

## SPSS Workshop ---Decision Tree/Rule Induction

**Objectives**

- Build a C5.0 rule model
- Browse and interpret the results

**Data**

We will use the dataset *churn.txt* that is on Black board. This data file contains information on 1477 customers of a telecommunication company who have at some time purchased a mobile phone. The customers fall into one of three groups: current customers, involuntary leavers and voluntary leavers. In this lesson, we use decision tree models to understand which factors influence group membership.

**Note:** To download the churn.txt from Blackboard, right click on “churn.txt”, then choose ‘save link as’, save the data to you destination folder.

**Introduction**

Rule induction or decision tree methods are capable of culling through a set of predictors by successively splitting a dataset into subgroups on the basis of the relationships between predictors and the target field.

We will use the C5.0 node to create a rule induction model. It contains the rule induction model in either decision tree or rule set format. By default, the C5.0 node is labeled with the name of the output field. The C5.0 model can be browsed and predictions can be made by passing new data through it in the Stream Canvas.

**Steps:**

1. Click **File...New Stream** → Place a **Var. File** Node on the Canvas → Double-click the node to input the data of **Churn.txt** from → then Click **OK**.
2. Place a **Type Node** from **Field Ops** palette to the right of **Churn.txt** node → Connect them → Double click the type node to define the role of data → Click the **measurement** for **ID** → select the **Typeless** → Click the **Role** for **CHURNED** → select the **Target** (as shown below).

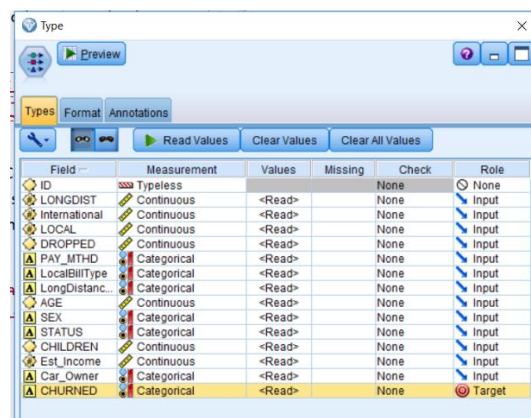


Figure 1

3. Place a **Partition Node** from **Field Ops** palette to the right of **Type** node → Connect the **partition node** to **Type node** → Double-click the partition node to decide the training and testing size → we choose **50% samples** go to training and **50% go to the testing**. (This proportion is subjective, you can decide for your own analysis. In order to get the same result, we recommend you use this proportion).

4. Place a **C5.0** node from the Modeling palette to the right of the **Partition** node → Connect the **Partition** node to the **C5.0** node → Double click the **C5.0** node → Click **Run** → then the generated **CHURNED** appear (the diamond-shaped icons) → Double Click the generated **CHURNED** → then the results will appear.

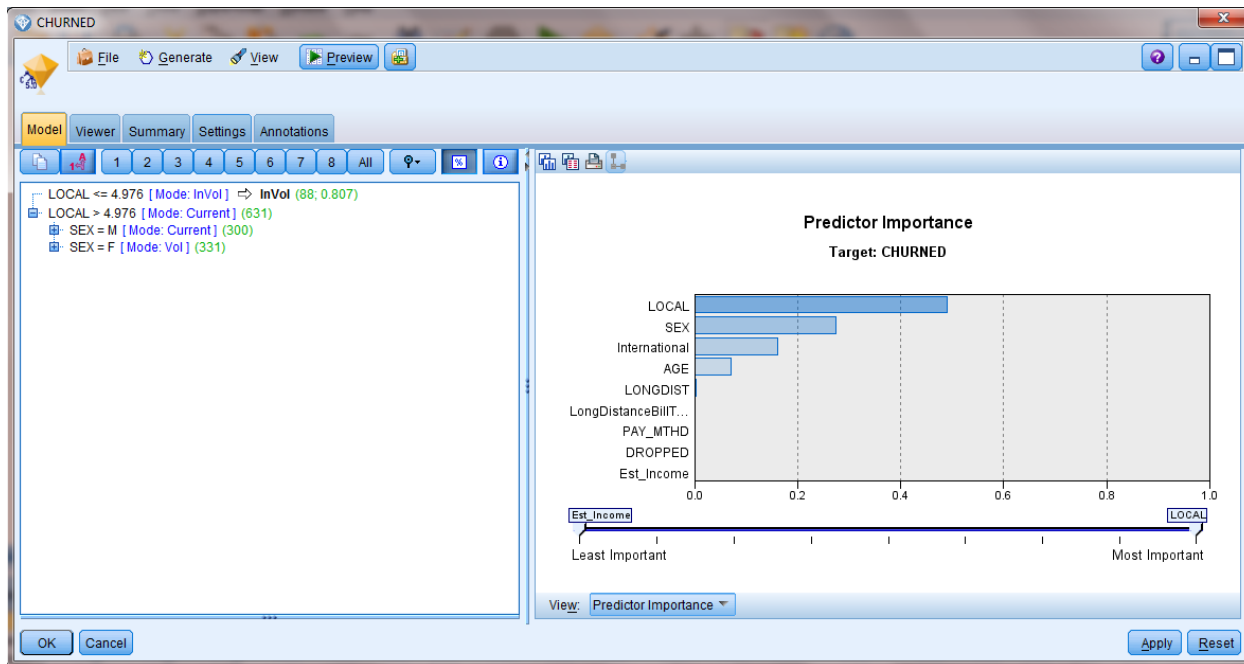


Figure 2

## Interpret the results


5. The Model Viewer window has two panes. The left one shows the root node of the tree and the first and second splits; the right pane displays a graph of predictor importance measures.

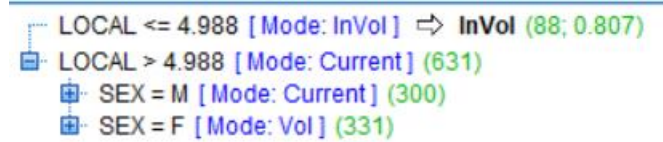
According to what we see of the tree so far, *LOCAL* is the first split in the tree. Further, we see that if *LOCAL* ≤ 4.976, the *Mode* value for *CHURNED* is *InVol*. The *Mode* is the modal (most frequent) output value for the branch, and it will be the predicted value unless there are other fields that need to be taken into account within that branch to make a prediction. When *LOCAL* ≤ 4.976 the branch terminates, visually apparent because of the arrow. So this means the prediction for all customers with this range of values on *LOCAL* is to be an involuntary churner.

In the second half of the first split where *LOCAL* > 4.976, the *Mode* value is *Current*. In this instance, no predictions of *CHURNED* are visible, and to view the predictions we need to further unfold the tree to *SEX* split.

The bar chart shows that the field *LOCAL*, used on the first split, is by far the most important in predicting *CHURNED*. However, we haven't seen the whole tree, and critically, we aren't yet ready to use the test partition data, so we won't examine predictor importance any further at the moment.


7. Click **Show or hide instance and confidence figures**  in the toolbar

8. Click the expand button  to **unfold** the branch **LOCAL > 4.976**

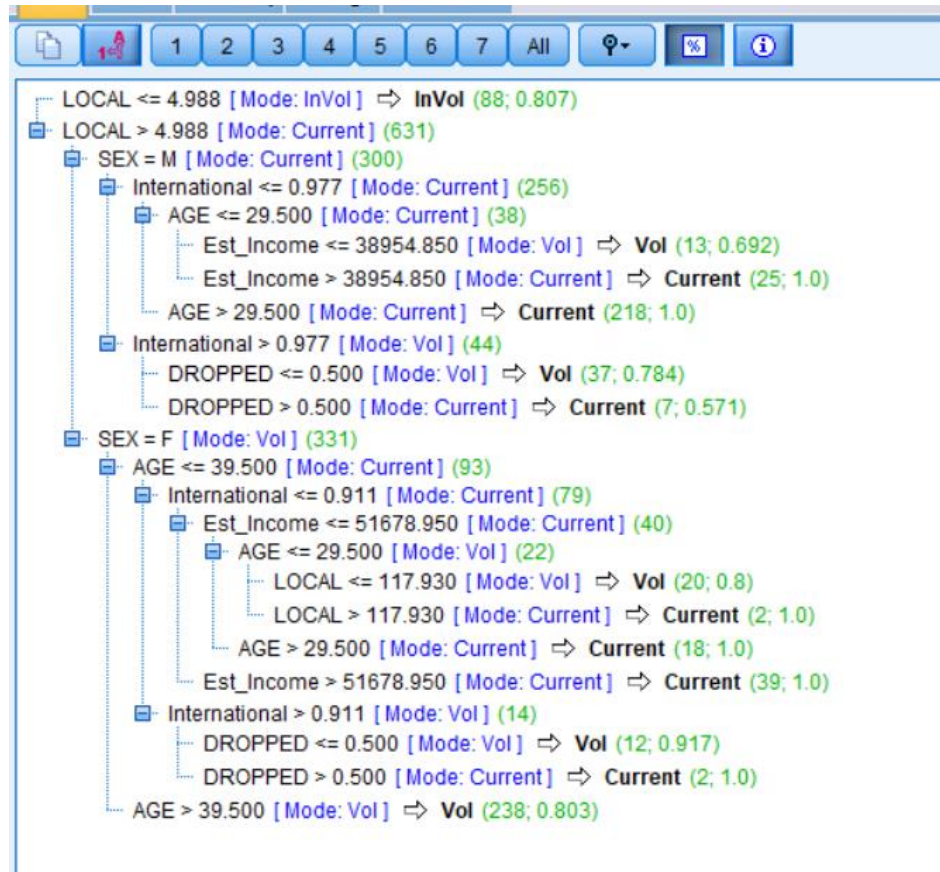


**Figure 3**

*SEX* is the next split field. Now we see that *SEX* is the best predictor for persons who spend more than 4.976 minutes on local calls. The Mode value for Males is *Current* and for Females is *Vol*. However, at this point we

still cannot make any predictions because there is a  symbol to the left of each value of *SEX* which means that other fields need to be taken into account before we can make a prediction. Once again we can unfold each separate branch to see the rest of the tree, but we will take a shortcut:

9. Click the **All** button in the Toolbar



**Figure 4**

We can see several nodes usually referred to as terminal nodes that cannot be refined any further. In these instances, the mode is the prediction. For example, if we are interested in the *Current Customer* group, one group we would predict to remain customers are persons where *LOCAL* > 4.976, *SEX* = M, *International* <= 0.905, and *AGE* > 29. To get an idea about the number and percentage of records within such branches we ask for more details.

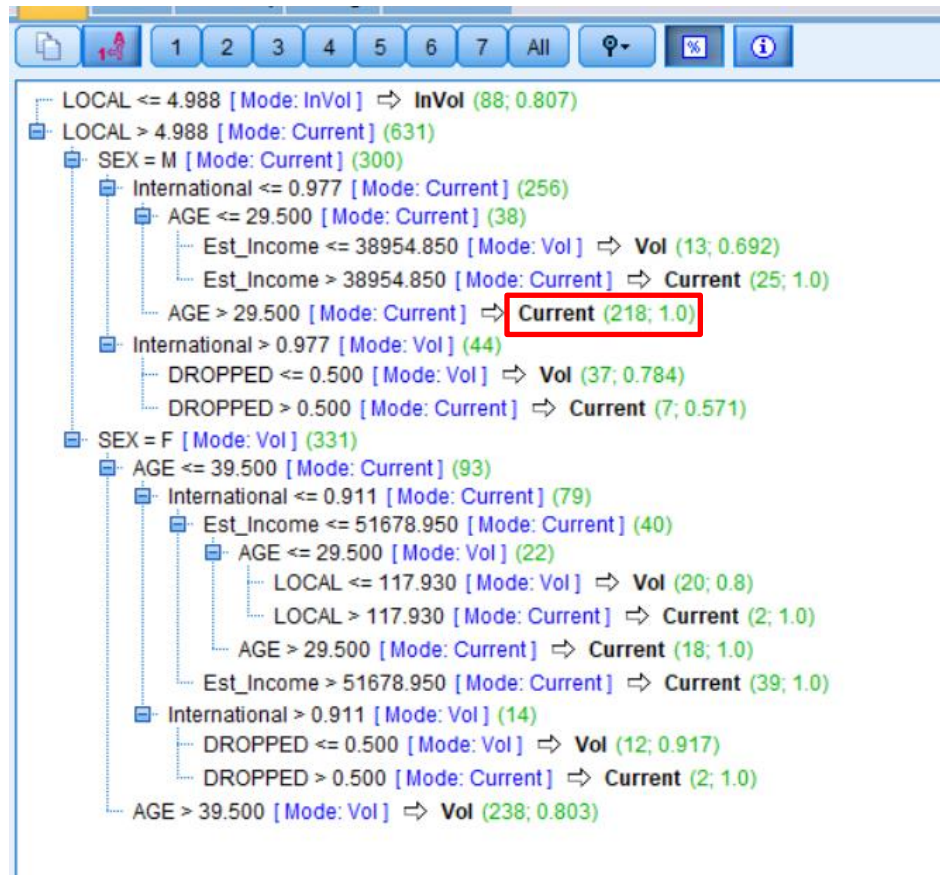


Figure 5

The *incidence* tells us that there are 218 persons who met those criteria. The *confidence* figure for this set of individuals is 1.0, which represents the proportion of records within this set correctly classified (predicted to be *Current* and actually being *Current*). That means it is 100% accurate on this group! If we were to score another dataset with this model, how would persons with the same characteristics be classified? Because SPSS (PASW) Modeler assigns the group the modal category of the branch, everyone in the new dataset who met the criteria defined by this rule would be predicted to remain *Current Customers*.

10. If you would like to present the results to others, an alternative format is available that helps visualize the decision tree. The Viewer tab provides this alternative format.

Click the **Viewer** tab → Click the **Decrease Zoom** tool (to view more of the tree). (You may also need to expand the size of the window.)

The root of the tree shows the overall percentages and counts for the three categories of CHURNED. The modal category is shaded in each node. We see that there are 719 customers in the training partition.

When it is not possible to view the whole tree at once, such as now, one of the more useful buttons in the toolbar is the Tree map button because it shows you the size of the tree. A red rectangle indicates the portion of the tree that is being displayed. You can then navigate to any portion of the tree you want by clicking on any node you desire in the Tree map window.

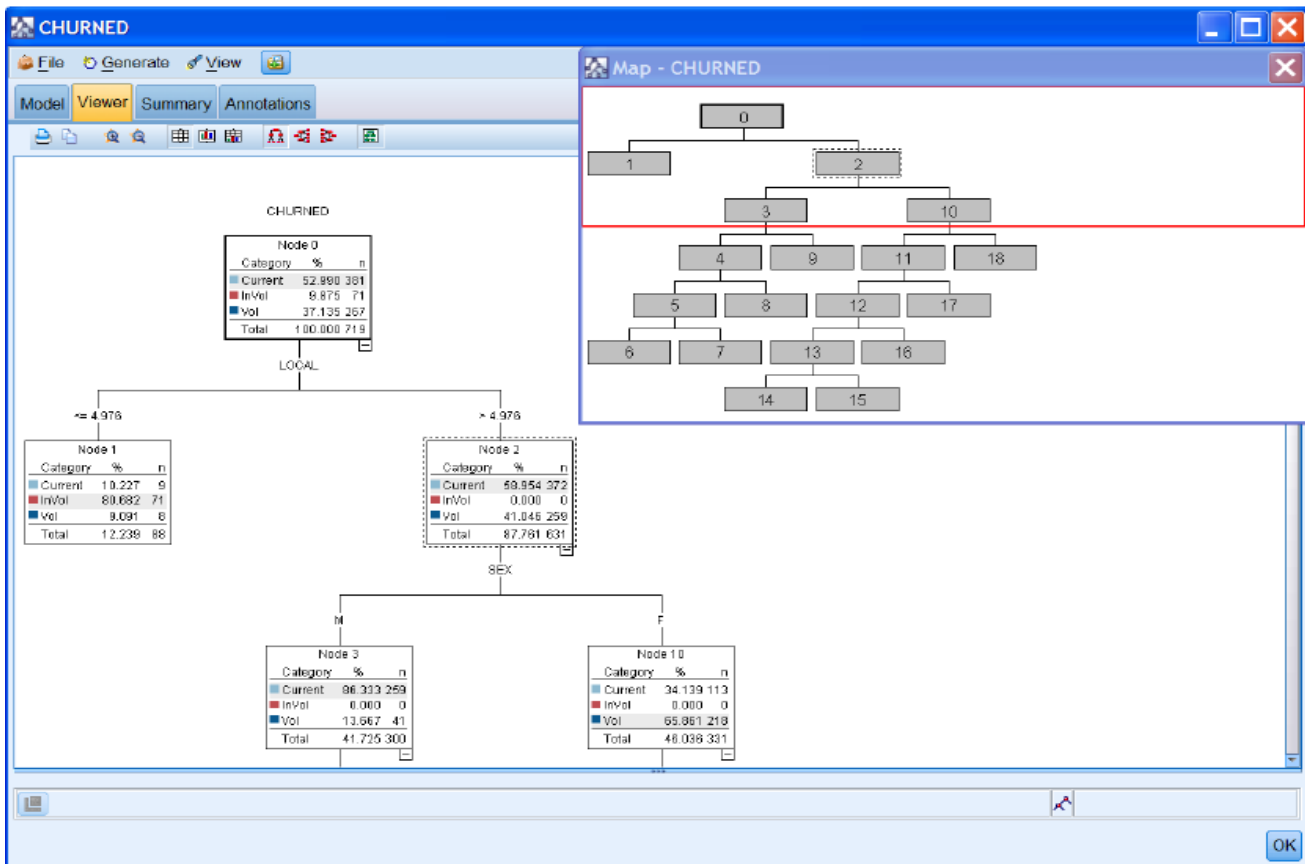


Figure 8

## Understanding the Rule and Determining Accuracy

The predictive accuracy of the rule induction model is not given directly within the C5.0 model node. To get that information, you can use Analysis nodes to determine how good the model is. **(The output of step 11 can also be obtained using the Matrix nodes from the output palette in SPSS modeler. See instructions below)**

### Using Analysis Node

11. Using the **Analysis** node: Place the **Analysis Node** from Output tab after the gold diamond, double click Analysis, and make sure to check '**Coincidence matrices (for symbolic targets)**', then click **Run**.

**Results for output field CHURNED**

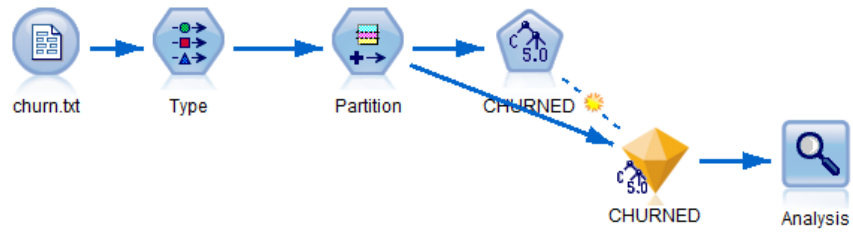
Comparing \$C-CHURNED with CHURNED

'Partition'	1_Training		2_Testing	
Correct	643	89.43%	647	85.36%
Wrong	76	10.57%	111	14.64%
Total	719		758	

Coincidence Matrix for \$C-CHURNED (rows show actuals)

'Partition' = 1_Training	Current	InVol	Vol
Current	314	9	58
InVol	0	71	0
Vol	1	8	258

'Partition' = 2_Testing	Current	InVol	Vol
Current	352	8	91
InVol	0	61	0
Vol	6	6	234



## The modeling stream for Rule Induction