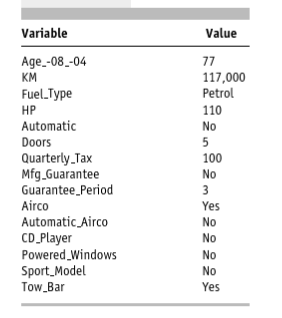
**Homework Three**

**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 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**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?
3. (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?
4. (10 points) How can we achieve predictions for the training set that are not equal to the actual prices?
5. (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?)
6. 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?
   2. (10 points) Predict the price, using the RT and the CT, of a used Toyota Corolla with the speciﬁcations listed in the following Table. 
   3. (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.

**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.**