



Institute
and Faculty
of Actuaries

B.A.U. for actuaries: **B**ig data, **A**nalytics & **U**nstructured data

Frankie Chan and Frank Devlin

Singapore Actuarial Society Big Data Working Party

Mudit Gupta (Chair), Clementine Vie, Colin Priest, David Menezes, Frank Devlin, Frankie Chan, Kate Chen, Xavier Conort

3 March 2016

Goals of the Big Data Working Party

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

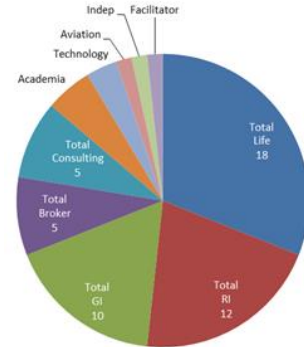
Who we are

The working party is made up of actuaries from life insurance, general insurance, and consulting backgrounds as well as data scientists based in Singapore and Hong Kong.

Activities of the Big Data Working Party


- Presentation at the 2015 SAS GI conference and developed case study
- Two workshops on machine learning using R held in September, 2015, led by Colin Priest in Singapore
- Published an article in HK actuarial society magazine on Data Analytics
- Advanced workshop and other CPD sessions being planned in 2016

Participants at the machine learning workshop (Total = 60)



3 March 2016

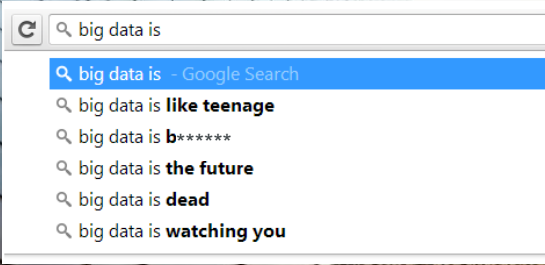
3



Institute
and Faculty
of Actuaries

Part I

Introduction to Big data

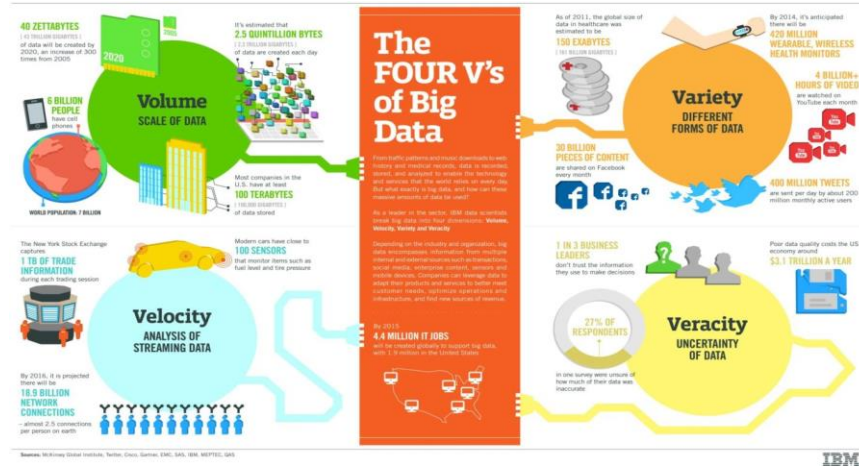


3 March 2016

2

What is big data?

- Often used to describe large volume of data being collected by organizations
- Lack of structure

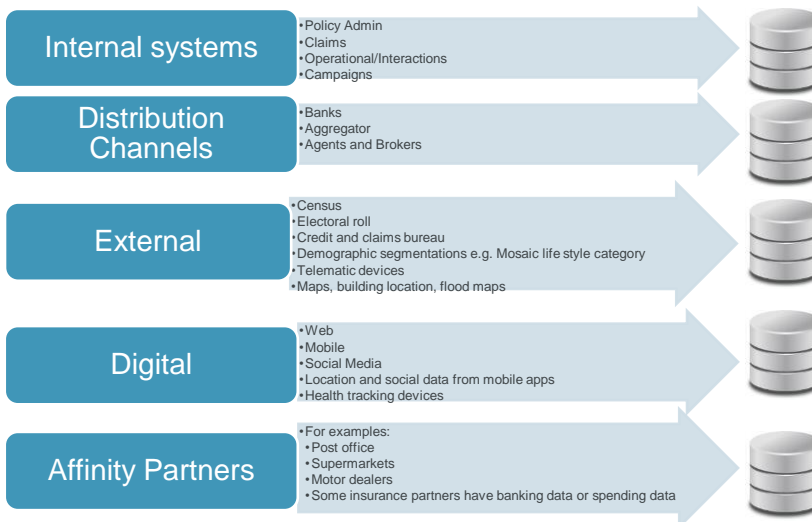


3 March 2016

Infographic from <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>

5

Where does data come from?



3 March 2016

6

Applications of Big Data in Insurance

Claims	Product & Pricing	Customer	Marketing & Distribution
<ul style="list-style-type: none"> • Fraud detection • Case Estimation • Claim handling • Loss adjustment 	<ul style="list-style-type: none"> • Telematics • Wearable • Connected homes • Underwriting Acceptance • Granular Pricing • Price Comparison Websites (PCWs) 	<ul style="list-style-type: none"> • Price Sensitivity • UpSell / Xsell • Take up and Churn • Personalised targeting and messages • Customer satisfaction 	<ul style="list-style-type: none"> • Agency Scoring • Orphan policies agent matching • Media effectiveness • Competitor analysis • Location-based targeting • Sentiment analysis

HR departments using big data – article in Financial Times 9 July 2014:

“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

3 March 2016

7

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.

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

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://www.wsj.com/articles/SB10001424052748704104104575622531084755588>

3 March 2016

8

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 including 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

3 March 2016

¹ Source: <http://andrewmcafee.org/2012/03/a-data-scientist-youve-never-heard-of-is-now-the-master-of-your-domain/>

9

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.

Demand for actuarial jobs flat while growth in analytics and predictive modelling jobs



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

CareerCast
JOBS RATED REPORT
THE BEST JOBS OF 2015

MATH MATTERS

1. ACTUARY
3. MATHEMATICIAN
4. STATISTICIAN
6. DATA SCIENTIST



Four math related jobs in Top 6

<http://www.careercast.com/jobs-rated/best-jobs-2015>

3 March 2016

10

Actuaries of the Fifth Kind?

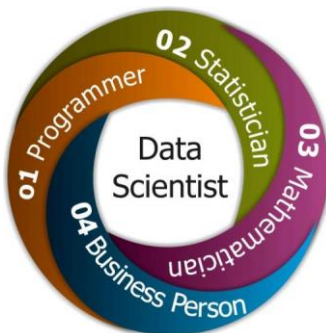
Hans Bühlmann 1987	Actuaries of the First Kind	• 17 th century: Life insurance, Deterministic methods
	Actuaries of the Second Kind	• Early 20 th century: General insurance, Probabilistic methods
	Actuaries of the Third Kind	• 1980s: Assets/derivatives, Contingencies Stochastic processes
Paul Embrechts 2005	Actuaries of the Fourth Kind	• Early 21 st century: ERM
Big Data Working Party	Actuaries of the Fifth Kind	• Second decade of 21 st century: Big Data



3 March 2016

11

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/>

3 March 2016

12

Need to Learn New Tools

Harvard Business Review advice to managers hiring data scientists¹:

“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 reference

- For a detailed comparison of software options, see presentation by Hugh Miller:
<http://www.actuaries.asn.au/Library/Events/GIS/2014/5CMillerSoftwarePres.pdf>

3 March 2016

¹ Source: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>

13

Where to Begin?

Beginner resources

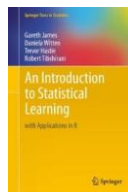
- Lots of online resources
- Attend a workshop

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>

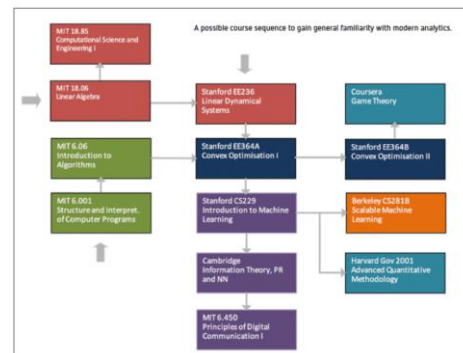
Textbook

- 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>



3 March 2016

14

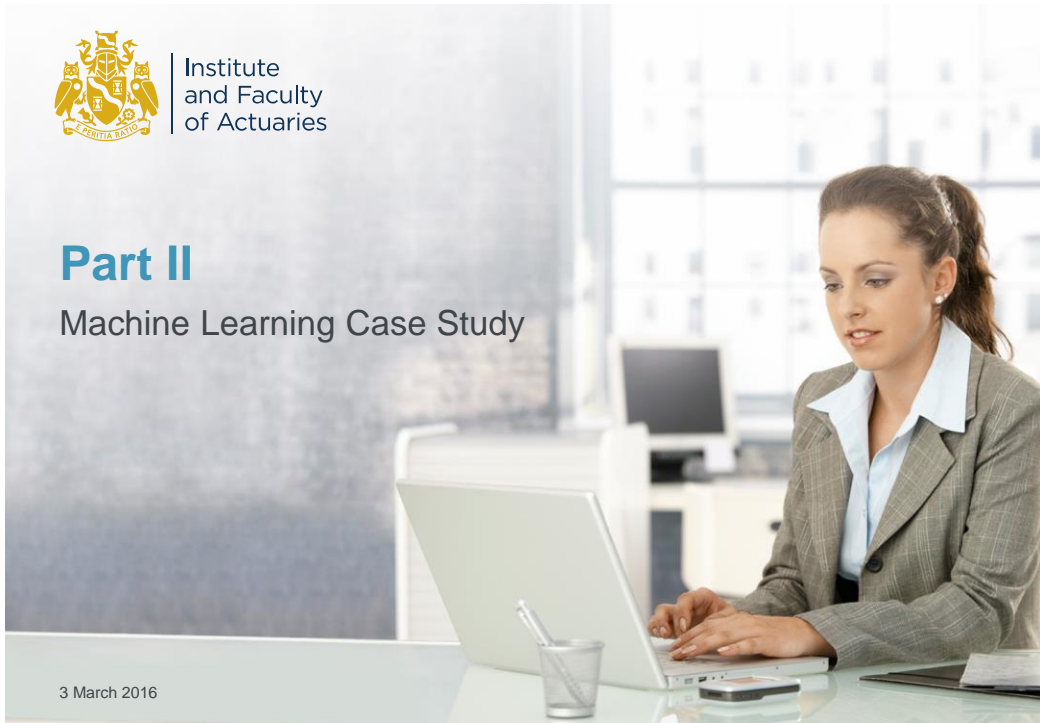


Institute
and Faculty
of Actuaries

Part II

Machine Learning Case Study

3 March 2016



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

VS

Machine Learning

- Automated
- Non-parametric or obscure form
- Predictive accuracy
- Training & validation process
- Control for overfitting
- Data hungry and evolve with computing power

3 March 2016

16

Why GLMs are Less Popular in a Big Data world?

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



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

3 March 2016

17

Case Study

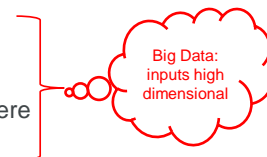
Background:

- Alarming 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

3 March 2016

18

Useful machine learning techniques for insurance

- **Linear models:** GLM & GAM (this case study focused on GLM)
- **Regularized GLM**
- **Decision Tree:** CART
- **Forests:** GBM & Random Forest (this case study focused on GBM)

So what are these methods doing?

3 March 2016

19

Generalized Linear Model (GLM)

- General linear regression model

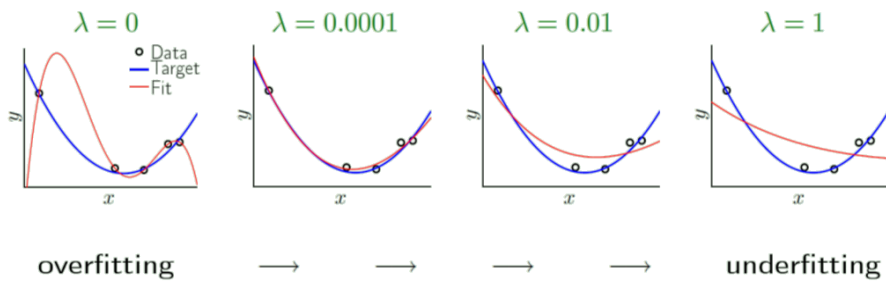
$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_n x_{in} + \varepsilon_i$$
 - Error distributed usually $\sim N(0, \sigma^2)$
- Extend to GLM
 - Linear predictor $\eta_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_n x_{in}$
 - Link function $g(\mu_i) = \eta_i$ where $E(y_i) = \mu_i$
 - Variance function $Var(y_i) = \varphi V(\mu_i)$ where φ is a dispersion factor
 - Errors distributed \sim exponential family
- Logistic Regression for binary outcome data is simple example
 - $Logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_n x_{in}$
- Standard approach in non-life and health, less so in life?

3 March 2016

20

Regularized GLM

- Combine statistical and ML techniques
- Regularization penalizes complexity of model
- Thereby controlling possible overfitting
- Lambda parameter controls the amount of regularization

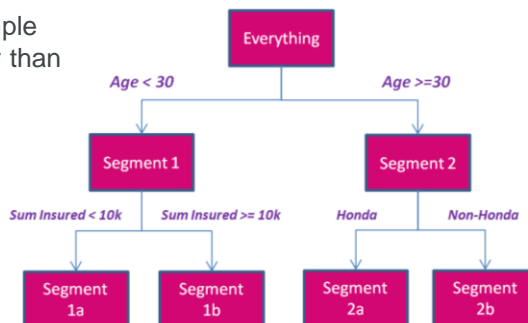


3 March 2016

21

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

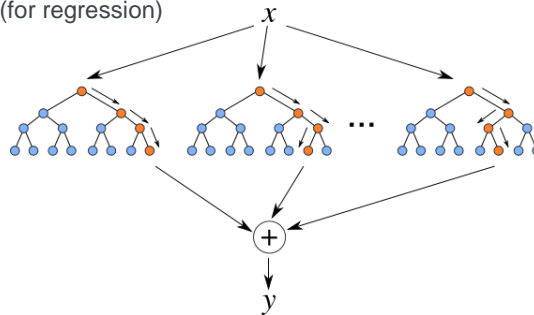


3 March 2016

22

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

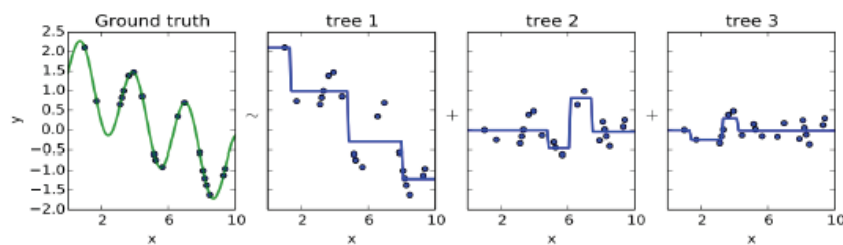


3 March 2016

Source: <http://kazoo04.hatenablog.com/entry/2013/12/04/175402>

23

Gradient Boosting Machine (GBM)



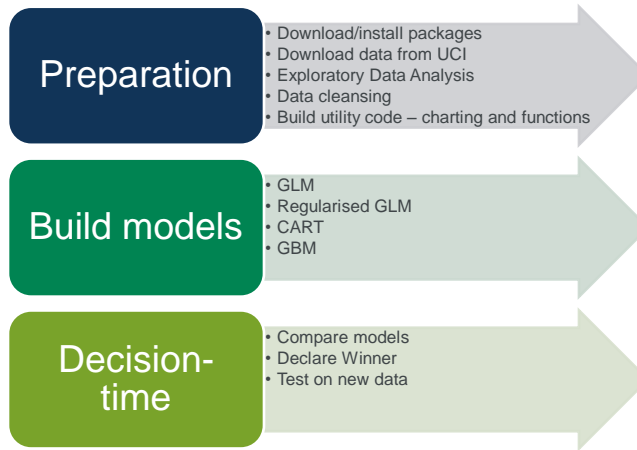
- Each extra tree focuses models the residuals from the existing model
- Runs quickly: running many small models/trees do not take much long run times than one big model
- Robust and combats overfitting
- Final model may be very complex

3 March 2016

24

What process did we follow

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



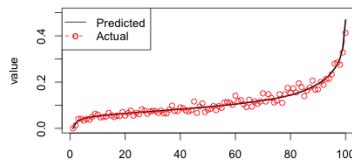
3 March 2016

25

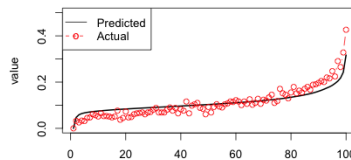
Diagnostic – Lift Chart, a Graphical A vs E

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

Lift chart for Best GLM / LogLoss= 0.329

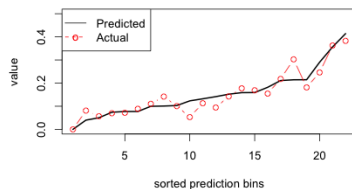


Lift chart for Best Regul. logistic / LogLoss= 0.331

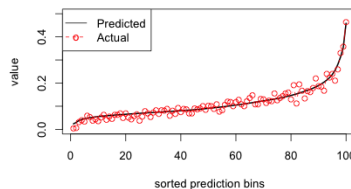


GBM performs best – good match across spectrum. GLMs reasonable. Again, CART fit is poor

Lift chart for Best CART / LogLoss= 0.331



Lift chart for Gradient Boosting / LogLoss= 0.328

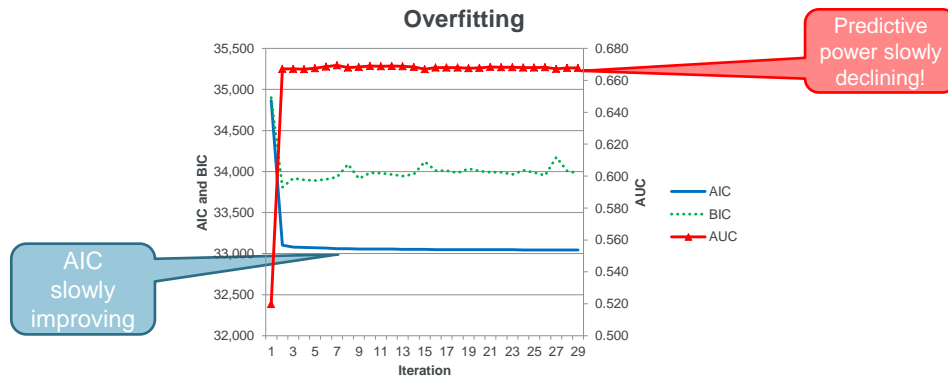


3 March 2016

26

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

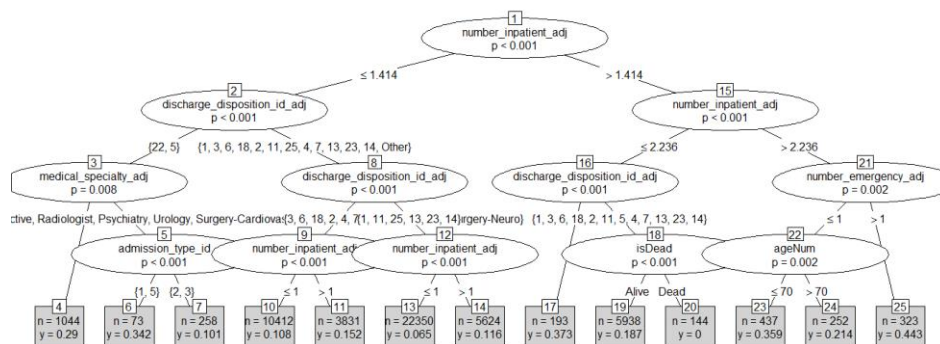


3 March 2016

27

Key Findings

- CART may not have been best, but...



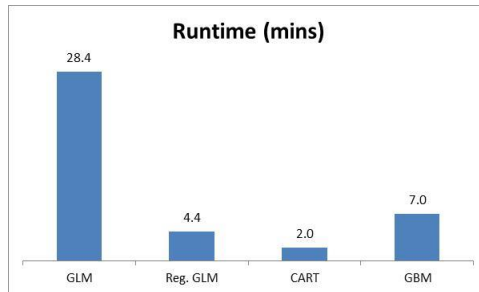
- ...It can be a powerful way to establish potential rating factors
- Also easy to understand and communicate to a non-technical audience

3 March 2016

28

Key Findings

- Time matters... here's a pure run time comparison:



- 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.

3 March 2016

29

Results of Models We Built



Most Likely (44% probability)

- Nickname: “Frequent Flyer”
- Number of inpatient visits ≥ 3
- Number of emergency visits ≥ 2

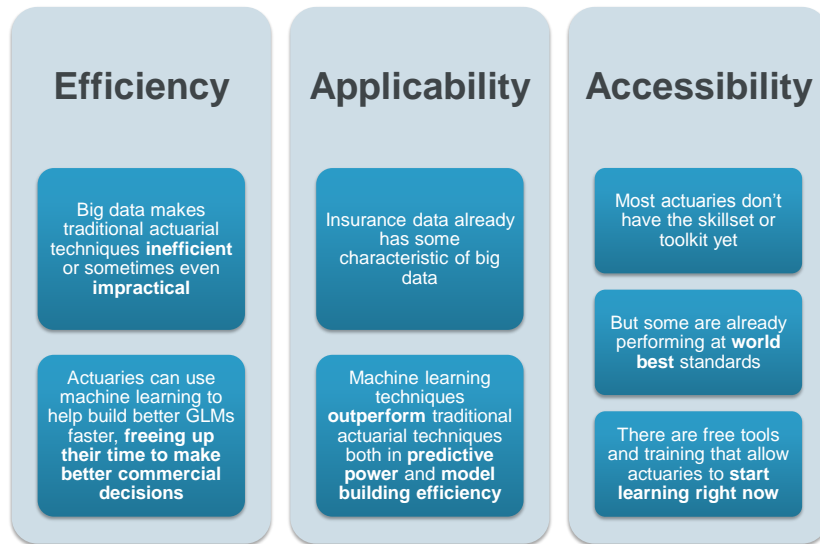
Least Likely (10% probability)

- Number of inpatient visits ≤ 1
- Transferred to a different inpatient or rehab facility
- Admission type is emergency or urgent

3 March 2016

30

Conclusions



3 March 2016

31



The views expressed in this presentation are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this [publication/presentation] and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this [publication/presentation].

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the IFoA [or authors, in the case of non-IFoA research].

3 March 2016

32