Article from

Predictive Analytics and Futurism

December 2016
Issue 14

Machine Learning: An Analytical Invitation to Actuaries

By Syed Danish Ali

This post highlights the various value-additions that machine learning can provide to actuaries in their analytical work for insurance companies. As such, a key problem of swapping specific risk for systematic risk in general insurance ratemaking is highlighted along with key solutions and applications of machine learning algorithms to various insurance analytical problems.

"In pricing, are we swapping specific risk for systematic risk?"[1]

The hypothesis is that in normal market conditions, premiums are kept at low levels to increase revenues and market share. The traditional approach requires precise figures (point estimates) and so leads to understatement of uncertainty. This keeps a comfort level for us, but the hidden risk of underpricing in our premium estimates is hardly given the attention it merits. This crops up from the rug it was shrugged in during stressed market conditions when high loss ratios then systematically prove the premium rates to be underpriced and unsustainable.

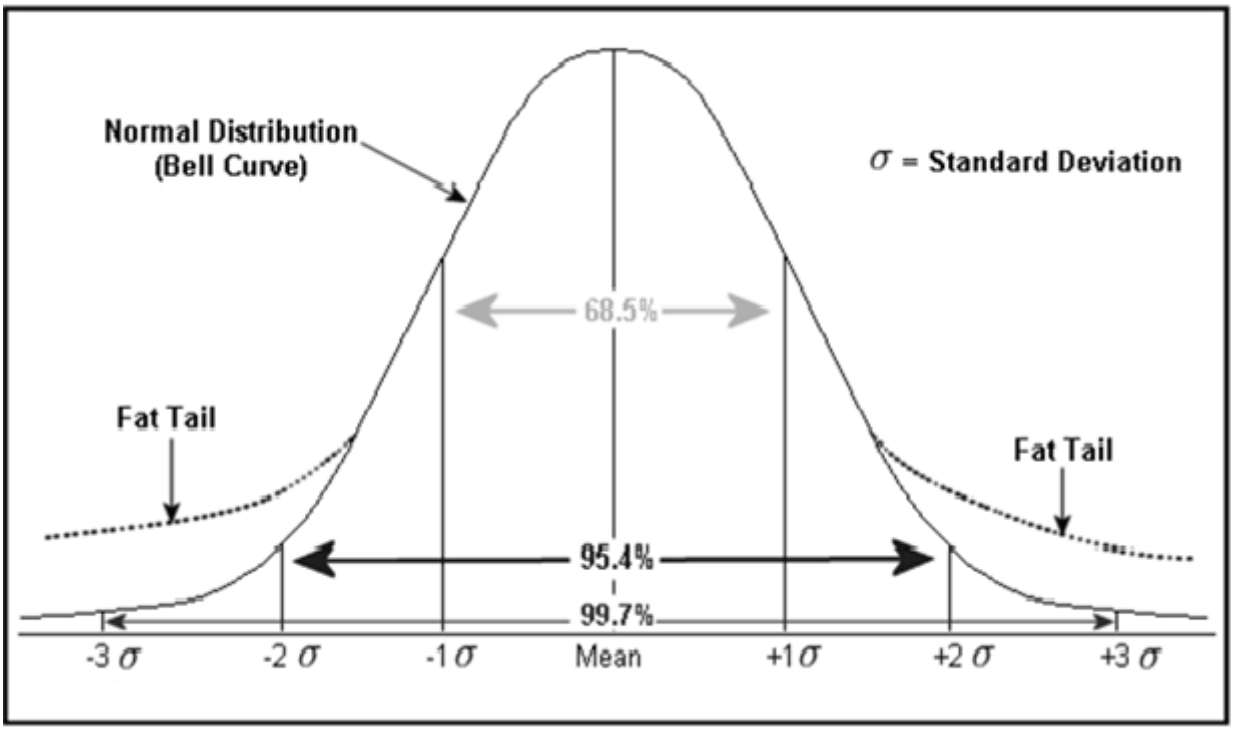
In other words, are we causing the fat tail problem[2] by our practices? Even if not, what can be done to reduce the fatness of such tails and bring the hidden uncertainties onto the surface explicitly?[3]

A fat tail exhibits large skew and kurtosis and so there is a higher probability for large losses compared to other distributions like normal distributions as shown by the diagram below. This higher loss tendency remains hidden under normal market conditions only to resurface in times of higher volatility. Complexity scientists call fat tails the signature of unrecognized correlations. Fat tails are an indicator that cascading risks are influencing the probability distribution.

While our discussion does not provide an exhaustive guide to the machine learning tools and algorithms available to the actuary, it provides an outline of them while supplying a context for them in the ratemaking process.

We argue that what was perceived as uncertain can now be made less uncertain with machine learning. Also the uncertainty should be captured from where it was partly generated like risky classes were underwritten which later lead to greater pricing uncertainty and so on.

Machine learning has brought about an explosion of algorithms in recent times. As actuaries are not traditionally trained for



Source: *MachineLearningMastery.com*

machine learning, and because there are so many algorithms, it can lead to 'paralysis through analysis' where one is confounded by so many choices (R's Caret package of machine learning has 147+models) and decides instead to do nothing but follow previous precedent. The mindmap above, still not exhaustive, made by Jason Brownlee at Machine Learning Mastery highlights a number of diverse classes and subclasses of algorithms and approaches applied in Machine Learning:[4]

Each of these models has a different bias, and hence its own strengths and weaknesses relative to other algorithms and areas of application. It is certainly not possible to discuss many of these algorithms so we will try to stick to "actionable insights" produced from focusing on a small number of relevant algorithms.

With regards to pricing uncertainty and ratemaking applications generally, machine learning can be applied in ratemaking in a number of ways:

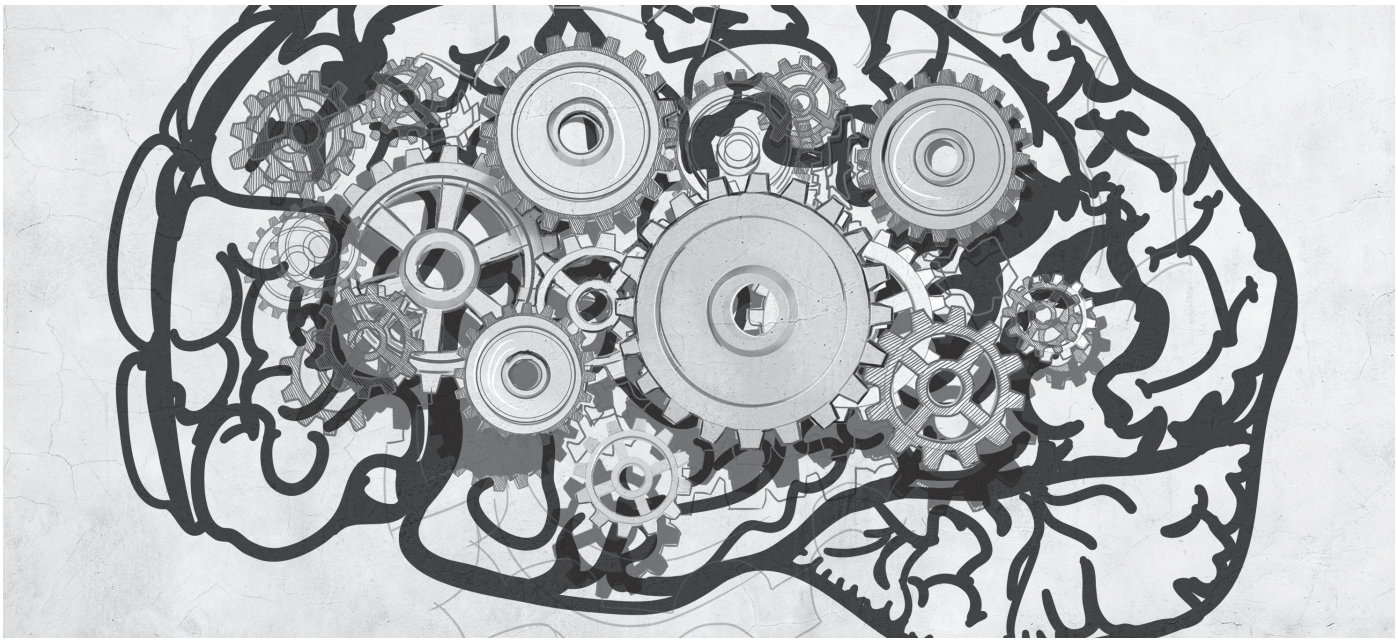
- Exploring our data;
- Predictive modeling; and
- Unstructured data mining and text analytics.

EXPLORING OUR DATA

Decision trees such as hidden decision trees or random forests can allow us to see the map and the critical paths upon which the data is proceeding. Thus, the trend and nature of even huge datasets can be understood through decision trees.[5] Decision trees are unsupervised methods of learning which means that they expose the trends within the data without relying only on what the analyst is interested in querying.

Clustering, especially K-means clustering, is an imperative algorithm that exposes different clusters operating within a given data.[6] This can tell us the groupings within claim registers and premium registers like one cluster can be that bodily injuries are associated with third parties that are associated with non-luxury vehicles that are commercial and so on.

For time series decomposition, there are R codes available for running this decomposition algorithm. Basically, decomposition of time series takes a real-data time series and breaks it down into: 1) trend (long term), 2) seasonal (medium term); and 3) random movements.[7] Such decomposition can have huge potential in understanding trends in data. For instance, claims data have trends



that follow an underwriting cycle and mimic the economic cycle closely. An instance for a seasonal trend can be higher sales of travel insurance in spring break and summer breaks and so on.

PREDICTIVE MODELING

Aside from exploring the data, various uncertain elements of risks can be captured by predictive modeling as well.

Generalized Linear Models (GLMs) can be applied to arrive at a distribution of frequency and severity of claims. Mostly Gamma or Lognormal distributions are fitted to severity data and Poisson or Negative Binomial to frequency. Another approach is to directly apply Tweedie distribution on pure premium.

GLMM is a natural extension to GLM models as the linear predictor now contains random effects as well to incorporate fuzziness and give a stochastic feel for enhanced pricing.[8]

Predictive modeling using GLMs and GLMMs can also be assigned to categorize a particular policy into its proper risk category, like into predictive risk for claim likelihood for a particular policy and so on (unacceptable risk, high risk, medium risk, low risk, etc.). Separate modeling can then be done for each major risk category so as to expose greater insight into the ratemaking process.[9] The results from the separate models can act as a feedback loop to the risk and underwriting categories of how valid and reliable these categories are, and promote greater cooperation between underwriting function and the claim/reserving function, which is vital to generating adequate risk-adjusted premiums.

While it is important to select optimum risks for predictive modeling and ratemaking on a broad level, it is also vital to take the notion of fairness into account. There have been a couple of backlashes around ratemaking such as a law not allowing the use of gender to quote prices, controversial social images of using credit scores to quote premiums and most recently, pricing optimization where customers and regulators have pointed out that simply market dynamics like price elasticity and consumer preferences should not lead to different premiums and that only risk factors (and not market factors) should lead to premium differentiation.[10]

Complexity scientists also favor use of power law distributions, like the Pareto-Levy distribution, for any modeling purpose. This should be tried by the actuary to apply it on severity data and compare its results with other distributions to see if any improvements have been achieved.[11]

UNSTRUCTURED DATA AND TEXT MINING

It is well known that 80 percent of data is unstructured. Unstructured data is the messy stuff every quantitative analyst tries to traditionally stay away from. It can include images of accidents, text notes of loss adjusters, social media comments, claim documents and review of medical doctors, etc. Unstructured data has massive potential, but has never been traditionally considered as a source of insight before. Deep learning is becoming the method of choice for its exceptional accuracy and capturing capacity for unstructured data. The traditional relational databases use rows and columns in handling data, but NoSQL (Not-Only-SQL) uses a number of other components such as giving unique key or hash tagging to every item in the data. Insurance companies can utilize NoSQL databases like MongoDB, Cloudera and Hadoop

because they capture so many elements of reserving that were deemed belonging to the domain of uncertainty before, as they were too messy and qualitative.[12]

Text mining utilizes a number of algorithms to make linguistic and contextual sense of the data. The usual techniques are text parsing, tagging, flagging and natural language processing.[13] There is a correlation between unstructured data and text mining as many unstructured data is qualitative free text like loss adjusters' notes, notes in medical claims, underwriters' notes, and critical remarks by claim administration on particular claims and so on. For instance, a sudden surge in homeowners' claims in a particular area might remain a mystery, but through text analytics, it can be seen that they are due to rapid growth in mold in those areas. Another useful instance is utilizing text analytics when lines have little data or are newly introduced, which is our research aim here.[14]

Sentiment analysis/opinion mining over expert judgment on level of uncertainty in reserves can also prove fruitful. Natural Language Processing (such as in Stanford 'CoreNLP' software available free for download[15]) is a powerful source of making sense out of the texts.

Claim professionals often have more difficulty in assessing loss values associated with claims that are commonly referred to as "creeping cats."[16]

These losses typically involve minor soft tissue injuries, which are reserved and handled as such. Initially, these soft tissue claims are viewed as routine. Over time, however, they develop negatively. For example, return-to-work dates get pushed back, stronger pain medication is prescribed, and surgery may take place down the road. Losses initially reserved at $8,000–$10,000 then become claims costing $200,000–$300,000 or more. Since these claims may develop over an extended time period, they can be difficult to identify. Creeping cat is a big problem for emerging liabilities because mostly, we do not fully know what we are dealing with. Emerging risks like cyber-attacks, terrorism, etc., have shown to have huge creeping cat potential.

As discussed, predictive models can review simulated claim data from agent-based modeling, network theory and other methods mentioned in this report for similarities and other factors shared by such losses, thereby alerting the claims professional to emerging risks that may have creeping cat potential. With this information, strategies and resources can be applied at a point in time where they can be most effective in an effort to achieve the best possible outcome and control cost escalation. Additional loading on premiums can also be given on areas with higher creeping cat potential.

In conclusion, by measuring and exposing areas of uncertainty that are traditionally not considered, we can reduce our chances of swapping specific risk for systematic risk in our ratemaking procedures and lessen fatness of the tails and handle emerging liabilities in a more resilient manner.

Moving these data collection policies and the uses of this data from the subconscious to our consciousness is a first step in the process of potentially applying big data in a business context. The use of big data and analytics has rapidly evolved from a back-room niche to a strategic core competency.[17]

In conclusion, actuaries will have to understand and appreciate the growing use of big data and the potential disruptive impacts on the insurance industry. Actuaries will also need to become more proficient with the underlying technology and tools required to use big data in business processes.[18] ■

Syed Danish Ali is a senior consultant at SIR consultants, a leading actuarial consultancy in the Middle East and South Asia. He can be reached at *sd.ali90@ymail.com.*

ENDNOTES

1 Idea adapted from The Economist "In Plato's Cave"; January 2009.

2 Fat Tail: Lexicon of financial times

3 Image from advisor analyst.com available here (*http://advisoranalyst.advisoranalystgr.netdna-cdn.com/wp-content/up...*)

4 Jason Brownlee at Machine Learning Mastery; Mindmap of machine learning algorithms

5 HR Varian, 2014; The Journal of Economic Perspectives. "Big Data: New Tricks for Econometrics."

6 Liu, D. R., Shih, Y.Y., 2005: The Journal of Systems and Software, 77 (2005) 181–191. "Hybrid Approaches to Product Recommendation Based on Customer Lifetime Value and Purchase Preferences"

7 Zucchini and Nenadic; R Vignette: Time Series analysis with R—Part I

8 University College London; Introduction to GLMM

9 Breton and Moore; SOA 14; Predictive Modeling for Actuaries: Predictive Modeling Techniques in Insurance

10 CAS Task Force (November 2015): Price Optimization White Paper.

11 Smith, L. and Tossani, L.S. CAS; "Applications of Advanced Science in the New Era of Risk Modeling"

12 IBM White paper 2013: "Harnessing the Power of Big Data and Analytics for Insurance"

13 Stanford Natural Language Processing Group; available at: *http://nlp.stanford.edu/software/*

14 CAS Ellingsworth and Balakrishnan: 2008. "Practical Text Mining in Insurance"

15 Stanford Natural Language Processing Group; available at: *http://nlp.stanford.edu/software/*

16 Lentz,W. GenRe Research (November 2013); "Predictive Modeling—An Overview of Analytics in Claims Management"

17 SOA; The Actuary Magazine, December 2013/January 2014 – Volume 10, Issue 6, Ferris, et. al. "Big Data."

18 Ibid