PM-11 Predictive Modeling

Data Mining vs Statistical Analysis

A presentation for 2011 CAS RPM Seminar
By Serhat Guven, FCAS, MAAA

March 22, 2011

TOWERS WATSON

Agenda

- Background
- Predictive Modeling Techniques
- Predictive Modeling for Insurance
- Conclusion

Background

- Purpose of modeling is to separate the signal from the noise

Response Variable = Systematic Component + Random Component

Signal:
Function of the Rating Factors/Predictors

Noise:
Reflects stochastic process

- Goals of a model
  - Predictive accuracy
  - Explanatory power
  - Individual parameter importance
  - Interactions among parameters

Background

- Types of Predictive Models
  - Descriptive
    - Summarization of data in convenient ways to increase understanding
  - Classification
    - Prediction of a categorical unknown value of a variable (response) given the known values of other variables (predictors)
  - Regression
    - Prediction of a continuous unknown value of a variable (response) given the known values of other variables (predictors)

Background

- Alternate taxonomies of predictive models
  - Parametric vs. Nonparametric
    - Parametric: a fitted value is described by a functional relationship applied to specific parameters
    - Nonparametric:
      - Techniques that do not rely on data belonging to any specific distribution
      - Structure of the model is not fixed and grows as the complexity of the data grows
  - Supervised vs. Unsupervised Learning
    - Supervised: training data consists of predictors and responses
    - Unsupervised: training data consists of predictors only

Background

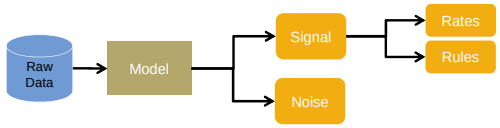
- Components of a predictive model

$$\text{Response Variable} = \text{Systematic Component} + \text{Random Component}$$

- Model structure
  - Determines the functional relationship between the predictors and response
- Score function
  - Quantifies the relationship between the model structure and the actual data
- Optimization Strategy
  - Calculate the resulting parameters

Background

- Predictive modeling for insurance



- Rates
  - Require parametric structure
  - Exceptions include boundaries, symboling, other high dimensional constructs
- Rules
  - Tier and company placement can be more tree based
- Residual analysis provide the framework

Predictive Modeling Techniques

Principle Component Analysis

- Goal: Identify a smaller number of dimensions as combinations of the original dimensions
- Components
  - Model Structure
    - Let Z be an n x c matrix of principle components where $Z=X\mu$
    - Such that μ is a p x c eigenvector matrix
  - Score Function
    - Find m such that Var(Z) is maximized subject to the constraint that $\mu^T\mu = 1$
  - Optimization Strategy
    - Solve the following systems of equations $R\mu = \lambda\mu$
    - Where R is the correlation matrix and λ is the eigenvalue
- Taxonomy – descriptive & unsupervised

Predictive Modeling Techniques

Principle Component Analysis

- Example: external geodemographic data

GEO_UNIT	GEO_POP_DENSITY	GEO_MED_AGE	GEO_UNEMP
A	100,000	34	5%
B	50,000	55	6%
…	…	…	…

x

GEO	PC1	PC2	PC3
GEO_POP_DENSITY	1.50	0.75	0.68
GEO_MED_AGE	0.40	1.20	0.34
GEO_UNEMP	2.00	3.00	1.50

GEO_UNIT	PC1_SCORE	PC2_SCORE	PC3_SCORE
A	150,013	75,040	68,011
B	75,022	37,566	34,018
…	…	…	…

Predictive Modeling Techniques

Principle Component Analysis

- Advantages
  - Scores are independent of each other
  - Models can use significantly fewer score dimensions than geo dimensions
  - Coefficients can be used to select similar geo
- Disadvantages
  - Interpretative
  - Implementation

Predictive Modeling Techniques

Clustering

- Goal: Produce predictive groupings that minimize within group heterogeneity while maximizing cross group heterogeneity
- Components
  - Model Structure
    - Given a set of data points find clusters such that each data point is assigned to a unique cluster
  - Score Function
    - Minimize within cluster heterogeneity and maximize the between cluster variation where variation is based on 'distance'
  - Optimization Strategy
    - Combinatorial optimization (allocate n objects into k classes)
    - Iterative heuristic search mechanisms
- Taxonomy – descriptive & both supervised and unsupervised

Predictive Modeling Techniques

Clustering

- Example: bucketing vehicles units into profile groups based on similarities in vehicle characteristics
- Advantages
  - Non parameteric
  - Interpretive
  - Data controls
- Disadvantages
  - Heavily dependent on choice of score functions
  - Difficult to incorporate dimensional contiguity constraints
  - Difficult to gauge success of clustering routine

Predictive Modeling Techniques

Spatial Analysis

- Goal: fitted values follow the principle of locality
- Components
  - Model Structure
    - Neighboring structure
  - Score Function
    - Bayesian applications distribution analysis
  - Optimization Strategy
    - Search routine
- Taxonomy – descriptive & supervised

Predictive Modeling Techniques

Spatial Analysis

- Examples
  - Geographic smoothing
  - Vehicle symboling
- Advantages
  - No requirement to select a distribution
  - Non parameteric solution for high dimensional variables
- Disadvantages
  - Difficult to judge the appropriateness of the smoothing function
  - Interpretive challenges for non geographic components

Predictive Modeling Techniques

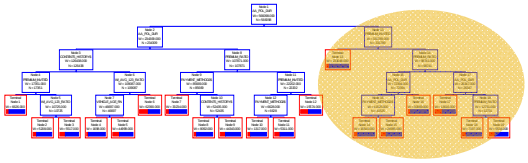
Classification and Regression Trees

- Goal: Identify importance of individual factors and/or determine whether localized models should be constructed
- Components
  - Model Structure
    - Decision Tree
  - Score Function
    - Cross Validation: Builds tree on training data and estimates misclassification on validation data
    - Different sets of training and validation data used to develop overall misclassification rate
  - Optimization Strategy
    - Greedy search over structures: recursively expands tree and prunes limbs.
- Taxonomy – classification/regression & supervised

Predictive Modeling Techniques

Classification and Regression Trees

- Example: indicating localization strategies



- Significantly different branch structure suggests data split and model localization

Predictive Modeling Techniques

Classification and Regression Trees

- Advantages
  - Easy to understand
  - No prior assumptions about the data (non parametric)
  - Able to process numerical and categorical data
- Disadvantages
  - Tree structure is unstable
  - Can be overly complex
  - Prediction can create irregularities in continuity

Predictive Modeling Techniques

Regression Techniques (Generalized Linear and Nonlinear Models)

- Goal: predication of an unknown response given the known values of the predictors
- Components
  - Model Structure
    - Functional relationship between the mean and variance
    - Link function applied to a combination of parameters
  - Score Function
    - Likelihood and quasi likelihood functions
  - Optimization Strategy
    - Maximize likelihood function
    - Newton Raphson routine
- Taxonomy – regression & supervised

Predictive Modeling Techniques

Generalized Linear and Nonlinear Models

- Examples
  - Loss cost modeling
  - Demand modeling
- Advantages
  - Closed form solution
    - Interpretive ease
    - Transparency
  - Robust in dealing with bad/missing data
- Disadvantages
  - Specification of the mean-variance function difficult to validate
  - High dimensional data vectors

Predictive Modeling Techniques

Regression Techniques (Neural Networks)

- Goal: predication of an unknown response given the known values of the predictors
- Components
  - Model Structure
    - Recursive logit link function
  - Score Function
    - Sum of squared error
  - Optimization Strategy
    - Newton Raphson routine
- Taxonomy – regression & supervised

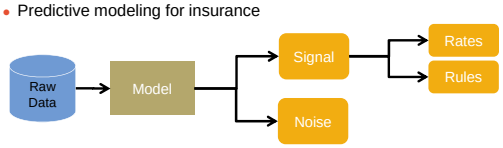
Predictive Modeling Techniques

Neural Networks

- Examples
  - Loss cost modeling
  - Demand modeling
- Advantages
  - No requirement to select a distribution
  - Robust model form to identify more complex structures
- Disadvantages
  - Tends to overfit
  - Responds less well to bad data
  - Interpretive challenges

Predictive Modeling for Insurance

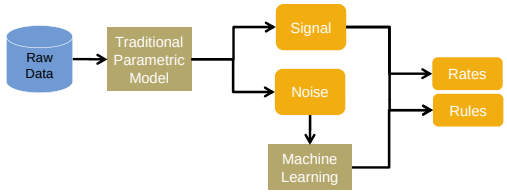
- Predictive modeling for insurance



- Need for a blend of parametric and non parametric techniques
  - Regulatory environment as well as consumer understanding requires a clean parametric solution
  - Ability to properly reflect the differences in loss costs with more complex interaction concepts

Predictive Modeling for Insurance

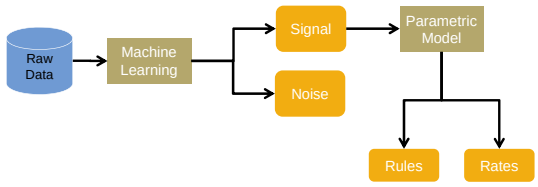
- Blending approaches – a study of residuals



- Residual modeling is useful when controlled

Predictive Modeling for Insurance

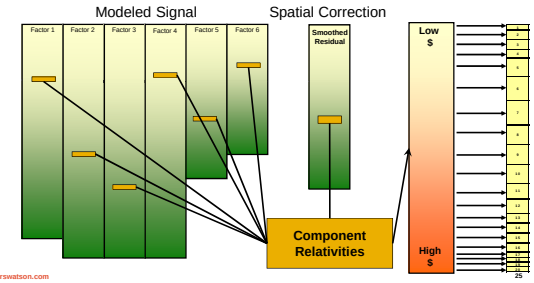
- What about this alternative?



- The machine can grow the signal in the wrong direction thereby influencing the rates/rules
- Difficult to trust the fitted values from the machine

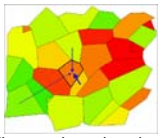
Predictive Modeling for Insurance

- Spatial Correction
  - Residual signal used to adjust score from the multivariate model

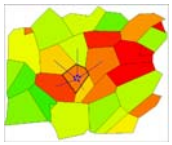


Predictive Modeling Methods

- Spatial Smoothing Methods
  - Uses knowledge of surrounding areas to enhance estimates of the underlying risk in each area based on "Principle of Locality".



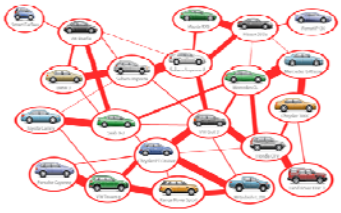
- Distance based methods
  - Simpler to implement and interpret
  - Does not consider natural boundaries
  - May over smooth urban areas and under smooth rural areas



- Adjacency based methods
  - Distribution assumptions allow prior knowledge of claims
  - Distance can still be incorporated
  - Considers natural boundaries

Predictive Modeling for Insurance

- Neighboring vehicles: instead of using latitude/longitude to build adjacency relationships, use vehicle dimensions



- Once neighbors are determined similar techniques used for territorial analysis can be applied

Predictive Modeling for Insurance

- Super profiling
  - Use alternate machine learning techniques to model the residual
    - CARTs
    - Neural Networks
    - Adaptive Networks
    - Etc
- Advantages
  - Very specific customer classes that could only be captured with several layers of scoring or with deep interaction terms
- Disadvantages
  - No clear explanatory effect of the structural output

Conclusion

- Variety of predictive modeling techniques are available to analysts
  - Different forms are useful for different situations
  - Consider the taxonomy and structure of each form
- Modeling loss costs should use a combination of parametric and non parametric methods
  - Parametric methods used the core rate algorithm'
  - Nonparametric methods used more for the rules
- Alternate modeling techniques should be applied to the residual
  - Allows greater control of the resulting model form
  - Over reliance of machine learning methods tend to overfit the data and should be further controlled to specific factors