Institute of Actuaries of Australia

The Rise and Rise of Hybrid Modelling

Prepared by Hugh Miller

Presented to the Institute of Actuaries of Australia
17th General Insurance Seminar
7 – 10 November 2010
Gold Coast

The Institute of Actuaries of Australia
Level 7 Challis House 4 Martin Place
Sydney NSW Australia 2000
Telephone: +61 2 9233 3466 Facsimile: +61 2 9233 3446
Email: actuaries@actuaries.asn.au Website: www.actuaries.asn.au

The Rise and Rise of Hybrid Modelling

Hugh Miller
Taylor Fry

Abstract

Hybrid modelling refers to the use of multiple modelling approaches to improve the overall accuracy of a fit. It is useful in areas where a high degree of prediction accuracy is desired. Typically it involves using approaches that are complementary, in the sense that they are able to detect different relationships in the data. This talk/paper includes a general introduction to the area, including discussion of how it relates to:

- Model averaging – combining two or more model predictions together
- Variable generation – using one model to generate new variables for use in the other
- Residual fitting – how a final model can be produced iteratively by feeding the results from one technique into another

There will also be an emphasis on some of the model types commonly used in general insurance, such as generalised linear models and decision trees. I explain why these two in particular are well suited for hybrid modelling and describe some new work to help maximise accuracy. Concepts and performance will be demonstrated on a workers compensation claim triage problem.

Keywords: Model averaging, gains chart, composite variables, residuals, decision trees, generalised linear models

1. Introduction

1.1 Background

There are many aspects of actuarial practice that are concerned with building predictive models. These are built with a focus on a particular response variable, denoted Y_i, that requires estimation from a collection of p predictor variables $X_i = (X_{i1}, \ldots, X_{ip})$. Thus we are interested in constructing a function f from the space of predictors that is in some way close to the response:

$$Y_i = f(X_i) + \text{error}_i \tag{1}$$

Such a model is built using a collection of n past observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, where both predictors and response are observed. It almost always makes use of a loss function, which measures the magnitude of departure of the observed responses from the fitted model, which is then minimised. These observations are often taken to be (and will be assumed to be for the purposes of this paper) independent, and are often drawn from some common distribution. The fit can then be applied to new observations, where the response is not available, to give insight into the problem being studied. While uses of predictive models are many and varied, common uses in insurance contexts include:

- *Risk pricing*: Given a series of risk characteristics about a customer, the task is to predict the expected cost of insuring that risk. For example in comprehensive car insurance predictors such as age, gender, car type and location are used to predict the average cost of claims in a given exposure period. This prediction is typically based on historical data and is a key driver in the premium charged to a customer.

- *Claims monitoring*: As claims develop, various pieces of information are collected, which can be used to predict the eventual duration and cost associated with the claims process. This can be particularly powerful for monitoring longer-tailed insurance lines such as workers compensation.
- *Classical reserving*: All traditional reserving models can be viewed as exercises in predictive modelling, attempting to forecast future claims cost based on historical information. In some cases the only predictors available might be accident period and development period (a classical claims triangle), but if the data is sourced as individual claim records then other variables may be available too.
- *Customer behaviour*: There has been an increased focus on marketing analytics in recent years, enabled by computerised collection of detailed consumer information. Using demographic, behavioural and economic predictors, complex models of consumer behaviour can be built, assisting to identify the value of different customers to a company, as well as their loyalty and responsiveness to promotions.

In some circumstances the need for highly accurate predictive models based on past experience is limited. For example, it may be the case that large changes in future behaviour caused by factors invisible to the model (for instance, legislative or industry developments) are likely to have a far greater impact on the response than the model error in (1). In such a circumstance incremental increases in model accuracy are likely to offer little practical value. However, in many other situations such incremental improvements are in fact highly valuable and contribute directly to a company's performance. For example, accurate risk pricing in a competitive insurance market may have large impacts in profitability, and accurate claims monitoring can lead to large improvements in the claims management process. It is in this context of attempting to achieve maximum accuracy that the present paper is written.

There is an extremely wide variety of tools available for modelling the relationship in (1). The recent textbook by Hastie et al. (2009) forms a good reference describing a broad range of approaches. Some common models/frameworks used in actuarial practice include:

- *Generalised linear models (GLMs):* These are a family of extensions to classical multivariate linear regression, accommodating many different types of response data. Book length treatments of generalised linear models include McCullagh and Nelder (1989) and Dobson (2002).
- *Decision trees:* These generally seek to partition the data into different regions through a series of bivariate splits, with each split defined by a rule involving a single predictor variable. The resulting regions are given separate predictions, usually based on the average response. Foundational papers include Breiman et al. (1984) and Quinlan (1993).
- *Neural networks:* These are a nonlinear extension to traditional linear and logistic models through the addition of hidden layers that relate to the predictors and response by some pre-defined function. Hertz et al. (1991), Bishop (1995) and Ripley (1996) are useful references here.
- *Credibility models:* Models built on credibility have long been a part of actuarial science. They attempt to balance the observed data with informative priors, which tend to reduce the weight of outliers supported by only limited amounts of data. Modern treatments of this topic include Bühlmann (2005) and Herzog (1999).
- *Classical reserving models:* As mentioned above, traditional techniques such as chain ladder methods are included the predictive modelling framework. Standard introductions include Taylor (2000).

Readers are directed to the above references for further information. The first two listed are described further below, as some familiarity of them is required to understand the remainder of the paper. In many ways GLMs and decision trees are complementary, in that they often detect different relationships in a dataset; GLMs can fit broad main effects that apply over an entire dataset, whereas decision trees can detect small regions where the response behaviour is markedly different.

Of course, the problem of constructing good predictive model is not unique to actuarial practice. There is a wide literature on the topic spanning a number of fields, most notably statistics and areas in computer science such as machine learning. Two truisms arising from the wealth of model building approaches are that:

1. No single approach will consistently perform best on different data types
2. Even for a given dataset, different approaches come with their own strengths and weakness in their efforts to produce accurate estimates.

The first of these points motivate an analyst to be familiar with a number of modelling approaches, and to be prepared to try more than one in situations where accuracy is important. The second is the key motivation for this paper, in that the analyst may want to attempt to incorporate the good characteristics of multiple models in an effort to boost performance. This general approach is what we define as hybrid modelling.

1.2 Purpose of this paper

This paper seeks to make two contributions. The first is to introduce, survey and summarise some of the existing work and ideas of hybrid modelling. Much of this discussion will draw on ideas and references from the statistical and machine learning literature as the use of such approaches in actuarial contexts is less widespread. Section 2 covers this material, and in particular explores three specific ways of using hybrid modelling; model averaging, variable generation and residual fitting. While the section is intended primarily as a survey, some inclusions are novel, such as the applicability of pseudo-residual modelling to generalised linear models. Also, variable generation as a formal procedure is poorly represented in the literature.

The second contribution of the paper is a real-data example of these methods in a claims triaging context for a workers compensation portfolio, presented in Section 3. This clearly demonstrates the usefulness of hybrid modelling in actuarial contexts, and attempts to make some of the concepts introduced more concrete.

The remainder of Section 1 introduces some notation and model assessment ideas that will be useful from the remainder of the paper. Some brief conclusions are given in Section 4.

1.3 Notation

We use x and y to denote the generic predictor and response variables respectively. These will have an underlying distribution from which we sample n historical observation pairs, $(X_1,Y_1)\ldots(X_n,Y_n)$. Notice here that Y_i is a scalar while $X_i=\{X_{i1},\ldots,X_{ip}\}$ is a p-vector. The accuracy of a model is almost always measured (explicitly or implicitly) by means of a loss function L, so that the loss between a prediction function $f(x)$ and response y,

$$L\{y,f(x)\} \tag{2}$$

The loss function should equal 0 whenever $y=f(x)$, and be non-decreasing as the distance between y and $f(x)$ increase. The optimal choice of f is that which minimises the average loss over the sampling distribution of the $\{x,y\}$. As this is not observable, f is instead estimated be minimising the empirical version of the loss,

$$n^{-1}\sum_{i=1}^{n} L\{Y_i,f(X_i)\}.$$

Table 1 presents some common loss functions with descriptions for the type of they are appropriate for.

Table 1: Common loss functions

Type of response	Loss function	$L\{y, f(x)\}$	Comments
Continuous	Squared loss	$\{y - f(x)\}^2$	Used in classical linear regression
	Absolute loss	$\lvert y - f(x) \rvert$	More robust to outliers than squared loss.
Count data	Poisson loss	$y\log\{f(x)\} - f(x)$	Used in a standard Poisson GLM with log link
Continuous, strictly positive	Gamma loss	$(\beta - 1)\log y - \frac{y\beta}{f(x)} - \Gamma(\beta) - \beta\log\{f(x)\}$	Used in a standard gamma GLM with log link. β is a scale parameter.
Binary (0-1) data	Logistic loss	$y\log\{f(x)\} + (1-y)\log\{1 - f(x)\}$	Used in standard logistic GLM logit link. $f(x)$ is the modelled probability of a positive response
	Misclassification rate	$I\{y \neq f(x)\}$	For when predictions are 0-1 classes, rather than probabilities

1.4 Generalised linear models

Generalised linear models are an extension to classical linear models where the response y is assumed to follow the distribution of a member of the exponential family (see Ch. 2, McCullagh and Nelder, 1989). This covers a broad range of distributions including the normal/Guassian, Poisson, Gamma, inverse Gaussian and Bernoulli (logistic regression) distributions. The mean of the response variable is estimated as a function of the linear combination of the predictors:

$$E[Y_i] = f(X_i) = h^{-1}(\beta_1 X_{i1} + \ldots + \beta_p X_{ip})$$

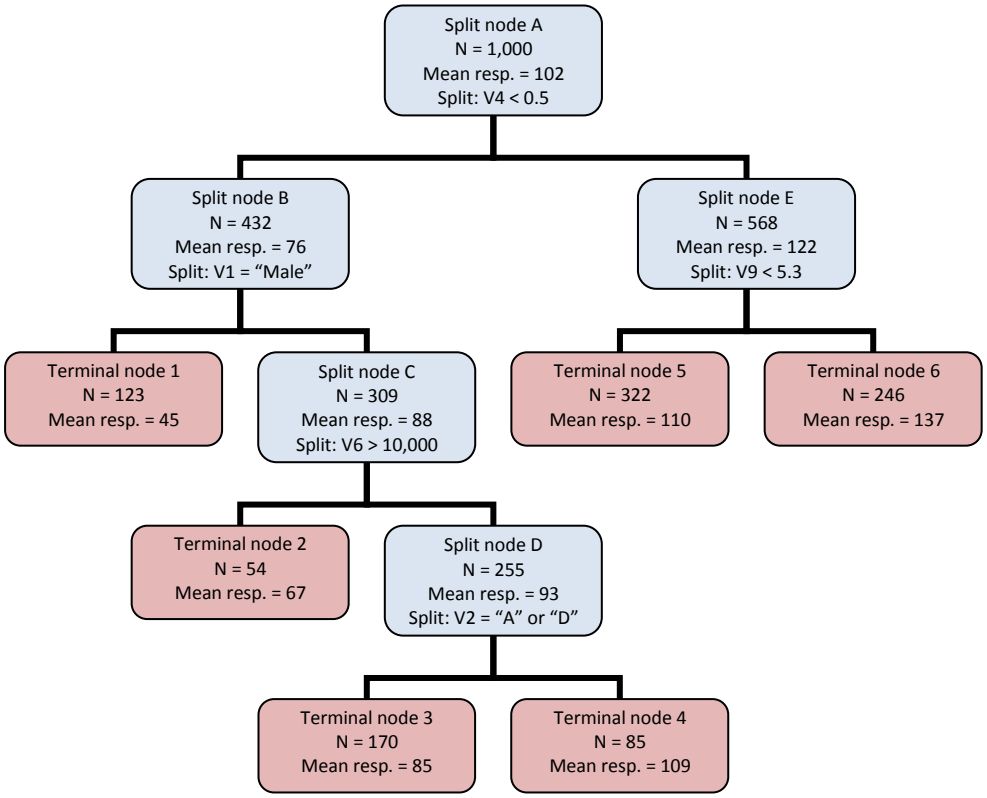
The function h is termed the link function and is useful in applying a linear model to a restricted response domain. For example, if $y > 0$ then using a log link ensures that all predictions are positive.

As with other models, the particular choice of parameters while fitting is performed with reference to a loss function, such as those in Table 1. In most cases these flow naturally out of maximum liklihood expressions relating to the assumed distributions. Thus the choice of loss function is often not explicit, but arises in a natural way.

1.5 Decision trees

A decision tree makes separate predictions on disjoint regions of the predictor domain. This is done by creating a set of splitting rules that separate the current region into two. In the case where the Y_i are continuous, each split aims to split the current observations into a set with higher than average response and the other with lower than average. Consider the mock-up of a decision tree fit on a dataset with 10 predictors and continuous response, presented diagrammatically in Figure 1 below.

Figure 1: Example decision tree



Starting with the top box, Split Node A, we have 1,000 observations with an average response of 102. The chosen splitting rule is whether an observation satisfies V4<0.5, where V4 denotes the fourth predictor variable. The 432 observations for which this is true are sent to the left, and the remainder sent to the right. Each split is chosen to minimise the resulting loss on the data, and if a "good" split cannot be found then a node becomes a terminal node and all observations belonging to it are given the mean response as a final prediction. So, for example, the 54 observations satisfying V4<0.5, V1 not "male" and V6 > 10,000 are given a final prediction of $f(X_i)$=67. Note here that not all variables are necessarily used, and each splitting rule depends on only one variable.

While many more comments could be made here about fitting decision trees, interested readers are referred to the references given in Section 1.1.

1.6 Model assessment

One of the most crucial decisions in setting up a model is deciding how to assess a model. The most basic way to validate a model is to calculate the average loss (2) on a dataset that has been kept separate from the modelling process. Such a dataset is often called a test dataset, and a model with lower average loss is to be considered superior. Calculating this loss on a test dataset as opposed to the original data is an important protection against over-fitting, where spurious patterns are included in the final model.

The analysis performed in Section 3 actually considered a range of tools in model selection, but the presentation will focus almost exclusively on the gains chart. This is a powerful tool that shows how well a set of predictions is ordered. Each point on the curve is defined by a prediction cut-off t. For a given value of t, the x and y coordinates of the curve are defined as

$$\left(\frac{\text{Number of } i \text{ such that } f(X_i) > t}{\text{Number of observations}}, \frac{\text{Sum of } Y_i \text{ for the } i \text{ satisfying } f(X_i) > t}{\text{Sum of all } Y_i} \right)$$

A better curve will move towards the top right of a plot, while a poor model tends towards a 45 degree line. Thus gains curves with larger areas underneath represent more powerful models. Generally these areas are difficult to improve upon, and are often measured in fractions of a percent. Examples are given in the data analysis in Section 3, where the gains charts of performance on a test dataset are reported. When applied to binary response data, the gains curve shares many characteristics with the ROC curve (see, for example, Zou et al. 2007).

One key advantage of the gains chart is that it is model (and loss function) independent, which makes it useful in situations where models are fit according to different loss paradigms.

2. Types of hybrid modelling

While the definition of "hybrid" is not universally consistent, we focus here on three classes of hybrids.

2.1 Model Averaging

Suppose that K different models have been fit on the dataset. This means there are K different prediction functions $\hat{f}_1, \ldots, \hat{f}_K$, each attempting to approximate f, the true relationship between predictors x and the response y. Rather than assessing each and choosing the single best model for use in prediction, performance can be improved by taking a weighted sum of these predictions,

$$\hat{f}_{\text{avg}}(x) = \sum_{k=1}^{K} \delta_k \hat{f}_k(x). \tag{3}$$

The simplest approach is to simply set the model weights to $\delta_k = K^{-1}$, so that each underlying model has equal influence in the final prediction. However, recognising that the set $\hat{f}_1(x), \ldots, \hat{f}_K(x)$ can be treated as a set of predictor variables themselves means that the full spectrum of regression techniques can be used to produce better choices of the δ_k.

The key feature of this approach is that the performance of the averaged model $\hat{f}_{\text{avg}}$ is generally better than any individual one. The formal statistical basis for this behaviour is not completely understood, although there are some efforts, particularly in the Bayesian literature (see, for instance, Hoeting et al., 1999, and references cited therein). The core intuition is that different models are able to capture different types of structure and dependencies in the data, which can prove complimentary.

A recent example of where model averaging has been used to great effect was in the recent Netflix Prize competition (Tőscher and Jahrer, 2009). Netflix is an American DVD rental company where customers choose movies over the internet. It is crucial for Netflix to be able to predict how different customers will enjoy different movies, so that they can recommend new movies to returning visitors, thus driving business. The competition allowed contestants to use Netflix's data to attempt to predict how customers would rate other movies in the future, with a $1m prize to the first team that improved on Netflix's accuracy by a set threshold. It was quickly recognised that model averaging was the key tool for achieving hyper-accurate predictions. Teams would produce a wide range of different models using

different techniques and attempt to combine them into a final model. In later stages of the competition, teams found that the easiest way to make further progress was to merge with other teams, thus allowing more models to be averaged. The eventual competition winners were a merger of three teams, who together combined 166 different models in a linear combination of the type in (3).

Other "out-of-the-box" model in the statistical and machine learning literature that make use of model averaging include bagging (Breiman, 1996) boosting (Freund and Schapire, 1997, and Friedman et al., 2000) and random forests (Breiman, 2001). Boosting in particular has had a surge of popularity in recent years, and Treenet, used in the example of Section 3, is an example of this; it fits a series of models by updating the weights at each stage to focus on the observations that are hardest to estimate. It also bears characteristics of residual fitting, discussed in Section 2.3.

2.2 Variable Generation

An alternative to averaging final predictions is to have a single model producing the final predictions, but to inform it using other models via means of variable generation.

One recent methodology of a variable generation approach which appears to have strong predictive ability is the RuleFit algorithm (Friedman and Popescu, 2008). In the variable generation stage, multiple decision trees are generated, and each node on the tree is converted into a "rule", which is a binary variable (1 if an observation is in the node, 0 otherwise) that is added to a new set of predictor variables. This prediction matrix is then used in a penalised regression (such as the lasso model of Tibshirani, 1996), with the final model using a small subset of the rules generated. A simpler version of this approach is also possible; take a single decision tree fitted on the data, and use it create a new categorical variable with levels corresponding to the distinct terminal nodes of the tree. These are essentially interactions which may be then tested for significance in the original model, usually a GLM.

The key advantage to variable generation is that the generating models may be able to detect features that are otherwise hard to identify. In the examples above, decision trees have a greater ability to detect pairwise (and higher level) interaction effects, which depend on more than one variable. These are often hard to identify in a traditional regression model or GLM, and so the generated variables can often yield genuine improvements to the model.

2.3 Residual Fitting

The third approach to hybrid modelling discussed in this paper is using the residuals of a model as a response variable for a subsequent model. The two models can then be combined to give a superior overall prediction. Naturally the process of residual fitting can be repeated over many iterations, resulting in a series of models which need to be combined (usually by some form of model averaging, above). In fact, boosting, mentioned in Section 2.1, is often viewed as a residual fitting algorithm, although the type of model fitted at each stage usually does no vary.

The simplest situation to concretely demonstrate residual fitting is in least squares regression. Here Y_i is continuous and squared loss $L\{Y_i, f(X_i)\} = \{Y_i - f(X_i)\}^2$ is minimised, so that the minimisation using equation (2) becomes

$$\sum_{i=1}^{n} Y_i - f(X_i)^2 . \tag{4}$$

If a model f_1 is estimated using the data, then the raw residuals,

$$r_{1i} = Y_i - f_1(X_i), \tag{5}$$

can then be used as a new response variable. The next model, f_2, is then found by attempting to minimise

$$\sum_{i=1}^{n} r_{1i} - f(X_i)^2 .$$

One reason why residual fitting is relatively straightforward in the least squares context is the transferability within the (squared) loss function:

$$L\{Y_i - f_1(X_i), f_2(X_i)\} = L\{Y_i, f_1(X_i) + f_2(X_i)\}. \quad (6)$$

This means that fitting f_2 to the residuals naturally gives a way to construct a final model $f_1(X_i) + f_2(X_i)$. Unfortunately this property does not hold for all loss functions, so simply subtracting off the current prediction is not a general recipe for residual modelling. One way to overcome this issue is through the use of pseudo-residuals.

2.3.1 Pseudo-residuals

Pseudo-residuals, or generalised residuals, for current estimate f_m are defined as

$$r_{mi} = \left[\frac{\partial L\{Y_i, f(X_i)\}}{\partial f(X_i)} \right]_{f=f_m} . \quad (7)$$

They attempt to capture the "direction of greatest gain", meaning that a model g_{m+1} is a good approximation for the residuals, then $f_{m+1} = f_m + \delta g_{m+1}$ for small enough δ should yield a model with decreased loss. These principles underpin much of the boosting literature. Notice that this definition coincides with the formulations in (4) and (5) above, up to a scalar factor, which implies that raw residual modelling under least squares loss is a special case of this generalised framework.

We include the factor δ here because in situations with complicated (non-convex) loss functions, allowing a full residual modelling step of $\delta = 1$ may result in overfitting or instability in the fitting procedure.

2.3.2 Pseudo-residuals for log link GLMs

One area for which the theory of pseudo-residuals does not immediately apply is to that of GLMs, and here we focus on those using making use of log links. While the derivative of the loss functions are directly calculable for the well known distributions such as the Poisson or gamma, many GLMs are fit using Tweedie distributions, which have no closed form formulae for their distributions, and hence their loss functions. Despite this, there is a reasonable choice for pseudo-residuals. If ϕ is the scale parameter of GLM and the variance of an observation for a given value of the mean μ is $V[\mu] = \phi\mu^p$, then defining

$$r_i = \frac{Y_i - f(X_i)}{\phi f(X_i)^p},$$

where $f(X_i)$ is the current estimate for the mean μ, correctly captures the dynamics of residual modelling for GLMs. Notice this is just the raw residual scaled by the variance, which has the intuitive interpretation of given less credence to observations that are expected to be more volatile. Further, this definition coincides precisely for those situations where the distribution function is known (normal, Poisson, gamma and inverse Gaussian). Thus a pseudo-residual based approach to GLM residual modelling is possible.

2.3.3 Other GLM based techniques

One final technique for incorporating an alternative model into a GLM structure is to include it as an offset in the GLM. If $f^*(X_i)$ is this alternate prediction, then the updated generalised linear model equation is

$$E[Y_i] = f(X_i) = h^{-1}(\beta_1 X_{i1} + \ldots + \beta_p X_{ip} - h\{f^*(X_i)\}).$$

The linear equation thus builds on the existing model in an effort to add extra power to the model. This technique is relatively straightforward to implement, and is demonstrated in Section 3.3.3 below.

3. Case study: Workers' Compensation Statistical Case Estimates

3.1 Data description

Some of the above ideas for hybrid modelling are demonstrated on a dataset provided by the Worksafe Victoria[1]. The task is to design a statistical case estimate model for the Impairment Benefits payment type. Thus for every claim without a paid impairment benefit, we seek to predict the likely outcome. One important portion of this project was estimating the probability that a claim would result in a Section 98C payment, the most prevalent type of impairment benefit. This is the model that will be investigated in the current analysis.

The data itself was essentially a snapshot of claims history at two dates – 31 December 2002 and 31 December 2005. All predictor variables were taken from the 2002 snapshot, and the binary response (whether a Section 98C payment was made) based on the 2005 data. Some of the important variables included in the model were:

- *Diagnostic:* location, means and type of injury
- *Weekly compensation details:* time off work, amount of payments
- *Demographic:* Age, salary
- *Medical:* medical costs, the dates of medical reports, days in hospital.
- *Time based:* Time since claim lodged, time to finalisation, duration since last payment etc

In total there were 41 predictor variables considered in the final models.

There were about 110,000 observations in the dataset, of which 1.5% had an eventual Section 98C payment. The response was thus a binary variable, with $Y_i = 1$ when such a payment occurred in the future. The data was randomly allocated into training and test datasets in the ratio of 75%-25%, with all models fitted using only the training data and all results reported concern prediction performance on the test dataset.

3.2 Initial models

Two different models were built on the data. First a GLM was built, with the usual logistic link and loss functions. The fitting process involved deleting insignificant variables, creating splines and indicator variables where appropriate, using the AIC to help determine the overall contribution of a parameter. The resulting model contained 34 parameters, including 5 pairwise interaction terms.

The second model was built using Treenet, propriety software from Salford Systems that builds what is essentially a weighted average of decision trees using boosting. Each tree had 6 terminal nodes, and cross-validation suggested using 170 trees in the final fit was about optimal.

[1] Data used with permission

Figure 2: Gains chart comparison for GLM and Treenet fits

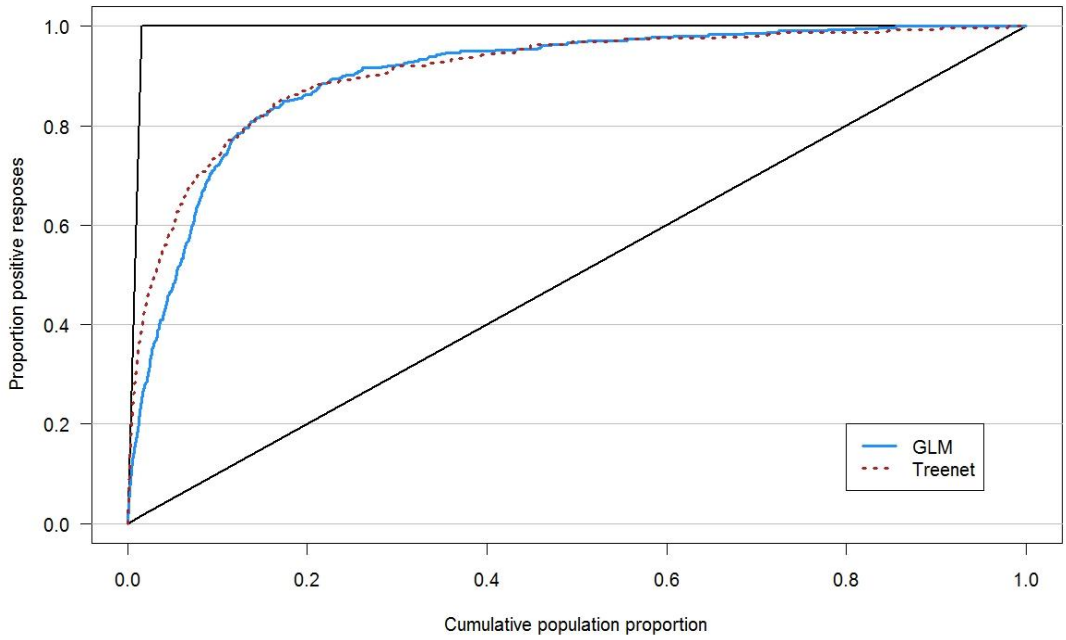


Figure 2 compares the resulting gains chart performance on the test data for each of the two models. Note that we have also plotted curves representing a perfectly predicting model (top left) and a random guessing model (the 45^{o} line). The most striking feature of the plot is that each model performs best in different areas of the plot. For the riskiest 15% of the population (as judged by the models), the Treenet model is much stronger at ordering observations as likely to be a genuine claim. However on the less risky portion of the population, seen on the right of the plot, the GLM has superior performance.

Such a result supports the use of hybrid models, since each approach is clearly identifying different patterns and features in the dataset, so there is the potential to leverage both these models in some sort of hybrid.

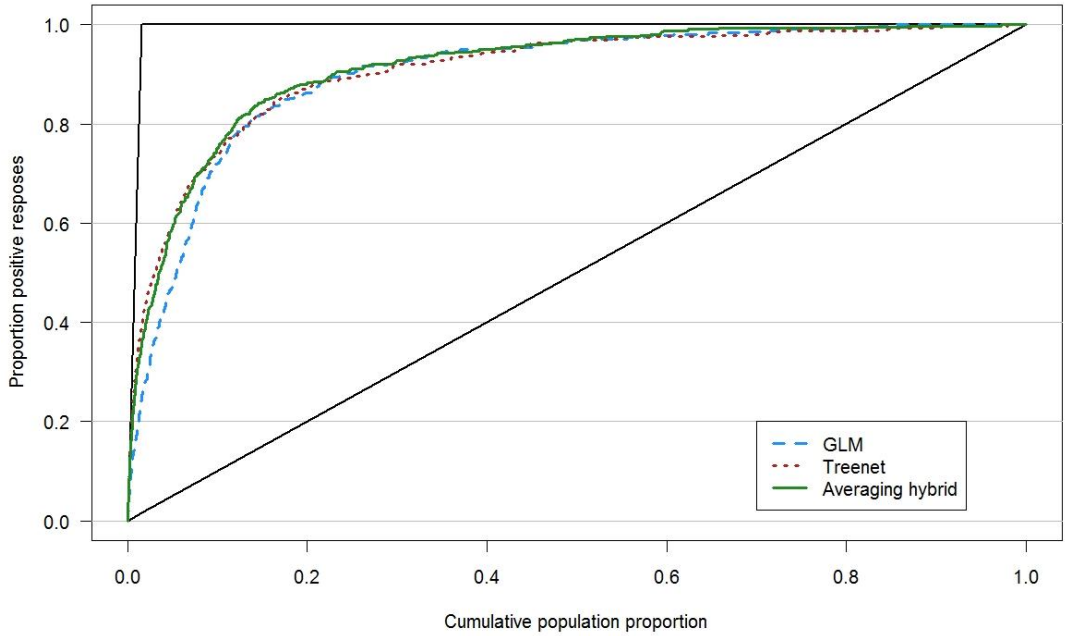
3.3 Hybrid models

In this section we explore the different ways to combine and improve the models of Section 3.2. In particular there is one hybrid for each of the three approaches discussed in Sections 2.1 to 2.3.

3.3.1 Model averaging

Here the hybrid model is formed by simply taking the straight average of the two prediction probabilities, so the models have equal weight. What is interesting is that the resulting gains curve, shown in Figure 3, is superior to the underlying models, rather than lying between the two. In fact the area under the gains curve for the hybrid model, a useful means of comparison, is 3.1% higher than the GLM and 1.4% higher than Treenet. Indeed the best features of both models are reflected in the new curve; it tracks the Treenet curve fairly accurately on the left, and then follows (or lies above) the GLM's curve on the right. Further, there is some out-performance near where the lines cross at 15% of the population, meaning substantially better detection of potential claims.

Figure 3: Gains chart comparison for model averaging hybrid approach

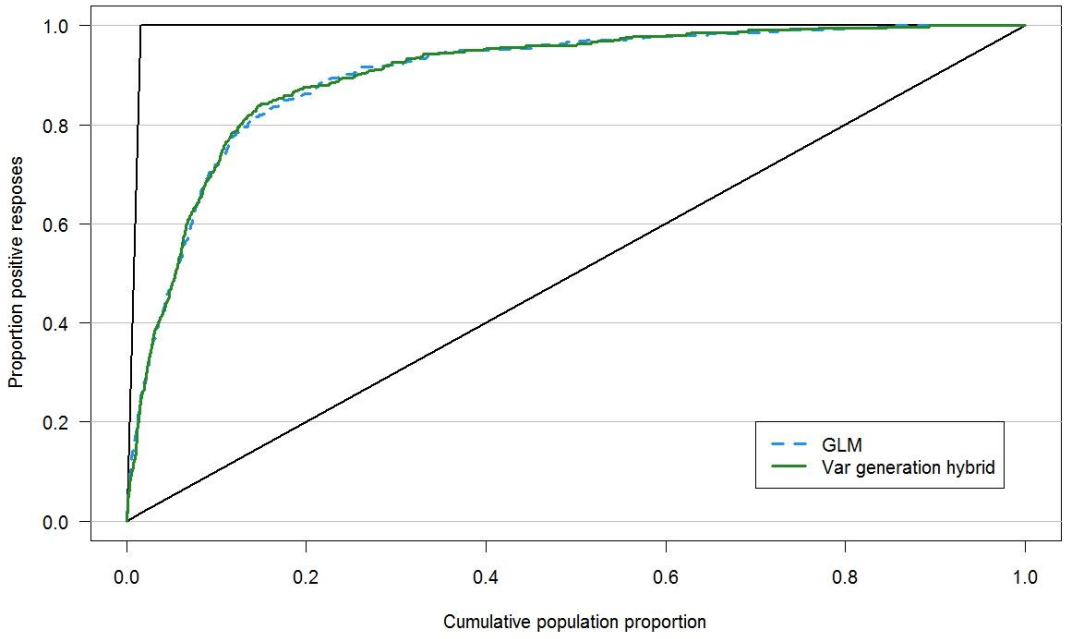


3.3.2 Variable generation

Rather than incorporating the Treenet model, here we attempt to improve the GLM by means of some simple variable generation. A single decision tree with 10 terminal nodes was fitted, and a categorical variable denoting which of these nodes an observation belonged was added to the GLM. This corresponds to adding 9 interactions to the model, ranging in complexity from 2-way up to 7-way (that is, interacting up to seven different variables together). Six of these interactions proved to be statistically significant, with the most important ones finding interesting patterns amongst days in hospital, the duration of weekly payments and the bodily location of the injury.

The resulting test dataset gains curve is presented in Figure 4, with the GLM curve provided as reference. An improvement is evident near 0.15 on the x-axis, while the area around 0.23 is slightly worse. The improvement in the area under the gains is only slight, at 0.3%, but this hybrid has the advantage of identifying interesting (and interpretable) interactions for incorporation into the GLM.

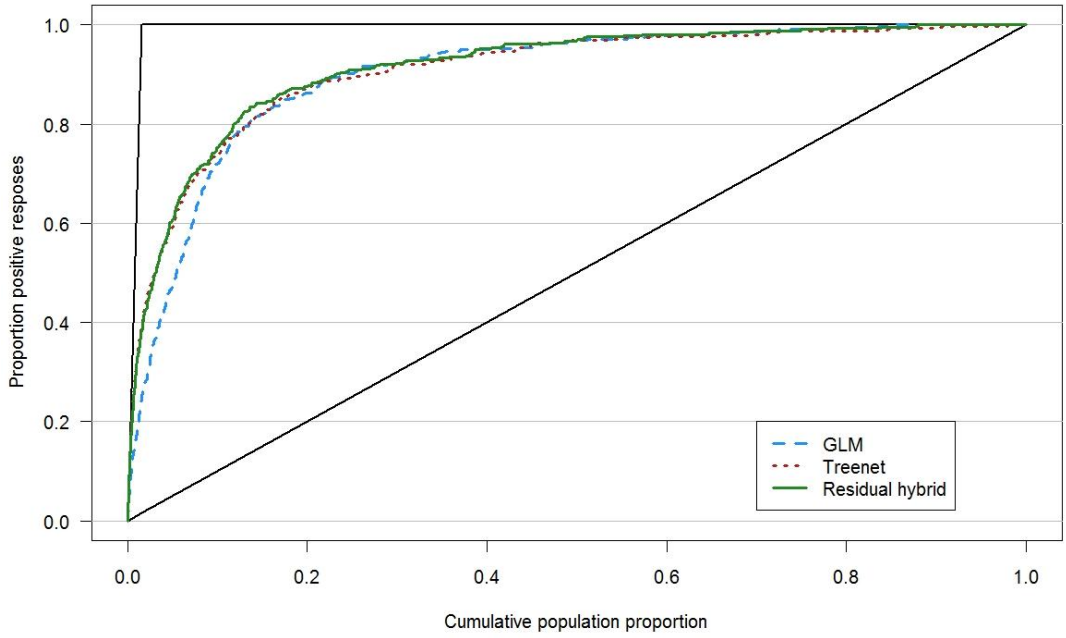
Figure 4: Gains chart comparison for variable generation hybrid approach



3.3.3 Residual modelling

In this case the Treenet predictions were used as an offset in the GLM, which was then fit using exactly the same parameterisation as the original GLM. The result is again a model which dominates the other two, but perhaps follows the Treenet curve more closely. The out-performance seen at around 15% of the population is once again visible here. The area under the gains curve is 3.2% and 1.5% better than the individual GLM and Treenet models respectively, making this the most accurate model by this measure.

Figure 5: Gains chart comparison for residual modelling hybrid approach



One other comment relevant here is that we did not change the structure of the GLM, even though the refitting suggested that some effects and splines were no longer significant. It should be possible to produce even better performance by remodelling these GLM effects, although such improvements are not pursued here.

4. Summary and conclusions

In situations were modelling accuracy is particularly important, hybrid modelling offers the possibility of very real gains. It is often straightforward to implement, can be done in a number of different ways, and allows a much broader range of approaches to come to bear on a particular problem. In contexts such as risk pricing, claims triaging and other situations where prediction accuracy is paramount, the use of hybrid models is likely to continue to grow.

Acknowledgements

I would like to thank Julie Evans at Worksafe Victoria for provision of data for the analysis. Also, thanks to Peter Mulquiney for useful suggestions at the peer review stage.

References

BISHOP, C. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

BÜHLMANN, H. AND GISLER, A. (2005). *A course in credibility theory and its applications*. Springer Verlag.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning*, **26**, 123–140.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth, New York.

BREIMAN, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

DOBSON, A. J. (2002). *An introduction to generalized linear models.* CRC Press LLC.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion), *Annals of Statistics* **28**, 337–307.

Friedman, J. and Popescu, B. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, **2**, 916–954.

HASTIE, T.,TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning, 2nd Ed.* Springer.

HERTZ, J., KROGH, A. and PALMER, R. (1991). *Introduction to the Theory of Neural Computation*. Addison Wesley, Redwood City, CA.

HERZOG, T. N. (1999). *Introduction to credibility theory*. Actex Publications.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E., and VOLINSKY, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, **14**, 382–417.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

QUINLAN, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

TAYLOR, G. C. (2000). *Loss reserving: an actuarial perspective*. Springer, Netherlands.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

TŐSCHER, A. and JAHRER, M. (2009). The BigChaos Solution to the Netflix Grand Prize. *Available at www.netflixprize.com*.

ZOU, K. H., O'MALLEY, A. J., and MAURI, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, **115**, 654–7.