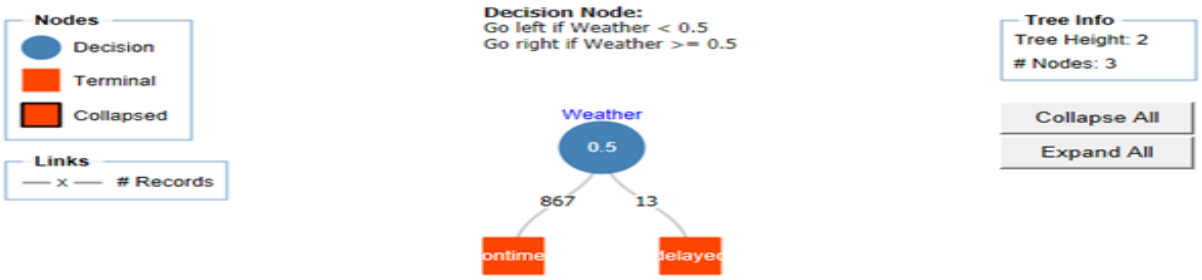
Flight Delays

Data Preprocessing. By creating dummies for day of week, carrier, departure airport, and arrival airport, It will give you 17 dummies. Bin the scheduled departure time into eight bins (in XLMiner use Transform a Bin Continuous Data and select equal width). After binning CRS DEP TIME into the 8 bins, this new variable should be broken down into dummies (because the effect will not be linear, due to the morning and afternoon rush hours). This will avoid treating the departure time as a continuous predictor, since it is reasonable that delays are related to rush-hour times. Partition the data into training and validation sets.

Here we preprocess the data by creating dummies for the mentioned variables above. We get 21 dummies. To remove the redundant information we created only 2 dummies for DEST variable keeping JFK as reference (i.e., if both EWR and LGA are zeroes in the table then we consider that as JFK). Similarly we do that with ORIGIN by keeping IAD as reference , for the DAY_WEEK with Sunday(Value =7) as reference and for CARRIER we keep US Airways as reference. So we get 2 dummies for DEST , 2 dummies for ORIGIN, 7 dummies for CARRIER and 6 dummies for Day of week. After this we bin the departure time into 8 bins and further add 7 more dummies to our data. So we will have a total of 24 dummies in our data. The final data can be seen in the sheet Final Encoding and partitioned data can be seen in the STD Partition sheet

a. Fit a classification tree to the flight delay variable using all the relevant predictors. Do not include DEP_TIME (actual departure time) in the model because it is unknown at the time of prediction (unless we are generating our predictions of delays after the plane takes off, which is unlikely). In the third step of the Classification Tree menu, choose "Maximum # levels to be displayed = 6." Use the best-pruned tree, setting the minimum number of observations in the final nodes to 1. Express the resulting tree as a set of rules.



Answers:

The above tree is the result of building a best pruned tree with the given FlightDelays dataset. The decision nodes Weather is the split value which is 0.5 and the rule given by this classification tree is:

IF(Weather<0.5) go left (ontime).

IF(Weather>=0.5) go right (delayed)

Training: Classification Summary

Confusion Matrix

Actual\Predicted	delayed	ontime
delayed	19	227
ontime	0	1075

Error Report

Class	# Cases	# Errors	% Error
delayed	246	227	92.27642276
ontime	1075	0	0
Overall	1321	227	17.18395155

Metrics

Metric	Value
Accuracy (#correct)	1094
Accuracy (%correct	82.81604845
Specificity	0.077235772
Sensitivity (Recall)	1
Precision	0.825652842
F1 score	0.904501472
Success Class	ontime
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	delayed	ontime
delayed	13	169
ontime	0	698

Error Report

Class	# Cases	# Errors	% Error
delayed	182	169	92.85714286
ontime	698	0	0
Overall	880	169	19.20454545

Metrics

Metric	Value
Accuracy (#correct)	711
Accuracy (%correct	80.79545455
Specificity	0.071428571
Sensitivity (Recall)	1
Precision	0.805074971
F1 score	0.89201278
Success Class	ontime
Success Probability	0.5

The above pictures show the error and accuracy of the training and validation datasets. The overall error is 17.18% for training and 19.2045% for validation datasets respectively. The accuracy of the training and validation models are seen as 82 and 80 % approximately.

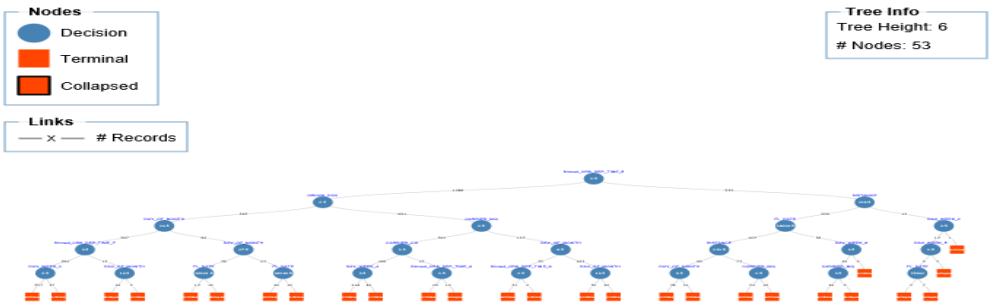
b. If you needed to fly between DCA and EWR on a Monday at 7 AM, would you be able to use this tree? What other information would you need? Is it available in practice? What information is redundant?

Answer:

No, we cannot use the above tree to use for the data to fly between DCA and EWR at 7 AM . We need to have information about weather to use the tree formed above to get the information on the flight status. The tree would be of use, practically, if we get the weather information too. The information of destination (EWR), day and time is redundant.

c. Fit another tree, this time excluding the Weather predictor. (Why?) Select the option of seeing both the full tree and the best-pruned tree. You will find that the best-pruned tree contains a single terminal node.

Fully Grown Tree



.

Best Prune Tree:

The above two trees doesn't have weather data available, so the best pruned data only has single terminal node which says that all the flights are on time. As Weather was the best predictor for the flight delays previously shown on the model built in (a), removing that gives us only the terminal node on time which says that without weather data available all the flights will be on time.

I. How is this tree used for classification? (What is the rule for classifying?)

Answer: The tree used here is a 6 level tree. There are about 26 classification nodes in the tree. The rule of less than or equal/greater than the split value is used over all the nodes in the tree. For example, at the decision node of CARRIER_MQ, if the value is less than 0.5, it goes to the left child node of CARRIER_CO and if the value is more than/ equal to 0.5, it goes to the right child node of DAY_OF MONTH in the level 3 left middle part of the tree in the CT_FullTree_F&P sheet.

II. To what is this rule equivalent?

Answer - The above mentioned rule is situated at the left middle part of the tree in sheet CT_FullTree_F&P.

IF(CARRIER_MQ < 0.5) then go left to CARRIER_CO OR IF(CARRIER_MQ >0.5) then go right to DAY_OF_MONTH.

IF(CARRIER_CO <0.5) AND IF(DAY_WEEK4 <0.5) then ontime

IF(CARRIER_CO >0.5) AND (IF(Binned_CRS_DEP_TIME_6<0.5) then ontime ELSE IF(Binned_CRS_DEP_TIME_6 >0.6 then delayed))

III. Examine the full tree. What are the top three predictors according to this tree?

Answer - The tree is of 6 levels. There are in total 26 decision nodes. The rule of greater than/ equal to or less than 0.5 is used all over. The top three predictors of the tree are Binned_CRS_DEP_TIME_5, ORIGIN_DCA and DISTANCE. So we can say that if the Binned_CRS_DEP_TIME_5 says that the flight may be in the time slot 1400 to 1600 hours, flights starting from the airport DCA and Distance of the flight journey mainly decides whether the flight will be on time or not.

IV. Why, technically, does the pruned tree result in a tree with a single node?

Answer - Weather is the best predictor that we get in the best pruned tree we got from the sheet CT_BestTree_Pruned sheet, we have excluded the weather for predicting this tree so this is the reason why we got a single terminal node ‘ontime’.

V. What is the disadvantage of using the top levels of the full tree as opposed to the best-pruned tree?

Answer: Decision trees are used for building regression or classification models. Pruning is reducing the size of the tree, and it creates a simpler tree structure. Having a complex tree structure refers to the process that the training data is fitted too well because of which any negative impact of the test data brings down the performance of the whole model. If the top levels of the tree is used, that may result in generalization ability and lower precision, the pruned decision tree uses validation set to choose sub-node, that results in better predictive accuracy.

VI. Compare this general result to that from logistic regression in the example in Chapter 10. What are possible reasons for the classification tree’s failure to find a good predictive model?

- Training and validation scores for the classification tree

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	delayed	ontime
delayed	0	246
ontime	0	1075

Error Report			
Class	# Cases	# Errors	% Error
delayed	246	246	100
ontime	1075	0	0
Overall	1321	246	18.62225587

Metrics	
Metric	Value
Accuracy (#correct)	1075
Accuracy (%correct)	81.37774413
Specificity	0
Sensitivity (Recall)	1
Precision	0.813777441
F1 score	0.897328881
Success Class	ontime
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	delayed	ontime
delayed	0	182
ontime	0	698

Error Report			
Class	# Cases	# Errors	% Error
delayed	182	182	100
ontime	698	0	0
Overall	880	182	20.68181818

Metrics	
Metric	Value
Accuracy (#correct)	698
Accuracy (%correct)	79.31818182
Specificity	0
Sensitivity (Recall)	1
Precision	0.793181818
F1 score	0.884664132
Success Class	ontime
Success Probability	0.5

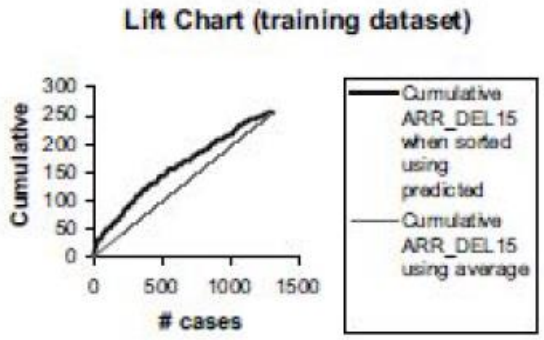
Training and Validation scores of Logistic Regression from Example in Chapter 10 Page no. 380 (from Data Mining for Business Intelligence_ Concepts, Techniques and Applications)

Training Data Scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	19	237
0	0	1065

Error Report			
Class	# Cases	# Errors	% Error
1	256	237	92.58
0	1065	0	0.00
Overall	1321	237	17.94



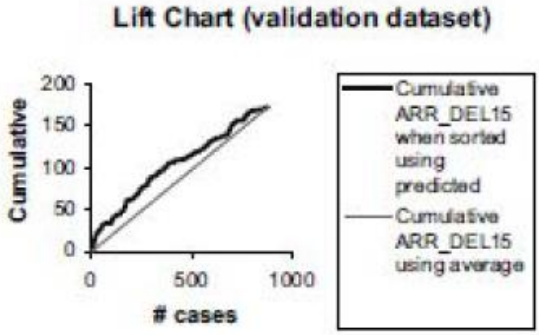
The table here represents 1 as delayed flight and 0 represents on-time flights.

Validation Data Scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	13	159
0	0	708

Error Report			
Class	# Cases	# Errors	% Error
1	172	159	92.44
0	708	0	0.00
Overall	880	159	18.07



The table here represents 1 as delayed flight and 0 represents on-time flights.

The training and validation scores for both classification tree and logistic regression are shown above. From the above results we can say that logistic regression has better training and validation scores when compared to that of the classification tree. Also, while comparing the accuracy of both the models we can say that logistic regression has 82.05% and 81.931% in training and validation whereas classification tree has 81.377% and 79.31% in training and validation respectively. We got the accuracy of logistic regression by using formaula TP+TN/(P+N), the data is taken from the confusion matrix as shown in the figures. So, we can say that logistic regression is better model when compared to the classification tree. When we compare the delayed error

percentage in classification tree. we can see that it has 100% error in both training and validation scores. Whereas in logistic regression we can see that the error in delayed is around 92% in both training and validation. From this we can say that the prediction error of classification tree is possible reason for failure to become a good prediction tree. Decision boundaries are generated differently by Trees and Logistic Regression. The boundaries to separate classes in tree bisects the space into smaller regions, whereas a single line divides the space for the two classes in Logistic Regression.