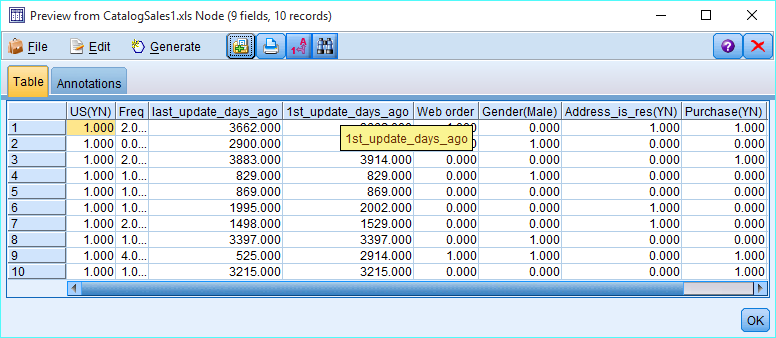
Suppose that we are trying to predict if a person will make a purchase or not with previous Catalog Sales data that has been collected and also the size of their purchase.

1. **Load in the data from the CatalogSales.xlsx file.**

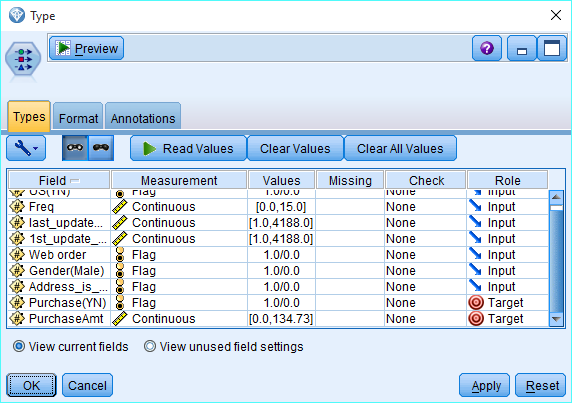


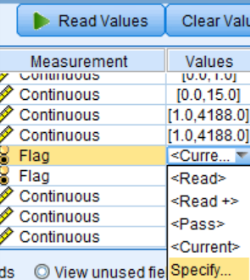
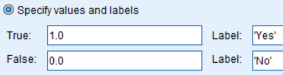
1. **Use a Type node to verify the Types of all the variables.**

**The predictor variables to be examined follow.**

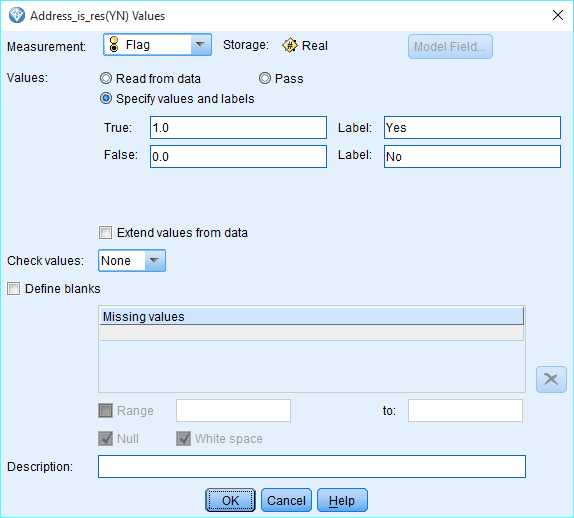
**(Note the Y/N variables are coded Y=1, N=0 – specify them all to flag variables.)**

|  |  |
| --- | --- |
| **Field** | **Description** |
| FREQ | # of transactions in the preceding year |
| LAST\_UPDATE\_DAYS\_AGO | # of days since last update to customer record |
| 1st\_UPDATE\_DAYS\_AGO | # of days since first update to customer record |
| WEB ORDER | whether customer purchased by Web order at least once – Y/N |
| GENDER(male) | male? – Y/N |
| ADDRESS\_IS\_RES(YN) | whether it is a residential address – Y/N |
| US(YN) | whether it is a US address – Y/N  Purchase(YN): (Y/N) 🡨 **TARGET** |
| PurchaseAmt | Amount of last purchase 🡨 **TARGET** |



1. **Use the Type note to change the 1/0 coding on the flag/categorical variables to values that are more ‘user friendly (like Yes or No, etc.).’** 
   1. GENDER: ‘Male’ if 1 and ‘Female’ if no
   2. PURCHASE: ‘Yes’ if 1 and ‘No’ if no.
   3. WEB: ‘Yes’ if 1 and ‘No’ if no.
   4. ADDRESS\_RES: ‘Yes’ if 1 and ‘No’ if no.
   5. US: ‘Yes’ if 1 and ‘No’ if no.

**NOTE**: To get the stream to actually pay attention to these, you have to change the settings in Tools, Stream Properties, Options and under the Options tab check the option **Display field and value labels in output.**



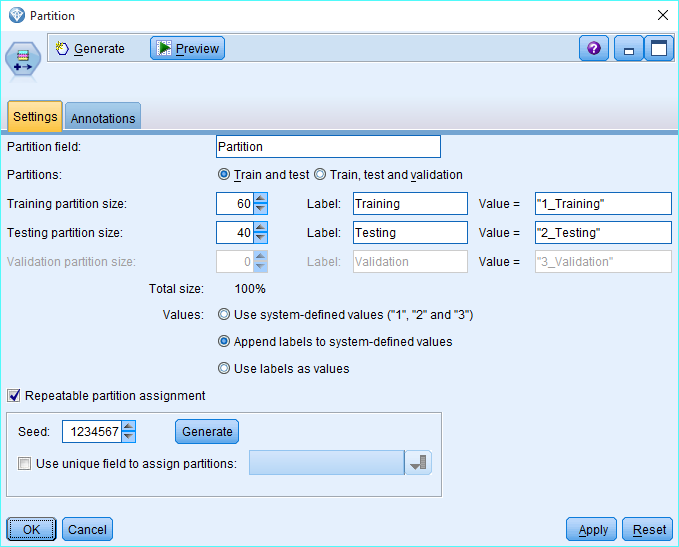
1. **This should leave you with 2 target variables and 3 continuous inputs and 4 flag inputs (9 in total). Use a Data Audit node to verify your changes.**

**Why are there so many 0 purchase amounts?**

There are about 1000 0 amounts because there was no purchase in these records.Purchase amount is greater than 0 only when Purchase is Yes.

1. **Create a stratified partition variable on Purchase with 60% training and 40% test data.**

**Note what values are saved in this variable to indicate training and the test data.**



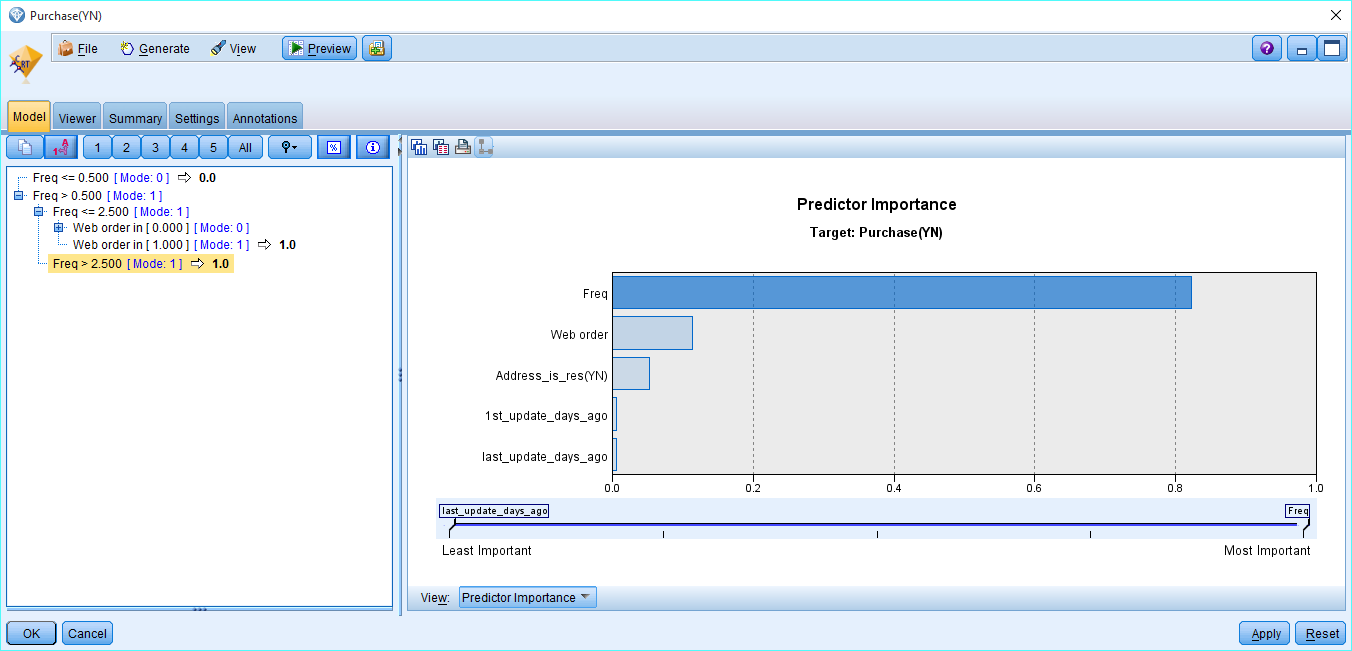
**OK – now we are ready to create some TREES with this dataset.**

1. **Create a CART decision tree model using the default settings predicting the target flag variable Purchase. (DO NOT INCLUDE PURCHASEAMT!!!) (Therefore, you should have 7 predictors).**

**Look at your tree – does it make sense?**

Answer:

Yes, the tree does make sense. Root node is split on Frequency parameter , which make sense because higher frequency is highly correlated with Purchase variable(Target). Also , other leaves are based on weborder and Adress\_is\_res variable , which are also good predictors for our target variables. For example: One of the rule tells us that if Freq<=0.5 then purchase is 0 or NO(100%). Another rule tells that if Freq>=0.5 and Freq <=2.5 and weborder=1 then Purchase is 1 or Yes (66% times or 66% chance).



1. **Create confusion matrices for the following situations:**
   1. ALL DATA – use Matrix node
   2. TRAINING DATA – use Select node (Partition = "1\_Training") and Matrix node
   3. TEST DATA – use Select node (Partition = "1\_Testing") and Matrix node

**ALL DATA % Error =**

|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | |
| **ACTUAL** | **Purchase** | **Not Purchase** |
| **Purchase** | **765** | **235** |
| **Not Purchase** | **231** | **769** |

**TRAINING DATA % Error =**

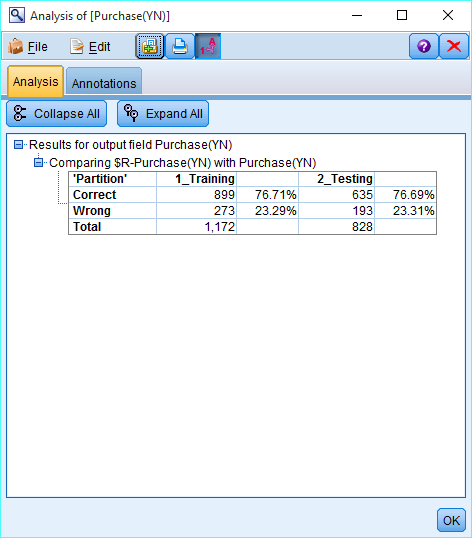
|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | |
| **ACTUAL** | **Purchase** | **Not Purchase** |
| **Purchase** | **441** | **128** |
| **Not Purchase** | **145** | **458** |

**TESTING DATA % Error =**

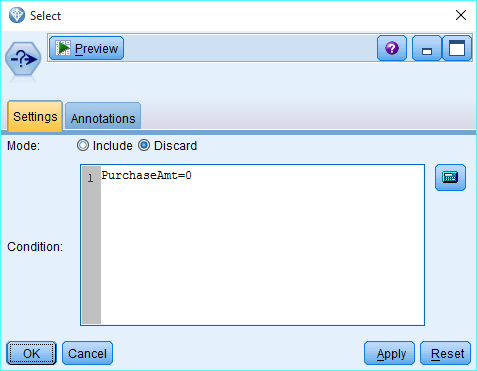
|  |  |  |
| --- | --- | --- |
|  | **PREDICTED** | |
| **ACTUAL** | **Purchase** | **Not Purchase** |
| **Purchase** | **324** | **107** |
| **Not Purchase** | **86** | **311** |

1. **Attach an Analysis node to your CART nugget. Did the results match your test and training errors?**

**Yes.**



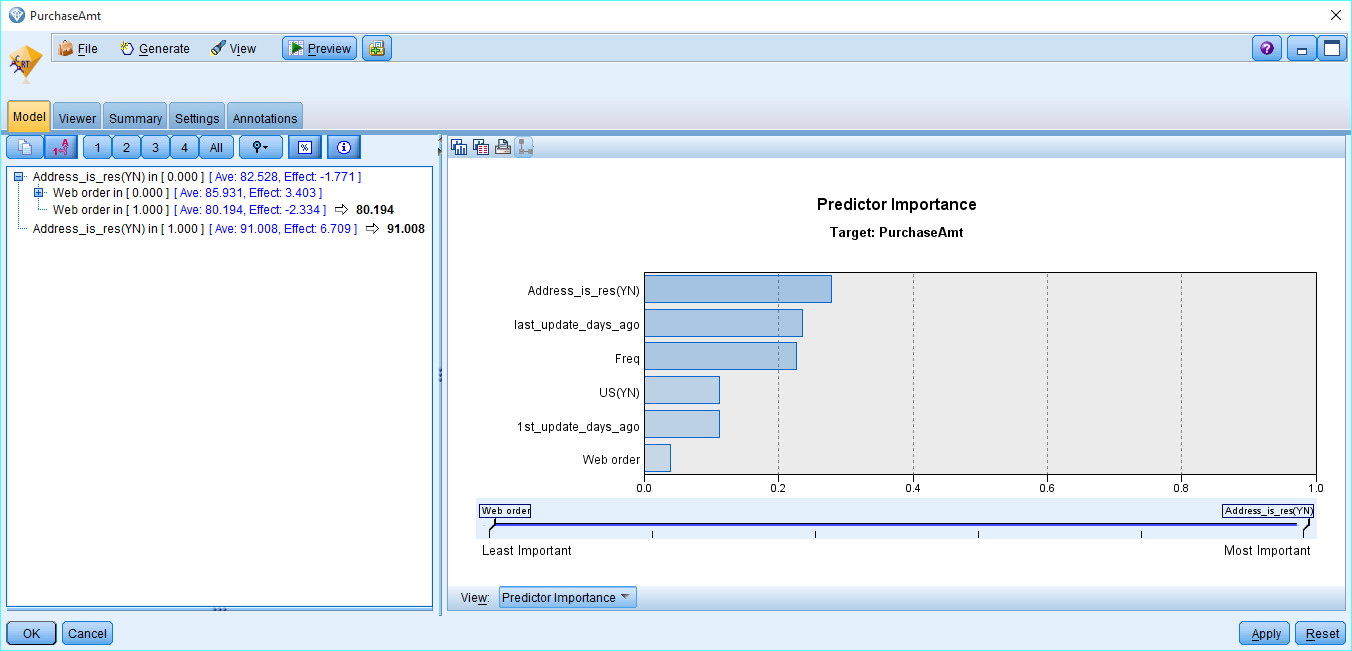
1. **Create a CART decision tree model using the default settings predicting the target continuous variable ‘PurchaseAmt’. (DO NOT INCLUDE PURCHASE (Y/N)!!!) BUT WAIT – we don’t want to include those records that have 0 purchase amounts.**



**Exclude those first with a Select Node before making the tree. (You should have 7 predictors).**

**Look at your tree – does it make sense?**

Yes, it does make sense. First split is on Address\_is\_res and one of the rule tells that if Address\_is\_us=0 and web\_order=1 then mean value for the purchase\_amount is 80.194, which make sense. The most valuable predictors are Address\_is\_seq ,last\_update\_days\_ago and freq.



1. **Attach an Analysis node to your CART nugget. Compare your errors between the training and test data. Is there a big difference?**

We can compare the predictive accuracy on the training and test data sets by comparing the Mean Absolute Error. There does not seem to be a big difference in errors (10.042 & 11.157), however, running an ANOVA to test if their difference is statistically significant we have:

