Session 3

Using Predictive Analytics to Solve Business Problems

Yi-Ling Lin, FSA, MAAA
What’s the Buzz is all About...

Many terms used interchangeably... confusing... exciting... and seem magical

**Big Data**
- Five “V”s...
- Structured/unstructured data...
- Traditional data...
  ...to provide value need...

**Data Analytics**
- Spectrum by function, business value and sophistication
  ...uses traditional and emerging algorithms...

**Statistical methods and machine learning**
- ML is part of AI, deep learning is part of ML
- Mathematically defined “learning” architecture plus optimization framework...
Applying Data Analytics to Business Problems

Spectrum of data analytics: hindsight to insight to foresight

Adapted from Gartner’s Data Analytics Maturity Model

Application of Data Analytics in Health

Current and emerging applications of data analytics in US healthcare

- Dashboards, descriptive statistics, and reporting
- Trends exploration and forecasting
- Pricing
- Claims reserving
- Risk scoring, risk stratification, and risk adjustment
- Plan design modeling
- Stress testing
- Care management and decease prediction
- Fraud detection and outliers analysis
- Assumptions setting

... and many more
Models and Business Problems

Data analytics toolbox: spectrum of traditional to emerging methodologies

- Disease progression
- Claims reserving
- Survival/ Markov models
- Generalized linear regression/logit
- Clustering/classification
- Time Series
- Deep Learning
- Fraud identification
- Targeted marketing
- High cost group stratification
- Provider referral patterns
- Trend forecasting
- Stress testing
- Trend forecasting
- Stress testing
- Signal processing
- Text processing

New Programming Paradigm: Machine Learning
- Humans input data & answers
- And how to “learn”… and what does it mean to be wrong…
- Example: clustering algorithm or neural networks or decision tree/Random Forest

Classical programming model
- Humans input data and set of rules/function on how to arrive at answers
- Also how close they want data to fit to the “model”...
- Example: linear regression or generalized linear regression

Traditional Modeling versus Machine Learning

Could computer automatically learn the rules by looking at data?
Model Evaluation and Validation

Model evaluation is an important part of any modeling project

- Relevance and importance of criteria
- Appropriate and consistent with purpose
- On "unseen" or "test" sample of data
- Examples of criteria/metrics
  - Standard statistical measures (R squared, RMSE, MAE, etc.)
  - Predictive Ratios on groups of interest (e.g. Diagnostic groups or age groups)
  - Tolerance curves
  - Sensitivity and specificity (confusion matrix)
  - ROC curves
  - Comparison with naïve and standard models
  …and many more

Cautionary tale!
Famous Anscombe's quartet: all four datasets have the same statistical properties, including R squared=0.82, means and variance of x and y, correlation and linear regression model: y=3+0.5x

Challenges

Exciting things often come with challenges and potential pitfalls

- Messy, often high-dimensional with missing values, data and data quality issues
- Potential bias in data
- Use of proxies
- Non-discrimination, security and confidentiality
- Transparency vs. “black box”
- Spurious correlations: correlation vs. causality
- Interpretability and replicability
- Overfitting and overreliance
- Business purpose appropriateness and applicability
  … and many more

Stories from the front lines...

- Quality of data. NYC taxi story: if it's too good to be true... it's probably isn't!
- Bias inherited from humans through training data (training facial analytics on male white faces?)
- Spurious correlations... (divorce rate in Maine correlates with per capita margarine consumption at 99.3%)
Risk Scoring Modeling in Healthcare

Application of data analytics in healthcare for risk scoring

Risk scoring = using individuals’ data to “predict”/estimate outcome of an event

Used in many areas: cost predictions, mortality hazard ratios, probability of disease
Credit scores, healthcare risk scores, mortality risk scores, life underwriting…

In healthcare risk scoring (cost-based) in US moving from traditional to emerging

- Methodologies: from linear regression models to machine learning algorithms
- Data: traditional claims and enrollment data (Rx, Dx, demographic, prior year costs, etc.) to new and emerging data, e.g. socio-economic status factors
- Types of risk scoring models: concurrent vs. prospective; off-the-shelf vs. custom

Health risk scores are used for variety of purposes

Many sources of uncertainty
Case Study: Descriptive Analytics

Dashboards, distributions, descriptive statistics

This is traditional analysis to inform what actually happened and the first step in any modeling project

In this case study babies under age of 2 were excluded, and population shown were enrolled at least for one month in both years

Case study for illustration purposes only

Average PMPM (Medical and Rx) by year and gender

Case Study: Diagnostic Analytics

Investigating and identifying trends & relationship

Diagnostics focused on uncovering patterns, relationships, trends, and potentially engineering predictive features

Claim costs are lognormally distributed; prior and current somewhat linearly correlated (0.54)

Case study for illustration purposes only
Case Study: Comparing Custom and Off-the-shelf

Great improvement on accuracy and predictive ratios

On Test Data:
- $R^2$ is 0.24, correlation 0.65, MAE=67% for off-the-shelf HCC
- $R^2$ is 0.48, correlation 0.70, MAE = 73% for residual custom-recalibrated HCC

Case study for illustration purposes only

Case Study: Decision-informing Analytics

Various uses of risk scoring in health care: population health and care management

Identifying best cases for care management

Assess characteristics of high-risk/low-cost group: potential for care management

From low to high

Prevention and wellness programs
Case Study: Health Plan Choice Modeling

An employer group wants to change the medical plans it offers to employees. Important to “predict” who will select various plans for financial and other modeling.

<table>
<thead>
<tr>
<th>Current Options</th>
<th>Actuarial Value</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMO</td>
<td>0.90</td>
<td>10%</td>
</tr>
<tr>
<td>High Value</td>
<td>0.86</td>
<td>42%</td>
</tr>
<tr>
<td>Medium Value</td>
<td>0.81</td>
<td>45%</td>
</tr>
<tr>
<td>CDHP</td>
<td>0.76</td>
<td>2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Options</th>
<th>Actuarial Value</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan A</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Plan B</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Plan C</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Plan D</td>
<td>0.68</td>
<td></td>
</tr>
</tbody>
</table>

Participants data → Modeling → Current choices/utility → Probability of selecting new options
Subject Matter Expertise is Critical

Feature engineering is often informed by subject matter expert

<table>
<thead>
<tr>
<th>ID</th>
<th>Medical Coverage</th>
<th>Coverage Tier</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Medium Value Option Employee + 1 Dependent</td>
<td>90043</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Medium Value Option Employee Only</td>
<td>95066</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Medium Value Option Employee Only</td>
<td>98053</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Medium Value Option Employee Only</td>
<td>60630</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>HMO Plan Employee Only</td>
<td>95121</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Medium Value Option Employee Only</td>
<td>33472</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>HMO Plan Employee + 1 Dependent</td>
<td>94610</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Medium Value Option Employee Only</td>
<td>94109</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Medium Value Option Employee Only</td>
<td>60478</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>HMO Plan Employee + 1 Dependent</td>
<td>94607</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Medium Value Option Employee + 1 Dependent</td>
<td>11377</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Medium Value Option Employee Only</td>
<td>94587</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Medium Value Option Employee Only</td>
<td>94109</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Medium Value Option Employee Only</td>
<td>94606</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Medium Value Option Employee Only</td>
<td>94133</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Medium Value Option Employee Only</td>
<td>60532</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Medium Value Option Employee + 2 or More Dependents</td>
<td>95635</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>HMO Plan Employee + 1 Dependent</td>
<td>91206</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Medium Value Option Employee Only</td>
<td>02458</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Medium Value Option Employee Only</td>
<td>94103</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Medium Value Option Employee Only</td>
<td>98121</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Medium Value Option Employee Only</td>
<td>60614</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Medium Value Option Employee Only</td>
<td>10604</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>HMO Plan Employee Only</td>
<td>94121</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Medium Value Option Employee Only</td>
<td>94122</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Medium Value Option Employee Only</td>
<td>95356</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Medium Value Option Employee Only</td>
<td>33026</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Medium Value Option Employee Only</td>
<td>94107</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Medium Value Option Employee Only</td>
<td>94123</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>HMO Plan Employee + 2 or More Dependents</td>
<td>94501</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Medium Value Option Employee + 2 or More Dependents</td>
<td>94595</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Medium Value Option Employee + 1 Dependent</td>
<td>32703</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Medium Value Option Employee Only</td>
<td>90046</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Medium Value Option Employee Only</td>
<td>91367</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Medium Value Option Employee Only</td>
<td>28173</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>HMO Plan Employee Only</td>
<td>95691</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Medium Value Option Employee Only</td>
<td>10573</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Medium Value Option Employee Only</td>
<td>94115</td>
<td></td>
</tr>
</tbody>
</table>

Data Visualizations and Feature Engineering

Data visualization helps with feature selection, model evaluation and much more...
Process of Predicting and Evaluating Choice

What are predicted costs for individuals and plan?

- Age / Stage in Life
- Risk Tolerance
- Premium / Contributions
- Expected Claims
- Plan Design

Individual Plan Elections → Individual Annual Cost → Total Employee Cost Sharing for the individual

Plan Cost

Case Study: Modeling Approach

Estimating Parameters

Heterogeneous Logit Model

- \( i \) – individuals
- \( j \) – plan options
- \( k \) - \# of attributes with weights \( \beta_k \)
- \( U_{ij} \) – utility of plan option \( j \) to person \( i \)
  \[ U_{ij} = X_{ij} \beta + \epsilon_{ij} \]
  \[ \beta_k = \beta_k + \sigma_k \mu_k \]
- Monte Carlo simulation and maximize log likelihood function
Case Study: Modeling New Choices

Probability of participants’ elections of options A, B, C, or D

Age / Stage in Life
Risk Tolerance
Premium / Contributions
Expected Claims
Plan Design

New choices
Know preferences ($\beta$ and $\sigma$) and now changing the attributes ($X$)

Logit Model

Case Study: Choice Modeling Results

Use results to inform business decisions and identify areas for further study

<table>
<thead>
<tr>
<th>Current Options</th>
<th>Actuarial Value</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMO</td>
<td>0.90</td>
<td>10%</td>
</tr>
<tr>
<td>High Value</td>
<td>0.86</td>
<td>42%</td>
</tr>
<tr>
<td>Medium Value</td>
<td>0.81</td>
<td>45%</td>
</tr>
<tr>
<td>CDHP</td>
<td>0.76</td>
<td>2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Options</th>
<th>Actuarial Value</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan A</td>
<td>0.87</td>
<td>57%</td>
</tr>
<tr>
<td>Plan B</td>
<td>0.81</td>
<td>22%</td>
</tr>
<tr>
<td>Plan C</td>
<td>0.75</td>
<td>17%</td>
</tr>
<tr>
<td>Plan D</td>
<td>0.68</td>
<td>5%</td>
</tr>
</tbody>
</table>

In this case study, people are more likely to choose a higher value plan option regardless of their current plan.

Plan A (97% AV) Plan B (81% AV) Plan C (75% AV) Plan D (68% AV)
Questions? Thoughts... Comments?

SOCIETY OF ACTUARIES®