Predictive Modelling: GLM vs Machine-Learning

Guanjun Jiang
Principal & Consulting Actuary
Milliman Limited
Agenda

- Introduction of Predictive Modelling
- Generalised Linear Model (GLM)
- Machine Learning (Eagle Eye Analytic)
- Case Study
- Summary
PREDICTIVE MODELLING

—— PREDICTIVE MODELLING IS THE PROCESS BY WHICH A MODEL IS CREATED OR CHOSEN TO TRY TO BEST PREDICT THE PROBABILITY OF AN OUTCOME.
Some Insurers……

Are happy doing what they have always done……

Others are happy being better and bigger than some……
Key Areas of Business Interaction

Knowledge gained results in competitive advantage through:

- Greater client satisfaction and retention
- Better risk selection
- Granular, targeted pricing
- More effective marketing
Usages of Predictive Modelling in Insurance

- Underwriting cycle management
- Profitability Analysis
- Reinsurance optimization
- M&A post-transaction analysis

- Target marketing
- New business acquisition
- Retention management
- Agency management

- Claims routing and prioritization
- High risk identification
- Loss control
- Reserve projection and estimation

- Accuracy & adequacy
- Competitiveness
- Adverse Selection
- Customer view
- Efficiency
GENERALISED LINEAR MODEL
Brief Introduction of GLM

- **Basic Structure:**
  
  $g(\mu) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e_i$

- **Y** --- $n \times 1$ Vector (measured), belonging to Exponential Family (Poisson, Gamma, Normal, Binomial, Inverse Gaussian, Negative Binomial, Tweedie)
- **Var(Yi) = f (E[Yi])**
- **\eta** --- \eta = X\beta
- **\beta** --- $p \times 1$ Vector (to be estimated)
- **X** --- $n \times p$ (Design Matrix)
- **g** --- Link Function
GLM: What is a Good Model?

- Consistent over time and withstand random sampling tests
- Strikes a balance between fitting well and over-fitting the data
- Various measures and tests can be done using a combination of:
  - AIC/BIC
  - Residual plots
  - Cramer’s V - test the correlation of two categorical factors
  - Deviance
  - Chi-square
  - Confidence interval of fitted values for each factor
  - Gini
GLM: Revealing the Risk Shape

Claim Frequency

- Actual
- "Model"
- LCL
- UCL

Exposure

- Actual
- Model
GLM Results: Does the Curve Fit?

GLM Output

Driver’s Age

Less than 24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65+

6+ Years
Less than 6

Relativities
0.8  0.9  1.0  1.1  1.2  1.3  1.4

Less than 24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65+

6+ Years
Less than 6
GLM Results: Does the Curve Fit?

Empirical Experience

<table>
<thead>
<tr>
<th>Driver’s Age</th>
<th>Relativities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 24</td>
<td>1.3</td>
</tr>
<tr>
<td>25-29</td>
<td>1.2</td>
</tr>
<tr>
<td>30-34</td>
<td>1.1</td>
</tr>
<tr>
<td>35-39</td>
<td>1.0</td>
</tr>
<tr>
<td>40-44</td>
<td>0.9</td>
</tr>
<tr>
<td>45-49</td>
<td>0.8</td>
</tr>
<tr>
<td>50-54</td>
<td>0.9</td>
</tr>
<tr>
<td>55-59</td>
<td>1.0</td>
</tr>
<tr>
<td>60-64</td>
<td>1.1</td>
</tr>
<tr>
<td>65+</td>
<td>1.2</td>
</tr>
</tbody>
</table>

6+ Years
Less than 6
MACHINE LEARNING
What is Machine Learning

- “a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases.” Wikipedia

Addresses the fundamental questions directly

- Where are we making money?
- Where are we losing money?
- Can we be confident?

Identifies risk segments that are credible and produce consistent results from year to year

- Iterative, artificial intelligence process
- User defines the degree of credibility within segments

Results are only as good as the algorithm

- A good algorithm will maximise the number of segments identified
### Technology/Modern Statistical Techniques is the Differentiator...

<table>
<thead>
<tr>
<th>Current Methods (like GLM/GAM)</th>
<th>Machine Learning (like Ensembles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Assumes that variables are independent <em>unless specifically defined otherwise</em></td>
<td></td>
</tr>
<tr>
<td>✓ “Optimal” predictors are based on assumptions</td>
<td></td>
</tr>
<tr>
<td>✓ Can’t solve what you don’t know</td>
<td></td>
</tr>
<tr>
<td>✓ The number of risk attribute/value interactions is too large for a human to investigate given real-world resource and time constraints, therefore only a very small subset is investigated</td>
<td></td>
</tr>
<tr>
<td>✓ Pricing models are done at a coverage level versus a customer level</td>
<td>✓ Allows data to interact naturally to find the patterns between characteristics within the data</td>
</tr>
<tr>
<td></td>
<td>✓ Finds the trade-off between over- and under-fitting automatically</td>
</tr>
<tr>
<td></td>
<td>✓ Does not require the user to specify the predictors and interactions to be included in the model - it discovers them!</td>
</tr>
<tr>
<td></td>
<td>✓ Extremely Fast and Efficient</td>
</tr>
<tr>
<td></td>
<td>✓ Performed at coverage, unit, or policy level</td>
</tr>
</tbody>
</table>
EEA Segmentation Analysis Types

Typical Uses:
- Rate plan improvement
- Underwriting rules
- Target marketing

- Partitions the whole “universe” into exhaustive and mutually exclusive segments
- Available model responses: loss ratio, pure premium, frequency, severity, profit, retention
- Segments:
  - Described by significant attributes
  - Plain English description, easy to understand and actionable
  - “Complex” compound variables
Pricing & UW: Find Errors

We found 60% of the exposures in their technical premiums had pricing errors greater than 10%.
  - Underpricing errors of up to 54%
  - Overpricing errors of up to 34%

The difference: Talon’s learning algorithms are designed specifically for insurance data.
Price & UW – How Talon finds the Errors

Private Passenger Auto
(Total Portfolio Loss Ratio = 71%)

**Best Customers**
- Loss Ratio = 35%
  - No: 59%
  - Yes: 87%
  - Tenure: 76%
  - Male: 74%
  - Female: 58%

**Worst Customers**
- Loss Ratio = 150%
  - Tenure: 90%
  - Safe Driver Discount: 63%
  - Passive Restraint: 62%
  - Unmarried Drivers: 83%
  - Females: 75%
  - Min Driver Age: 78%
  - Vehicle Age: 60%

Identify New Patterns in the Data

Some of the best customers are overpriced
*Unique Pattern*: Combining
- Safe Driver Discount
- 10 year old policy, or older
- No Passive Restraints produces lowest loss ratio of 35%, 36 points lower than carrier average.

Some of the worst customers are Underpriced
*Unique Pattern*: Combining 5 unique variables, including customer tenure, marital status and vehicle age, identifies unprofitable business with loss ratios at 2x carrier average.
Pricing - Main Concept for Telematics

- Rapid Pricing Diagnostics using Machine Learning:

Combine regular policy pricing with Telematics data analysis*:

This needs new modeling technology!

Price Difference = Loss Ratio = Telematics Claims / TP (Telematics)

Technical Price = TP (Standard Policy)

*Such analysis cannot be done with classical methods like GLMs because
a) Cost effecting, complex interactions within the Telematics data can only be detected automatically (through Machine Learning)
b) The price difference cannot be fitted by a GLM-Distribution
### Pricing – Machine Learning for Telematics

Auto Telematics Product  
(Total Portfolio Loss Ratio = 106%)

<table>
<thead>
<tr>
<th>Best Clients</th>
<th>Worst Clients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Loss Ratio</strong></td>
<td><strong>Loss Ratio</strong></td>
</tr>
<tr>
<td>&lt;20T</td>
<td>&gt;=20T</td>
</tr>
<tr>
<td>45%</td>
<td>90%</td>
</tr>
<tr>
<td>&lt;75%</td>
<td>&gt;=75%</td>
</tr>
<tr>
<td>74%</td>
<td>48%</td>
</tr>
<tr>
<td>&lt;100</td>
<td>&gt;=100</td>
</tr>
<tr>
<td>52%</td>
<td>72%</td>
</tr>
<tr>
<td>#Trips/Year</td>
<td>#Trips/Year</td>
</tr>
<tr>
<td>Daylight rides</td>
<td>Daylight rides</td>
</tr>
<tr>
<td>48%</td>
<td>74%</td>
</tr>
</tbody>
</table>

#### Identify very profitable and unprofitable segments

Some of the best customers might be overpriced

**Unique Pattern:** Combining
- High Mileage
- Mostly Day light
- Many trips produces lowest loss ratio of 38%,

Some of the worst customers are underpriced and might be unexpected from their univariate patterns and can lie close to their good counterparts!
CASE STUDY

FROM MODELS TO RESULTS
# China Motor Tariff

## Rating Factor

<table>
<thead>
<tr>
<th>Description</th>
<th>Factor</th>
<th>Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>With designated driver</td>
<td>C1a</td>
<td>0.9</td>
</tr>
<tr>
<td>Undesignated driver</td>
<td>C1b</td>
<td>1</td>
</tr>
<tr>
<td>Younger than 25 years old</td>
<td>C2a</td>
<td>1.05</td>
</tr>
<tr>
<td>[25,30)</td>
<td>C2b</td>
<td>1</td>
</tr>
<tr>
<td>[30,40)</td>
<td>C2c</td>
<td>0.95</td>
</tr>
<tr>
<td>[40,60)</td>
<td>C2d</td>
<td>1</td>
</tr>
<tr>
<td>At least 60 years old</td>
<td>C2e</td>
<td>1.05</td>
</tr>
<tr>
<td>Male</td>
<td>C3a</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>C3b</td>
<td>0.95</td>
</tr>
<tr>
<td>Less than 1 year</td>
<td>C4a</td>
<td>1.05</td>
</tr>
<tr>
<td>[1,3)</td>
<td>C4b</td>
<td>1.02</td>
</tr>
<tr>
<td>At least 3 years</td>
<td>C4c</td>
<td>1</td>
</tr>
<tr>
<td>Domestic</td>
<td>C5a</td>
<td>1</td>
</tr>
<tr>
<td>Within province</td>
<td>C5b</td>
<td>0.95</td>
</tr>
<tr>
<td>Routine</td>
<td>C5c</td>
<td>0.92</td>
</tr>
<tr>
<td>Less than 30,000 km/p.a.</td>
<td>C6a</td>
<td>0.9</td>
</tr>
<tr>
<td>(30000,50000) km/p.a.</td>
<td>C6b</td>
<td>1</td>
</tr>
<tr>
<td>At least 50,000 km/p.a.</td>
<td>C6c</td>
<td>1.1-1.3</td>
</tr>
<tr>
<td>No liable traffic ticket record in previous year</td>
<td>C8a</td>
<td>0.9</td>
</tr>
<tr>
<td>Liable traffic ticket record (s) in previous year</td>
<td>C8b</td>
<td>1</td>
</tr>
</tbody>
</table>

## Private Vehicle

### Own Damage

<table>
<thead>
<tr>
<th>Seats</th>
<th>Fixed premium</th>
<th>Rate (%)</th>
<th>Fixed premium</th>
<th>Rate (%)</th>
<th>Fixed premium</th>
<th>Rate (%)</th>
<th>Fixed premium</th>
<th>Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 1 year</td>
<td>539</td>
<td>1.28</td>
<td>513</td>
<td>1.22</td>
<td>508</td>
<td>1.21</td>
<td>523</td>
<td>1.24</td>
</tr>
<tr>
<td>1-2 years</td>
<td>646</td>
<td>1.28</td>
<td>616</td>
<td>1.22</td>
<td>609</td>
<td>1.21</td>
<td>628</td>
<td>1.24</td>
</tr>
<tr>
<td>2-6 years</td>
<td>646</td>
<td>1.28</td>
<td>616</td>
<td>1.22</td>
<td>609</td>
<td>1.21</td>
<td>628</td>
<td>1.24</td>
</tr>
<tr>
<td>6+ years</td>
<td>646</td>
<td>1.28</td>
<td>616</td>
<td>1.22</td>
<td>609</td>
<td>1.21</td>
<td>628</td>
<td>1.24</td>
</tr>
</tbody>
</table>
Significant Improvement on the Tariff
A Southern China Branch

Loss Ratio Lift: 69.6-145.9% = 2.1x
Significant Improvement on GLM
A Southern China Branch

- High loss ratio means GLM underpriced and vice-versa
- Low loss ratio means GLM over-priced
Segmentation Result Drill Down

Worse Segment
Segmentation Result Drill Down
Worse Segment

- Worse Segment here means GLM has Under Priced the risk

<table>
<thead>
<tr>
<th>Segment</th>
<th>Branch2</th>
<th>Driver Age</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Over 20 branches in this segment!</td>
<td>18 to 35 (inclusive)</td>
<td>&gt; RMB 100,000</td>
</tr>
</tbody>
</table>

Segment 1 (Worst Seg) → Make (Toyota) → cc (2200-2498)

- Vehicle Age
- NCD
- Driver Age

All Ages Poor
Most exposure in 0%-10% NCD
Most ages but particularly 25-29
Improvement on GLM – Using ML Results

- Introduced new interactions based
- Introduced new rating variable

Results
- ✓ AIC Improved
- ✓ BIC Improved
- ✓ Gini Improved
- ✓ Chi-square – just as good
- ✓ Deviance Improved
Recap

- Need for Predictive Modelling is Today
- GLM is a robust pricing approach

BUT

- Machine Learning will
  - Tackles GLM’s shortcomings
  - Identify critical hidden “gems” and “pitfalls”
  - Speed up the model build process systematically
The Value

**Most Predictive Signal**
- Lift curves of 2x-4x or more over other methods
- 4-6 way or more data interactions
- Non-linear interactions
- Local effects
- High correlations, over 90%

**Fast**
- Hundreds of iterations produced in a few hours
- Results in 60-90 days
- Real-Time Scoring Service supports real time decision-making

**Actionable**
- Understandable Segments & Scores
- Forward looking Management tools for Enterprise-wide application
- Approved rate filings in regulated markets
Why EagleEye Analytics?
From the perspective of clients

Most powerful and actionable predictive signal
After a failed attempt at getting a different, larger multivariate software provider to produce results specific to our company, we abandoned them. With Talon we have already implemented the model results and are seeing the changes come to fruition.”

Speed to business impact allowing for real time excellence
Talon is extremely fast and efficient. It allows us to process analyses in a matter of minutes or hours. We now have the ability to quickly implement and maintain a sustainable competitive advantage.”

Complete vision
“EagleEye Analytics” solution suite gives us a common platform from which to dialogue regarding analytics and business performance throughout the enterprise into such areas as pricing, underwriting, claims and marketing. We now have a common and robust analytical foundation being used across our entire portfolio by multiple constituencies.

Proven results
We correctly determined that the cost of not utilizing EagleEye’s solution suite was too great to ignore. It is the most innovative, unique and powerful approach to driving profits, avoiding adverse selection, and improving our competitive advantage.”
Questions?