# Handling Transformation and Outliers

Dr. Goutam Chakraborty

# Outline For This Session

- Data exploration and data preparation before data mining
  - Handling missing (and/or unexpected) values,
  - Handling transformation ,
  - Handling extreme values (outliers)

# Data Transformation

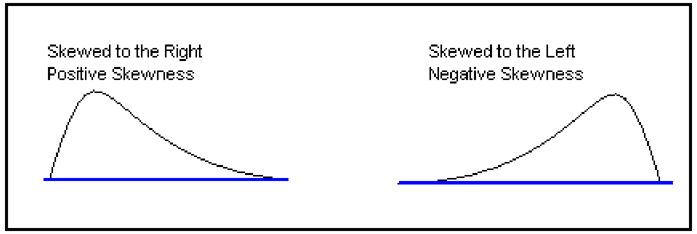
- Why should you consider transforming data?
  - Skewed distribution of numerical variables create problems in many modeling algorithms.
  - Numerical variables with very high variance is likely to emerge as more important in some modeling algorithms.
  - For categorical variables, it is often impractical to use a very large number of classes in models.
  - Often, predictive ability of models improve (particularly when model is applied on unseen data) when independent variables are transformed to be more symmetric (or, Normal).
  - In segmentation problems, skewed variables often create many small segments
- Is there a downside of transformations?

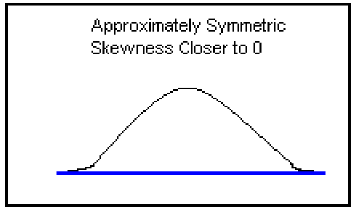
# Normalization or Scaling Transformation

- Also, called r**ange-standardization** or **linear scaling**
- Transformed X= (X- Min. X)/(Max. X – Min. X)
- Range of New X is (0,1)
  - Produces no distortion in the data as it retains relative position of values (if X1>X2, then NewX1 > NewX2)
  - Does not change shape of distribution (if X was skewed so is Transformed X)
  - Some modeling tools may do it automatically (Neural Net)
    - Others could benefit from it!
- What if we get out-of-range values in new data?
  - Clip it to maximum/minimum
  - Use special scaling such as SoftMax

# (Distributive) Transformation for Numerical Variables

- Power Series Transformation : Goal is to apply a transformation such that the transformed variable closely resemble a Normal distribution. Changes the **shape as well as range** of the distribution:
  - Log — Variable is transformed by taking the natural log of the variable.
  - Square Root — Variable is transformed by taking the square root of the variable.
  - Inverse — Variable is transformed by using the inverse of the variable.
  - Square — Variable is transformed by using the square of the variable.





# Useful (Distributive) Transformations in SAS EM

- **Best Power Transformations** : These transformations are a subset of the general class of transformations that are known as Box-Cox transformations. The Transform Variables node in SAS EM supports the following best power transformations:
  - **Maximize Normality** — This method chooses the transformation that yields sample quantiles that are closest to the theoretical quantiles of a normal distribution.
  - **Maximize Correlation with Target** — This method chooses the transformation that has the best squared correlation with the target. **This method requires an interval target**.

# Useful (Distributive) Transformations in SAS EM

- **Binning Transformations** : These transformations enable you to collapse an interval variable, such as debt to income ratio, into an ordinal grouping variable. There are three types of binning transformations.
  - **Bucket** — Buckets are created by dividing the data values into equally spaced interval based on the difference between the maximum and the minimum values.
  - **Quantile** — Data is divided into groups that have approximately the same frequency in each group.
  - **Optimal Binning for Relationship to Target** — Data is binned in order to optimize the relationship to the target. This method requires a **binary target**.
  - **Group Rare Levels Transformation** **:**This is available for **class variables only**. This method combines the rare levels (if any) into a separate group, _OTHER_. You can use the **Cutoff Value** property to specify the cutoff percentage. Rare levels are the variable values that occur less than the specified cutoff value.

# Transform Node (Modify Tab)

- This node enables you to create new variables that are transformations of existing variables in a data.
  - Transformations can be used to stabilize variances, remove nonlinearity, and correct non-normality in variables to help improve model performance.
- The default settings for this node is no transformation
  - You choose the variables you want to transform
  - You choose the type of transformations

# Demonstration of Transform

- Add a **Transform** node (under **Modify** tab) to the right of the Metadata node.
- Connect the Metadata node to the Transform node
- Check property panel
  - Under Sample properties, change Method to Random, Size to Max.
  - Under Score, **change Hide to No**.
- Right-Click Transform node, select edit variables, Select GiftCntCardAll,. Select Explore
  - Note the skewness in the variables. Close Explore.
  - Change Method from Default to **Max Normal**.
  - Change Hide to No under Score
- Right-click on Transform node, select Run. Examine results.
- Now, change method to **log** (as indicated by Max Normal) and rerun the node
- Explore distributions of original and transformed variables via another Metadata node

# Handling Extreme Values (Filter)

- Extreme values in variables are problematic because they may have undue influence on model.
- Often in predictive modeling applications, training data set is filtered to exclude observations, such as outliers or other extreme observations, that we **do not want** to include in model building.
- Filtering extreme values from the training data tends to produce better models because the parameter estimates are more stable.
  - Sometimes transformation fixes extreme values; other times filtering extreme values may negate the need for transformation.
- Filtering observations typically should not be used to in the **validation, or score** data sets. Why?
  - Since the validation are used for model assessment, you should not filter observations in these data sets.
  - You typically are interested in scoring the outcome of every observation in the score data set. Therefore, you should not filter the score data set.

# General Methods of Outlier Detection

- Univariate outliers are cases that have an unusual value for a single variable.
  - For numerical variables, consider values beyond ±3SD or, ±4SD or, ±5SD away from mean as potential outliers.
    - What's the rationale?
- Multivariate outliers are cases that have an unusual combination of values for a number of variables. The value for any of the individual variables may not be a univariate outlier, but, in combination with other variables, the case is an outlier. This can be detected by
  - Density approach
  - Distance (Mahalanobis) approach

# After Outlier Detection

- Change outlier observation
  - Winsorize it
    - Change the value to the highest or lowest range that you want to consider
    - Use the winsorized data in your analysis
- Delete (keep aside) outlier observation
  - Sample size gets reduced

# Filter Node (Sample Tab)

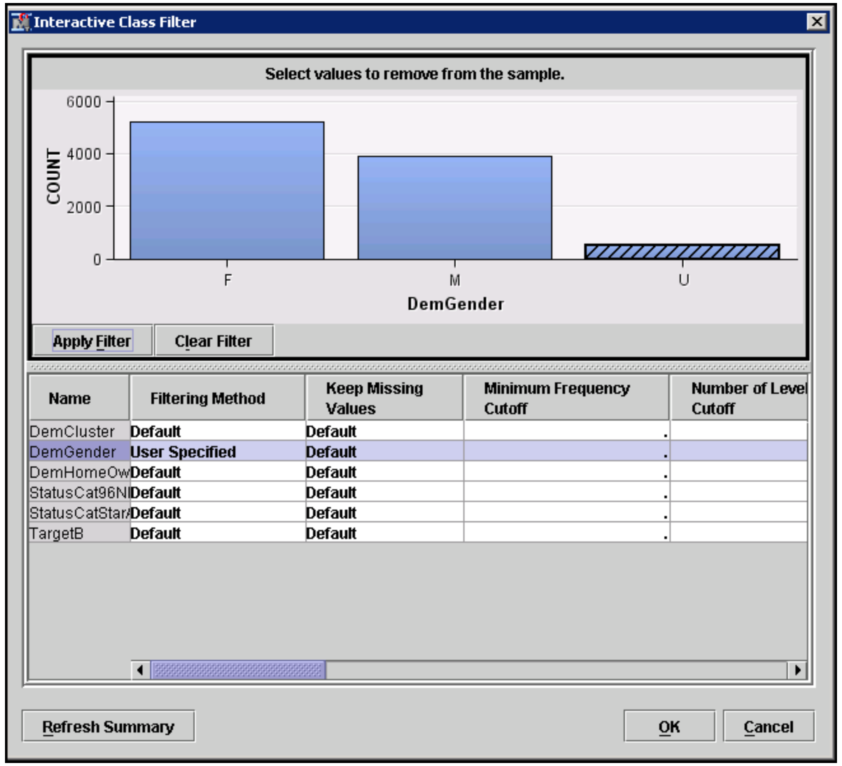
- Use the Filter node to create and apply filters to your training data set
  - You can use filters to exclude extreme outliers and errant data that you do not want to include in your analysis so that you can build a more stable (robust) model.
- Filtering can be applied either:
  - Automatically on your data (you can choose different rules for interval/class variables)
  - Or, *interactively* with one variable at a time

# Demonstration of Filter Node

- Add a **Filter** node (under **Sample** tab) to the right of the metadata node.
- Connect the meta data node to the Filter node
- Check property panel
  - Look at default filtering for interval variables (3 std.dev from mean) and class variables (rare values less than 0.01%)
- Right-Click Filter node, select Run. Examine results.

# Interactive Filtering of a Class Variable (Self Study)

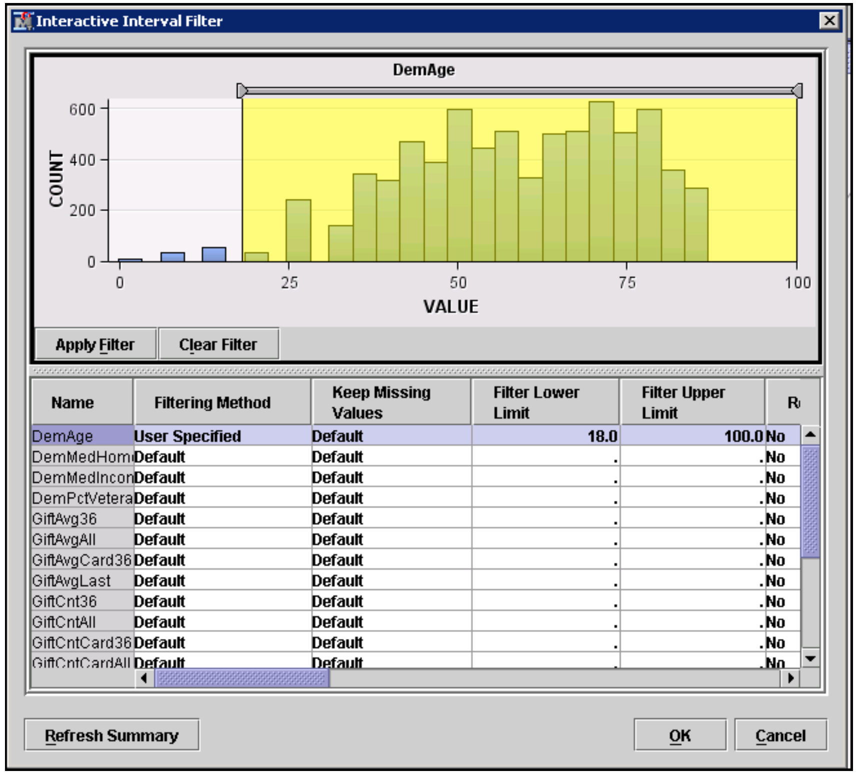
- Click on the **Class Variables ellipsis button** in the property sheet for Filter node
- Select DemGender variable.
- Suppose we want to filter out the U values of this variable.
  - Click on the U value in the graph (it should get shaded).
  - Click Apply Filter button. Note that the Filtering Method changes from Default to User Specified.
  - Click OK.
  - Run Filter node again and take a look at the results



| Name | Filtering Method | Keep Missing Values | Minimum Frequency Cutoff | Number of Level Cutoff |
|---|---|---|---|---|
| DemCluster | Default | Default | . | |
| DemGender | User Specified | Default | . | |
| DemHomeOw | Default | Default | . | |
| StatusCat96NI | Default | Default | . | |
| StatusCatStar/ | Default | Default | . | |
| TargetB | Default | Default | . | |

# Interactive Filtering of an Interval Variable (Self Study)

- Click on the **Interval Variables ellipsis button** in the property sheet for Filter node
- Select DemAge variable.
- Suppose we want to filter out all values less than 18 for this variable.
  - Click on the slider on top of the graph and drag the left slider to more than 18. The shaded area of the graph values will be kept after filtering.
  - Click Apply Filter button. Note that the Filtering Method changes from Default to User Specified. Note also the Filter lower limit (we could have directly enter the number in the limit)
  - Click OK.
  - Run Filter node again and take a look at the results



| Name | Filtering Method | Keep Missing Values | Filter Lower Limit | Filter Upper Limit | R |
|---|---|---|---|---|---|
| DemAge | User Specified | Default | 18.0 | 100.0 | No |
| DemMedHom | Default | Default | . | . | No |
| DemMedIncon | Default | Default | . | . | No |
| DemPctVetera | Default | Default | . | . | No |
| GiftAvg36 | Default | Default | . | . | No |
| GiftAvgAll | Default | Default | . | . | No |
| GiftAvgCard36 | Default | Default | . | . | No |
| GiftAvgLast | Default | Default | . | . | No |
| GiftCnt36 | Default | Default | . | . | No |
| GiftCntAll | Default | Default | . | . | No |
| GiftCntCard36 | Default | Default | . | . | No |
| GiftCntCardAll | Default | Default | . | . | No |