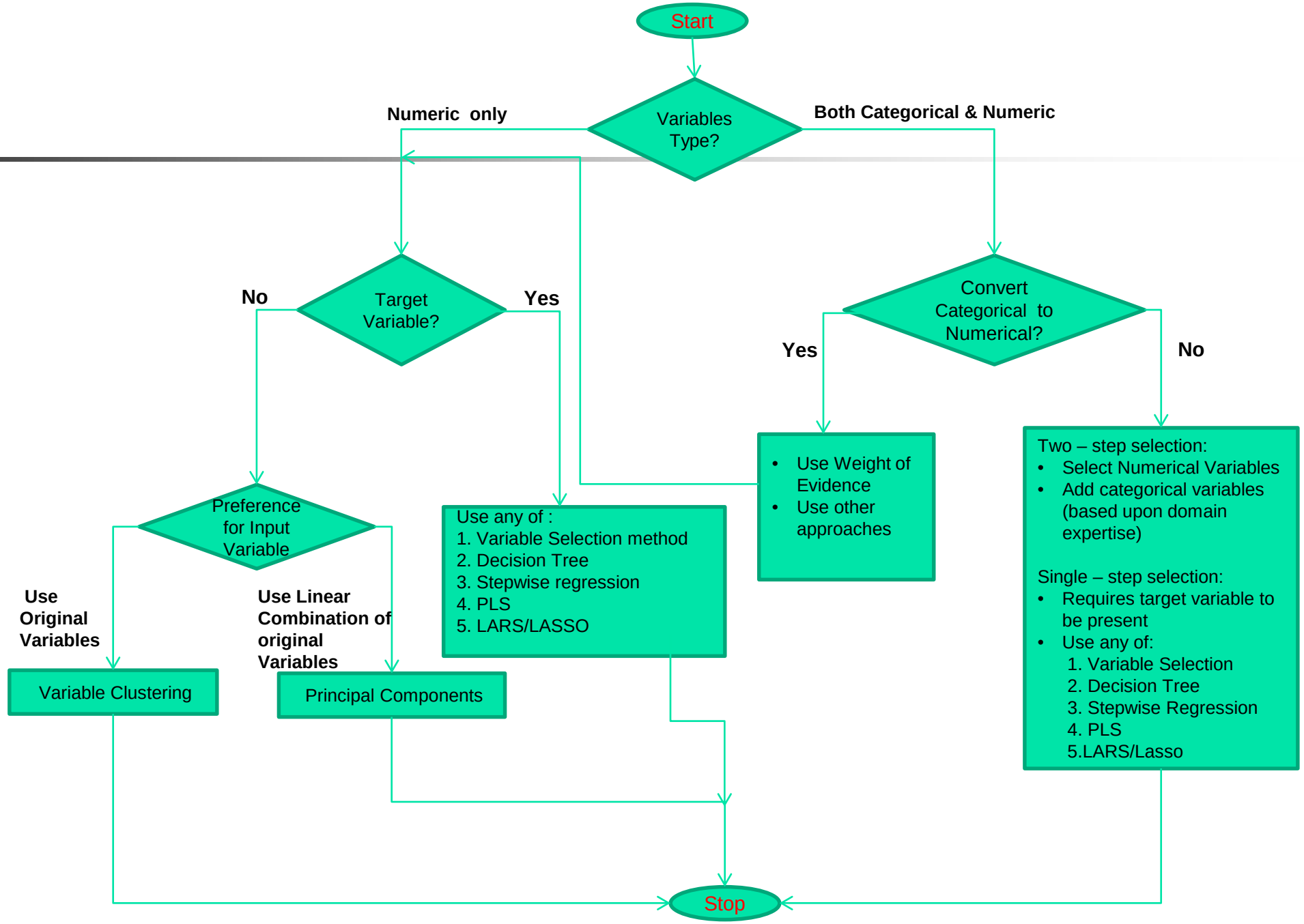
Variable Selection

Dr. Goutam Chakraborty

Variable Reduction versus Variable Selection

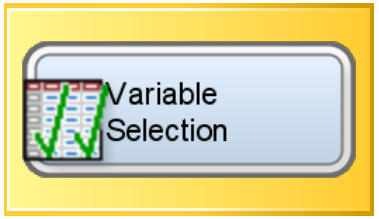
- In data science, both approaches are useful.
  - For predictive modeling
  - For segmentation (clustering)
- Variable reduction is typically **unsupervised**
  - Requires no target variable
  - Reduction due to redundancy of information contained in X-variables
- Variable selection is typically **supervised**
  - Aims at finding a smaller subset of X's that have the best chance of predicting the **target**
  - Requires presence of target variable
    - For clustering (segmentation), a pseudo-target variable will suffice


Start
Variables Type?
Numeric only
Both Categorical & Numeric
Target Variable?
No
Yes
Convert Categorical to Numerical?
Yes
No
Preference for Input Variable
Use Original Variables
Use Linear Combination of original Variables
Variable Clustering
Principal Components
Use any of :
1. Variable Selection method
2. Decision Tree
3. Stepwise regression
4. PLS
5. LARS/LASSO
• Use Weight of Evidence
• Use other approaches
Two – step selection:
• Select Numerical Variables
• Add categorical variables (based upon domain expertise)
Single – step selection:
• Requires target variable to be present
• Use any of:
1. Variable Selection
2. Decision Tree
3. Stepwise Regression
4. PLS
5.LARS/Lasso
Stop

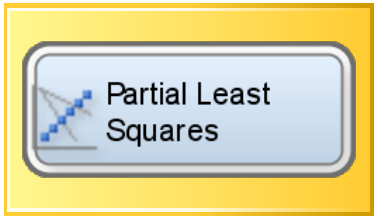
...

Variable Selection Alternatives For This Session

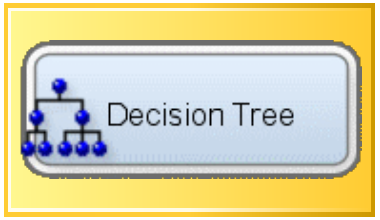
Sequential selection



Univariate + forward selection (R-square)
Tree-like selection (chi-square)



Variable Importance in the projection (VIP)



Split-search selection

Demo

- Follow handout titled “Variable Selection SAS EM_handout”
- Connect Impute Node to Variable Selection Node (Explore tab). **Select** Target model to **R-Square**. Run and view results.
- Connect another Variable Selection node to the Impute node. **Select** Target model to **Chi-Square**. Run and View results.
- Connect Impute Node to PLS Node (Model tab). **Select** Export Selected Variables to **Yes**. Run and view results.
- Connect a **Decision Tree** node to the **Impute** node. Rename the Decision Tree node **Selection Tree.** Enter **1** as the Number of Surrogate Rules value. Select **Subtree** ⇨ **Method** ⇨ **Largest**. Run and view results.

Variable Selection Summary

Method	Number of Variables	Names of Variables
Variable Selection (R Square)	6	G_DemCluster, GiftTimeLast, LOG_GiftAvgAll, LOG_GiftCnt36, LOG_GiftCntCard36, REP_StatusCat96NK
Variable Selection (Chi Square)	15	DemCluster, DemPctveterans, GiftTimeFirst, GiftTimeLast, IMP-Dem_Age, IMP-LOG-GiftAvgCard36, ……..
PLS	7	GiftTimeLast, LOG_GiftAvgAll, LOG_GiftCnt36, LOG_GiftCntAll, LOG_GiftCntCard36, LOG_GiftCntCardAll, PromCntAll
Decision Tree	9	LOG_GiftCnt36, LOG-GiftCntCard36, PromCnt12, DemCluster, DemMedHomeValue, LOG_GiftAvg36, LOG_Gift_AvgLast, GiftTimeLast, IMP-LOG_GiftAvgcard36