Predictive Modeling and its applicability in Life Insurance

Tim Heslin and Nabaneeta Sirkar
AIG
Life, Health and Disability Business

For the Actuaries Club of the Southwest

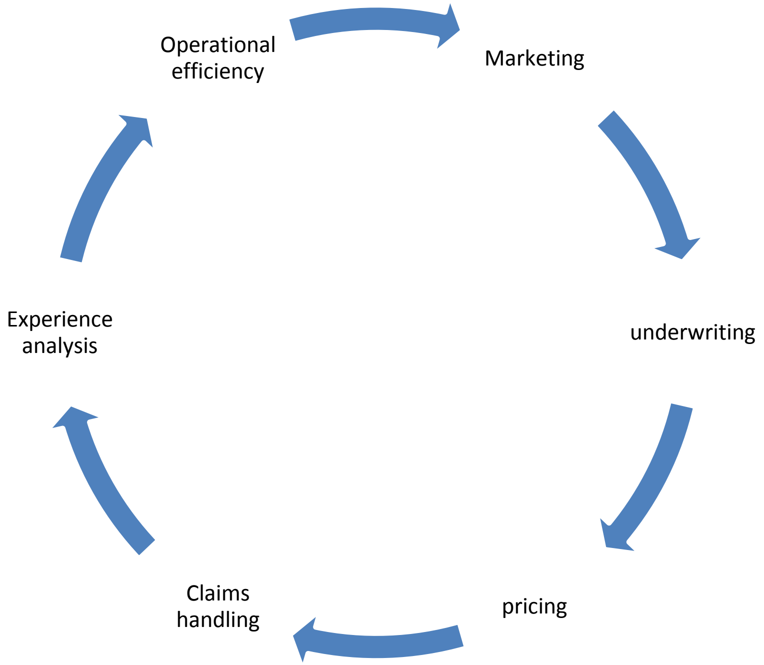
What is Predictive Modeling?

Predictive Modeling can be defined as the analysis of large data sets to make inferences or identify meaningful relationships, and the use of these relationships to better predict future events . It uses statistical tools to separate systematic patterns from random noise, and turns this information into business rules, which should lead to better decision making.

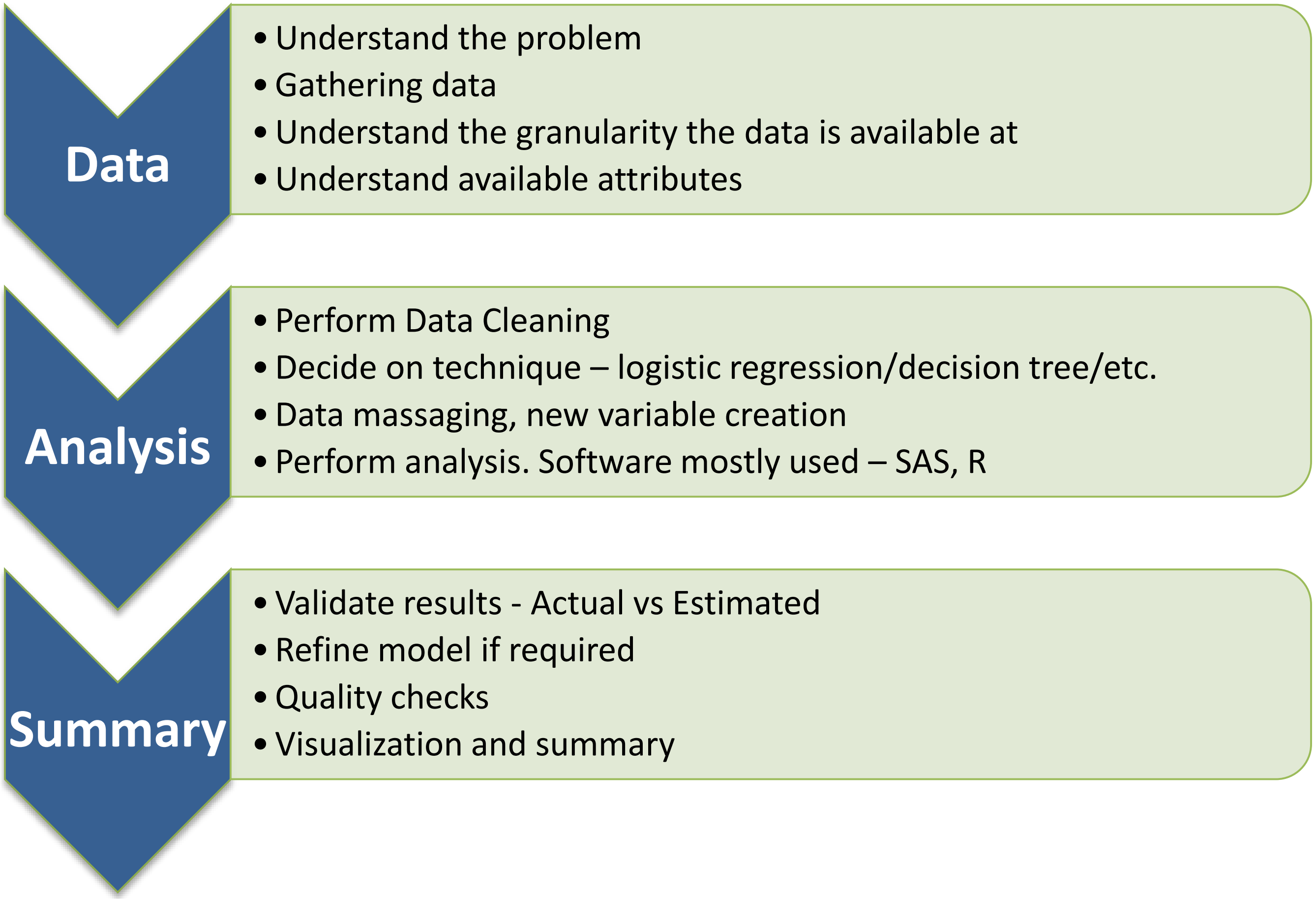
Credit scoring is the classic example of predictive modeling. Credit scores were initially developed to more accurately and economically underwrite and determine interest rates for home loans. Personal auto and home insurers subsequently began using credit scores to improve their risk selection and pricing of personal auto and home risks.

Predictive Modeling Applications

- Database Marketing
- Risk selection in Underwriting
- Fraud detection
- Claims Handling
- Pattern recognition
- Operational Efficiencies



Predictive Modeling Process



Data

- Understand the problem
- Gathering data
- Understand the granularity the data is available at
- Understand available attributes

Analysis

- Perform Data Cleaning
- Decide on technique – logistic regression/decision tree/etc.
- Data massaging, new variable creation
- Perform analysis. Software mostly used – SAS, R

Summary

- Validate results - Actual vs Estimated
- Refine model if required
- Quality checks
- Visualization and summary

It all starts with the DATA

- Cleanliness of Data
- At what level it is available
- What is the size
- Static or Dynamic

As covered on Slide 4

Target Variables

- Binary (Yes/No) – eg - Mortality
- Multiple Classes – eg – Different demographic segments
- Continuous – eg – Size of claim

As covered on Slide 4

Modeling Methods

Depending on the type of target variable, the methods will vary

- Generalized Linear Models
- Decision Trees
- Clustering and other segmentation frameworks
- Machine Learning methods

As covered on Slide 4

What is a GLM?

Generalized Linear Models (GLM) is a generalization of the ordinary Linear Regression Model that allows the response variables to have some other non-normal error distribution. GLM is often used in general insurance for determining premiums.

A Generalized Linear Model (GLM) consists of three elements –

- A probability distribution (this may be extended to any distribution from the exponential family – Binomial or Poission)
- A linear predictor that is a function of covariates' i.e. $\eta = \mathbf{X\beta}$.
- A link function g such that $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\eta)$.

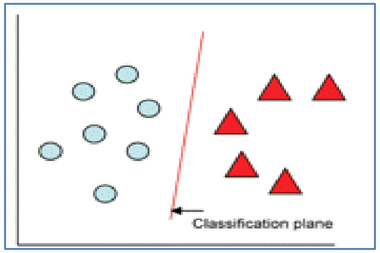
Logistic Regression : When the response variable is binary (say 0 and 1), then the distribution function is chosen to be Bernoulli and the interpretation of μ is the probability of the response variable taking ‘1’. This special case of GLM is called Logistic Regression.

What is Machine Learning?

- It deals with the construction and study of algorithms that can learn from and make predictions on data.
- The broad categories of machine learning are :

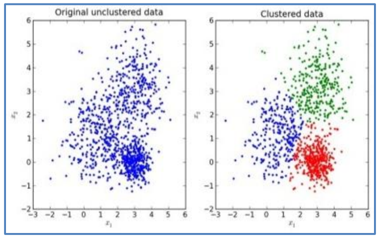
Supervised Learning

- The learning algorithm is presented with inputs and their desired outputs
- The goal is to learn a general rule that maps inputs to outputs
- E.g. Decision trees, Neural Networks, SVM, etc.



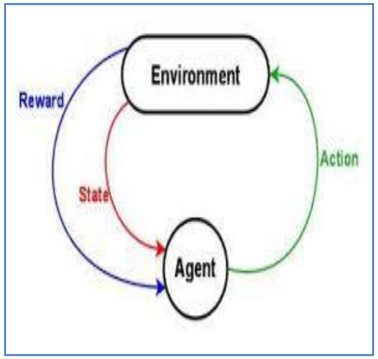
Unsupervised Learning

- No labels are given to the algorithm
- The goal is to discover the hidden patterns in its inputs
- E.g. Clustering, PCA, SVD, etc.



Reinforcement Learning

- It is the problem of getting an agent to act in the world to optimize its rewards/losses.
- In other words, optimizing a reward/loss function rather than the actual problem in hand, e.g. teaching a dog a new trick: you cannot tell it what to do, but you can reward/punish it if it does the right/wrong thing.

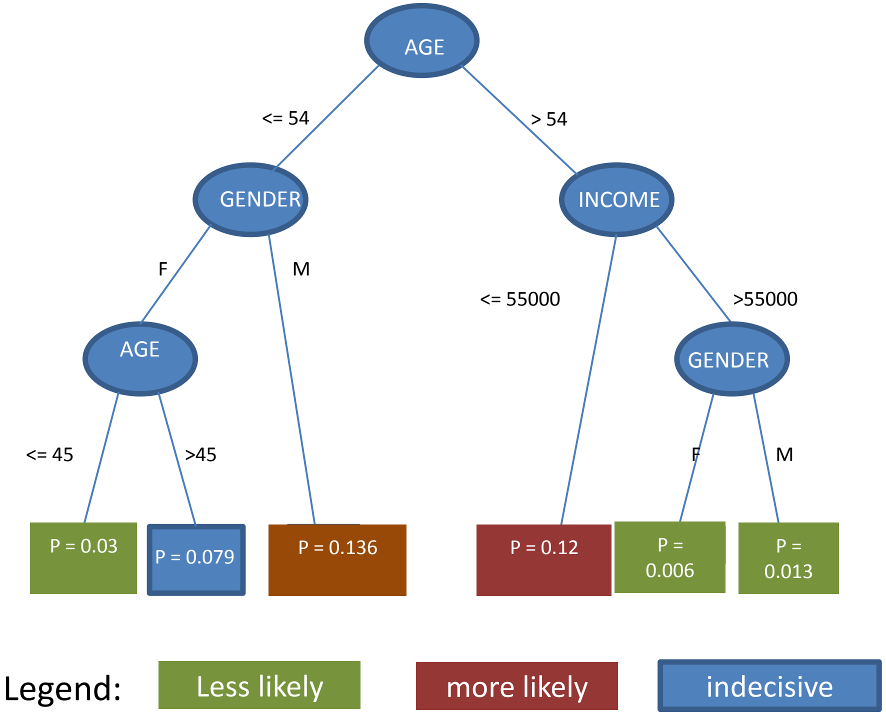


Illustrative example of decision tree

Decision tree is an algorithm which helps us to identify various ways of splitting a data set into branch-like segments. It allows to investigate relationships that exist locally within subgroups of data.

Decision Tree on mortality

- Classify the data based on high or low risk of mortality.
- Dependent variable:-
  - Death = “yes”
- Independent variables:-
  - Age
  - Gender
  - Income



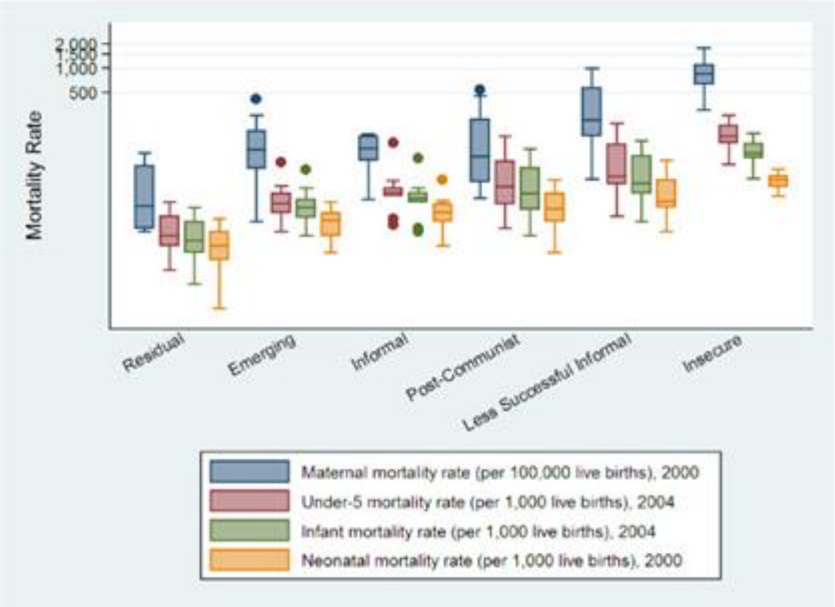
Example of clustering exercise

- **Cluster Analysis**: Clustering is a task of grouping a set of objects, based on information found in the data, so that objects in the same group(cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). There are various ways to form clusters and hierarchical clustering & non-hierarchical clustering are two mostly used methods.
  - **Hierarchical Clustering:** One method is each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Another method is all observations start in one cluster and splits are performed recursively as one moves down the hierarchy.
  - **Non-Hierarchical Clustering:** One of the famous method is *k-means clustering* where a set of points, called cluster seeds, is selected as a first guess of the centroids of the clusters. Each observation is assigned to the nearest seed to form temporary clusters. The seeds are then replaced by the centroids of the temporary clusters, and the process is repeated until no further changes occur in the clusters.

Using Gross National Product per capita 113 countries are clustered into 6 groups to conduct further study on mortality.

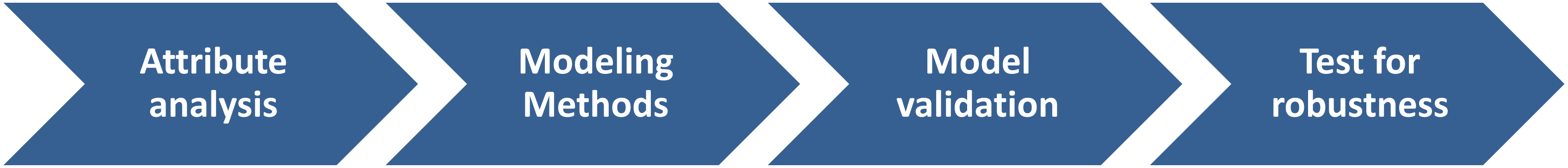
Taxonomy of labour market clusters by national income

	More Equal	← Labour Market →	Less Equal
Middle-Income	Residual	Emerging	Informal
	The Bahamas, Croatia, Czech Rep, Hong Kong, Hungary, Jamaica, Korea Rep, Latvia, Lithuania, Poland, Russian Fed, Singapore, Slovak Rep, Slovenia, Thailand, Uruguay	Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Fiji, Kuwait, Malaysia, Mexico, Panama, Paraguay, Peru, South Africa, Trinidad and Tobago, Venezuela	Bahrain, Belize, Botswana, El Salvador, Lebanon, Oman, Saudi Arabia, Tunisia, Turkey
Low-Income	Post-Communist	Less Successful Informal	Insecure
	Albania, Armenia, Belarus, Bolivia, Bulgaria, Cambodia, China, Ghana, Indonesia, Moldova, Mongolia, Papua New Guinea, Philippines, Romania, Tajikistan, Ukraine, Uzbekistan, Viet Nam	Algeria, Cape Verde, Cote d'Ivoire, Dominican Rep, Egypt, Equatorial Guinea, Guatemala, Guyana, Honduras, India, Iran, Jordan, Mauritania, Morocco, Nicaragua, Nigeria, Pakistan, Sri Lanka, Sudan, Swaziland, Syrian Arab Rep, Yemen Rep	Bangladesh, Benin, Burkina Faso, Burundi, Cameroon, Central African Rep, Chad, Comoros, Congo Dem Rep, Congo Rep, Eritrea, Ethiopia, Gambia, Guinea-Bissau, Haiti, Kenya, Lao PDR, Madagascar, Malawi, Mali, Mozambique, Namibia, Nepal, Niger, Rwanda, Senegal, Tanzania, Togo, Uganda, Zambia, Zimbabwe



Source: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3349504/#!po=4.16667*

Steps in Predictive Modeling



Attribute analysis

- Data visualization
- Degree of influence
- Relationship with response
- Machine learning

Modeling Methods

- Binary/Continuous
- Linear vs. non-linear
- Normal error vs. Generalized model

Model validation

- Test statistics
- Validity of result, p-value
- Relativity check

Test for robustness

- K-fold validation
- Out of sample validation

Over fitting

Model complexity: As the complexity of models and number of parameters increases, accuracy of model on training sets improves but makes the model more unstable

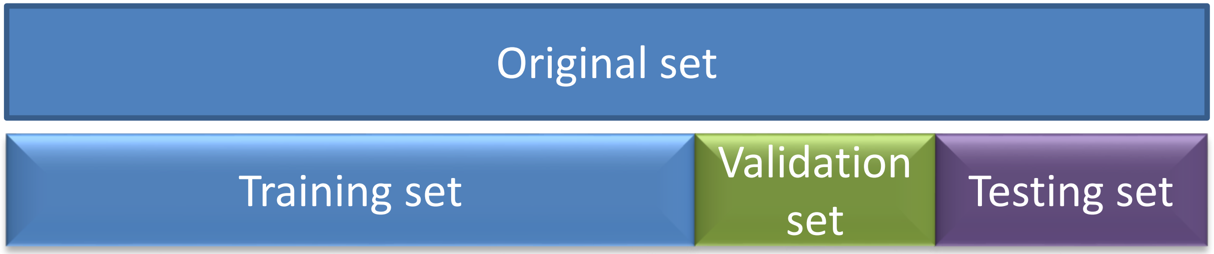
How to check and avoid over fitting?

Out of sample validation and k-fold validation are the two most commonly used methods to test for over fitting

Out of Sample Validation

- Objective : This is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.

  — In a prediction problem, the 'original dataset' is split in two parts :
    - Training + Validation data (90%) : This data is further split into training and validation data in the ratio of 80:20. The model is built on the training data and it is tested on the validation data.
    - Out-of-sample data (10%) : If the model fits well on both training and validation data then it tested on the out-of-sample data.



Limitation: the test for over fitting relies completely on a single set of validation data. K-fold validation can overcome this limitation.

K-fold cross validation

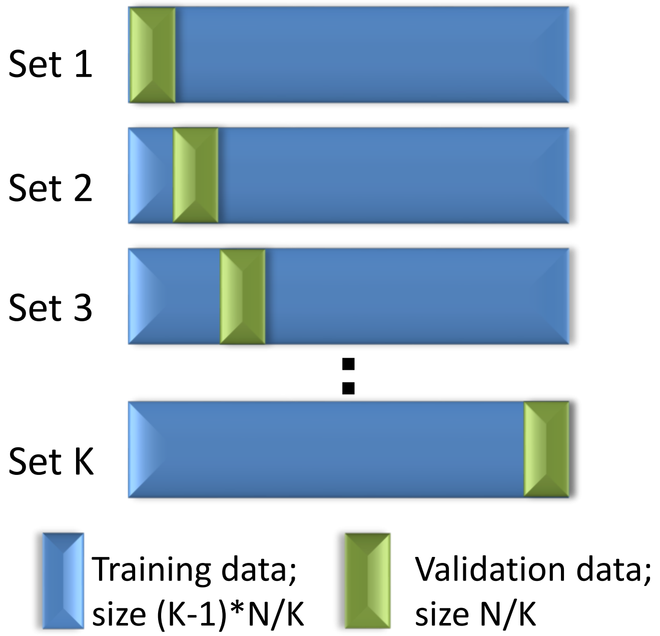
K-fold validation utilizes complete data set by performing multiple splits. Therefore it is more suited where the size of the data is small

Case 1: N is sufficiently large

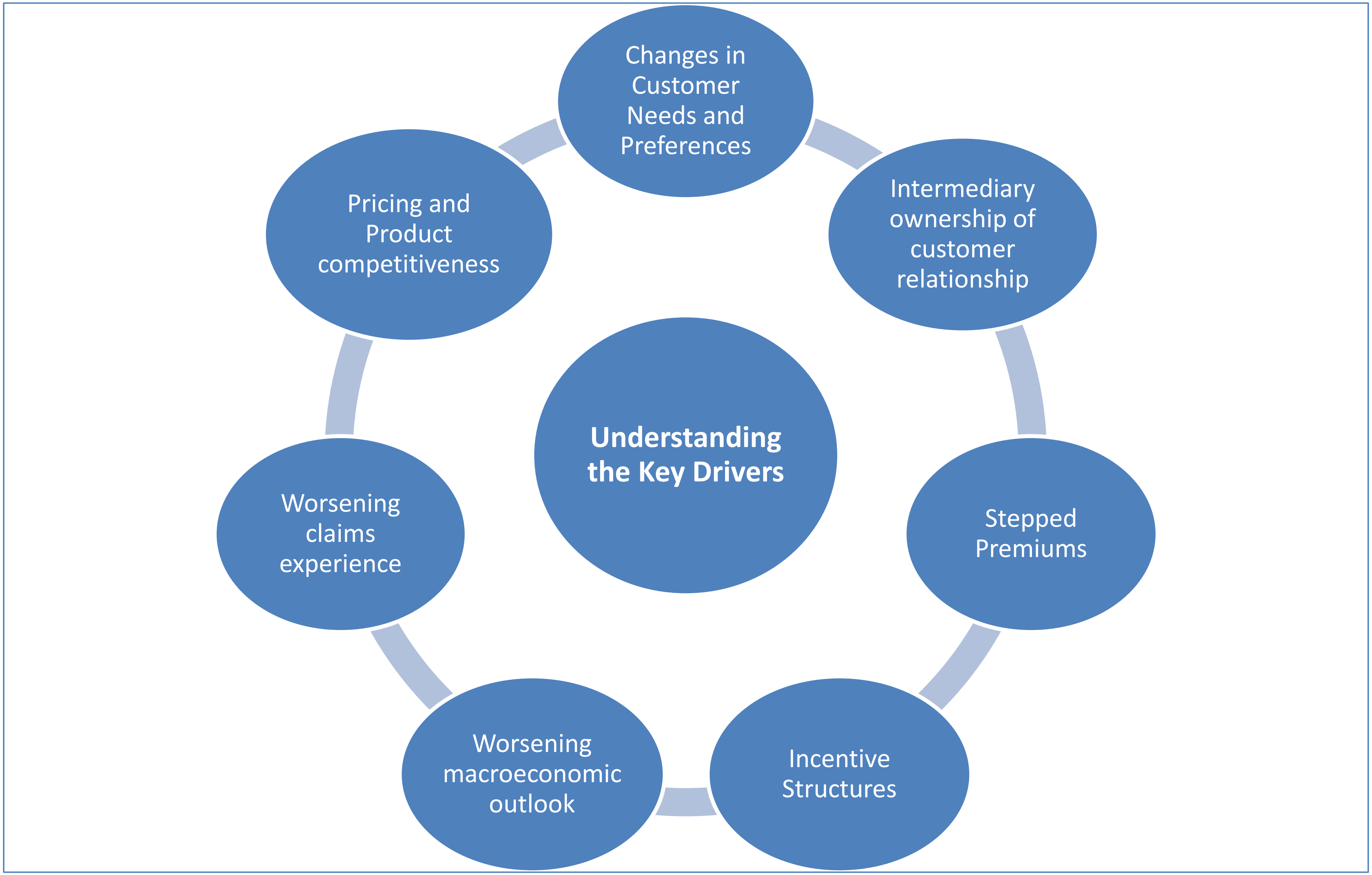
Build K different model on the K training data sets; test for model accuracy on validation set.

Case 2: sparse data

Build K different model on the K training data; test for robustness on parameter estimate over the K model



Example : Customer Retention - It’s Drivers



Understanding the Key Drivers

Changes in Customer Needs and Preferences

Intermediary ownership of customer relationship

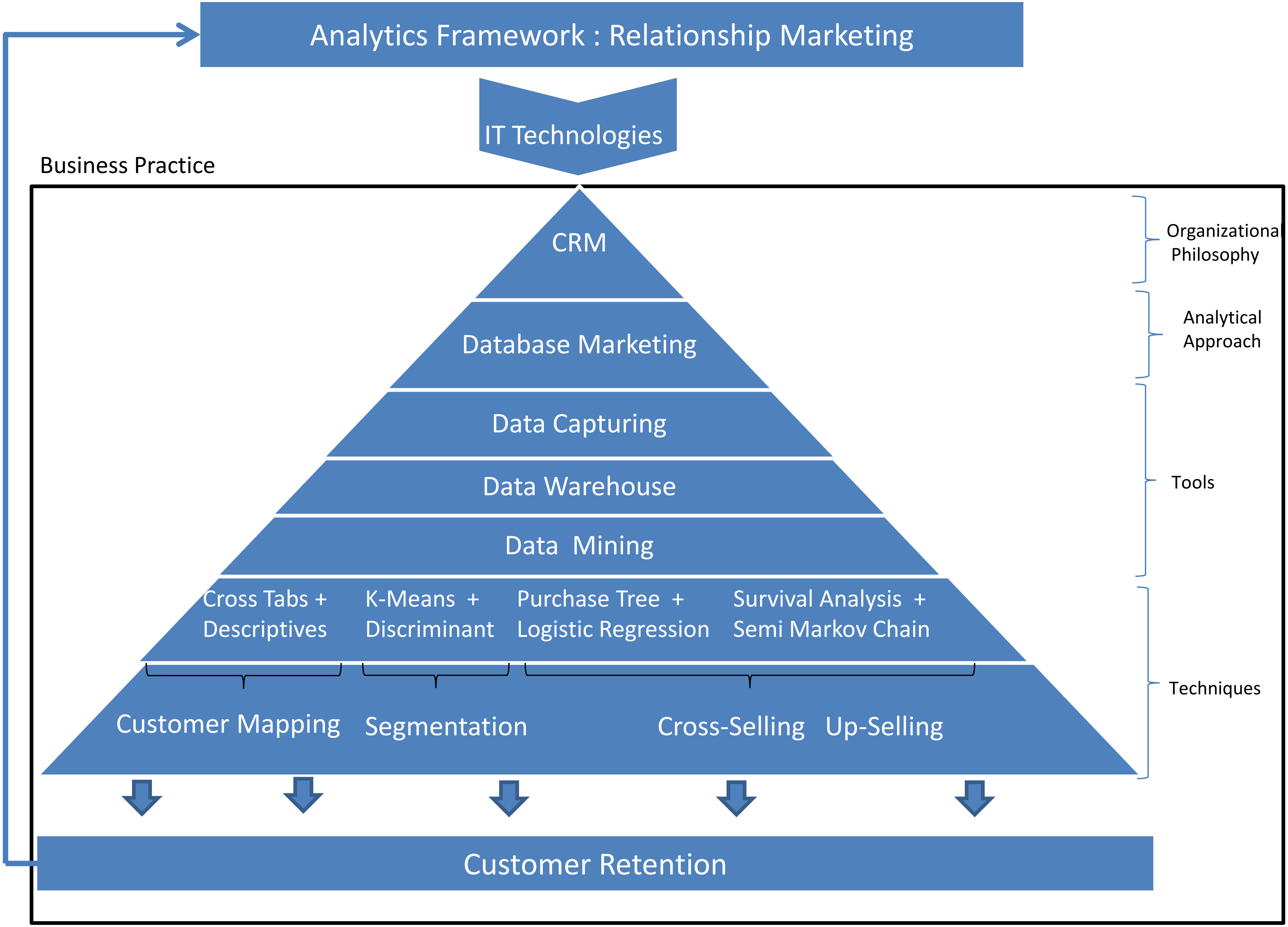
Stepped Premiums

Incentive Structures

Worsening macroeconomic outlook

Worsening claims experience

Pricing and Product competitiveness



Analytics Framework : Relationship Marketing

IT Technologies

Business Practice

CRM

Database Marketing

Data Capturing

Data Warehouse

Data Mining

Cross Tabs + Descriptives

K-Means + Discriminant

Purchase Tree + Logistic Regression

Survival Analysis + Semi Markov Chain

Customer Mapping

Segmentation

Cross-Selling Up-Selling

Customer Retention

Organizational Philosophy

Analytical Approach

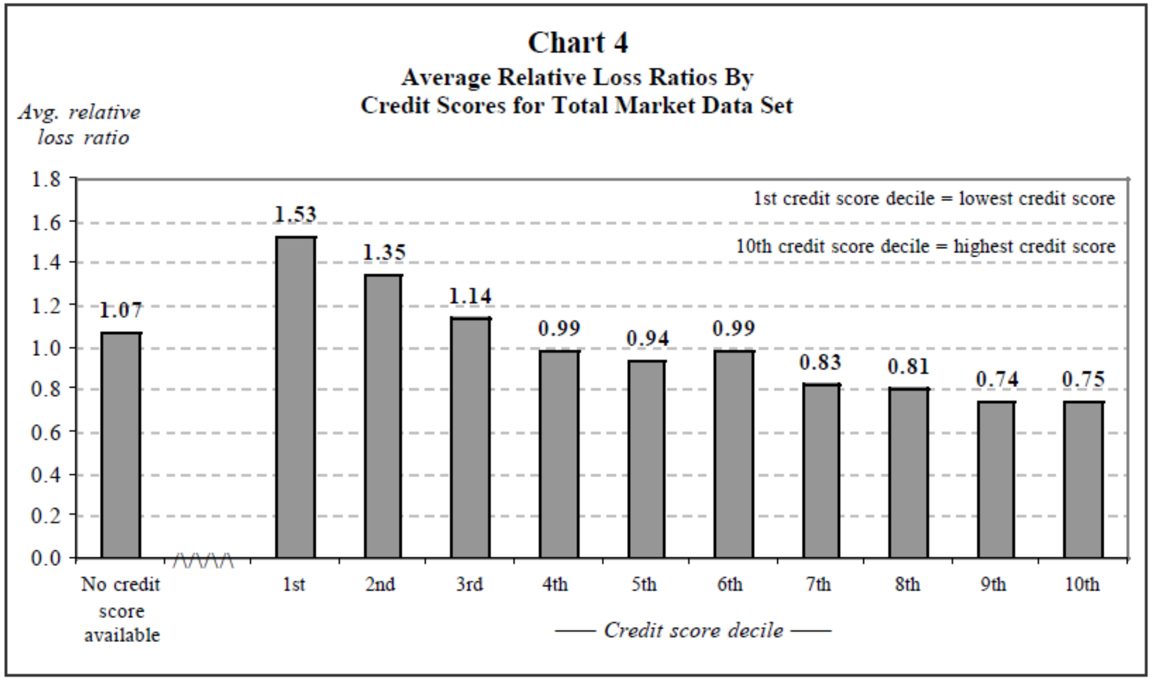
Tools

Techniques

Source: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/4448/Salazar2010.pdf;jsessionid=A31599E97285661B7D3301A3EA512B39?sequence=2

Credit score use in Auto insurance

The Chart below displays a strong correlation between credit scores and relative loss ratios



Source: http://www.progressive.com/content/PDF/shop/UTCreditStudy.pdf