BAN 5733 – Individual Exercise 9 (10 Points)

Exercise Description:

A telecom operator, which has successfully launched a fourth generation (4G) mobile telecommunications network, would like to make use of existing customer usage and demographic data to identify which customers are likely to switch to using their 4G network. The target categorical variable is "Customer_Type" (3G/4G). A 4G customer is defined as a customer who has a 4G Subscriber Identity Module (SIM) card and is currently using a 4G network compatible mobile phone. You are asked to assist them in examining the data.

In your first meeting with the company, the clients kept talking about following issues.

- They would like to know how much an issue missing data is in their data set and what approach you took to deal with it. **(1 point)**

- The company would like to ensure the information is from a valid sample of their customers. The company would like to see a training data set that is 70% of their customer set. **(1 point)**

- They believe that classifying people correctly is the most important metric to examine. **(1 point)**

- The telecom company would like to know just how well they could determine those who are really going to change to 4G and those will not. **(2 points)**

- They would also like to know exactly what the rules are for classifying people, how to use them and how well they will work for them. **(3 points)**

You are provided with the telecommunications data (filename: **Telecom_Fall2021.jmp**) and asked to create a manager-friendly report that covers results of your analysis. You will use **12345 for any random seed** you need. The data set contains 18 variables and over 6,000 observations. You report should be restricted to 7-pages maximum. Please back up your assertions with appropriate statistics and graphs as needed. The variables in the data set are shown in the table below with the appropriate roles and levels.

Variable Description:

Attribute Name	ROLE	LEVEL	Description
Age	INPUT	INTERVAL	In Years
Contract_flag	INPUT	NOMINAL	Contract ownership flag (Y/N)
customer_class	INPUT	NOMINAL	Codes indicating VIP, Individual, Corporate, Government, Under 21, Foreigner, etc.
customer_type	TARGET	BINARY	TARGET FIELD: 4G Customer Flag (3G/4G)
Gender	INPUT	NOMINAL	Male or Female
HS_AGE	INPUT	INTERVAL	Handset Age in Months
ID_change_flag	INPUT	NOMINAL	1 if the customer change the ID in the last 6 month
Line_tenure	INPUT	INTERVAL	Line tenure in days
Marital_status	INPUT	NOMINAL	Marital status
Nationality	INPUT	NOMINAL	Nationality
occup_CD	INPUT	NOMINAL	Occupation code
pay_metd	INPUT	NOMINAL	Payment method code such as Credit card, Cash, etc. as of last billing cycle in the 6 months
subplan	INPUT	INTERVAL	Current Subscription Plan Type
serial_num	ID	NOMINAL	Record Index (ID)
top1_int_cd	INPUT	NOMINAL	Top 1 international country
top2_int_cd	INPUT	NOMINAL	Top 2 international country
top3_int_cd	INPUT	NOMINAL	Top 3 international country

Deliverables (please follow these instructions):

- *As you complete the exercise, create a report in Microsoft Word. In this report, answer the questions in the exercise description.*
- *Make sure you comment or explain and not just provide snapshots of data.*
- *Limit your report to no more than **7 pages** including tables and diagrams.*
- *Copy and paste or screen shot supporting tables/diagrams as needed to justify any of your answer. You may need to shrink your table/ diagrams but please ensure they are readable.*
- *Include any required data sets or codes/projects as requested as separate files.*
- *Make sure you print your name, student ID#, student email on the cover page of the report and turn-in the report as communicated by your instructor.*
- *Please also put a running header/footer with your name, on each page of your exercise solution report.*

Failure to follow these instructions will result in deduction of points

4G Telecommunications Network Customer Report Customer Report

This report will summarize the findings of customer usage and demographic information. The report will also convey the accuracy of the decision tree models created to predict whether a customer would switch to 4G service or maintain their 3G service. Management wanted to ensure accuracy of sampling methods, the best optimization method for their classification business problem, and that resulting models are accurate and specific.

To address the first concern, a missing values analysis on the 6,000 records in the data set was conducted. Occupation Code had the highest number of missing data with 63% of the records missing information. The next variable only had 6.5% missing values. Fortunately, decision trees are very good at working with missing data and treats the missing information as its own field.

Table 1: Missing Values Analysis

Variable	Count of Missing	Percent of Records
Occupation Code	3,804	63.4%
Marital Status	388	6.5%
Payment Method	313	5.2%
Contract Flag	268	4.5%
Age	192	3.2%
Nationality	5	0.1%

Next, the supplied data set had a breakdown of 50% 3G customers and 50% 4G customers. A stratified random sample and a 70%/ 30% partitioning split were employed to ensure the similar proportions were obtained from the training and validation data sets to reflect the overall population of customers. Table 2 summarizes this process.

Table 2: Sampling and Partitioning Outcomes

Customer Type	Full Dataset		Training		Validation	
	Count	% of col count	Count	% of col count	Count	% of col count
3G	3,000	50.0%	2,100	50.0%	900	50.0%
4G	3,000	50.0%	2,100	50.0%	900	50.0%
Total	6,000	100%	4,200	70.0%*	1,800	30.00%*

*indicates the percent is of the full data set total instead of the column total

A decision tree models was created using the misclassification rate as the model assessment statistic. This statistic was selected because the business problem aims to classify customers and misclassification will indicate how well the prediction worked. Specifically, it will examine the training and validation data to see if the predicted category the model develops is what occurred in reality. The lower the rate the better the model is at classify customer type (3G/ 4G status).

The decision tree conducted in JMP Pro 14 resulted in 13 terminal leaves to be interpreted. Figure 1 shows the split history for the resulting decision tree and indicates that this is the best number of terminal leaves for this data set. These leaves ultimately become business rules for the company to enact during the future scoring process to determine if someone will switch to 4G status. The company can then target only those people who fit the business rules (or leaf) that results in high percentages of 4G customer prediction.

Figure 1: Split History for Decision Tree

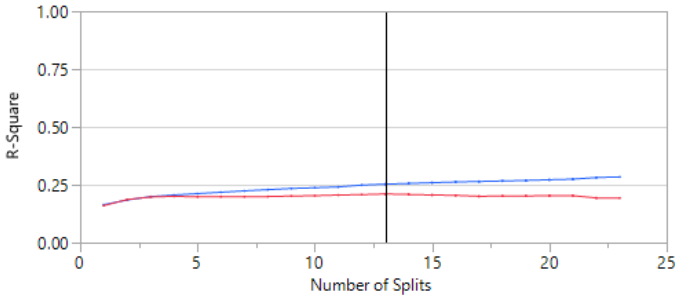


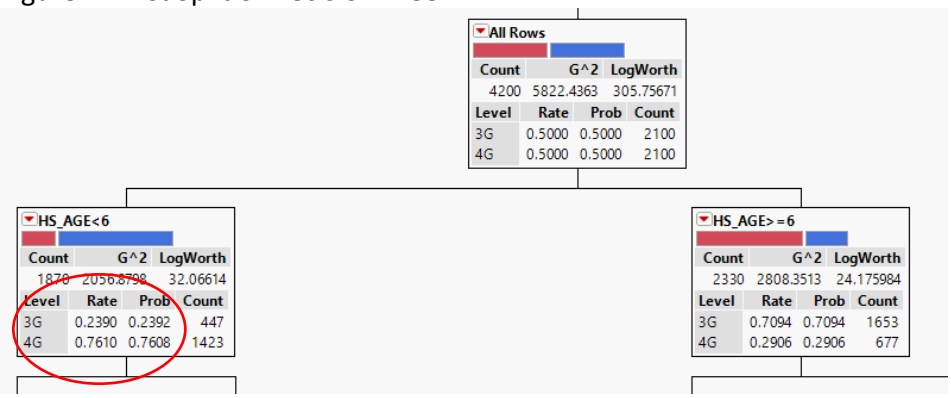
Table 3 indicates the variables that contribute most to this classifying decision tree. The age of the handset (HS_Age) was found to be the most important variable in this model and accounted for the first split followed by subplan, customer's age, marital status and the contract_flag. These results indicate that the age of the handset and the subplan type are more predictive in determining 3G and 4G services. Figure 2 shows the first split of this decision tree and indicates that if the handset age is less than 6 months than the customer is more likely to switch to 4G.

Table 3: Column Contributions

Column Contributions

Term	Number of Splits	G^2	Portion
HS_AGE	4	1148.84072	0.7789
SUBPLAN	5	159.527645	0.1082
AGE	2	118.916639	0.0806
MARITAL_STATUS	1	29.6656781	0.0201
CONTRACT_FLAG	1	18.0848785	0.0123
GENDER	0	0	0.0000
NATIONALITY	0	0	0.0000
OCCUP_CD	0	0	0.0000
LINE_TENURE	0	0	0.0000
CUSTOMER_CLASS	0	0	0.0000
PAY_METD	0	0	0.0000
TOP1_INT_CD	0	0	0.0000
TOP2_INT_CD	0	0	0.0000
TOP3_INT_CD	0	0	0.0000

Figure 2: First Split of Decision Tree



The misclassification rate for this model was 24.9% indicating that about 25% of the time a person would be misclassified into 4G when they were really not planning on moving or vice versa. Further, the model will accurately predict 4G customers 73% of the time (sensitivity). The model will also accurately predict those who will not switch to 4G service 71% of the time.

Table 4: Confusion Matrix for Validation Decision Tree

Actual Customer Type	Predicted Customer Type		Total
	3G	4G	
3G	636	264	900
4G	241	659	900
Total	877	923	1,800

Actual Total Positive = True Positive + False Negative
Sensitivity = True Positive/ Actual Total Positive = 659/ (241+659) = 659/900=.7322 = 73%
Specificity = True Negative/ Actual Total Negative = 636/ (636+264) = 636/900 = 0.7067 = 71%

Additional work could be done to reduce the misclassification rate but this is a good start and better than a random guess at whether a customer will switch.

For the customer to accurately pinpoint those customers who are likely to switch to 4G service, they will need to examine the leaf report for the rules to take action on. Figure 3 gives a concise look at this information sorted in descending fashion for 4G service. The first rule indicates that a person with a handset age is between 6 and 8 months, customer age is greater than or equal to 36 or not missing, and they are not under contract then the person is highly likely to change to 4G service (prob = .944). However, this only accounts for 9 people so this is not a group that I would recommend targeting.

The second rule is more promising. This rule indicates that persons with handsets less than 1 month and are single instead of married are more likely to change to 4G service. This rule is based on the largest count of people and would be relatively stable. Even if a person is married, they are more likely to switch if their handset is new as evidenced by the 3rd rule.

Figure 3: Leaf Report Showing Validation Probability of Service Type

Leaf Label of Untitled 6	Probability of 3G Service	3G Count	Probability of 4G Service	4G Count
HS_AGE>=6&AGE>=36 not Missing&HS_AGE<8&CONTRACT_FLAG(0)	5.6%	0	94.4%	9
HS_AGE<6&HS_AGE<1&MARITAL_STATUS(S)	8.4%	46	91.6%	506
HS_AGE<6&HS_AGE<1&MARITAL_STATUS(, M)	20.7%	83	79.3%	320
HS_AGE<6&HS_AGE>=1&AGE<43 or Missing	29.5%	206	70.5%	492
HS_AGE>=6&AGE<36 or Missing&HS_AGE>=8&SUBPLAN<2105&SUBPLAN>=2102	33.2%	40	66.8%	81
HS_AGE>=6&AGE<36 or Missing&HS_AGE<8	46.7%	111	53.3%	127
HS_AGE>=6&AGE>=36 not Missing&HS_AGE>=8&SUBPLAN<2105&SUBPLAN>=2102	47.1%	38	52.9%	43
HS_AGE<6&HS_AGE>=1&AGE>=43 not Missing	51.6%	112	48.4%	105
HS_AGE>=6&AGE>=36 not Missing&HS_AGE<8&CONTRACT_FLAG(1)	64.9%	115	35.1%	62
HS_AGE>=6&AGE<36 or Missing&HS_AGE>=8&SUBPLAN>=2105&SUBPLAN>=2120	67.4%	410	32.6%	198
HS_AGE>=6&AGE<36 or Missing&HS_AGE>=8&SUBPLAN<2105&SUBPLAN<2102	74.0%	106	26.0%	37
HS_AGE>=6&AGE>=36 not Missing&HS_AGE>=8&SUBPLAN<2105&SUBPLAN<2102	83.2%	135	16.8%	27
HS_AGE>=6&AGE<36 or Missing&HS_AGE>=8&SUBPLAN>=2105&SUBPLAN<2120	88.1%	106	11.9%	14
HS_AGE>=6&AGE>=36 not Missing&HS_AGE>=8&SUBPLAN>=2105	88.2%	592	11.8%	79