Exploring RapidMiner

Decision Trees

PL09

[Hugo Peixoto] – 2018/2019, Universidade do Minho

Material

http://hpeixoto.github.io/dc

CONTEXT AND PERSPECTIVE

Richard works for a large online retailer.

His company is **launching a next-generation eReader** soon, and they want to **maximize the effectiveness** of their marketing.

Richard 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.

CONTEXT AND PERSPECTIVE

Richard believes that by mining the **customers' data regarding general consumer behaviors on the web site**, he'll be able to figure out which customers will buy the new eReader early, which ones will buy next, and which ones will buy later on.

He hopes that by predicting when a customer will be ready to buy the next-gen eReader, he'll be able to time his target marketing to the people most ready to respond to advertisements and promotions.

BUSINESS UNDERSTANDING

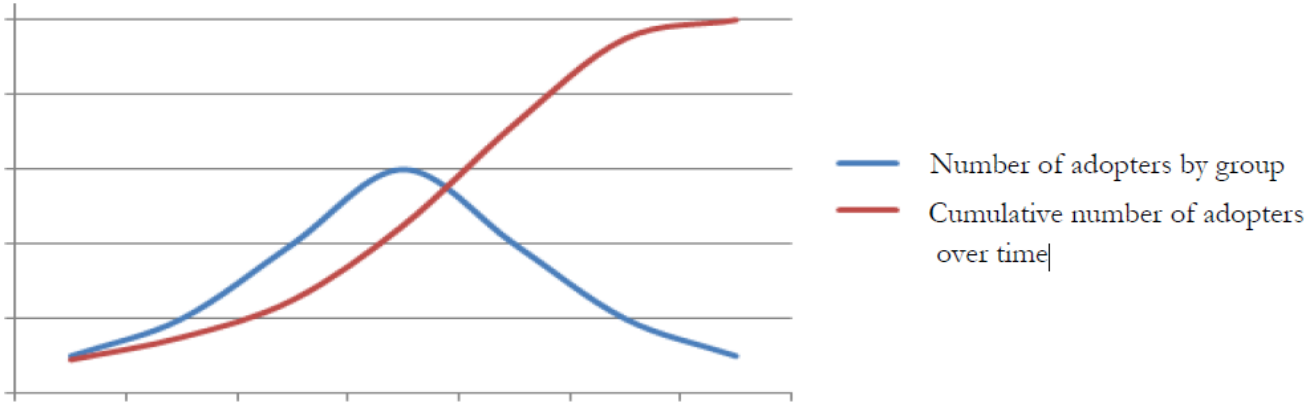


Figure 10-1. Everett Rogers' theory of adoption of new innovations.

Innovators
Early Adopters
Early Majority
Late Majority

BUSINESS UNDERSTANDING

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

Two datasets:

The training data set contains the web site activities of customers who bought the company's previous generation reader, and the timing with which they bought their reader.

The second is comprised of attributes of current customers which Richard hopes will buy the new eReader.

DATA UNDERSTANDING

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. This is calculated to the nearest year by taking the difference between the system date and the person's birthdate as recorded in their account.

Marital_Status: The person's marital status as recorded in their account. People who indicated on their account that they are married are entered in the data set as 'M'. Since the web site does not distinguish single types of people, those who are divorced or widowed are included with those who have never been married (indicated in the data set as 'S').

DATA UNDERSTANDING

Website_Activity: This attribute is an indication of how active each customer is on the company's web site. Working with Richard, we used the web site database's information which records the duration of each customers visits to the web site to calculate how frequently, and for how long each time, the customers use the web site. This is then translated into one of three categories: Seldom, Regular, or Frequent.

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 (such as MP3 music) in the past year and a half. This attribute does not include digital book purchases.

Bought_Digital_Books: Richard believes that as an indicator of buying behavior relative to the company's new eReader, this attribute will likely be the best indicator. Thus, this attribute has been set apart from the purchase of other types of digital media. Further, this attribute 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 data set. 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 csv:
http://hpeixoto.github.io/dc/pl09/pl09.dataset.scoring.csv

http://hpeixoto.github.io/dc/pl09/pl09.dataset.training.csv

Import csv to rapidminer repository.

You do not need to worry about attribute data types because the **Decision Tree operator can handle all types of data.**

DATA PREPARATION

Rename to "Training" / "Scoring":

DATA PREPARATION

User_ID ?????

Select Atributes (aula 04)

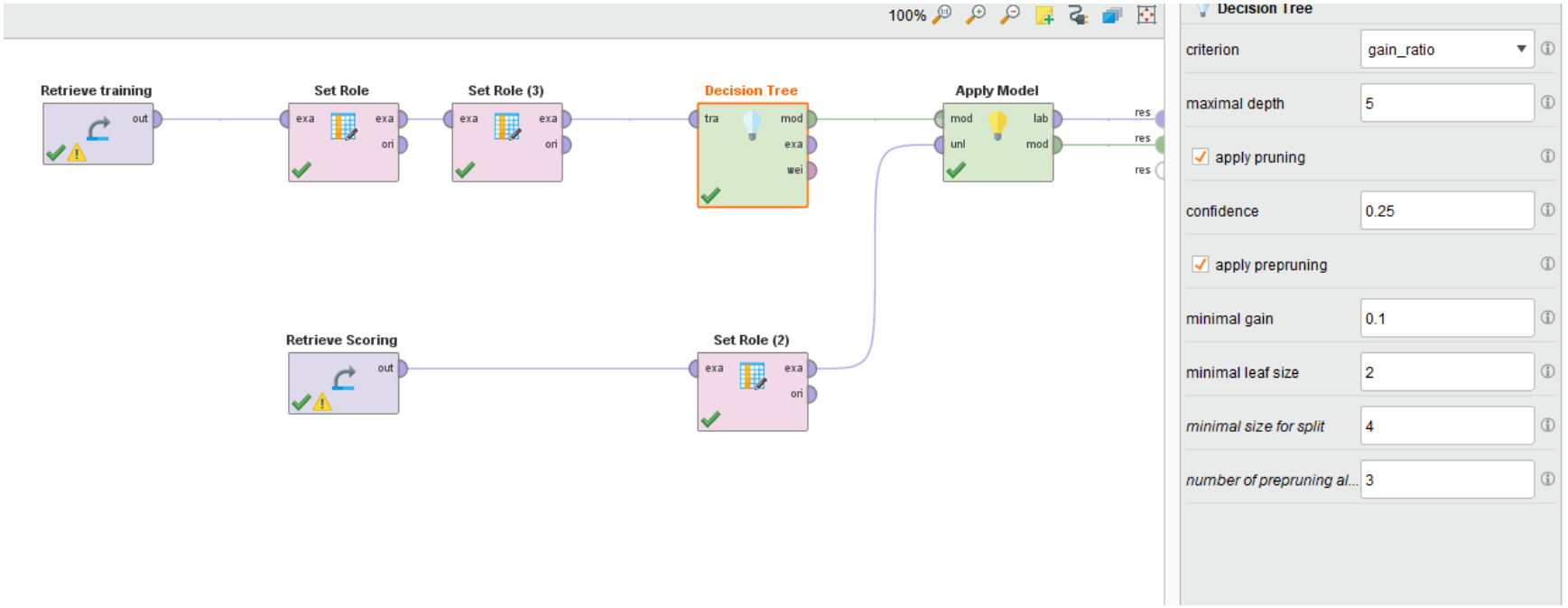
Set Role – novo método

DATA PREPARATION

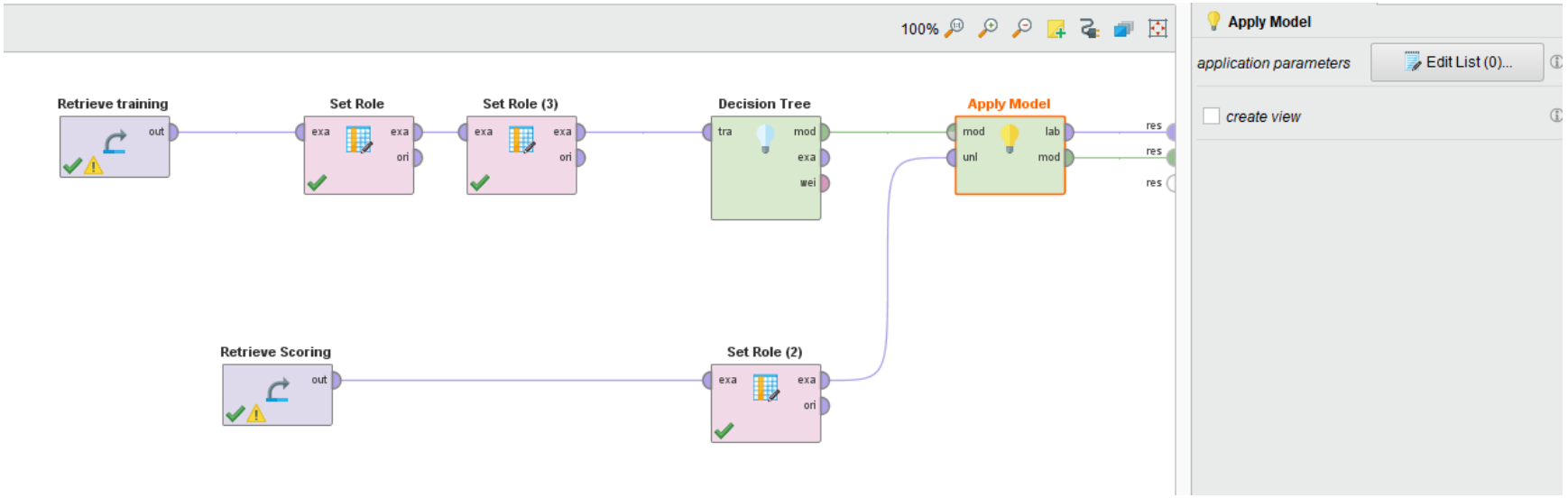
Set USER_ID role to ID with "Set Role"

DATA PREPARATION

Add "Decision Tree" and adjust the parameters



MODELING

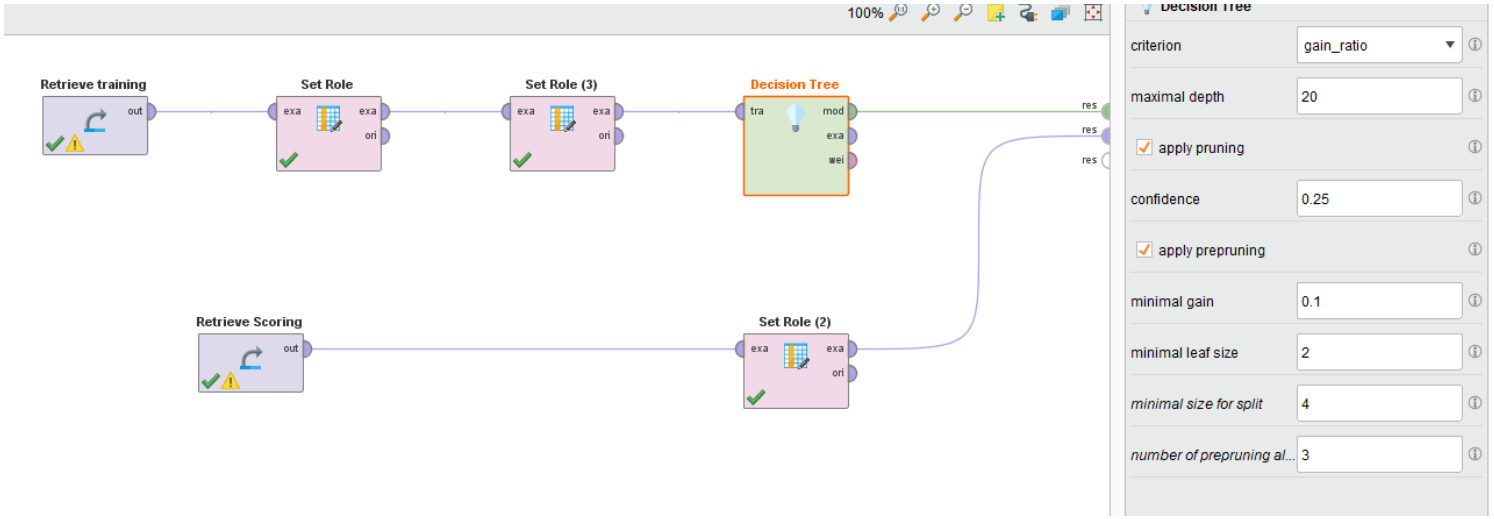

100%
Retrieve training
out
Set Role
exa
exa
ori
Set Role (3)
exa
exa
ori
Decision Tree
tra
mod
exa
wei
Apply Model
mod
lab
unl
mod
res
res
res
Retrieve Scoring
out
Set Role (2)
exa
exa
ori
Apply Model
application parameters
Edit List (0)...
create view

MODELING

Row No.	User_ID	prediction(eReader_Adoption)	confidence(Late Majority)	confidence(l...	confidence(...	confidence(...	Gender	Age	Marital_Stat...	Website_Ac...	Browsed_El...	Bought_Elec...	Bought
1	56031	Innovator	0.030	0.576	0.318	0.076	M	57	S	Regular	Yes	Yes	Yes
2	25913	Early Adopter	0	0	1	0	F	51	M	Regular	Yes	Yes	No
3	19396	Late Majority	0.751	0.021	0.053	0.175	M	41	M	Seldom	Yes	Yes	Yes
4	93666	Early Majority	0.250	0	0	0.750	M	66	S	Regular	Yes	Yes	Yes
5	72282	Late Majority	0.751	0.021	0.053	0.175	F	31	S	Seldom	Yes	No	Yes
6	64466	Early Majority	0.250	0	0	0.750	M	68	M	Regular	Yes	Yes	Yes
7	76655	Late Majority	0.751	0.021	0.053	0.175	F	51	S	Seldom	Yes	No	No

MODELING

Change maximal depth to 20



EVALUATION

- Take a minute to explore around the tree model.

SUMMARY

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 are made of nodes and leaves (connected by labeled branch arrows), representing the best predictor attributes in a data set.

These nodes and leaves lead to confidence percentages based on the actual attributes in the training data set, and can then be applied to similarly structured scoring data in order to generate predictions for the scoring observations.

FE07

Exploring RapidMiner

Decision Trees

PL09

[Hugo Peixoto] – 2018/2019, Universidade do Minho