PREDICTIVE MODELLING FOR COMMERCIAL INSURANCE

General Insurance Pricing Seminar
13 June 2008
London

James Guszcza, FCAS, MAAA
jguszcza@deloitte.com
General Themes
Predictive modelling: 3 Levels of Discussion

- **Strategy**
  - Profitable Growth
  - Right-pricing
  - Improved retention ...

- **Methodology**
  - Model design (actuarial)
  - Modelling process (modern machine learning POV)

- **Technique**
  - GLM vs classification trees vs neural networks ...
Methodology vs Technique

- Technique is only one facet of overall methodology.

- It’s not enough to be statisticians – we must be *actuarial* statisticians.

- **How does predictive modelling need actuarial science?**
  - Variable creation
  - Model design
  - Model validation

- **How does actuarial science need predictive modelling?**
  - Advances in computing, modelling techniques
  - Ideas from other fields can be applied to insurance problems
Semantics: Data Mining vs Predictive Modelling

- **Data Mining**: “knowledge discovery”, often in large industrial databases – “KDD”
  - Data exploration techniques (some brute force)
  - Data visualization
  - e.g. discover strength of credit variables

- **Predictive Modelling**: Application statistical techniques (like GLM) after knowledge discovery phase is completed.
  - Quantify & synthesize relationships found during KDD phase
  - e.g. build a credit model
Aside:
A Famous Example of KDD in Insurance

- Mid-90’s: insurers discovered a strikingly powerful relationship between personal credit score and personal motor / homeowners claim propensity.

- The reason “why” was (is?) mysterious.

- The discovery – and the business benefit – did not hinge on particularly advanced statistical techniques.

- A dramatic illustration of the business value of the data mining / KDD paradigm.

- KDD is “fact-finding”.

Copyright © 2008 Deloitte Development LLC. All rights reserved.
Commercial Insurance vs Personal Insurance

- Personal insurance modelling is a “nice” statistical problem.
  - Many data points
  - Straightforward exposure base (car-year)
  - Many well understood pricing factors
  - In the UK’s liberal market especially, prices can be determined scientifically
    - GLM-based loss cost modelling
    - Elasticity modelling, price optimisation
    - Controlled pricing experiments
Commercial Insurance vs Personal Insurance

- Commercial insurance modelling is a “messy” statistical problem.
  - Fewer data points – especially for new business
  - Often lower frequency / higher severity
  - Heterogeneous risks
    - The corner bakery vs the suburban über-market
  - Complex exposure bases (sales, payroll, feet$^2$)
  - Messy data
  - Risk selection/pricing often a “free for all”
  - *Underwriter Subjectivity*
Strategy:
Why Undertake a Modelling Project?
The Parable of Moneyball  
(Or:  How Underwriting is Like Baseball)

- In 1999 Billy Beane (manager of the Oakland Athletics) found a novel use of data mining.
  - A’s not a wealthy team: ranked 12th (out of 14) in payroll
  - How could the A’s compete with the rich teams?

- Beane hired a junior statistician (Paul dePodesta) to analyze statistics advocated by baseball guru Bill James.

- Using predictive analytics, Beane was able to hire excellent players undervalued by the market.
  - A year after Beane took over, the A’s ranked 2nd!
The Implication

- Beane *quantified* how well a player would do.
  - Not perfectly, just better than his peers
  - *He realized that statistical regularities are more reliable than baseball scouts’ subjective, expert judgments.*

- Implication:
  - Be on the lookout for fields where an expert is required to reach a decision based on judgmentally synthesizing quantifiable information across many dimensions.
  - *(Does this sound like commercial insurance underwriting?)*
  - *Maybe a predictive model can beat the human expert.*
Mental Accounting

- Take a guess: which is a worse EL risk?... and by how much?

<table>
<thead>
<tr>
<th>Flower shop</th>
<th>Pub</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 employees</td>
<td>10 employees</td>
</tr>
<tr>
<td>5 year-old business</td>
<td>15 year-old business</td>
</tr>
<tr>
<td>2 EL claims in past 5 years</td>
<td>Most recent EL claim: 4 years ago</td>
</tr>
<tr>
<td>Credit: 70th %ile</td>
<td>Credit: 90th %ile</td>
</tr>
</tbody>
</table>

- Unlike a human decision-maker, a predictive algorithm “knows” how much weight to give each consideration.
  - Just as the A’s used models to select players, commercial insurers use models to select and price risks.
  - Humans are “predictably irrational” …
    … but models don’t engage in “creative mental accounting”.

Copyright © 2008 Deloitte Development LLC. All rights reserved.
### Keeping Score

<table>
<thead>
<tr>
<th>Billy Beane</th>
<th>CEO who wants to run the next Progressive Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beane’s Scouts</td>
<td>Commercial Insurance Underwriters</td>
</tr>
<tr>
<td>Potential Team Member</td>
<td>Potential Policyholder</td>
</tr>
<tr>
<td>Bill James’ stats</td>
<td>Innovative collection of predictive variables</td>
</tr>
<tr>
<td>Billy Bean’s Super Cruncher</td>
<td>You and me</td>
</tr>
</tbody>
</table>
The Moral of Our Parable

- Billy Beane has arguably transformed US professional sports by introducing the strategic use of predictive analytics to baseball.
  - The way Beane crunched his numbers was determined by his business strategy:
    - Exploit an inefficient and subjective market for baseball players.

- Similarly in the commercial insurance domain:
  - Start off by trying to understand the business/strategic context.
  - *Allow the modelling strategy to conform to the business strategy, not vice versa.*
Competing on Analytics

- In “Competing on Analytics”, Tom Davenport defines:
  - “An analytic competitor [is] an organization that uses analytics extensively and systematically to outthink and out-execute the competition.”
  - Think of predictive modelling as a strategic capacity… not just another actuarial tool.

- The most valuable modelling projects are an integral part of a company’s core strategy.
More Business Considerations

- Davenport: truly analytic competitors promulgate an “analytic” and “fact-based” culture from the top down.
  - A related point: culture change is often a critical part of implementing a predictive model.
  - A model can be worse than nothing if it is implemented improperly and/or if critical users do not buy into it.

- Building models is only a one phase of a “predictive modelling” project.
  - Planning, data scrubbing, project management, IT implementation, business implementation often dwarf the modelling part of the project.
  - Modelling is the fun part, not the hard part!
  - Highly multi-disciplinary process.
Methodology:
Integrating Concepts from Statistics, Actuarial Science, Machine Learning
Concepts from Modern Statistics

- Generalized Linear Models
- Goodness-of-fit measures – $R^2$, AIC, BIC, …
- Nested models, analysis of deviance, $F$-tests, …
- Graphical analysis of model fit
- Graphical residual analysis
- Variance estimators
- Bayesian credibility
- Bootstrapping, simulation

(...you know the drill)

- But these doesn’t exhaust modern “predictive modelling”
Concepts from Modern Machine Learning

- Data Mining and KDD
  - Brute-force search techniques

- Scoring engines
  - A “predictive model” by any other name

- Lift Curves
  - *Operationally meaningful* measure of “predictive power”

- Out-of-sample model tests, cross-validation
  - Ideally yield unbiased estimates of “predictive power”
  - Alternative to AIC, BIC
Scoring Engines

- Scoring engine: (non)linear function of multiple predictors:
  \[
  \text{score} = f(X_1, X_2, \ldots, X_N)
  \]
- Used for segmentation.
- The \( X_1, X_2, \ldots, X_N \) are as important as the \( f() \)
  - Major reason why actuarial expertise is necessary.
- A large part of the modelling process consists of variable creation and selection
  - Often possible to generate 100’s of variables
  - Steepest part of the learning curve
  - Data scrubbing / variable creation is time-consuming
Model Evaluation – the Lift Curve

- Sort data by model score
- Break the dataset into 10 equal pieces
  - Best “decile”: lowest score → lowest LR
  - Worst “decile”: highest score → highest LR
  - Difference: “Lift”
- Lift = segmentation power
- Lift → ROI of the modelling project
Out-of-Sample Model Validation

- Randomly divide data into 3 pieces
  - Training data, Test data, Validation data

- Use Training data to fit models

- Score the Test data to create a lift curve
  - Perform the train/test steps iteratively until you have a model you’re happy with
  - Test data is implicitly used in building the final model
    - test lift is overly “optimistic”
  - During this iterative phase, validation data is set aside in a “lock box”

- Once model has been finalized, score the Validation data and produce a lift curve
  - Unbiased estimate of future performance
Credit Scoring is a Classic Example

- All four of our machine learning concepts apply to Credit Scoring.
  - Knowledge discovery in databases (KDD)
  - Scoring engine
  - Lift Curve evaluation $\rightarrow$ translates to LR improvement $\rightarrow$ ROI
  - Blind-test validation

- Credit scoring has been the insurance industry’s segue into the modern synthesis of classical statistics with machine learning concepts.
  - Very useful paradigm in the context of commercial insurance modelling.
Concepts from Actuarial Science

- Overall design of model / analysis
  - What are we trying to predict? At what level?

- Predictive variable creation
  - Calls on subject-matter expertise of insurance

- Target variable creation
  - Loss development and trending
  - Whether/how to use premium
  - Deductibles, claim/claimant level, etc …
  - Considerations of time periods

- Analysis file creation
  - “Level” of the analysis – risk, policy, account, …
  - Inclusions / exclusions
What are we Trying to Predict?

- Pricing: Pure Premium
- Underwriting: Profitability
- Premium audit: Additional / returned premium
- Retention models
- Cross-sell models
- Elasticity models
- Agent/agency profitability
- Target marketing
- Fraud detection

Again… the modelling strategy should follow the business strategy.
  - No one-size-fits-all answer
Variable Creation

- Research possible data sources
- Extract/purchase data
- Check data for quality (QA)
  - Messy! (we are still toiling deep in the data mines)
- Create Predictive and Target Variables
  - Opportunity to quantify tribal wisdom
  - …and come up with new ideas
  - Can be a very big task!
- Steepest part of the learning curve
Types of Predictive Variables

- Behavioral
  - Prior claims, bill-paying, credit ...

- Policyholder
  - Business class, age, # employees ...

- Policy specifics
  - Number of buildings, Construction Type ...

- Territorial
  - Geo-demographic, economic, weather ...
Data Exploration & Variable Transformation

- 1-way analyses of predictive variables
  - Weed out weak / redundant variables

- Correlation study of predictive variables
  - Avoid multicollinearity – further weeding out

- Exploratory Data Analysis (EDA)
  - Advanced techniques can be helpful
  - Data Visualization very helpful here

- Use EDA to cap / transform predictive variables
  - Extreme values, missing values, etc
Modeling Process

1. Finalize set of transformed predictive variables

2. Iterative training / testing of candidate models
   - Build candidate models on “training data”
   - Evaluate on “test data”
   - Many things to tweak
     - Different target variables
     - Different predictive variables
     - Different modelling techniques
     - # NN nodes, hidden layers; tree splitting rules; tuning parameters …

3. Select & validate final model
   - Use as-yet untouched validation data

Copyright © 2008 Deloitte Development LLC. All rights reserved.
Some Pragmatic Considerations

- Do signs / magnitudes of parameters make sense? Statistically significant?

- Is the model biased for/against certain types of policies? Regions? Policy sizes? Business classes? ...
  - If so, is that an appropriate thing, or not?

- Predictive power holds up for larger policies?

- Continuity
  - Are there small changes in input values resulting in large score swings?
  - Could an agent or underwriter “game” the model?
Model Analysis & Implementation

- Perform model analytics
  - Necessary for client to gain comfort with the model

- Calibrate Models
  - Create user-friendly “scale” – client dictates

- Implement models
  - **Technical**: IT skills are critical here
  - **Business**: *Culture change* can be critical

- Monitor performance
  - Distribution of scores over time, predictiveness, usage of model...
  - Plan model maintenance
Technique: Regressions and its Relations

Artificial Neural Networks
MARS
CART
Regression and its Relations

- **GLM**: relaxes some regression assumptions
  - Assume linearity on link function scale
  - Variance is *modeled* as a function of expected value

- **MARS & Neural Networks**
  - Clever ways of *automatically* transforming and *interacting* input variables
  - Why: sometimes the “true” relationships aren’t linear
  - Universal approximators: model any functional form

- **CART is simplified MARS**
Uses of “Advanced” Techniques

- Alternatives to GLM models
- Provide benchmarks for GLM models
- Exploratory data analysis (especially CART)
- Variable selection
- Detection of interaction terms
- Detection of optimal variable transformations
A neural net models $Y$ as a complicated non-linear function of $X$.

- **Lingo**
  - Green: “input layer”
  - Red: “hidden layer”
  - Yellow: “output layer”

- The $\{a, b\}$ numbers are “weights” to be estimated.

- The network architecture and the weights constitute the model.
Neural Networks: Functional Form

\[ Z_1 = \frac{1}{1 + e^{a_{01} + b_{11}x_1 + b_{21}x_2 + b_{31}x_3}} \]

\[ Z_2 = \frac{1}{1 + e^{a_{02} + b_{12}x_1 + b_{22}x_2 + b_{32}x_3}} \]

\[ Y = \frac{1}{1 + e^{b_0 + b_1z_1 + b_2z_2}} \]

- These look like logit models.
- NN is thus related to GLM.
MARS

- **Multivariate Adaptive Regression Splines**
- Automatically searches a space of “basis functions” for the right combination to model complex, multi-dimensional, non-linear patterns.
- Basis functions look like “hockey sticks”
- MARS model is a linear model of hockey sticks and interactions of hockey sticks.
- Cross-validation is built into the core MARS algorithm.

**Linear model offers a poor fit**

**MARS considers basis function transformations**
MARS Result

- MARS performs a stepwise search and the prunes back.
  - Cross-validation is used to determine optimally complex model.

- The final MARS model is:
  \[ y^* = 0.29 + 0.02*x - 0.086\max(0,x-35) + 0.084\max(0,x-65) \]

- This is a GLM model!
  - A more complex example would have multiple variables and interactions.
CART: Recursive Partitioning

- Classification And Regression Trees
- Key idea: recursive partitioning
  - Take all of the data.
  - Consider *all* possible values of *all* variables.
  - Select the variable/value \((X=t_i)\) that produces the greatest “separation” in the target.
  - \((X=t_i)\) is called a “split”.
  - If \(X< t_i\) then send the data to the “left”; otherwise, send data point to the “right”.
  - Now repeat same process on these two “nodes”.

- You get a tree-structured model.
- As with MARS, cross-validation is used to “prune back”.

Commercial Insurance Example

- Suppose you have 3 variables:
  - # vehicles: \{1,2,3...10^+\}
  - Age category: \{1,2,3...6\}
  - Liability-only: \{0,1\}

- At each iteration, CART tests all 15 splits.
  - (#veh<2), (#veh<3),..., (#veh<10)
  - (age<2),..., (age<6)
  - (lia<1)

- Select split resulting in greatest increase in purity.
  - Perfect purity: each split has either all claims or all no-claims.
  - Perfect impurity: each split has same proportion of claims as overall population.

- Then iterate – grow the tree out... then prune back
Example of a Split

- Commercial Auto Dataset
  - 57,000 policies
  - 34% claim frequency

- Predict likelihood of claim
  - Classification Tree using Gini splitting rule

- First split:
  - Policies with ≥5 vehicles have 58% claim frequency
  - Else 20%
  - Big increase in purity
Bringing it All Back Home

- Remember that a **MARS** model is a GLM model fit on basis-function-transformed variables.
  - ... as well as interactions thereof

- A **CART** model is like a **MARS** model in which the “hockey stick” basis functions are replaced with \{0,1\} step functions.
  - “tree-structured regression”

- Thus – like **MARS** and **NNET** models – **CART** models are relatives of regression models.
  - “Only connect.” – E.M. Forster
References

For Beginners:

Data Mining Techniques
--Michael Berry & Gordon Linhoff

For Mavens:

The Elements of Statistical Learning
--Jerome Friedman, Trevor Hastie, Robert Tibshirani
PREDICTIVE MODELLING FOR COMMERCIAL INSURANCE

General Insurance Pricing Seminar
13 June 2008
London

James Guszcza, FCAS, MAAA
jguszcza@deloitte.com