Feature Selection Methods

Data mining to pick predictive variables

In Focus: Cutting Edge Tools For Pricing and Underwriting Seminar
Baltimore, MD
October, 2011

Ravi Kumar ACAS, MAAA
Mark Richards, Director, ISO Analytics

Milliman

ISO

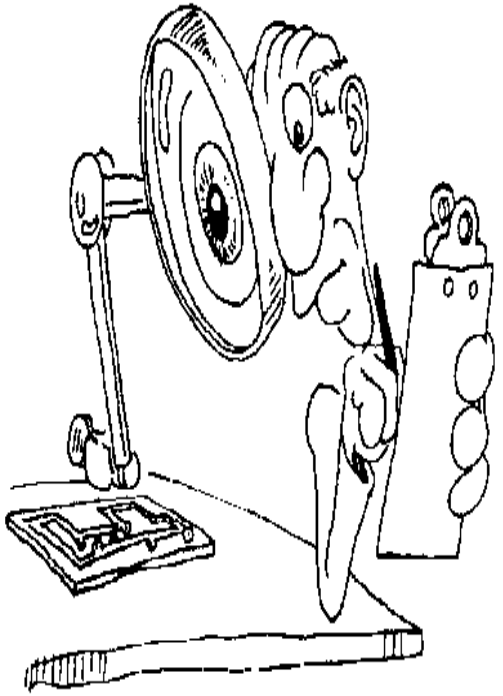
Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

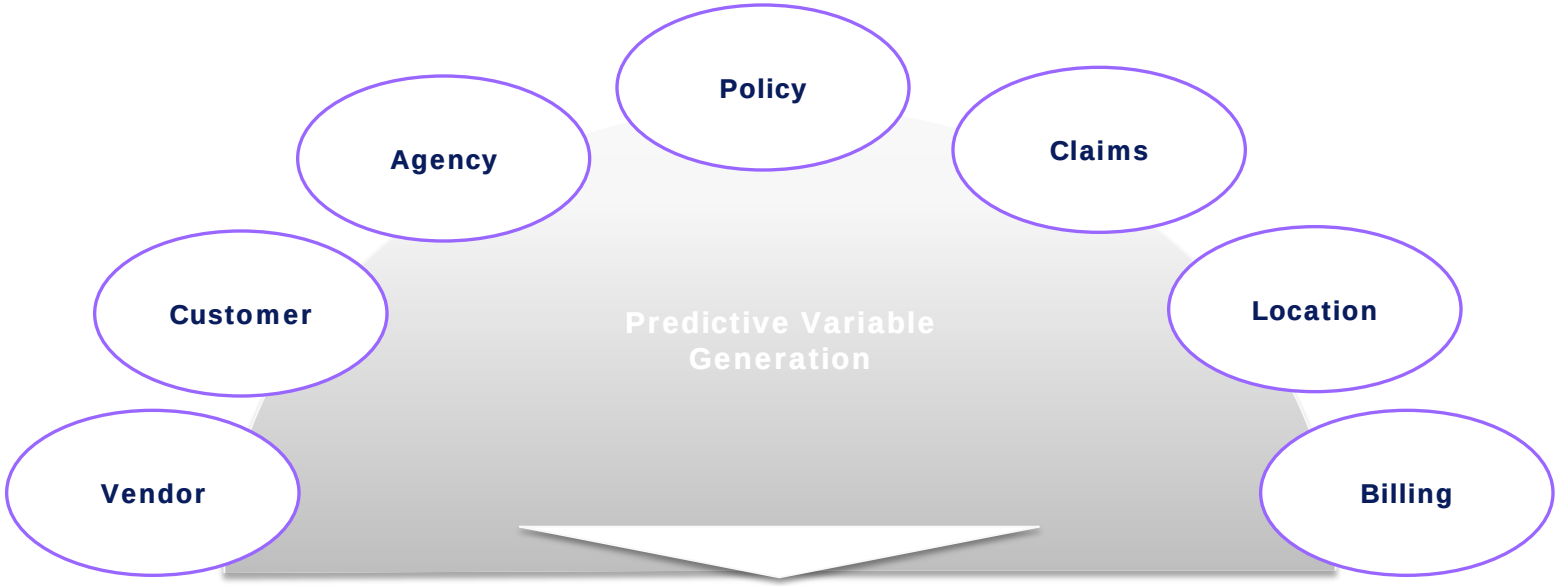
Topics

- Overview
- General Approach
  - Filters
  - Data visualization
  - Wrappers
- Conclusion

Overview



Predictive Modeling



100s of Predictive Variables

Feature Selection

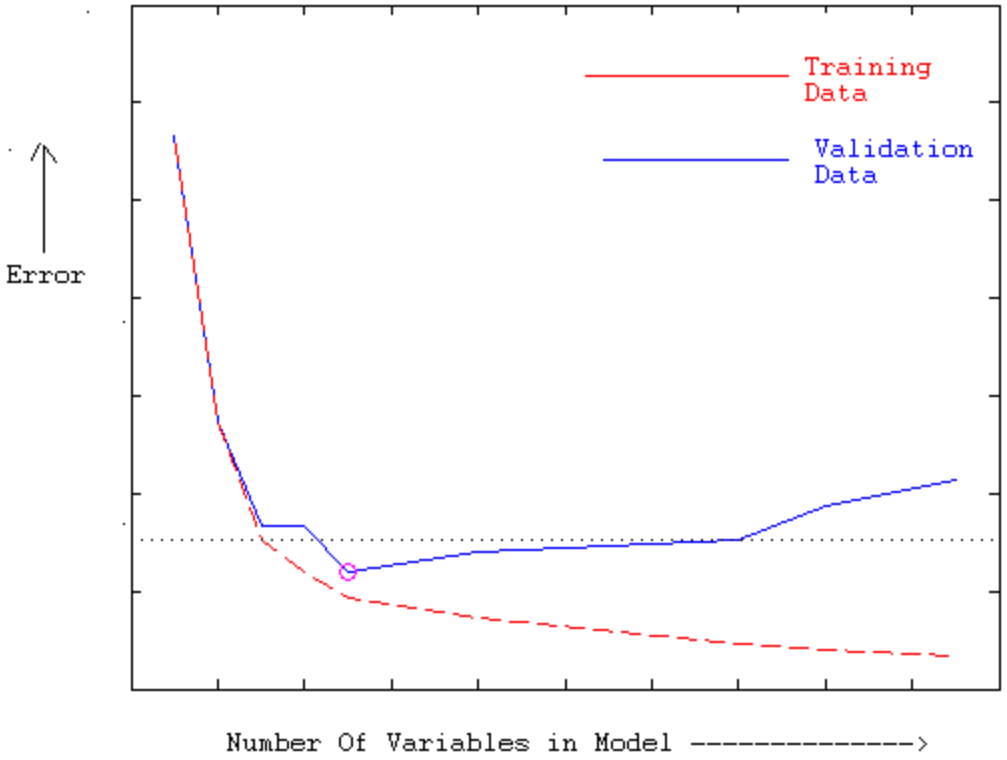
Drivers of Profitablity, Expenses,Severity, Litigation Rates, Fraud etc.

Some Definitions

- Target Variable Y
  - What we are trying to predict.
    - Profitability (loss ratio, LTV), Retention, …
- Features or Predictive Variables $\{X_1, X_2, \ldots, X_N\}$
  - "Covariates" used to make predictions.
    - Policy Age, Credit, #vehicles….
- Placebo Variables
  - Random variables used to validate variable selection methodology
- Predictive Model $Y = f(X_1, X_2, \ldots, X_N)$

Reason for Feature Selection: Curse of Dimensionality

- Using too many features reduces predictive performance



Feature Selection : Things to Ponder

- A Highly Predictive Variable
  - May not translate into a useful variable in a multivariate model
- A seemingly useless variable
  - Can become very useful when used with other variables
- Two highly correlated variables
  - May bring complementary information to a model
- Thus, an optimal model cannot be guaranteed just by looking at variables one at a time, two at a time , or even a few at a time.

Feature Selection: Not a trivial task

- Feature Selection problem is actually a model selection problem
- Feature Selection is a NP-hard problem
  - Cannot be solved in polynomial time $O(n^c)$
  - 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
- A definitive solution is lacking
- Need to have a Validation strategy

Validation strategy using Placebo Variables

Original List of Predictive Variables

AGT01 AGT02 AGT03 AGT04 AGT05 AGT06 AGT07 AGT08 AGT09 PFM01
PFM02 PFM03 PFM04 PFM05 PFM06 PFM07 RSK01 RSK01R RSK02 RSK03 RSK04
RSK05 RSK06 RSK07 RSK08 RSK09 RSK10 RSK11 RSK12 RSK13 RSK14 RSK15
RSK16 RSK17 RSK18 RSK19 RSK20 RSK21 RSK22 RSK23 RSK24 RSK25 RSK26
RSK27 RSK28 RSK29 RSK30 RSK31 RSK32 RSK33 RSK34 RSK35 RSK36 RSK37
RSK38 RSK39 RSK40 RSK41 RSK42 Zip01 Zip02 Zip03 Zip04 Zip05 Zip06 Zip07
Zip08 Zip09 Zip10 Zip11 Zip12 Zip13 Zip14 Zip15 Zip16 Zip17 Zip18 Zip19

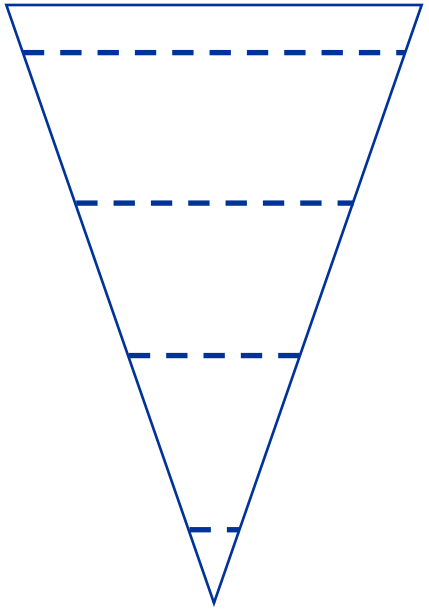
With Additional Placebo Variables

AGT01 AGT01R AGT02 AGT02R AGT03 AGT03R AGT04 AGT04R AGT05 AGT05R
AGT06 AGT06R AGT07 AGT07R AGT08 AGT08R AGT09 AGT09R PFM01 PFM01R
PFM02 PFM02R PFM03 PFM04 PFM04R PFM05 PFM05R PFM06 PFM06R PFM07
PFM07R Ran01R Ran02R Ran03R Ran04R Ran05R RSK01 RSK01R RSK02 RSK02R
RSK03 RSK03R RSK04 RSK04R RSK05 RSK05R RSK06 RSK06R RSK07 RSK07R
RSK08 RSK08R RSK09 RSK09R RSK10 RSK10R RSK11 RSK11R RSK12 RSK12R
RSK13 RSK13R RSK14 RSK14R RSK15 RSK15R RSK16 RSK16R RSK17 RSK17R
RSK18 RSK18R RSK19 RSK19R RSK20 RSK20R RSK21 RSK21R RSK22 RSK22R
RSK23 RSK23R RSK24 RSK24R RSK25 RSK25R RSK26 RSK26R RSK27 RSK27R
RSK28 RSK28R RSK29 RSK29R RSK30 RSK30 RSK30R RSK30R RSK31 RSK31R
RSK32 RSK32R RSK33 RSK33R RSK34 RSK34R RSK35 RSK35R RSK36 RSK37
RSK37R RSK38 RSK38R RSK39 RSK39R RSK39RR RSK40 RSK40R RSK41 RSK41R
RSK42 RSK42R Zip01 Zip01R Zip02 Zip02R Zip03 Zip03R Zip04 Zip04R Zip05
Zip05R Zip06 Zip06R Zip07 Zip07R Zip08 Zip08R Zip09 Zip09R Zip10 Zip10R
Zip11 Zip11R Zip12 Zip12R Zip13 Zip13R Zip14 Zip14R Zip15 Zip15R Zip16
Zip16R Zip17 Zip17R Zip18 Zip18R Zip19 Zip19R

A placebo variable is a random variable that has the same distribution as another real variable

A good feature selection methodology should NOT pick the placebo variables

General Approach

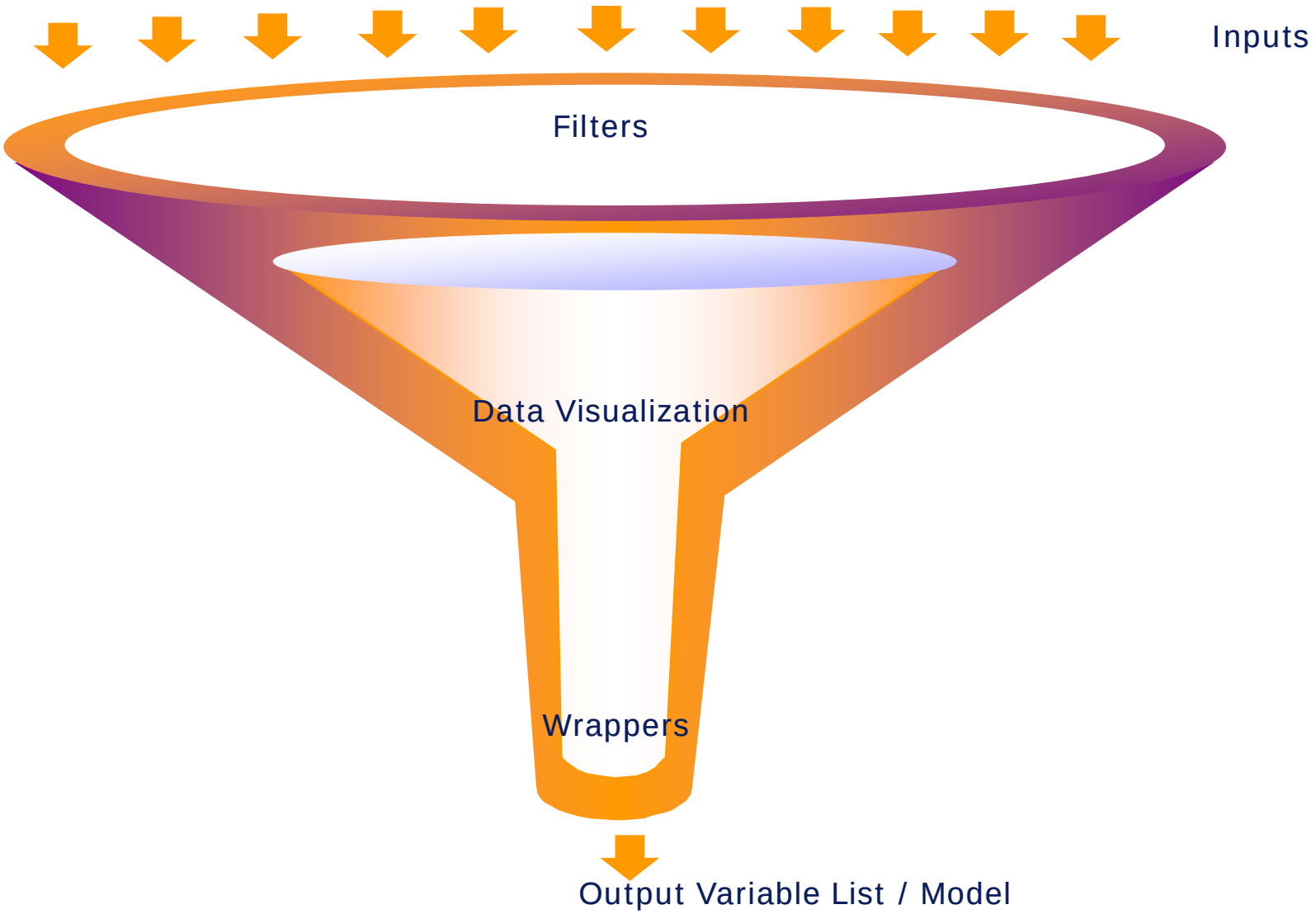


Using the Funnel approach for Feature Selection

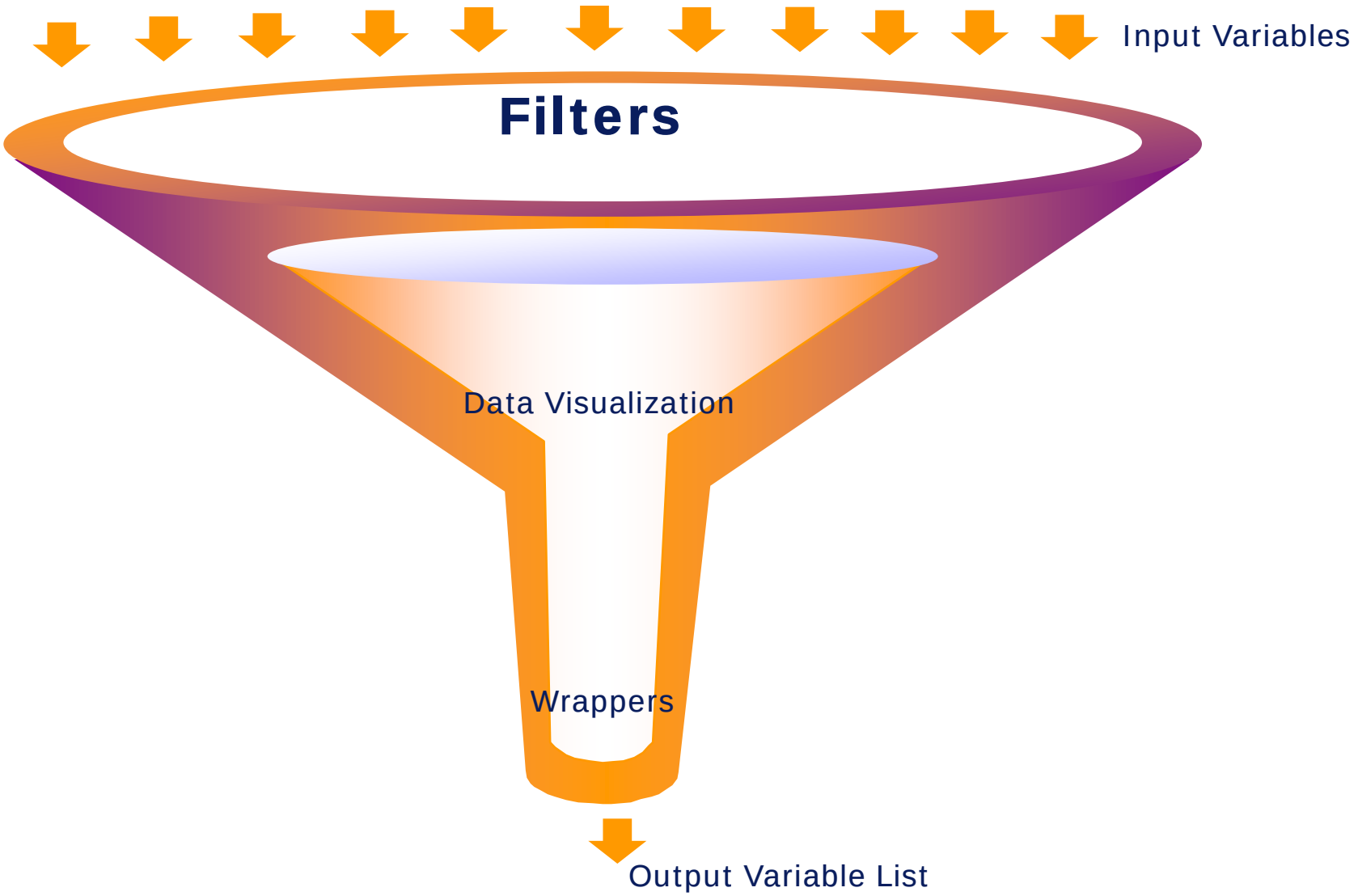
Milliman

ISO

Feature Selection: General Approach



Feature Selection: General Approach



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**

Filters: Variable Selection Criteria

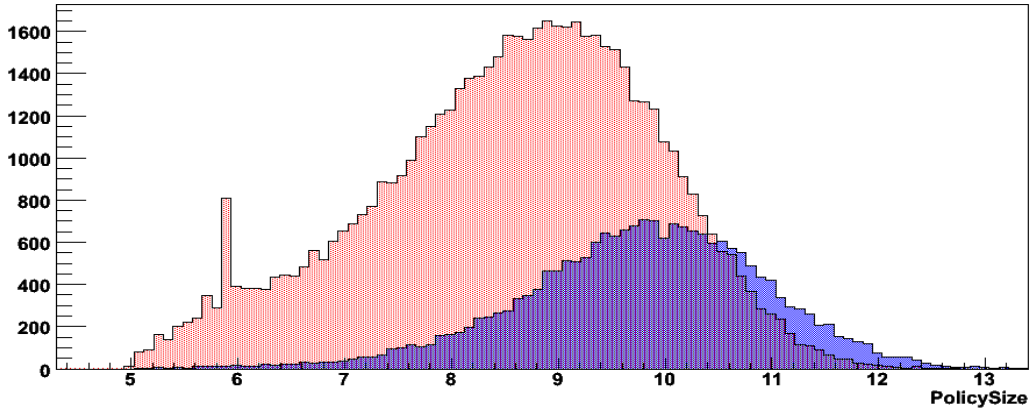
- A priori Business/Reliability knowledge
- Variable performance in Univariate analysis
  - K-S Tests
- Variable performance in simple, fast performing models
  - Stepwise Regression
  - Decision Trees
- Selection methods validated by the use of placebo variables

Filters: Variable performance in Univariate analysis

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

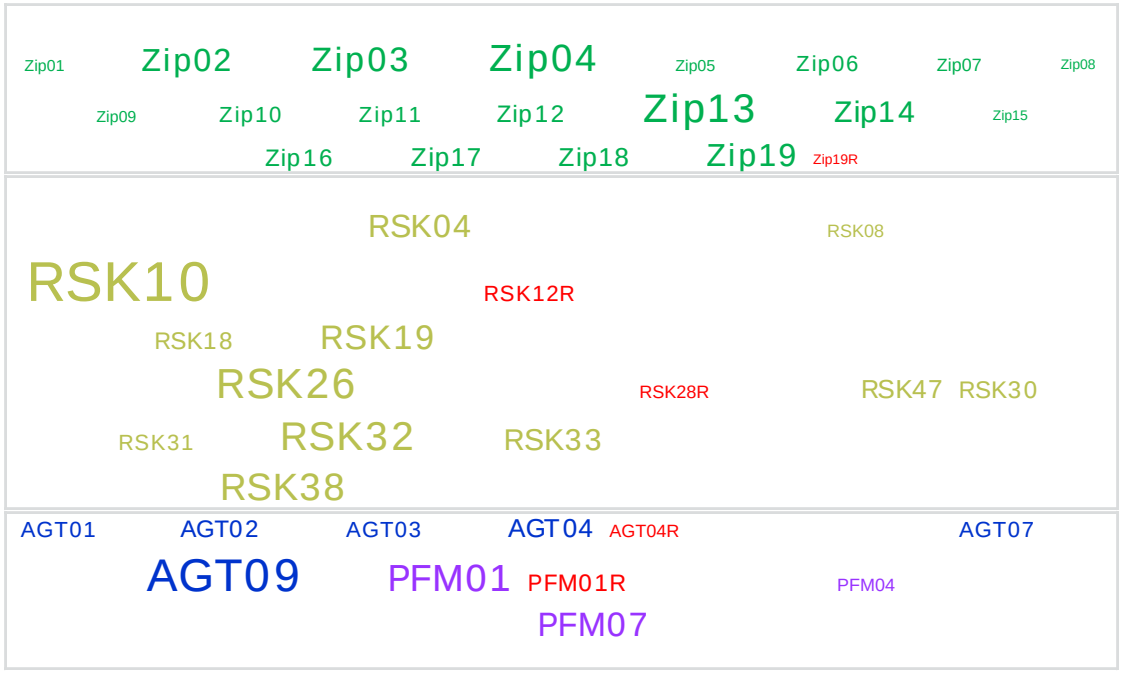
Divide data into two samples based on a Binary Target
(Example: NoClaim policies vs. Others)
Compare the distribution of Xs in these two samples
Rank the Xs based on K-S test

Focus on features with highest ranks



Filters: Variable Performance in Univariate Analysis

- Sample Rank of variables that influence Agent Performance



- Placebo variables are used to validate the method

Simple Models: Stepwise Regression

Pros

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

Cons

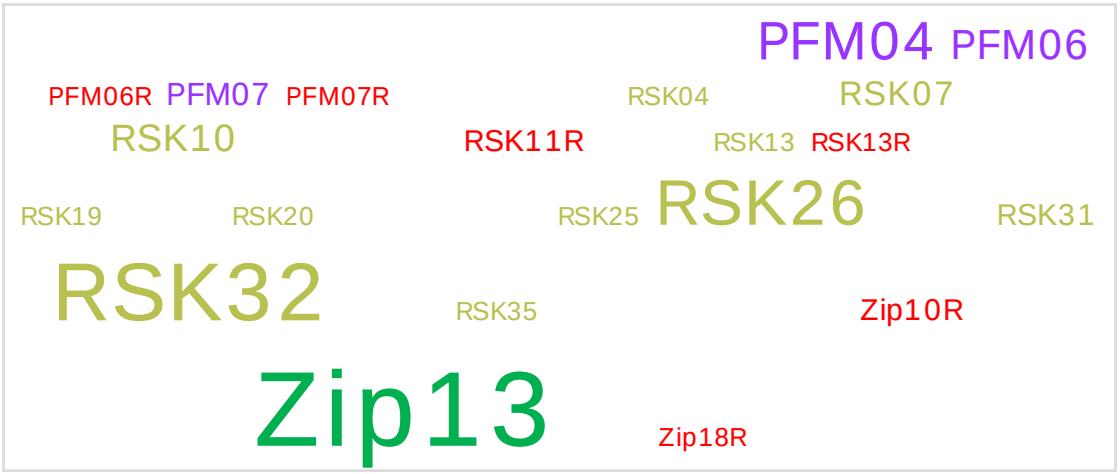
- Variables are picked based on Training data only
- No penalty for picking too many variables

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

Filters: Stepwise Regression

- Sample Rank of variables that influence Agent Performance



- Placebo variables are used to validate the method

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

Filters: Decision Trees

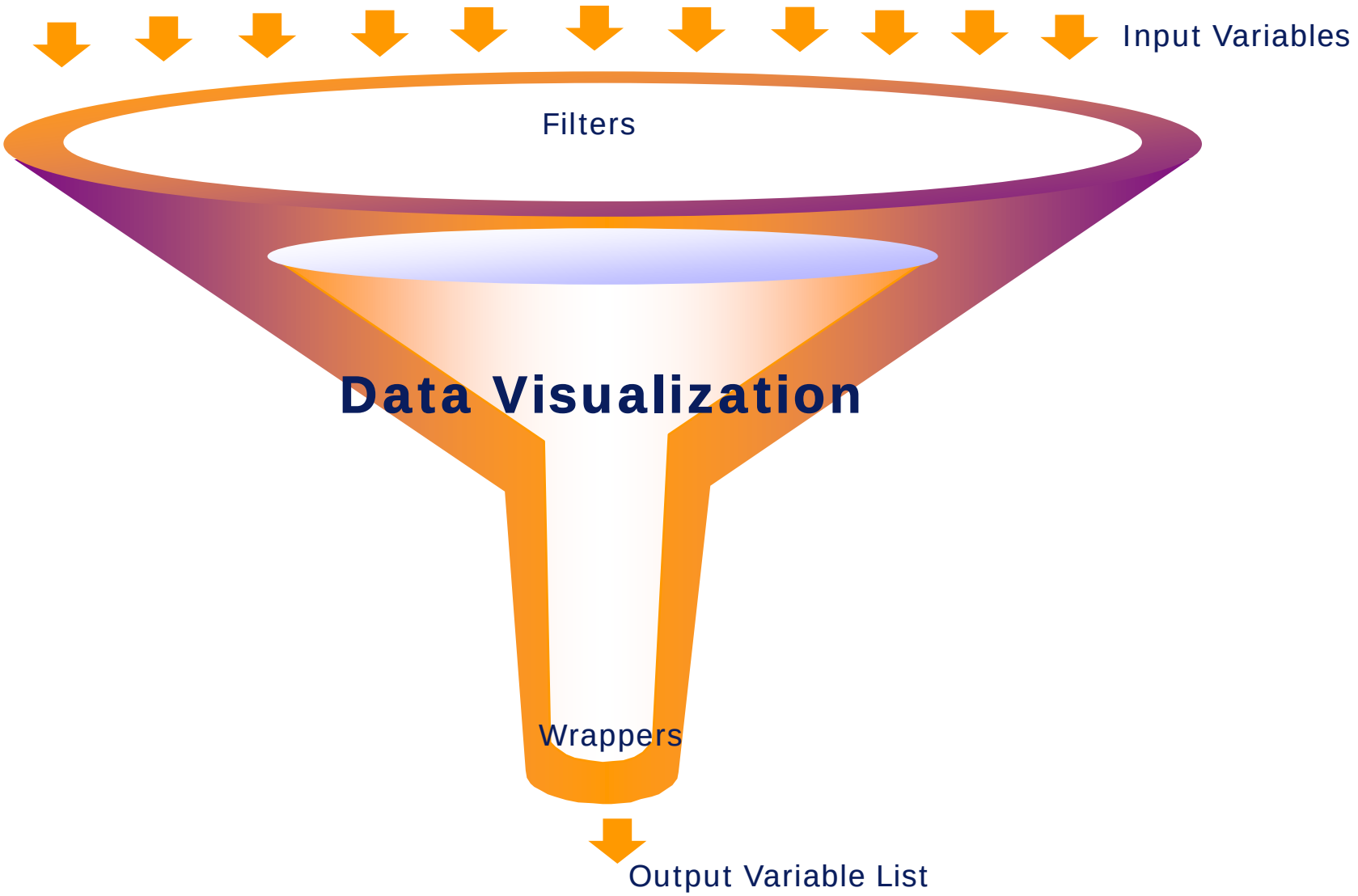
- Sample Rank of variables that influence Agent Performance

- Pruning the tree based on performance on test data reduced the chance of placebo variables being selected
  - Good to try Regularization methods

Filters: How to get most out of filters?

- Mix in some random number based Placebo Variables
  - For validating variable selection methodology
- Use many different Ranking techniques
  - K-S Statistics, Linear Models, Decision Trees, etc.
  - Different techniques have different strengths & weaknesses
- 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/Renew, Policy Size, Restaurant Class, etc.
  - Data Sampling?

Feature Selection: General Approach



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

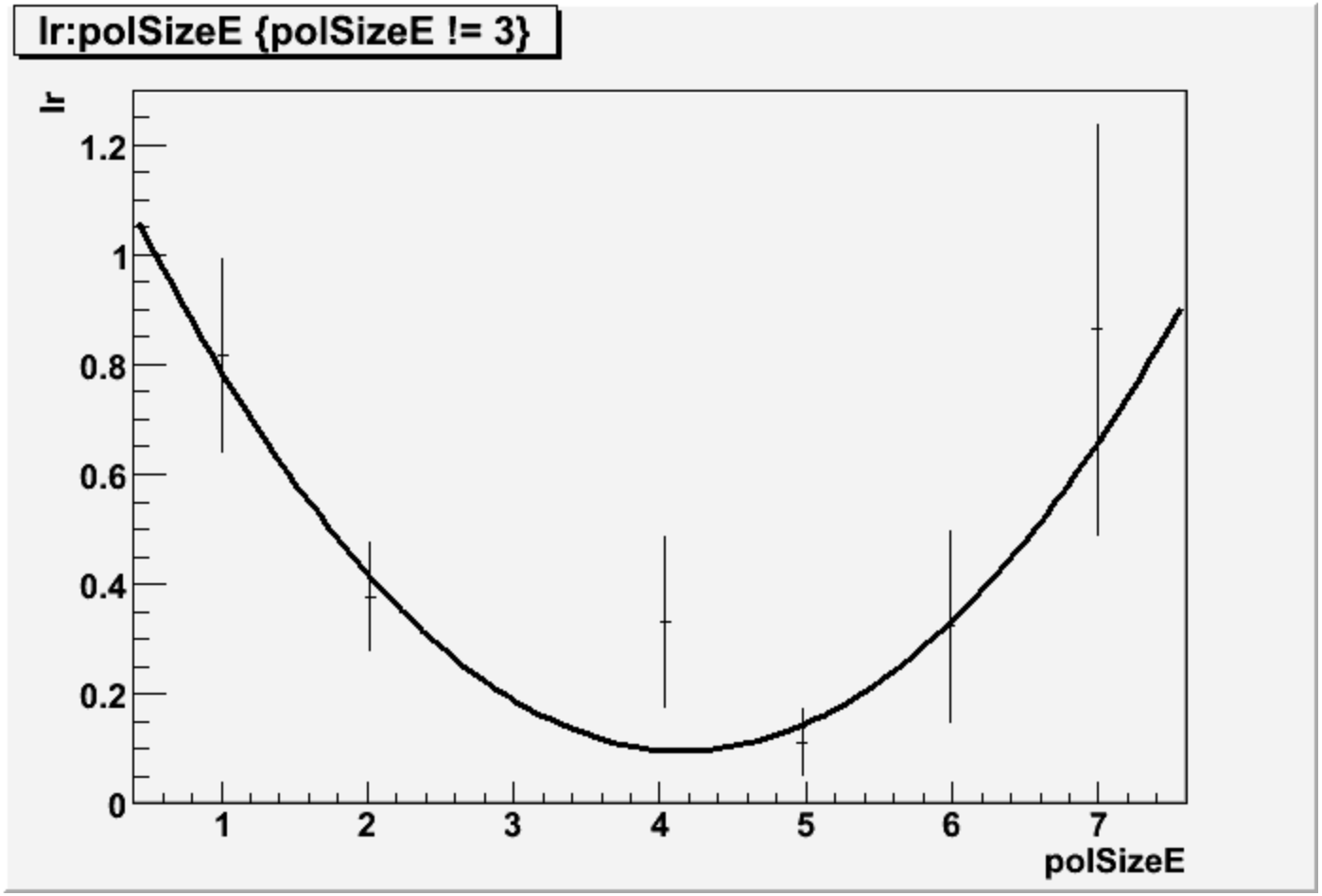
- to take care of correlation with other variables

- to identify interaction terms

Useful Tools

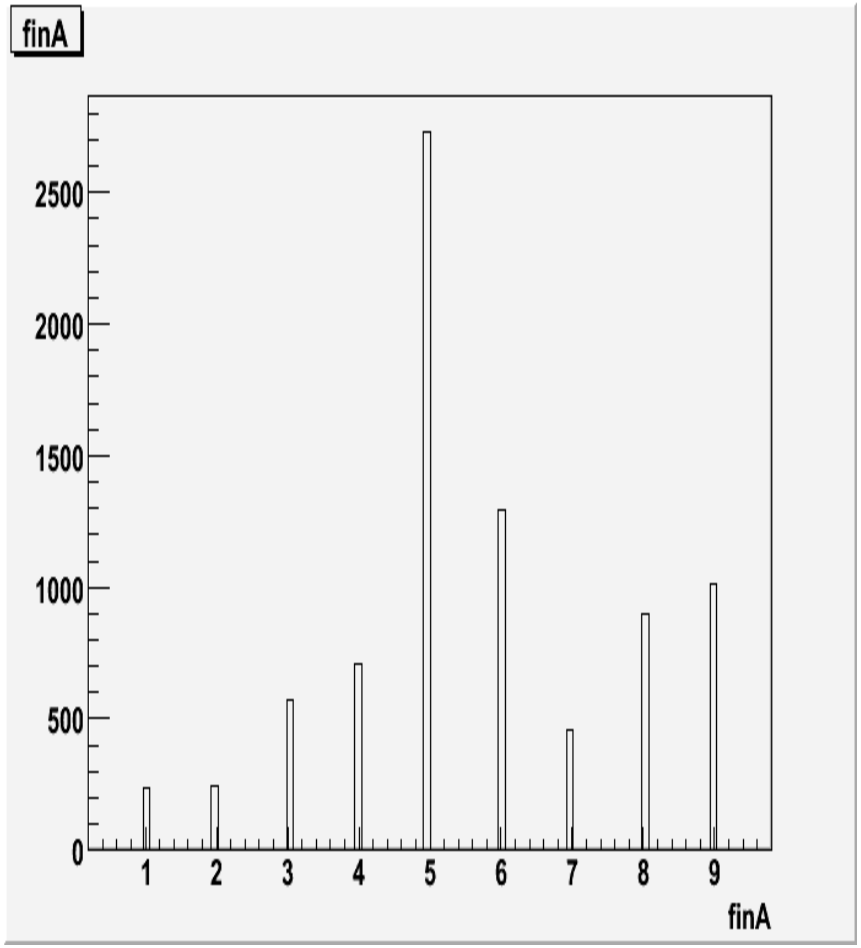
- Profile Plots, Scatter Plots
- Analysis in MARS
- Correlation & Principal Component Analysis

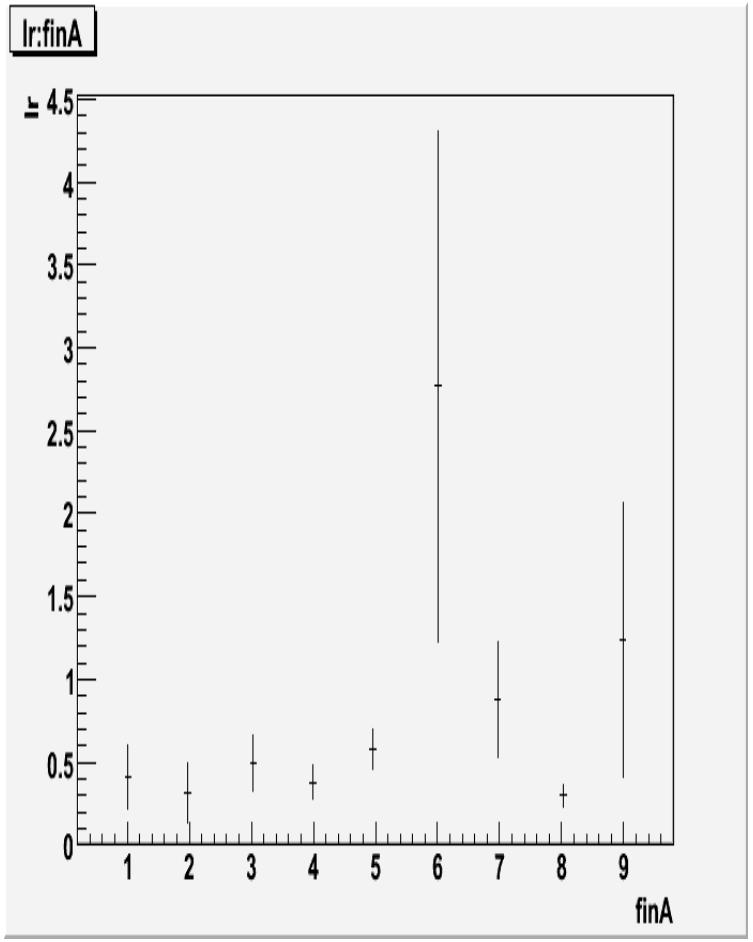
Data Visualization: Non linear relationships



Consider adding a squared term for variable polSizeE

Data Visualization: Indicator Variables

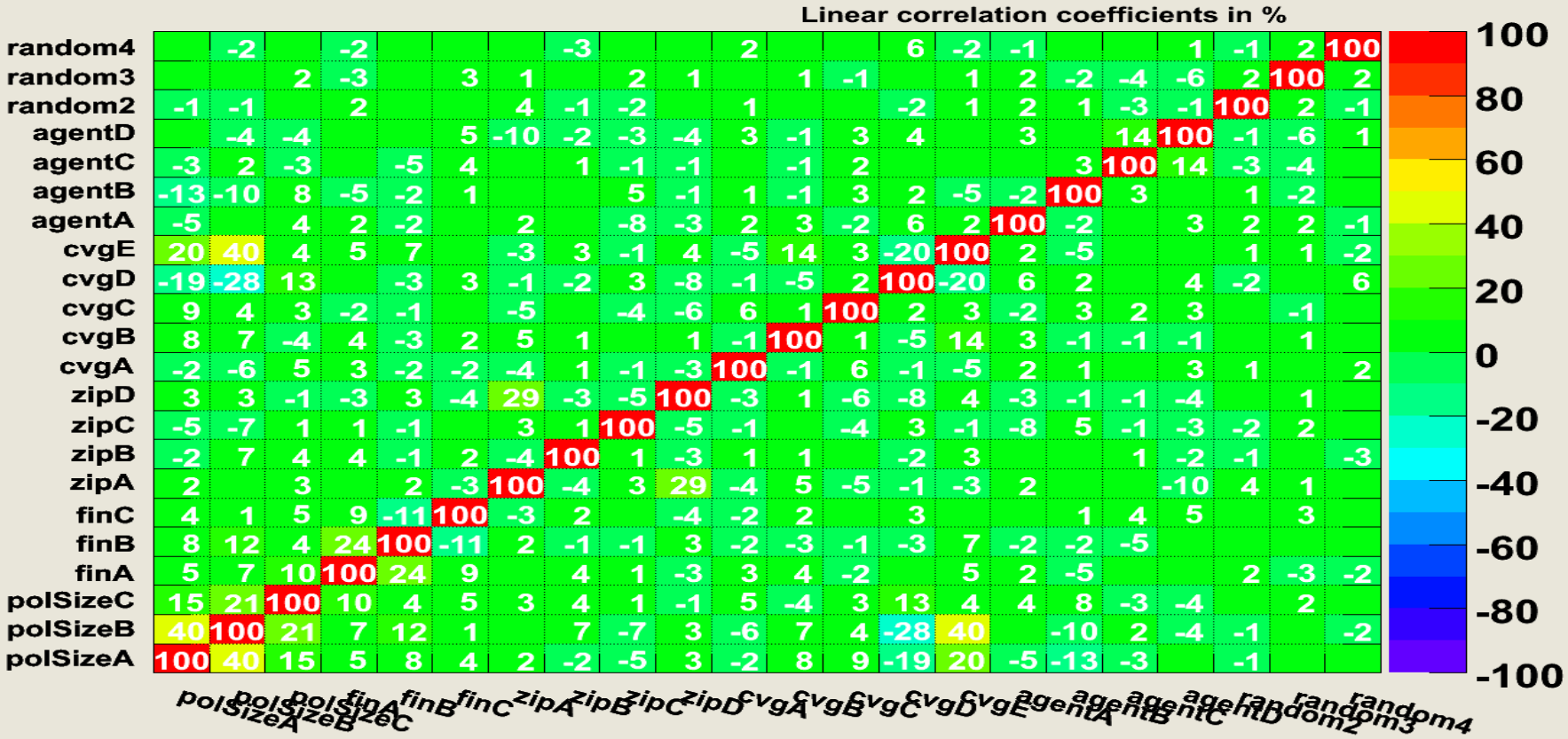




Consider finA=5 as a reference variable and creating indicator variables for finA=6 and for finA=8

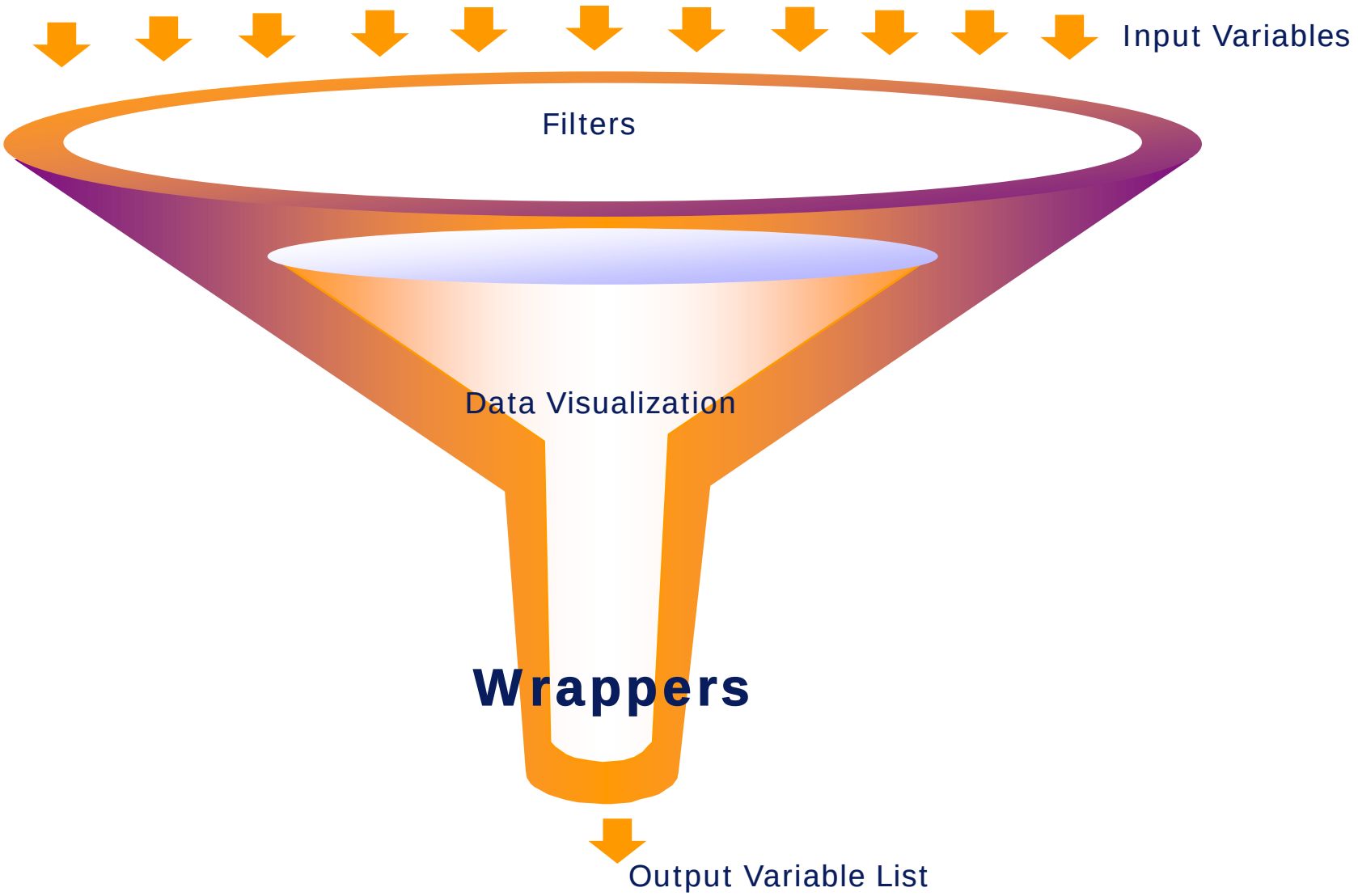
Data Visualization: Correlations

Correlation Matrix



Consider constructing Principal Components for highly correlated variables

Feature Selection: General Approach



Methods Discussed

- Stepwise Selection
- Computationally Intensive Methods
  - Tabu Algorithms
  - Simulated Annealing
  - Genetic Algorithms
- Regularization Methods
  - Ridge Regression
  - The Lasso
  - Elastic Net
- Adjunct Methods
  - Trees

28

History (Ancient?)

1. “The problem of determining the "best" subset of variables *has long been of interest* to applied statisticians and, primarily because of *the current availability of high-speed computations*, this problem has received considerable attention in the recent statistical literature.”
   - R. R. Hocking (1976). “*The Analysis and Selection of Variables in Linear Regression*.” *Biometrics*, Vol. 32, No. 1, pp. 1-49.

2. “Selection of regressors is *an old and important problem* in econometrics…”
   - Takeshi Amemiya (1980). “Selection of Regressors.” *International Econometric Review*, Vol. 21, No. 2, pp. 331-354.

Wrappers: Stepwise Selection

- A combination of Forward and Backward selection.
  - Alternates between forward selection and backward elimination steps.
- Used when exhaustive search not feasible.
  - In combinatorial optimization terms, basically a “hill climber”.
- Many criticisms (some strident).
  - Can be “trapped” in local (not global) optima.
  - Selection Bias (inflates |β|), multiple comparisons / over-fitting (p-values), F-statistic Distribution (sequential testing), exaggerates multi-colinearity problem, etc.
- Still a useful tool (when caution is exercised)
  - Do not apply blindly.
  - Prefer objective functions (AIC and BIC) that penalize complexity over F.
  - When possible, multiple starting points (in SAS, First= option).
  - Watch correlated variables
  - Use *in conjunction with* other approaches.

Wrappers: Computationally Intensive Methods

- Treats variable selection as an optimization problem
  - Examples: Tabu, Simulated Annealing, Genetic algorithms

- Advantages
  - Flexible: Many model forms and objective functions
  - Efficient: Less likely to get trapped in local optima
  - Can have rules to penalize complexity

- Disadvantages
  - Computationally intensive
  - Can over-fit the training data
  - Sensitive to selection of key tuning values
    - Example Population Size or Cooling Schedule

31

Tabu Algorithms

- Heuristics, rules applied that “make sense” but lack theoretical bounds

- Selection methods
  - Steepest ascent
    - a hill climber similar to stepwise
  - Mildest descent
    - if no allowable change results in an improvement of the objective function, then select the “least bad” move in the other direction (in order to avoid being trapped in a local optima)
  - Diversification and Intensification
    - Keeps a history of solutions, encourages large changes early (many added/dropped attributes) but then smaller moves when in neighborhood of “best” solutions
  - Aspiration
    - If a variable on the Tabu List results in “best” solution compared to history of solutions, accept it

- Recent attributes added to (expiring) Tabu List: to avoid cycling back to previous solution when mildest descent is accepted

Simulated Annealing

- Simulated Annealing
  - Physics analogy, but Markovian interpretation.
  - Important/relevant concepts: *Neighborhoods (allowable moves)* and *Cooling Schedule* (convergence, reduction of temperature "τ").

- ***j*** stages {0, 1, 2, … J}, with a fixed temperature per stage of $\boldsymbol{\tau_j}$
- *length of each stage is:* $\boldsymbol{m_j}$ (large and increasing in *j*)

For m = 1 to m_j

Select θ^* in neighborhood of $\theta^{(t)}$

Let $\theta^{(t+1)} = \theta^*$ with Probability $\min(1, \exp\{[f(\theta^{(t)}) - f(\theta^*)]/\boldsymbol{\tau_j}\}$

increment *j*

Note: this is a *minimization problem!*

Add or drop a variable from the model

If it improves model, accept the change

If new model is worse, accept it *sometimes*

As temperature goes down, less likely to accept new model that is worse

Typically, total # of iterations in the thousands!

- If 1 minute per model, it takes ~17 hours per 1,000 iterations to run.

Genetic Algorithms

- Genetic Algorithms: the basics (the birds and bees)

Phenotype

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_3}x_3 + \widehat{\beta_4}x_4$$

Individual (organism)

Genotype (Chromosome)

1 0 0 1 1 0

gene, position is its *locus*

***allele* is the gene's potential values {0,1}**

- Genetic Operator: Crossover

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_3}x_3 + \widehat{\beta_4}x_4$$

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_5}x_5$$

1 0 0 1 1 0

1 0 0 0 0 1

Randomly select *crossover point(s)* between two adjacent loci

- Offspring

1 0 0 1 0 1

1 0 0 0 1 0

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_3}x_3 + \widehat{\beta_5}x_5$$

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_4}x_4$$

- Genetic Operator: Mutation

1 0 0 0 1 0 $\xRightarrow{p}$ 1 0 1 0 1 0

$p << 1$ (e.g. 0.01)

Genetic Algorithms (cont.)

- Populations: collections of individuals.
  - Size P (e.g. $C \leq P \leq 2C$, where C is chromosome length).
  - The "fittest" P offspring from the previous generation(s).
- Fitness and the objective function *f (e.g. –AIC)*.
  - *Slow convergence* vs. (fast) *convergence to local optima*.
  - A function of *f*, not f itself (e.g. rank).
- Updating Generations
  - The *generation gap* (G) is the proportion of generation *t* to be replaced in generation *t+1.* $1/P \leq G \leq 1$
- Selection Mechanisms
  - Simple: select two unique individuals with probability determined by fitness. Repeat GP or GP/2 times. (Some strategies differ slightly).
  - Tournament: randomly partition population and select fittest individuals from each partition. Repeat (w/ new random partitions).
  - Other strategies (including different types of tournament).

Regularization Methods

- Finds a model that will provide the best fit to the *data it will encounter in the future*, not just to the training data.
  - A model that over-fits to the training data will not perform well on new data (validation, testing or out of sample holdout sets)
  - An over-fit model has *high variance*

- Designed to reduce prediction errors
  - By introducing a small amount of *bias*
  - Shrinks the parameter estimates towards (or in some cases to) zero

- Motivating concept: “The Bias-Variance Tradeoff”

36

Regularization: Penalized Least Squares

- Ridge Regression and The Lasso impose a penalty that biases the least squares solution (slightly) in order to *minimize prediction error*. The penalties only differ slightly:

Ridge (L2 penalty) $\sum_j \beta_j^2 \leq t$ Lasso (L1 penalty) $\sum_j |\beta_j| \leq t$

- The penalty is imposed on the sum of the magnitude of the parameter estimates

- Searches for solutions on the training data that fit well, but where the parameter estimates are shrunk towards zero (penalized).
  - If t is large, the solution is closer to the unbiased OLS solution, as t gets smaller estimates are more biased.
    - Some Ridge parameter estimators become non-significant.
    - Some Lasso parameter estimates shrink to zero.

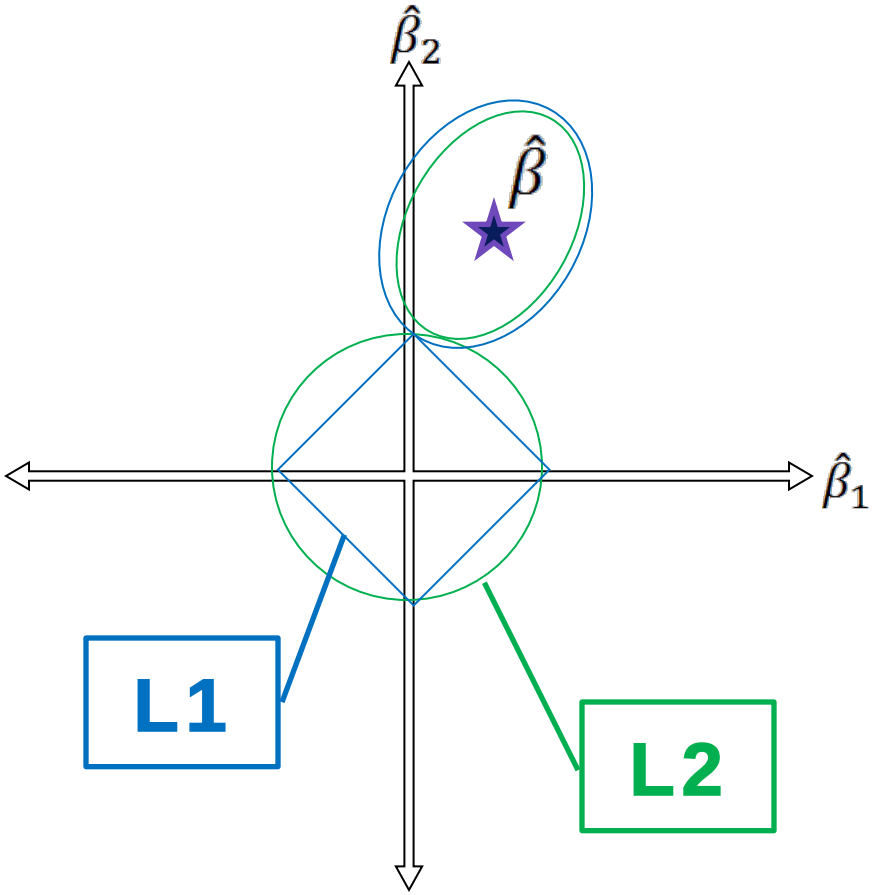
Regularization: Penalized Least Squares (cont.)

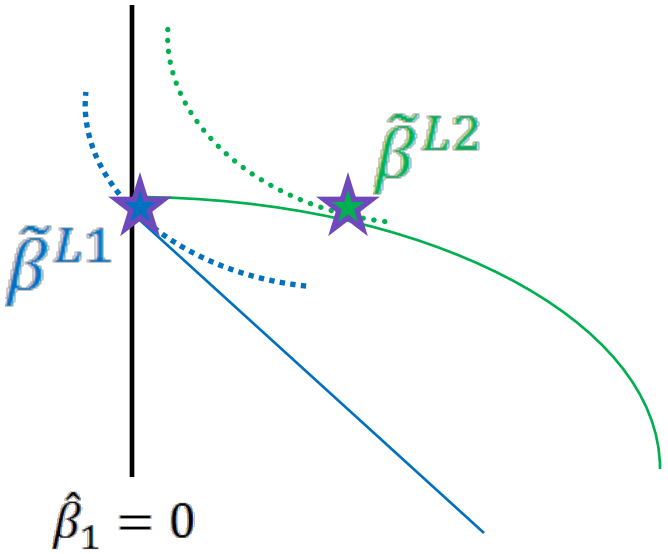
- K-fold cross validation can be used to search for the value of t that gives the *smallest prediction error* across a range of values

- A hybrid approach called the *Elastic Net* is a weighted average of the L1 and L2 penalties
  - Cross Validation can be used to determine weight (as well as penalty) that minimizes prediction error

Geometry of the L1 and L2 Norms in 2-D

How Parameter Estimates are Shrunk





39

Regularization Methods

- Advantages
  - Directly addresses prediction error

- Disadvantages
  - All candidate variables must be normalized (scaled)
  - Limited to Least Squares models (linear regression)
    - The constraint above is particularly restrictive in many typical P&C Insurance model applications.
    - Can sometimes define “good enough” approximate model

Trees for Variable Selection

- Before a candidate model is selected
  - Use Trees as an adjunct method
    - A sanity check on other methods
    - For initial variable selection (Filter)
  - Build deep binary trees
    - Do variables that rank high in the Tree appear in the models you select via your other method(s)?
    - Do variables appear in the model(s) that are not in the Tree? Why?
- After a candidate model is selected
  - Fit Trees with shallow, multi-way splits to your residuals
  - Identify interaction terms and non-linearities

Trees for Variable Selection

- Advantages
  - Fast
  - Flexible
  - Widely available

- Disadvantages
  - Interpretation
    - The Tree model cannot be interpreted in terms of parametric models
    - Provides ranking of variables, not significance measures
  - Over fits to training data
    - Cross validation available

42

Variable Selection issues specific to P&C Insurance

- Model forms are often more complex than are available / feasible in methods discussed here (e.g. Compound Poisson).
  - Often use approximate models (e.g. select frequency and severity separately at first)

- Commonly have existing models (with known variables and parameters), and are looking to select from a distinct set of new variables
  - "Offset" existing models and select variables that explain the residuals

- Over dispersion: often present in P&C Loss data
  - When present (and not accounted for), over dispersion inflates the significance of parameter estimates

- Large Samples: can strain some software (need to sample)

43

Comparison of methods

Method	Efficient Search	Speed	Control over-fitting	Model Specification
Stepwise	some – trapped in local	moderate	complexity penalty (AIC)	limited
Tabu	yes	slow	AIC	flexible
Simulated Annealing	yes	very slow	AIC	flexible
Genetic Algorithm	yes	slow	AIC	flexible
Ridge, Lasso & ENet	yes	fairly fast	regularization	very limited (PLS)
Trees	yes	very fast	cross validation	N/A

44

Wrapper Software

- Stepwise Selection
  - SAS Procs: REG, LOGISTIC, GLMSELECT.
  - "*step*" function in R.
- Computationally Intensive Methods
  - Tabu Algorithm: "*bnlearn*" package in R.
  - Simulated Annealing: "*GenSA*" or "*subselect*" in R.
  - Genetic Algorithms: "*glmulti*" package in R.
- Regularization Methods
  - "*glmnet*" package in R can be used to fit Elastic Net, Ridge Regression or The Lasso.
  - In SAS: Ridge Regression can be performed in *Proc REG*
    - The Lasso is available via *Proc GLMSELECT* (V9.2 or higher)
- Trees
  - SAS Enterprise Miner
  - "*rpart*" package in R.

45

Appendix: K-fold Cross Validation

1. Partition data (randomly) into K mutually exclusive and equal sized samples.

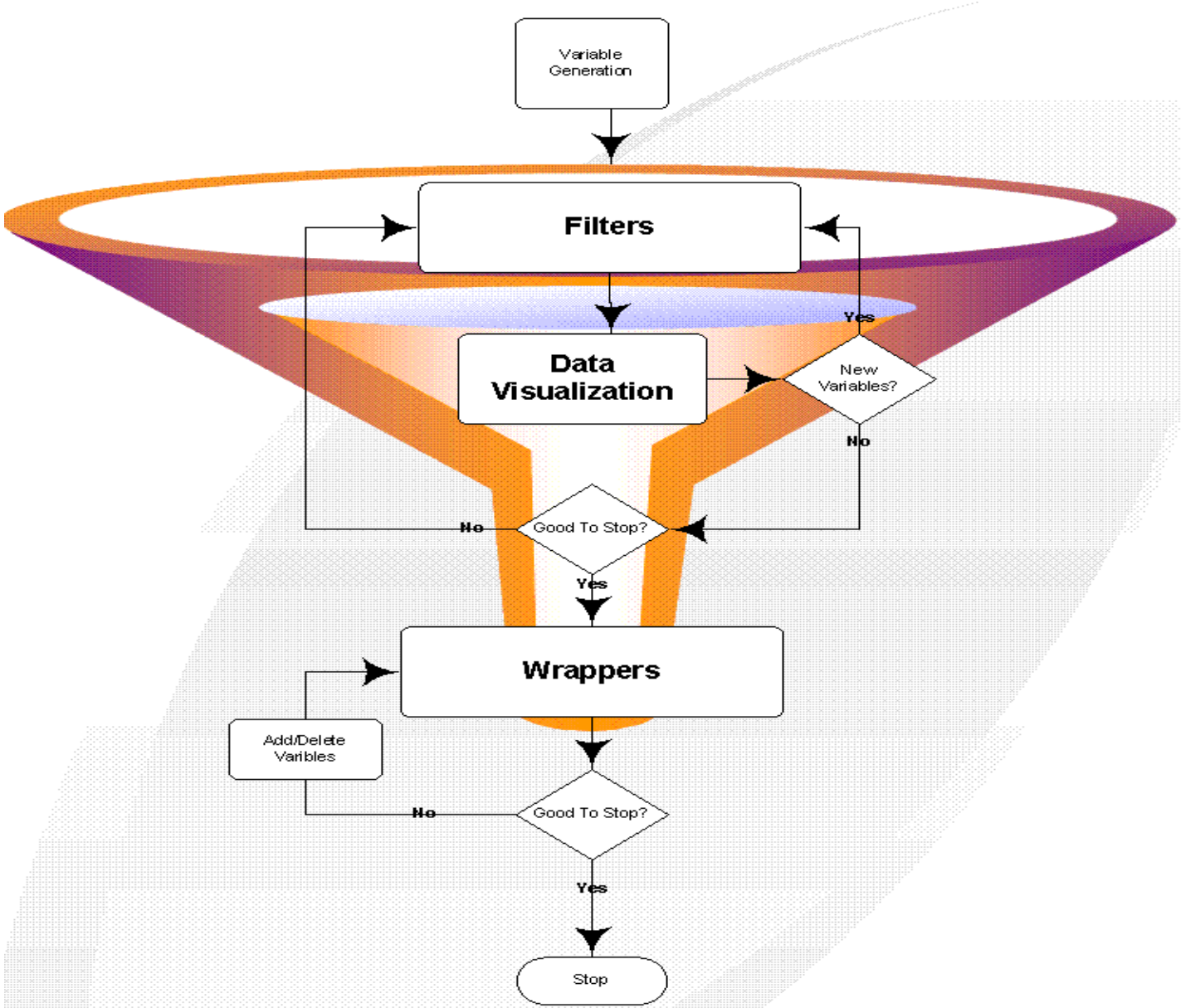
2. Fit models (estimate parameters) K times, omitting one sample each time.

3. Fit the models on the omitted sample and calculate the prediction error.

4. Average the prediction error across the K samples.

    - Leo Breiman's "Out of Bag Estimation" is a related and useful approach.

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

References

• Givens and Hoeting (2005). *Computational Statistics*. Wiley Chapter 3: Combinatorial Optimization.

• Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso*. J. Royal Statistical Society B, 58 (1) pp. 267-288.

• Zou and Hastie (2005). Regularization and Selection via the Elastic Net. J. Royal Statistical Society B, 67 (2) pp. 301-320.

•SAS/STAT 9.3 User's Guide, 2nd Ed.

•Documentation on R http://cran.r-project.org/

• Feature Selection Methods
  •Ravi Kumar and Peter Wu, CAS PM seminar Oct 2008