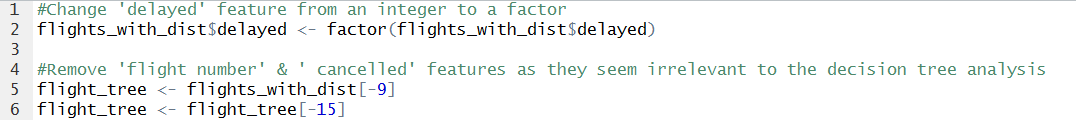
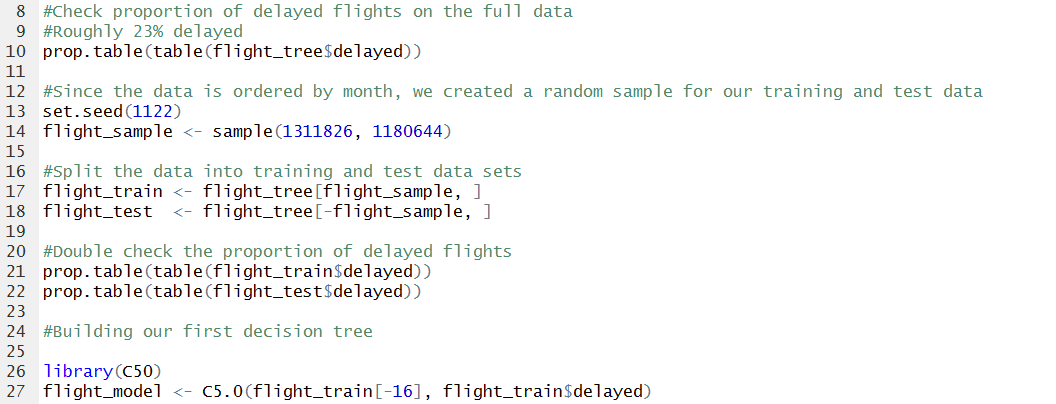
We decided to perform a decision tree analysis on the data to get an understanding of what causes a flight to be delayed. Since our created variable “delayed” is binary (an integer), we need to change this to a factor to run a decision tree. We also removed two variables, ‘flight number’ and ‘cancelled’ as they seemed irrelevant in determining a delayed flight. The final variables used for the decision tree analysis include: DayofMonth, DayOfWeek, DepTime, CRSDepTime, ArrTime, CRSArrTime, UniqueCarrier, ActualElapsedTime, CRSElapsed Time, DepDelay, Origin, Dest, Distance, and Diverted.

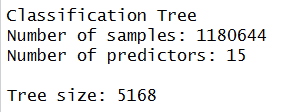


Before running the decision tree analysis, we can see that 22.7% of the total flights were delayed and the remaining 77.3% were on time. We’ll use this is a reference point for any future manipulation of the data. The data also appeared to be ordered by month, so we created random samples of the data before creating our training and test data to avoid any bias. The full data had roughly 1.3 million observations, so allocating 90% of the data to training and the remaining 10% to test results in data sets of size 1.18 million and 131,000 respectively. Checking the proportion of delayed flights within our training and test data still results in roughly 23% delayed for both, so we can move on to the analysis portion.



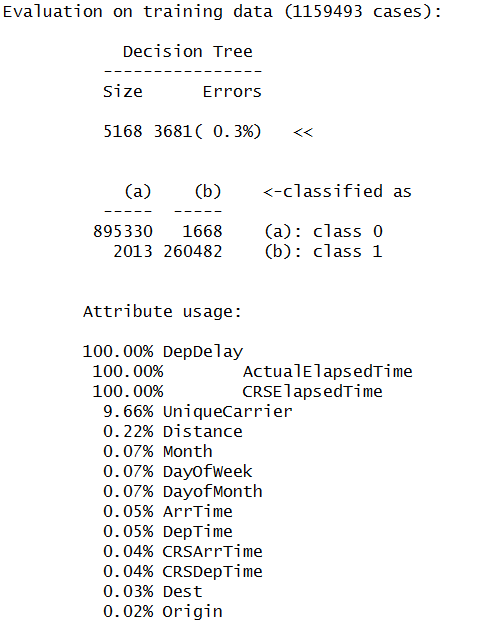
Since we are dealing with a fairly large data set that includes 15 predictor variables, the tree size ended up being 5,168.



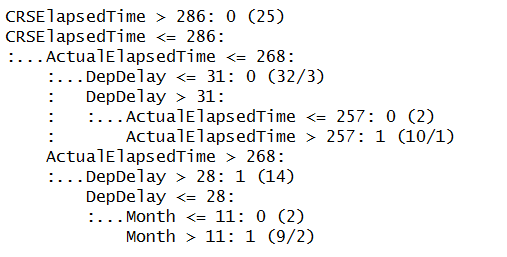


After running a summary on the model, the error rate on the training data was an extremely low 0.3%. Departure delay, actual & scheduled elapsed time also were the main contributors to whether or not a flight was delayed. Airline carrier also seemed to play an important role in the decision making process, more so than flight distance or the time of year.





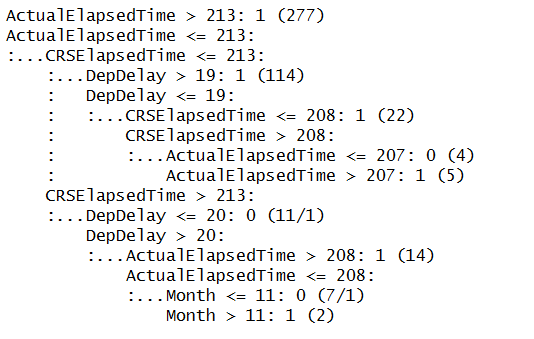
Since there ended up being 524 sub-trees and we cannot look at them all, here is a snippet of a few decision trees:



* If the scheduled elapsed time was greater than 286 minutes then the flight was unlikely to be delayed.
* But if the scheduled elapsed time was less than or equal to 286 minutes and …
* The actual elapsed time was less than or equal to 268 minutes and …
* The departure delay was greater than 31 minutes and…
* The actual elapsed time ended up being greater than 257 minutes then the flight was likely to be delayed.

This makes sense intuitively because if you have a scheduled 4hr 45min flight, with a 31 minute departure delay, and the pilot was only able to make up a maximum of 11 minutes of lost time in the air, this still meets our criteria of being a delayed flight (ie: at least a 15 minute arrival delay).

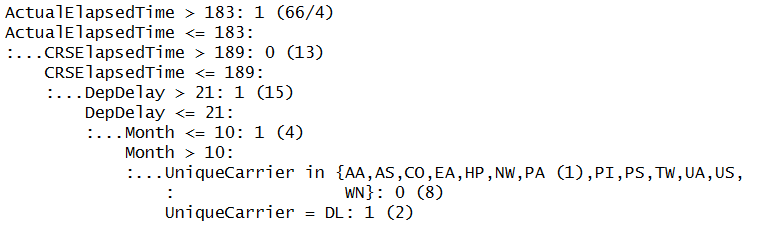
Here is a look at another sub-tree:



* If the actual elapsed time was less than or equal to 213 minutes and …
* The scheduled elapsed time was greater than 213 minutes and …
* The departure delay was greater than 20 minutes and …
* The actual elapsed time was less than or equal to 208 minutes and …
* The month was greater than 11 (ie: December) the flight was likely to be delayed.
* But if it were any other month of the year then the flight was unlikely to be delayed.

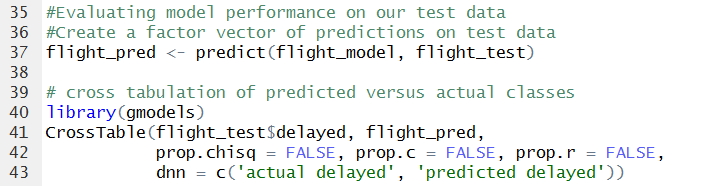
So, a December flight that had a departure delay of more than 20 minutes, with a scheduled time of more than 3hr 33min that actually took less than 3hr 28min was still likely to be delayed.

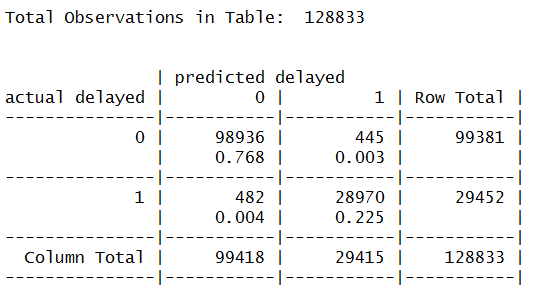
Last example of a different sub-tree, this time involving airline carriers:



* If the actual elapsed time is less than or equal to 183 minutes and …
* The scheduled elapsed time was less than or equal to 189 minutes and …
* The departure delay was less than or equal to 21 minutes and …
* The month was either November or December and …
* The carrier was Delta Air Lines then the flight was likely to be delayed.
* But, under the same conditions, if the airline carrier was one of 13 different carriers (ie: American Airlines, Alaska Airlines, Continental Air Lines, United Air Lines etc.) then it was unlikely to be delayed.

So now that we’ve looked at a few decision trees from the training data set, let’s see how our test data performs.





Out of nearly 130,000 observations, only 482 flights were actually delayed when predicted otherwise, and 445 predicted delayed when they were actually not delayed. These results are great given the error rate of only 0.7%. Granted the model performed a lot better on our training data set (only 0.3% error rate), but it is expected for the decision tree to perform worse on unseen data. Whether or not these results are ideal is up for debate. There are potentially millions of flights per year, so 0.7% would still account for a large chunk of incorrectly classified flights.