Universidade do Minho
Escola de Engenharia

Knowledge Discovery

José Machado

Cristiana Neto

DECISION TREES WITH RAPIDMINER

CONTEXT AND PRESPECTIVE

Ricardo works for a large online retail store. The company will launch a next generation eReader soon and they want to maximize the effectiveness of their marketing.

Ricardo has noticed that certain types of people were the most anxious to get the previous generation device, while other folks seemed to content to wait to buy the electronic gadget later.

He’s wondering what makes some people motivated to buy something as soon as it comes out, while others are less driven to have the product.

CONTEXTO E PRESPECTIVA

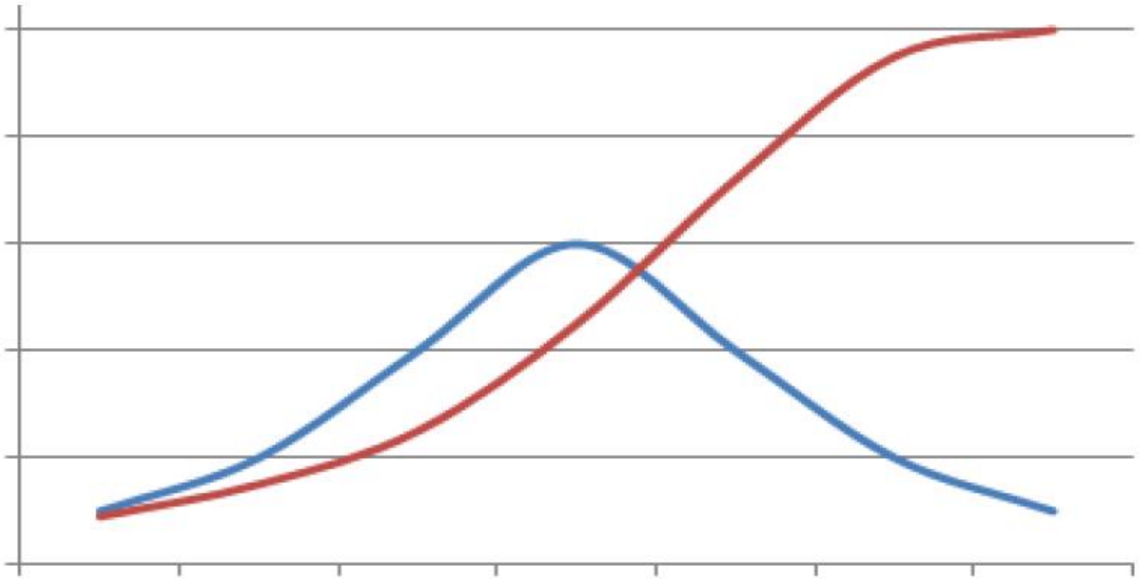
The company where Ricardo works also sells other products, such as books (paper and digital), music and electronic products of various types. Ricardo believes that by extracting customer data on the general consumer behavior on the website, he will be able to find out which customers will buy the new eReader early, which ones will buy next, and which ones will buy later on.

Data Mining can help Ricardo predict when a customer will be ready to buy the next generation eReader, thus being able to target his marketing to the people most ready to respond to ads and promotions.

BUSINESS UNDERSTANDING

Ricardo also wants to understand how customer behavior on his company's website can indicate the moment of purchase of the new eReader.

Diffusion of Innovations Theory (Rogers, 1960s)



Number of adopters by group
Cumulative number of adopters over time

Rogers surmised that the adoption of a new technology or innovation tends to follow an ‘S’ shaped curve, with a smaller group of the most enterprising and innovative customers adopting the technology first, followed by larger groups of middle majority adopters, followed by smaller groups of late adopters.

BUSINESS UNDERSTANDING

Understanding Rogers' theory, Richard believes that he can categorize his company's customers into one of four groups that will eventually buy the new eReader:

Innovators → Early Adopters → Early Majority → Late Majority

He hopes that by watching the customers' activity on the company's web site, he can anticipate approximately when each person will be most likely to buy an eReader. He feels like data mining can help him figure out which activities are the best predictors of which category a customer will fall into. Knowing this, he can time his marketing to each customer to coincide with their likelihood of buying.

DATA UNDERSTANDING

Training Dataset

Contains the web site activities of customers who bought the company's previous generation reader, and the timing with which they bought their reader.

Scoring Dataset

Comprised of attributes of current customers which Richard hopes will buy the new eReader.

Ricardo hopes to figure out which category of adopter each person in the scoring dataset will fall into based on the profiles and buying timing of those people in the training dataset.

DATA UNDERSTANDING

The *datasets* present the following attributes:

- **User_ID:** A numeric, unique identifier assigned to each person who has an account on the company's web site.
- **Gender:** The customer's gender, as identified in their customer account. In this data set, it is recorded a 'M' for male and 'F' for Female. The Decision Tree operator can handle non-numeric data types.
- **Age:** The person's age at the time the data were extracted from the web site's database.
- **Marital_Status**: The person's marital status as recorded in their account. 'M' – married and S – single.
- **Website_Activity**: This attribute is an indication of how active each customer is on the company's web site (Seldom, Regular, or Frequent);

DATA UNDERSTANDING

- **Browsed_Electronics_12Mo:** This is simply a Yes/No column indicating whether or not the person browsed for electronic products on the company's web site in the past year.
- **Bought_Electronics_12Mo:** Another Yes/No column indicating whether or not they purchased an electronic item through Richard's company's web site in the past year.
- **Bought_Digital_Media_18Mo:** This attribute is a Yes/No field indicating whether or not the person has purchased some form of digital media in the past year and a half.
- **Bought_Digital_Books:** indicates whether or not the customer has ever bought a digital book, not just in the past year or so.

DATA UNDERSTANDING

- **Payment_Method**: This attribute indicates how the person pays for their purchases. In cases where the person has paid in more than one way, the mode, or most frequent method of payment is used. There are four options:
  - ✓ Bank Transfer—payment via e-check or other form of wire transfer directly from the bank to the company
  - ✓ Website Account—the customer has set up a credit card or permanent electronic funds transfer on their account so that purchases are directly charged through their account at the time of purchase.
  - ✓ Credit Card—the person enters a credit card number and authorization each time they purchase something through the site.
  - ✓ Monthly Billing—the person makes purchases periodically and receives a paper or electronic bill which they pay later either by mailing a check or through the company web site's payment system.

DATA UNDERSTANDING

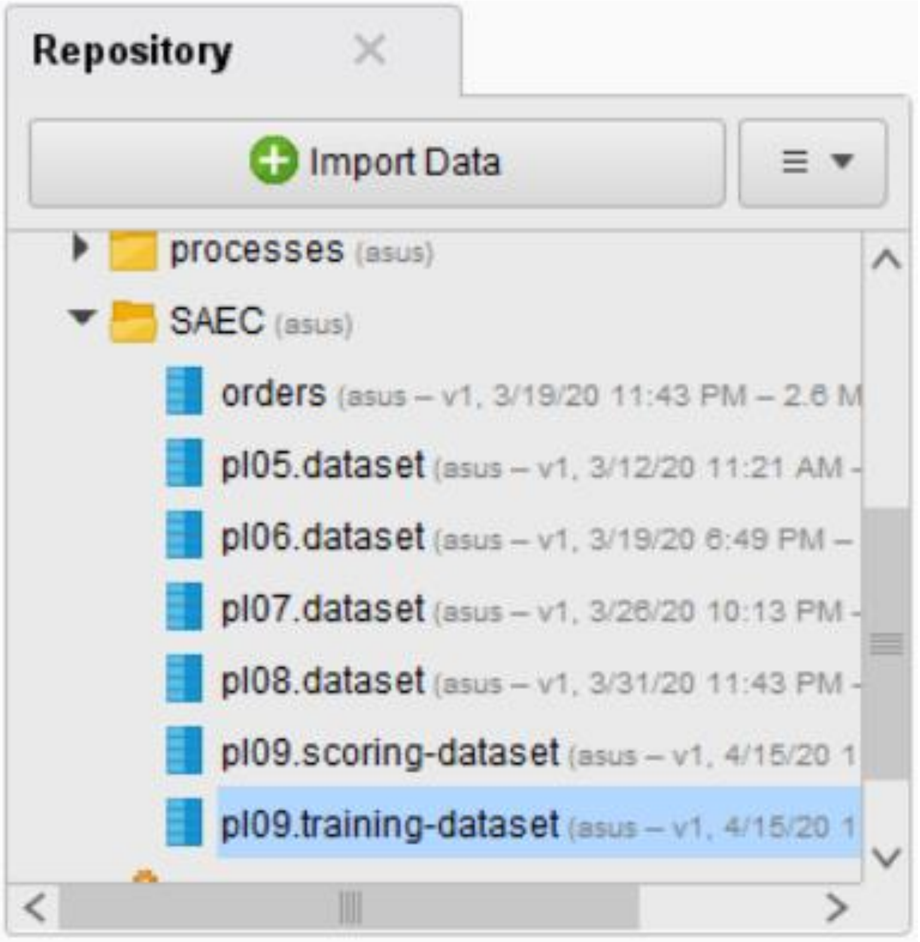
- **eReader_Adoption:** this attribute exists only in the training dataset. It consists of data for customers who purchased the previous-gen eReader. Those who purchased within a week of the product's release are recorded in this attribute as 'Innovator'. Those who purchased after the first week but within the second or third weeks are entered as 'Early Adopter'. Those who purchased after three weeks but within the first two months are 'Early Majority'. Those who purchased after the first two months are 'Late Majority'. This attribute will serve as our label when we apply our training data to our scoring data.

DATA PREPARATION

Download datasets: pl09.training-dataset.csv
pl09.scoring-dataset.csv

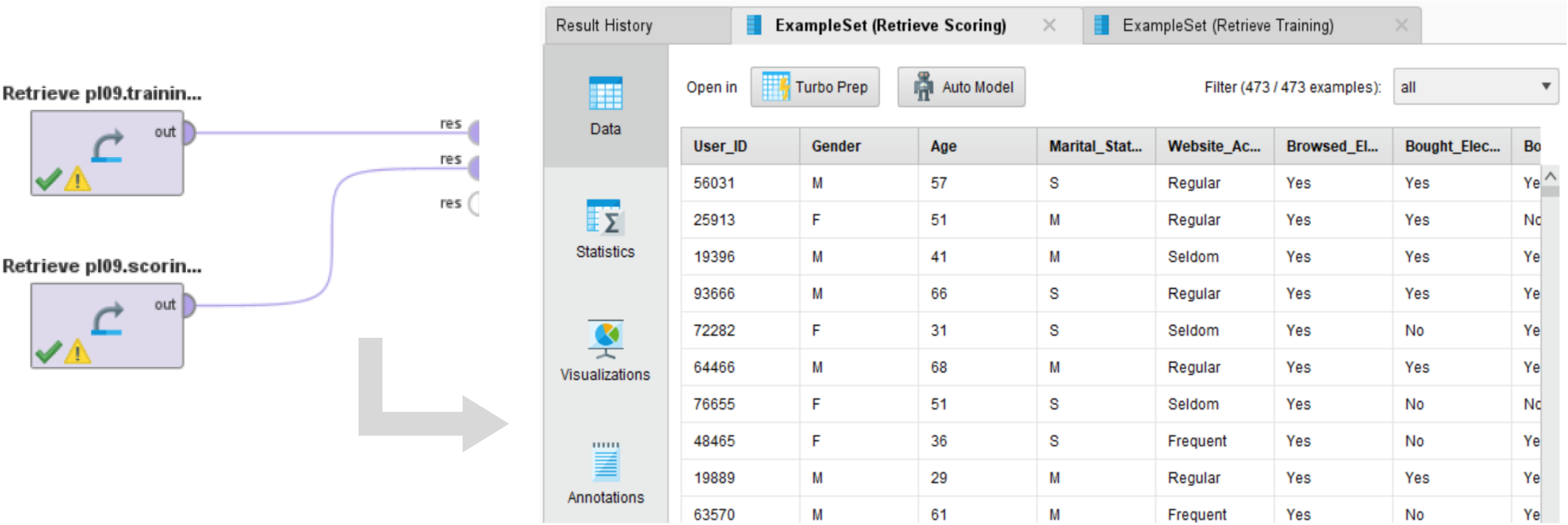
1. Import the datasets to the rapidminer repository (Import Data -> My Computer).

2. Check the Results view and inspect the imported CSV data. You don't have to worry about the attribute's data types because the Decision Tree operator can manipulate all types of data.



DATA PREPARATION

3. Connect both *out* ports to the *res* ports, as shown in the figure and the run the model. Examine the data and familiarize yourself with the attributes shown in the table.



User_ID	Gender	Age	Marital_Stat...	Website_Ac...	Browsed_El...	Bought_Elec...	Bo
56031	M	57	S	Regular	Yes	Yes	Ye
25913	F	51	M	Regular	Yes	Yes	No
19396	M	41	M	Seldom	Yes	Yes	Ye
93666	M	66	S	Regular	Yes	Yes	Ye
72282	F	31	S	Seldom	Yes	No	Ye
64466	M	68	M	Regular	Yes	Yes	Ye
76655	F	51	S	Seldom	Yes	No	No
48465	F	36	S	Frequent	Yes	No	Ye
19889	M	29	M	Regular	Yes	Yes	Ye
63570	M	61	M	Frequent	Yes	No	Ye

DATA PREPARATION

While there are no missing or apparently inconsistent values in the data set, there is still some data preparation yet to do.

1º User_ID attribute

It serves only as a customer identifier in the dataset and therefore should not be included in the model as an independent variable. Solution?

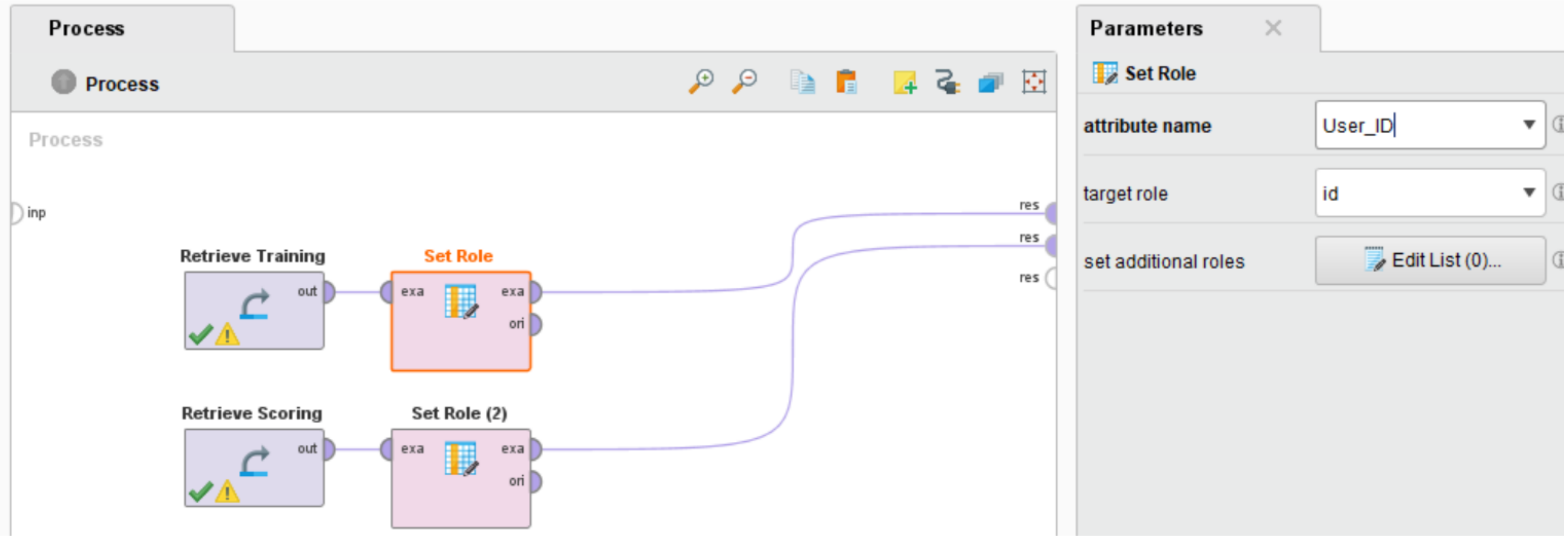
Select Attributes
Removes the attribute

OU

Set Role
New way of handling a non-predictive attribute

DATA PREPARATION

4. Locate and add two *Set Role* operators to each one of the streams (training and scoring). In the parameters on the right side, set the role of the User_ID attribute to 'id' (for the 2 Set Role operators). This causes the attribute to remain in the dataset, but it is not considered as a predictor for the label attribute.

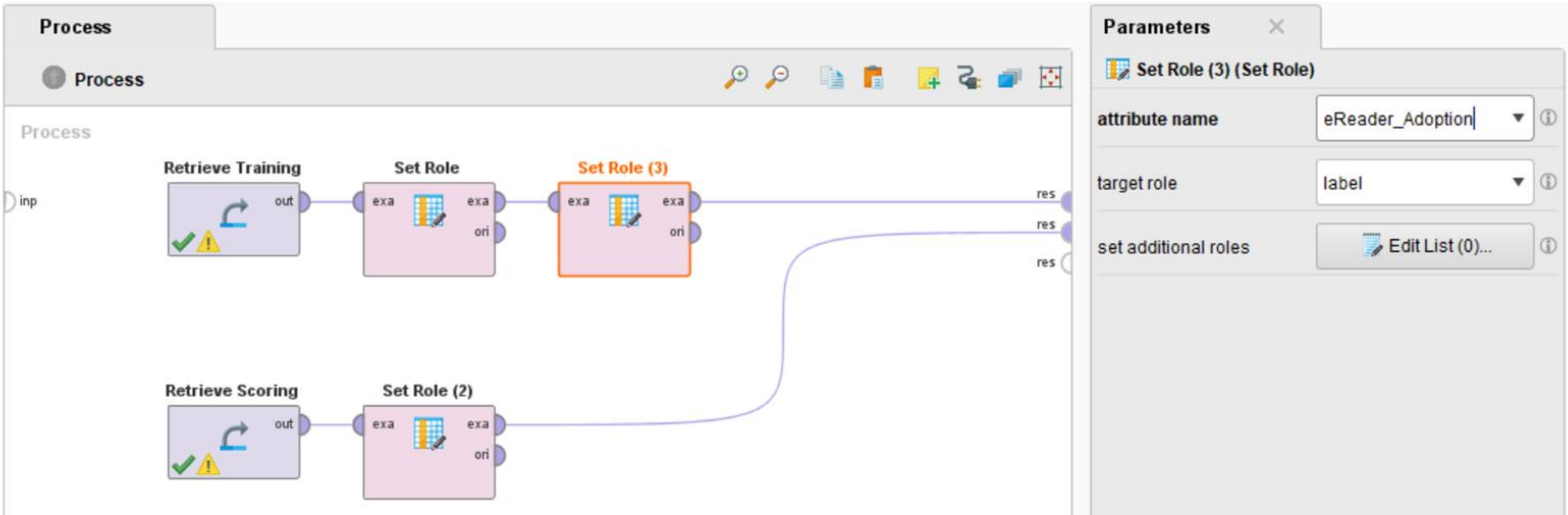


DATA PREPARATION

2º Label attribute

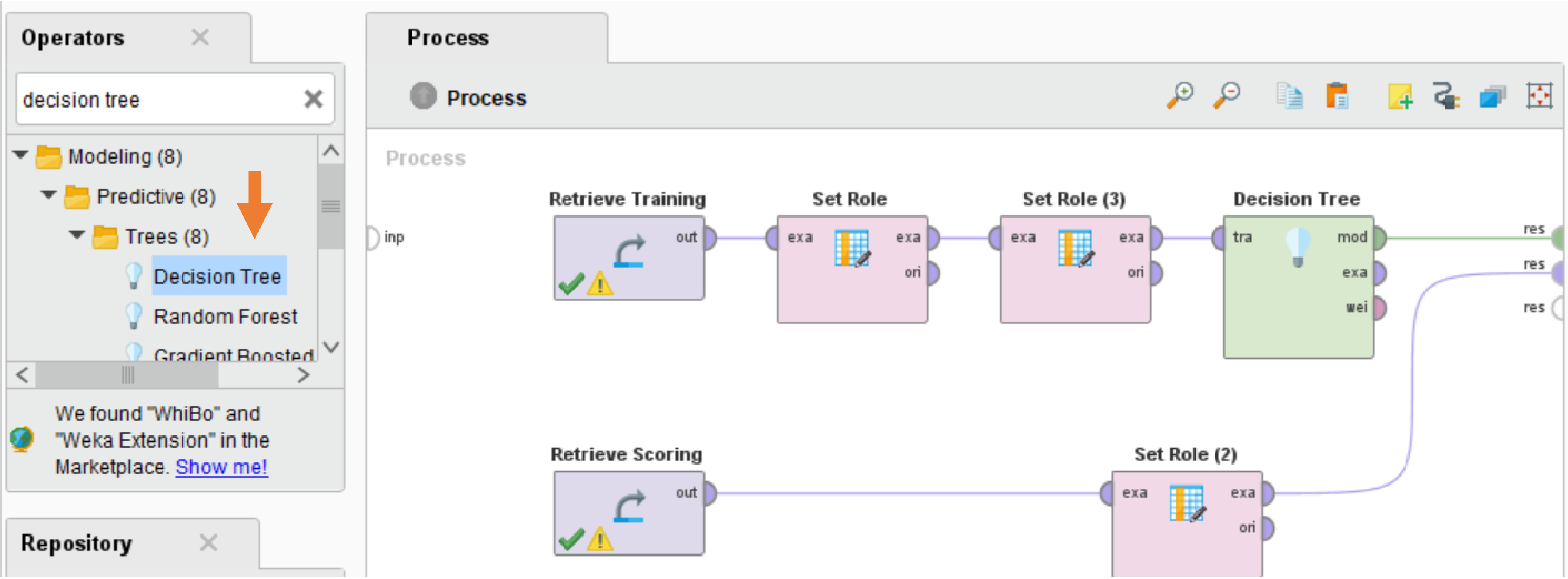
As with other predictive models, it is necessary to define the label attribute.

5. Add a Set Role operator to the training flow and set the “eReader_Adoption” attribute to ‘label’.



DATA PREPARATION

6. Next, search the Operators for “Decision Tree”. Select the basic operator from the Decision Tree and add it to your training flow.



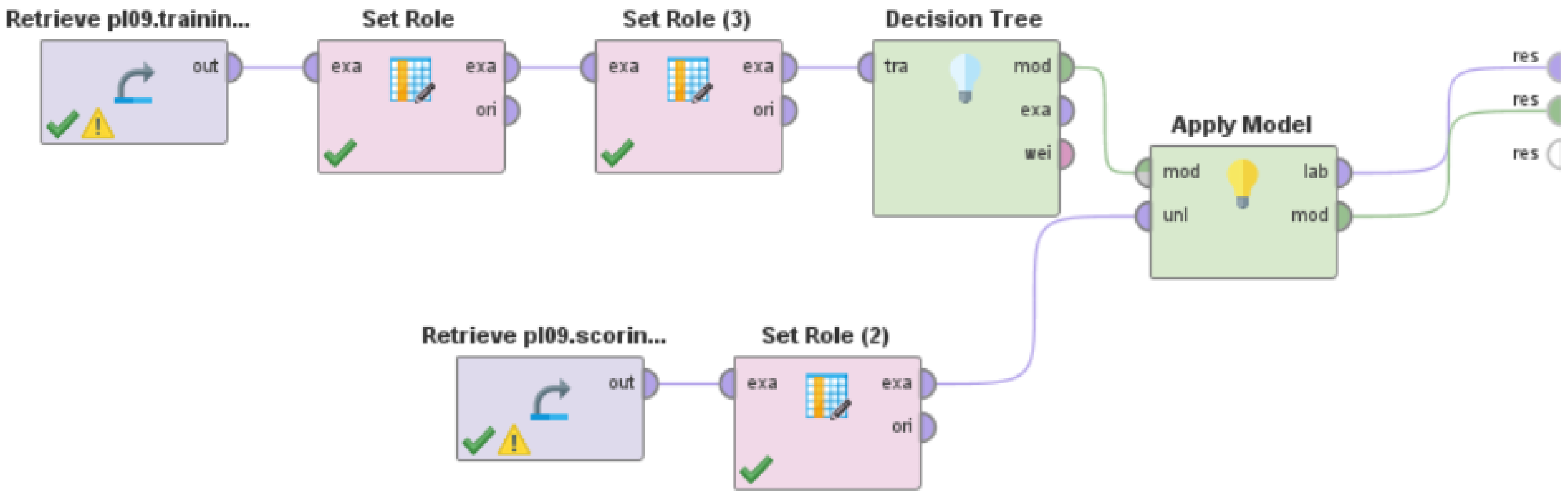
DATA PREPARATION

7. Run the model and switch to the *Tree (Decision Tree)* tab in the Results perspective. You will be able to see our preliminary tree, consisting of **nodes** (completely gray rectangles) and **leaves** (gray rectangles with a colored line in the background).

Nodes are attributes that serve as good predictors for the label attribute. The leaves are the multicolored end points that show us the distribution of categories from our label attribute that follow the branch of the tree to the point of that leaf.

MODELING

1. Switch to the Design perspective. On the Operators tab look for the ‘Apply Model’ operator and drag it to the process window, joining the training and scoring flows. Make sure that both the *lab* ports and the *mod* ports are connected to the *res* ports in order to generate the desired results.



MODELING

2. Run the model. Click the ‘ExampleSet’ tab next to the ‘Tree’ tab. The generated tree has been applied to our scoring data. Also, confidence attributes have been created by RapidMiner, along with a prediction attribute.

Id **User_ID**	Integer	0	Min 10153	Max 99694	Average 54647.074
Prediction **prediction(eReader_Adoption)**	Polynominal	0	Least Innovator (37)	Most Early Adopter (153)	Values Early Adopter (153), Late Majority (14
Confidence_Early Majority **confidence(Early Majority)**	Real	0	Min 0	Max 1	Average 0.287
Confidence_Late Majority **confidence(Late Majority)**	Real	0	Min 0	Max 1	Average 0.294
Confidence_Early Adopter **confidence(Early Adopter)**	Real	0	Min 0	Max 1	Average 0.288
Confidence_Innovator **confidence(Innovator)**	Real	0	Min 0	Max 1	Average 0.131
Gender	Polynominal	0	Least F (221)	Most M (252)	Values M (252), F (221)

MODELING

3. Change to the 'Data View' option where the prediction for each customer's adoption group is displayed, along with the confidence percentages for each prediction. There are four attributes of trust, corresponding to the four possible values on the label (eReader_Adoption).

Row No.	User_ID	prediction(e...	confidence(...	confidence(...	confidence(...	confidence(I...	Gender	Age	Marital_Stat...	Website_A
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57	S	Regular
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51	M	Regular
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41	M	Seldom
4	93666	Early Majority	1	0	0	0	M	66	S	Regular
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31	S	Seldom
6	64466	Early Majority	0.750	0.250	0	0	M	68	M	Regular
7	76655	Late Majority	0.065	0.879	0.056	0	F	51	S	Seldom
8	48465	Innovator	0	0.111	0	0.889	F	36	S	Frequent
9	19889	Late Majority	0	0.500	0.500	0	M	29	M	Regular
10	63570	Early Majority	1	0	0	0	M	61	M	Frequent
11	63239	Early Adopter	0.273	0.045	0.545	0.136	M	47	S	Regular
12	67603	Early Majority	0.950	0	0	0.050	F	62	S	Regular

MODELING

How can we interpret these values?

The confidence percentages add up to a total of 100% and measure how confident we are that the prediction will come true. The prediction corresponds to the category that produced the highest percentage of confidence.

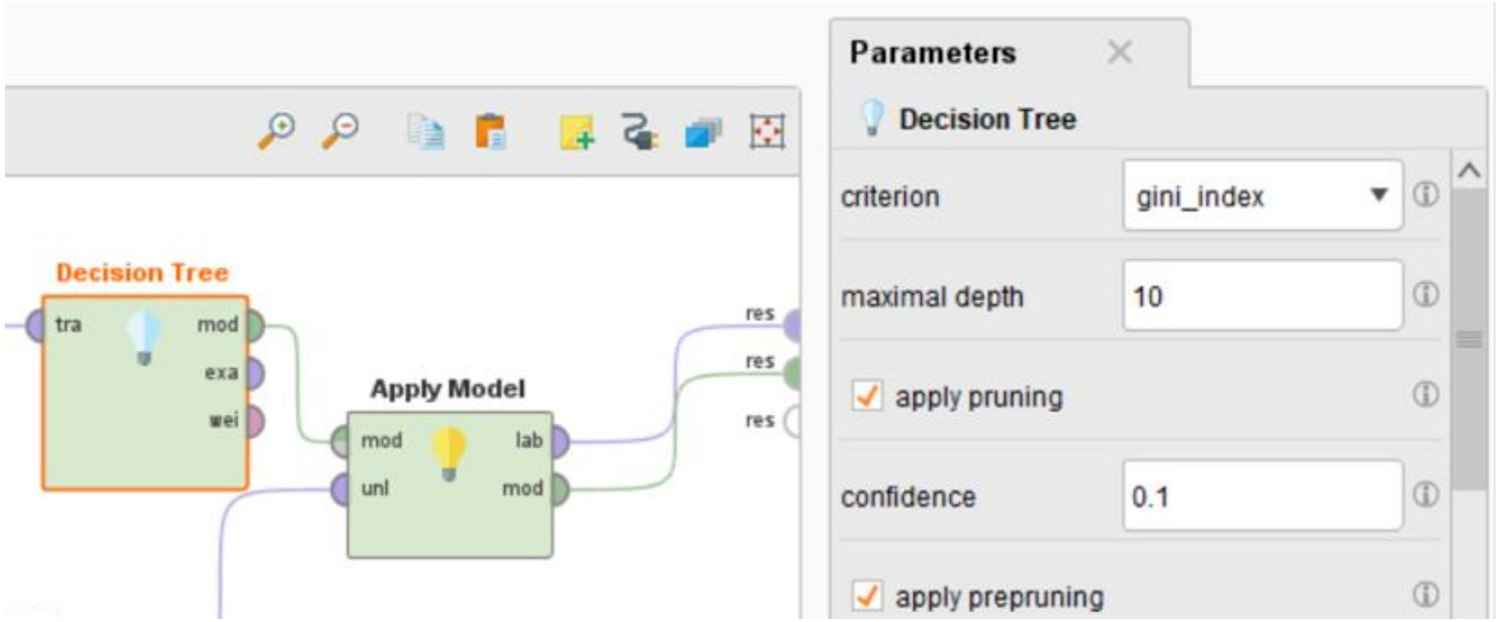
Row No.	User_ID	prediction(eReader_Adoption)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
5	72282	Late Majority	0.061	0.879	0.030	0.030
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.065	0.879	0.056	0

RapidMiner is quite (but not 100%) convinced that the 64466 person (line 6) is going to be a member of the 'early majority' (75%). Despite some uncertainty, RapidMiner is completely convinced that this person is not going to be an ‘early adopter’ (0%) nor a member of ‘innovator’ (0%).

MODELING

Remember that CRISP-DM is cyclical in nature, and that in some modeling techniques, especially those with less structured data, some back and forth trial-and-error can reveal more interesting patterns in data.

4. Switch back to design perspective, click on the Decision Tree operator, and in the Parameters area, change the ‘criterion’ parameter to ‘gini_index’. Save the process with “pl09-classification” name and execute the model.



EVALUATION

Through the analysis of the results we see that the tree has even more detail when using the gini_index criterion.

We could further modify the tree by going back to the Design window and changing the minimum number of items to form a node (minimal size for split) or the minimum size for a leaf (minimal leaf size).

Even with the default values for these parameters, we can see that the Gini algorithm itself is more sensitive than the Gain Ratio algorithm in identifying nodes and leaves.

EVALUATION

1. Switch to the ‘ExampleSet’ tab and choose the ‘Data View’ option. Changing the algorithm underlying the tree has changed, in some cases, our confidence in the prediction.

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Innovator	0.333	0	0	0.667	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.750	0.250	0	0	M	68
7	76655	Late Majority	0.333	0.667	0	0	F	51
8	48465	Innovator	0	0.250	0	0.750	F	36
9	19889	Early Majority	0.500	0	0.500	0	M	29
10	63570	Early Majority	1	0	0	0	M	61
11	63239	Early Majority	0.667	0	0.167	0.167	M	47
12	67603	Early Majority	0.917	0	0.042	0.042	F	62

EVALUATION

Take the customer on line 2 (ID 25913) as an example. According to the Gain Ratio criterion, this customer was calculated to have at least some percentage of probability of landing in any of the four categories of adopters. There was 54.5% certainty that he would be an early adopter, but almost 27.3% certainty that he could also become a member of the early majority.

Gain Ratio

Row No.	User_ID	prediction(...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51

Gini Index

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51

O Ricardo terá de decidir durante a fase de implementação a qual das categorias o cliente pertence. Mas talvez usando o critério *Gini Index*, seja possível ajudá-lo a decidir.

EVALUATION

According to the Gini Index criterion, this customer has an 87.5% chance of being an early adopter and only 12.5% of being an innovator. Note that the chances of him becoming part of the early majority and late majority have dropped to zero.

Although client 25913 may not be at the top of Ricardo's list when the implementation is released, it will probably be positioned higher than it would be if it were under the Gain Ratio criteria.

Gain Ratio

Row No.	User_ID	prediction(...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51

Gini Index

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51

EVALUATION

Note that, although the Gini Index criterion has changed some of the predictions, it has not affected all. Double check person ID 64466. There is no difference in this person's predictions under any of the algorithms. Sometimes the level of confidence in a forecast through a decision tree is so high that a more sensitive underlying algorithm does not change the values of that forecast at all.

Gain Ratio

Row No.	User_ID	prediction(eReader_Adoption)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.065	0.879	0.056	0

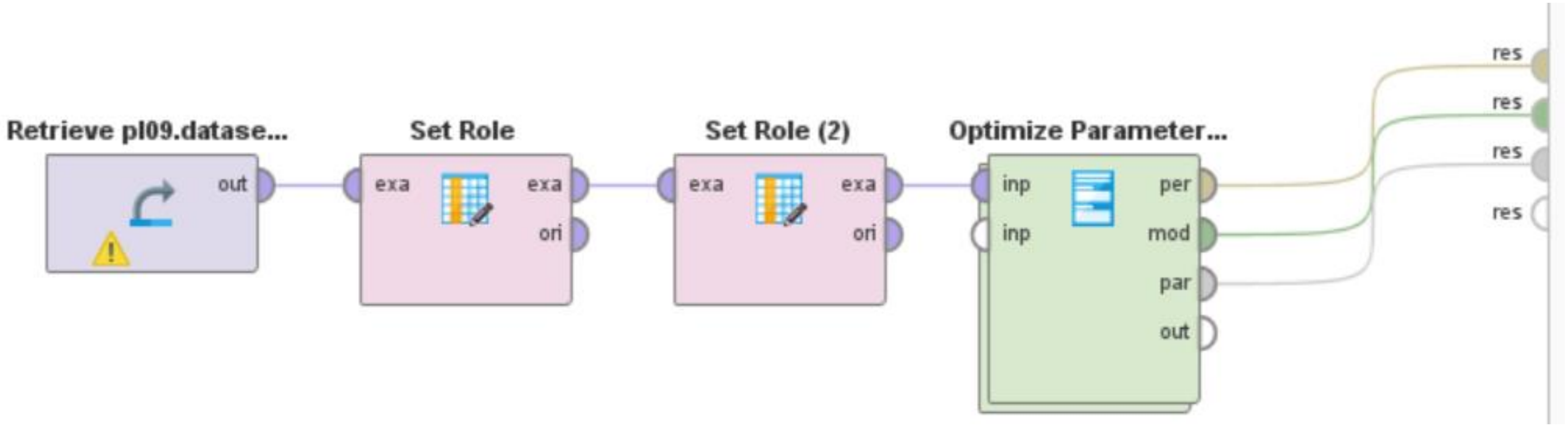
Gini Index

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.333	0.667	0	0

EVALUATION

After this initial approach, it is important to realize that the parameters that were defined in the Decision Tree operator are probably not the most suitable to achieve the best possible result. Thus, it is important to try to find the best values that the parameters can have in order to maximize the performance of the model.

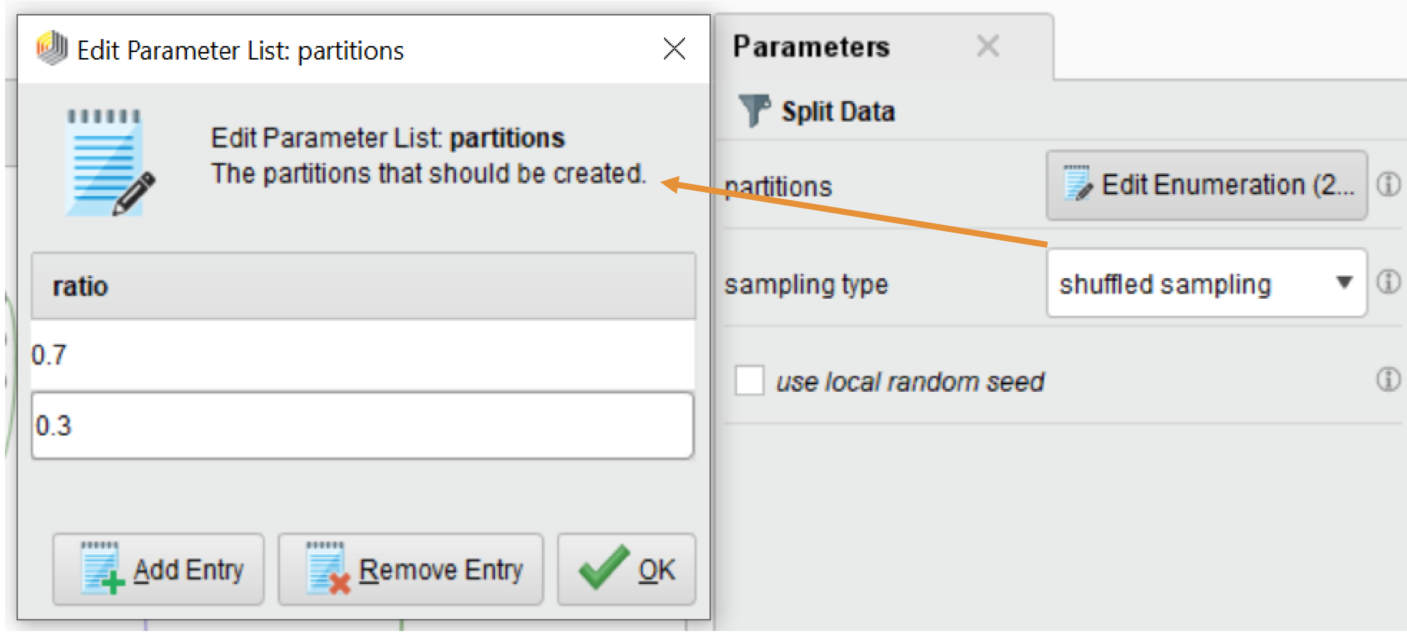
2. Create and save a new process (named OptimizeParameters) and drag the training dataset into that process. Use the two Set Role operators (one for the ID and one for the label attribute) again, as before. Look for the Optimize Parameters (Grid) operator and drag it into the process. Connect the first 3 ports of this operator to the *res* ports.



EVALUATION

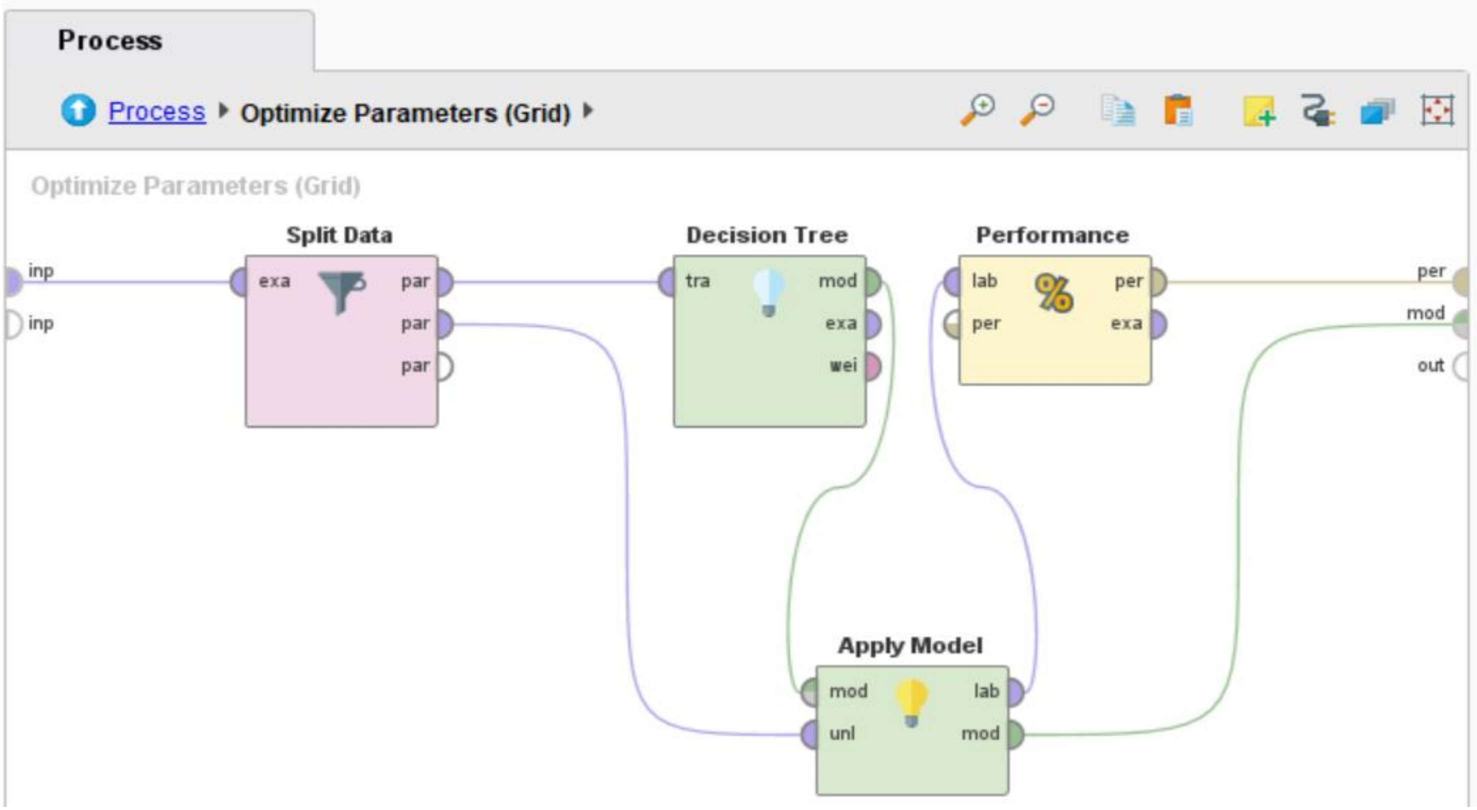
The *Optimize Parameters (Grid)* operator is a nested operator. It executes the subprocess it incorporates for all combinations of values of the selected parameters and then returns the ideal values of these parameters. We now need to incorporate the sub-process that we want to repeat, within our optimization operator, that is, our classification with Decision Tree.

3. Double-click the *Optimize Parameters (Grid)* operator. A new sub-process window will open. Start the sub-process with a Split Data operator, since for this case it will be necessary to split the dataset, so that later it is possible to evaluate the model's accuracy. Set the parameters as in the image.



EVALUATION

4. Then, add the *Decision Tree* and *Apply Model* operators as shown in the figure. This time, a Performance operator will be added that will allow to statistically evaluate the performance of the classification model. By default, this assessment will be done through accuracy.



EVALUATION

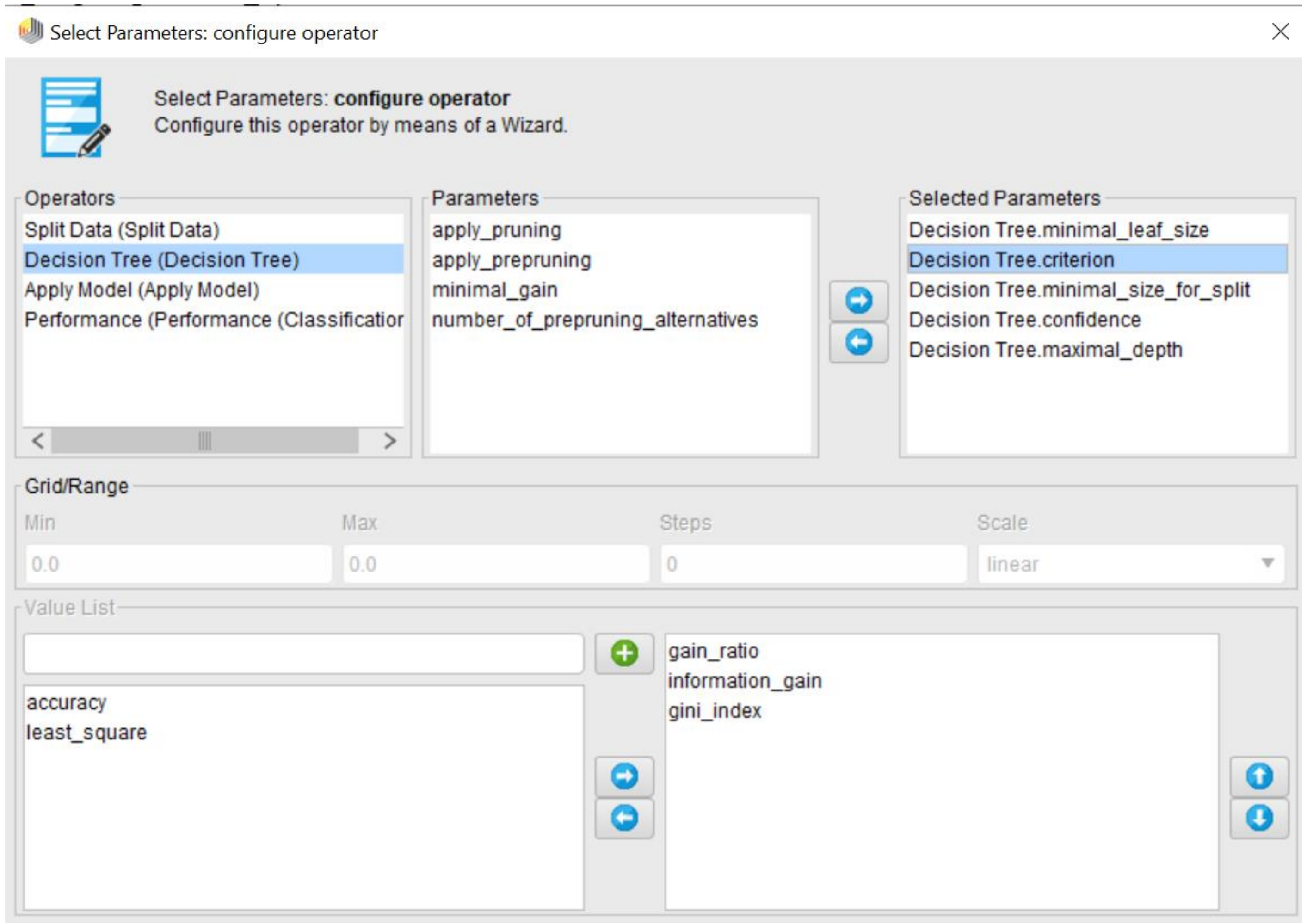
Now it is necessary to indicate in the optimization operator which parameters we want to optimize, in this case, the parameters associated with the decision tree.

5. Go back by clicking on “Process” in the process bar.

6. Click once on the *Optimize Parameters (Grid)* operator and in the parameters panel, on the right side, click *Edit Parameters Settings.*

EVALUATION

A new window opens where we can choose the parameters to be optimized. We will first choose the Decision Tree from the list of Operators and then choose the desired Parameters, sending them to the list on the right. It is important to keep in mind that regarding the criterion parameter, it is necessary to remove the accuracy and least_square values from the list, as they are not suitable for our model.



EVALUATION

7. Finally, run the model. Note that the more optimization parameters you choose, the slower it will run. In the *Results* window we have several tabs. The *ParameterSet* tab shows the best result (accuracy) obtained during all the iterations that were made, and what are the values of the parameters used to obtain this result. The *Optimize Parameters* tab shows the iterations made for each parameter.

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 74.40%
ConfusionMatrix:
True:   Early Majority  Late Majority   Early Adopter   Innovator
Early Majority: 59      5       21      5
Late Majority:  3       62      4       2
Early Adopter:  9       1       46      9
Innovator:      1       1       3       19
]
Decision Tree.criterion = gain_ratio
Decision Tree.minimal_size_for_split    = 21
Decision Tree.maximal_depth     = 39
Decision Tree.confidence        = 0.25000005
```

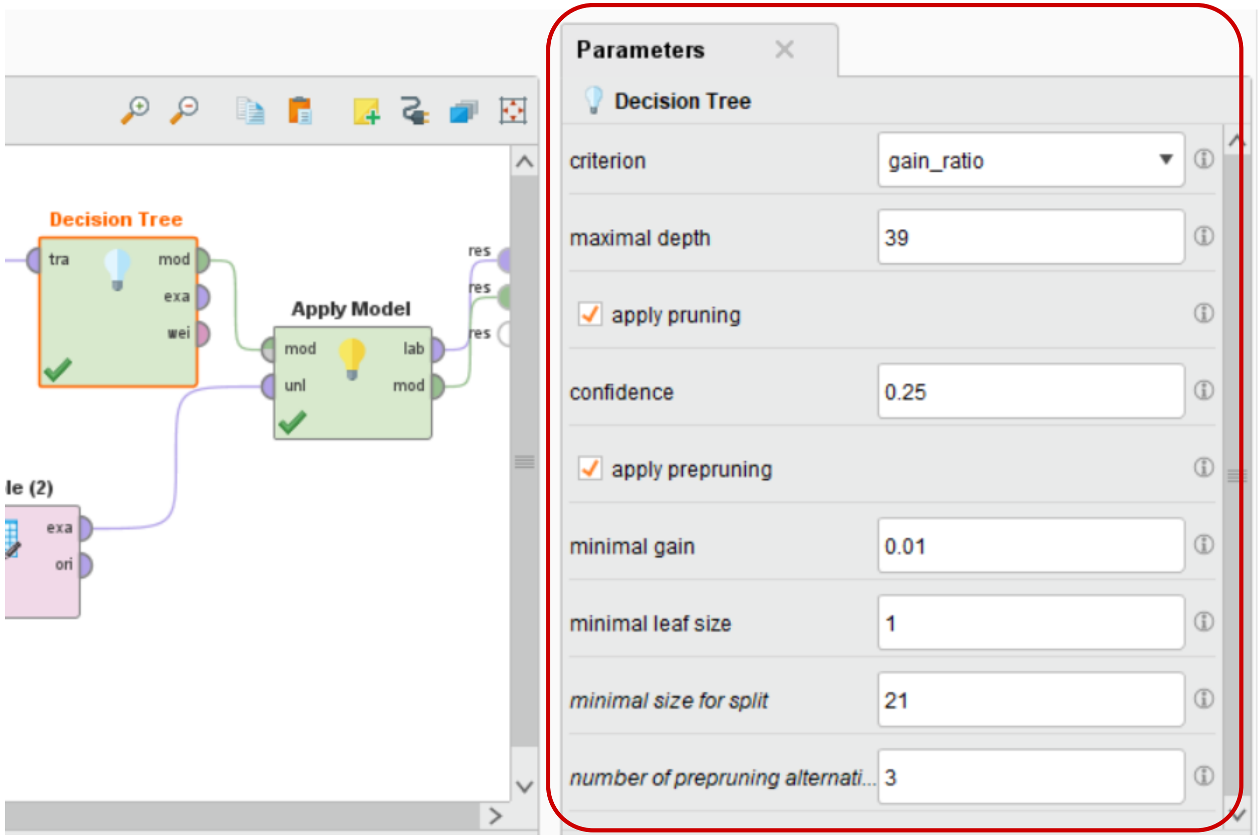
Optimize Parameters (Grid) (3993 rows, 6 columns)

iteration	Decision Tree.criterion	Decision Tree.minimal_size_...	Decision Tree.maximal_...	Decision Tree.confidence	accuracy
1001	information_gain	31	80	0.100	0.700
501	gini_index	11	39	0.050	0.632
1002	gini_index	31	80	0.100	0.632
1	gain_ratio	1	-1	0.000	0.636
502	gain_ratio	21	39	0.050	0.688
1003	gain_ratio	41	80	0.100	0.652
503	information_gain	21	39	0.050	0.664
2	information_gain	1	-1	0.000	0.588
1004	information_gain	41	80	0.100	0.672
504	gini_index	21	39	0.050	0.684
1005	gini_index	41	80	0.100	0.656
505	gain_ratio	31	39	0.050	0.696
1006	gain_ratio	51	80	0.100	0.688
3	gini_index	1	-1	0.000	0.588

EVALUATION

Now that we have discovered the optimized values of the Decision Tree operator parameters, we can go back to the previous process (pl09-classification) to try to obtain better results in the classification of the test dataset.

9. Back to the other process (pl09-classification), replace the values of the parameters criterion, minimal_size_for_split, maximal_depth and confidence with the values found.



EVALUATION

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Innovator	0.333	0	0	0.667	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.750	0.250	0	0	M	68
7	76655	Late Majority	0.333	0.667	0	0	F	51
8	48465	Innovator	0	0.250	0	0.750	F	36

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Innovator	0	0	0.357	0.643	M	57
2	25913	Early Adopter	0	0	0.800	0.200	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Early Majority	1	0	0	0	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.815	0.111	0	0.074	M	68
7	76655	Late Majority	0.063	0.874	0.049	0.014	F	51
8	48465	Innovator	0	0.111	0	0.889	F	36

EVALUATION

Updating the values of the parameters of the decision tree according to the data found, resulted, as expected, in an increase in the confidence of the predictions made by the tree.

This is in line with the logic behind the CRISP-DM methodology that argues that the Data Mining process is cyclical, being possible to go back as many times as necessary to redo and readjust the final model in order to obtain satisfactory results.

With these results, Ricardo now has the information and knowledge necessary to achieve his goals.

DEPLOYMENT

Ricardo’s original desire was to be able to figure out which customers he could expect to buy the new eReader and on what time schedule, based on the company’s last release of a high-profile digital reader. The decision tree has enabled him to predict that and to determine how reliable the predictions are. He’s also been able to determine which attributes are the most predictive of eReader adoption,

But how will he use this new found knowledge?

The simplest and most direct answer is that he now has a list of customers and their probable adoption timings for the next-gen eReader. These customers are identifiable by the User_ID that was retained in the results perspective data but not used as a predictor in the model. He can segment these customers and begin a process of target marketing that is timely and relevant to each individual.

DEPLOYMENT

Those who are most likely to buy the product early (**early adopter**) can be contacted and encouraged to buy as soon as the new product comes out and may even want the option to pre-order the new device.

Those who are less likely (predicted early majority) might need some persuasion, perhaps a free digital book or two with eReader purchase or a discount on digital music playable on the new eReader.

The least likely (predicted late majority), can be marketed to passively, or perhaps not at all if marketing budgets are tight and those dollars need to be spent incentivizing the most likely customers to buy.

On the other hand, perhaps very little marketing is needed to the predicted innovators, since they are predicted to be the most likely to buy the eReader in the first place.

DEPLOYMENT

Ricardo now has a tree that shows him which attributes matter most in determining the likelihood of buying for each group.

New marketing campaigns can use this information to focus more on increasing web site activity level, or on connecting general electronics that are for sale on the company's web site with the eReaders and digital media more specifically.

These types of cross-categorical promotions can be further honed to appeal to buyers of a specific gender or in a given age range.

Ricardo has much that he can use in this rich data mining output as he works to promote the next-gen eReader.

RESUME

Decision trees are excellent **predictive models** when the target attribute is **categorical** in nature, and when the data set is of **mixed types**.

Decision trees have the advantage of effectively handling attributes that have missing or inconsistent values that are not handled—decision trees will work around such data and still generate usable results.

Decision trees are made of **nodes** and **leaves** (connected by labeled branch arrows), representing the best predictor attributes in a dataset. These nodes and leaves lead to **confidence** percentages based on the actual attributes in the training dataset, and can then be applied to similarly structured scoring data in order to generate predictions for the scoring observations.

Decision trees tell us **what is predicted**, **how confident** we can be in the prediction, and **how we arrived** at the prediction. The 'how we arrived at' portion of a decision tree's output is shown in a **graphical view** of the tree.