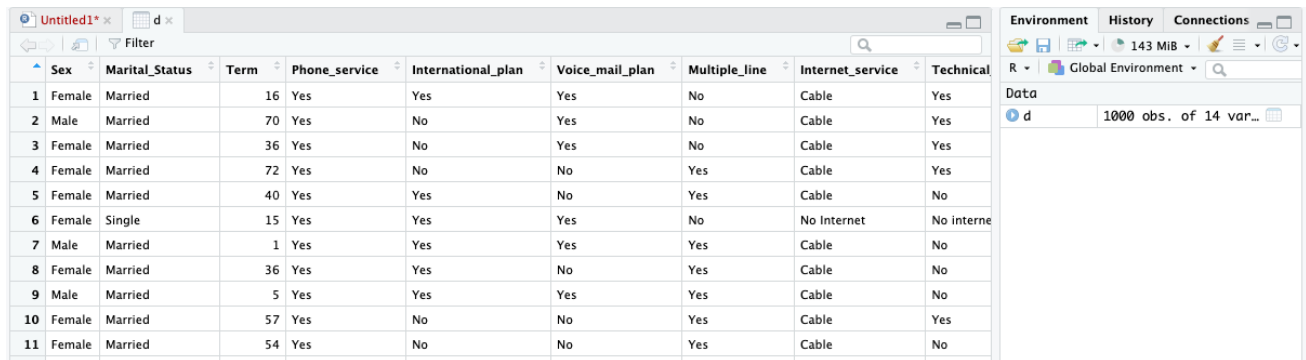


Question 1:

1.1 Import the customer churn data set that has been provided in D2L.



	Sex	Marital_Status	Term	Phone_service	International_plan	Voice_mail_plan	Multiple_line	Internet_service	Technical
1	Female	Married	16	Yes	Yes	Yes	No	Cable	Yes
2	Male	Married	70	Yes	No	Yes	No	Cable	Yes
3	Female	Married	36	Yes	No	Yes	No	Cable	Yes
4	Female	Married	72	Yes	No	No	Yes	Cable	Yes
5	Female	Married	40	Yes	Yes	No	Yes	Cable	No
6	Female	Single	15	Yes	Yes	Yes	No	No Internet	No internet
7	Male	Married	1	Yes	Yes	Yes	Yes	Cable	No
8	Female	Married	36	Yes	Yes	No	Yes	Cable	No
9	Male	Married	5	Yes	Yes	Yes	Yes	Cable	No
10	Female	Married	57	Yes	No	No	Yes	Cable	Yes
11	Female	Married	54	Yes	No	No	Yes	Cable	No

1.2 Tabulate the outcome variable Churn. What percentage of customers in the sample have switched vendors?

The percentage of customers who switched vendors is 25.9%

1.3, 1.4 Use the `set.seed()` command to set the starting value to 123.

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.

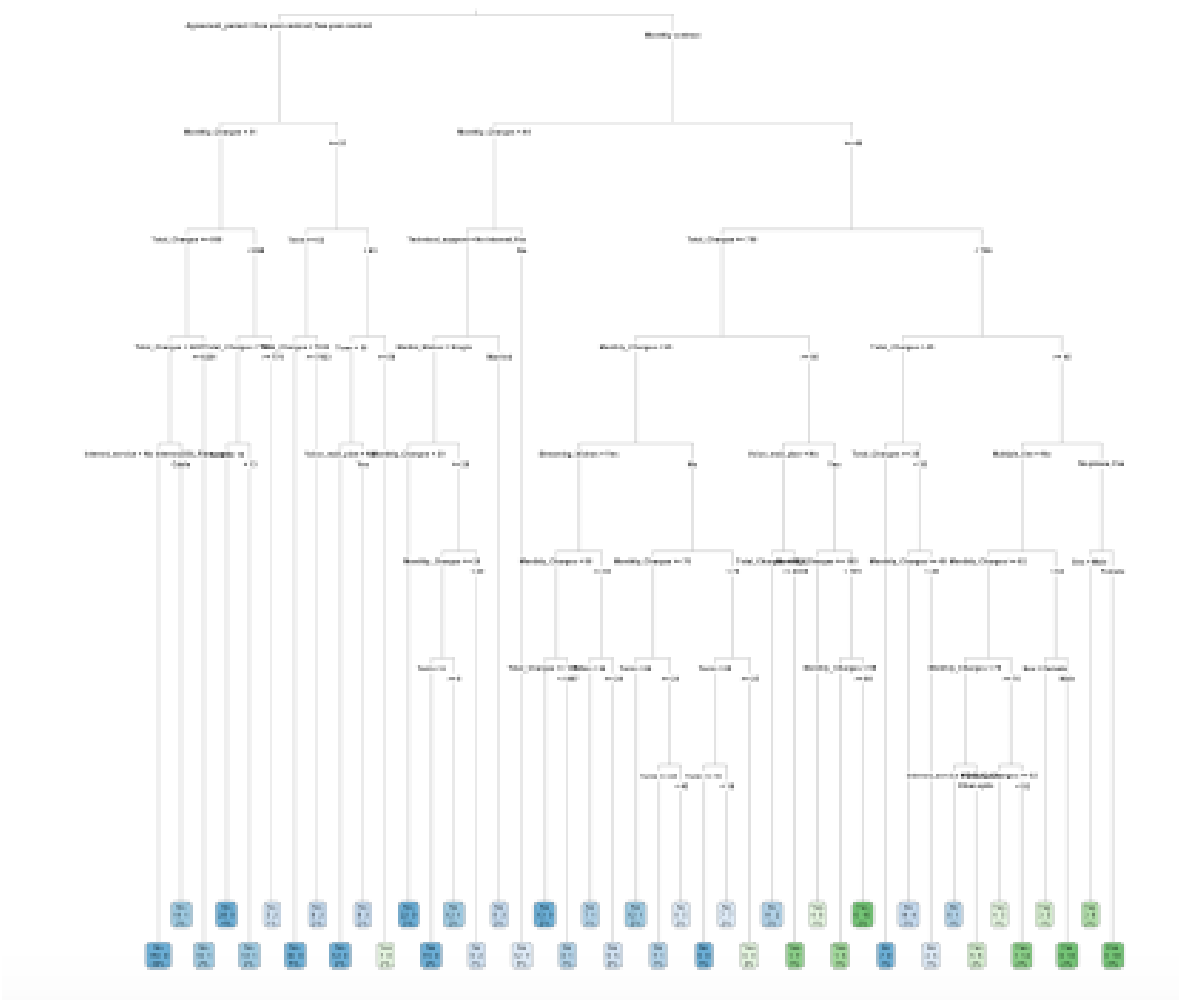
I've successfully created a training dataset consisting of 70% of the observations and a testing dataset consisting of 30% of the rest observations.

```
> dim(train_data) # Should be approximately 70% of the total data
[1] 700 14
> dim(test_data)  # Should be approximately 30% of the total data
[1] 300 14
```

1.5 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.

	CP	nsplit	rel error	xerror	xstd
1	0.066298343	0	1.0000000	1.0000000	0.06400227
2	0.024861878	5	0.5911602	0.6795580	0.05563049
3	0.006906077	7	0.5414365	0.6629834	0.05509054
4	0.005524862	11	0.5138122	0.7182320	0.05684359
5	0.001841621	14	0.4972376	0.7292818	0.05717860
6	0.000000000	17	0.4917127	0.7348066	0.05734423
7	-1.000000000	42	0.4917127	0.7348066	0.05734423

Decision Tree for Churn



1.6 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.

1.7 What is the variable importance in the full decision tree? You can use the summary() command to get the ordered list.

Variable importance			
Total_Charges	Term	Monthly_Charges	Agreement_period
19	17	15	13
Technical_support	Streaming_Videos	Internet_service	Marital_Status
10	8	8	4
Multiple_line	Phone_service	Voice_mail_plan	Sex
2	2	2	1
International_plan			
1			

From the summary result, we have the following top features:

Total_Charges: 19

Term: 17

Monthly Charges: 15

As well as some less important features:

Sex, International_plan: 1

Phone_service, Voice_mail_plan, Multiple_line: 2

Question 2:

2.1 Estimate a pruned decision tree model using the training data set with the following parameters:

maxdepth = 3

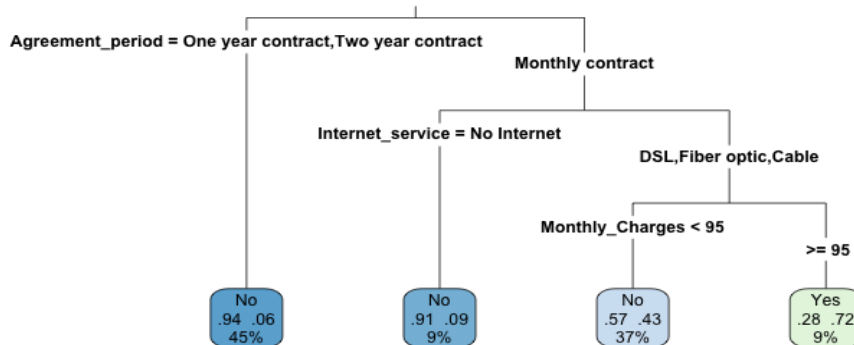
minsplit = 2

minbucket = 2

	CP	nsplit	rel error	xerror	xstd
1	0.05156538	0	1.0000000	1.0000000	0.06400227
2	0.01000000	3	0.8453039	0.9779006	0.06353454

2.2 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.

Pruned Decision Tree (Training Dataset)



2.3 What is the height of the estimated tree?

The height of the estimated tree is 3.

2.4 Which feature is used as the highest ranking feature for the churn decision?

From the summary, agreement period is used as the highest-ranking feature for the churn decision.

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

From the plot, the following variables were used in constructing the decision tree:

- Agreement period: this variable was used in the root node to separate contracts into “One year contract” and “Two-year contract” vs. “Monthly contract”.
- Internet service: this variable was used in the 2nd level under “Monthly agreement” to split nodes based on whether charges No Internet vs. DSL, Fiber optic and Cable.
- Monthly charges: this variable is used on 3rd level (under monthly charges) to split nodes based on whether charges are ≥ 95 or < 95 .

Question 3:

3.1, 3.2 Use the pruned decision tree model to predict the churn outcomes for the customers in the testing data set. Display the confusion matrix/classification table for the pruned model.

Classification table:

```
> class_table
      testing_outcome_preds
      No Yes
No    218  4
Yes    58 20
```

3.3 Based on the classification table calculate the accuracy rate for the pruned model.

The accuracy percentage of the pruned table is 79.33%

3.4 Now, again using the pruned model, predict the churn probabilities for the customers in the testing data set.

Below is a few first probability predictions:

	No	Yes
1	0.9426752	0.05732484
2	0.5697674	0.43023256
3	0.9426752	0.05732484
4	0.9062500	0.09375000
5	0.5697674	0.43023256
6	0.9426752	0.05732484
7	0.2812500	0.71875000

3.5 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.

```
[1] "Number of customers in the target group:"
> print(high_risk_count)
[1] 24
```

Question 4

4.1 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)
```

```

      OOB estimate of  error rate: 15.57%
Confusion matrix:
      No Yes class.error
No  484  35  0.06743738
Yes  74 107  0.40883978

```

4.2 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?

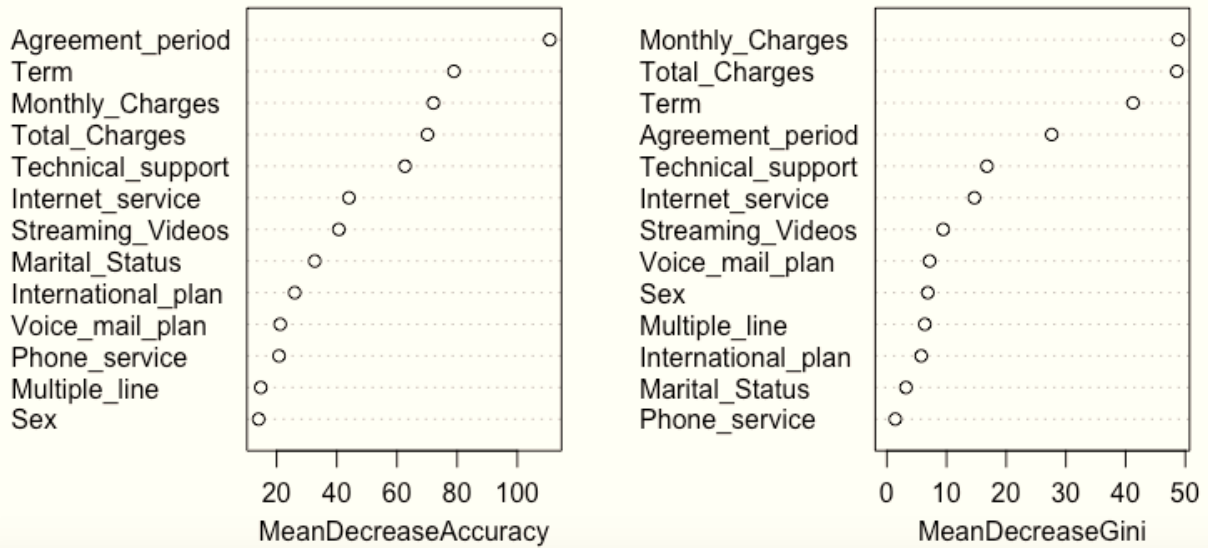
```

> # Check variable importance
> importance(rf_model)

```

	No	Yes	MeanDecreaseAccuracy	MeanDecreaseGini
Sex	2.352268	19.160833	14.01599	6.863892
Marital_Status	24.871234	13.461802	32.66338	3.208798
Term	47.583636	51.663790	78.92037	41.253619
Phone_service	18.407642	6.001359	20.79435	1.409227
International_plan	10.319503	26.820516	26.00716	5.759570
Voice_mail_plan	15.273104	14.916601	21.17254	7.169351
Multiple_line	7.806648	12.005386	14.71930	6.364139
Internet_service	23.828277	39.869342	44.02924	14.672925
Technical_support	25.225084	75.182479	62.65735	16.768837
Streaming_Videos	23.784702	35.487946	40.66856	9.442823
Agreement_period	53.915005	113.741382	110.78453	27.611305
Monthly_Charges	36.887714	65.682682	72.10239	48.771916
Total_Charges	50.652880	31.579841	70.11513	48.580144

rf_model



Based on the returned Mean Gini index, the top 4 important features are:

Monthly_Charges: 48.771916

Total_Charges: 48.580144

Term: 41.253619

Agreement_period: 27.611305

4.3 Use the random forest model to predict the churn outcomes for the customers in the testing data set.

A few first predictions of churn outcomes for the customers in the testing data set:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
No	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	Yes	No
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
No	No	No	No	No	No	No	No	No	No	Yes	Yes	No	No	No	Yes	Yes	No	No
39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	Yes	No	No	Yes	Yes

4.4 Display the confusion matrix/classification table for the random forest model.

Classification table:

	rf_outcome_preds	
	No	Yes
No	201	21
Yes	33	45

4.5 Based on the classification table calculate the accuracy rate for the random forest model.

Accuracy percentage of the random forest model is 82%

```
[1] "Accuracy percentage of the random forest model:"  
> print(rf_accuracy)  
[1] 82
```

4.6 Now, again using the random forest model, predict the churn probabilities for the customers in the testing data set.

Below is a few first probability predictions:

```
> rf_prob_preds  
      No      Yes  
1  0.9060 0.0940  
2  0.7498 0.2502  
3  0.8852 0.1148  
4  0.9940 0.0060  
5  0.7230 0.2770  
6  0.7252 0.2748  
7  0.3084 0.6916  
8  0.2898 0.7102  
9  0.9452 0.0548  
10 0.8674 0.1326
```

4.7 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.

```
[1] "Number of customers in the target group:"  
> print(high_risk_count2)  
[1] 29  
.
```

4.8 Compared to the pruned decision tree model does the random forest model allocate more or less customers to the high risk group?

Compared to the pruned decision tree model, the random forest model allocates more customers to the high-risk group ($29 > 24$).