Data Warehousing and Data Mining:

Assignment 1: Analysis of Czech Bank Data

1.0 *Learning Outcomes*

This assignment will be assessed upon your ability to:

LO1. Describe and critically evaluate the role and relevance of [analytical investigation] to the solution of business information problems.

LO2. Explain the concepts that underpin the subject area of [data mining] making reference to main established concepts and some developing areas.

LO3. Apply concepts and justify decisions when modelling, designing and constructing practical examples or paper descriptions of applications in this area.

1.1 *Deadline:*

Your completed assignment should be submitted electronically, in word format, through the assessment areas on Blackboard, at or before **23:59** on **Tuesday** the **17th of January 2017**. Failure to submit the assignment on time, without a valid extension, will result in zero marks being awarded.

1.2 *Assessment Criteria*

Marks are indicated next to each question.

This is the first assignment and consists of an individual piece of work that contributes 50% to the final marks of this module. It takes the form of a well-structured report of 2,500 words in length. Marks will be deducted if it is outside the 10% allowance.

1.3 *Notes*

- You are required to submit this assignment twice once through turn-it-in, to generate a turn-it-in report and again, through the SHU assignment submission area so that the work can be easily marked. These will be made available to you on blackboard in due course.
- All references must follow the APA‡ format.

‡ Help on this can be found in the library resource: http://libguides.shu.ac.uk/referencing You must be logged onto the library for this link to work).

1.4 *Problem outline*

For this assignment you are required to analyse a data set taken from the data mining competition prior to the third international conference of Principles and Practices of knowledge discovery in data bases (PKDD). This conference was held in Prague in 1999[§]. One of the challenges given for the competition was a set of datasets concerning financial transactions and details for customers at a Czech bank. Several of these files have been combined to give 4 500 observations of various financial and personal details relating to different accounts. Details of the variables included are given below.

Details of the relationships are shown below:

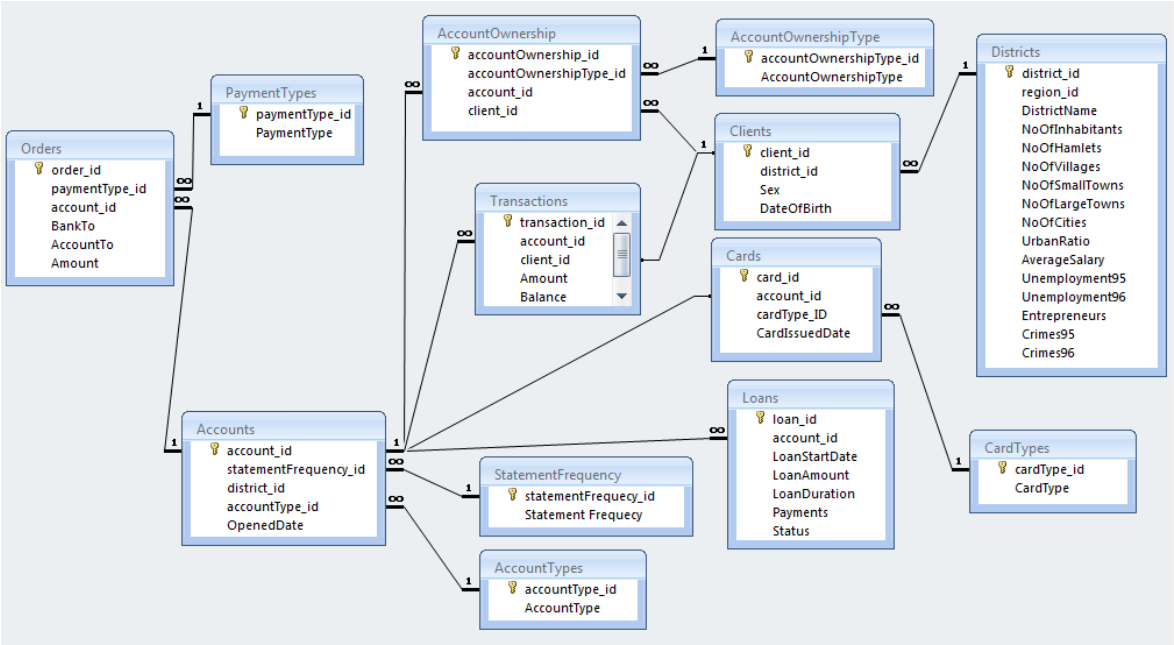

Figure 1: Database Relationships for Czech Bank Data

You are required to analyse one table resulting from a query from this database as detailed below. Full details of the fields in this table are given below and in the appendix.

[§] http://lisp.vse.cz/pkdd99/

1.5 *Data Provided*

The final query is saved as a SAS dataset for use in Enterprise Miner. It is called *czech16.sas7bdat.* It is available on the SHU server in the path: ***D:\SASMetaShare\!SharedData\DWDM.*** You will need to create a library to access the data.

1.6 *Details of the Query and resulting data*

We wish to investigate if there are any groups of accounts with similar properties. Also we wish to determine which accounts have a loan attached to that account. For this purpose a subset of variables (fields) are selected from the final query on the database for each account. These variables can be seen to represent, for each account different types of credits and withdrawals that take place:

Credits (this is where customers pay into their account):

- Cash
- Salary
- Pension
- Interest

Withdrawal (taking money out i.e. customers spend their money):

- Cash
- Insurance payment
- Credit Card
- Overdraft Penalty
- Statement Payment
- Household Payment
- Other bank withdrawal

All amounts are measured in Czech Koruna (Kč). For each of these the mean size of the payment is calculated as the total value of all payments over the total number of payments. So for example the variable for mean insurance payments (***minsure***) is calculated as the total value of insurance payments divided by the number of insurance payments. For an account which has made ten monthly insurance payments of £12, and then a further 10 of £15 (perhaps their premium went up), the value of ***minsure*** would be:

$$\textbf{\textit{minsure}} = \frac{\overbrace{12+12+12\ldots+12}^{10}+\overbrace{15+15+\ldots+15}^{10}}{20} = \frac{120+150}{20} = \frac{270}{20} = 13.5\text{Kč}$$

In addition to the mean for each of the variables above the average gap between cash credits (***mcashcrcgap***) and cash withdrawals(***mcashwdcgap***) is also included. (These are **NOT** included for the other credit/debit methods). So if an account shows a cash payment in (cash credit) each week on the same day, then the value for ***mcashcrcgap*** would be 7 days, an account with no cash credits or only one would have ***mcashcrcgap***=0. Similarly, suppose an account shows a gap of 3 days between the first cash withdrawal and the second withdrawal. Then if subsequent cash withdrawals occur again after 5 further days, 4 further days and finally after a further 10 days; then the value of ***mcashwdcgap*** $= \frac{3+5+4+10}{4} = 5.5$ days.

Finally additional information is held about each account:

Account id, **Age** of primary account holder, if they have a credit **card** or not (with this bank), number of **days** account open, if they have a **loan** or not, if there is a **second** user of the account and the gender of the main account holder (***sex***). This gives the set of variables as shown in the appendix

1.7 *Questions*

The bank wishes to see if different customers have similar profiles and have therefore asked that the ***czech16*** data be clustered. They are looking for about five clusters. You will need to build an Enterprise Miner diagram and use appropriate nodes to answer these questions. A typical diagram is shown in Figure 3 on page 7.

1) Since cluster analysis requires the use of fields that are symmetrical as possible you should first investigate each of the **interval** fields in the ***czech16*** data.

   a) Produce suitable summary measures (means, standard deviation, skewness etc.) of the data and fully interpret your results. (No marks will be awarded for discussing the kurtosis). For example, what is the largest average type of credit? Which is the most consistent? What about withdrawals? Are there any other interesting features?

   (7 marks)

   b) Produce suitable plots of each of these fields. Are there any unusual features to any of your plots? By examining the actual data records can you discern why these features might occur? Are they what you might expect?

   (7 marks)

   (You may initially find it best to achieve this with the “explore” feature available within Enterprise Miner, to obtain detailed statistics use a StatExplore node with loan set as input. Remember to reset loan back to target for building the tree in question 9) below).

2) Also investigate the remaining binary variables by producing suitable plots. Fully discuss your results.

(2 marks)

3) Use the transform node in Enterprise Miner and the "Maximum Normal" option for interval variables to find suitable transformations of the interval variables. The aim is to make them as symmetrical as possible. You should ensure that in your settings, you still retain a copy of the original variables (set both **Hide** and **Reject** to "no").

   a) Explain what actual transformations the software has picked.

   (3 marks)

   b) Produce further plots of the transformed variables and use these to present evidence of whether the transformations have been successful. Comment clearly on your results. (Please note that it may not be possible to make all variables totally symmetrical). Consequently, state **for each** interval variable whether subsequent analysis should use the original (untransformed) variable **or** the new transformed variable. Hence explain which set of interval variables you would use for clustering.

   (6 marks)

4)

   a) Clustering of these data has been carried out using code resulting in the dendrogram on the following page (Figure 2). Fully discuss this dendrogram and explain how many clusters you might pick. You should give all reasonable possibilities and then explain what your final choice is likely to be.

   (4 marks)

   b) Use a clustering node to cluster the interval variables only. You should use the variable set that you specified in 3)b) above. Ensure that the resulting set of interval variables are suitably standardised. Set the “Initial Cluster Seed” options as follows:
   - Seed Initialization Method➔ MacQueen
   - Minimum Radius➔1
   - Drift during Training➔1

   Hence obtain a CCC plot for cluster solutions of the interval variables.

   (1 mark)

   c) Fully discuss the CCC plot and clearly explain how many clusters you would pick.

   (5 marks)

5) Change the settings of the clustering node so as to fit **five** clusters. Discuss the resulting "segment size" plot and "Mean statistics" in the output.

(6 marks)

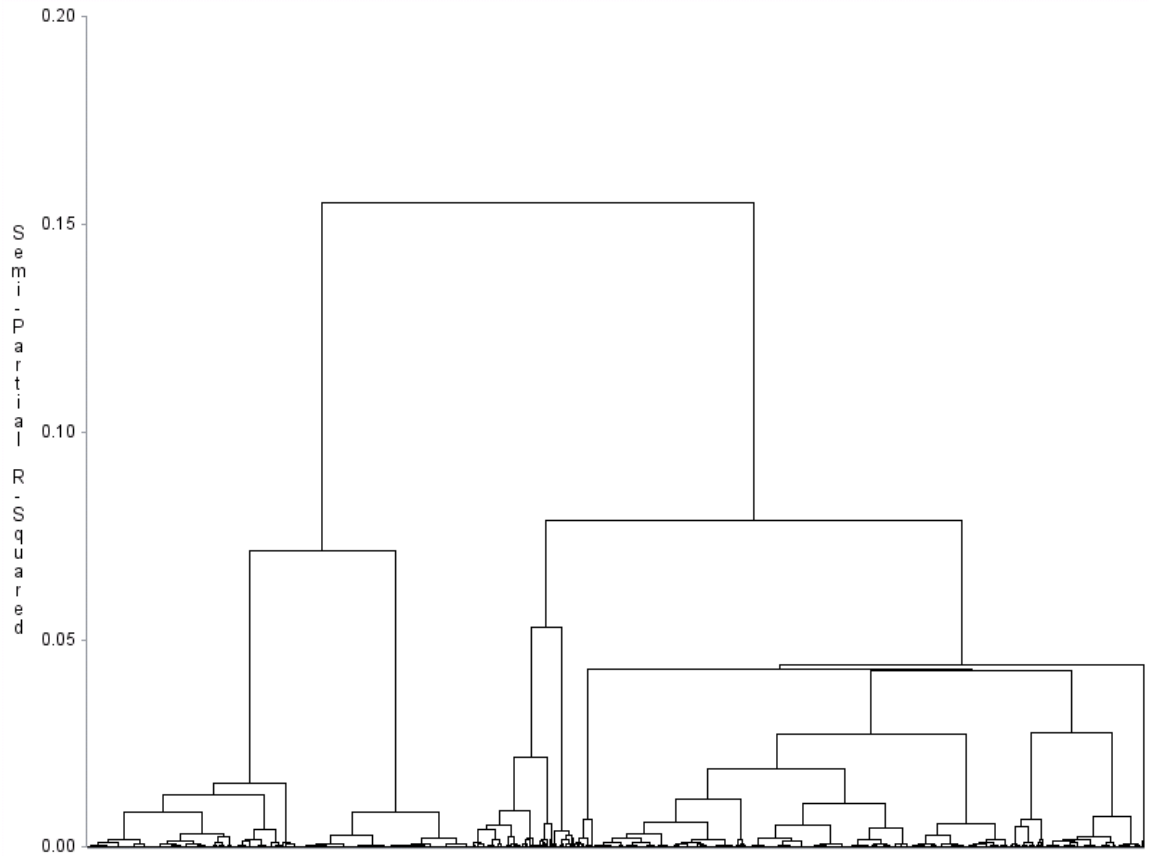

Figure 2: Dendrogram Produced Using Interval Variables Only

6) Produce suitable plots to investigate the nature of each cluster. Hence illustrate the validity of your cluster solution by profiling one of your clusters and interpret the factors that make it unique. Draw any conclusions regarding your cluster solution.

(8 marks)

7)

a) From within the cluster results, obtain the decision tree for your cluster solution. Discuss this tree explaining how many leaves it has, its depth and any other relevant features. (Little credit will be given for just writing the tree out in words.) Give **two** examples of how the results of the tree support your answer to question 6) above.

(6 marks)

b) Explain the disadvantages and advantages of using this tree rather than a new tree, built by adding a tree node to the stream.

(3 marks)

8) Use all of your results above to discuss how the cluster solution may be utilised by the bank. How may the results of any of the analyses above be utilised to carry out further supervised data mining? What other possible targets might be appropriate?

(8 marks)

9) Add a partition node and a tree node to your diagram. Set the field ***loan*** as the target in the input (***czech16)*** node. Connect the input (***czech16)*** to the partition node and the partition node to the tree node. A typical final Enterprise Miner diagram is shown in Figure 2

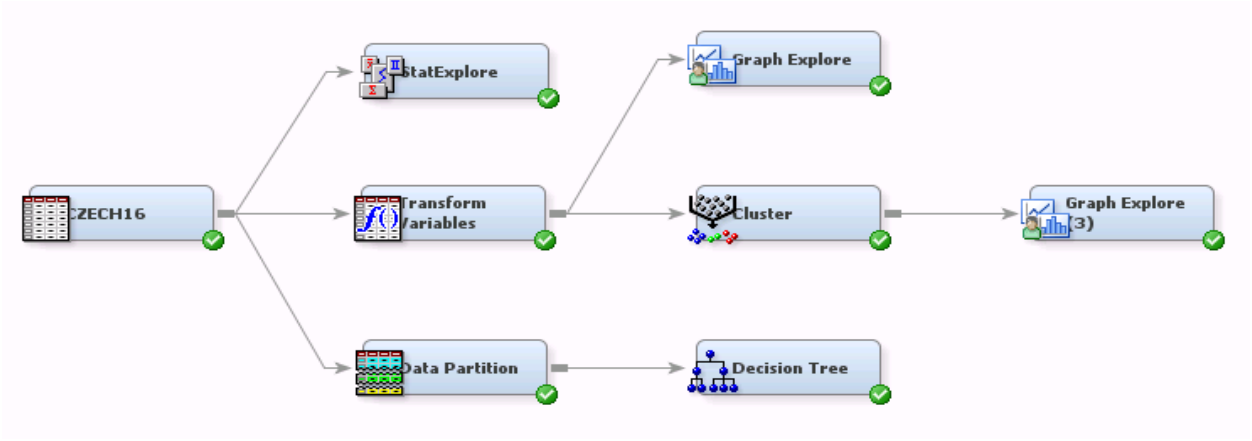

Figure 3

a) Run the tree node and examine the results.

(1 mark)

b) Explain the type of customer that may be more likely to have a loan. How might this information be used by the bank? How reliable is this tree likely to be on new data?

(8 marks)

c) A customer has the following profile:

Age =30 years, ***card*** = "N", ***days*** = 500 days, ***mInsure*** =0Kč, ***mInterest*** = 3Kč, ***moverdt*** = 0Kč, ***mpension*** = 0Kč, ***Mregsal*** = 15,000Kč, ***mstment*** = 15Kč, ***mcardwdl*** = 5Kč, ***mcashcr*** = 0Kč, ***mcashcrcgap*** = 40days, ***Mcashwd*** = 7,000Kč, ***mcashwdcgap*** = 30 days, ***mhousehold*** = 10Kč, ***mothbwd*** = 0Kč, ***second*** ="N", ***Sex*** = "F".

Use your tree to find out the probability they will have a loan. Hence would you predict that they would have a loan? Justify your answer.

(5 marks)

10) Find an example in the academic literature (i.e. an article or paper from a reputable academic journal) where data mining has been used. You should select an article where either clustering or a decision tree or another data mining technique has been used. The article must be clearly referenced using the APA** format now used at SHU. In no more than 500 words discuss your chosen article. Briefly describe the situation in which it was applied, what was discovered and whether (with reasons) you think data mining was used effectively.

(10 marks)

** Help on this can be found in the library resource: http://libguides.shu.ac.uk/referencing You must be logged onto the library for this link to work).

1.8 *Report*

Although this assignment comprises various data analyses together with a discussion of a piece of academic literature, you should write all your findings in a technical report. This should include the technical reasoning of your findings but be written in a way that would be suitable for the data science section of the bank. It should include any relevant output either in the main body of the report or in a suitable appendix (as appropriate). Marks will be lost for aspects that depart from a professional report such as:

- using an inappropriate style for this purpose (e.g. use of first person),
- poor English,
- lack of numbered headings and sub-headings,
- lack of figure labels,
- being untidy,
- inappropriate screenshots
- no page numbers
- inappropriate balance between the appendix and the main report,
- including the questions in the assignment
- incorrect length

The report should not be of more than 3000 words **including** question 10.

(10 marks)

1.9 *Submission*

You must submit your final work electronically to TWO areas on Blackboard. One is via Turn-It-In so as to generate a TURN-IT-IN report. The other is as a "SHU assignment" which will be the copy that is actually marked, but the electronic turn-it-in reports will be checked online for plagiarism.

Failure to submit **both** forms of your work will mean that it will not be marked.

Total Marks available: 100 marks
(Will contribute 50% to final module mark)

APPENDIX 1

item	meaning	remark
account_id	Identification of the account	
age	Age of primary account holder	
card	Whether the account holder has a credit card with this bank	
days	Total number of days between first transaction and last	Over time period of database which is from 1/1/93 - 31/12/98.
loan	=N if the account has a loan =Y if the account does not have a loan	
mInsure	Average of Insurance payments	Measured over the period of the transaction data base (see comment above). The shaded rows represent withdrawals (or debits) and the un-shaded credits (money paid into the account).
mInterest	Average value of interest payments	
moverdt	Average value of overdraft payments	
mpension	Average value of pension payments	
mregsal	Average value of regular salary payments	
mstment	Average value of statement payments	
mcardwdl	Average value of credit card withdrawals	
mcashcr	Average value of cash credits	
mcashcrcgap	Average gap between cash credits	
mcashwd	Average value of cash withdrawals	
mcashwdcgap	Average gap between cash withdrawals	
mhousehold	Average value of household payments	
mothbwd	Average value of other bank payments	
second	=N if no second account holder =Y if there is a secondary account holder	
sex	Gender of primary account holder	