B.A.U. for actuaries: Big data, Analytics & Unstructured data

Big Data Working Party
Mudit Gupta (Chair), Colin Priest, David Menezes, Frank Devlin, Frankie Chan, Xavier Conort

SAS – GI Conference 2015
Goals of the Big Data Working Party

To explore the future of big data, analytics and unstructured data in Asia and what actuaries need to do to have the right skillsets that will be in demand for such work.

- Introduction at the GI conference and develop case study
- Workshop on machine learning using R
- CPD sessions
Part I

Introduction to Big data
What is big data?

- Often used to describe large volume of data being collected by organizations
- Lack of structure
Where does data come from?

Internal systems
- Policy Admin
- Claims
- Operational/Interactions
- Campaigns

Distribution Channels
- Banks
- Aggregator
- Agents and Brokers

External
- Census
- Electoral roll
- Credit and claims bureau
- Demographic segmentations e.g. Mosaic life style category
- Telematic devices
- Maps, building location, flood maps

Digital
- Web
- Mobile
- Social Media
- Location and social data from mobile apps
- Health tracking devices

Affinity Partners
- For examples:
  - Post office
  - Supermarkets
  - Motor dealers
  - Some insurance partners have banking data or spending data
Applications of Big Data

<table>
<thead>
<tr>
<th>Insurance</th>
<th>Banking</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Identify which claim to assign to which handler</td>
<td>• Credit scoring</td>
</tr>
<tr>
<td>• Fraud detection</td>
<td>• Cross selling</td>
</tr>
<tr>
<td>• Pricing</td>
<td>• Better target customers for marketing campaigns</td>
</tr>
<tr>
<td>• Underwriting acceptance</td>
<td>• Analysis of transaction and spending habits to identify preferences</td>
</tr>
<tr>
<td>• Better target customers for marketing campaigns</td>
<td>and risk appetite</td>
</tr>
<tr>
<td>• Brochure design and targeted marketing messages</td>
<td></td>
</tr>
<tr>
<td>• Telematics</td>
<td></td>
</tr>
<tr>
<td>• Fitness tracking wrist bands</td>
<td></td>
</tr>
<tr>
<td>• Identifying claims likely to become large claims</td>
<td></td>
</tr>
<tr>
<td>• Cross selling</td>
<td></td>
</tr>
<tr>
<td>• Matching orphan policyholders to replacement insurance agents</td>
<td></td>
</tr>
</tbody>
</table>

HR departments using big data – article in Financial Times:

“Employees who are members of one or two social networks were found to stay in their job for longer than those who belonged to four or more social networks”

Potential for behavioral change

Privacy and ethical considerations
Predictive Underwriting using External Information

Life insurance in Thailand

• Swiss Re built a predictive underwriting model for a major Thailand life insurance company together with a local bank whereby the model predicts using banking information a prospective customer’s chance of being a good/bad risk.
• Using the model they are able to select customers with good predicted underwriting risk, and offer them insurance without any additional underwriting.
• A Swiss Re blog on data analytics describes some valuable sources of data:
  – **Banks** have heavily invested in data and are exceptionally well placed to take advantage of their data
  – **Third party data sources** can have very strong predictive power in some markets
  – **Loyalty card / supermarket data** is frequently as strong – if not stronger – than banking data. The challenge is persuading these providers to extract/share their data.

Aviva in USA

• Aviva USA had 60k life insurance applicants which it had underwritten in the traditional way – including blood and urine tests – and categorised accordingly.
• Deloitte took 30k applications and built a predictive model based on insurance application forms, industry information (past insurance applications and motor vehicle reports) and consumer-marketing data from Equifax Inc (hundreds to attributes per individual e.g. hobbies, income, TV-viewing habits).
• Tested predictive model on other 30k to see if could replicate underwriters’ traditional assessments.
• "The use of third-party data was persuasive across the board in all cases," said John Currier, chief actuary for Aviva USA

Source:
http://cgd.swissre.com/risk_dialogue_magazine/Healthcare_revolution/Data_Analytics_in_life_insurance.html

Source:
http://www.wsj.com/articles/SB100014240527487041041045756225310847555588
Non-Actuaries Outperforming Actuaries in this Field

HCF customer retention initiative

- In 2013, Australian health insurer HCF (through Deloitte) invited data scientists to analyze their data to identify policyholders most likely to lapse
- 300 data scientists from Kaggle were invited from around the world from which three submissions were selected for closer examination to use in building a “predictive algorithm that allows them to tailor their health cover more closely to member needs”

Liberty Mutual fire loss prediction

- In 2014, Liberty Mutual ran a contest on Kaggle to predict fire losses to enable more accurate assessment of policyholder’s risk exposure
- 634 entries were submitted included 19 from Liberty Mutual employees. The best Liberty Mutual entry was ranked 36th in the competition
- In a similar competition run by Allstate in 2011, the participants were able to achieve a 340% improvement over Allstate’s ability to predict bodily injury insurance. And that too, with anonymized data and not knowing true makes and models of the cars.¹

Our working party member, Xavier, won both of these competitions demonstrating that it is possible for actuaries to excel in this field

¹ Source: http://andrewmcafee.org/2012/03/a-data-scientist-youve-never-heard-of-is-now-the-master-of-your-domain/
Job Trends & Employer Demand

- Demand for traditional actuarial roles expected to remain strong in Asia driven by market growth and regulatory developments.
- Outside of Asia, in developed markets, predictive modelling and analytics are growing much faster than traditional actuarial jobs. This trend may extend to Asia in the long term.

Source: Presentation by Morand & Troceen, DW Simpson at ICA 20141

What I fear is complacency. When things always become better, people tend to want more for less work.

Lee Kuan Yew
www.geckoandfly.com
Actuaries of the Fifth Kind?

- 17th century: Life insurance, Deterministic methods
- Early 20th century: General insurance, Probabilistic methods
- 1980s: Assets/derivatives, Contingencies Stochastic processes
- Early 21st century: ERM
- Second decade of 21st century: Big Data
Skills Required

Actuaries

• Possess good computing skills
• Are good at math & statistics
• Have deep understanding of business

Actuaries as managers or modelers have a niche in the data science arena

Need to upgrade skillset with emerging tools and techniques relevant to analyze big data

• **Management**: to understand the process, what questions to ask, what skillset to hire
• **Modelling**: to build skillsets that are growing in importance

Source: http://www.edureka.co/blog/who-is-a-data-scientist/
Need to Learn New Tools

Harvard Business Review advise to managers hiring data scientists\(^1\):

“Don’t bother with any candidate who can’t code”

Excel is excellent for learning & visualization but has limitations

- Data size
- Complex analysis becomes difficult (e.g. GLMs)

Tools for big data analytics

- Good first step: R, Python
- Longer term: Revolution R, Hadoop, Microsoft Azure, DataRobot

Useful references


Where to Begin?

Beginner resources
• Lots of online resources
• Attend the workshop

Online courses
• Online course by Caltech: https://work.caltech.edu/telecourse.html
• Online course by Andrew Ng, Stanford University: https://www.coursera.org/course/ml

Textbook
• An Introduction to Statistical Learning with applications in R: http://www-bcf.usc.edu/~gareth/ISL/

In depth learning
• Dimitri Semenovish provides a sample learning pathway (shown below) using courses available online
• Refer to his article in Actuary Australia for more detail: http://actuaries.asn.au/Library/AAArticles/2014/Actuaries191JULY2014p22t25.pdf
Forever Learning
Forever Young
Part II

Machine learning case study
Tools & Techniques

GLMs
- User defined
- Clear model form
- Learning and insight
- Goodness of fit statistics
- Easy to overfit
- Invented in the 70s with limited data

Machine Learning
- Automated
- Non-parametric or obscure form
- Predictive accuracy
- Training & validation process
- Control for overfitting
- Data hungry and evolve with computing power
Why GLMs are Less Popular in a Big Data world?

<table>
<thead>
<tr>
<th>GOOD</th>
<th>BAD</th>
<th>UGLY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognized as a standard in the banking and insurance industry</td>
<td>Need to pre-process data (missing values, outliers, dimension reduction)</td>
<td>GLMs is prone to overfitting while used with large amount of features or features with a large number of categories</td>
</tr>
<tr>
<td>Accommodate responses with skewed distributions</td>
<td>GLMs do not automatically capture complexity in the data. It can take weeks or months to go through the GLM iterative modelling process</td>
<td></td>
</tr>
<tr>
<td>Simple mathematical formula easy to implement and easy to interpret</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Machine Learning based techniques have become the techniques of choice for many industries
Case Study

Background:
• Alarmingly high risk of hospital readmission for diabetes patients in USA

Data:
• UCI machine learning website contained Diabetes hospitalization data from USA
• 10 years’ data; 100,000 records
• 50 columns (variables) for each record detailing patient demographics; treatment; hospitalization, etc.
• For each variable, typically a number of categorical outcomes were observed; for some more than 20 potential outcomes...

Our mission:
• To develop a model that predicts if a patient will need to be readmitted to hospital for treatment within 30 days of leaving
Useful machine learning techniques for insurance

1. **Linear models**: GLM & GAM (this case study focused on GLM)

2. **Regularized GLM**

3. **Decision Tree**: CART

4. **Forests**: GBM & Random Forest (this case study focused on GBM)

So what are these methods doing?
Regularized GLM

- Combine statistical and ML techniques
- Regularization penalizes complexity of model
- Thereby controlling possible overfitting
- Lambda parameter controls the amount of regularization
Classification & Regression Tree (CART)

- Start with a “Target” and split population into 2 groups that are different to each other, using simple rules – e.g. typically higher/lower than a threshold

- is immune to outliers & handles missing values automatically

- generally finds the optimal split

- fast and easy to build

- simple to communicate to non-technical audiences

- Can be unstable.
Random Forest

- Fit trees to random subsets of data, with random choices of explanatory variables
- Use a linear combination of the trees’ predictions
  - Voting (for classification)
  - Averaging (for regression)
- More stable than Decision Tree alone
- Generally higher predictive accuracy
- Much longer runtime

Source: http://kazoo04.hatenablog.com/entry/2013/12/04/175402
Gradient Boosting Machine (GBM)

- Each extra tree focuses models the residuals from the existing model
- Runs quickly: running many small models/trees do not take much long run times than one big model
- Robust and combats overfitting
- Final model may be very complex
What process did we follow

• Solution developed in R – it’s free, so no excuses!
• A series of scripts developed:

  Preparation
  • Download/install packages
  • Download data from UCI
  • Exploratory Data Analysis
  • Data cleansing
  • Build utility code – charting and functions

  Build models
  • GLM
  • Regularised GLM
  • CART
  • GBM

  Decision-time
  • Compare models
  • Declare Winner
  • Test on new data
Diagnostic – **Lift Chart, a Graphical A vs E**

- *Scatterplot of actual readmissions test set vs predictions (ordered ascending)*

GBM performs best – good match across spectrum. GLMs reasonable. Again, CART fit is poor.
Diagnostics - How GLMs Overfit

• Optimising for lowest AIC (as we were taught to do in statistics classes) can cause overfitting with zero or negative gain in predictive power

• Traditional GLM pricing approaches can produce suboptimal models

![Overfitting graph](image)

- Predictive power slowly declining!
- AIC slowly improving
Key Findings

• CART may not have been best, but...

• It can be a powerful way to establish potential rating factors

• Also easy to understand and communicate to a non-technical audience
Key Findings

• Time matters... here’s a pure run time comparison:

![Runtime (mins)](image)

- GLM: 28.4 mins
- Reg. GLM: 4.4 mins
- CART: 2.0 mins
- GBM: 7.0 mins

• In reality many GLMs were tested. So the figure shown is understated.

• Worse, many of the attempted “refined” GLMs failed to produce better models than the initial attempts.
Results of Models We Built

Most Likely (44% probability)

• Nickname: “Frequent Flyer”
• Number of inpatient visits >= 3
• Number of emergency visits >= 2

Least Likely (10% probability)

• Number of inpatient visits <= 1
• Transferred to a different inpatient or rehab facility
• Admission type is emergency or urgent
Conclusions

**Efficiency**

Big data makes traditional actuarial techniques **inefficient** or sometimes even **impractical**

Actuaries can use machine learning to help build better GLMs faster, **freeing up their time to make better commercial decisions**

**Applicability**

Insurance data already has some characteristic of big data

Machine learning techniques **outperform** traditional actuarial techniques both in **predictive power** and **model building efficiency**

**Accessibility**

Most actuaries don’t have the skillset or toolkit yet

But some are already performing at **world best standards**

There are free tools and training that allow actuaries to **start learning right now**
Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it, so everyone claims they are doing it...

(Dan Ariely)