Hands on Labs: Data Analytics Part 1, 2, 3

To prepare for these Assignments

- Read the RapidMiner Manual, Sections 2.1 to 2.3.
- Watch all of the following three videos:
    - The quick tour video demonstrates some of the features of RapidMiner (RapidMiner, 2010c).
    - The RapidMiner GUI Intro (RapidMiner, 2010b). This will show you the user interface of RapidMiner. It will explain to you how data analysis processes can be designed. It will also describe the concept of 'Operators' and introduce you to a data mining process.
    - The data import and Repositories introduction (RapidMiner, 2010a). This video will show you how data is imported into RapidMiner and stored in 'Repositories'. These repositories facilitate automatic metadata propagation and can perform some automated checks on data.
- Recommended: Read Chapter 3, Steps 6 to 21, of the eBook *Data Mining for the Masses* (North, 2012). You may skip the material about handling missing data, as there are no missing data entries in the data used for this project.

Datasets used in these labs

- Review datasets available at Kaggle https://www.kaggle.com/datasets
- As an additional assignment, one can work with different datasets that are available in an extra material folder

Hands on Lab 03: Data Analysis, Part 3

In marketing, 'upselling' is the practice of convincing your customer to buy a higher priced service from your organisation. For example, when you check into a hotel or rent a car, the person helping you may ask if you would like a better room or nicer car. In many cases, front desk staff are encouraged to upsell rooms and cars to increase net profits. Using Big Data, you can better predict what kind of upselling is likely to be successful with various customers.

In this Assignment, you will attempt to predict client responses to upselling attempts. The data set you will use is related to a data set from a direct marketing campaign of a banking institution. The marketing campaign was based on phone calls. It may seem sensible to try upselling by telephoning every customer; however, this is expensive.

Your task is to compare two analytics techniques: neural networks MLPs (multilayer perceptrons) and decision trees. Decide which of these techniques is the best for predicting whether clients will respond positively to an upselling telephone call.

Dataset used in this lab

This lab will use datasets specially prepared for you and available on Google Drive in the folder rapidminer-hol-datasets.

You need to download datasets hol03dm-BankTestData.csv and hol03dm-BankTrainingData.csv. Follow instructions on how to retrieve it in RapidMiner.

Read/review the data set description.

To complete this Assignment

Compile a single document with answers to all the questions in this Assignment. As you complete the Assignment, you may wish to refer to *Data Mining for the Masses* (North, 2012).

Data set description

There are no missing entries in this data set.

Attribute Information: (Moro, Cortez and Rita, 2014)

- 1 - age (numeric)
- 2 - job: type of job (categorical: 'admin.', 'blue collar', 'entrepreneur', 'housemaid', 'management',
  'retired','self-employed','services','student','technician','unemployed','unknown')
- 3 - marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown')
- 4 - education (categorical: 'basic.4y', 'basic.6y','basic.9y', 'high.school', 'illiterate', 'professional.course', university.degree', 'unknown')

- 5 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- 6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- 7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')
- # Data from previous communications with the client
- 8 - contact: contact communication type (categorical: 'cellular', 'telephone')
- 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- 10 - day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
- 11 - duration: last contact call duration, in seconds (numeric)
- 12 - campaign: number of contacts performed previously for this client (numeric, includes last contact)
- 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- 14 - previous: number of contacts performed before this campaign and for this client (numeric)
- 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')
- 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
- 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
- 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- 19 - euribor3m: euribor 3 month exchange rate - daily indicator (numeric)
- 20 - nr.employed: number of employees - quarterly indicator (numeric)

Target Attribute (label):

- 21 – responded? – whether the client chose to buy the upselling product. 1 = ‘yes’, 0 = ‘no’

Reference:

Moro, S., Cortez, P. & Rita, P. (2014) ‘A data-driven approach to predict the success of bank telemarketing’, *Decision Support Systems*, 62, pp.22-31.

Assignment questions

Question 1

Create a new process.

Import the file ‘BankTrainingData.csv’ into a repository. Set the attribute ‘responded?’ to be the label and change its type to be ‘nominal’ (as you are performing a classification task).

When you import the file, RapidMiner automatically guesses what type the attributes are. Ignore the fact that some of the attributes on the right-hand side are assigned the type of ‘numeric’ or ‘real’ because for the task you are currently performing, they can be considered to be the same.

However, RapidMiner has incorrectly guessed at the type of one of the other attributes. Which attribute is it, why is it incorrectly typed and what should the correct type be?

Change the attribute to the correct type.

Then import the file 'BankTestData.csv' into a repository and change the type of the attribute that RapidMiner has guessed the wrong type for to the correct type (the training and test data should always have the correct coding).

If you cannot spot the variable that is the wrong type or how it should be reformulated, simply import the two files into their respective repositories and go straight to Question 2.

Question 2

Drag both of your new repositories into the process view.

MLPs can only process input attributes that are numeric. This means that you have to transform the nominal and polynomial attributes in the data set to numeric values. To do this:

Drag a 'Nominal to Numerical' operator into the process view and connect it to the operator that retrieves your training data repository.

You must now select which attributes are to be converted from nominal and polynomial format to numeric format. To do this the operator will convert polynomial attributes such as 'married' 'single' or 'divorced' into 'dummy variables'. In this example, this would mean creating three new attributes, each of which is set to either 1 or zero. For example, 'married' would be transformed into [1, 0, 0], 'single' into [0, 1, 0] and 'divorced' into [0, 0, 1] .

Click on the 'Nominal to Numerical' operator.

On the right-hand side of the screen set 'Attribute Filter Type' to 'subset'.

Click on 'Select Attributes'.

You will then be presented with a list of attributes. Select every attribute except for 'responded?' Attribute (your target label), and click on the arrow in the centre to select them for transformation.

Click 'Apply'.

Now you have transformed the nominal and polynomial attributes in the training data set to numeric values, do the same for the test set by dragging another 'Nominal to Numerical' operator into the process view and connecting it to the operator that retrieves your test data repository. Select the same attributes to be transformed as you did above.

Drag a ‘Neural Net’ operator into the process and connect it to the ‘Nominal to Numerical’ operator that takes input from the training data.

Drag an ‘Apply Model’ operator into the process. Connect its input ‘mod’ (model) port to the ‘mod’ port on the neural net. Connect its ‘unl’ (unlabelled) port to the ‘exa’ (examples) port of the ‘Nominal to Numerical’ operator that takes input from the test data. Connect its output ‘mod’ port to a ‘res’ port on the right-hand side of the screen.

Drag a ‘Performance (classification)’ operator into the process. Connect its ‘lab’ port to the ‘lab’ port on the ‘Apply Model’ operator. Connect the output ports of the Performance (classification)’ operator to the results ports on the right-hand side of the screen.

Copy and paste a screenshot of the process into your Assignment submission.

Question 3

Now you are going to change some parameters of the Neural Net.

Click on the Neural Net operator. Leave the ‘hidden layers’ parameter alone for the time being (RapidMiner will automatically set the number of units in the hidden layer).

Leave the momentum and learning rate as they are but set ‘training cycles’ to be 100.

Make sure the ‘shuffle’ (randomly presents the data to the Neural Net) and ‘normalize’ (scales all of the input attributes so that they lie in the same range) boxes are checked, but the ‘decay’ box is unchecked.

Run the model (this will take some time, depending on how powerful your computer is).

Click on the ‘Performance Vector (Performance)’ tab and copy and paste a screenshot of all the results there into your Assignment submission.

Question 4

For this run, keep all the parameters for the Neural Net the same, except for the ‘decay’ parameter (the box for this parameter should be checked).

Run the model again.

Paste a screenshot of your results into your Assignment submission.

Question 5

The results using the ‘decay’ parameter selected are better than when this parameter is not selected. Why is this?

Question 6

Save your old process and create a new one.

In this process, you are going to use **k-fold cross validation** and **ROC curves** to compare the results from two analytic techniques (MLPs and decision trees). Because you are using k-fold cross validation, you can use all of the data (training and test data combined) to train the analytic models.

Import the file 'BankAllData.csv' into a repository.

Set the attribute 'responded?' to be the label, and change its type to be 'nominal' (as you are performing a classification task).

Change the attribute that has the incorrect type to the correct type (if you did not spot this variable in Question 1, just carry on to the next step).

Drag the repository into the process view.

Drag a 'Nominal to Numerical' operator, make the appropriate connections and change the type of all the attributes except the label 'responded'.

Drag a 'Compare ROCs' operator into the process and connect its 'exa' port to the 'exa' port of the 'Nominal to Numerical' operator. Connect both of the output ports of the 'Compare ROCs' operator to results ports.

The 'Compare ROCs' operator is a nested operator. This means that you can embed one or more analytics techniques inside it (refer to the RapidMiner documentation on nested operators). If you double click on the 'Compare ROCs' operator, you will be able to see the subprocess inside it.

Drag a Neural Net operator into this subprocess, and connect its 'tra' port to the 'tra' port on the left-hand side of the subprocess and its 'mod' port to the 'mod' port on the right-hand side of the process. Set the Neural Net parameters to those you used before, i.e. training cycles = 100 and the 'decay' checkbox selected.

Drag a 'Decision Tree' operator into the subprocess. Connect its 'tra' port to the 'tra' port on the left-hand side of the subprocess and its 'mod' port to the 'mod' port on the right-hand side of the process. You do not have to change any of the parameters of the decision tree.

In the right-hand side panel, set the 'number of folds' to 5 (i.e., the *k* in k-fold cross validation is set to 5). Leave the 'split ratio' at 0.7. This means that during each fold, 70% of the data will be used for training and 30% for testing.

Set 'sampling type' to 'shuffled sampling'. This means that during each fold, random subsets of the data will be placed in the training and test sets.

Leave the ‘roc bias’ to ‘neutral’.

Copy and paste a screenshot of your subprocess into your Assignment report.

Question 7

Run the process. This may take a long time as the neural network is being trained on the whole data set and five different cross validation folds are being performed. However, RapidMiner allows you to carry on working with another package (such as a word processor or Web browser) while it is operating, and the RapidMiner icon will flash when the training has finished.

When training has finished, post a screenshot of your ROC comparison curve into your Assignment report.

Question 8

Based on the ROC curves, which analytics technique would you use to perform the ‘upselling’ task and why did you choose this technique?