Minxia Ji

**ISGB 7967 Data Mining for Business**

SPSS Modeler Assignment 1 - Rule Induction/Decision Tree

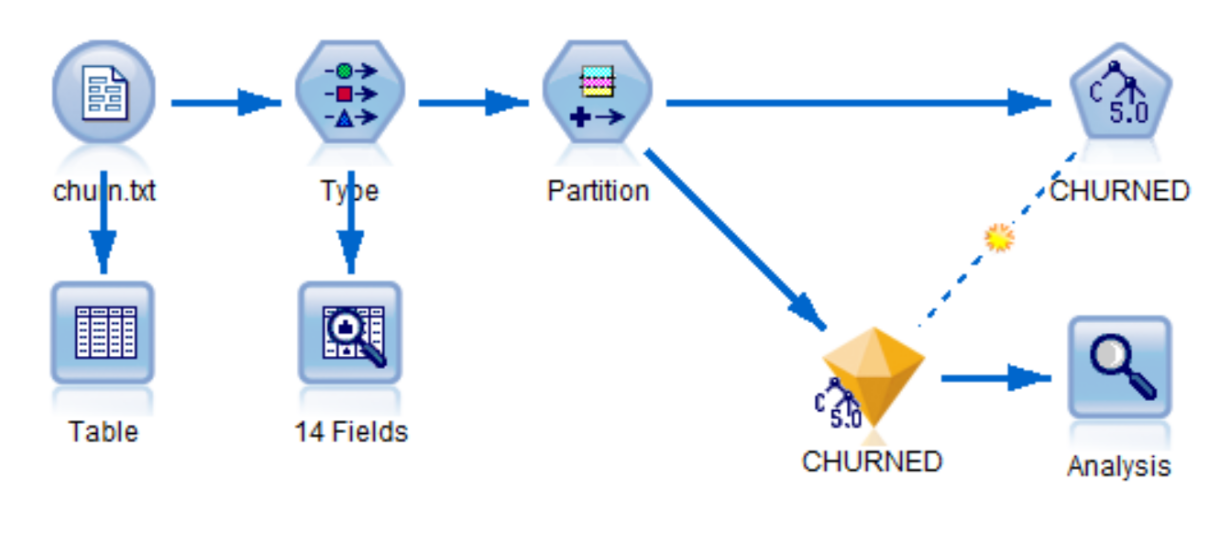
**PART A**

**Problem statement**

We want to use the sample data, which contains records of 1477 customers, to infer all customers.The customers fall into one of three groups: current customers, involuntary leavers and voluntary leavers.We need to determine ***what are the features*** of these 3 groups.

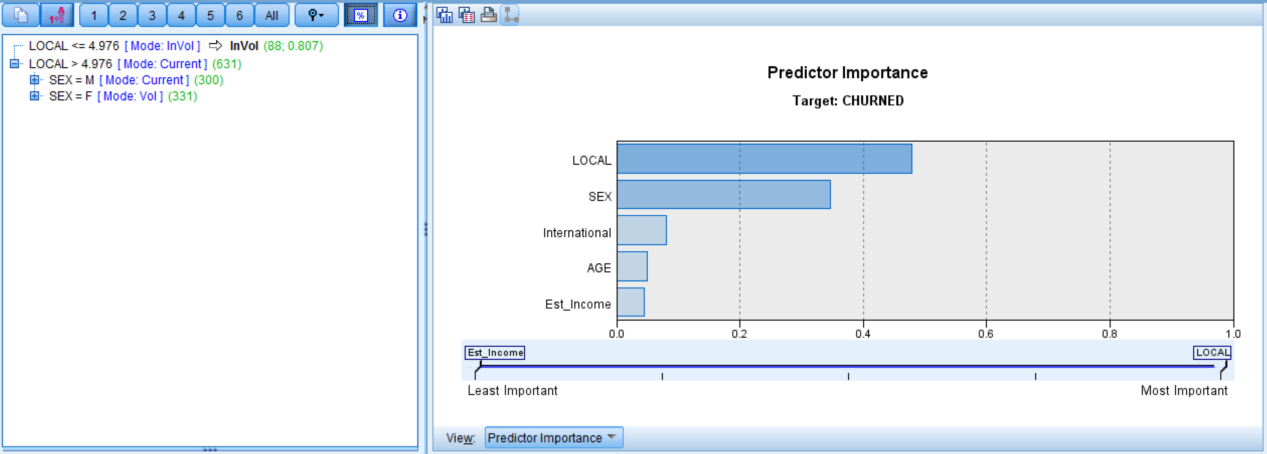
According to these different features, we can tell who is ***voluntary leavers***.Thus, the company can make strategies to attract those customers.

**What was done and analysis**



**Figure 1.1**

**Figure 1.1** Use var.file to load churn.txt,connect the var.file to type bottom,set ID to ‘none’(because ID is meaningless to the analysis),target ‘Churn’.Set 50% training data and 50% testing data.Then, build a C5.0 model.The pruning severity is 75 and minimum records per child branch is 3 at C5.0 node.

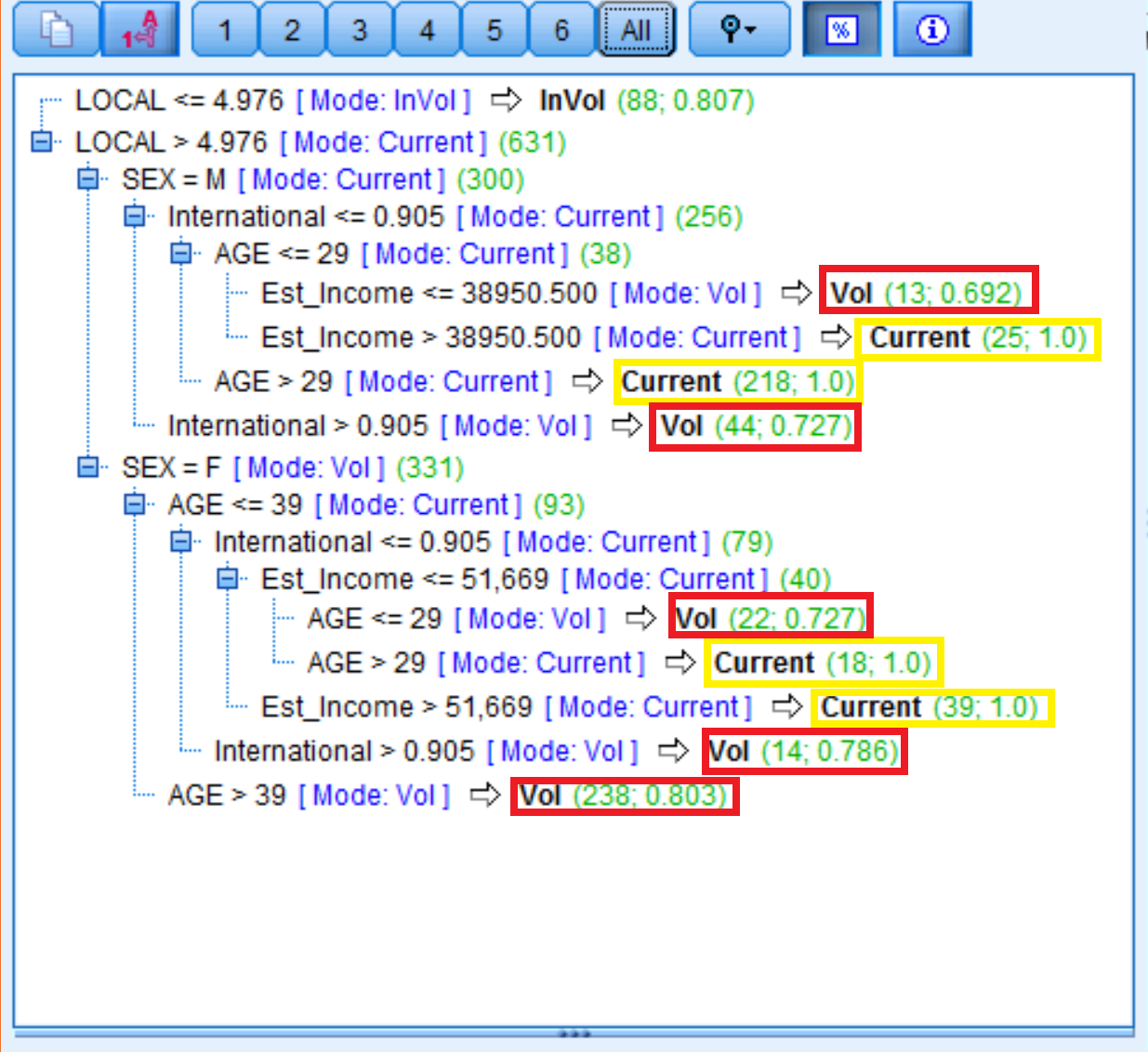


**Figure 1.2**

**Figure 1.2** The bar chart on the right side indicates that ***LOCAL*** is the most important predictor to predict ***CHURNED***.

On the left side, ***LOCAL***is the **first split** in the tree.We can tell that local is the most important predictor to distinguish involuntary leavers from the rest customers.If *LOCAL* <= 4.976, the *Mode* value for *CHURNED* is *InVol*.

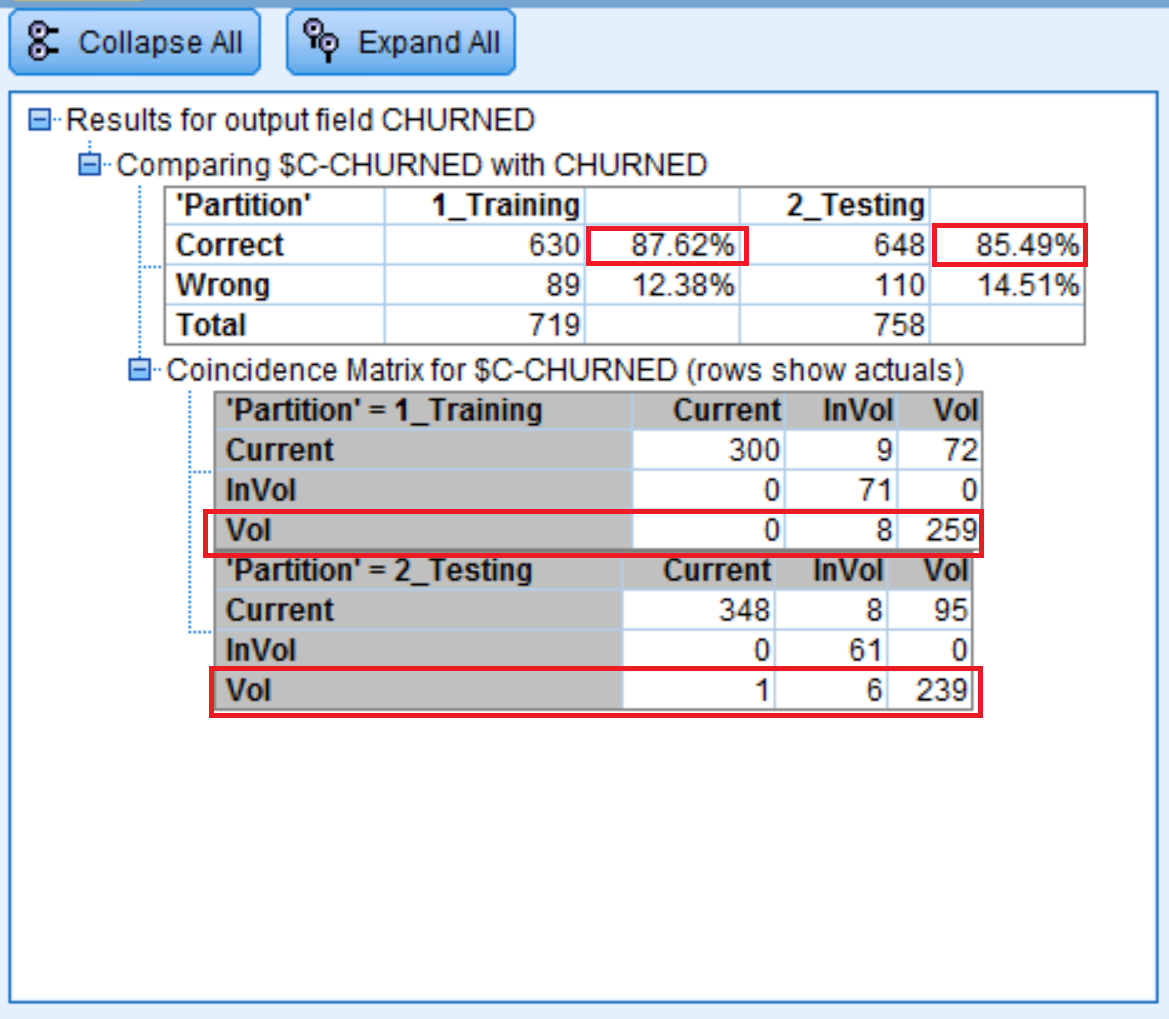
**SEX** is the **second split**.Among the rest customers,the most important predictor to distinguish VOL and CURRENT is SEX.



**Figure 1.3**

**Figure 1.3** The highlight parts(in yellow) of figure 1.3 show that the accuracy is 100% on these groups.

The highlight parts(in red) of figure 1.3 include all VOL.The biggest group of VOL has these features:LOCAL>4.976,SEX=F,AGE>39.



**Figure 1.4**

**Figure 1.4** The model predicts about **87.6%** of the *whole data* category correctly.Since ***87.6%*** is much greater than ***56%***(the percent of the main group), the model is not useless.The results in testing data are similar to the training data.

The 2 long red rectangles suggest that the group of voluntary leavers has been well classified,both in training and testing data.It’s a good model to tell who is voluntary leaver.

Overall, the results of testing suggest that the model will perform well with new data.

**Managerial conclusions**

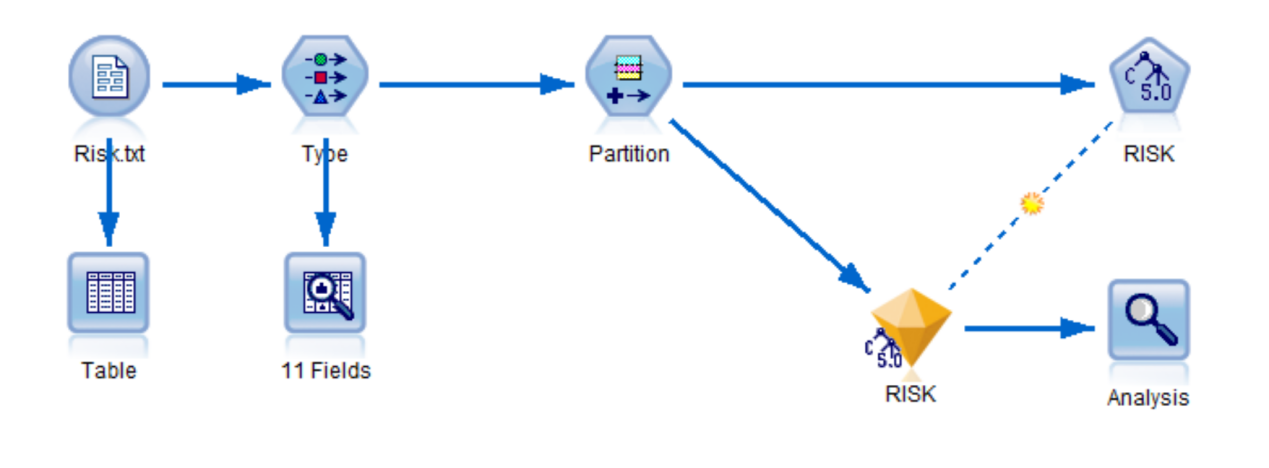
For the biggest VOL group(LOCAL>4.976,SEX=F,AGE>39), the company should offer special discount or other attractions to keep them.For the second biggest VOL group(LOCAL>4.976,SEX=M,International>0.905), the company could provide special discount on international calls to them.

PART B

**Problem statement**

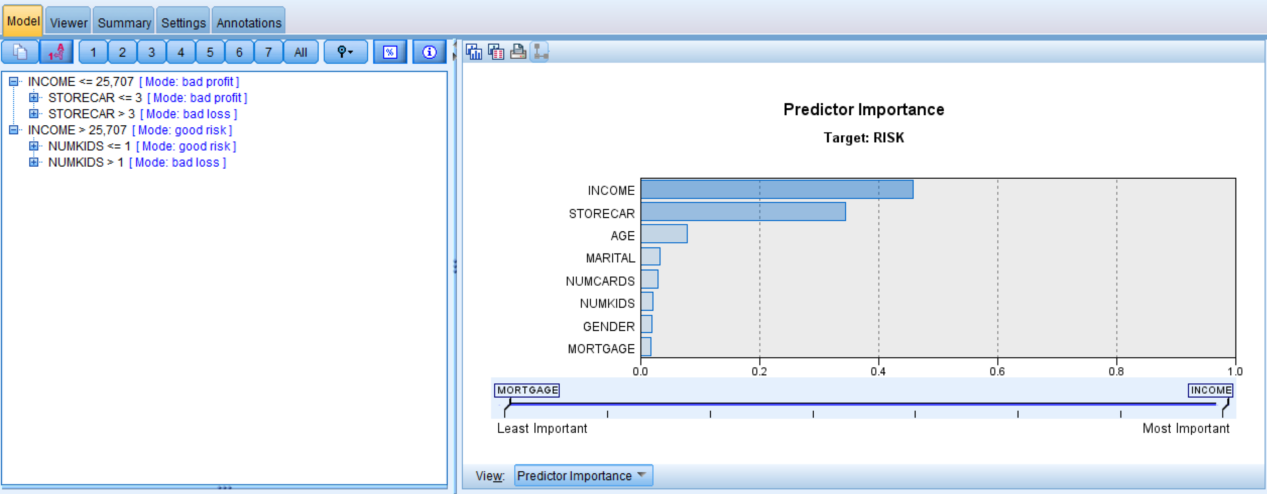
We want to predict credit risk from the demographic fields.Thus, we need to build a decision tree model from the data, a risk study.The customers fall into one of three groups: good risk, bad risk-profitable and bad risk-loss.We need to determine ***what are the features*** of these 3 groups.

**What was done and analysis**



**Figure 2.1**

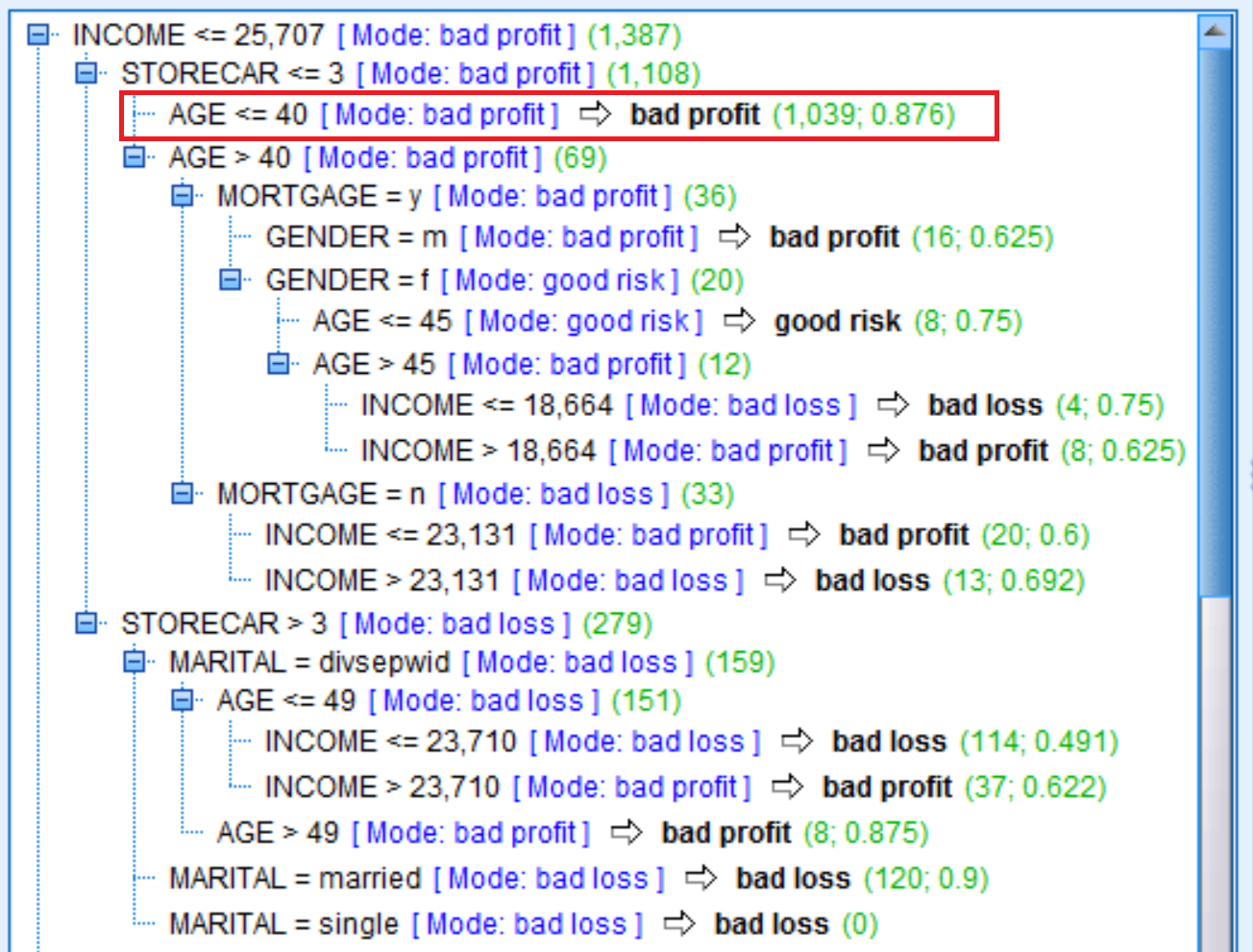
**Figure 2.1** Use var.file to load Risk.txt,connect the var.file to type bottom,set ‘none’ to the ID(because ID is meaningless to the analysis),target ‘Risk’.Set 50% training data and 50% testing data.Then, build a C5.0 model.The pruning severity is 75 and minimum records per child branch is 3 at C5.0 node.

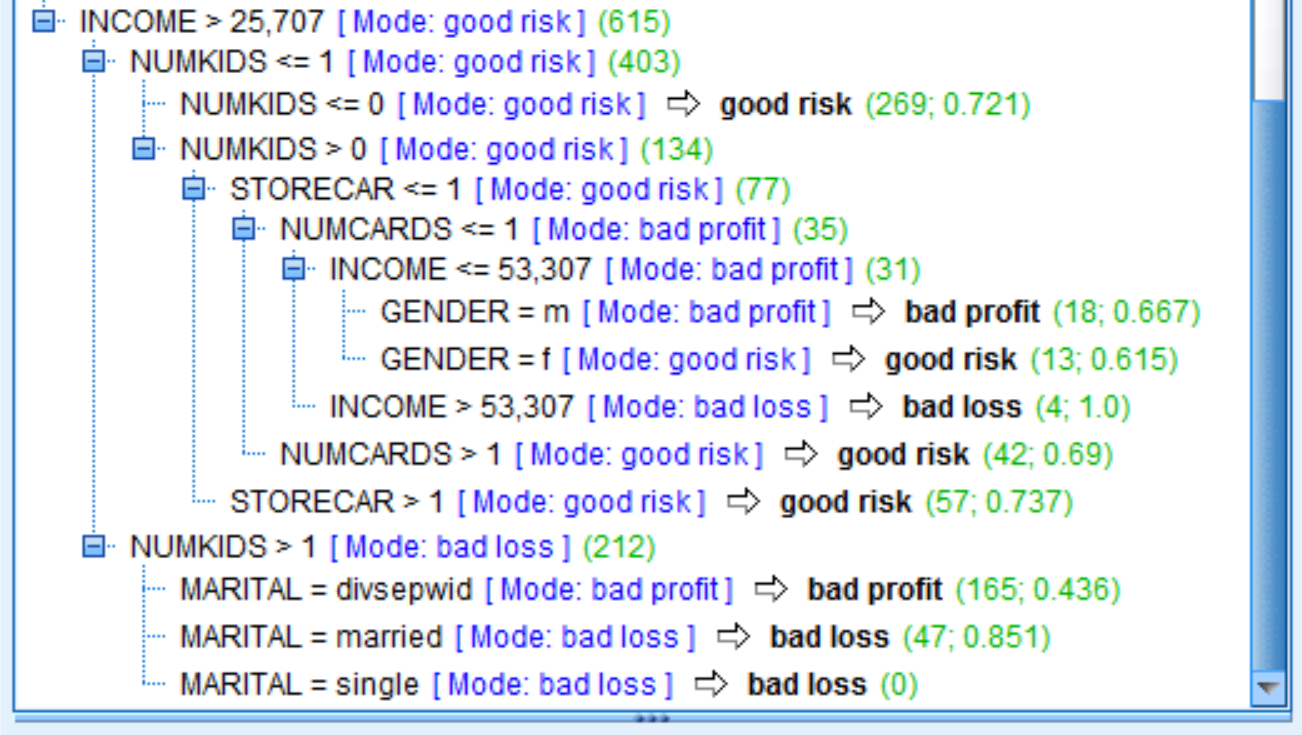


**Figure 2.2**

**Figure 2.2** The bar chart on the right side indicates that ***INCOME*** is the most important predictor to predict ***Risk***.

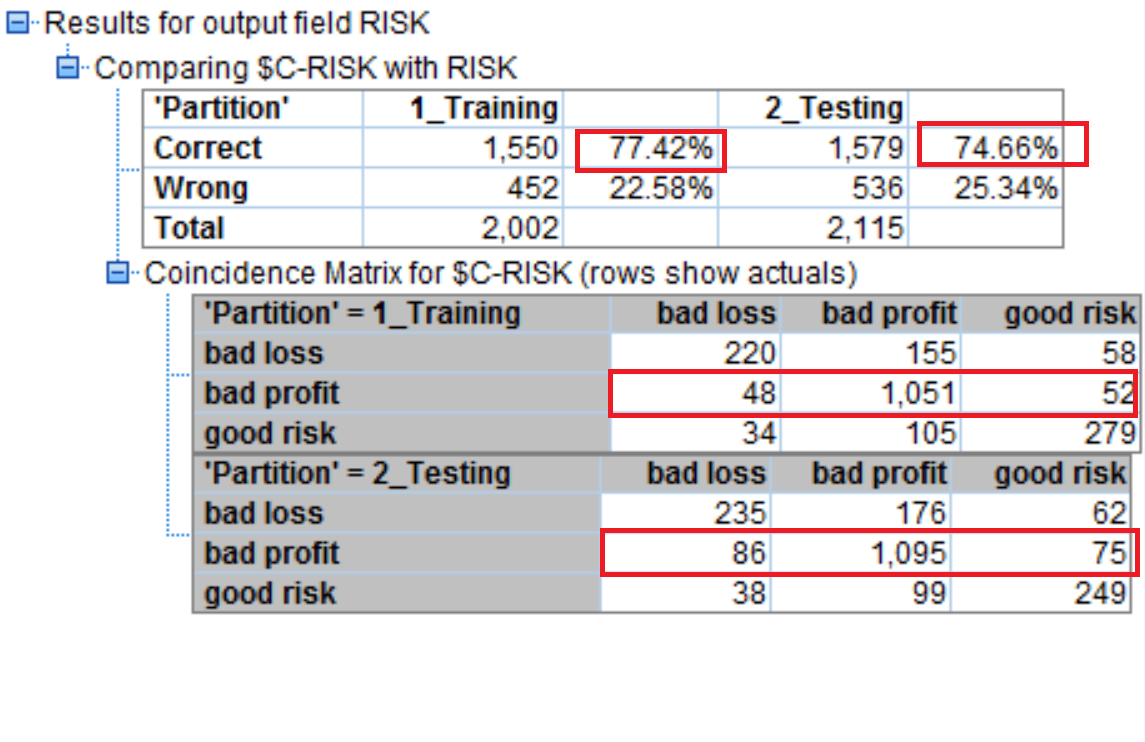
On the left side, ***INCOME***is the **first split** in the tree.***STORECAR*** and ***NUMKIDS*** are the second important predictor in the decision tree.





**Figure 2.3**

**Figure 2.3** we can tell from the highlighted rectangle that the majority of customers belongs to the group ***bad profit.***Their features:INCOME<=25,707, STORECAR<=3 and AGE<=40.87.6% of customers in this group is correctly classified.



**Figure 2.4**

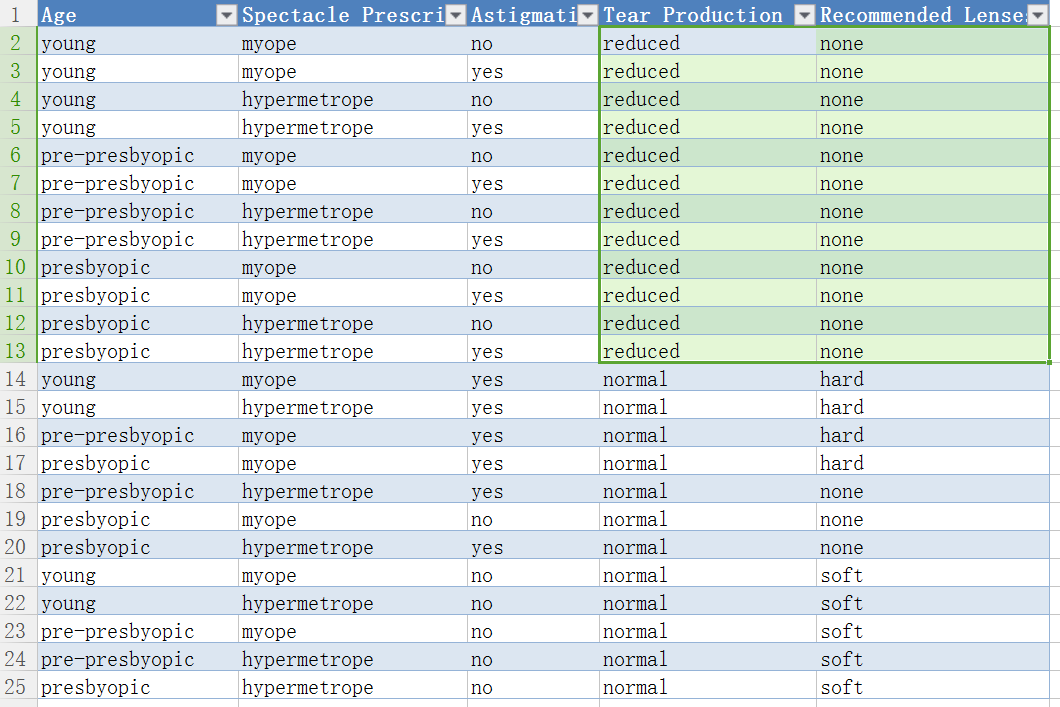
**Figure 2.4** The model predicts about **77.4%** of the *whole data* category correctly.Since ***77.4%*** is much greater than ***58%***(the percent of the main group,bad profit,data from data audit node), the model is not useless.The results in testing data are similar to the training data.The 2 long red rectangles suggest that the group of bad profit has been well classified,both in training and testing data.

***But***,the classification of bad loss and good risk is not good enough.The confusion matrix for ***testing*** result shows that this model predicts about 49.7% *bad loss* category correctly, 87.2% of *bad profit* and 65.5% of *good risk* correctly. In addition, Precision(*bad loss*)= 65.5%, Precision(*bad profit*)=79.9% and Precision(good *risk*)=64.5%. Regarding the cost of making the wrong decisions, wrongly classifying *bad loss* as *bad profit* or *good risk* is much more costly than making a mistake oppositely. Overall, this model will perform well in predicting *bad profit*, but not perform well in predicting *bad loss* and *good risk*.

**Managerial conclusions free of technical jargon**

This model is limited in predicting bad loss and good risk.If we want to use the model to predict credit risk from the demographic fields, we need to be careful about bad loss.In order to minimize the loss, under the circumstance that we can’t decide whether the customer belongs to good risk or bad loss, we can put him/her into bad loss.The accuracy of bad profit is good so far. People with income lower or equal to 25,707, number of store credits cards less of equal to 3 and age lower or equal to 40 might fall into *bad profit* group.

**PART C**



**Figure 3.1**

**Figure 3.1** We should choose ***Tear Production*** for split at the root of the tree.

As we can see in Figure 3.1, when the column Tear Production was sorted, all ‘**reduced**’ correspond to the value ‘**none**’ in the column Recommend Lenses.According to that, we can get a **pure leaf node**:if Tear Production=reduced, Recommend Lenses should all be ‘none’.