Assumptions for Data-Preprocessing

- Delay time is supposed to be within 15 minutes to decide whether the flight is on-time or delayed. Delay time can be calculated as subtraction scheduled departure time from actual time. When I examined the data, some of the flight are classified as delayed even if they are landing within the 15 minutes according to scheduled time. Hence, my assumption is possible misclassification of these rows when they are entered to dataset. I move on by turning these rows into on-time. This part is a bit controversial, then it can be turned into comment line.
- Flight number and tail number feature are not needed to examine. They are unique numbers to the flight; hence I cannot decide the flight status by looking at them.
- Even if all the carrier names are not written in the dataset, I believe that the carrier are related to flights delayed. I cannot predict about the carrier that are not in the dataset.
- Flight date and the days of the month are the same features, then one of them can be deleted, so I will be moving with the days of month feature.
- Actual time will not be given for future dataset. We want to predict the delays for the specific flight without looking at actual time of these variables. Hence, I deleted that feature.
- Scheduled departure time should be turned into minutes. It is currently given in the form of hour and minutes. Hence, they are in the form of minutes after processing.
- After turning flight status class labels into 1 and 2. 2 represents the ‘ontime’, 1 represents the ‘delayed’.
- Trees are obtained for Gini Index and Entropy separately.
- To fix the data separation, rng function is used.

PART A&B - Classification Tree with Gini Index

Classification tree after creating with the usage of gini index seems as follows. There are 11 prune level in this graph.

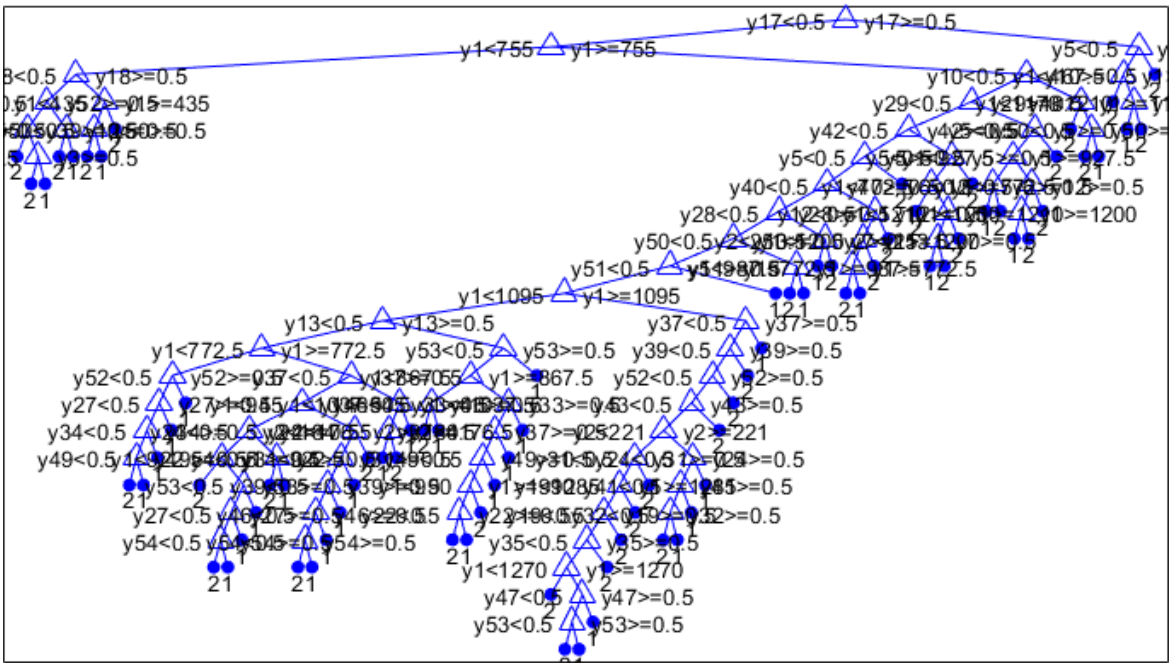


Figure 1: Classification Tree

Error rates for each pruned tree can be seen on the right. Y-coordinates show the error rates and the x-coordinates shows the pruned trees. 1 shows the tree with the 14 levels, then level is decreasing while x is increasing. The best error rate is 0.1114 and the pruning level is 4.

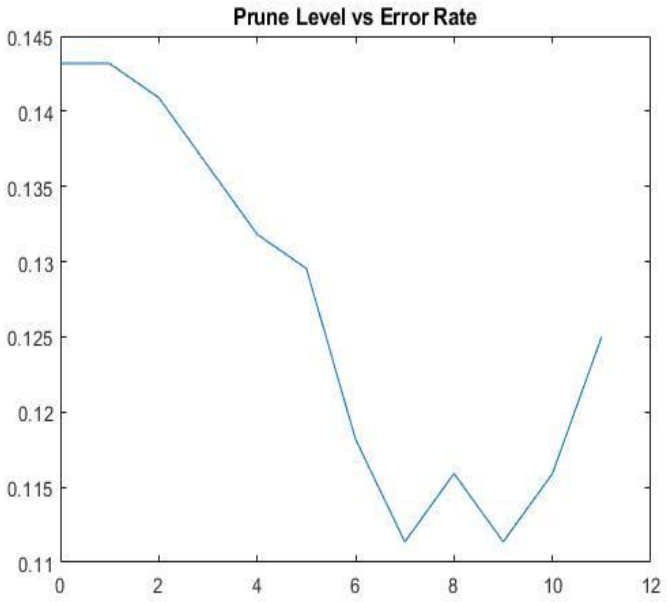


Figure 2: Pruning Level vs Error Rates

Error rate is firstly, declining, then increasing from some level.

The pruned tree can be seen at below. Pruning level is 7 as mentioned above.

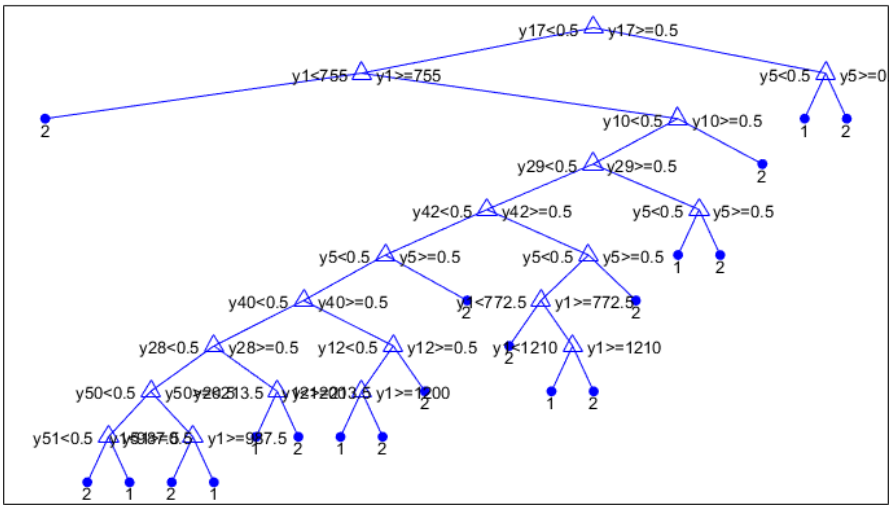


Figure 3: Pruned Tree

PART C - Classification Tree with Entropy

Classification tree after creating with the usage of entropy seems as follows. Pruned level for this tree is 14.

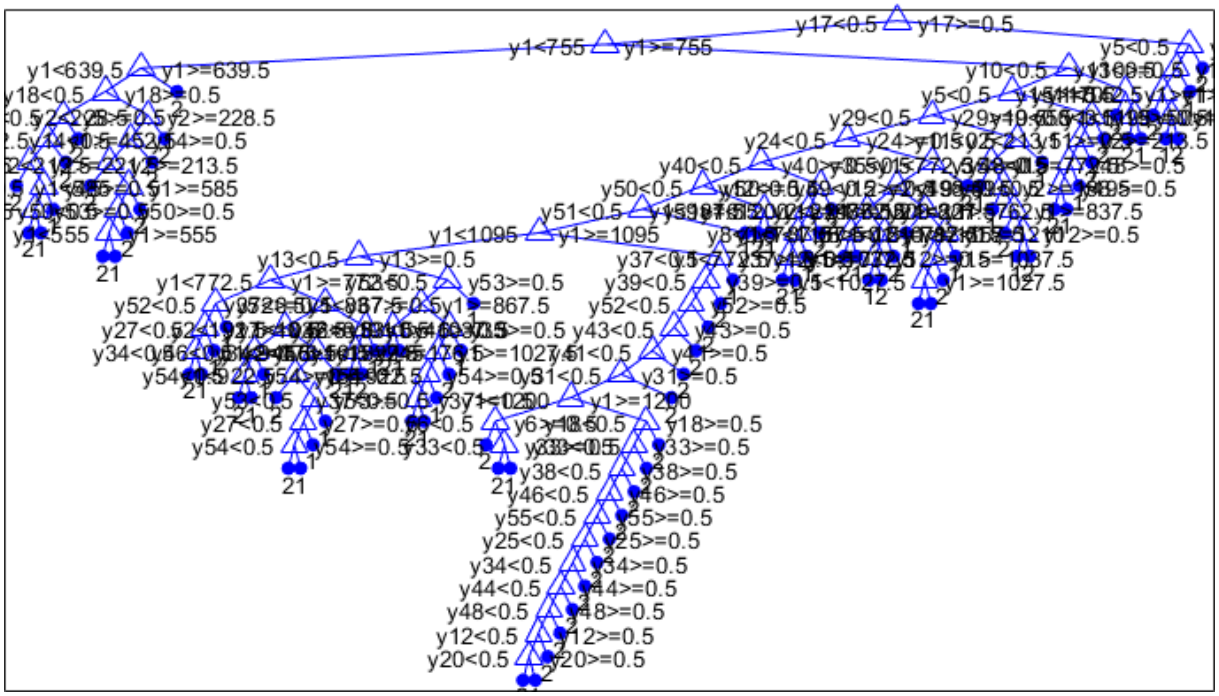


Figure 4: Classification Tree with Entropy

Error rates for each pruned tree can be seen on the right. Y-coordinates show the error rates and the x-coordinates shows the pruned trees. 1 shows the tree with the 14 levels, then level is decreasing while x is increasing. The best error rate is 0.1114 and the pruning level is 12.

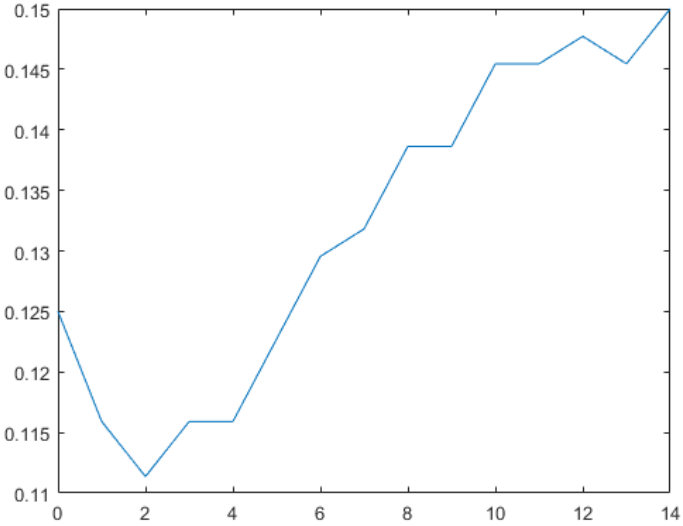


Figure 5: Pruning Level vs Error Rate

The pruned tree can be seen at below. Pruning level is 12 as mentioned above.

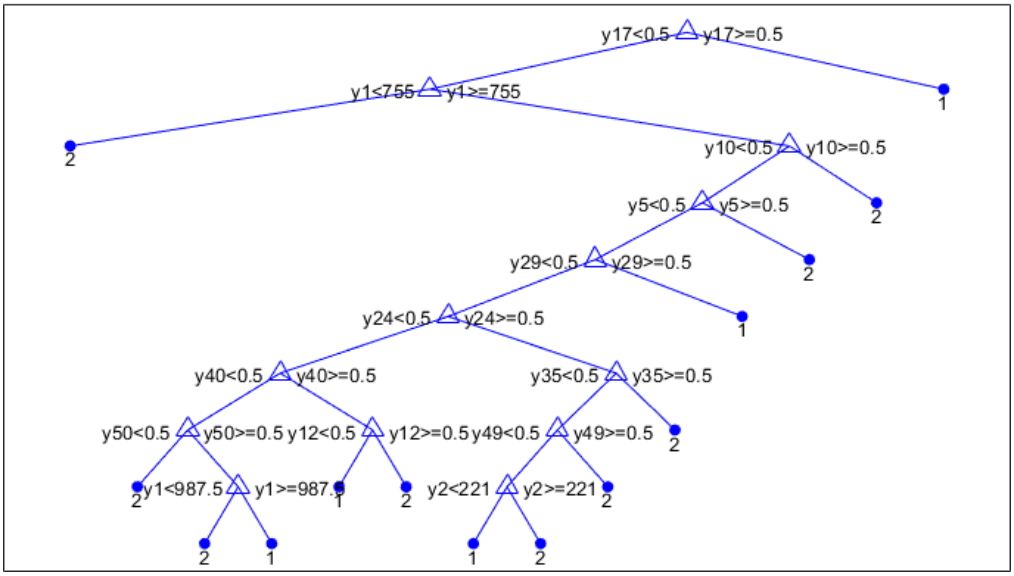


Figure 6: Pruned Tree with Entropy

To compare in the next part, the cost of misclassification is as follows:

Cost

Table 1: Cost Table

	Delayed	Ontime
Delayed	12	43
Ontime	6	379

Total Cost = 50*43+6*5 = 2180$

Part D

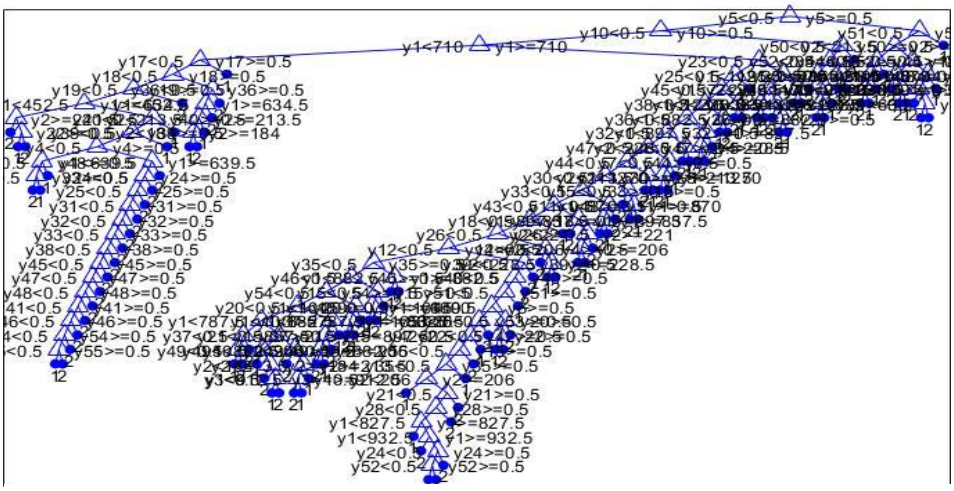
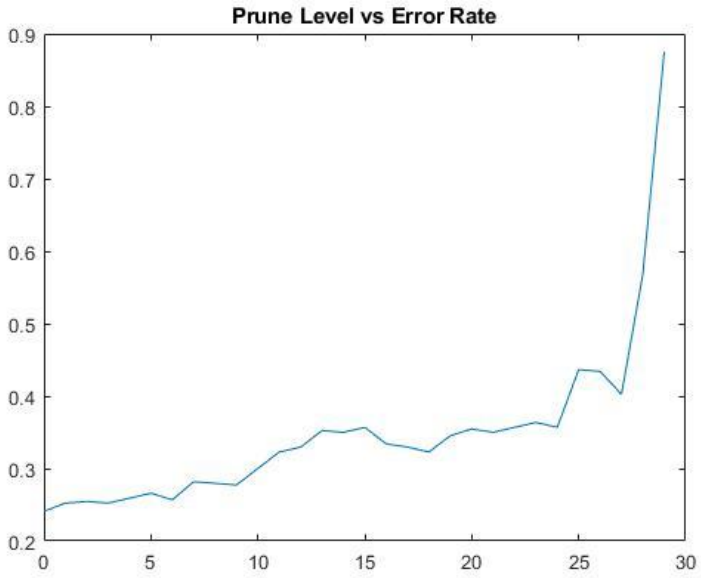


Figure 7: Classification Tree After Cost Implemented

When implemented the misclassification costs to the model with gini index, classification tree is shown as above.



Error rate is calculated as 0.2409. It is the least error among the rates after implemented the misclassification cost to the model.

Figure 8: Prunin Level vs Error Rate

Moreover, total cost of the model as follows:

Table 2: Cost Table 2

	Delayed	Ontime
Delayed	29	26
Ontime	80	305

Total Cost = 26*80+80*5 = 1700$

- Total cost is decreased after new cost entries are implemented.