Predictive Modeling

And Its Application in Insurance

Richard Xu, **PhD FSA**

VP & Actuary, Head of Data Science

Global R&D, RGA

June, 2015 **India**

Contents

- PM Introduction
- Statistical Models
- Data and Modeling
- PM Case Study

What is Predictive Modeling

1. **Data**
   High quality data
2. **Model**
   Statistical model
3. **Prediction**
   Business decisions

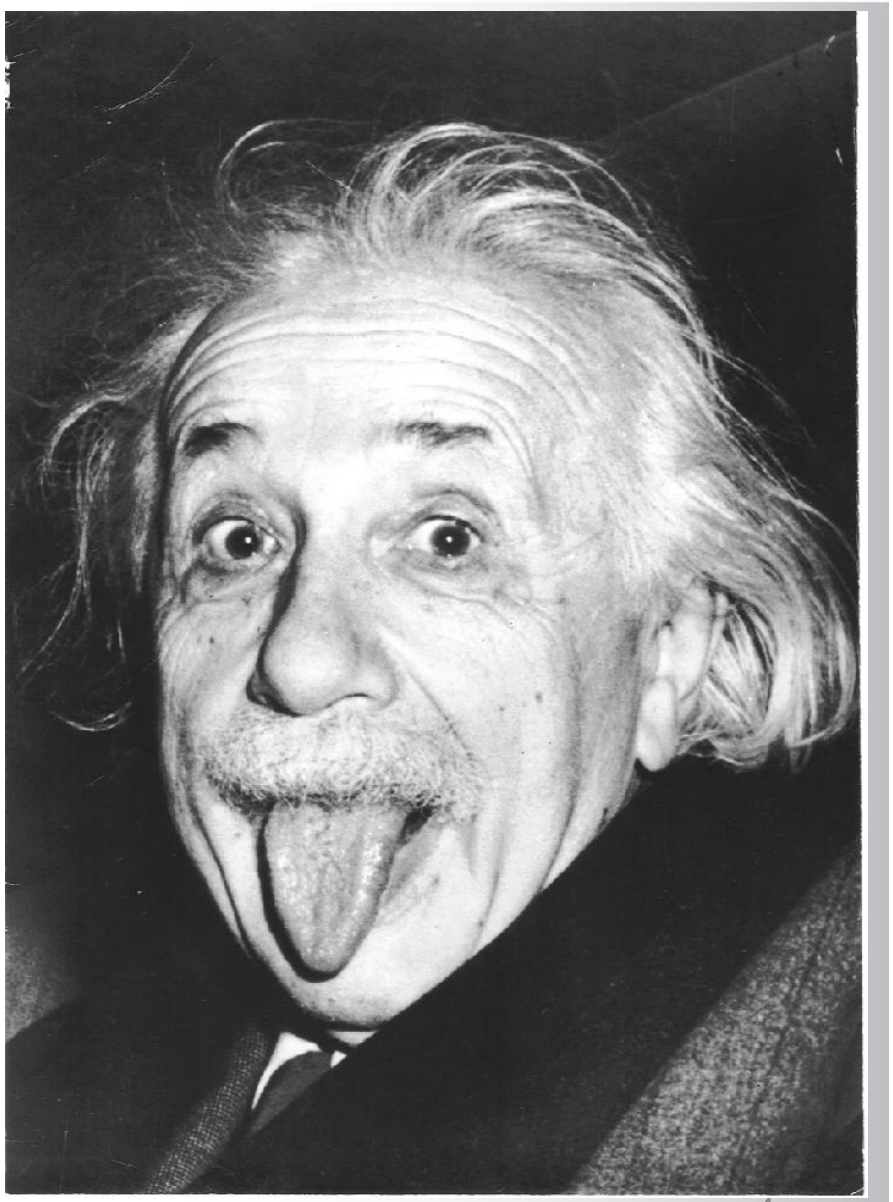
- Grow Business
  - Identify potential customers, new & existing
  - Enhance sales process
- Improve efficiency
  - Accurate & granular view of driven factors of experience
  - Better risk segmentation
- An advantage hard for competitors to replicate

PM is about statistics, but more about data & business

Introduction

Know where to find the information and how to use it - That's the secret of success

Albert Einstein



Linear Regression Model

- Linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon = \Sigma_i \beta_i X_i + \varepsilon, \varepsilon \in (0, \sigma^2)$$

  Ordinary Least Squares(OLS) or Maximum Likelihood Estimation (MLE)

$$\hat{\beta} = arg\min(\Sigma_i(\hat{y_i} - y_i)^2) \ or \ = arg\max(L(\beta_i, y_i))$$

- Widespread applications in various fields
  - Inherently linear process, or well-approximated by LM
  - Effective and efficient with data
  - Easy to understand and communicate
- Great! But …there are issues when applying LM
  - Non-linear, normality, unbounded data, etc.
- How about insurance application?

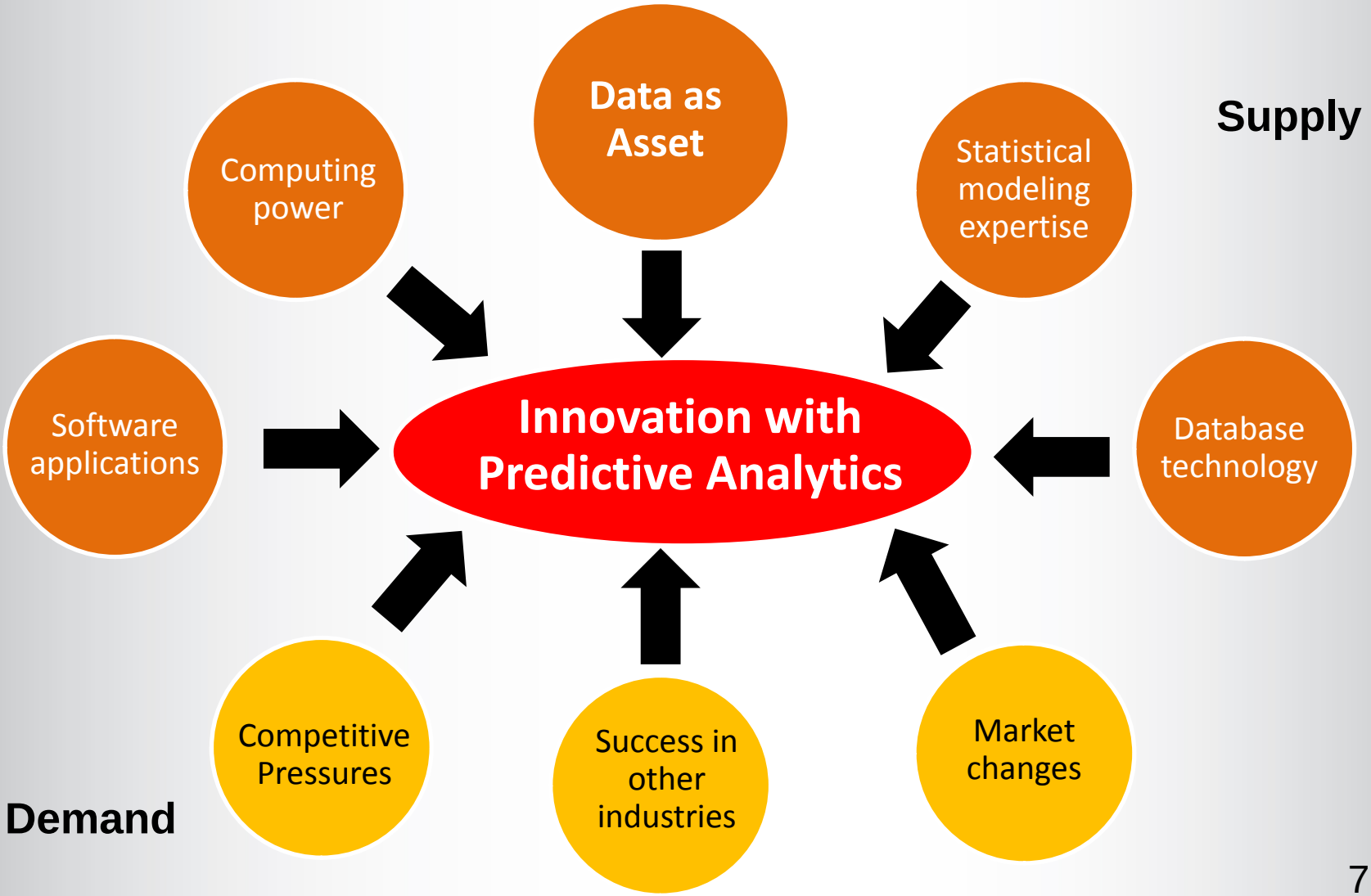
  Distribution of data, variance structure, for example
  - Poisson for claim count, ~ mean

Why PM, Why Now

Why PM, Why Now

Why are Predictive Analytics becoming so Popular?

Increasing ease of access to enablers plus demand pull



Generalized Linear Model

- Generalized Linear Model(GLM)
  - Main focus of PM in insurance industry
  - Inclusion of most distributions related to insurance data
    - Normal, binomial, Poisson, Gamma, inverse-Gaussian, Tweedie
  - Ordinary Least Square (OLS) is a special case of GLM
  - Easy to understand/communicate
  - Multiplicative model intuitive & consistent with current insurance practice

- OLS(LM) $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \Sigma_i \beta_i X_i$
- GLM $g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \Sigma_i \beta_i X_i$

Generalized Linear Model

- Compare OLS and GLM

	Random	Systematic	Link
OLS	Normal only	$\eta_i = \sum_j x_{ij}\beta_j$	$E(y_i) = \eta_i$
GLM	Various distrib.		$g(E(y_i)) = \eta_i$

- Link function

	Identity	Log	Logit	Reciprocal
$g(\mu_i)$	x	$\ln(x)$	$\ln(\frac{x}{1-x})$	$1/x$
$g^{-1}(\eta_i)$	x	e^x	$\frac{e^x}{1+e^x}$	$1/x$

  - Log is unique in insurance application s.t. all parameters are multiplicative

$$\mathrm{E}(y) = \exp(\textstyle\sum_j x_{ij}\beta_j) = \prod_j \exp(x_{ij}\beta_j) = \prod_j \exp(\beta_j)^{x_{ij}} = \prod_j f_j^{x_{ij}}$$

  - Intuitively easy to understand and communicate

Generalized Linear Model

Distribution	V(μ)	Link	Sample Application
Normal	1	Identity	(LM) General Application
Poisson	μ	Log	Claim frequency/count, experience
Binomial	$\mu(1-\mu)$	Logistic	Retention, cross-sell, UW, experience
Gamma	μ^2	Log	Claim severity
Compound	$\mu^p, p \in (1,2)$	Log	Claim Cost & Premium
Inverse-Gaussian	μ^3	Log	Claim cost

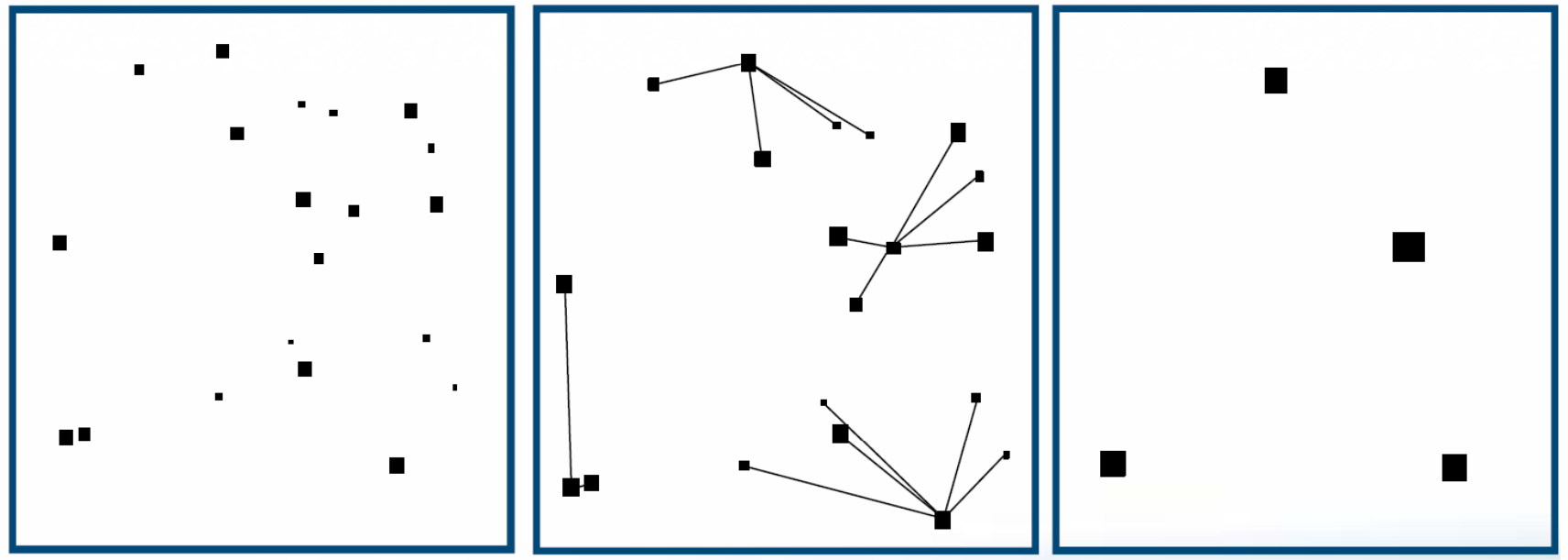
- "Bread and Butter" for PM in insurance
  - ✓ Great flexibility in variance structure
  - ✓ Multiplicative model intuitive & consistent
  - ✓ Non-linear function between variables
  - ✓ Weights & offset to be more flexible

Beyond GLM

- Regression study
  - Linear, general linear, nonlinear
  - Generalized LM(GLM)
  - Survival Models (Cox Proportional Hazard)
  - Generalized Additive Models (GAM)
  - Multilevel/Hierarchical Linear Model(HLM)
- Time series analysis
- Some other advanced tools
  - Data clustering
  - Decision tree based: CART, Random Forest, MARS, Ada-Boosting, etc.
  - Other machine learning algorithms
    - Neural network, Genetic programming, Support vector machine, Bayesian inference, Cluster analysis, K-nearest neighbor

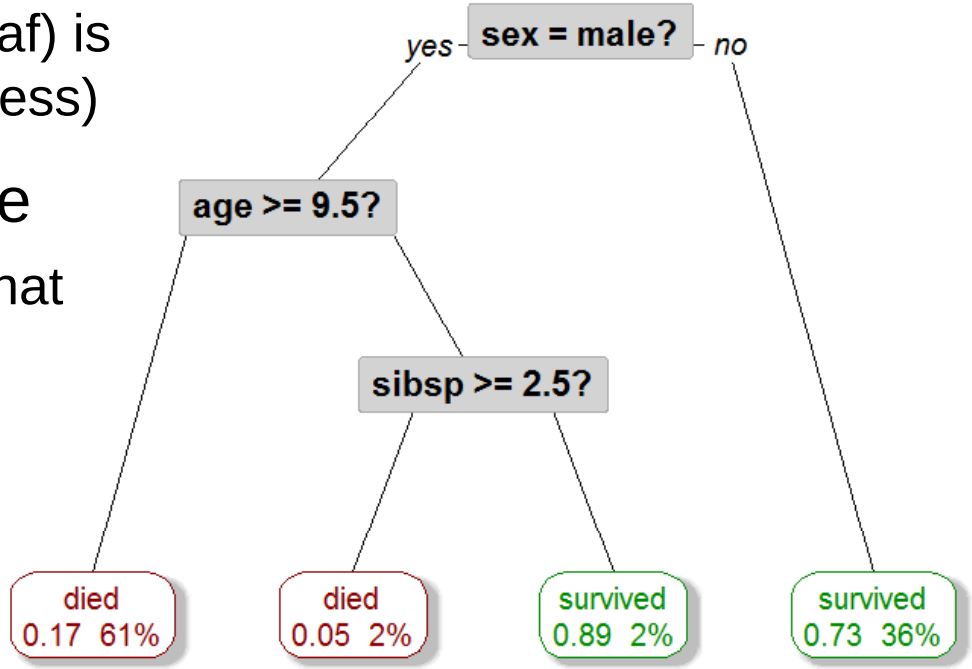
Data Clustering

- Clustering algorithm
  - ✓ Find similarities in data according to features found in data and group similar objects into clusters
  - ✓ Unsurprised (no pre-defined), classification, non-parametric
  - ✓ How to measure similarities/dissimilarities, e.g. distance
    - Numeric, categorical, and ordinal variables
  - ✓ Partitioning (k-means), Hierarchical, Density-based, etc.



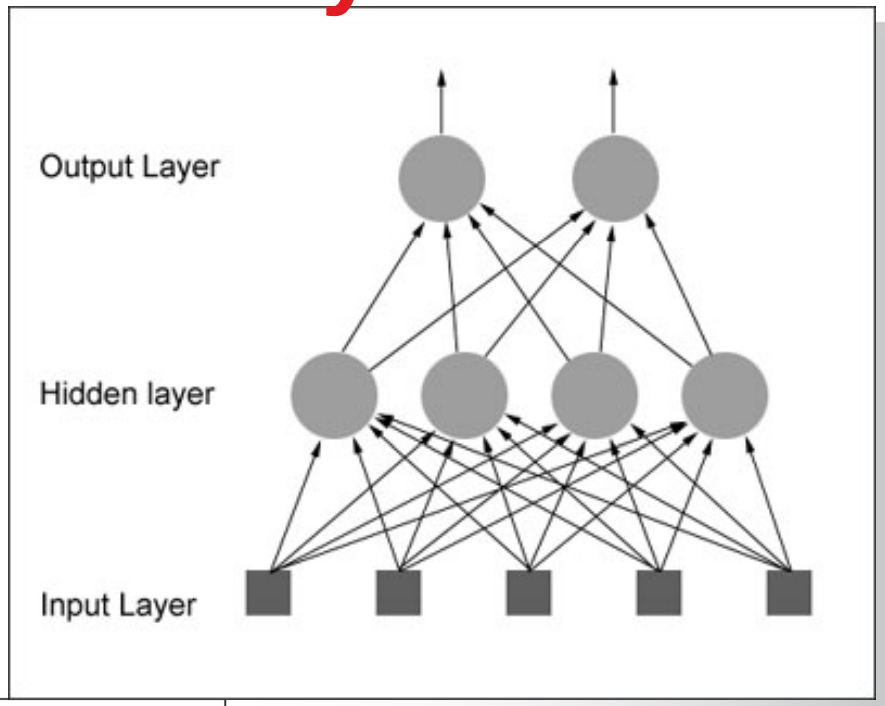
CART Model

- Classification And Regression Tree (CART)
  - ✓ Both classification and regression
  - ✓ Non-parametric approach (no insight in data structure)
- CART tree is generated by repeated partitioning of data set
  - ✓ Data is split into two partitions (binary partition)
  - ✓ Partitions can also be split into sub-partitions (recursive)
  - ✓ Until data in end node(leaf) is homogeneous (more or less)
- Results are very intuitive
  - ✓ Identify specific groups that deviate in target variable
  - ✓ Yet, algorithm is very sophisticated

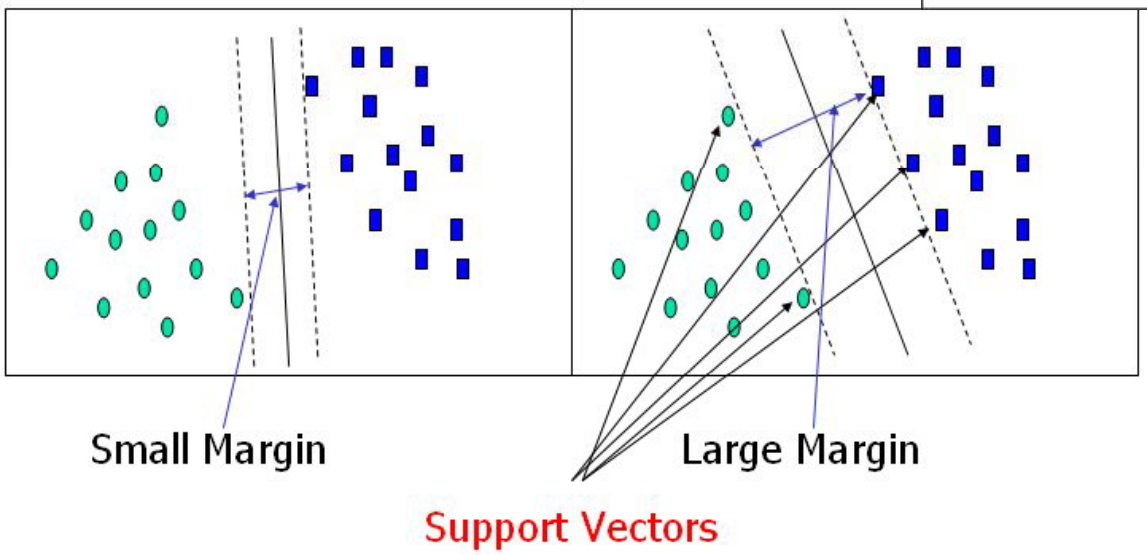


Beyond GLM

- Neural network
  - Powerful
  - Black box approach



- Support vector machine
  - Classification
  - Regression



- Decision tree based
  - CART
  - Ada-Boosting
  - Random Forest

PM terminology

Supervised vs. Unsupervised Learning

- Supervised: estimate expected value of *Y* given values of *X*.
  - GLM, Cox, CART, MARS, Random Forests, SVM, NN, etc.
- Unsupervised: find interesting patterns amongst *X;* no target variable
  - Clustering, Correlation / Principal Components / Factor Analysis

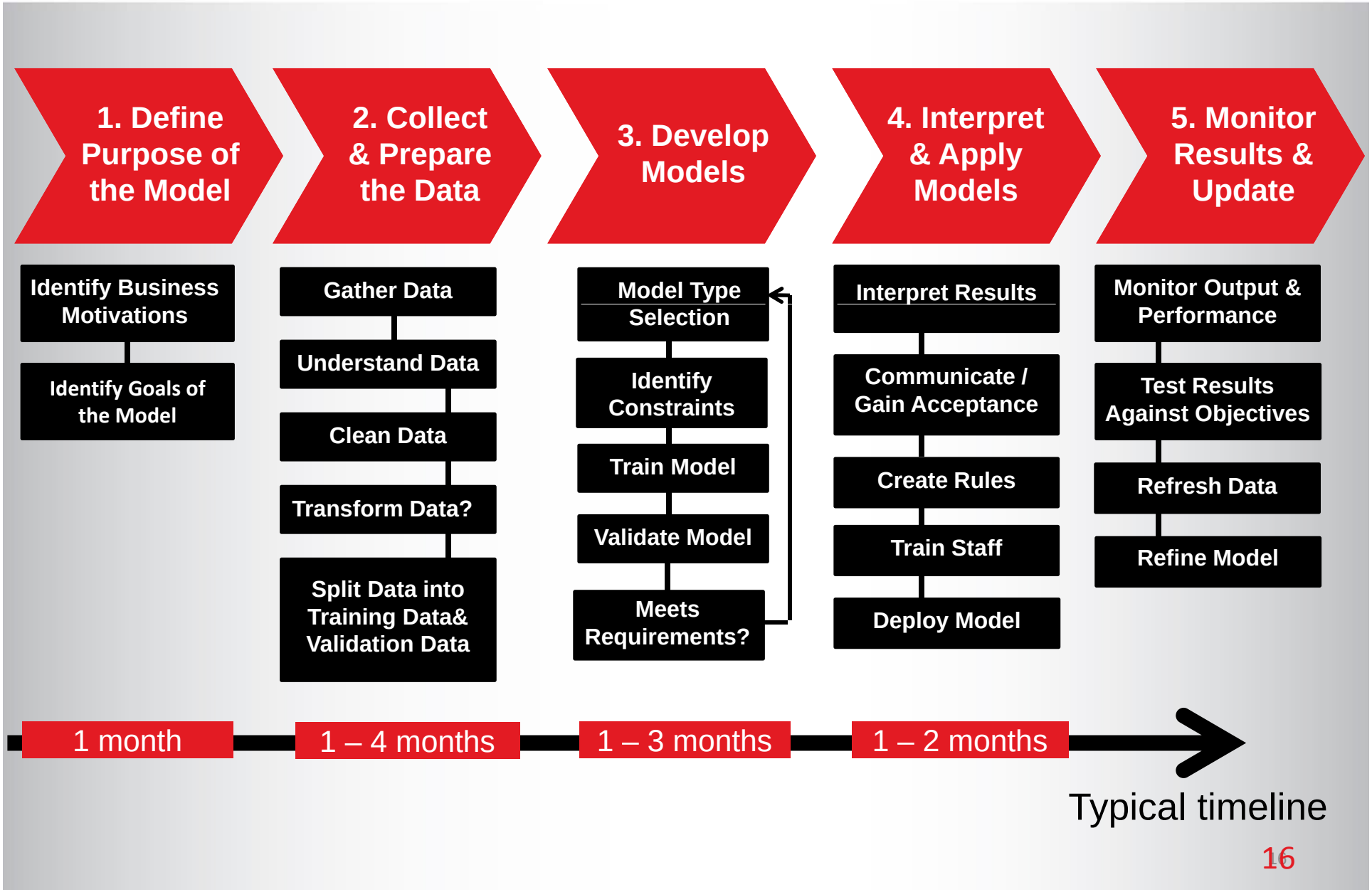
Classification vs. Regression

- Classification: to segment observations into 2 or more categories
  - fraud vs. legitimate, lapsed vs. retained
- Regression: to predict a continuous amount.
  - Dollars of loss for a policy, Ultimate size of claim

Parametric vs. Non-Parametric

- Parametric Statistics: probabilistic model of data
  - Poisson Regression(claims count), Gamma (claim amount)
- Non-Parametric Statistics: no probability model specified
  - classification trees, NN

Predictive Modeling Process



Optimize the Value of Data

Financial & Socioeconomic Data

- Property Ownership
- Income & Wealth
- Banking Data
- Financial Credit

Medical Data

- Prescription Histories
- Laboratory Profiles
- Third-Party Data (MIB)
- Physician Reports/EHR

(Potential) Customer

Insurance Data

- Application Disclosures
- Underwriting Data
- Policy values
- Claim/Lapse Experience

Behavioral & Lifestyle Data

- Consumer Spending
- Driving Violations
- Hobbies & Interests
- Physical Activity

Build Models & Make Predictions

- ✓ Likelihood to buy new product
- ✓ Probability to qualify for new product by SI/GI
- ✓ Likelihood to lapse
- ✓ Likelihood to be fraudulent
- ✓ …and more

Opportunity

Sales & Marketing

- Customer response modeling ("propensity to buy/renew")
- Recommendations
- Agent recruiting, quality assessment & monitoring

Risk Selection / Risk Scoring

- Predictive Underwriting
- Underwriting triage
- Risk quantifying & risk segmentation
- Improve placement rates

Pricing / Product Development

- More pricing variables & more accurate
- Better incorporated interaction
- Price optimization
- More accurate formula-driven assumptions

Experience Analysis

- Mortality, lapse, incidence, etc
- True multivariate approach
- Efficient use of data
- Handle low-credibility data
- Create assumptions

In-force Policy Management

- Customer retention model "propensity to lapse/persist
- Customer lifetime value

Claims Administration

- Claim risk scoring
- Claims triage
- Fraudulent claims
- Rescinded claims

Case Study 1: UW Model

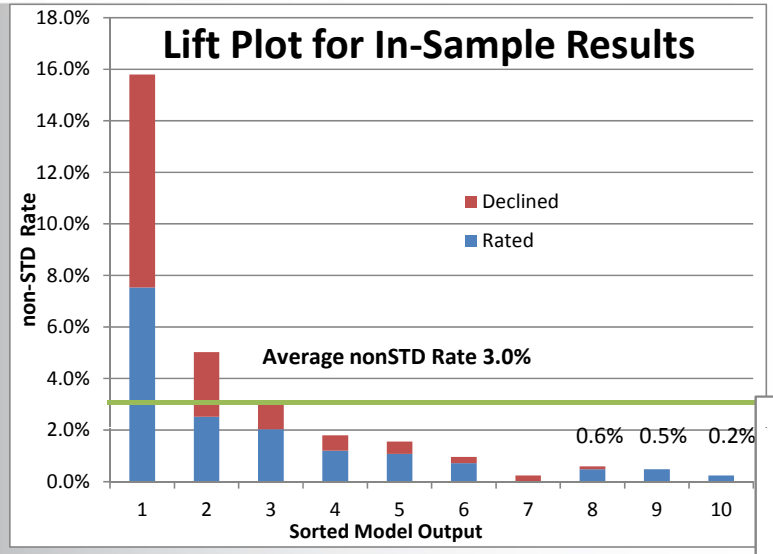
- Goal: to predict UW decisions on its existing customers
  - Bancassurance in Asia with full UW life products, but low penetration
  - ✓ Identify certain pre-qualified existing customers, & offer guaranteed issue (GI) or simplified issue (SI) without medical UW
  - ✓ Acquisition costs will be significantly reduced
  - ✓ Market penetration will be deeper, and sales will increase
- Bancassurance is unique for PM
  - ✓ Financial/demographic information about customers
- Major challenge - very limited data
  - ✓ A total of about 8k-9k full UW cases
  - ✓ Target variable UW decision, with very low declined/rated cases, ~3.0%
  - ✓ Many missing values due to old time, especially for sub-STD
  - ✓ Not all information collected at the time of UW

Key Variables

- GLM with binomial and logistic link function
- Model uses about a dozen of variables that are statistically significant for prediction and readily available in client database
- Here are a few key predictor variables
  - ✓ “Positive” means the probability to be STD increases if the value goes up; otherwise, it is ”Negative”

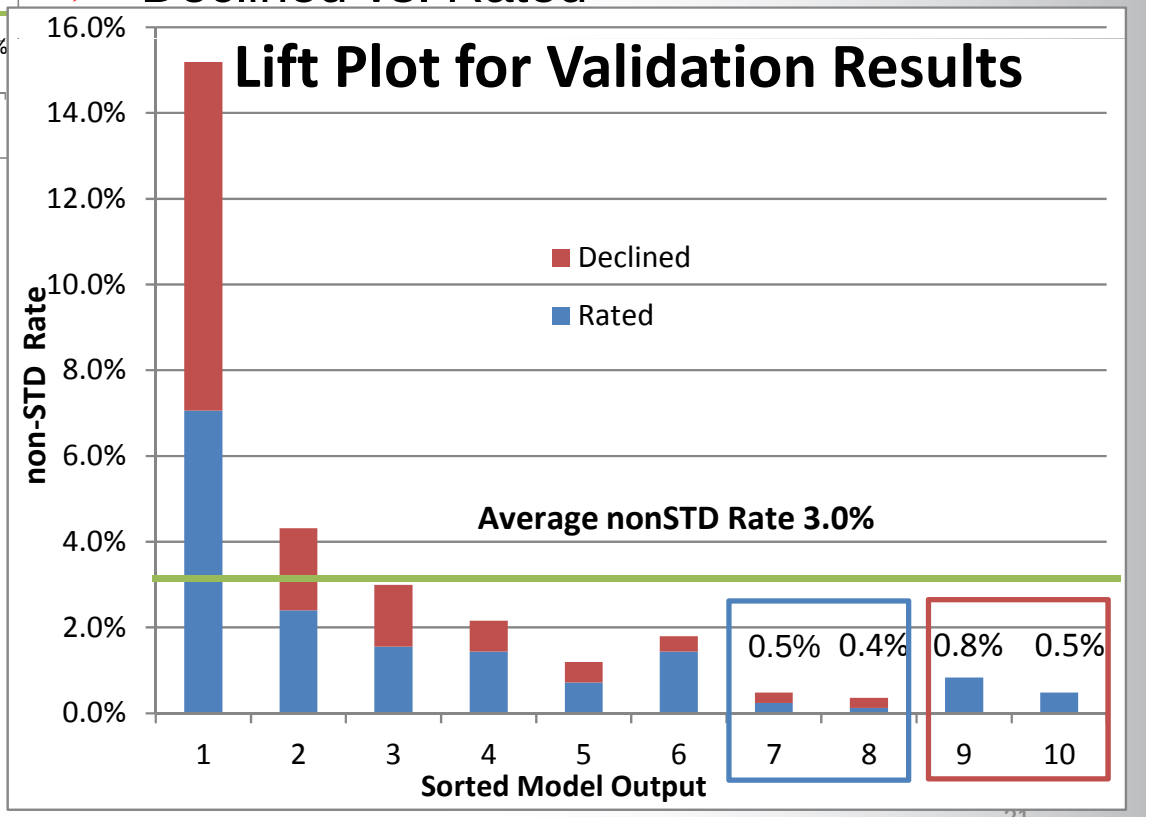
Name	Type	Note
Age_At_Entry	Numeric	Negative; less likely to qualify for STD as age goes up
Branch	Categorical	Proxy of geographic locations
AUM	Numeric	Positive; more likely to qualify for STD with large AUM
Customer_Segment	Categorical	Positive for Premier, negative for non-Premier
Nationality	Categorical	Positive for domestic; negative for certain others

Model Results



- Validation results are a better test of model performance in real business
- 0.6% sub-STD in the top 30% of model outputs, 80% reduction compared to random 3%
- Declined vs. Rated

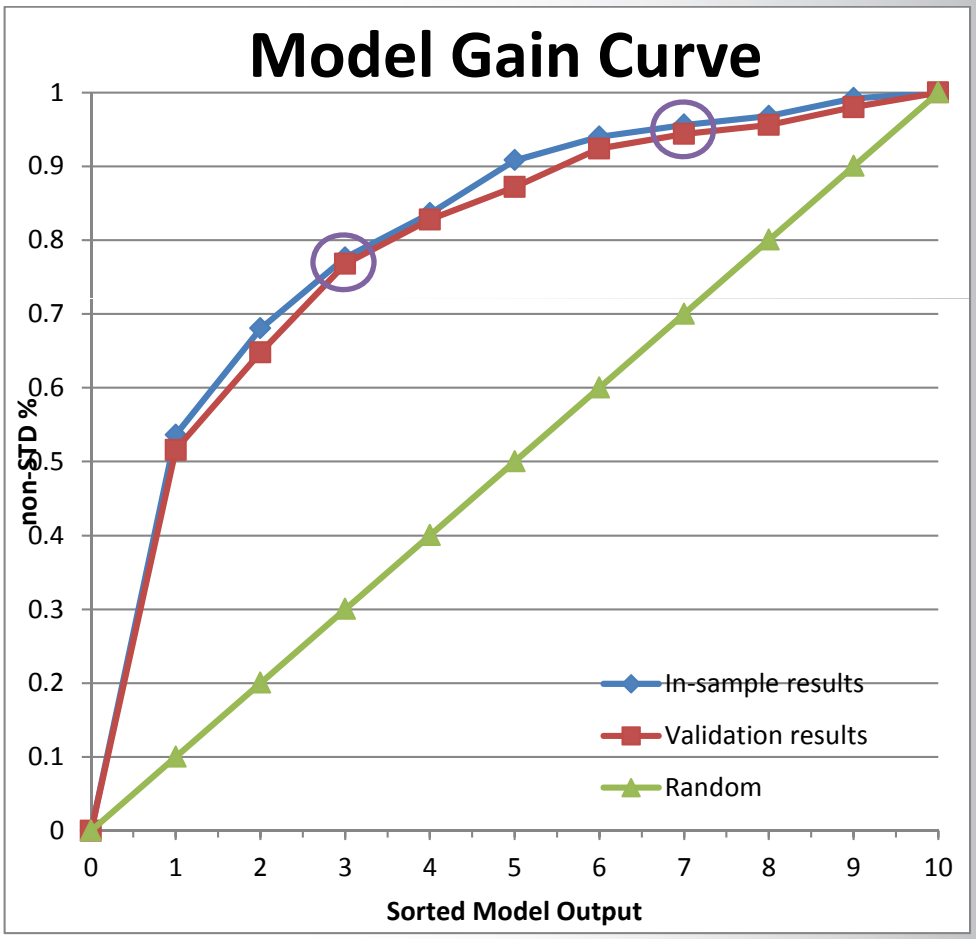
- In-sample results show model performance under optimal condition
- May have over-fitting issue
- 0.5% of sub-STD in top 30% of model output



Model results

- Another way to understand model capability to differentiate STD from sub-STD
- Best 30% of model outputs contains about 5% of total non-STD
- Lowest 30% captures about 75% of bad risks

- Current Status
  - ✓ Model results have been delivered to the client
  - ✓ Final implementation stage



Case Study 2: Medical Product Upsell

Objectives

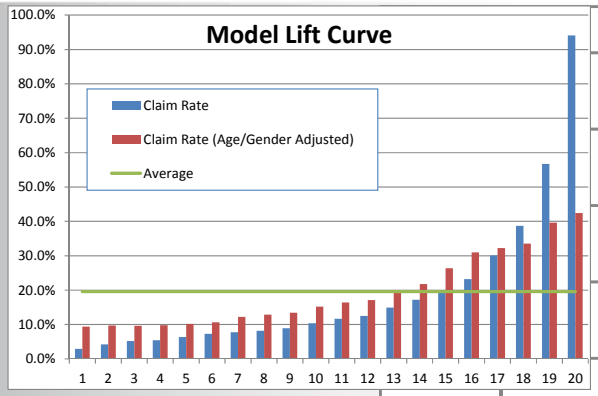
- Increase sales by upselling new PHI products to a very large portion(>50%) of the in-force medical policyholders with reduced underwriting.
- By using predictive modeling to significantly simplify the underwriting approach; current UW process is intrusive, long and expensive
- Make the sales process simpler and quicker for both customers and agents thereby reducing acquisition expenses and maximize sales.
- Increase take-up by reducing underwriting requirements for the best risks.

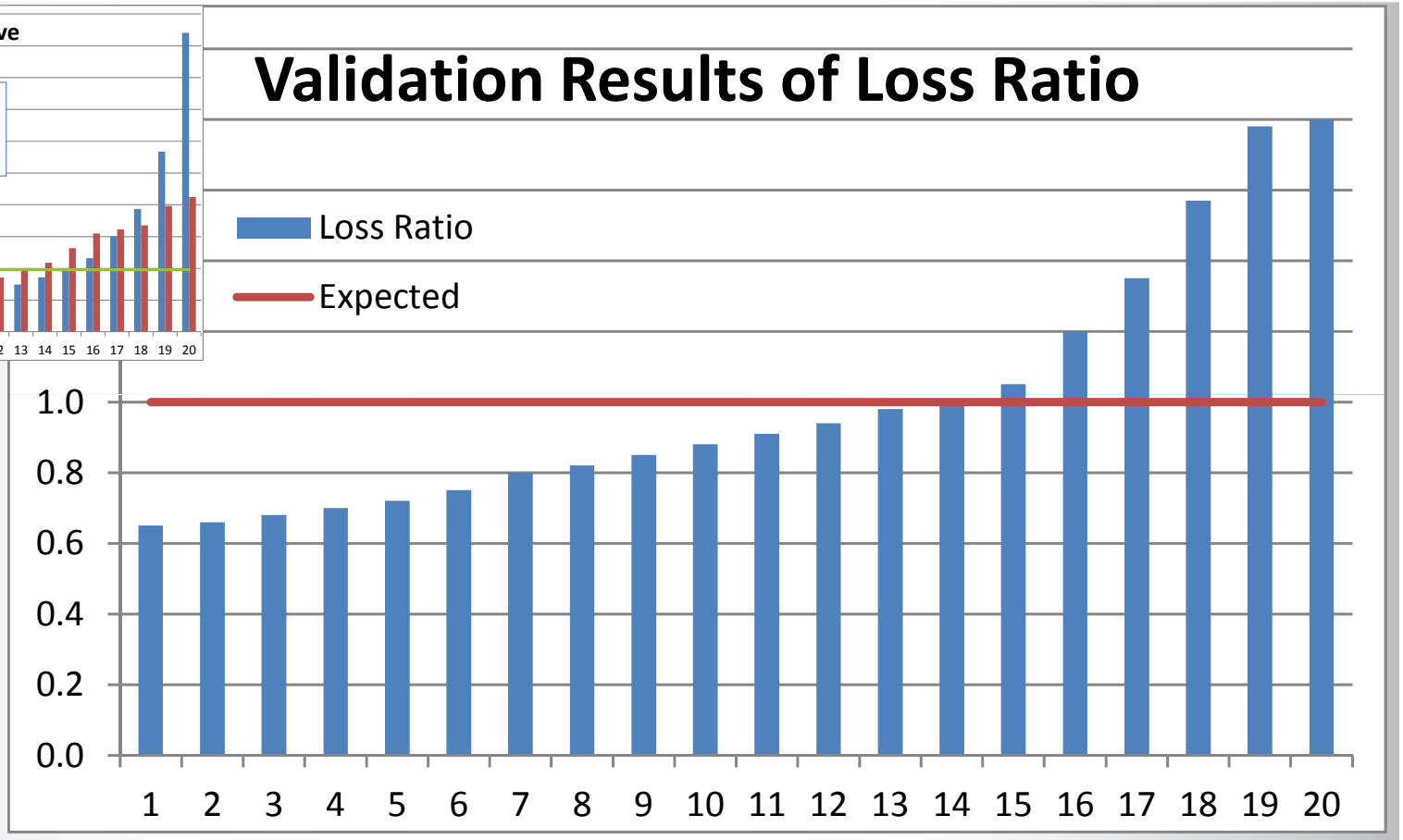
Data Analysis

- Large data set with detailed individual policyholder profile and claims information (>2m base policies and associated claims; >4m riders with claims).
- Use demographic, socio-economic and policy information as predictor variables with exposure/claims history as the target variables
- Use customer risk profile to determine the expected claim amount or loss ratio that at in-force medical policyholder will incur in the future.
- Use model validation to estimate how accurately our predictive model will perform in practice.

23

Medical Product Upsell





Model Results

- Validation results reflect the predictive power of the model in real business

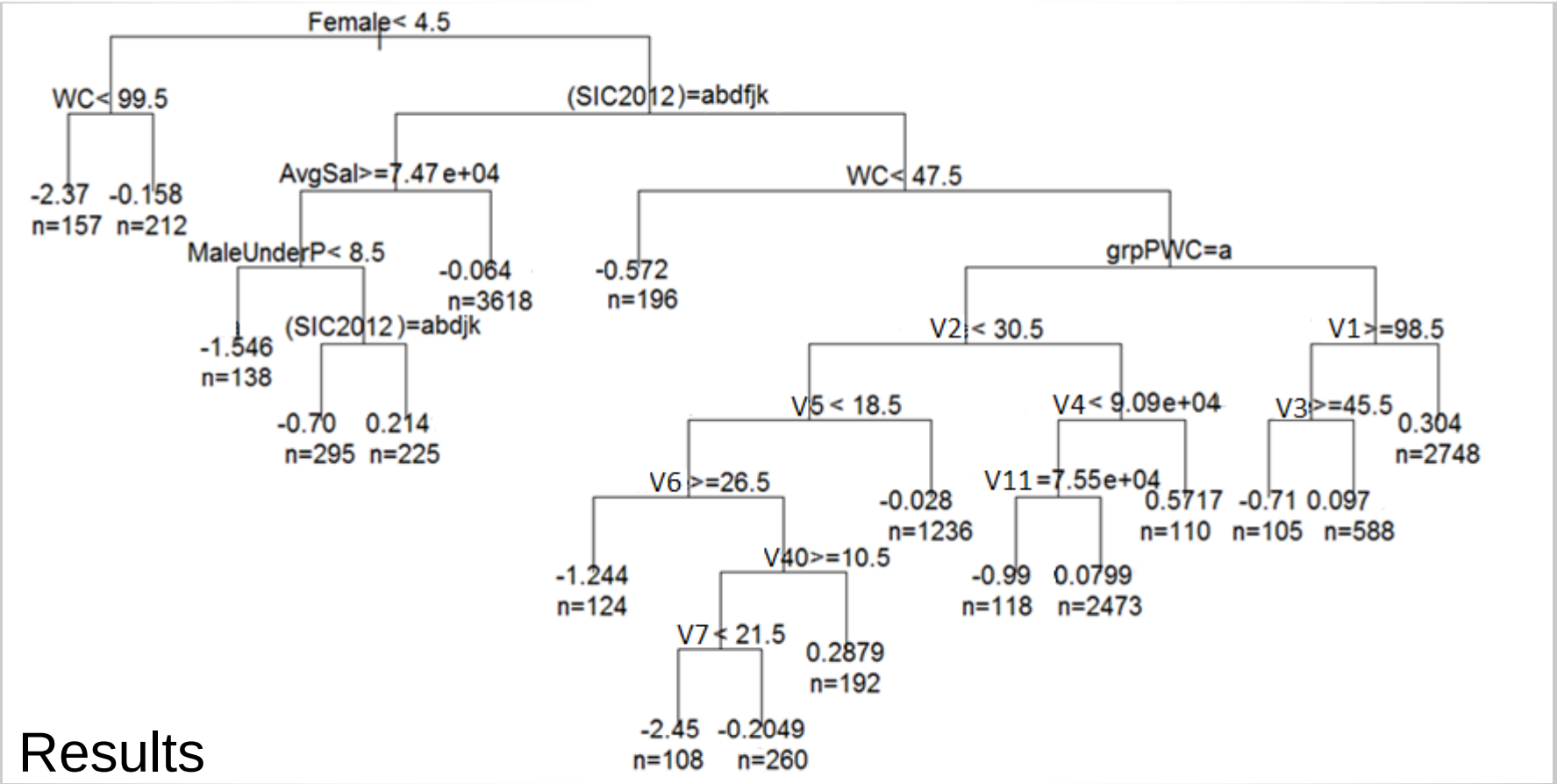
- By selecting the best risks, the UW requirements will be significantly reduced from 3-4 pages questionnaire to one single question.
- The worst 20% of model output has an average LR more than 70% higher than expected; the best 20% will have a expected LR at about 65%

Sample results presented to protect client confidentiality

Case Study 3: LTD Pricing

- Business: US group Long-Term Disability(LTD)
  - ✓ About 13k policies, with lives per policies from 10 to 30k
  - ✓ Current pricing variables: about 30-40
  - ✓ Experience data of past 5 years with >80 variables
  - ✓ Major pricing variables: age, gender, industry, location, benefit structure
- Objective
  - ✓ To determine additional pricing variables and possible interaction terms (for pricing)
  - ✓ To identify groups with experience deviating from pricing assumptions (for UW)
- Client has experience with PM
  - ✓ Minimum efforts on business & data understanding
  - ✓ CART model

CART Model results



Results

- ✓ Easy to develop, interpret and understand; business insights
- ✓ Not efficient for linear function; sensitive to noise; over-fitting

CART Model results

- Results improve profit margin and pricing accuracy
  - ✓ Useful tool for both pricing and UW of group LTD business
- Model implementation
  - ✓ Client is very interested in model results; approved by management team
  - ✓ Implemented in Q2'13

Quartile	# of cases	Actual EPM	Model Predicted EPM
1	3230	(0.28)	(0.32)
2	3230	(0.088)	(0.060)
3	3230	0.063	0.020
4	3230	0.017	0.14

Case Study 4: Risk Segmentation

- Foreign travel increased exponentially in past 50 years; associated risk impacted on life insurance
- There are mortality/morbidity differences between countries; location of residence can make large difference in mortality
- Objective
  - Assess foreign travel & residence risk
  - Compare all countries around the world on uniform basis
  - Data-driven conclusions based on facts and data, not popular opinion & preconceptions

Risk Segmentation

Life Expectancy (years)

Rank	Country	Life Expectancy	Rank	Country	Life Expectancy
1	Monaco	89.6	204	Chad	49.1
2	Japan	84.2	203	South Africa	49.5
3	Singapore	84.1	202	Guinea Bissau	49.5
4	San Marino	83.1	201	Swaziland	50.0
5	Andorra	82.6	200	Afghanistan	50.1
6	Switzerland	82.3	199	Central African Republic	50.9
7	Hong Kong	82.2	198	Somalia	51.2
8	Australia	82.0	197	Zimbabwe	51.5
9	Italy	82.0	196	Namibia	52.0
10	Liechtenstein	81.6	195	Gabon	52.2

Infant Mortality (deaths before age 1 per 1,000 live births)

Rank	Country	Infant Mortality	Rank	Country	Infant Mortality
1	Monaco	1.8	189	Malawi	77.0
2	Japan	2.2	190	Burkina Faso	78.3
3	Bermuda	2.5	191	Angola	81.8
4	Singapore	2.6	192	Niger	88.0
5	Sweden	2.7	193	Chad	91.9
6	Hong Kong	2.9	194	Guinea Bissau	92.7
7	Iceland	3.2	195	Central African Republic	95.0
8	Italy	3.3	196	Somalia	101.9
9	France	3.3	197	Mali	106.5
10	Spain	3.4	198	Afghanistan	119.4

Risk Segmentation

- Data for all 205 countries/regions w/ 25 fields

  Life Expectancy(1), Maternal Mortality(2), Infant Mortality(3), Underweight Children(4), Adult Obesity(5), HIV Prevalence(6), Communicable Disease Death Rate(7), Physician Density(8), Sanitation(9), Drinking Water(10), Hospital Beds(11), Traffic(12), Homicide(13), Military Conflicts(14), Foreign Deaths(15), Occupational Accidents(16), Carbon Dioxide(17), Particulate Matter concentration(18), Internet Users(19), Mobile Phone(20), Road Density(21), GDP Per Capita (PPP)(22), Corruption(23), Education-Expected Years of School(24), Gini Index(25)

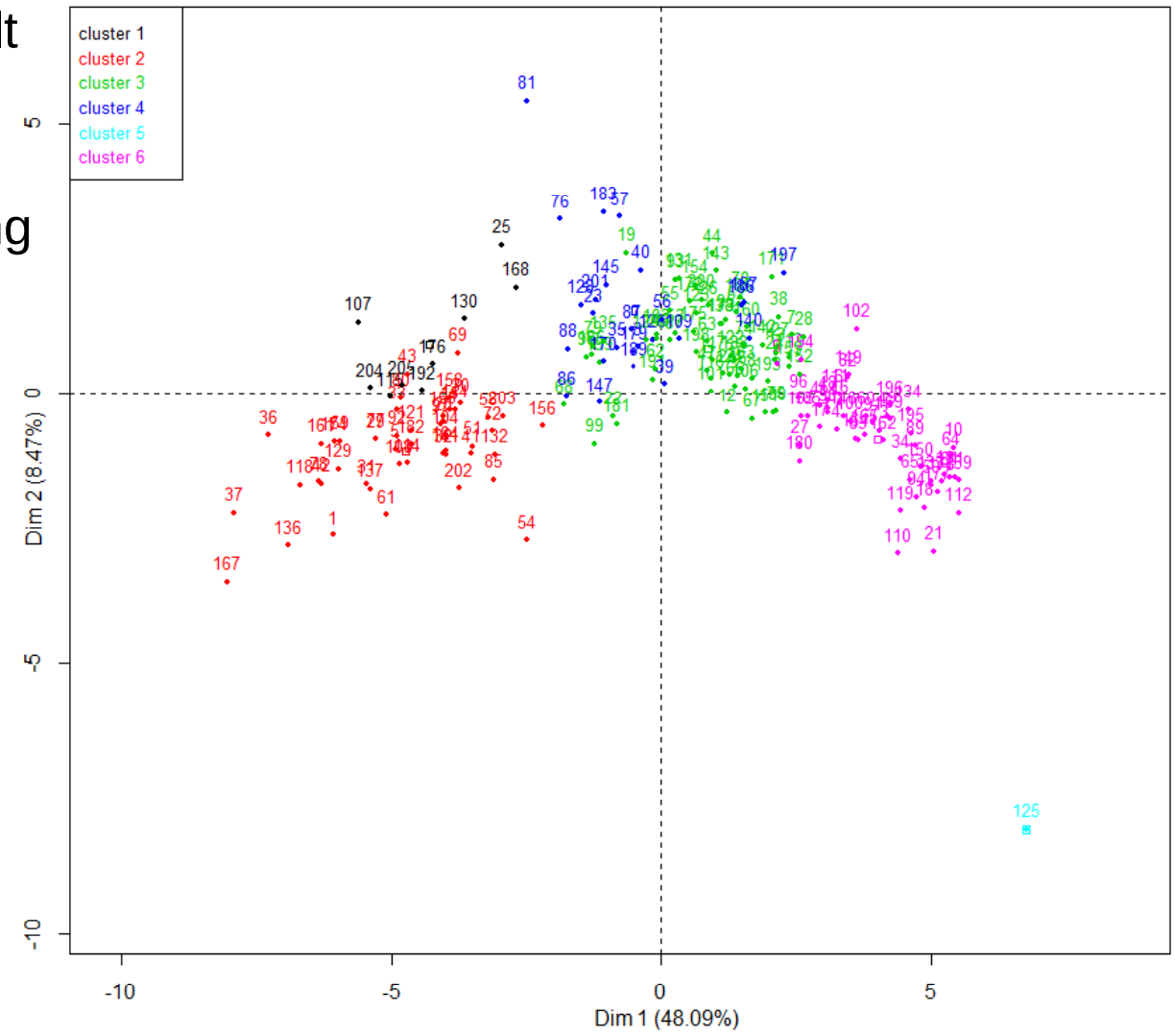
- Data sources

  CIA, WHO, World Economic Forum, World Bank, UN, Center for Systemic Peace, U.S. State Department, pueblo.gsa.gov, Elsevier, Transparency International

- Main challenges
  - Many missing values
  - Different weights on certain fields, e.g. life expectancy

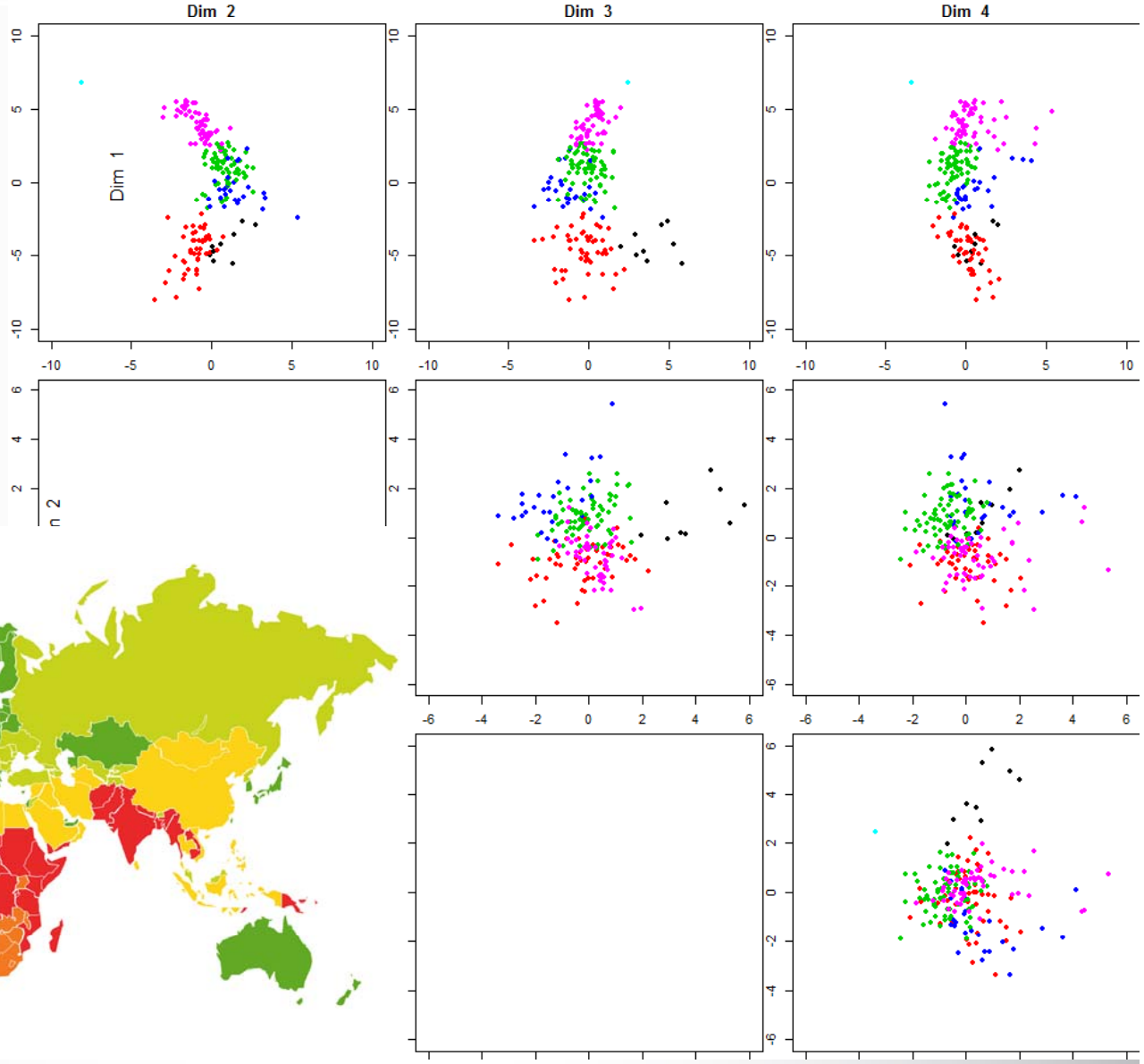
Risk Segmentation

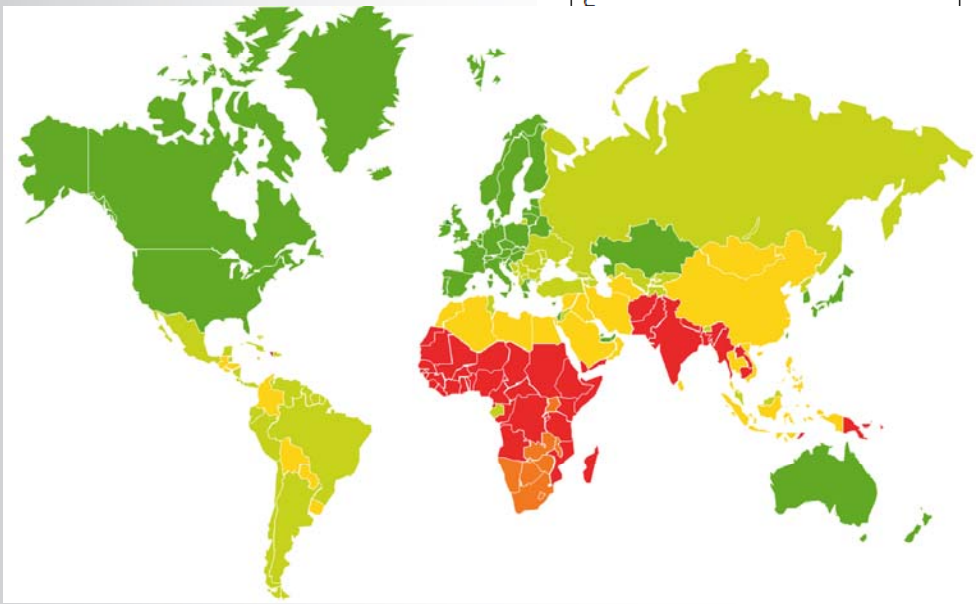
- Missing values are dealt with at algorithm level
- PCA analysis followed by hierarchical clustering
  - ✓ Principle Component Analysis – explain variance in data
  - ✓ Weights are based on judgment, and considered at hierarchical clustering
- Results on 6 clusters
  - ✓ Number of clusters is a free parameters



Risk Segmentation

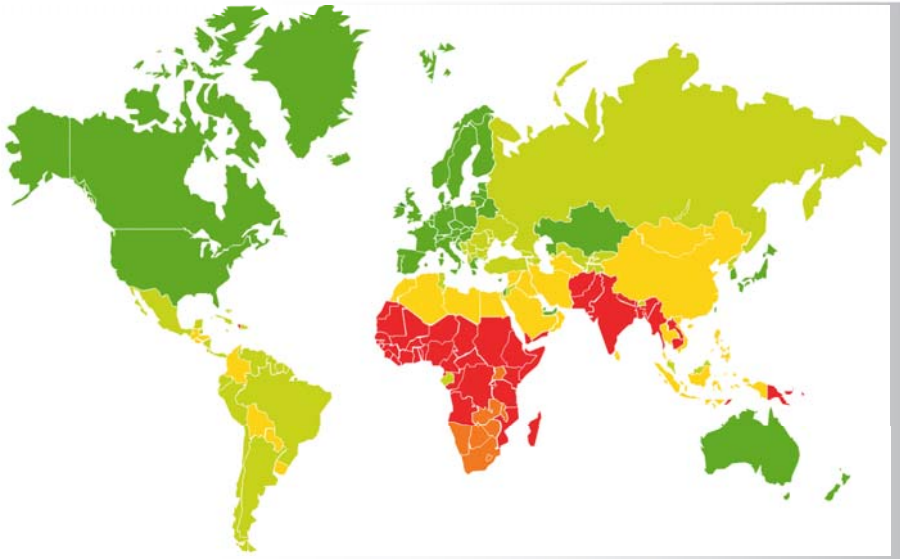
- Data visualization
  - Scatter plot of first 4 components
  - World map for relative risk index





Risk Segmentation

- Data visualization
  - Examples of first 4 components
- Major PCA analysis followed by hierarchical clustering
  - ✓ Principle Component Analysis
- Results on 5 clusters
  - ✓ Number of clusters is a free parameters



Predictive Modeling

And Its Application in Insurance

Richard Xu, **PhD FSA**

VP & Actuary, Head of Data Science

Global R&D, RGA

June, 2015

India