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 01: Data Analysis, Part 1

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 dataset hol01dm-creditDataset.csv and retrieve it in RapidMiner as instructed below.

To complete this assignment

Carefully read the Project Description below and submit a single document in which you document your work on all assignment steps.

Project Description:

The focus of this project is for you to become familiar with practical application of the analytics concepts introduced in Lecture Notes 4. To complete this project, you will use RapidMiner analytics software.

This project uses Linear Regression, Logistic Regression and Supervised Learning to predict whether to approve or reject individuals' credit card applications. Because this data set is in the public domain, many attribute names (columns in the data set) have been changed to meaningless characters (such as 'j' or 'aa') to protect confidentiality of the data. While this data set was not generated by Big Data resources, using it for modelling provides you a strong introduction to analytics with RapidMiner.

It is not standard practice to train analytic models with the complete data set (we will discuss this in the lecture on classification). However, for simplicity in this first analytics project, you will train models on the whole data set.

Description of the Data Set

There are 690 records in the data set. There are no missing values in the data.

There are 13 attributes (columns in the data) and a final column representing the target class label (whether to allow the individual to have a credit card or to decline his or her application).

There are 6 numerical and 8 categorical attributes. You will use these as inputs for your regression models. The original character values have been changed to make them easy to use by an analytic algorithm. For example, Column 4 originally had one of three character labels: 'p', 'g', or 'gg'. These have been changed to the values 1, 2 or 3.

The last column (Column 14) is the target you are trying to predict. The value of '1' indicates that your organization should allow this individual to have a credit card. The value of '0' indicates that this individual's application should be rejected.

Attribute	Values	Type	Original Labels
A1	0, 1	CATEGORICAL	a or b
A2	continuous	continuous	Col_2
A3	continuous	continuous	Col_3
A4	1, 2, 3	CATEGORICAL	p, g or gg
A5	1, 2,3,4,5, 6,7,8,9,10,11,12,13, 14	CATEGORICAL	ff, d, i, k, j, aa, m, c, w, e, q, r, cc or x
A6	1, 2, 3, 4, 5, 6, 7, 8, 9	CATEGORICAL	ff, dd, j, bb, v, n, o, h or z
A7	continuous	continuous	Col_7
A8	0, 1	CATEGORICAL	t or f
A9	continuous	continuous	Col_9
A10	0, 1	CATEGORICAL	qt or rs
A11	1, 2, 3	CATEGORICAL	s, g or p
A12	continuous	continuous	Col_12
A13	continuous	continuous	Col_13
A14	0, 1	CATEGORICAL	Target Classification, 1 or 0

The data is in Comma Separated Variable (*.csv) format. This data format is easily imported into spreadsheets. You may want to inspect it before you use RapidMiner (though this is not essential).

Project steps and questions

Step 1: From the Design Perspective, create a new repository called 'CreditRepository' in Rapidminer. Paste a screenshot of the repositories into your Project submission to show that you have achieved this step.

Step 2: Import the creditData.csv data file into the repository and name it 'CreditData'. Make sure that you set the last column ('Target') to be a nominal 'label', as this is what you will be trying to predict. Paste a screenshot of the Data Import Wizard into your Project submission to show that you have achieved this step.

Step 3: In the Design Perspective, drag the 'CreditData' into the Process View to create a 'Retrieve Credit Data' operator. Drag a 'Linear Regression' operator into the process and connect it to the 'Retrieve Credit Data' operator. Paste a copy of the process view into your Project submission to show that you have achieved this.

Step 4: Add 'Apply Model' and 'Performance (classification)' operators and join them all together. Join the last operator to the results ('res') output. Paste a copy of the process view into your Project submission to show that you have achieved this.

Step 5: Run the model. If you have built it successfully, a Results Perspective will appear. Paste a copy of the Results Perspective into your Project submission to show that you have achieved this.

There will be a matrix in the centre of the Results Perspective. What is this type of matrix called? With reference to this type of matrix, what do the terms 'True Positives', 'True Negatives', 'False Positives', 'False Negatives', 'Class Recall' and 'Class Precision' mean? How are these values calculated?

Step 6: Now build a logistic regression model for the same data and interpret the results in the same way as for the linear regression model. Is there a difference in performance between the two types of model? If so, what does this difference in performance imply? Paste a copy of both the process view and the results view into your Project submission.

Step 7: Imagine a hypothetical situation in which the data set has a lot of missing entries in the columns for attributes A1, A2 and A5. If there were too many missing values to simply discard the data rows containing them (i.e. discarding these rows would make the data set too small to practically model), what would you do with these missing data entries?

Step 8: Is there any pre-processing of the data that could improve either the accuracy of the model or the speed at which the model arrives at a solution? (Do not implement any of this pre-processing in RapidMiner or build a model using this pre-processing.)