Wellness Tourism Package - Customer Analysis

Visit with us Tourism Company

Business Overview and Solution Approach

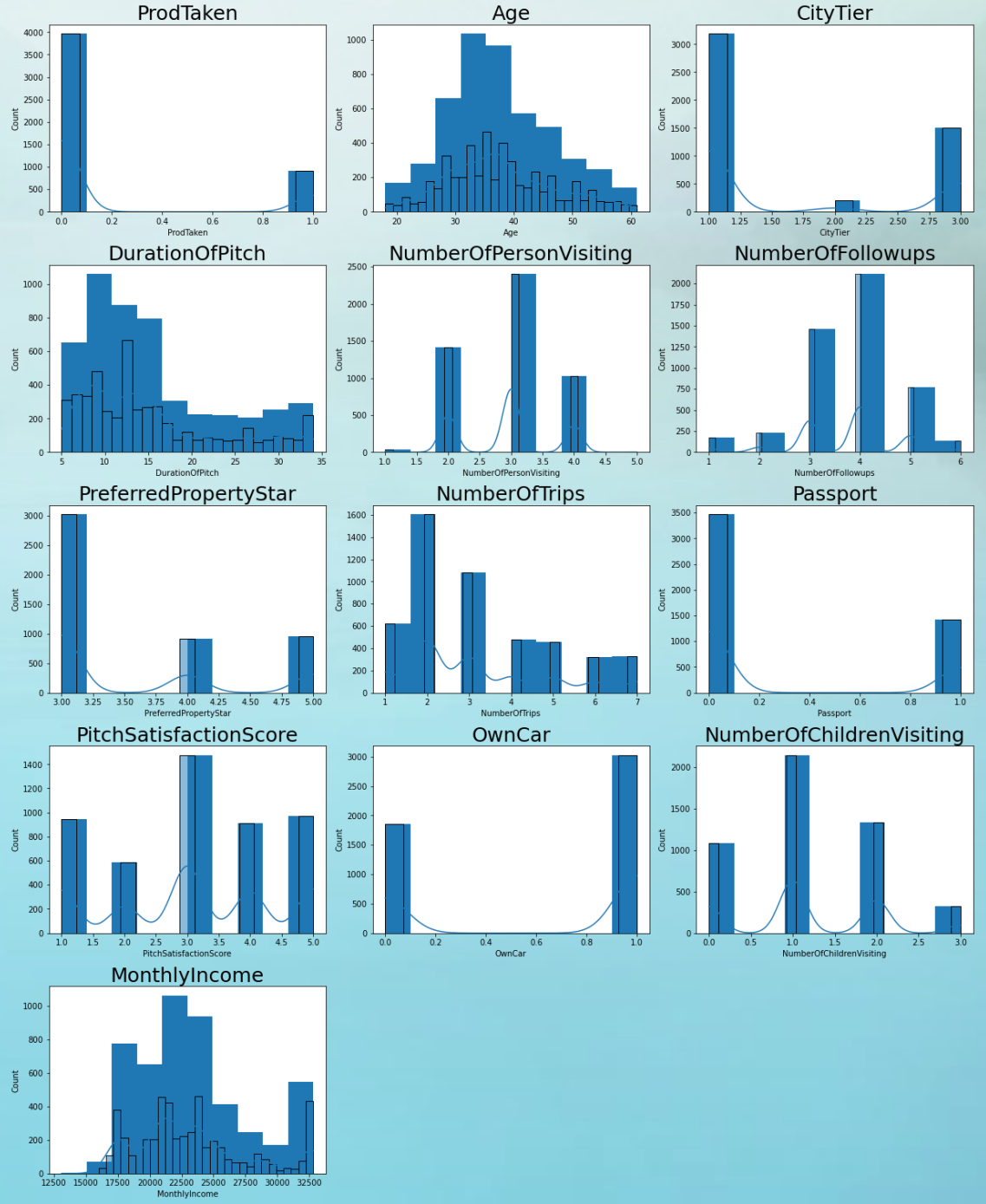
- ‘Visit with us’ is a tourism company currently offering a variety of somewhat traditional travel packages. Management wants to offer a new travel package called ‘Wellness Tourism Package’.
- Customers currently purchase packages 18% of the time.
- Customers were surveyed at random and asked various questions, without reviewing currently available information.
- Management wants a model developed that will help predict what customers will be most likely to purchase the new package, as well as provide suggestions to the Marketing Team and policy makers.

Data Set Summary

- *CustomerID* – unique identifier for individual customers
- *ProdTaken* – target variable – denotes whether or not customer purchased package
- *TypeofContact* – whether or not the company called the customer or the customer called the company
- *Occupation* – Type of employment, salaried, small business, large business or free lance
- *Gender* – male or female
- *PackagePitched* – which of 5 packages were pitched to customer
- *MaritalStatus* – majority of customers were married
- *Designation* – Professional designation, most common was ‘Executive’
- *CityTier* – sorted by level of development, population, etc
- *NumberofPersonsVisiting* – number of individuals planning on traveling
- *NumberOfFollowups* – frequency of follow up phone calls
- *PreferredPropertyStar* – when traveling, what star rated accommodations customer prefers
- *NumberOfTrips* – number of travel trips customer experienced during a year
- *Passport* – whether or not customer has valid passport
- *OwnCar* – whether or not customer owns vehicle
- *PitchSatisfactionScore* – survey scores regarding satisfaction of sales pitch
- *NumberOfChildrenVisiting* – Number of children traveling under the age of 5
- *Age*- current age of customer
- *DurationOfPitch* – length of sales pitch
- *MonthlyIncome* - how much money a customer earns in a month

EDA - Univariate Analysis

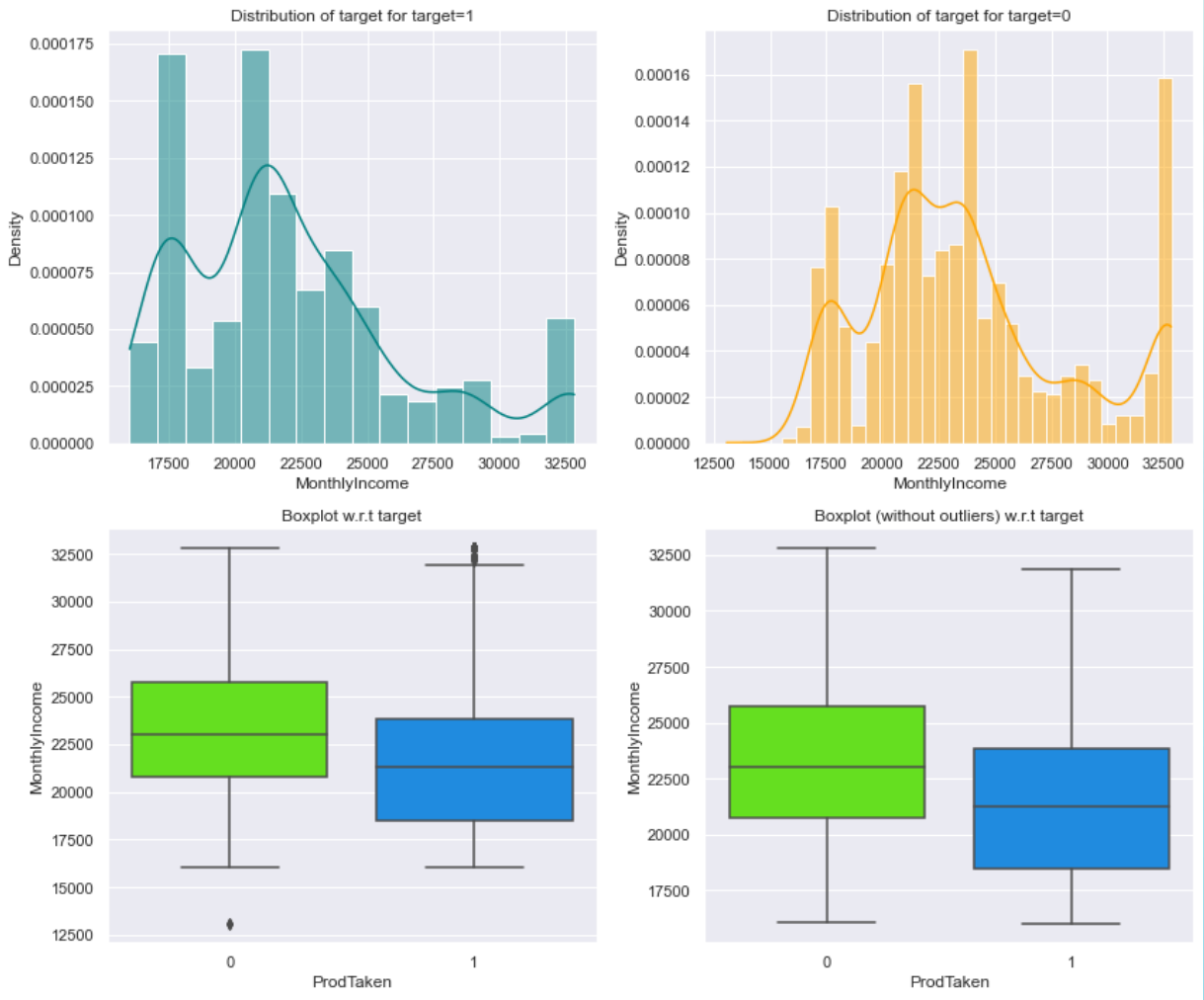
- More customers did not take a package than did, only 18% made purchase
- Age fairly normally distributed
- More customers in 1st tier city – more developed and populated
- Duration of pitch ranged from 5-35 minutes
- Most frequently 3 people were in travel party
- 4 follow up calls were most often made
- More customers preferred 3 star accommodations
- Majority of customers traveled twice a year
- More customers had passports than didn’t
- There was an average of 3 scored on pitch satisfaction
- More customers owned vehicles than didn’t
- 1-2 children were most frequently traveling
- Monthly income ranged from 12,500 to 32,500

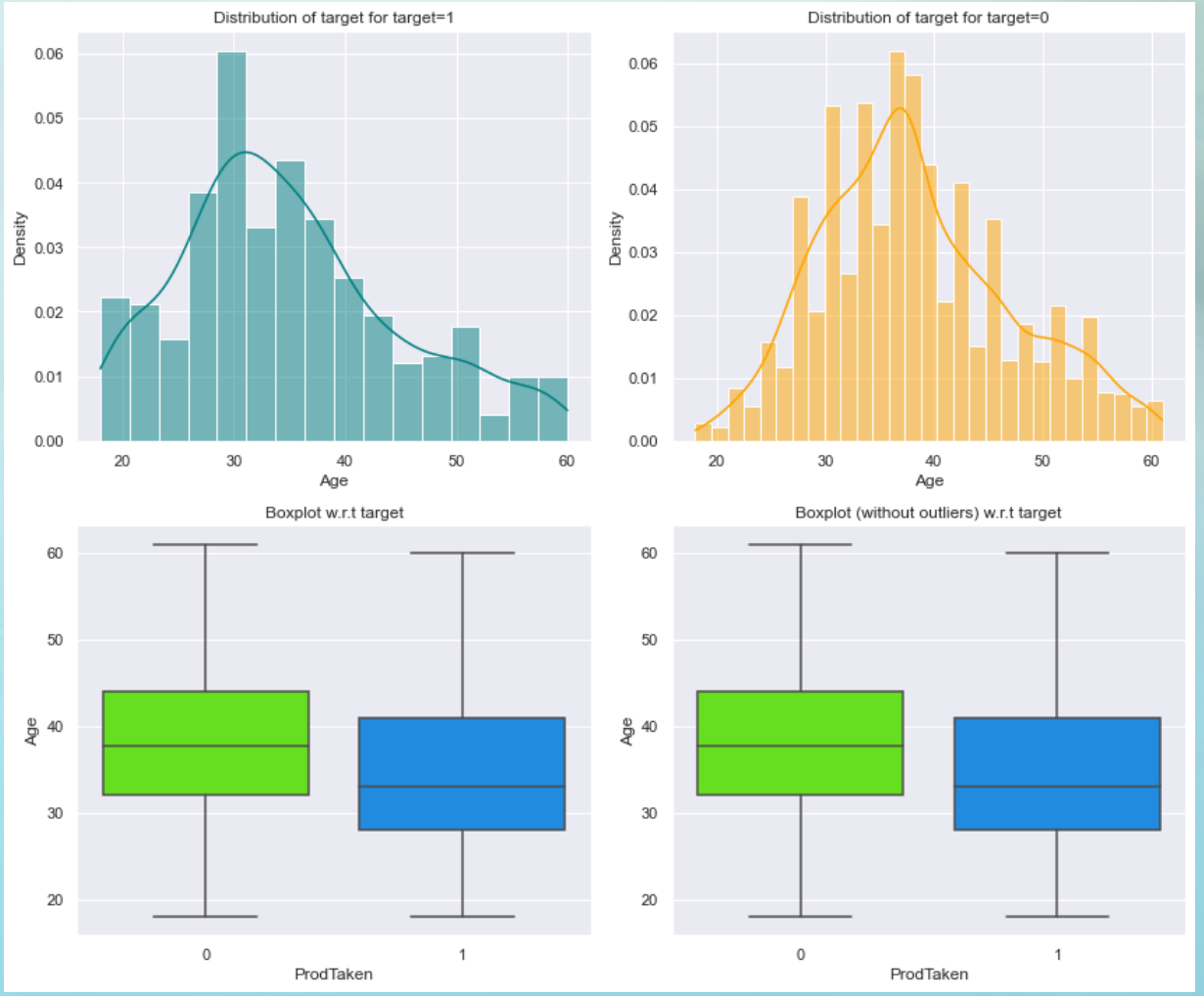


EDA - Bivariate Analysis - Correlations

- The features with the highest correlation of 0.61 are Number of Children Visiting and Number of Persons Visiting – which is reasonable
- Age and Monthly Income have a correlation of 0.47 – again reasonable
- Interestingly there is a negative correlation of -0.14 between Product Taken and both Monthly Income and Age

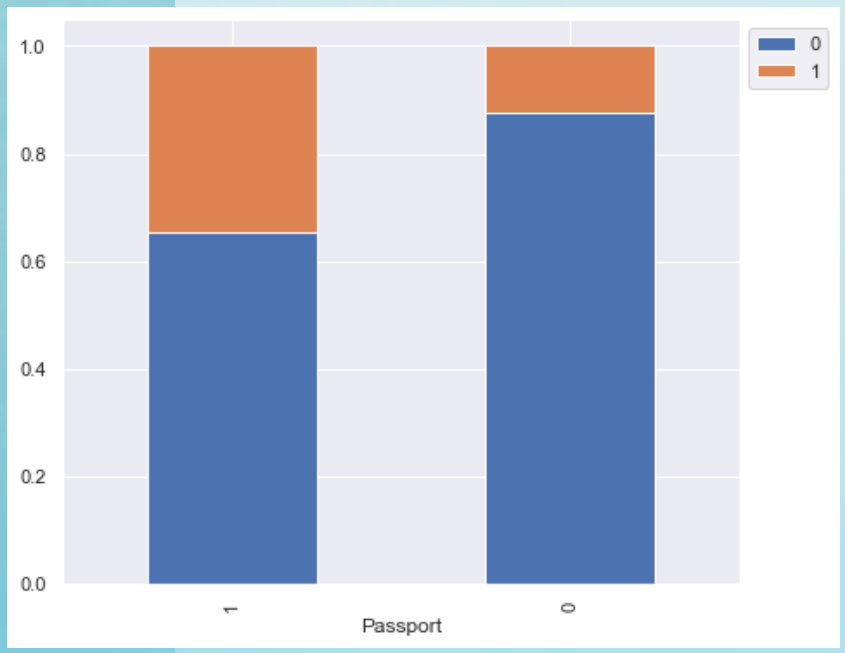
EDA - Bivariate Analysis



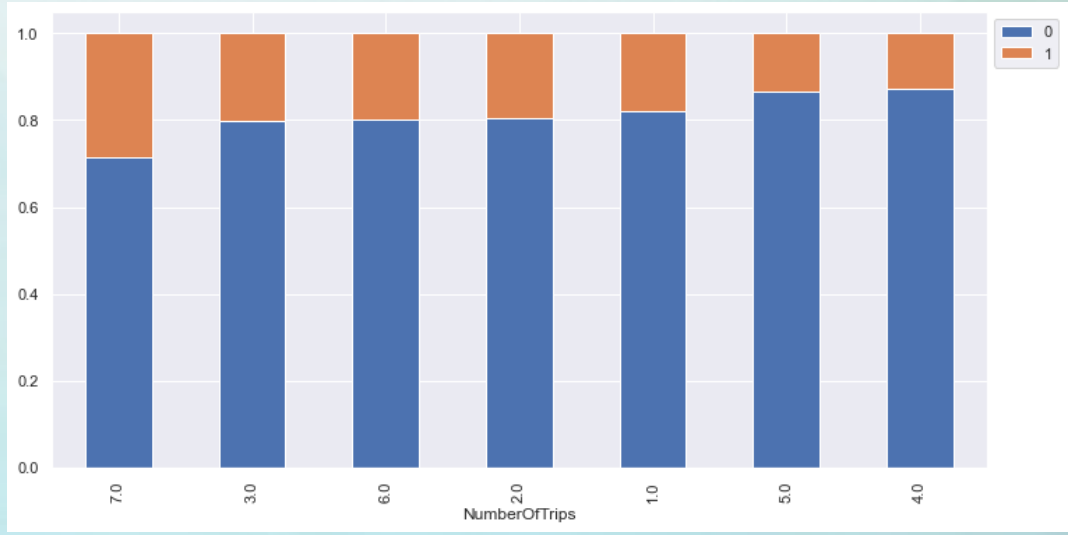


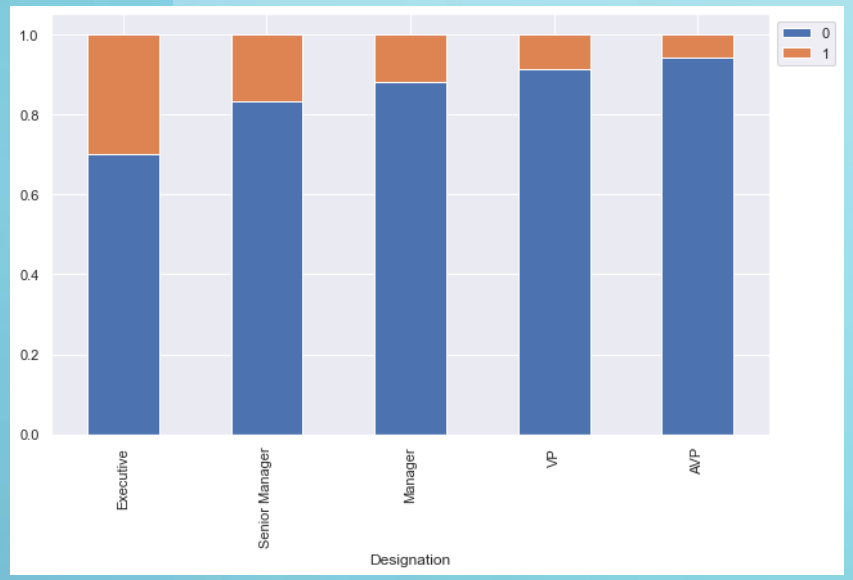
- The median monthly income of customers who purchased packages is lower than those who did not. This would explain the negative correlation
- The median age of customers who purchased packages is also lower than those who did not. Again, this would explain the negative correlation.

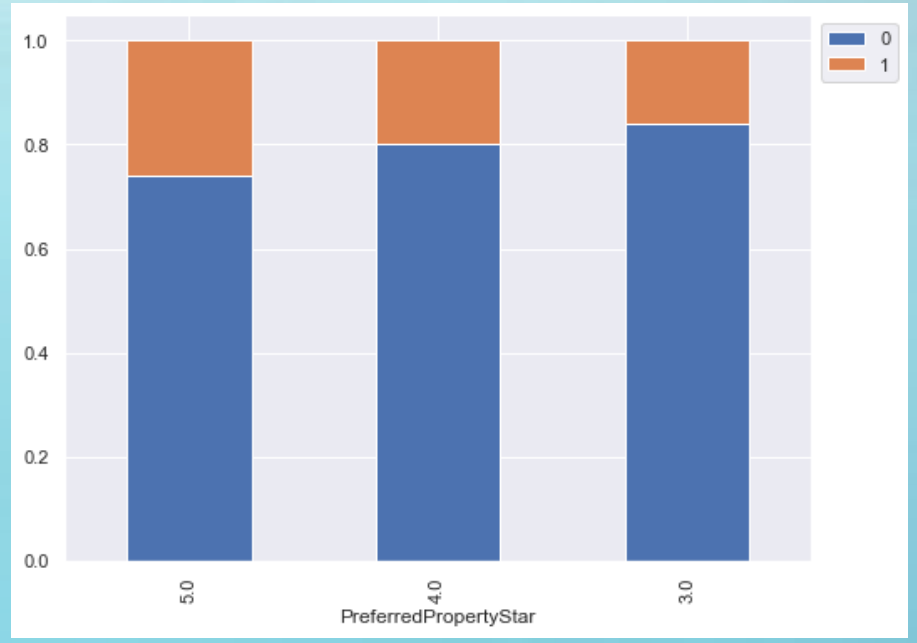
EDA-Bivariate Analysis



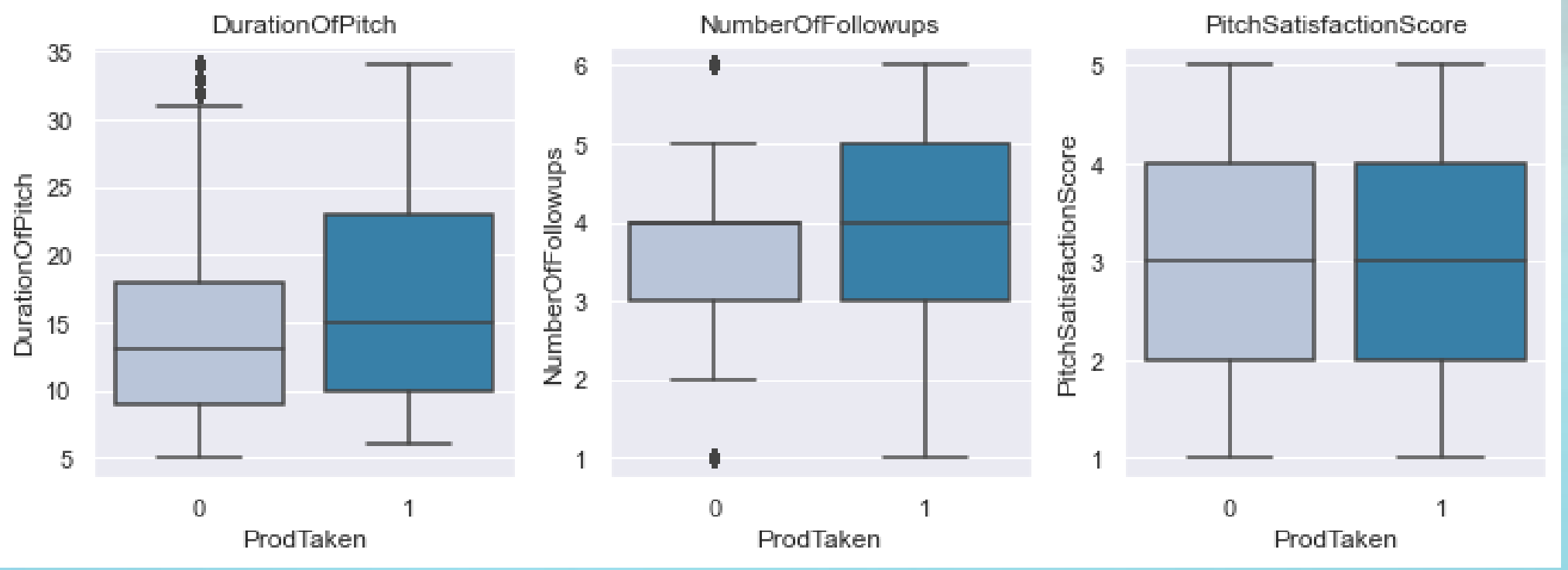
- Customer who had passports purchased more frequently than those who didn't
- Those with professional designation of 'Executive' purchased more frequently
- Customers who traveled 3 times during the year were more frequent than most and also purchased packages more often
- Customers who purchased packages more often preferred 5-star accommodations







EDA - Bivariate Analysis



- Duration of sales pitch that resulted in sale ranged most frequently between 10 & 23 minutes
- The number of follow up calls that resulted in sales ranged between 3-5 calls
- Pitch satisfaction scores were very middle of the road with no difference in those who did and did not purchase packages

Data Cleaning and Preparation for Modeling

- Missing Data and Outliers
  - 'Age‘ is fairly normally distributed - the mean was used to replace all missing values.
  - 'TypeofContract' is a categorical feature - the mode was used for missing values
  - 'DurationofPitch' is heavily right skewed with extreme outliers - the median was used for missing values. Outliers were capped at 1.5 times the upper quartile figure.
  - 'NumberOfFollowups' is a discrete number - mode was used for missing values.
  - 'PreferredPropertyStar' is also a discrete number and will use mode to replace missing values.
  - 'NumberOfTrips' is also discrete - mode was used for replacement. This feature is heavily right skewed with outliers that were capped at 1.5 times the upper quartile figure.
  - 'NumberOfChildrenVisiting' is again, discrete and mode will be used for replacement.
  - 'MonthlyIncome' is reasonably normally distributed but with outliers that do not appear to have much affect on the mean, therefore mean will be used. This feature were capped at 1.5 times the upper quartile figure.

Data Cleaning and Preparation for Modeling

After Exploratory Data Analysis, the following columns were dropped:

- CustomerID – unique to each customer, not useful for modeling
- Gender – while there were more male than female customers, percentagewise, those purchasing packages were closely split between genders and therefore this feature did not hold much value for analysis
- OwnCar – this feature didn’t appear to have much influence on purchasing a package as percentage wise, the acceptance rate for those who did and didn’t own a vehicle were very similar
- All customer interaction data was dropped because, while it is useful for sales and marketing strategies, it is not very meaningful for determining a customer profile the model will predict.

Model Evaluation Criteria

Possible Incorrect Predictions:

1. Customer predicted not to purchase travel package but does make purchase
2. Customer predicted to buy travel package but does not make purchase

More Critical Incorrect Prediction:

Predicting a customer will purchase a travel package, but does not, is more important than predicting a customer will not purchase a travel package but does. This is due to the cost factor in targeting a specific demographic with marketing and sales campaigns.

Model Strategy:

To avoid predicting a customer will purchase a package when they do not, the Precision metric should be maximized. For this model, Precision compares the number of customers who were predicted to purchase a travel package and actually did to the total number of customers who were predicted to purchase a travel package. This model will help assess the features and target demographics to focus future sales and marketing campaigns thereby maximizing the financial investment in those campaigns.

Model Comparison – Training Data Set

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	Stacking Classifier
Accuracy	1.0	0.845075	1.0	0.998538	0.990938	1.0	0.844782	1.0	0.884244	0.912599	0.995031	0.990354
Recall	1.0	0.181677	1.0	0.992236	0.953416	1.0	0.310559	1.0	0.464286	0.594720	0.975155	0.959627
Precision	1.0	0.975000	1.0	1.000000	0.998374	1.0	0.696864	1.0	0.854286	0.909739	0.998410	0.988800
F1	1.0	0.306283	1.0	0.996103	0.975377	1.0	0.429646	1.0	0.601610	0.719249	0.986646	0.973995

A little more than half of the models are essentially perfectly fit to the training data. The tuned Decision Tree, both the AdaBoost Classifiers and both the Gradient Boost classifiers are the models that are not perfectly fit.

Performance Metric Overview

While Precision is the primary metric to optimize, the four essential metrics were computed for reach model.

- *Accuracy* is the percentage of correct predictions, both True Positives and True Negatives
- *Recall* is the percentage of True Positives in relation to all actual positives.
- *Precision* is the percentage of True Positives of all predicted positives. It measures how precise the model is out of all predicted positives.
- *F1* is the weighted average of Precision and Recall. It is used when the goal is a balance between Precision and Recall and there is an uneven class distribution.

Model Comparison – Testing Data Set

	Decision Tree	Decision Tree Estimator	Random Forest Estimator	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	Stacking Classifier
Accuracy	0.854806	0.833674	0.880709	0.895706	0.886162	0.893661	0.830948	0.884799	0.849352	0.858896	0.882754	0.892297
Recall	0.619565	0.123188	0.489130	0.619565	0.547101	0.576087	0.282609	0.539855	0.369565	0.413043	0.547101	0.612319
Precision	0.612903	0.944444	0.798817	0.780822	0.782383	0.803030	0.609375	0.780105	0.684564	0.716981	0.762626	0.768182
F1	0.616216	0.217949	0.606742	0.690909	0.643923	0.670886	0.386139	0.638116	0.480000	0.524138	0.637131	0.68145

- While the Decision Tree Estimator has a higher Precision and Accuracy score than all the other models, the Recall and subsequently F1 score are very low. This is caused by very low True Positives and False Positives. The issue is comparatively high False Negatives, which suggests the model can't guarantee to identify True Positives on future data sets. Even though the original goal was to achieve the highest Precision score, for this model, other parameters prove it is not the best model.
- The Bagging Estimator Tuned is the best model.
- This model is followed closely by the Random Forest Tuned model. The Random Forest Tuned model has lower Precision but higher Recall and F1 than the Bagging Estimator Tuned model.

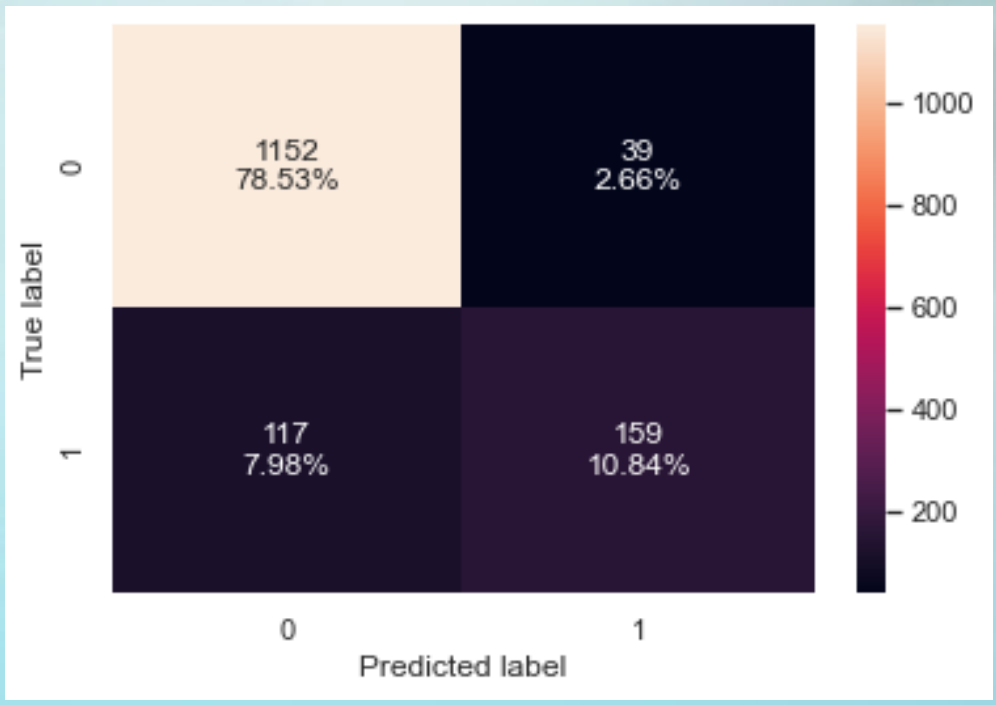
Best Model - Bagging Estimator Tuned Model

Performance Metrics

Training Performance			
Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0
Testing Performance			
Accuracy	Recall	Precision	F1
0.894	0.576	0.803	0.671

While the model is a perfect fit to the training data, it also does the best of all other models to fit the testing data with a Precision score of 80.3%.

Confusion Matrix

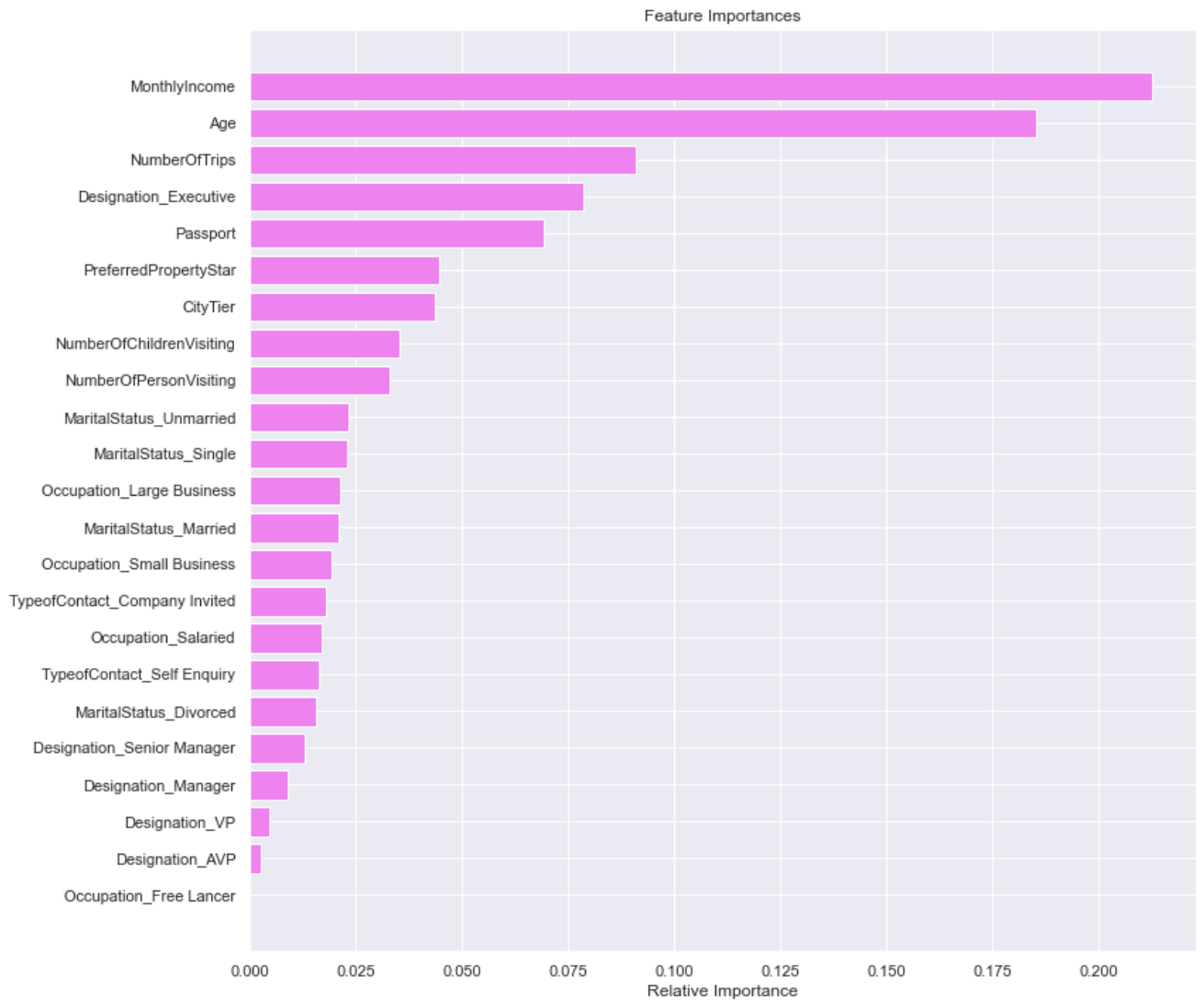


True Positives are greater than False Negatives and False Positives are acceptably low

Confusion Matrix Overview

- *True Positive* (1,1) - Customer purchased a vacation package and model predicted customer would make purchase
- *False Positive* (1,0) - Customer DID NOT purchase a vacation package and model predicted that customer would make purchase
- *True Negative* (0,0) - Customer did not purchase vacation package and model predicted that customer would not make purchase
- *False Negative* (0,1) - Customer purchased a vacation package and model predicted customer WOULD NOT make purchase

Feature Importance



The most important features for the customer profile model, according to the tuned Bagging Estimator model, are Monthly Income, followed by Age, then Number of Trips, employment designation as an Executive and if the customer has a passport. Of less significance are star rating accommodation preference and which tiered city the customer resides. Coincidentially, these features are quite similar across almost all models.

Business Insights and Recommendations

Insights

The type of customers most likely to purchase a travel package have the following characteristics:

- Median monthly income between 20,000 and 22,500. This is lower than the median income of the entire data set because there is a slight negative correlation between customers who purchased packages and monthly income.
- Median age approximately 34 yrs old. This is also lower than the median age of the entire data set as there is also a slight negative correlation between age and customers who purchased packages.
- There is a sweet spot with customers who traveled 3 times per year. This category was the 2nd most common of all number of trips taken and also were the 2nd most frequent group to purchase travel packages.
- Customers with the professional designation of 'Executive' more frequently purchase packages than those with other designations.
- Customers who have passports are more often purchase a travel package.
- More frequent travel package purchasers prefer 5-star properties
- Customers living in the least populated or developed regions (Tier 3) purchase travel packages more often than more populated areas.

When developing sales and marketing strategies attention should be focused on the following areas:

- The average acceptable length of an initial sales pitch should range from 10-23 minutes.
- The range of successful follow-up calls should be between 3 and 5 calls.

Recommendations

- Generate different models based on geographic locations
- Create packages specific to business Executives – business travel could be an overlooked market
- Create a class of packages devoted to Wellness Tourism, both in pricing tiers as well as subcategories such as healthy lifestyle, well-being, ecotourism, etc.

THE END