MIS772

Predictive Analytics

Workshop: Text Mining

Data preparation, text mining, dimensionality reduction with weight-select method, Decision Tree Classification

DEAKIN BUSINESS SCHOOL

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

Workshop Plan

Objectives:

Your task is to create a predictive text mining model, to parse text of real estate ads, create a classification model which utilises attributes derived from text to predict the price of properties (i.e. using word descriptions to predict price classification (expensive or affordable)).

In addition to text mining, conversion of attribute types will be illustrated.

Data Set:

Use the dataset "Melb Real Train.csv" available on the unit site.

Method:

Attend the workshop, follow the tutor's demo and instructions, take notes. Note that the online seminar will be recorded for later reference.

This seminar requires the ***RapidMiner Text Processing 9.4*** *extension or later*

1 **Acquire data**

(a) Load the real estate data and unzip

(b) Convert data types (nominal to text, date to numerical)

(c) Replace missing values

(d) Filter out examples with missing values

(e) Discretize price (affordable vs expensive) as label

(f) Read and explore the data set

2 **Create a Text Mining process**

(a) Add text processing (process documents from data)

(b) Transform, Tokenize, Stem, Filer out stopwords and tokens by length

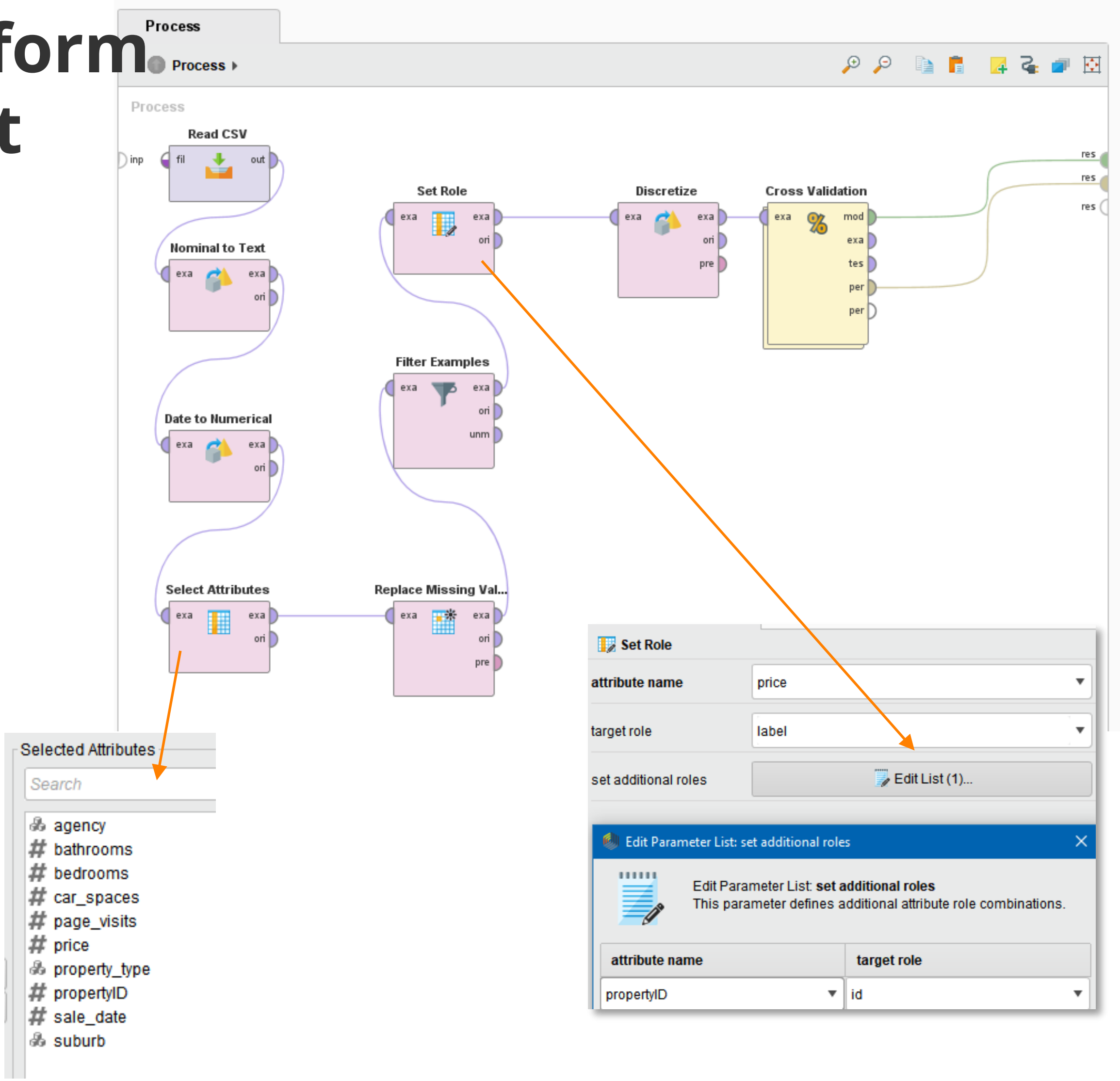
(c) Use weight-select dimensionality reduction

(d) Use k-fold Cross validation and train a decision tree

(e) Investigate performance

Acquire data, transform and prepare for text mining

Turn nominal attributes "description" into text. Then split date into month. Select attributes (no text).
Set missing car_spaces to zero. Check missing values and eliminate them by replacement or filtering (check which is more appropriate).
Check price distribution and discretise it into Affordable / Expensive classes, breaking at $700,000.
Set role of price (label), and ensure the role of PropertyID is set to "id".



Selected Attributes

Search

- agency
- bathrooms
- bedrooms
- car_spaces
- page_visits
- price
- property_type
- propertyID
- sale_date
- suburb

Set Role

attribute name	price
target role	label
set additional roles	Edit List (1)...

Edit Parameter List: set additional roles

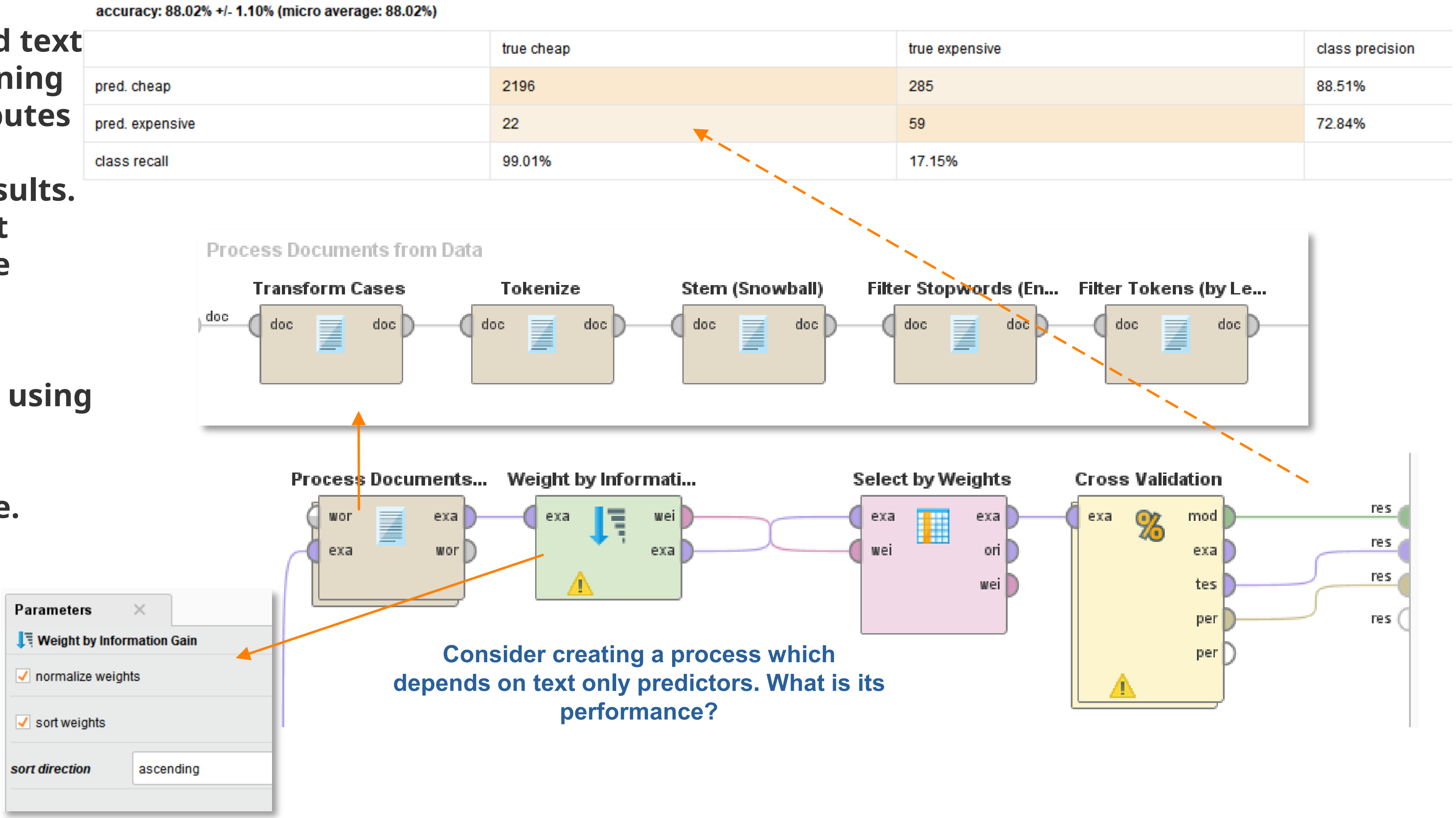
Edit Parameter List: **set additional roles**
This parameter defines additional attribute role combinations.

attribute name	target role
propertyID	id

Create a Model With Text Mining

Add text attributes. Add text processing (TF-IDF, pruning 2-40%, select text attributes weight(top k, k=50). Run and analyse the results. Check if addition of text attributes improved the model performance.

Reduce dimensionality using weight-select method. Consider models: k-NN and Decision Tree.



accuracy: 88.02% +/- 1.10% (micro average: 88.02%)

	true cheap	true expensive	class precision
pred. cheap	2196	285	88.51%
pred. expensive	22	59	72.84%
class recall	99.01%	17.15%	

Parameters

Weight by Information Gain

☑ normalize weights

☑ sort weights

sort direction: ascending

Consider creating a process which depends on text only predictors. What is its performance?