**MAX522/MSC615 Data Preparation Exercise Due Sept 15th, 6:00pm CDT**

1. Load in credit.hmwk.sav [Statistics] using an appropriate data source node.
2. Insert a Type node Verify that the Target is Credit Risk and the Inputs exclude ID. Are the type classifications acceptable? If not, what change(s) have you made? Connect a Table node to instantiate the data set OR Read Data in TYPE. Explain your actions.
3. Eliminate any duplicate records if appropriate. Describe your actions briefly.
4. Check for missing data. Resolve as you deem appropriate. Briefly explain your steps.
5. Check INCOME variable for outliers. Resolve as you deem appropriate either using the Supernode in Data Audit OR Specify/Coerce functions in TYPE node.. Briefly explain your actions.
6. Review the Income data item for skewness? What adjustment? Explain your action.
7. Balance the input data file such that risk=2.0 class will have 700 records (or 0.475). Be careful re: syntax, e.g., risk = 1.0. Hint—Partition AFTER performing Balancing of input data.
8. Connect a Partition node to the ? Node and specify a 60% training and 40% test partition.
9. Any nodes to be cached? Cache ON freezes the data records contained at this node. If you want to rerun a model with the SAME data; turn Cache ON with the On/Off switch found in the list of actions when you right click on the node.
10. Remember that nodes only impact a subsequent modeling node IF chained together in a single, sequential path!
11. Perform Data Preparation Sensitivity Analyses per team number identification: Use **another** copy of credit\_hmwk.sav Data Source node with ONLY TYPE and DISCTINCT nodes connected. READ ALL values, Close, then eliminate any duplicate records and proceed with the appropriate data preparation task below. For each of the Task Groups 1-4 below, connect a TABLE node to the last node that you are working with in your stream. After you make your required modification, RUN the Table node BEFORE you record the requested variable values.

**Teams 1-5**: Compare these 3 methods for addressing Missing data—(1) delete records that have missing data for any variable, (2) replace missing income data with the mean value of income, and (3) do not delete any records or impute an values.. For each of these 3 approaches, report the mean, standard deviation and skewness of income (in Data Audit node).  
**Teams 6-10:** Change the method for addressing outlier values for Income: (1) In TYPE node for Income, Click in the Income—Missing column cell—SPECIFY Then change the value for Upper and Lower limits to mean +/- 2 std. deviations using COERCE—what are the mean and std dev. Values for Income? (2) Repeat #1, but change the UL and LL values to 2.5 std. deviations Record the Income mean and std dev values, and (3) Repeat the operation (2), but set the UL and LL values to 3.5 std dev’s. Record the mean and std dev values for Income.   
**Teams 11-15:** Change the method for addressing skewness of Income variable: (1) use the natural log transformation and record the mean, std dev and skewness values for income, (2) use the log10 transformation and record the mean, std dev and skewness values for income, (3) repeat (1), but choose sq root transform and record requested values,

**Teams 16-19**: perform the same Balance reduction operation as indicated in Requirement 7 above and Insert a Data Audit node between the Balance node and Table node but each time change the reduction factor in the Balance node to: (1) 0.50, (2) 0.425, and (3) 0.40. After you Run the Table node for each of the 3 operations, record the mean, std dev. and skewness values for risk from the Data Audit node located After the Balance node.

**Team 20**: Complete the above Data Preparation Sensitivity analyses listed for Teams 16-19 AND Teams 6-10 using your 3 person additional capacity.

1. Submit your str and doc(x) files including your team number in the filename. Create a 4X4 or 4X2 table to summarize the values displayed from your Data Preparation Sensitivity Analyses outcomes, e.g.,

|  |  |  |  |
| --- | --- | --- | --- |
| **Conditions** | **Mean** | **Std dev** | **skewness** |
| **Option1** |  |  |  |
| **Option2** |  |  |  |
| **Option 3** |  |  |  |

1. **FOR R Homework OPTION (FYI: this option will require more effort)**
   1. Copy all your lines of code into a WORD docx file
   2. Include the output or results from executing commands (copy/paste into WORD after the command line for execution.
   3. Use Comment lines as appropriate to label the next set f R code lines performing a task, e.g., deduplication of rows
   4. Perform the appropriate above Data Preparation Sensitivity Analyses and document results—could be located separate from the command code lines