Scoring Models for Insurance Risk Sharing Pool Optimization

Nicolas Chapados
Dép. d'informatique et recherche opérationnelle
Université de Montréal, Canada
and ApSTAT Technologies
chapados@iro.umontreal.ca

Charles Dugas
Dép. de mathématiques et de statistique
Université de Montréal, Canada
and ApSTAT Technologies
dugas@dms.umontreal.ca

Pascal Vincent
Dép. d'informatique et recherche opérationnelle
Université de Montréal, Canada
and ApSTAT Technologies
vincentp@iro.umontreal.ca

Réjean Ducharme
ApSTAT Technologies Inc.
Montréal, Canada
ducharme@apstat.com

Abstract

We introduce a flexible scoring model that can be used by Property and Casualty insurers that have access to a risk-sharing pool to better select the insureds to transfer to the pool. The model discriminates between insureds whose transfer is likely to be profitable under the pool regulations against those paying a fair premium. This model makes use of feature selection methods to automatically discover the most relevant model inputs, yet is robust to overfitting due to the use of a rank averaging technique. By analogy to the knapsack problem, we show what should be the most suitable sorting criterion depending on the pool regulations. We illustrate the performance of the approach by testing against the historical data of a mid-sized Canadian insurer.

1. Introduction

Most automobile insurers use underwriting criteria to decide who can obtain coverage with their company. For people with the poorest driving records, simply finding an insurer that will accept to offer a quote can be very hard. Since automobile insurance is mandatory in many jurisdictions, some legislators have devised *facilities* or *pools* which act as insurers (or reinsurers) for these poor risks. Losses incurred by the pools are then shared among insurers that operate in the jurisdiction.

Rules vary widely from one pool to another. For some, insurers are allowed to simply *choose*, within their book of business, a certain portion of their risks and cede these risks to the pool. From a data-mining perspective, this possibility of choice that is sometimes granted is paramount as this creates the need, sometimes overlooked by the insurers themselves, to maximize the profitability of the transactions with the pool, given the rules set forth by the legislator.

In this application paper, we consider the pools of the two largest Canadian provinces, Québec and Ontario, with volumes of 2.8 and 9.6 billion Canadian dollars[1], respectively. We first describe the characteristics of Québec's pool, then consider those that prevail in Ontario and show that they lead to a different ranking criterion of the policies to be ceded.

Québec's automobile insurers are allowed to cede up to 10% of the volume of their book of business. Each insurer chooses the risks to cede to the pool. For those risks, the pool reimburses all claims. In return, the insurer must pay 75% of the premiums that were charged to the ceded risks. The remaining 25% are kept by the insurer to cover for administrative expenses and commissions related to the policy. Thus, insureds for which the mathematical expectation of the total claims is above 75% of the premium charged should be ceded by the insurer since we expect that claims will exceed what the pool charges to cover them.

An important figure that is often used to evaluate the profitability of an insurer's operations is the *loss ratio*, defined as the ratio of claims to written premiums (during a given period). This ratio is available from insurers' financial statements. If an insurer has a high loss ratio, then other expenses must remain low for the insurer to remain profitable. Conversely, insurer often measure the *permissible loss ra-*

[1] 2006 direct written premiums. Source: Canadian Insurance, 2007 Annual Statistical Issue. As of July 2, 2008, C$1 = €0.6213.

tio, i.e. the maximum loss ratio that, given fixed values for other expenses, would lead to a zero profit. The loss ratio is directly tied to the pool optimization problem described above: in the context of the Québec pool for instance, risks should be ceded if their expected loss ratio is above 75%. Thus, looking at a series of insurers financial results, one can readily see whether identifying risks to cede to the pool should be an easy task or not: insurers with global loss ratios close to or above 75% very likely have numerous risks that can be pinpointed, using state-of-art data-mining techniques, as having expected loss ratios above the 75% target. The goal of this paper is to introduce an effective scoring model to perform such a task.

The use of data mining techniques, including scoring models, for insurance risk estimation has been explored in the past, particularly in the areas of pure premium estimation (ratemaking) [6], fraud detection [2, 17] and underwriting [1]. However, specific applications to risk-sharing pool optimization has not, to the best of our knowledge, been previously investigated.

Insurance databases typically include a large number of variables, with important redundancies between them. Many of these variables were included for purposes other than ratemaking and are irrelevant from the risk-assessment perspective that is of interest here. For this reason, an initial feature selection stage is warranted in order to retain those features most likely to provide improvements in predictive accuracy. Limiting the number of features helps alleviate the curse of dimensionality and improve generalization and may also provide a better understanding of the underlying data-generating process.

Many feature selection algorithms have been proposed in order to build linear and/or nonlinear models. According to the taxonomy set forth by Guyon and Elisseeff [10], these fall in one of three categories: wrappers, embedded methods or filters. Wrappers [13] compare different sets of features on the basis of their predictive performance. One needs to specify how the space of all possible feature subsets should be searched. Greedy search strategies are often used and can be of two types: according to forward selection strategies, features are progressively selected and added to an initial empty set. Pursuit-type algorithms [3, 8, 14, 18] and the more recent LARS [7] algorithm fall in this category. Backward elimination strategies proceed the other way around: from an initial set of features including all variables, the least promising are progressively removed.

Embedded methods can either estimate or directly evaluate changes in the objective function in order to select features. Such methods include regularization or shrinkage approaches for which the objective function includes a goodness-of-fit term to be maximized as well as a regularization term that penalizes complexity and is to be minimized. The regularization term is often expressed as the $\mathcal{L}_p$-norm (usually, $p \in \{0, 1, 2\}$) of the model parameters. The Lasso [16] is a popular algorithm that falls in that category. Finally, filters correspond to a feature selection preprocessing that is conducted independently of the learning machine to be used.

Our approach falls within the category of wrappers with greedy forward selection search. A notable aspect of the approach is the fact that the dictionary of candidate variables is dynamically increased, as features are selected. Here, we consider linear models but the feature selection algorithm can equally be applied to nonlinear models. Another innovation, from a practical viewpoint, is the use of the loss ratio, rather than the claims level, as the target.

This paper is organized as follows: in section 2 we detail the scoring model used, the feature selection methodology and the scoring criteria that are applicable depending on the pool regulations. In section 3 we explain the performance evaluation methodology, and give experimental results in section 4. Finally section 5 concludes.

2. Models

The basic scoring model consists of a linear combination of *automatically-discovered features* to predict the loss ratio of each insured given its profile. Given the profile of an insured known at the start of a policy, $\mathbf{x} = (x_1, \ldots, x_K)^T$, consisting of K raw variables, we model the loss ratio for the duration of the policy as

$$\mathrm{LR}(\mathbf{x}; \boldsymbol{\beta}) = \sum_{i=1}^{N} \beta_i \phi_i(\mathbf{x}) + \epsilon, \qquad (1)$$

where $\boldsymbol{\beta}$ is a vector of model parameters, $\phi_i(\cdot)$ are features extracted from the raw input profile, and ϵ is a noise term.

Given a fixed set of features, the training criterion tries to fit the empirical *weighted loss ratio* of each policy in the training set, where the weight is given by the duration of the policy. Let $\mathbf{x}_j$ be the insured profile of training policy j, as observed when the *policy is written* (i.e. when a ratemaking or pool transfer decision can be made from the insurer standpoint), d_j be the duration the policy j (fraction of year), p_j the premium amount and c_j the observed claim amount. Thus, for profile $\mathbf{x}_j$, the observed loss ratio is c_j/p_j and this value is used as the target, within the context of supervised learning. The model parameters are obtained by the standard ridge estimator [12], which minimizes a regularised squared error,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{\sum_j d_j (\mathrm{LR}(\mathbf{x}_j; \boldsymbol{\beta}) - \frac{c_j}{p_j})^2}{\sum_j d_j} + \lambda \sum_i \beta_i^2 \right\}, \qquad (2)$$

where the hyperparameter λ can be determined empirically.

2.1. Stepwise Feature Selection for Scoring Model

The features $\phi_i(\cdot)$ are selected using a well-known stepwise forward selection procedure, from a dictionary of possible transformations depending on the variable type.

Algorithm 1 shows the algorithm in pseudocode. It starts with an initial set of features, Ψ, described below. Those features are *candidates* that can be retained in the final model, which has selected features Φ. The maximum number of features, N, is fixed *a priori*.[2] The algorithm iterates greedily to find, at each round, the best new feature to add to the selected set, in the sense of most reducing the mean-squared error on the training data, keeping fixed the features selected in previous iterations. After a feature ϕ_{best} has been found at a given iteration, it is removed from the set of candidates, and interaction (product) terms between it and previously-selected features are added to the set of candidates.[3] To constrain model capacity, only interaction terms up to a maximum order M are allowed.[4]

For automobile insurance modeling, standard raw variables include the driver's age, sex, vehicle type and color, driving experience, accident history and geographical location. The initial set of features consists of an encoding of *individual raw variables*, determined by the variable type. In other words, each element of Ψ is a function of a single raw variable only; the stepwise algorithm synthesizes higher-order features as low-order ones are selected.

The encoding performed by the initial features depends on the variable type, as follows:

- **Continuous** variables (e.g. age) are standardized to have zero-mean and unit-standard deviation,

  $$\phi_k(\mathbf{x}) = \frac{x_k - \hat{\mu}_k}{\hat{\sigma}_k},$$

  where $\hat{\mu}_k$ and $\hat{\sigma}_k$ are the mean and standard deviation of raw variable k estimated on the training data.

- **Discrete unordered** variables (e.g. vehicle color) are encoded as one-hot (dummy variables), with one less variable than the number of levels. Suppose that variable k has ℓ levels, denoted $1, 2, \ldots, \ell$ for simplicity; we introduce $\ell - 1$ features defined as

  $$\phi_{k,m}(\mathbf{x}) = I[x_k = m], \qquad m = 1, \ldots, \ell - 1,$$

  where $I[\cdot]$ is the indicator function.

Algorithm 1. Stepwise feature selection with automatic higher-order feature synthesis

```
def StepwiseSelection(model-train, data):
    $\Phi \longleftarrow \emptyset$
    $\Psi \longleftarrow \{\phi_k\}$        # initial set of features

    while $|\Phi| < N$ and $|\Psi| > 0$:
        best $\longleftarrow \infty$
        $\phi_{best} \longleftarrow$ None

        # Find the best feature among remaining ones
        for $\phi$ in $\Psi$:
            $\hat{\Phi} \longleftarrow \Phi \cup \phi$
            train-error $\longleftarrow$ model-train(data, $\hat{\Phi}$)
            if train-error $<$ best:
                best $\longleftarrow$ train-error
                $\phi_{best} \longleftarrow \phi$

        # Add best feature to working set. Add candidate
        # interactions with previous features to dictionary
        $\Phi \longleftarrow \Phi \cup \phi_{best}$
        $\Psi \longleftarrow \Psi - \phi_{best}$
        for $\phi$ in $\Phi$:
            $\hat{\phi} = \phi \times \phi_{best}$
            if order($\hat{\phi}$) $\leq M$:
                $\Psi \longleftarrow \Psi \cup \hat{\phi}$

    return $\Phi$
```

- **Discrete ordered** variables (e.g. number of accidents) are encoded in "thermometer form". Suppose that variable k has ℓ ordered levels, denoted $1, 2, \ldots, \ell$. We introduce $\ell - 1$ features defined as

  $$\phi_{k,m}(\mathbf{x}) = I[x_k \geq m], \qquad m = 2, \ldots, \ell.$$

For discrete variables, each level in the one-hot or thermometer encoding is considered independent and separately added to the initial feature set Ψ.

It should be emphasized that although the current scoring model is linear, non-linear extensions (e.g. feed-forward neural networks) are easy to accommodate in the framework introduced above. An interesting middle ground is to include in the set Ψ features that are formed by the *kernel evaluation* between two raw variables, similarly to the Kernel Matching Pursuit algorithm [18].

[2] In our experiments below, we used $N = 25$.

[3] The notation $\phi_1 \times \phi_2$ denotes a feature consisting of the product between features ϕ_1 and ϕ_2.

[4] We restricted interactions up to $M = 2$, for otherwise computational complexity quickly explodes and severe overfitting problems are likely to be encountered.

2.2. Sorting criteria

Assuming that, according to our models, more than the allowed 10% of the volume of business should be ceded, then we must rank and sort the policies in order to choose which will actually be ceded. How this should be done bears close ties with the well-known knapsack problem of combinatorial optimization, e.g. [4]. The knapsack problem considers the situation where a certain number of essentials, up to a maximum weight, are to be put in a bag and taken on a trip. There are different versions of the problem, leading to a different algorithmic solutions. Since each automobile insurance policy is either ceded in full or not at all, this corresponds to the *0–1 version* which would require us to use a dynamic programming approach. But, since each individual policy is very small compared to the volume of business of an insurer, we revert to the *fractional knapsack* version allowing us to use a greedy algorithm according to which policies are ceded from the first in rank until we reach 10% of the insurer's volume. As an approximation, the last ceded policy can be assumed to only be partially ceded so that precisely 10% of the insurer's total volume ends up in the pool. This situation corresponds to the regulations of the Québec pool, which puts a constraint on the *total premium volume* that can be ceded.

According to the greedy solution to the fractional knapsack problem, essentials are sorted in decreasing order of their value-to-weight ratio so that the total value, for a given maximum weight, is maximized. In our case, the value is the profit associated to ceding a policy and that corresponds to the expected claims level less 75% of the premium charged. The weight associated to each policy is the proportion of the entire book volume it represents, i.e. its premium divided by total volume. Thus, we obtain

$$\begin{aligned}\frac{\text{value}}{\text{weight}} &= \frac{\text{expected claims level} - 75\%\,\text{premium}}{\text{premium / book volume}} \\ &= \text{book volume}(\text{expected loss ratio} - 75\%).\end{aligned}$$

Removing constants shared by all policies and do not affect ranking, we conclude that policies should be *sorted in decreasing order of their expected loss ratio*.

Let us now consider the case of the Ontario pool where the limit is set to 5% of the *number of policies* insured. For this reason, all policies can be considered as having equal weight. Here again, each policy's value is its expected cession profit, obtained through a slightly different formula. The end result is that in order to maximize total profit, policies must be *sorted in decreasing order of their expected cession profit*.

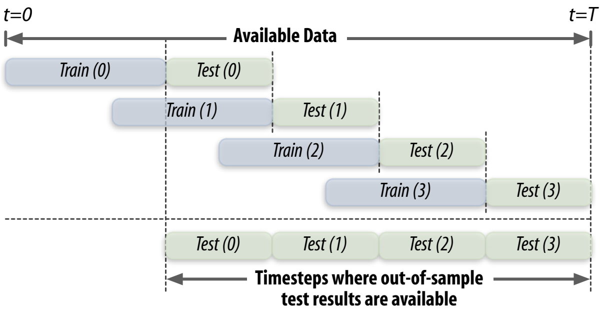

Figure 1. Illustration of the sequential validation procedure, where training and testing stages are interleaved through time.

3. Experimental Procedure

Insurance data is affected by a certain level of nonstationarities: the average premium level tends to vary across time according to market conditions. To better evaluate model performance in this context, we relied on the technique of *sequential validation* [9], which is an empirical testing procedure that aims to emulate the behavior of a rational decision maker updating a model as well as possible across time.

Sequential validation (also known as a rolling window or simulated out-of-sample testing) is inspired by the well-known technique of cross-validation [15, 11], and is appropriate when the elements of a dataset cannot be permuted freely, such as is the case for sequential learning tasks. One can intuitively understand the procedure from the illustration in figure 3. One trains an initial model from a starting subset $Train(0)$ of the data,[5] which is tested out-of-sample on a data subset $Test(0)$ immediately following the end of the training set. This test set is then added to the training set for the next iteration, a new model is trained and tested on a subsequent test set, and so forth. As the figure shows, at the end of the procedure, one obtains out-of-sample test results for a large fraction of the original data set, all the while always testing with a model trained on "relatively recent data".

The raw variables include all standard ratemaking variables typically used for automobile insurance, including driving experience, sex of driver, vehicle type and color, accident history, and so forth. We performed sequential validation by using the *most recent twelve months* of policy data to use as the training set, testing on the following month, and rolling forward by one month (i.e. adding the test month to the new training set, and deleting the oldest

[5] "Starting" is meant in the temporal sense.

one). In our dataset, a total of 117 raw variables were usable as inputs; after encoding of discrete variables (as outlined in section 2.1), 1726 variables were included in the initial set (Ψ in algorithm 1) for stepwise variable selection.

3.1. Controlling Overfitting

Given such a large set of variables and the repeated selection that occurs within the sequential validation procedure, caution is advisable lest overfitting proves seriously detrimental.[6] To this end, we do not perform a complete selection of features at each training iteration, but emphasize features that have worked well in the past. As before, let N be the total number of features that should be kept in the final model. We proceed as follows:

- The first three years of data are used as a **warmup phase**. Sequential validation is performed as usual, but the test performance results are discarded. In this phase, a total of 25 models are trained, and the identity and rank (from 0 to $N-1$) of selected features, for each of the 25 models, are recorded. The set of N features arising out of warmup stage w is denoted by Φ_w.

- Next comes the **validation phase**, where model performance is evaluated as in figure 3. Consider step s of sequential validation. We decompose this phase into three steps:

  i. *(Feature generation for current stage).* We apply algorithm 1 to find the top N features given the training set $\mathit{Train}(s)$. Denote by Φ_s the set of features selected at this step.

  ii. *(Rank averaging).* We compute the average rank of all features that have been generated in both the warmup phase and all previous validation phases. More specifically, for all features ϕ, we compute

  $$\overline{\mathrm{Rnk}}_\phi = \frac{1}{W+s}\Big(\sum_{w=1}^{W} \mathrm{Rnk}_\phi(\Phi_w) + \sum_{s'=1}^{s} \mathrm{Rnk}_\phi(\Phi_{s'})\Big),$$

  where $\mathrm{Rnk}_\phi(X)$ denotes the rank of feature ϕ in set X (which can range from 0 to $|X|-1$), or $|X|$ if ϕ is not part of X. In words, this step considers *all features* that have been generated so far, and establishes a "quality measure" based on the average rank that each feature gets across all prior and current training sets.

  iii. *(Final selection).* We form the subset of $\overline{N}$ features having the *highest average rank*, and apply again algorithm 1 to that subset only to select N features,[7] without allowing the creation of new higher-order features within the algorithm. This final set of features is used for sequential validation stage s.

The purpose of this procedure is to ensure a certain stability within selected features as we proceed with the sequential validation. We consistently observed that simply applying plain feature selection from anew at each sequential validation stage s produces quite unstable features that likely overfit the current training set.

4. Results

Our data was provided by a mid-sized Canadian automobile insurer and ranges from October 1999 until May 2008. The total data consists of nearly one million policy records and associated claim data. The out-of-sample results presented here were obtained on the last 34 distinct test months, and follow the experimental procedure set forth in section 3; for reference, over these 34 months, the total claims filed were C\$88M, the total premium volume was C\$147M, for an overall loss ratio of nearly 60%.

Table 1 shows the extracted features having the highest average rank across the complete validation period, where a total of $N = 25$ features were retained in the selection procedure. In addition to standard ratemaking variables commonly found in automobile risk estimation (such as age of driver, driving experience, accident and claim history), some interesting higher-order features emerge as significant, such as the squared annual distance driven with the vehicle, or an interaction between the driver's experience and the premium paid on the policy. It should be emphasised that these are high-ranking features *on average*; any specific iteration within sequential validation will use a slightly different set of features.

As emphasised in section 2.2, the appropriate sorting criterion (either decreasing order of predicted loss ratio or expected cession profit) depends on the regulatory context in which a specific pool operates. For the Québec risk-sharing pool, which is volume-constrained, the appropriate criterion is the loss ratio; this is the case that we are considering with the current data.

To illustrate the practical difference between the two criteria, table 2 shows the out-of-sample model performance (averaged over all 34 sequential validation test sets) at various cession percentage, for both sorting criteria. The performance measures are (i) the loss ratio of the *ceded insureds*, (ii) the total net profit earned by ceding over the test

[6] A delicate balance must be struck between the size of the sliding training set used in sequential validation, and the level of nonstationarities encountered in the data—a manifestation of the classical bias-variance dilemma.

[7] In our experiments, $\overline{N}$ is fixed at 100.

Table 1. Top 18 of the most significant features selected by Algorithm 1 for the dataset under study. Some internal ratemaking variables are omitted for confidentiality reasons. It is significant to note that a number of "meaningful" (to actuaries) second-order terms are selected.

Rank	Feature	Rank	Feature
1.	Number of years vehicle has been owned	10.	Age of vehicle
2.	Annual km	11.	Whether insured is in a large city
3.	Number of years without claims	12.	(Number of years driver has a license) × (Premium paid)
4.	Number of years driver has a license	13.	Amount of deductible
5.	Reason for last policy change	14.	Number of years without accidents
6.	(Annual km)×(Annual km)	15.	Death or mutilation endorsement
7.	Age of policy	16.	(Death or mutilation endorsement) × (Death or mutilation endorsement)
8.	Number of responsible claims in past ten years	17.	Age of main driver
9.	(Number of years driver has a license) × (Annual km to drive to work)	18.	Number of years of ownership of vehicle

period, and (iii) the net profit as a proportion of the total premium volume, expressed in basis points (one hundredth of one percent). Recall that, under the Québec pool rules, it is profitable to cede an insured whose loss ratio is greater than 75%.

At the outset, we note that the proposed scoring model, under both sorting criteria, is very profitable across a wide range of cession percentages. The loss ratio of ceded insureds—a direct measure of model performance—exceeds both the profitability threshold (75%) and the average book loss ratio over the period (60%). Moreover, with a net profit ranging from 18 to 44 basis points over the period, the model can prove of significant benefits to enhance operational profitability.

The table also lists 95% confidence intervals for the loss ratios, obtained by a bootstrap resampling (with 5000 iterations) of the claim and premium amounts using the percentile method [5].

A closer inspection reveals that the loss ratio criterion strictly dominates the predicted profit on this task, both in terms of loss ratio of ceded insureds and earned net profit. The confidence intervals on the loss ratio generally exclude the 75% threshold for the loss ratio sorting criterion, which allows rejecting the null hypothesis that the model cannot cede profitably at the $p = 0.05$ level. In contrast, the lower bound of the intervals obtained under the predicted profit criterion are systematically lower and generally include the 75% point (albeit slightly), another indication of its lower performance against the alternative sorting criterion.

These results are illustrated in graphical form in figures 2–4, all of which are expressed as a function of the percentage of ceded insureds extending up to 10% (the transfer limit in the Québec pool), and separately show the performance under the two sorting criteria: by loss ratio and predicted profit.

First, figure 2 plots the average loss ratio over the test period of insureds transferred to the pool. Comparing against loss ratio profitability threshold of 75% (dotted horizontal line), we see that both models profitably cede, on average, up to the applicable pool limit. The curve variability at low cession levels (below one percent) corresponds to a small number of very large claims in the data, which become progressively ceded as we go along; it is noteworthy that these risks are ceded very early on, which is an indication of the model's ability to efficiently pick out those risks. However, the converse also holds in that it must be emphasized that the considerable variability in this region indicates that profitability cannot be guaranteed at very low cession levels (e.g. smaller than 0.25%). We also note that the loss ratio of ceded risks is significantly higher than the average global book loss ratio, which is around 60% in this period (solid red horizontal line), further suggestive of the model's ability to discriminate likely candidates for pool transfer.

Figure 3 shows the average *excess loss ratio*, defined as the *ratio* of the loss ratio of ceded risks to the global book loss ratio. This differs from the pure loss ratio (shown in figure 2) in that the denominator (the global book loss ratio) is not taken to be a constant across the whole test set, but is separately computed for each test month in the sequential validation. Hence, this figure corrects for possible seasonal effects (e.g. there are more accidents during winter, which affect all insureds) that would be occluded by a global average. We note that the average excess loss ratio always lies above the unity line at all cession levels, which indicates

Table 2. Performance as a function of the percentage of ceded premium volume, for the loss ratio (left) and the predicted profit (right) sorting criteria. In each case are given the average loss ratio of the ceded insureds (95% bootstrap confidence intervals on the loss ratio), the net cession profit, and the proportional profit, which is expressed as a fraction (in basis points) of the total written premium volume, where 100 basis points = 1%.

	Sorted by Loss Ratio			Sorted by Predicted Profit		
Ceded %	Loss Ratio	Profit (k$)	Prop. Profit	Loss Ratio	Profit (k$)	Prop. Profit
1%	0.937 (0.7263–1.1679)	272.14	18.5	0.930 (0.7167–1.1669)	261.56	17.8
2%	0.923 (0.7770–1.0784)	509.52	34.6	0.914 (0.7754–1.0675)	482.12	32.8
3%	0.883 (0.7649–0.9938)	588.75	40.0	0.857 (0.7492–0.9723)	474.71	32.3
4%	0.859 (0.7725–0.9545)	647.41	44.0	0.827 (0.7370–0.9239)	454.98	30.9
5%	0.838 (0.7615–0.9160)	649.40	44.1	0.825 (0.7371–0.9150)	554.29	37.7

that the results of figure 2 are not due to a few "lucky" test months, but occur consistently and robustly across periods.

Finally, figure 4 shows the average net profit earned by the insurer as a result of ceding up to the indicated percentage of premium volume, expressed as a percentage of written premiums. This profit, strikingly, is positive across the whole cession range; it is also flatter over a larger range for the loss-ratio sorting criterion than the predicted-profit one. This behavior is very reassuring since it suggests that beyond a base cession level (of slightly over 3%) the bottom-line performance is fairly insensitive to the precise choice of threshold, quite a desirable quality given insurers' steep risk aversion to variance in pool performance.[8]

5. Conclusion

This paper demonstrated the effectiveness of a scoring model for insurance risk sharing pool optimization. The model is based on stepwise feature selection and can automatically synthesize higher-order features based on already-selected ones. We also showed that the regulatory constraints imposed by the jurisdiction in which the pool operates affect the optimal sorting criterion, and this has practical consequences both in terms of loss ratio of ceded insureds and overall net profitability. The approach is robust over a wide range of operation points, and can therefore easily integrates within larger "pool transfer rules" of an insurer (which can consume a fixed portion of the total allowed transfer budget).

A compelling prospect for future work in this area consists in embedding this model within larger operational constraints: in some jurisdictions, pool regulations strictly control the moments where insureds can be ceded, such as when a policy is renewed or when there is significant changes to the coverage (e.g. a new car). When operating close to the transfer limit allowed by the pool, it may be advantageous, in some cases, to avoid transferring up to the limit, in order to reserve space for better prospects that will be anticipated to come along in the future. Such an method could further enhance the profitability of the approach in strongly budged-constrained situations.

Acknowledgements

The authors wish to thank Xavier Saint-Mleux for assistance with the software implementation and analytical work.

References

[1] C. Apte, E. Grossman, E. P. D. Pednault, B. K. Rosen, F. A. Tipu, and B. White. Probabilistic estimation-based data mining for discovering insurance risks. *IEEE Intelligent Systems*, 14(6):49–58, November 1999.

[2] M. Artís, M. Ayuso, and G. Montserrat. Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk & Insurance*, 69(3):325–340, September 2002.

[3] S. Chen. *Basis Pursuit*. PhD thesis, 1995.

[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

[8] Since a ceded insured accounts for less revenue from an insurer's standpoint, for the insurer to accept taking on the risk of transferring the insured to the pool, there must be strong indications that this transfer would more than compensate through an expected reduction in expenses arising from an avoided claim.

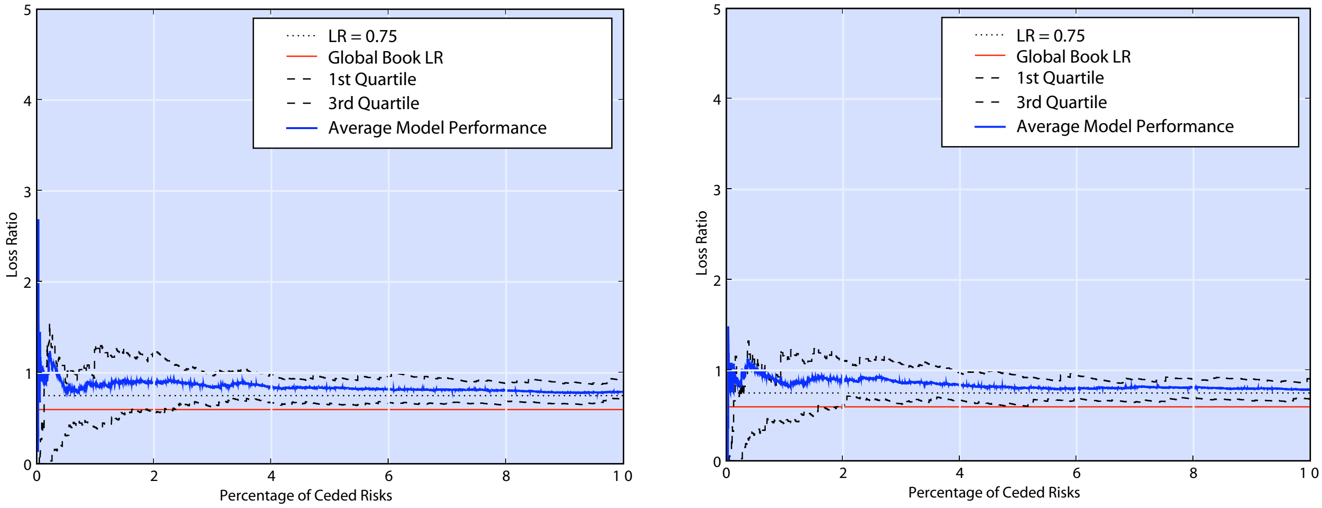

Figure 2. Loss ratio as a function of the percentage of premium volume ceded to the pool. Sorting criterion is the loss ratio (left) and the predicted profit (right). The dotted horizontal line is the 75% loss ratio threshold, above which ceding to the pool is profitable. The solid red line immediately below is the overall loss ratio for the entire book of business. We note that, regardless of the sorting criterion, the model can cede profitably up the the 10% maximum level allowed by the Québec pool regulations.

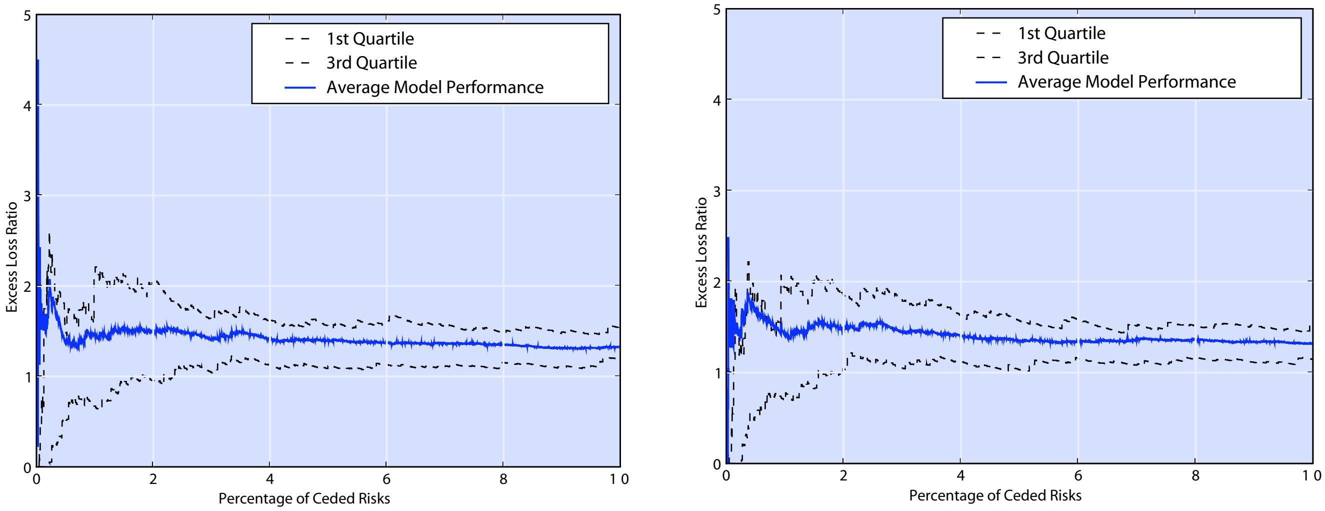

Figure 3. Excess loss ratio as a function of the percentage of premium volume ceded to the pool. Sorting criterion is the loss ratio (left) and the predicted profit (right).

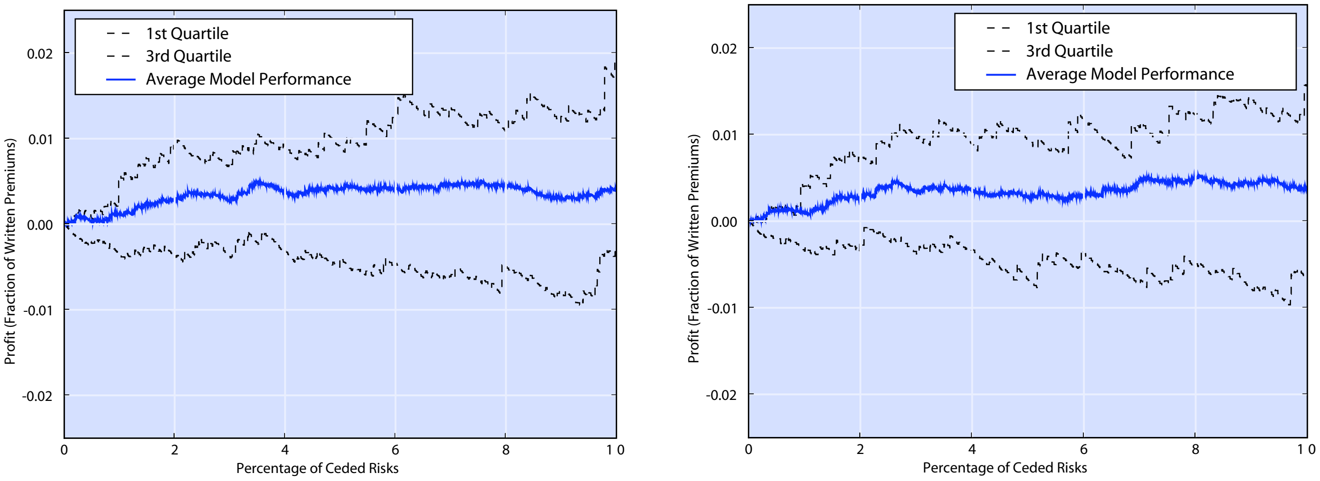

Figure 4. Proportional profit as a function of the percentage of premium volume ceded to the pool. Sorting criterion is the loss ratio (left) and the predicted profit (right).

[5] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK, 1997.

[6] C. Dugas, Y. Bengio, N. Chapados, P. Vincent, G. Denoncourt, and C. Fournier. Statistical learning algorithms applied to automobile insurance ratemaking. In A. F. Shapiro and L. C. Jain, editors, *Intelligent and Other Computational Techniques in Insurance: Theory and Applications*, Series on Innovative Intelligence, 6, pages 137–197. World Scientific Publishing Company, 2003.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[8] J. Friedman and J. Tuckey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23(9), 1974.

[9] F. Gingras, Y. Bengio, and C. Nadeau. On out-of-sample statistics for time-series. Technical Report 2002s-51, CIRANO, 2002.

[10] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Reasearch*, 3:1157–1182, 2003.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Berlin, New York, 2001.

[12] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[13] R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence*, 97(1–2):273–324, 1997.

[14] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[15] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B*, 36(1):111–147, 1974.

[16] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

[17] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk & Insurance*, 69(3):373–421, September 2002.

[18] P. Vincent and Y. Bengio. Kernel matching pursuit. *Machine Learning*, 48:169–191, 2002.