**Data Mining, Tutorial 1: Working through CRISP-DM**

**Business Scenario:**

You have been requested to analyse a dataset to determine how a bank can identify whether or not a customer is a good or bad risk for a loan.

The bank want to create a model which will read in a persons details, and predict if that person is a good or bad risk.

**Dataset**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **age** | **gender** | **income** | **other loans** | **amount on deposit** | **value of loan** | **risk** |
| 123 | 25 | M | € 30,000.00 | N | €20,000.00 | €25,000.00 | Good risk |
| 124 | 26 | M | € 25,000.00 | Y | €1,000.00 | €25,000.00 | Bad risk |
| 125 | 50 | C | € 50,000.00 | Y | €40,000.00 | €10,000.00 | Good risk |
| 126 | 54 |  | € 25,000.00 | N | €20,000.00 | €50,000.00 | Bad risk |
| 127 | 30 | F | € 25,000.00 | N | €20,000.00 | €50,000.00 | Good risk |

**Stage 1: Business Understanding**

**1.1** What are the business objective(s)?

(Do they just want to understand their customer base better, or do they want to make predictions? ) Is there one or more than one objective? Are they very clear about that objective, or is further brainstorming necessary?)

**1.2** Assess the situation. Assume that there is sufficient historical data from which to build a model, that access has been approved (internal politics), and ethical/privacy considerations had been addressed. Also assume there is adequate time to complete the project on time and within budget.

**1.3** What are the data mining objectives?

For each business objective above, what is its problem type:

|  |  |
| --- | --- |
| **Problem Type** | **Tick if applicable** |
| Clustering |  |
| Classification |  |

**1.4** Produce a project plan. (Not needed for this exercise).

***Stage 2: Data Understanding***

**2.1** Collect the data – available above.

**2.2** Describe the data.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute name | Data Type:  Character or numeric?  Is it a key field? | Min/Max values; range | Missing values? |
| ID |  |  |  |
| Age |  |  |  |
| Gender |  |  |  |
| Income |  |  |  |
| Other Loans |  |  |  |
| Amount on deposit |  |  |  |
| Value of loan |  |  |  |
| Risk |  |  |  |

**2.3** Explore the data

At this stage, do some variables look more predictive than others? Which ones look most promising?

\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2.4** Verify the data:

a) quantity of data – There are 7 attributes, is there over 70 rows of data? \_\_\_\_\_\_\_\_\_\_\_\_

b) quality of data – are there missing values? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

c) Do value ranges seem reasonable for each attribute? \_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Stage 3: Data Preparation***

**3.1** Select the data. Based on exploration done in the last phase, are there variables that will not be useful which can be discarded at this stage?

Attributes to discard, if any: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**3.2** Clean the data: Are there rows with too many missing values that should be discarded at this stage? Are there columns with too many missing values that should be discarded at this stage. Is it possible to deduce missing values from other data in the data set.

Attributes to discard, if any: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**3.3** Construct the data. Are there additional attributes to be added which can be derived from existing data. Is there a value that could be calculated from the dataset which may be useful?

**\_\_\_\_\_\_\_\_\_\_\_\_\_**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Stage 4: Model the data**

**4.1** Decide which model to use: Decision Tree

**4.3** Generate the model: Create a number of **if..then** Statements to predict values, starting with the most predictive variable, e.g.

1. If income > 50,000 then good risk
2. Else if . . . .

**4.4** Assess the model.

Is the model reasonably accurate? (data set too small to determine this)

Has the model generated met the data mining objective(s)? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Stage 5: Evaluate the model**

Has the model met the business objective(s)? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_