Question 1

(38 pts)

1. (4 pts) Import the customer churn data set that has been provided in D2L. You will also find a Word doc that contains the description of the variables in the data set. Upon a successful import you should see 14 variables and 1000 observations on the R environment pane.

A screenshot of a computer

Description automatically generated

2. (4 pts) Tabulate the outcome variable Churn. What percentage of customers in the sample have switched vendors?

A close up of a computer screen

Description automatically generated

The percentage of customers in the sample that have switched vendors is 25.9%.

3. (4 pts) Use the set.seed() command to set the starting value to 123. This ensures that any random

sample generation starts at a pre-determined value (in this case from 123), causing R to generate

exactly the same random sample every time your code runs. This is very useful to ensure consistent

results during model building.

4. (8 pts) Once you set the seed, split the data set randomly into two parts: a training data set, consisting of 70% of the observations, and a testing data set, consisting of 30% of the observations. As in class, use the sample.split() function in the caTools library to do this:

sample < − sample.split(d$Churn, SplitRatio=0.70)

You should end up with exactly 700 observations for the training set and 300 for the testing set.

5. (6 pts) Estimate a full decision tree model using the training data set with Churn as the categorical dependent variable. Include all the independent variables in the data set as features/predictors. Please note that since this is a full decision tree model you need to set the complexity parameter cp to -1, allowing the algorithm to freely grow the decision tree as it sees fit.



6. (4 pts) What is the number of splits in the full decision tree? You can use the summary() command to get this value.

The number of splits in the full decision tree is 42.

A screenshot of a computer error

Description automatically generated

7. (8 pts) What is the variable importance in the full decision tree? You can use the summary()

command to get the ordered list.

A close-up of a screen

Description automatically generated

Variable importance ordered list

1. Total Charges
2. Term
3. Monthly Charges
4. Agreement period
5. Technical support
6. Streaming Videos
7. Internet service
8. International plan
9. Marital Status
10. Multiple line
11. Voice mail plan
12. Phone service

Question 2 (30 pts) The full decision tree is a good starting point but, due to its complicated structure, it is not super helpful as an interpretable model for decision making. This is known as the overfitting problem. Because of this, it is almost always the case that a full decision tree has to be fine tuned into a simpler and more interpretable form. This process is called pruning. In this question we will do a pruning exercise using the maxdepth, minsplit, and minbucket parameters.

1. (8 pts) Estimate a pruned decision tree model using the training data set with the following parameters: maxdepth = 3 minsplit = 2 minbucket = 2 (1)

2. (8 pts) Use rpart.plot() to plot the pruned decision tree that you estimated. Please include a copy- /screenshot of the plot in your Word doc submission.



3. (4 pts) What is the height of the estimated tree?

the height of the estimated tree is 4

4. (4 pts) Which feature is used as the highest ranking feature for the churn decision?

The highest ranking feature for the churn decision is: Agreement\_period

5. (6 pts) Which variables were used in the tree construction? You can answer this either by looking at the plot or by using the printcp() command.

Variables actually used in tree construction:

[1] Agreement\_period International\_plan Internet\_service Marital\_Status

[5] Monthly\_Charges Multiple\_line Sex Streaming\_Videos

[9] Technical\_support Term Total\_Charges Voice\_mail\_plan

Question 3

Now that we have an interpretable model lets use it to calculate some statistics and identify a group of high risk customers to target with a marketing campaign. We will use the testing data set that we had set aside before for this purpose.

1. (6 pts) Use the pruned decision tree model to predict the churn outcomes for the customers in the testing data set.

A number grid with numbers

Description automatically generated with medium confidence

2. (6 pts) Display the confusion matrix/classification table for the pruned model.

A blue and white text

Description automatically generated

3. (6 pts) Based on the classification table calculate the accuracy rate for the pruned model.

"The accuracy rate for the pruned model is: 0.79"

4. (6 pts) Now, again using the pruned model, predict the churn probabilities for the customers in the testing data set.

A number of numbers on a white background

Description automatically generated

5. (6 pts) If the company decides to target the customers (in the testing data set) who have a greater than 70% probability of churning, how many customers would be in the target group? (an easy way to answer this would be to sort the probabilities in descending order)

"The number of customers in the target group is: 24"

Question 4

Finally, lets see if we can do better than our pruned decision tree model if we are to go with a more advanced algorithm like a Random Forest. In R, we will utilize the randomForest library to do this.

1. (8 pts) Estimate a random forest model for the training data using 5000 trees. Limit the mtry parameter to 3 to ensure that the algorithm randomizes over 3 different features/variables during each random tree formation. Please note that if you directly include the Churn variable as randomForest(Churn ∼ ., data=train data, ntree = 5000, mtry = 3, importance = TRUE) R will complain about it saying that you have given it a ”non-numeric argument to binary operator”. A quick workaround is to provide the Churn variable as a factor: randomForest(as.factor(Churn) ∼ ., data=train data, ntree = 5000, mtry = 3, importance = TRUE)

2. (4 pts) According to the random forest model you estimated, what are the top 4 important features based on the mean decrease that they provide to the Gini index?



Monthly\_Charges, Total\_Charges, Term, Agreement\_Period

3. (6 pts) Use the random forest model to predict the churn outcomes for the customers in the testing data set.

A number and text on a white background

Description automatically generated

4. (6 pts) Display the confusion matrix/classification table for the random forest model.

A white background with black text

Description automatically generated

5. (6 pts) Based on the classification table calculate the accuracy rate for the random forest model.

"The accuracy rate for the random forest model is: 0.82"

6. (6 pts) Now, again using the random forest model, predict the churn probabilities for the customers in the testing data set.

A screenshot of a computer

Description automatically generated

7. (6 pts) If the company decides to target the customers (in the testing data set) who have a greater than 70% probability of churning, how many customers would be in the target group?

"The number of customers in the target group is: 29"

8. (4 pts) Compared to the pruned decision tree model does the random forest model allocate more or less customers to the high risk group?

Yes, random forest model allocates 29 customers whereas the pruned decision tree allocates 24 customers.