Visit With Us

Travel Package Purchase Prediction

July 2021

Brian Hall – DSBA – Project 5

Problem

Visit With Us is working to expand their customer base by creating a new product offerings called the Wellness Tourism Package. Currently there are 5 types of packages the company is offering and a marketing campaign from the previous year showed an 18% customer purchase rate, however the costs for the campaign were high. Visit With Us would like to utilize the data from the previous marketing campaign to predict which customers will potentially purchase the newly introduced travel package.

Objective

Utilizing data obtained from the previous successful campaign we will create customer profiles for each product & build machine learning models with the potential to predict what types of customers Marketing could be targeted in a new campaign.

Recommendations (Next Campaign)

Utilize the new Customer Profiles by Product and the individuals predicted within each to create new marketing guidelines for outreach and products to pitch. This would allow better interaction and increased purchase rates with existing customers, while potentially reducing the number of follow ups and duration of pitch. This would also allow for the ability to better market individual packages to appropriate new audiences and increase success of outreach for higher profit tourism packages. Continued next slide

Recommendations continued

- Without additional information on the Wellness Tourism Package, such as expense / benefits, it is difficult to determine which customers may opt for that package. Once the new package is classified along with the Customer Profiles by Product it should be easier to predict who will purchase each product.
- Additional data on customers lifestyles could help improve / create models to predict who will opt for the new Wellness Tourism Package.
- Additional research on what type of individual is more likely to be interested in wellness and travel would also be advised. Possibly introduce a 'Wellness Score' to existing data and rework profiles / re-train the models.
- Monthly Income, Age & Duration of Pitch are showing as the most important features for the model in predicting who will purchase a product, EDA mostly agrees with this.
- Product Pitched is also showing as an important feature and EDA supports this, however the Product Pitched maps perfectly to Designation and has some correlation with Income & Age. Using Designation as the sole means of determining what product to pitch to a given customer is not likely a good practice as many other factors, outside the scope of this data, could impact someones designation. A better practice could be to determine which product to pitch based on the newly developed Customer Profiles by Product. This may also reduce the slight negative correlation that ProductPitched has with ProdTaken.
- Unmarried customers account for ~ 13% of observations and prefer the mid to lower priced products, it would be good to better understand who makes up this group.
- Determine why the Standard package is not pitched more often as it has a higher purchase rate than the Deluxe package which is pitched more than 2x as often.
- Self Enquiry has a 17% product purchase rate, only 4% lower than company invited, but accounts for 70% of observations. Determine if the cost of outreach is worth the additional 4% purchase rate and if putting those dollars to better targeted marketing (predicted purchasers) would make more sense.
- Increase marketing of Standard and Deluxe packages to single customers.
- Reworking the data & data dictionary to better indicate if NumberOfPersonsVisiting include the customer or is in addition. No customers are showing to travel alone, which makes this suspect.
- Collecting additional data on products pitched once new guidelines for how products are pitched are in place will likely increase the predictive power of these models.
- Including additional data such as family size, reason for trip, destination (in country, out of country, distance from home), better details on what drives the satisfaction score, how purchases are made (online, etc), what drove a Self Enquiry, etc. could greatly increase the outcomes of future analysis and marketing plans.

Customer Profiles by Product (Travel Package)

Basic

Age	Pitched to all with a concentration < 25
MonthlyIncome	Concentration of pitches between 17k & 23k
MaritalStatus	Most popular with Single
CityTier	Most popular in Tier 3 cities
Gender, NumberOfPersonVisiting, NumberOfChildrenVisiting, Occupation, OwnCar & Passport are fairly equal in distribution	

Standard

Age	Concentration of pitches between 23k & 30ktween 20 & 50
MonthlyIncome	Concentration of pitches between 17k & 23k
MaritalStatus	Most popular with Unmarried
CityTier	Most popular in Tier 1 cities
Gender, NumberOfPersonVisiting, NumberOfChildrenVisiting, Occupation, OwnCar & Passport are fairly equal in distribution	

Deluxe

Age	Pitched to all with a concentration between 25 & 42
MonthlyIncome	Concentration of pitches between 18k & 26k
MaritalStatus	Most popular in Tier 1 citiesarried
CityTier	Most popular in Tier 1 cities
Gender, NumberOfPersonVisiting, NumberOfChildrenVisiting, Occupation, OwnCar & Passport are fairly equal in distribution	

Super Deluxe

Age	Pitched to customer > 39 with a concentration between 42 & 55
MonthlyIncome	Concentration of pitches between 18k & 26k
MaritalStatus	Most popular in Tier 1 citiesarried
CityTier	Most popular in Tier 1 cities
Gender, NumberOfPersonVisiting, NumberOfChildrenVisiting, Occupation, OwnCar & Passport are fairly equal in distribution	

King

Age	Pitched to customer > 40 with a concentration between 42 & 55
MonthlyIncome	Concentration of pitches between 33k & 38k
MaritalStatus	Most popular with Single, Divorced & Married
CityTier	Most popular in Tier 3 cities
Gender, NumberOfPersonVisiting, NumberOfChildrenVisiting, Occupation, OwnCar & Passport are fairly equal in distribution	

Wellness Travel

Unknown.
No information was given about the attributes of the new package.

EDA Overview

4888 Observations and 20 Features

8 Features had missing values, all features imputed based on EDA. No observations were deleted from the data.

Variable - impute value

Age - median
DurationOfPitch - median
NumberOfFollowups - median
PreferredPropertyStar - median
NumberOfTrips - median
NumberOfChildrenVisiting - median
MonthlyIncome - mean
TypeofContact - 'Self Enquiry'

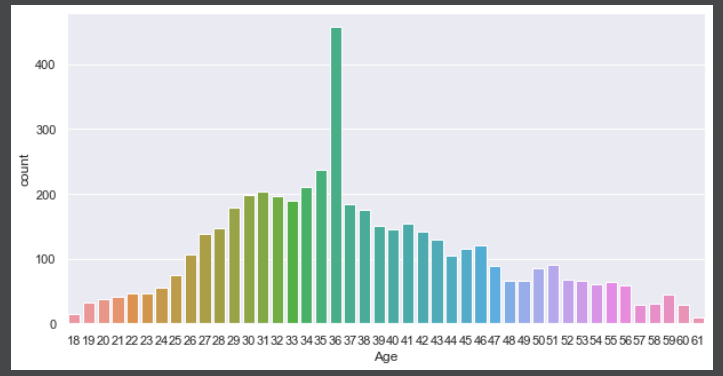
	count	mean	std	min	25%	50%	75%	max
Age	4662.0	37.622265	9.316387	18.0	31.0	36.0	44.0	61.0
CityTier	4888.0	1.654255	0.916583	1.0	1.0	1.0	3.0	3.0
DurationOfPitch	4637.0	15.490835	8.519643	5.0	9.0	13.0	20.0	127.0
NumberOfPersonVisiting	4888.0	2.905074	0.724891	1.0	2.0	3.0	3.0	5.0
NumberOfFollowups	4843.0	3.708445	1.002509	1.0	3.0	4.0	4.0	6.0
PreferredPropertyStar	4862.0	3.581037	0.798009	3.0	3.0	3.0	4.0	5.0
NumberOfTrips	4748.0	3.236521	1.849019	1.0	2.0	3.0	4.0	22.0
PitchSatisfactionScore	4888.0	3.078151	1.365792	1.0	2.0	3.0	4.0	5.0
NumberOfChildrenVisiting	4822.0	1.187267	0.857861	0.0	1.0	1.0	2.0	3.0
MonthlyIncome	4655.0	23619.853491	5380.698361	1000.0	20346.0	22347.0	25571.0	98678.0

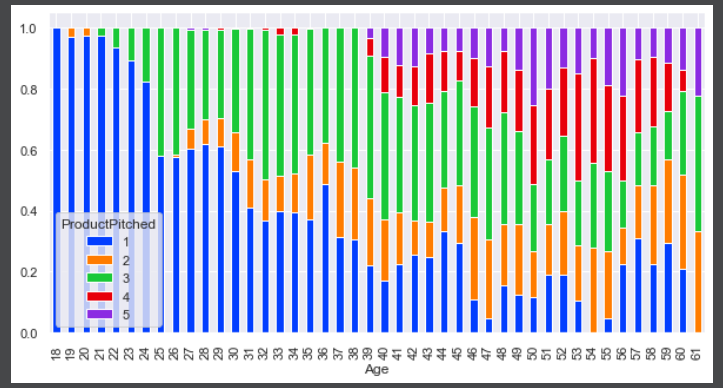
	count	unique	top	freq
ProdTaken	4888	2	0	3968
TypeofContact	4863	2	Self Enquiry	3444
Occupation	4888	4	Salaried	2368
Gender	4888	3	Male	2916
ProductPitched	4888	5	Basic	1842
MaritalStatus	4888	4	Married	2340
Passport	4888	2	0	3466
OwnCar	4888	2	1	3032
Designation	4888	5	Executive	1842

Visit With Us

EDA Overview

Monthly Income & Age were found to be the 2 most important features through modeling and EDA



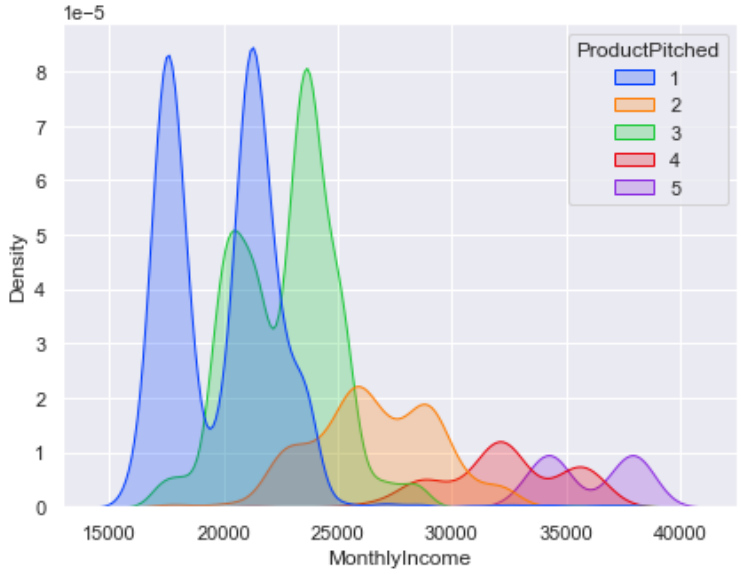


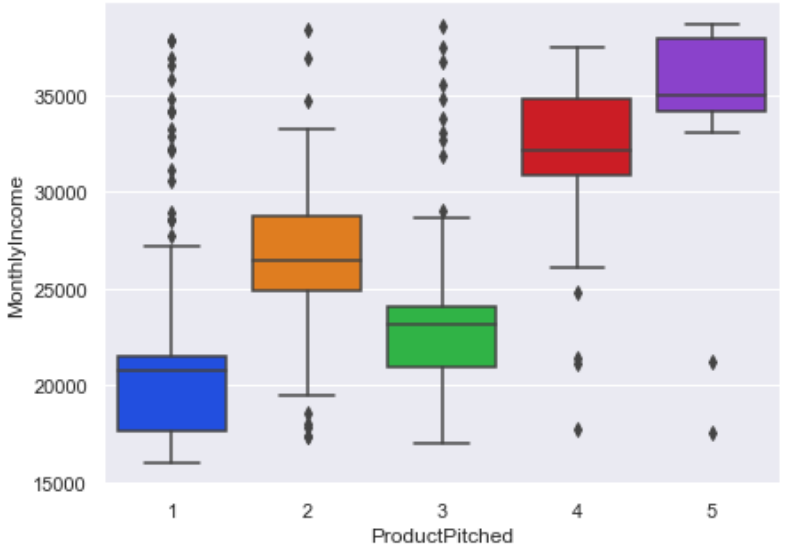
1. Basic
2. Standard
3. Deluxe
4. Super Deluxe
5. King

July 2021

Key Observations

- **ProdTaken is our Target Variable for prediction** and has a 18.8% purchase rate which matches the product brief.
- The most products pitched were Basic (38%) and Deluxe (35%)
- Most ages are between 25 - 48 with 36 being nearly double that of other ages
- 65% of all customers reside in a Tier 3 (highest) city, surprisingly only 4% reside in a Tier 2 city
- The average pitch time is 15 minutes with the majority being less than 17 minutes
- Most contact types are Self Enquiry (71%)
- No customers are planning trips on their own, it is possible this is due to a typo in the data description where NumberOfPersonsVisiting is not in addition to the customer but total persons traveling
- The vast majority of planned trips have > 2 additional persons which likely indicates traveling with family
- Majority of trips per year is 2 & 3
- The basic package has the highest pitch rate with Deluxe a close second.
- The basic package has the highest purchase rate overall with ~ 30% of customers purchasing
- The Deluxe package has about 12% customers purchasing
- The Standard package has about 16% purchase rate with a less than half of the pitches of either basic or deluxe.





EDA Overview

ProductPitched maps perfectly to **Designation**, this was verified in the original data.

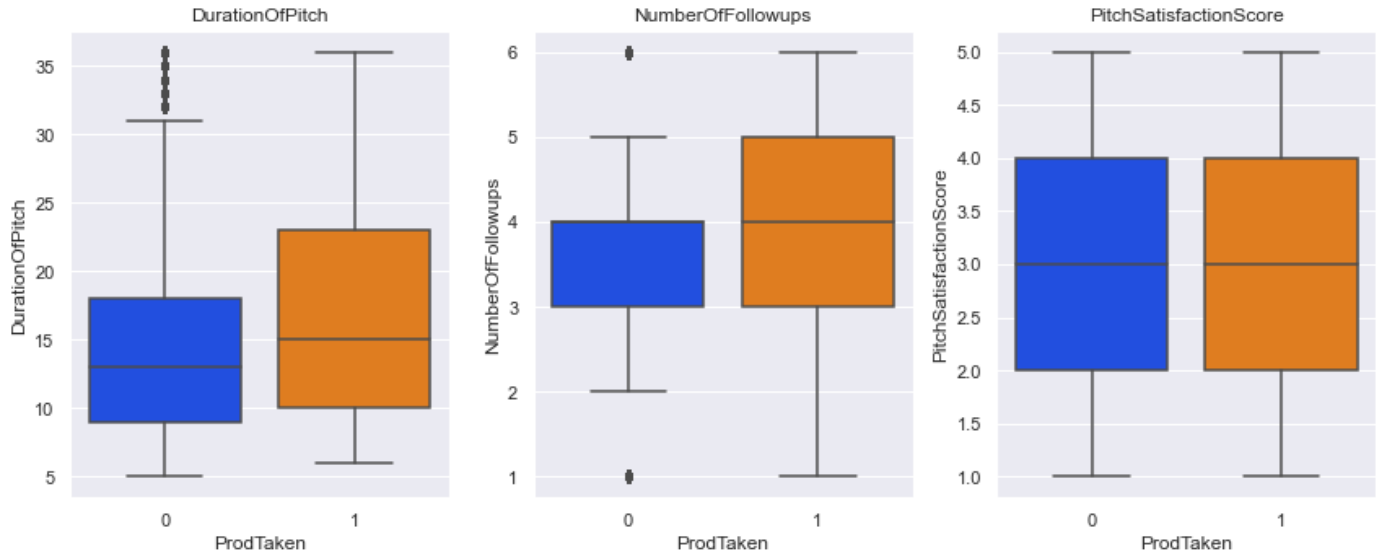
It is possible that this is a data collection error, it is unclear if this is a collection error or simply existing marketing guidelines

The level package pitched does not seem to match the level of designation / income.

The mapping is:
Manager - Deluxe
Senior Manager - Standard
Executive - Basic
AVP - Super Deluxe
VP – King

Key Observations

- Most customers (65%) reside in tier 3 cities with very few residing in tier 2 cities (4%)
- Duration of pitch has the most successes at 19 and 31 at about 37% purchase rate
- Most customers had a 3 satisfaction
- Both 3 & 5 Satisfaction have a purchase rate of ~ 21%
- All other satisfaction scores have about a 16% purchase rate
- Most customers take 2 or 3 trips throughout the year and have about a 19% purchase rate
- Customers who travel more, 7 or 8 times, are more likely to purchase with ~ 27% purchase rate
- Customers between the ages of 29 and 40 are more likely to purchase a product
- Number of Persons or Children Visiting are equally likely to purchase as not
- Salaried and Small Business are roughly equal at 17% for purchasing and make up 91% of all observations
- Self Enquiry has a 17% product purchase rate, only 4% lower than company invited, but accounts for 70% of observations



Model Performance

By developing Machine Learning models that are based on historical data we can predict within a range of certainty who is more likely purchase a travel package and what factors play a key role.

16 models trained with F1 Score as the metric

Conclusion

The best performing model was the **Random Forest Tuned & Weighted** and could predict the likelihood of product purchase 72% of the time.

This model had an ~ 6% lower F1 Test Score than the highest performer, however it is more generalized and likely to perform better when additional data is introduced

- F1 on training set : 0.923647
- **F1 on test set : 0.720755**

- The XGBoost Tuned model performed the best on Test F1 Score (0.785) but is overfitting on the training data, F1 score (0.992)
- Nearly every model is overfitting on the training data.
- Models that are not overfitting have a poor F1 Test Score

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_F1-Score	Test_F1-Score
12	XGBoost Tuned	0.997077	0.919564	1.000000	0.782609	0.984709	0.788321	0.992296	0.785455
14	Stacking - GBM Tuned	0.996492	0.917519	1.000000	0.775362	0.981707	0.783883	0.990769	0.779599
5	Random Forest Tuned	1.000000	0.927744	1.000000	0.663043	1.000000	0.933673	1.000000	0.775424
3	Bagging Classifier Tuned	1.000000	0.926380	1.000000	0.644928	1.000000	0.946809	1.000000	0.767241
15	Stacking - Random Forest Tuned	0.996492	0.906612	0.998447	0.735507	0.983180	0.760300	0.990755	0.747698
2	Bagging Classifier	0.994738	0.916155	0.972050	0.641304	1.000000	0.880597	0.985827	0.742138
11	XGBoost	0.999415	0.913429	0.996894	0.637681	1.000000	0.866995	0.998445	0.734864
6	Random Forest Tuned Weighted	0.969892	0.899114	0.967391	0.692029	0.883688	0.751969	0.923647	0.720755
13	Stacking - XGM Tuned	0.987431	0.877301	1.000000	0.826087	0.937409	0.633333	0.967693	0.716981
4	Random Forest	1.000000	0.907975	1.000000	0.550725	1.000000	0.932515	1.000000	0.692483
0	Decision Tree	1.000000	0.879346	1.000000	0.695652	1.000000	0.673684	1.000000	0.684492
8	Adaboost Tuned	0.973107	0.884117	0.872671	0.572464	0.982517	0.752381	0.924342	0.650206
10	Gradient Boosting Tuned	0.915522	0.873892	0.594720	0.442029	0.931873	0.797386	0.726066	0.568765
9	Gradient Boosting	0.882490	0.867757	0.451863	0.394928	0.855882	0.801471	0.591463	0.529126
1	Decision Tree Tuned	0.821397	0.841172	0.341615	0.369565	0.540541	0.633540	0.418649	0.466819
7	Adaboost	0.849167	0.847307	0.340062	0.333333	0.706452	0.696970	0.459119	0.450980

Data Overview

The data is assumed to be a random subset of customers targeted during a previous marketing campaign.
Data has been provided via Excel file (Tourism.xlsx | 456k)

Customer details:
- **CustomerID**: Unique customer ID
- **ProdTaken (Target Variable)**: Whether the customer has purchased a package or not (0: No, 1: Yes)
- **Age**: Age of customer
- **TypeofContact**: How customer was contacted (Company Invited or Self **Inquiry**)
- CityTier: City tier depends on the development of a city, population, facilities, and living standards. The categories are ordered i.e. Tier 1 > Tier 2 > Tier 3
- **Occupation**: Occupation of customer
- **Gender**: Gender of customer
- **NumberOfPersonVisiting**: Total number of persons planning to take the trip with the customer
- **PreferredPropertyStar**: Preferred hotel property rating by customer
- **MaritalStatus**: Marital status of customer
- **NumberOfTrips**: Average number of trips in a year by customer
- **Passport**: The customer has a passport or not (0: No, 1: Yes)
- **OwnCar**: Whether the customers own a car or not (0: No, 1: Yes)
- **NumberOfChildrenVisiting**: Total number of children with age less than 5 planning to take the trip with the customer
- **Designation**: Designation of the customer in the current organization
- **MonthlyIncome**: Gross monthly income of the customer

Customer interaction data:
- **PitchSatisfactionScore**: Sales pitch satisfaction score
- **ProductPitched**: Product pitched by the salesperson
- **NumberOfFollowups**: Total number of follow-ups has been done by the salesperson after the sales pitch
- **DurationOfPitch**: Duration of the pitch by a salesperson to the customer

The Data Contains:

- 4888 rows and 20 columns
- There are 8 variables with missing values
- There are no Duplicate Rows
- 15 variables require data type changes
  - 8 to category
  - 7 to integer

Data Treatment (pre-model):

- All missing values imputed, no observations deleted
  - Age - median
  - DurationOfPitch - median
  - NumberOfFollowups - median
  - PreferredPropertyStar - median
  - NumberOfTrips - median
  - NumberOfChildrenVisiting - median
  - MonthlyIncome - mean
  - TypeofContact - 'Self Enquiry
- CityTier remapped from 1>2>3 to 3>2>1

It is assumed that CityTier refers to the city that the customer resides in

Note: ProductPitched maps perfectly to Designation, it is unclear if this is a collection error or simply existing marketing guidelines