Assignment 2

Customer data analysis: combining managerial assumptions and Computations

Computational techniques and tools for the task:

- *MS Excel* modules (pivoting, visualiasation graphs)
- *Statistica* advanced models (Neural networks, association rules). The Statistica software can be installed from the licenced university resources **inet.muni.cz.** (index, software offer, https://inet.muni.cz/app/soft/licence
- In the first stage of the assignment the customer data set is created from initial data file by applying Pivot analysis in Excel
- During the second stage of assignment the created customer data set is inserted to Statistica software as the data sheet for further computational analysis (Neural neworks, Association rules).
- *Viscovery SoMine (SOM, clustering, cluster research).* The software is installed from the internet, as trial version http://www.viscovery.net/. The same data set created in Excel is used for analysis.

The task is performed in teams of 2-3 people

.

The initial data file is the same for everyone : CRM_data_for_analysis.xls

Data mining procedures:

1. Exploration (data cleansing, e.g. dealing with negative values, transformation of the provided transaction data set by creating new derived variables for building customer data set.
2. Model building by creating variables and applying suggested computational techniques for their analysis
3. Deployment: using the elaborated models for generating insights

Pivot analysis (MS Excel)

Pivot analysis functions are used for creating variables for analysis, making analytical reports and for visualisation.

1. The analysis is perfomed by using all variables computed during the data exploration (transformation) phase. Please Among these variables are:
   1.1. Tenure (duration of life cycle in days)
   1.2. Average purchase during one visit
   1.3. Average number of days between visits
   1.4. Standard deviation of the selected variables
   1.5. Indicators of „good“, „bad“ or other types of customers
   1.6. categorical variables for entitling meaningful subsets, frequency of routes, etc.

   1.7. Variable R (Recency) show the number of days since the last visit till the date set for analysis (you can use the „Today“ or any other date (e.g. end of the year).
   1.8. Variable F indicator is equal to the number of visits of the customer. In our task F is equal to the number of completed transactions without reimbursements).
   1.9. The M is equal to the total value of purchases during all the history of communication.
2. Customer analysis by comparing two segments: individuals and firms. Please use various diagrams for graphical visualisation
3. RFM analysis (Recency, Frequency, Monetary value). The three indicators of R, F and M are computed for each customer.
4. Loyalty and Churn analysis – create variables and define the formulas for their values suitable for analysis of the level of loyalty (defined by yourself) of the customers. Select two variables for defining loyalty (frequency of visits or other variables). and to suggets formula for computing level of risk for churn.
5. Pareto analysis (testing of the statement if the turnover generated by 20% of best customers equals to 80% of total turnover of the enteprise).
6. „Whale curve“ analysis: in order to plot it you have to sort customers by descending order of their turnover values, to compute thier cumulative percent values and to plot to Y axis. In X axis you plot the cumulative percent of the number of customers (e.g. if the enterprise has 10 customers, each of them makes 10% of the enterprise customers, second line will show cumulative of 2 customers which make 20 cumulative percent, etc. Till last line showing 100%). The Whale curve shows what percent of customers (plotted as cumulative from 0 to 100 % in axis) generate related part of the total enterprise turnover (plotted in % in Y axis from 0 to 100).
7. Analysis by popularity among individuals and business travellers, by showing their loyalty breakdown
8. Whale curve analysis and interpetation. Goal seek for desired monetary values.

Each task please fulfil in separate Sheet of Excel. In each sheet please provide your calculation and interpretation of the results.
The outcome – Excel data file including initial data table, its pivot transformations, comments of applied analysis and the customer data table for further computational analysis

Neural network analysis

Copy the prepared customer data from Excel to Statistics
Assign the names in the first line as Variable names
Statistics-Neural Networks menu (SNN)
Select analysis :
The intelligent problem solver can select the best network for you, but we shall analyse the NN types by selection, therefore we‘ll use „Custome network designer“.
Quick menu:
Problem type- classification
Variables: define which are input and which are output variables. Variable types: here you have to define the type of variable as you created them: continuous for historical numeric values, categorical for the meaningful subsets as you called them (e.g. loalty category 1,2 or 3) .
Compare the ability to classify the data set for three types of networks:
-Multilayer perceptron

- Radial basis function
-Probabilistic neural network
When you get results, please analyse them by interpretation of numeric and visual reports (Quick menu)

a) What is the train, select and test performance ? (similar in each phase, overfit, etc)
b) What are predictions?
c) Descriptive statistics (recognize correctly or wrongly)
d) Sensitivity (importance of variables for classification)
e) Response surface for interpretation of visualized results

Select the best neural network type. You can see the suummary of all models in *Networks/Ensemble*
For illustration of your report use *Edit-Screen catcher-Capture rectangle*
Register this NN model *Select model*
In *Statistics-Neural Networks menu (SNN)* take „Run existing model“ . Here you can apply the selected best model for analysing the other data subset of the same problem.

Association rules

Make sure that the data set has at least four categorical variables. Use the appropriate menu item „Association rules“. Modify the required level of confidence, correlation and support of the rules. Define the strongest rules of your data set.

Viscovery Somine (optional)

Import the datafile from Excel Spreadsheet (you may need to have this file saved in Excel 2003 format.)
The Self organizing map will separate data in the number of cluster fdefine by you, only in this case the clusters will mean data group according to the similarity revealed by the SOM algorithm, not by the output variables created by yourselves.

- Explore the variable value ranges withing clusters
- Find separate customers in the clusters and their characteristics.
- Input the characteristics of „new“ customer record and find out the cluster where it belongs.
- Analyse clusters by the impact of separate variables. Are there any clusters which you can decribe according to articular value ranges of your variables in the data set