Estimating Car Insurance Premia: a Case Study in High-Dimensional Data Inference

DIRO Technical Report #1199

Nicolas Chapados, Yoshua Bengio, Pascal Vincent,
Joumana Ghosn, Charles Dugas, Ichiro Takeuchi,
Linyan Meng

University of Montreal, Dept. IRO, CP 6128, Succ. Centre-Ville,
Montréal, Québec, Canada, H3C3J7
{chapados,bengioy,vincentp,ghosn,dugas,takeuchi,mengl}@iro.umontreal.ca

3 July 2001

Abstract

Estimating insurance premia from data is a difficult regression problem for several reasons: the large number of variables, many of which are discrete, and the very peculiar shape of the noise distribution, asymmetric with fat tails, with a large majority zeros and a few unreliable and very large values. We introduce a methodology for estimating insurance premia that has been applied in the car insurance industry. It is based on mixtures of specialized neural networks, in order to reduce the effect of outliers on the estimation. Statistical comparisons with several different alternatives, including decision trees and generalized linear models show that the proposed method is significantly more precise, allowing to identify the least and most risky contracts, and reducing the median premium by charging more to the most risky customers.

1 Introduction

A successful application of learning algorithms to real-world industrial problems often requires considerable creativity in modifying the "textbook" methods to suit the application. Still, in many cases, this effort pays off handsomely in producing a system whose performance far exceeds that of existing solutions. This paper describes a most fruitful embodiment of this principle, which our group developed as part of a large data mining challenge with the insurance industry.

The concrete problem we were tackling is that of automobile insurance premia estimation: how much should a driver pay for car insurance, given information about the driver's status (age, type of car, etc.), driving history (convictions, accidents, and so forth), and type of insurance coverage. As we shall see below, this problem is nothing

more than a classical case of conditional mean estimation, or in other words, a problem of regression.

Unfortunately, as we discovered, the usual algorithms for regression that we tried on this data (standard and generalized linear models (McCullagh and Nelder, 1989), ordinary feed-forward neural networks, CHAID decision trees (Kass, 1980), support vector machines (Vapnik, 1998)) did not perform very well. As it shall become clear in the subsequent sections, the reasons for this failure are threefold. Firstly, there is a statistical difficulty: most drivers will not file a claim with the insurance company in any given year, and the premia for all drivers must be inferred from those few drivers (whom might be called **outliers** but provide essential information) who do file for a claim. Note also that the distribution of claim amounts is asymmetric with with a heavy tail, which makes estimation difficult (and bars the use of classical robust regression methods based on downweighting or eliminating outliers). Secondly, there is another statistical difficulty, due to the large number of variables (mostly discrete) and the fact that many interactions exist between them. Thus methods based on tabulating average claim amounts for each combination of values are quickly hurt by the curse of dimensionality, unless they make hurtful independence assumptions (Bailey and Simon, 1960). Finally, there is a computational difficulty: we had access to a large database of $\approx 8 \times 10^6$ examples, and the training effort and numerical stability of some algorithms can be burdensome.

This paper is organized as follows: we start by describing the mathematical criteria underlying insurance premia estimation (section 2), followed by a brief review of the learning algorithms that we consider in this study, including our best-performing mixture of positive-output neural networks (section 3). We then highlight our most important experimental results (section 4), and in view of them conclude with an examination of the prospects for applying statistical learning algorithms to insurance modeling (section 5).

2 Mathematical Objectives

The goal of insurance premia modeling is to estimate the *expected claim amount* for a given insurance contract for a future one-year period (here we consider that the amount is 0 when no claim is filed). Let $X \in \mathbf{R}^m$ denote the customer and contract *input profile*, a vector representing all the information known about the customer and the proposed insurance policy before the beginning of the contract. Let $A \in \mathbf{R}_+$ denote the amount that the customer claims during the contract period; we shall assume that A is non-negative. Our objective is to estimate this claim amount, which is the *pure premium* p_{pure} of a given contract x:[1]

$$p_{\text{pure}}(x) = E[A|X = x]. \tag{1}$$

The Precision Criterion. In practice, of course, we have no direct access to the quantity (1), which we must estimate. One possible criterion is to seek the *most*

[1]The pure premium is distinguished from the premium actually charged to the customer, which must account for the risk remaining with the insurer, the administrative overhead, desired profit, and other business costs.

precise estimator, which minimizes the mean-squared error (MSE) over a data set $D = \{\langle x_\ell, a_\ell \rangle\}_{\ell=1}^{L}$. Let $\mathcal{P} = \{p(\cdot;\theta)\}$ be a function class parametrized by the parameter vector θ. The MSE criterion produces the most precise function (on average) within the class, as measured with respect to D:

$$\theta^* = \arg\min_\theta \frac{1}{L} \sum_{i=1}^{L} (p(x_i;\theta) - a_i)^2. \tag{2}$$

The Fairness Criterion. However, in insurance policy pricing, the precision criterion is not the sole part of the picture; just as important is that the estimated premia do not systematically discriminate against specific segments of the population. We call this objective the *fairness criterion*. We define the *bias of the premia* $b(P)$ to be the difference between the average premium and the average incurred amount, in a given population P:

$$b(P) = \frac{1}{|P|} \sum_{\langle x_i, a_i \rangle \in P} p(x_i) - a_i, \tag{3}$$

where $|P|$ denotes the cardinality of the set P, and $p(\cdot)$ is some premia estimation function. A possible fairness criterion would be based on minimizing the norm of the bias over every subpopulation Q of P. From a practical standpoint, such a minimization would be extremely difficult to carry out. Furthermore, the bias over small subpopulations is not statistically significant. We settle instead for an approximation that gives good empirical results. After training a model to minimize the MSE criterion (2), we define a finite number of disjoint subsets (subpopulations) of the test set P, $P_k \subset P, P_k \cap P_{j \neq k} = \emptyset$, and *verify* that the absolute bias is not significantly different from zero. The subsets P_k can be chosen at convenience; in our experiments, we considered 10 subsets of equal size delimited by the deciles of the test set premium distribution. In this way, we verify that, for example, for the group of contracts with a premium between the 5th and the 6th decile, the average premium matches the average claim amount.

3 Models Evaluated

An important requirement for any model of insurance premia is that it should produce *positive* premia: the company does not want to charge negative money to its customers! To obtain **positive outputs neural networks** we have considered using an exponential activation function at the output layer but this created numerical difficulties (when the argument of the exponential is large, the gradient is huge). Instead, we have successfully used the "softplus" activation function (Dugas et al., 2001):

$$\mathrm{softplus}(s) = \log(1 + e^s)$$

where s is the weighted sum of an output neuron, and $\mathrm{softplus}(s)$ is the corresponding predicted premium. Note that this function is convex, monotone increasing, and can be considered as a smooth version of the "positive part" function $\max(0, x)$.

The best model that we obtained is a **mixture of positive outputs neural networks**. The idea is to reduce the undesirable interactions between the estimation of

premia associated to small claims and large claims (the latter having much more variability). We consider a categorical variable C that corresponds to three components of the mixture: $C = 0$ for zero claims, $C = 1$ for non-zero claims $< \$10000$, $C = 2$ for claims $> \$10000$. The overall expectation is decomposed accordingly:

$$E[A|X] = \sum_i P(C = i|X)E[A|X, C = i]$$

Since the categorical variable C is observed (it is a deterministic function of the claim amount A), each component can be trained separately. A *gater network* (Jacobs et al., 1991) with softmax outputs is trained by maximum likelihood to estimate $P(C = i|X)$. The case of $C = 0$ is handled trivially since $E[A|X, C = 0] = 0$. Two regressions are separately estimated. First, using only the examples corresponding to small claims ($C = 1$), we estimate $E[A|X, C = 1]$. Second, using only the examples corresponding to large claims ($C = 2$), we estimate $E[A|X, C = 2]$. Note that two out of these three models are estimated on data that excludes the "outlier" large values ($P(C|X)$ and $E[A|X, C = 1]$), thereby "protecting" the estimation of their parameters from the variability introduced by the large claims. It should be noted that this setup also a significantly reduces the computational cost for the regressors, since the vast majority of the contracts (about 90%) have a claim amount $a = 0$; thus the two regression models are each trained on a small fraction of the data. As for the gater, it is trained on all the data but we found that it does not need to be trained for as many iterations as a single NN regression model in order to obtain good results. Using the validation set, we have made a choice of architecture for all three models, among the following: linear (or multivariate logistic for the gater), neural network with one hidden layer, with linear, exponential, or softplus output activation. The validation set was also used to choose a weight decay penalty and the number of hidden units.

The mixture model was compared to other models. The **constant model** only has intercepts as free parameters. The **linear model** corresponds to a ridge linear regression (with weight decay chosen with the validation set). **Generalized linear models** (GLM) estimate the conditional expectation from $f(x) = e^{b+w.x}$ with parameters b and w. Again weight decay is used and tuned on the validation set. There are many variants of GLMs and they are popular for building insurance models, since they provide positive outputs, interpretable parameters, and can be associated to parametric models of the noise.

Decision trees are also used by practitioners in the insurance industry, in particular the **CHAID**-type models (Kass, 1980; Biggs, Ville and Suen, 1991), which use statistical criteria for deciding how to split nodes and when to stop growing the tree. We have compared our models with a CHAID implementation based on (Biggs, Ville and Suen, 1991), adapted for regression purposes using a MANOVA analysis. The threshold parameters were selected based on validation set MSE.

Regression **Support Vector Machines** (SVM) (Vapnik, 1998) were also evaluated but yielded disastrous results for two reasons: (1) SVM regression optimizes an L1 criterion that finds a solution close to the conditional median, whereas the MSE criterion is minimized for the conditional mean, and because the distribution is highly asymmetric the conditional median is far from the conditional mean; (2) because the

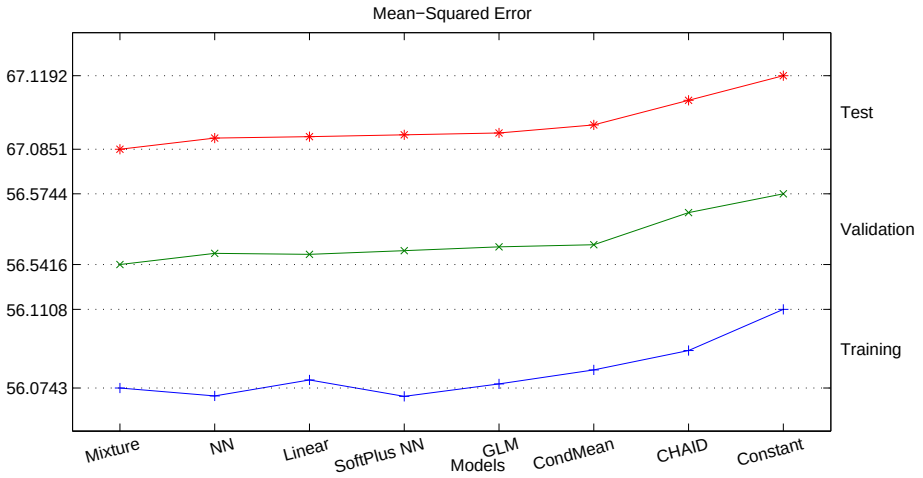

Figure 1: *MSE results for eight models. Models have been sorted in ascending order of test results. The training, validation and test curves have been shifted closer together for visualization purposes (the significant differences in MSE between the 3 sets are due to "outliers"). The out-of-sample test performance of the Mixture model is significantly better than any of the other. Validation based model selection is confirmed on test results. CondMean is a constructive greedy version of GLM.*

output variable is difficult to predict, the required number of support vectors is huge, also yielding poor generalization.

Finally, we compared the best statistical model with a proprietary table-based and rule-based premium estimation method that was provided to us as the **benchmark** against which to judge improvements.

4 Experimental Results

Data from five types of losses were included in the study (i.e. a sub-premium was estimated for each type of loss), but we report mostly aggregated results showing the error on the total estimated premium. The input variables contain information about the policy (e.g., the date to deal with inflation, deductibles and options), the car, and the driver (e.g., about past claims, past infractions, etc...). Most variables are subject to discretization and binning. Whenever possible, the bins are chosen such that they contain approximately the same number of observations. For most models except CHAID, the discrete variables are one-hot encoded. The number of input random variables is 39, all discrete except one, but using one-hot encoding this results in an input vector x of length $m = 266$. An overall data set containing about 8 million examples is randomly permuted and split into a training set, validation set and test set, respectively of size 50%, 25% and 25% of the total. The validation set is used to select among models (including the choice of capacity), and the test set is used for final statistical comparisons. Sample-wise paired statistical tests are used to reduce the effect of huge per-sample variability.

Table 1: *Statistical comparison of the prediction accuracy difference between several* individual *learning models and the best Mixture model. The p-value is given under the null hypothesis of no difference between Model #1 and the best Mixture model. Note that* ***all differences are statistically significant****.*

Model #1	Model #2	Mean MSE Diff.	Std. Error	Z	p-value
Constant	Mixture	3.40709e-02	3.32724e-03	10.2400	**0**
CHAID	Mixture	2.35891e-02	2.57762e-03	9.1515	**0**
GLM	Mixture	7.54013e-03	1.15020e-03	6.5555	**2.77e-11**
Softplus NN	Mixture	6.71066e-03	1.09351e-03	6.1368	**4.21e-10**
Linear	Mixture	5.82350e-03	1.32211e-03	4.4047	**5.30e-06**
NN	Mixture	5.23885e-03	1.41112e-03	3.7125	**1.02e-04**

Table 2: *MSE difference between benchmark and Mixture models across the 5 claim categories and the total claim amount. In all cases except category 1, the Mixture model is* ***statistically significantly*** *($p < 0.05$) more precise than the benchmark model.*

Claim Category (Kind of Loss)	MSE Difference Benchmark minus Mixture	95% Confidence Interval Lower	Higher
Category 1	20669.53	(−4682.83	– 46021.89)
Category 2	1305.57	(1032.76	– 1578.37)
Category 3	244.34	(6.12	– 482.55)
Category 4	1057.51	(623.42	– 1491.60)
Category 5	1324.31	(1077.95	– 1570.67)
Total claim amount	60187.60	(7743.96	– 112631.24)

Table 4 and Figure 3 summarize the comparison between the test MSE of the different tested models. *NN* is a neural network with linear output activation whereas *Softplus NN* has the *softplus* output activations. The *Mixture* is the mixture of softplus neural networks. This result identifies the mixture model with softplus neural networks as the best-performing of the tested statistical models. Our conjecture is that the mixture model works better because it is more robust to the effect of "outliers" (large claims). Classical robust regression methods (Rousseeuw and Leroy, 1987) work by discarding or downweighting outliers: they cannot be applied here because the claims distribution is highly asymmetric (the extreme values are always large ones, the claims being all non-negative). Note that the capacity of each model has been tuned on the validation set. Hence, e.g. CHAID could have easily yielded lower training error, but at the price of worse generalization.

Table 4 shows a comparison of this model against the rule-based benchmark. The improvements are shown across the five types of losses. In all cases the mixture improves, and the improvement is significant in four out of the five as well as across the sum of the five.

A qualitative analysis of the resulting predicted premia shows that the mixture model has smoother and more spread-out premia than the benchmark. The analysis

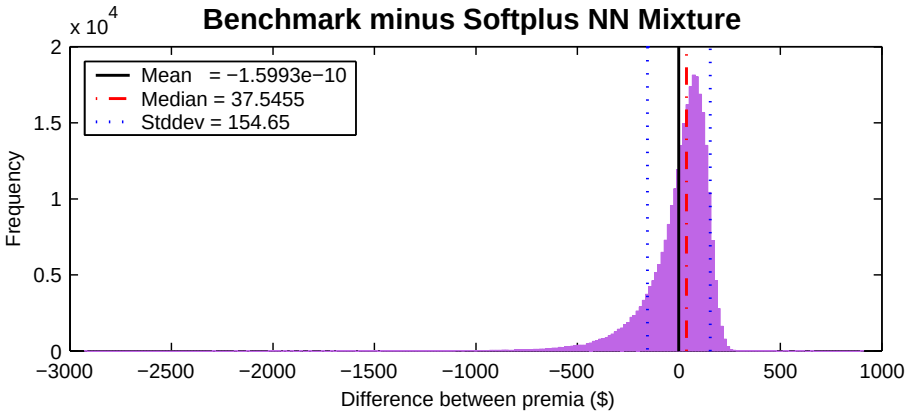

Figure 2: *The premia difference distribution is **negatively skewed**, but has a **positive median** for a mean of zero. This implies that the benchmark Model undercharges risky customers, while overcharging typical customers.*

(figure 4) also reveals that the difference between the mixture premia and the benchmark premia is negatively skewed, with a positive median, i.e., the typical customer will pay less under the new mixture model, but the "bad" (risky) customers will pay much more.

To evaluate **fairness**, as discussed in the previous section, the distribution of premia computed by the best model is analyzed, splitting the contracts in 10 groups according to their premium level. Figure 4 shows that the premia charged are fair for each subpopulation.

5 Conclusion

This paper illustrates a successful data-mining application in the insurance industry. It shows that a specialized model (the mixture model), that was designed taking into consideration the specific problem posed by the data (outliers, asymmetric distribution), performs significantly better than existing and popular learning algorithms. It also shows that such models can significantly improve over the current practice, allowing to compute premia that are lower for less risky contracts and higher for more risky contracts, thus reducing the cost of the median contract.

Future work should investigate in more detail the role of temporal non-stationarity, how to optimize fairness (rather than just test for it afterwards), and how to increase further the robustness of the model with respect to large claim amounts.

References

Bailey, R. A. and Simon, L. (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin*, 1(4):192–217.

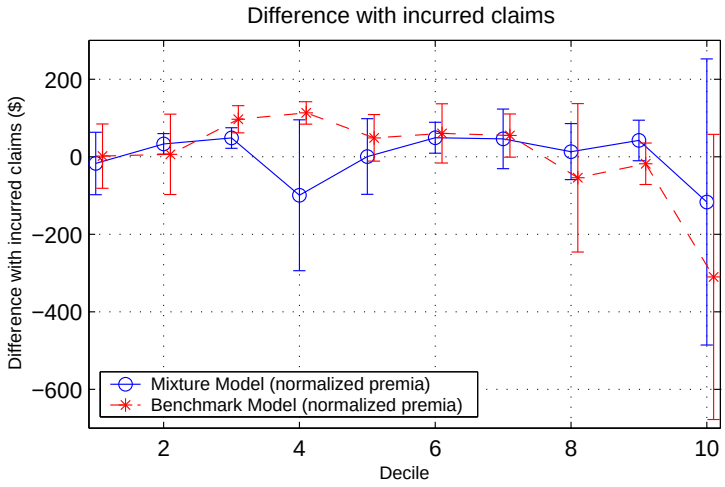

Figure 3: *We ensure fairness by comparing the average incurred amount and premia within each decile of the premia distribution; both models are* ***generally fair*** *to subpopulations. The error bars denote 95% confidence intervals.*

Biggs, D., Ville, B., and Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18(1):49–62.

Dugas, C., Bengio, Y., Bélisle, F., and Nadeau, C. (2001). Incorporating second order functional knowledge into learning algorithms. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 472–478. The MIT Press.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixture of local experts. *Neural Computation*, 3:79–87.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons Inc.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, Lecture Notes in Economics and Mathematical Systems, volume 454.