Higher-Order Feature Synthesis for
Insurance Scoring Models
Charles Dugas\textsuperscript{1,2}, Nicolas Chapados\textsuperscript{1,2}, Réjean Ducharme\textsuperscript{1},
Xavier Saint-Mleux\textsuperscript{1}, Pascal Vincent\textsuperscript{1,2}

\textsuperscript{1}ApSTAT Technologies, Montréal, Canada
\textsuperscript{2}Université de Montréal, Montréal, Canada

CRM Working Paper #3285
August 21, 2009
Abstract

In many jurisdictions, automobile insurers have access to risk-sharing pools to which they can transfer some of their worst risks. Better selection of these risks in order to maximize profitability is the application we consider in this paper. To that end, different feature selection and modeling approaches are tested and compared against the historical data of a mid-sized Canadian insurer. In particular, we introduce a flexible scoring model that estimates each risk’s loss ratio, a target that is better suited to the application considered than the usual claims level. We also devise a feature selection method that 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 pool regulations. Given the sequential structure of insurance data, model selection through cross validation with free permutation of instances must be discarded. Instead, we use a similar technique but that is coherent with the sequential structure. We explain how different software maturity levels lead to different levels of information being available at the time a decision must be made which in turn, impacts profitability.
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 (re)insurers for these poor risks. Insurers can then accept these otherwise unsuitable risks and simply cede (i.e. transfer) them to the pool. In such cases, policy administration usually remains a duty of the insurer for which it keeps part of the insurance premium but the risk is borne by the pool, i.e. insurance claims incurred on ceded risks are paid by the pool. Losses in the pool 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 paper, we consider the pools of the two largest Canadian provinces, Qu´ebec and Ontario, with volumes of 2.8 and 9.6 billion Canadian dollars, respectively. We first describe the characteristics of Qu´ebec’s pool, then consider those that prevail in Ontario and later (subsection 2.5) show that they lead to a different ranking criterion of the policies to be ceded.

Qu´ebec’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.

In Ontario [1], up to 5% of the insured units1 can be ceded. Risks are not ceded in full, rather the insurer retains a 15% share of each ceded risk. The amount an insurer can charge for expenses may vary from one risk to another, depending on the premium structure that has been filed with the legislator, but are limited to 30% of the premium. Thus, in most cases, an insurer can transfer 85% of the risk by paying 59.5% (85%×(1-30%)) of the premium. Here, risks with an expected claims level above 70% of their premium should be ceded.

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, insurers often measure the permissible loss ratio, i.e. the loss ratio that, given fixed values for other expenses, would lead to targeted values for profit and return on equity. The loss ratio is directly tied to the pool optimization problem described above: in the context of the Qu´ebec 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

---

1A unit is defined as one year of coverage for one vehicle and is also referred to as a car year
been explored in the past, particularly in the areas of pure premium estimation (ratemaking) [8], fraud detection [3, 23] and underwriting [2]. However, specific applications to risk-sharing pool optimization has not, to the best of our knowledge, been previously investigated.

1.1 Importance of Feature Selection

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, improve generalization performance, 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 [15], these fall in one of three categories: wrappers, embedded methods or filters. Wrappers [18] 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 [5, 12, 19, 24] and the more recent LARS [9] 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 [22] 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.

This paper proposes an approach, called Basis Selection Regression, that 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 regression target.

1.2 Overview

This paper is organized as follows: in section 2 we detail the scoring models compared in this paper, including 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. In section 5 we examine reliability considerations that must be faced by models deployed in the field, and in section 6 we quantify one dimension of the data mining process maturity on the achievable financial benefits. Finally section 7 examines avenues for future research.
2 Models

The profit resulting from the decision of ceding a particular insured to the pool is the amount paid in claims for the insured’s policy (and that the pool will pay instead of the insurer) minus what the insurer must pay to the pool, which corresponds to a premium \( p' \) that is a percentage of the annual premium rate \( p \) charged to the insured: \( p' = \alpha p \) (in Québec, \( \alpha = 0.75 \)). Thus, ceding a policy of duration \( d \) will yield a profit of \( c - dp' \) where \( c \) is the total claim amount incurred over the coverage period. Our goal is to identify insureds for which transfer to the pool would be, on average, the most profitable.

All models considered here can be interpreted as scoring models: given an input profile \( x \), the model outputs a real-valued score quantifying the desirability of ceding to the pool the insured having the given profile. The objective, obviously, is to assign a higher score to profiles that are more likely to be profitable. In a batch implementation (see section 6 for alternatives), the scoring process proceeds as follows:

1. All profiles that are considered in a given batch (e.g. one month worth of new or renewed policies) are scored separately;

2. Profiles in the batch are sorted in decreasing order, as a function of the score;

3. Higher-scoring profiles are ceded up to a target limit, which is a function of pool regulations and insurer business objectives.

The models that we compare in the experimental section are described next. We follow with a description of two possible sorting criteria (used in step 2, above), both of which are based on the model score.

2.1 Frequency/Severity Approach

Usually, insurers estimate the expected claims level, conditional on input profile \( x \), assumed to be known at the start of a policy, through a two-staged approach often referred to as frequency/severity modeling:

\[
E[c|x] = E[c|y = 0, x]P(y = 0|x) + E[c|y = 1, x]P(y = 1|x)
\]

where \( y \) is a binary variable identifying, for a given policy, the occurrence of an accident \( (y = 1) \) or not \( (y = 0) \). Since, by definition, \( E[c|y = 0, x] \equiv 0 \), the above equation simplifies to the following:

\[
E[c|x] = E[c|y = 1, x]P(y = 1|x).
\]

(1)

The first term, \( E[c|y = 1, x] \) (called severity factor), estimates the claims given that an accident occurred, while the second term, \( P(y = 1|x) \) (called frequency factor), gives the probability that an accident occurs.

The severity part can be addressed as a simple regression problem; in the experiments, we considered Weighted Partial Least Squares (WPLS) regression (see [11] and [20]). A WPLS model search for the multidimensional directions in the input space that explain the maximum multidimensional variance directions in the target space.
The frequency part corresponds to a two-class classification formulation. We make use of a generative model to estimate class-conditional probabilities $P(x|y=k)$, $k=0,1$. Then, given prior class probabilities $P(y=k)$, $k=0,1$, Bayes’ theorem is used to obtain posterior class probabilities:

$$P(y=k|x) = \frac{P(x|y=k)P(y=k)}{P(x)} \quad k=0,1$$

$$P(x) = \sum_i P(x|y=k)P(y=k).$$

Prior probabilities are estimated by maximum likelihood, as the training set proportions of each class (the large quantity of data for insurance problems—several hundreds of thousands or millions of training examples—makes regularization unnecessary). As a baseline, class-conditional densities can be modeled using multivariate Gaussian distributions. We first considered tied distributions whereby both class-conditional densities share the same covariance matrix. Within the context of classification, this approach leads to a linear decision boundary, and is therefore referred to as the Linear Discriminant Analysis (LDA). Relaxing the homoscedastic hypothesis, i.e., using different covariance matrices for each class, lead to a quadratic decision boundary, hence the term Quadratic Discriminant Analysis (QDA). Although we are not performing classification but conditional probability estimation, we shall retain both LDA and QDA labels for later references.

### 2.2 Basis Selection Regressors

As an alternative to the standard frequency/severity approach, we consider models that consist in a linear combination of automatically-discovered features to predict the loss ratio of each insured given its profile. Using the loss ratio as the target of our model rather than the claims level is motivated by the application we consider for which profitability depends on the loss ratio (see subsection 2.5).

Given the profile $x$, we model the loss ratio for the duration of the policy as

$$\text{BSR}(x; \beta) = \sum_{i=1}^{N} \beta_i \phi_i(x) + \epsilon,$$

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

Given a fixed set of features, the training criterion targets the empirical weighted loss ratio of each policy in the training set, where the weight is given by the duration of the policy. Let $x_j$ 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 $x_j$, 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 $[17]$, which minimizes a regularised squared error,

$$\hat{\beta} = \arg \min_{\beta} \left\{ \frac{\sum_j d_j (\text{BSR}(x_j; \beta) - \frac{c_j}{p_j})^2}{\sum_j d_j} + \lambda \sum_i \beta_i^2 \right\},$$

where the hyperparameter $\lambda$ can be determined empirically.

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, $\Psi$, described below. Those features are candidates that can be retained in the final model, which has selected features $\Phi$. The maximum number of features, $N$, is fixed a priori.\textsuperscript{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 $\phi_{\text{best}}$ 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.\textsuperscript{3} To constrain model capacity, only interaction terms up to a maximum order $M$ are allowed.\textsuperscript{4}

\begin{algorithm}
\caption{Stepwise feature selection with automatic higher-order feature synthesis}
\begin{algorithmic}
\Function{StepwiseSelection}{model-train, data}\
\State $\Phi \leftarrow \emptyset$
\State $\Psi \leftarrow \{\phi_1\}$ \Comment{initial set of features}
\While{$|\Phi| < N$ and $|\Psi| > 0$}\
\State $\text{best} \leftarrow \infty$
\State $\phi_{\text{best}} \leftarrow \text{None}$
\Comment{Find the best feature among remaining ones}
\For{$\phi$ in $\Psi$:}\
\State $\text{train-error} \leftarrow \text{model-train}(\text{data}, \hat{\Phi})$
\If{$\text{train-error} < \text{best}$}\
\State $\text{best} \leftarrow \text{train-error}$
\State $\phi_{\text{best}} \leftarrow \phi$
\Comment{Add best feature to working set. Add candidate}
\Comment{interactions with previous features to dictionary}
\State $\Phi \leftarrow \Phi \cup \phi_{\text{best}}$
\State $\Psi \leftarrow \Psi \setminus \phi_{\text{best}}$
\For{$\hat{\phi}$ in $\Phi$:}\
\State $\hat{\phi} = \phi \times \phi_{\text{best}}$
\If{$\text{order}(\hat{\phi}) \leq M$}\
\State $\Psi \leftarrow \Psi \cup \hat{\phi}$
\EndIf
\EndFor
\EndIf
\EndFor
\EndWhile
\State \Return $\Phi$
\EndFunction
\end{algorithmic}
\end{algorithm}

For automobile insurance modeling, standard raw variables include the driver’s age, sex, vehicle type, 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 $\Psi$ 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:

\textsuperscript{2}In our experiments below, we used $N = 75$.
\textsuperscript{3}The notation $\phi_1 \times \phi_2$ denotes a feature consisting of the product between features $\phi_1$ and $\phi_2$.
\textsuperscript{4}We restricted interactions up to $M = 2$, for otherwise computational complexity quickly explodes and severe overfitting problems are likely to be encountered.
• **Continuous** variables (e.g. age) are standardized to have zero-mean and unit-standard deviation,
\[
\phi_k(x) = \frac{x_k - \hat{\mu}_k}{\hat{\sigma}_k},
\]
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 type) are encoded as one-hot (dummy variables), with one less variable than the number of levels. Suppose that variable \(k\) has \(\ell\) levels, denoted \(1, 2, \ldots, \ell\) for simplicity; we introduce \(\ell - 1\) features defined as
\[
\phi_{k,m}(x) = I[x_k = m], \quad m = 1, \ldots, \ell - 1,
\]
where \(I[\cdot]\) is the indicator function.

• **Discrete ordered** variables (e.g. number of accidents) are encoded in “thermometer form”. Suppose that variable \(k\) has \(\ell\) ordered levels, denoted \(1, 2, \ldots, \ell\). We introduce \(\ell - 1\) features defined as
\[
\phi_{k,m}(x) = I[x_k \geq m], \quad m = 2, \ldots, \ell.
\]

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

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 \(\Psi\) features that are formed by the kernel evaluation between two raw variables, similarly to the Kernel Matching Pursuit algorithm [24].

For reference, the above model is referred to as the **Basis Selection Regressor**.

### 2.3 Linear Regressor

For comparison purposes, we experimented with a **Linear Regressor** model, but rather than simply use it on the raw input variables, we applied the model on all the initial features generated by the **Basis Selection Regressor** model of subsection 2.2, without performing any feature selection; this corresponds to the initial set \(\Psi\) in algorithm 1 and excludes higher-order features. As will be seen in section 3, with such a large number of features, it is crucial to adequately regularize the model, which was accomplished by penalizing with a squared \(L_2\) norm penalty.

As for the Basis Sélection Regressor, the linear regressor expresses a predicted loss ratio as a linear combination of features of the same form as eq. (2).

\[
\text{LR}(x; \beta) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + \epsilon,
\]
where \(\beta = (\beta_0, \ldots, \beta_K)'\) and \(x = (x_1, \ldots, x_K)'\), and \(\epsilon\) is a noise term.

The training procedure minimizes the following objective:

\[
\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{j=1}^{P} d_j p_j \left( \text{LR}(x_j; \beta) - \frac{c_j}{d_j p_j} \right)^2 + \lambda \sum_{i=0}^{K} \beta_i^2 \right\}
\]

where \(j\) iterates over all the \(P\) training profiles, \(\lambda\) is the regularization coefficient (weight decay), and \(c_j/(d_j p_j)\) is the regression target corresponding to the loss ratio for profile \(j\).
In addition to this linear regression using all features, we included in the model set to be compared an elementary pruning procedure aimed at eliminating coefficients with low statistical significance. For this purpose, the training phase is carried out in three steps:

1. Linear regression on all the features, as above;
2. Pruning the less informative features, according to the criterion given next;
3. Linear regression on the remaining features.

The pruning phase keeps a fixed number of the most significant features (the exact parameters are listed in section 4). We use the t-ratio of the regression coefficients as the discriminant factor. It is computed as [14]:

\[ t = \frac{\hat{\beta}_i}{\sqrt{S_{ii}}} \]  

where \( S_{ii} \) are the elements on the diagonal of the \( S \) matrix, the asymptotic sampling variance of the estimator \( \hat{\beta} \),

\[ S = \text{Var}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1} \]  

and \( \hat{\sigma}^2 \) is an unbiased estimator of the noise variance

\[ \hat{\sigma}^2 = \frac{1}{P-K} \sum_{j=1}^{P} \left( \text{LR}(x_j, \hat{\beta}) - \frac{c_j}{d_j p_j} \right)^2. \]  

### 2.4 Bagging Model

Finally, due to the high noise level in the data, we experimented with a Bagging model, which is known to often improve performance in similar situations [4]. The idea is quite simple: starting with a training set of \( P \) examples, we generate \( B \) new training sets (called bags) using bootstrap resampling [10]. Then, we train the \( B \) bags using the chosen sub-model, and we combine the solutions by averaging the outputs (for regression) or by voting (for classification).

In our experiments, we applied the Bagging meta-model on the Linear Regressor model (with and without basis pruning).

### 2.5 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. [6]. 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 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 Qu´ebec 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

\[
\frac{\text{value}}{\text{weight}} = \frac{\text{expected claims level} - 75\% \text{ premium}}{\text{premium} / \text{book volume}} = \text{book volume}(\text{expected loss ratio} - 75\%).
\]

Removing constants shared by all policies that do not affect ranking, we conclude that policies should be sorted in decreasing order of their expected loss ratio. Models of subsections 2.2, 2.3, and 2.4 provide expected loss ratios as outputs and these can be readily used for sorting. On the other hand, the frequency/severity approach of subsection 2.1 leads to models that output an expected claims level. Dividing the output by the premium level, we obtain the desired 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. In that case, 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.

The impact of choosing one sorting criterion over the other, for the Québec pool, is examined in subsection 4.1.

3 Experimental Procedure and Minimum-Rank Basis Selection

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 [13], 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 prequential analysis, rolling window or simulated out-of-sample testing) is inspired by the well-known technique of cross-validation [21, 16], 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 1. One trains an initial model from a starting subset \textit{Train}(0) of the data,\(^5\) which is tested out-of-sample on a data subset \textit{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 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, accident history, and so forth. We

\(^5\)“Starting” is meant in the temporal sense.
Figure 1: Illustration of the sequential validation procedure, where training and testing stages are interleaved through time.

performed sequential validation by using the most recent 12 to 48 months (depending on the model) 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 one). In our dataset, a total of 117 raw variables were usable as inputs; after encoding of discrete variables (as outlined in subsection 2.2), 1726 variables were included in the initial set (Ψ in algorithm 1) for stepwise variable selection.

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. To this end, we adapt the basis selection regression algorithm of subsection 2.2 as we do not perform a complete selection of features at each training iteration, but rather 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 \( \Phi_w \).

- Next comes the validation phase, where model performance is evaluated as in Figure 1. 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 \( \text{Train}(s) \). Denote by \( \Phi_s \) the set of features selected at this step.

---

\(^6\)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.
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 $\phi$, we compute

$$
\text{Rnk}_\phi = \frac{1}{W + s} \left( \sum_{w=1}^{W} \text{Rnk}_\phi(\Phi_w) + \sum_{s'=1}^{s} \text{Rnk}_\phi(\Phi_{s'}) \right),
$$

where $\text{Rnk}_\phi(X)$ denotes the rank of feature $\phi$ in set $X$ (which can range from 0 to $|X| - 1$), or $|X|$ if $\phi$ 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 $N$ features having the highest average rank, and apply again algorithm 1 to that subset only to select $N$ features, 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.

This selection procedure is called the Minimum-Rank Basis Selection Procedure (BSR-MinRank).

## 4 Results

Our data was provided by a mid-sized Canadian automobile insurer and ranges from October 1999 until December 2007. The total data consists of nearly one million policy records and associated claim data. The out-of-sample results presented here are obtained on the last 36 distinct test months, and follow the experimental procedure set forth in section 3; for reference, over these 36 months, the total claims filed were C$97M, the total premium volume was C$160M, for an overall loss ratio of nearly 60%.

Table 1 shows a summary of the model performances. For each model, we present the loss-ratio and the profit for 2% and 5% cession percentages, and also the cession percentage (between 1% and 5%) that gives the maximum profit.

Here is the brief description of each model. In what follows, all hyperparameters were selected by cross-validation.

- **FreqSev-LDA**: Frequency/Severity model with LDA model for the frequency part. Training set is 48 months long. Weight decay of $10^{-3}$ for the LDA.
- **FreqSev-QDA**: Frequency/Severity model with QDA model for the frequency part. Training set is 48 months long. Weight decay of $10^{-6}$ for the QDA.
- **FreqSev-WPLS**: Frequency/Severity model with WPLS model for the frequency part. Training set is 12 months long.
- **BSR**: Basis Selection Regressor model with $N = 75$. Training set is 12 months long.

In our experiments, $\overline{N}$ is fixed at 100.
Table 1: Summary of model performances. Both the loss-ratio and the proportional profit (measured in basis points) are provided, at three cession levels: 2%, 5%, and the best level for the model (indicated in the column %).

<table>
<thead>
<tr>
<th>Model</th>
<th>Ceding 2%</th>
<th>Ceding 5%</th>
<th>Best Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loss-Ratio</td>
<td>Profit</td>
<td>Loss-Ratio</td>
</tr>
<tr>
<td>FreqSev-LDA</td>
<td>0.833</td>
<td>16.5</td>
<td>0.846</td>
</tr>
<tr>
<td>FreqSev-QDA</td>
<td>0.747</td>
<td>-0.5</td>
<td>0.703</td>
</tr>
<tr>
<td>FreqSev-WPLS</td>
<td>0.801</td>
<td>10.1</td>
<td>0.815</td>
</tr>
<tr>
<td>BSR</td>
<td>0.930</td>
<td>36.0</td>
<td>0.932</td>
</tr>
<tr>
<td>BSR-MinRank</td>
<td>0.904</td>
<td>30.7</td>
<td>0.893</td>
</tr>
<tr>
<td>LR</td>
<td>0.947</td>
<td>39.4</td>
<td>0.843</td>
</tr>
<tr>
<td>Bagging-LR</td>
<td>0.911</td>
<td>32.1</td>
<td>0.852</td>
</tr>
<tr>
<td>LR-Pruning</td>
<td>0.963</td>
<td>42.5</td>
<td>0.840</td>
</tr>
<tr>
<td>Bagging-LR-Pruning</td>
<td>0.856</td>
<td>21.1</td>
<td>0.860</td>
</tr>
</tbody>
</table>

- BSR-MinRank: *Min-Rank Basis Selection* model with $N = 75$ and $\bar{N} = 100$. Training set is 12 months long.
- LR: *Linear Regressor* model with $\lambda = 2 \times 10^7$. Training set is 12 months long.
- Bagging-LR: *Bagging of $B = 5$ Linear Regressor* model with $\lambda = 2 \times 10^7$. Training set is 12 months long.
- LR-Pruning: *Linear Regressor with Pruning* model with $\lambda = 2 \times 10^7$. Keep best 500 features. Training set is 12 months long.
- Bagging-LR-Pruning: *Bagging of $B = 5$ Linear Regressor with Pruning* model with $\lambda = 2 \times 10^7$. Keep best 250 features. Training set is 12 months long.

For each of the models described above, we present two plots showing the model performance relative to the percentage of ceded insureds (between 0 and 10%). These plots appear in Figures 2 through 10. The left-hand plot shows curves of the model loss-ratio. In addition to the average (over the 36 test months) loss-ratio, we also show the average loss-ratio of the first and third quartile. Finally, for information purpose, we also added two constant “curves”, the global loss-ratio of the book and the 0.75 line. This last one is very important since it represents the limit between winning and losing money. The right-hand plot shows the first and third quartile, and the average model profit, expressed as a proportion of total written premiums, and measured in basis points (one hundredth of one percent).

Figures 2, 3, and 4 show results for the frequency/severity approach using WPLS, QDA, and LDA models for frequency, respectively. The WPLS and LDA frequency models yield very similar results while the QDA model has a clear detrimental impact on profitability. At low cession percentages, profit rises faster with the Basis Selection Regressor model (Figure 5) than with the frequency/severity models. Profit attains its peak value earlier but then slowly decreases back to a value close to zero at 10% cession. In Figure 6, we see that the Min-Rank procedure has a very
Figure 2: Frequence/Severity model with WPLS.  
Left: Loss ratio.  
Right: Proportional profit.

Figure 3: Frequence/Severity model with QDA.  
Left: Loss ratio.  
Right: Proportional profit.

Figure 4: Frequence/Severity model with LDA.  
Left: Loss ratio.  
Right: Proportional profit.
Figure 5: Basis Selection Regressor model.  
**Left:** Loss ratio.  
**Right:** Proportional profit.

Figure 6: Min-Rank Basis Selection model.  
**Left:** Loss ratio.  
**Right:** Proportional profit.

Figure 7: Linear Regressor model.  
**Left:** Loss ratio.  
**Right:** Proportional profit.
Figure 8: Linear Regressor with Pruning model.  
**Left:** Loss ratio.  
**Right:** Proportional profit.

Figure 9: Bagging on Linear Regressor model.  
**Left:** Loss ratio.  
**Right:** Proportional profit.

Figure 10: Bagging on Linear Regressor with Pruning model.  
**Left:** Loss ratio.  
**Right:** Proportional profit.
Table 2: 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.

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

positive effect on profit. For cession percentages above 5%, the BSR-MinRank outperforms all other candidate models. In Figures 7 through 10, we consider linear models with or without bagging and with or without pruning of the selected sets of bases. In the experiments shown here, the impact of bagging is to lower the profit at low cession percentages but to increase it at higher cession percentages. Pruning of the bases has very little impact on the results. Clearly, the Min-Rank Basis Selection model obtain the best results in our experiments.

Table 2 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.

4.1 Impact of Sorting Criterion

As emphasized in subsection 2.5, 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, we compared them under the Basis Selection Regression (BSR) model with $N = 75$ selected bases. Table 3 shows the out-of-sample model performance at various cession percentages, for both sorting criteria (averaged over 5000 bootstrap resamples of policies and claims obtained on 36 consecutive months [7]). The performance
Table 3: 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 (with, in parenthesis, the 95% bootstrap confidence intervals on the loss ratio), and the proportional profit, which is expressed as a fraction (in basis points) of the total written premium volume, where 100 basis points = 1% (and, in parenthesis, the 95% bootstrap confidence intervals). All results are obtained from 5000 bootstrap resamples.

<table>
<thead>
<tr>
<th>Ceded %</th>
<th>Sorted by Loss Ratio</th>
<th></th>
<th>Sorted by Predicted Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loss Ratio</td>
<td>Proportional Profit (bps)</td>
<td>Loss Ratio</td>
</tr>
<tr>
<td>1%</td>
<td>0.996 (0.836 – 1.172)</td>
<td>35.6 (12.5 – 60.9)</td>
<td>0.922 (0.798 – 1.060)</td>
</tr>
<tr>
<td>3%</td>
<td>0.889 (0.805 – 0.979)</td>
<td>58.5 (23.2 – 95.6)</td>
<td>0.890 (0.812 – 0.972)</td>
</tr>
<tr>
<td>5%</td>
<td>0.884 (0.815 – 0.955)</td>
<td>90.4 (43.6 – 138.9)</td>
<td>0.852 (0.792 – 0.914)</td>
</tr>
<tr>
<td>7%</td>
<td>0.858 (0.801 – 0.917)</td>
<td>99.4 (47.2 – 153.6)</td>
<td>0.840 (0.787 – 0.895)</td>
</tr>
<tr>
<td>9%</td>
<td>0.835 (0.787 – 0.885)</td>
<td>98.8 (43.1 – 157.6)</td>
<td>0.812 (0.766 – 0.860)</td>
</tr>
</tbody>
</table>

Measures are (i) the loss ratio of the ceded insureds, and (ii) the net profit as a proportion of the total premium volume, expressed in basis points. Recall that, under the Québec pool rules, it is profitable to cede an insured whose loss ratio is greater than 75%.

Consistent with previous results, we note that the model is very profitable across a wide range of cession percentages under both sorting criteria. 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%).

A closer inspection reveals that the loss ratio criterion performs slightly better than 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 always excludes the 75% threshold, which allows rejecting the null hypothesis that the model cannot cede profitably at the $p = 0.05$ level.

These results are illustrated in graphical form in Figures 11–13, 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 11 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 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 12 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 11) in that the
shows the average net profit earned by the insurer as a result of ceding up to
are not due
11
0 2 4 6 8 1 0
Percentage of Ceded Risks
0
1
2
3
4
5Loss Ratio
LR = 0.75
Global Book LR
1st Quartile
3rd Quartile
Average Model Performance
0
13
14
30x454
0 2 4 6 8 1 0
Percentage of Ceded Risks
0
1
2
3
4
5Loss Ratio
LR = 0.75
Global Book LR
1st Quartile
3rd Quartile
Average Model Performance
[54x66]compensate through an expected reduction in expenses arising from an avoided claim.
[54x127]give approximately the same performance. Beyond this level, an advantage emerges for the loss-
shows it according to the loss ratio measure. Up to about 5% of ceded risks, both sorting criteria
shows the performance difference according to the proportional profit measure, whereas Figure
bootstrap repetitions to obtain estimates of the distribution of performance differences. Figure
measures: proportional profit and loss ratio of ceded risks. We then tally these results over 5000
month test duration), we compute the loss-ratio/predicted-profit difference for the two performance
More specifically, for each bootstrap resample of the test-set profiles and claims (over the same 36-
and predicted profit sorting criteria, we compare them in a pairwise fashion in a bootstrap test.
fairly insensitive to the precise choice of threshold, quite a desirable quality given insurers’ steep risk
aversion to variance in pool performance.
bottom-line performance 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
above the unity line at all cession levels, which indicates that the results of Figure 11 are not due
to a few “lucky” test months, but occur consistently and robustly across periods.
Finally, Figure 13 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.\footnote{\textsuperscript{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.} This property is further examined in the next section.

To arrive at a more definitive assessment of the performance difference between the loss ratio
and predicted profit sorting criteria, we compare them in a pairwise fashion in a bootstrap test.
More specifically, for each bootstrap resample of the test-set profiles and claims (over the same 36-
month test duration), we compute the loss-ratio/predicted-profit difference for the two performance
measures: proportional profit and loss ratio of ceded risks. We then tally these results over 5000
bootstrap repetitions to obtain estimates of the distribution of performance differences. Figure 14
shows the performance difference according to the proportional profit measure, whereas Figure 15
shows it according to the loss ratio measure. Up to about 5\% of ceded risks, both sorting criteria
give approximately the same performance. Beyond this level, an advantage emerges for the loss-
ratio sorting criterion, with the difference statistically significant at the 90\% level (under both

\textsuperscript{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 12: 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 13: 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).
Figure 14: Proportional profit difference between sorting by the loss-ratio and the predicted profit criteria (in basis points), as a function of the percentage of ceded risks. From 4% and up, the loss-ratio criterion exhibits better performance, significantly so (at the 90% confidence level) above around 8.5%.

performance measures) beyond approximately 8.5% ceded risks.

5 Robust Choice of RSP Operation Points

As mentioned, model selection for RSP purposes is based on out-of-sample performance, using a strict sequential validation framework. In this section, we attempt to replicate an operational constraint of insurers by including a three-month gap between the end of the training set and the start of the test set. The reason for this gap is that it can take up to three months for automobile insurance claims to fully develop.\(^9\) Thus, models must be trained with data at least three months old to avoid systematic risk underestimation. Since virtually all accidents are declared within a few days, frequency estimation could easily be performed using one month old data, within the context of the frequency/severity approach. On the other hand, effect on severity would be material since, as mentioned, up to three months may be necessary before the exact claim amount is known.

Although careful attention is put to ensure this out-of-sample performance evaluation process

\(^9\)In other words, the time it takes, on average, for the damage to occur to the car, filing the claim and fully assessing the value of the damage.
Figure 15: Loss ratio difference between sorting by the loss-ratio and the predicted profit criteria, as a function of the percentage of ceded risks. The 90% confidence bands are outlined, and from above 5%, the loss ratio sorting criterion generally significantly dominates the predicted profit criterion.
Figure 16: Bootstrap resampling of test set. Ten bootstrap resamples of the test set were drawn. The profit curve for the test set (solid blue) is plotted along with first and third quartiles (dashed black) against the percentage of cession. On this particular test set, the profit as predicted by the model (solid green) underestimated the realized profit (solid blue).

is free of estimation bias, it remains very noisy. Noise in profit estimation is mainly caused by the presence of a few very large claims; a way to measure this noise and derive confidence intervals for the profit, at different cession percentages, is to draw bootstrap resamples of the test set. For each sample, a profit curve is obtained. Statistics for the profit, at each cession percentage of interest, are obtained from the set of all curves.

Test set bootstrap resampling allows to better assess the robustness of performance of models across data sets. Of course, other factors must be accounted for so that a complete view of the range of performances that can be expected to occur is displayed. In particular, confidence intervals obtained through this process of test set bootstrap resampling assume stationarity of the underlying data generating process which, as mentioned before, is not supported by empirical evidence. Confidence intervals must therefore be observed with caution. Figure 16 gives an example of an out-of-sample profit curve with confidence intervals obtained through the process of test set bootstrap resampling. Here, we used a model of the class defined in subsection 2.4, i.e. by bagging \((B = 5)\) linear regression models.
6 Robust Choice of Evaluation Framework

6.1 Pool Regulation and Fairness Considerations

Insurance contracts tend to get modified over time, often at moments when insurance companies
would not normally be allowed to make the decision of putting the contract in the risk-sharing pool:
while contracts can be taken out of the pool at any moment, ceding contracts is permitted in some
situations only, like the renewal or the initial enactment of a contract. The information that is
available to our models at the moment it needs to make a decision does not necessarily reflect the
reality at that moment and may even depend on the exact way that the models are deployed.

From the way the data is recorded and then supplied to a modeler, and considering the internal
rules of the risk-sharing pool, it is obvious that finding the right way of preparing this data to
be fed to machine-learning algorithms is a non-trivial task. Any change to a contract may affect
the final performance of a model, but the model has to learn as if it was making a decision prior
to knowing about such changes and then consider future modifications to calculate performance.
Moreover, our initial experimentations have shown that the exact way in which we view a contract
and its different profiles (wherein every change to a contract starts could start a new profile) may
affect learning algorithms by indirectly “leaking” information from the future in very subtle ways,
and therefore exhibit greater performance than what could actually be achieved.

In this section, we examine ways to create a data representation that would be as close as
possible to the reality while remaining relatively simple and avoiding any leakage of information
from the future. We also consider different scenarios to reflect both a batch implementation—where
data is received monthly and recommendations are made in batch—and an online implementation
where recommendations are made instantaneously as soon as new information is received from a
client. We also quantify, in financial terms, the differences that these two implementations entail.

6.2 Impact of the Data Mining Process Maturity Level on Profit

By process maturity level, we refer to the integration of a data mining model within the information
technology infrastructure of the insurance company. As the two opposite ends of the maturity scale
lie the batch and online implementations mentioned above. As we make clear in this subsection,
the maturity level of the system that is used to make cession decisions has a clear impact on the
operational framework. This in turn should be reflected in the training and sequential validation
procedures used to perform model selection and subsequent performance assessment. A significant
aspect here is that, by comparing the performances obtained under different maturity level hypothe-
ses, one can easily derive a business case for an IT project that involves bringing the system to a
higher maturity level, simply by using the estimated difference in the risk-sharing pool profitability
as a measure of the expected benefits of the project.

Underwriters of insurance companies generate information that is fed to what is known as a
transactional database. This transactional database is modified in real time. Actuaries, often oper-
atin a separate department, are usually provided with a monthly update of a processed version
of this transactional database that is called the analytical database. Implementing cession models at
the transactional level so that cession recommendations can be given in real time within the under-
writing interface (the online implementation) is a much more involved project than simply having
the model interact with the analytical database and output recommendations on a monthly basis,
in batch mode (the batch implementation). The main difference, from a data mining perspective,
is the information that is available at the time when a decision must be made. The ideal situation
would be to have access to the most recent information, including all modifications up to the minute
Figure 17: **Optimistic** evaluation methodology, that assumes that a profile can be transferred and that up-to-date information is available when a decision is to be made. **Realistic** evaluation methodology, that considers a predicted proportional profit calculated at the time a transfer decision is made.

at which the decision is made.

Modifications to insurance contracts may come in several different ways, like the addition of a secondary driver, a change of address or the revision of specific clauses, and most of these modifications affect the premium that the insured is being charged for this coverage. The premium and the way it gets modified are therefore important features when it comes to predicting the potential savings of putting a contract in the risk-sharing pool and our models tend to quickly learn about that and then refine their predictions from the rest of the available data. Testing using different ways of viewing this premium and the associated profile or contract is therefore desirable both to reflect different implementation scenarios and to assess the robustness of our models.

When a contract gets renewed, we would like to make a decision using the updated information supplied by the customer at renewal time. On the other hand, the reality of interaction with insurance companies means that this information may not be available to the model at the moment it needs to make a decision, so we have also experimented with models which make a decision using the information from the last profile of the previous contract; in this case, we have also calculated the performance of the model using either the premium as estimated just before renewal or the premium actually paid from the beginning to the end of the contract.

We also experimented with different units to transfer to the pool: full contracts, individual profiles or even parts of profiles. Since a new profile is created every time a claim is filed, the full-contract approach allows to cede more than one claim at once, but it also represents the full premium instead of just part of it, so the potential loss is also greater. The per-profile approach allows finer-grain control, as if contracts were regularly taken out of the pool, but does not reflect reality since the decision to cede a contract cannot normally be made while the contract is active. The choice of the approach also changes the claim frequency, namely the proportion of the examples (contracts or profiles) for which one or more claims have been filed.

To illustrate the effect of different approaches, the following plots show the predicted profit, as a proportion of written premiums, made by similar models trained and tested with different views of the data. All models in this section share the same basic architecture: a Basis Selection Regressor with Bagging (where the number of bags $B = 5$ is used).

In Figure 17, the left-hand side plot is obtained using a scenario that assumed that any profile
can be ceded and that up-to-date information is always available when the decision is to be made. This is the most optimistic scenario (corresponding to an online implementation) and our test results confirm this. On the right-hand side of Figure 17, we consider a scenario according to which fresh information is not available and the decision to cede has to be made even before we know whether a contract will be renewed, and this decision has to stand for the full length of the contract (this corresponds, roughly, to a batch implementation). The plot shows the predicted proportional profit calculated from the approximate premium known upon making a decision. In Figure 18, profit is calculated from the actual premium paid for the full contract, including all modifications. It is interesting to note that the right-hand side of Figure 17 and Figure 18 are quite similar: it shows that our model is robust to modifications to the premium made while the contract is active. Also, the scenario illustrated on the left-hand side of Figure 17 was deemed more optimistic than reality while the scenarios of the latter two plots are more pessimistic; we could therefore expect the most realistic of models to give a performance that falls between the per-profile and the full-contract models.

This exercise has been useful to give a measure of how model behaves in the face of operational risks: What if pool rules are suddenly changed? What if only partial information is available at the
time of making a decision? What if there is a delay between making and implementing a decision? The answers we obtained suggest that the models do degrade gracefully in the face of difficulties; in all cases, this is considerably more desirable than one that would exhibit catastrophic failure when operating under any other than optimal conditions.

7 Conclusion

In this paper, we compared the performance of different feature selection and modeling approaches through their performance on the task of selecting risks to be ceded to automobile insurance risk-sharing pools. At low cession percentages (ca. 2%), linear regression models perform best. But as we consider higher cession percentages (5-10%), the proposed BSR-MinRank approach outperforms all others. Furthermore, if one is allowed to select the cession percentage optimally (profit maximizing), then BSR-MinRank again provides the best results. We 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. Through bootstrap resampling of the test set, we were able to obtain a distribution of performance of the models, rather than a single profit curve. This allows to assess the robustness, against data set variations, of the results obtained. Finally, we described how software maturity level affects the information that can be used for cession decision purposes. This in turn affects profitability. By measuring differences in profitability between software maturity levels, one can easily derive sound business cases for IT projects targeted at bringing software to a higher maturity level.

As a prospect for future work, we wish to consider situations where an insurer is operating close to the transfer limit allowed by the pool. In some jurisdictions, pool regulations strictly control the moments where insureds can be ceded so that, in some cases, it may be advantageous to avoid transferring up to the limit, in order to reserve space for better risks that will be anticipated to come along in the future. Such a method could further enhance the profitability of the approach in strongly budget-constrained situations.

References


