**Homework Three \_ Team 21**

**Instruction:** Consider the scenario from homework 1, where we explored the dataset ToyotaCorolla. It has been pointed out that there is an extreme value in the CC variable – 16,000. This error has been corrected to 1600. The categorical Fuel-type variable has been converted into three dummy variables. The data has been divided into training (50%), validation (30%), and test (20%) datasets. The updated Toyota Corolla dataset, named PreToyotaCorollaII.xls, will be used for this homework. Create a new Word document and save it as HW3Answers\_X (where X is your team number). Where required, write your answers and/or paste screenshots of XLMiner results into this Word document. Your response should not exceed 100 words for each below question. Write every member’s full name and participation on the first page of the Word document as follows. You need to submit this Word document and Excel file with XLMiner solution.

|  |  |  |  |
| --- | --- | --- | --- |
| Participant | Complete the Assignment before the Meeting (Y/N) | Percentage of Contribution | Justification |
| Anchal Atlani | Y | 100 |  |
| Nikhil Jadhav | Y | 100 |  |
| Soham Dhodapkar | Y | 100 |  |
|  |  |  |  |

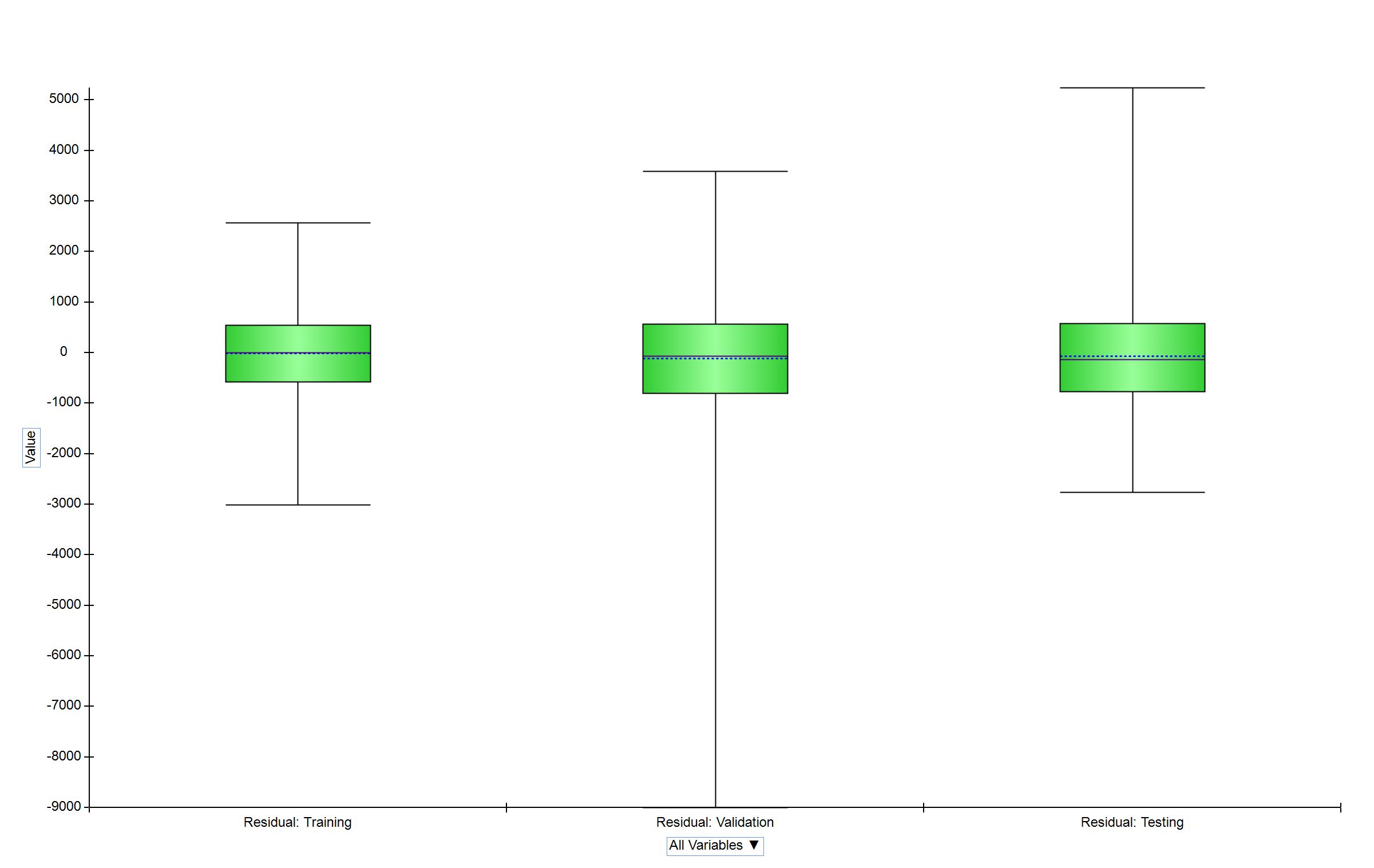
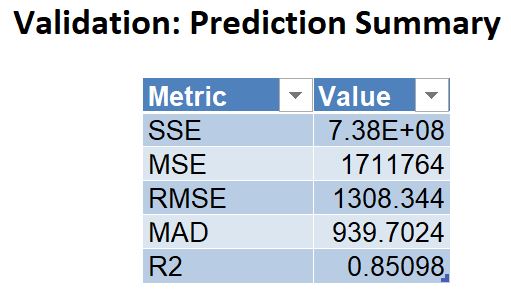
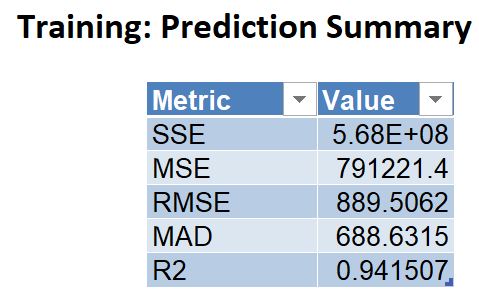
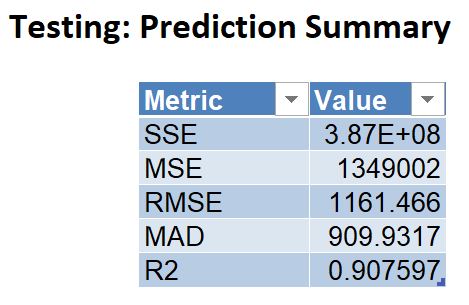
**Tasks:**

1. Run a regression tree (RT) using the Prediction menu in XLMiner with the output variable Price and input variables Age 08 04, KM, Fuel Type, HP, Automatic, Doors, Quarterly Tax, Mfg Guarantee, Guarantee Period, Airco, Automatic Airco, CD Player, Powered Windows, Sport Model, and Tow Bar. Keep the minimum number of records in a terminal node to 1, and the maximum number of tree levels to 10, and the scoring option to Full Tree, to make the run least restrictive.
2. (10 points) Which appear to be the three or four most important car speciﬁcations for predicting the car’s price? (Nikhil)

Ans. Following are the four important car specification:



1. Age\_08\_04
2. KM
3. HP
4. Quarterly\_Tax
5. (20 points) Compare the prediction errors of the training, validation, and test sets by examining their RMS error and by plotting the three boxplots. What is happening with the training set predictions? How does the predictive performance of the test set compare to the other two? Why does this occur? (Soham)



Ans: By looking at the train, test and validate errors in the tables above, we observe that the training error is quite low as compared to the validation error. This means that the model is overfitting the data. We have a good R2 value for the test set which means that the model is producing good accuracy nonetheless. The R2 value for validation set is lower than those of test and training set. This clearly indicates that the model fit too well on the training data and not so well on the validation data, i.e. overfitting. The value drops again for the test set, but not as much as for the validation which indicates that the model fit better on the test data than the validation data. Higher values of R2 and lower value of error indicate that the training cases were ‘easy’ cases to predict hence there was ‘no difficulty’ for the model to fit the data causing a high R2 value and a low error value. In contrast, the data in the validation set could be harder to generalize from or harder to fit the model according to the data that it learned from. Similarly, in the test cases, the model is fitting well but not as good as training data. Since test data is anyway holdout, the achieved R2 value is closer to the training value indicating that the model is fitting better on the training and testing samples than validation sample.

1. (10 points) How can we achieve predictions for the training set that are not equal to the actual prices?

Ans: As the level of the tree increases the overall error of the tree should decrease until the point of overfitting which usually happens in the case of training data. However, for new data, the overall error is expected to decrease until the point where the tree models the relationship between class and the predictors. After that, the tree starts to model the noise in the training set, and we expect the overall error for the validation set to start increasing. Two ways to try and avoid exceeding this level, thereby limiting overfitting, are by setting rules to stop tree growth, or alternatively, by pruning the full-grown tree back to a level where it does not overfit. A method like CHAID uses chi-square statistical test for independence to assess whether splitting node improves the purity significantly. The predictor which has the strongest association with the target is used splitting and the association is measured by the p-value of the chi-square test. Splitting stops when purity improvement is not statistically significant. \*\*

Another such method is pruning which has methods like CART and C4.5. In C4.5, the training data is used both for growing and pruning. In CART, the validation data is used to prune back the tree that is used from the training data. The idea behind pruning is that a very large tree is likely to be overfitting training data and the weakest branches which hardly reduce the error rate should be pruned thus increasing the performance and reducing the error. \*

1. (10 points) Create a best-pruned tree using the same data partitioning. If we used the full tree instead of the best-pruned tree to score the validation set, how would this affect the predictive performance for the validation set? (Hint: does the full tree use the validation data?) (Anchal)

Sheet name for validation score using Full Grown and Best-Pruned Tree

RT\_ValidationScore1- Best-Pruned Tree

RT\_ValidationScore- Full Grown Tree

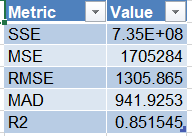
Ans: In Full Tree Pruning is not done on validation set. The “full” classification tree create decision node till it reaches the data which belong to only one class and there are no impurities, that means complete set belongs to only one class.

When we add decision nodes at first the classification error decreases on the validation set, too many decision nodes overfits the classification tree to the training data and results in increased error on the validation set.

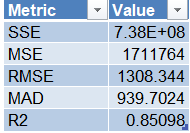
In the best pruned tree begins with the full classification tree and prune decision rules until the classification error on the validation set increases.

And as we can see from below the RMSE value of validation set of best Prune and Full Tee, the RMSE value decreased in best -pruned tree.

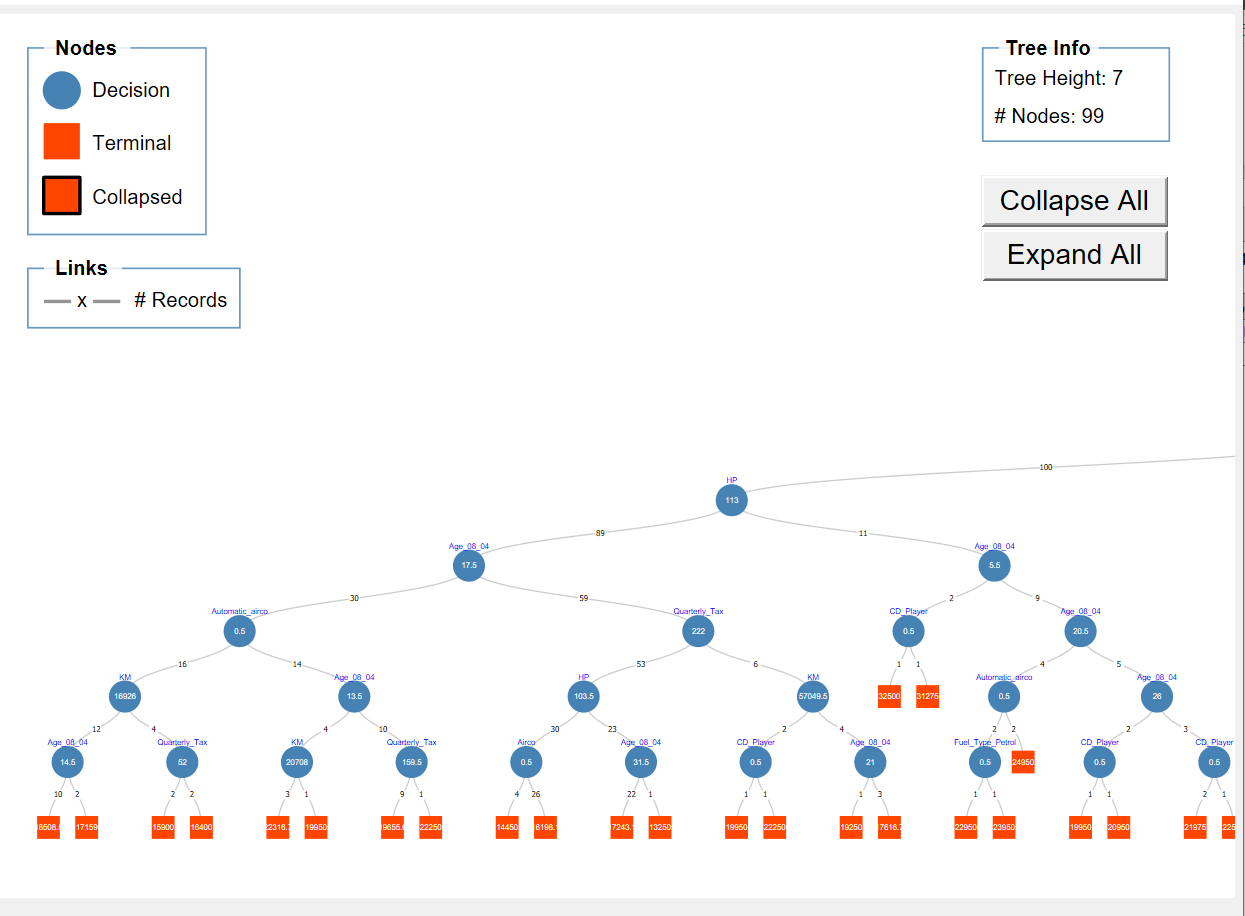
**Best-Pruned Validation score:**



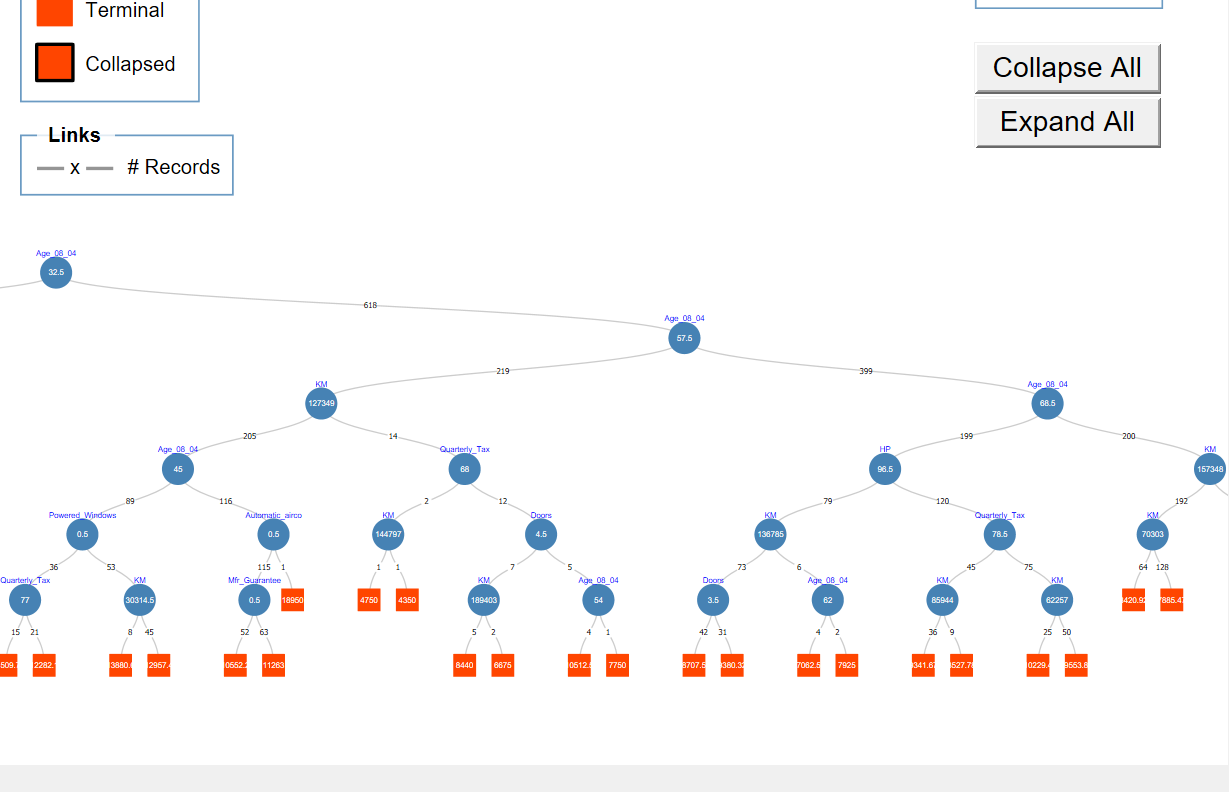
**Full Tree**



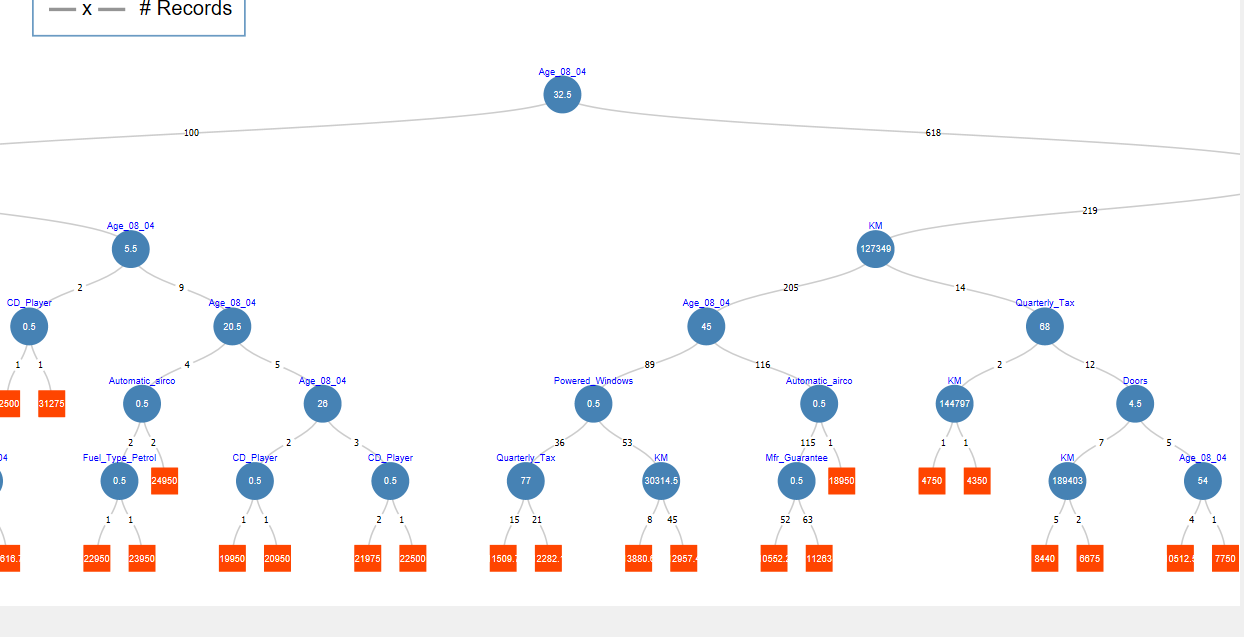
Left side of Full Tree:



Right side of Full Tree:

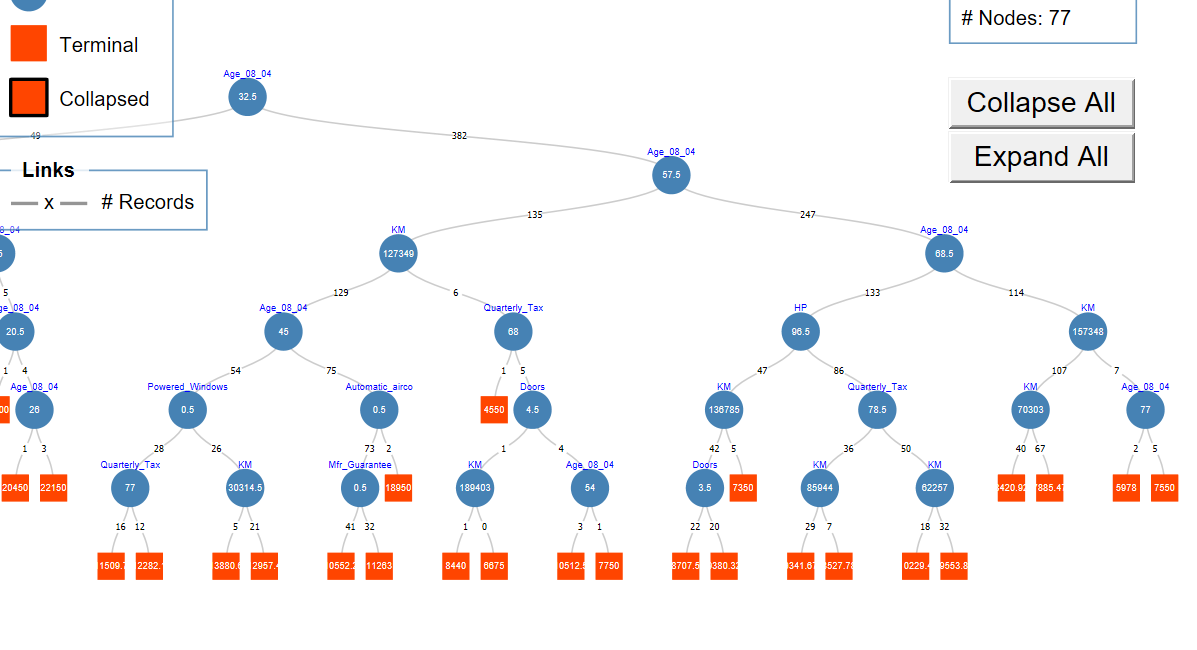


Middle of Full Tree:

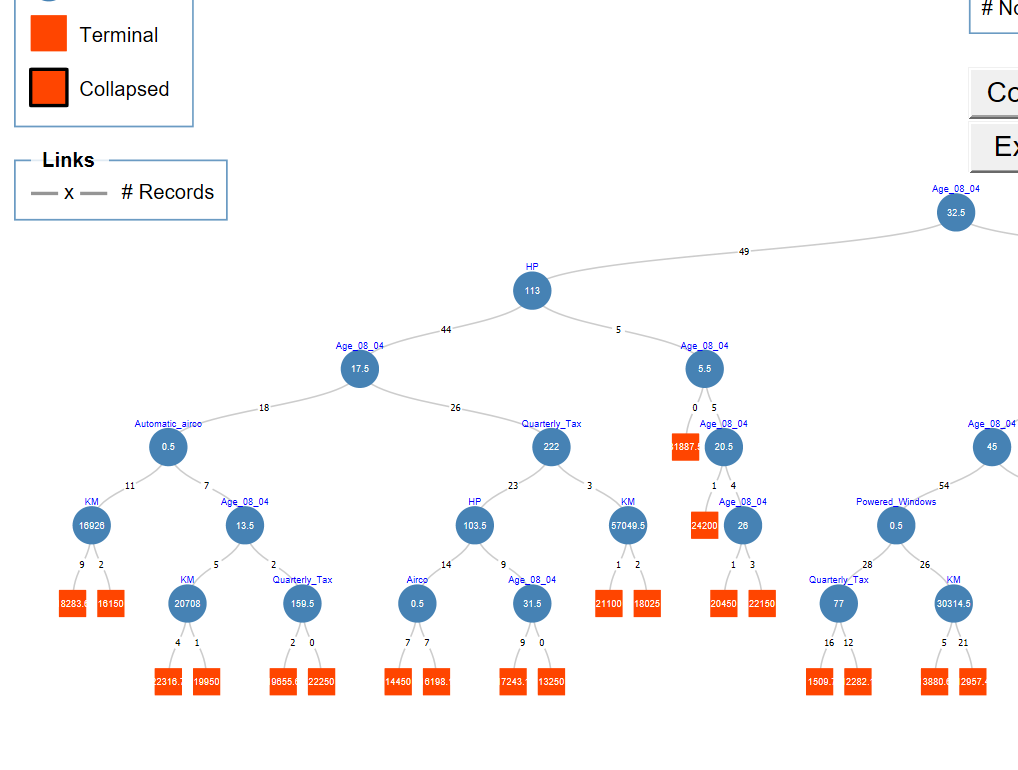


Best-Pruned Tree:

Right:



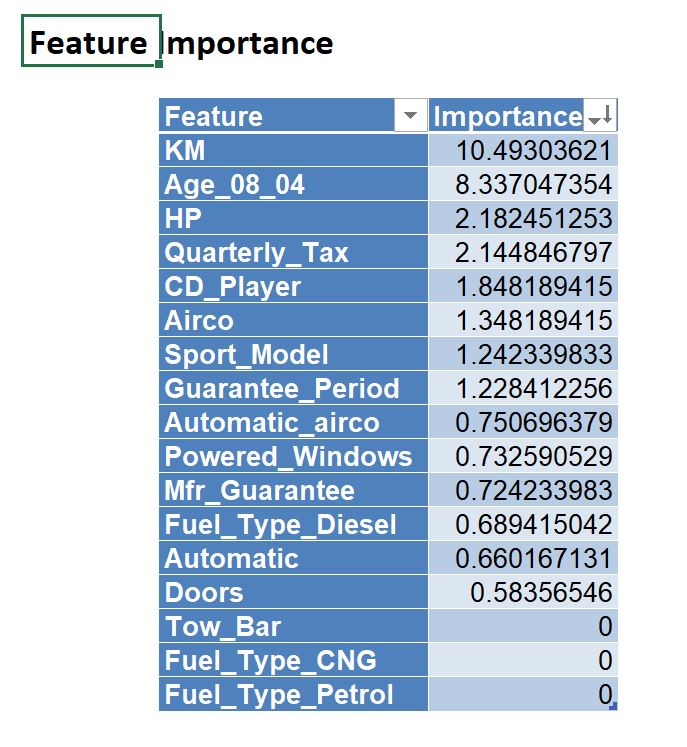
Left:



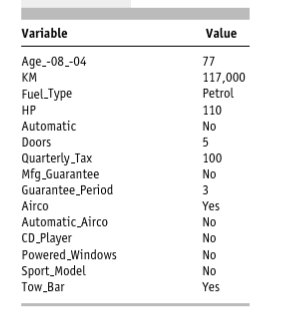
1. Let us see the effect of turning the price variable into a categorical variable. First, create a new variable that categorizes price into 20 bins from encoding data. Use Transform → Bin continuous data to categorize Price into 20 bins of equal counts (leave all other options at their default). Next, repartition the data keeping Binned Price instead of Price. Run a classiﬁcation tree (CT) using the Classiﬁcation menu of XLMiner with the same set of input variables as in the RT, and with Binned Price as the output variable. Keep the minimum number of records in a terminal node to 1 and uncheck the Prune Tree option, to make the run least restrictive.  
   1. (25 points) Compare the tree generated by the CT with the one generated by the RT. Are they different? (Look at structure, the top predictors, size of tree, etc.) Why? (Soham)

Ans - Both the trees are different, differing in their size and order of top predictors. The Regression Tree has more nodes than that of Classification Tree although the levels are same. This means that there are a greater number of splits in the Regression Tree than the Classification tree. This is because Classification and Regression trees use different error metrics for node splits. The cost function in regression predictive modelling is the sum of squared errors which fall in the rectangle. In classification, the Gini index function is used, which calculates the purity of the nodes. The regression tree has 99 nodes whereas the classification tree has 89 nodes even though both have 7 levels.

Also, KM, Age\_08\_04, HP, Quarterly\_Tax are the top predictors. These remain the same in both the cases of classification and regression tree, the order of Age and Km changes. (position 1 and 2)



These predictors remain the same even though the order changes. The classification has terminal nodes closer to the root than the regression tree, because of the difference in the error metric it uses.

* 1. (10 points) Predict the price, using the RT and the CT, of a used Toyota Corolla with the speciﬁcations listed in the following Table. (Nikhil) 

Bin 2-

Traversing the Regression Tree(RT) and the Classification Tree(CT) using the provided value we get the following prices. Using Classification tree we get **6900-7400** as the predicted price of the vehicle since the input data(training) is divided into bins in the classification tree the predicted price is also in the range of the bins given. Using the Regression Tree we get **7885.47** as the predicted price of the vehicle.

* 1. (15 points) Compare the predictions in terms of the predictors that were used, the magnitude of the difference between the two predictions, and the advantages and disadvantages of the two methods. (Anchal)

Advantages:

1. Simplicity:

In Both Tree we value can be predicted or categorized by following if and else rule, by traversing from root to bottom to find the correct terminal node. This method is quite simple and easy to explain to any business world

2. In both methods, we don’t have to normalize the data like we have to do in Multiple Regression.

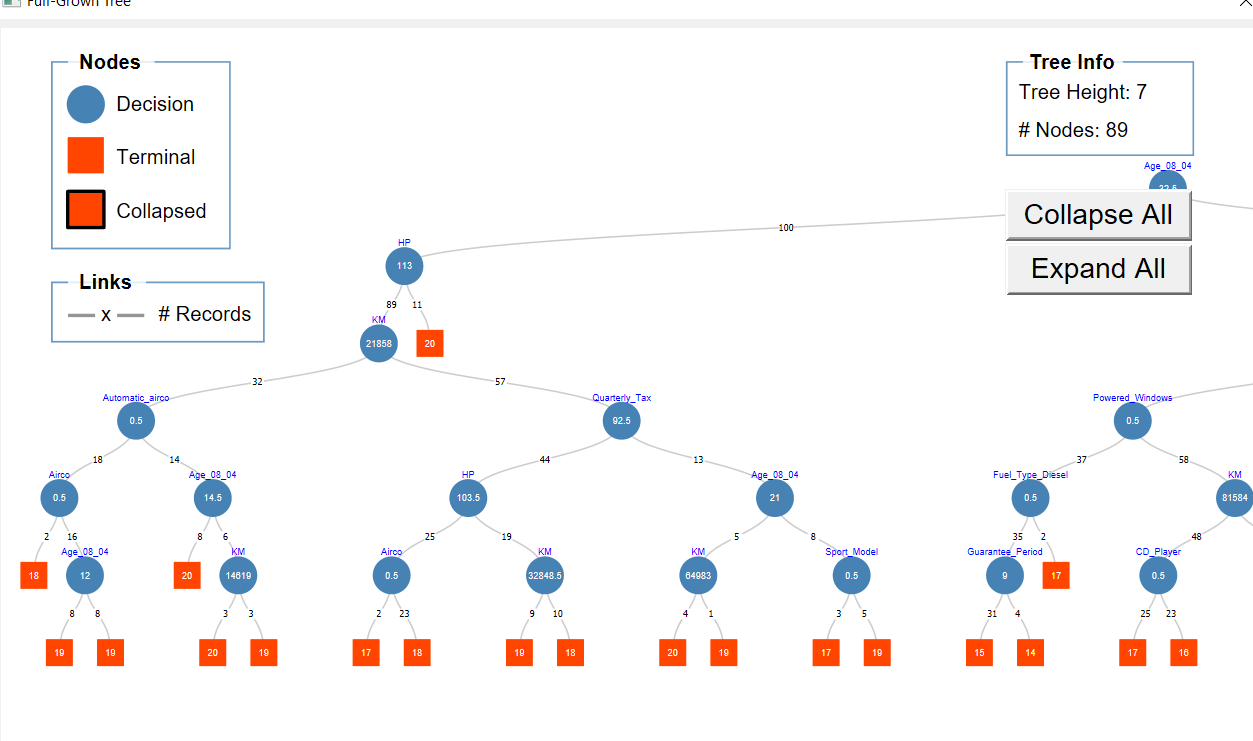
3. In both Trees, the relation between the output variable and predictors is non-linear and nonparametric

4. Both trees handle missing data without adding any value or deleting any missing record.

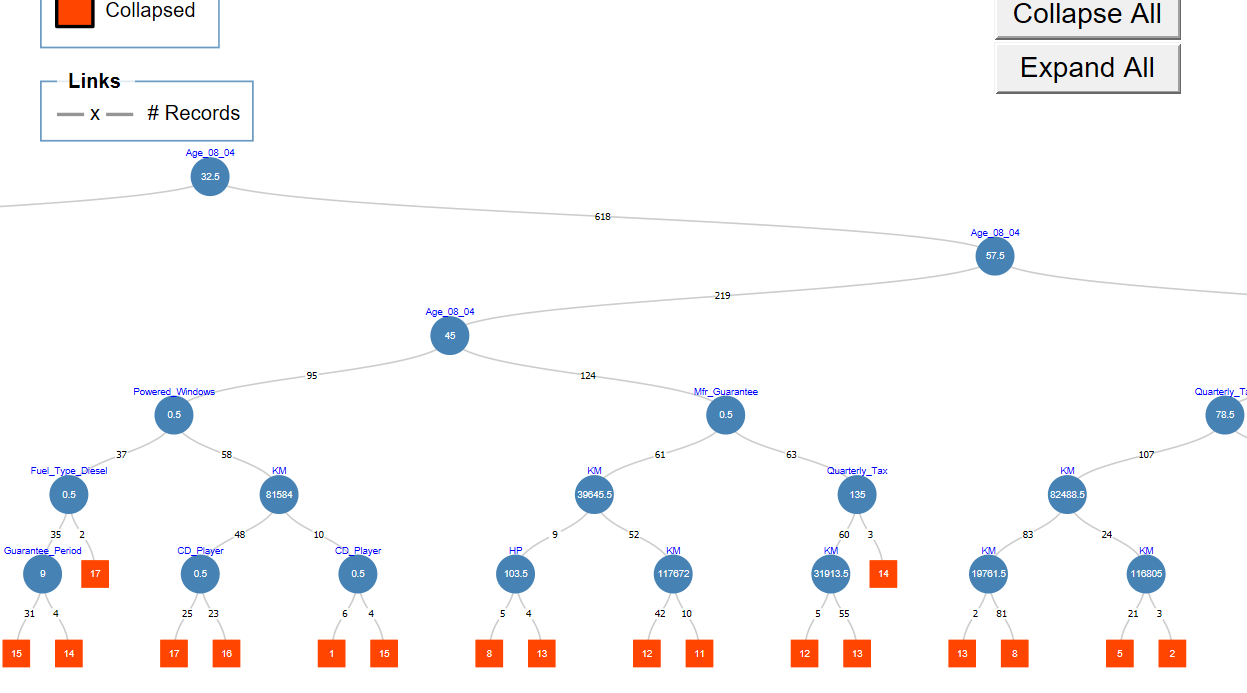
Disadvantages:

1. To perform well both Trees require large data set to choose the right decision node.
2. Both Type of tree can sometime become bias against some predictors, specially the one having more data points to split, like if we see in our example Age\_08\_04 variable appeared many times in our tree.

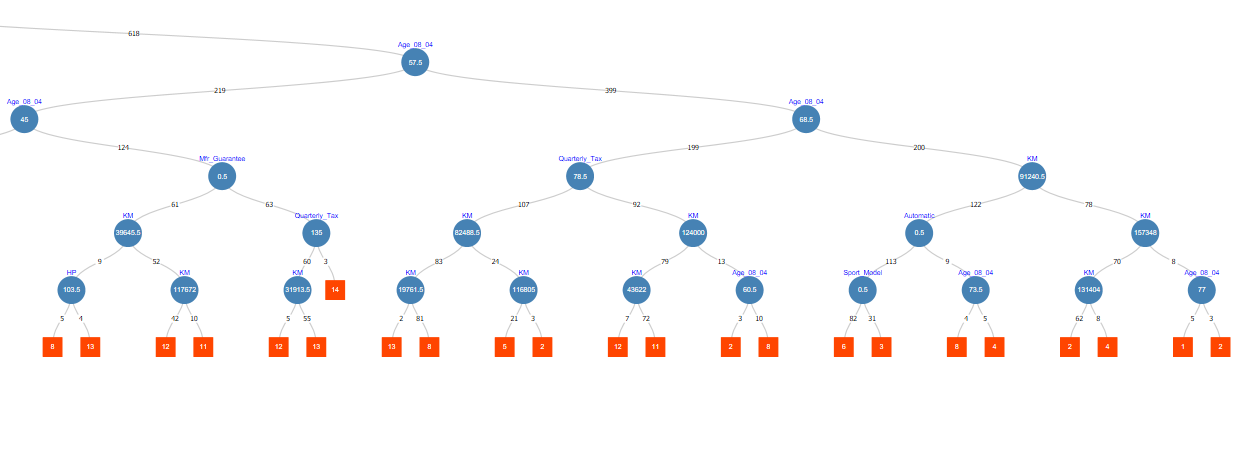
Left



Right:



Extreme Right:



In Regression Full Tree if we see in part A question 4 tree, in left side the the second decision node is HP with value <113 and >113, and in Classification Tree as well same decision node with same magnitude is selected. After that in CT, right Tree is Pruned and next decision node of left is KM, but in RT the tree is not pruned and next decision nodes are Age\_08\_04, the data set having value of HP less than 113 are 89 and in this set the Age\_08\_04 is taken as next decision node with magnitude 17.5, and in right for 11 records again Age\_08\_04 is taken as next decision node with magnitude 5.5.

Similarly, if we compare right side of the data sets having Age\_08\_04 >32.5

the next decision node in both trees is same and value of split is also same, but after that the decision nodes are different.

In RT the next decision node for data records having Age\_08\_04<57.5 is KM with magnitude 127349, but in CT at same level the decision node is again Age\_08\_04 with splitting value 45.

So, if we by looking into tree after level 1 the decision nodes are changed .

**Important submission instructions**

Save your Word file and Excel file. Use the link “Homework 3” to upload these files. **Due by 11.59 P.M. Mar. 24, 2019.**

**\* - Taken from “Classification and Regression Trees Using XLMiner.pptx”**

**\*\* - Taken from “Data Mining for Business Analytics”**