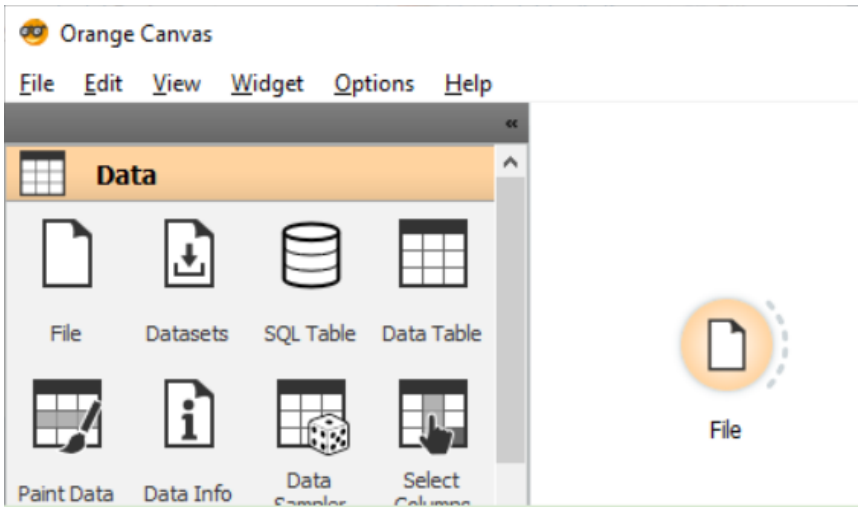
CSCU9M5 Practical 3

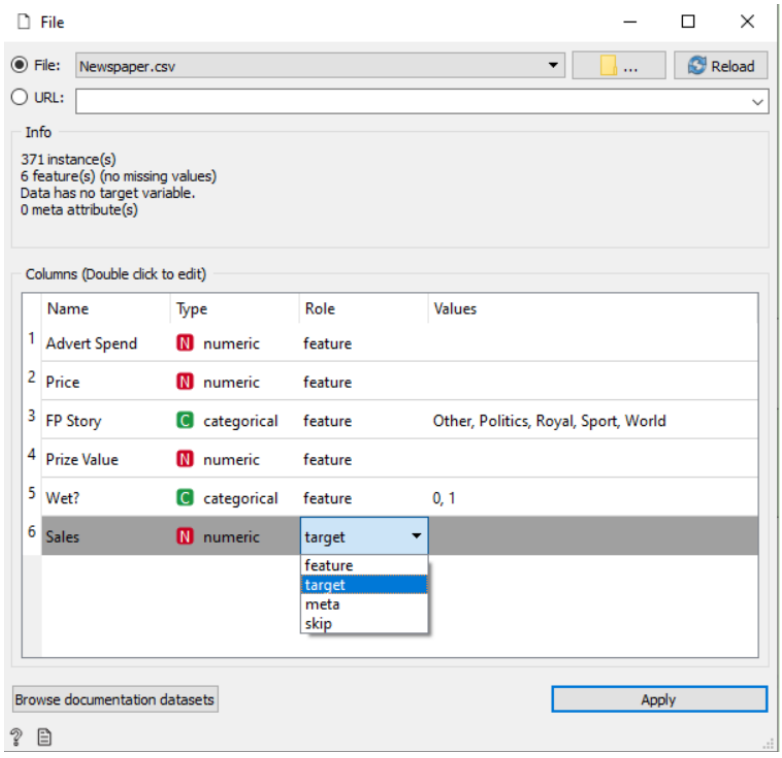
Linear Regression Models

In this exercise, you will build the regression models from the newspaper data discussed in the videos. Download this .csv file with the data we will be working with: Newspaper.csv

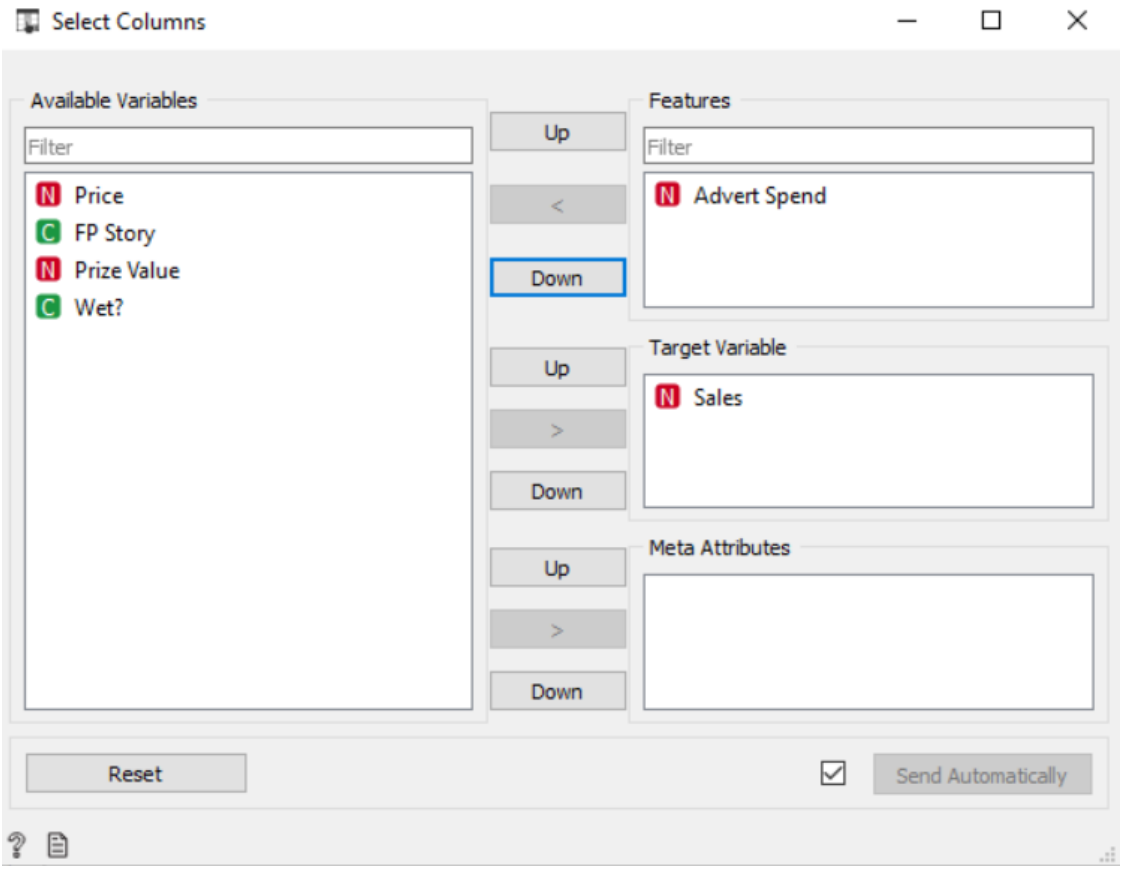
Run Orange and then open the file like this:



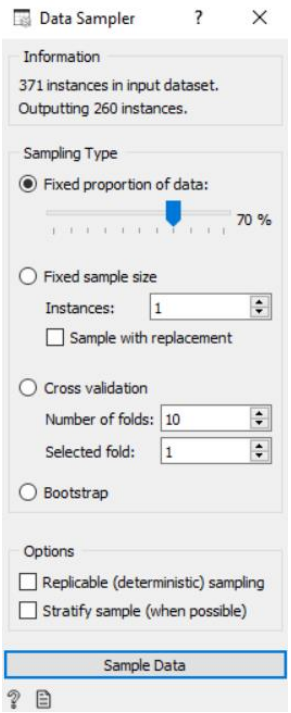
Now click on the File widget to choose the newspaper file. Set the *Sales* variable to be a target as shown.



In this example, we only need the Advert Spend variable, so we use a Select Columns widget to select that.

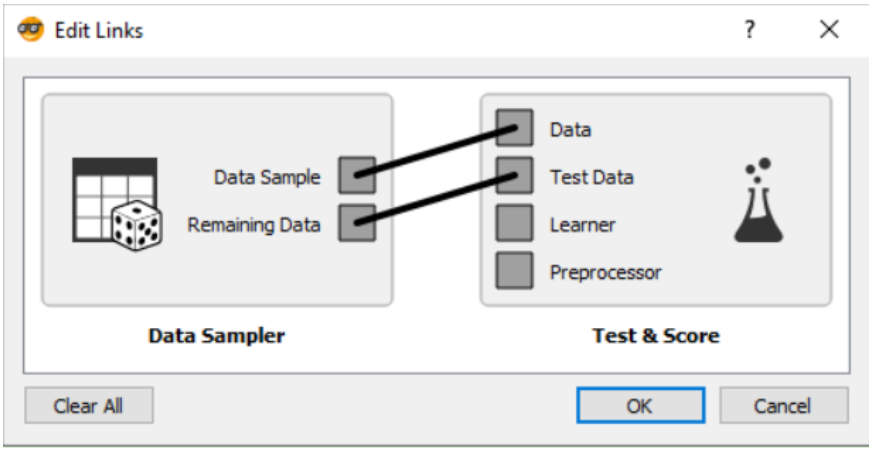


The next thing to do, as always, is to separate some test data. Do that using the data sampler. Here we extract 70% for testing.

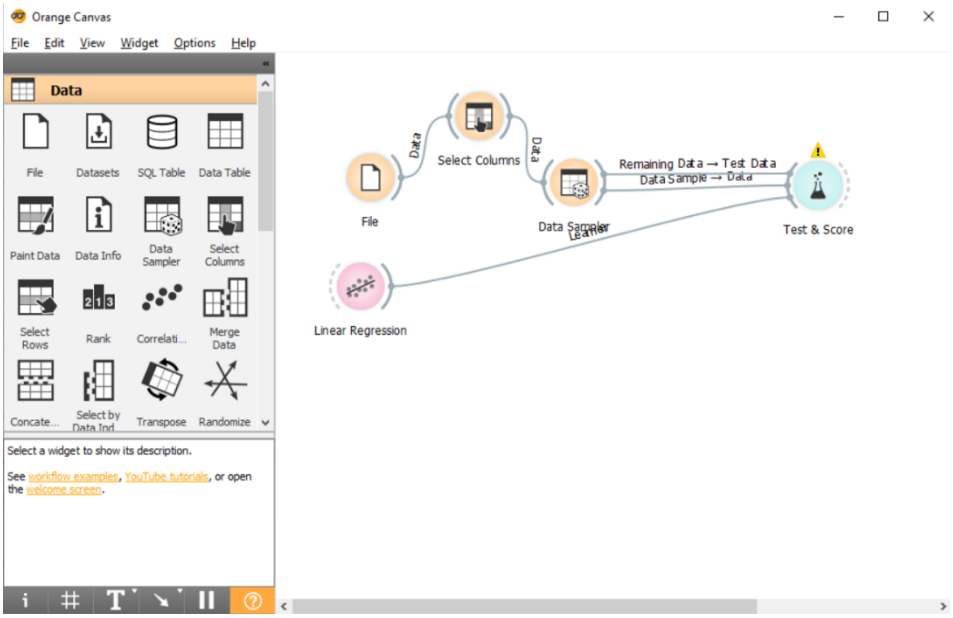


Now we choose Linear Regression and a Test and Score widget for validation.

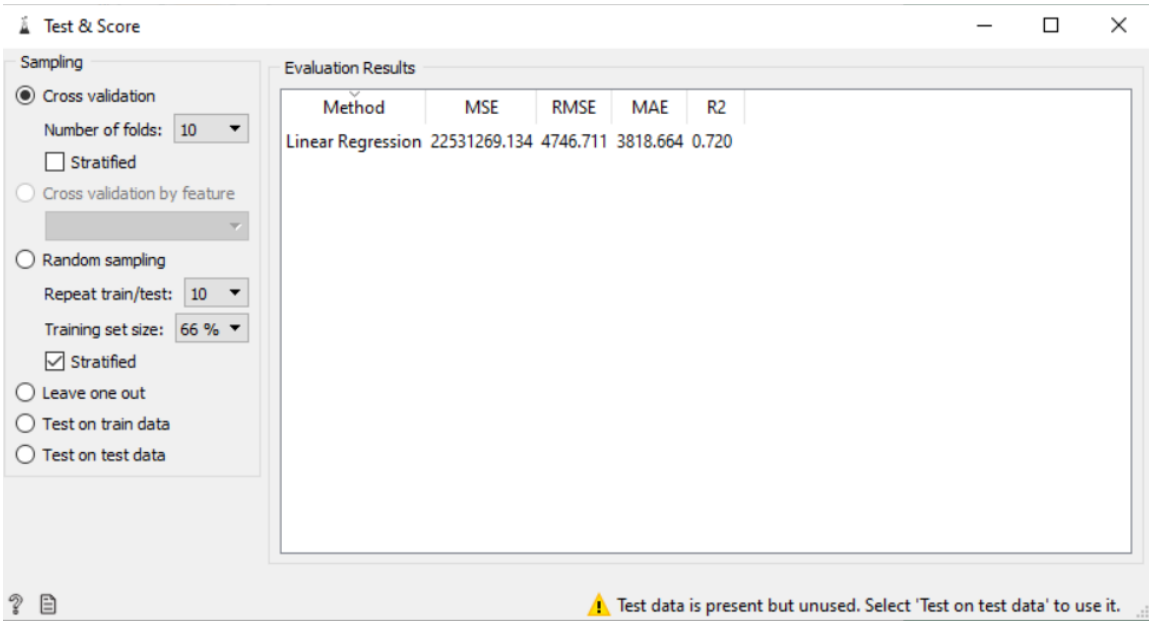
Click on the link between the Data Sampler and the Test & Score widgets and set up how the train / test split works like this:



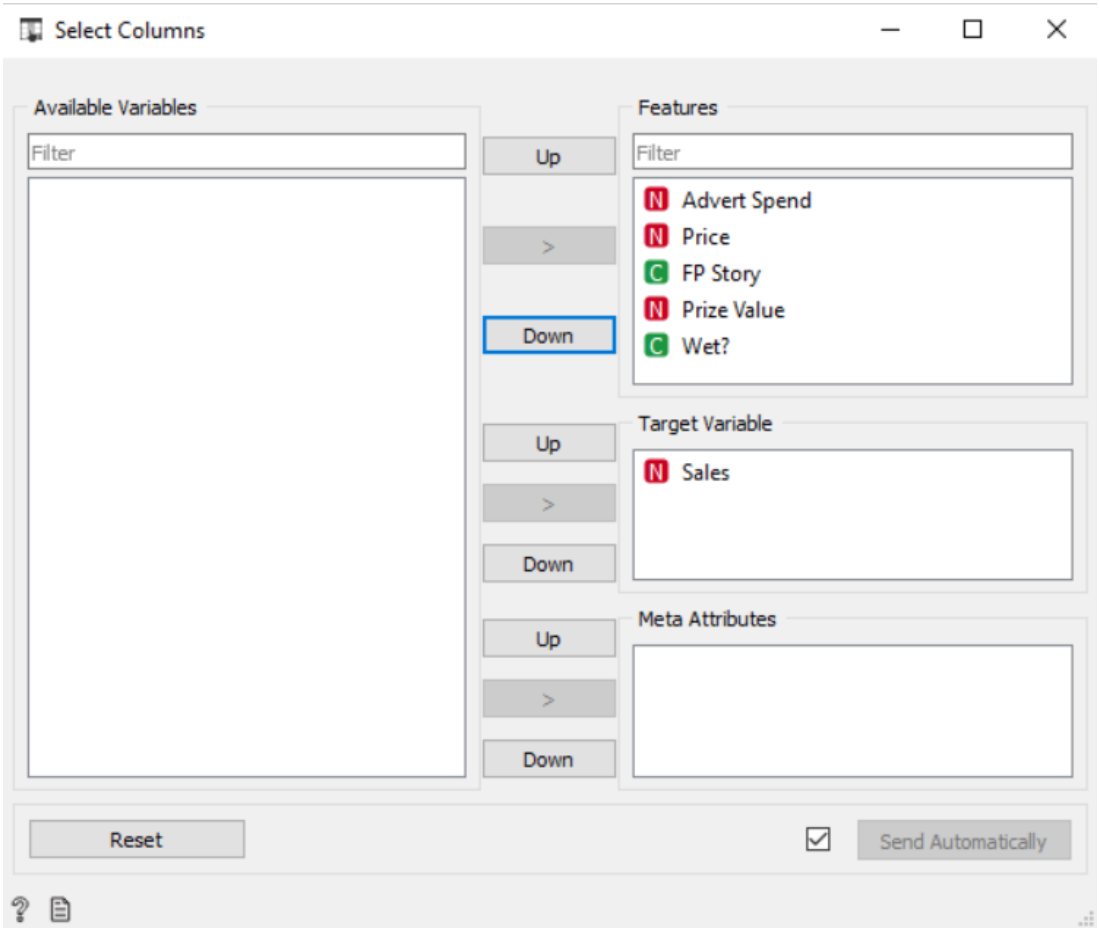
Your graph should now look like this:



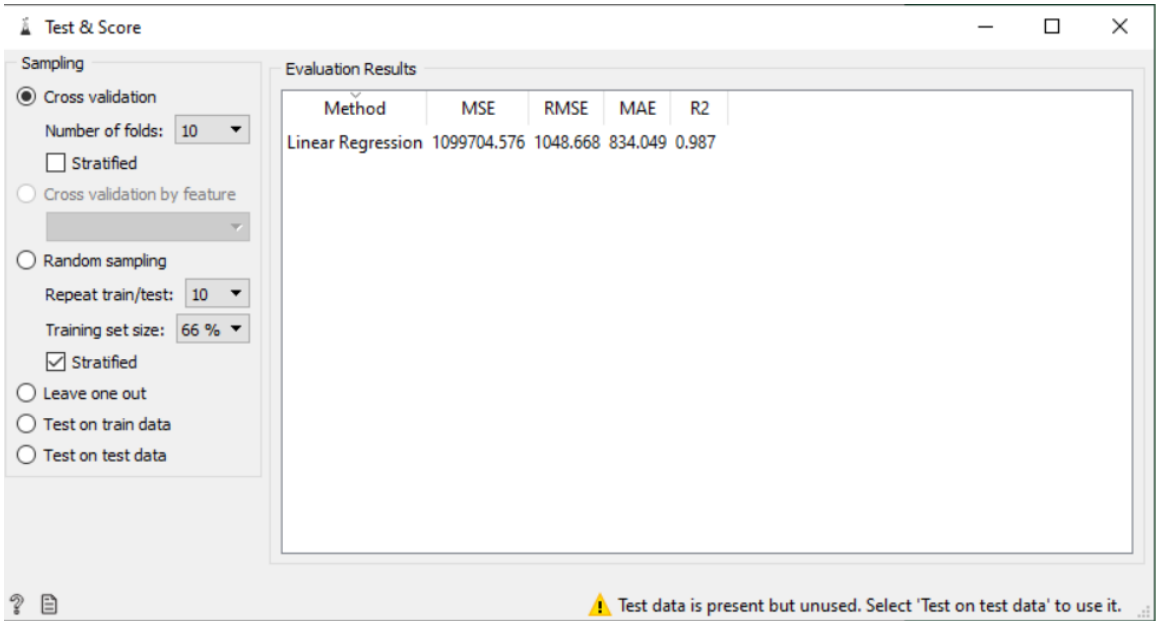
Now we look in the Test & Score results to see how well the model performs. With Cross validation selected, we get an R-squared value of 0.72



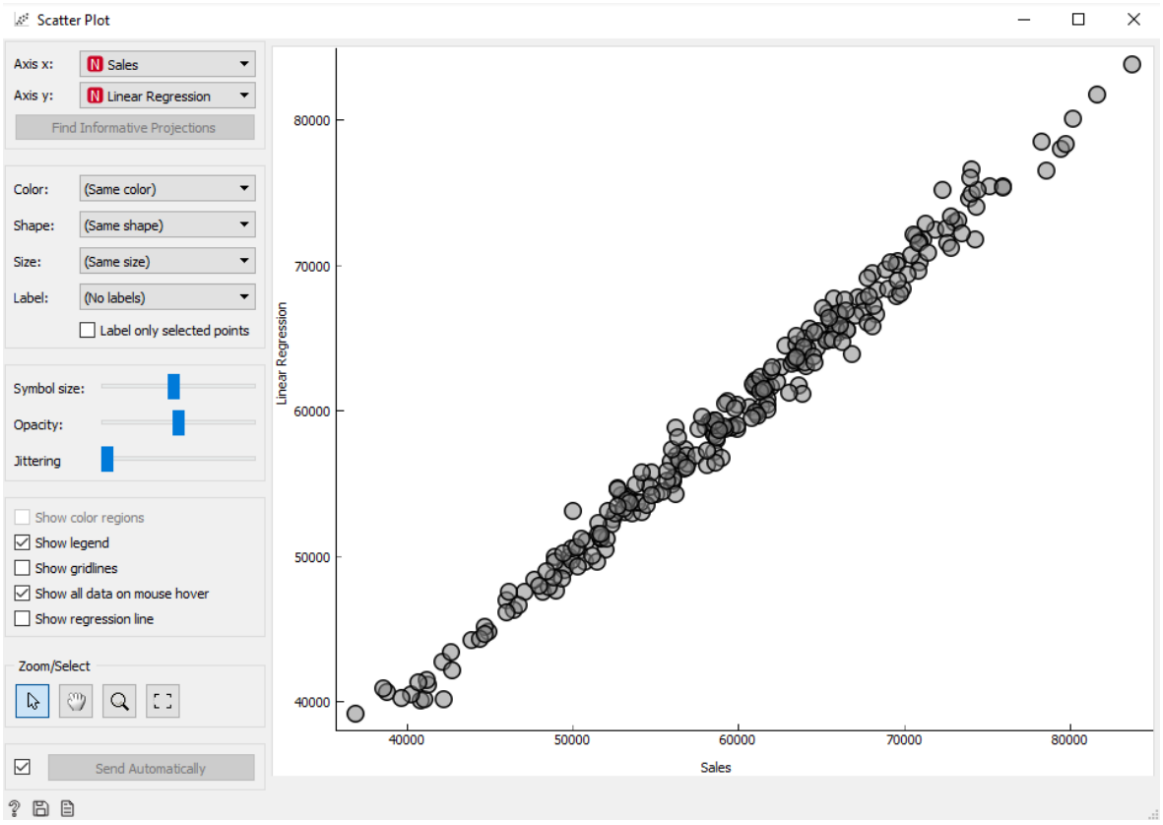
Now let's try a multiple linear regression with all the variables. We return to the Select Columns widget and select all the other inputs.



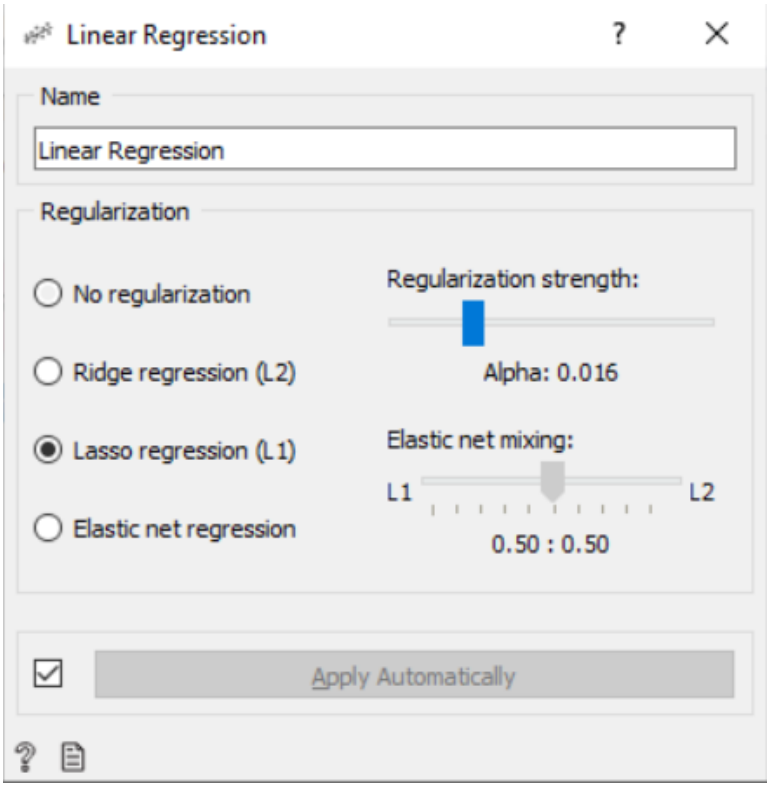
Now we look in the Test & Score results again to see how well the new model performs. With Cross validation selected, we get an R-squared value of 0.987, which is an improvement over the simple model we built before.



We can see the predicted outputs plotted against the real sales values using a scatter plot from the Visualize tab. Select Sales for the x axis and Linear Regression for the y axis:

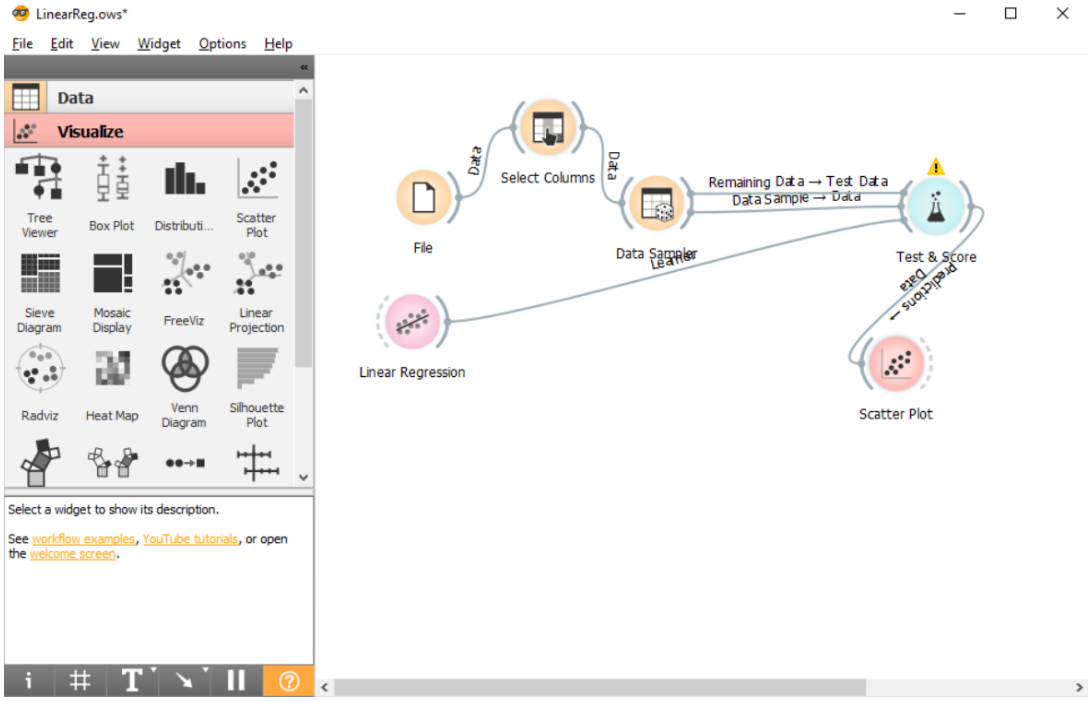


To add regularisation, we open the Linear Regression widget and choose L1 (Lasso) or L2 (Ridge) or Elastic net regression, which we have not discussed in this course, but which is a mixture of L1 and L2.



The data you have here does not need much regularisation, as the model is already very good, but you can experiment with a few values. See what happens if you set the regularisation strength to maximum.

The final project looks like this:



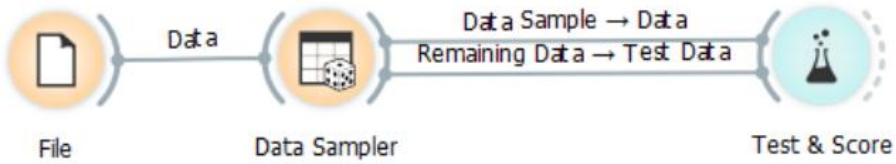
Now you know how to build and regularise regression models in Orange.

Logistic Regression

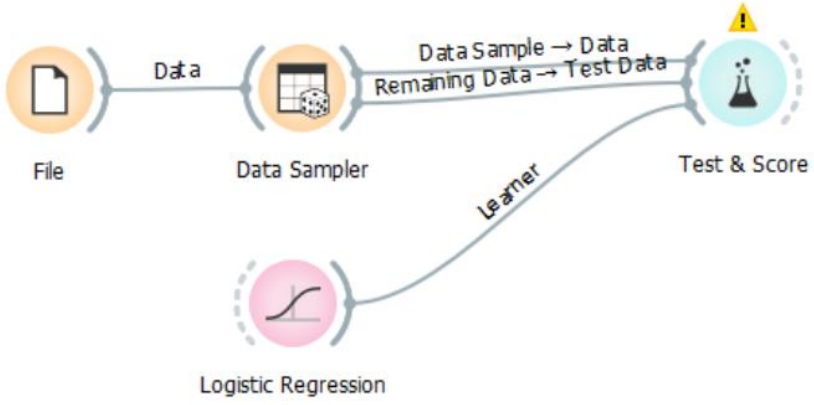
In this example, we will see how to build a logistic regression model in Orange. To run the example, you will also need this file of data: Newspaper_SoldOut.csv

The data has the numeric inputs from the newspaper sales prediction example, and a binary output that is 1 if the newspaper sold out that day and 0 otherwise.

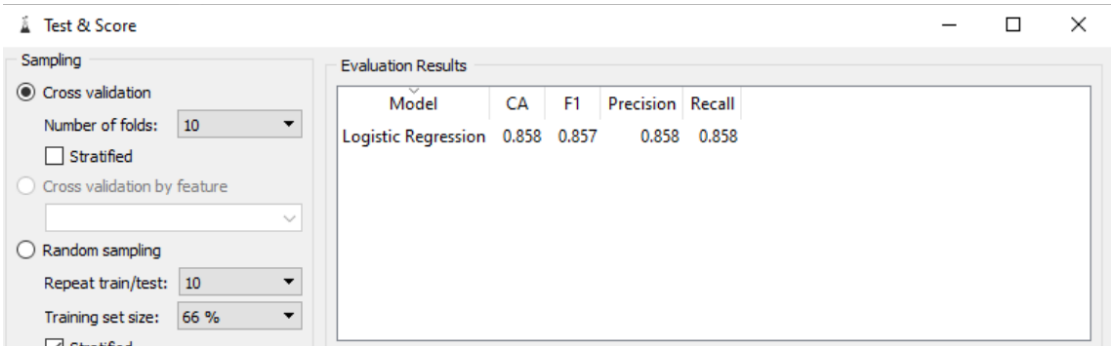
Load the data into Orange (I'm sure you know how by now, but look back at the linear regression example if you need a reminder). Select all the columns except the last one as inputs. The last column is called 'SoldOut' and is the target output. Add a Test & Score widget and your graph should look like this:



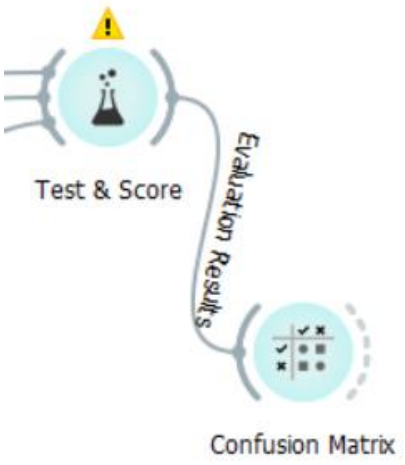
Now add a Logistic Regression widget and connect it to Test & Score.



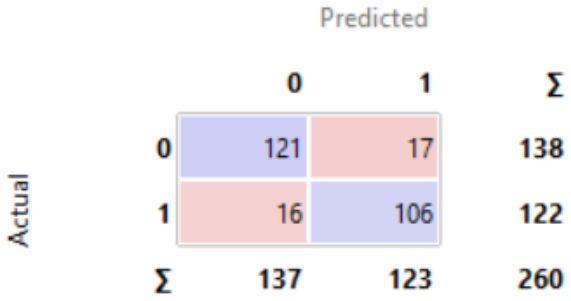
If we now look at the results in Test & Score, we see a number of measures of model quality, known as metrics. These are explained next, but look at the column titled CA, which means classification accuracy. The model is correct around 86% of the time.



We can see the confusion matrix by adding the appropriate widget to the graph:

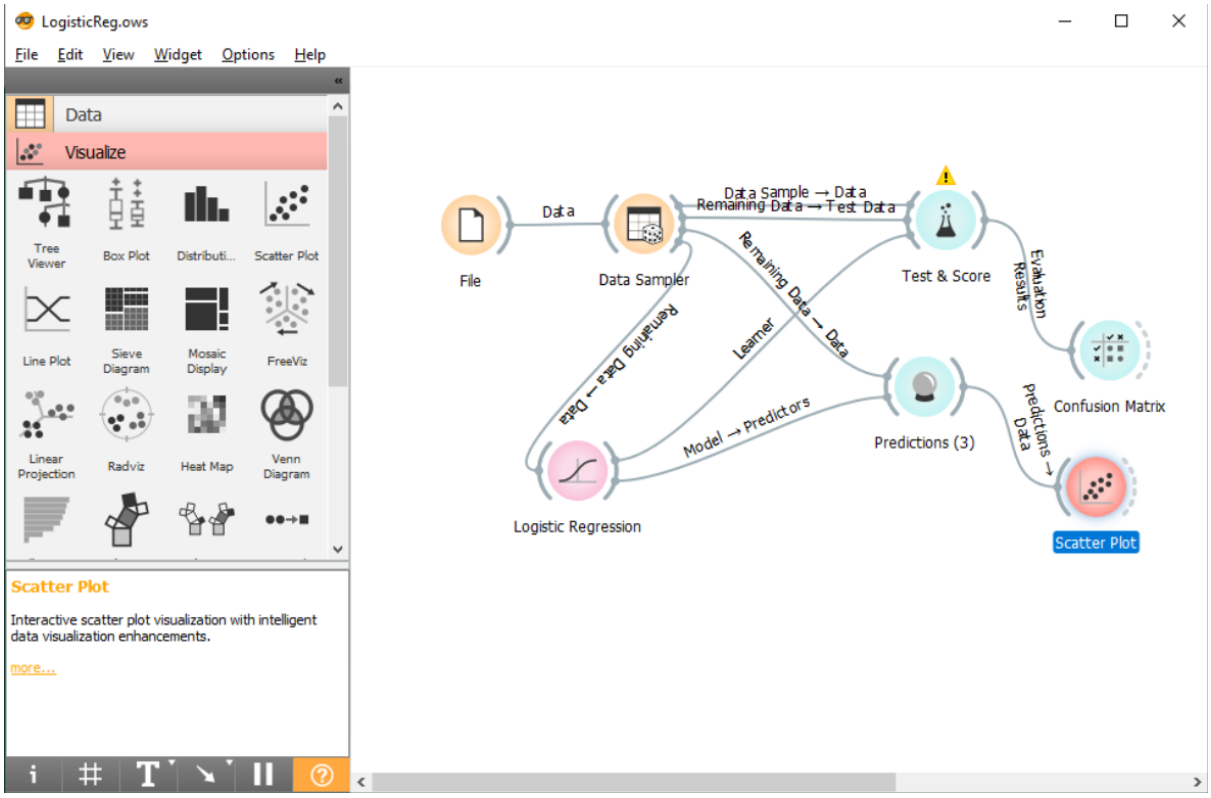


Which gives us this - the blue boxes count the ways the model is correct with its classifications and the red boxes count the two types of error.

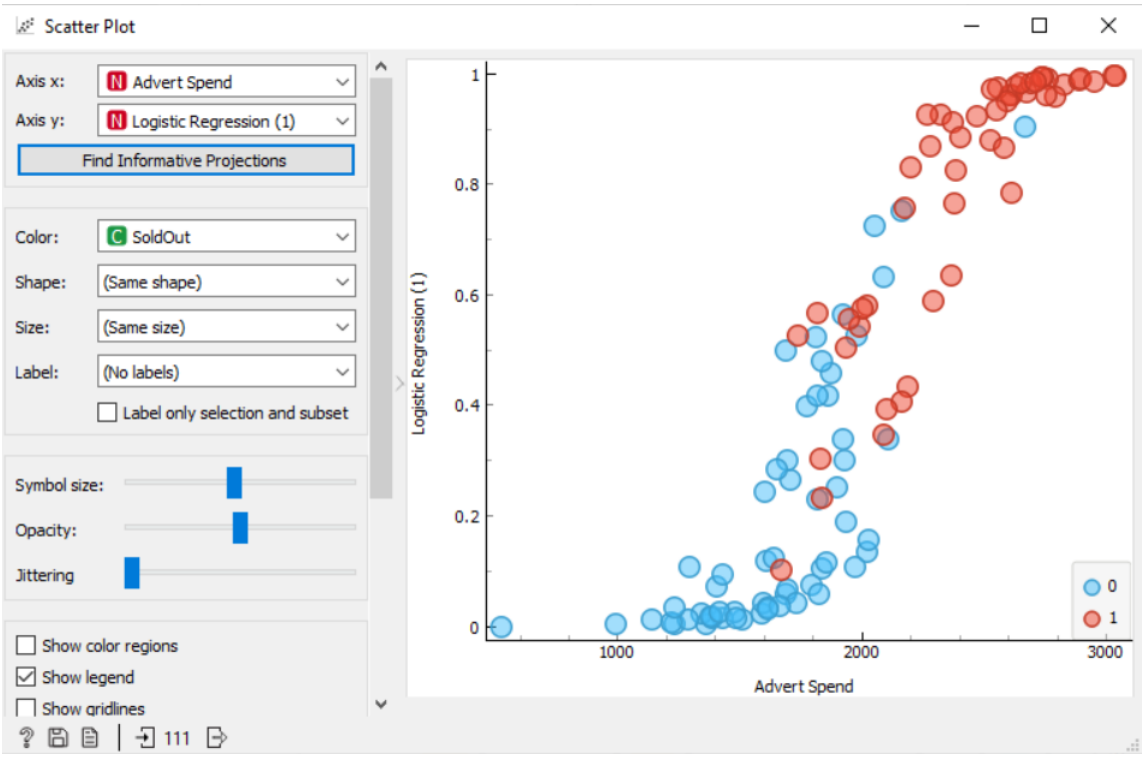


Finally, we can visualise the relationship between the different input variables and the probability associated with the positive class using a Predictions widget and a Scatter Plot.

Here is the final workflow:



Here is the plot of probability of selling out against advert spend. The colour of the points reflects the true outcomes in the target data. You can see the logistic shape, and the fact that the errors are made mostly in the middle region where there is more uncertainty. To see the relationship for other variables, select a different Axis x: choice in the first drop down box.



Now you know what logistic regression is for, and how to apply it to a dataset in Orange.