RapidMiner

Lesson 3

Kyle Goslin

College of Computing Technology

In this lesson, we will attempt to perform a simple data mining exercise. The process we will be using the K-Nearest neighbour algorithm.

An important thing to note about this process is that we will have two different datasets that we will be using. First we will be using what is called a “Training set”. This set can be seen as a collection of data that has already been classified.

After this we will be using a “Test set”. This set contains all of the information that we are providing, but it does not have a classification added. We will be using the knn algorithm to attempt and classify these records based upon what it has already learned from the training set.

Below is an example of a training set, you will notice that the “Non_comp” field has values added.

	A	B	C	D	E	F	G
1	married	children	getting_welfare	more_than_1	offered_courses	refused	non_comp
2	1	1	1	0	3	1	1
3	0	1	1	0	2	1	1
4	1	2	1	0	1	0	1
5	1	3	1	0	1	0	0
6	1	1	1	1	1	0	0
7	1	1	1	1	1	0	0
8	1	1	1	1	2	2	0
9	0	1	1	0	3	3	1
10	0	2	1	1	2	2	1
11	0	3	1	1	1	1	1
12	0	1	1	1	0	1	0
13							

Below is the test set, you will notice it does not have values for the “Non_comp” field. These are the values that we want to predict!

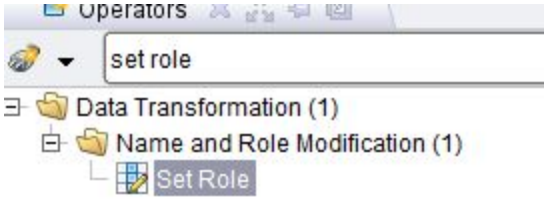
A	B	C	D	E	F	G	H	I
married	children	getting_welfa	more_than_1	offered_cours	refused	non_comp		
1	1	1	0	3	1			
0	1	1	0	2	1			
1	2	1	0	1	0			

To read in both of these files into RapidMiner, we first need to add two of the “Read Excel” operators onto the main canvas. This is done the same as Lesson 2, where they are clicked and dragged across from the toolbox on the left hand side of the screen.

Once you have done this, the first operator should be pointing to the “training” excel file and the second operator should be pointing to the “Test” excel file.

After the two excel files have been read in the next step which needs to be done is adding two of the “Set Role” operators. The purpose of this is to tell RapidMiner which column we are most interested in.

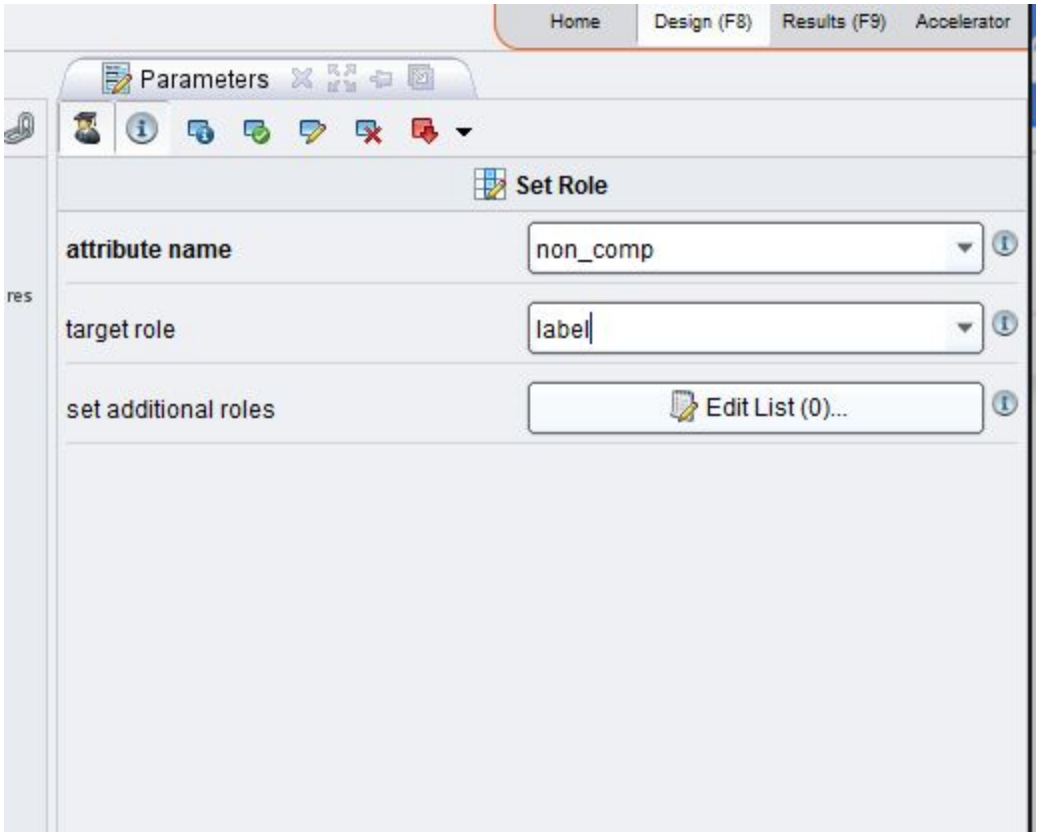
Click and drag two of these across onto the main canvas.



Once you have added these, your canvas should look like the picture below. After you have done this, click the first Set Role operator and then look to the right hand side where we will set the preferences for this operator.

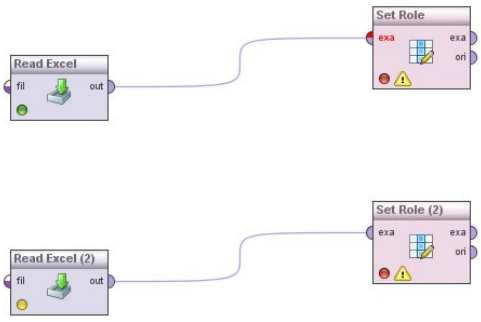
We need to tell RapidMiner that we want to set the attribute column titled “non_comp” and we want to set it as the “label” which is added to the second box.

This will tell RapidMiner that this is the most important column we are interested in.

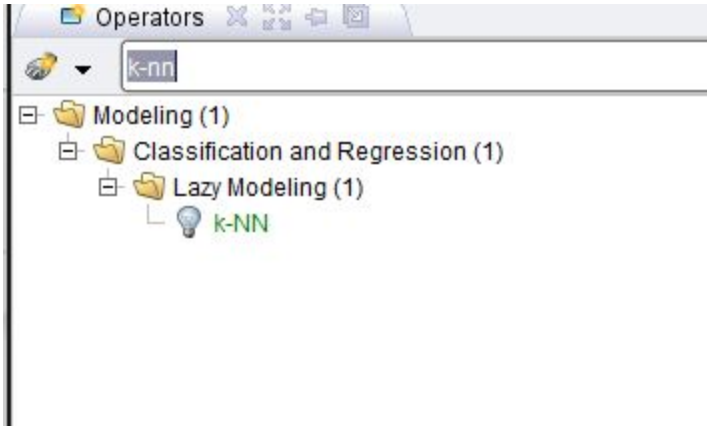


Once you have done this, repeat this step for the second “Set Role” operator.

After we have added our read Excel for both of the files and then we have added the Set Role, we can then connect the Excel Reader to the set role operator for both of them as shown below.

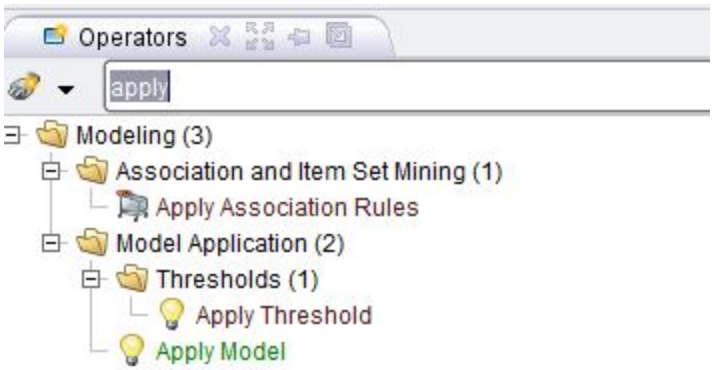


Next we will then begin the process of added the type of classifier we want to used, which is titled K-nn inside RapidMiner. Search the toolbox on the left hand side for this operator.



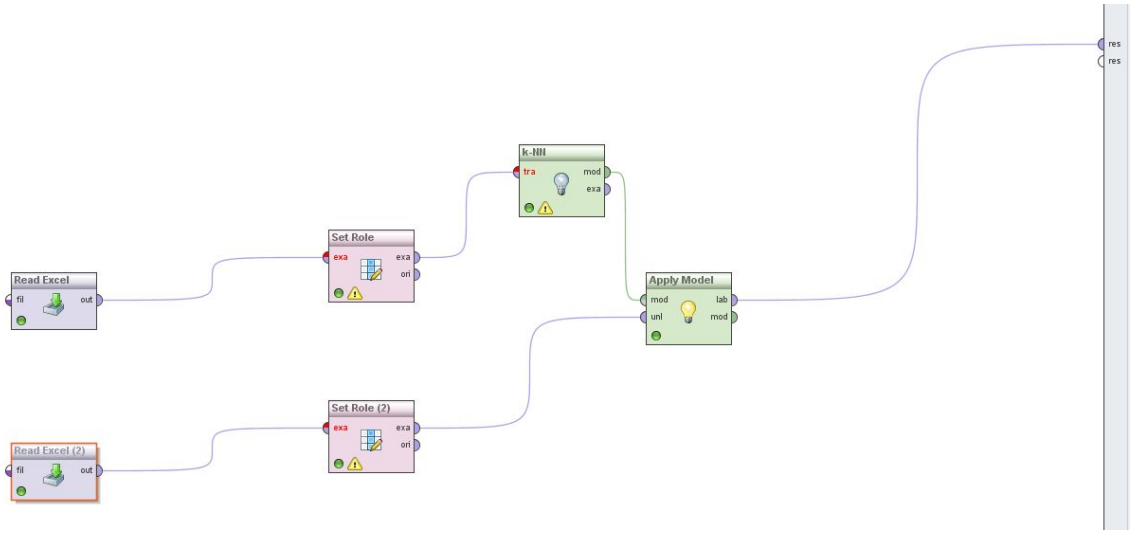
Click and drag this across to the main canvas.

Next we will need to include an “Apply Model” operator. This is used for actually applying the model to the data containing a column we want to predict. Search the toolbox for this operator also.



Click and drag the “Apply Model” operator across to the main canvas.

After you have added all the operators, connect them together so your canvas looks like the image below:



When you have finished, click the blue play button

After pressing the play button, the results window will open. You will notice an additional column will be added titled “prediction” this is the value that the system has predicted for this column based upon the training set that you have given it.

Row No.	non_comp	prediction(n...	married	children	getting_welf...	more_than...	offered_cou...	refused
1	?	1	1	1	1	0	3	1
2	?	1	0	1	1	0	2	1
3	?	1	1	2	1	0	1	0