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 02: Data Analysis, Part 2

In this Assignment, you will examine a business data set and create decision tree models. You will use these models to both predict the income group of each of your customers and to build visualisations of the important factors that influence this business process.

In this Assignment, you will become familiar with practical application of analytic concepts related to classification algorithms. Additionally, you will gain more familiarity with the RapidMiner software. If you have difficulty using the RapidMiner software, you may discuss the problems you are having with the package in the Discussion folder, as long as you do not directly discuss the Assignment questions (this is an Individual Assignment).

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 the datasets hol02dm-TreeTestData.csv and hol02dm-TreeTrainingData.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 the eBook *Data Mining for the Masses* (North, 2012).

Data Set Description

There are two data sets you will use for this Assignment: a training data set and a testing data set. In both the training and testing data set, the attributes are as follows:

- Age: Integer
- Workplace: Private, Self-employed-not-incorporated, Self-employed-incorporated, Federal-government, Local-government, State-government, Without-pay, Never-worked.
- Population_Density_Location (the population density of the area your customer lives in): Integer
- Education (highest level of qualification): Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- Years_of_Education (number of years of secondary education): Integer

- Marital_Status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-speciality, Handlers-cleaners, Machine-op-inspectors, Admin-clerical, Farming-fishing, Transport-moving, Private-house-services, Protective-services, Armed-Forces.
- Relationships (family relationships): Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- Ethnic_Group: White, Asian-Pac-Islander, American-Indian or Eskimo, Other, Black.
- Gender: Female, Male.
- Financial_Attribute_1 (an anonymised financial attribute)
- Financial_Attribute_2 (an anonymised financial attribute)
- Hours Worked Per Week: Integer
- Birth_Country (where the customer was born): United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc.), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad and Tobago, Peru, Hong Kong, Holland-Netherlands.

The target label you will try to predict is: Income_Group (Normal Income or High Income).

There are no missing entries in the data sets.

There are 30,162 records in the training data set TreeTrainingData.csv.

There are 1,560 records in the test data set TreeTestData.csv.

Assignment questions

Question 1

In RapidMiner, create a repository and load the training data set TreeTrainingData.csv into this repository. Make sure that you set the last attribute Income_Group to be the label that you are going to predict.

Drag this data set into the main RapidMiner Design Perspective so that there is a ‘Retrieve TreeTrainingDataOperator’ in the window.

Drag a ‘Decision Tree’ operator into the process window and connect it to the ‘Retrieve TreeTrainingDataOperator’

Connect all of the outputs of the ‘Decision Tree’ operator to the ‘Results’ ports on the right hand side of the Design Perspective. Once you have connected one output, another results port will appear.

Paste a screenshot of the Design Perspective into your Project submission to show that you have achieved these steps.

Question 2

Click on the 'Decision Tree' operator, and on the right-hand side of the screen set the parameter 'Minimal Leaf Size' to 20 (this controls the minimum number of customers needed to generate a new leaf of the tree).

Run the process.

Click on the 'Tree (Decision Tree)' tab in the Results Perspective and inspect the tree.

Click the 'Save Image' button on the left-hand side of the screen showing the decision tree and save the tree in a convenient format (such as *.png).

Import the image of the decision tree into your Project submission to show that you have achieved these steps. Once you have imported the image, you may have to resize it to make it easily readable.

Question 3

Place an 'Apply Model' operator after the 'Decision Tree' operator and connect the ports. Add a 'Performance (Classification)' operator after the 'Apply Model' operator and connect its input ports to the previous operator and its output ports to two results ports.

Run the model.

When the model has finished training, click on the 'Performance Vector (Performance)' tab. Write the value of the accuracy into your Project submission.

Question 4

So far, you have generated an accuracy estimate on the training data. However, a more realistic evaluation of accuracy would involve using the model to predict the income group of customers whose data has not been used to train the model.

Create a new repository and load the test data set TreeTestData.csv into this repository. Make sure that you set the last attribute Income_Group to be the label that you are going to predict.

Drag this data set into the main RapidMiner Design Perspective so that there is a 'Retrieve TreeTestData' operator in the window.

Connect the output port on the 'Retrieve TreeTestData' to the 'unl' port of the 'Apply Model' operator (this means that once the decision tree model has been built on the training data, the tree model will be applied to the test data).

Paste a screenshot of the Design Perspective into your Project submission to show that you have achieved this step.

Train the model and write the value of the accuracy on the test set into your report.

Question 5

From Question 4, you now have the accuracy for the test set when the 'Minimal Leaf Size' is 20.

Now run a series of models with the parameter 'Minimal Leaf Size' set to different values. Each time you do this, you will have to click on the 'Decision Tree' operator and change the value of 'Minimal Leaf Size' on the right-hand side of the screen and then run the model.

Try these values for 'Minimal Leaf Size', and each time record the accuracy on the test set:

16, 15, 12, 10, 7, 5, 4, 1

Then add the accuracy on the test set for 'Minimal Leaf Size' 20 that you calculated in Question 4 to the front of your list of results for these different models.

Question 6

What phenomena do you observe when inspecting your results from Question 5?

Question 7

Run the model again with 'Minimal Leaf Size' set to 10 and inspect the values in the 'Performance Vector (Performance)' tab. Inspect the values in the matrix. There is a difference in the accuracy of prediction for the two different classes 'Normal_Income' and 'High Income'. What do you think is causing this?

Question 8

Think of a way to improve the accuracy for the less frequent class. You can do this either by manipulating your model in RapidMiner or by manipulating one of the data sets provided to you and then running the model you have constructed in the previous questions using this new data set.

If you have manipulated the model in RapidMiner, paste a copy of your design perspective into your Project submission followed by a copy of the results matrix generated by your model.

If you have manipulated the data set, just describe how you have manipulated the data set followed by a copy of the result matrix generated by your model.

Question 9

Are there other approaches to solve the problem you have attempted to remedy in Question 8?