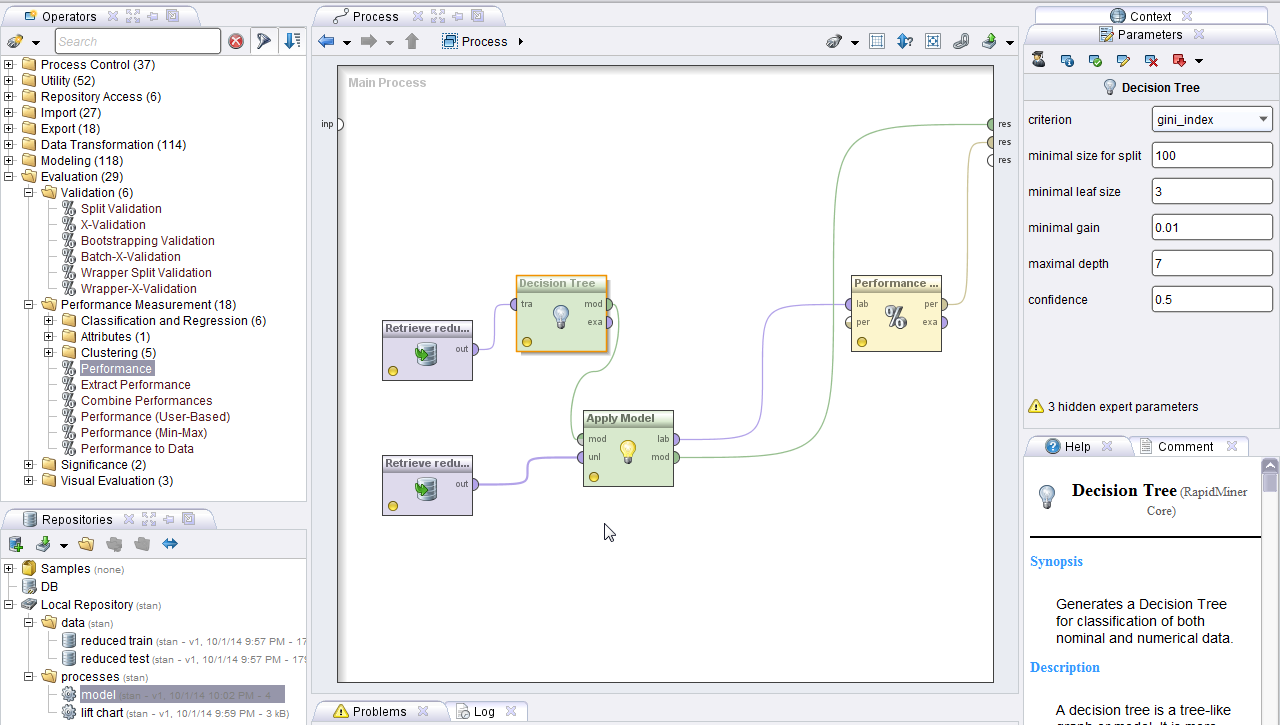
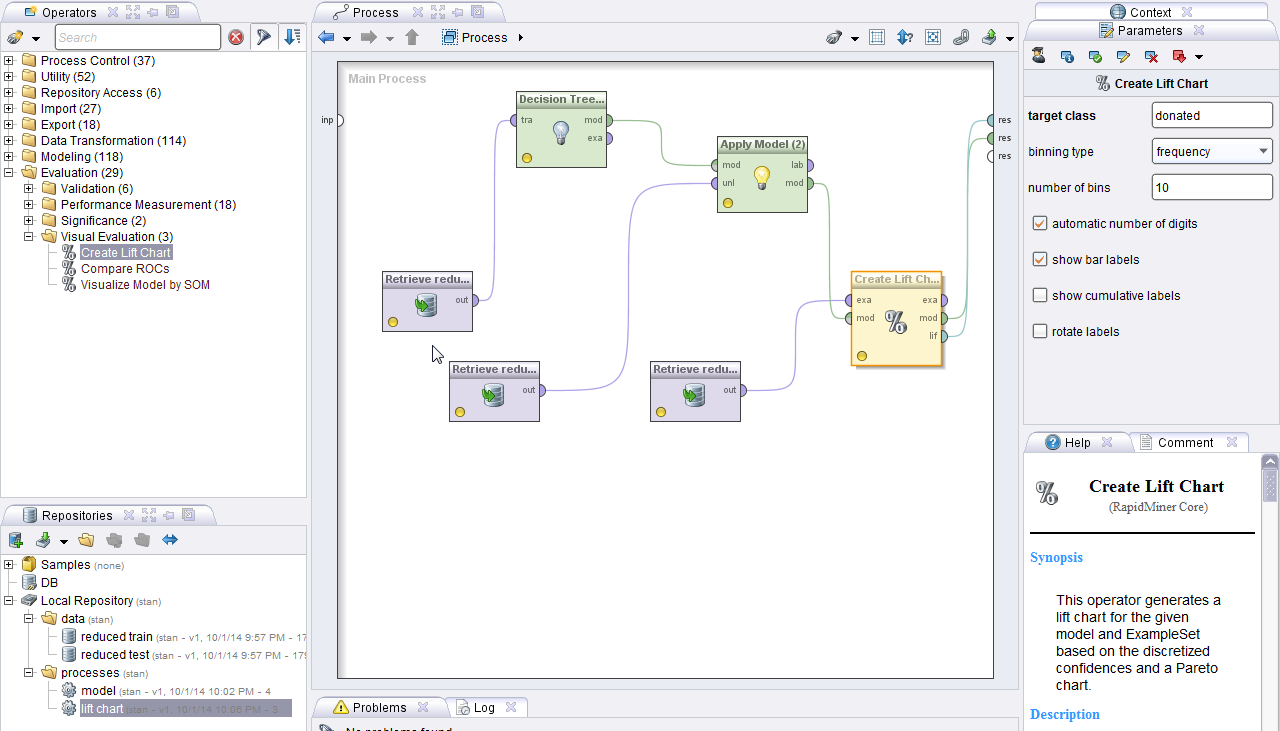
1. In EXCEL convert attribute DONIND (binary) into an attribute label (binary string). I.e. 0=”not donated” and 1=”donated”. This is done because some algorithms can not use integer values as labels.
2. We cannot use DONAMT as an attribute in our modeling as there is a 1:1 relation between this attribute and investigated label. I.e. if we used it in the decision tree induction algorithm would simply check if DONAMT>0 and if so ascribe the label: “donated”.
3. Because we use SPSS for data reduction purposes (to obtain correlation matrix as well as Principal Factor Analysis) we have to import our data into SPSS. We will also use SPSS to produce a tree induction model. SPSS handles train and test samples differently than RapidMiner so we have to combine the data into one sheet in the following way:
   1. Create a copy of data.xls and save it as data spss.xlx.
   2. Add a new attribute to „train” sheet called „train” and input 1 (int) for all records.
   3. Copy whole “train” data sheet to test sheet (ie. put the records under records of test)
   4. Test records need to be given a value for the “train” attribute so add 0 for all test records.
   5. Save
4. Now we can import data into SPSS:
   1. File->Open Database->New Query
   2. Select data spss file. Finish
5. Data reduction:
6. Correlation Matrix:
   1. Correlation matrix = Analyze->Correlate-> add variables (Excluding REGNUM)->OK
   2. See “Content of outputs” document Table 1. We observe high correlation between 1) FRQRES and TIMELR, 2) ANNDON and AVGDON, 3) LSTDON and AVGDON and 4) ANNDON and LSTDON.
   3. 1) It is natural that if somebody is a frequent donator (high FREQRES) we can expect to hear from him recently (low TIMELR). However, frequency measures overall (long time or historic) responsiveness to our mailings, while time since last response reflects rather “swiftness” of the most recent mailing. As such, for now we do not remove the attributes.
   4. 2) The correlation between annual average donation and average donation per mailing is more problematic. The difference between the attributes is in the unit of measurement: year and a mailing. We don’t know yet which of those two attributes is best to remove, but we know one of them is ok to stay. The question is what distinguishes the two averages? Annual average may possibly be associated with mental yearly budget a person attributes to charity. Per mailing average may reflect the responsiveness of a donator to a mailing stimuli. However, increasing the amount of mailings to a person (e.g. those with high AVGDON) is likely to decrease his donations per mailing since his yearly budget was already reduced. On the face of it I would argue that due to prevalence of mental budgeting theories in behavioral economics ANNDON works as a better indicator of a donator capability to donate. Thus, AVGDON is removed.
   5. 3) As a result of d. 3) is solved.
   6. 4) Since one’s last donation would be a part of a sum of all his donations in the past years LSTDON lingers into ANNDON calculation. LSTDON removed.
   7. The remaining attributes: TIMELR, TIMECL, FRQRES, MEDTOR, ANNDON.
7. Principal Factor Analysis:
   1. Before we complete PFA we have to standardize variables. Analyze->Descriptive Statistics. Add the 5 variables and check the box “Save standardized values as variables” above the OK button.
   2. Analyze->Dimension Reduction->Factor. Add the five standardized variables [ZScore(Variable)].
   3. Select “Extraction” option. Check in Display “Scree plot” and select method Principal Axis Factoring.
   4. See Scree plot and Factor Matrix in SPSS “output” file (you have to open data first). PFA suggests two latent factors influence the dependent variable. This might suggest the tree can have two main stages. A person with high first factor would have low TIMELR (high negative factor loading) and high FRQRES (high positive factor loading). Factor 1 can be interpreted as indentifying a highly engaged person who donates frequently and has been recently active. Factor 2 is more difficult to interpret.
   5. High communalities values (Extraction column) of TIMELR and FRQRES indicate low uniqueness of those attributes – they are largely dependent on others. Interestingly MEDTOR is a very unique attribute.
   6. We also see that the two factors account for only aprox 62% of the variation in the data. This is quite low. 95% is achieved by 4 factors what may suggest 4 layers of nodes in the decision tree.
8. SPSS CHAID model:
   1. Analyze->Classify->Tree. A popup will appear. Select: Define variable properties. Select “label” as a variable to scan and next set its Role to Target.
   2. Dependent variable is label, the independent are the five variables.
   3. In Validation select Split-sample validation->Use variable-> Split sample by->train. (as obtained in 1).
   4. Criteria are left unchanged. Default Decision tree size is three which is conveniently between 2 and 4 – the same as the amount of factors we predicted in PFA.
   5. In Save select Predicted values and Predicted Probabilities.
   6. Run. Zombies approach.
   7. See “outputs” doc for detailed calculation of ratios from the result matrix.
9. We want to present SPSS model as an intuitive one i.e. one that can be easily read and interpreted so that people can have a rough idea of groups of donators an organization can have. E.g. Node 22 may correspond to frequent donators who take more time to respond to a mailing but eventually do donate. Node 27 can be read along the lines: infrequent donator (unengaged) with usually high time of response (delaying donation) and short time as a client (so not a proven long time donator) is much less likely to donate than a similar donator with a longer tenure (Node 28).
10. There is no automatic way to produce a lift chart in SPSS (at least I can’t find it) but it can be computed with enough time put into it. Here would be a procedure:
    1. Assume you sent letters to all people in your test sample i.e. 4080 people. Out of those 1409 will be donators. So on average if you send a letter you expect to get 0,345 donator. At 10 letters 3 donators at 30 letters 10 donators and so on. In a graph this yields a straight line.
    2. Lets assume you decide to send letters to 100 people. The best option is to select those records whose probability of being classified as a donator is the highest. Say in your model there are only 60 people with 75% probability and only 40 people with 70% probability. You then can expect (on average) 60\*0,75+40\*0,70=73 people to be donators. If you used the “naïve” predictor out of sample of 100 people you would expect 35 donators. This mechanism creates the lift.
    3. First x-axis coordinate of a lift chart is the (cumulative) number of people in the highest probability group (e.g x=60). Y-axis coordinate is expected number of people to come out as donators out of the sample specified by x-coordiante (y=45 for 60 people with 75%). Next x-coordinate is sum of the previous coordinate (60) and number of people with second highest probability of being donators (60+40=100).Next y-coordinate is sum of the previous y-coordinate (45) and product of 2nd x-coordinate and respective probability (45+40\*0,70=73).
    4. And then you continue to move along each axis with each new record containing a previous record.
11. Rapid Miner:
    1. Import data to RM. Set REGNUM as id, “label” as label, remove AVGDON, LSTDON.
    2. Set the model like this:



* 1. Set up Lift chart like this:



The lift chart looks like this:

