1)

Chart

Description automatically generated

Chart, diagram

Description automatically generated

Chart, histogram

Description automatically generated

True value of cancellation is 0.366, predicted by model is 0.365962.

Chart, line chart

Description automatically generated

This is the ROC plot comparing the model only dependent on lead rate (blue line) with model dependent on both lead time and average daily rate. Clearly, the model with 2 parameters have area under the curve (AUC) higher than that of first. So, we can say that the is model is more equipped in predicting whether the cancellation will occur or not. AUC for first model is 0.671 and for second model is 0.6826.

1b)

Chart, line chart, scatter chart

Description automatically generated

If we add all four predictors, namely lead time, average daily rate, previous cancellations and if the person is repeated guest, our model improves. The AUC improves from 0.6826 to 0.7128. This is evident from the curve above.

2a)

Calendar

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Party, Trump Approval, Education , Climate change have strong interactions with people beliefs on safety of vaccines.

2b) Based on the data that we observe, 64.2 % was classified correctly. The model seems to classify all the data as “Strongly Agree”. Infact , major bulk of data corresponds to “strongly agree” class. We chose party and education as the two response variable.

I am not sure what the question was being asked. I suppose, the question is to create a 6 model with 6 different response variables and determine which one gives maximum accuracy. So, creating my model based on that.

|  |  |
| --- | --- |
| Parameter | Accuracy(%) |
| Age | 64.2 |
| Party | 64.2 |
| Education | 64.2 |
| Books | 63 |
| Trump Approval | 64.2 |
| Climate Change | 64.2 |

From the table we see that accuracy is same when we have one parameter model. Only the books, have lower accuracy compared to others. Hopefully, I have answered the required question.

2c) I selected all the 6 response variable to create a model and got an accuracy of 63.2 %. This is on expected lines as my response variable book has lower accuracy, so it is dragging the prediction of my overall model. Also, the other variables is lifting the predictions by the model containing only books.

2d) The classifier with “party” and “education” as classifier seemed to work good for our model. This is because, if we add all the response variable, the response variable “books” seems to be making our model worse. The sensitivity and specificity is defined for a two level model. Here, we have five levels. If we consider “strongly agree” as level 1 and Positive and rest others as level 2 and Negative.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | Predicted |  |
| Truth | Other than “Strongly Agree” | Strongly Agree | Total |
| Other than “Strongly Agree” | 0 | 358 | 358 |
| Strongly Agree | 0 | 642 | 642 |
| Total | 0 | 1000 | 1000 |

For two parameter model

Sensitivity = 642/642 = 1

Specificity = 0/358 = 0

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | Predicted |  |
| Truth | Other than “Strongly Agree” | Strongly Agree | Total |
| Other than “Strongly Agree” | 9 | 349 | 358 |
| Strongly Agree | 14 | 628 | 642 |
| Total | 23 | 977 | 1000 |

For all parameter model

Sensitivity = 628/642 = 0.978

Specificity = 9/358 = 0.025

So, we see that by adding parameters, the sensitivity decreases but specificity increases. However, the overall accuracy decreases.