Holiday Package Purchase Prediction

Project Report

Geetha Ragiphani

**Table of Contents**

1. **Project Description**
   * Introduction to Trips & Travel.Com's mission and the introduction of the Wellness Tourism Package.
2. **Business Questions**
   * Specific questions addressed by the project.
3. **Dependent & Independent Variables**
   * Overview of the variables used in the analysis.
4. **Data Preprocessing**
   * Handling missing data and summary of changes made.
5. **Summary Characteristics**
   * Mean, median, standard deviation, and identification of outliers.
6. **Correlation Analysis**
   * Explanation and comments on the correlation between variables.
7. **Visualization**
   * Histograms, scatterplots, and boxplots of important variables.
8. **Logistic Regression Model**
   * Description of the model, variable selection, model output, interpretation of coefficients, and training and validation summary, including lift charts and model performance metrics.
9. **Classification Tree Model**
   * Description of the model, variable usage, model output, interpretation of rules, and training and validation summary, including lift charts and model performance metrics.
10. **Neural Network Model**
    * Description of the attempted model, variable usage, model output, training summary, and training and validation summary, including lift charts and model performance metrics.
11. **Comparison Between Models**
    * Evaluation of model performance and comparison of metrics.
12. **Conclusions**
    * Decision on the best model, answers to business questions, and recommendations.
13. **Summary**
    * Reflections on the analysis process, and lessons learned.
14. *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.

1. *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.
2. *Dependent & Independent Variables:*

A table with blue text and white text

Description automatically generated

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.

1. *Data Preprocessing (refer to milestone 2, please revise according to my comments)*
   * Handling missing data

# Output Records: 4128

#Records Deleted:760

1. cords had missing values which were deleted.
2. Summary characterizes (mean, median, sd), any outliers?

A screen shot of a computer

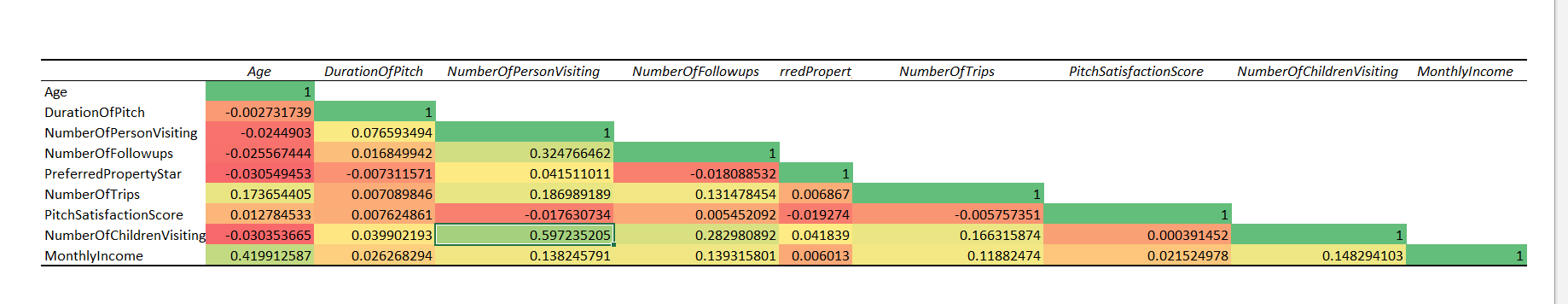
Description automatically generatedA screen shot of a number

Description automatically generatedA black and white screen with numbers and percentages

Description automatically generated

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

1. Correlation table (and comments/explanations)



NumberOfChildrenVisiting is highly correlated with NumberOfPersonVisiting

Secondly, MonthlyIncome is highly correlated with Age

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

A graph of age and age

Description automatically generatedA graph with a red bar

Description automatically generated with medium confidenceA graph with red bars

Description automatically generatedA graph of a number of income

Description automatically generatedA graph with numbers and a bar

Description automatically generated

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.

A graph with numbers and symbols

Description automatically generated with medium confidenceA graph with red dots

Description automatically generatedA graph with numbers and lines

Description automatically generatedA graph with red dots

Description automatically generatedA graph with red dots

Description automatically generated

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.

A green rectangle with white lines

Description automatically generatedA green bar graph with numbers and a line

Description automatically generatedA green rectangle with white lines

Description automatically generatedA green square with white lines

Description automatically generatedA green and black graph

Description automatically generated

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.

1. *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)



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.

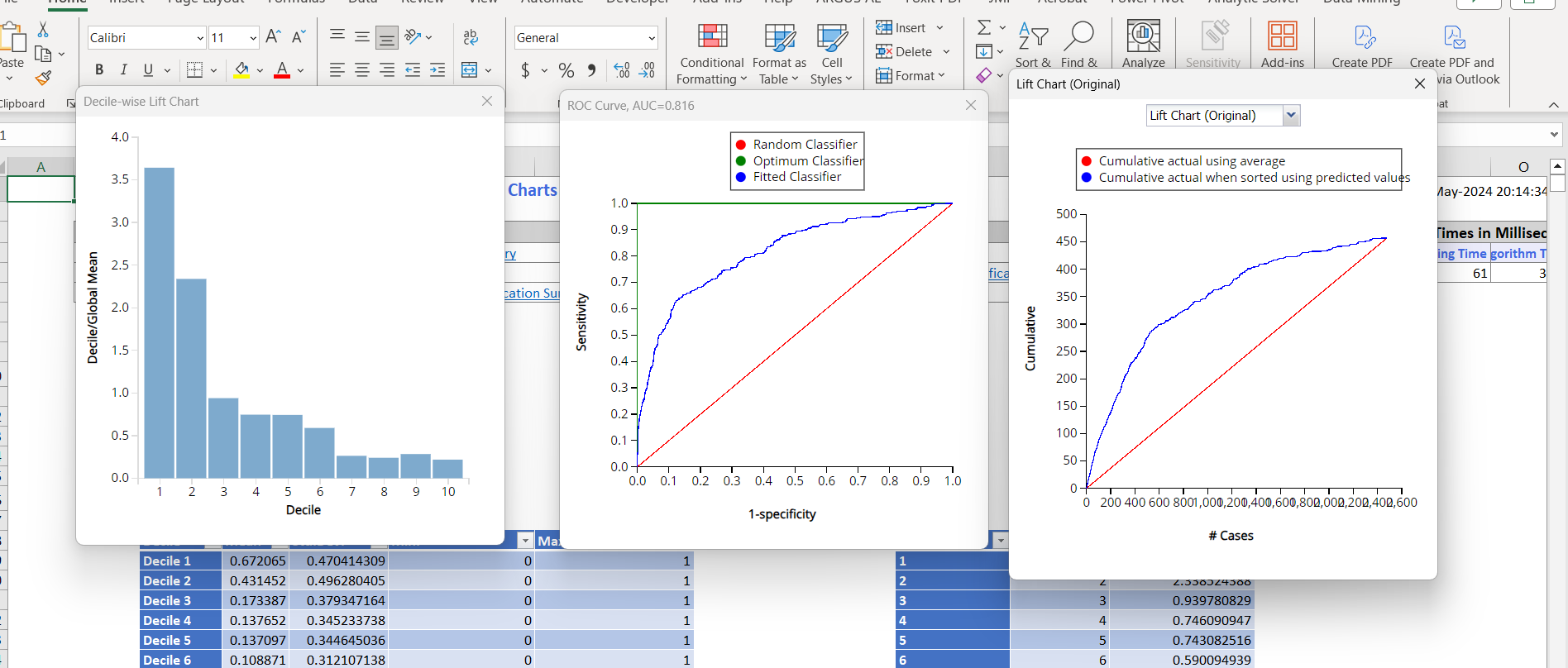
ProductPitched\_Deluxe: Pitching the deluxe product decreases the odds of taking the product significantly. The odds decrease by approximately 0.295 times when the deluxe product is pitched, holding other variables constant.

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.

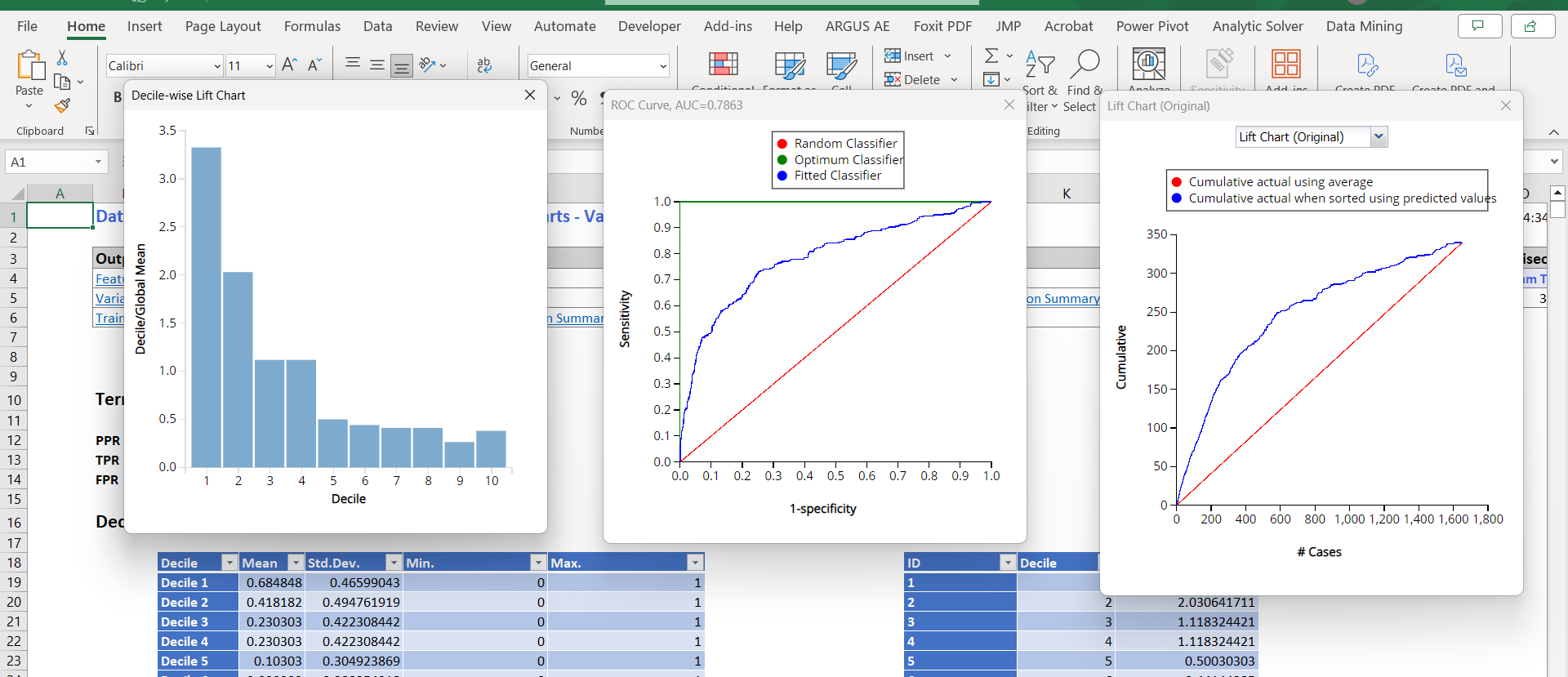
MartialStatus\_Married: Married individuals are also less likely to take the product compared to others. The odds decrease by about 0.372 times for individuals who are married, holding other variables constant.

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

Training:



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.

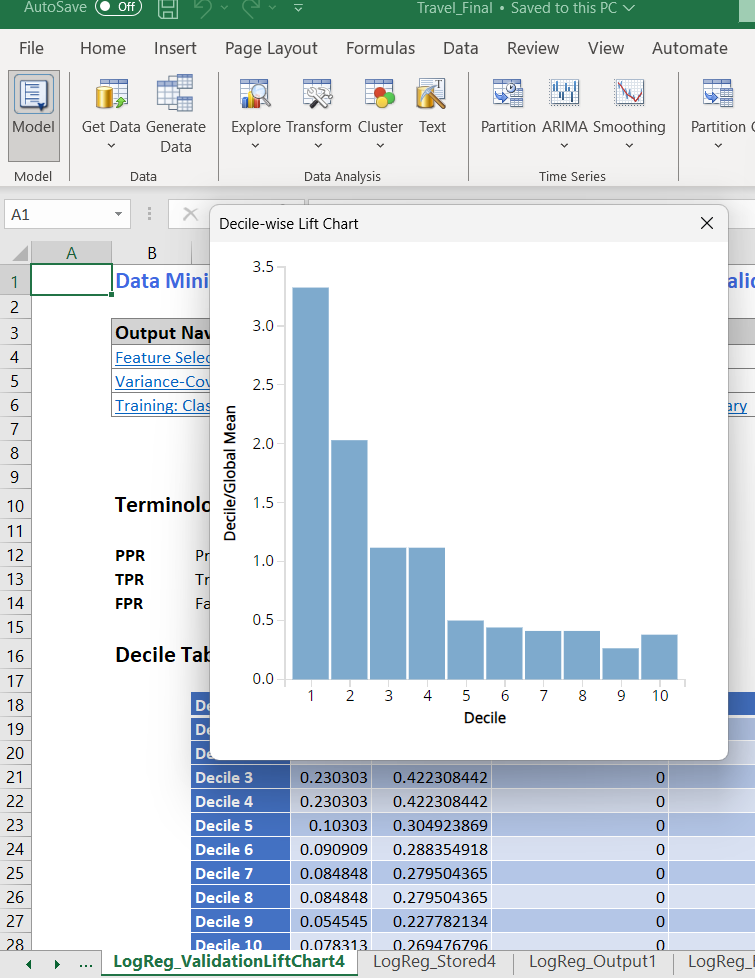
The training AUC value of 0.816 and the validation AUC of 0.7863 are closely aligned, indicating consistency in performance between the training and validation datasets. This similarity suggests no apparent overfitting issues, and that the model exhibits good predictive capability and generalizability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

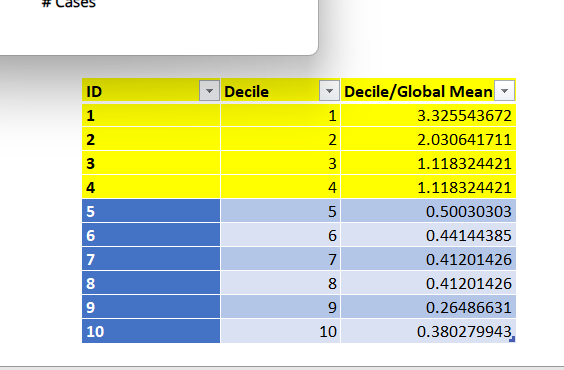


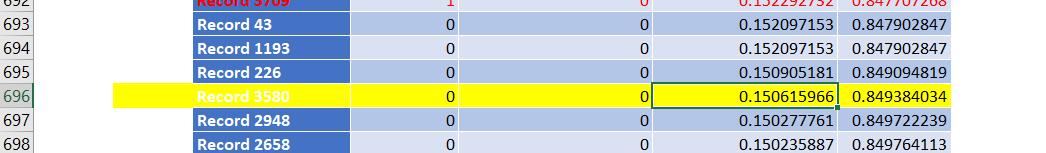
The Overall Accuracy of the model is 70.68%. The Error rate of Class 0 is 30.66%, The Error rate of Class 1 is 24.12% and the overall error rate is 29.32%.

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.





* + 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)^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.

1. *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:

A screenshot of a computer screen

Description automatically generated

Best Pruned Tree:

A screenshot of a computer screen

Description automatically generated

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:

A screenshot of a computer

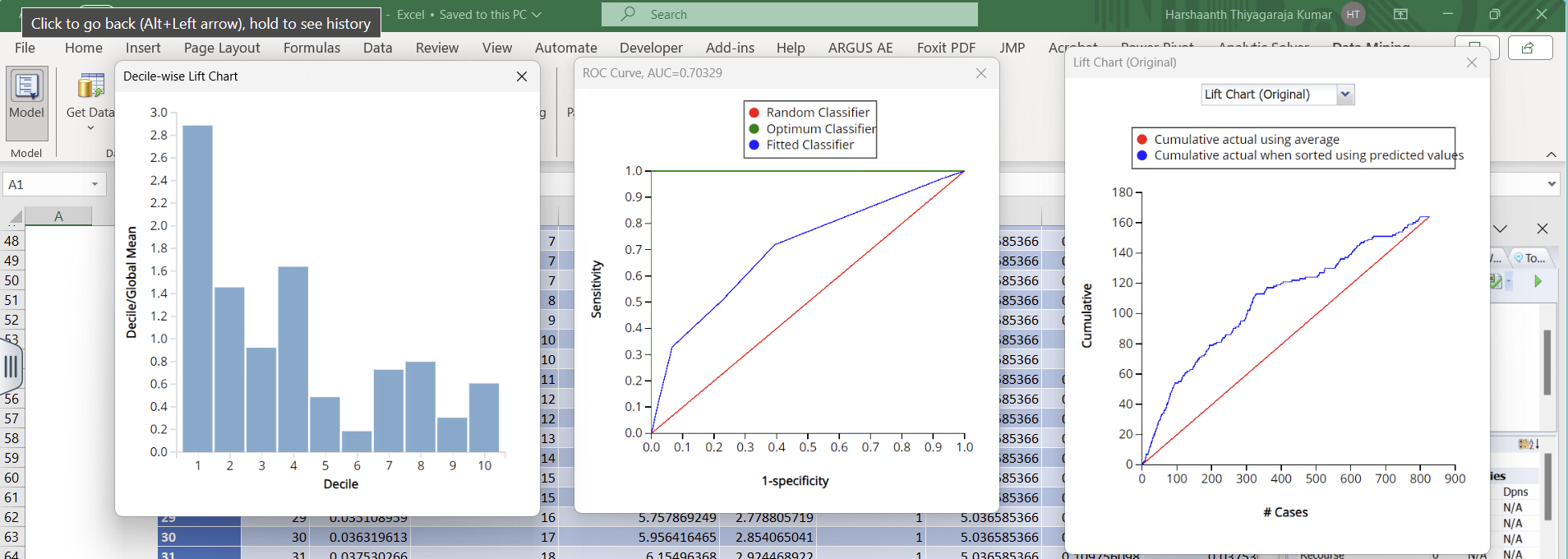
Description automatically generated

Validation Lift Chart

A screenshot of a computer

Description automatically generated

Test Lift Chart:



A screenshot of a computer

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

A yellow line with red and black numbers

Description automatically generated

* + 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.

1. *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)

A screenshot of a computer

Description automatically generated

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?

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

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.

A screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

* + 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.

1. 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.

1. 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:

1. 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.
2. 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
3. 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.

1. *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.