Big Data in Practice, and the challenges of Big MI
Darren McCauley, Chief Underwriting Officer, Tesco Underwriting
Richard Bretton, Pricing Director, AXA Direct and Partnerships

Common experiences

IT Prioritisation  Customer expectations  “Believability”
Correlations  Model disconnects  Personal Lines (plus?)
Expectation management  Stakeholder management

What is big data?
• Definition
• A data set whose size is beyond the ability of databases on data management tools to
  – Capture
  – Store
  – Manage
  – Analyse
Three Categories of big data problems

• Volume
• Velocity
• Variety

Why look at big data

• Quote data explosion
  – Phone
  – Internet
  – Aggregation
  – Enrichment
So is our data big data?

… no not really yet!

What did we do?

• Lots of reading and research

• Lots of conferences
  – The first things the conferences tell you is this is really hard

  … its really not if you work with the right people

Sandbox exercise

• What am I talking about

• What did we do

• What did we learn
Problem Articulation

- What did we need to achieve?

- Six key data requirements
  - We were not looking to mine social media data
  - Nothing here was difficult about designing the reports
  - Nothing earth shattering in terms of data required

  ... but was by far the most important phase of the project

Impact on Impact Analyses

- Hopefully we've all progressed through a programme of increasingly predictive modelling and built confidence in ability to predict impacts of pricing actions

- New factor introduction means one cannot create fully populated quote batches or future renewal batches on which to model. Usually we only have a subset of populated historic data on which to rely.

- Moreover, behavioral/competitiveness models will not contain the new factors

- So ... We need to stakeholders to be:
  - sufficiently convinced of the benefits to go live with lower certainty of outcome
  - OR
  - patient enough to invest in the wiring then wait to build live understanding

- We also need to avoid creation of an industry of trying (in vain) to tailor analyses to every new data source’s particular correlations/peculiarities

- We must remain aware that our models have high statistical variance, yet we wish to rely on their central estimates more than “gut feel” of stakeholders, whilst not over-engineering analyses relative to these inherent uncertainties
The timing factor

- We also like to put sales data back into our risk models to give an indication of net effects of rating actions on profitability. We are missing "lag data" for past periods. Even with backfill, it's ordinary to see the same data we would have collected at the time it presumably didn't have.

A focus is brought to the proper treatment of "unknowns" in modelling. Reasons for unknown levels change, and the volumes coming through in future may not be the same. With more uncertainly than in the past, the new data must be more carefully considered. Most of the time we assumed the renewal data, but with more "lag", it's unlikely to be the same data we would have attached at the time (it possibly didn't exist!)

- Issues we overcame
  - Communication

- Supplier selection and contracts
  - Final solution required five suppliers, and eight contracts

- Internal challenges
  - IT skill sets
  - Internal IT
  - Executive and Board
What did we achieve?

- Speed
- Curiosity
- Innovation
- Data sources
- Data accessibility

What did we achieve that we weren’t expecting

- Control of data throughout organisation
  - Standardisation of definitions
  - Reconciled data
  - Standardisation of reports
- Much more data was available for analysis
  - Portal data
  - Sub peril data
  - Fraud

What does it allow

- Two routes in:
  - traditional SAS coding access
  - Visual Analytics tool
- In memory analytics on
  - 25m quotes
  - 3m policies
  - >400k claims
Living with the consequences

- There will be no "prevailing" mix of new factor conversion/retention rates from which to anchor views of appropriate trading KPIs. Any mix-stability MI is rendered immediately pretty useless. It takes time for conversion/retention and especially LR to evolve.
- The usual trap of wanting all the benefits of the new models but the mix from old (underperforming?) rates is heightened when deploying multiple new factors.
- Therefore: We need tools to communicate unusual performance versus inherent noise to stakeholders.

Illustration of loss ratio variability for 500k earned vehicle years (even with heavy capping)

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Quote exposure</th>
<th>Conversion</th>
<th>Exception level</th>
<th>Exception breach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor level 1</td>
<td>3.7%</td>
<td>3.23%</td>
<td>2.50%</td>
<td>29%</td>
</tr>
<tr>
<td>Factor level 2</td>
<td>1.4%</td>
<td>4.35%</td>
<td>3.48%</td>
<td>25%</td>
</tr>
<tr>
<td>Factor level 3</td>
<td>12.1%</td>
<td>1.34%</td>
<td>1.11%</td>
<td>21%</td>
</tr>
</tbody>
</table>

Since we've been fighting against the temptation to unnecessarily fiddle with rates, we should already have this.

Finally, if currently adjusting conversion levels based on market info about general quote to sale ratios to get a more insightful MI pack, the difficulties of enriching external quote data with our new factors means we find ourselves back in a distorted position.

Questions

Expressions of individual views by members of the Institute and Faculty of Actuaries and its staff are encouraged.

The views expressed in this presentation are those of the presenter.