# Holiday Package Purchase Prediction

Our objective is to develop a statistical model using supervised machine learning, employing Excel's Analytical Solver to predict the likelihood of a person purchasing a product, while also identifying the factors influencing their buying decision.

Group 14 | Charlie | Harsha | Geetha

# Business Description

Trips & Travel.Com wants to grow its customer base by offering new packages. Currently, they have five types of packages - Basic, Standard, Deluxe, Super Deluxe, and King. They observed that 18% of customers bought these packages last year, but marketing costs were high because they reached out randomly. Now, they plan to launch a Wellness Tourism Package, focusing on promoting healthy lifestyles. They aim to use customer data to make marketing more efficient.

Dataset: https://www.kaggle.com/datasets/susant4learning/holiday-package-purchase-prediction

# Business Questions

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

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.

# Missing Data Handling

- Output Records: 4128
- Records Deleted:760
- 760 records had missing values which were deleted.

# 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<br>• Duration Of Pitch<br>• Number Of Person Visiting<br>• Number Of Follow ups<br>• Preferred Property Star<br>• Number Of Trips<br>• Pitch Satisfaction Score<br>• Number Of Children Visiting<br>• Monthly Income | • Passport<br>• Own Car<br>• Typeof Contact<br>• CityTier<br>• Occupation<br>• Gender<br>• Product Pitched<br>• Marital Status<br>• Designation |

# Statistical modelling - Classification

## EDA

- Missing Data Handling
- Checked distributions of variables with histogram and box plot analysis
- Used scatter plots to measure relationship to the target variable

## Models Used

- Logistic Regression
- Decision Tree
- Neural Network

## Model Validation & Performance

- Precision, Accuracy, Recall, Error rates and F1 Scores
- AUC Score and ROC Curve
- Lift and Decile Charts

# Logistic Regression

**Our target variable is "product taken," which is a categorical variable. Therefore, we employed logistic regression, classification trees, and neural networks to evaluate and compare model performance in achieving our objective. I'll begin with logistic regression.**

| 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 |

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..

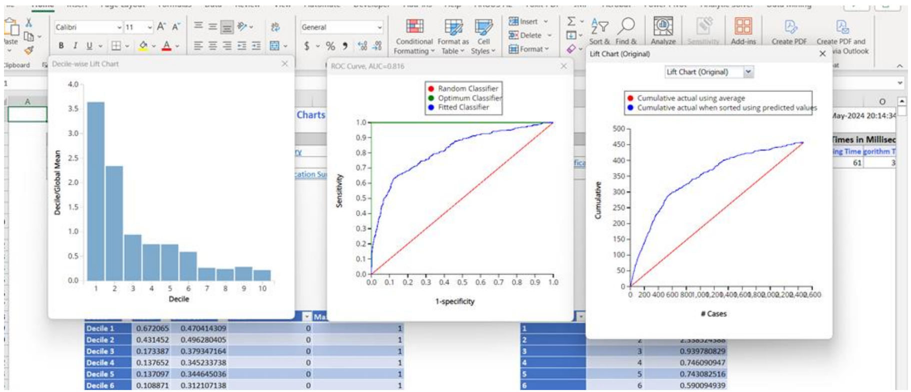
**variable selection techniques used:**

- Stepwise Selection
- Significance Based on P-Value
- Coefficient Close to Zero

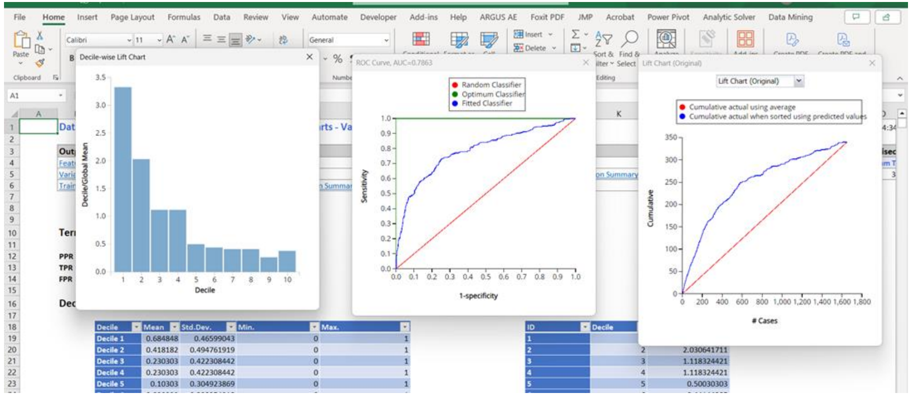
Following feature selection in logistic regression, we have identified Subset 13, comprising the following selected variables:

DurationOfPitch, NumberOfFollowups, PreferredPropertyStar, Passport, MonthlyIncome, TypeofContact, CityTier, Occupation_Salaried, Occupation_SmallBusiness, ProductPitched, MaritalStatus_Married, and MaritalStatus_Divorced.

## Training:

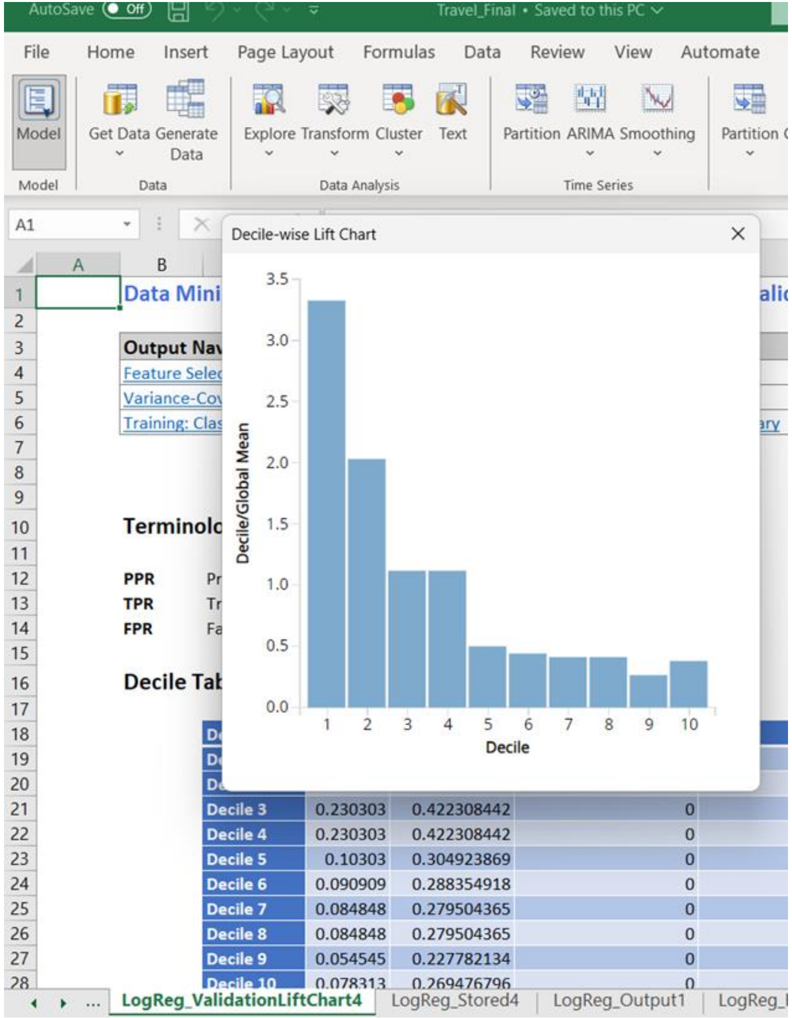


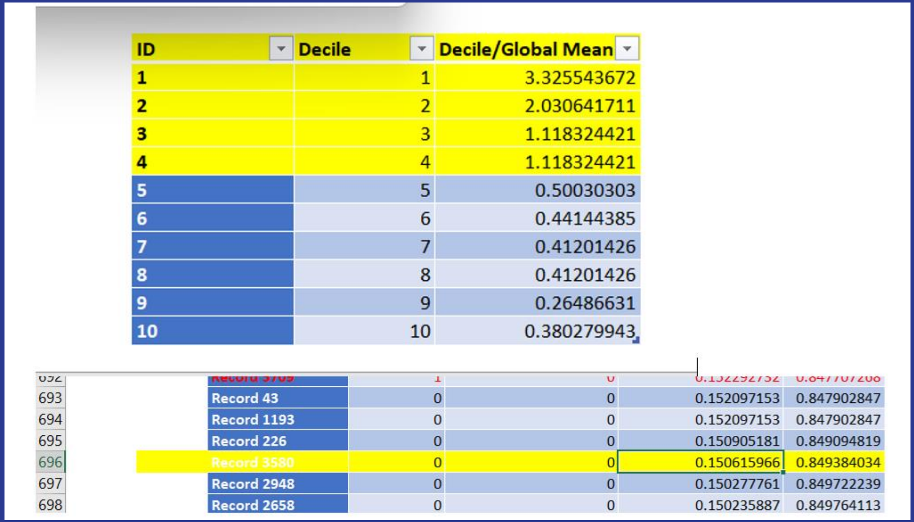

## Validation:




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.



| 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 |

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.

## 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 predicted to buy the product 69.3% bought the product.

The Sensitivity (Recall) is 0.7588, which means that out of all the persons who actually bought the product 75.88% were correctly predicted by the model.

| 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

**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))

| 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 |

**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

**Age:** Each additional unit of age decreases the odds of taking the product by about 0.972, all else constant.

DurationOfPitch: Increasing pitch duration boosts odds of product uptake by around 1.038 times, keeping other factors constant.

NumberOfFollowups: Each extra follow-up raises odds of product uptake by roughly 1.531 times, holding other variables steady.

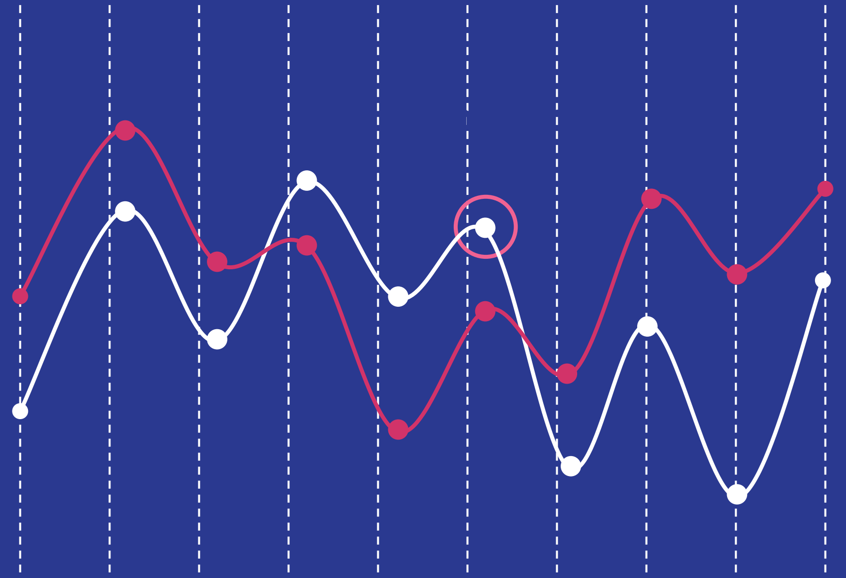
PreferredPropertyStar: Every increase in star rating preference elevates odds of product uptake by approximately 1.398 times, with other factors held constant.

**Passport: Holding a passport increases odds of product uptake by about 5.583 times, other variables remaining constant.**

**MonthlyIncome:** Monthly income has minimal impact on product uptake odds, with a coefficient close to zero (0.0001).

**TypeofContact_SelfEnquiry:** Self-enquiry contacts decrease odds of product uptake by about 0.673 times compared to company-initiated contacts, other variables held constant.

# Impact of coefficients



# Impact of coefficients

**CityTier_3: Being from City Tier 3 boosts odds of product uptake by roughly 2.245 times compared to other tiers, holding other variables constant.**

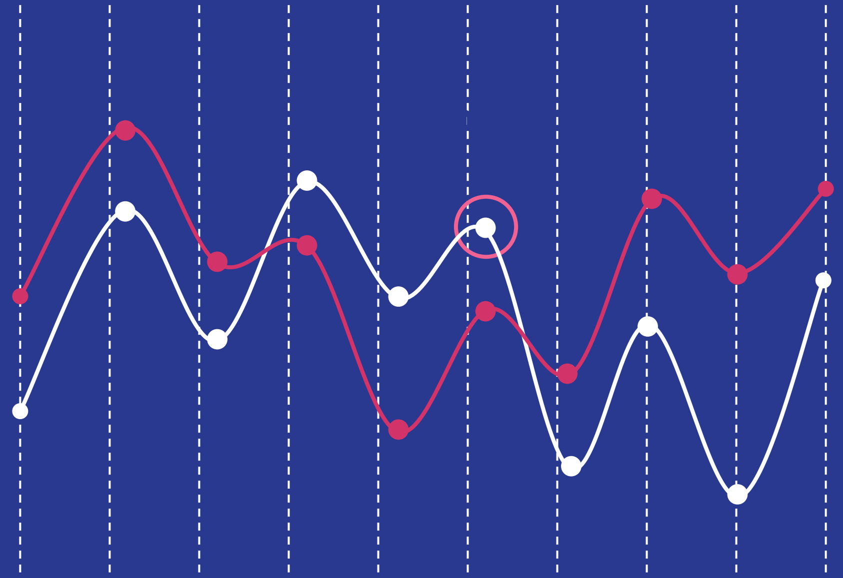
**Occupation_Salaried:** Salaried individuals have lower odds of product uptake, decreasing by around 0.7 times compared to other occupations, all else constant.

**Occupation_SmallBusiness:** Small business owners have even lower odds of product uptake, decreasing by about 0.55 times compared to other occupations, all else constant.

**ProductPitched_Deluxe: Pitching the deluxe product significantly decreases odds of product uptake by approximately 0.295 times, holding other variables constant.**

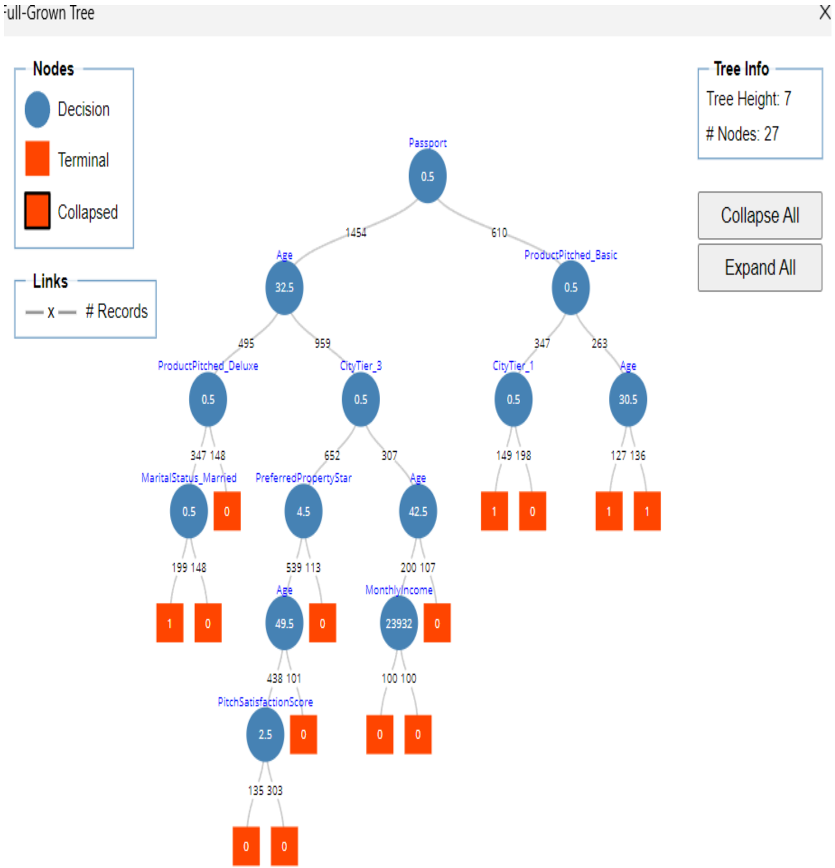
**MartialStatus_Divorced: Divorced individuals have lower odds of product uptake, decreasing by about 0.352 times compared to other marital statuses, holding other variables constant.**

MartialStatus_Married: Married individuals have lower odds of product uptake, decreasing by about 0.372 times compared to other marital statuses, holding other variables constant.

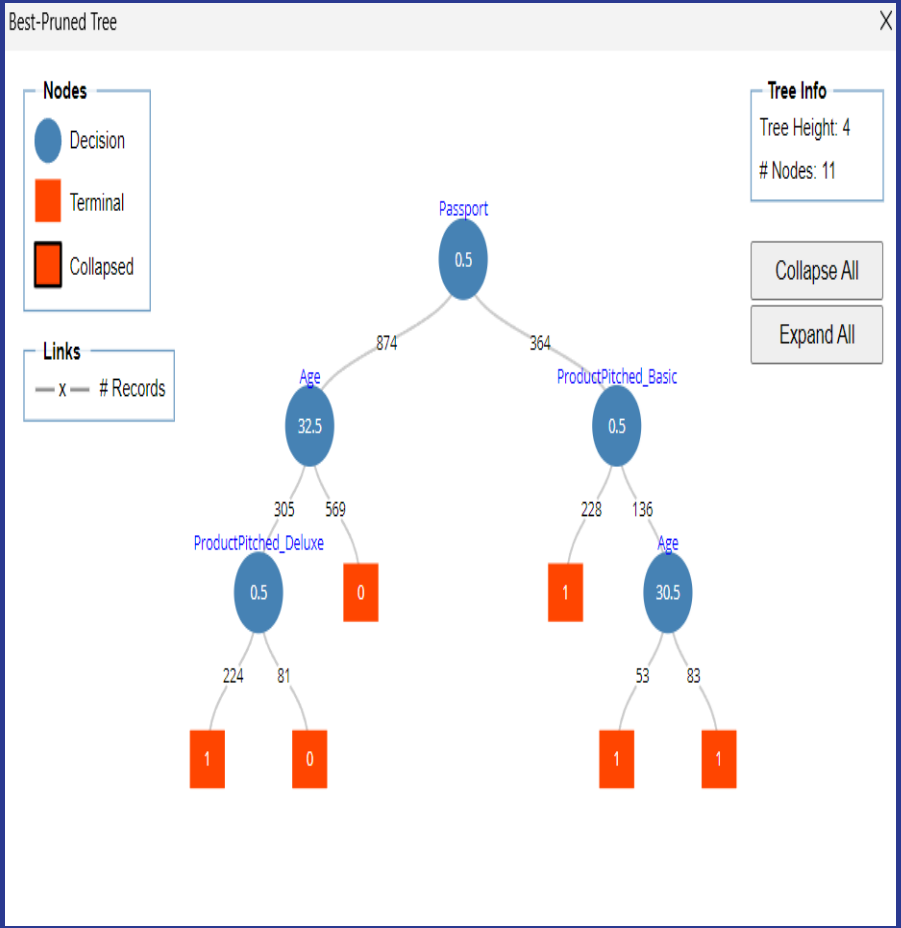


# Classification Tree

Since we  have a categorical predictor variable , we used a predictive model that organizes data into a hierarchical structure of decisions based on input features, enabling the classification of new data points into predefined categories.



ull-Grown Tree
X
Nodes
Decision
Terminal
Collapsed
Links
x # Records
Tree Info
Tree Height: 7
# Nodes: 27
Collapse All
Expand All
Passport
0.5
1454
610
Age
32.5
ProductPitched_Basic
0.5
495
959
347
263
ProductPitched_Deluxe
0.5
CityTier_3
0.5
CityTier_1
0.5
Age
30.5
347 148
652
307
149 198
127 136
MaritalStatus_Married
0.5
0
PreferredPropertyStar
4.5
Age
42.5
1
0
1
1
199 148
539 113
200 107
1
0
Age
49.5
0
MonthlyIncome
23932
0
438 101
100 100
PitchSatisfactionScore
2.5
0
0
0
135 303
0
0

Best-Pruned Tree
X
Nodes
Decision
Terminal
Collapsed
Links
x # Records
Tree Info
Tree Height: 4
# Nodes: 11
Collapse All
Expand All
Passport
0.5
874
364
Age
32.5
ProductPitched_Basic
0.5
305
569
228
136
ProductPitched_Deluxe
0.5
0
1
Age
30.5
224
81
53
83
1
0
1
1

- **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**

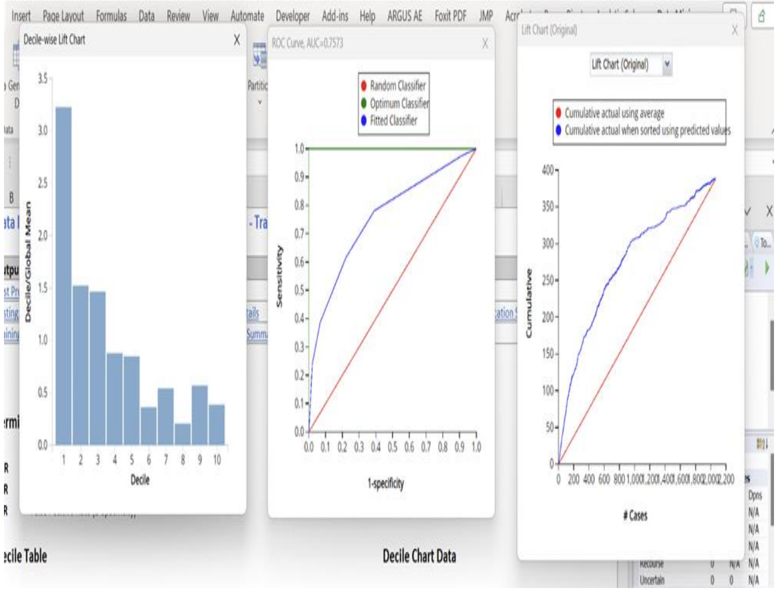
This rule applies when a person has a passport, the product pitched is basic, and their age is 30.5 or older. In this case, the class is 1.

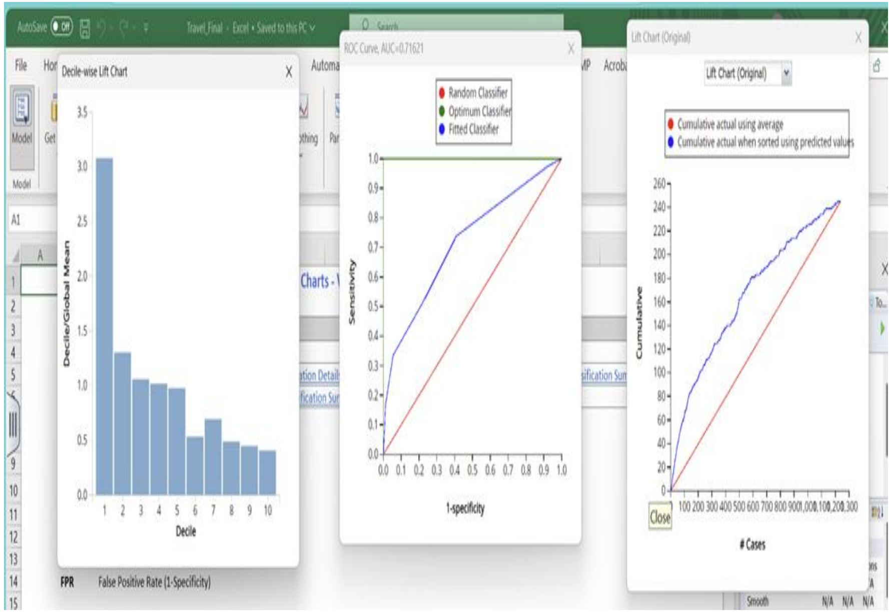
- **If Passport >= 0.5 and productpitched_basic >= 0.5 and Age < 30.5, class = 1**

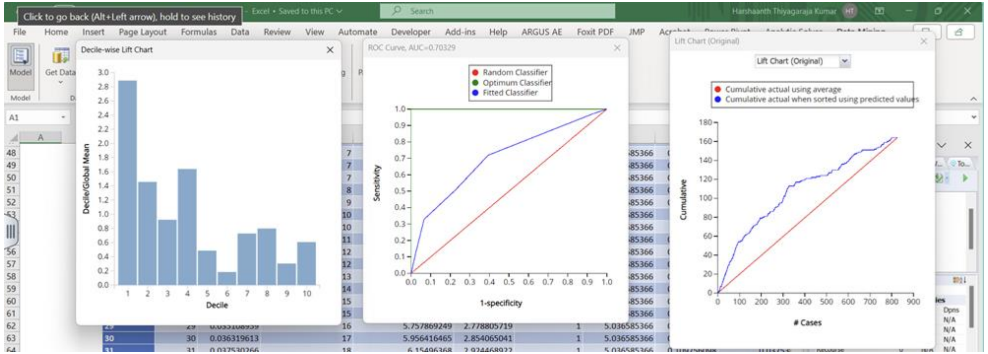
Similar to the previous rule, but here the age condition is less than 30.5.

- **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.


Decile-wise Lift Chart
Decile/Global Mean
Decile
ROC Curve, AUC=0.7573
Random Classifier
Optimum Classifier
Fitted Classifier
Sensitivity
1-specificity
Lift Chart (Original)
Cumulative actual using average
Cumulative actual when sorted using predicted values
Cumulative
# Cases
Decile Table
Decile Chart Data


Decile-wise Lift Chart
Decile/Global Mean
Decile
ROC Curve, AUC=0.71621
Random Classifier
Optimum Classifier
Fitted Classifier
Sensitivity
1-specificity
Lift Chart (Original)
Cumulative actual using average
Cumulative actual when sorted using predicted values
Cumulative
# Cases
FPR
False Positive Rate (1-Specificity)


Click to go back (Alt+Left arrow), hold to see history
Decile-wise Lift Chart
Decile/Global Mean
Decile
ROC Curve, AUC=0.70329
Random Classifier
Optimum Classifier
Fitted Classifier
Sensitivity
1-specificity
Lift Chart (Original)
Cumulative actual using average
Cumulative actual when sorted using predicted values
Cumulative
# Cases

## 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 |

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 validation and test datasets, suggesting that it will perform well on new datasets too.

The validation AUC value of 0.71 and the test AUC of 0.70 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.

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 predicted to buy the product 60.6% bought the product.

The Sensitivity (Recall) is 0.72, which means that out of all the persons who actually bought the product 72% were correctly predicted by the model.

# Neural Networks

**Neural Networks describe relationships in data other models can’t and thus work well for classification analysis. Since our target variable is categorical we have employed a neural network for our analysis.**

# Immediate Issues

- We initially had issues with the model performance
- We would see the model classify no positive values after a run of the neural network
- This led to many runs of manual neural networks with different parameters
- Settled on one with high ROC score and changed the cutoff value

**Network Structures Tested:**
1 hidden layer 5 nodes
1 hidden layer 10 nodes
2 hidden layers each with 5 nodes
1 hidden layer with 2 nodes

**Training: Classification Summary**

**Confusion Matrix**

| Actual\Predicted | 0 | 1 |
|---|---|---|
| 0 | 1676 | 0 |
| 1 | 388 | 0 |

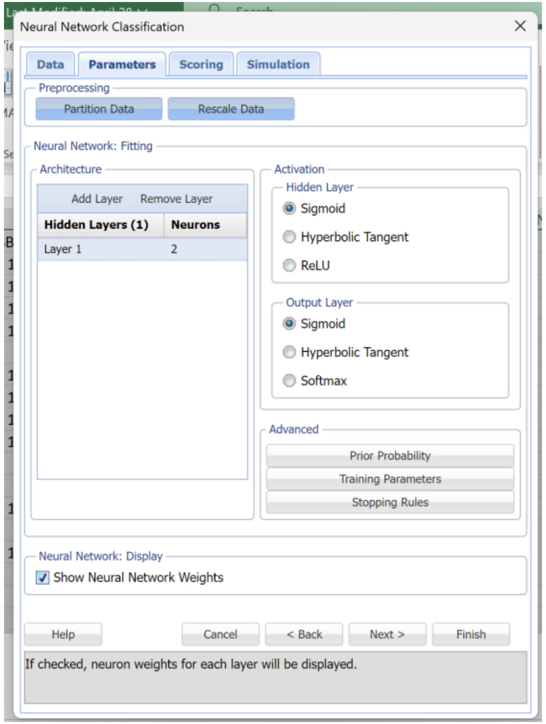
**Error Report**

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

**Metrics**

| Metric | Value |
|---|---|
| Accuracy (#correct) | 1676 |
| Accuracy (%correct) | 81.2015504 |
| Specificity | 1 |
| Sensitivity (Recall) | 0 |
| Precision | #N/A |
| F1 score | #N/A |
| Success Class | 1 |
| Success Probability | 0.5 |

# Data Preprocessing




Initially we had problems getting the network to work with our data

Ultimately for our model we used all the variables in the dataset to construct it

Split the dataset into 60% training and 40% test variables

Standardized the variables

# Cutoff Selection

**Training: Classification Summary**

| Confusion Matrix | | |
|---|---|---|
| Actual\Predicted | 0 | 1 |
| 0 | 1676 | 0 |
| 1 | 388 | 0 |

| 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 |
| Accuracy (%correct) | 81.2015504 |
| Specificity | 1 |
| Sensitivity (Recall) | 0 |
| Precision | #N/A |
| F1 score | #N/A |
| Success Class | 1 |
| Success Probability | 0.5 |

412.8 =.2*2064

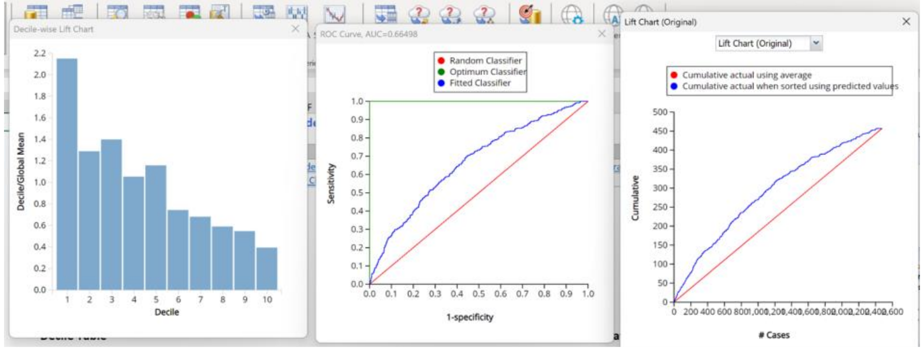
new cutoff = .3099

| | | | | |
|---|---|---|---|---|
| Record 1727 | 0 | 0 | 0.689912603 | 0.310087397 |
| Record 3773 | 0 | 0 | 0.68993101 | 0.31006899 |
| Record 2381 | 1 | 0 | 0.6899882 | 0.3100118 |
| Record 1760 | 0 | 0 | 0.690058895 | 0.309941105 |
| Record 2100 | 0 | 0 | 0.690070467 | 0.309929533 |
| Record 2397 | 0 | 0 | 0.690099033 | 0.309900967 |
| Record 3279 | 0 | 0 | 0.690135398 | 0.309864602 |

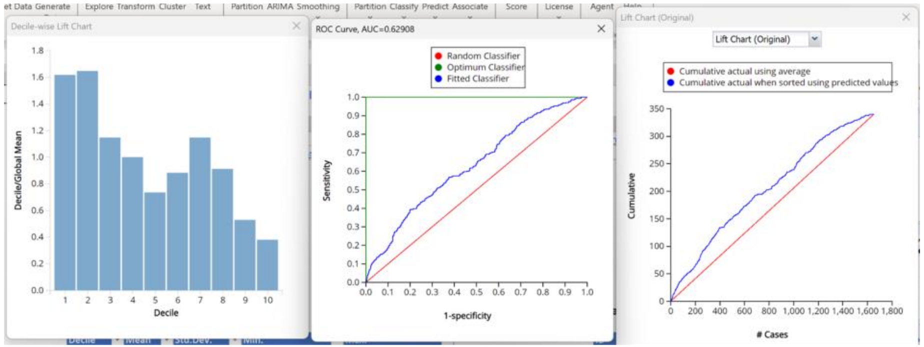
- **We decided the best network to use was one with an ROC score of .6636**
- **To fix the classification issues we were having in earlier runs of the neural network we decided to select a new cutoff value**
- **There was a massive drop off in the decile charts at decile 2 so we selected the probability of that value as our cutoff**
- **The new cutoff value selected is: .3099**

# Model Results

## Training




## Validation




We had a mixed result running this model on our data

The accuracy and specificity of our model are both high, as is the ROC score

However, the recall, precision and F1 scores of this model are all very low which shows poor performance of our model

## Training

**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

**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 |

# Model Comparison

- The training AUC of 0.816 and validation AUC of 0.7863 closely align, indicating consistent performance in logistic regression, outperforming the classification tree and neural networks.
- The model achieves an overall accuracy of 70.68%, 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 sensitivity (recall) of 0.7588 indicates that the model correctly identifies 75.88% of actual buyers. The specificity of 69.3% suggests that 69.3% of predicted non-buyers do not purchase the product.
- These metrics enable us to make informed decisions on marketing investments, effectively and efficiently targeting specific customer groups.

# 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 Logistic Regression We can answer our Business Questions:

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

**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.

**Example:** 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)=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

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.

# 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.