**Context:**

You work for a large travel agency called Traveligo. The company is a travel supplier, based in the Montreal area.

By this very fact, your clientele is made up of 70% of Quebecers. This is winter, and the company is aiming to grow international sales, the company runs monthly promotions to entice their customers to buy a plane ticket to a certain destination.

As a Data Analyst in Traveligo, you are responsible for defining the right promotions to offer each month. To do this, you use your knowledge of data mining techniques.

Part I: Classification tree

You decide to build a predictive model to better target the people who are most likely to go to Vinales. You then build 3 models: a classification tree model, a random forest model and a neural network model.

Here is your final classification tree:

Timeline

Description automatically generated

1. For this classification tree, describe the model rules that lead to the prediction of a trip to Vinales. With these results, briefly describe what the campaign would consist of to attract new customers to this destination.

* Rule 1 :

27% of customers **aged 35 and older with an income over 60k** will buy the plane ticket to Vinales. (300 people; 300/1560 = **19.2%** of all clients)

* Rule 2 :

89% of customers **aged 24 and younger with income over 60k** will buy the plane ticket to Vinales. (100 people; 100/1560 = **6.4%** of all clients)

* Rule 3:

13% of customers whose **income is between 20k and 60k** will buy the plane ticket to go to Vinales. (200 people; 200/1560 = **12.8%** of all customers)

* Rule 4:

95% of clients whose **income is less than 20k and who have a university degree and who have traveled more than twice a yea**r will buy the plane ticket to go to Vinales. (560 people; 560/1560 = **36%** of all customers)

* Rule 5:

18% of customers with an **income of less than 20k and who have a university degree** and who have traveled twice or less per year will buy the plane ticket to go to Vinales. (100 people; 100/1560 = **6.4%** of all clients)

* Rule 6:

15% of customers whose **income is less than 20k and who do not have a university degree** will buy the plane ticket to go to Vinales. (300 people; 300/1560 = **19.2%** of all customers)

Based on the results of this tree, to increase the Vinales ticket sale, the campaign would better prepare promotions for rule 4 customers who build up 36% of the company's customer base.

b) Using the information in the classification tree output, determine the misclassification rate. By comparing the models’ misclassification rate, please comment on the usefulness of the model for Traveligo.

Here are two summary tables of the measurements for the three models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Training (75 000 observations)** | **Misclassification rate** | **Cumulative lift, 4th decile** | **ROC** | **Sensitivity** |
| Neural network | 15% | 3,2 | 0.81 | 51% |
| Decision tree | 18% | 2,1 | 0.86 | 63% |
| Random forest | 21% | 2,7 | 0.8 | 76% |
| **Validation (40 000 observations)** | **Misclassification rate** | **Cumulative lift, 4th decile** | **ROC** | **Sensitivity** |
| Neural network | 18% | 2,7 | 0.77 | 45% |
| Decision tree | 18% | 2 | 0.85 | 63% |
| Random forest | 21% | 2,7 | 0.8 | 75% |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Rule 1 | Rule 2 | Rule 3 | Rule 4 | Rule 5 | Rule 6 | sum |
| % of customers who bought the ticket | 27% | 89% | 13% | 95% | 18% | 15% |  |
| # of individuals in the leaf | 300 | 100 | 200 | 560 | 100 | 300 | 1560 |
| # of customers who traveled to Vinales | 81 | 89 | 26 | 532 | 18 | 45 |  |
| # of customers who didn't travel to Vinales | 219 | 11 | 174 | 28 | 82 | 255 |  |
| #misplaced individuels | 81 | 89 | 26 | 28 | 18 | 45 | 287 |

Misclassification rate (MCR) = (81 + 89 + 26 + 28 + 18 + 45) / 1560 = 209/1560 = 13.4%

Natural error rate (NER) = 287 / 1560 = 22%

MCR < NER 🡪 The tree will predict wrong 13.4% of the time whereas before the model, the model was prediction wrong results 22% of the time. So, with the tree, we have a gain of 8.6% (22% -13.4% =8.6%) on the error i.e., that the tree learns from data.

1. Knowing that you have a greater importance in predicting the people who have traveled, which is the best model among the 3 models (justify your answer)?

In both cases (75k observations and 40k observations), the misclassification rates (MCR) of the three models are smaller than the natural error rate (NER). So, all three models are doing a good job in predicting the results, but that doesn't mean that a model whose MCR is much smaller than the TNE is necessarily the best model. Here we try to correctly predict the 1s of target (people who are going to travel); therefore, special attention must be paid to the sensitivity which is an indicator of the model’s strength in terms of predicting the 1s.

Neural network (RN) has a smaller MCR compared to the other models in both cases (75k observations and 40k observations), and its accumulated lift is also greater than other models, but this model has a low sensitivity, especially for the 40k observation where the sensitivity is lower than 50%. This means that the RN is not very good at predicting customers who will buy a ticket (the 1s). Therefore, we reject using the RN model.

On the other hand, the sensitivity of the random forest is the best of the others in both cases (75% and 76%). Also, its cumulative lift value is greater than that of decision tree. The high ROC value of the decision tree shows that the decision tree is more powerful in predicting the customer who won’t buy a ticket (not those who buy a ticket).

Also, it is important to note that all models are exposed to over-learning but ensemble methods are less exposed to over-learning and Random forest, which is an ensemble method has lower chance of over-learning.

For reasons mentioned above, I choose the random forest model to predict the customer who will buy a ticket.