Individual Assignment 6

Topics: Classification and Regression Trees

YOUR NAME (in English): YUCHIH, SHIH

STUDENT ID: 103034034

Submission instruction:

1. Upload a single PDF to Canvas. It should include all the screenshots and answers (a)-(i).
2. Hand in the printed copy of your document in class.

Data Details and Goal:

The file "purchase_item.txt" contains transaction-level information for over 200,000 purchases of restaurant prepaid vouchers and coupons sold on EZTABLE. Each transaction includes the purchase of one or more vouchers/coupons. Our goal in this assignment is to build a predictive model for 'quantity' (=number of units sold in a single transaction).

Data Preparation:

We will use the data from Assignment 3 (linear regression). Reminder of what we did:

- Filtered the data and kept only the records for the "most popular" restaurant.
- Created derived variables: day-of-week of the transaction date (Excel function =WEEKDAY), and dummy variables for *type* and for *day-of-week (DOW)*
- Handled missing values for categorical columns by creating a new category NA, and for numerical columns removed rows if there are only a few.

Create Binned Quantity:

Bin Quantity into 20 bins with equal counts. Keep this Binned_Quantity variable for Model 2.

R: use *cut()* or, in package binr use function *bins.quantiles()*

Data Partitioning:

Partition the data into training, validation, and test sets (XLMiner: use default %. R: use 50%-30%-20%). Use random seed 12345.

(a) Using the information in the training set, what is a *naive prediction* for Quantity in a future purchase? Quantity = 1.

(b) RMSE of this prediction on the test set = 13.0755.

Model Building: Regression Tree

We are going to model Quantity as a function of price, type, and DOW.

(c) For *type*, if we include both dummies (type_coupon, type_voucher), is it possible to get a different tree compared to including only type_coupon? Yes

Why?because coupon only includes 80% of the type data, and there is only 2 types.

(d) For *DOW*, if we include all 7 dummies (DOW_1,…, DOW_7), is it possible to get a different tree compared to including only 6 dummies? No.

Why?Because Saturday only includes under 5% of the data. I think it can barely affect the result.

Model 1: Quantity. Run a regression tree with Quantity as the outcome and price, type and DOW as predictors. We will build three trees: a deep *full tree*, a *pruned* (=minimum validation error) *tree* , and a *best-pruned tree* (smallest tree within 1 standard error from pruned tree).

XLMiner: *Predict > Regression Tree*.

- **In Step 2:** Min #Records in a terminal node: change to 100 (bigger full tree)
- **In Step 3:** Maximum #levels to display = 7. Check Full tree, Best Pruned Tree, Minimum Error Tree, and all three Detailed Report options.

R:

- Use *rpart()* to run a tree and *prp()* to plot a tree. See Figure 9.7 and https://www.statmethods.net/advstats/cart.html
- Argument *method="anova"* gives a regression tree
- For a full tree, use arguments *minsplit=1,* and *cp=0.0001* (see Figure 9.10)
- To find the pruned tree and best pruned tree compute *xerror* and *xstd* (see Table 9.4)

(e) How many terminal nodes does each of the trees have?

Terminal nodes in Full tree _____73_____

Terminal nodes in Pruned (Min Error) tree _____70_____

Terminal nodes in Best Pruned tree _____20_____

Variable Selection

Examine the regression trees.

(f) Which are the two main predictors of Quantity? _price_ and _type_

(g) Look at the top four levels of the full, pruned, and min-error trees. Are they identical?

Yes. Why? _because they are using the same dataset. pruned trees are based on a full tree and prune those overfitted epochs._

Model Evaluation

(h) Compare the prediction errors of the training, validation, and test sets by examining their RMS Error (RMSE) and by plotting the three box plots.

XLMiner Hint: for side-by-side boxplots, place all three columns of residuals in one column and add another column with label "training/validation/test". Copy these two columns to Tableau for an easy boxplot (*Analysis*> uncheck Aggregate Measures)

Fill in the following table:

	Training	**Validation**	**Test**
RMSE	11.75461	15.7575	11.35505
Which tree used? (full, min err, best pruned)	best pruned	best pruned	best pruned

[Include your boxplots here: they should all have the same y-axis range]

(i) For good predictive accuracy of new records, which tree should we use? (Look at holdout error. XLMiner: worksheet RL_PruneLog. R: xerror)

Pruned

Model 2: Classification Tree on Binned Quantity.

Run a classification tree with outcome Binned_Quantity, and predictors price, type and DOW.

XLMiner: *Classify > Classification Tree*.

- **In Step 2:** Min #Records in a terminal node: change to 100 (bigger full tree)
- **In Step 3:** Maximum #levels to display = 7. Check Full tree, Pruned Tree, Minimum Error Tree.

R: Same instructions as before, but use argument *method="class"* for classification tree

(j) Compare the Classification Tree with the Regression Tree by filling the table:

	Regression Tree	**Classification Tree**
Number of Terminal nodes in full tree	73	252
Number of Terminal nodes in best pruned tree	70	5
Predictor in first split	price	price
Value of split on first predictor	price = 374	
Predictors in top 3 layers of full tree	1.price 2.weedays 3.type	1.price 2.weedays 3.type
Predictors in top 3 layers of best pruned tree	1.price 2.weedays 3.type	1.price 2.weedays 3.type
Predict the Quantity for a purchase of a Voucher on DOW = 5, at price = $2000. Use best pruned tree.		