Deloitte.

Feature Selection Methods

Data mining to pick predictive variables

Ravi Kumar ACAS, MAAA

CAS Predictive Modeling Seminar

San Diego

October, 2008

Audit.Tax.Consulting.Financial Advisory.

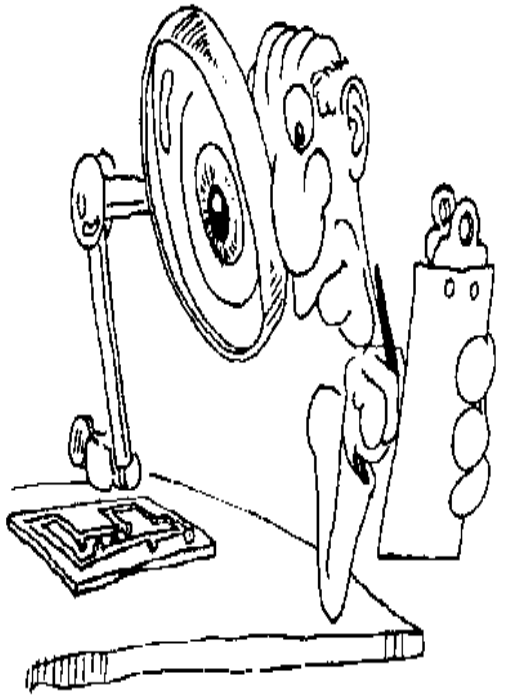
Topics

- Overview
- Funnel Approach
  - Filters
  - Data Visualization
  - Wrappers
- Conclusion

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Deloitte.

Overview



Definition of Feature Selection

Semantics: Data Mining vs Predictive Modeling

- Data Mining
  - KDD: Knowledge Discovery in Databases
  - EDA: Exploratory Data Analysis
  - Open-ended
  - "cast the net wide"
  - "Let the data speak for itself"

- Predictive Modeling
  - Build a model tailored to achieve a pre-specified goal
  - Build on:
    - Results of data mining
    - **Domain expertise! (actuarial & insurance knowledge)**

Actuarial science needs data mining…

… but data mining *also* needs actuarial science

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Some Definitions

- Raw Variables
  - Variables available in the data
- Features
  - Variables constructed from the Raw variables
- Target Variable Y
  - What we are trying to predict.
    - Profitability (loss ratio, LTV), Retention, …
- Predictive Variables $\{X_1, X_2, \ldots, X_N\}$
  - "Covariates" used to make predictions.
    - Policy Age, Credit, #vehicles….
- Predictive Model $Y = f(X_1, X_2, \ldots, X_N)$

Copyright © 2008 Deloitte Development LLC. All rights reserved.

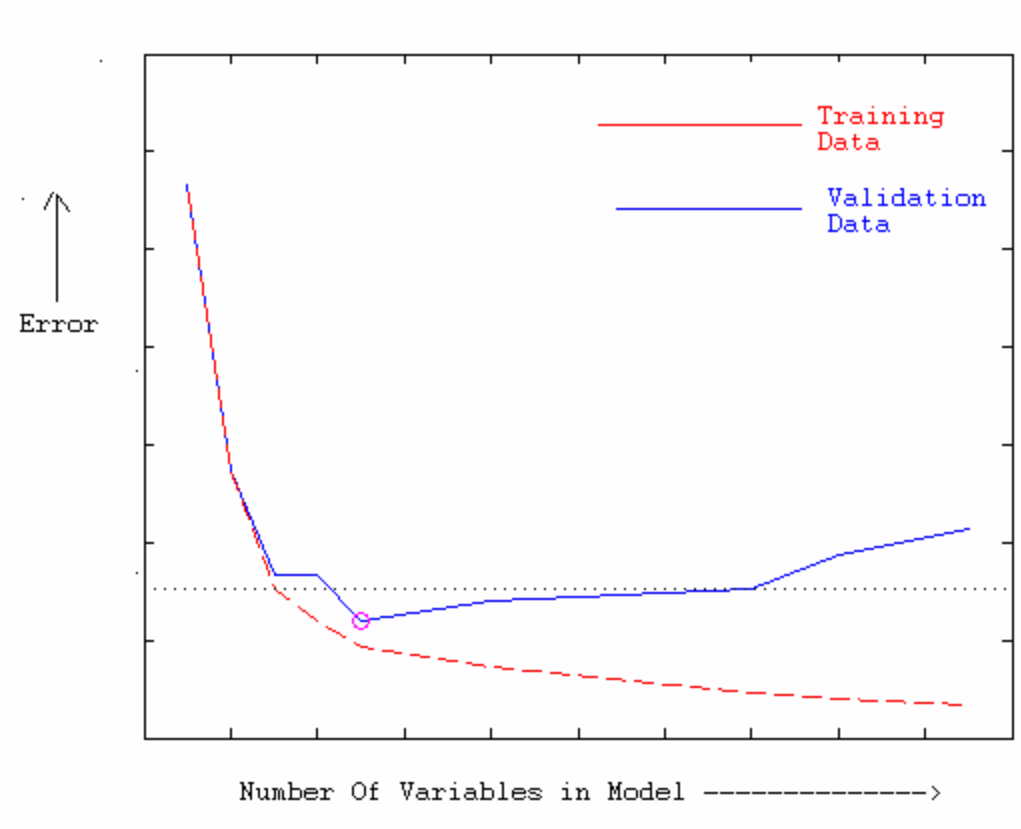
Casting the net wide

- Internal Data Sources
  - Policy Administration Systems
  - Claim Administration Systems
  - Stat Records
  - Billing Systems, Agency Systems, Loss Control data
- External Data Sources
  - Demographic data
  - Credit & Financial Information
  - MVR, Accident records . . . .
- Create 100s of predictive variables from the above data sources
- Feature Selection: From the 100s of variables, pick the best combination of variables that explains the business best

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Reason for Feature Selection: Curse of Dimensionality

- Using too many features reduces predictive performance



Copyright © 2008 Deloitte Development LLC. All rights reserved.

Feature Selection : Things to Ponder

- A Highly Predictive Variable
  - May not translate into a useful variable in a multivariate model
- A useless variable
  - Can become very useful when used with other variables
- Two highly correlated variables
  - May bring complementary qualities to a model

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Feature Selection

- Feature Selection problem is actually a model selection problem
  - NP-hard problem (Cannot be solved in polynomial time $O(n^c)$)
- Unifying theoretical framework is thus lacking
- Example: Selecting the best model from just 20 Variables
  - Number of models to consider: 20+(20*19/2) +(20*19*18/6)+ …
    - More than **1 Million** variable combinations to choose from

Copyright © 2008 Deloitte Development LLC. All rights reserved.

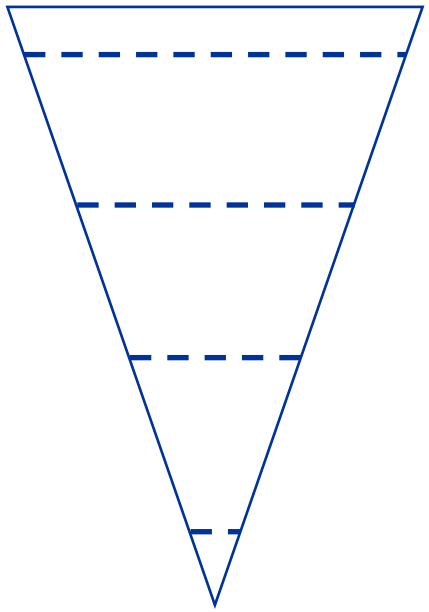
Objectives of Feature Selection Methods

- Improve understanding of underlying business
  - Ease of interpretation/modeling
- Improve Efficiency
  - Measurement Costs
  - Storage Costs
  - Computation Costs
- Improve Prediction Performance of the predictors in the model
  - Improve goodness of fit
  - Reduce the number of variables in model
    - Defy the curse of dimensionality

Copyright © 2008 Deloitte Development LLC. All rights reserved.

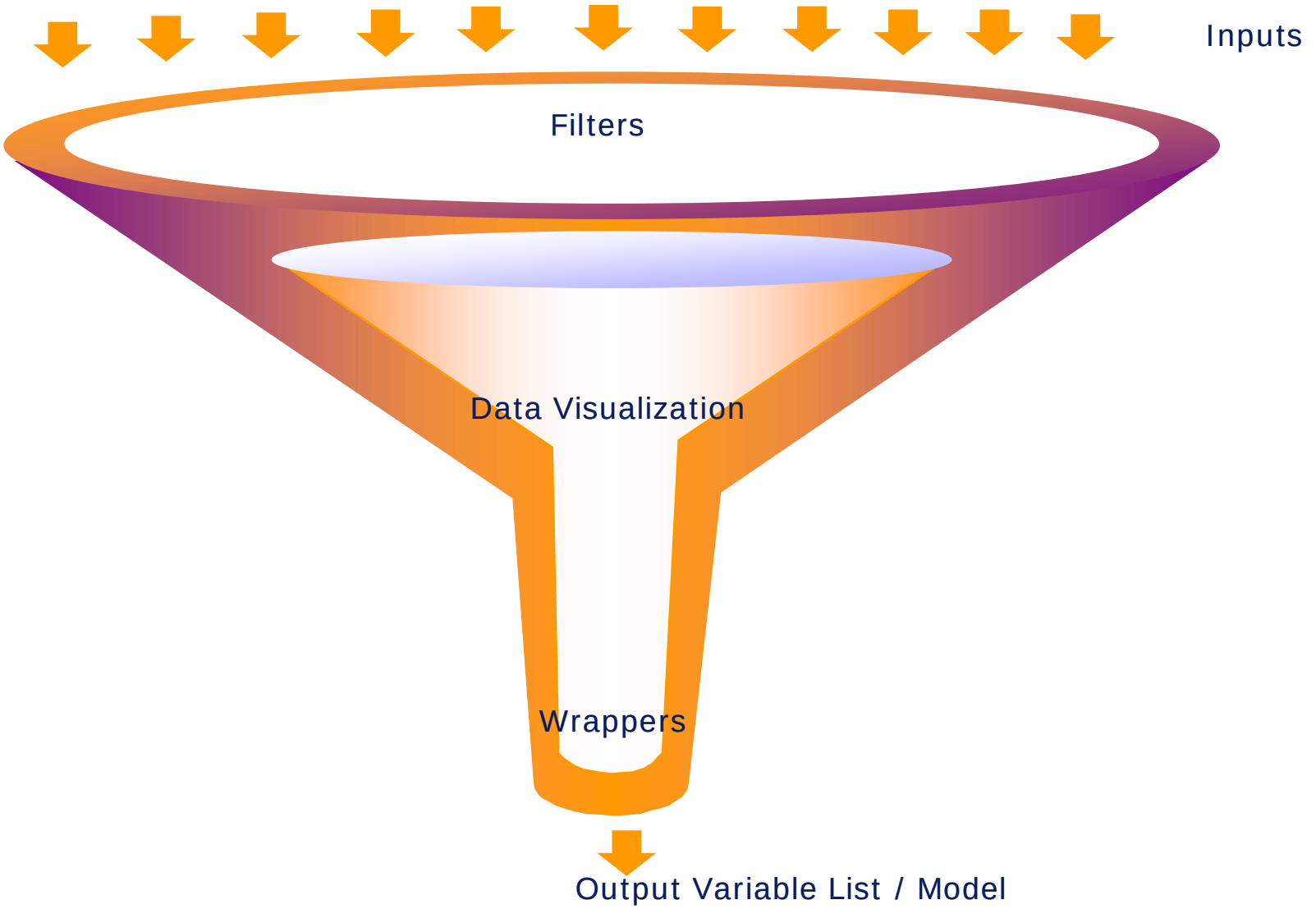
Deloitte.

Funnel Approach



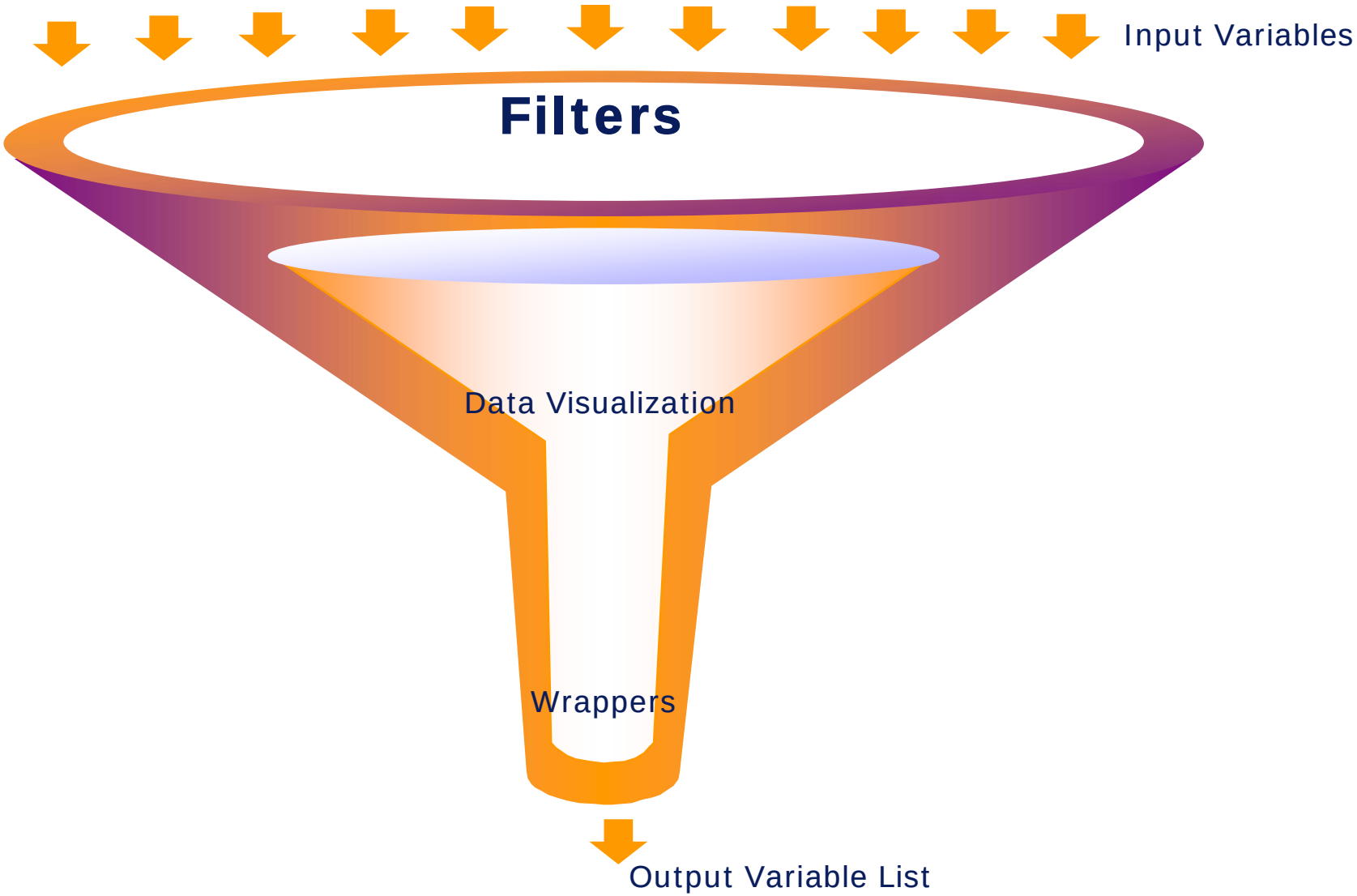
Using the Funnel approach for Feature Selection

Feature Selection: Funnel Approach



Copyright © 2008 Deloitte Development LLC. All rights reserved.

Feature Selection: Funnel Approach



Copyright © 2008 Deloitte Development LLC. All rights reserved.

Filters

- Filters are methods that rank variables based on usefulness
- Used as a preprocessing step
- Uses fast algorithms
- Can be independent of Target Variable
- **Designed to improve understanding of underlying business**

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Filters: How to get most out of filters?

- Simplify the target variable
  - Use a binary target variable? Examples:
    - High/Low Claim propensity
    - Zero/non-Zero claims
    - High/Low Severity
    - High/Low Profitability
- Focus on different subsets of data
  - Examples: New Business, Renewal business, Restaurant Class, Medium size Policies etc.
  - Data Sampling?
- Use many different Ranking techniques
  - K-S Statistics, Linear Models, Decision Trees, etc.
  - Different techniques have different strengths & weaknesses
- Mix in some random number based Placebo Variables
  - For validating variable selection methodology

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Filters: Variable Selection Criteria

- A priori Business/Reliability knowledge
- Variable performance in various simple models
- Correlation Analysis

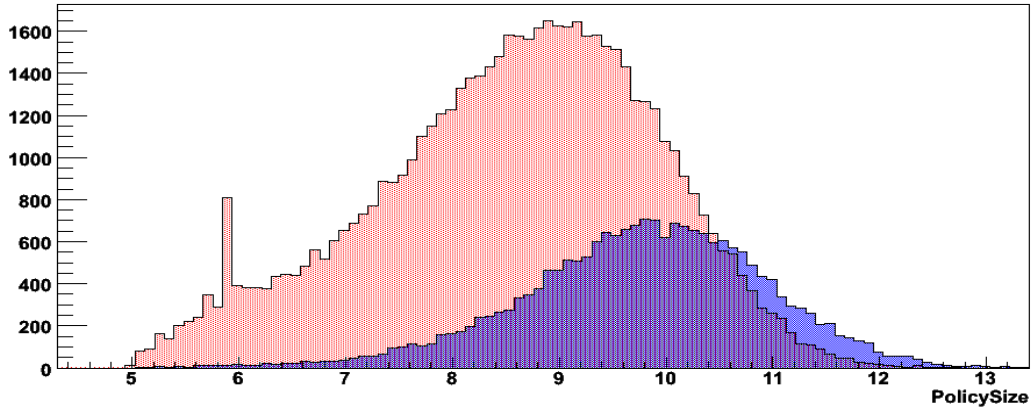
Copyright © 2008 Deloitte Development LLC. All rights reserved.

Filters: Test for Equality of Distribution

- Kolmogorov – Smirnov Two-Sample Test
  - Non-parametric test
  - Tests if distribution of a variable is same across two samples

Divide data into two samples based on a Binary Target
(Example: NoClaim policies vs. Others)
Compare the distribution of Xs in these two samples
Rank the Xs based on K-S test

Focus on features with highest ranks



Copyright © 2008 Deloitte Development LLC. All rights reserved.

Filters: Test for Equality of Distribution

- Sample Rank of variables that influence “Zero claims”

Rank	_VAR_	_KSA_
1	PolicysizeA	41.54114767
2	CvgE	29.45908575
3	FinC	25.15868948
4	FinA	18.72773123
5	PolicySizeB	16.65784683
6	CvgA	16.09490114
7	AgentA	14.53964193
8	ZipD	14.45720589
9	CvgB	14.22495807
10	PolicyYear	11.38747008
11	ZipA	11.31822673
12	.	10.93840937
13	.	9.590276851
14	.	8.138575461
.	.	7.638420449
.	.	7.621321237
185	random6	0.338630276
	.	
	.	
203	random7	0.075358981
217	random8	0.072322166

- Placebo variables are used to validate the method

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Simple Models: Stepwise Regression

Pros

- Ease of use
- Does give some useful insights about the data

Cons

- Variables are picked based on Training data only
- No penalty for picking too many variables

Few tricks

- Try different target variables
- Run it separately for various variable groups
- Include random variables (as X's) to understand if the method works for the problem
- Good idea to run Stepwise Regression multiple times, each time removing the top few variables from the previous run

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Filters: Stepwise Regression

- Sample Rank of variables that influence “Zero claims”

Step	Variable Entered	Variable Removed	Partial R-Square
1	PolSizeA		0.008
2	PolSizeB		0.007
3	FinB		0.005
4	AgentB		0.003
5		PolSizeB	
6	FinA		0.002
.			
.			
.			
15	random3		0.001
16	ZipD		0.001

- Placebo variables are used to validate the method

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Simple Models: Decision Trees

Pros

- Ease of use
- Non Parametric
- Not Sensitive to outliers in data
- Great way to explore/visualize the data
- Variables picked based on performance on Test data
- Can apply Penalty for picking too many variables
- Can give insights on variable interactions

Cons

- Does not pick linear relationships easily
- Unstable models in the presence of correlated variables

- Few tricks
  - Try different splitting rules (Gini,Entrpoy, Twoing etc)
  - Try different cost complexities for pruning the tree

Copyright © 2008 Deloitte Development LLC. All rights reserved.

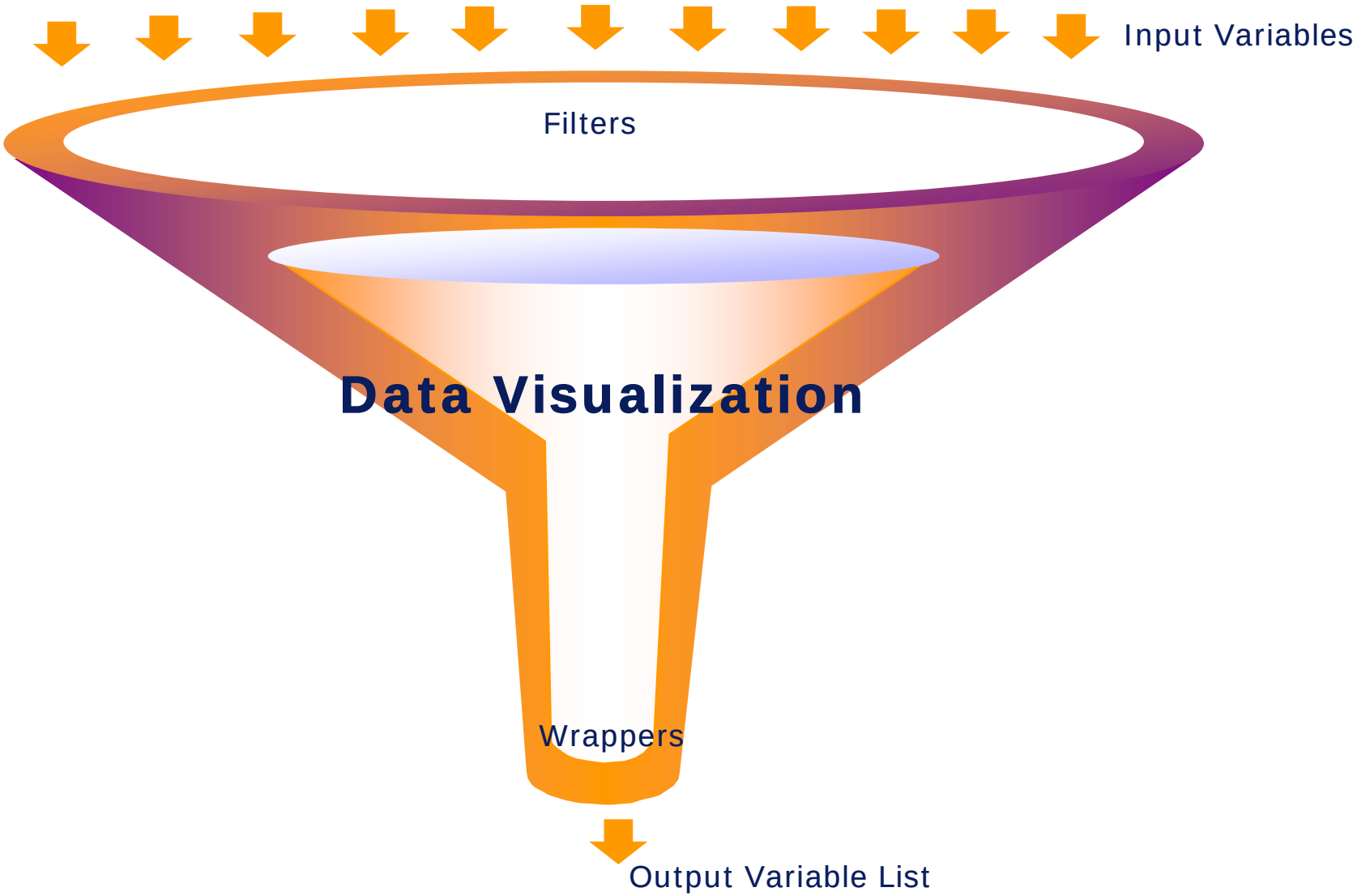
Filters: Decision Trees

- Sample Variable Importance report from CART

Variable	Score
PolSizeA	100.00
PolSizeB	66.34
CvgA	22.67
FinA	15.90
.	
FinC	5.67
random3	0.03

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Feature Selection: Funnel Approach



Copyright © 2008 Deloitte Development LLC. All rights reserved.

Data Visualization

Visualization is important

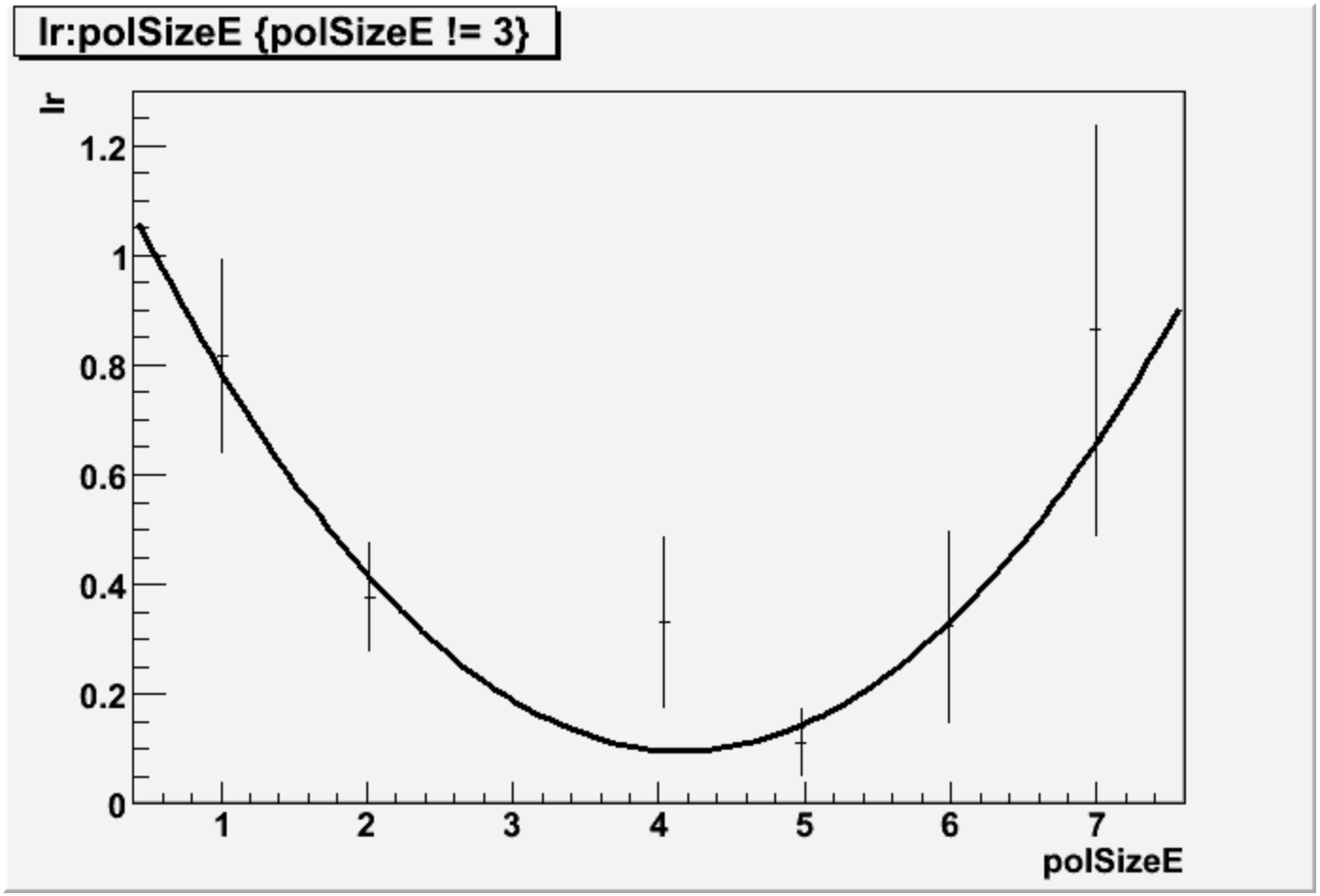
- to take care of non-linear relationship with target
    - Example: add Log(X), X^2 or other polynomial terms
- to take care of extreme values
- to take care of missing values
- to create indicator variables
- to take care of correlation with other variables
- to identify interaction terms

Useful Tools

- Profile Plots, Scatter Plots
- Analysis in MARS
- Correlation & Principal Component Analysis

Copyright © 2008 Deloitte Development LLC. All rights reserved.

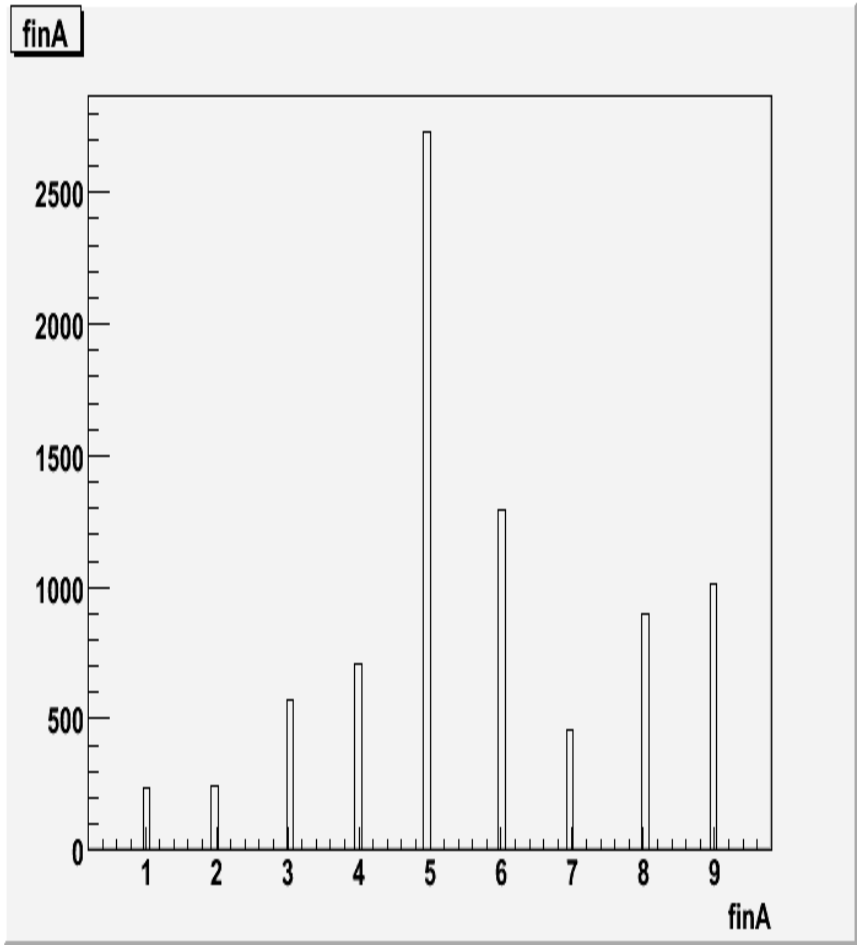
Data Visualization: Non linear relationships

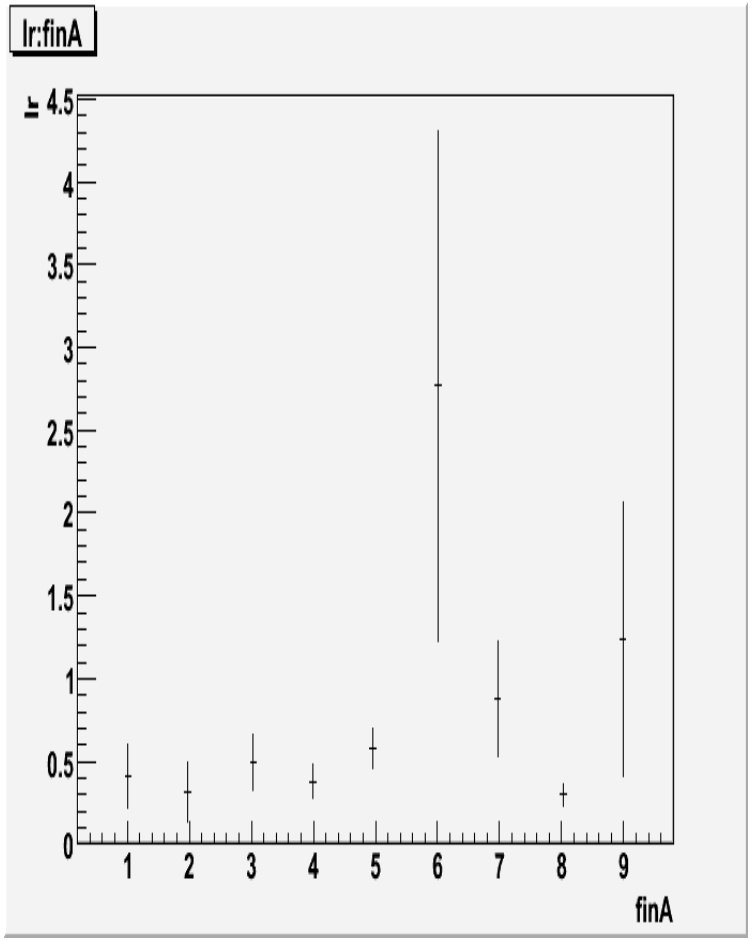


Consider adding a squared term for variable polSizeE

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Data Visualization: Indicator Variables

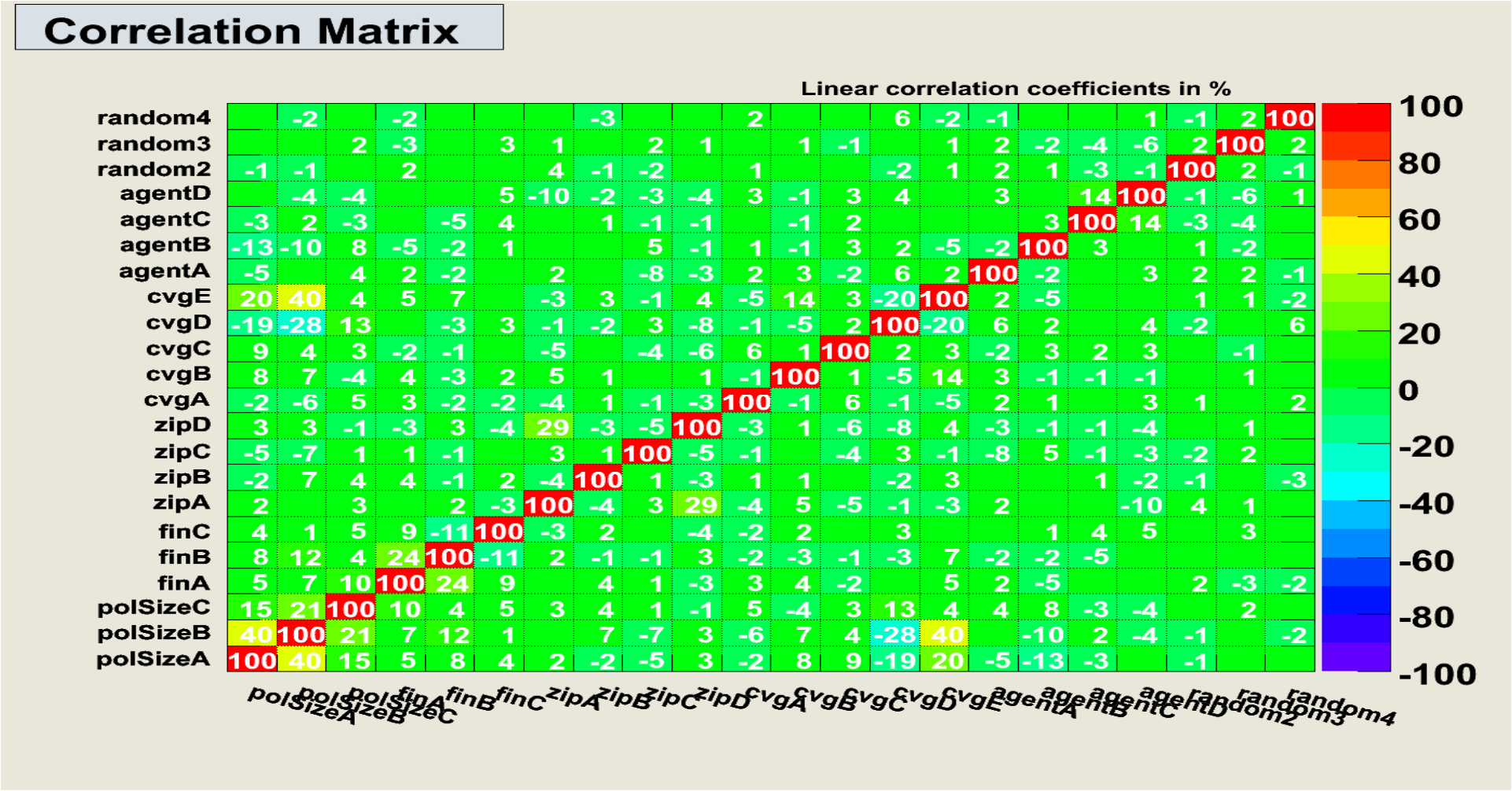




Consider creating indicator variables for finA=5 and for finA=6

Copyright © 2008 Deloitte Development LLC. All rights reserved.

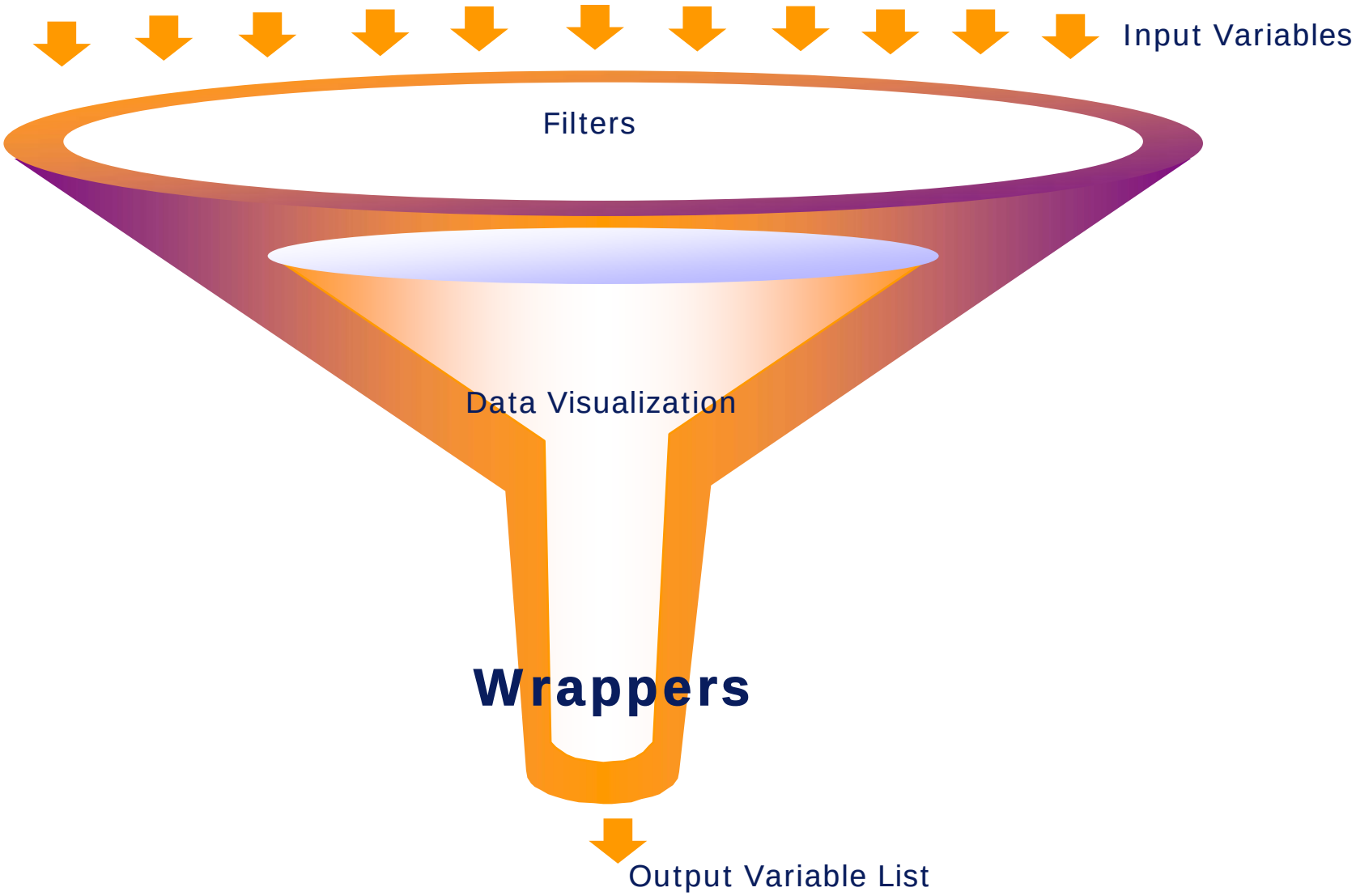
Data Visualization: Correlations



Consider constructing Principal Components for highly correlated variables

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Feature Selection: Funnel Approach



Copyright © 2008 Deloitte Development LLC. All rights reserved.

Wrappers

- Evaluate subset of variables based on predictive power
  - Focus is on Variable Selection
- Independent of the statistical techniques used in modeling
  - Try Multiple Learning Techniques
- Can also be embedded into the modeling process
- Can be Computer intensive
  - Need to start with manageable number of variables

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Wrappers: Machine Learning Techniques

Linear Models with Cross Validation

- Data is randomly divided in to K groups
- Score one group based on model fitted from other K-1 groups
- Repeat this K times, once for each group

- Variables are chosen based on performance of model on test

Neural Networks

- Non-Linear statistical modeling tool
- A good tool to understand variable importance
- Built in Train-Test Concept
- Variable importance is one of the outputs from the model

Copyright © 2008 Deloitte Development LLC. All rights reserved.

Wrappers: Machine Learning Techniques . . .

Boosted Decision Trees

- Many trees based on different error weighting schemes
  - harder to classify points are given a boost
- Majority vote over a number of decision trees
- Produces very stable results
- Available in CART and ROOT packages

RuleFit or (M) Rule Based Ensembles

- Combines Regression and Classification models

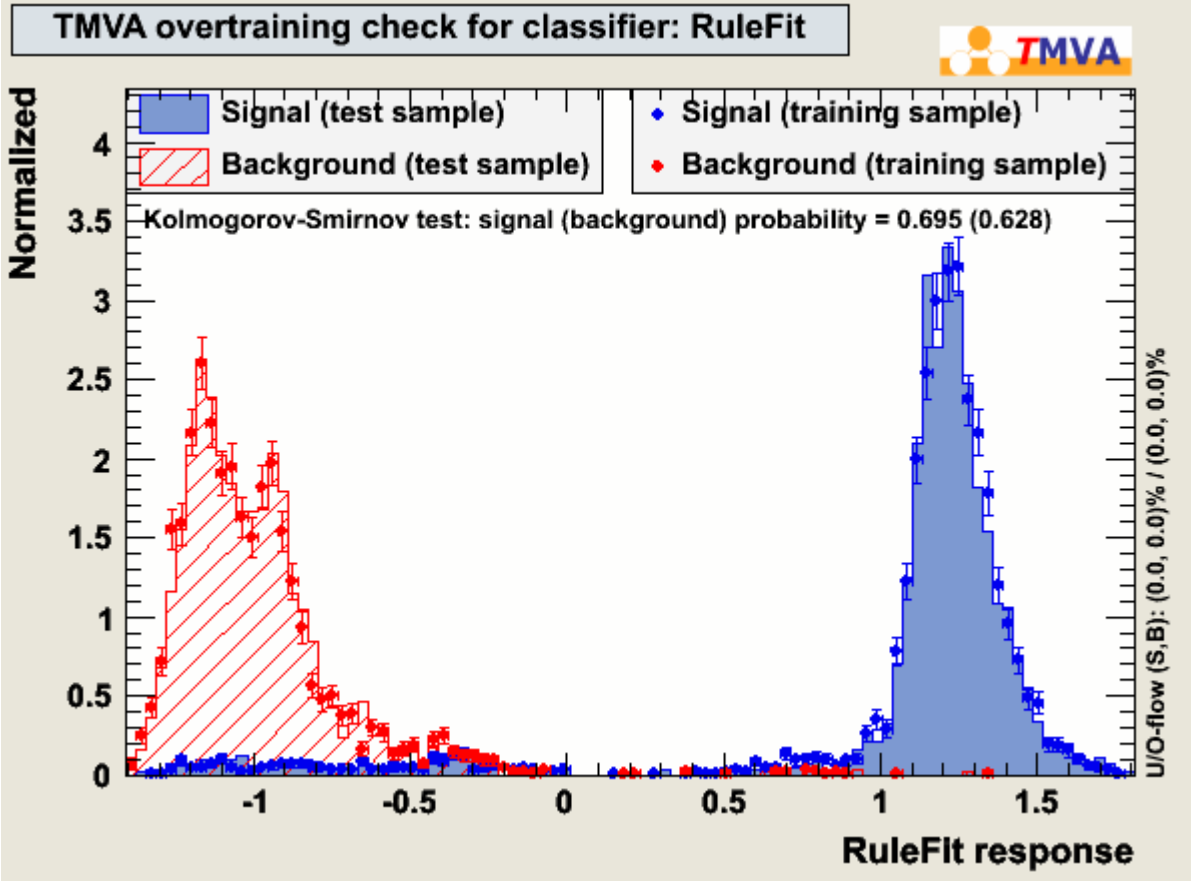
$$Y = a_0 + a_1 f_1(X) + a_2 f_2(X) + \ldots + a_M f_M(X)$$

- Easy to explain
- Available in R and ROOT

Copyright © 2008 Deloitte Development LLC. All rights reserved.

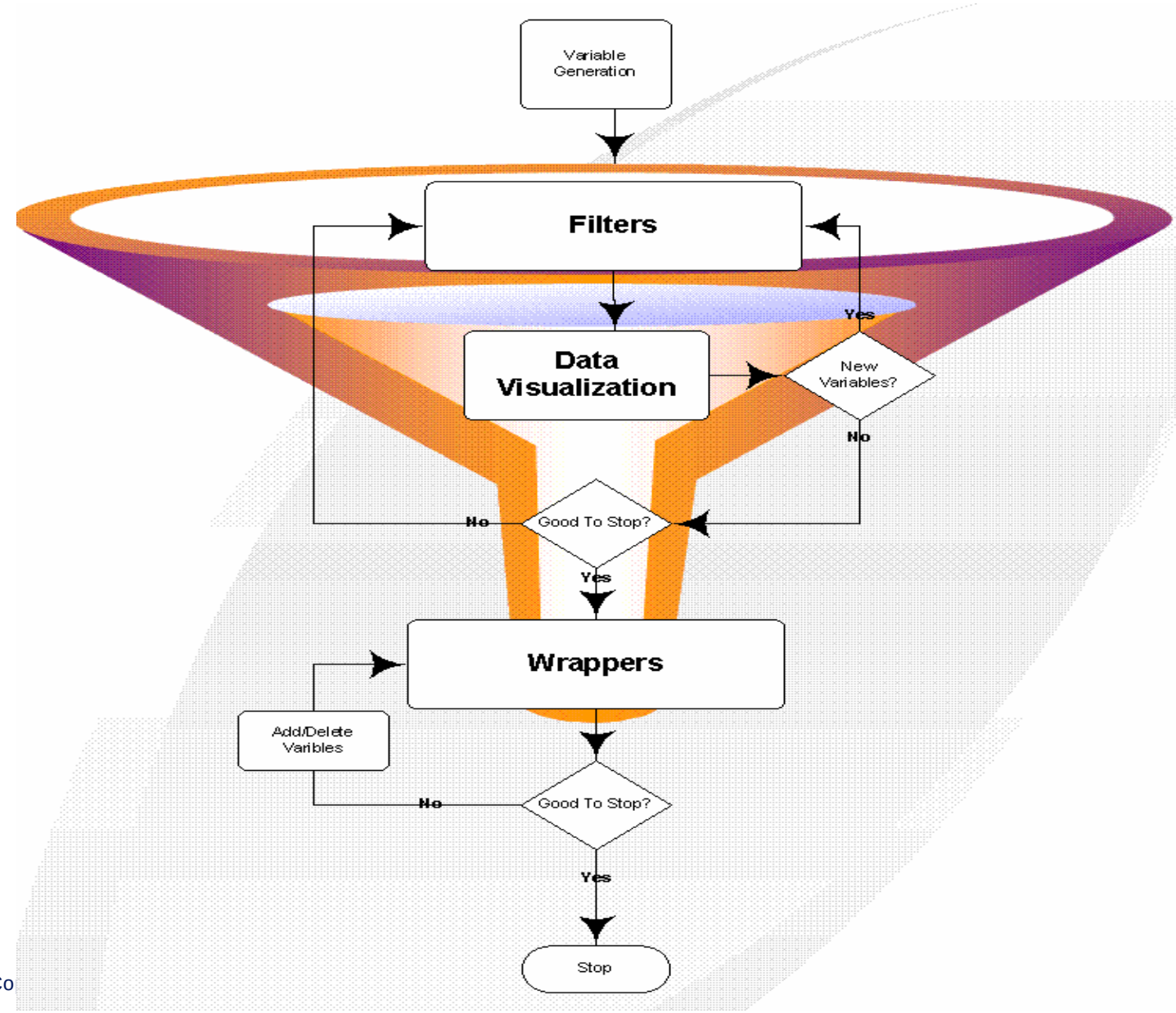
Wrappers: Variable Selection Criteria

1. Performance
2. Consistency between Train and Test



Copyright © 2008 Deloitte Development LLC. All rights reserved.

Putting It All Together



Feature Selection: Conclusion

- There is no perfect algorithm for Feature Selection problem
- **Keep it Simple – Principle of Parsimony**
- **Visualizing the data is very important**
- **Embed Validation into your methodology**
- Work with subsets of data for additional insights

Copyright © 2008 Deloitte Development LLC. All rights reserved.

References

- Documentation on ROOT
  - www.root.cern.ch
- Documentation on R
  - www.r-project.org
- Documentation on SAS
  - sas.com

Copyright © 2008 Deloitte Development LLC. All rights reserved.