**Decision Tree Process**

This process guide shows the steps from SAS Enterprise guide coded data, through an Excel formatting spreadsheet, to the R code required to produce a decision tree, the display is then return to Excel for further use.

**Create the SAS code**

Write the SAS code to include the Rating Factors, Premium and Claims information, export as a .csv in the folder of your choice. Where possible replace blank values with 0 (for numeric) or ‘N/A’ (for strings, note don’t use ‘NA’ as this will cause errors in R).

In the SAS process flow, right click on the table that was exported and selected Properties. Select columns and Copy to clipboard.

A screenshot of a computer

Description automatically generated

**Open the Excel spreadsheet**

In the Excel spreadsheet go to the ‘1a. Original Dataset Format’ tab, select call A1 and press ctrl+v to paste the details from SAS.

A screenshot of a computer

Description automatically generated

Select the ‘1.Set Files’ tab.

A screenshot of a computer

Description automatically generated

In cells C3 and C4 write in the file path, which can be copied from the Explorer path. C4 has the name of the file exported from SAS, C3 just has the folder location.

Cells D3 and D4 contain formulas which create the R code, select and copy the cells.

**Open R code part 1**

Open R code ‘1. Import Data’

A screenshot of a computer

Description automatically generated

Select lines 6 and 7 and paste. On line 23 write in the file path to save the result. Use ctrl+alt+r to run the code.

Open the resulting .csv, which will be in the path you specified, named ‘1b Result’. Copy all of the data and return to the Excel spreadsheet, select ‘1b. Dataset in R’ tab and paste in the new data.

A screenshot of a data sheet

Description automatically generated

Select ‘2. Data Constraints’ tab, this can be used to set Where clauses.

A screenshot of a computer

Description automatically generated

Leaving blank is fine.

**Open R code part 2**

Open R code ‘2. Data Constraints’.

A screenshot of a computer program

Description automatically generated

Press ctrl+alt+r to run the code. A .csv file will be saved in the folder specified in step 1, open this file and copy the data.

Go back to the Excel spreadsheet and select the ‘2b. Dataset in R’ tab. Paste the information from the .csv file into cell A1. The columns F to H are formulas and will populate automatically.

A screenshot of a computer

Description automatically generated

Check column G to see if there are any errors. If there are this is due to data formats not matching, if the format shown in column B is correct, then go back to the ‘1a. Original Dataset Format’ and amend the format shown there. If the format is wrong in column B then check the data in .csv, replace any unusual or blank values then re-run R code part 1 and 2. Re-copy and paste the data for step 2.

Once Column G all says either Fine or Ignore then the Drop Flag and Hotkey Flag columns (I and J) can be completed. If there are any Variables that are not required then put a Y in the Drop Flag column (reference ID’s aren’t required, other reasons to remove is if the data used is used for multiple reports).

For the Variables that are to be included, any which are Character format will need to be hotkeyed, put a Y in the Hotkey Flag column.

**Open R code part 3**

Open R code ‘3. Data Prep’

A screenshot of a computer

Description automatically generated

Return to the Excel spreadsheet and click on the ‘Columns to Hotkey’ button. From the ‘3a. Character Columns to Hotkey’ tab copy the data in column B. Back in the R code Paste the data in the ‘3a. Character Columns To Hotkey’ section.

Go back to the Excel spreadsheet and return to the ‘2b. Dataset in R’ tab, click on the ‘Columns to Drop’ button. From the ‘3b. Columns to Drop’ tab, copy the data in column B, return to the R code and scroll down to the section commented as ‘3b Columns to Drop’, paste the data here.

Return again to the Excel spreadsheet and select the ‘2b. Dataset in R’ tab, click on the ‘Format Columns’ button. From the ‘3c. Format Columns’ tab copy the data in column B. Go back to the R code and scroll down to the section commented as ‘3c. Format columns’, paste the data in here. Press ctrl+alt+r to run the code.

**Open R code part’s 4 and 5**

After this code has run open R code ‘4. Train Set’, Press ctrl+alr+r to run then open R code ‘5. Decision Tree’.

The R code run in step 3 will have created a .csv file called ‘4. New Dataset’ in the folder specified in step 1. Open this file and copy the data.

Go back to the Excel spreadsheet and select the ‘4. New Dataset’ tab. Paste the data into cell A1.

A screenshot of a computer

Description automatically generated

Put a Y next to the factors to include in column F. The factor that the Variables will be summarised by does not need to be included here, also Character variables can’t be included, the Hotkey process will have turned all Character variables into extra variables with a 1/0 outcome, the original character variable will still show, but do not include.

Click on the ‘Factors to Include’ button. Go to the ‘6. Factors to Include’ tab and copy the data in column B. In the part 5 R code paste this data into the section commented as ‘6. Factors to Include’.

A screenshot of a computer

Description automatically generated

Above this section is the Target Variable to be used.

Below the ‘6. Factors to Include’ section is a line which starts with roc\_tree, next to the $ sign enter the target variable again.

Press ctrl+alr+r to run. The decision tree will display in the bottom right corner.

**A diagram of a network

Description automatically generated**