Holiday Package Purchase Prediction

Project Report

Harshaanth Thiyagaraja Kumar | Geetha Raghiphani | Charles Schlissel

Table of Contents

1. *Project description*

Trips & Travel.Com is on a mission to expand its clientele through innovative package offerings. Presently, their repertoire boasts five distinct packages: Basic, Standard, Deluxe, Super Deluxe, and King. However, despite observing an 18% purchase rate among customers last year, their marketing expenditure soared due to random outreach strategies. To address this, they are gearing up to introduce a Wellness Tourism Package, aimed at promoting healthy lifestyles. Their strategy now revolves around leveraging customer data to streamline marketing efforts.

Our primary goal is to construct a statistical model utilizing supervised machine learning techniques, harnessing the power of Excel's Analytical Solver. This model will predict the likelihood of a customer purchasing a product while uncovering the key determinants influencing their buying behavior.

To achieve this, we've employed the dataset available at: https://www.kaggle.com/datasets/susant4learning/holiday-package-purchase-prediction from Kaggle.

2. *Business questions*

- Predict whether a specific customer will purchase the tour package.
- Determine which types of customers are most and least likely to buy a travel package.
- Identify groups of people for targeted marketing to reduce marketing costs.

3. *Dependent & Independent Variables:*

Target Variable	Numerical - Predictors	Categorical - Predictors
The target variable in this case is **ProdTaken**, which is a binary categorical variable describing whether or not the customer would buy the product	• 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	• Passport • Own Car • Typeof Contact • CityTier • Occupation • Gender • Product Pitched • Marital Status • Designation

Since our overall goal (Target) is to predict whether a customer will purchase the travel package based on various influencing factors (Predictor Variables), we have determined that we need to work with classification models.

4. *Data Preprocessing (refer to milestone 2, please revise according to my comments)*

- Handling missing data

Output Records: 4128

#Records Deleted:760

760 cords had missing values which were deleted.

5. Summary characterizes (mean, median, sd), any outliers?

	Age	DurationOfPitch	NumberOfFollowups	PreferredPropertyStar
count	4128.000	4128.000	4128.000	4128.000
mean	37.232	15.585	3.742	3.578
std	9.175	8.398	1.007	0.795
min	18.000	5.000	1.000	3.000
25%	31.000	9.000	3.000	3.000
50%	36.000	14.000	4.000	3.000
75%	43.000	20.000	4.000	4.000
max	61.000	127.000	6.000	5.000

	NumberOfTrips	PitchSatisfactionScore	NumberOfChildrenVisiting		MonthlyIncome
count	4128.000	4128.000	4128.000	count	4128.000
mean	3.295	3.061	1.224	mean	23178.464
std	1.856	1.363	0.853	std	4506.615
min	1.000	1.000	0.000	min	1000.000
25%	2.000	2.000	1.000	25%	20751.000
50%	3.000	3.000	1.000	50%	22418.000
75%	4.000	4.000	2.000	75%	25301.000
max	22.000	5.000	3.000	max	98678.000

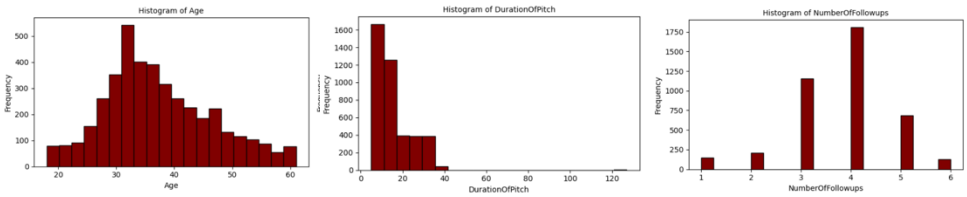
Here we have the summary statistics for our selected numeric variables. We can see that there are outliers in DurationOfPitch, NumberOfFollowups, NumberOfTrips, and MonthlyIncome.

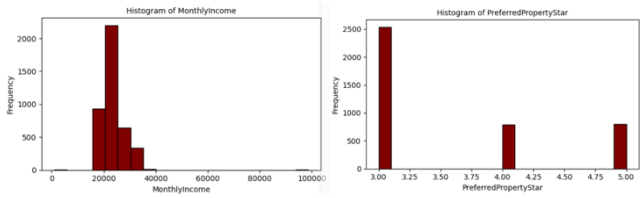
6. Correlation table (and comments/explanations)

	Age	DurationOfPitch	NumberOfPersonVisiting	NumberOfFollowups	rredPropert	NumberOfTrips	PitchSatisfactionScore	NumberOfChildrenVisiting	MonthlyIncome
Age	1								
DurationOfPitch	-0.002731739	1							
NumberOfPersonVisiting	-0.0244903	0.076593494	1						
NumberOfFollowups	-0.025567444	0.016849942	0.324766462	1					
PreferredPropertyStar	-0.030549453	-0.007311571	0.041511011	-0.018088532	1				
NumberOfTrips	0.173654405	0.007089846	0.186989189	0.131478454	0.006867	1			
PitchSatisfactionScore	0.012784533	0.007624861	-0.017630734	0.005452092	-0.019274	-0.005757351	1		
NumberOfChildrenVisiting	-0.030353665	0.039902193	0.597235205	0.282980892	0.041839	0.166315874	0.000391452	1	
MonthlyIncome	0.419912587	0.026268294	0.138245791	0.139315801	0.006013	0.11882474	0.021524978	0.148294103	1

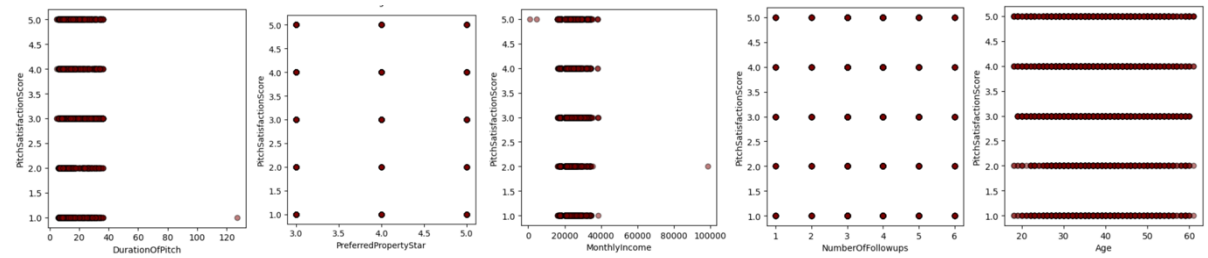
NumberOfChildrenVisiting is highly correlated with NumberOfPersonVisiting
Secondly, MonthlyIncome is highly correlated with Age

7. Histogram, scatterplot, boxplot (dependent variable vs. important independent variables) (and comments/explanations)

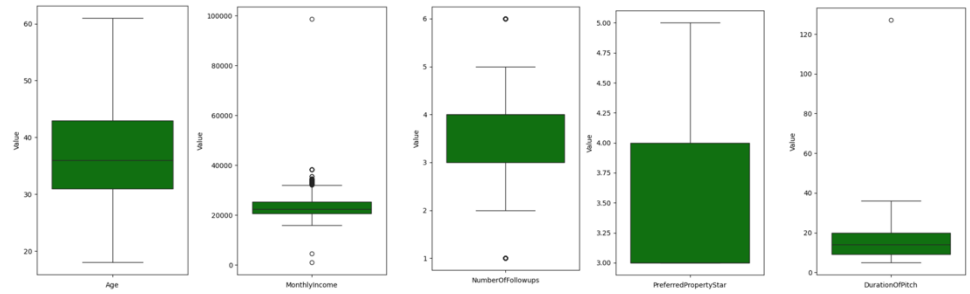




Here we have histograms of the numeric variables so that we can check their distributions to see how that may affect our analysis. It seems that for the most part these variables are not skewed except for the distributions of DurationOfPitch and PreferredPropertyStar, which seem to be skewed hard to the right. There seem to be no real outliers with any of these variables, and some even appear to be more categorical than strictly numeric.



Here we have scatter plots for these same numeric variables as they relate to a second independent variable that is numeric, PitchSatisfactionScore. There seems to be no real trend between it and the rest of our highlighted numeric variables, this may be due to the fact that it is on a set scale and so the values must be from this set scale, altering the results.



Here are the box plots for the numeric variables. From these we can see that some variables here are distributed fairly normally although as we mentioned earlier DurationOfPitch and PreferredPropertyStar are hard right distributed. With these boxplots however we can see more clearly that some of these variables have major outliers. MonthlyIncome, NumberofFollowups, and DurationOfPitch all have some major outliers, with MonthlyIncome having the most cases of outliers.

8. Logistic Regression Model

- The model you use? Why?

Our main focus is on the product taken, a categorical variable. Hence, we employed Logistic Regression, Classification Trees, and Neural Networks to assess and compare model performance in achieving our objective. Subsequently, after conducting Logistic Regression, we selected the final model (Subset 13) due to its optimal Mallow's Cp, which closely matched the number of coefficients, and it exhibited the highest probability among all models evaluated.

Best Subsets Details				
Subset ID	#Coefficients	RSS	Mallows's Cp	Probability
Subset 1	1	3247.695	279.017337	2.38288E-46
Subset 2	2	3137.29	187.3948862	4.63699E-30
Subset 3	3	3102.633	160.0060026	2.99479E-25
Subset 4	4	3080.128	142.9215319	2.82928E-22
Subset 5	5	3051.156	120.3541051	2.56329E-18
Subset 6	6	3029.211	103.7448906	1.99346E-15
Subset 7	7	3007.724	87.52370425	1.2868E-12
Subset 8	8	2980.397	66.35054814	5.33424E-09
Subset 9	9	2955.44	47.18698536	7.53904E-06
Subset 10	10	2944.751	40.1235389	0.000103611
Subset 11	11	2933.163	32.29643413	0.001695462
Subset 12	12	2927.081	29.13946015	0.005509771
Subset 13	13	2921.562	26.4594302	0.015125678

- Which variables are used? Why?

Following feature selection in logistic regression, we have identified Subset 13, comprising the following selected variables: DurationOfPitch, NumberOfFollowups, PreferredPropertyStar, Passport, MonthlyIncome, TypeofContact, CityTier, Occupation_Salaried, Occupation_SmallBusiness, ProductPitched, MaritalStatus_Married, and MaritalStatus_Divorced

- Any variable selection techniques used?

Stepwise Selection: This method involves iteratively adding or removing variables from the model to select the best subset of predictors. In this case, stepwise selection resulted in choosing a subset of 13 variables for the model. (Chosen based on High Probability and least difference between mallow's CP and number of variables).

Significance Based on P-Value: Variables with p-values less than 0.05 are typically considered statistically significant and are included in the model. This approach helps in determining which predictors have a significant association with the outcome variable.

Coefficient Close to Zero: Additionally, a coefficient close to zero (0.0001) for MonthlyIncome

indicates that this variable has minimal impact on the odds of taking the product, even though it may have been included in the model based on the selection techniques mentioned above.

- Report the model output, the equation (if applied), the explanation of coefficients (if applied)

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-0.650209376	-1.679325347	0.378906594	0.521936484	0.525068817	1.533465474	0.215593083
Age	-0.028476419	-0.043105141	-0.013847698	0.971925212	0.007463771	14.55640658	0.000136025
DurationOfPitch	0.037613788	0.023484714	0.051742863	1.03833014	0.007208844	27.22468716	1.81131E-07
NumberOfFollowups	0.426197245	0.299562312	0.552832178	1.531422811	0.064610847	43.51213659	4.21338E-11
PreferredPropertyStar	0.335334674	0.193132289	0.477537059	1.398408318	0.07255357	21.36187383	3.80258E-06
Passport	1.719748818	1.474966882	1.964530753	5.583125907	0.124891038	189.6127304	3.86246E-43
MonthlyIncome	-0.000114108	-0.00014902	-7.91965E-05	0.999885898	1.78125E-05	41.0378226	1.49312E-10
TypeofContact_Self Enquiry	-0.396214571	-0.649395022	-0.143034121	0.672862303	0.129176073	9.407987513	0.002160422
CityTier_3	0.808712679	0.540838671	1.076586687	2.245016069	0.136672924	35.01253317	3.2759E-09
Occupation_Salaried	-0.361525515	-0.755963194	0.032912165	0.696612822	0.201247412	3.227136224	0.072427241
Occupation_Small Business	-0.588170868	-0.991195092	-0.185146644	0.555342151	0.205628382	8.181650083	0.004231625
ProductPitched_Deluxe	-1.218086844	-1.514390588	-0.9217831	0.295795529	0.151178157	64.9199848	7.8002E-16
MaritalStatus_Divorced	-1.042662382	-1.387886724	-0.697438039	0.352514902	0.176138105	35.04134516	3.22778E-09
MaritalStatus_Married	-0.988038877	-1.249145195	-0.726932559	0.372306114	0.133219957	55.00592694	1.20167E-13

Logit (Prodtaken=1)= b0+b1x1+b2x2+b3x3…bqxq

Logit (Prodtaken=1)= -0.65-0.028Age+0.038DurationOfPitch+0.426NumberOfFollowups+0.335PreferredPropertyStar+1.720Passport-0.0001MonthlyIncome-0.396TypeofContact_SelfEnquiry+0.809CityTier_3-0.362Occupation_Salaried-0.588Occupations_SmallBusiness-1.219ProductPitched_Deluxe-1.043MartialStatus_Divorced-0.999MartialStatus_Married

Odds(Prodtaken=1)=e^(b0+b1x1+b2x2+b3x3…bqxq)

Odds(Prodtaken=1)=e^(-0.65-0.028Age+0.038DurationOfPitch+0.426NumberOfFollowups+0.335PreferredPropertyStar+1.720Passport-0.0001MonthlyIncome-0.396TypeofContact_SelfEnquiry+0.809CityTier_3-0.362Occupation_Salaried-0.588Occupations_SmallBusiness-1.219ProductPitched_Deluxe-1.043MartialStatus_Divorced-0.999MartialStatus_Married)

Odds(Prodtaken=1)=Odds0*(Odds1)^x1*(Odds2)^x2*(Odds3)^x3*……*(Oddsq)^xq

Odds(Prodtaken=1)=0.522*(0.972)^Age*(1.038)^DurationOfPitch*(1.531)^NoOfFollowups*(1.398)^PrefferedPropertyStar*(5.583)^Passport*(1)^MonthlyIncome*(0.673)^TypeofContact_SelfEnquiry*(2.245)^CityTier_3*(0.7)^Occupation_Salaried*(0.55)^Occupation_SmallBusiness*(0.295)^ProductPitched_Delux*(0.352)^MaritialStatus_Divorced*(0.372)^MaritialStatus_Married

P(Prodtaken=1)=1/(1+ e^-(b0+b1x1+b2x2+b3x3…bqxq)
P(Prodtaken=1)=1/(1+ e^-(-0.65-0.028Age+0.038DurationOfPitch+0.426NumberOfFollowups+0.335PreferredPropertyStar+1.720Passport-0.0001MonthlyIncome-0.396TypeofContact_SelfEnquiry+0.809CityTier_3-0.362Occupation_Salaried-0.588Occupations_SmallBusiness-1.219ProductPitched_Deluxe-1.043MartialStatus_Divorced-0.999MartialStatus_Married))

Age: For each additional unit increase in age, the odds of taking the product decrease by a factor of approximately 0.972, holding other variables constant.

DurationOfPitch: With each unit increase in the duration of the pitch, the odds of taking the product increase by about 1.038 times, keeping other factors constant.

NumberOfFollowups: For each additional follow-up, the odds of taking the product increase by approximately 1.531 times, holding other variables constant.

PreferredPropertyStar: Customers preferring higher star-rated properties are more likely to take the product. For each unit increase in preferred property star rating, the odds of taking the product increase by around 1.398 times, with other variables held constant.

Passport: Customers with passports are significantly more likely to take the product. Having a passport increases the odds of taking the product by approximately 5.583 times, holding other variables constant.

MonthlyIncome: The coefficient for monthly income is very close to zero (0.0001), indicating that monthly income has minimal impact on the odds of taking the product.

TypeofContact_SelfEnquiry: Customers who initiate contact themselves are less likely to take the product compared to those contacted by the company. The odds of taking the product decrease by about 0.673 times for self-enquiry contacts, holding other variables constant.

CityTier_3: Customers from City Tier 3 are more likely to take the product compared to other tiers. Being from City Tier 3 increases the odds of taking the product by approximately 2.245 times, holding other variables constant.

Occupation_Salaried: Salaried individuals are less likely to take the product compared to others. The odds of taking the product decrease by around 0.7 times for individuals with a salaried occupation, holding other variables constant.

Occupations_SmallBusiness: Individuals engaged in small businesses are even less likely to take the product. The odds decrease by about 0.55 times for individuals with a small business occupation, holding other variables constant.

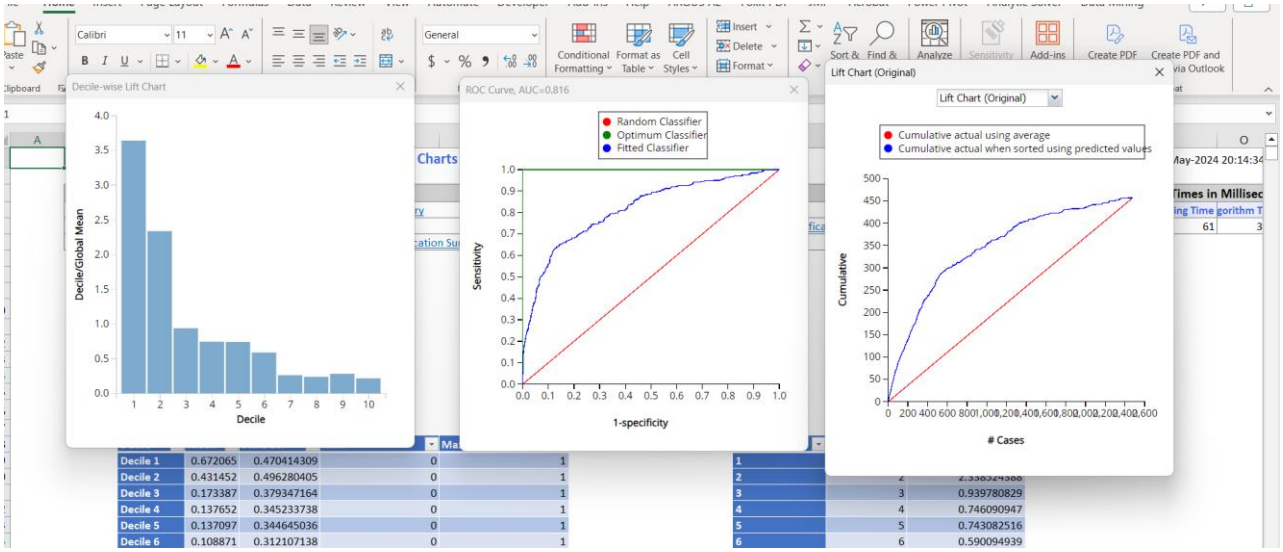
ProductPitched_Deluxe: Pitching the deluxe product decreases the odds of taking the product significantly. The odds decrease by approximately 0.295 times when the deluxe product is pitched, holding other variables constant.

MartialStatus_Divorced: Divorced individuals are less likely to take the product compared to others. The odds decrease by about 0.352 times for individuals who are divorced, holding other variables constant.

MartialStatus_Married: Married individuals are also less likely to take the product compared to others. The odds decrease by about 0.372 times for individuals who are married, holding other variables constant.

Report Training and validation (and test) data summary report and lift charts. How's the model work?

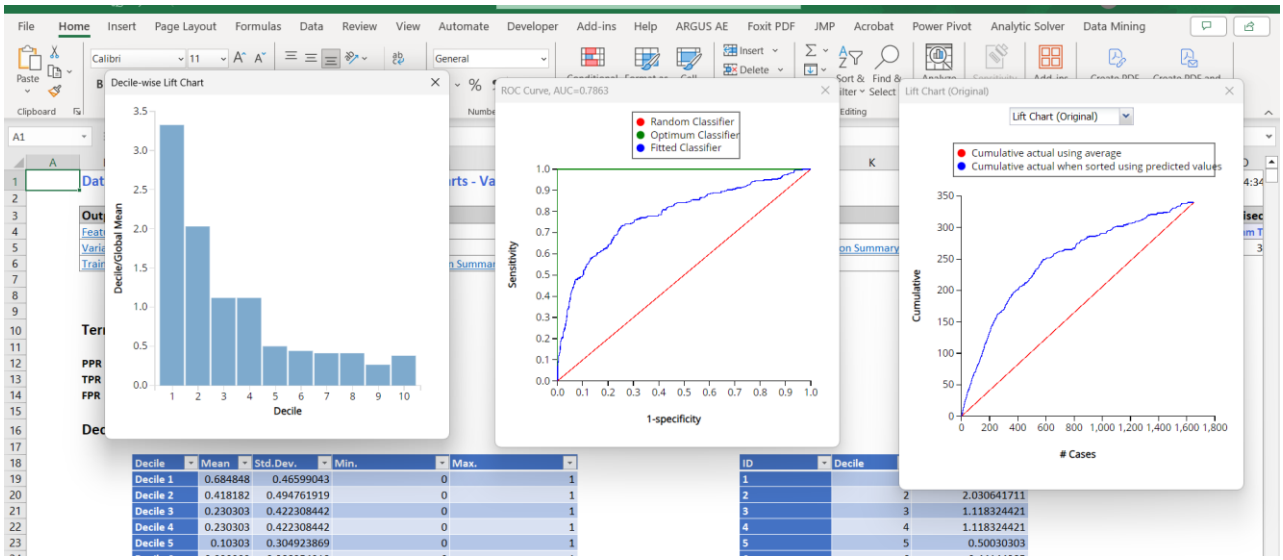
Training:



Decile 1	0.672065	0.470414309	0	1
Decile 2	0.431452	0.496280405	0	1
Decile 3	0.173387	0.379347164	0	1
Decile 4	0.137652	0.345233738	0	1
Decile 5	0.137097	0.344645036	0	1
Decile 6	0.108871	0.312107138	0	1

1		
2	2	2.338324388
3	3	0.939780829
4	4	0.746090947
5	5	0.743082516
6	6	0.590094939

Validation:



Decile	Mean	Std.Dev.	Min.	Max.
Decile 1	0.684848	0.46599043	0	1
Decile 2	0.418182	0.494761919	0	1
Decile 3	0.230303	0.422308442	0	1
Decile 4	0.230303	0.422308442	0	1
Decile 5	0.10303	0.304923869	0	1

ID	Decile	
1		
2	2	2.030641711
3	3	1.118324421
4	4	1.118324421
5	5	0.50030303

This model works well; the AUC is larger than 0.5, indicating that the model performs better than the benchmark. Furthermore, the performances are quite similar in both the training and validation datasets, suggesting that it will perform well on new datasets too.

The training AUC value of 0.816 and the validation AUC of 0.7863 are closely aligned, indicating consistency in performance between the training and validation datasets. This similarity suggests no apparent overfitting issues, and that the model exhibits good predictive capability and generalizability to new data.

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	909	402
1	82	258

Error Report

Class	# Cases	# Errors	% Error
0	1311	402	30.66361556
1	340	82	24.11764706
Overall	1651	484	29.31556632

Metrics

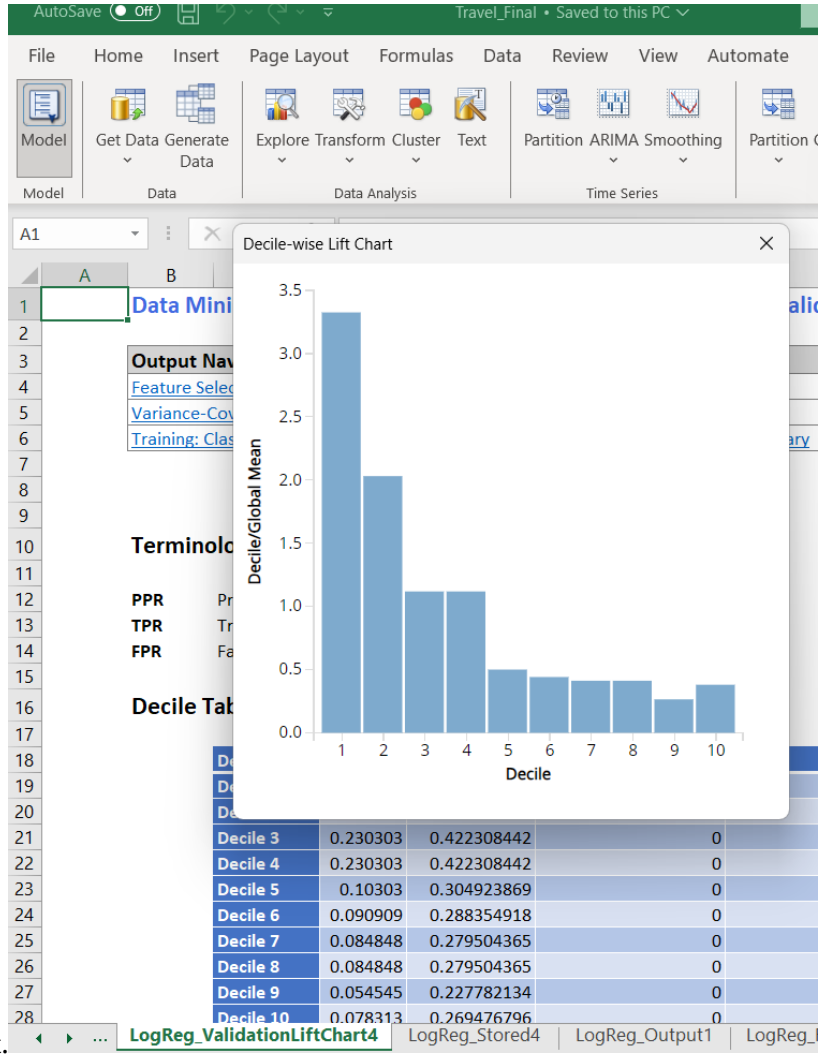
Metric	Value
Accuracy (#correct)	1167
Accuracy (%correct)	70.68443368
Specificity	0.693363844
Sensitivity (Recall)	0.758823529
Precision	0.390909091
F1 score	0.516
Success Class	1
Success Probability	0.1506

The Overall Accuracy of the model is 70.68%. The Error rate of Class 0 is 30.66%, The Error rate of Class 1 is 24.12% and the overall error rate is 29.32%.

The Specificity is 0.693, which means that out of all the persons who didn't the product 69.30% were correctly identified by the model. The Sensitivity (Recall) is 0.7588, which means that out of all the persons who bought the product 75.88% were correctly identified by the model.

- If answering a classification question, based on your results, how do you choose the cutoff value?

In Logistic Regression, due to a notably low recall rate when using a 0.5 default cutoff probability for the success class, we opted to adjust it to 0.1506. Based on the decile chart. We decided by arranging the post-probability 1 values in descending order and selecting the probability associated with the 660th record from



the entire dataset.

Cases

ID	Decile	Decile/Global Mean
1	1	3.325543672
2	2	2.030641711
3	3	1.118324421
4	4	1.118324421
5	5	0.50030303
6	6	0.44144385
7	7	0.41201426
8	8	0.41201426
9	9	0.26486631
10	10	0.380279943

693	Record 43	0	0	0.152097153	0.847902847
694	Record 1193	0	0	0.152097153	0.847902847
695	Record 226	0	0	0.150905181	0.849094819
696	Record 3580	0	0	0.150615966	0.849384034
697	Record 2948	0	0	0.150277761	0.849722239
698	Record 2658	0	0	0.150235887	0.849764113

- How do the results answer your questions?

Most likely to purchase the product:

Customers who are younger, have been pitched the product for a longer duration, have been followed up with more times, prefer higher star-rated properties, have a passport, and are from City Tier 3, are more likely to take the product.

Least likely to purchase the product:

Customers who are older, don't have a passport, who self-enquired, salaried individuals, individuals with small businesses, those who were pitched the deluxe product, and those in a divorced marital status are less likely to take the product. Married individuals, as being married also decreases the odds of taking the product.

- Make an example of one new record and make prediction/classification

If a person is 27 years old, type of contact is self-inquiry, city tier is 2, duration of pitch is 10, occupation is salaried, gender is male, number of person visiting is 3, number of follow-ups is 3, product pitched is basic, preferred property star is 4, marital status is single, number of trips is 2, passport is 0, pitch satisfaction is 5, owns a car is 1, number of children visiting is 0, designation is manager, and monthly income is 30,000. Will the person buy the product?

Odds(Prodtaken=1)=e^(b0+b1x1+b2x2+b3x3…bqxq)

Odds(Prodtaken=1)=Odds0*(Odds1)^x1*(Odds2)^x2*(Odds3)^x3*……*(Oddsq)^xq

Odds(Prodtaken=1)=0.522*(0.972)^Age*(1.038)^DurationOfPitch*(1.531)^NoOfFollowups*(1.398)^PreferedPropertyStar*(5.583)^Passport*(1)^MonthlyIncome*(0.673)^TypeofContact_SelfEnquiry*(2.245)^CityTier_3*(0.7)^Occupation_Salaried*(0.55)^Occupation_SmallBusiness*(0.295)^ProductPitched_Delux*(0.352)^MaritialStatus_Divorced*(0.372)^MaritialStatus_Married

Odds(Prodtaken=1)=0.522*(0.972)^27*(1.038)^10*(1.531)^3*(1.398)^4*(5.583)^0*(1)^30,000*(0.673)^0*(2.245)^0*(0.7)^1*(0.55)^0*(0.295)^0*(0.352)^0*(0.372)^0

Odds(Prodtaken=1)= 2.675

P = Odds/1+Odds = 2.675/1+2.675 = 0.727

Since predicted Probability (0.727) > Success class cutoff probability (0.1506), Classification 1

So, the probability of the person taking the product is approximately 0.727. Since this probability is greater than the cutoff probability of 0.1506, the person is predicted to buy the product.

9. *Classification Tree Model*

- The model you use? Why?

We used a classification tree to further evaluate and check which model suits this dataset. We Fit a classification tree using all predictors by splitting the data into training, validation and test datasets using a 50%, 30%, and 20% ratio. To avoid overfitting, set the minimum number of records in a leaf node to 100. Also, we set the maximum number of levels to be displayed at 7, Displayed both the full tree and the best-pruned tree.

- Which variables are used? Why?

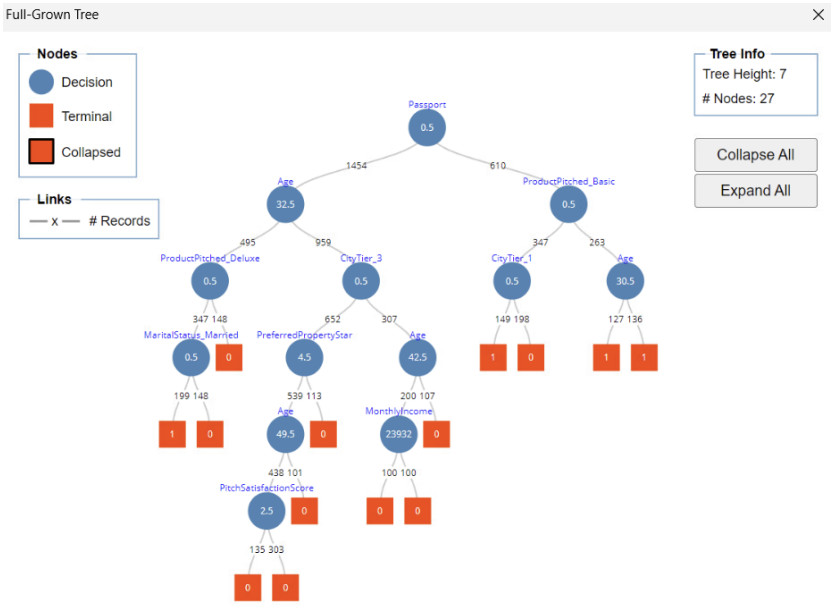
  Passport, Age, ProductPitched_Basic, productPitched_Deluxe.

- Any variable selection techniques used?

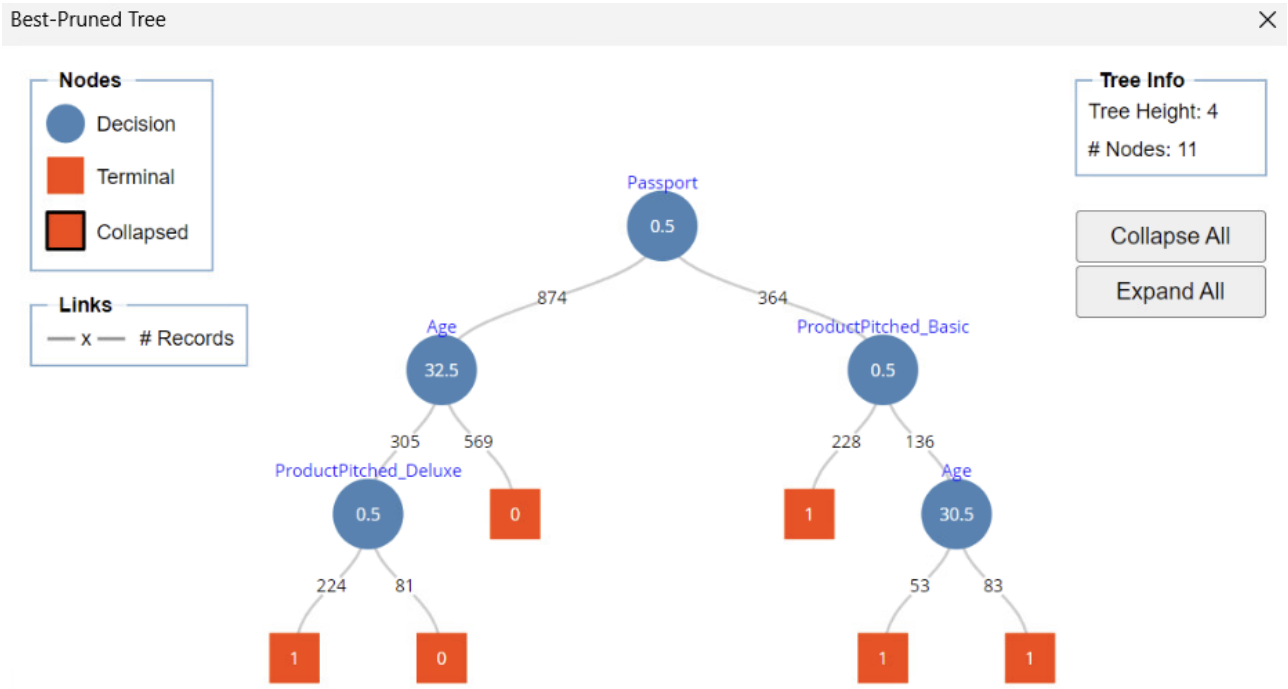
  We selected all the dependent variables and target variable to run the tree.

- Report the model output, the equation (if applied), the explanation of coefficients (if applied)

Full tree:



Best Pruned Tree:



So, we have 6 rules as we have 6 leaf nodes in the best-pruned tree:

If Passport < 0.5 and Age >= 32.5, class = 0

This rule states that if a person does not have a passport and their age is greater than or equal to 32.5, they belong to class 0.

If Passport < 0.5 and Age < 32.5 and productpitched_deluxe >= 0.5, class = 0

This rule applies if a person does not have a passport, their age is less than 32.5, and the product pitched is deluxe. In this case, the class is 0.

If Passport < 0.5 and Age < 32.5 and productpitched_deluxe < 0.5, class = 1

Here, if a person lacks a passport, their age is less than 32.5, and the product pitched is not deluxe, the class assigned is 1.

If Passport >= 0.5 and productpitched_basic >= 0.5 and Age >= 30.5, class = 1

This rule applies when a person has a passport, the product pitched is basic, and their age is 30.5 or older. In this case, the class is 1.

If Passport >= 0.5 and productpitched_basic >= 0.5 and Age < 30.5, class = 1

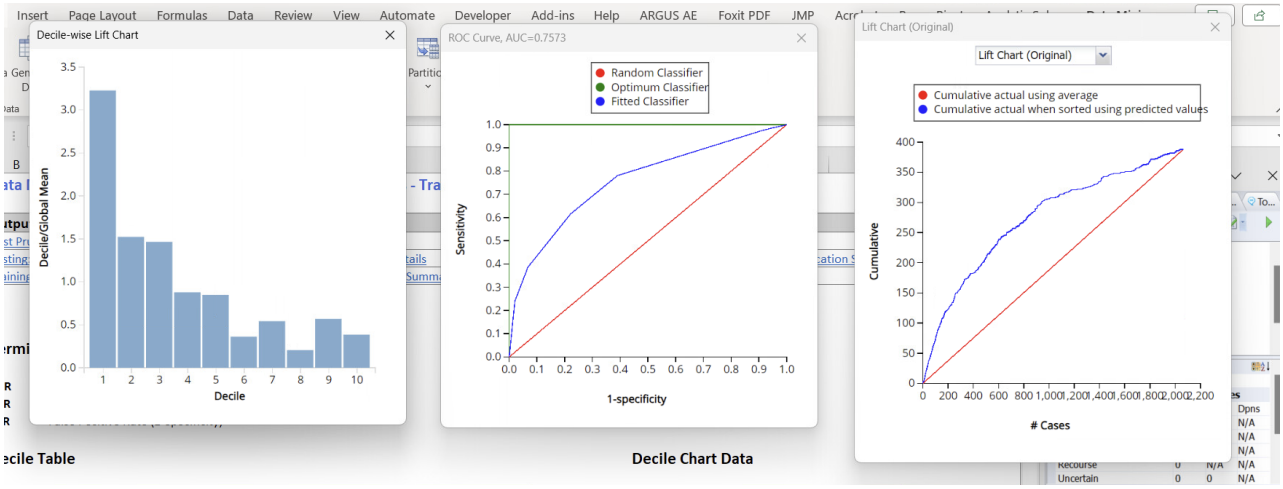
Similar to the previous rule, but here the age condition is less than 30.5.

If Passport >= 0.5 and productpitched_basic < 0.5, class = 1

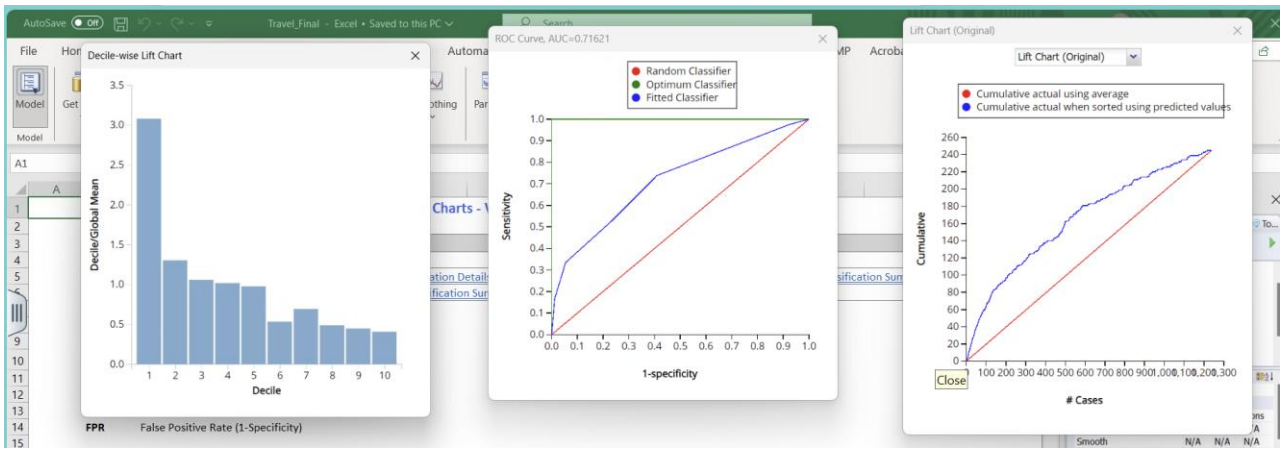
This rule applies if a person has a passport and the product pitched is not basic, assigning them to class 1.

Report Training and validation (and test) data summary report and lift charts. How's the model work?

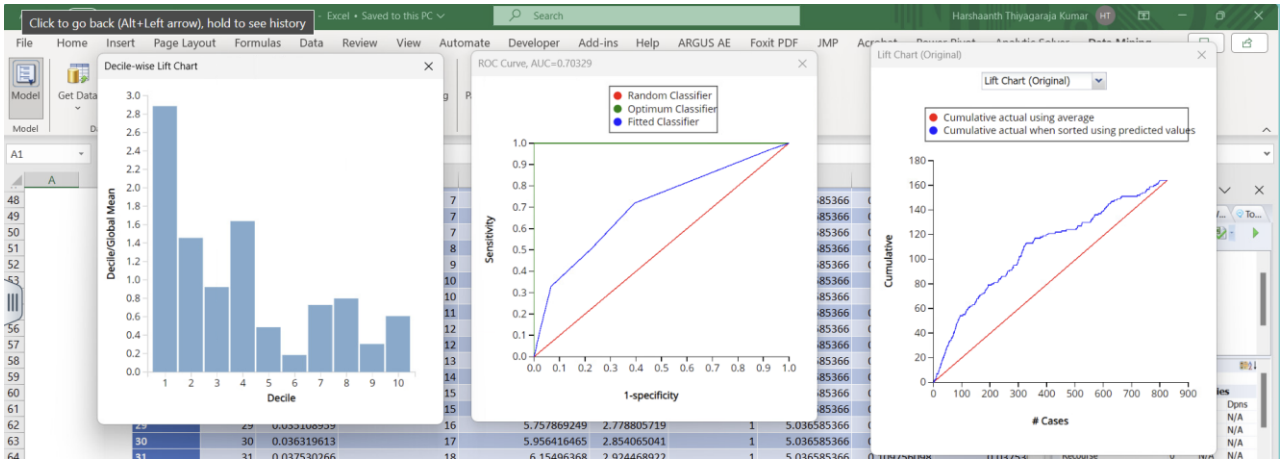
Training Lift Chart:



Validation Lift Chart



Test Lift Chart:



Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	401	261
1	46	118

Error Report			
Class	# Cases	# Errors	% Error
0	662	261	39.42598187
1	164	46	28.04878049
Overall	826	307	37.16707022

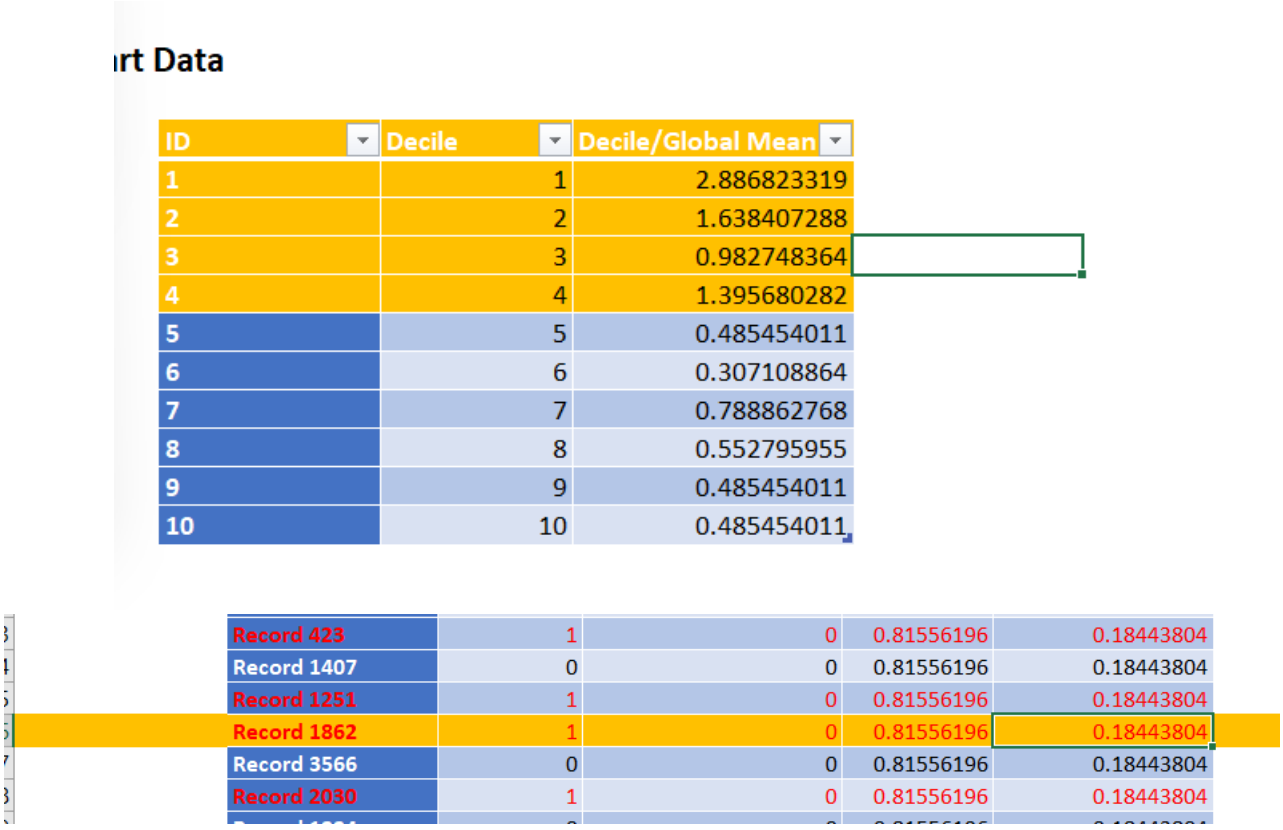
Metrics	
Metric	Value
Accuracy (#correct)	519
Accuracy (%correct)	62.83292978
Specificity	0.605740181
Sensitivity (Recall)	0.719512195
Precision	0.311345646
F1 score	0.434622468
Success Class	1
Success Probability	0.1844

The Overall Accuracy of the model is 62.83%. The Error rate of Class 0 is 39.43%, The Error rate of Class 1 is 28% and the overall error rate is 37.17%.

The Specificity is 0.606, which means that out of all the persons who didn't the product 60.60% were correctly identified by the model. The Sensitivity (Recall) is 0.72, which means that out of all the persons who bought the product 72% were correctly identified by the model.

- If answering a classification question, based on your results, how do you choose the cutoff value?

In the classification tree, due to a notably low recall rate when using a 0.5 probability threshold for the success class, we opted to adjust it to 0.1844. This decision was made by arranging the post-probability 1 values in descending order and selecting the probability associated with the 330th record from the entire dataset, as indicated by the decile chart.



ırt Data

ID	Decile	Decile/Global Mean
1	1	2.886823319
2	2	1.638407288
3	3	0.982748364
4	4	1.395680282
5	5	0.485454011
6	6	0.307108864
7	7	0.788862768
8	8	0.552795955
9	9	0.485454011
10	10	0.485454011

Record 423	1	0	0.81556196	0.18443804
Record 1407	0	0	0.81556196	0.18443804
Record 1251	1	0	0.81556196	0.18443804
Record 1862	1	0	0.81556196	0.18443804
Record 3566	0	0	0.81556196	0.18443804
Record 2030	1	0	0.81556196	0.18443804

- How do the results answer your questions?

A person who possesses a passport will always opt for the holiday package. Conversely, an individual without a passport who is older than 32.5 years will not choose the holiday package. Furthermore, someone without a passport and younger than 32.5 years, who purchases a deluxe product pitched pack, will be classified as class 0. Finally, an individual without a passport, aged less than 32.5 years, and opting for a non-deluxe product will always purchase the holiday package.

- Make an example of one new record and make prediction/classification

If a person is 27 years old, type of contact is self-inquiry, city tier is 2, duration of pitch is 10, occupation is salaried, gender is male, number of person visiting is 3, number of follow-ups is 3, product pitched is basic, preferred property star is 4, marital status is single, number of trips is 2, passport is 0, pitch satisfaction is 5, owns a car is 1, number of children visiting is 0, designation is manager, and monthly income is 30,000. Will the person buy the product?

If Passport < 0.5 and Age < 32.5 and productpitched_deluxe < 0.5, class = 1

Classification: 1, The person will buy the product.

10. Neural Network Model

- The model you use? Why?

Since we were attempting to conduct a classification analysis we decided to also go with a neural network for our data. We used a neural network because they often work well with classification techniques and can describe relationships in the data other models can't explain. We used 60/40 training and validation data partition and standardized variables for this network. We also used a couple different versions of the neural network. We rand a few different versions of a neural network with 1 hidden layer and 5 nodes, 2 hidden layers each with 5 nodes, and 2 hidden layers with 10 nodes each. We then ran an automatic neural network to find a good structure. And after failing to find a network that really stood out we then ran a neural network with 1 hidden layer and 2 nodes, and after selecting a new cutoff value of .3099 this model seemed to work, although not well, and thus we chose this model to be our neural network. The model has very low precision, recall and F1 scores, but has a solid accuracy score, AUC score, and lift charts. We chose this because these metrics were much better than the low or lack of metric scores that were returned with the other attempts of building the model.

- Which variables are used? Why?

The variables used in this neural network are all of the variables in the dataset. We did this because we were having massive issues with getting any sort of neural network to run properly so we tried many different combinations of variables. Each yielded different results but the resounding conclusion we drew was that none of the selections we chose were helping the model run properly. So we tried it one more time with all the variables and after selecting a new cutoff value we found that we could get the model to work with these variables and thus they became the variables we used for this neural network.

- Any variable selection techniques used?

For this neural network we tried many different selections of variables. We used the variables from the feature selection with the logistic regression, the variables from the best pruned tree and a selection of all the variables in the dataset. Most of the models that used these selection techniques failed to work properly, so ultimately, we ended up going with all the variables in the dataset and not using any variable selection techniques.

- Report the model output, the equation (if applied), the explanation of coefficients (if applied)

Neuron Weights: Input Layer - Hidden Layer 1

Neurons	Age	DurationOfPitch	NumberOfPersonVisiting	NumberOfFollowups	PreferredPropertyStar	NumberOfTrips	Passport	PitchSatisfactionScore	OwnCar	NumberOfChildrenVisiting
Neuron 1	0.0186916	0.055987505	-0.378955593	0.099726896	0.0966965	0.125883317	-0.193632	0.134396444	-0.04984	0.221827586
Neuron 2	-0.155052	-0.167200274	-0.135113359	0.097770539	-0.227299418	-0.137744983	0.158157	-0.061418544	-0.07517	-0.320512322

OwnCar	NumberOfCh	MonthlyIncome	TypeofContact_Company Invi	TypeofContact_Self Enqu	CityTier_1	CityTier_2	CityTier_3	Occupation_Free Lancer	Occupation	Occupation_Salaried
-0.049837	0.2218276	-0.110597011	-0.105492114	0.032194418	-0.002975951	-0.167934092	-0.075058	-0.18970296	0.148129	-0.018142538
-0.075173	-0.320512	-0.013618273	-0.073525795	-0.311612608	0.304153521	0.002072495	0.333726	0.081848761	0.110992	-0.038124922

Occupation	Gender_Fe M	Gender_Female	Gender_Male	ProductPitched_Basic	ProductPitched_Deluxe	ProductPitched_Ki	ProductPitc	ProductPitched_Super De	MaritalStat	MaritalStatus_Married
-0.091304	0.0263483	0.047919348	-0.148977698	-0.03123372	0.088232516	-0.128748391	-0.044276	0.094053597	0.291752	0.146881402
-0.44272	-0.406146	0.259265483	0.06603951	0.180367022	-0.157301833	0.11722195	0.032888	-0.143656247	-0.08573	-0.207662271

`	MaritalStatu	Designation_AVP	Designation_Executive	Designation_Manager	Designation_Senior Mana	Designation_VP	Bias
-0.27378	-0.002282	0.32236914	0.025527528	0.029722894	0.052219193	0.019141858	0.137696
0.054282	-0.037705	-0.197744055	0.192385105	0.081172788	0.032037737	-0.154147876	0.011001

Neuron Weights: Hidden Layer 1 - Output Layer

Neurons	Neuron 1	Neuron 2	Bias
0	0.1415662	0.629655552	0.526047324
1	-1.819243	0.489555847	-0.375608327

These are the neuron outputs for the neural network, there are a lot of inputs into our hidden layer with 2 nodes as we could only get the model working by using all the variables in the dataset. Each of these shows the weight on the neuron from each input variable and also the bias values for those neurons. Then we can see in the output layer the weights of the neuron inputs and the bias values of the output layer.

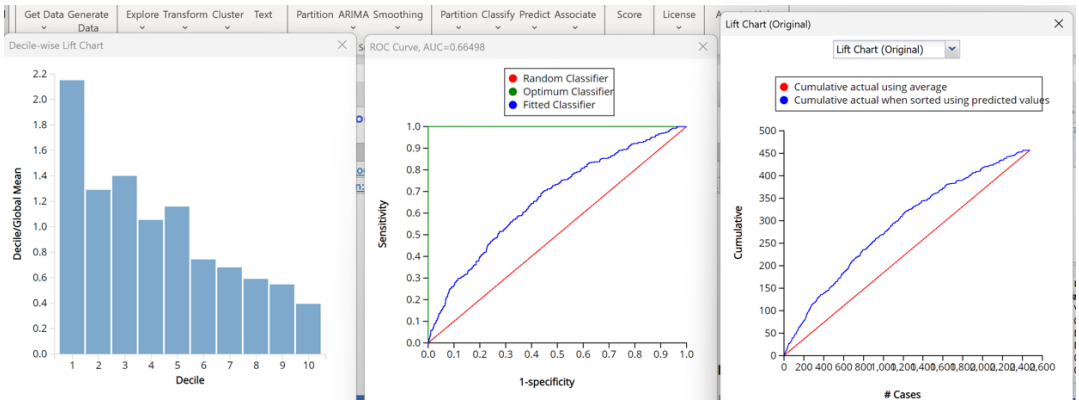
- Report Training and validation (and test) data summary report and lift charts. How's the model work?

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	1710	310
1	307	150

Error Report			
Class	# Cases	# Errors	% Error
0	2020	310	15.34653465
1	457	307	67.17724289
Overall	2477	617	24.90916431

Metrics	
Metric	Value
Accuracy (#correct)	1860
Accuracy (%correct)	75.0908357
Specificity	0.84653465
Sensitivity (Recall)	0.32822757
Precision	0.32608696
F1 score	0.32715376
Success Class	1
Success Probability	0.3099

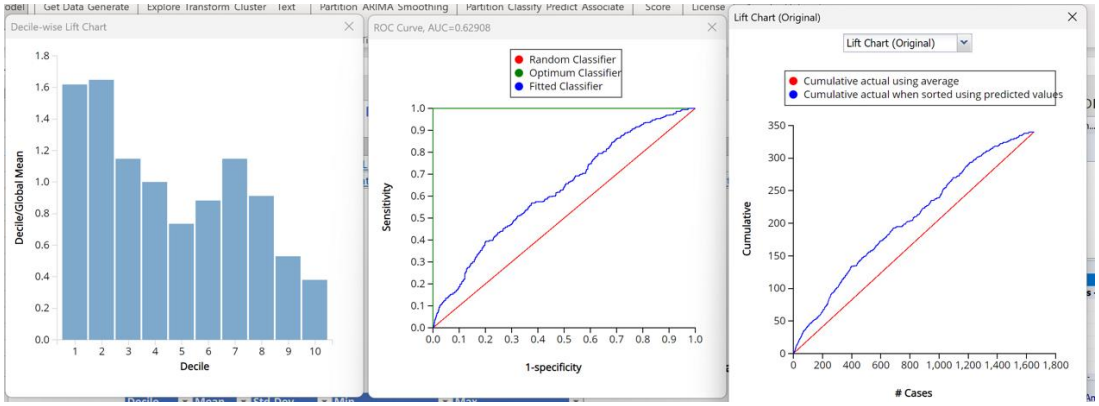


Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	1113	198
1	238	102

Error Report			
Class	# Cases	# Errors	% Error
0	1311	198	15.10297483
1	340	238	70
Overall	1651	436	26.40823743

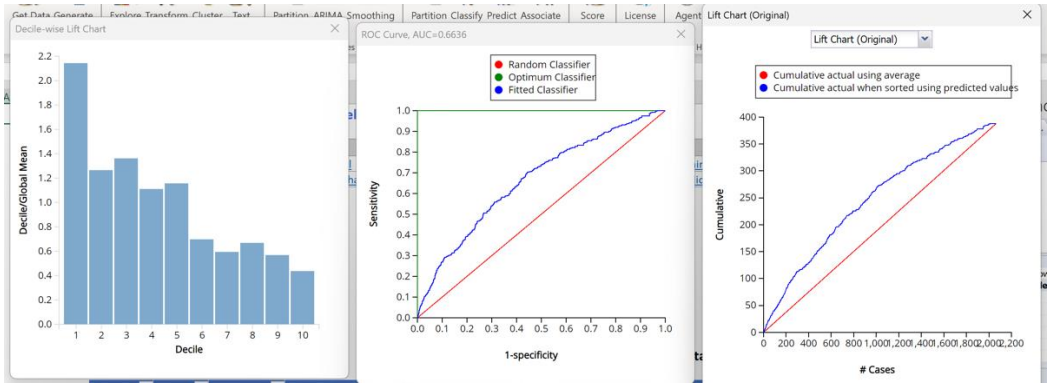
Metrics	
Metric	Value
Accuracy (#correct)	1215
Accuracy (%correct)	73.5917626
Specificity	0.84897025
Sensitivity (Recall)	0.3
Precision	0.34
F1 score	0.31875
Success Class	1
Success Probability	0.3099



This model works but not very well. We wanted to get a model that would not return errors in the major scoring metrics so when we chose a new cutoff value and it worked, we chose this model. Further testing revealed that other models using different cutoffs were still not yielding proper results, so this is the model we went with. The metrics for it are in some cases decent and in other cases really not great. We had a decent ROC curve and AUC score for the model, as well as a decent lift chart which showed preliminarily that our model worked okay. Further examination of the scoring metrics showed that we had good accuracy and specificity scores at 73% and .84 respectively, but seemingly awful precision, recall and F1 scores. These last three are very telling for the overall performance of the model and their values of .3, .34, and .319 respectively show that the model really does not work well at all.

- If answering a classification question, based on your results, how do you choose the cutoff value?

In our analysis preliminary runs of neural networks proved unsuccessful as most network structures we tried yielded awful or errored results. However, our last attempt, a network using all the variables in the dataset and having 1 hidden layer with 2 nodes, finally yielded a decent result for us to work with. This model still did not return any good results in terms of scoring metrics but from it we could chose a new cutoff value to try and run a new model with. Based on the decile chart output featured below, we could see that there is a huge drop off around decile 2, so using the total number of cases 2064 and multiplying that by .2 for the second decile, we arrived at our cutoff, using the probability for the 413th value, being around .3099.



Record 2381	1	0	0.6899882	0.3100118
Record 1760	0	0	0.690058895	0.309941105
Record 2100	0	0	0.690070467	0.309929533

Error Report			
Class	# Cases	# Errors	% Error
0	1676	0	0
1	388	388	100
Overall	2064	388	18.79844961

Metrics	
Metric	Value
Accuracy (#correct)	1676

412.8 =.2*2064
new cutoff = .3099

- How do the results answer your questions?

The results of this model do not help much when answering our questions as it does not do the best job at accurately classifying who is most likely to purchase a travel package. If we want to answer the question of which customers will most want to buy our travel package, this model will not accurately give us a set of specifications for the way we should market that. What it may tell us is that predicting which types of customers will buy a travel package is not so complex that we need a deep learning model to effectively describe which types of people will buy our package and which types of people will not.

11. Comparison between your models.

The training AUC of 0.816 and validation AUC of 0.7863 closely align, indicating consistent performance across logistic regression, outperforming classification trees, and neural networks. Overall, the model achieves 70.68% accuracy, with a Class 0 error rate of 30.66%, a Class 1 error rate of 24.12%, and an overall error rate of 29.32%. The Specificity is 0.693, which means that out of all the persons who didn't have the product 69.30% were correctly identified by the model. The Sensitivity (Recall) is 0.7588, which means that out of all the persons who bought the product 75.88% were correctly identified by the model. These metrics allow us to make informed decisions on marketing investments, targeting specific customer groups effectively and efficiently.

12. The conclusions

Based on model comparison, we have decided that the logistic regression model is the best choice for our analysis. It demonstrates superior performance metrics compared to the other models tested, selects the most relevant variables for analysis, and overall, it is the optimal model for predicting whether any given customer will purchase a travel package from us.

Based on the logistic regression analysis:

a) Predict whether a specific customer will purchase the tour package: Yes, you can predict whether a specific customer will purchase the tour package by calculating the probability of them taking the product using the logistic regression equation provided.
b) Determine which types of customers are most and least likely to buy a travel package:

- Most likely to buy:
    - Younger customers
    - Customers who have been pitched the product for a longer duration
    - Customers who have been followed up with more times
    - Customers who prefer higher star-rated properties
    - Customers with a passport
    - Customers from City Tier 3
- Least likely to buy:
    - Older customers
    - Customers who don't have a passport
    - Customers who self-enquired
    - Salaried individuals
    - Individuals with small businesses
    - Those who were pitched the deluxe product
    - Those in a divorced marital status
    - Married individuals

c) Identify groups of people for targeted marketing to reduce marketing costs: Based on the identified characteristics of customers who are most and least likely to buy the travel package, you can tailor your marketing efforts accordingly. For example:

Target younger customers, especially those from City Tier 3, who prefer higher star-rated properties, and have passports, with focused marketing campaigns. Avoid investing heavily in marketing towards older individuals, those without passports, self-enquiring customers, salaried individuals, individuals with small businesses, those pitched the deluxe product, and those in divorced or married marital statuses, as they are less likely to purchase the product.

13. Summary

In this analysis, we learned a few valuable lessons. We learned how to apply the models we've studied in class to real, actionable models that will help determine who is most likely to buy a travel package. We learned how to use metric scores to determine which of those models works best for the objectives we are looking to accomplish. We learned how to change models such that if they don't work in one iteration they can then perform reasonably well in another. When working with this data we also have found a few things we would recommend the authors of the dataset change. They may want to add a few more columns to the dataset to expand their understanding of what drives people to buy travel packages. Perhaps some information about the types of rental properties available or specific details of the vacation plans offered would be helpful.

Please also attach a table of how you split the tasks among group members.

Harshaanth Thiyagaraja Kumar	Logistic Regression	Data description, Business Questions, Correlation Analysis, Model comparison, Conclusion
Geetha Raghiphani	Classification Tree	Data description, Business Questions, Correlation Analysis, Model comparison, Conclusion
Charles Charles Schlissel	Neural Networks	Data Visualization, Summary statistics, Summary