ASTIN, AFIR/ERM and IACA Colloquia
Innovation & Invention
23-27 August 2015 | Amora Hotel Jamison Sydney

Actuaries Institute

Why High Dimensional Modeling in Actuarial Science?

Prepared by Simon Lee, Katrien Antonio

Presented to the Actuaries Institute
ASTIN, AFIR/ERM and IACA Colloquia
23-27 August 2015
Sydney

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9233 3466 **f** +61 (0) 2 9233 3446
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

WHY HIGH DIMENSIONAL MODELING IN ACTUARIAL SCIENCE?

Simon CK Lee[*1] and Katrien Antonio[†1]

[1] Faculty of Economics and Business, KU Leuven, Belgium

May 1, 2015

Abstract. We describe the Generalized Additive Family which covers Generalized Linear Model, Generalized Additive Model, Neural Network and Boosting and more. We outline the theoretical features of each member, along with the corresponding implications on actuarial applications. A complete procedure that details the data-scrubbing, preliminary data diagnosis, variable selection, model selection and model diagnosis is presented for each family member to model claim frequency for a set of real life data. This paper aims to provide a robust framework to actuaries on how multiple modeling techniques can help enhance the credibility of popular pricing models.

1. Introduction

Predictive modeling has gained immense popularity in information analytics. In marketing, predictive models are used to derive cross-sales or up-sales opportunities, and catalogs are aligned to enhance customers' shopping experience. In hotel or flt bookings, predictive models are used to vary prices by the time booked or by classes of the product. In search engines, models are built to best match what users need; for example, time spent on the selected site is heavily analyzed as a benchmark of matching. In Bio-statistics, models are used to predict clinical results, and quality of life has significantly improved for many patients because of accurate prediction of required formula ingredients. In short, the ability of information analytics and predictive modeling to make accurate predictions has transformed society.

Many predictive modeling techniques have been developed to fit the great variety of applications that have arisen. Popular modeling choices include Support Vector Machines, Boosting, Random Forests, Artificial Neural Networks and Classification and Regression Trees, with each

[*]email: chunking.lee@student.kuleuven.be

[†]email: Katrien.Antonio@kuleuven.be

option diff somewhat in how the modeling problem is framed and how the prediction is derived. Support Vector Machines, for example, transform the original data space into a higher dimensional space so that the data is linear decomposable. Classification and Regression Tree provides decisions through simple branch type fl w charts (Figure 2). Boosting and Random Forests *ensemble* base learners, e.g. decision trees, into a predictive model. Artificial Neural Networks attempt to mimic how a neural system would process the information. All these models aim to extract patterns between the explanatory variables and the response; the real pattern is generally high dimensional, meaning that some combinations among explanatory variables exert unique infl on the response. The modeling techniques are thus sometimes called data mining, machine learning or high dimensional modeling.

Information analytics in actuarial science are evolving as well. According to the 2013 predictive modeling benchmarking suvery by Towers Watson, 71 percent, compared to 67 percent in 2012, of North American personal insurers indicated that some form of predictive analytics are either in place will be in place in the next year. The numbers have been increasing over time, and the numbers are even higher in Europe due to more competitive operational environments.

While actuaries believe they are fully embracing the advanced modeling technologies, discussions in the insurance industry are still heavily biased towards the application of Generalized Linear Models (GLM). Publications on how to apply GLM for pricing, reserving, demand, economic capital models are numerous [1, 8, 20] and techniques that help reduce extreme predictions from GLMs are popular topics at actuarial conferences. However, high dimensional techniques are not yet prevalent among actuaries.

Several factors contribute to the actuarial bias towards GLM instead of higher dimensional techniques. Firstly, insurance is a highly regulated industry in North America and Asia. The regulators have a mandate to assess the reasonableness of proposed pricing processes, but this assessment is diffi if regulators do not have a working understanding of the submitted process. GLM is a fairly mature topic in Statistics and standardized packages are available in many statistical engines like SAS and R; there are even software packages specifically designed to engineer GLM models for insurance applications. So, while GLM modeling choices and selections must still be supported and/or defended, there is enough documentation of the process itself that regulators can be convinced of its reasonableness. In contrast, the "black box" nature of data mining process makes those assessments diffi Algorithms are not easily presentable as they can be quite complicated, and may

be counter-intuitive from a non-statistical perspective. These methods also lack theoretical diagnostic tools due to their non-parametric nature. Even within each technique there are many subtle variations that are not fully reviewed by the statistical community. All of these issues contribute to regulatory resistance to high-dimensional modeling.

Secondly, the insurance industry includes many professionals other than actuaries. Underwriting, Marketing, Sales and other functions are also involved in the pricing process, with varying levels of statistical training, such that all of the concerns outlined above for regulators can also apply within the company itself.

Lastly, most insurers' rating engines are only capable of reading a limited number of tables. While algorithms from data mining techniques can generally be decomposed into tables, the number can easily go into the thousands. The space required to store these tables and the capacity to audit table accuracy is a significant operational challenge to the IT infrastructure.

Despite these drawbacks, there is tremendous value in exploring high dimensional modeling in the insurance industry; data mining techniques can significantly outperform GLM if there are strong interactions among variables, which is commonly the case for insurance data. Data mining can detect interactions, select variables and handle missing values simultaneously within the modeling process. The degree of improvement of data mining over GLM varies with the experience and judgment of the modeler, but leveraging features of data mining techniques will make the pricing process more efficient and effective. The practice also resonnates the Statement of Principles regarding Property and Causaulty Insurance Ratemaking by Casualty Actuarial Society: *A number of ratemaking methodologies have been established by precedent or common usage within the actuarial profession. Since it is desirable to encourage experimentation and innovation in ratemaking, the actuary need not be completely bound by these precedents.*

All discussions about modeling techniques starts with a formal definition of the problem to be solved. Section 2 defi that problem. Section 3 contains a more detailed description on the selected models, highlighting the diff and similarities between GLM and data mining.

Using real life pricing data as an illustration, Sections 4 and 5 outline how a modeling process is adapted to combining insights from various modeling techniques, as well as describe our approach to data scrubbing, variable selection, model selection and diagnostics.

2. The problem: Function Approximation

Actuarial ratemaking applications attempt to model key values such as claim frequency and severity, conversion and retention. These models are classified as supervised learning, which means a *response* is to be predicted. Non-supervised learning models do exist, such as Clustering, but our discussion will focus on supervised learning with a clearly defi response.

Mathematically, a system of data contains entries with response variables, y, and corresponding predictive co-variates, $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$. The co-variates and response are assumed to be linked by an unobserved mapping F and a user-specifi strictly monotonic *link function*, $g(\cdot)$. The goal is to fi an estimate function F^* that minimizes a specified loss function $\Phi(y, F(\mathbf{x}))$, mathematically represented as

$$F^*(\mathbf{x}) = \underset{F(\mathbf{x})}{\operatorname{argmin}} E_{\mathbf{x}}[E_y(\Phi(y, F(\mathbf{x}))|\mathbf{x})] \tag{1}$$

Not every function is a loss function. A loss function should fulfill the following conditions.

Definition 1. *A function, $\Phi(y, F(\boldsymbol{x}))$, is a loss function if it satisfies all the following conditions.*

(1) ***Identifiable:*** *if $\Phi(y, F_1(\boldsymbol{x})) = \Phi(y, F_2(\boldsymbol{x})) \ \forall\, y,\ F_1(\boldsymbol{x}) = F_2(\boldsymbol{x})$.*
(2) ***F-convex:*** *$\Phi(y, F(\boldsymbol{x}))$ is convex on $F(\boldsymbol{x})$ and is strictly convex at $F_{min}(\boldsymbol{x})$ where $F_{min}(\boldsymbol{x}) = \underset{F(\boldsymbol{x})}{\operatorname{argmin}}\Phi(y, F(\boldsymbol{x}))$. In the problem of function estimation, $F_{min}(\boldsymbol{x}) = g(y_i)$.*
(3) ***Y-convex:*** *$\Phi(y, F(\boldsymbol{x}))$ is convex on y.*
(4) ***Closed:*** *The set where $\Phi(y, \cdot)$ is defined is closed.*

Condition (1) ensures identifiability which is a property that a model must satisfy in order for precise inference to be possible. Condition (2) and (3) guarantee that any loss function is a *measure of distance*. If $F_1(\mathbf{x}) > F_2(\mathbf{x}) \geq y \geq F_3(\mathbf{x}) > F_4(\mathbf{x})$, then $\Phi(y, F_1(\mathbf{x})) > \Phi(y, F_2(\mathbf{x}))$ and $\Phi(y, F_3(\mathbf{x})) < \Phi(y, F_4(\mathbf{x}))$. Condition (4) is necessary to guarantee the end points are included in the parameter space.

The prediction $\hat{y}$ is equal to $g^{-1}(F^*(\mathbf{x}))$. Loss functions can be generic or specific to various types of problems. For example, Random Forest uses squared error for all types of problems whereas gradient boosting allows for Huber loss, deviance, absolute error and many others as loss functions [16]. The most commonly used loss functions in actuarial predictive modeling is deviance. Deviance, D, is a negative

linear transformation of loglikelihood, ll.

$$D = -2ll + C \tag{2}$$

where C is a constant. Thus, minimizing the deviance is equivalent to maximizing the log-likelihood. Since our goal is to provide a platform to compare model performance in actuarial applications, deviance will be used throughout this paper. Interested readers can refer to Lee [30] for a more detailed study of alternative loss functions.

3. Model Candidates: Family of Generalized Additive Models

The family of generalized additive models covers many existing solutions to the function approximation problem due to its generality. Mathematically, the class consists of an algorithm that is represented in the following form:

$$F^*(\mathbf{x}) = g(\sum_{j=1}^{J} f_j(\mathbf{x})) \tag{3}$$

Each family member uniquely has its unique way to specify the basis function and assigns parameters, which distinguish itself from other members. The basis function, also known as the base learner, and parameters are combined to form $f_j(\mathbf{x})$. Key features of the members used in the modeling are described in the following subsections. Each subsection will end with some suggested reference for interested readers.

3.1. **Generalized Linear Models (GLM).** As explained in Section 1, GLM is the most popular predictive modeling technique in actuarial community. It was formally introduced in 1972 by John Nelder and Robert Wedderburn [35] and was intended to serve as a generalization of linear regression, logistic regression and Poisson regression, the major statistical models back then. Minor extensions, including generalized estimating equations, generalized linear mixed model and generalized linear interaction model, are proposed to address issues when independence assumption is violated.

The basis function is the individual variable x_i and the parameter β_i is estimated base on maximum likelihood approached. Together, $f_j(\mathbf{x}) = \beta_j x_j$ and,

$$F^*(\mathbf{x}) = g(\sum_{j=1}^{J} \beta_j x_j) \tag{4}$$

GLM is the only parametric member among all candidates we discuss from the Generalized Additive Family. Parametric modeling assumes the data comes from a type of probability distribution and makes statistical inferences about the parameters, with the form known apriori, of the distribution [18]. Independence and linearity assumptions are made to derive the parameters. If all the assumptions are correct, GLM can produce precise estimates. However, those assumptions are in general violated in actuarial pricing and results can be misleading in this situation.

Suggested reference: Anderson et. al. [1], Geisser and Johnson [18], Haberman and Renshaw [20], Hardin and Hilbe [22], Hastie et. al. [24], McCullagh and Nelder [32], Nelder and Wedderbun [35]

3.2. **Generalized Additive Models (GAM).** Developed by Hastie and Tibshirani [23], GAM were to extend GLM. The evolution is motivated by occasional unsatisfactory performance of GLM due to the linearity constraint. This constraint creates extreme predictions when the value of the explanatory variable at both ends of the the range and thus not desirable in extrapolation. Regularization techniques, LASSO, Ridge Regression and Elastic Net, are established alternatives that penalize high betas that trigger the issue. However, the improvement comes with the sacrifice of the overall predictive accuracy.

The GAM version of Equation 4 is

$$F^*(\mathbf{x}) = g(\sum_{j=1}^{J} \beta_j f(x_j)) \tag{5}$$

where $f(x_j)$ is some *smooth* function to be estimated. The most commonly used smoothing function is the spline function. A spline is made up of piece-wise polynomials that satisfy certain criteria. Details regarding the criteria can be found in the suggested reference. Since the smooth function can be in numerous forms, including a straight line, the assumption of linearity is relaxed. The parameter estimation method in Hastie and Tibshirani [23] was a back-fitting algorithm. In back-fitting algorithm, the formula and variables are specified prior to the model fi The values of parameters are iteratively adjusted until the loss/error converges. On the contrary, forward-fitting algorithm iteratively adds variables and adjustment to the formula and thus does not require upfront specifications. The mechanism is described in more details in Section 3.4

Suggested reference: Hastie and Tibshirani [23], Hastie et. al. [24], Wood [45, 46]

3.3. **Artificial Neural Network (ANN).** ANN has its origin in 1943. As indicated by its name, ANN is inspired by how our neural network works. Information/Input is decomposed into the various explanatory variables and stored in the neurons. The information from input neurons are then aggregated and processed by neuron in the next layer. The receiving neuron will serve as the input neuron for the following layer and the process will continue until the output layer is reached. Any layers between the input and output layers are called hidden layers.

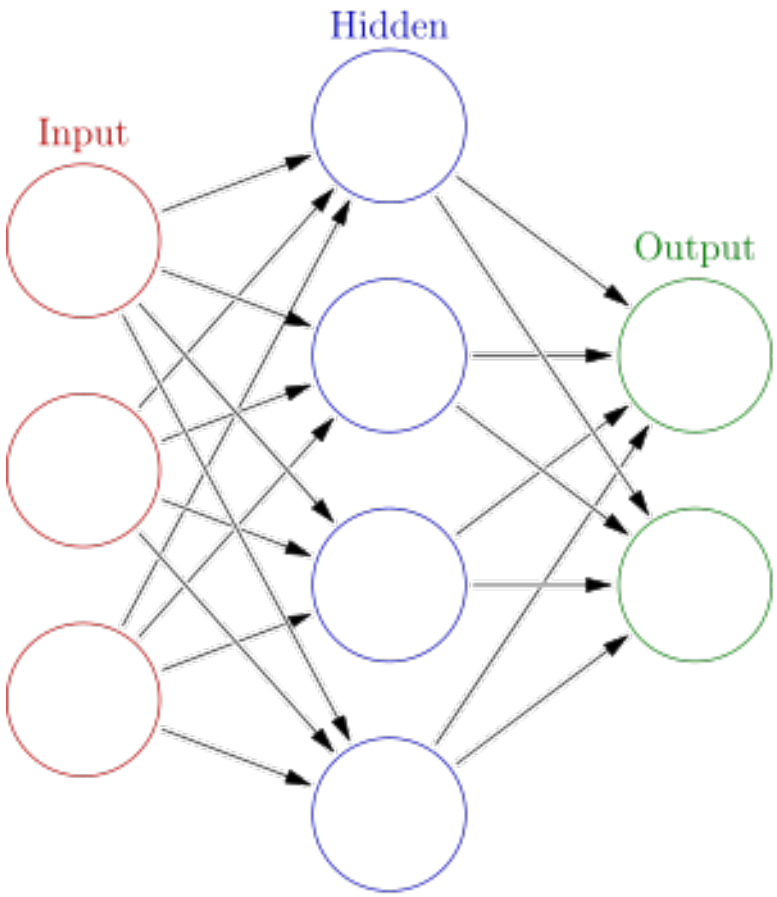

Figure 1. Concept of Neural Network from Glosser.ca

Equation 6 shows the mathematics for the information processing at a layer, with $g(\cdot)$ the activation function and w_i the information weight.

$$f_k^*(\mathbf{x}) = g(w_{k0} + \sum_{j=1}^{J} w_{kj}x_j) \tag{6}$$

$$F^*(\mathbf{x}) = g(w_0 + \sum_{k=1}^{K} w_j f_k^*(\mathbf{x})) \tag{7}$$

This candidate has had tremendous successes in the fi of pattern recognition and machine learning. Many of the international competition winners are variations of the original ANN. http://www.kurzweilai.net/how-bio-inspired-deep-learning-keeps-winning-competitions captures an interview with Jrgen Schmidhuber on the eight competitions won using ANN. The popularity is spread to Finance, Economics and Actuarial Science as well [2, 3, 11, 25, 26]. Vendor solutions are also available to actuaries.

Suggested Reference: Hastie et. al. [24], McCulloch and Pitts [33]

3.4. **Gradient Boosting Machine (GBM).** Adaptive Boosting [12] is the fi success of boosting algorithms. It became a popular classification tool in late 90's to early 00's. Brieman [5, 6] later explained the algorithm as a gradient descent approach with numerical optimization and statistical estimation. Friedman et. al. [15] further extend the idea and introduces a few similar sibling models for comparison. The crux of the algorithm is to iteratively minimize a transformed distance, deviance or other loss functions, between the actual observation and the corresponding prediction.

Friedman [16] proposes a boosting method called Gradient Boosting Machine. It extends the boosting capacity by featuring solutions to regression problems. It is considered to be a significant breakthrough as boosting was limited to classification before GBM. The algorithm successfully includes statistical elements, such as additive modeling and maximum-likelihood, in the modeling technique. By doing so, the authors were able to derive diagnostics to assess the quality of the predictions. The existence of the statistic based diagnostics substantially blurs the boundary between machine learning and traditional statistical modeling.

It is also shown in Hastie et. al. [24], using empirical examples, that GBM is the top-tier predictive model among data mining techniques. In today's world where computing power is less an issue, predictive power is clearly the top concern. The simplicity of the algorithm, critical in predictive modeling, is also inspiring more research that could lead to even more powerful extensions.

The GBM version of Equation 4 is

$$F^*(\mathbf{x}) = \sum_{t=1}^{T} f_t^*(\mathbf{x}) = \sum_{t=1}^{T} \beta_t h(\mathbf{x}; \mathbf{a}_t) \tag{8}$$

Although any weak rule, $f_t^*(\mathbf{x})$, alone would not be strong enough to make accurate predictions on all observations, it is possible to combine many of those rules to produce a highly accurate model. This idea is known as the **the strength of weak learnability** [39]. It was originated in the machine learning community with the introduction of *AdaBoost* [12, 13], the first major success of boosting algorithms.

The estimation of the parameters, β_t and $\mathbf{a}_t$, in (8) amounts to solving

$$\underset{\beta_t, \mathbf{a}_t}{\operatorname{argmin}} \sum_{i=1}^{N} \Phi\left(y_i, \sum_{t=1}^{T} \beta_t h(\mathbf{x}_i; \mathbf{a}_t)\right) \tag{9}$$

Boosting adopts the approach of forward stage-wise method that solves (9) by sequentially fitting a single weak learner and adding it to the expansion of previously fitted terms. The corresponding solutions of each new fitted term is not readjusted as new terms are added into the model. This characteristic is commonly called adaptive and is outlined in Algorithm 1 [15].

Algorithm 1 Forward Stagewise Additive Modeling

1: Initialize $F_0(\mathbf{x})$
2: **for** $t = 1$ to T **do**
3: Estimate β_t and $\mathbf{a}_t$ by minimizing $\sum_{i=1}^{N} \Phi(y_i, F_{t-1}(\mathbf{x}_i) + \beta_t h(\mathbf{x}_i; \mathbf{a}_t))$
4: Update $F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \beta_t h(\mathbf{x}; \mathbf{a}_t)$
5: **end for**
6: Output $\hat{F}(\mathbf{x}) = F_T(\mathbf{x})$

Suggested Reference: Brieman [7], Freund and Schapire [12], Friedman et. al. [15], Friedman [16, 17], Hastie et. al. [24], Lee [30], Ridgeway [37], Schapire [39], Sun et. al. [40]

3.5. **Delta Boosting.** Lee [30] proposes Delta Boosting Machine (DBM), a modification to GBM, that better utilizes the base learner. Lee proves that DBM is the optimal boosting algorithm for many common distributions and asymptotically optimal for the rest. A few empirical examples are illustrated in the paper to show how DBM outperforms GBM.

The primary diff between DBM and GBM is their sorting rules. In GBM, the gradient of each observation is used as the sorting element:

$$r_i = -\left[\frac{\partial \Phi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{t-1}(\mathbf{x})}, \quad i = \{1, \ldots, M\}$$

whereas DBM attempts to reduce the deviance to the maximum extent at each iteration and thus the minimizer also called delta is used in sorting:

$$\delta_i = \underset{s}{\operatorname{argmin}}\, \Phi(y, F_{t-1}(\mathbf{x}_i) + s), \quad i = \{1, \ldots, M\}$$

The ways on how the data is partitioned also vary slightly. Interested readers can fi details in Lee [30]

3.6. **Classification and Regression Tree.** Since its introduction in Brieman et. al. [4], Classification and Regression Tree(CART) has been incorporated to some degree in almost all analytical fi Either used standalone or as elements of ensemble methods, CART creates predictive model through iteratively partition the data through logical decisions. Figure 2 represents a typical CART output.

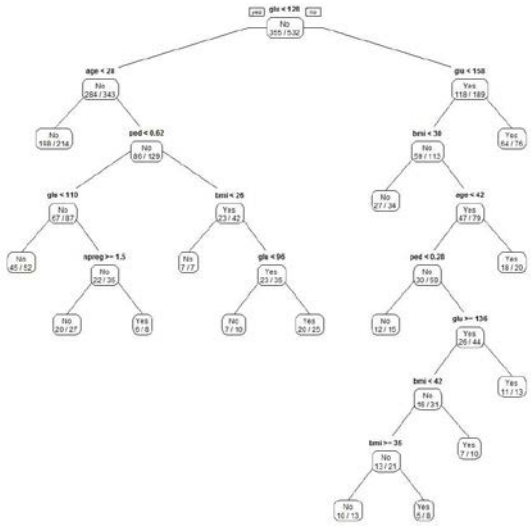

Figure 2. Typical CART output

The vast popularity of this technique is due to its appealing features. It is simple to understand and interpret, requires little data preprocessing, handles numeric, categorical and missing data, performs well on large data and provides statistical diagnostics.

Suggested Reference: Brieman et. al. [4], Hastie et. al. [24], Quinlan [36], Rokach and Maimon [38]

3.7. **Other modeling techniques.** There are many more techniques that actuaries can fi useful. To name a few, K-nearest neighbor (KNN), Multivariate Adaptive Regression Splines(MARS), Naive Bayes, Random Forest (RF) and Support Vector Machines (SVM) all have strong presence in various fi of application. However, we limit the scope of the paper to the analysis to the candidates described in the previous subsections.

4. The Data

This paper uses real-life data from a Canadian insurer. It consists of policy and claim information at the vehicle level for Collision coverage in a particular province. Collision coverage covers insured from the cost of repairing or replacing their vehicles in the event that the covered vehicles hit another vehicle, or any object in or on the ground.

The data set includes the experience for calendar/accident years 2001 to 2005. The response to be predicted is the claim frequency. The data includes 290,147 earned exposures and an overall claim frequency of 4.414%. Although the number falls into the typical industry range of 4% to 8%, this represents an imbalanced or skewed class distribution for the target variable under most of the modeling standard. This commonly hinders the detection of claim predictors and eventually decreases the predictive accuracy of the model [40]. Thus, a rigorous exploratory data analysis and variable selection process are more infl ential to the outcome of the modeling.

4.1. **Exploratory Data Analysis (EDA).** Actuaries should should consider an EDA as a significant component in any modeling process. As Tukey [41, 42, 43] defi EDA is an attitude, a fl y, and a reliance on display, NOT a bundle of techniques. It contrasts with standard predictive modeling techniques, also known as confirmatory data analysis, where the process is easier to computerize. The heart of EDA is the willingness to look for what can be seen, which cannot be replaced by any fancy predictive models.

As stated in the philosophy, there is no set rules on how to formalize the process. However, actuaries are encouraged to consult experts from underwriting, sales, IT and claims to get a comprehensive understanding of the inputs. Readers can fi more details about the questions actuaries commonly ask on [14, 44]. Actuarial Standards of

Practice Document Number 23 also describes actuaries' responsibility when using data from external sources.

The visual tools required are dependent on the problems and style of actuaries. Whilst it is not meant to be exhaustive, actuaries may fi the following tools handy for typical personal auto pricing purposes.

Scatter Plots: Laying the response variable against each explanatory variable is the most intuitive way to visualize the *dependence effects*. The graphs also provide insights to actuaries whether linear assumption is appropriate and if not, whether a transformation is appropriate.

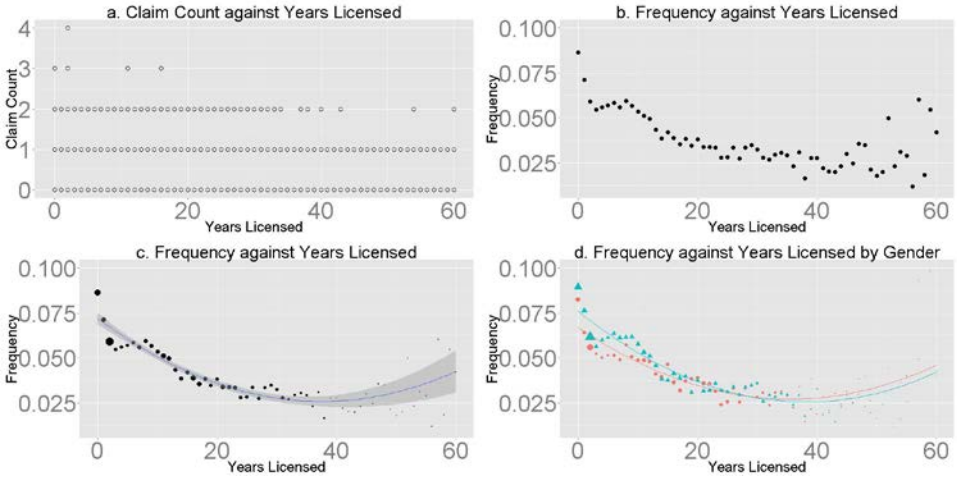

Figure 3. Scatter Plots

One key shortcoming of standard scatter plots stand out when a large data set is used. Figure 3a displays the scatter plot of the claim count against years licensed of insured. The graph does not reveal much information except for the range of claim counts (from 0 to 4) and years licensed (0 to 60). Figure 3b presents the same relationship at a summarized level. The claims counts aggregated at each level of years licensed. The pattern is more meaningful: the negative trend prevails until the far right end of the spectrum.

With the advancement of plotting techniques, actuaries can extract more information on the same plot. Figure 3c utilizes bubble plotting and simple smoothing techniques. The size of the each bubble (point) is proportional to the exposures aggregated. The scale is shown to further facilitate the assessment. Smoothing provides a simple visual illustration of the trend. Actuaries can choose the degree of polynomial to be fi In Figure 3c, a degree 2 polynomial is used. The 95%

confidence interval is also plotted in gray. Using these tools, actuaries can more confidently conclude that most of the observations follow the trend and the reversal of pattern may not be statistically credible.

We stretch the capacity further by plotting the pattern by gender of the insured in Figure 3d. The pattern reveals that female drivers generally incur fewer claims at every years licensed, but that the gap between men and women shrinks from 0 years licensed to around 35, where it begins to expand again.

While we can always add more information to the standard plots, it is dangerous to abuse the graphing capacity. If the graphs convey too much information to a point that actuaries are distracted from the key messages, the original purpose of using graphs as a simple and quick tool is lost.

Time Consistency Plots: One of the goals of actuarial pricing is to derive a breakeven loss cost for any risk that the insurer will underwrite in the future. The data usually comes from 3 to 5 years of experience; for catastrophic risk, the experience period can be extended to more than 10 years. Mahler [31] shows that the correlation of results drops fairly significantly as the time diff between the experience and prediction period increases. Thus, time consistency plot can be essential to decide whether the general pattern between explanatory variables and the response deteriorates through time. If the relationship is not consistent through time, it is customarily discarded from the modeling.

From Figure 4, the pattern is very consistent from AY3 to AY5. AY is an abbreviation of Accident Year and 5 indicates Year 2005. Since the consistency exists in the latest experience years, we are comfortable to select the negative trend for this factor. We should also consider if 3 years of experience should be used instead of 5 when similar pattern is observed for other key rating variables.

Histograms: Contrary to scatter plots, histograms are used to analysis the *exposure distribution*. Actuaries can examine if observations are clustered at a few levels or widely spanned, which helps actuaries to group observations if necessary.

From the assessment in Figure 3 and 5, actuaries may consider grouping observations with years licensed over 35 years. Readers should be warned that one-way analysis like EDA only provides a preliminary insight of how a variable should be handled. The pattern can be contaminated by dis-proportionate distribution of other variables. Using years licensed effect as an example, the reversal at the right end may likely be contributed by more senior drivers who are less physical sharp

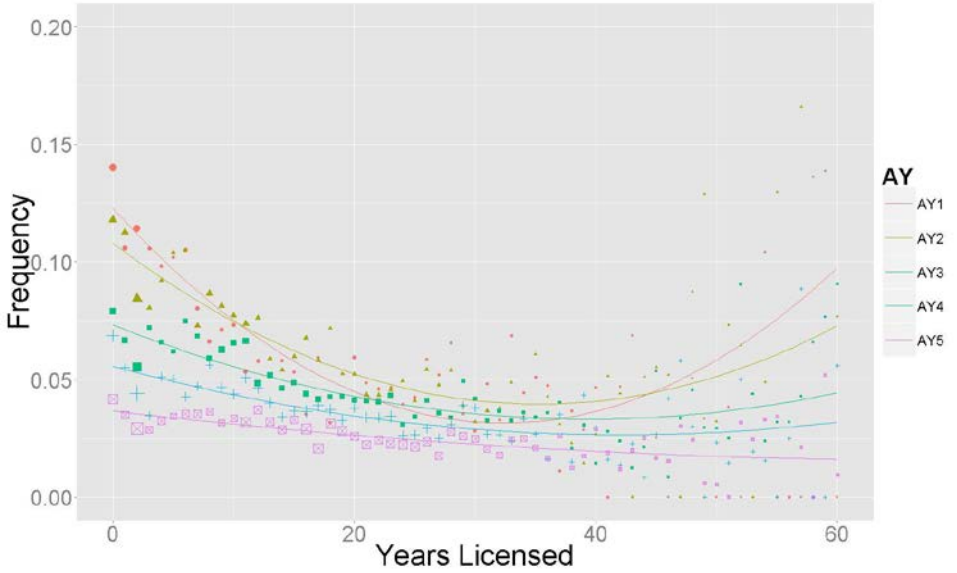

Figure 4. Time Consistency Plot for Years Licensed Effect

to handles road emergencies. If the model contains both variables, the reversal effect will be offset by increasing loss tendency by age.

Correlation Plots: Correlation Plots are sometimes known as scatter plot matrices. It is a collection of scatter plots, histograms and correlations. It aims at condensing all the mentioned graphs into a few pages. A popular correlation plot looks like the plot below.

Similar to the logic we adopted for scatter plots, we aggregated the exposure by the horizontal axes to help visual the underlying trend between each pair of variables. Driving record is derived based on years licensed, conviction and accident history. Thus, it is expected that the driving record is positively correlated with years licensed. Driver's age also trends the same direction with years licensed. The plot confirms the intuition nicely.

Principal Components: Just as shown in the Figure 6, many explanatory variables are correlated in auto pricing. This phenomenon adversely affects the quality of modeling in general as it likely induces extremely positive and negative coefficients that offset each other. Principal components(PC) are created to transform elements into linearly uncorrelated variables. The transformation is done through iterative eigen decomposition such that the early components always explains more variability of the data. Actuaries are thus able to compress the list of variables going into modeling stage by dropping the variables do not significantly explain the variability of data.

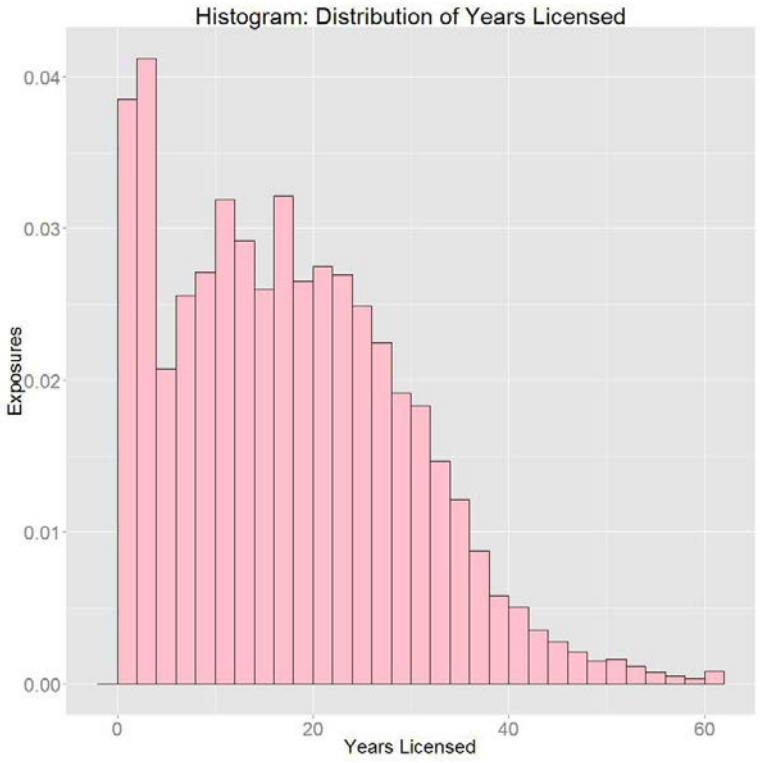

Figure 5. Histogram for Years Licensed Effect

Using the variables in correlation matrix as an example, the principal components derived are

	PC 1	PC 2	PC 3
Years Licensed	-0.67	0.73	0.12
Driving Record	-0.07	0.11	-0.99
Age	-0.74	-0.67	-0.02
Variability Captured	0.88	0.11	0.01

The fi 3 rows are the linear transformation vectors from original data to the Principal components. The last row shows that the fi PC accounts for 88% of the variability and the third PC only accounts for 1%. The exercise can be expanded to all numeric variables. Actuaries should note that PCA assume all variables are in the same scale. Using driving record as a counter example, while the numbering indicates the relative order, the diff between levels has no meaningful interpretation. Using PCA on this variable without adjustment may result in distortion of prediction.

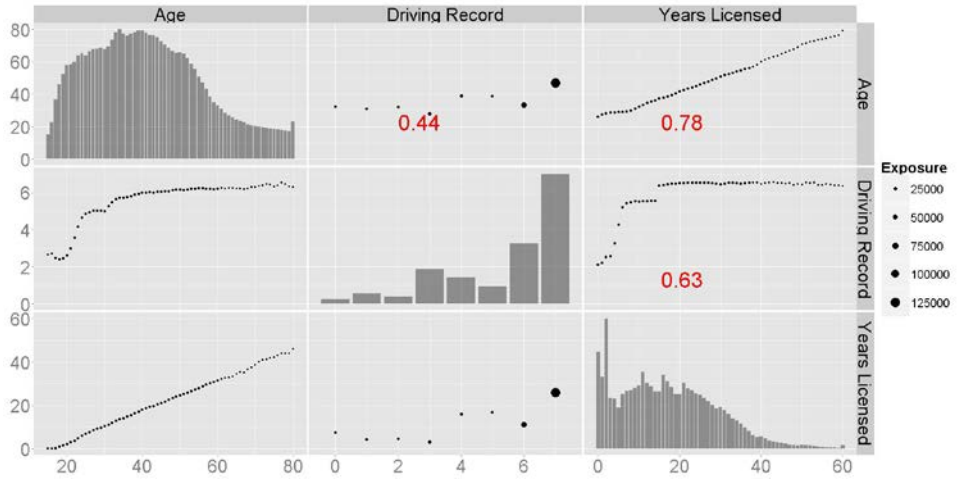

Figure 6. Correlation Plots for Years Licensed, Driving Record and Driver's Age Effect

4.2. **Preliminary Variable Selection.** A good EDA gives actuaries a clear picture about the data and which variables should be selected for modeling. A predictive variable usually exhibits a clear pattern against the response. The pattern needs not to be straight line or monotonic. For example, it is well known that driving behavior improves when insured ages from early stage of life due to the accumulation of driving experience. The behavior is reversed at older stage as senior drivers have a slower reaction time. We shortlist around 30 out of more than 200 variables for the modeling.

It should be noted that while this is not the only place that for variable selections, a robust selection should leave actuaries with only a handful of variable combinations for modeling. Leaving too many options to modeling stage will significantly lengthen the modeling process and likely result in overfitting eventually.

4.3. **Data with unusual values.** Real-life data is seldom perfectly clean. Contamination comes from inaccurate information from the insured, omission from the agents, system errors, mismatch from external source etc. The contamination in turns will affect the quality of results. Thus, a robust process to validate the quality of data is necessary.

Many mistakes can be easily spotted, negative years licensed and ages for examples. Thus, the fi step of validity check is the check if any levels are not prescribed in the systems. Other checks may require more understanding of data. The issues only become apparent when a few fi are jointly considered. For example, in most of the

jurisdictions, there is a minimum age requirement for obtaining driver license. In the jurisdiction that we are modeling, the minimum age is 16. Thus, an observation with a 20 year old driver with 8 years of driving experience should be investigated. While each fi looks reasonable when considered individually, obtaining license at age 12 does not look reasonable in this case. Other checks include, but are not limited to, checking

(1) the number of vehicles insured to confirm the discount on insuring multiple vehicles.
(2) the age of vehicle when it was purchased to confirm the model year is accurate.
(3) the number of years insured to confirm the years licensed.
(4) the address to confi the territory of the insured.

Missing Values: Some predictive modeling tools, built-in GLM package in R for example, require all fi to be complete. Actuaries thus have to react to the situation when the data consists of some missing values. There are a few common rules of thumb that attempt to solve the problem: Replacing the missing values with the fi average, creating a new variable that indicates the value for the observation is missing, fi a proxy to approximate the value from other variables or even deleting the observations. In the data that we are using, there are around 1200 observations, less than 0.15%, that have missing fi The observations are deemed insignifi t and deleted for this modeling purpose.

5. Collision Frequency Models

We use GLM, GAM, ANN, GBM, DBM and CART as the candidates for the frequency model. To compare the performance of the competing techniques, we partitioned the data into train (80%) and test (20%) data sets through random sampling. The train data is used for modeling while the test data is used as an independent source to verify the performance. For models that requires intermediate validation data, the train data is further split into pure train (80%) and validation (20%) of the train data. Since we have a ex-ante belief that the claim count follows the Poisson distribution, Poisson deviance is used as the basis of performance. The best model should have lowest deviance among the candidates in the holdout data.

The following table displays the variables selected for the modeling.

The names of all variables are masked in the rest of the paper, but that has no significant impact on understanding the results presented.

Table 1. Overview of Loss Cost predictors

Driver Characteristics	Accident/Conviction History	Policy characteristics	Vehicle characteristics
DC1. Age of principal operator	AC1. Number of chargeable accidents (6 years)	PC1. Years since policy inception	VC1. Vehicle Rate Group
DC2. Years licensed	AC2. Number of not chargeable accidents (6 years)	PC2. Presence of multi-vehicle	VC2. Age of Vehicle
DC3. Age licensed	AC3. Number of minor driving convictions (3 years)	PC3. Collision Deductible	VC3. Horse Power
DC4. Driving Record	AC4. Number of major driving convictions (3 years)	PC4. Use of Vehicle	VC4. Weight
DC5. Dwelling unit type	AC5. Number of criminal driving convictions (3 years)	PC5. Group business	VC5. Suggested Retailed Price of Vehicle
DC6. Aggregated Credit score	AC6. (Total, At-fault) Claims free years	PC6. Presence of multi-line	VC6. Number of Seats
	AC7. (Total, Minor, Major, Criminal) Convictions free years		VC7. Wheelbase
			VC8. Vehicle Age at Purchase

Table 2. Normalized deviance of competing models

Model	Train Deviance
GLM	0.00
GAM	-359.69
GBM	-808.60
DBM	-1173.52
ANN	4132.64
CART	1343052.49

5.1. **Initial Run.** Table 2 illustrates the deviance for all the candidate models and the result falls within expectation. CART in general serves as a quick solution for simple decision. While regularization techniques like pruning and conditional can significantly improve the log-likelihood, CART can seldom rank fi in the group. GLM has the most restricted assumptions among other candidates, leading to lowest test likelihood in general. GAM still assumes independence among explanatory variables, except specifically included prior to the modeling. However, the joint effect can be significant in many real-life data.

Readers might be surprised that ANN has a higher deviance than GLM. One main limitation of ANN is the infl y of loss function assignment. ANN models minimize least squares and thus the model

is not necessarily minimizing the Poisson deviance. Anderson et. al. [1] illustrates an example on how diff t loss functions impact the estimation of parameters.

Boosting performs best among the candidates, with DBM better than GBM. It resonates the results shown in Lee [30]. We can utilize the likelihood ratio test to give a sense on how to interpret the magnitude of deviance. For two competing models, one nested by another, the diff of deviance between the models follows chi-square distribution with the degree of freedom equal to the diff in number of parameters. For the diff between GLM and DBM considered to be statistically immaterial at 5% significance, DBM has to have at least 850 more parameters than GLM. For the case between GLM and GAM, GAM has to have at least 318 more parameters.

In addition to the predictive performance, actuaries should also examine the diagnostic tools available for each technique. For boosting, Lee [30] presents a rich variety of diagnostics that help actuaries to assess the diff t aspects of the model. Contrarily, ANN has only very few assessment tools, if any.

We can also assess the performance of the candidates by utilizing the lift plot. Lift is a popular diagnostic tool in predictive modeling due to its intuition. To derive the statistic of a model, we sort the prediction and group the observations into 10 deciles. Lift is defi to be the ratio of the mean of the response in the top decile and the mean of the bottom decile. A high lift implies the model's ability to diff tiate observations. In addition, lift plot is a plot of the average responses against the average prediction over the 10 deciles. If the points are aligned with the line $y = x$, the model has a high predictive performance. Thus, the slope and R^2 of the plot should both be assessed.

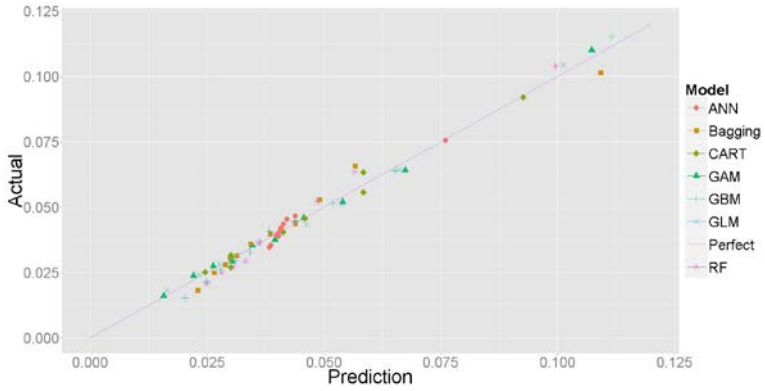

Figure 7. Lift Plots for Train data

Figure 7 shows that ANN has a high lift. However, overfitting exists at the tail. The model has a average prediction of 13.4% on the top decile. The corresponding actual frequency is 11.2%. Thus, results of ANN may be adequate for underwriting purposes as only the relative order among insured is required. However, good pricing requires fair assessment of all risks throughout the book of business. Thus, both the relative order and exact magnitude of the loss cost will be required in this case. Over pricing of risks will likely result in clients leaving the book in a competitive market and under pricing of risks will result in adverse selection. The same issue also applies on CART.

GLM, GAM, GBM and DBM predict the frequency fairly well on each decile as all the points are close to the $x = y$ line. GBM and DBM have a higher lift than GAM which in turn having a higher lift than GLM. The result is consistent with what the likelihood table suggests.

Since boosting, and ANN in certain aspects, performs better than GLM, it would be desirable if actuaries can use them to improve on the GLM model. However, ANN models data through multiple layers of neutrons and activation functions and inference on individual elements can hardly be extracted from this type of operation in a scientific way. Fortunately, there are many diagnostics available from boosting that are comparable to those commonly used in GLM.

Variable Importance: After assessing the overall performance of each model, actuaries should then focus on which variables exert more infl on the model performance. In GLM, we use the t-statistics, which is the ratio between the coefficients and the corresponding standard errors. However, it is not a perfect tool for categorical variables with $n > 2$ levels. In such case, $n - 1$ coefficients are derived, some maybe significant and some maybe not. Thus, the statistics describe the significance of parameters rather than variables themselves. Some actuaries may argue that it is a good enough proxy; however, in the case where all its levels have medium significance, the variable can become a significant one. Removing the variable simply based on individual value can results in significant deterioration of model predictive power.

The boosting models, on the contrary, assess the elements at variable level. The likelihood improvement at each iteration is assigned to the each variable and improvements of all iterations are then aggregated.

Figure 8 is a standard representation of relative importance of the boosting members. The importance is normalized such that the sum equals 100 for easier comparison. To see how the predictive performance by variables are aligned by diff t models, we put the corresponding ranking of GLM to the variable importance table of DBM.

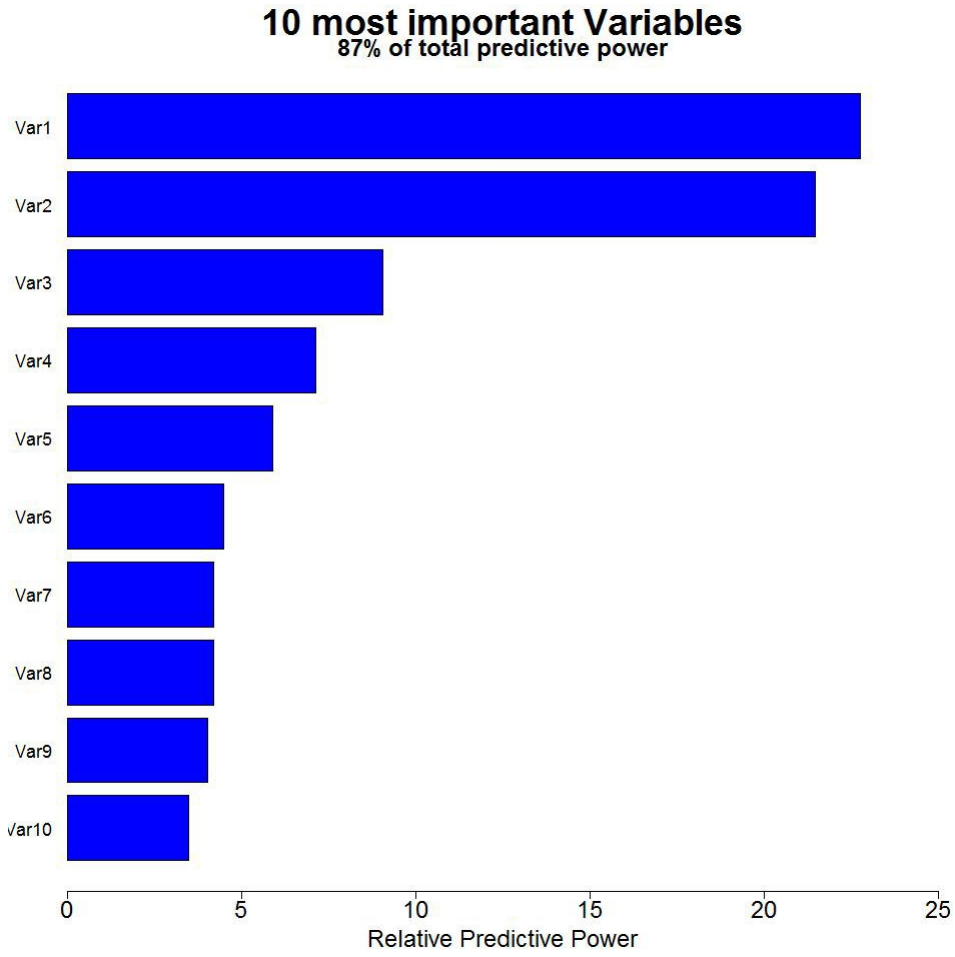

Figure 8. Lift Plots for Train data

Table 3. DBM Importance and corresponding GLM ranking

Variable	Importance	GLM Rank
Var1	22.80	7
Var2	21.50	1
Var3	9.10	20
Var4	7.10	3
Var5	5.90	2
Var6	4.50	15
Var7	4.20	11
Var8	4.00	6
Var9	3.50	13
Var10	3.10	9

Perhaps surprisingly, the rankings of the variables are fairly inconsistent. It gives actuaries a great opportunity to improve the GLM

models by comparing the diff tial plots of those variables that show vastly diff t rankings between models.

Marginal Differential Plot: We illustrate the diff tial plot of a variable that shows significantly diff ent ranking between GLM and DBM in Figure 9. The diff ntial diff very dramatically. It is likely due to the existence of another variable that has an co-linearity effect to this variable. GLM and GAM are not effective in this type of situation because coefficients are derived by inverting matrices. If the columns of the matrix are linearly dependent, then the matrix is singular and cannot be inverted. In the case where the column vectors are almost linearly dependent, then the inverse will have a high coefficients similar to what the graph indicates.

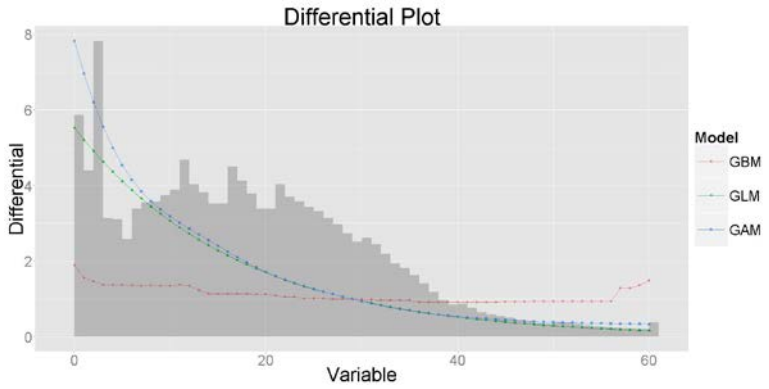

Figure 9. Differential Plot

A thorough investigation confirms that there are 2 other variables that jointly co-linear to the variable. They each has a high diff rential relative to what suggested by boosting. A solution is to create an interaction effect in GLM and GAM that tempers the multiplication effect.

In some occasions, the signs of the parameters may be counter intuitive. Actuaries should be extra cautious on this. There are two main causes of this situation: insignifi t magnitude or co-linearity. The fi cause can be handled by artificial assignment of fi factors after the modeling exercise is complete. The second cause requires more effort. If all of the highly correlated variables have to be kept in the rating algorithm, an interaction term may need to be introduced. An alternative is to impose an offset on one of the variables and re-run the model with the offset. Actuaries may also fi redefining or removing the variables necessary to guarantee an interpretable algorithm.

Joint Differential Plot: Besides co-linearity, boosting also has a built-in mechanism to addresses the interaction effect in the data. Friedman [16, 17] invents a H-statistic that quantify the interaction among variables. Using this mechanism, we calculate H-statstics for each pair of variables. The below shows the two that has the highest H-statistics.

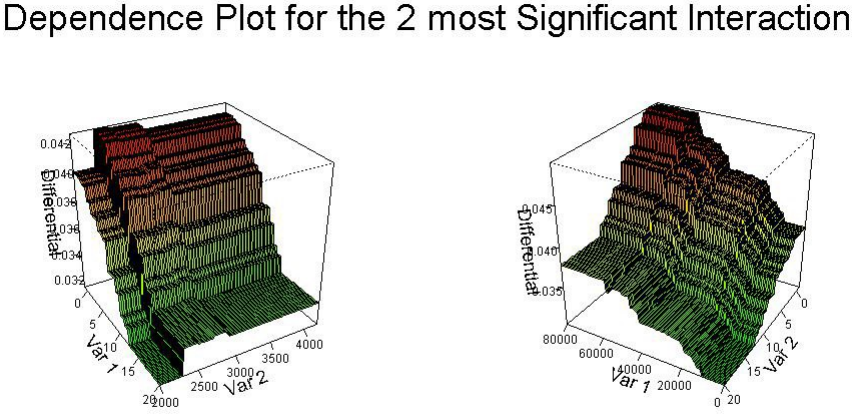

Figure 10. Lift Plots for Train data

The left panel of Figure 10 shows that the spread of var2 diff tial is sharply increasing as var1 increases. Adding a parameter that captures this effect will likely improve the predictive power of the GLM.

The right panel shows another type of interaction. While the dependence is negligible in most part of the range, the diff tial performs diff tly when var 1 is high and var 2 is low. Adding a parameter that for the block [70000,80000] X [5,20] should essentially capture the interaction.

5.2. **Final Run.** The magnitude of improvement decreases as we drill down to the less significant variables or interactions. Over-fitting may also result when one attempts to temper the model too much. We consider reviewing 5 to 10 most significant variables and including 2 to 4 interactions should warrant a model that is adequate for most pricing problems. Actuaries may also fi that several iterations of revision may be necessary as parameters of all variables will be revised to reflect new composition of the formula.

We run the GLM again following the process described in the previous subsections according to the iterative philosophy. At each iteration,

improvement in deviance is observed after modifi ions. Using likelihood ratio test, where it is applicable, improvements are significant. The model turns out having considerably fewer parameters than the initial one due to removal of many highly correlated variables to reduce the co-linearity effect.

We can now apply the results for all the candidates on the holdout data. Readers should be warned test data is the one and only independent benchmark in comparing all candidates and thus should not be accessed until results of all the model candidates are fi Else, actuaries selection will be biased by what is indicated by the test data. The original intent of setting aside the data will be lost.

Table 4. Deviance of competing models by partition of data

Model	Train Deviance	Test Deviance
GLM	0.00	0.00
GAM	-359.78	-70.92
GBM	-808.60	-148.21
DBM	-1173.52	-188.66
ANN	4132.64	1054.41
CART	1343052.49	336436.83
GLM Final	-82.93	-30.00

As Table 4 suggests, the deviance of fi GLM model improves significantly over the one at the initial run. For readers reference, for the improvment to be deemed as statistically insignifi t at 5%, the fi version has to have 70 more parameters than the original one, assuming the original is nested by the fi one. However, as stated, the fi version indeed has fewer parameters than the original version. The improvement in test data also confirms the superiority of the fi model. Actuary should conclude the modeling with one last re-run. The fi run models the full set of data to maximize the utility of the data.

6. Concluding Remark

We provide a simplified framework for actuarial pricing which involves data cleaning, exploring and modeling. Data quality is the key to the success of actuarial pricing. Data cleaning is the fi gate for actuaries to understand the inputs. It requires actuaries to have assess to various functional experts within the their organizations.

Actuaries should not ignore the importance of exploratory data analysis. Many tools are available to actuaries to visualize the interdependence among data at no cost. The analysis will provide crucial clues on

whether certain variables should be included and how they should be included, whether with transformation or interaction, in the modeling.

The complexity of data modeling is usually overlooked. Simply relying on the output of one model is generally insuffi t. Inference from other models can help actuaries to tailor the pricing algorithm to best describe the data behavior. With careful investigation and modifi the improvement can be highly significant. In addition, the modeling procedure is usually recursive; modifi of the model should be done one variable at a time. It helps actuaries to visualize the impact of each change and actuaries may be rewarded by observing new inference that can further improve the modeling accuracy. Once comfortable with the fi model, actuaries should verify the results with an independent holdout data. It provides an additional layer of confidence to actuaries if the results go as expected.

While the framework of a pricing process is presented, it is not meant to be exhaustive. Actuaries will fi modifi are necessary to reflect the nature of diff rent problems. For example, analysis of rating territories can be best visualized through a contour plot. Compound modeling is required where the geo-spatial residual is used for spatial smoothing. When new deductibles or limits are introduced, actuaries will be required to produce exposure rating or increase limit factor analysis to derive the appropriate diff tials for the new levels.

References

[1] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D. Schirmacher, E. and Thandi, N. (2007), *A practitioner's guide to generalized linear models*. CAS Exam Study Note Casualty Actuarial Society, CAS Exam 9 Syllabus 2010

[2] Boero, G. and Cavalli, E. (1996), *Forecasting the Exchange Rage: A Comparison Between Econometric and Neural Network Models*. AFIR, Volume II, 981

[3] Brockett, P., Cooper, W., Golden, L. and Pitaktong, U. (1994) *A Neural Network Method for Obtaining an Early Warning of Insurer Insolvency,*. Journal of Risk and Insurance, Volume 61, **3**, 402.

[4] Brieman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, CRC Press.

[5] Brieman, L. (1996), *Bagging Predictors*. Machine Learning, **24**, 123–140

[6] Brieman, L. (1998), *Arcing classifiers (with discussion).*. Annal. Statistic. , **26**, 801–849

[7] Brieman, L. (2001), *Statistical modeling: the two cultures*. Statistical Science, **16**, 199–231

[8] Brockman, M. and Wright, T. (1992), *Statistical motor rating: making effective use of your data*. Journal of the Institute of Actuaries, **119**, 457–543

[9] Chan, P. and Stolfo, S. (1998), *Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection*. Proceedings of ICKDDM, **4**, 164–168

[10] Chapados, N. Bengio, Y., Vincent, P., Ghosn, J., Dugas, C., Takeuchi, I. and Meng, L. (2001), *Estimating car insurance premia: a case study in high-dimensional data inference*. University of Montreal, DIRO Technical Report, **1199**

[11] Francis, L. (2001), *Neural networks demysnfied*. Casualty Actuarial Society Forum, Winter 2001, 252–319.

[12] Freund, Y. and Schapire, R. (1996), *Experiments with a new boosting algorithm*. Proceedings of ICML, **13**, 148–156

[13] Freund, Y. and Schapire, R. (1997), *A decision-theoretic generalization of online learning and an application to boosting*. J. Comput. System Sciences, **55**

[14] Friedland, J. (2010), *Estimating Unpaid Claims Using Basic Techniques*. Casualty Actuarial Society, Third Version, July 2010. Chapter 3–6

[15] Friedman, J, Hastie, T. and Tibshirani, R. (2000), *Additive logistic regression: a statistical view of boosting*. The Annals of Statistics,

28, 337–407
[16] Friedman, J. (2001), *Greedy function approximation: a gradient boosting machine*. The Annals of Statistics, **29**, 1189–1232
[17] Friedman, J. (2002), *Stochastic gradient boosting*. Computational Statistics & Data Analysis, **38**, 367–378
[18] Geisser, S.; Johnson, W. (2006), *Modes of Parametric Statistical Inference*, John Wiley & Sons
[19] Guelman, L. (2012), *Gradient boosting trees for auto insurance loss cost modeling and prediction*. Expert Systems with Applications, **39**, 3659-3667.
[20] Haberman, S. and Renshaw, A. (1996), *Generalized linear models and actuarial science*. Journal of the Royal Statistical Society, Series D, **45**, 407–436
[21] Hand, D., Blunt, G., Kelly, M. and Adams, N. (2000) *Data Mining for fun and profit*. Statistical Science, **15**, 111-131
[22] Hardin, J. and Hilbe, J. (2001) *Generalized Linear Models and Extensions*. Stata Press
[23] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*. Chapman & Hall/CRC
[24] Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The elements of statistical learning*, Springer.
[25] Huang, C., Dorsey, E. and Boose, M. (1995), *Life Insurer Financial Distress Prediction: A Neural Network Model*, Journal of Insurance Regulation, Winter, Volume 13, **2**, 131–167.
[26] Ismael, M. (1999), *Prediction of Mortality and In-hospital Complications for Acute Myocardial Infarction Patients Using Artificial Neural Networks*. Ph.D. Dissertation, Duke University.
[27] Kolyshkina, I., Wong, S. and Lim, S. (2004), *Enhancing generalised linear models with data mining*. Casualty Actuarial Society 2004, Discussion Paper Program
[28] Lee, C.K. Simon and Lin, Sheldon. (2010), *Modeling and Evaluating Insurance Losses via Mixtures of Erlang Distributions*. North American Actuarial Journal, **14**, 107–130
[29] Lee, C.K. Simon and Lin, Sheldon. (2012), *Modeling dependent risks with multivariate Erlang mixtures*. ASTIN Bulletins, **42**, 1–28
[30] Lee, C.K. Simon (2014), *Delta Boosting Machine*. under revision
[31] Mahler, H. (1990), *An Example of Credibility and Shifting Risk Parameters*. PCAS LXXVII, 225–282
[32] McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman and Hall.

[33] McCulloch, W. and Pitts, W. (1943). *A Logical Calculus of Ideas Immanent in Nervous Activity* Bulletin of Mathematical Biophysics, 5, **4**, 115-133.

[34] Meyers, G. and Cummings, D. (2009), *"Goodness of Fit" vs. "Goodness of Lift"*. The Actuarial Review, **36–3**, 16–17

[35] Nelder, John and Wedderburn, Robert (1972) *"Generalized Linear Models"*. Journal of the Royal Statistical Society. Series A 135, **3**, 370-384

[36] Quinlan, J. , (1986), *Induction of Decision Trees*. Kluwer Academic Publishers

[37] Ridgeway, G. (2007), *Generalized boosted models: a guide to the gbm package*. http://cran.r-project.org/web/packages/gbm/index.html

[38] Rokach, Lior; Maimon, O. (2008), *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc.

[39] Schapire, R. (1990), *The strength of weak learnability*. Machine Learning, **5**, 197–227

[40] Sun, Y., Kamel, M., Wong, A. and Wang, Y. (2007), *Cost-sensitive boosting for classification of imbalanced data*. Pattern Recognition, **40**, 3358–3378

[41] Tukey, John W. (1972), *Some Graphic and Semigraphic Displays*. Iowa State University Press

[42] Tukey, John W. (1977), *Exploratory Data Analysis*. Addison-Wesley. ISBN 0-201-07616-0.

[43] Tukey, John W. (1980), *We need both exploratory and confirmatory*. The American Statistician 34, **1**, 23–25

[44] Werner, G, and Modlin, C. (2010), *Basic Ratemaking*. Casualty Actuarial Society, Fourth Edition, October 2010, Chapter 4

[45] Wood, S. (2000), *Modelling and smoothing parameter estimation with multiple quadratic penalties*. Journal of the Royal Statistical Society: Series B 62, **2**, 413–428

[46] Wood, S. (2006), *Generalized Additive Models: An Introduction with R*. JAn Introduction with R. Chapman & Hall/CRC