
Machine Learning vs. Traditional Methods

Summary Document

March 2009

by:
Dr. Paul Beinat



The most commonly used analytics software in the Property and Casualty (P&C) industry, the Generalized Linear Model (GLM) is over 30 years old. During these 30 years there has been a considerable advance in technologies in all areas of human endeavour. What is particularly relevant is the development of totally new analytic fields such as machine learning¹. Raising the question: Do the methods from machine learning produce a stronger signal and better fit than GLM?

To get an answer to this question we:

- Selected a personal automobile portfolio.
- Had an independent actuarial consulting firm rigorously apply a GLM to each of the claim types (7) without having the constraint of having to file rates. The GLM software employed was the most commonly used GLM software (in the US).
- Machine learning methods² were then applied to the premiums derived using the seven GLMs in order to find any residual signal.
- The machine learning methods were applied with the objectives of deriving:
 - Discrete values.
 - Continuous values (Insurance Scores).

Derivation of Discrete Values

In this experiment, there are two years of data. One year of data will form the training data while the other will be used for validation. Only the variables present in the original data have been used and the data is not supplemented in any way with additional variables.

Table 1 below displays the results on the training data. One can see that the segment loss ratios can be regarded as a piecewise constant function, in much the same manner as most GLM relativities.

¹ “Machine learning is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to improve their performance over time based on data, such as from sensor data or databases. A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from data. Hence, machine learning is closely related to fields such as data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science.” Wikipedia, http://en.wikipedia.org/wiki/Machine_learning

² The primary methods applied came from machine learning, however, no machine learning method that we are aware of can be applied directly to insurance data, things like the variability of exposures, different record for exposures and classism, the noise in the data, credibility requirements, etc, generally preclude “native” machine methods from working. What was applied in each instance was a cocktail of machine learning, mathematics, statistics and actuarial science.

Segment	Exposure	GLM Total Premium	Claims	Claim Count	Training LR
1	40,088	9,677,889	7,223,230	5,730	75%
8	26,642	8,770,620	7,454,508	4,717	85%
3	35,946	8,036,238	7,298,945	5,178	91%
4	20,954	6,699,637	6,353,455	3,664	95%
6	26,212	6,754,957	6,534,512	4,127	97%
0	29,558	7,868,872	8,109,686	5,018	103%
9	20,049	5,636,667	5,935,182	3,576	105%
2	33,043	10,830,010	11,614,780	6,287	107%
7	23,203	8,181,896	10,125,938	4,356	124%
5	30,163	7,419,663	9,590,068	5,081	129%

Table 1

There is a significant loss ratio lift across these segments (segments have been given numbers by the software which are not related to their loss ratio performance). Given the number of claims in each segment, this loss ratio signal is likely to be related to some fundamental feature that is contained in the data. However, it could be that the algorithm has chosen the segment boundaries so that they overfit the training data. The only way to determine whether this is a feature of this insurance portfolio or merely a chance alignment of the segment boundaries is to test the signal on validation data. Table 2 shows the same statistics for the validation data.

Segment	Exposure	GLM Total Premium	Claims	Claim Count	Validation LR
1	39,262	9,511,229	7,767,501	5,913	82%
8	20,083	6,415,686	5,565,564	3,784	87%
3	35,105	7,505,323	6,283,145	5,073	84%
4	15,379	4,749,230	4,195,864	2,822	88%
6	29,387	6,935,811	7,187,731	4,688	104%
0	33,141	8,311,156	8,171,977	5,761	98%
9	20,488	5,266,095	5,748,663	3,720	109%
2	34,729	10,911,435	12,336,791	6,768	113%
7	24,679	8,140,954	9,532,883	4,641	117%
5	25,717	5,925,355	7,358,740	4,570	124%

Table 2

An inspection of the performance on the validation data shows reasonably similar loss ratios. The correlation between the training and validation loss ratios for the segments is 0.93; the loss ratios found in training are a consistent feature of this portfolio. Figure 1 shows graphically the two sets of loss ratios.

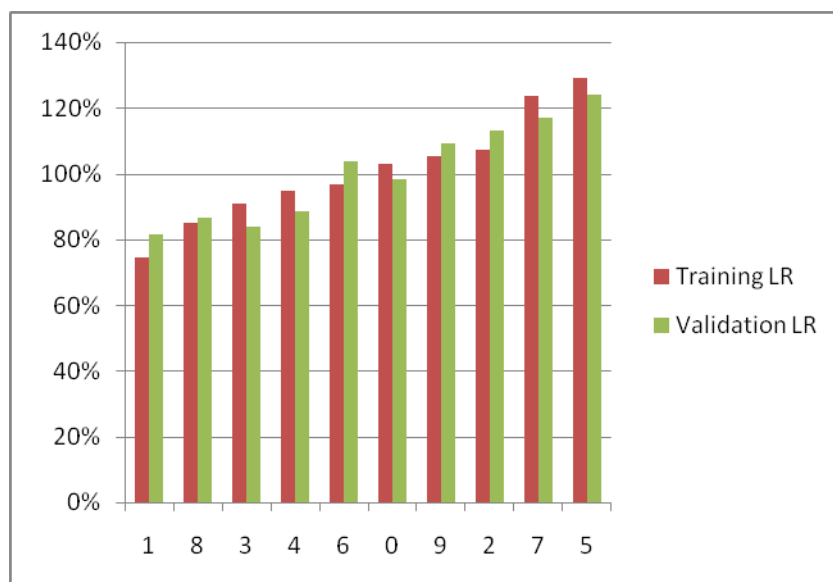


Figure 1

The chart shows that the lift in the training loss ratios is larger than for the validation loss ratios. This has been measured by taking the exposure weighted standard deviation of the two sets of loss ratios. The training data standard deviation is 16.4% of loss ratio while for the validation data it is 14.5%. This demonstrates that while the segment boundaries have been derived from the training data, and are in this sense tuned to the patterns in the training data, they lose very little of their discriminating power on the validation data.

It is reasonable to expect that by using relativities calculated from the training data loss ratios and applying them to derive new premiums for the validation data that these new premiums should fit better than the premiums derived by the series of GLMs (note that the GLMs were derived using all of this data and have an unfair advantage as they will have been influenced by the validation data).

Segment	Risk Cost	GLM Performance				Machine Learning Performance			
		GLM Premium	Deviance	Squared error	Chi squared error	Estimated Premium	Deviance	Squared Error	Chi squared error
1	197.83753	242.25011	44.41259	1972.478	8.14232	180.8068	17.0307	290.0446	1.604168
8	277.12184	319.45131	42.32948	1791.785	5.608944	271.5147	5.607135	31.43996	0.115795
3	178.98199	213.797	34.81501	1212.085	5.669326	194.182	15.19998	231.0394	1.189809
4	272.83309	308.81537	35.98228	1294.724	4.192551	292.8583	20.02523	401.0097	1.369296
6	244.59126	236.01866	8.572601	73.48949	0.311372	228.3163	16.27496	264.8742	1.16012
0	246.58141	250.78101	4.199596	17.6366	0.070327	258.4557	11.87434	140.9998	0.545547
9	280.58117	257.02797	23.55321	554.7536	2.15834	270.6401	9.941116	98.82579	0.365156
2	355.23425	314.19152	41.04272	1684.505	5.361396	336.9586	18.27563	333.9986	0.991216
7	386.2787	329.87683	56.40188	3181.172	9.643514	408.2565	21.97783	483.0251	1.183141
5	286.14786	230.41004	55.73782	3106.705	13.48338	297.8097	11.66185	135.9988	0.456663

Table 3

The GLM Premium column represents the per exposure premium as calculated by the collection of GLMs. The Estimated Premium column contains premiums calculated by applying a relativity derived from the training data to the GLM premium, and it is also expressed as per exposure value. The risk cost column contains the total claims incurred divided by the corresponding exposure. One can see that the deviance, squared error and chi squared error columns (that have the exposure weighted values for these statistics) have more favorable values for the estimated premiums than for the GLM premiums. Table 4 details the exposure weighted average of these measures.

	Deviance	Squared error	Chi squared error
GLM Premiums	34.15388	1463.838	5.47708
Estimated Premiums	15.02064	243.8955	0.946702

Table 4

These statistics indicate that the premiums developed by the inclusion of one additional set of relativities, derived from the segments found from the training data, produce premiums that fit the losses better than the premiums derived from the collection of GLMs. The deviance is more than halved, while the other measures show a much greater improvement.

Derivation of Continuous Values (Insurance Scores)

To develop continuous values, we again use the automobile portfolio from the previous experiment on derivation of discrete values. The same training and validation data sets are used. The goal will also be the same, to find a signal in the residuals left by the collection of GLM and we will similarly express this as a loss ratio based on the GLM premiums. In this experiment, the quasi-continuous output is mapped into a range [0, 1000] with 0 being the highest loss ratio and 1000 being the lowest loss ratio. The outputs have been binned into ranges in order to aggregate a sufficient amount of experience so that reliable estimates can be calculated.

Table 5 below shows the output ranges and corresponding experience data for the training data.

Output Range	Exposure	Premium	Claims Cost	Claim Count	Loss Ratio
0 - 302	7324	1,802,287	3,949,415	1301	219%
303 - 414	14569	4,079,891	5,779,738	2440	142%
415 - 553	20790	5,756,714	8,394,295	3537	146%
554 - 704	173637	52,046,796	51,689,134	26266	99%
705 - 722	27797	7,020,762	4,640,641	3295	66%
723 - 736	19114	4,439,842	3,151,909	2100	71%
737 - 749	14143	3,092,726	1,724,326	1395	56%
750 - 1000	8743	1,707,370	969,907	769	57%

Table 5

It can be seen that the resulting loss ratios have a significantly wider range than those for the derivation of discrete variables experiment. There can be reversals of loss ratio by output values. If the output value increases then so does the loss ratio. This can be seen between the second and third rows in Table 10. This could be masked by altering the boundaries between output ranges, but no attempt has been made to do this. Finally, the distribution of exposures is not uniform between the output ranges, and the number of claims is also sparser at the extremes. Figure 2 below shows the distribution of exposure from the training data with respect to the output ranges.

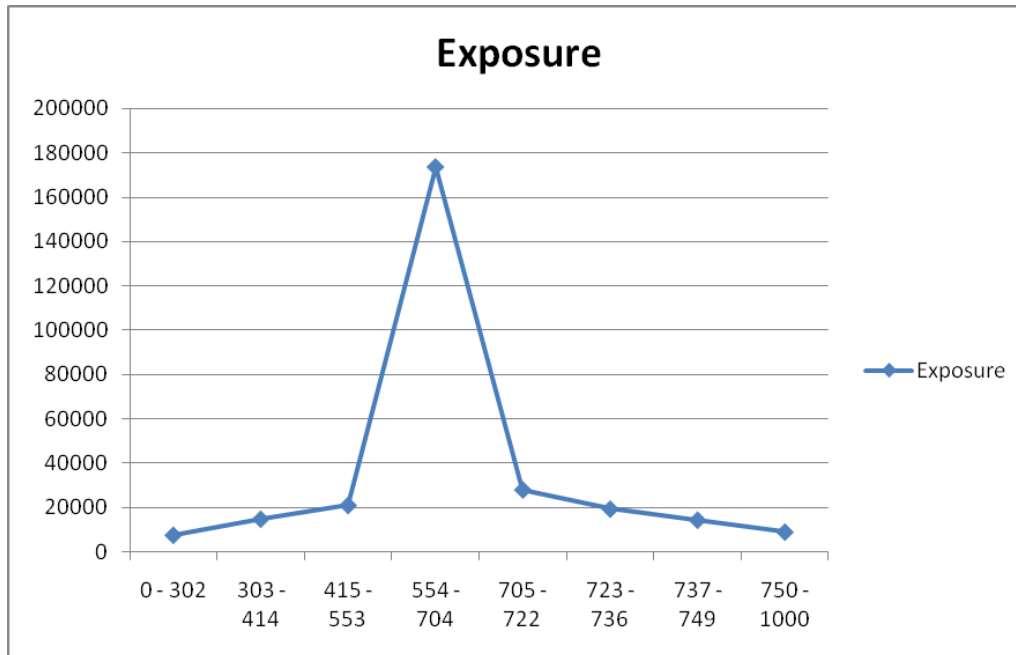


Figure 2

The output has been divided into a series of ranges using two competing criteria – the largest range in signal (in this case this being loss ratio) and a quasi-normal distribution of exposure.

The performance of the signal found on the training data is tested against the validation data. Table 6 below contains the performance of the output ranges on the validation data. Figure 3 shows the loss ratios for given output ranges between training and validation.

Output Range	Exposure	Premium	Claims Cost	Claim Count	Loss Ratio
0 - 302	9598	2,599,238	5,468,544	1901	210%
303 - 414	13563	3,804,456	6,090,286	2666	160%
415 - 553	18323	4,985,779	6,285,874	3142	126%
554 - 704	163770	46,299,639	44,418,782	24868	96%
705 - 722	28430	6,788,571	5,254,515	3429	77%
723 - 736	20059	4,400,802	3,421,952	2305	78%
737 - 749	14880	3,081,390	2,109,845	1569	68%
750 - 1000	9644	1,788,207	1,156,329	899	65%

Table 6

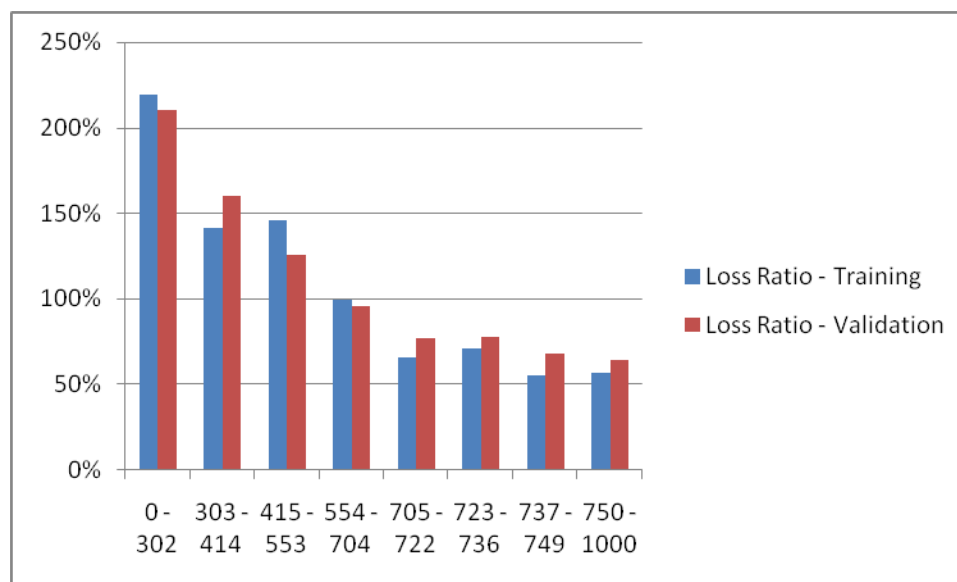


Figure 3

The correlation of the loss ratios is 0.98 and the standard deviation of the loss ratios is 30% and 29% for training and validation respectively. The trends found in training are therefore very evident in the validation data, but the validation data exhibits a slightly more moderate loss ratio signal. The other measure that we are interested in is how well the premiums fit the losses. Similarly to the previous experiment, the exposure weighted deviance, squared error and chi squared error are computed for the pure premiums estimated by the collection of GLMs versus the additional relativities that we calculate from the training data and apply to the validation data. These statistics are calculated using the output ranges from the tables above. Table 7 below shows the comparison of these statistics.

	Deviance	Squared Error	Chi Squared Error
GLMs	44.75	5722	21.74
Output Ranges	18.23	528	1.99

Table 7

The weighted deviance measure is improved by approximately 60% for the output ranges experiment, while the other measures are far more significantly improved (about 90% improvement). The stratification of the portfolio by the quasi-continuous output produces relativities that can be employed on validation data to demonstrate that these new premiums fit better than the ones originally developed by the collection of GLMs, for which the validation was used for training.

Final Considerations

GLMs are attractive. There is a considerable body of work on these methods; there are practitioners who are experienced at using them; and there is a set of model statistics which describe the nature of the fit to the training data. The issue is not whether what was originally a pencil and paper method has been extended and exploited using modern computers, but the accuracy of the results produced by the

modern version of GLMs. The results presented indicate that other methods can be used to gauge how accurately GLMs can model insurance data.

One potential reason for the relatively poor fit of GLMs, compared to other methods, is that the assumptions underlying GLMs may not be sufficiently realistic. There is no requirement for the real world to behave in any particular manner. Claims incidence need not be Poisson distributed, regardless of how aesthetically desirable we find the Poisson function. Similar remarks can be made for the Gamma distribution which forms the basis of claim severity models.

The use of machine learning methods as a supplement to GLM derived theoretical pure premiums demonstrates that compound variables are a very real feature of insurance modelling. Traditionally, compound variables have been derived by combining all the values, or bins of values, of two or more variables to derive a new compound variable. In contrast, the results presented are driven by locally acting compound variables, where only specific values of variables have been found to act in concert with specific values of other variables. In most cases, a significant number of variables have been found to act together in this manner; the involvement of five or more variables in forming a locally compound variable has been found frequently. The predictive accuracy of these locally compound variables has been verified using a validation approach and they offer very real improvements in accuracy.

A modification of machine learning methods to produce a quasi continuous output has also been tested. While the true nature of variable interactions has been lost due to the process used, the signal found has a greater contrast at the extremes and fits the validation better than the results of the discrete derivation. This also demonstrates, for the portfolio data used, that the GLM modelling has not captured a significant fraction of the signal present in the underlying data.

Conclusion

The methods originating from machine learning provide a better signal and fit than the optimally fitted GLM. Essentially, the common sense test applies – *“modern methods work better than traditional methods”*.

If you would like more information please contact:

Sales Department

803-726-7214

salesinfo@eeanalytics.com