfPitch emerged as strongest features
- True Positive = 187 (higher better)
- False Negative = 89 (lower better)

MODEL PERFORMANCE
SUMMARY

COMPARING TREE, FOREST, AND BAGGED MODELS - Performance Summary

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	1.000	0.897	1.000	0.714	1.000	0.732
Tuned Decision Tree	0.750	0.752	0.681	0.688	0.403	0.406
Bagging Classifier	0.994	0.909	0.972	0.609	0.997	0.870
Weighted Bagging Classifier	0.995	0.899	0.977	0.543	0.995	0.872
Tuned Bagging Classifier	1.000	0.918	1.000	0.620	1.000	0.919
Random Forest	1.000	0.919	1.000	0.605	1.000	0.944
Weighted Random Forest	1.000	0.909	1.000	0.547	1.000	0.944
Tuned Random Forest	1.000	0.930	1.000	0.678	1.000	0.935

- Decision Tree produced the best recall scores (0.714) but is overfitting the training data
- Tuned Decision Tree model had slightly lower recall scores (training 0.681, test 0.688) but seems to be a better generalized model
- An argument could be made for the Tuned Random Forest Model because it has really high Accuracy and Precision scores with just a slightly lower Recall.

  However, there would need to be additional hyperparameter tuning work done to ensure the model is not overfitting.

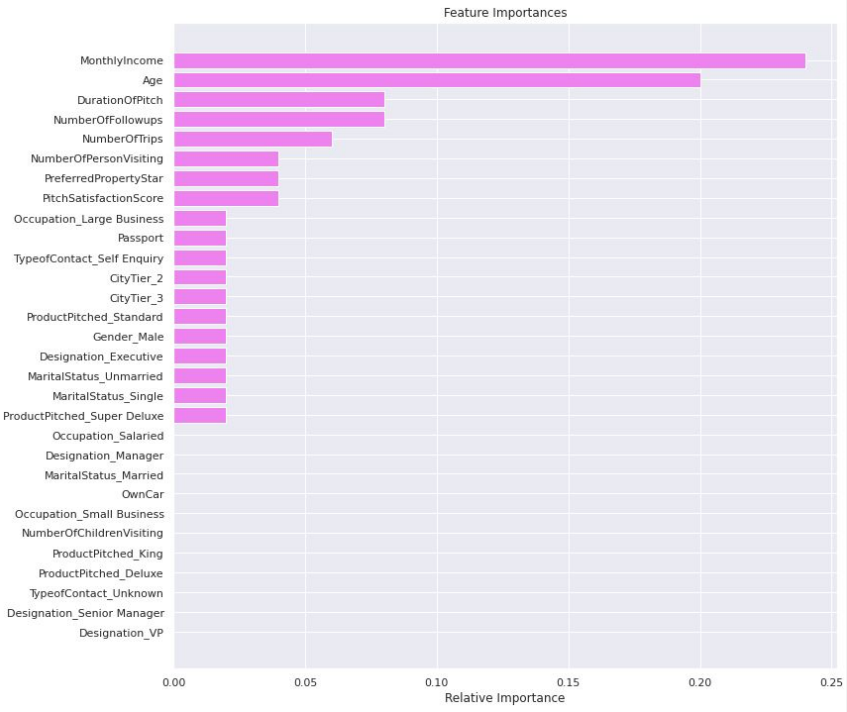
05

MODEL BUILDING

ADABOOST, GRADIENT BOOST, XGBOOST & STACKING CLASSIFIERS

ADABOOST CLASSIFIER
MODEL PERFORMANCE SUMMARY

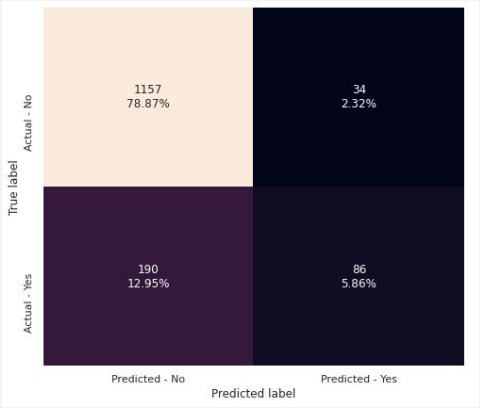
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	0.845	0.847
Recall	0.315	0.312
Precision	0.695	0.717

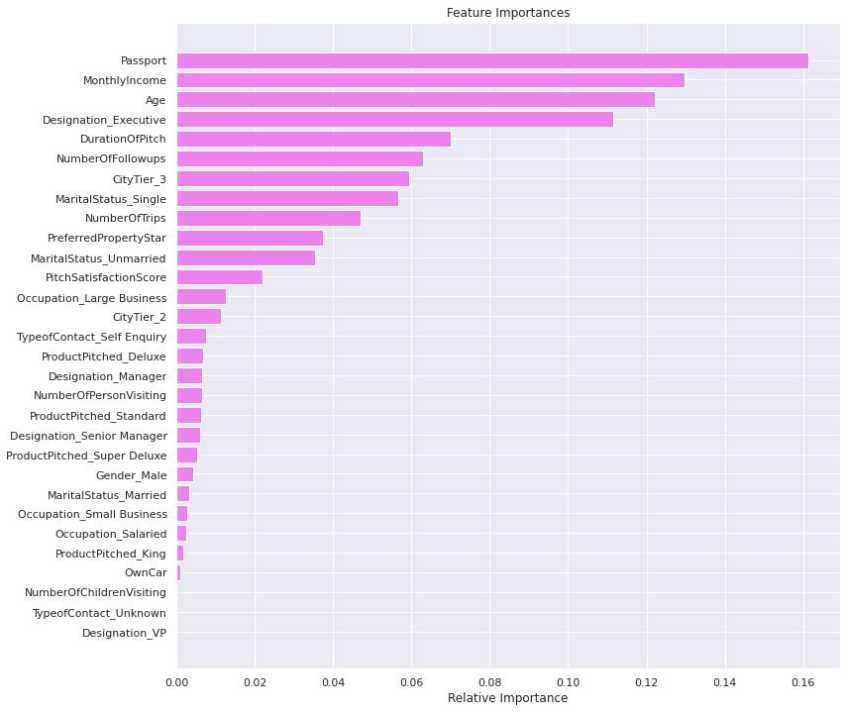
CONFUSION MATRIX



- All metrics significantly underperforms the Decision Tree and Random Forest Models
- MonthlyIncome, Age, and DurationOfPitch are again the strongest features
- True Positive = 86 (higher better)
- False Negative = 190 (lower better)

GRADIENT BOOST CLASSIFIER
MODEL PERFORMANCE SUMMARY

FEATURE IMPORTANCE



METRICS

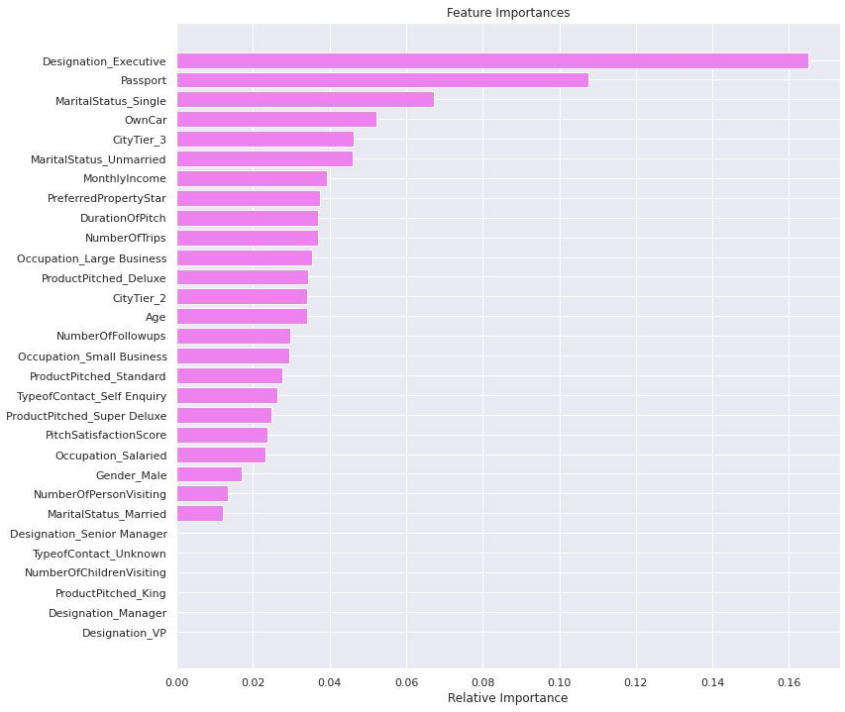
Metric	Train	Test
Accuracy	0.889	0.865
Recall	0.468	0.394
Precision	0.890	0.784

CONFUSION MATRIX

- This model does not perform as well as previous models
- Passport is the strongest feature
- True Positive = 109 (higher better)
- False Negative = 167 (lower better)

XGBOOST CLASSIFIER
MODEL PERFORMANCE SUMMARY

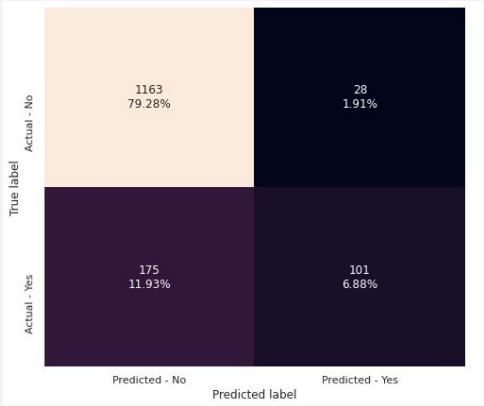
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	0.879	0.861
Recall	0.417	0.365
Precision	0.876	0.782

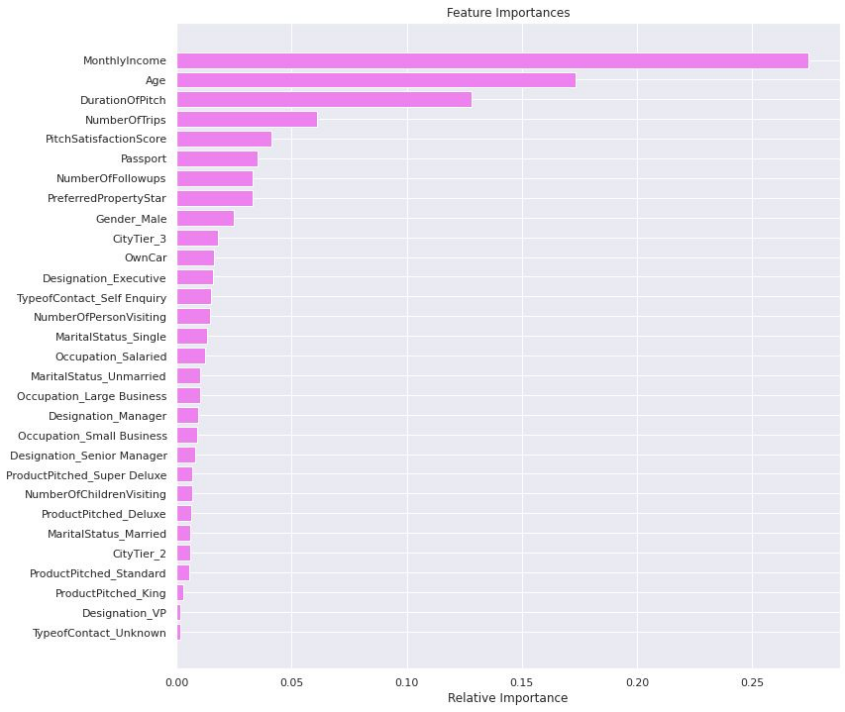
CONFUSION MATRIX



- Recall does not perform as well as previous models
- Designation_Executive has emerged as strongest feature along with Passport and MaritalStatus_Single
- True Positive = 101 (higher better)
- False Negative = 175 (lower better)

ADABOOST CLASSIFIER - Tuned
MODEL PERFORMANCE SUMMARY

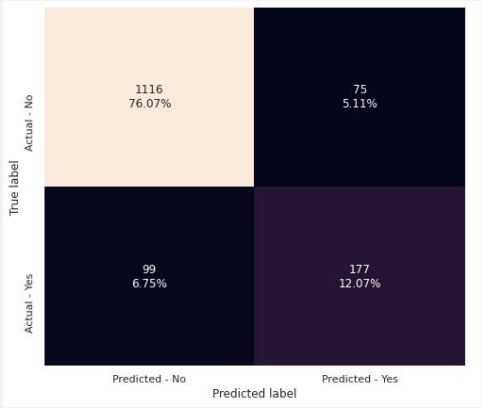
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	0.992	0.930
Recall	0.972	0.641
Precision	1.000	0.935

CONFUSION MATRIX

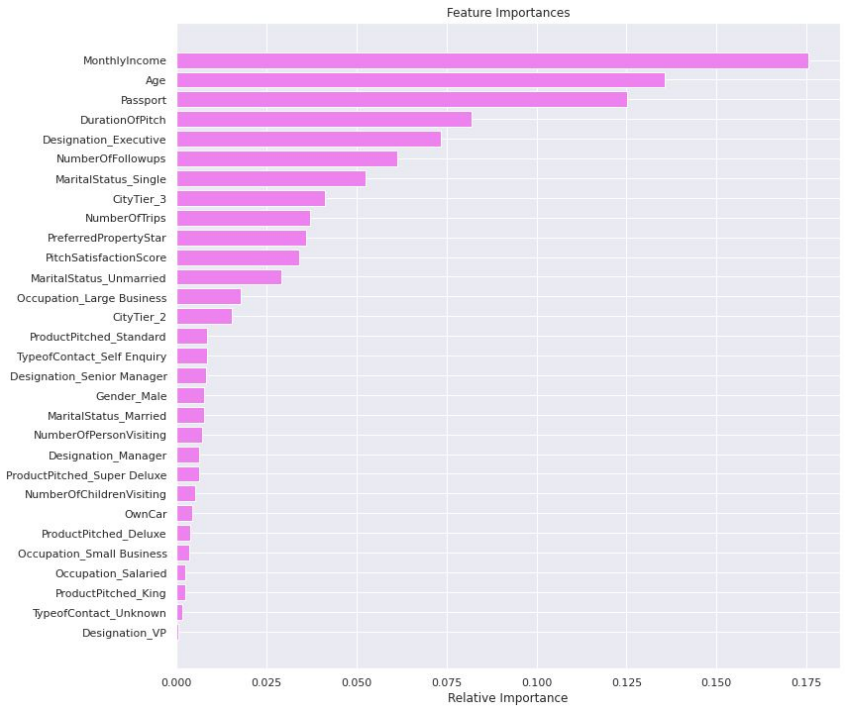


- Recall improved significantly from the non-tuned Adaboost model 64% vs 31%
- This is a good generalized model but other models perform better
- True Positive = 177 (higher better)
- False Negative = 99 (lower better)

GRADIENT BOOST CLASSIFIER - Tuned

MODEL PERFORMANCE SUMMARY

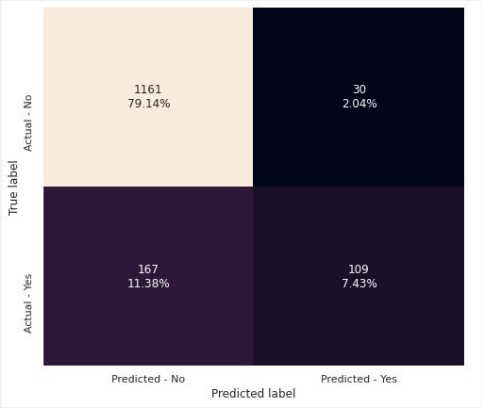
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	1.000	0.884
Recall	0.631	0.503
Precision	0.939	0.812

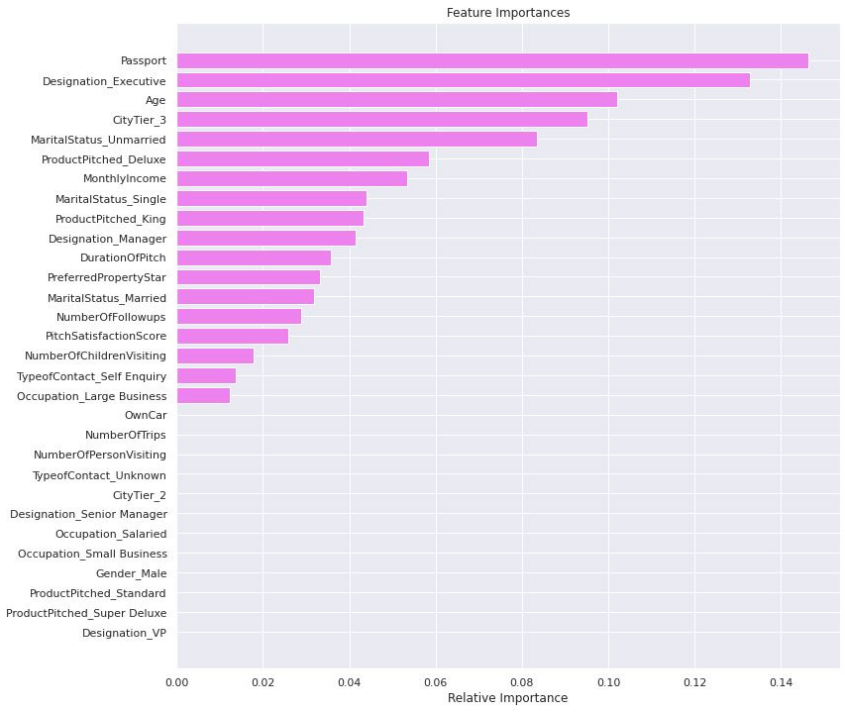
CONFUSION MATRIX



- Model is better than non-tuned but still underperforms the Decision Tree
- True Positive = 139 (higher better)
- False Negative = 137 (lower better)

XGBOOST CLASSIFIER - Tuned
MODEL PERFORMANCE SUMMARY

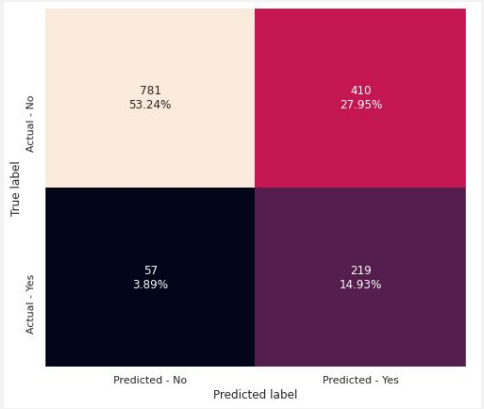
FEATURE IMPORTANCE



METRICS

Metric	Train	Test
Accuracy	0.667	0.681
Recall	0.768	0.793
Precision	0.333	0.348

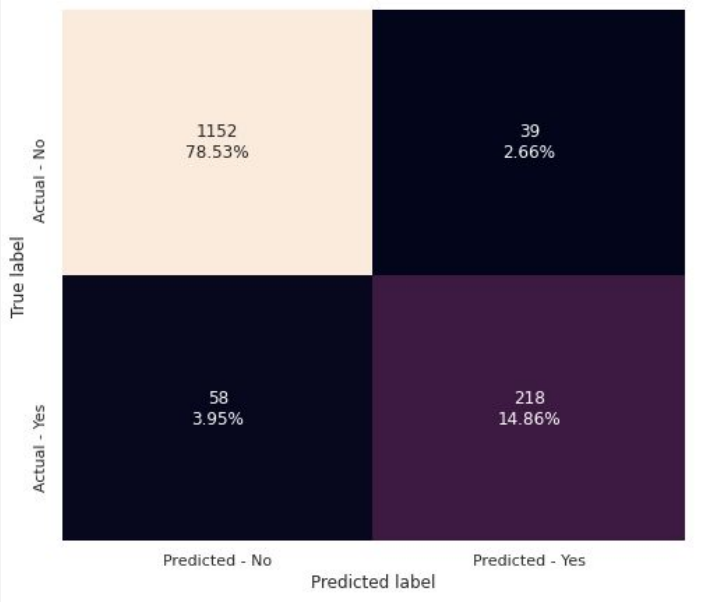
CONFUSION MATRIX



- XGBoost Tuned appears to be best overall model
- 0.793 is the highest Recall of any of the models
- The models appears to be generalizing well and not overfitting
- Passport ad Designation_Executive are most prominent features
- Model as produced the highest True Positives and lowest False Negatives of any model thus far
- True Positive = 219 (higher better)
- False Negative = 57 (lower better)

STACKING CLASSIFIER
MODEL PERFORMANCE SUMMARY

CONFUSION MATRIX



METRICS

Metric	Train	Test
Accuracy	1.000	0.933
Recall	1.000	0.789
Precision	1.000	0.848

- Overall performance is very strong (second highest Recall) but there is overfitting
- True Positive = 218 (higher better)
- False Negative = 58 (lower better)

MODEL PERFORMANCE
SUMMARY

COMPARING ADABOOST, GRADIENTBOOST, XGBOOST, AND STACKING

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
ADABoost with default parameters	0.845	0.847	0.315	0.312	0.695	0.717
ADABoost Tuned	0.993	0.881	0.972	0.641	0.991	0.702
Gradient Boosting with default parameters	0.889	0.866	0.469	0.395	0.891	0.784
Gradient Boosting with AdaBoost	0.891	0.868	0.477	0.395	0.895	0.807
Gradient Boosting Tuned	0.923	0.885	0.632	0.504	0.940	0.813
XGBoost with default parameters	0.879	0.862	0.418	0.366	0.876	0.783
XGBoost Tuned	0.667	0.682	0.769	0.793	0.334	0.348
Stacking Estimator	1.000	0.934	1.000	0.790	1.000	0.848

- Decision Tree produced the best recall scores (0.714) but is overfitting the training data
- Tuned Decision Tree model had slightly lower recall scores (training 0.681, test 0.688) but seems to be a better generalized model
- An argument could be made for the Tuned Random Forest Model because it has really high Accuracy and Precision scores with just a slightly lower Recall.

  However, there would need to be additional hyperparameter tuning work done to ensure the model is not overfitting.

COMPARING TREE, FOREST, AND BAGGED MODELS

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	1.000	0.897	1.000	0.714	1.000	0.732
Tuned Decision Tree	0.750	0.752	0.681	0.688	0.403	0.406
Bagging Classifier	0.994	0.909	0.972	0.609	0.997	0.870
Weighted Bagging Classifier	0.995	0.899	0.977	0.543	0.995	0.872
Tuned Bagging Classifier	1.000	0.918	1.000	0.620	1.000	0.919
Random Forest	1.000	0.919	1.000	0.605	1.000	0.944
Weighted Random Forest	1.000	0.909	1.000	0.547	1.000	0.944
Tuned Random Forest	1.000	0.930	1.000	0.678	1.000	0.935

COMPARING ADABOOST, GRADIENTBOOST, XGBOOST, AND STACKING

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
ADABoost with default parameters	0.845	0.847	0.315	0.312	0.695	0.717
ADABoost Tuned	0.993	0.881	0.972	0.641	0.991	0.702
Gradient Boosting with default parameters	0.889	0.866	0.469	0.395	0.891	0.784
Gradient Boosting with AdaBoost	0.891	0.868	0.477	0.395	0.895	0.807
Gradient Boosting Tuned	0.923	0.885	0.632	0.504	0.940	0.813
XGBoost with default parameters	0.879	0.862	0.418	0.366	0.876	0.783
XGBoost Tuned	0.667	0.682	0.769	0.793	0.334	0.348
Stacking Estimator	1.000	0.934	1.000	0.790	1.000	0.848

MODEL PERFORMANCE
SUMMARY

- XGBoost Tuned has the highest Recall and generalizes well making it the best model
- Stacking Estimator had the second highest Recall but is overfitting

06

KEY INSIGHTS & RECOMMENDATIONS

KEY INSIGHTS

OBJECTIVES

1. **To predict which customers are more likely to purchase new travel package products or not.**
   Key Insight: Young male, unmarried executives with higher incomes and live in third tier cities and have passport purchase the most travel packages and should be targeted.

2. **Identify customer and market characteristics which will help improve the company's marketing effectiveness and efficiency.**
   Key Insight: Sales people with longer sales pitches and more follow ups are most successful. Customers like to be pitch expensive packages but are more likely to purchase standard or deluxe packages.

   Key indicators: Passport, Executive, Young Age, Tier 3 Cities, Unmarried or Single

BUSINESS RECOMMENDATIONS

1. **Key Insight: Young unmarried male executives purchase the most travel packages**
   Recommendation: Target these customers with packages tailored to their needs such as adventure travel

2. **Key Insight - Customers who live in tier 3 cities are more likely to travel**
   Recommendation: Target these customers with trips to tier 1 cities for weekend getaways

3. **Key Insight - Deluxe and Standard packages are the most popular**
   Recommendation: Develop more budget travel and all inclusive packages for budget travelers to increase travel frequency

4. **Key Insight - Customer who are pitched slowly and followed up with often are more likely to purchase**
   Recommendation: Ensure sales staff are properly trained on their sales and marketing materials. Invest in a customer relationship management system with follow up reminders, automated emails and sales tracking to help sales staff build long term relationships with customers

**BONUS - Young married couples with high incomes and passports are a strong secondary market which could be developed